From Noise to Models to Numbers: Evaluating Negative Binomial Models and Parameter Estimations in Single-Cell RNA-seq

Yiling Wang*[1], Zhanpeng Shu*[2], Zhixing Cao†[1,3], and Ramon Grima‡[4]

[1] *State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, 200237, China*
[2] *School of Electrical Engineering, Shanghai Dianji University, Shanghai, 201308, China*
[3] *Department of Chemical Engineering, Queen's University, Kingston K7L 3N6, Canada*
[4] *School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, UK*

Abstract

The Negative Binomial (NB) distribution effectively approximates the transcript count distribution in many single-cell RNA sequencing (scRNA-seq) datasets. This has led to its widespread use in various computational tools for scRNA-seq analysis. However, the underlying reasons for its ubiquitousness remain unclear. Here, we use a computationally efficient model selection technique to precisely map the relationship between the choice of the best-fit models — Beta-Poisson (Telegraph), NB and Poisson — and the kinetic parameters that control the stochasticity of gene expression. We find that the NB distribution is an excellent approximation to simulated data, that accounts for both biological and technical noise, in an intermediate range of an effective parameter — the sum of the gene activation and inactivation rates normalized by the mRNA degradation rate. The size of this range increases with decreasing mean expression, increasing technical noise, and increasing sample size (number of cells). These findings have important implications: (i) excellent NB fits span diverse parameter regimes and are not exclusive indicators of transcriptional bursting; (ii) for small sample sizes, biological noise generally becomes the primary factor shaping the NB characteristics of the count distribution, even when technical noise is significant; (iii) under the assumption of steady-state conditions, gene-specific parameters (burst size and frequency) estimated in regions where the NB model fits well, typically show large relative errors, even after corrections for technical noise; (iv) gene ranking by burst frequency remains accurate, indicating that burst parameter magnitudes are often only relatively informative.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) facilitates the quantitative characterization of cellular heterogeneity by providing genome-wide transcriptomic profiles at single-cell resolution across hundreds or thousands of individual cells [1–4]. The use of unique molecular identifiers (UMIs) has greatly improved the reliability of sequencing data by reducing (though not eliminating) artifacts introduced during sample processing, leading to molecular counts that better reflect the true biological abundance [5]. The transcript count distribution of most genes is unimodal and well fit by a negative binomial (NB) distribution (also called a Gamma-Poisson distribution) [6–11]. The significance of the NB distribution is highlighted by its extensive application in various computational tools for scRNA-seq analysis [10, 12–19]. The fitting of the NB distribution to data yields two parameters for each gene, providing a convenient numerical signature of its observed stochasticity. However, the origin of the universality of the NB distribution in single-cell sequencing data remains unclear. In fact, this distribution is also commonly found using single-molecule fluorescence in situ hybridization (smFISH), suggesting that its ubiquity cannot be simply dismissed as stemming from the significant technical noise of sequencing technologies [20].

Over the past two decades, using fluorescence-based single-cell transcriptomics, it has been shown that biological noise for each gene is well described by the two-state telegraph model [21–27] or by a three or more-state extension of this model [28–30]. Here, we focus on the telegraph model because it has been shown to very well approximate the steady-state distribution of mRNA counts predicted by models with a higher number of states [31–33]. The telegraph model describes a gene that switches between an active state (G) and an inactive state (G^*). Transcription occurs

*Equally contributed.
†z.cao@queensu.ca
‡ramon.grima@ed.ac.uk

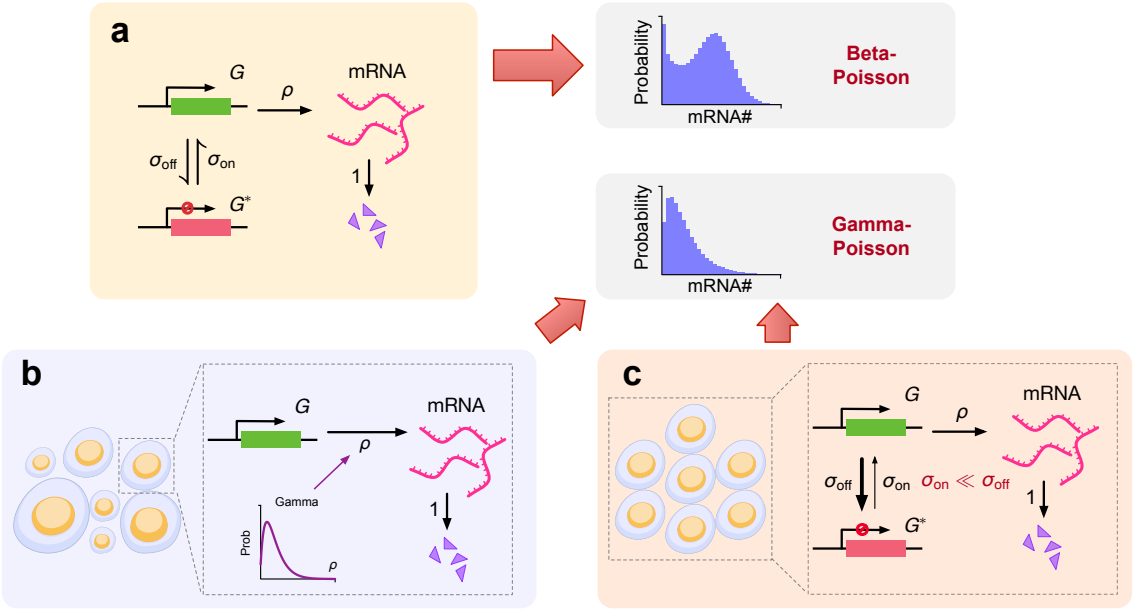

Figure 1: (a) Schematic illustrating the telegraph model of gene expression. A gene switches between active (green) and inactive states (red) with rates σ_{on} and σ_{off}. Synthesis of transcripts occurs from the active state with rate ρ. The transcripts are subsequently degraded with rate 1. The rates are all normalized by the degradation rate. The steady-state distribution of transcript numbers is a Beta-Poisson compound distribution; (b) Schematic showing the special case where the gene is always in the active state and the transcription rate ρ varies from cell to cell according to a Gamma distribution. In this case, the mRNA distribution predicted by the telegraph model reduces to a Gamma-Poisson compound distribution (an NB distribution); (c) Schematic showing the special case where the gene spends most of its time in the inactive state ($\sigma_{\text{on}} \ll \sigma_{\text{off}}$) which leads to transcription occurring in short-lived bursts that are well separated from each other. All cells are identical, i.e. the rate constants do not vary from cell to cell. In this case, the mRNA distribution predicted by the telegraph model also reduces to an NB distribution.

only from the active state G, and subsequently mRNA decay occurs with first-order kinetics. For an illustration, see Fig. 1a. The solution of the chemical master equation that describes the telegraph model in steady-state conditions leads to a Beta-Poisson distribution of mRNA counts [34–36]. Typically, scRNA-seq measurements detect a significantly smaller fraction of transcripts than smFISH and hence extending the telegraph model to also account for technical noise is important. It has been shown that even after this modification, the steady-state distribution of transcript counts in a single cell follows a Beta-Poisson distribution [37, 38] — technical noise simply leads to a rescaling of the effective transcription rate by the probability of transcript capture in the cell, but does not change the type of distribution. Depending on the parameter values, the Beta-Poisson distribution can be unimodal with a peak at zero, unimodal with a peak at a non-zero value, or bimodal with peaks at zero and non-zero values [39]. All of these distribution shapes have been measured using smFISH [20, 40, 41]. Hence, clearly, the classical model of stochastic gene expression, even when extended to account for technical noise, does not generally predict an NB distribution. There are two known conditions under which the Beta-Poisson distribution reduces to an NB distribution.

Case (i). Say that a gene is always in the active state, i.e. there is no gene state switching. The telegraph model predicts that, in each cell, the transcript counts are sampled from a Poisson distribution with a parameter given by the effective transcription rate divided by the degradation rate [42]. If the effective transcription rate, which is equal to the product of the transcription rate and the probability of transcript capture, is the same in each cell, then the distribution of transcript counts across all cells is necessarily Poisson. However, if the effective transcription rate varies from cell to cell according to a Gamma distribution, then it immediately follows that the observed distribution is a Gamma-Poisson compound distribution, which is an NB distribution. Hence, in this case, the NB character of the distribution stems from extrinsic (cell-to-cell variation in the transcription rate) or technical noise (cell-to-cell variation in the transcript capture probability), and not from the intrinsic dynamics of a gene. For an illustration, see Fig. 1b. This case is unlikely to be the main reason for the universality of the NB in scRNA-seq data because the visualization of transcription in living cells using live-cell imaging conclusively shows alternating periods of gene activity and inactivity [43]. This is due to many factors such as reversible binding of transcription factors, enhancer-promoter interactions, the clustering dynamics of Pol II and the opening and closing of chromatin [44–48].

Case (ii). Now consider the case where the gene spends most of its time in the inactive state and synthesizes

mRNA in a burst, i.e. in the short time that the gene is active. This occurs when the gene inactivation rate is much larger than the gene activation rate. This phenomenon is often referred to as transcriptional bursting [36]; for an illustration, see Fig. 1c. If the effective transcription rate is the same in each cell, then the Beta-Poisson distribution of the telegraph model has been shown to reduce to an NB distribution [49]. This case has sometimes been used to explain how the NB distribution of mRNA counts in sequencing data can arise from a physical model of gene expression [11]. However, this is also unlikely to explain the universality of the NB distribution in scRNA-seq data because the constraint that the activation rate is much lower than the inactivation rate represents a very small part of the available parameter space. Also because the assumption that the effective transcription rate is the same in all cells implies negligible variation in the probability of mRNA capture from one cell to another — which is difficult to reconcile with the wide distribution of total UMI counts (from all genes) per cell in typical scRNA-seq datasets [17, 50, 51].

In summary, there is no strong theoretical reason behind the universality of the NB distribution in scRNA-seq data. The two existing results in the literature on stochastic gene expression can only explain the NB distribution by invoking strong assumptions that are not realistic. Furthermore, it remains unclear what biological or biophysical meaning is to be imparted to the two parameters of the NB distribution. Currently, this meaning is only clear when gene expression is bursty (as in Case (ii)), in which case after correcting for technical noise, the two parameters can be used to extract the burst frequency (the rate at which mRNA bursts are transcribed) and the burst size (the mean number of mRNA produced when the gene is active) [52].

In this paper, we overcome these limitations by deriving new parametric conditions under which the observed distribution of transcript counts is well approximated by an NB distribution. We find that these conditions include, but are not limited to, transcriptional bursting. We also show that while the absolute values of the burst frequency and the burst size cannot generally be reliably extracted from the two parameters of the NB distribution, nevertheless these still contain useful information about differential gene expression.

2 Results

2.1 The NB distribution can provide a good approximation to the mRNA distribution of the telegraph model even when transcriptional bursting is absent

Consider an idealized scenario where each cell is identical to each other, i.e. for a given gene, there is no cell-to-cell variation in the parameters of the telegraph model (Fig. 1a). This implies that all cells are of the same type. Furthermore, assume that all transcripts from each cell can be detected. In this case, the distribution of mRNA counts is given by the steady-state solution of the chemical master equation of the telegraph model:

$$P(n) = \frac{\rho^n}{n!} \frac{(\sigma_{\text{on}})_n}{(\sigma_{\text{on}} + \sigma_{\text{off}})_n} {}_1F_1(\sigma_{\text{on}} + n, \sigma_{\text{on}} + \sigma_{\text{off}} + n, -\rho), \tag{1}$$

where n represents the mRNA count, $(\bullet)_n$ denotes the Pochhammer symbol, and ${}_1F_1$ is the Kummer confluent hypergeometric function [34]. Note that σ_{on} is the gene activation rate, σ_{off} is the gene deactivation rate, ρ is the transcription rate; these rates are non-dimensional because they are normalized by the degradation rate (this convention will be used throughout the paper). Eq. (1) is equivalent to the compound Beta-Poisson distribution because the transcript numbers are distributed as follows:

$$n \sim \text{Poisson}(\rho x), \quad x \sim \text{Beta}(\sigma_{\text{on}}, \sigma_{\text{off}}). \tag{2}$$

As mentioned previously, it is well known that if gene expression is bursty then Eq. (1) is well approximated by an NB distribution, or its continuous counterpart, the Gamma distribution [36, 49]; related derivations for the distribution of protein numbers can be found in Refs. [53, 54]. Specifically, transcriptional bursting occurs in the limit $\sigma_{\text{off}} \to \infty$ taken such that ρ/σ_{off} remains constant [49, 55]. A simple way to see how this limit leads to an NB distribution makes use of the Beta-Poisson formulation of the telegraph model. From Eq. (2), it follows that

$$\lim_{\sigma_{\text{off}} \to \infty} \text{Poisson}\left(\frac{\rho}{\sigma_{\text{off}}} \sigma_{\text{off}} \text{Beta}(\sigma_{\text{on}}, \sigma_{\text{off}})\right) = \text{Poisson}\left(\frac{\rho}{\sigma_{\text{off}}} \text{Gamma}(\sigma_{\text{on}}, 1)\right) = \text{NB}\left(\sigma_{\text{on}}, \frac{1}{1 + \rho/\sigma_{\text{off}}}\right), \tag{3}$$

where in the second step we used the standard statistical result: $\lim_{\sigma_{\text{off}} \to \infty} \sigma_{\text{off}} \text{Beta}(\sigma_{\text{on}}, \sigma_{\text{off}}) = \text{Gamma}(\sigma_{\text{on}}, 1)$.

It is presently unclear if the converse is true: Does an excellent NB fit to the mRNA count distribution of the telegraph model imply transcriptional bursting?

To answer this question, we proceed as follows. It is straightforward to show using Eq. (1) that the mean and variance of mRNA counts for the telegraph model are given by

$$\langle n\rangle_{\text{tele}} = \frac{\rho\sigma_{\text{on}}}{\sigma_{\text{on}}+\sigma_{\text{off}}}, \qquad \text{Var}_{\text{tele}}(n) = \langle n^2\rangle_{\text{tele}} - \langle n\rangle_{\text{tele}}^2 = \frac{\rho\sigma_{\text{on}}[(\sigma_{\text{on}}+\sigma_{\text{off}})(\sigma_{\text{on}}+\sigma_{\text{off}}+1)+\rho\sigma_{\text{off}}]}{(\sigma_{\text{on}}+\sigma_{\text{off}})^2(\sigma_{\text{on}}+\sigma_{\text{off}}+1)}, \tag{4}$$

where $\langle\bullet\rangle$ is the averaging operator. In contrast, the mean and variance of an NB distribution $\text{NB}(r,p)$ are

$$\langle n\rangle_{\text{NB}} = \frac{r(1-p)}{p}, \qquad \text{Var}_{\text{NB}}(n) = \frac{r(1-p)}{p^2}. \tag{5}$$

By equating $\langle n\rangle_{\text{tele}} = \langle n\rangle_{\text{NB}}$ and $\text{Var}_{\text{tele}}(n) = \text{Var}_{\text{NB}}(n)$ in Eqs. (4) and (5) and solving for r and p, we can construct an NB distribution that has the same first and second moments as the mRNA count distribution of the telegraph model:

$$\text{NB}\left(\frac{\sigma_{\text{on}}(\sigma_{\text{on}}+\sigma_{\text{off}}+1)}{\sigma_{\text{off}}}, \frac{(\sigma_{\text{on}}+\sigma_{\text{off}})(\sigma_{\text{on}}+\sigma_{\text{off}}+1)}{(\sigma_{\text{on}}+\sigma_{\text{off}})(\sigma_{\text{on}}+\sigma_{\text{off}}+1)+\rho\sigma_{\text{off}}}\right). \tag{6}$$

This distribution is hereafter referred to as the effective NB distribution denoted as $\text{NB}(r_{\text{e}}, p_{\text{e}})$. Of course, this distribution and the telegraph model potentially can differ in their third and higher moments, and hence there is no guarantee that the shapes of the two distributions will generally be similar.

In Fig. 2, we compare the distributions given by Eq. (1) and Eq. (6) for four distinct parameter sets. In each case, the effective NB distribution is practically perfectly aligned with the corresponding telegraph distribution. However, note that only one of these four cases (Fig. 2a) corresponds to the classical case of $\sigma_{\text{off}} \gg \sigma_{\text{on}}$. The three counterexamples (Fig. 2b-2d) clearly show that excellent fitting of the telegraph model to an NB distribution does NOT necessarily indicate transcriptional bursting.

2.2 An alternative set of conditions under which the NB distribution provides a good approximation to the telegraph model distribution

We start by reparameterizing the gene switching rates as

$$\sigma_{\text{on}} = f_{\text{on}}N_\sigma, \qquad \sigma_{\text{off}} = (1-f_{\text{on}})N_\sigma, \tag{7}$$

where $N_\sigma = \sigma_{\text{on}} + \sigma_{\text{off}}$ represents the timescale of the transition rates relative to the rate of mRNA degradation and f_{on} is the fraction of time spent in the active state, which is given by

$$f_{\text{on}} = \frac{\sigma_{\text{on}}}{\sigma_{\text{on}}+\sigma_{\text{off}}}. \tag{8}$$

Furthermore, for convenience, we denote the telegraph model distribution, Eq. (1), by $\text{Tele}(\rho, \sigma_{\text{on}}, \sigma_{\text{off}})$. Given these definitions, we state the following theorem:

Theorem 1: In the limit $N_\sigma \to \infty$, the telegraph model distribution $\text{Tele}(\rho, f_{\text{on}}N_\sigma, (1-f_{\text{on}})N_\sigma)$ and the effective NB distribution $\text{NB}(r_{\text{e}}, p_{\text{e}})$ (Eq. (6)) converge to the same distribution.

The key idea behind proving this theorem is to show that the Beta and Gamma distributions converge to the same limiting distribution – specifically, a normal distribution – as $N_\sigma \to \infty$. This then implies a convergence of the Beta-Poisson (telegraph model) distribution to a Gamma-Poisson (NB) distribution. This convergence is not immediately obvious because the Beta distribution is defined on $(0,1)$ while the Gamma distribution is defined on $(0,\infty)$. Details of the proof can be found in Methods Section 4.1.

This proof makes more precise the statement in the caption of SI Fig. 4 of Ref. [56] which states that the mRNA count distribution of the telegraph model is bimodal (and therefore not similar to an NB distribution) when $\sigma_{\text{on}} < 1$ and $\sigma_{\text{off}} < 1$; it is unimodal and resembles an NB distribution when $\sigma_{\text{on}} > 1$ and $\sigma_{\text{off}} > 1$. In particular, Theorem 1 shows that $\sigma_{\text{on}} > 1$ and $\sigma_{\text{off}} > 1$ are not sufficient by itself to guarantee an excellent fit of the NB distribution to that of the telegraph model. We also note that the conditions identified by the theorem are different from those given by the classical result that requires $\sigma_{\text{on}} \ll \sigma_{\text{off}}$. In fact, while the classical results cannot explain the excellent fits shown in Fig. 2, these can be explained by Theorem 1, because in each of these cases N_σ is sufficiently large.

