bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pathway dynamics can delineate the sources of transcriptional noise in gene expression

Lucy Ham[a,✉], **Marcel Jackson**[b], **and Michael P.H. Stumpf**[a]

[a]School of BioSciences and School of Mathematics and Statistics, University of Melbourne, Parkville VIC 3010, Australia
[b]Department of Mathematics and Statistics, La Trobe University, Melbourne VIC 3086, Australia

Single-cell expression profiling is destructive, giving rise to only static snapshots of cellular states. This loss of temporal information presents significant challenges in inferring dynamics from population data. Here we provide a formal analysis of the extent to which dynamic variability from within individual systems ("intrinsic noise") is distinguishable from variability across the population ("extrinsic noise"). Our results mathematically formalise observations that it is impossible to identify these sources of variability from the transcript abundance distribution alone. Notably, we find that systems subject to population variation invariably inflate the apparent degree of burstiness of the underlying process. Such identifiability problems can be remedied by the dual-reporter method, which separates the total gene expression noise into intrinsic and extrinsic contributions. This noise decomposition, however, requires strictly independent and identical gene reporters integrated into the same cell, which can be difficult to implement experimentally in many systems. Here we demonstrate mathematically that, in some cases, the same noise decomposition can be achieved at the transcriptional level with non-identical and not-necessarily independent reporters. We use our result to show that generic reporters lying in the same biochemical pathways (e.g. mRNA and protein) can replace dual reporters, enabling the noise decomposition to be obtained from only a single gene. Stochastic simulations are used to support our theory, and show that our "pathway-reporter" method compares favourably to the dual-reporter method.

stochastic gene expression | parameter identifiability | noise decomposition
Correspondence: *lucy.ham@unimelb.edu.au*

Introduction

Noise is a fundamental aspect of every biochemical network, and has integral functional roles in many cellular processes (1). Accurate prediction and understanding of these processes is therefore only possible through understanding their stochasticity. Noise arising in gene expression has arguably attracted most of the attention so far. Generally speaking, gene expression noise is separable into two sources of variability, as pioneered by Swain et al. (2). *Intrinsic noise* is generated by the dynamics of the gene expression process itself. The process, however, is often influenced by other external factors, such as the availability of promoters and of RNA polymerase, the influence of long noncoding RNA as a transcriptional regulator (3), as well as differences in the cellular environment. Such sources of variability contribute *extrinsic noise*, and reflect the variation in gene expression and transcription activity across the cell population. As such, understanding extrinsic noise lies at the heart of understanding cell-population heterogeneity.

So far, elucidating the sources of gene expression noise from transcriptomic measurements alone has proven difficult (4, 5). The fundamental hindrance lies in the fact that single-cell RNA sequencing is destructive, so that datasets reflect samples from across a population, rather than samples taken repeatedly from the same cell. As temporal information is lost in such measurements (6), it may be impossible to distinguish temporal variability within individual cells (e.g. burstiness), from ensemble variability across the population (i.e. extrinsic noise). A number of numerical and experimental studies have suggested this confounding effect (7–9), showing that systems with intrinsic noise alone exhibit behaviour that is indistinguishable from systems with both extrinsic and intrinsic noise. This is examined more formally in (10), where it shown that the moment scaling behaviour and transcript distribution may be indistinguishable from situations with purely intrinsic noise. The limitations in inferring dynamics from population data are becoming increasingly evident, and a number of studies that seek to address some of these problems have emerged (11, 12) .

Here we provide a significantly more detailed analysis of the extent to which sources of variability are identifiable from population data. We demonstrate mathematically that it is in general impossible to identify the sources of variability, and consequently, the underlying transcription dynamics, from the observed transcript abundance distribution alone. Systems with intrinsic noise alone can always present identically to similar systems with extrinsic noise, and moreover, extrinsic noise is shown to invariably distort the apparent degree of burstiness of the underlying system. This has important ramifications for parameter inference, and highlights the requirement for additional information, beyond the observed copy number distribution, in order to constrain the space of possible dynamics that could give rise to the same distribution.

This seemingly intractable problem can at least partially be resolved with a brilliantly simple approach: the dual-reporter method (2). In this approach, noise can be separated into extrinsic and intrinsic components, by observing correlations between the expression of two independent, but identically distributed fluorescent reporter genes. Dual-reporter assays have been employed experimentally to study the noise contributed by both global and gene-specific effects (13–15). A particular challenge, however, is that dual reporters are rarely identically regulated (15, 16), and are not straightforward to set up in many systems. More recently, there have been some efforts in developing alternative methods for decom-

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

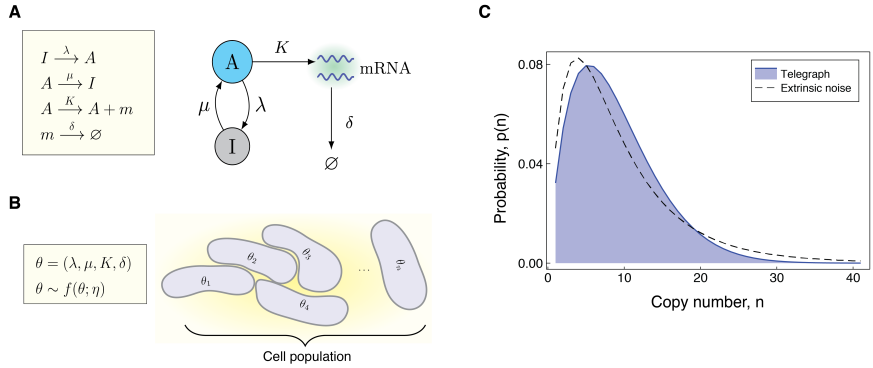

Fig. 1. Modelling the effects of both intrinsic and extrinsic noise. (**A**) A schematic of the Telegraph process, with nodes A (active) and I (inactive) representing the state of the gene. Transitions between the states A and I occur stochastically at rates μ and λ, respectively. The parameter K is the mRNA transcription rate, and δ is the degradation rate. (**B**) The compound model incorporates extrinsic noise by assuming that parameters θ of the Telegraph model vary across an ensemble of cells, according to some probability distribution $f(\theta;\eta)$. (**C**) Variation in the parameters across the cell population leads to greater variability in the mRNA copy number distribution.

posing noise (16, 17). Here we develop a direct generalisation of the original contribution of (2), that enables an analogous noise decomposition to be obtained from non-identical and not-necessarily independent reporters. Our result shows that measurements taken from the same biochemical pathway (e.g. mRNA and protein) can replace dual reporters, enabling the noise decomposition to be obtained from a single gene. This completely circumvents the requirement of strictly independent and identically regulated reporter genes. The results obtained from our "pathway-reporter" method are also borne out by stochastic simulations, and compare favourably with the dual-reporter method. In the case of constitutive expression, the results obtained from our decomposition are identical to those obtained via dual reporters. For bursty systems, we show that our approach provides a satisfactorily close approximation, except in extreme cases.

Modelling Intrinsic and Extrinsic Noise

A simple model of stochastic mRNA dynamics is the *Telegraph model*: a two-state model describing promoter switching, transcription, and mRNA degradation. In this model, all parameters are fixed, and gene expression variability arises due to the inherent stochasticity of the transcription process. As discussed above, this process will often be influenced by extrinsic sources of variability, and so modifications to account for this additional source of variability are required.

The Telegraph Model. The Telegraph model was first introduced in (18), and since then has been widely employed in the literature to model bursty gene expression in eukaryotic cells (19–22). In this model, the gene switches probabilistically between an active state and an inactive state, at rates λ (on-rate) and μ (off-rate), respectively. In the active state, mRNA are synthesized according to a Poisson process with rate K, while in the inactive state, transcription does not occur, or possibly occurs at some lower Poisson rate $K_0 \ll K$. Degradation of RNA molecules occurs independently of the gene promotor state at rate δ. A schematic of the Telegraph model is given in Fig. 1A. Throughout the paper, we will assume that all parameters of the Telegraph model are normalised by the mRNA degradation rate so that $\delta = 1$. The steady-state distribution of mRNA copy number can be explicitly described as (23):

$$\tilde{p}_T(n;\theta) = \frac{K^n \lambda^{(n)}}{n!(\mu+\lambda)^{(n)}} {}_1F_1(\lambda+n, \lambda+\mu+n, -K). \quad \textbf{(1)}$$

Here θ denotes the parameter vector $(\mu, \lambda, K, \delta)$, the function ${}_1F_1$ is the confluent hypergeometric function (24), and, for real number x and positive integer n, the notation $x^{(n)}$ abbreviates the rising factorial of x (also known as the Pochhammer function). Throughout, we will refer to the probability mass function $\tilde{p}_T(n;\theta)$ as the *Telegraph distribution with parameters* θ.

One limiting case of the Telegraph model is *constitutive* gene expression, which arises when the off-rate μ is 0, so that the gene remains permanently in the active state. In this case, the Telegraph distribution simplifies to a Poisson distribution with rate equal to K; that is, $\mathrm{Pois}(K)$.

At the opposite extreme is *instantaneously bursty* gene expression in which mRNA are produced in very short bursts with periods of inactivity in between. This mode of gene expression has been frequently reported experimentally, particularly in mammalian genes (15, 19, 21, 22). Transcriptional bursting may be treated as a limit of the Telegraph model, where the off-rate μ has tended toward infinity, while the on-rate λ remains fixed. In this limit, it can be shown (7, 10) that the Telegraph distribution converges to the negative binomial distribution $\mathrm{NegBin}(\lambda, \frac{K}{\mu+K})$.

The Compound Distribution. We account for random cell-to-cell variation across a population by way of a compound distribution (25)

$$\tilde{q}(n;\eta) = \int \tilde{p}(n;\theta) f(\theta;\eta)\, d\theta, \quad \textbf{(2)}$$

where $\tilde{p}(n;\theta)$ is the stationary probability distribution of a system with fixed parameters θ and $f(\theta;\eta)$ denotes the multivariate distribution for θ with hyperparameters η. Often we will take $\tilde{p}(n;\theta)$ to be the stationary probability distribution of the Telegraph model (Eq. (1)), and refer to Eq. (2) as the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Copy number distribution $\tilde{q}(n;\eta)$	Underlying distribution $\tilde{p}(n;\theta)$	Noise distribution $f(\theta;\eta)$
$\mathrm{Tele}(\lambda,\mu',K')$	$\mathrm{Tele}(\lambda,\mu,K)$	$K\sim\mathrm{Beta}_{K'}(\lambda+\mu,\mu'-\mu)$
$\mathrm{Tele}(\lambda,\mu',K')$	$\mathrm{Pois}(K)$	$K\sim\mathrm{Beta}_{K'}(\lambda,\mu')$
$\mathrm{NegBin}(\lambda,\frac{\beta}{\beta+1})$	$\mathrm{Tele}(\lambda,\mu,K)$	$K\sim\mathrm{Gamma}(\lambda+\mu,\beta)$
$\mathrm{NegBin}(\lambda,\frac{\beta}{\beta+1})$	$\mathrm{Pois}(K)$	$K\sim\mathrm{Gamma}(\lambda,\beta)$
$\mathrm{NegBin}(\lambda',\frac{c}{c+1})$	$\mathrm{NegBin}(\lambda,\frac{\alpha}{\alpha+1})$	$\alpha\sim\mathrm{BetaPrime}_c(\lambda-\lambda',\lambda')$

Table 1. Summary of the non-identifiability results. The results given in lines 1, 3 and 5 are our contributions, while the remaining representations (lines 2 and 4) are known and can be obtained as special cases of our results. Note that here we use $\mathrm{Tele}(\lambda,\mu,K)$ to denote a Telegraph distribution with parameters λ,μ,K. In lines 3 and 4, the parameter $\beta>0$ can be chosen freely and determines the mean burst intensity in the resulting compound system. In line 5 the parameters $c,\alpha>0$ are again mean burst intensities, and c can be chosen freely in the determination of the distribution of α.

compound Telegraph distribution. Sometimes $\tilde{p}(n;\theta)$ will be the Poisson distribution or the negative binomial distribution, depending on the underlying mode of gene activity.

The compound distribution is valid in the case of static environmental heterogeneity, that is, static parameter values for individual cells, but which vary across an ensemble of cells according to the distribution $f(\theta;\eta)$. This model is also a valid approximation for individual cells with dynamic parameters, provided these change sufficiently slower than the transcriptional dynamics. Fig. 1B gives a pictorial representation of the compound distribution.

