Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production

Gennady Gorin[1] and Lior Pachter[2]

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, 91125
[2]Division of Biology and Biological Engineering & Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125
[*]Address correspondence to Lior Pachter (lpachter@caltech.edu)

September 25, 2020

Abstract

Intrinsic and extrinsic noise sources in gene expression, originating respectively from transcriptional stochasticity and from differences between cells, complicate the determination of transcriptional models. In particularly degenerate cases, the two noise sources are altogether impossible to distinguish. However, the incorporation of downstream processing, such as the mRNA splicing and export implicated in gene expression buffering, recovers the ability to identify the relevant source of noise. We report analytical copy-number distributions, discuss the noise sources' qualitative effects on lower moments, and provide simulation routines for both models.

1 Background

Recent improvements in single-cell transcriptomics, including increasingly sensitive fluorescence- and sequencing-based methods, have begun to provide data useful for discriminating between competing biophysical models. One immediate application of interest is that of *intrinsic* and *extrinsic* cellular gene expression noise, which has already been studied directly from mRNA reporter statistics [1, 2]. While experimental and statistical methods for measuring the relative contributions of intrinsic and extrinsic noise are relatively advanced [3–6], microscopic models of cell-to-cell variability are less well developed. These models are necessary in light of recent methods for measuring the molecular state of cells, which offer routes to better mechanistic understanding, but present a number of new challenges in controlling noise sources.

While the introduction of single-cell RNA sequencing (scRNA-seq) data with unique molecular identifiers (UMIs) provides measurements of a substantial fraction of transcripts in individual cells [7], the resulting copy-number data are discrete, and thus challenging to model with existing methods that largely focus on continuous-valued fluorescence readouts. The biochemistry of scRNA-seq also generally relies on the capture of polyadenylated sequences in fixed media [8], which limits the

scope of assays, and is not directly compatible with *in vivo* experimental methods relying on the integration of multiple fluorescent reporters to distinguish between the sources of noise [3]. Furthermore, the analysis of lower moments of gene expression has been shown to be insufficient for the identification of biophysical parameters even for purely intrinsic noise models [9], suggesting that full copy-number distributions are necessary for modeling more complex systems with multiple sources of noise.

Another challenge lies in theory; ideally, analytical results will be available to provide qualitative interpretability and guide computational approaches, but many current methods are purely numerical. For example, while methods for the explicit description of extrinsic noise are formally available, in the context of a transcriptional model, the incorporation of extrinsic noise typically corresponds to the construction of a mixture model with parameter values drawn from a distribution [2, 3, 10]. Under this construction, full analytical solutions are only available in the simplest cases.

2 Two models for gene expression

It is well-known that the common two-state model of gene expression [11] gives rise to a negative binomial (NB) distribution of mRNA counts in the short-burst limit [12]. However, a recent study shows that constitutive transcription in a cell population with a gamma-distributed production rate parameter also yields a negative binomial distribution of mRNA counts [13]. Although there are both experimental and theoretical arguments favoring a bursting model for eukaryotic transcription [14–17] – current theories posit that superstructure modifications are responsible for occlusion and exposure of the gene locus [18, 19] – a comprehensive model should account for all relevant sources of noise, as well as provide both a quantitative and qualitative understanding of their effects.

A current limitation of existing models is that processes downstream of eukaryotic mRNA production, such as export and/or splicing processes [20, 21], are generally ignored. However, promising new technologies and experiments, based on fluorescence [22, 23] and sequencing [24] methods, can distinguish nascent from mature mRNA molecules based on spatial or intronic readouts, thus providing essential data for studying model of increasing complexity. In particular, the maturation of these methods and the availability of resulting multimodal data naturally suggests the potential of fitting otherwise poorly identifiable models [9].

As a first step, and to gain a qualitative understanding of the effects of intrinsic and extrinsic noise in the context of downstream processing, we compare solutions of two simple two-stage models of transcription that include downstream processing at steady state. Both models assume that nascent mRNA (unspliced or pre-mRNA) is converted to mature mRNA (spliced mRNA) after an exponentially-distributed delay, corresponding to splicing. This is followed by another exponentially-distributed delay that models the mature mRNA being degraded. The splicing rate β and degradation rate γ are deterministic. The gene locus dynamics are modeled by either bursts, with stochastic burst size $B \sim Geom(b)$ and deterministic burst *i*nitiation frequency k_i [20], or constitutive, with stochastic but constant transcription rate $K \sim Gamma(\alpha, \eta)$. The model parametrizations are illustrated in Figure 1. We calculate lower moments and cross-moments, and show how these can be used to differentiate between distributions and statistics resulting from the two models.

