

 Open access • Journal Article • DOI:10.1137/19M1253198

Kinetic Foundation of the Zero-Inflated Negative Binomial Model for Single-Cell RNA Sequencing Data — [Source link](#)

Chen Jia

Published on: 28 May 2020 - [Siam Journal on Applied Mathematics](#) (Society for Industrial and Applied Mathematics)

Topics: Overdispersion

Related papers:

- [Exact protein distributions for stochastic models of gene expression using partitioning of Poisson processes.](#)
- [Inferring steady state single-cell gene expression distributions from analysis of mesoscopic samples](#)
- [Length Biases in Single-Cell RNA Sequencing of pre-mRNA](#)
- [Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics](#)
- [Contrastive latent variable modeling with application to case-control sequencing experiments](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/kinetic-foundation-of-the-zero-inflated-negative-binomial-2tl181u25y>

Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data

Chen Jia^{1,2}

¹ Division of Applied and Computational Mathematics, Beijing Computational Science Research Center, Beijing 100193, China.

² Department of Mathematics, Wayne State University, Detroit, MI 48202, U.S.A.

Email: chenjia@wayne.edu

Abstract

Single-cell RNA sequencing data have complex features such as dropout events, over-dispersion, and high-magnitude outliers, resulting in complicated probability distributions of mRNA abundances that are statistically characterized in terms of a zero-inflated negative binomial (ZINB) model. Here we provide a mesoscopic kinetic foundation of the widely used ZINB model based on the biochemical reaction kinetics underlying transcription. Using multiscale modeling and simplification techniques, we show that the ZINB distribution of mRNA abundance and the phenomenon of transcriptional bursting naturally emerge from a three-state stochastic transcription model. We further reveal a nontrivial quantitative relation between dropout events and transcriptional bursting, which provides novel insights into how and to what extent the burst size and burst frequency could reduce the dropout rate. Three different biophysical origins of over-dispersion are also clarified at the single-cell level.

Keywords: dropout, over-dispersion, transcriptional bursting, stochastic gene expression, chemical master equation, multiscale modeling, model simplification

1 Introduction

Gene expression in living cells is a complex stochastic process, resulting in spontaneous random fluctuations in mRNA and protein abundances [1]. Recent technological advances in single-cell RNA sequencing (scRNA-seq) have made it possible to measure mRNA expression and provide transcriptome profiles at the single-cell level [2–5]. Compared with traditional bulk RNA sequencing which measures the average mRNA expression levels across millions of cells, scRNA-seq enables the dissection of gene expression heterogeneity in different cell populations and tissues, and thus allows the investigation of many fundamental biological questions such as the identification of novel cell types, the classification of cell subtypes, and the reconstruction of cellular developmental trajectories [6].

Stochasticity in gene expression measurements has two fundamental origins: (i) the intrinsic noise due to small copy numbers of biochemical molecules and random collisions between them, giving rise to various probabilistic chemical reactions [1], and (ii) the extrinsic noise due to limitations of current experimental techniques. Although scRNA-seq provides a new level of data resolution, it also produces a much higher noise level than bulk-level measurements. A remarkable characteristic of scRNA-seq data is the high frequency of zero read counts [7, 8]. Given the excessive amount of zero observations in scRNA-seq data, it is important to distinguish between (i) the structural (true) zeros where the genes are truly unexpressed and (ii) the dropout (false) zeros where the genes are actually expressed but fail to be detected [9–13]. While the former is due to intrinsic biological variability, the latter, which is referred to as dropout events, is due to extrinsic technical reasons.

Due to the tiny amount of mRNA in an individual cell, the input material needs to be captured with low efficiency and go through many rounds of amplification before being sequenced. This results in low mRNA capture rate and strong amplification bias, as well as dropout events [14]. To be more specific, the workflow of scRNA-seq experiments includes the following steps: isolation of single cells, cell lysis while preserving mRNA, mRNA capture, reverse transcription of primed RNA into cDNA, cDNA amplification, library preparation, and sequencing [15]. During these steps, possible technical reasons leading to dropouts include mRNA degradation after cell lysis, low efficiency of mRNA capture, reverse transcription, and cDNA amplification, library size differences, and sequencing depth [10]. Recent studies [15] have shown that the efficiency for poly-adenylated mRNA species to be captured, converted into cDNA, and amplified can range between 10% and 40%, depending on the study. This means that if the starting transcripts in an individual cell are in low amount, there is a certain probability that they will not be detected by current scRNA-seq methods.

Besides the dropout effect, other characteristics of scRNA-seq data include over-dispersion [16] and high-magnitude outliers [17] due to the stochastic nature of gene expression at the single-cell level and the related phenomenon of transcriptional bursting [15]. Given these complex features of scRNA-seq data, recent studies have highlighted the need to develop novel statistical and computational methods in data analysis, especially differential expression analysis [5]. When handling dropout events, a popular perspective held by the bioinformatic field is that the complicated probability distributions of mRNA abundances in a cell population need to be explicitly characterized by a global zero-inflation parameter, resulting in various zero-inflated models [18, 19]. Among these statistical models, the zero-inflated negative binomial (ZINB) model is the most widely used [20–31], where the zero-inflated part describes dropouts and the negative binomial part accounts for over-dispersion. Some other commonly used models are listed in Sec. 7.

Modern sciences emphasize quantitative characterization of experimental observations, which is widely known as mathematical modeling. Along this line, two types of modeling methods should be distinguished: data-driven and mechanism-based modeling [32]. The former explains experimental phenomena in terms of data analysis based on various mathematical formulas and statistical models, while the latter understands the world in terms of mathematical deductions based on various mechanisms and scientific laws. The ZINB model of scRNA-seq data proposed in previous studies belongs to the former category.

In the present work, we provide a mesoscopic kinetic foundation of the widely used ZINB model based on the stochastic biochemical reaction kinetics underlying transcription. In fact, many stochastic models of transcription dynamics have been proposed [33–40]. Although some previous models could provide a clear explanation of over-dispersion, very few of them have incorporated the dropout effect into their model assumptions. So far, there is still a lack of a kinetic basis for the ZINB distribution of mRNA abundance. In addition, it is widely believed that the complex features of scRNA-seq data are closely related to the phenomenon of transcriptional bursting. However, the quantitative relationship among dropout events, over-dispersion, and transcriptional bursting still remains unclear. The present paper addresses these issues.