In general, the compound Telegraph distribution $\tilde{q}(n;\eta)$ will be more dispersed than a Telegraph distribution to account for the uncertainty in the parameters; see Fig. 1C. Such dispersion is widely observed experimentally, and as demonstrated in (10) often reflects the presence of extrinsic noise.

Identifiability Considerations

Decoupling the effects of extrinsic noise from experimental measurements has been notoriously challenging. In the context of Eq. (2), the distribution $f(\theta;\eta)$ reflects the population heterogeneity, but experimental data provides samples only of $\tilde{q}(n;\eta)$. How much can we deduce of the underlying dynamics (that is, $\tilde{p}(n;\theta)$), and the population heterogeneity ($f(\theta;\eta)$), from measurements of transcripts from across the cell population ($\tilde{q}(n;\theta)$)?

Of course even though we may be able to infer the parameter η from experimental data, the expression $\tilde{p}(n;\theta)$ is really a family of distributions, parameterised by θ. This presents two challenges. The first is the possibility that there are different families of distributions $\tilde{p}(n;\theta)$ that can yield the same compound distribution $\tilde{q}(n;\eta)$ by way of different noise distributions, $f(\theta;\eta)$. The second challenge is that, even for a fixed family of distributions $\tilde{p}(n;\theta)$ it may be possible that different choices of the noise distribution $f(\theta;\eta)$ could still yield the same compound distribution $\tilde{q}(n;\eta)$. In these situations, we cannot hope to infer a unique solution for the noise distribution. This *identifiability problem* has important ramifications for the interpretability of parameter estimates obtained from experimental data. Indeed, if two or more model parameterisations are observationally equivalent (in this case, in the form of the transcript abundance distribution $\tilde{q}(n;\eta)$), then not only does this cast doubt upon the suitability of the model to represent (and subsequently predict) the system, but also obstructs our ability to infer mechanistic insight from experimental data.

An example of the first identifiability problem arises from a well-known example of a compound distribution. Referring to Eq. (2), when $f(\theta;\eta)$ is the gamma distribution and $\tilde{p}(n;\theta)$ is a Poisson distribution, corresponding to constitutive gene expression, the resulting compound distribution $\tilde{q}(n;\eta)$ is a negative binomial distribution (26). But this is the same distribution as that arising from instantaneously bursty gene expression (10). Such identifiability instances may be circumvented if there is confidence in the basic mode of gene activity i.e., if there is reason to believe that a gene is not constitutive, for example. We find, however, that there are numerous instances that can present insurmountable identifiability problems.

Bursty Gene Expression. We first observe that any Telegraph distribution with fixed parameters can identically be obtained from a Telegraph distribution with parameter variation. As shown in the SI Appendix, Section 1.A, any Telegraph distribution $\tilde{p}_T(n;\lambda,\mu',K')$ can be written as:

$$\tilde{p}_T(n;\lambda,\mu',K')=\int_0^{K'}\tilde{p}_T(n;\lambda,\mu,t)f_{K'}(t;\lambda+\mu,\mu'-\mu)\,dt, \tag{3}$$

where $\mu<\mu'$ and $f_{K'}(t;\lambda+\mu,\mu'-\mu)$ is the probability density function for a scaled beta distribution $\mathrm{Beta}_K(\lambda+\mu,\mu'-\mu)$ with support $[0,K']$. Thus, any Telegraph distribution can be obtained by varying the transcription rate parameter on a narrower Telegraph distribution (i.e., with a smaller off-rate) according to a scaled beta distribution. Fig. 2A (top panel) compares the representation obtained in Eq. (3) with the corresponding fixed-parameter Telegraph distribution for two different sets of parameters.

When $\mu=0$ the representation given in Eq. (3) simplifies to the well-known Poisson representation of the Telegraph distribution in terms of the scaled beta distribution (27).

Instantaneously Bursty Gene Expression. The previous result extends to instantaneously bursty systems. The copy number distribution of an instantaneously bursty system can be obtained from both bursty and instantaneously bursty systems, provided that there is appropriate parameter variation. The SI Appendix, Section 1.B, contains the relevant derivations. In the following, we let $\tilde{p}_{\mathrm{NB}}(n;r,\beta)$ denote the probability mass function of a $\mathrm{NegBin}(r,\frac{\beta}{\beta+1})$ distribution, where $\beta>0$. Then for any negative binomial distribution of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

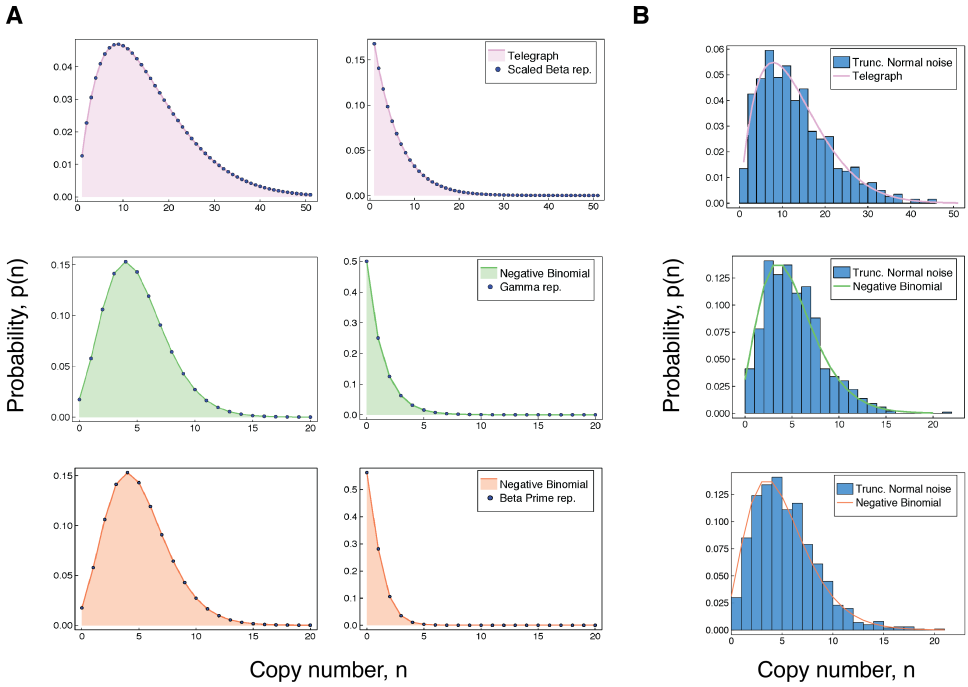

Fig. 2. (**A**) Accuracy of our representations for the Telegraph and negative binomial distribution. For each of the results in Eq. (3)-(5), we compare the (fixed-parameter) Telegraph and negative binomial distributions with their respective compound representations for two different sets of parameter values. The top panel (shown in pink) is shows comparisons for Eq. (3). Referring to Table 1, the parameter values for the top panel are (left) $\lambda = 2$, $\mu' = 12$, $K' = 100$, $\mu = 3$, and $K \sim \mathrm{Beta}_{K'}(5, 9)$, and (right) $\lambda = 1$, $\mu' = 20$, $K' = 100$, $\mu = 2$ and $K \sim \mathrm{Beta}_{K'}(3, 18)$. The middle panel (green) gives comparisons for Eq. (4), with parameter values (left) $\lambda = 10$, $\beta = 2$, $\mu = 2$ and $K \sim \mathrm{Gamma}(12, 2)$ and (right) $\lambda = 1$, $\beta = 1$, $\mu = 2$ and $K \sim \mathrm{Gamma}(3, 1)$. The bottom panel (coral) gives comparisons for Eq. (5). The parameter values (left) are $\lambda' = 10$, $\lambda = 15$ and $c = 2$ and (right) are $\lambda' = 2$, $\lambda = 5$ and $c = 3$. (**B**) The top figure compares a $\mathrm{Telegraph}(2, 4, 60)$ distribution with samples from a compound Telegraph distribution with normal noise $\mathrm{Norm}(37, 10)$ on the transcription rate parameter. The middle figure compares a $\mathrm{NegBin}(5, 0.5)$ with samples from a compound Telegraph distribution with normal noise $\mathrm{Norm}(5.5, 2.3)$ on the transcription rate parameter. The bottom figure compares a $\mathrm{NegBin}(5, 1)$ distribution with samples from a compound negative binomial distribution with normal noise $\mathrm{Norm}(2.3, 0.6)$ on the burst intensity parameter.

form $\mathrm{NegBin}(\lambda, \frac{\beta}{\beta+1})$ we have,

$$\tilde{p}_{\mathrm{NB}}(n; \lambda, \beta) = \int_0^\infty \tilde{p}_T(n; \lambda, \mu, t) f(t; \lambda + \mu, \beta)\, dt, \qquad \textbf{(4)}$$

where $f(t; \lambda + \mu, \beta)$ is the probability density function of a $\mathrm{Gamma}(\lambda + \mu, \beta)$ distribution. This generalises the aforementioned well-known Poisson representation of the negative binomial distribution (26), which corresponds to the particular case of $\mu = 0$. In Fig. 2A (middle panel), we compare the representation obtained in Eq. (4) with the corresponding fixed-parameter negative binomial distribution for two different sets of parameters.

We also have the following representation for a negative binomial distribution in terms of a scaled beta prime distribution,

$$\tilde{p}_{\mathrm{NB}}(n; \lambda', c) = \int_{\frac{1}{c}}^\infty \tilde{p}_{\mathrm{NB}}(n; \lambda, \theta) f_c(\theta c - 1; \lambda - \lambda', \lambda')\, d\theta \qquad \textbf{(5)}$$

where $f_c(c\theta - 1; \lambda - \lambda', \lambda')$ is the probability mass function of a scaled beta prime $\mathrm{BetaPrime}_c(\lambda - \lambda', \lambda)$ distribution, where $c > 0$ and $\lambda > \lambda'$. This equivalently corresponds to scaled beta noise $\mathrm{Beta}_c(\lambda - \lambda', \lambda')$ on the inverse of the expected burst intensity. Thus, the distribution of any instantaneously bursty system with mean burst intensity c can be obtained from one with greater burst frequency, by varying the mean burst intensity θ according to a shifted beta prime distribution. Fig. 2A (bottom panel), compares the representation obtained in Eq. (5) with the associated fixed-parameter negative binomial distribution for two different sets of parameters.

An exception: constitutive expression. It has long been known (28) that a compound Poisson distribution uniquely determines the compounding distribution. In the context of Eq. (2), this means the full extrinsic noise distribution $f(\theta, \eta)$ is identifiable from $\tilde{q}(n; \eta)$.

Implications for Parameter Inference. Estimates of kinetic parameters from experimental data suggest that gene expression is often either bursty or instantaneously bursty (i.e., $\mu \gg \lambda$). In turn, the assumption that gene-inactivation events occur far more frequently than gene-activation events is often used to derive other models of stochastic mRNA dynamics (29–31). The representations given in Eq. (3)–(5), however, show that both estimating parameters and the underlying dynamics from the form of the copy number distribution alone can be misleading. Noise on the transcription rate will invariably produce copy number data that is suggestive of a more bursty model. To illustrate this, consider an example in which the underlying process is a (mildly) bursty Telegraph system with distribution $\tilde{p}_T(n; 2, 3, K)$. Now assume that noise on the transcription rate parameter K follows the scaled Beta distribution on the interval $[0, 100]$ with

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$\alpha = \lambda + \mu = 2 + 3 = 5$ and $\beta = \mu' - \mu = 12 - 3 = 9$. The shape of this noise distribution closely resembles a slightly skewed bell curve, with the majority of transcription rates between around 10 and 60. This noise on the transcription rate K within the Telegraph system $\tilde{p}_T(n; 2, 3, K)$ will present identically to the significantly burstier system $\tilde{p}_T(n; 2, 12, 100)$. We remark that while the non-identifiability results (summarised in Table 1) are dependent on specific noise distributions, for practical purposes any similar distribution will produce a similar effect. To demonstrate this, we replaced the various noise distributions required for the representations in Eq. (3)–(5), with suitable normal distributions truncated at 0. In each case, we sampled 1000 data points from the corresponding compound distribution, and compared this with the associated fixed-parameter copy number distribution. The results are shown in Fig. 2B. The truncated normal distribution is not chosen on the basis of biological relevance, but rather to demonstrate that even a symmetric noise distribution (except for truncation at 0) produces qualitatively similar results to the distributions used in the precise non-identifiability results. In every case, the effect of a unimodal noise distribution on the transcription rate or burst intensity parameter is to produce a copy number distributions that are generally consistent with systems that appear burstier.