In Fig. 3 top left corner, we use the Kullback-Leibler (KL) divergence to measure the proximity of the effective NB distribution to the telegraph model distribution. For fixed values of ρ and f_{on} , increasing N_σ results in a rapid decrease in the KL divergence, which agrees with Theorem 1. In the same figure, we also show the KL divergence

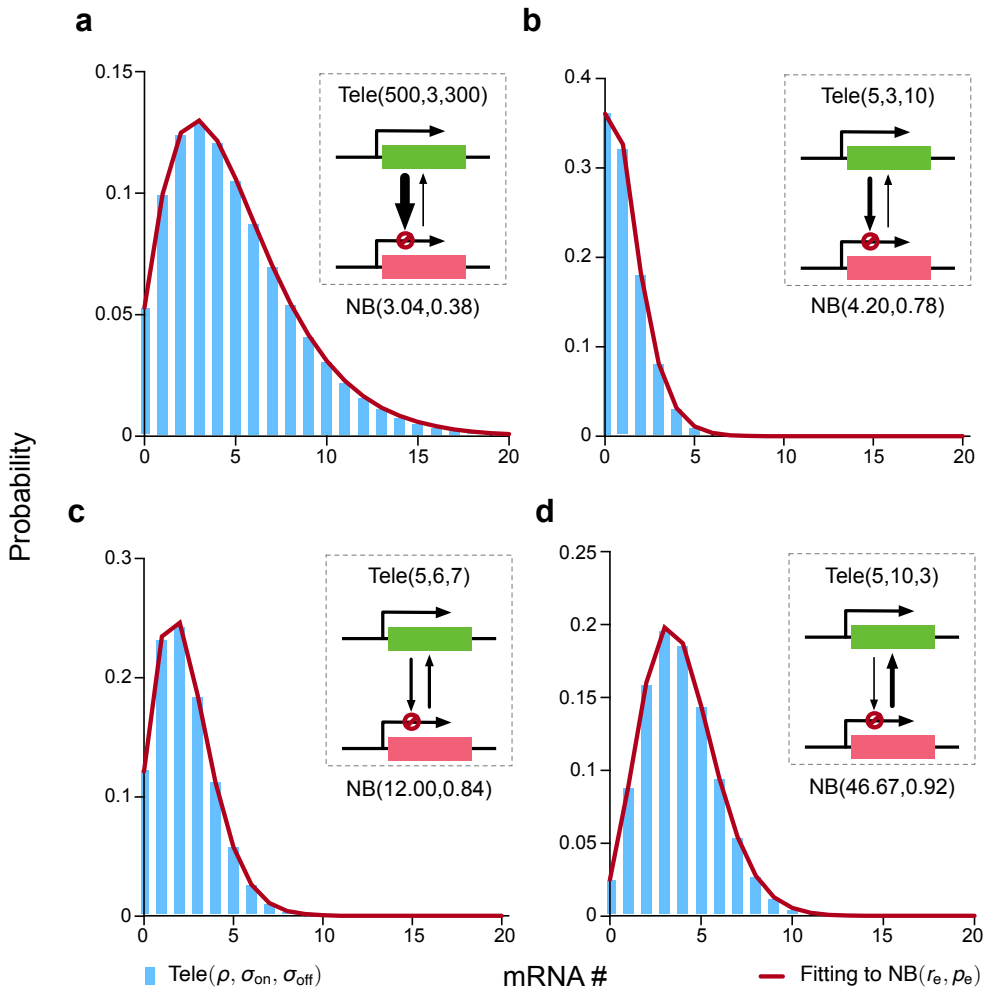

Figure 2: Comparison of the steady-state mRNA distribution of the telegraph model (Eq. (1) denoted as Tele$(\rho, \sigma_{\rm on}, \sigma_{\rm off})$) with the effective NB distribution (Eq. (6) denoted as NB$(r_{\rm e}, p_{\rm e})$) for the case of perfectly identical cells (parameters do not vary from cell-to-cell). The two parameters of the effective NB distribution are chosen so that its first and second moments of mRNA counts exactly agree with those of the telegraph model (the values of the two parameters, $r_{\rm e}$ and $p_{\rm e}$, are stated to 2 decimal places in the figure). In all four parameter cases, the effective NB distribution exceptionally well fits the corresponding telegraph distribution, and yet only in case (a) we have $\sigma_{\rm on} \ll \sigma_{\rm off}$ (the classical case of transcriptional bursting). This demonstrates that a good fit of an NB distribution to the telegraph model distribution does not imply the presence of transcriptional bursting.

between the Poisson distribution (with its parameter set to the mean of the telegraph model distribution) and the telegraph model distribution. It can be deduced that the fit of the Poisson distribution to the telegraph model distribution also becomes more accurate as N_σ increases, although it is always a poorer fit than the effective NB distribution. Of course, in the limit of large N_σ, the KL divergence is very small for both effective NB and Poisson distributions and hence, practically speaking, in this limit both will provide an excellent approximation to the telegraph model distribution. These observations are further supported by comparing in Fig. 3 the mRNA count distributions predicted by the telegraph model with the effective NB distribution and the Poisson distribution for 5 different values of N_σ (these correspond to the points A-E in the plot of the KL divergence versus N_σ in the top left corner of Fig. 3).

2.3 The NB distribution provides an optimal approximation to the telegraph model distribution for an intermediate range of the sum of the gene switching rates

The distribution comparison in Fig. 3 suggests that the effective NB distribution is an optimal fit to the telegraph model distribution in the intermediate range of N_σ. This is because: (i) for small N_σ, Theorem 1 does not hold. As well, in this case, it is already known that small $\sigma_{\rm on}$ and $\sigma_{\rm off}$ imply bimodal distributions of the telegraph model, which clearly cannot be well fitted by the (unimodal) effective NB distribution [56, 57]; (ii) for large N_σ,

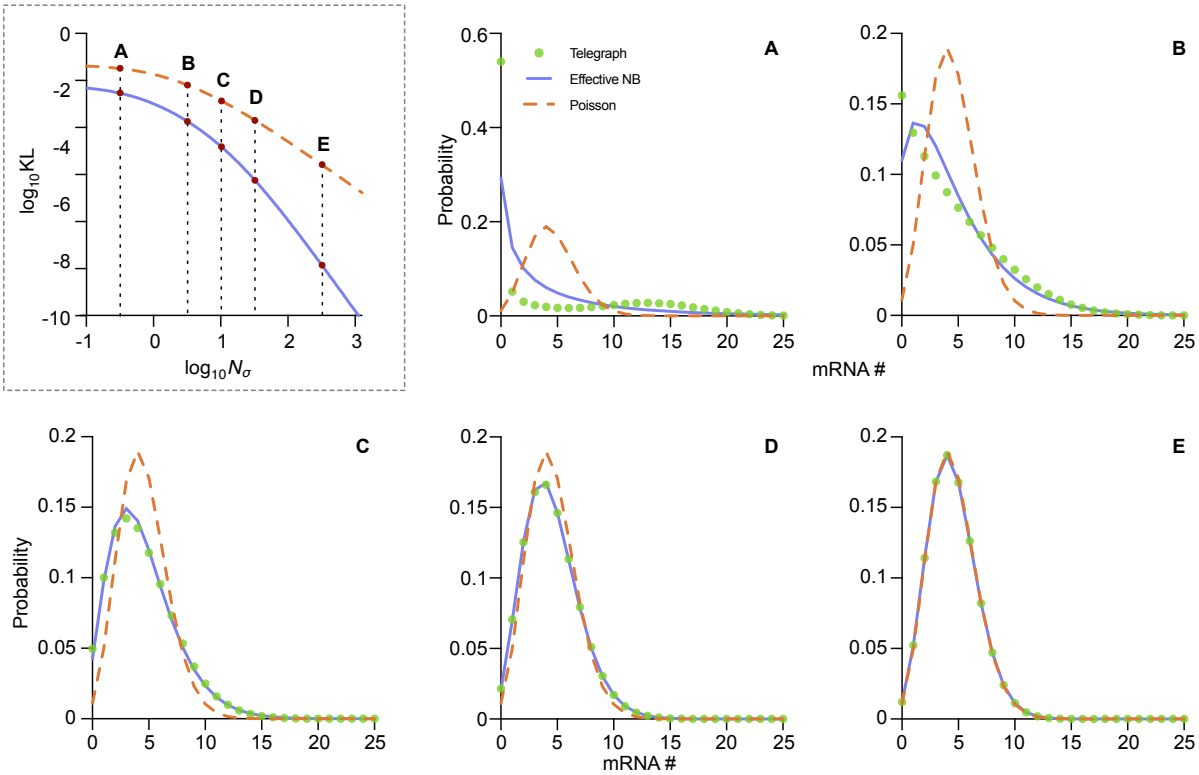

Figure 3: The mRNA distribution of the telegraph model converges to that of effective NB distribution as the sum of the gene-state switching rates relative to the mRNA degradation rate (N_σ) increases. In the dashed box, we show that the effective NB distribution (blue solid line) exhibits a lower KL divergence to the telegraph model distribution compared to the Poisson approximation (orange dashed line) as N_σ grows. The distributions of the telegraph model (green dots), effective NB (blue solid lines), and Poisson (orange dashed lines) are shown for Points A-E, as indicated in the KL divergence plot. The other parameters are fixed at $\rho = 15$ and $f_{\text{on}} = 0.3$. Point A: $N_\sigma = 0.3$, Point B: $N_\sigma = 3$, Point C: $N_\sigma = 10$, Point D: $N_\sigma = 30$, Point E: $N_\sigma = 300$.

the Poisson distribution practically provides an equally good fit as the effective NB distribution, but is preferable by the principle of Occam's razor because it has one less parameter than the effective NB approximation.

To make this observation more rigorous, we will use a model selection approach. An often used method to select which of two models provides a best fit to the data involves (i) maximizing the likelihood function of each model on the sample data; (ii) evaluating the Bayesian Information Criterion (BIC) for each model which is a function of both the maximum value of the likelihood and the number of parameters; (iii) the optimal model is selected to be the one with the smallest BIC score.

In particular, the BIC is defined as

$$\text{BIC}(\mathcal{M}, \mathcal{X}) = |\mathcal{M}| \ln |\mathcal{X}| - 2 \sum_{i=1}^{|\mathcal{X}|} \ln P_{\mathcal{M}}(n_i|\theta_{\text{MLE}}), \tag{9}$$

where $\mathcal{X}$ represents the set of count data $\{n_i\}$ for $i = 1, \cdots, n_c$, with $|\mathcal{X}| = n_c$ being the size of the dataset (number of cells). The proposed model is denoted by $\mathcal{M}$, its number of parameters by $|\mathcal{M}|$ and the distribution that defines the model (its likelihood function) by $P_{\mathcal{M}}(n_i|\theta)$ where θ represents the model parameters. The parameter θ_{MLE} is the maximum likelihood estimate (MLE) obtained by maximizing the log-likelihood function

$$\mathcal{L}_\theta = \sum_{i=1}^{|\mathcal{X}|} \ln P_{\mathcal{M}}(n_i|\theta).$$

In our case, given a parameter set $\{\sigma_{\text{on}}, \sigma_{\text{off}}, \rho\}$, the count data is simulated by randomly sampling the telegraph model distribution n_c times. The difficulty with this approach is that for each parameter set, there is an infinite number of datasets that can be simulated, and in principle, the algorithm can select a different best model for each dataset. Rather, what we want is to associate a unique best model with each set of parameters.

Hence, in what follows, we propose a new measure, the approximate expected Bayesian information criterion (aeBIC), as an estimator of the average BIC over repeated samples. The aeBIC is defined as

$$\mathrm{aeBIC}(\mathcal{M}, n_c) = |\mathcal{M}| \ln n_c + 2 n_c \varepsilon_{\mathrm{c}}(\mathcal{M}_{\mathrm{MLE}}, \mathcal{G}). \tag{10}$$

Here, ε_{c} represents the cross-entropy between two distributions (representing the model and the ground truth), defined as

$$\varepsilon_{\mathrm{c}}(\mathcal{M}, \mathcal{G}) = -\sum_{n=0}^{\infty} P_{\mathcal{G}}(n) \ln P_{\mathcal{M}}(n|\theta), \tag{11}$$

where $P_{\mathcal{G}}(n)$ and $P_{\mathcal{M}}(n|\theta)$ are the probabilities of observing n molecules in distributions $\mathcal{G}$ (the ground-truth distribution which for us is the telegraph model distribution) and $\mathcal{M}$ (the proposed model which can be telegraph, NB or Poisson with parameters θ), respectively. Note that the subscript MLE denotes that the kinetic parameters of the proposed distribution are estimated via maximum likelihood. Specifically, $\mathcal{M}_{\mathrm{MLE}}$ represents the model whose parameters θ minimize Eq. (11). If $\mathcal{M} = \mathcal{G}$, the cross entropy $\varepsilon_{\mathrm{c}}(\mathcal{M}, \mathcal{G})$ reduces to the information entropy $\varepsilon(\mathcal{G})$. Additionally, the cross entropy can be decomposed as [58]

$$\varepsilon_{\mathrm{c}}(\mathcal{M}, \mathcal{G}) = \varepsilon(\mathcal{G}) + \mathrm{KL}(\mathcal{M}||\mathcal{G}), \tag{12}$$

where $\mathrm{KL}(\mathcal{M}||\mathcal{G})$ is the KL divergence between $\mathcal{M}$ and $\mathcal{G}$. This decomposition implies that the closer the proposed distribution $\mathcal{M}$ is to the ground-truth distribution $\mathcal{G}$, the smaller the KL divergence becomes, and consequently, the cross entropy decreases as well.

In SI Sec 4.2 we show that the expectation of the BIC over an infinite number of independent samples (each of size n_c) is related to the aeBIC by:

$$\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathrm{BIC}(\mathcal{M}, \mathcal{X})] \leq \mathrm{aeBIC}(\mathcal{M}, n_c). \tag{13}$$

This shows that aeBIC serves as an upper bound for the expectation of the BIC. It can also be shown that the equality in Eq. (13) is achieved as $n_c \to \infty$.

However, in practice, we find that the difference between $\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathrm{BIC}(\mathcal{M}, \mathcal{X})]$ and $\mathrm{aeBIC}(\mathcal{M}, n_c)$ is negligible. To demonstrate this, for a fixed set of parameters $\{\rho, \sigma_{\mathrm{on}}, \sigma_{\mathrm{off}}\}$, we randomly sampled n_c counts from the telegraph model distribution $\mathrm{Tele}(\rho, \sigma_{\mathrm{on}}, \sigma_{\mathrm{off}})$ (using its Beta-Poisson formulation in Eq. (2)) and repeated this step $n_{\mathrm{trial}} = 10^3$ times. For each trial, we fitted the sampled counts to both Poisson and NB distributions using MLE and computed the BIC for both distributions. The BIC values were then averaged across all n_{trial} datasets to estimate the expectation of the BIC. Additionally, we computed the aeBIC using Eq. (10) for both distributions for each parameter set of the telegraph model distribution. The relative error of aeBIC compared to the expectation of BIC was calculated for each set of parameters. This comparison process is illustrated by a cartoon in Fig. 4a and the results for numerous parameter sets that cover the range $N_\sigma \in [0.1, 1000]$, $f_{\mathrm{on}} \in [0.1, 0.9]$ and $\rho \in [1, 15]$ are shown in Fig. 4b and Fig. S1a,b (these parameter sets correspond to a mean number of transcripts that ranges between 0 and 15). Note that, independent of the type of the fitted distribution, the relative error decreases with increasing sample size n_c. Notably, even for small sample sizes, the magnitude of the relative error is very small, and thereby for all intents and purposes, we can equate the aeBIC with the average of the BIC computed over many independent samples. This is very convenient because it is far faster to compute the aeBIC than the BIC since, for the former, the maximum likelihood estimation only needs to be done once using Eq. (11).

Next, we use Eqs. (10)-(11) to compute the aeBIC and identify the best-fitted model (the one with the smallest aeBIC) among Poisson, NB and telegraph model distributions across the $N_\sigma - f_{\mathrm{on}}$ parameter space (while keeping the transcription rate ρ constant), given that the ground-truth data is generated by the telegraph model. Note that this model selection can be performed for different sample sizes because the aeBIC is a function of n_c. Note also that the mean mRNA is proportional to f_{on} since the mean mRNA of the telegraph model is ρf_{on} (Eq. (4)) and ρ is fixed in our analysis. The results are shown in Fig. 4c — the model associated with the smallest aeBIC score corresponds to the model most frequently selected by the BIC calculated over many independent samples of the same size, thus providing a further test of the validity of the aeBIC approach (Table 2). In particular, aeBIC preferentially selects the telegraph model distribution for small N_σ, while the Poisson distribution is preferred for large N_σ. This agrees with the distribution comparison in Fig. 3 which suggested that the effective NB distribution is an optimal fit to the telegraph model distribution in an intermediate range of N_σ. The NB distribution is more likely to be selected as f_{on} decreases, which is consistent with the fact that the telegraph model distribution converges to an NB distribution when the gene spends most of its time in the inactive state (f_{on} is small). Interestingly, the shape and area of the crescent-shaped region where the NB distribution is an optimal model remains largely unchanged

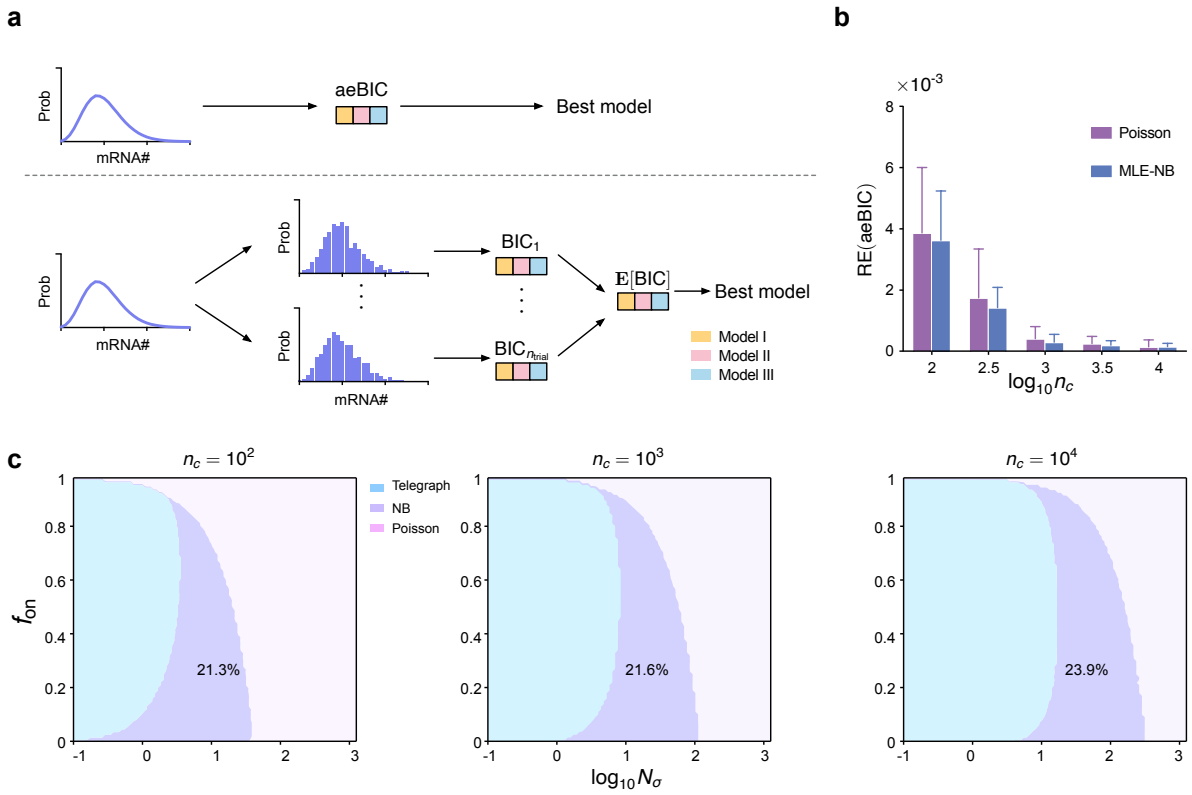

Figure 4: (a) Cartoon illustrating a computational approach to compare the aeBIC (top) with $\mathbf{E}[\text{BIC}]$ — the expected value of the BIC (bottom). The aeBIC utilizes a single score to select the best model distribution (telegraph, NB or Poisson) given that the ground-truth mRNA distribution is that of the telegraph model. The BIC method assigns a score to each different sample of simulated data from the telegraph model and then all these scores are averaged leading to $\mathbf{E}[\text{BIC}]$. (b) The relative error (RE) of aeBIC compared to $\mathbf{E}[\text{BIC}]$ for two distributions (Poisson and NB) as a function of sample size n_c for 10 parameter sets (see Table 1 for the values of N_σ and f_{on}; ρ is fixed to 15). Error bars show the standard error of the mean. (c) Phase diagram showing the regions of parameter space where the telegraph, NB and Poisson distributions are selected as optimal by the aeBIC, given that the ground-truth mRNA distribution is that of the telegraph model. Here n_c is the sample size, N_σ is the sum of gene-state switching rates normalised by the degradation rate of mRNA, and f_{on} is the fraction time spent in the active state. The fraction of the total parameter space occupied by the region where the NB distribution is optimally selected is shown on the plots. Note that the transcription rate is fixed to $\rho = 15$ which implies that the maximum mean number of transcripts in the phase plots is 15.

as the sample size increases, but its position shifts horizontally. In particular, the fraction of space where the NB distribution is the optimal model increases monotonically from 21.3% to 30% as the sample size increases over four orders of magnitude, from 10^2 to 10^6 cells (results for 10^6 cells are not shown).