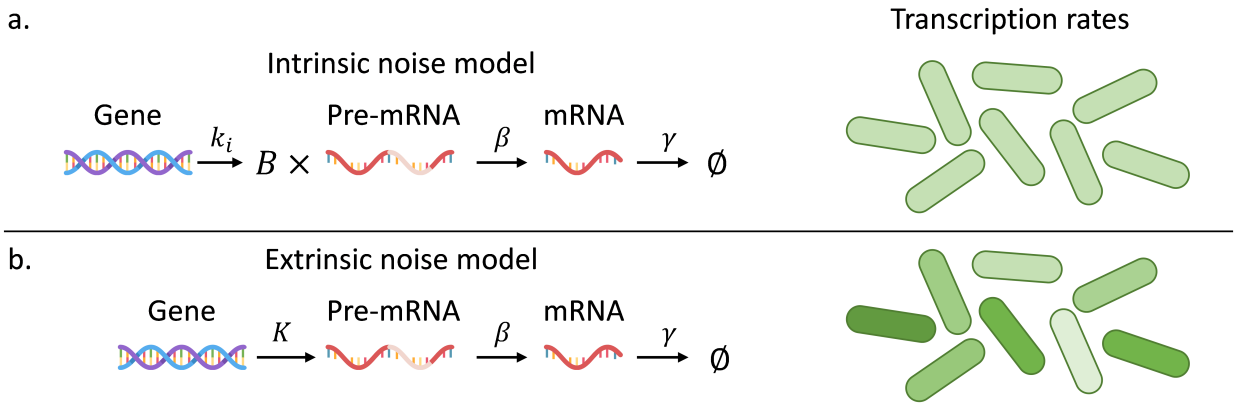

Figure 1: (a) Schema of the intrinsic noise model (k_i: burst frequency; B: burst size drawn from a geometric distribution; β: pre-mRNA splicing rate; γ: mRNA degradation rate. Uniform shade of green indicates identical parameter values across all cells). (b) Schema of the extrinsic noise model (K: transcription rate; β: pre-mRNA splicing rate; γ: mRNA degradation rate. Different shades of green indicate different values of K across cells).

3 Notation

3.1 Model parametrization

Model-independent quantities and statistics are defined in Table 1. The two models' parameters are defined in Tables 2 and 3. Finally, x_z, where x is a statistic computed from data moments (e.g., μ_M, σ^2_M, ρ, γ) and $z \in \{i, e\}$ refers to the predicted value of that statistic based on either the $i$ntrinsic or $e$xtrinsic noise model. For example, $\mu_{M,i}$ refers to the predicted mean mature mRNA copy number under the intrinsic noise model, while ρ_e refers to the predicted nascent–mature correlation under the extrinsic noise model.

A probability mass function (PMF) associated with a discrete-valued random variable X is equivalently denoted by $P(\cdot;\cdot)$ or $P(X = k;\cdot)$. A probability density function (PDF) associated with a continuous-valued random variable is denoted by $f(\cdot;\cdot)$.

3.2 Probability distributions

The geometric distribution is defined as follows: if $X \sim Geom(p)$, $P(X = k; p) = (1-p)^k p$, where $p \in (0, 1]$ and $k \in \mathbb{N}_0$. The geometric distribution is well-known to arise in the short-burst limit of the two-state transcription model [25].

The negative binomial distribution is defined as follows: if $X \sim NegBin(r, p)$, $P(X = k; r, p) = \frac{\Gamma(r+k)}{k!\Gamma(r)}(1-p)^r p^k$, where $p \in [0, 1]$ and $r > 0$. We note that MATLAB and the NumPy library take the opposite convention, with a $\tilde{p}$ parameter defined as $1 - p$.

The gamma distribution is defined as follows: if $X \sim Gamma(\alpha, \eta)$, $f(x; \alpha, \eta) = \frac{\eta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\eta x}$. This is the shape/rate parametrization. We note that MATLAB and the NumPy library take the opposite shape/scale parametrization with parameter $\theta = \eta^{-1}$. Furthermore, the rate η is usually given the variable name "β"; however, we use the current convention to preclude confusion with the splicing rate parameter.

Table 1: Observation variables

Parameter	Definition
n	Number of nascent mRNA
m	Number of mature mRNA
N	Random variable denoting number of nascent mRNA; $N \in \mathbb{N}_0$
M	Random variable denoting number of mature mRNA; $M \in \mathbb{N}_0$
γ	Degradation rate
μ_z, $\mathbb{E}[Z]$	Expectation of species $Z \in \{N, M\}$
μ	Expectation of an arbitrary distribution
σ_z^2	Variance of species $Z \in \{N, M\}$
σ^2	Variance of an arbitrary distribution $Z \in \{N, M\}$
$Cov(N, M)$	Covariance nascent and mature copy numbers
ρ	Pearson correlation coefficient
q	Computed statistic $\frac{\sigma_N^2}{\mu_N} - 1$
$P(n, m; \cdot)$	Joint PMF of nascent and mature mRNA, shorthand for $P(N = n, M = m; \cdot)$
$P(\cdot; \cdot)$	PMF of an arbitrary parametrized discrete random variable
$f(\cdot; \cdot)$	PDF of an arbitrary parametrized continuous random variable

Table 2: Intrinsic noise model parameters

Parameter	Definition
k_i	Burst frequency
B	Geometric random variable denoting burst size
b	Mean of B
β	Splicing rate
γ	Degradation rate
f	$\frac{\beta}{\beta+\gamma}$, non-dimensional splicing rate

Table 3: Extrinsic noise model parameters

Parameter	Definition
K	Gamma-distributed random variable denoting transcription rate
α	Shape of gamma distribution
η	Rate of gamma distribution
β	Splicing rate
γ	Degradation rate
f	$\frac{\beta}{\beta+\gamma}$, non-dimensional splicing rate

4 Preliminaries

4.1 Intrinsic noise model

4.1.1 Probability mass function

The full joint distribution for the burst model requires numerical integration and Fourier transformation [20]. To our knowledge, no analytical solution exists, although approximations in terms of hypergeometric functions are available [26].
The nascent marginal is distributed per $NegBin(\frac{k_i}{\beta}, \frac{b}{1+b})$. The mature marginal is NB-distributed in the limit of low β and Poisson-distributed in the limit of high β. Although the distribution in the intermediate region is qualitatively similar to NB, it does not appear to be *exactly* representable as NB. Furthermore, even the determination of the closest NB approximation according to some divergence metric is an open problem, although method of moments approximations may be satisfactory for some purposes.