2 A novel three-state model of transcription

Based on the central dogma of molecular biology, the transcription of a gene in an individual cell has a standard two-stage representation involving the switching of the gene between an active and an inactive epigenetic state and the synthesis of the mRNA from the gene [1]. In the active state, the gene produces the mRNA. When the gene is inactive, the process of mRNA synthesis is terminated. Due to various technical factors in scRNA-seq experiments such as low mRNA capture rate, amplification bias, and sequencing depth, at a particular time, the mRNA expression in a single cell can be either detectable or undetectable [31]. As a result, it is reasonable to assume that the gene of interest can exist in a third state, referred to as the dropout state, where the mRNA expression of this gene cannot be detected due to technical reasons. Here the dropout state should not be regarded as an epigenetic conformation of the gene. Instead, it characterizes an undetectable state where the transcriptional signal of the gene is missing. These considerations lead to the three-state transcription model illustrated in Fig. 1(a), where a transcript can be synthesized with rate s or be degraded with rate v , and the gene can switch among the active, inactive, and dropout states with certain switching rates a_i and b_i , $i = 1, 2, 3$. Compared with the classical two-state transcriptional model without the dropout state [1], the cyclic structure of gene state switching will remarkably increase the theoretical complexity, as we shall see.

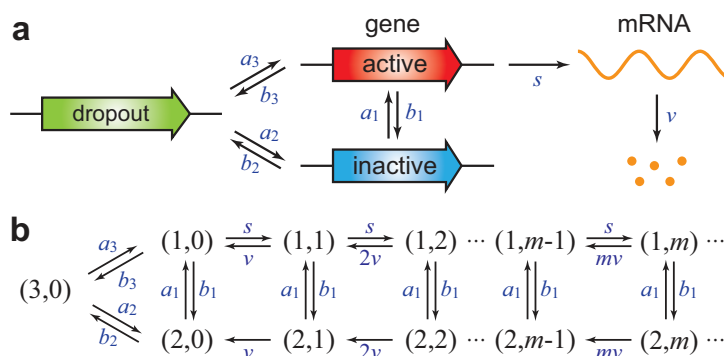


Fig. 1. **Stochastic transcription kinetics in individual cells with dropout events.** (a) A three-state transcription model involving gene switching among the active, inactive, and dropout states. Here the dropout state characterizes the detection state where the mRNA expression of this gene is undetectable. (b) Transition diagram of the Markovian model whose dynamics is governed by the chemical master equation.

From the chemical perspective, the microstate of the gene of interest can be described by an ordered pair (i, m) : the state i of the gene and the copy number m of detectable transcripts, where $i = 1, 2, 3$ correspond to the active, inactive, and dropout states, respectively. Then the stochastic dynamics of our three-state transcription model can be described by the Markov jump process (continuous-time Markov chain) with transition diagram illustrated in Fig. 1(b). Since the transcriptional signal is missing when a dropout occurs, it is reasonable to assume that the dropout state can only exist with zero detectable transcript, described by the microstate $(3, 0)$.

Experimentally, it was widely observed that the dropout rate for a given cell strongly depends on its expression level, with dropouts being more frequent for cells with low mRNA expression levels [17]. In general, the total content of mRNA in a single cell is low (0.01-2.5pg per cell) [41] and most genes only transcribe a small copy number of mRNA [42]. Due to the tiny amount of mRNA in an individual cell, the input material needs to be captured with low efficiency and go through many rounds of amplification before being sequenced. This results in low mRNA capture rate and strong amplification bias, as well as

dropout events [26]. As a result, microstates $(1, m)$ and $(2, m)$ with tiny mRNA abundance m are more likely to convert to the dropout microstate $(3, 0)$. In our Markovian model, for simplicity, we assume that $(3, 0)$ can only be reached from $(1, m)$ and $(2, m)$ with $m = 0$ (Fig. 1(b)). In Sec. 7, a removal of this assumption will be discussed and a more realistic model will be given.

There is another reason leading us to consider the three-state transcription model. Recent single-cell experiments have provided evidence that for many genes, more than two states may participate in the transcription process [43–47]. In fact, if a gene can only switch between the active and inactive states, then the sojourn times in the active and inactive states should be exponentially distributed. However, recent single-cell time-lapse measurements in eukaryotic cells [43, 44] have indicated that the sojourn time in the inactive state may have a non-exponential peaked distribution. This indicates that the gene dynamics in the inactive state may contain at least two exponential steps, so that in sum the gene would undergo a three-state switching process.

In particular, in two recent studies, the authors monitored gene expression dynamics in mouse fibroblasts [43] and Chinese hamster ovary cells [47] using single-cell time-lapse microscopy and found that both data sets were well described by a three-state gene expression model involving gene switching among an active, an inactive (reversibly silent), and a refractory (irreversibly silent) state. The difference between the inactive and refractory states is that the former has a good chance to switch back to the active state, while the possibility for the latter to switch back is much lower. In the inactive or refractory state, RNA polymerases could either be absent from the promoter or present in a paused state. Therefore, the dropout state in our three-state transcription model may have two different interpretations: It may either correspond to an undetectable state due to purely technical factors or correspond to a refractory state due to real biological factors.

Let $p_{i,m}(t)$ denote the probability of having m detectable transcripts at time t when the gene is in state i . Then the evolution of the Markovian model is governed by the chemical master equation

$$\begin{cases} \dot{p}_{1,0} = a_1 p_{2,0} + a_3 p_{3,0} + v p_{1,1} - (b_1 + b_3 + s) p_{1,0}, \\ \dot{p}_{2,0} = b_1 p_{1,0} + a_2 p_{3,0} + v p_{2,1} - (a_1 + b_2) p_{2,0}, \\ \dot{p}_{3,0} = b_3 p_{1,0} + b_2 p_{2,0} - (a_2 + a_3) p_{3,0}, \\ \dot{p}_{1,m} = a_1 p_{2,m} + s p_{1,m-1} + (m+1) v p_{1,m+1} - (b_1 + s + m v) p_{1,m}, \quad m \geq 1, \\ \dot{p}_{2,m} = b_1 p_{1,m} + (m+1) v p_{2,m+1} - (a_1 + m v) p_{2,m}, \quad m \geq 1. \end{cases}$$

Here s is the transcription rate; v is the degradation rate of the mRNA; a_i and b_i , $i = 1, 2, 3$ are the switching rates of the gene among the three states. Since (i, m) represents the microstate of having m transcripts in a single cell when the gene is in state i and each transcript can be degraded with rate v , the transition rate from microstate (i, m) to microstate $(i, m-1)$, which represents the total degradation rate of the m transcripts, should be $m v$ (Fig. 1(b)) [1]. In addition, since the dropout state could describe a refractory state, which has a lower chance to switch back to the active state than the inactive state, it is natural to assume $a_1 > a_3$ in our model.