Resolving non-identifiability

The results of the previous section show that additional information, beyond the observed copy number distribution, is required to constrain the space of possible dynamics that could give rise to the same distribution. One way to narrow this space of possibilities, is to determine the intrinsic and extrinsic contributions to the total variation in the system.

The Dual-Reporter Method. The total gene expression noise, as measured by the squared coefficient of variation η^2, can be decomposed exactly into a sum of intrinsic and extrinsic noise contributions (2). The decomposition applies to dynamic noise (32), and generalises to higher moments in (33). As shown in (32), the noise decomposition is equivalent to the normalised Law of Total Variance (34). Indeed, if X is the random variable denoting the number of molecules of a certain species (eg. mRNA or protein) in a given cell, then we can decompose the total noise by conditioning X on the state $\mathbf{Z} = (Z_1, \ldots, Z_n)$ of the environmental variables $Z_1, \ldots, Z_n$:

$$\eta_X^2 = \frac{\mathrm{E}(\mathrm{Var}(X;\mathbf{Z}))}{\mathrm{E}(X)\,\mathrm{E}(Y)} + \frac{\mathrm{Var}(\mathrm{E}(X;\mathbf{Z}))}{\mathrm{E}(X)\,\mathrm{E}(Y)} \equiv \eta_{\text{int}}^2 + \eta_{\text{ext}}^2. \quad \textbf{(6)}$$

It has been shown (2, 32) that if X_1 and X_2 are random variables denoting the expression levels of independent and identically distributed gene reporters, then the extrinsic noise contribution η_{ext}^2 in Eq. (6) can be identified by the normalised covariance between X_1 and X_2:

$$\eta_{\text{ext}}^2 = \frac{\mathrm{Cov}(X_1, X_2)}{\mathrm{E}(X_1)\,\mathrm{E}(X_2)} \quad \textbf{(7)}$$

Decomposing Noise with Non-Identical Reporters. The dual-reporter method requires distinguishable measurements of transcripts or proteins from two independent and identically distributed reporter genes integrated into the same cell. In practice however, dual reporters rarely have identical dynamics, which is widely considered to be a significant challenge to interpreting experimental results (16). We show that, under certain conditions, the decomposition in Eq. (6) can alternatively be obtained from non-identically distributed and not-necessarily independent reporters.

Our result relies on the observation that the covariance of any two variables can be decomposed into the expectation of a conditional covariance and the covariance of two conditional expectations (the Law of Total Covariance (34)). If X and Y denote, for example, the numbers of molecules of two chemical species (eg. mRNA and protein) in a given cell, then the covariance of X and Y can be decomposed by conditioning on the state $\mathbf{Z} = (Z_1, \ldots, Z_n)$ of the environmental variables $Z_1, \ldots, Z_n$:

$$\mathrm{Cov}(X,Y) = \underbrace{\mathrm{E}(\mathrm{Cov}(X,Y;\mathbf{Z}))}_{\text{intrinsic}} + \underbrace{\mathrm{Cov}(\mathrm{E}(X;\mathbf{Z}), \mathrm{E}(Y;\mathbf{Z}))}_{\text{extrinsic}}. \quad \textbf{(8)}$$

We will see that, in many cases of interest, the random variable $\mathrm{E}(X;\mathbf{Z})$ (as a function of $\mathbf{Z}$) *splits across common variables* with $\mathrm{E}(Y;\mathbf{Z})$. By this we mean that $\mathrm{E}(X;\mathbf{Z}) = f(\mathbf{Z}_X)h_X(\mathbf{Z}')$ and $\mathrm{E}(Y;\mathbf{Z}) = g(\mathbf{Z}_Y)h_Y(\mathbf{Z}')$, where $\mathbf{Z}_X$ are the variables of $\mathbf{Z}$ that appear in $\mathrm{E}(X;\mathbf{Z})$ but not in $\mathrm{E}(Y;\mathbf{Z})$, and dually, $\mathbf{Z}_Y$ are those in $\mathrm{E}(Y;\mathbf{Z})$ that are not in $\mathrm{E}(X;\mathbf{Z})$. The variables $\mathbf{Z}'$ are those variables from $\mathbf{Z}$ not in $\mathbf{Z}_X \cup \mathbf{Z}_Y$. In these cases, the component of $\mathrm{Cov}(X,Y)$ that is contributed by the variation in $\mathbf{Z}$ may be written as the covariance of the functions $h_X(\mathbf{Z}')$ and $h_Y(\mathbf{Z}')$. Conveniently, in the cases of interest here, the two functions h_X and h_Y coincide, and this is the form we use in the following decomposition principle. The SI Appendix (Section 2) contains the proof of this result.

The Noise Decomposition Principle (NDP). *Assume that there are measurable functions f, g and h such that $\mathrm{E}(X;\mathbf{Z})$ and $\mathrm{E}(Y;\mathbf{Z})$ split across common variables by way of $\mathrm{E}(X;\mathbf{Z}) = f(\mathbf{Z}_X)h(\mathbf{Z}')$ and $\mathrm{E}(Y;\mathbf{Z}) = g(\mathbf{Z}_Y)h(\mathbf{Z}')$. Then provided that the variables $Z_1, \ldots, Z_m$ are mutually independent, the normalised covariance of $\mathrm{E}(X;\mathbf{Z})$ and $\mathrm{E}(Y;\mathbf{Z})$ will identify the total noise on $h(\mathbf{Z}')$ (i.e. $\eta^2_{h(\mathbf{Z}')}$).*

As we show in the next section, there are many situations where the random variable $\mathrm{E}(X;\mathbf{Z})$ is precisely the common part of $\mathrm{E}(Y;\mathbf{Z})$ and $\mathrm{E}(X;\mathbf{Z})$ (i.e., $h(\mathbf{Z}') = \mathrm{E}(X;\mathbf{Z})$), and the normalised intrinsic contribution to the covariance is either zero or is negligible. In these cases, the normalised covariance of X and Y will identify precisely the extrinsic noise contribution η_{ext}^2 to the total noise η_X^2. To see this, consider the situation where $\mathrm{E}(Y;\mathbf{Z}) = f(\mathbf{Z}_Y)\,\mathrm{E}(X;\mathbf{Z})$. Then provided $f(\mathbf{Z}_Y)$ and $\mathrm{E}(X;\mathbf{Z})$ are independent random variables, the extrinsic contribution to the covariance of X and Y

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

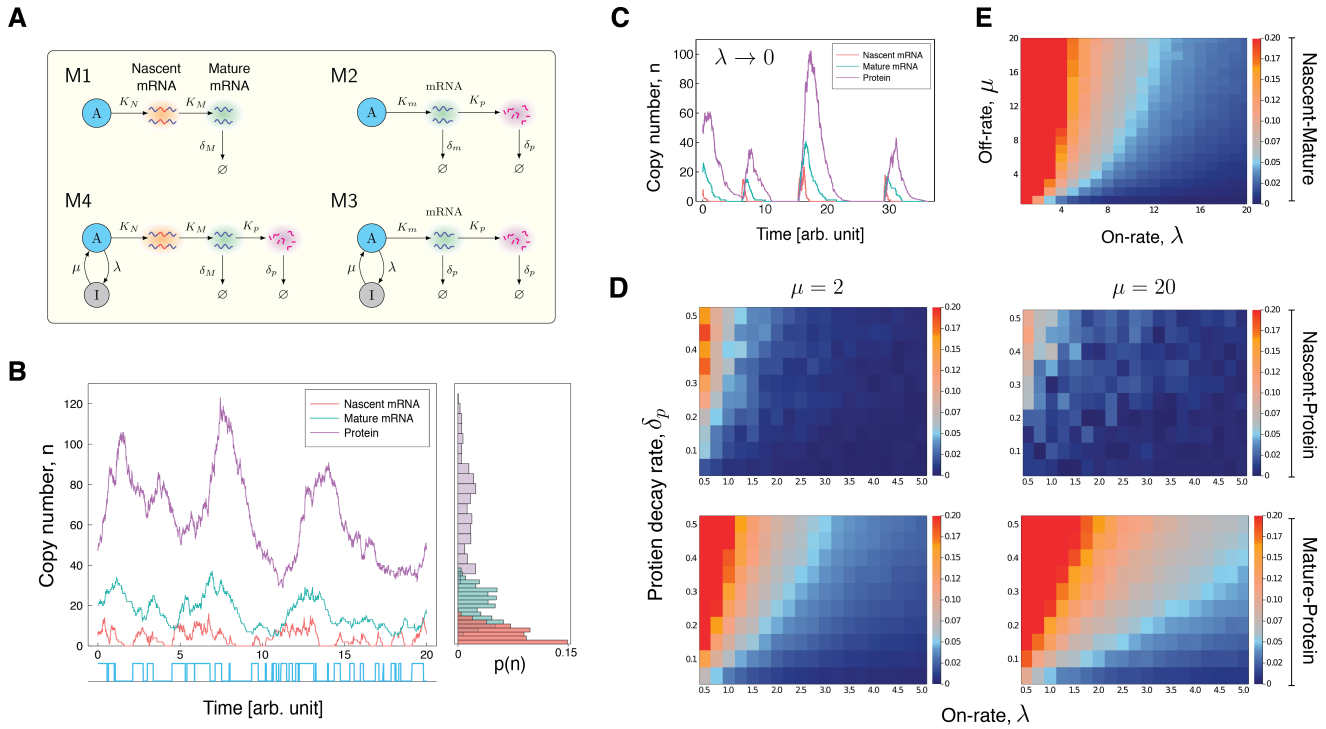

Fig. 3. (**A**) Stochastic models of gene expression. The model M1 is the simplest model of mRNA maturation. Here nascent (unspliced) mRNA are shown in red/blue wavy lines; the blue segments represent introns and the red segments represent the exons. Nascent mRNA are synthesized at the rate K_N, and spliced at rate K_M. Degradation of the mature mRNA occurs at rate δ_M. The model M2 is the well-known two-stage model of gene expression. In this model, mRNA (shown in blue) are produced at the rate K_m, which are later transcribed into protein (shown in pink) at rate K_p. The degradation of mRNA and protein occur at rates δ_m and δ_p, respectively. The model M3 is the extension of the two-stage model to include promoter switching. The nodes A (active) and I (inactive) represent the state of the gene, with transitions between states occurring at rates λ and μ. The remaining parameters are the same as those in the model M2. The model M4 extends the model M3 by incorporating mRNA maturation. Here K_N is the transcription rate parameter, and K_M is the maturation rate. All other parameters are the same as in M3. (**B**) Time series simulation of the copy number and activity state of a gene modelled by M4. For ease of visualisation, the parameters were artificially chosen as $\lambda = 2$, $\mu = 2.5$, $K_N = 40$, $K_M = 4$, $K_p = 4$ and $\delta_p = 1$, with all parameters scaled relative to $\delta_m = 1$. (**C**) As λ approaches 0, we see a higher correlation in the copy numbers of nascent mRNA, mature mRNA and protein. Again, the parameters are artificially chosen to be $\lambda = 0.1$, $\mu = 2.5$, $K_N = 80$, $K_M = 4$, $K_p = 4$ and $\delta_p = 1$, with all parameters scaled relative to $\delta_m = 1$. (**D**) Heatmaps for the intrinsic contribution to the covariance, which estimates the level of overshoot in the pathway-reporter approach for the nascent-protein and mature-protein reporters; blue regions show an overshoot of less than $\simeq 0.05$. Here the intrinsic contribution is calculated using stochastic simulations of the model M4. In the left panels, the parameter $\mu = 2$ is fixed, while δ_p and the on-rate λ are varied between 0.01 and 1 and 0.5 and 4, respectively. In the right panels, the parameter $\mu = 20$ is fixed, while δ_p and the on-rate λ vary between 0.01 and 1, and 1 and 10, respectively. In all cases, the parameters are scaled so that $\delta_M = 1$. The maturation rate is fixed at 20, with the parameters K_N and K_P are chosen to produce a mean protein level of 1000, a mean nascent mRNA level of 5 and a mean mature mRNA level of 50. (**E**) The nascent-mature reporter concerns only mRNA and so is independent of all protein-related parameters. The heatmap shows the intrinsic contribution for values of λ and μ between 0.1 and 20, with the same parameter selections for K_N, K_M as in (D). Similar simulations for average nascent mRNA levels of 3 and of 8, and mature mRNA levels of 30 and of 160 produced almost identical heatmaps.

is given by:

$$\begin{aligned}\mathrm{Cov}(\mathrm{E}(X;\mathbf{Z}),\mathrm{E}(Y;\mathbf{Z})) &= \mathrm{Cov}(\mathrm{E}(X;\mathbf{Z}), f(\mathbf{Z}_Y)\,\mathrm{E}(X;\mathbf{Z}))\\ &= \mathrm{E}(f(\mathbf{Z}_Y))\,\mathrm{Cov}(\mathrm{E}(X;\mathbf{Z}),\mathrm{E}(X;\mathbf{Z}_Y))\\ &= \mathrm{E}(f(\mathbf{Z}_Y))\,\mathrm{Var}(\mathrm{E}(X;\mathbf{Z})). \qquad \textbf{(9)}\end{aligned}$$

If the normalised intrinsic contribution to the covariance is either zero or is negligible, it then follows from Eq. (8) that

$$\frac{\mathrm{Cov}(X,Y)}{\mathrm{E}(X)\,\mathrm{E}(Y)} \approx \frac{\mathrm{Var}(\mathrm{E}(X;\mathbf{Z}))}{\mathrm{E}(X)^2} = \eta^2_{\text{ext}}. \qquad \textbf{(10)}$$

Thus, under certain conditions, measuring the covariance between two non-identically distributed and not-necessarily independent reporters can replace dual reporters.