4.1.2 Moments

Per the results from Singh and Bokes [20]:

$$\begin{aligned}
\mu_N &= \frac{k_i b}{\beta} \\
\mu_M &= \frac{k_i b}{\gamma} \\
\sigma_N^2 &= \mu_N(1+b) = \frac{k_i b}{\beta}(1+b) \\
\sigma_M^2 &= \mu_M\left(1+\frac{b\beta}{\beta+\gamma}\right) = \frac{k_i b}{\gamma}\left(1+\frac{b\beta}{\beta+\gamma}\right) \\
Cov(N,M) &= \frac{\mu_N b\beta}{\beta+\gamma} = \frac{k_i b}{\beta}\frac{b\beta}{\beta+\gamma} = \frac{k_i b^2}{\beta+\gamma},
\end{aligned}$$

yielding the following Pearson correlation coefficient:

$$\begin{aligned}
\rho &:= \frac{Cov(N,M)}{\sigma_N \sigma_M} \\
&= \frac{\frac{k_i b^2}{\beta+\gamma}}{\sqrt{\frac{k_i b}{\beta}(1+b)\frac{k_i b}{\gamma}\left(1+\frac{b\beta}{\beta+\gamma}\right)}} \\
&= b\sqrt{\frac{f(1-f)}{(1+b)(1+bf)}},
\end{aligned}$$

where f controls the relationship between the splicing and degradation timescales.

4.2 Extrinsic noise model

4.2.1 Probability mass function

The full time-dependent copy-number probability distribution under constitutive production is well-known and represents one of the most valuable and general results in chemical master equation (CME) analysis [27]. In the relevant steady-state regime, the solution $\tilde{P}(n,m)$ giving the probability of a state with a given number of nascent and mature molecules is the product of independent Poisson distributions. Given a production rate K, splicing rate β, and degradation rate γ,

$$\tilde{P}(n,m;K/\beta,K/\gamma)=\left(\frac{(K/\beta)^n e^{-K/\beta}}{n!}\right)\left(\frac{(K/\gamma)^m e^{-K/\gamma}}{m!}\right)$$

Therefore, marginalizing over the gamma-distributed production rate K:

$$\begin{aligned}
P(n,m;\alpha,\eta)&=\int_0^\infty P(n,m;x)f(x;\alpha,\eta)dx\\
&=\int_0^\infty\left(\frac{(x/\beta)^n e^{-x/\beta}}{n!}\right)\left(\frac{(x/\gamma)^m e^{-x/\gamma}}{m!}\right)\frac{\eta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\eta x}dx\\
&=\frac{\eta^\alpha}{\Gamma(\alpha)n!m!\beta^n\gamma^m}\int_0^\infty x^{n+m-\alpha-1}e^{-x\left(\eta+\frac{1}{\beta}+\frac{1}{\gamma}\right)}dx\\
&=\frac{\Gamma(\alpha+n+m)}{\Gamma(\alpha)n!m!}\left(\frac{\eta}{\eta+\frac{1}{\beta}+\frac{1}{\gamma}}\right)^\alpha\left(\frac{1}{\beta\left(\eta+\frac{1}{\beta}+\frac{1}{\gamma}\right)}\right)^n\left(\frac{1}{\gamma\left(\eta+\frac{1}{\beta}+\frac{1}{\gamma}\right)}\right)^m\\
&=\frac{\Gamma(\alpha+n+m)}{\Gamma(\alpha)n!m!}\left(\frac{1}{C}\right)^\alpha\left(\frac{\frac{1}{\eta\beta}}{C}\right)^n\left(\frac{\frac{1}{\eta\gamma}}{C}\right)^m,
\end{aligned}$$

where $C:=1+\frac{1}{\eta}\left(\frac{1}{\beta}+\frac{1}{\gamma}\right)$. This is the multivariate negative binomial (MVNB) distribution [28]. For the sake of completeness, we show that the marginal distributions take the expected negative binomial form:

$$
\begin{aligned}
P(n;\alpha,\eta) &= \int_0^\infty P(n;x) f(x;\alpha,\eta) dx \\
&= \int_0^\infty \left(\frac{(x/\beta)^n e^{-x/\beta}}{n!} \right) \frac{\eta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\eta x} dx \\
&= \frac{\eta^\alpha}{\Gamma(\alpha) n! \beta^n} \int_0^\infty x^{n-\alpha-1} e^{-x\left(\eta+\frac{1}{\beta}\right)} dx \\
&= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha) n!} \left(\frac{\eta}{\eta+\frac{1}{\beta}} \right)^\alpha \left(\frac{1}{\beta\left(\eta+\frac{1}{\beta}\right)} \right)^n \\
&= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha) n!} \left(\frac{1}{C_N} \right)^\alpha \left(\frac{\frac{1}{\eta\beta}}{C_N} \right)^n ; \\
P(m;\alpha,\eta) &= \int_0^\infty P(m;x) f(x;\alpha,\eta) dx \\
&= \int_0^\infty \left(\frac{(x/\gamma)^m e^{-x/\gamma}}{m!} \right) \frac{\eta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\eta x} dx \\
&= \frac{\eta^\alpha}{\Gamma(\alpha) m! \gamma^m} \int_0^\infty x^{m-\alpha-1} e^{-x\left(\eta+\frac{1}{\gamma}\right)} dx \\
&= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha) m!} \left(\frac{\eta}{\eta+\frac{1}{\gamma}} \right)^\alpha \left(\frac{1}{\gamma\left(\eta+\frac{1}{\gamma}\right)} \right)^m \\
&= \frac{\Gamma(\alpha+m)}{\Gamma(\alpha) m!} \left(\frac{1}{C_M} \right)^\alpha \left(\frac{\frac{1}{\eta\gamma}}{C_M} \right)^m
\end{aligned}
$$

where $C_N := 1 + \frac{1}{\eta\beta}$ and $C_M := 1 + \frac{1}{\eta\gamma}$. The two marginals' NB parameters are $r = \alpha$, $p_N = \frac{1}{\eta\beta+1}$, and $p_M = \frac{1}{\eta\gamma+1}$.