3 Model simplification via decimation

One of the most important reasons for over-dispersion of bulk and single-cell RNA-seq data is transcriptional bursting, also known as transcriptional pulsing [15], which describes the phenomenon of relatively short transcriptionally active and high expression periods followed by longer transcriptionally

silent and low expression periods [43, 48], resulting in spontaneous temporal fluctuations of transcript levels (Fig. 2(a)).

In general, transcriptional bursting results from multiple time scales underlying the transcription process [49]. In fact, the mechanism of transcriptional bursting has been described by Paulsson in his review paper [1], “If genes are mostly inactive but transcribe a large number of mRNAs while in the active state, transcription could occur in bursts”. Intuitively, if we require the gene to be mostly inactive, the switching rate b_1 of the gene from the active to the inactive state should be much larger than the reverse switching rate a_1 from the inactive to the active state. On the other hand, if we require the gene to transcribe a large number of transcripts while in the active state, the transcription rate s should be very large, at least at the same order of magnitude as the switching rate b_1 . These considerations lead to the following biochemical conditions for transcriptional bursting: $b_1 \gg a_1$ and s/b_1 is finite. Here, by saying that s/b_1 is finite, we mean that s and b_1 are roughly at the same order of magnitude. When the gene is active, the large transcription rate s will give rise to fast accumulation of mRNA. Once the gene becomes inactive, the transcription process is terminated and transcripts will be degraded until the gene becomes active again. We stress here that the above biochemical conditions imposed on the rate constants are consistent with a recent single-cell experiment on transcriptional bursting [43], where the authors monitored the transcription kinetics in mouse fibroblasts using time-lapse bioluminescence imaging and found that the three rate constants a_1 , b_1 , and s across different genes are typically at the magnitude of 0.01/min, 0.1/min, and 1/min, respectively.

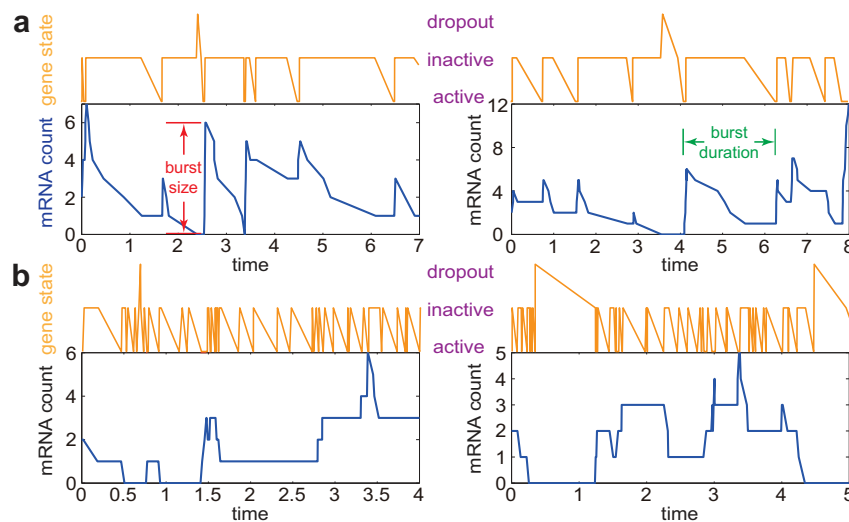


Fig. 2. Numerical simulations of the stochastic trajectories of the gene state and mRNA copy number based on the original Markovian model under two sets of biologically relevant parameters. (a) Two typical trajectories when the mean burst size h and maximum burst frequency λ are moderate. The model parameters are chosen as $h = 3$, $\lambda = 1.5$, $s = hb_1$, $v = 1$, $a_1 = \lambda v$, $b_1 = 25$, $a_2 = 1$, $b_2 = 4$, $a_3 = 0$, $b_3 = 1$. (b) Two typical trajectories in the limiting case of $h \rightarrow 0$ and $\lambda \rightarrow \infty$, while $\lambda h = \gamma$ is kept as a constant. The model parameters are chosen as $h = 0.2$, $\lambda = 10$, $s = hb_1$, $v = 1$, $a_1 = \lambda v$, $b_1 = 100$, $a_2 = 1$, $b_2 = 4$, $a_3 = 0$, $b_3 = 1$.

Due to the timescale separation of the underlying biochemical reaction kinetics, our Markovian model can be simplified to a much simpler one. To see this, let $\beta = b_1/a_1 \gg 1$ denote the ratio of the switching rates between the active and inactive states. Moreover, let $q_{(i,m),(i',m')}$ denote the transition rate of the Markovian model from microstate (i, m) to microstate (i', m') and let

$$q_{(i,m)} = \sum_{(i',m') \neq (i,m)} q_{(i,m),(i',m')}$$

denote the rate at which the system leaves microstate (i, m) , which is defined as the sum of transition rates from (i, m) to other microstates. Since $\beta \gg 1$, we say that (i, m) is a fast state if

$$\lim_{\beta \rightarrow \infty} q_{(i,m)} = \infty$$

and we say that (i, m) is a slow state if

$$\lim_{\beta \rightarrow \infty} q_{(i,m)} < \infty.$$

If (i, m) is a fast state, then the time that the system stays in this state will be very short. Since $b_1 \gg a_1$ and s/b_1 is finite, we write $b_1 = \beta a_1$ and $s = \beta a_1 (s/b_1)$, where $\beta \gg 1$ and we treat a_1 and s/b_1 as constants. Here b_1 and s are the only two model parameters depending on β and all other model parameters are independent of β . It is easy to check that the leaving rates of all microstates are given by

$$\begin{cases} q_{(1,0)} = b_1 + b_3 + s = \beta a_1 (1 + s/b_1) + b_3, \\ q_{(2,0)} = a_1 + b_2, \\ q_{(3,0)} = a_2 + a_3, \\ q_{(1,m)} = b_1 + s + mv = \beta a_1 (1 + s/b_1) + mv, \quad m \geq 1, \\ q_{(2,m)} = a_1 + mv, \quad m \geq 1, \end{cases}$$

which shows that

$$\lim_{\beta \rightarrow \infty} q_{(1,m)} = \infty, \quad \lim_{\beta \rightarrow \infty} q_{(2,m)} < \infty, \quad \lim_{\beta \rightarrow \infty} q_{(3,m)} < \infty.$$

Therefore, all active microstates $(1, m)$ are fast states and all other microstates $(2, m)$ and $(3, m)$ are slow states (Fig. 3(a)). By using a classical simplification method of two-time-scale Markov jump processes called decimation [50–55], the original Markovian model can be simplified to a reduced one by removal of all fast states.