The Pathway-Reporter Method

We show that for some reporters X and Y lying in the same biochemical pathway, the covariance of X and Y continues to identify the extrinsic, and subsequently intrinsic, noise contributions to the total noise. Throughout this section, we assume that extrinsic noise sources act independently i.e., the environmental variables $Z_1, \ldots, Z_n$ of $\mathbf{Z}$ are mutually independent. Additionally, our modelling is focused only on a single gene copy, though the same analysis applies to multiple but indistinguishable gene copies; see SI Appendix, Section 3.A.

Measuring Noise from a Constitutive Gene. We consider first the simplest case where the underlying process is constitutive. We begin with a stochastic model of mRNA maturation, which we denote by $\mathbf{M}_1$; Fig. 3A (top left) gives a schematic of the model. In this model, the gene continuously produces nascent mRNA according to a Poisson process at constant rate K_N, which are subsequently spliced into mature mRNA according to the rate K_M. Degradation of the mature mRNA occurs as a first-order Poisson process with rate δ_M. The model $\mathbf{M}_1$ and extensions have been considered

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in a number of recent studies (30, 35–37), including a recent preprint (12), where information about the marginal nascent and mature distribution is used to distinguish between intrinsic and extrinsic versions of the model.

The model $\mathbf{M}_1$ has a known solution for the stationary joint probability distribution (38), given by:

$$\tilde{p}(n,m;\theta) = \frac{e^{-\frac{K_n}{K_M}}(\frac{K_N}{K_M})^n}{n!}\frac{e^{-\frac{K_n}{\delta}}(\frac{K_N}{\delta_M})^m}{m!}. \quad \textbf{(11)}$$

Here n is the number of nascent mRNA, m is the number of mature mRNA and $\theta = (K_N, K_M, \delta_M)$.

We let X_N denote the number of nascent mRNA, let X_M denote the number of mature mRNA produced from the same constitutive gene and let $\mathbf{Z} = (K_N, K_M, \delta_M)$. To simplify notation, we will abbreviate the variables in $\mathbf{Z}_{X_N}$ as $\mathbf{Z}_N$, and similarly for $\mathbf{Z}_{X_M}$. It follows immediately from Eq. (11) that X_N and X_M are independent conditional on $\mathbf{Z}$ and so the intrinsic contribution to the covariance of X_N and X_M (the first term of Eq. (8)) is 0. It is also easy to see from Eq. (11) that $\mathrm{E}(X_N;\mathbf{Z}) = f(\mathbf{Z}_N)K_N$ and $\mathrm{E}(X_M;\mathbf{Z}) = g(\mathbf{Z}_M)K_N$, where $f(\mathbf{Z}_N) = \frac{1}{K_M}$ and $g(\mathbf{Z}_M) = \frac{1}{\delta_M}$. Since the extrinsic noise sources are assumed to act independently, it follows that the Noise Decomposition Principle (NDP) of the previous section holds. We then have that $\mathrm{Cov}(X_N, X_M) = \eta^2_{K_N}$, where $\eta^2_{K_N}$ is the total noise on the transcription rate parameter K_N. Thus, measuring $\mathrm{Cov}(X_N, X_M)$ can replace dual reporters to decompose the gene expression noise at the transcriptional level.

To support our mathematical findings, we simulated the model $\mathbf{M}_1$ subject to parameter variation using the stochastic simulation algorithm (SSA). Table 2 (left) compares the extrinsic noise contributions found from various simulations with the corresponding theoretical values. In each simulation, the degradation rate δ_m is fixed at 1, with the other parameters scaled accordingly. The maturation rate K_M is sampled according to a $\mathrm{Gamma}(8, 0.0125)$ distribution, which has coefficient of variation 0.125. We considered different noise distributions on K_N, producing a range of noise strengths. Our theory predicts that pathway-reporters will identify the total noise on K_N. Overall, we observe an excellent agreement between the results obtained by the pathway-reporter method, the dual-reporter method and the theoretical noise.

To explore the pathway-reporter method more widely, we considered 36 different parameter combinations to produce varying mean copy numbers consistent with those reported in experimental papers. We also considered different noise distributions taken from the scaled Beta distribution family in order to produce a range of noise strengths; see SI Table 1. The pathway-reporter method performed favourably to the dual-reporter method calculated from mature mRNA, and consistently outperforms the dual-reporter method on nascent mRNA. We note that both the pathway-reporter and dual-reporter methods are slow to converge for systems with very low nascent mean copy numbers (mean ≤ 1); see SI Table 2.

Next, we considered the simplest stochastic model of gene expression that includes both mRNA and protein dynamics: the "two-stage model" of gene expression. We denote this model by $\mathbf{M}_2$; see Fig. 3A (top right) for a schematic of this model. The two-stage model, and its three-stage extension to include promoter switching, have been considered in numerous studies (14, 15, 39–44). In this model, mRNA are synthesized according a Poisson process at rate K_m, which are then later translated into protein at rate K_p. Degradation of mRNA and protein occur as first-order Poisson processes with rates δ_m and δ_p, respectively. . Let X_m denote the number of mRNA and X_p denote the number of proteins produced from the same constitutive gene and let $\mathbf{Z} = (K_m, K_p, \delta_m, \delta_p)$. Then the stationary means and covariance are given by (39, 41):

$$\mathrm{E}(X_m;\mathbf{Z}) = \frac{K_m}{\delta_m}, \quad \mathrm{E}(X_p;\mathbf{Z}) = \frac{K_p}{\delta_p}\frac{K_m}{\delta_m} \quad \text{and}$$
$$\mathrm{Cov}(X_m, X_p;\mathbf{Z}) = \frac{K_m K_p}{\delta_m(\delta_m+\delta_p)}. \quad \textbf{(12)}$$

It is easily seen that $\mathrm{E}(X_p;\mathbf{Z}) = f(\mathbf{Z}_p)\mathrm{E}(X_m;\mathbf{Z})$, where $f(\mathbf{Z}_p) = \frac{K_p}{\delta_p}$. Thus, it follows from the NDP, the normalised contribution of $\mathrm{Cov}(X_m, X_p)$ contributed by $\mathbf{Z}$ will identify the extrinsic noise contribution to the total noise on X_m. Now, if we assume that δ_m is fixed across the cell-population, and all parameters are scaled so that $\delta_m = 1$, we have the following expression for the intrinsic contribution to the covariance of X_m and X_p (see SI Appendix, Section 3.A),

$$\frac{\mathrm{E}(\mathrm{Cov}(X_m, X_p;\mathbf{Z}))}{\mathrm{E}(X_m)\mathrm{E}(X_p)} = \frac{\alpha}{\mathrm{E}(K_m)}, \quad \text{where } \alpha = \frac{\mathrm{E}(1/(\delta_p+1))}{\mathrm{E}(1/\delta_p)}. \quad \textbf{(13)}$$

Since mRNA tends to be less stable than protein, we have that $\delta_p < 1$, and often $\delta_p \ll 1$ (45, 46). So, we can expect $\alpha \ll 1$. Further, for many genes we can expect the number of mRNA per cell (K_m) to be in the order of tens, so $1/E(K_m) < 1$. It follows that $\mathrm{E}(\mathrm{Cov}(X_m, X_p;\mathbf{Z})) \ll 1$, so that $\mathrm{Cov}(X_m, X_p)$ will closely approximate the extrinsic noise at the transcriptional level.

We tested our theory using stochastic simulations of the model $\mathbf{M}_2$ subject to parameter variation. Table 2 (right) gives a comparison of the results of the mRNA-protein reporters and dual reporters. In each case, we varied K_p according to a $\mathrm{Gamma}(5, 0.4)$ distribution and δ_p according to a $\mathrm{Gamma}(8, 0.125)$ distribution; the corresponding noise strengths are 0.20 and 0.125, respectively. We considered different noise distributions on K_m, which produce a range of noise strengths, and the noise distribution parameters were selected to produce a mean mRNA of approximately 50 and a mean number of approximately 1000 proteins in each simulation. As our theory predicts, the mRNA-protein reporters identify the extrinsic noise contribution to the total noise on X_m. Again, we can see from Table 2 (right) that there is excellent agreement between the results of the pathway reporters and the dual reporters. A larger exploration of the parameter space reveals similar results; these results are provided in SI Table 3. Thus, despite mRNA and protein numbers not being strictly independent, they can, for practical

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Theory		Simulation	
η^2_{ext}	Noise (K_N)	PR	DR (Mat)
0.00	$K_N = 50$	0.00 ± 0.01	0.00 ± 0.00
0.10	$\text{Beta}_{133.\dot{3}}(6, 10)$	0.10 ± 0.01	0.10 ± 0.01
0.20	$\text{Gamma}(5, 10)$	0.20 ± 0.02	0.20 ± 0.01
0.50	$\text{Beta}_{300}(1.5, 7.5)$	0.49 ± 0.04	0.50 ± 0.03

Theory		Simulation	
η^2_{ext}	Noise (K_m)	PR	DR (Mat)
0.00	$K_m = 50$	0.00 ± 0.01	0.00 ± 0.00
0.10	$\text{Beta}_{133.\dot{3}}(6, 10)$	0.10 ± 0.01	0.10 ± 0.01
0.20	$\text{Gamma}(5, 10)$	0.20 ± 0.02	0.20 ± 0.01
0.50	$\text{Beta}_{300}(1.5, 7.5)$	0.51 ± 0.04	0.50 ± 0.03

Table 2. A comparison of the pathway-reporter method and the dual-reporter method for constitutive expression. Here PR gives the results of the nascent and mature pathway reporters, while DR (Mat) gives the results of dual reporters calculated from the mature mRNA. We considered noise on both the transcription rate (K_N) and the maturation rate (K_M). The decay rate is fixed at one, with the other parameters scaled accordingly. In each case, the maturation rate K_M is varied according to a $\text{Gamma}(8, 1.25)$ distribution, which has coefficient of variation 0.125. The values given are the average of 100 simulations, each calculated from 500 copy number samples, and the errors are $\pm$ one standard deviation. Our theory predicts that pathway-reporters will identify the noise on the nascent transcription rate K_N (η^2_{ext}). The noise distribution parameters were chosen to produce an average nascent mRNA copy number of approximately 5 and an average mature mRNA copy number of approximately 50.

purposes, replace dual reporters to decompose the noise at the transcriptional level.

Measuring Noise from a Facultative (bursty) Gene. The most common mode of gene expression that is observed experimentally is burstiness (7, 15, 19, 21, 22, 47), in which mRNA are produced in bursts with periods of inactivity in between. One example is gene regulation via repression, which naturally leads to periods of gene inactivity. Here we consider a four-stage model of bursty gene expression, which incorporates both promoter switching and mRNA maturation; we denote this model by $\mathbf{M}_4$; see Fig. 3A (bottom left). This model has recently been considered in (30), where the marginal distributions are solved in some limiting cases. In this model, the gene switches probabilistically between an active state (A) and an inactive state (I), at rates λ (on-rate) and μ (off-rate), respectively. In the active state, nascent mRNA are synthesized according to a Poisson process at rate K_N, while in the inactive state, transcription does not occur. The nascent mRNA are spliced into mature mRNA at rate K_M, which are later translated into protein, according to the rate K_P. Degradation of mRNA and protein occurs independently of the promotor state according to the rates δ_M and δ_P, respectively.