Note that an alternative (and faster) way to obtain the phase space plots in Fig. 4c is to calculate the aeBIC with the value of the cross-entropy computed using moment matching between the proposed and the ground-truth distribution (instead of MLE). Specifically, the aeBIC using moment-matching is defined as

$$\text{aeBIC}(\mathcal{M}, n_c) = |\mathcal{M}| \ln n_c + 2 n_c \varepsilon_{\text{c}}(\mathcal{M}_{\text{MOM}}, \mathcal{G}), \tag{14}$$

where $\varepsilon_{\text{c}}(\mathcal{M}_{\text{MOM}}, \mathcal{G})$ is determined by evaluating Eq. (11) with parameters θ determined by moment-matching between the proposed distribution ($P_{\mathcal{M}}(n|\theta)$) and the ground-truth distribution ($P_{\mathcal{G}}(n)$). In Fig. S2 top left corner, we demonstrate that these two methods lead to very similar minimum values for the cross entropy and hence there is no appreciable difference on the model selection phase plots in Fig. 4c.

2.4 Cell-to-cell variability in transcript capture probability significantly affects the region of parameter space where the NB distribution is optimally selected

Thus far we have assumed that the scRNA-seq data can be well explained by the telegraph model. In reality, not all transcripts are captured, and hence there is a need to account for technical noise in the data. A simple way to do this is to assume that the number of zeros in the data is artificially high and to generate simulated scRNA-seq data using a zero-inflated telegraph model (a distribution that is a sum of a Dirac-delta function and

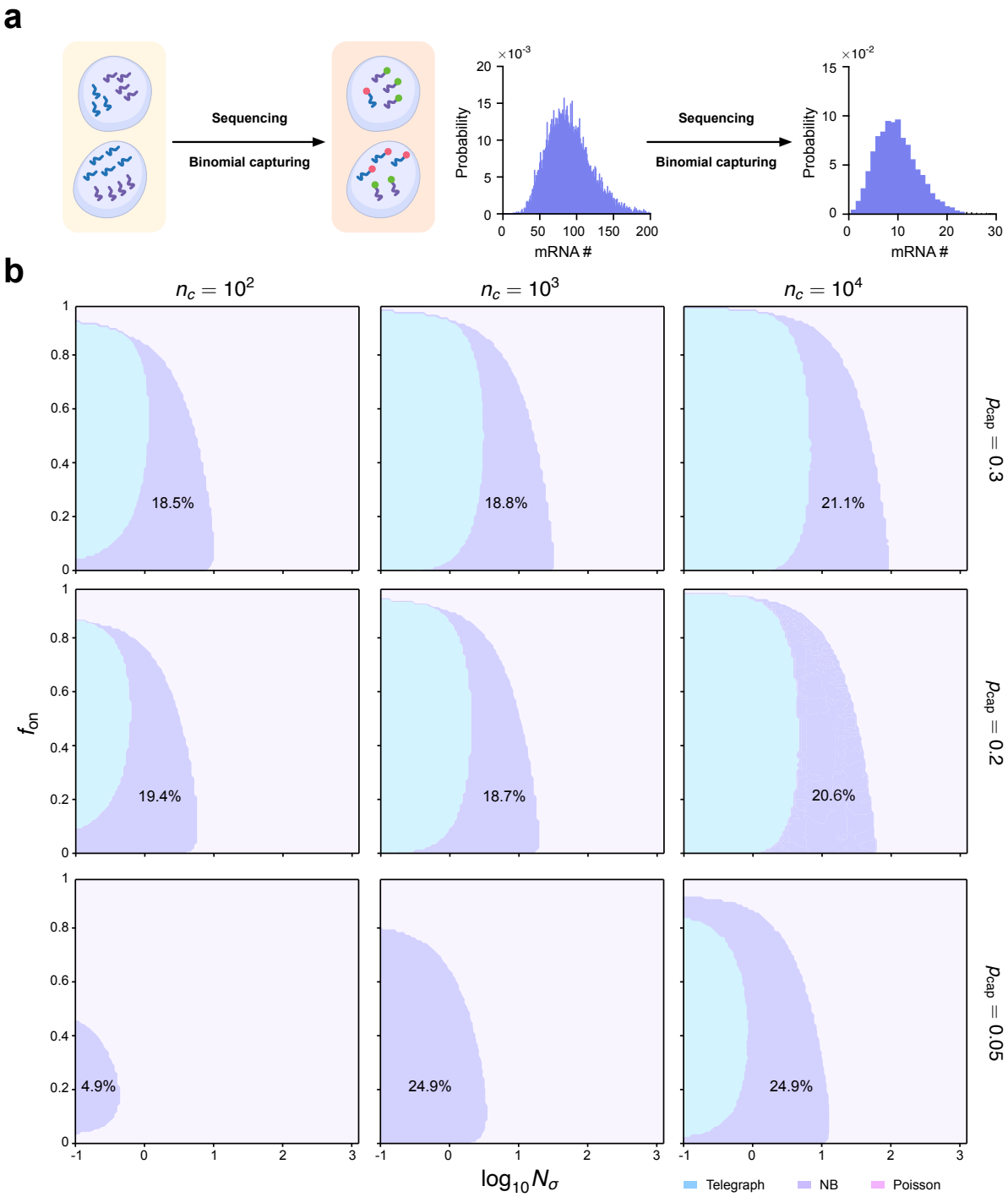

Figure 5: (a) Schematic illustrating the binomial capture model for scRNA-seq. Transcripts in each cell are captured with some probability p_{cap}. This causes a downsampling of the distribution of mRNA counts. (b) Phase diagram showing the regions of parameter space where the telegraph, NB and Poisson distributions are selected as the optimal ones by the aeBIC, given that the ground-truth mRNA distribution is that of the telegraph model. The phase diagrams are shown for three different values of p_{cap} (values stated next to the plots). Here n_c is the sample size, N_σ is the sum of gene-state switching rates normalised by the degradation rate of mRNA, and f_{on} is the fraction time spent in the active state. The fraction of the total parameter space occupied by the region where the NB distribution is optimally selected is shown on the plots. Note that the transcription rate is fixed to $\rho = 15$ which implies that the maximum mean number of transcripts in the phase plots is $15p_{\text{cap}} = 4.5, 3$ and 0.75 for the phase plots in rows 1, 2 and 3, respectively. These phase plots do not appreciably change if aeBIC is determined using moment-matching instead of MLE (Fig. S2).

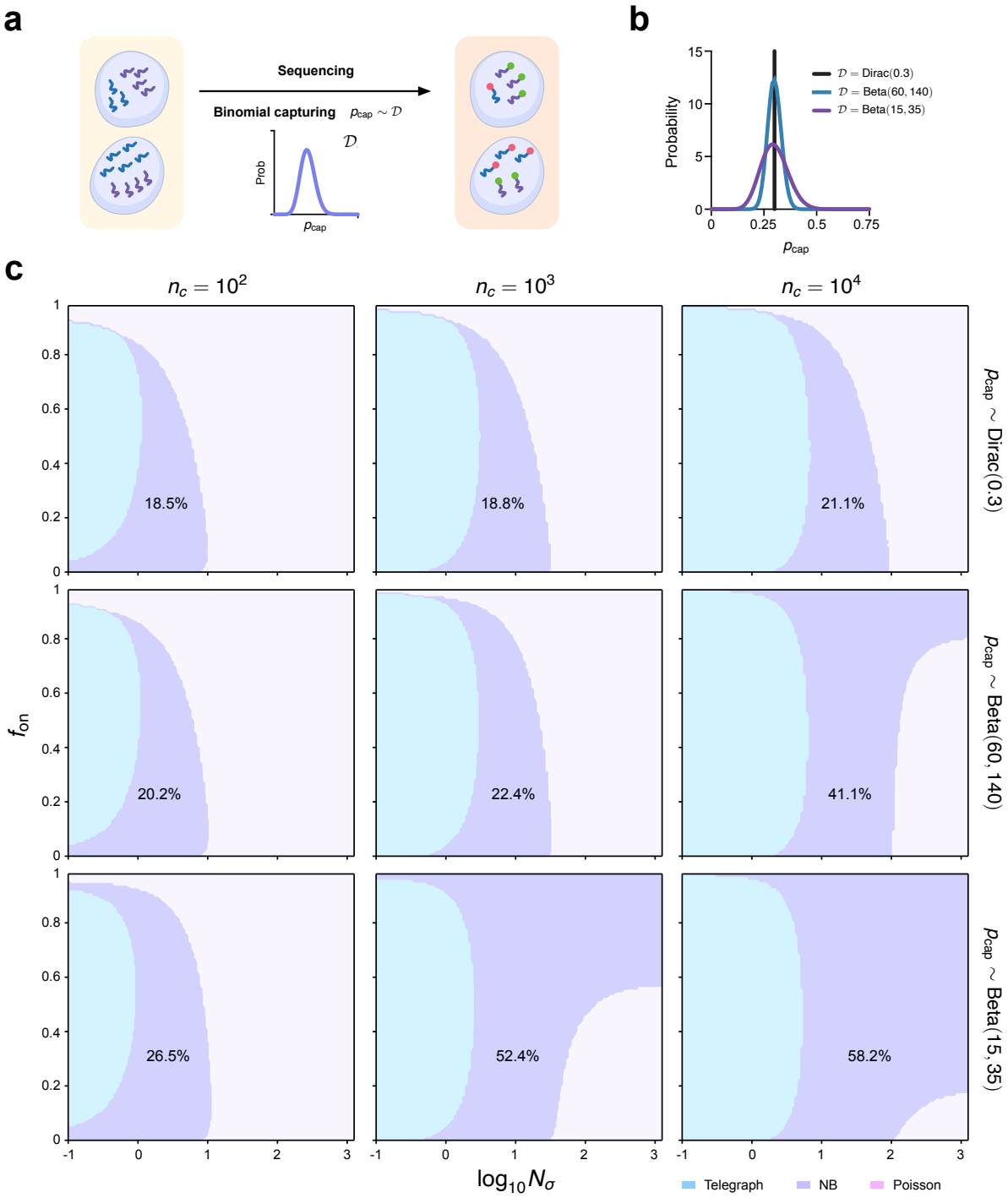

Figure 6: (a) Schematic illustrating the binomial capture model for scRNA-seq with a probability of mRNA capture, p_{cap}, that varies between cells according to some distribution. (b) We consider three different distributions all with mean $\langle p_{\text{cap}} \rangle = 0.3$ but with varying coefficient of variation (CV): (i) Dirac(0.3) with CV = 0; (ii) Beta(60, 140) with CV = 0.11; (iii) Beta(15, 35) with CV = 0.21. (c) Phase diagram showing the regions of parameter space where the telegraph, NB and Poisson distributions are selected as the optimal ones by the aeBIC, given that the ground-truth mRNA distribution is that of the telegraph model with effective transcription rate ρp_{cap} where p_{cap} is sampled from the 3 distributions mentioned above. Here n_c is the sample size, N_σ is the sum of gene-state switching rates normalised by the degradation rate of mRNA, and f_{on} is the fraction time spent in the active state. The fraction of the total parameter space occupied by the region where the NB distribution is optimally selected is shown on the plots. Note that the transcription rate is fixed to $\rho = 15$ which implies that the maximum mean number of transcripts in the phase plots is $15\langle p_{\text{cap}} \rangle = 4.5$.

the telegraph model distribution). While this approach or variants of it are widespread [14, 18, 19, 59] and attempts have been made to explain zero-inflation using mechanistic models [60], it is not ideal because (i) the downsampling of transcripts due to imperfect capture results in an increase of not only zeros but also 1's, 2's, etc; (ii) there is evidence that scRNA-seq data with unique molecular identifiers (UMIs) barely suffer from zero-inflation [6, 61]. A more principled and increasingly used model to explain the effect of imperfect capture is the binomial capture model [10, 38, 62]. This model assumes that each transcript is captured and observed with a probability p_{cap} (Fig. 5a). For current standard droplet-based sequencing technologies, the capture probability typically ranges between 0.05 and 0.3, depending on the sequencing technique used [11, 63–65].

For simplicity, first we consider the case where the capture probability does not vary from cell to cell and from gene to gene, i.e. the distribution of the capture probability is a Dirac-delta function, $\text{Dirac}(p_{\text{cap}})$. We investigated the impact of p_{cap} on model selection outcomes (using aeBIC) and the results are shown in Fig. 5b. Note that in this case $\mathcal{G}$ in Eq. (10) is the telegraph model distribution Eq .(1) with ρ rescaled to ρp_{cap} [37, 38]. Comparing Fig. 4c with Fig. 5b, we see that for a fixed sample size, the fraction of parameter space where the NB distribution is preferentially selected is almost constant for capture probabilities in the range $p_{\text{cap}} = 0.2 - 1.0$. The accuracy of scRNA-seq techniques is constantly improving, with commonly used platforms such as 10x having a capture probability that exceeds 0.2 [65]. Hence, we can conclude that if the capture probability is roughly the same from one cell to another, then the universality of the NB distribution in scRNA-seq datasets has little to do with the low capture efficiency.

Of course, for some scRNA-seq methods the capture probability may be very low and then in that case the technical noise will be so large that there will be a significant impact on model selection. For example, in the case of $p_{\text{cap}} = 0.05$ and sample size $n_c = 10^2$ in Fig. 5b, the Poisson distribution emerges as the optimal model across most of the parameter space, with the NB distribution being selected only in a limited region. The telegraph model, however, is never identified as the optimal choice under these conditions. This is expected because the bimodal character of the telegraph model distribution at small N_σ becomes less obvious at lower capture probabilities — downsampling causes the two peaks of the distribution to become closer together or even to merge and hence the best-fit distributions are exclusively unimodal (Poisson or NB distributions).

The wide distribution of the total counts (from all genes) per cell in typical scRNA-seq datasets (see for e.g. Fig. 1A of [17]) strongly suggests that the capture probability varies considerably across cells (Fig. 6a). To address this, we simulate scRNA-seq data by sampling from the underlying mRNA distribution given by Eq. (2) with ρ rescaled to ρp_{cap}, and p_{cap} that varies between cells according to three different distributions (Dirac-delta function and two different Beta distributions), all of which have mean $\langle p_{\text{cap}} \rangle = 0.3$ but with different coefficients of variation (CV = 0 for the Dirac-delta function, and 0.11 and 0.21 for the two different Beta distributions). See Fig. 6b for a plot of the three distributions. Note that the Beta distribution was chosen because it samples numbers in the range $(0, 1)$, a necessity given p_{cap} is a probability. The aeBIC was used to select between the standard telegraph model distribution (no assumption of variability of rates or capture probability between cells), NB and Poisson models with the ground-truth (measured) mRNA distribution $\mathcal{G}$ given by that of the telegraph model with p_{cap} distributed according to a Beta distribution (SI section 4.3). Note that the aeBIC remains an accurate predictor of the expectation of the BIC computed from MLE and hence suitable for model selection even when the transcript numbers are heavily influenced by technical noise (Fig S1c). The phase plots generated by aeBIC-based model selection are shown in Fig. 6c. We find that for fixed sample size n_c, an increase in the CV of the Beta distribution, leads to very small changes in the fraction of parameter space where the telegraph model distribution is selected, but the fraction of parameter space where the NB or Poisson distributions are selected changes significantly. In particular, an increase in the cell-to-cell variability in the capture probability tends to favor the selection of the NB distribution over the Poisson distribution. This is because variability in the capture probability increases the Fano factor (variance divided by the mean) of mRNA distributions (SI Section 4.4) — Fano factors larger than 1 can be captured by an NB distribution but the Poisson distribution has a fixed Fano factor of 1. For example, for a sample size of 10^3 cells, as the CV of the Beta distribution of p_{cap} is increased from 0 to 0.21 (keeping the mean constant at 0.3), the fraction of space where the telegraph model distribution is optimal decreases slightly from 32.1% to 29.1%, where the NB distribution is optimal increases from 18.8% to 52.4%, and where the Poisson distribution is optimal decreases from 49.1% to 18.5%.

2.5 In NB-optimal parameter space, burst parameter estimation accuracy is typically low, but ranking genes by burst frequency remains reliable

Finally, we seek to understand if any biologically relevant interpretation can be imparted to the two parameters of the NB distribution fitted to the scRNA-seq data. Consider the ideal scenario in which we can perfectly correct

for technical noise. In regions of parameter space where the NB distribution is optimally selected, can we obtain accurate estimates of the burst size and burst frequency from the two parameters of the best fitted NB distribution?

The aeBIC Eq. (10) used for model selection depends on the choice of the ground-truth model describing the distribution of the data ($\mathcal{G}$) and the model(s) ($\mathcal{M}$) to be fitted to this data. In our case, $\mathcal{G}$ is the telegraph model with effective transcription rate ρp_{cap}, and p_{cap} distributed according to a Beta distribution, Beta(a, b). Previously, we assumed that the models $\mathcal{M}$ are given by the standard telegraph, NB and Poisson distributions. These models do not have any information on the distribution of p_{cap} and hence this is the case where model selection using aeBIC is done without correcting for technical noise. In contrast, a perfect correction for technical noise is possible if instead the models $\mathcal{M}$ are corrected for technical noise, i.e. the mRNA distributions of the telegraph model, NB and Poisson models are integrated over the Beta(a, b) distribution (this of course assumes perfect knowledge of the parameters a and b characterizing the technical noise). For a derivation of these distributions, see Section 4.3. The differences between the conventional models and those with p_{cap} sampled from a distribution are illustrated in Fig. 7a.