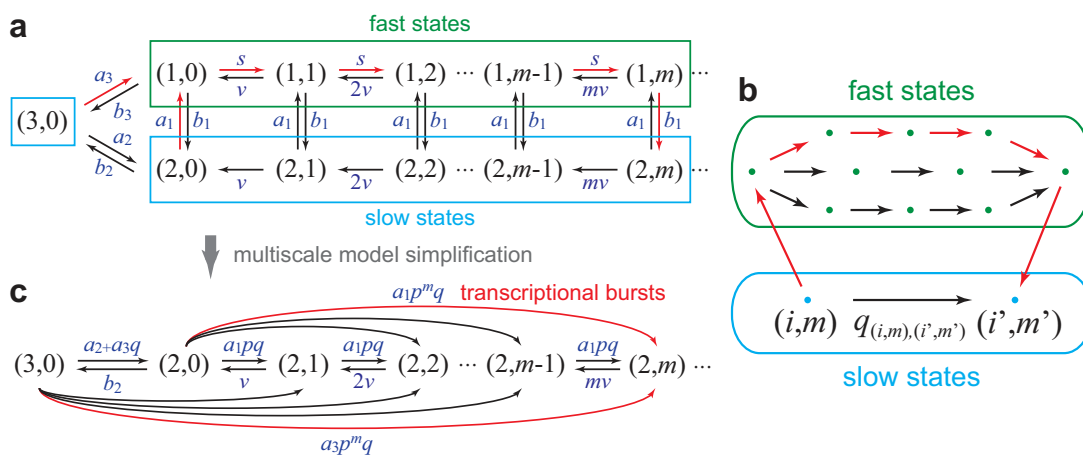


Fig. 3. Multiscale model simplification of the Markovian model. (a) Fast (green) and slow (blue) states of the original Markovian model. (b) Schematic diagram of the decimation method of model simplification. The effective transition rate from microstate (i, m) to microstate (i', m') is the superposition of the direct transition rate and the contribution of indirect transitions via all fast transition paths. (c) Transition diagram of the reduced Markovian model when $b_1 \gg a_1$ and s/b_1 is finite. The red arrows in (a)-(c) point the directions of typical fast transition paths, which correspond to random transcriptional bursts.

The remaining question is to determine the transition diagram and effective transition rates of the reduced model. This process is described as follows. Suppose that the original model jumps from microstate (i, m) to another microstate a particular time. When $\beta \gg 1$, the transition probability from microstate (i, m) to another microstate (i', m') is given by

$$w_{(i,m),(i',m')} = \lim_{\beta \rightarrow \infty} \frac{q_{(i,m),(i',m')}}{q_{(i,m)}}.$$

Let $(i_1, m_1), \dots, (i_n, m_n)$ be a sequence of microstates. We say that

$$c : (i, m) \rightarrow (i_1, m_1) \rightarrow \dots \rightarrow (i_n, m_n) \rightarrow (i', m')$$

is a fast transition path from (i, m) to (i', m') if the intermediate states $(i_1, m_1), \dots, (i_n, m_n)$ are all fast states. Moreover, the probability weight w_c of the fast transition path c is defined as

$$w_c = q_{(i,m),(i_1,m_1)} w_{(i_1,m_1),(i_2,m_2)} \cdots w_{(i_n,m_n),(i',m')}.$$

According to the decimation theory [50–55], the effective transition rate from (i, m) to (i', m') is given by

$$\tilde{q}_{(i,m),(i',m')} = q_{(i,m),(i',m')} + \sum_c w_c,$$

where c ranges over all fast transition paths from (i, m) to (i', m') . This formula shows that the effective transition rate from (i, m) to (i', m') is the sum of two parts: the direct transition rate $q_{(i,m),(i',m')}$ and the contribution of indirect transitions via all fast transition paths, as illustrated in Fig. 3(b).

Since the intermediate states of a fast transition path c are all fast states, in order for the path to have a positive probability weight, all the intermediate transitions along this path must satisfy

$$\lim_{\beta \rightarrow \infty} q_{(i_1,m_1),(i_2,m_2)} = \cdots = \lim_{\beta \rightarrow \infty} q_{(i_n,m_n),(i',m')} = \infty.$$

By using this criterion, it is easy to see that the original model only has two types of fast transition paths with positive probability weights, which are given by

$$(2, m) \rightarrow (1, m) \rightarrow (1, m+1) \rightarrow \cdots \rightarrow (1, m') \rightarrow (2, m'), \quad m' > m, \quad (1)$$

and

$$(3, 0) \rightarrow (1, 0) \rightarrow (1, 1) \rightarrow \cdots \rightarrow (1, m) \rightarrow (2, m), \quad m \geq 0, \quad (2)$$

as illustrated by the red arrows in Fig. 3(a). To proceed, we defined two constants p and q as

$$p = \frac{s}{s + b_1}, \quad q = \frac{b_1}{s + b_1}.$$

When $\beta \gg 1$, the transition probabilities along the above two fast transition paths are given by

$$w_{(1,m),(1,m+1)} = \lim_{\beta \rightarrow \infty} \frac{s}{q_{(1,m)}} = p,$$

$$w_{(1,m),(2,m)} = \lim_{\beta \rightarrow \infty} \frac{b_1}{q_{(1,m)}} = q.$$

Therefore, the probability weight of the path (1) is given by $a_1 p^{m'-m} q$ and the probability weight of the path (2) is given by $a_3 p^m q$. Since there is no direct transition, the effective transition rate from $(2, m)$ to $(2, m')$ is the indirect transition rate via the fast transition path (1):

$$\tilde{q}_{(2,m),(2,m')} = a_1 p^{m'-m} q.$$

Moreover, the effective transition rate from $(3, 0)$ to $(2, m)$ is the sum of the direct transition rate and the indirect transition rate via the fast transition path (2):

$$\tilde{q}_{(3,0),(2,m)} = q_{(3,0),(2,m)} + a_3 p^m q = \begin{cases} a_2 + a_3 q, & m = 0, \\ a_3 p^m q, & m \geq 1. \end{cases}$$

The above two formulas indicate that the reduce model may produce large jumps of mRNA abundance within a very short period, which correspond to transcriptional bursts. Each random burst corresponds to a fast transition path of the original model (see the red arrows in Fig. 3(a)). So far, we have obtained all effective transition rates of the reduced model, whose transition diagram is depicted in Fig. 3(c).