Three natural candidates for pathway reporters arise from this model: (a) nascent and mature mRNA (b) mature mRNA and protein, and (c) nascent mRNA and protein reporters. We will find that nascent-protein reporters yield consistently acceptable estimates of the extrinsic noise contribution η^2_{ext}, while nascent-mature and mature-protein reporters are reliable in some restricted cases. We begin by showing that each of the reporter pairs (a), (b) and (c) satisfy the Noise Decomposition Principle. We then demonstrate, computationally, that despite a lack of independence between these reporter pairs, the pathway-reporter method can still be used to decompose the total gene expression noise at the transcriptional level. Throughout, we let X_N denote the number of nascent mRNA, we let X_M denote the number of mature mRNA, and let X_P denote the number of proteins produced from the same gene. We also let $\mathbf{Z} = \{\lambda, \mu, K_N, K_M, K_P, \delta_M, \delta_P\}$.

As in (15), we now assume that the transcription K_N is large relative to the other parameters. We further assume that the maturation rate K_M is fast (i.e. $K_M > \delta_M$), which is supported experimentally (30). Then using the results of (15) and arguments similar to those in (30), it can be shown that the stationary means for the nascent mRNA, mature mRNA and protein levels are given by,

$$\text{E}(X_N;\mathbf{Z}) = \frac{K_N}{K_M}\frac{\lambda}{(\lambda+\mu)},\ \text{E}(X_M;\mathbf{Z}) = \frac{K_N}{\delta_M}\frac{\lambda}{(\lambda+\mu)} \text{ and } \text{E}(X_P;\mathbf{Z}) = \frac{K_P}{\delta_P}\frac{K_N}{\delta_M}\frac{\lambda}{(\lambda+\mu)} \quad \textbf{(14)}$$

respectively. The means for the mature mRNA and protein levels are given in (15) for the well-known "three-stage model" of gene expression; see Fig. 3A (bottom right). The SI Appendix provides a more detailed justification of these expressions. We consider first the nascent-mature pathway reporters, case (a). From Eq. (14), it is easily seen that $\text{E}(X_N;\mathbf{Z}) = f(\mathbf{Z}_N)K_N\frac{\lambda}{(\lambda+\mu)}$ and $\text{E}(X_M;\mathbf{Z}) = g(\mathbf{Z}_M)K_N\frac{\lambda}{(\lambda+\mu)}$, where $f(\mathbf{Z}_N) = \frac{1}{K_M}$ and $g(\mathbf{Z}_M) = \frac{1}{\delta_M}$. So the NDP holds, and the normalised covariance of $\text{E}(X_N;\mathbf{Z})$ and $\text{E}(X_M;\mathbf{Z})$ will identify the extrinsic noise on the transcription component $K_N\frac{\lambda}{(\lambda+\mu)}$. Consider now the mature-protein reporters, case (b). Again, we can see from Eq. (14) that $\text{E}(X_M;\mathbf{Z}) = f(\mathbf{Z}_P)\text{E}(X;\mathbf{Z})$, where $f(\mathbf{Z}_P) = \frac{K_P}{\delta_P}$. Thus the NDP holds, and so the normalised covariance of $\text{E}(X_M;\mathbf{Z})$ and $\text{E}(X_P;\mathbf{Z})$ will identify the total noise on $\text{E}(X_M;\mathbf{Z})$ (the extrinsic noise on X_M). For the nascent-protein reporters, case (c), it is easy to see that $\text{E}(X_N;\mathbf{Z}) = f(\mathbf{Z}_N)K_N\frac{\lambda}{(\lambda+\mu)}$, where $f(\mathbf{Z}_N) = \frac{1}{K_M}$, and $\text{E}(X_P;\mathbf{Z}) = g(\mathbf{Z}_P)K_N\frac{\lambda}{(\lambda+\mu)}$, where $g(\mathbf{Z}_P) = \frac{K_P}{\delta_M\delta_P}$. Thus, again the NDP holds, and the normalised covariance of $\text{E}(X_N;\mathbf{Z})$ and $\text{E}(X_P;\mathbf{Z})$ will identify the noise on the transcriptional component $K_N\frac{\lambda}{(\lambda+\mu)}$.

In order for the pathway-reporter method to provide a close approximation to the extrinsic noise in cases (a), (b) and (c), we require that the normalised intrinsic contribution to the covariance is either zero or negligible. This condition will hold provided there is sufficiently small correlation between the reporter pairs. In the case of (prokaryotic) mRNA and protein, this lack of correlation has been been verified experimentally in *E. Coli* (48). More generally, it is possible to provide an intuition about the conditions under which the lack of correlation might hold. The time series of copy numbers for

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mean					Simulation		
λ	μ	K_N	K_P	δ_P	PR (MP)	PR (NP)	DR (Mat)
0.5	1	150	2	0.1	0.46 ± 0.06	0.38 ± 0.07	0.32 ± 0.07
1	2	150	2	0.1	0.39 ± 0.05	0.34 ± 0.07	0.32 ± 0.05
1	20	1050	2	0.1	0.66 ± 0.15	0.52 ± 0.22	0.50 ± 0.15
2	2	100	6	0.3	0.35 ± 0.04	0.29 ± 0.05	0.27 ± 0.03
2	20	550	6	0.3	0.61 ± 0.09	0.47 ± 0.15	0.47 ± 0.09
10	10	100	6	0.3	0.29 ± 0.03	0.27 ± 0.04	0.27 ± 0.02

Table 3. A comparison of the pathway-reporter method and dual-reporter method for bursty expression. Here PR (NP) gives the results of the nascent and protein pathway reporters, PR (MP) gives the results of the mRNA and protein reporters, while DR (Mat) gives the results of the dual reporters calculated from the mature mRNA. We consider noise on all of the parameters except for δ_M and K_M; see discussion in main text. The values given are the average of 100 simulations, each calculated from 500 copy number samples, and the errors are $\pm$ one standard deviation. Our theory predicts that pathway-reporters will identify the noise at both the promoter level (λ, μ) and transcriptional level (K_N, δ_m); the total extrinsic noise in each case is given by η^2_{ext}. As before, the noise distribution parameters were chosen to produce an average nascent mRNA copy number of 5 and an average mature mRNA copy number of 50, and an average number of 1000 proteins.

each of nascent mRNA, mature mRNA and protein broadly follow each other, each with delay from its predecessor (Fig. 3B). Parameter values that reduce this delay will tend to increase correlation, and thereby increase the normalised intrinsic contribution to the covariance. The primary example of this effect is seen when δ_p approaches, or even exceeds δ_m (or for nascent-mature reporters, when δ approaches the maturation rate). A further contributor to high correlation between mRNA and protein, is when the system undergoes long timescale changes. In this situation the copy numbers tend to drop to very low values for extended periods. The primary parameter influencing this type of behaviour is λ, specifically, when λ tends to 0 (Fig. 3C). An illustrative example of this can be seen by considering a Telegraph system in the limit of slow switching, which produces a copy number distribution that converges to a scaled Poisson-Bernoulli compound distribution: even without any extrinsic noise, the pathway reporter method will identify the η^2 value of the corresponding scaled Bernoulli distribution.

An extensive computational exploration of the parameter space (SI Table 4) supports our intuition, though the strength of the effect varies varies across the three different reporter pairs. For nascent-protein reporters, the normalised intrinsic contribution to the covariance is satisfactorily small, except in cases of fast protein decay coupled with low values of λ; changes in μ have a less significant effect. This is further corroborated by the heatmaps shown in Fig. 3D (top row), which for two fixed values of μ and a broad spectrum of δ_p and λ values, give the intrinsic term η_{int} in Eq. (8), for fixed $\mathbf{Z}$. These provide only an estimate for the overshoot error in the pathway-reporter approach. Blue pixels represent an overshoot estimate of less than 0.05. The heatmaps show that an estimate for the overshoot error in the pathway reporter approach is satisfactorily small (less than 0.05) for most reasonable parameter values: only failing for (a) high values of δ_p in unison with (b) low values of λ (less than 1, though lower values are acceptable if δ_p is small). Note that here the parameters have been scaled so that $\delta = 1$. Similar results are obtained for the intermediate case of $\mu = 10$ (SI Appendix, Fig. S1).

The assumption (a) that $\delta_P < \delta_M$ is known to be true for a large number of genes, and is justified by the difference in the mRNA and protein lifetimes. While there is of course variation across genes and organisms, values of $\delta_P \leq 0.5\delta_M$ and even $\delta_P \leq 0.2\delta_M$ seem reasonable for the majority of genes. In *E. Coli* (48) and yeast (49), for example, mRNA are typically degraded within a few minutes, while most proteins have lifetimes at the level of 10s of minutes to hours. For mammalian genes (46), it is reported that the median mRNA decay rate δ_M is (approximately) five times larger than the median protein decay rate δ_P, determined from $4,200$ genes. The assumption (b) requires that the gene is sufficiently active. In a recent paper by Larsson et al. (22), the promoter-switching rates λ, μ and the transcription rate K of the Telegraph model were estimated from single-cell data for over $7,000$ genes in mouse and human fibroblasts. Of those genes with mean mRNA levels greater than 5, we found that over 90% have a value for λ of at least 0.5, and over 65% have a value for λ greater than 1. In Cao and Grima (30), a comprehensive list of genes (ranging from yeast cells through to human cells) with experimentally-inferred parameter values are sourced from across the literature (see Table 1 in (30)). After scaling the parameter values of the 26 genes reported there, we find that around 88% have a value for λ of at least 0.5, and approximately 58% have a value for λ greater than 1. Thus, the heatmaps given in Fig. 3D (top panel) suggest that nascent-protein reporters will provide a satisfactory estimate of the extrinsic noise level for a substantial number of genes.

The mature-protein reporters are less reliable, with the requirement of slow protein decay and faster on-rate being more pronounced than for the nascent-protein reporters; this is evident from Fig. 3D (second panel). The performance of the nascent-mature reporter is of course independent of δ_p, but is only viable in the case of fast on-rate (Fig. 3E).

We tested our approach for each of the pathway reporter pairs (a), (b) and (c) against dual reporters using stochastic simulations. Table 3 shows the results from six simulations across a spectrum of behaviours from moderately slow switching, to fast switching as well as a range of levels of burstiness. For each of the parameters $\mu, \lambda, K_P, \delta_P$ we selected a scaled $\text{Beta}(5, 6)$ distribution, with squared coefficient of variation $\eta^2 = 0.1$; the scaling was chosen in each case to achieve a

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mean value equal to the parameter value given in Table 3. It is routinely verified that scaling these distributions does not change the value of η^2. The parameter K_N was given the slightly heavier noise distribution $\mathrm{Beta}(3,6)$, with $\eta^2 = 0.2$. In order to achieve direct benchmarking against the dual reporters, the parameter K_M was kept fixed at 10. This is because the nascent-protein pathway reporter estimates noise on the value of $K_N \frac{\lambda}{\lambda+\mu}$, while the mature mRNA dual-reporter measures noise on $\frac{K_N}{K_M}\frac{\lambda}{\lambda+\mu}$, and these coincide only when K_M is fixed. The mean values of K_N and K_P were chosen to achieve approximate average nascent mRNA levels, mature mRNA levels and protein levels at 5, 50 and 1000 respectively, given the chosen values of λ, μ, δ_P.

The results for the nascent-protein reporters given in Table 3 show comparable performance to dual reporters, with only modest overshoot; even in the worst performing case of $\lambda = 0.5$, $\mu = 1$ the result of the pathway reporters is within one standard deviation, in a very tight distribution. The error heatmaps of Fig. 3 provide a surprisingly accurate estimate of the overshoot in the nascent-protein results in Table 3. As an example, the first row is most closely matched by the heatmap at top left of Fig. 3D, which at $\lambda = 0.5$ and $\delta_P = 0.1$ is suggestive of an error around the boundary between blue and red (around 0.06). The same accuracy is obtained for the other rows. As predicted, the mature-protein reporters show significantly more overshoot, especially with the less active genes. Improved accuracy can again be obtained by subtracting the estimated overshoot given in the error heatmaps from the obtained value. Thus for example, the error heatmap for $\mu = 2$ (Fig. 3D lower left) gives an error approximately 0.07 for $\lambda = 1, \delta_P = 0.1$, which agrees very closely to the actual overshoot of 0.07 shown in the corresponding row of Table 3. An overshoot of approximately 0.06 is suggested by the $\mu = 2$ error heatmap when $\lambda = 2, \delta_P = 0.3$, which leads to a correction from 0.35 in Table 3 to a value of 0.29. This is quite consistent with the dual reporter benchmark of 0.27. As expected, based on Fig. 3E, nascent-mature reporters do not perform well on bursty systems except for high λ and so the values are not included in Table 3; only in the case of $\lambda = \mu = 10$ was the result approaching the dual reporter value, returning 0.32 ± 0.03.