We note that the Poissonian framework due to Jahnke and Huisinga [27] yields the solutions for arbitrary graphs representing sources, sinks, and reaction channels. This is sufficient, for example, to construct a directed acyclic graph representing alternative splicing of a constitutively expressed gene. Adding extrinsic noise to these graphs is trivial and immediately follows from the definitions of the corresponding Poisson rate constants.

4.2.2 Moments

The moments and variances of the marginals follow immediately from standard identities for the NB distribution:

$$\mu = \frac{rp}{(1-p)}$$
$$\mu_N = \frac{\alpha\frac{1}{\eta\beta+1}}{\frac{\eta\beta}{\eta\beta+1}} = \frac{\alpha}{\eta\beta}$$
$$\mu_M = \frac{\alpha\frac{1}{\eta\gamma+1}}{\frac{\eta\gamma}{\eta\gamma+1}} = \frac{\alpha}{\eta\gamma}$$
$$\sigma^2 = \frac{\mu}{1-p}$$
$$\sigma_N^2 = \frac{\mu_N}{1-p_N} = \frac{\alpha}{\eta\beta}\frac{\eta\beta+1}{\eta\beta} = \frac{\alpha(\eta\beta+1)}{(\eta\beta)^2}$$
$$\sigma_M^2 = \frac{\mu_M}{1-p_M} = \frac{\alpha}{\eta\gamma}\frac{\eta\gamma+1}{\eta\gamma} = \frac{\alpha(\eta\gamma+1)}{(\eta\gamma)^2}$$

The moment-generating function (MGF) of the MVNB distribution is $\phi(x,y) = \left(C - \frac{e^x}{\eta\beta} - \frac{e^y}{\eta\gamma}\right)^{-\alpha}$ [28]. Differentiating the expression with respect to x and y yields $\frac{\alpha(\alpha+1)}{\eta^2\beta\gamma}\left(C - \frac{e^x}{\eta\beta} - \frac{e^y}{\eta\gamma}\right)^{-\alpha-2}$. Evaluating at $x = y = 1$ yields the cross moment $\mathbb{E}[NM] = \frac{\alpha^2+\alpha}{\eta^2\beta\gamma}$. Therefore, the covariance is $Cov(N,M) = \mathbb{E}[NM] - \mu_N\mu_M = \frac{\alpha^2+\alpha}{\eta^2\beta\gamma} - \frac{\alpha^2}{\eta^2\beta\gamma} = \frac{\alpha}{\eta^2\beta\gamma}$. This result yields the following Pearson correlation coefficient:

$$\rho := \frac{Cov(N,M)}{\sigma_N\sigma_M}$$
$$= \frac{\frac{\alpha}{\eta^2\beta\gamma}}{\sqrt{\frac{\alpha(\eta\beta+1)}{(\eta\beta)^2}\frac{\alpha(\eta\gamma+1)}{(\eta\gamma)^2}}}$$
$$= \frac{1}{\sqrt{(\eta\gamma+1)(\eta\beta+1)}}$$

5 Discriminating between intrinsic and extrinsic noise models

Using the above computations, we can show that steady-state information about the nascent and mature distributions is sufficient to distinguish between the two models. We start from an *a priori* non-identifiable negative binomial nascent mRNA distribution, and demonstrate disagreement between statistics predicted for the mature mRNA. For the purposes of illustration, we assume data-based constraints upon μ_M and upon σ_M^2, motivated by the existence of experimental methods for determining these quantities [1, 29]. However, we restrict our analysis to analytical distributions to avoid the details of any particular observation or statistical inference method.

The intrinsic and extrinsic noise models are respectively parametrized by $\{b, k_i, \beta, \gamma\}$ and $\{\alpha, \eta, \beta, \gamma\}$. However, at steady state, the time variable is not independently identifiable. Therefore, the absolute scaling of the rate variables $\{k_i, \eta, \beta, \gamma\}$ is not feasible to determine. This non-identifiability is self-evident from the functional forms of the distributions, e.g. the moment dependence on $\frac{k_i}{\beta}$ and $\frac{k_i}{\gamma}$ in the intrinsic noise model and on $\eta\beta$ and $\eta\gamma$ in the extrinsic noise model. Therefore, we set β to 1 with no loss of generality.

Further, the nascent marginal is governed by the two-parameter NB distribution. In the context of model fitting, this implies that two of the parameters of the joint distribution are fully determined by the nascent distribution, and only one degree of freedom remains to be determined by the mature mRNA data.

Crucially, given a negative binomial distribution of nascent mRNA, with mean μ_N and variance σ_N^2, the intrinsic and extrinsic noise models are not distinguishable. Using the intrinsic noise model uniquely identifies $b = \frac{\sigma_N^2}{\mu_N} - 1$ and $k_i = \frac{\mu_N}{b}$. Conversely, using the extrinsic noise model uniquely identifies $\eta = \left(\frac{\sigma_N^2}{\mu_N} - 1\right)^{-1}$ and $\alpha = \mu_N \eta$.

5.1 Case of constrained γ or μ_M

Constraining γ is equivalent to fixing the mature mRNA means:

$$\mu_{M,i} = \frac{k_i b}{\gamma} = \frac{\mu_N}{\gamma}$$
$$\mu_{M,e} = \frac{\alpha}{\eta\gamma} = \frac{\mu_N}{\gamma}$$
$$\frac{\mu_{M,i}}{\mu_{M,e}} = 1$$