The above calculations can be understood intuitively as follows. Since $b_1 \gg a_1$ and s/b_1 is finite, the process of mRNA synthesis followed by gene silencing is essentially instantaneous. Once the gene becomes active, it can either produce a transcript with probability $p = s/(s + b_1)$ or switch to the inactive state with probability $q = 1 - p = b_1/(s + b_1)$. Therefore, the probability that the gene produces k transcripts in a single burst before it is finally silenced will be $p^k q$, which follows a geometric distribution. This consideration again leads to the reduced model illustrated in Fig. 3(c). The evolution of the reduced model is governed by the chemical master equation

$$\begin{cases} \dot{p}_{2,0} = v p_{2,1} + (a_2 + a_3 q) p_{3,0} - (a_1 p + b_2) p_{2,0}, \\ \dot{p}_{3,0} = b_2 p_{2,0} - (a_2 + a_3) p_{3,0}, \\ \dot{p}_{2,m} = \sum_{k=0}^{m-1} a_1 p^{m-k} q p_{2,k} + (m+1) v p_{2,m+1} \\ \quad + a_3 p^m q p_{3,0} - (a_1 p + m v) p_{2,m}, \quad m \geq 1. \end{cases} \quad (3)$$

Since the burst size of the mRNA, which is defined as the number of transcripts produced in a single burst (Fig. 2(a)), is geometrically distributed, its expected value is given by

$$h = \sum_{k=0}^{\infty} k p^k q = \frac{p}{q} = \frac{s}{b_1}.$$

4 Theoretical foundation for the ZINB model

Although the topological structure of the reduced model is complicated, its steady-state probability distribution can be solved analytically. To see this, let $p_{(i,m)}^{ss}$ denote the steady-state probability of microstate (i, m) . At the steady state, the probabilities of all microstates are time-independent and thus the left side of (3) must equal zero, giving rise to a set of linear equations. Interestingly, this set of linear equations can be solved explicitly with its solution being given by (see Appendix)

$$\begin{cases} p_{2,0}^{ss} = A \cdot \frac{a_1}{\tilde{a}_1}, \\ p_{3,0}^{ss} = A \cdot \frac{a_1 b_2}{\tilde{a}_1 (a_2 + a_3)}, \\ p_{2,m}^{ss} = A \cdot \frac{p^m (a_1/v)_m}{m!}, \quad m \geq 1, \end{cases} \quad (4)$$

where A is a normalization constant, \tilde{a}_1 is a constant given by

$$\tilde{a}_1 = a_1 + \frac{b_2 a_3}{a_2 + a_3}, \quad (5)$$

and $(x)_m = x(x+1)\cdots(x+m-1)$ is the Pochhammer symbol. Since all steady-state probabilities add up to one, the normalization constant A can be determined as

$$A = \left[\frac{a_1}{\tilde{a}_1} \left(1 + \frac{b_2}{a_2 + a_3} \right) + q^{-a_1/v} - 1 \right]^{-1}.$$

Let p_m^{ss} denote the steady-state probability of having m copies of detectable transcripts. Then we obtain

$$\begin{cases} p_0^{ss} = p_{2,0}^{ss} + p_{3,0}^{ss} = A \cdot \frac{a_1}{\tilde{a}_1} \left(1 + \frac{b_2}{a_2 + a_3} \right), \\ p_m^{ss} = p_{2,m}^{ss} = A \cdot \frac{p^m (a_1/v)_m}{m!}, \quad m \geq 1. \end{cases}$$

Since the probabilities p and q can be represented by the mean burst size h as

$$p = \frac{h}{1+h}, \quad q = \frac{1}{1+h},$$

the steady-state distribution of mRNA abundance can be written in a unified way as

$$\begin{aligned} p_m^{ss} &= w\delta_0(m) + (1-w) \frac{(\lambda)_m}{m!} \left(\frac{h}{1+h} \right)^m \left(\frac{1}{1+h} \right)^\lambda \\ &= wp_m^{\text{zero-inflated}} + (1-w)p_m^{\text{NB}}, \end{aligned}$$

where $\delta_0(m)$ is Kronecker's delta function which takes the value of 1 when $m = 0$ and takes the value of 0 otherwise, and $\lambda > 0$ and $0 < w < 1$ are two constants given by

$$\begin{aligned} \lambda &= \frac{a_1}{v}, \\ w &= \frac{\frac{a_1}{\tilde{a}_1} \left(1 + \frac{b_2}{a_2 + a_3} \right) - 1}{\frac{a_1}{\tilde{a}_1} \left(1 + \frac{b_2}{a_2 + a_3} \right) + (1+h)^\lambda - 1}. \end{aligned}$$

Here $0 < w < 1$ is a result of our model assumption $a_1 > a_3$. This is exactly the ZINB distribution of mRNA abundance widely used in scRNA-seq data analysis [20–31]. Specifically, the ZINB distribution is the mixture of two distributions: the zero-inflated part

$$p_m^{\text{zero-inflated}} = \delta_0(m)$$

is a single-point distribution concentrated at zero and the negative binomial part

$$p_m^{\text{NB}} = \frac{(\lambda)_m}{m!} \left(\frac{h}{1+h} \right)^m \left(\frac{1}{1+h} \right)^\lambda$$

is a negative binomial distribution. The ZINB distribution is determined by three parameters with clear biological implications: the dropout rate w which characterizes the proportion of the zero-inflated part due to both technical and biological effects, the mean burst size h which describes the average number of transcripts synthesized in a single burst, and the maximum burst frequency λ which represents the maximum number of occurrence of random bursts per mRNA lifetime. A more detailed discussion on the burst frequency will be given in the next section.