Discussion

Despite the proliferation of experimental methods for single-cell profiling, the ability to extract transcriptional dynamics from measured distributions of mRNA copy numbers is limited. In particular, the multiple factors that contribute to mRNA heterogeneity can confound the measured distribution, which hinders analysis. Theoretical contributions that can concretise these observable effects are therefore of great importance. In this work, we have demonstrated, through a series of mathematical results, that it is impossible to delineate the relative sources of heterogeneity from the measured transcript abundance distribution alone: multiple possible dynamics can give rise to the same distribution. Our approach involves establishing integral representations for distributions that are commonly encountered in single-cell data analysis, such as the negative binomial distribution and the stationary probability distribution of the Telegraph model. We show that a number of well-known representations can be obtained from our results. A particular feature of our non-identifiability results is that population heterogeneity inflates the apparent burstiness of the system. It is therefore necessary to obtain further information, beyond measurements of the transcripts alone, in order to constrain the number of possible theoretical models of gene activity that could represent the system. In particular, additional work may be required to determine the true level of burstiness of the underlying system.

Further, we have developed a theoretical framework for extracting estimates of the level of extrinsic noise, which can assist in resolving the non-identifiability problems given in the first part of this work. The well-known dual reporter method of Swain et al. (2) already provides one such approach, however is experimentally challenging to set up in many systems, requiring strictly identical and independent pairs of gene reporters. We present a Noise Decomposition Principle, that directly generalises the theoretical underpinnings of the dual reporter method and use it to identify a practical approach—the pathway-reporter method—for obtaining the same noise decomposition. Our approach allows us to use measurements of two different species from the transcriptional pathway of a single gene copy in replacement of dual reporters. The accuracy of the pathway-reporter method is provably identical for constitutive gene expression, and in the case of nascent-mature mRNA reporters, the measurements are readily obtained from current single-cell data (11, 35, 50). For bursty systems, the method in general provides only an approximation of the extrinsic noise. We are, however, able to demonstrate computationally, that one of the proposed pathway reporters provides a satisfactory estimate of the extrinsic noise for most genes. The other pathway reporters also provide viable estimates of the extrinsic noise in some cases.

Despite the generality of our theoretical contribution, our pathway-reporter approach has some caveats. In particular, the approach relies on the assumption that extrinsic noise sources act independently. Experimentally, however, these may be correlated. For example, it has been suggested (51, 52) that the transcription and translation rates in E. coli anticorrelate. Additional work, however, is required to determine degree to which the independence of noise sources is a reasonable assumption.

A further potential limitation of the pathway-reporter approach in the case of bursty genes is that experimental implementation requires simultaneous measurements of nascent mRNA and protein from a single cell. Methods for obtaining such measurements typically have low throughput and can only profile a few genes and proteins at a single time. There are, however, a number of methods emerging, such as CITE-seq (53) and REAP-seq (54), that can now measure both transcripts and protein levels simultaneously for thousands of single cells.

The results that we have presented here will assist in setting better practices for model fitting and inference in the analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of single-cell data.

Data Availability

Simulations of the models used in the paper were performed using Gillespie's Stochastic Simulation Algorithm (SSA) implemented in Julia. The simulation code is available in the GitHub repository https://github.com/leham/PathwayReporters. The data used in the paper are provided in SI Tables.

ACKNOWLEDGEMENTS

We gratefully acknowledge Rowan D. Brackston for helpful discussions in the early stages of this research. L.H. and M.P.H.S. were supported by the University of Melbourne DVCR fund.

Bibliography

1. V. Shahrezaei and P. S. Swain. The stochastic nature of biochemical networks. *Curr. Opin. Biotechnol.*, 19(4):369–374, 2008. doi: 10.1016/j.copbio.2008.06.011.
2. P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci.*, 99(20):12795–12800, 2002. doi: 10.1073/pnas.162041399.
3. J. A. Goodrich and J. F. Kugel. Non-coding-RNA regulators of RNA polymerase ii transcription. *Nat. Rev. Mol. Cell. Biol.*, 7:612–616, 2006.
4. J. Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–418, 2004. doi: 10.1038/nature02257.
5. J. M. Pedraza and J. Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339–343, 2008. ISSN 0036-8075. doi: 10.1126/science.1144331.
6. M. Komorowski, Maria J. Costa, David A. Rand, and Michael P. H. Stumpf. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences of the United States of America*, 108(21):8645–8650, May 2011. doi: 10.1073/pnas.1015814108.
7. D. Jones and J. Elf. Bursting onto the scene? exploring stochastic mrna production in bacteria. *Curr. Opin. in Microbiol.*, 45:124–130, 2018. doi: 10.1016/j.mib.2018.04.001.
8. D. L. Jones, R. C. Brewster, and R. Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536, 2014. doi: 10.1126/science.1255301.
9. C.J. Zopf, K. Quinn, J. Zeidman, and N. Maheshri. Cell-cycle dependence of transcription dominates noise in gene expressions. *PLoS Comput. Biol.*, 9(7):e1003161, 2013. doi: doi.org/10.1371/journal.pcbi.1003161.
10. L. Ham, R. D. Brackston, and M. P.H. Stumpf. Extrinsic noise and heavy-tailed laws in gene expression. *Phys. Rev. Lett.*, 124:108101, 2020. doi: 10.1101/623371.
11. SO Skinner, H. Xu, S. Nagarkar-Jaiswal, P.R. Freire, T.P. Zwaka, and I. Golding. Single-cell analysis of transcription kinetics across the cell cycle. *Elife*, 10.7554/eLife.12175, 2016. doi: 10.1038/nprot.2013.066.
12. G. Gorin and L. Pachter. Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mrna production. *bioRxiv*, 2020. doi: 10.1101/2020.09.25.312868.
13. M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002. doi: 10.1126/science.1070919.
14. J. M. Raser and E. K. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 304(5678):1811–1814, 2004. doi: 10.1126/science.1098641.
15. A. Raj, C. S Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, 4(10):e309, 2006. doi: 10.1371/journal.pbio.0040309.
16. Tyler Quarton, Taek Kang, Vasileios Papakis, Khai Nguyen, Chance Nowak, Yi Li, and Leonidas Bleris. Uncoupling gene expression noise along the central dogma using genome engineered human cell lines. *Nucleic Acids Research*, 10.1093/nar/gkaa668, 08 2020. ISSN 0305-1048. gkaa668.
17. A. Singh. Transient changes in intercellular protein variability identify sources of noise in gene expression. *Biophys J.*, 107(9):2214–2220, 2014. doi: 10.1016/j.bpj.2014.09.017.
18. M. S. H. Ko. A stochastic model for gene induction. *J. Theor. Biol.*, 153(2):181–194, 1991. doi: 10.1016/S0022-5193(05)80421-7.
19. K. B. Halpern, S. Tanami, S. Landen, M. Chapal, L. Szla, A. Hutzler, and A. Nizhberg. Bursty gene expression in the intact mammalian liver. *Mol. Cell*, 58(1):147–156, 2015. doi: 10.1016/j.molcel.2015.01.027.
20. L.-H. So, A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding. General properties of transcriptional time series in escherichia coli. *Nature Genetics*, 43(6):554–560, 2011. doi: 10.1038/ng.821.
21. D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011. ISSN 0036-8075. doi: 10.1126/science.1198817.
22. A. JM. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(251–254), 2019. doi: 10.1038/s41586-018-0836-1.
23. J. Peccoud and B. Ycart. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.*, 48:222–234, 1995. doi: 10.1006/tpbi.1995.1027.
24. M. Abramowitz and I. A. Stegun. *Handbook Of Mathematical Functions With Formulas, Graphs, And Mathematical Tables*. U.S. Govt. Print. Off., 1965. ISBN 9780486612720.
25. L. Ham, R. D. Schnoerr, D.mand Brackston, and M. P. H. Stumpf. Exactly solvable models of stochastic gene expression. *J. Chem. Phys.*, 152(14):144106, 2020.
26. M. Greenwood and G. Udny Yule. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 83(2): 255–279, 1920.
27. I-B. Srividya, F. Hayot, and C. Jayaprakash. Transcriptional pulsing and consequent stochasticity in gene expression. *Phys. Rev. E*, 79(3):031911, 2009. doi: 10.1103/PhysRevE.79.031911.
28. W. Feller. On a general class of contagious distributions,. *Annals of Mathematical Statistics*, 14(4):389–400, 1943.
29. C. Jia and R. Grima. Small protein number effects in stochastic models of autoregulated bursty gene expression. *The Journal of Chemical Physics*, 152(8):084115, 2020. doi: 10.1063/1.5144578.
30. Z. Cao and R. Grima. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences*, 117(9): 4682–4692, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1910888117.
31. C. H. L. Beentjes, R. Perez-Carrasco, and R. Grima. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. *Phys. Rev. E*, 101: 032403, Mar 2020. doi: 10.1103/PhysRevE.101.032403.
32. A. Hilfinger and J. Paulsson. Separating intrinsic fluctuations in dynamic biological systems. *PNAS*, 108(29):12167–12172, 2011.
33. A. Hilfinger, M. Chen, and J. Paulsson. Using temporal correlations and full distributions to separate intrinsic and extrinsic fluctuations in biological systems. *Phys Rev Lett.*, 109: 248104, 2012. doi: doi:10.1103/PhysRevLett.109.248104.
34. S. Ross. *A first course in probability 9th edition*. Pearson, 2014.
35. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018. doi: 10.1038/s41586-018-0414-6.
36. G. Gorin and L. Pachter. Special function methods for bursty models of transcription. *Phys. Rev. E*, 102:022409, Aug 2020. doi: 10.1103/PhysRevE.102.022409.
37. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 10.1038/s41587-020-0591-3, 2020.
38. T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *J. Math. Biol.*, 54(1):1–26, 2007. doi: 10.1007/s00285-006-0034-x.
39. M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001. doi: 10.1073/pnas.151588598.
40. Nir Friedman, Long Cai, and X. Sunney Xie. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys. Rev. Lett.*, 97:168302, 2006. doi: 10.1103/PhysRevLett.97.168302.
41. A. Singh and J. P. Hespanha. Stochastic analysis of gene regulatory networks using moment closure. *2007 American Control Conference*, 10.1109/ACC.2007.4282604:1299–1304, 2007.
42. V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci.*, 105(45):17256–17261, 2008. doi: 10.1073/pnas.0803850105.
43. P. Bokes, John R. King, Andrew T. A. Wood, and Matthew Loose. Exact and approximate distributions of protein and mrna levels in the low-copy regime of gene expression. *J. Math. Biol.*, 64:829–854, 2012. doi: 10.1007/s00285-011-0433-5.
44. N Molina, D. M. Suter, R. Cannavo, B. Zoller, I. Gotic, and F. Naef. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proceedings of the National Academy of Sciences*, 110(51):20563–20568, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1312310110.
45. J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao, and S. N. Cohen. Global analysis of mrna decay and abundance in escherichia coli at single-gene resolution using two-color fluorescent dna microarrays. *Proceedings of the National Academy of Sciences*, 99(15): 9697–9702, 2002. ISSN 0027-8424. doi: 10.1073/pnas.112318199.
46. B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473 (7347):337–342, 2011. doi: 10.1038/nature10098.
47. I. Golding, J Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005. doi: 10.1016/j.cell.2005.09.031.
48. Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying E. Coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010. ISSN 0036-8075. doi: 10.1126/science.1188308.
49. A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *PNAS*, 103(35):13004–13009, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0605420103.
50. S. Shah, Takei Y., W. Zhou, E. Lubeck, J. Yun, C.L. Eng, N. Koulena, C. Cronin, C. Karp, E.J. Liaw, M. Amin, and L. Cai. Dynamics and spatial genomics of the nascent transcriptome by intron seqfish. *Cell.*, 174(2):363–376, 2018. doi: 10.1016/j.cell.2018.05.035.
51. A. Hilfinger, T.M. Norman, and J. Paulsson. Fluctuations to identify kinetic mechanisms in sparsely characterized systems. *Cell Syst.*, 2(4):251–259, 2016. doi: 10.1016/j.cels.2016.04.002.
52. J. Cole and Z. Luthey-Schulten. A careful accounting of extrinsic noise in protien expression reveals correlations among its sources. *Phys. Rev. E*, 95(6):062418, 2017.
53. M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017. doi: 10.1038/nmeth.4380.
54. V.M. Peterson, K.X. Zhang, N. Kumar, J. Wong, L. Li, D.C. Wilson, R. Moore, T.K. McClanahan, S. Sadekova, and J.A. Klappenbach. Multiplexed quantification of proteins and

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transcripts in single cells. *Nat. Biotechnol.*, 35(10):936–939, 2017. doi: 10.1038/nbt.3973.