In Fig. 7b we contrast model selection using conventional and technical-noise-corrected models for the case where p_{cap} is distributed according to a Beta$(60, 140)$ distribution which has a mean of $\langle p_{\text{cap}} \rangle = 0.3$ and CV$= 0.11$. We find that the regions of parameter space where the conventional and corrected telegraph models are selected are very similar. However, the region where the corrected NB distribution is selected (rainbow-colored region) is smaller than the region where the conventional NB distribution is selected — in fact, there is a region (shown in grey) where the conventional NB distribution and the corrected Poisson distribution are selected, i.e. in this region the apparent NB character of the mRNA count distribution is purely due to technical noise. The discrepancies between the two approaches are negligible for small sample sizes (10^2) but significant for large samples 10^4.

Next, in the parameter space region where the corrected NB distribution is selected as the optimal model, we estimate the two parameters of the corrected NB distribution by the method of moments. We equate the first two moments of the latter distribution with those of the model simulating the data, i.e. the corrected telegraph model. This mimics the fitting of the corrected NB distribution to the data using the method of maximum likelihood in the limit of large sample sizes. In Section 4.4 we show that this procedure leads to NB(r, p) where $r = r_{\text{e}}$ and $p = p_{\text{e}}$ are exactly as given by the effective NB distribution in Eq. (6). Clearly, because r_{e} and p_{e} do not depend solely on the true burst frequency σ_{on} and burst size ρ/σ_{off} of the telegraph model [66], the latter two parameters cannot be directly estimated by moment matching.

This issue can be circumvented by interpreting the best-fitting corrected NB distribution NB$(r_{\text{e}}, p_{\text{e}})$ as equal to the steady-state NB distribution solution NB $\left(\hat{\beta}_f, 1/(1+\hat{\beta}_s)\right)$ of the bursty gene expression circuit:

$$G \xrightarrow{\hat{\beta}_f} G + sM, \quad M \xrightarrow{1} \varnothing, \tag{15}$$

where s is a geometrically distributed random burst size with mean $\hat{\beta}_s$ and the burst frequency is $\hat{\beta}_f$ [67]. This model has experimental support [26] and its steady-state distribution is exactly the same as that of the telegraph model in the limit of transcriptional bursting, i.e. when the inactivation rate is much larger than the activation rate.

Hence the described inference procedure leads to estimated burst frequency and size that are given by

$$\hat{\beta}_f = r_{\text{e}} = \frac{f_{\text{on}}(1+N_\sigma)}{1-f_{\text{on}}}, \quad \hat{\beta}_s = \frac{1}{p_{\text{e}}} - 1 = \frac{\rho(1-f_{\text{on}})}{1+N_\sigma}. \tag{16}$$

Since we know that the true burst frequency of the telegraph model is σ_{on} and the true burst size is ρ/σ_{off}, it follows that the relative errors are given by

$$\text{RE}(\hat{\beta}_s) = \left|\frac{(\hat{\beta}_s - \rho/\sigma_{\text{off}})}{\rho/\sigma_{\text{off}}}\right| = 1 - \frac{(1-f_{\text{on}})^2 N_\sigma}{1+N_\sigma}, \quad \text{RE}(\hat{\beta}_f) = \left|\frac{(\hat{\beta}_f - \sigma_{\text{on}})}{\sigma_{\text{on}}}\right| = \frac{1+f_{\text{on}}N_\sigma}{(1-f_{\text{on}})N_\sigma}. \tag{17}$$

In Fig. 7b we show a heatmap of the relative errors in the NB-optimal region (using the aeBIC with technical-noise-corrected models). The magnitude of these errors vary significantly and are smallest when f_{on} is small, i.e. when $\sigma_{\text{off}} \gg \sigma_{\text{on}}$ and $\sigma_{\text{off}} \gg 1$. Hence, this implies that even though the NB distribution may be optimally selected, nevertheless this does not guarantee the accuracy of the estimated burst parameters. Indeed, the lack of accuracy is not very surprising given that the NB distribution of the reaction scheme in 15 is only a good approximation to the telegraph model distribution when $\sigma_{\text{on}} \ll \sigma_{\text{off}}$ [49] and that we have previously established that it is possible to have excellent NB fits even when this inequality does not hold (Fig. 2). Note that the burst size can generally be

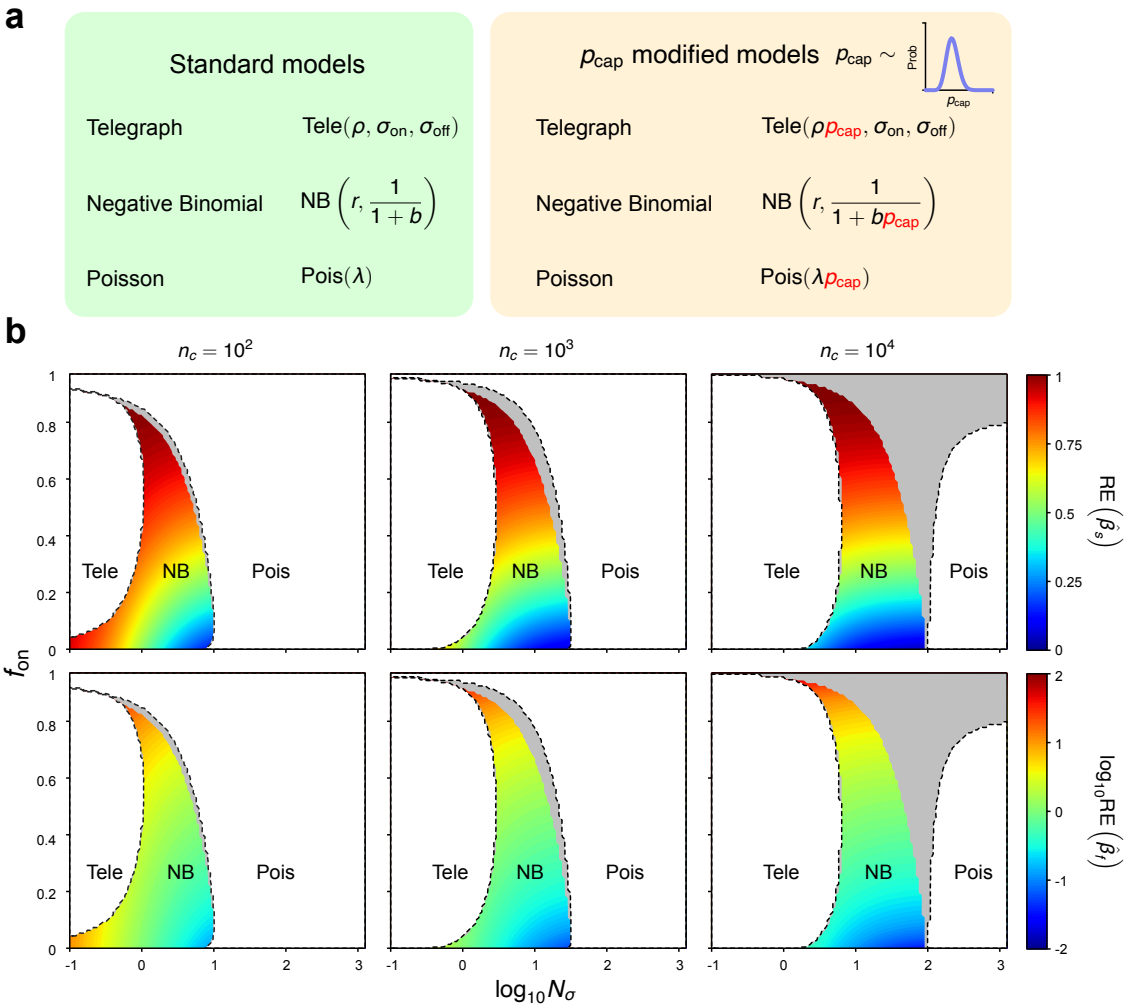

Figure 7: (a) Illustration of the differences between the standard and technical-noise-corrected models; (b) Phase diagrams produced by aeBIC model selection based on standard or corrected models. For both, the ground-truth model for observed data is the telegraph model with p_{cap} distributed according to the Beta$(60, 140)$ distribution which has mean $\langle p_{\text{cap}} \rangle = 0.3$ and CV=0.11. The transcription rate ρ is fixed to 15; the maximum mean number of transcripts is 4.5. The labels "Tele", "NB" and "Pois" denote the regions selected using the aeBIC procedure with corrected models. The dashed lines demarcate the same regions but using the aeBIC procedure with standard models. The "Pois" area is divided into a white part (where both aeBIC procedures select the Poisson distribution) and a grey part (where the aeBIC with standard models selects the NB distribution while the aeBIC with corrected models selects the Poisson distribution). The heatmap shows the magnitude of the relative errors in the estimated burst frequency ($\hat{\beta}_f$) and burst size ($\hat{\beta}_s$) in the NB-optimal region (using the aeBIC with corrected models). The errors are computed using Eq. (17) — note that this approach assumes full knowledge of the distribution of probability capture, an ideal case. In the plots, n_c denotes sample size, N_σ is the sum of gene-state switching rates normalised by the degradation rate of mRNA, and f_{on} is the fraction time spent in the active state.

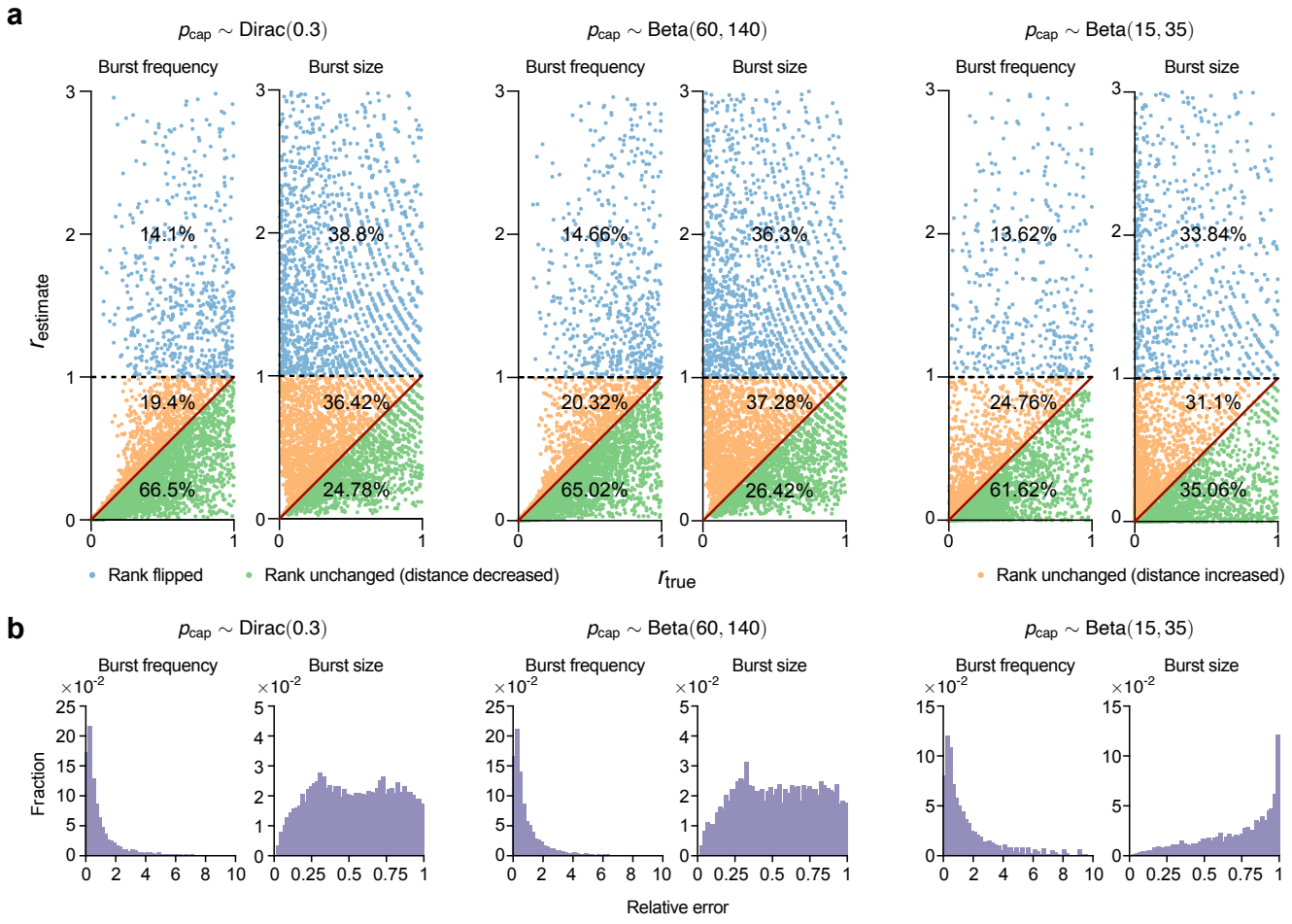

Figure 8: (a) Scatter plot of the estimated ratio of burst parameters r_{estimate} of a pair of genes and of the ground-truth ratio of the burst parameters r_{true} of the same pair of genes. All parameter sets are sampled from the region of parameter space where the technical-noise-corrected NB distribution is optimally selected using the aeBIC approach (purple regions in Fig. S3). Points marked as blue are those pairs of genes for which the estimation led to an incorrect ordering of genes by the size of the burst parameter. The order was correctly inferred for gene pairs corresponding to orange and green points. Perfect ratio estimation is shown by the solid red line; gene pairs corresponding to orange (green) points overestimate (underestimate) the distance between the burst parameters of the gene pairs. (b) Distributions of the relative errors in burst frequency and burst size for those pairs of genes for which the order was correctly inferred (orange and green points) in (a). Note that the inference and model selection approach here assumes full knowledge of the distribution of probability capture, an ideal case.

estimated more reliably than the burst frequency — this is also apparent from Eq. (17) which implies $\mathrm{RE}(\hat{\beta}_s) < 1$ and $\mathrm{RE}(\hat{\beta}_f) < \infty$.

Next, we sought to understand whether reliable information of some type can be extracted from the burst parameters, even though their absolute values are generally inaccurate. In particular, we seek to understand whether ranking of genes by their burst parameter values can be accurate since this is based on relative rather than absolute information. To answer this question, we used the following protocol:

1. Randomly sample two pairs of parameter sets $(f_{\mathrm{on}}^{(1)}, N_\sigma^{(1)})$ and $(f_{\mathrm{on}}^{(2)}, N_\sigma^{(2)})$, each associated with a different gene, from the parameter space region where the NB distribution (corrected for technical noise) is selected as the optimal model for p_{cap} distributed according to the Dirac $(0, 3)$, Beta$(60, 140)$ and Beta$(15, 35)$ distributions. For a sample size of $n_c = 10^4$, these regions are shown in purple in Fig. S3. Note that the transcription rate ρ is fixed to 15 in all cases.

2. Calculate the true burst frequency and size for each gene. For the first gene, we denote the true burst frequency and burst size as $\beta_f^{(1)} = \sigma_{\mathrm{on}}^{(1)} = f_{\mathrm{on}}^{(1)} N_\sigma^{(1)}$ and $\beta_s^{(1)} = \rho/\sigma_{\mathrm{off}}^{(1)} = \rho/(N_\sigma^{(1)}(1 - f_{\mathrm{on}}^{(1)}))$, respectively; for the second gene, $\beta_f^{(2)} = \sigma_{\mathrm{on}}^{(2)} = f_{\mathrm{on}}^{(2)} N_\sigma^{(2)}$ and $\beta_s^{(2)} = \rho/\sigma_{\mathrm{off}}^{(2)} = \rho/(N_\sigma^{(2)}(1 - f_{\mathrm{on}}^{(2)}))$, respectively. These relationships follow from Eq. (7).

3. For each gene, use the method of moments to fit an NB distribution integrated over the distribution of p_{cap} to the distribution of the data, i.e. telegraph model distribution integrated over the distribution of p_{cap}. This leads to the estimated burst frequency-size pairs for each gene: $(\hat{\beta}_f^{(1)}, \hat{\beta}_s^{(1)})$ and $(\hat{\beta}_f^{(2)}, \hat{\beta}_s^{(2)})$ which are given by Eq. (16).

4. Compute the ground-truth ratios of burst sizes for the pair of genes as $r_{\mathrm{true}} = \min(\beta_s^{(1)}, \beta_s^{(2)})/\max(\beta_s^{(1)}, \beta_s^{(2)})$. Note that by placing the smaller burst size in the numerator, we guarantee $r_{\mathrm{true}} < 1$. The ground-truth ratio for the burst frequencies can be computed similarly.

5. Compute the estimated ratio of burst sizes for the pair of genes r_{estimate} , preserving the original pair order used for the ground-truth ratio of burst sizes. That is, if $r_{\mathrm{true}} = \beta_s^{(1)}/\beta_s^{(2)}$ then $r_{\mathrm{estimate}} = \hat{\beta}_s^{(1)}/\hat{\beta}_s^{(2)}$; if $r_{\mathrm{true}} = \beta_s^{(2)}/\beta_s^{(1)}$ then $r_{\mathrm{estimate}} = \hat{\beta}_s^{(2)}/\hat{\beta}_s^{(1)}$. The estimated ratio for the burst frequencies can be computed similarly.

In Fig. 8a we show a plot of r_{true} versus r_{estimate}. Each point in this scatter plot corresponds to a randomly selected pair of genes. Points are classified by their color: (i) blue if $r_{\mathrm{estimate}} > 1$ indicates a reversal of pair order which implies that burst parameter inference fails to preserve the relative magnitude of the burst parameters; (ii) orange if $r_{\mathrm{estimate}} < 1$ and $r_{\mathrm{estimate}} > r_{\mathrm{true}}$ indicates that the order of the estimated burst parameters is the same as that of the ground truth parameters but that the estimation leads to an amplification of the difference between the burst parameters of the two genes; (iii) green if $r_{\mathrm{estimate}} < 1$ and $r_{\mathrm{estimate}} < r_{\mathrm{true}}$ indicates that the order of the estimated burst parameters is the same as that of the ground truth parameters but that the estimation leads to a reduction of the difference between the parameters of the two genes.

The results show that in many instances the ranking of a pair of genes by the size of their burst frequency is correctly estimated — these are the orange and green points in Fig. 8a which imply gene ranking accuracy by burst frequency in $86 - 87\%$ of cases. However, the ranking of genes by burst size is significantly less accurate: it is correct in merely $58 - 69\%$ of all cases, though it is better than expected by chance. In Fig. 8b we show the relative errors of the burst parameters for the pairs of genes that were correctly ranked (orange and green points in (a)). This verifies that generally relative errors and ranking accuracy are not related to each other since the ranking can be accurate and yet the relative errors can be large.