However, the higher moments disagree. Defining the statistic $q := \frac{\sigma_N^2}{\mu_N} - 1 > 0$ and recalling that $f = \frac{1}{1+\gamma}$ for $\beta = 1$:

$$\sigma_{M,i}^2 = \mu_M\left(1 + \frac{b}{1+\gamma}\right) = \frac{\mu_N}{\gamma}\left(1 + \frac{q}{1+\gamma}\right) = \frac{\mu_N}{\gamma}\left(\frac{1+\gamma+q}{1+\gamma}\right)$$
$$\sigma_{M,e}^2 = \mu_M \frac{\eta\gamma+1}{\eta\gamma} = \frac{\mu_N}{\gamma}\frac{q^{-1}\gamma+1}{q^{-1}\gamma}$$
$$\frac{\sigma_{M,i}^2}{\sigma_{M,e}^2} = \frac{1+\gamma+q}{1+\gamma}\frac{q^{-1}\gamma}{q^{-1}\gamma+1} = \frac{(1+\gamma)q^{-1}\gamma+\gamma}{(1+\gamma)q^{-1}\gamma+\gamma+1} < 1$$
$$\rho_i^2 = b^2\frac{f(1-f)}{(1+b)(1+bf)} = q^2\frac{f(1-f)}{(1+q)(1+qf)}$$
$$\rho_e^2 = \frac{1}{(\eta\gamma+1)(\eta+1)} = \frac{1}{(q^{-1}\gamma+1)(q^{-1}+1)}$$
$$\frac{\rho_i^2}{\rho_e^2} = q^2\frac{f(1-f)(q^{-1}\gamma+1)(q^{-1}+1)}{(1+q)(1+qf)} = \frac{f(1-f)(\gamma+q)(1+q)}{(1+q)(1+qf)} = \frac{\gamma}{(1+\gamma)^2}\frac{\gamma+q}{1+\frac{q}{1+\gamma}}$$
$$= \frac{\gamma(\gamma+q)}{(1+\gamma)(1+\gamma+q)} < \frac{(1+\gamma)(1+\gamma+q)}{(1+\gamma)(1+\gamma+q)} = 1$$

Therefore, the extrinsic noise model is overdispersed with respect to the intrinsic noise model, but its nascent and mature copy numbers are more highly correlated.

5.2 Case of constrained σ_M^2

Using these expressions, it is straightforward to extend the analysis to the scenario of fixing σ_M^2:

$$\sigma^2_{M,i} = \frac{\mu_N}{\gamma_i}\left(1 + \frac{q}{1+\gamma_i}\right)$$
$$\sigma^2_{M,e} = \frac{\mu_N}{\gamma_e}\left(1 + \frac{1}{q^{-1}\gamma_e}\right) = \frac{\mu_N}{\gamma_e}\left(1 + \frac{q}{\gamma_e}\right)$$

The physical solutions for γ_i and γ_e are given by positive roots of quadratic equations. For the intrinsic noise model:

$$\sigma^2_{M,i} = \frac{\mu_N}{\gamma_i}\left(1 + \frac{q}{1+\gamma_i}\right)$$
$$\gamma_i(1+\gamma_i)\sigma^2_{M,i} = \mu_N(1+\gamma_i) + \mu_N q$$
$$\sigma^2_{M,i}\gamma_i^2 + (\sigma^2_{M,i} - \mu_N)\gamma_i - \mu_N(q+1) = 0$$
$$\gamma_i = \frac{1}{2\sigma^2_{M,i}}\left[(\mu_N - \sigma^2_{M,i}) \pm \sqrt{(\mu_N - \sigma^2_{M,i})^2 + 4\sigma^2_{M,i}\mu_N(q+1)}\right]$$

Since $4\sigma^2_{M,i}\mu_N(q+1) > 0$, the physical solution ($\gamma_i > 0$) is given by:

$$\gamma_i = \frac{1}{2\sigma^2_{M,i}}\left[(\mu_N - \sigma^2_{M,i}) + \sqrt{(\mu_N - \sigma^2_{M,i})^2 + 4\sigma^2_{M,i}\mu_N(q+1)}\right]$$

Further, for the extrinsic noise model:

$$\sigma^2_{M,e} = \frac{\mu_N}{\gamma_e}\left(1 + \frac{q}{\gamma_e}\right)$$
$$\gamma_e^2\sigma^2_{M,e} - \mu_N\gamma_e - \mu_N q = 0$$
$$\gamma_e = \frac{1}{2\sigma^2_{M,e}}\left[\mu_N \pm \sqrt{\mu_N^2 + 4\sigma^2_{M,e}\mu_N q}\right]$$

Again, since $4\sigma^2_{M,e}\mu_N q > 0$, the physical solution ($\gamma_e > 0$) is given by:

$$\gamma_e = \frac{1}{2\sigma^2_{M,e}}\left[\mu_N + \sqrt{\mu_N^2 + 4\sigma^2_{M,e}\mu_N q}\right]$$

Imposing equal variances:

$$\sigma^2_{M,e} = \sigma^2_{M,i} = \sigma^2_M \implies$$
$$\gamma_i = \frac{1}{2\sigma^2_M}\left[(\mu_N - \sigma^2_M) + \sqrt{(\mu_N - \sigma^2_M)^2 + 4\sigma^2_M\mu_N(q+1)}\right]$$
$$\gamma_e = \frac{1}{2\sigma^2_M}\left[\mu_N + \sqrt{\mu_N^2 + 4\sigma^2_M\mu_N q}\right]$$

Consider $\chi := \frac{\sigma_M^2}{\mu_N}$. This definition yields:

$$\gamma_i = \frac{1}{2\chi}(1 - \chi + \sqrt{(1-\chi)^2 + 4\chi(q+1)})$$
$$\gamma_e = \frac{1}{2\chi}(1 + \sqrt{1 + 4\chi q})$$
$$\frac{\gamma_i}{\gamma_e} = \frac{1 - \chi + \sqrt{(1-\chi)^2 + 4\chi(q+1)}}{1 + \sqrt{1+4\chi q}}$$
$$= \frac{1 - \chi + \sqrt{1 + \chi^2 - 2\chi + 4\chi + 4\chi q}}{1 + \sqrt{1+4\chi q}}$$
$$= \frac{1 - \chi + \sqrt{(1+\chi)^2 + 4\chi q}}{1 + \sqrt{1+4\chi q}}$$