55. F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.
56. N. Saad and R. L. Hall. Integrals containing confluent hypergeometric functions with applications to perturbed singular potentials. *Journal of Physics A: Mathematical and General*, 36(28):7771–7788, jul 2003. doi: 10.1088/0305-4470/36/28/307.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Note 1: Derivations of the Non-identifiability Results

In the main text we provide a number of examples of when the compound distribution does not have a unique representation. We here provide the full details of the derivations of these results.

A. Bursty Expression: Telegraph Representation. Consider first a Telegraph distribution $\tilde{p}_T(n;\lambda,\mu',K')$ and the probability density function of the scaled beta distribution $\mathrm{Beta}_{K'}(\lambda+\mu,\mu'-\mu)$, given by

$$f_{K'}(t;\lambda+\mu,\mu'-\mu)=\frac{\Gamma(\lambda+\mu')}{\Gamma(\lambda+\mu)\Gamma(\mu'-\mu)}K'^{1-\lambda-\mu'}t^{\lambda+\mu-1}(K'-t)^{\mu'-mu-1},$$

Note that this distribution has support $[0,K']$. In the main text we claim that the Telegraph distribution $\tilde{p}_T(n;\lambda,\mu',K')$ can be obtained from compounding a Telegraph distribution by a scaled beta distribution with pdf $f(t;\lambda+\mu,\mu'-\mu)$. In other words, that $\tilde{p}_T(n;\lambda,\mu',K')$ can be written as:

$$\tilde{p}_T(n;\lambda,\mu',K')=\int_0^{K'}\tilde{p}_T(n;\lambda,\mu,t)f_{K'}(t;\lambda+\mu,\mu'-\mu)\,dt \tag{15}$$

This is Eq. (3) in the main text. Starting from the right hand side of Eq. (15), we have

$$\begin{aligned}&\int_0^{K'}\tilde{p}_T(n;\lambda,\mu,t)f_{K'}(t;\lambda+\mu,\mu'-\mu)\,dt\\&=\frac{1}{n!}\frac{\Gamma(\lambda+\mu')}{\Gamma(\lambda+\mu)\Gamma(\mu'-\mu)}\frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\frac{\Gamma(\lambda+\mu)}{\Gamma(\lambda+\mu+n)}\int_0^{K'}{}_1F_1(\lambda+n,\lambda+\mu+n,-t)(K')^{1-\lambda-\mu'}t^{\lambda+\mu+n-1}(K'-t)^{\mu'-\mu-1}\,dt\end{aligned} \tag{16}$$

Substituting $t=K'T$ and simplifying, the left hand side of Eq. (16) becomes

$$\frac{1}{n!}\frac{\Gamma(\lambda+\mu')}{\Gamma(\lambda+\mu)\Gamma(\mu'-\mu)}\frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\frac{\Gamma(\lambda+\mu)}{\Gamma(\lambda+\mu+n)}(K')^n\int_0^1{}_1F_1(\lambda+n,\lambda+\mu+n,-KT)(K')^{1-\lambda-\mu'}T^{\lambda-\mu-1}(K'-T)^{\mu'-\mu-1}\,dT$$

Now, using the integral representation (55)[13.4.2] of the confluent hypergeometric function, this becomes

$$\begin{aligned}&\frac{1}{n!}\frac{\Gamma(\lambda+\mu')}{\Gamma(\lambda+\mu)\Gamma(\mu'-\mu)}\frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\frac{\Gamma(\lambda+\mu)}{\Gamma(\lambda+\mu+n)}(K')^n\frac{\Gamma(\lambda+\mu+n)\Gamma(\mu'-\mu)}{\Gamma(\mu'+\lambda+n)}{}_1F_1(\lambda+n,\lambda+\mu'+n,-K')\\=&\frac{\Gamma(\lambda+\mu')}{\Gamma(\lambda+\mu'+n)}\frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\frac{(K')^n}{n!}{}_1F_1(\lambda+n,\lambda+\mu'+n,-K')\\=&\tilde{p}_T(n;\lambda,\mu',K'),\end{aligned}$$

which is the left hand side of Eq. (15). Hence, we have that Eq. (15) holds.

B. Instantaneously Bursty Expression: Negative Binomial Representations. We consider first the representation given in Eq. (4) in the main text. Recall that we let $\tilde{p}_{\mathrm{NB}}(n;r,\beta)$ denote the probability mass function of a $\mathrm{NegBin}(r,\frac{\beta}{\beta+1})$ distribution, where $\beta>0$. In the main text we claim that for any negative binomial distribution of the form $\mathrm{NegBin}(\lambda,\frac{\beta}{\beta+1})$ (where $\beta>0$) we have,

$$\tilde{p}_{\mathrm{NB}}(n;\lambda,\beta)=\int_0^\infty\tilde{p}_T(n;\lambda,\mu,t)f(t;\lambda+\mu,\beta)\,dt, \tag{17}$$

where $f(t;\lambda+\mu,\beta)$ is the probability mass function of a $\mathrm{Gamma}(\lambda+\mu,\beta)$ distribution. Beginning with the right hand side of Eq. (17) we have,

$$\begin{aligned}&\int_0^\infty\tilde{p}_T(n;\lambda,\mu,t)f(t;\lambda+\mu,\beta)\,dt\\&\qquad=\frac{1}{n!}\frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\frac{\Gamma(\lambda+\mu)}{\Gamma(\lambda+\mu+n)}\frac{\beta^{(\lambda+\mu)}}{\Gamma(\lambda+\mu)}\int_0^\infty{}_1F_1(\lambda+n,\lambda+\mu+n,-t)t^{\lambda+\mu+n-1}e^{-\beta t}\,dt\end{aligned} \tag{18}$$

We now apply the following identity given in (56)[Appendix 1] with $a=\lambda+n$, $b=\lambda+\mu+n$, $k=-1$, $d=\lambda+\mu+n$ and $h=\beta$.

$$\int_0^\infty{}_1F_1(a,b,kt)t^{d-1}e^{-ht}=\frac{\Gamma(d)}{h^d}(1-\frac{k}{h})^{-a}. \tag{19}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

So the left hand side of Eq. (18) becomes

$$\frac{1}{n!}\frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\frac{\Gamma(\lambda+\mu)}{\Gamma(\lambda+\mu+n)}\frac{\beta^{(\lambda+\mu)}}{\Gamma(\lambda+\mu)}\frac{\Gamma(\lambda+\mu+n)}{\beta^{\lambda+\mu+n}}\left(1+\frac{1}{\beta}\right)^{-(\lambda+n)}=\frac{1}{n!}\frac{\Gamma(\lambda+n)}{\Gamma(\lambda)}\frac{\beta^{\lambda}}{(\beta+1)^{\lambda+n}}=\tilde{p}_{\mathrm{NB}}(n;\lambda,\beta),$$

which is the right hand side of Eq. (17). Hence, we have that Eq. (17) holds. We now consider the representation given in Eq. 5 of the main text. Here we claim that any negative binomial distribution of the form $\mathrm{NegBin}(\lambda',\frac{c}{1+c})$ (where $c>1$) can be written as,

$$\tilde{p}_{\mathrm{NB}}(n;\lambda',c)=\int_{\frac{1}{c}}^{\infty}\tilde{p}_{\mathrm{NB}}(n;\lambda,\theta)f_c(\theta c-1;\lambda-\lambda',\lambda')\,d\theta \tag{20}$$

where $f_c(c\theta-1;\lambda-\lambda',\lambda')$ is the probability mass function of a scaled beta prime $\mathrm{BetaPrime}_c(\lambda-\lambda',\lambda)$ distribution, where $c>0$ and $\lambda>\lambda'$. Starting from the right hand side of Eq. (20), we have

$$\begin{aligned}\int_{\frac{1}{c}}^{\infty}\tilde{p}_{\mathrm{NB}}(n;\lambda,\theta)f_c(\theta c-1;\lambda-\lambda',\lambda')\,d\theta&=\int_{\frac{1}{c}}^{\infty}\frac{\Gamma(\lambda+n)}{\Gamma(n+1)\Gamma(\lambda)}\left(1-\frac{\theta}{\theta+1}\right)^{n}\left(\frac{\theta}{\theta+1}\right)^{\lambda}\frac{c\Gamma(\lambda)}{\Gamma(\lambda-\theta)\Gamma(\theta)}\frac{(c\theta-1)^{\lambda-\lambda'-1}}{(c\theta)^{\lambda}}\\&=\frac{c^{1-\lambda}\Gamma(\lambda+n)}{\Gamma(n+1)\Gamma(\lambda')\Gamma(\lambda-\lambda')}\int_{\frac{1}{c}}^{\infty}\frac{(c\theta-1)^{\lambda-\lambda'-1}}{(1+\theta)^{\lambda+n}}\end{aligned}$$

Now substituting $H=c\theta-1$ and simplifying, we obtain

$$\frac{c^{n}\Gamma(\lambda+n)}{\Gamma(n+1)\Gamma(\lambda')\Gamma(\lambda-\lambda')}\int_{0}^{\infty}\frac{H^{\lambda-\lambda'-1}}{(H+c+1)^{\lambda+n}}\,dH \tag{21}$$

Now, letting $y=c+1$ and simplifying we have that

$$\int_{0}^{\infty}\frac{H^{\lambda-\lambda'-1}}{(H+c+1)^{\lambda+n}}\,dH=\frac{1}{(c+1)^{\lambda'+n}}\int_{0}^{\infty}\frac{y^{\lambda-\lambda'-1}}{(y+1)^{\lambda+n}}\,dy \tag{22}$$

$$=\frac{1}{(c+1)^{\lambda'+n}}\frac{\Gamma(\lambda-\lambda')\Gamma(\lambda'+n)}{\Gamma(\lambda+n)} \tag{23}$$

Here we used the fact that the integral in the variable y is the probability density function of a $\mathrm{BetaPrime}(\lambda-\lambda',\lambda'+n)$ distribution. Thus, Eq. 21 simplifies to

$$\frac{\Gamma(\lambda'+n)}{\Gamma(n+1)\Gamma(\lambda')}\frac{c^{n}}{(c+1)^{\lambda'+n}}=\tilde{p}_{\mathrm{NB}}(\lambda',c), \tag{24}$$

which is the right hand side of Eq. (20). So Eq. (20) holds.

Supplementary Note 2: Proof of the Noise Decomposition Principle

Here we provide the proof of the Noise Decomposition Principle given in the main text. For convenience, we restate the principle below.