The ZINB distribution can exhibit three different types of shapes, as illustrated in Fig. 4. To clarify the conditions for the three types of shapes, we notice that the mode (maximum point) of the negative binomial part p_m^{NB} is given by

$$\mu_{\text{mode}} = \begin{cases} 0 & \text{when } \lambda < 1, \\ [(\lambda - 1)h] & \text{when } \lambda \geq 1, \end{cases}$$

where $[x]$ denotes the integer part of x . In fact, the first type of shape occurs when $p_0^{ss} \leq p_1^{ss}$, that is,

$$(\lambda - 1)h \geq 1 + \frac{w}{1-w}(1+h)^{\lambda+1}.$$

In this case, the dropout rate is small and the mode of the negative binomial part is large. Then the ZINB distribution peaks at the non-zero mode $[(\lambda - 1)h]$ with no zero-inflation (Fig. 4(a)). The second type of shape occurs when $p_0^{ss} > p_1^{ss}$ and $\mu_{\text{mode}} \leq 1$, that is,

$$(\lambda - 1)h < \min\{1 + \frac{w}{1-w}(1+h)^{\lambda+1}, 2\}.$$

In this case, the dropout rate is large and the mode of the negative binomial part is small. Then the ZINB distribution peaks at zero with apparent or inapparent zero-inflation (Fig. 4(b)). The third type of shape occurs when $p_0^{ss} > p_1^{ss}$ and $\mu_{\text{mode}} \geq 2$, that is,

$$2 \leq (\lambda - 1)h < 1 + \frac{w}{1-w}(1+h)^{\lambda+1}.$$

In this case, both the dropout rate and the mode of the negative binomial part are large. Then the ZINB distribution becomes bimodal and peaks at both zero and the non-zero mode $[(\lambda - 1)h]$ with apparent or inapparent zero-inflation (Fig. 4(c)).

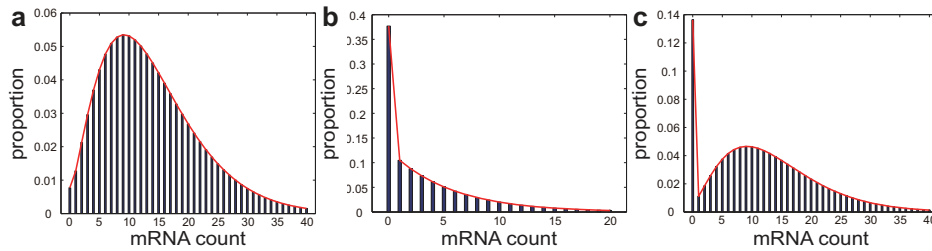


Fig. 4. Three different types of shapes for the ZINB distribution of mRNA abundance. (a) The distribution peaks at a non-zero mode with no zero-inflation. (b) The distribution peaks at zero with apparent or inapparent zero-inflation. (c) The distribution exhibits bistability and peaks at both zero and a non-zero mode with apparent or inapparent zero-inflation.

Three special cases should be paid special attention to. The first case occurs when the mean burst size $h \rightarrow 0$ and the maximum burst frequency $\lambda \rightarrow \infty$, while $\lambda h = \gamma$ is kept as a constant. In this case, we have

$$\begin{aligned} (\lambda)_m \left(\frac{h}{1+h} \right)^m &\rightarrow \gamma^m, \\ \left(\frac{1}{1+h} \right)^\lambda &= \left(1 - \frac{h}{1+h} \right)^\lambda \rightarrow e^{-\gamma}. \end{aligned}$$

Then the ZINB distribution of mRNA abundance reduces to the zero-inflated Poisson (ZIP) distribution

$$p_m^{ss} = w\delta_0(m) + (1-w)\frac{\gamma^m}{m!}e^{-\gamma}.$$

In fact, the ZIP distribution is also extensively applied in scRNA-seq data analysis [27] and its kinetic mechanism has been clarified in previous studies [37, 40]. Our analytic theory shows that the ZIP model also naturally emerges from our three-state transcription model.

The second special case occurs when $b_2 = b_3 = 0$, which means that the switching from the active or inactive state to the dropout state is forbidden. In this case, the three-state model reduces to the

classical two-state model without the dropout state [1]. It is easy to verify that $\tilde{a}_1 = a_1$ and $w = 0$ in this regime. This shows that the dropout rate will vanish in the absence of the dropout state.

The last special case occurs when $a_3 = 0$, which means that the switching from the dropout state to the active state is forbidden. This is especially biologically relevant when the dropout state is understood to be the refractory (irreversibly silent) state found in recent single-cell experiments [43, 47]. In this case, we also have $\tilde{a}_1 = a_1$ and thus the dropout rate is given by

$$w = \frac{K_2}{K_2 + (1 + h)^\lambda},$$

where $K_2 = b_2/a_2$ is the equilibrium constant of gene switching between the inactive and dropout states. An increased equilibrium constant K_2 will result in a larger fraction of cells being in the dropout state and thus is expected to enhance the dropout rate w . Interestingly, our theory reveals a nontrivial quantitative relation between dropout events and transcriptional bursting: an increased mean burst size h or maximum burst frequency λ will give rise to a decline in the dropout rate w . This relation provides novel insights into how and to what extent the burst size and burst frequency of the mRNA could reduce the dropout rate. The basic reason of such dependency is that an increase in the burst size and burst frequency will both promote rapid accumulation of mRNA from a low to a higher level, which is unfavorable to the occurrence of dropouts.

5 Mean burst duration and burst frequency

It has been shown that the mean burst size of the mRNA is given by $h = s/b_1$. Here we present a more detailed discussion on the burst frequency. In this section, we assume that the time-dependent mRNA abundance in an individual cell could be measured at a series of successive time points, and due to various technical factors, the mRNA expression is undetectable during some periods. Recall that each transcriptional burst is featured by a short transcriptionally active period followed by a long transcriptionally silent period. Mathematically, the mean burst duration, which is defined as the average time needed to complete a single burst (Fig. 2(a)), can be computed as the inverse of the total probability flux between the active microstates and other (inactive and dropout) microstates [56, 57]. From (4), the total flux J between the active microstates and other microstates is given by

$$J = \left[\sum_{m=0}^{\infty} p_{2,m}^{ss} \right] a_1 + p_{3,0}^{ss} a_3 = \alpha a_1,$$

where $0 < \alpha \leq 1$ is a constant given by

$$\alpha = \frac{\frac{a_1}{\tilde{a}_1} + \frac{a_3 b_2}{\tilde{a}_1 (a_2 + a_3)} + (1 + h)^\lambda - 1}{\frac{a_1}{\tilde{a}_1} + \frac{a_1 b_2}{\tilde{a}_1 (a_2 + a_3)} + (1 + h)^\lambda - 1},$$

and thus the mean burst duration is given by

$$\tau_{\text{burst}} = \frac{1}{J} = \frac{1}{\alpha a_1}.$$

Since the mRNA lifetime is the inverse of the mRNA degradation rate v , the mean burst frequency λ_0 of the mRNA, which is defined as the average number of occurrence of random bursts per mRNA lifetime, is given by the quotient of the mRNA lifetime $1/v$ and the mean burst duration τ_{burst} :

$$\lambda_0 = \frac{1}{v \tau_{\text{burst}}} = \frac{\alpha a_1}{v} = \alpha \lambda,$$

where $\lambda = a_1/v$ is the maximum burst frequency defined previously. Since $0 < \alpha \leq 1$, the true mean burst frequency λ_0 is always smaller than the maximal burst frequency λ .