We may investigate the case where this quantity is equal to 1:

$$1 + \sqrt{1 + 4\chi q} = 1 - \chi + \sqrt{(1+\chi)^2 + 4\chi q}$$
$$\sqrt{1 + 4\chi q} = \sqrt{(1+\chi)^2 + 4\chi q} - \chi$$

No values of $q, \chi > 0$ yield this equality. This is straightforward because even the more general equation $\sqrt{1+C} = \sqrt{(1+x)^2 + C} - x$ is nowhere satisfied for $x, C > 0$. Therefore, $\frac{\gamma_i}{\gamma_e}$ is never 1. From the quadratic equation solution, we know that γ_e and γ_i are both constrained to be positive; therefore, $\frac{\gamma_i}{\gamma_e} > 0$. Using the test case $\chi = q = 1$, we yield $\frac{\gamma_i}{\gamma_e} = \frac{\sqrt{4+4}}{1+\sqrt{1+4}} \approx 0.87 < 1$. Since γ_e is nonzero and $\gamma_i(\chi, q)$ is continuous with respect to both variables, $\frac{\gamma_i}{\gamma_e}(\chi, q)$ is continuous. Finally, we conclude that $\frac{\gamma_i}{\gamma_e}(\chi, q)$ is always constrained to $(0, 1)$ and $\gamma_e > \gamma_i$ whenever σ_M^2 is fixed. This result matches the intuition of the provided by the finding that $\sigma_{M,e}^2 > \sigma_{M,i}^2$ whenever γ is fixed: to compensate for increased dispersion in the extrinsic noise model, the degradation rate must be increased. It trivially follows that $\mu_{M,i} > \mu_{M,e}$.

6 Experimental opportunities and limitations

Multiple experimental approaches are available for the collection of nascent and mature mRNA data. We focus on the most prevalent technologies and their relevance to the modeling question at hand.

Fluorescence microscopy methods are broadly divided between spatial transcriptomics and intron counting. Spatial transcriptomics leverages relative positions of fluorescently-labeled mRNA and DNA to identify DNA-localized nascent mRNA [22, 29]. Intron counting directly detects intron-targeted fluorescent probes [23]. These methods are rather complex and impractical to perform on a genome-wide scale. Furthermore, we are unaware of any studies combining them with dual-reporter assays to directly estimate intrinsic and extrinsic noise. Finally, the discrimination of nascent and mature mRNA aside, dual-reporter assays are in general impractical to scale to large numbers of genes.

Sequencing methods are are broadly divided between labeling and bioinformatics. Labeling refers to spiking the live media with a nucleoside analogue and distinguishing older and newer mRNA

molecules based on characteristic mutations [30–33]. Purely computational methods do not require labeling, but identify nascent mRNA based on intron-aligned reads [24, 34]. These methods yield genome-wide information; however, they are not amenable to reporter duplication on the same scale. Commercially-available short-read methods present the problem of isoform indistinguishability if introns interest are outside the read region [24]. Finally, both short- and long-read methods tend to rely on the capture of polyadenylated tails [7, 35, 36], which are not present in nascent mRNA, introducing the potential of technical bias against the nascent molecules of interest. Off-target priming at intronic polyadenine sites [24, 37] and experimental methods including poly(A) ligation [31] facilitate the capture and identification of nascent transcripts, but the magnitude of technical biases is as of yet uncharacterized.

Parenthetically, we note that the motivating study by Ham et al. [13] describes a purely data-based approach to the identification of extrinsic effects, based upon the identification of heavy distribution tails. This approach appears to be quite powerful based on the provided demonstration. However, certain aspects are potentially problematic. The validation compares the tail behavior of the telegraph model to the compound telegraph model. However, even relatively simple telegraph models suffer from parameter non-identifiability issues [38, 39], so the robustness of the method is unclear. The specific fit method and metric are not reported; it is not clear that the conventional choices are appropriate when tail behavior is significant. Recent work in extreme value theory proposes several Rényi divergence alternatives. [40]. Finally, we note that the underlying data is from Zheng et al. [7], which is the earliest version of the 10X Genomics single-cell RNA sequencing platform. Since the underlying mammalian physiology has export and splicing processes [41], but 10X sequencing explicitly focuses on exonic reads [7], it is unclear that the choice of a one-stage model is justified. More problematically, raw count data are rarely used in scRNA-seq analyses [42], with substantial debate and disagreement regarding the appropriate approach to normalization [43–46]. Therefore, it is conceivable that technical biases may, in part, explain the 15-25 cells with extremely high expression that control the kernel density in the tail region, used to support the hypothesis of extrinsic noise.

7 Discussion

In spite of the indistinguishability of negative binomial distributions produced by intrinsic and extrinsic noise, the behavior of downstream processed gene products is substantially divergent. Specifically, we report inequality between the two models' lower moments. Even given identical nascent marginals, it is impossible to produce identical mature marginals, and by extension full joint distributions, using the two models. More dramatically, it is impossible for the solutions' mature marginals even to share more than one low-order moment.