The Noise Decomposition Principle (NDP). *Assume that there are measurable functions f, g and h such that $\mathrm{E}(X;\mathbf{Z})$ and $\mathrm{E}(Y;\mathbf{Z})$ split across common variables by way of $\mathrm{E}(X;\mathbf{Z})=f(\mathbf{Z}_X)h(\mathbf{Z}')$ and $\mathrm{E}(Y;\mathbf{Z})=g(\mathbf{Z}_Y)h(\mathbf{Z}')$. Then provided that the variables $Z_1,\ldots,Z_m$ are mutually independent, the normalised covariance of $\mathrm{E}(X;\mathbf{Z})$ and $\mathrm{E}(Y;\mathbf{Z})$ will identify the total noise on $h(\mathbf{Z}')$ (i.e., $\eta^2_{h(Z')}$).*

Consider first the covariance of $\mathrm{E}(X;\mathbf{Z})$ and $\mathrm{E}(Y;\mathbf{Z})$. We have

$$\begin{aligned}\mathrm{Cov}(\mathrm{E}(X;\mathbf{Z}),\mathrm{E}(Y;\mathbf{Z}))&=\mathrm{Cov}(f(\mathbf{Z}_X)h(\mathbf{Z}'),g(\mathbf{Z}_Y)h(\mathbf{Z}'))\\&=\mathrm{E}(f(\mathbf{Z_X})g(\mathbf{Z_Y})(h(\mathbf{Z}'))^2)-\mathrm{E}(f(\mathbf{Z_X})h(\mathbf{Z}'))\,\mathrm{E}(g(\mathbf{Z_Y})h(\mathbf{Z}'))\\&=\mathrm{E}(f(\mathbf{Z_X}))\,\mathrm{E}(g(\mathbf{Z_Y}))\left[\mathrm{E}((h(\mathbf{Z}'))^2)-(\mathrm{E}(h(\mathbf{Z}')))^2\right]\end{aligned} \tag{25}$$

Note that here we used the fact that the variables in $\mathbf{Z}=(Z_1,\ldots,Z_n)$ are mutually independent. Now using the fact that $\mathrm{E}(X)=\mathrm{E}(\mathrm{E}(X;\mathbf{Z}))$ and $\mathrm{E}(Y)=\mathrm{E}(\mathrm{E}(Y;\mathbf{Z}))$ (the Law of Total Expectation), and then normalising we obtain

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{aligned}\frac{\mathrm{Cov}(\mathrm{E}(X;\mathbf{Z}),\mathrm{E}(Y;\mathbf{Z}))}{\mathrm{E}(X)\,\mathrm{E}(Y)} &= \frac{\mathrm{E}(f(\mathbf{Z_X}))\,\mathrm{E}(g(\mathbf{Z_Y}))\left[\mathrm{E}((h(\mathbf{Z}'))^2)-(\mathrm{E}(h(\mathbf{Z}')))^2\right]}{\mathrm{E}(f(\mathbf{Z_X}))\,\mathrm{E}(g(\mathbf{Z_Y}))(\mathrm{E}((h(\mathbf{Z}')))^2} \\ &= \frac{\mathrm{Var}(h(\mathbf{Z}'))}{(\mathrm{E}(h(\mathbf{Z}')))^2} \\ &= \eta^2_{h(\mathbf{Z}')} \end{aligned} \tag{26}$$

Hence, under certain conditions the normalised covariance of $\mathrm{E}(X;\mathbf{Z})$ and $\mathrm{E}(Y;\mathbf{Z})$ will identify the total noise on $h(\mathbf{Z}')$.

Supplementary Note 3: Pathway reporters

A. Justification for simulating only one copy of the gene. Our simulations and theory have been based over reporters from a single gene copy, whereas in practice there may be multiple copies of the gene that contribute to the overall mRNA. If there are mechanisms in place to distinguish the mRNA or protein of one gene copy from another, then the theory and analysis we have developed in this paper holds without change. In the case where it is not possible to distinguish mRNA and protein (respectively) from the multiple gene copies, then we now observe that the general theory continues to hold, provided there is independence between the gene copies; this assumption has been verified experimentally (11). Here we demonstrate this in the case of two gene copies, though the general case for more than two genes is essentially identical but is notationally cumbersome. We are considering a situation where the variable X in the Noise Decomposition Principle is the sum of two independent variables X_1, X_2 and Y is the sum of two independent variables Y_1, Y_2. We assume common dependence on the environmental variables $\mathbf{Z}$ so that $E(X_1;\mathbf{Z}) = E(X_2;\mathbf{Z})$, $E(Y_1;\mathbf{Z}) = E(Y_2;\mathbf{Z})$. Using these equalities and the independence of X_1, X_2 and Y_1, Y_2 in $X = X_1 + X_2$, $Y = Y_1 + Y_2$, we find the numerator of

$$\frac{\mathrm{Cov}(\mathrm{E}(X;\mathbf{Z}),\mathrm{E}(Y;\mathbf{Z}))}{\mathrm{E}(X)\,\mathrm{E}(Y)}$$

is simply $4\,\mathrm{Cov}(\mathrm{E}(X_1;\mathbf{Z}),\mathrm{E}(Y_1;\mathbf{Z}))$, while the denominator is $4\,\mathrm{E}(X_1)\,\mathrm{E}(Y_1)$. Thus the noise decomposition coincides with that for the single copy X_1, Y_1. Further work may be required to consider systems where there is independence between, or there is significant deviations in the gene copies.

B. Upper Bound for the Intrinsic Contribution to the Covariance: Constitutive Expression. In the main text, we claim that the error in the pathway-reporter approach in the case of mRNA-protein reporters is negligible (i.e. the error is $\ll 1$). Our argument relies on the expression given in Eq. 13 in the main text. We here provide full details of the derivation of this expression. First let X_m and X_p be the number of mRNA and protein produced from the same constitutive gene modelled by the "two-stage" model, $\mathbf{M}_2$ (see Figure 3A (top right) of the main text). Also, let $\mathbf{Z} = \{K_m, \delta_m, K_p, \delta_p\}$.
We restate here the expression for the intrinsic contribution to the covariance of X_m and X_p given in Eq. 13 of the main text.

$$\frac{\mathrm{E}(\mathrm{Cov}(X_m, X_p;\mathbf{Z}))}{\mathrm{E}(X_m)\,\mathrm{E}(X_p)} = \frac{\alpha}{\mathrm{E}(K_m)}, \quad \text{where } \alpha = \frac{\mathrm{E}(1/(\delta_p+1))}{\mathrm{E}(1/\delta_p)}. \tag{27}$$

We require the following expressions for the stationary mean mRNA level and protein level of the two-stage model (39, 41).

$$\mathrm{E}(X_m;\mathbf{Z}) = \frac{K_m}{\delta_m}, \quad \mathrm{E}(X_p;\mathbf{Z}) = \frac{K_p}{\delta_p}\frac{K_m}{\delta_m} \quad \text{and} \quad \mathrm{Cov}(X_m, X_p;\mathbf{Z}) = \frac{K_m K_p}{\delta_m(\delta_m+\delta_p)}. \tag{28}$$

Assuming that δ_m is fixed across the cell-population, and all parameters are scaled so that $\delta_m = 1$, it follows from Eq. (28), that

$$\mathrm{E}(\mathrm{Cov}(X_m, X_p;\mathbf{Z})) = \mathrm{E}\left(\frac{1}{\delta_p+1}\right)\mathrm{E}(K_m)\,\mathrm{E}(K_p). \tag{29}$$

Using the Total Law of Expectation, we also have,

$$\mathrm{E}(X_m) = \mathrm{E}(K_m) \quad \text{and} \quad \mathrm{E}(X_p) = \mathrm{E}\left(\frac{1}{\delta_p}\right)\mathrm{E}(K_m)\,\mathrm{E}(K_p). \tag{30}$$

Thus, it follows that

$$\frac{\mathrm{E}(\mathrm{Cov}(X_m, X_p;\mathbf{Z}))}{\mathrm{E}(X_m)\,\mathrm{E}(X_p)} = \frac{\alpha}{\mathrm{E}(K_m)}, \quad \text{where } \alpha = \frac{\mathrm{E}(1/(\delta_p+1))}{\mathrm{E}(1/\delta_p)}. \tag{31}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.30.319814; this version posted September 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

C. Determination of the Marginal Means in the Model $\mathbf{M}_4$. To derive the marginal means for nascent and mature mRNA and for protein, we first observe that the nascent mRNA population may treated identically to that of mRNA in general (that is, no distinction between nascent and mature), as in (23), except that mRNA decay is replaced by the sum of decay and maturation. As in the work of Cao and Grima (30), the assumption of fast maturation allows us to ignore decay completely in the nascent phase, so that the marginal distribution is identical to that of (23), except with decay replaced by maturation. This leads to a marginal nascent mRNA mean of

$$\mathrm{E}(X_N;\mathbf{Z}) = \frac{K_N}{K_M}\frac{\lambda}{(\lambda+\mu)}.$$

The marginal means for mature mRNA and protein are derived in (15) under the assumption that the transcription rate parameter K_N is large relative to the other parameters. The expressions are given by

$$\mathrm{E}(X_M;\mathbf{Z}) = \frac{K_N}{\delta_M}\frac{\lambda}{(\lambda+\mu)} \quad \text{and} \quad \mathrm{E}(X_P;\mathbf{Z}) = \frac{K_P}{\delta_P}\frac{K_N}{\delta_M}\frac{\lambda}{(\lambda+\mu)}.$$

Formally, the marginal means in (15) are for the three-stage model $\mathbf{M}_3$, which ignores the downstream processing of mRNA, such as splicing. The assumption of fast maturation however, justifies the treatment of the nascent phase of mRNA as an ephemeral step within the Poissonian modelling of mRNA transcription.
There are a number of possibly compounding assumptions on the parameters here, but simulations show that there is a lot of tolerance, with even only moderate maturation and transcription still returning sample means consistent with the formulas.

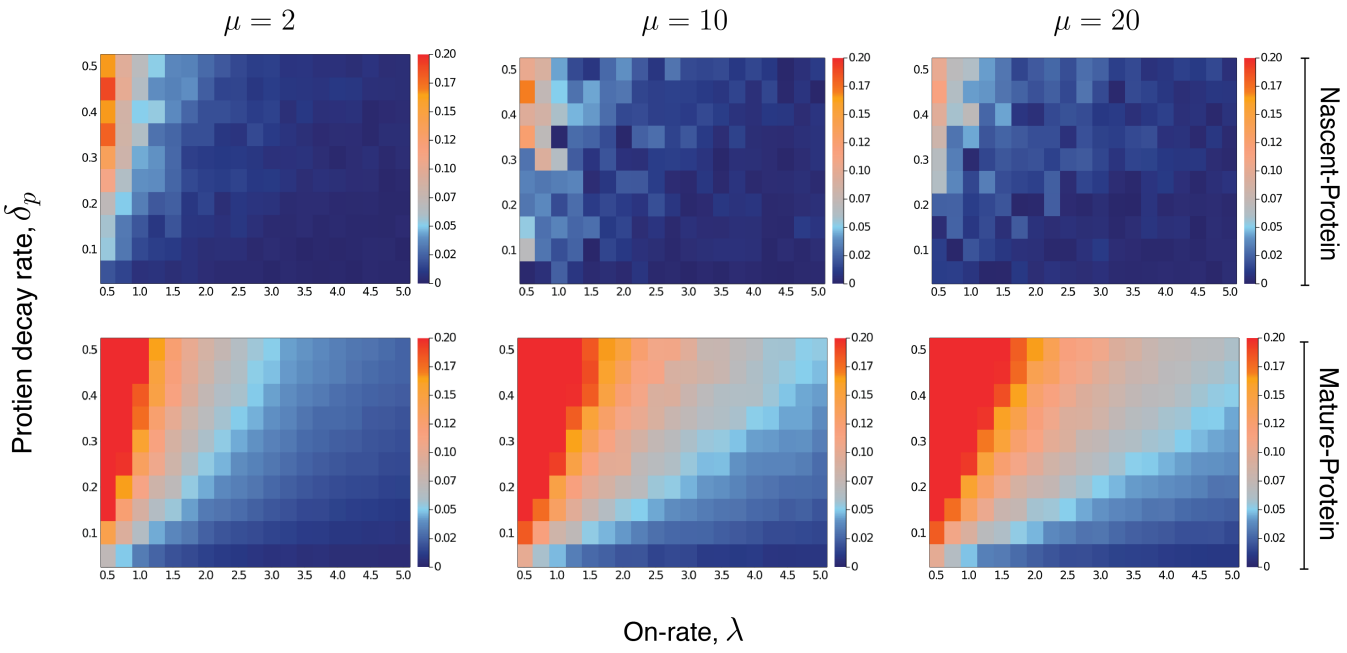

Fig. 4. An extension of Fig. 3D of the main text. The heatmaps shown for the intrinsic contribution to the covariance estimate the level of overshoot in the pathway-reporter approach, for both the nascent-protein and mature-protein reporters. For more details refer to the caption given in Fig. 3. Here we include the additional case of $\mu = 10$. The remaining parameter values are the same as in the main text. Each individual pixel is generated from a sample of size 3000, though there is still some instability in the convergence for the nascent-protein reporter, particularly as the overshoot estimation starts to increase, and particularly as μ is larger. To produce more accurate values, the case of $\mu = 2$ was averaged over two full experiments while $\mu = 20$ was averaged over three. This was also done for the mature-protein reporter, however for these images there was almost no visible difference between the various runs of the experiment and their averages. Each of the three μ values takes roughly 7–10 hours of computation, depending on lead in time before sampling within a simulation.