The main limitations of our approach are that we assume: (i) perfect knowledge of the distribution of capture probability; (ii) the information in the first two moments of the transcript count distribution is enough to accurately infer the burst parameters; (iii) we exclusively perform parameter inference using the technical-noise corrected NB model. To overcome these limitations, in Section 4.5, we describe an approach which simulates the workflow of an actual scRNA-seq experiment: (i) synthetic count data is generated for thousands of genes in n_c cells, which includes noise from both biological and technical sources; technical noise is modeled by a Beta distributed transcript capture probability; (ii) a maximum likelihood-based approach is used to infer the burst frequency and size, and to select between the technical-noise corrected telegraph, NB and Poisson models derived in Section 4.3. Note that this approach relies on using the measured total counts from all genes in a cell as a proxy for the (unknown) capture probability for that cell; the approach also does not simply use the first two moments of the observed distribution of counts, but the full distribution information and hence overcomes some of the known limitations of

moment-matching approaches [68]; (iii) for those genes for which the technical-noise corrected NB model is selected as optimal by the BIC, the burst frequency and size are estimated separately from the technical-noise corrected telegraph and NB models. These are compared to the ground-truth values to calculate the relative errors. In addition, the percentage of pairs of genes that are correctly ranked according to the magnitude of their burst frequency and size, are calculated for each model. The whole methodology is illustrated in Fig. S4a. We considered three variations of this procedure, according to the generation of synthetic data in step (i):(a) $n_c = 10^3$ cells, $N_\sigma = 10$ and the distribution of p_{cap} is fixed to Beta$(15, 35)$; (b) $n_c = 10^4$ cells, $N_\sigma = 10$ and the distribution of p_{cap} is fixed to Beta$(60, 140)$; (c) $n_c = 10^4$ cells, $N_\sigma = 20$ and the distribution of p_{cap} is fixed to Beta$(60, 140)$.

The results for (a) are shown in Fig. S4 and for (b) and (c) are shown in Fig. S5. We find that for each fixed value of N_σ, the relative errors of the burst frequency and size increase with the fraction of time that the gene spends in the active state f_{on}. This agrees with the results shown in Fig. 7. Interestingly, even though the genes analyzed are those for which the technical-noise corrected NB model is selected as the optimal model by BIC, the parameter estimates are more accurate using the technical-noise corrected telegraph model (compare blue and green bars in Fig. S4(b), Fig. S5(a) and Fig. S5(c)). In addition, a high percentage of pairs of genes are correctly ranked according to the magnitude of their burst frequency, both by the technical noise-corrected NB and telegraph models; however, ranking by the burst size was much less reliable (see the piecharts in Fig. S4(c), Fig. S5(b) and Fig. S5(d))). The rankings also agree with the results shown in Fig. 8. Hence, overall, the results obtained from the MLE-based inference and BIC verify the accuracy of those previously obtained using moment-matching and aeBIC.

3 Discussion

In this study, we have investigated why the NB distribution effectively approximates the transcript count distribution observed in scRNA-seq experiments and evaluated the biological implications of its two parameters. Presently, the universality of the NB distribution in UMI count data is puzzling given that a decade of analytical studies have unearthed only one stringent condition under which the distribution of the telegraph model of stochastic gene expression is well approximated by an NB distribution: genes must be predominantly inactive in a population of identical cells — conditions that are unlikely met in scRNA-seq experiments.

We first showed using theory and a new model selection criterion (aeBIC) that in the ideal case where the noise in the data is purely of biological origin, the NB distribution provides an accurate and optimal approximation of the transcript count distribution of a gene in a crescent-shaped region of its f_{on} vs N_σ parameter space where f_{on} is the fraction of time spent in the active state, while N_σ is the sum of the activation and deactivation rates (normalized by the degradation rate). This region occupied about 20% of the scanned parameter space, showing little dependence with the sample size (number of cells). In particular, the NB distribution is an excellent approximation in an intermediate range of N_σ and the size of this range increases with decreasing f_{on}. The classical parameter regime where the NB distribution is thought to be a good approximation to the telegraph model, i.e. when f_{on} is small (genes mostly inactive) and transcriptional bursting is apparent [49], is a subset of the parameter regime that we have identified. In fact, this new region encompasses genes whose f_{on} ranges from small to large, thus showing that a good fit of the NB to the measured distribution of transcript counts is generally not indicative of transcriptional bursting.

Next, we investigated the more realistic case where the noise, i.e. the variability of mRNA numbers across cells, is due to both biological and technical noise. We simulated observed scRNA-seq data by downsampling the counts of the telegraph model of gene expression, assuming that the transcript capture probability for each cell is either constant or beta distributed. The selection of models through aeBIC allowed us to quantify how the proportion F of the parameter space where the NB distribution optimally fits the data varies with sample size and the shape of the capture probability distribution. We found that for a sample size of 100 cells where all noise is biological noise, $F \approx 21\%$; where technical noise is added assuming that each cell has a transcript capture probability of 0.3, $F \approx 19\%$; where technical noise is added assuming that the transcript capture probability is sampled from a Beta distribution with mean 0.3 and CV= 0.21, $F \approx 27\%$. In contrast, for a sample size of 10^4 cells, the parameter space fractions for the aforementioned three cases were $F \approx 24\%$, $F \approx 21\%$ and $F \approx 58\%$. Note that a mean capture probability of 0.3 is typical of 10x Genomics Chromium Single Cell 3' version 3 reagent chemistry [65]. From these comparisons, we can deduce the following. If the capture probability distribution is: (i) narrow, the results are similar to those obtained under the assumption that all noise is biological; (ii) wide but the sample size is small (order 100 cells) then the results are also similar to those obtained under the assumption that all noise is biological; (iii) wide but for sample sizes of the order of thousands of cells, we find a significant enhancement of the region of parameter space where the NB distribution is the optimal model compared to the case where all noise is biological.

Specifically, the proportion of the parameter space favoring the telegraph model remains largely constant, while the proportion favoring the NB distribution increases at the expense of the Poisson distribution. Hence, given that the capture probability distribution is wide in many cases (reflected by wide distributions of the total UMI counts per cell [17, 50, 51]), our results suggest that for small sample sizes, biological noise from those genes with an intermediate-sized N_σ parameter is the main determinant of the NB character of the count distribution, while for large sample sizes, technical noise leads to NB distributions in regions of parameter space where biological noise cannot.

A previous study by Tang et al. [37] generated a simulated data set consisting of 500 cells and 7000 genes by sampling from the steady-state distribution of the telegraph model (Eq. (1)). For each gene, the Akaike Information Criterion (AIC) scores were calculated for the telegraph, NB, and Poisson models, and the model with the lowest AIC was selected as the optimal fit. The set of genes for which the Poisson and NB models were favored tended to have a lower mean expression; however, the distributions of the means of the genes in the three groups (see Fig. 3a of [37]) overlapped significantly, indicating that the mean alone is not enough to explain the model that best fits the expression of each gene. In the phase diagrams in Fig. 4c we show the regions where the three models are selected as a function of f_{on} and N_σ — clearly N_σ, not f_{on}, is the primary determinant of the chosen model. Since the mean mRNA is proportional to f_{on} (the mean mRNA of the telegraph model is ρf_{on} and ρ is fixed in these phase plots), it follows that the selection of the optimal model is only weakly determined by the mean RNA. This explains why Tang et al. [37] did not observe a strong relationship between the selected model and the mean mRNA count. Another study [69] has also shown that the NB distribution is often complex enough to describe the simulated scRNA-seq data generated from the telegraph model but did not map the relationship between the choice of the best-fit model and the regions of parameter space, or account for technical noise, as we have done here. The complete mapping that we have reported was only possible due to the use of the novel model selection criterion, aeBIC, which is a computationally efficient and accurate estimator of the model most frequently selected by the standard maximum likelihood procedure followed by BIC.

We also investigated the reliability of the burst frequency and burst size estimated from the two parameters of the best-fitting NB distribution in the parameter space region where this model is selected as the optimal one. These are commonly estimated by interpreting the NB distribution as that arising from a simpler mechanistic model than the telegraph model, namely the bursty gene expression model (with reaction scheme Eq. (15)), which was first studied in [67] and is now commonly used as the basis for more sophisticated stochastic models of gene expression [70–75] including those used to fit scRNA-seq data [6, 52, 76–79]. This model can be derived from the telegraph model under the assumption that the gene inactivation rate is much larger than the gene activation rates [49], i.e. the transcriptional bursting regime. However, our analysis shows that when the parameter N_σ is large, the steady-state distribution of the telegraph model is well approximated by the negative binomial (NB) distribution, a scenario that applies beyond the bursting regime. This means that estimating the burst frequency and size of a gene by fitting the NB distribution solution of the bursty model Eq. (15) to scRNA-seq measurements of its expression (after correcting for technical noise) can lead to potentially large errors, even if the data is excellently fit by a NB distribution. Our analysis supports this notion — we find that in the region of parameter space where the distribution of the telegraph model is best fit by an NB distribution, the errors in the estimated burst frequency and size are generally large, easily exceeding 50% for the burst frequency and at most 100% for the burst size. However, interestingly, we observe that the relative magnitudes of the burst parameters still hold some validity (as also suggested by a recent study of the cell-cycle dependence of bursty gene expression [6]). In particular, we find that the ranking of two genes by their burst frequency estimate is correct in about $86 - 87\%$ of the cases; for the burst size, the ranking is correct in $58 - 69\%$ of the cases, showing less reliability. We also found that in the parameter space region where the NB model is preferentially selected, the limitations discussed above remain largely the same if instead one estimates the burst size and burst frequency from the three parameters obtained by fitting the distribution of the telegraph model (after correcting for technical noise) to the data using the method of maximum likelihood [35, 37, 80–83].

Our study also has some limitations. We have accounted for cell-to-cell variability in transcript counts due to intrinsic noise and differences in the effective transcription rate from one cell to another due to a varying transcript capture probability. Any variation in the effective transcription rate due to variability in the transcription rate (extrinsic noise on the transcription rate) between cells is indistinguishable from variability in the transcript capture probability and hence is automatically accounted for in our present method. However, we have assumed that there is no cell-to-cell variation in the activation and inactivation rates, which is, of course, a simplification of biological reality. Within our phase diagram methodology, these effects can be properly incorporated using a ground-truth distribution $\mathcal{G}$ in Eq. (10) that is derived via an integration of the distribution of the telegraph model over the distributions of the activation and inactivation rates — these are typically unknown but could be approximated

by gamma or lognormal distributions as in [52]. The issues we have identified regarding the inaccuracy of the absolute values of the burst parameters are possibly because we have exclusively used the steady-state distributions of stochastic gene expression models. These are most commonly used for two reasons: (i) scRNA-seq data often come from a single snapshot measurement; (ii) it is much easier to solve the chemical master equation of gene expression models in steady-state conditions than in time [84]. However, the median mRNA half-life in mammalian cells is a sizeable fraction of the cell-cycle duration (for e.g. ≈ 7 hours [85] compared to ≈ 13 hours [86] in mouse embryonic stem cells), and hence it is unlikely that a steady state assumption generally holds in this case. To overcome this issue, one can first use methods like DeepCycle [87] or VeloCycle [88] to assign a cell age θ (which varies between 0 at the beginning of the cell-cycle and 1 at cell division) to each cell and then fit an age-dependent model of gene expression to the age-resolved scRNA-seq data [6]. Alternative ways to possibly circumvent the issues we have reported using steady-state models is to instead fit models that account for different cell states due to differentiation [89] or models that use a variety of single-cell data (transcriptomic, proteomic and epigenomic) [90]. These approaches are computationally challenging and are actively under investigation.

In conclusion, our findings establish that the NB distribution is a robust approximation of scRNA-seq transcript count distributions across diverse noise conditions and parameter regimes. Notably, its broad applicability challenges the conventional theoretical link to bursty gene expression, since both biological and technical noise can induce NB-like behavior even outside of classical conditions. Our analysis cautions against direct burst parameter estimation from NB or telegraph model fits and calls for nuanced model selection and parameter estimation approaches in single-cell genomics.

Data Availability Statement

The code for this paper is available at `https://github.com/Vivian0911unique/aeBIC`.

Acknowledgements

Z.C. acknowledges support from a Shanghai Action Plan for Technological Innovation Grant (23S41900500) and a Natural Science and Engineering Research Council of Canada Discovery Grant (RGPIN-2024-06015). R.G. acknowledges support from the Leverhulme Trust (RPG-2024-082). The authors are grateful to Augustinas Sukys for his valuable feedback on previous drafts of this work.

Author contributions

Conceptualisation, Z.C. and R.G.; Methodology, Z.C. and R.G.; Software, Y.W. and Z.S.; Analysis, Y.W. and Z.S.; Visualisation, Y.W. and Z.S.; Writing – Original Draft, Z.C. and R.G.; Writing – Review & Editing, Z.C. and R.G.; Supervision, Z.C. and R.G.

4 Methods

4.1 Proof of Theorem 1

It is well known that the probability distribution of the Beta-Poisson process

$$x \sim \mathrm{Beta}(\sigma_{\mathrm{on}}, \sigma_{\mathrm{off}}), \qquad n \sim \mathrm{Pois}(\rho x),$$

is identical to the steady-state distribution of the telegraph model given by Eq. (1) [35]. Another standard result is that the probability distribution of the Gamma-Poisson process

$$y \sim \mathrm{Gamma}\left(r_{\mathrm{e}}, \frac{p_{\mathrm{e}}}{1-p_{\mathrm{e}}}\right), \qquad n \sim \mathrm{Pois}(y)$$

is identical to that of a negative binomial distribution $\mathrm{NB}(r_{\mathrm{e}},\ p_{\mathrm{e}})$. In what follows we shall fix these two parameters to be

$$r_{\mathrm{e}} = \frac{\sigma_{\mathrm{on}}(\sigma_{\mathrm{on}}+\sigma_{\mathrm{off}}+1)}{\sigma_{\mathrm{off}}}, \quad p_{\mathrm{e}} = \frac{(\sigma_{\mathrm{on}}+\sigma_{\mathrm{off}})(\sigma_{\mathrm{on}}+\sigma_{\mathrm{off}}+1)}{(\sigma_{\mathrm{on}}+\sigma_{\mathrm{off}})(\sigma_{\mathrm{on}}+\sigma_{\mathrm{off}}+1)+\rho\sigma_{\mathrm{off}}}. \tag{18}$$

Note also that the definition of the Gamma distribution, $\mathrm{Gamma}(\alpha, \beta)$, that we use is $f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y},\ y > 0$.

Proposition 7 in Ref. [91] states that for two mixed Poisson distributions, denoted as $\mathrm{MP}(g_1)$ and $\mathrm{MP}(g_2)$, the convergence $\mathrm{MP}(g_1) \to \mathrm{MP}(g_2)$ holds if and only if $g_1 \to g_2$ where $\to$ denotes convergence in distribution. Consequently, to prove Theorem 1, it suffices to demonstrate that the distribution of the random variable x converges to that of the random variable y/ρ as $N_\sigma = \sigma_{\mathrm{on}} + \sigma_{\mathrm{off}} \to \infty$.

To proceed, we first determine the density function of the transformed random variable $\bar{y} = y/\rho$. This is computed as

$$\pi(\bar{y}) = \pi(\rho\bar{y}) \times \frac{\mathrm{d}(\rho\bar{y})}{\mathrm{d}\bar{y}} = \frac{\rho}{\Gamma(r_{\mathrm{e}})}\left(\frac{p_{\mathrm{e}}}{1-p_{\mathrm{e}}}\right)^{r_{\mathrm{e}}} (\rho\bar{y})^{r_{\mathrm{e}}-1}\, \mathrm{e}^{-\frac{p_{\mathrm{e}}\rho\bar{y}}{1-p_{\mathrm{e}}}} = \frac{1}{\Gamma(r_{\mathrm{e}})}\left(\frac{\rho p_{\mathrm{e}}}{1-p_{\mathrm{e}}}\right)^{r_{\mathrm{e}}} (\bar{y})^{r_{\mathrm{e}}-1}\, \mathrm{e}^{-\frac{p_{\mathrm{e}}\rho\bar{y}}{1-p_{\mathrm{e}}}},$$

which precisely corresponds to the density function of $\mathrm{Gamma}(r_{\mathrm{e}}, \frac{\rho p_{\mathrm{e}}}{1-p_{\mathrm{e}}})$.

To show $x \to \bar{y}$, we need to establish the convergence

$$\mathrm{Beta}(\sigma_{\mathrm{on}}, \sigma_{\mathrm{off}}) \to \mathrm{Gamma}\left(r_{\mathrm{e}}, \frac{\rho p_{\mathrm{e}}}{1-p_{\mathrm{e}}}\right), \tag{19}$$

as $N_\sigma = \sigma_{\mathrm{on}} + \sigma_{\mathrm{off}} \to \infty$. Rather than directly proving Eq. (19), we instead demonstrate that both distributions converge to a common limiting distribution, denoted as Dist, i.e.,

$$\mathrm{Gamma}\left(r_{\mathrm{e}}, \frac{\rho p_{\mathrm{e}}}{1-p_{\mathrm{e}}}\right) \to \mathrm{Dist}, \qquad \mathrm{Beta}(\sigma_{\mathrm{on}}, \sigma_{\mathrm{off}}) \to \mathrm{Dist}. \tag{20}$$

It turns out that Dist is a normal distribution.