In practice, if experimental joint or marginal copy-number distributions are available, it is possible to use relative likelihood testing to choose the better-fitting model. The relevant test statistic is $\lambda = \frac{\mathcal{L}_i(\hat{\Theta}_i)}{\mathcal{L}_e(\hat{\Theta}_e)}$, where $\mathcal{L}_z(\hat{\Theta}_z)$, $z \in \{i, e\}$ is the value of the likelihood function of the intrinsic or extrinsic model at the maximum likelihood joint parameter estimate. No closed-form joint maximum likelihood estimators are available for either model; however, estimation by numerical optimization is straightforward, especially starting at the moment-based estimates reported above. The two models' qualitative behaviors are illustrated in Figure 2. We use the Gillespie algorithm [47] to simulate both systems given identical nascent distributions ($r = 1.8$, $p = \frac{12}{13}$) and downstream processing rates ($\beta = 0.5$, $\gamma = 0.4$). The solutions are dramatically different. As expected from

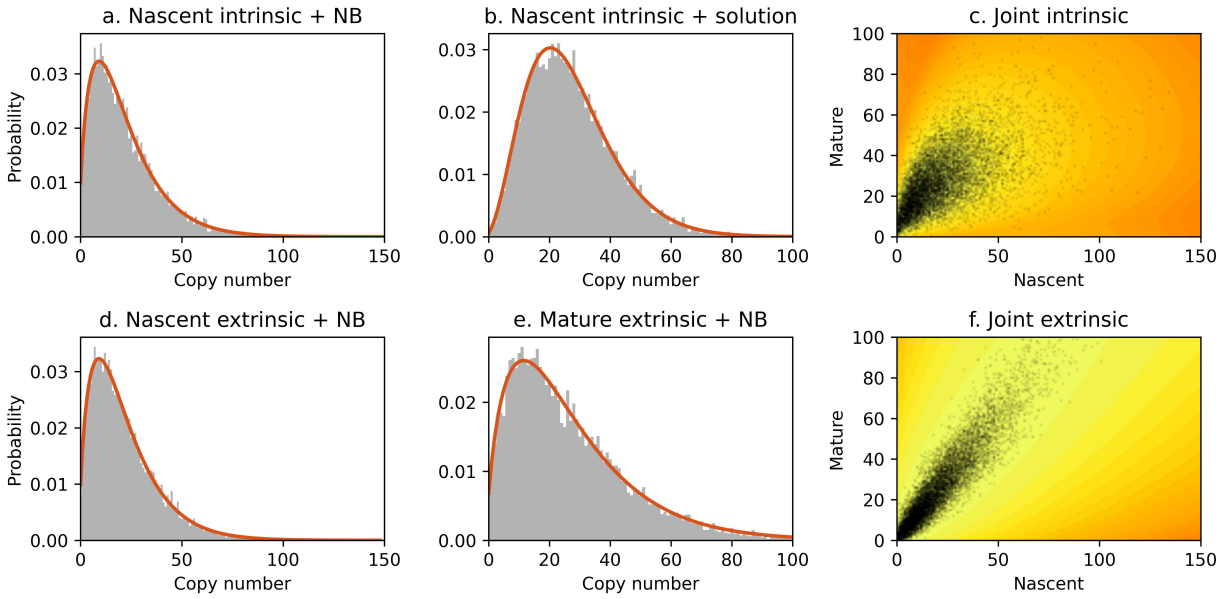

Figure 2: (a-c) Distributions for the intrinsic noise model with $b = 12$, $k_i = 0.9$, $\beta = 0.5$, and $\gamma = 0.4$, generated using 10^4 simulations. (a) Nascent marginal (gray region: copy-number histogram; orange line: analytical solution). (b) Mature marginal (gray region: copy-number histogram; orange line: analytical solution). (c) Joint distribution (points: cells; color: $\log_{10}$ analytical solution, lighter color corresponds to higher probability mass). (d-f) Distributions for the extrinsic noise model with $\alpha = k_i\beta^{-1}$, $\eta = (b\beta)^{-1}$, $\beta = 0.5$, and $\gamma = 0.4$, generated using 10^4 simulations. (d) Nascent marginal (gray region: copy-number histogram ; orange line: analytical solution). (e) Mature marginal (gray region: copy-number histogram; orange line: analytical solution). (f) Joint distribution (points: cells; color: $\log_{10}$ analytical solution, lighter color corresponds to higher probability mass).

the analytical moments, the extrinsic noise model gives a much more correlated joint distribution. However, despite identical marginal mature expectations, the extrinsic model has a longer tail, yielding a higher variance for that species. This drastic disagreement between distributions confirms that multimodal data is sufficient to distinguish between the two hypothesized sources of stochasticity. In addition to the theoretical and qualitative results, we provide simulation routines for both noise models. Furthermore, to facilitate comparison with discrete copy-number data, we report analytical marginal and joint distributions implied by the formulation of the system with extrinsic noise; their agreement with the simulation is shown in Figure 2. These distributions, along with moment-based initial parameter estimates, can be directly used for inference and hypothesis testing against other models.

Given the modeling-based insight into model identifiability, we suggest that multimodal data collection presents a valuable route to the identification of noise models. Specifically, we anticipate increased relevance for single-cell RNA sequencing, which has been challenging to integrate with experimental controls for the noise sources. Therefore, we suggest that experimental improvements in the detection of the nascent transcriptome, as well as theoretical improvements in the modeling of technical noise, would allow identification of sources of biological stochasticity on a genome-wide

scale. Finally, the discrete modeling framework we discuss is immediately interpretable in terms of biophysical parameters.

8 Code Availability

MATLAB and Python code that can be used to reproduce Figure 2, including the simulation and plotting routines, is available at `https://github.com/pachterlab/GP_2020_2`.

9 Acknowledgments

The DNA, pre-mRNA, and mature mRNA illustrations used in Figure 1, modified from [26], are derivatives of the DNA Twemoji by Twitter, Inc., used under CC-BY 4.0. G.G. and L.P. are partially funded by NIH U19MH114830.

References

[1] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186, 2002.

[2] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, October 2002.

[3] A. Hilfinger and J. Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–12172, July 2011.

[4] Marc S. Sherman, Kim Lorenz, M. Hunter Lanier, and Barak A. Cohen. Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. *Cell systems*, 1(5):315–325, November 2015.

[5] Erik van Nimwegen. Inferring intrinsic and extrinsic noise from a dual fluorescent reporter. Preprint, bioRxiv: 049486, April 2016.

[6] Audrey Qiuyan Fu and Lior Pachter. Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. *Statistical Applications in Genetics and Molecular Biology*, 15(6), January 2016.