We next focus on three special cases. In the limiting case of $h \rightarrow 0$ and $\lambda \rightarrow \infty$, while $\lambda h = \gamma$ is kept as a constant, we have $\lambda_0 \rightarrow \infty$. In this regime, random bursts occur very frequently but each burst only contributes a very small burst size. Due to large burst frequencies, the gene switches very rapidly between the active and inactive states, giving rise to a large number of “futile” switches (Fig. 2(b)).

In the special case of $b_2 = b_3 = 0$, the three-state model reduces to the classical two-state model without the dropout state [1]. In this regime, we have $\alpha = 1$ and the mean burst frequency attains its maximum $\lambda_0 = \lambda$. In the presence of the dropout state, we have $b_2 > 0$ and $\alpha < 1$. This shows that dropout events will lead to a reduction of the burst frequency by prolonging the transcriptionally silent periods.

The last special case occurs when $a_3 = 0$, which means that the switching from the dropout state to the active state is forbidden. In this case, we have $\tilde{a}_1 = a_1$ and

$$\alpha = \frac{(1+h)^\lambda}{K_2 + (1+h)^\lambda} = 1 - w$$

is the proportion of the negative binomial part. Then the mean burst frequency is given by

$$\lambda_0 = (1-w)\lambda.$$

This quantitative relation reveals how the dropout rate could affect the burst frequency.

6 Over-dispersion of scRNA-seq data

The simplest kinetic model of transcription is the classical birth-death process, which describes the synthesis and degradation of the mRNA. The steady-state distribution of the birth-death process turns out to be a Poisson distribution, whose mean and variance are equal. In bulk or single-cell RNA-seq experiments, read counts are always over-dispersed relative to Poisson: the variance is higher than the mean [16, 58].

In the literature, the dispersion, sometimes referred to as noise, in mRNA abundance within a cell population is often characterized by the Fano factor $\eta = \sigma^2/\langle m \rangle$, which is defined as the ratio of the variance σ^2 and the mean $\langle m \rangle$. A dispersion greater than one reveals a deviation from the Poisson distribution and thus serves as a characteristic signal of over-dispersion. Strictly speaking, the dispersion captures all sources of variation between samples, including contributions from technical factors leading to dropouts as well as real biological variation.

To calculate the mean and variance of mRNA abundance, we consider the generating function of the ZINB distribution:

$$F(z) = \sum_{m=1}^{\infty} p_m^{ss} z^m = w + (1-w) \frac{q^\lambda}{(1-pz)^\lambda}.$$

Then the mean and variance can be recovered by taking the derivatives of the generating function:

$$\begin{aligned} \langle m \rangle &= F'(1) = (1-w)\lambda h, \\ \sigma^2 &= F''(1) + F'(1) - F'(1)^2 = (1-w)[w\lambda^2 h^2 + \lambda h^2 + \lambda h]. \end{aligned} \tag{6}$$

Therefore, the dispersion in mRNA abundance is given by

$$\eta = \frac{\sigma^2}{\langle m \rangle} = 1 + h + w\lambda h,$$

where the constant term 1 is the dispersion of a Poisson distribution arising from individual births and deaths of the mRNA, the middle term h describes the dispersion due to transcriptional burst sizes, and the last term $w\lambda h$ characterizes the dispersion due to the interaction between dropout events and transcriptional bursting. When there are no dropouts, the dispersion reduces to $\eta = 1 + h$, which does not depend on the burst frequency [1]. Interestingly, in the presence of dropout events, the dispersion positively depends on the three parameters: the dropout rate w , mean burst size h , and maximum burst frequency λ . This clearly reveals three different biophysical origins of over-dispersion.

Statistically, the three parameters involved in the ZINB distribution can be estimated in several different ways. The maximum likelihood estimation has been discussed in [30]. Here we provide two additional approaches. In fact, the first three moments of the ZINB distribution can be recovered from the generating function as

$$\begin{aligned}\langle m \rangle &= F'(1) = (1-w)\lambda h, \\ \langle m^2 \rangle &= F''(1) + F'(1) = (1-w)[\lambda(\lambda+1)h^2 + \lambda h] \\ \langle m^3 \rangle &= F'''(1) + 3F''(1) + F'(1) \\ &= (1-w)[\lambda(\lambda+1)(\lambda+2)h^3 + 3\lambda(\lambda+1)h^2 + \lambda h].\end{aligned}$$

By analyzing scRNA-seq data, the first three moments of mRNA abundance can be estimated. Solving the above set of polynomial equations give the moment estimates of w , h , and λ .

When the mRNA levels across cells are relatively high, there is still another method to estimate the three parameters. From (4), when $m \gg 1$, we have

$$\frac{p_{m+1}^{ss}}{p_m^{ss}} = \frac{h}{h+1} \cdot \frac{m+\lambda}{m+1} \approx \frac{h}{h+1}.$$

This suggests that for any $k \geq 1$,

$$p_{m+k}^{ss} \approx \left(\frac{h}{h+1}\right)^k p_m^{ss}.$$

Taking logarithm on both sides gives rise to

$$\log p_{m+k}^{ss} \approx k \log \left(\frac{h}{h+1}\right) + \log p_m^{ss},$$

which is a linear relation with respect to k . Therefore, we only need to calculate the logarithm of the steady-state probabilities at large mRNA copy numbers and then carry out a linear regression analysis with respect to the copy number difference k . The slope of the linear regression provides an estimate of the mean burst size h . Once h is known, we can solve (6) to obtain the estimates of the dropout rate w and maximum burst frequency λ :

$$\begin{aligned}w &= \frac{\langle m \rangle}{\langle m \rangle + \eta - h - 1}, \\ \lambda &= \frac{(\eta - h - 1)(\langle m \rangle + \eta - h - 1)}{\langle m \rangle h}.\end{aligned}$$

where $\eta = \sigma^2/\langle m \rangle$ is the dispersion.