Next, we show that the Gamma distribution in Eq. (20) converges to a normal distribution as $N_\sigma \to \infty$. To achieve this, we introduce the moment generating function (MGF) for the random variable $\bar{y}$,

$$M_{\bar{y}}(t) = \langle \mathrm{e}^{t\bar{y}} \rangle.$$

Consider the linear transformation

$$\tilde{y} = \frac{\bar{y}-\mu}{\sigma},$$

where

$$\mu = f_{\mathrm{on}}, \qquad \sigma = \sqrt{\frac{f_{\mathrm{on}}(1-f_{\mathrm{on}})}{N_\sigma+1}},$$

and $\sqrt{r_{\mathrm{e}}} = \mu/\sigma$. The MGF of the transformed random variable $\tilde{y}$ is

$$M_{\tilde{y}}(t) = \langle \mathrm{e}^{t\tilde{y}} \rangle = \langle \mathrm{e}^{t(\frac{\bar{y}}{\sigma} - \sqrt{r_{\mathrm{e}}})} \rangle = \mathrm{e}^{-\sqrt{r_{\mathrm{e}}}t} \langle \mathrm{e}^{\frac{t\bar{y}}{\sigma}} \rangle = \mathrm{e}^{-\sqrt{r_{\mathrm{e}}}t} M_{\bar{y}}\left(\frac{t}{\sigma}\right).$$

Using the MGF of the Gamma distribution, this simplifies to

$$M_{\tilde{y}}(t) = \mathrm{e}^{-\sqrt{r_\mathrm{e}}t}\left(1 - \frac{t}{\sqrt{r_\mathrm{e}}}\right)^{-r_\mathrm{e}}.$$

As $N_\sigma \to \infty$, $r_\mathrm{e} \to \infty$, the expression further reduces to

$$\lim_{N_\sigma \to \infty} M_{\tilde{y}}(t) = \mathrm{e}^{\frac{t^2}{2}},$$

which is the MGF of the standard normal distribution $\mathcal{N}(0,1)$. Therefore, we conclude that

$$\lim_{N_\sigma \to \infty} \mathrm{Gamma}\left(r_\mathrm{e}, \frac{\rho p_\mathrm{e}}{1-p_\mathrm{e}}\right) = \mathcal{N}(\mu, \sigma).$$

Finally, we show that $\mathrm{Beta}(\sigma_\mathrm{on}, \sigma_\mathrm{off}) = \mathrm{Beta}(f_\mathrm{on}N_\sigma, (1-f_\mathrm{on})N_\sigma) \to \mathcal{N}(\mu, \sigma)$ as $N_\sigma \to \infty$. Note that we use the definition of the Beta distribution, $\mathrm{Beta}(\alpha, \beta)$, given by

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < x < 1, \tag{21}$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. The density function of this Beta distribution at the point $x = \mu + \sigma t$ is given by

$$f(\mu+\sigma t) = \underbrace{\frac{\Gamma(N_\sigma)}{\Gamma(f_\mathrm{on}N_\sigma)\Gamma((1-f_\mathrm{on})N_\sigma)}\mu^{f_\mathrm{on}N_\sigma-1}(1-\mu)^{(1-f_\mathrm{on})N_\sigma-1}}_{\mathcal{A}}\underbrace{\left(1+\frac{\sigma t}{\mu}\right)^{f_\mathrm{on}N_\sigma-1}\left(1-\frac{\sigma t}{1-\mu}\right)^{(1-f_\mathrm{on})N_\sigma-1}}_{\mathcal{B}}. \tag{22}$$

Using Stirling's formula, $\Gamma(x) \to \sqrt{2\pi/x}(x/\mathrm{e})^x$ as $x \to \infty$, the term $\mathcal{A}$ in Eq. (22) reduces to

$$\begin{aligned}\lim_{N_\sigma \to \infty} \mathcal{A} &= \frac{\sqrt{\frac{2\pi}{N_\sigma}}(N_\sigma/\mathrm{e})^{N_\sigma} {f_\mathrm{on}}^{f_\mathrm{on}N_\sigma-1}(1-f_\mathrm{on})^{N_\sigma(1-f_\mathrm{on})-1}}{\sqrt{\frac{2\pi}{f_\mathrm{on}N_\sigma}}(f_\mathrm{on}N_\sigma/\mathrm{e})^{f_\mathrm{on}N_\sigma}\sqrt{\frac{2\pi}{N_\sigma - f_\mathrm{on}N_\sigma}}[(N_\sigma - f_\mathrm{on}N_\sigma)/\mathrm{e}]^{N_\sigma - f_\mathrm{on}N_\sigma}} \\ &= \sqrt{\frac{N_\sigma}{2\pi f_\mathrm{on}(1-f_\mathrm{on})}} \\ &\approx \frac{1}{\sqrt{2\pi}\sigma}.\end{aligned} \tag{23}$$

Next, we rewrite the term $\mathcal{B}$ in Eq. (22) as

$$\mathcal{B} = \exp\left[(f_\mathrm{on}N_\sigma - 1)\ln\left(1+\frac{\sigma t}{\mu}\right) + (N_\sigma - f_\mathrm{on}N_\sigma - 1)\ln\left(1-\frac{\sigma t}{1-\mu}\right)\right].$$

Noting that $\sigma/\mu \to 0$ and $\sigma/(1-\mu) \to 0$ as $N_\sigma \to \infty$, we expand $\mathcal{B}$ around $t = 0$ using a Taylor series expansion, yielding

$$\begin{aligned}\mathcal{B} = &\exp\left[\underbrace{\sigma t\left(\frac{f_\mathrm{on}N_\sigma - 1}{\mu} - \frac{N_\sigma(1-f_\mathrm{on})-1}{1-\mu}\right)}_{\mathcal{B}_1} - \underbrace{\frac{\sigma^2 t^2}{2}\left(\frac{f_\mathrm{on}N_\sigma - 1}{\mu^2} + \frac{N_\sigma(1-f_\mathrm{on})-1}{(1-\mu)^2}\right)}_{\mathcal{B}_2}\right] \\ &\times \exp\left[\underbrace{O\left(\frac{(f_\mathrm{on}N_\sigma - 1)\sigma^3}{\mu^3} - \frac{(N_\sigma(1-f_\mathrm{on})-1)\sigma^3}{(1-\mu)^3}\right)t^3}_{\mathcal{B}_3}\right],\end{aligned} \tag{24}$$

where

$$\mathcal{B}_1 = \frac{t}{\sqrt{N_\sigma + 1}}\left(\sqrt{\frac{\sigma_\mathrm{on}}{\sigma_\mathrm{off}}} - \sqrt{\frac{\sigma_\mathrm{off}}{\sigma_\mathrm{on}}}\right),$$

$$\mathcal{B}_2 = \frac{t^2}{2(N_\sigma+1)}\left(N_\sigma - \frac{1-f_{\rm on}}{f_{\rm on}} - \frac{f_{\rm on}}{1-f_{\rm on}}\right) = \frac{t^2}{2} + O\left(\frac{1}{N_\sigma}\right)t^2,$$

and

$$\mathcal{B}_3 = O\left(\frac{1}{\sqrt{N_\sigma}}\right)t^3.$$

Using Eqs. (22)-(24), we obtain

$$f(x) = f(\mu+\sigma t) \approx \frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}\sigma} = \frac{\mathrm{e}^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma},$$

which is the probability density function of $N(\mu,\sigma)$. This completes the proof of Theorem 1.

4.2 Mathematical relationship between aeBIC and BIC

Since the data in set $\mathcal{X}$ (the set of count data $\{n_i\}$ for $i=1,\cdots,n_c$) are randomly sampled from the ground-truth distribution $\mathcal{G}$, the expectation of the BIC (Eq. (9)) is given by

$$\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathrm{BIC}(\mathcal{M},\mathcal{X})] = |\mathcal{M}|\ln n_c - 2\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}\left[\max_\theta \mathcal{L}_\theta\right] \le |\mathcal{M}|\ln n_c - 2\max_\theta \mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathcal{L}_\theta], \tag{25}$$

where the inequality follows from the fact that the expectation of the maximum is greater than or equal to the maximum of the expectation. A proof of this inequality is as follows.

For any fixed θ, it is true that

$$\mathcal{L}_\theta \le \max_{\theta'} \mathcal{L}_{\theta'}.$$

So taking expectations on both sides yields

$$\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathcal{L}_\theta] \le \mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\max_{\theta'} \mathcal{L}_{\theta'}].$$

Because this is true for any θ, it follows that

$$\max_\theta \mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathcal{L}_\theta] \le \mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\max_\theta \mathcal{L}_\theta].$$

Furthermore, we note that

$$\max_\theta \mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathcal{L}_\theta] = \max_\theta\left[\sum_{n=0}^{\infty}\sum_{i=1}^{n_c}\mathbf{1}_{n_i=n}\ln P_{\mathcal{M}}(n|\theta)\right] = n_c\max_\theta\left[\sum_{n=0}^{\infty}P_{\mathcal{G}}(n)\ln P_{\mathcal{M}}(n|\theta)\right] = -n_c\varepsilon_{\rm c}(\mathcal{M}_{\rm MLE},\mathcal{G}), \tag{26}$$

according to Eq. (11). Note that the first step in Eq. (26) is a reformulation of the log-likelihood function: the term $\mathbf{1}_{n_i=n}$ is an indicator function equal to 1 if and only if $n_i=n$, and hence $P_{\mathcal{G}}(n)=\sum_{i=1}^{n_c}\mathbf{1}_{n_i=n}$. From Eqs. (10), (25), and (26), it follows that

$$\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}[\mathrm{BIC}(\mathcal{M},\mathcal{X})] \le \mathrm{aeBIC}(\mathcal{M},n_c), \tag{27}$$

indicating that aeBIC serves as an upper bound for the expectation of the BIC. For a fixed distribution $\mathcal{M}$ with fixed parameters, $\mathcal{L}_\theta \to -n_c\varepsilon_{\rm c}(\mathcal{M},\mathcal{G})$ as $n_c\to\infty$, which further suggests

$$\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}\left[\max_\theta \mathcal{L}_\theta\right] \to \mathbf{E}_{\mathcal{X}\sim\mathcal{G}}\left[-\max_\theta n_c\varepsilon_{\rm c}(\mathcal{M},\mathcal{G})\right].$$

By noting that

$$\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}\left[-\max_\theta n_c\varepsilon_{\rm c}(\mathcal{M},\mathcal{G})\right] = -\max_\theta n_c\varepsilon_{\rm c}(\mathcal{M},\mathcal{G}),$$

one can conclude that

$$\mathbf{E}_{\mathcal{X}\sim\mathcal{G}}\left[\max_\theta \mathcal{L}_\theta\right] \to n_c\varepsilon_{\rm c}(\mathcal{M}_{\rm MLE},\mathcal{G})$$

as $n_c\to\infty$.

4.3 Derivation of the observed count distributions under the assumption of Beta-distributed capture rates

4.3.1 Telegraph model

The probability generating function (PGF) of the steady-state mRNA distribution of the telegraph model is given by

$$G_{\text{tele}}(z) = \sum_n z^n P(n) = {}_1F_1(\sigma_{\text{on}}, \sigma_{\text{on}} + \sigma_{\text{off}}, \rho(z-1)),$$

where $P(n)$ is given by Eq. (1). Assuming the transcript capture process in sequencing can be adequately described by the binomial capture model, i.e. assuming that each transcript can be captured with probability p_{cap} [10, 62], the PGF of the observed counts becomes

$$G_{\text{obs}}(z) = {}_1F_1(\sigma_{\text{on}}, \sigma_{\text{on}} + \sigma_{\text{off}}, p_{\text{cap}}\rho(z-1)). \tag{28}$$

This is because downsampling a general distribution with generating function $G(z)$ by the binomial capture model leads to the new generating function $G(1 - p_{\text{cap}}(1-z))$ (see [49], SI P.9 where it is shown for the special case of $p_{\text{cap}} = 1/2$; also see [37]).

If the capture rate p_{cap} follows a Beta distribution, $p_{\text{cap}} \sim \text{Beta}(a, b)$, defined by

$$P(p_{\text{cap}}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_{\text{cap}}^{a-1}(1-p_{\text{cap}})^{b-1},$$

the PGF of the observed counts is given by integrating Eq. (28) over the distribution of p_{cap}, which yields a generalized hypergeometric function

$$G_{\text{obs}}(z) = \int_0^1 {}_1F_1(\sigma_{\text{on}}, \sigma_{\text{on}} + \sigma_{\text{off}}, p_{\text{cap}}\rho(z-1))P(p_{\text{cap}})dp_{\text{cap}} = {}_2F_2(a, \sigma_{\text{on}}; a+b, \sigma_{\text{on}} + \sigma_{\text{off}}; \rho(z-1)). \tag{29}$$

Finally, we obtain the closed-form probability of observing n counts

$$P(n) = \frac{1}{n!}\frac{d^n}{dz^n}G_{\text{obs}}(z)\bigg|_{z=0} = \frac{\rho^n}{n!}\frac{(\sigma_{\text{on}})_n a_n}{(\sigma_{\text{on}} + \sigma_{\text{off}})_n (a+b)_n} {}_2F_2(a+n, \sigma_{\text{on}} + n; a+b+n, \sigma_{\text{on}} + \sigma_{\text{off}} + n; -\rho). \tag{30}$$

Note that this method of accounting for technical noise, i.e. via compound distributions, is essentially the same as that used to account for static extrinsic noise [92, 93].

4.3.2 Negative binomial model

The PGF of negative binomial distribution, $\text{NB}(r, p)$, is given by

$$G_{\text{NB}}(z) = (1 - (1-p)(z-1)/p)^{-r}.$$

If p_{cap} is the same in all cells, the PGF of the observed counts becomes

$$G_{\text{obs}}(z) = (1 - p_{\text{cap}}(1-p)(z-1)/p)^{-r}.$$

If we have a $\text{Beta}(a, b)$ distribution of p_{cap} then we integrate the above expression over the Beta distribution, yielding

$$G_{\text{obs}}(z) = \int_0^1 (1 - p_{\text{cap}}(1-p)(z-1)/p)^{-r} P(p_{\text{cap}})dp_{\text{cap}} = {}_2F_1(a, r; a+b; (1-p)(z-1)/p). \tag{31}$$

Finally, we obtain the probability distribution of the observed counts

$$P(n) = \frac{1}{n!}\frac{d^n}{dz^n}G_{\text{obs}}(z)\bigg|_{z=0} = \frac{\rho^n}{n!}\frac{(r)_n (a)_n}{(a+b)_n} {}_2F_1(a+n, r+n; a+b+n; -(1-p)/p).$$

4.3.3 Poisson model

The PGF of the Poisson distribution, Pois(λ), is given by

$$G_{\text{pois}}(z) = \exp(\lambda(z-1)).$$

If p_{cap} is the same in all cells, the PGF of the observed counts becomes

$$G_{\text{obs}}(z) = \exp(p_{\text{cap}}\lambda(z-1)).$$

If we have a Beta(a, b) distribution of p_{cap} then we integrate the above expression over the Beta distribution, yielding

$$G_{\text{obs}}(z) = {}_1F_1(a; a+b; \lambda(z-1)).$$

Finally, we obtain the probability distribution of the observed counts

$$P(n) = \frac{1}{n!}\frac{d^n}{dz^n}G_{\text{obs}}(z)\bigg|_{z=0} = \frac{\lambda^n}{n!}\frac{(a)_n}{(a+b)_n}{}_1F_1(a+n, a+b+n, -\lambda). \tag{32}$$

4.4 Inference of burst parameters assuming ideal conditions

We consider ideal conditions: (i) infinite sample size so that the measured moments of mRNA numbers do not suffer from finite sample size effects and thus the method of moments can be used for reliable inference; (ii) knowledge of the distribution describing cell-to-cell variability of the capture probability; this means that we can completely correct for technical noise.

4.4.1 Moments of the measured mRNA counts in scRNA-seq data

We assume that the distribution of the measured counts from a gene of interest is given by the telegraph model with capture probability sampled from a Beta(a, b) distribution, i.e. Eq. (30).

Given a PGF of observed counts $G_{\text{observed}}(z)$, the first two orders of moments can be derived as

$$\langle n\rangle = \frac{d}{dz}G_{\text{observed}}(z)\bigg|_{z=1}, \qquad \text{var}(n) = \frac{d^2}{dz^2}G_{\text{observed}}(z)\bigg|_{z=1} + \langle n\rangle - \langle n\rangle^2. \tag{33}$$

Substituting Eq. (29)into Eq. (33), we obtain expressions for the mean and the variance of the measured count distribution

$$\langle n\rangle_{\text{tele,observed}} = \frac{a\rho\sigma_{\text{on}}}{(a+b)(\sigma_{\text{on}}+\sigma_{\text{off}})}, \tag{34}$$

and

$$\text{Var}_{\text{tele,observed}}(n) = \frac{a(a+1)\rho^2\sigma_{\text{on}}(1+\sigma_{\text{on}})}{(a+b)(a+b+1)(\sigma_{\text{off}}+\sigma_{\text{on}})(1+\sigma_{\text{off}}+\sigma_{\text{on}})} + \frac{a\rho\sigma_{\text{on}}}{(a+b)(\sigma_{\text{off}}+\sigma_{\text{on}})} - \left(\frac{a\rho\sigma_{\text{on}}}{(a+b)(\sigma_{\text{off}}+\sigma_{\text{on}})}\right)^2. \tag{35}$$

It follows that the Fano factor of mRNA counts (variance divided by the mean) is given by

$$\text{FF}_{\text{observed}} = 1 + \langle p_{\text{cap}}\rangle\frac{\rho\sigma_{\text{off}}}{(\sigma_{\text{on}}+\sigma_{\text{off}})(1+\sigma_{\text{on}}+\sigma_{\text{off}})} + \text{CV}^2_{p_{\text{cap}}}\langle p_{\text{cap}}\rangle\frac{\rho(1+\sigma_{\text{on}})}{1+\sigma_{\text{on}}+\sigma_{\text{off}}}, \tag{36}$$

where $\langle p_{\text{cap}}\rangle$ and $\text{CV}_{p_{\text{cap}}}$ are the mean and the coefficient of variation (standard deviation divided by the mean) of the distribution of the capture probability, Beta(a, b).

4.4.2 Moments of the best fit model

We assume that we are in a region of parameter space where aeBIC selects the NB distribution with Beta-distributed capture probability as the optimal model. Substituting Eq. (31) into Eq. (33), we find expressions for the mean and variance of this model

$$\langle n\rangle_{\text{NB,observed}} = \frac{ar(1-p)}{(a+b)p}, \tag{37}$$

and

$$\text{Var}_{\text{NB,observed}}(n) = \frac{a(a+1)(1-p)^2r(r+1)}{(a+b)(a+b+1)p^2} + \frac{ar(1-p)}{(a+b)p} - \left(\frac{ar(1-p)}{(a+b)p}\right)^2. \tag{38}$$

4.4.3 Fitting the model to the data and parameter inference

Since we are assuming that the parameters of the Beta distribution, i.e. a and b, are known, the only parameters to be inferred are r and p. These two unknown parameters can be determined exactly by matching the mean and variance of the optimal model Eqs. (37)-(38) with the mean and variance of the data Eqs. (34)-(35). This leads to two simultaneous equations for r and p which when solved lead to expressions for the inferred two parameters of the NB distribution

$$r = \frac{\sigma_{\text{on}}(\sigma_{\text{on}} + \sigma_{\text{off}} + 1)}{\sigma_{\text{off}}}, \qquad p = \frac{(\sigma_{\text{on}} + \sigma_{\text{off}})(\sigma_{\text{on}} + \sigma_{\text{off}} + 1)}{(\sigma_{\text{on}} + \sigma_{\text{off}})(\sigma_{\text{on}} + \sigma_{\text{off}} + 1) + \rho\sigma_{\text{off}}}.$$

Note that these expressions do not have any dependence on the parameters of the technical noise (a and b), which indeed show that the inference method has perfectly corrected for technical noise. In fact, r and p determined in this way are none other than those parameterizing the effective NB distribution Eq. (6) that best fits the conventional telegraph model.

4.5 Parameter inference without exact knowledge of the capture rate distribution

Here, we describe an inference approach that simulates the workflow of an actual scRNA-seq experiment including the subsequent analysis of the data. Note that in the inference and model selection parts of the procedure, we do not assume that the distribution of the capture probability is known — rather we use the distribution of the measured total counts per cell as a proxy.

4.5.1 Simulating genomic data

We generated synthetic count data for a genome using the following procedure:

- For each cell $j = 1, \ldots, n_c$ ($n_c = 1000$), sample the capture rate $p_{\text{cap},j}$ from a Beta$(15, 35)$ distribution which has a mean of $\langle p_{\text{cap}} \rangle = 0.3$ and a CV of 0.21. This accounts for cell-to-cell variability in the transcript capture probability.
- Choose the fraction of time spent by genes in the active state from the set $f_{\text{on}} \in \{0.1, 0.2, \ldots, 0.7\}$, and fix $N_\sigma = 10$ and $\rho = 15$. Use the telegraph model as the ground-truth generative model for all genes.
- For each value of f_{on} and each cell j, sample the transcript counts of gene i, $x_{ij} \sim \text{Tele}(\rho, f_{\text{on}}N_\sigma, (1-f_{\text{on}})N_\sigma)$, independently for 400 genes. This results in transcript numbers per cell for a total of $n_g = 2800$ genes.
- Simulate observed counts in each cell by downsampling the counts according to the Binomial distribution: $y_{ij} \sim \text{Binomial}(x_{ij}, p_{\text{cap},j})$.