[7] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, April 2017.

[8] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4):631–643.e4, February 2017.

[9] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, 115(29):7533–7538, 2018.

[10] Christoph Zechner and Heinz Koeppl. Uncoupled Analysis of Stochastic Reaction Networks in Fluctuating Environments. *PLoS Computational Biology*, 10(12):e1003942, December 2014.

[11] V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, November 2008.

[12] Johan Paulsson and Måns Ehrenberg. Random Signal Fluctuations Can Reduce Random Fluctuations in Regulated Components of Chemical Regulatory Networks. *Physical Review Letters*, 84(23):5447–5450, June 2000.

[13] Lucy Ham, Rowan D. Brackston, and Michael P.H. Stumpf. Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Physical Review Letters*, 124(10):108101, March 2020.

[14] R. D. Dar, B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, October 2012.

[15] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty Gene Expression in the Intact Mammalian Liver. *Molecular Cell*, 58(1):147–156, April 2015.

[16] Damien Nicolas, Nick E. Phillips, and Felix Naef. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, 13(7):1280–1290, 2017.

[17] Anton J. M. Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R. Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, January 2019.

[18] A. Klindziuk, B. Meadowcroft, and A. B. Kolomeisky. Mechanochemical Model of Transcriptional Bursting. Preprint, bioRxiv: 802751, October 2019.

[19] Bhaswati Bhattacharyya, Jin Wang, and Masaki Sasai. Stochastic Epigenetic Dynamics of Gene Switching. Preprint, bioRxiv: 2020.03.18.996819, March 2020.

[20] Abhyudai Singh and Pavol Bokes. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal*, 103(5):1087–1096, September 2012.

[21] Qianliang Wang and Tianshou Zhou. Alternative-splicing-mediated gene expression. *Physical Review E*, 89(1):012713, January 2014.

[22] Mengyu Wang, Jing Zhang, Heng Xu, and Ido Golding. Measuring transcription at a single gene copy reveals hidden drivers of bacterial individuality. *Nature Microbiology*, 4:2118–2127, September 2019.

[23] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, Christopher Cronin, Christoph Karp, Eric J. Liaw, Mina Amin, and Long Cai. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell*, 174(2):363–376.e16, July 2018.

[24] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018.

[25] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036, December 2005.

[26] Gennady Gorin and Lior Pachter. Special function methods for bursty models of transcription. *Physical Review E*, 102(2):022409, August 2020.

[27] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26, December 2006.

[28] E. Al-Hussaini and K. Ahmad. On the identifiability of finite mixtures of distributions (Corresp.). *IEEE Transactions on Information Theory*, 27(5):664–668, September 1981. Conference Name: IEEE Transactions on Information Theory.

[29] Heng Xu, Samuel O. Skinner, Anna Marie Sokac, and Ido Golding. Stochastic Kinetics of Nascent RNA. *Physical Review Letters*, 117(12):128101, 2016.

[30] Erin M. Wissink, Anniina Vihervaara, Nathaniel D. Tippens, and John T. Lis. Nascent RNA analyses: tracking transcription and its regulation. *Nature Reviews Genetics*, 20:705–723, August 2019.

[31] Heather L. Drexler, Karine Choquet, and L. Stirling Churchman. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Molecular Cell*, 77(5):985–998.e8, March 2020.

[32] Junyue Cao, Wei Zhou, Frank Steemers, Cole Trapnell, and Jay Shendure. Characterizing the temporal dynamics of gene expression in single cells with sci-fate. Preprint, bioRxiv: 666081, June 2019.

[33] Qi Qiu, Peng Hu, Xiaojie Qiu, Kiya W. Govek, Pablo G. Cámara, and Hao Wu. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods*, August 2020.

[34] Páll Melsted, A. Sina Booeshaghi, Fan Gao, Eduardo Beltrame, Lambda Lu, Kristján Eldjárn Hjorleifsson, Jase Gehring, and Lior Pachter. Modular and efficient pre-processing of single-cell RNA-seq. Preprint, bioRxiv: 673285, June 2019.

[35] Seyed Yahya Anvar, Guy Allard, Elizabeth Tseng, Gloria M. Sheynkman, Eleonora de Klerk, Martijn Vermaat, Raymund H. Yin, Hans E. Johansson, Yavuz Ariyurek, Johan T. den Dunnen, Stephen W. Turner, and Peter A. C. 't Hoen. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biology*, 19(1):46, December 2018.

[36] Tetsutaro Hayashi, Haruka Ozaki, Yohei Sasagawa, Mana Umeda, Hiroki Danno, and Itoshi Nikaido. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nature Communications*, 9(1):619, December 2018.

[37] D. K. Nam, S. Lee, G. Zhou, X. Cao, C. Wang, T. Clark, J. Chen, J. D. Rowley, and S. M. Wang. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences*, 99(9):6152–6156, April 2002.

[38] Eugenio Cinquemani. Identifiability and Reconstruction of Biochemical Reaction Networks from Population Snapshot Data. *Processes*, 6(9):136, August 2018.

[39] Suresh Kumar Poovathingal and Rudiyanto Gunawan. Global parameter estimation methods for stochastic biochemical systems. *BMC Bioinformatics*, 11(1):414, 2010.

[40] Jose Blanchet, Fei He, and Karthyek R. A. Murthy. On distributionally robust extreme value analysis. Preprint, arXiv: 1601.06858, June 2020.

[41] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of Transcript Variability in Single Mammalian Cells. *Cell*, 163(7):1596–1610, December 2015.

[42] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019.

[43] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, October 2018.

[44] Peng Qiu. Embracing the dropouts in single-cell RNA-seq data. Preprint, bioRxiv: 468025, November 2018.

[45] Tallulah Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7:1740, 2019.

[46] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, February 2020.

[47] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976.