4.5.2 Parameter inference and model selection

- Compute the total observed count for each cell j as

$$t_j = \sum_{i=1}^{n_g} y_{ij}.$$

- Estimate the normalization factor β_j using the known mean capture rate $\langle p_{\text{cap}} \rangle$ (e.g., obtained via spike-in controls) [37]:

$$\beta_j = \frac{\langle p_{\text{cap}} \rangle \, t_j}{\langle t_j \rangle}.$$

- Use kernel density estimation (KDE) to estimate the distribution $p(\beta)$ of the normalization factors.
- For each gene i, estimate model parameters by minimizing the negative log likelihood across cells:

$$J(\theta_i) = -\sum_{j=1}^{n_c} \log P(y_{ij} \mid \theta_i), \tag{39}$$

where the likelihood $P(y_{ij} \mid \theta_i)$ is computed by marginalizing over β:

$$P(y_{ij} \mid \theta_i) = \int_0^\infty p(y_{ij} \mid \theta_i, \beta)\, p(\beta)\, \mathrm{d}\beta. \tag{40}$$

where $p(y_{ij} \mid \theta_i, \beta)$ for each candidate model is given by

- Telegraph: $\mathrm{Tele}(\rho\beta, \sigma_{\mathrm{on}}, \sigma_{\mathrm{off}})$, with $\theta_i = \{\rho, \sigma_{\mathrm{on}}, \sigma_{\mathrm{off}}\}$;
- Negative Binomial: $\mathrm{NB}(r, 1/(1 + b\beta))$, with $\theta_i = \{r, b\}$;
- Poisson: $\mathrm{Pois}(\lambda\beta)$, with $\theta_i = \{\lambda\}$.

- Evaluate the integral in Eq. (40) using Gauss quadrature for computational efficiency. Minimize the negative log likelihood using the Nelder-Mead algorithm.
- Compute the Bayesian Information Criterion (BIC) for model selection based on the log-likelihood in Eq. (39).
- If the NB model is selected, estimate the burst size and the burst frequency from the parameter estimates of the telegraph and negative binomial models. For the telegraph model, the burst size is $\rho/\sigma_{\mathrm{off}}$ and the burst frequency is σ_{on}. For the NB model, the estimates are given by $\hat{\beta}_f = r$ for the burst frequency and $\hat{\beta}_s = b$ for the burst size.
- Calculate the relative errors in the burst parameter estimates given the actual values used for simulating the genomic data. Randomly select 10^3 sets of two pairs of burst size and burst frequency estimates. In each set, rank genes by the magnitude of the burst frequency and separately by the burst size. Calculate the percentage of rankings that are flipped compared to the true ranking.

Supplementary Figures

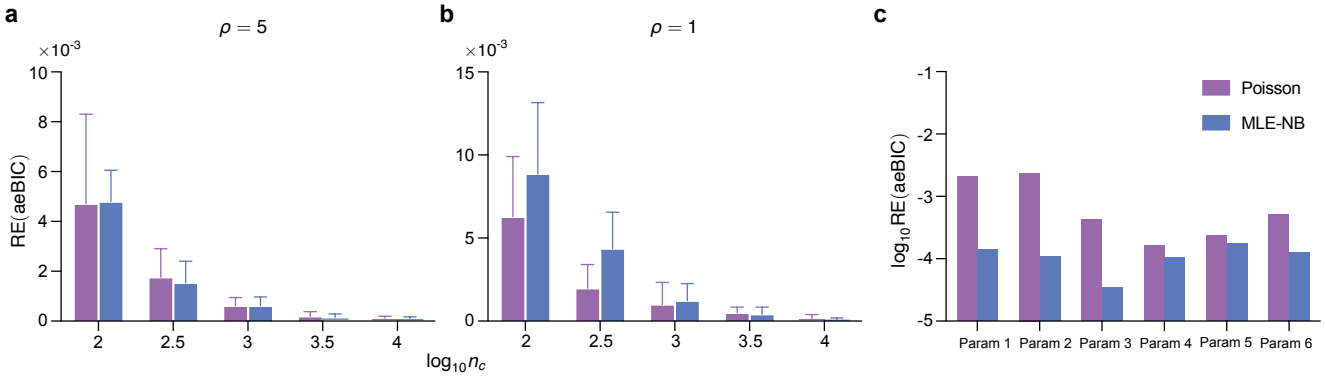

Figure S1: (a)(b) The relative error of aeBIC compared to the expectation of BIC ($\mathbf{E}$[BIC]) for two distributions (Poisson and NB) as a function of sample size n_c for 10 parameter sets (see Table 1 for the values of N_σ and f_{on}) with (a) $\rho = 5$ and (b) $\rho = 1$. Error bars show the standard error of the mean. (c) Relative error of aeBIC with respect to $\mathbf{E}$[BIC] for Poisson and NB distributions, evaluated for 6 parameter sets sampled from the plot corresponding to $n_c = 10^4$, $\rho = 15$ and $p_{\text{cap}} \sim \text{Beta}(60, 140)$ in Fig. 6c. The parameter sets are: $f_{\text{on}} = 0.8, N_\sigma = 3.8$ (Param 1), $f_{\text{on}} = 0.88, N_\sigma = 1.5$ (Param 2), $f_{\text{on}} = 0.5, N_\sigma = 38$ (Param 3), $f_{\text{on}} = 0.96, N_\sigma = 758.6$ (Param 4), $f_{\text{on}} = 0.7, N_\sigma = 380.2$ (Param 5), $f_{\text{on}} = 0.1, N_\sigma = 631$ (Param 6).

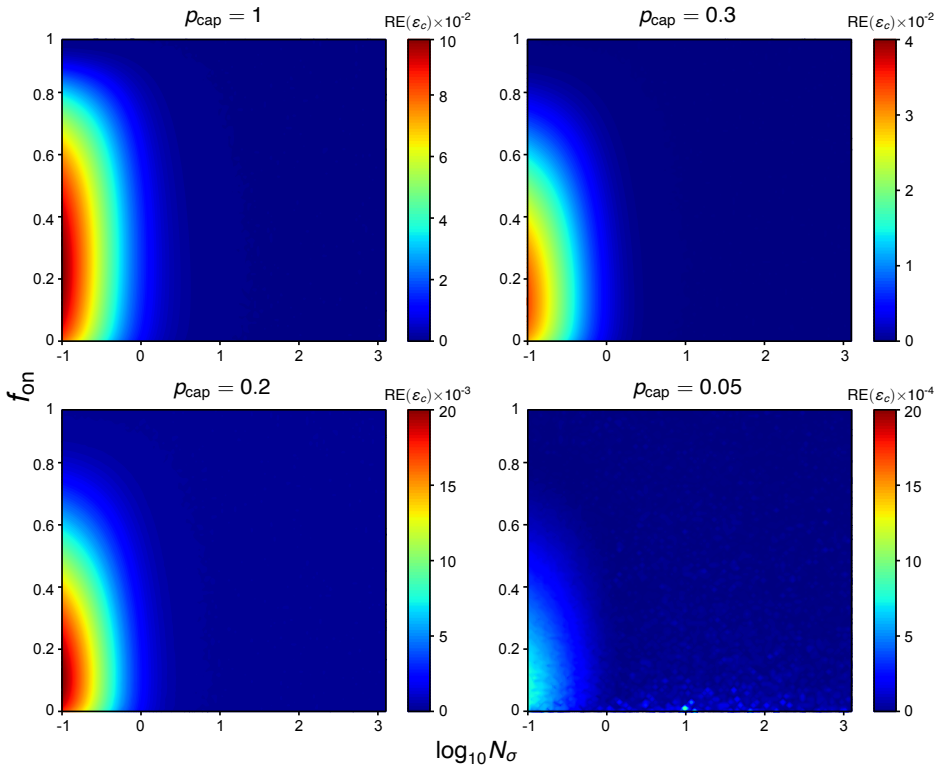

Figure S2: The relative error between the minimum cross-entropy determined by MLE and moment matching. The cross-entropy is that between the NB distribution as the proposed model ($P_{\mathcal{M}}(n|\theta)$) and the telegraph model distribution Eq. (1) with $\rho \to \rho p_{\text{cap}}$ as the ground-truth distribution ($P_{\mathcal{G}}(n)$), assuming all cells have an identical mRNA capture probability p_{cap}. The minimum cross-entropy using MLE is obtained by finding the two parameters of the NB distribution which minimize Eq. (11). The minimum cross-entropy using moment-matching is obtained by directly evaluating Eq. (11) using the effective NB distribution (Eq. (6)) with $\rho \to \rho p_{\text{cap}}$. As the heat maps show, the relative error between the two minimum cross-entropies is very small, regardless of the values of p_{cap}, f_{on} and N_σ. Note that ρ is in all cases fixed to 15.

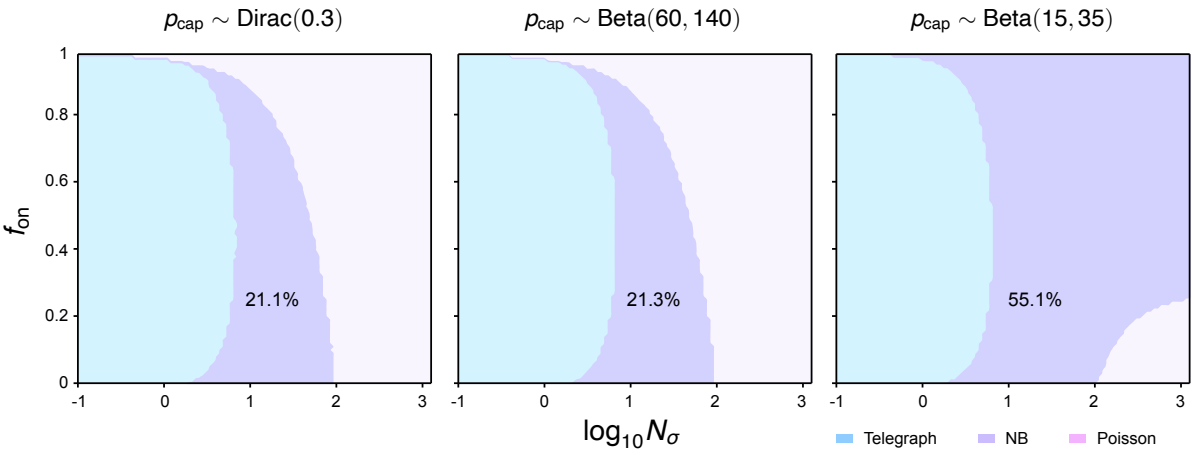

Figure S3: Phase diagram showing the regions of parameter space where the technical-noise corrected telegraph, NB and Poisson models are selected as the optimal models by the aeBIC, given that the ground-truth mRNA distribution is that of the corrected telegraph model, i.e. the telegraph model with p_{cap} sampled from Dirac(0.3), Beta(60, 140) and Beta(15, 35) — these distributions have mean 0.3 and CV = 0, 0.11 and 0.21, respectively. Here N_σ is the sum of gene-state switching rates normalised by the degradation rate of mRNA, and f_{on} is the fraction time spent in the active state. The fraction of the total parameter space occupied by the region where the corrected NB distribution is optimally selected is shown on the plots. Note that the transcription rate is fixed to $\rho = 15$. The maximum mean number of transcripts in the phase plots is 4.5. The sample size is fixed to $n_c = 10^4$.

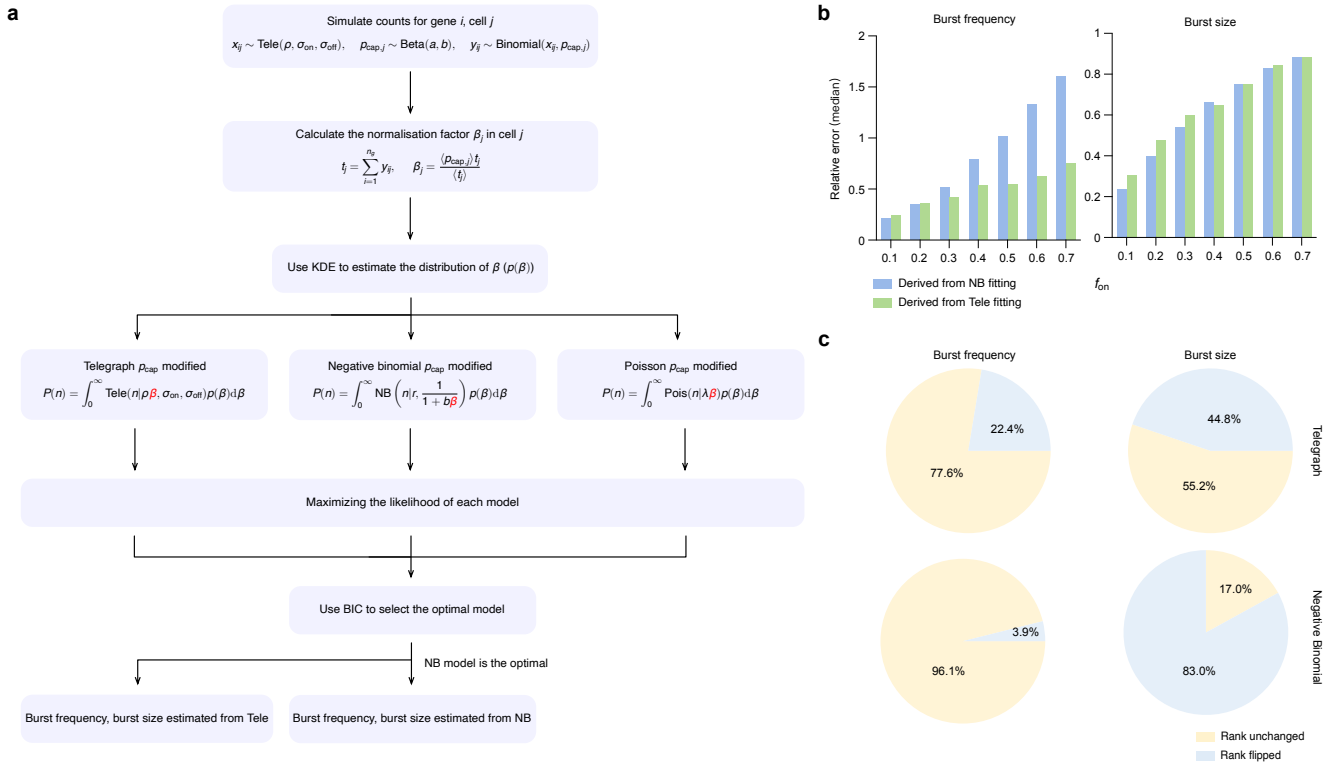

Figure S4: (a) Flowchart of the procedure described in Section 4.5 which simulates the workflow of a scRNA-seq experiment. A synthetic scRNA-seq dataset (1000 cells and 2800 genes) is simulated — technical noise is account for by sampling the capture probability for cell j ($p_{\text{cap},j}$) from a Beta$(15, 35)$ distribution. The data is then used to perform MLE-based parameter inference and BIC-based model selection. Notably, the latter is done without exact knowledge of the capture rate distribution, which mimics the real scenario. Only knowledge of the mean capture probability p_{cap} is assumed which can be estimated using spike-in controls or by comparison of scRNA-seq data with smFISH data. (b) Plots of the median relative error of the estimates of the burst frequency and burst size as a function of the true fraction of time spent in the active state f_{on} by genes. The estimates are for those genes whose distribution of counts per cell is best fit by an NB distribution (after correcting for technical noise). The estimates are obtained by maximizing the likelihood of the technical-noise corrected NB and telegraph models (blue and green bars, respectively). (c) 10^3 sets of two pairs of burst size and burst frequency estimates are randomly selected. In each set, genes are ranked by the magnitude of the burst frequency and separately by the magnitude of the burst size. The piecharts show the percentage of rankings that are flipped compared to the true ranking. The rankings are separately done using the parameter estimates obtained by maximizing the likelihood of the technical-noise corrected NB and telegraph models (piecharts in the bottom and top rows, respectively).

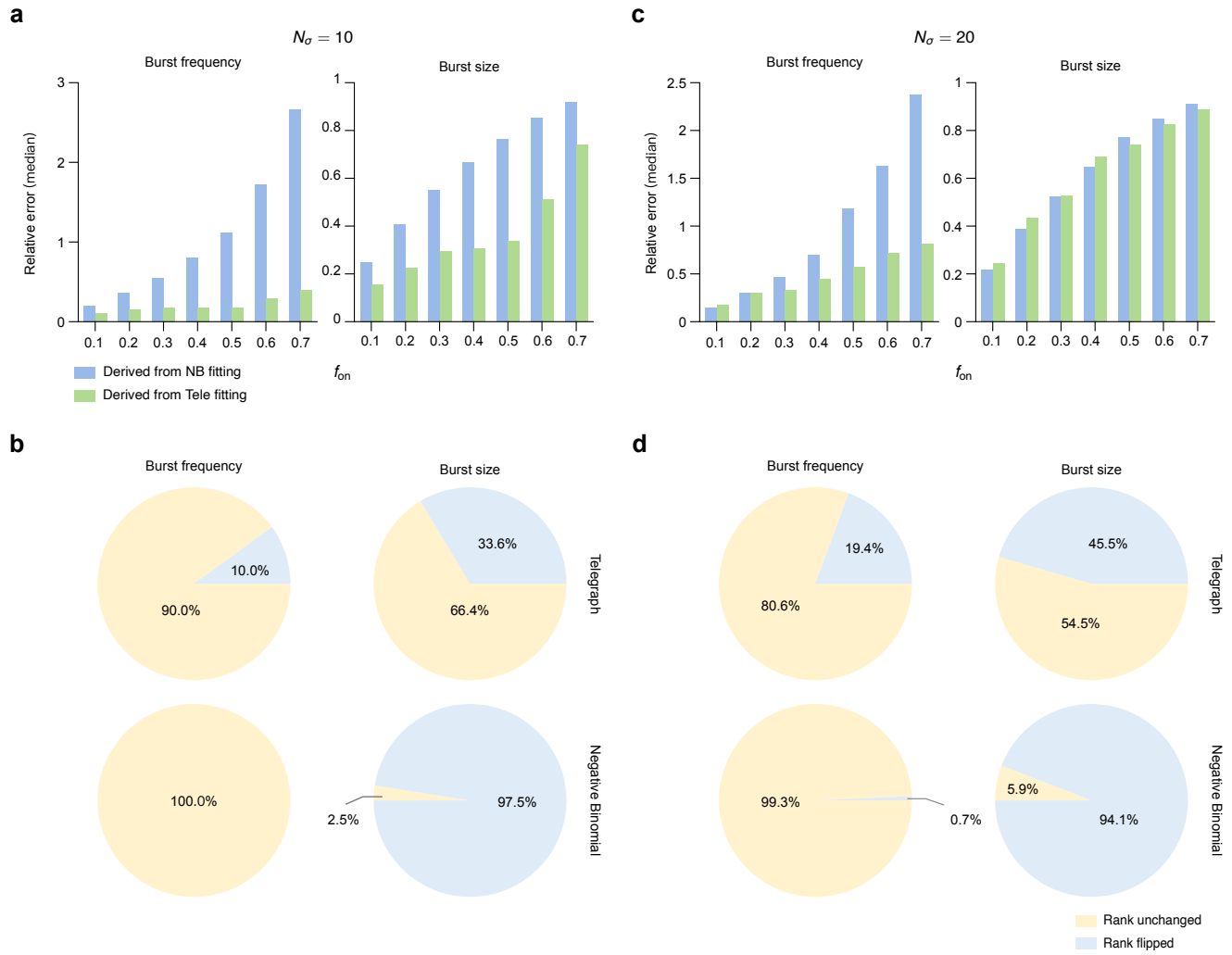

Figure S5: (a)(b)Recalculation of the statistics shown in Fig. S4(b)(c) using the procedure described in Section 4.5 with the difference that the number of cells is $n_c = 10^4$ and the transcript capture distribution is sampled from Beta$(60, 140)$. Note that $N_\sigma = 10$. (b)(d) Recalculation of the statistics shown in (a) and (b) with the difference that $N_\sigma = 20$. Note that 10^3 sets of two pairs of burst size and burst frequency estimates were used to compute the piecharts for the gene rankings in (b) and (d).

Table 1: Parameter sets used in Fig. 4b and Fig. S1a,b.

	Parameters	
	f_{on}	N_σ
Set 1	0.1	1
Set 2	0.4	10
Set 3	0.2	10
Set 4	0.7	100
Set 5	0.5	1
Set 6	0.9	0.1
Set 7	0.3	1000
Set 8	0.8	100
Set 9	0.6	0.1
Set 10	0.6	1000

Table 2: Comparison of the models selected by the aeBIC score and the BIC score for Fig. 4c with $n_c = 100$ cells. The minimum aeBIC score directly leads to a unique model that is shown in the first column. The BIC scores for the telegraph, NB and Poisson models are computed using MLE for each of 100 independent samples of size n_c generated by sampling from the Beta-Poisson formulation of the telegraph model Eq. (2) for the parameters shown in the 2nd and 3rd columns and $\rho = 15$; for each sample, the model with the smallest BIC score is selected (% of models selected are shown in the last 3 columns).

Region	Parameters		Percentage of selected models by BIC		
	f_{on}	N_σ	Telegraph	Negative Binomial	Poisson
Tele	0.8	0.3	100%	0%	0%
	0.5	0.1	100%	0%	0%
NB	0.4	10	7%	88%	5%
	0.2	7	7%	93%	0%
Pois	0.9	100	0%	1%	99%
	0.1	400	0%	1%	99%

References

[1] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, *et al.*, "mRNA-seq whole-transcriptome analysis of a single cell," *Nature Methods*, 6(5): 377–382, 2009.

[2] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lönnerberg, and S. Linnarsson, "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq," *Genome Research*, 21(7): 1160–1167, 2011.

[3] D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, *et al.*, "Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells," *Nature Biotechnology*, 30(8): 777–782, 2012.

[4] M. Hagemann-Jensen, C. Ziegenhain, P. Chen, D. Ramsköld, G.-J. Hendriks, A. J. Larsson, O. R. Faridani, and R. Sandberg, "Single-cell RNA counting at allele and isoform resolution using smart-seq3," *Nature Biotechnology*, 38(6): 708–714, 2020.

[5] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale, "Counting absolute numbers of molecules using unique molecular identifiers," *Nature Methods*, 9(1): 72–74, 2012.

[6] A. Sukys and R. Grima, "Cell-cycle dependence of bursty gene expression: Insights from fitting mechanistic models to single-cell RNA-seq data," *Nucleic Acids Research*, 53(7): gkaf295, 2025.

[7] S. Mangiola, E. A. Thomas, M. Modrák, A. Vehtari, and A. T. Papenfuss, "Probabilistic outlier identification for RNA sequencing generalized linear models," *NAR Genomics and Bioinformatics*, 3(1): lqab005, 2021.

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

[9] B. Vieth, C. Ziegenhain, S. Parekh, W. Enard, and I. Hellmann, "Powsimr: Power analysis for bulk and single cell RNA-seq experiments," *Bioinformatics*, 33(21): 3486–3488, 2017.

[10] W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, and V. Shahrezaei, "Baynorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data," *Bioinformatics*, 36(4): 1174–1181, 2020.

[11] D. Grün, L. Kester, and A. Van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nature Methods*, 11(6): 637–640, 2014.

[12] C. A. Vallejos, J. C. Marioni, and S. Richardson, "Basics: Bayesian analysis of single-cell sequencing data," *PLoS computational biology*, 11(6): e1004333, 2015.

[13] X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, and C. Trapnell, "Single-cell mrna quantification and differential analysis with census," *Nature methods*, 14(3): 309–315, 2017.

[14] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell rna-seq denoising using a deep count autoencoder," *Nature communications*, 10(1): 390, 2019.

[15] T. S. Andrews and M. Hemberg, "M3drop: Dropout-based feature selection for scrnaseq," *Bioinformatics*, 35(16): 2865–2867, 2019.

[16] M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang, "Saver: Gene expression recovery for single-cell rna sequencing," *Nature methods*, 15(7): 539–542, 2018.

[17] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression," *Genome Biology*, 20(1): 296, 2019.

[18] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature methods*, 15(12): 1053–1058, 2018.

[19] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature methods*, 11(7): 740–742, 2014.

[20] Z. S. Singer, J. Yong, J. Tischler, J. A. Hackett, A. Altinok, M. A. Surani, L. Cai, and M. B. Elowitz, "Dynamic heterogeneity and DNA methylation in embryonic stem cells," *Molecular Cell*, 55(2): 319–331, 2014.

[21] S. O. 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*, 5: e12175, 2016.

[22] A. Senecal, B. Munsky, F. Proux, N. Ly, F. E. Braye, C. Zimmer, F. Mueller, and X. Darzacq, "Transcription factors modulate c-fos transcriptional bursts," *Cell Reports*, 8(1): 75–83, 2014.

[23] X. Fu, H. P. Patel, S. Coppola, L. Xu, Z. Cao, T. L. Lenstra, and R. Grima, "Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions," *eLife*, 11: e82493, 2022.

[24] H. Ochiai, T. Sugawara, T. Sakuma, and T. Yamamoto, "Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells," *Scientific Reports*, 4(1): 7125, 2014.

[25] K. B. Halpern, S. Tanami, S. Landen, M. Chapal, L. Szlak, A. Hutzler, A. Nizhberg, and S. Itzkovitz, "Bursty gene expression in the intact mammalian liver," *Molecular Cell*, 58(1): 147–156, 2015.

[26] 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.

[27] E. Tunnacliffe and J. R. Chubb, "What is a transcriptional burst?" *Trends in Genetics*, 36(4): 288–297, 2020.

[28] B. Zoller, D. Nicolas, N. Molina, and F. Naef, "Structure of silent transcription intervals and noise characteristics of mammalian genes," *Molecular Systems Biology*, 11(7): 823, 2015.

[29] 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.

[30] T. Zhou and J. Zhang, "Analytical results for a multistate gene model," *SIAM Journal on Applied Mathematics*, 72(3): 789–818, 2012.

[31] Z. Cao, T. Filatova, D. A. Oyarzún, and R. Grima, "A stochastic model of gene expression with polymerase recruitment and pause release," *Biophysical Journal*, 119(5): 1002–1014, 2020.

[32] F. Jiao, J. Li, T. Liu, Y. Zhu, W. Che, L. Bleris, and C. Jia, "What can we learn when fitting a simple telegraph model to a complex gene expression model?" *PLOS Computational Biology*, 20(5): e1012118, 2024.

[33] A. G. Nicoll, J. Szavits-Nossan, M. R. Evans, and R. Grima, "Transient power-law behaviour following induction distinguishes between competing models of stochastic gene expression," *Nature Communications*, 16(1): 2833, 2025.

[34] J. Peccoud and B. Ycart, "Markovian modeling of gene-product synthesis," *Theoretical Population Biology*, 48(2): 222–234, 1995.

[35] J. K. Kim and J. C. Marioni, "Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data," *Genome Biology*, 14: 1–12, 2013.

[36] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mRNA synthesis in mammalian cells," *PLoS Biology*, 4(10): e309, 2006.

[37] W. Tang, A. C. S. Jørgensen, S. Marguerat, P. Thomas, and V. Shahrezaei, "Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics," *Bioinformatics*, 39(7): btad395, 2023.

[38] P. Trzaskoma, S. Jung, A. Pkowska, C. H. Bohrer, X. Wang, F. Naz, S. DellOrso, W. D. Dubois, A. Olivera, S. V. Vartak, *et al.*, "3d chromatin architecture, brd4, and mediator have distinct roles in regulating genome-wide transcriptional bursting and gene network," *Science Advances*, 10(32): eadl4893, 2024.

[39] F. Jiao, Q. Sun, M. Tang, J. Yu, and B. Zheng, "Distribution modes and their corresponding parameter regions in stochastic gene transcription," *SIAM Journal on Applied Mathematics*, 75(6): 2396–2420, 2015.

[40] D. Zenklusen, D. R. Larson, and R. H. Singer, "Single-RNA counting reveals alternative modes of gene expression in yeast," *Nature Structural & Molecular Biology*, 15(12): 1263–1271, 2008.

[41] M. Wang, J. Zhang, H. Xu, and I. Golding, "Measuring transcription at a single gene copy reveals hidden drivers of bacterial individuality," *Nature Microbiology*, 4(12): 2118–2127, 2019.

[42] B. Munsky, G. Neuert, and A. Van Oudenaarden, "Using gene expression noise to understand gene regulation," *Science*, 336(6078): 183–187, 2012.

[43] I. Brouwer and T. L. Lenstra, "Visualizing transcription: Key to understanding gene expression dynamics," *Current Opinion in Chemical Biology*, 51: 122–129, 2019.

[44] Z. Liu and R. Tjian, "Visualizing transcription factor dynamics in living cells," *Journal of Cell Biology*, 217(4): 1181–1191, 2018.

[45] C. R. Bartman, S. C. Hsu, C. C.-S. Hsiung, A. Raj, and G. A. Blobel, "Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping," *Molecular Cell*, 62(2): 237–247, 2016.

[46] J. Viñuelas, G. Kaneko, A. Coulon, E. Vallin, V. Morin, C. Mejia-Pous, J.-J. Kupiec, G. Beslon, and O. Gandrillon, "Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts," *BMC Biology*, 11: 1–19, 2013.

[47] H. Chen, M. Levo, L. Barinov, M. Fujioka, J. B. Jaynes, and T. Gregor, "Dynamic interplay between enhancer–promoter topology and gene activity," *Nature Genetics*, 50(9): 1296–1303, 2018.

[48] W.-K. Cho, N. Jayanth, B. P. English, T. Inoue, J. O. Andrews, W. Conway, J. B. Grimm, J.-H. Spille, L. D. Lavis, T. Lionnet, *et al.*, "RNA polymerase ii cluster dynamics predict mRNA output in living cells," *eLife*, 5: e13617, 2016.

[49] 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.

[50] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, 8(1): 14049, 2017.

[51] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, 161(5): 1202–1214, 2015.

[52] R. Grima and P.-M. Esmenjaud, "Quantifying and correcting bias in transcriptional parameter inference from single-cell data," *Biophysical Journal*, 123(1): 4–30, 2024.

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

[54] N. Friedman, L. Cai, and X. S. Xie, "Linking stochastic dynamics to population distribution: An analytical framework of gene expression," *Physical Review Letters*, 97(16): 168302, 2006.

[55] C. Jia, "Simplification of markov chains with infinite state space and the mathematical theory of random gene expression bursts," *Physical Review E*, 96(3): 032402, 2017.

[56] Q. F. Wills, K. J. Livak, A. J. Tipping, T. Enver, A. J. Goldson, D. W. Sexton, and C. Holmes, "Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments," *Nature Biotechnology*, 31(8): 748–752, 2013.

[57] P. Thomas, N. Popovi, and R. Grima, "Phenotypic switching in gene regulatory networks," *Proceedings of the National Academy of Sciences*, 111(19): 6994–6999, 2014.

[58] K. Öcal, G. Sanguinetti, and R. Grima, "Model reduction for the chemical master equation: An information-theoretic approach," *The Journal of Chemical Physics*, 158(11): 2023.

[59] R. Jiang, T. Sun, D. Song, and J. J. Li, "Statistics or biology: The zero-inflation controversy about scRNA-seq data," *Genome Biology*, 23(1): 31, 2022.

[60] C. Jia, "Kinetic foundation of the zero-inflated negative binomial model for single-cell rna sequencing data," *SIAM Journal on Applied Mathematics*, 80(3): 1336–1355, 2020.

[61] Y. Cao, S. Kitanovski, R. Küppers, and D. Hoffmann, "Umi or not umi, that is the question for scRNA-seq zero-inflation," *Nature Biotechnology*, 39(2): 158–159, 2021.

[62] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, 161(5): 1187–1201, 2015.

[63] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard, "Comparative analysis of single-cell RNA sequencing methods," *Molecular Cell*, 65(4): 631–643, 2017.

[64] R. Salomon, D. Kaczorowski, F. Valdes-Mora, R. E. Nordon, A. Neild, N. Farbehi, N. Bartonicek, and D. Gallego-Ortega, "Droplet-based single cell RNAseq tools: A practical guide," *Lab on a Chip*, 19(10): 1706–1727, 2019.

[65] 10x Genomics, *What fraction of mRNA transcripts are captured per cell?* `https://kb.10xgenomics.com/hc/en-us/articles/360001539051-What-fraction-of-mRNA-transcripts-are-captured-per-cell`, Accessed: April 2, 2025, 10x Genomics Knowledge Base, 2025.

[66] 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, 2012.

[67] J. Paulsson, O. G. Berg, and M. Ehrenberg, "Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation," *Proceedings of the National Academy of Sciences*, 97(13): 7148–7153, 2000.

[68] B. Munsky, G. Li, Z. R. Fox, D. P. Shepherd, and G. Neuert, "Distribution shapes govern the discovery of predictive models for gene regulation," *Proceedings of the National Academy of Sciences*, 115(29): 7533–7538, 2018.

[69] L. Amrhein, K. Harsha, and C. Fuchs, "A mechanistic model for the negative binomial distribution of single-cell mrna counts," *bioRxiv* : 657619, 2019.

[70] A. Singh and P. Bokes, "Consequences of mrna transport on stochastic variability in protein levels," *Biophysical journal*, 103(5): 1087–1096, 2012.

[71] N. Kumar, T. Platini, and R. V. Kulkarni, "Exact distributions for stochastic gene expression models with bursting and feedback," *Physical review letters*, 113(26): 268105, 2014.

[72] N. Kumar, A. Singh, and R. V. Kulkarni, "Transcriptional bursting in gene expression: Analytical results for general stochastic models," *PLoS computational biology*, 11(10): e1004292, 2015.

[73] C. Jia and R. Grima, "Small protein number effects in stochastic models of autoregulated bursty gene expression," *The Journal of chemical physics*, 152(8): 2020.

[74] G. Gorin and L. Pachter, "Special function methods for bursty models of transcription," *Physical Review E*, 102(2): 022409, 2020.

[75] C. Jia and R. Grima, "Frequency domain analysis of fluctuations of mrna and protein copy numbers within a cell lineage: Theory and experimental validation," *Physical Review X*, 11(2): 021032, 2021.

[76] G. Gorin and L. Pachter, "Modeling bursty transcription and splicing with the chemical master equation," *Biophysical Journal*, 121(6): 1056–1069, 2022.

[77] B. Wu, J. Holehouse, R. Grima, and C. Jia, "Solving the time-dependent protein distributions for autoregulated bursty gene expression using spectral decomposition," *The Journal of Chemical Physics*, 160(7): 2024.

[78] G. Gorin, S. Yoshida, and L. Pachter, "Assessing markovian and delay models for single-nucleus rna sequencing," *Bulletin of Mathematical Biology*, 85(11): 114, 2023.

[79] T. Chari, G. Gorin, and L. Pachter, "Biophysically interpretable inference of cell types from multimodal sequencing data," *Nature Computational Science*, 4(9): 677–689, 2024.

[80] A. J. 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(7738): 251–254, 2019.

[81] S. Luo, Z. Zhang, Z. Wang, X. Yang, X. Chen, T. Zhou, and J. Zhang, "Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model," *Royal Society Open Science*, 10(4): 221057, 2023.

[82] S. Luo, Z. Wang, Z. Zhang, T. Zhou, and J. Zhang, "Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics," *Nucleic Acids Research*, 51(1): 68–83, 2023.

[83] D. Ramsköld, G.-J. Hendriks, A. J. Larsson, J. V. Mayr, C. Ziegenhain, M. Hagemann-Jensen, L. Hartmanis, and R. Sandberg, "Single-cell new rna sequencing reveals principles of transcription at the resolution of individual bursts," *Nature Cell Biology*, 26(10): 1725–1733, 2024.

[84] D. Schnoerr, G. Sanguinetti, and R. Grima, "Approximation and inference methods for stochastic biochemical kineticsa tutorial review," *Journal of Physics A: Mathematical and Theoretical*, 50(9): 093001, 2017.

[85] L. V. Sharova, A. A. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M. S. Ko, "Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells," *DNA research*, 16(1): 45–58, 2009.

[86] A. Waisman, F. Sevlever, M. Elías Costa, M. S. Cosentino, S. G. Miriuka, A. C. Ventura, and A. S. Guberman, "Cell cycle dynamics of mouse embryonic stem cells in the ground state and during transition to formative pluripotency," *Scientific reports*, 9(1): 8051, 2019.

[87] A. Riba, A. Oravecz, M. Durik, S. Jiménez, V. Alunni, M. Cerciat, M. Jung, C. Keime, W. M. Keyes, and N. Molina, "Cell cycle gene regulation dynamics revealed by rna velocity and deep-learning," *Nature communications*, 13(1): 2865, 2022.

[88] A. R. Lederer, M. Leonardi, L. Talamanca, D. M. Bobrovskiy, A. Herrera, C. Droin, I. Khven, H. J. Carvalho, A. Valente, A. Dominguez Mantes, *et al.*, "Statistical inference with a manifold-constrained rna velocity model uncovers cell cycle speed modulations," *Nature methods* : 1–16, 2024.

[89] M. Fang, G. Gorin, and L. Pachter, "Trajectory inference from single-cell genomics data with a process time model," *PLOS Computational Biology*, 21(1): e1012752, 2025.

[90] S. Peidli, T. D. Green, C. Shen, T. Gross, J. Min, S. Garda, B. Yuan, L. J. Schumacher, J. P. Taylor-King, D. S. Marks, *et al.*, "Scperturb: Harmonized single-cell perturbation data," *Nature Methods*, 21(3): 531–540, 2024.

[91] D. Karlis and E. Xekalaki, "Mixed poisson distributions," *International Statistical Review/Revue Internationale de Statistique*, 73(1): 35–58, 2005.

[92] L. Ham, R. D. Brackston, and M. P. Stumpf, "Extrinsic noise and heavy-tailed laws in gene expression," *Physical review letters*, 124(10): 108101, 2020.

[93] L. Ham, M. Jackson, and M. P. Stumpf, "Pathway dynamics can delineate the sources of transcriptional noise in gene expression," *Elife*, 10: e69324, 2021.