7 Discussion

In this work, we present a comprehensive analysis of a three-state transcription model with dropout events and over-dispersion based on the biochemical reaction kinetics underlying transcription. Using the multiscale simplification technique of decimation, we simplify the original Markovian model to a reduced one by removal of all fast states. It turns out that transcriptional bursts exactly correspond to the fast transition paths of the original model. Although the reduced model has a complicated topology, we obtain its steady-state analytic solution. The widely used ZINB or ZIP model of scRNA-seq data naturally emerges as the steady-state distribution of the reduced model. This provides a mesoscopic kinetic foundation of these statistical models. We further clarify the biological implications of the three parameters involved in the ZINB distribution: the dropout rate w , mean burst size h , and maximum burst frequency λ . In addition, we discover a nontrivial relation between dropout events and transcriptional bursting, which quantitatively reveals how and to what extent the burst size and burst frequency could reduce the dropout rate. Another relation reveals how dropout events could lower the burst frequency by prolonging the transcriptionally silent periods. The dispersion of scRNA-seq data is also investigated at the single-cell level and three different biophysical origins of over-dispersion are found. Finally, two statistical methods are given to estimate the three parameters involved in the ZINB distribution.

Our three-state transcription model is a minimal kinetic model that could account for the ZINB distribution of mRNA abundance. Recently, there has been some discussion on the role of various technical and biological effects on the apparent zero-inflation in scRNA-seq data [27, 59, 60]. In our minimal three-state model, zero-inflation is realized by the introduction of a dropout state, which may be either interpreted as an undetectable state due to technical factors or interpreted as a refractory state due to biological factors. In other words, our three-state model cannot distinguish whether zero-inflation is a consequence of technical or biological effects. If we would like to empirically decide between the two interpretations, a more realistic model that takes into account more complex features of stochastic transcription dynamics must be developed.

If the dropout state is interpreted as a refractory state due to biological factors, then a more realistic model would be the Markovian model illustrated in Fig. 5(a), where microstates $(3, m)$, $m \geq 1$ are incorporated and transitions between microstates $(1, m)$, $(2, m)$, and $(3, m)$ are allowed. Here $(3, m)$ represents the microstate of having m transcripts in an individual cell when the gene is in the refractory state. In fact, the minimal kinetic model depicted in Fig. 1(b) can be viewed as an approximation of the more realistic model when $a_2, a_3 \ll v$. This can be understood intuitively as follows. Since $a_2, a_3 \ll v$, the degradation of the mRNA is fast and the switching of the gene from the refractory state to the active or inactive state is slow. Once the gene is in the refractory state, before it could switch to the active or inactive state, the microstates $(3, m)$, $m \geq 0$ are already in rapid pre-equilibrium due to fast mRNA degradation and thus most of the probability is concentrated on microstate $(3, 0)$.

If the dropout state is interpreted as an undetectable state due to technical factors, then a more realistic model would be the Markovian model illustrated in Fig. 5(b), where the microstate of the gene of interest is described by an ordered triple (i, m, k) : the activity i of the gene with $i = 1, 2$ corresponding to the active and inactive states, respectively, the copy number m of the mRNA, and the detection state k of the transcriptional signal with $k = 1, 0$ corresponding to the detectable and undetectable states, respectively. In scRNA-seq experiments, the variable of interest is the copy number

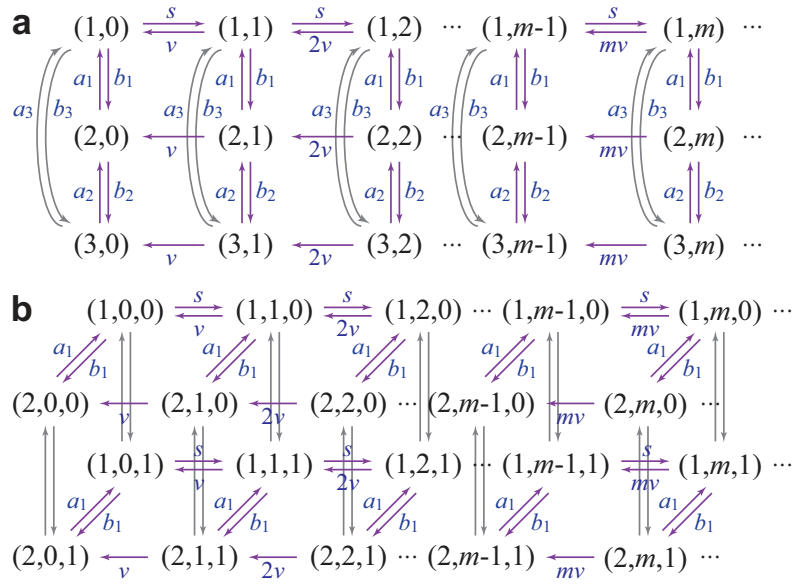


Fig. 5. More realistic models of transcription. (a) A Markovian model of stochastic transcription involving gene switching among an active, an inactive, and a refractory state. The microstate of the gene of interest is described by an ordered pair (i, m) : the activity i of the gene and the copy number m of the mRNA. Here $i = 1, 2, 3$ correspond to the active, inactive, and refractory states, respectively. (b) A Markovian model of stochastic transcription with dropout events. The microstate of the gene of interest is described by an ordered triple (i, m, k) : the activity i of the gene, the copy number m of the mRNA, and the detection state k of the transcriptional signal. Here $i = 1, 2$ correspond to the active and inactive states, respectively, and $k = 1, 0$ correspond to the detectable and undetectable states, respectively.

of detectable transcripts, which is given by

$$N(i, m, k) = \begin{cases} m, & \text{if } k = 1, \\ 0, & \text{if } k = 0. \end{cases}$$

This Markovian model allows transitions between detectable microstates $(i, m, 1)$ and undetectable microstates $(i, m, 0)$. Since dropouts are more frequent for cells with low mRNA expression levels [17], the transition rate from $(i, m, 1)$ to $(i, m, 0)$ should be a decreasing function of m and the transition rate from $(i, m, 0)$ to $(i, m, 1)$ should be an increasing function of m . Within this framework, the minimal kinetic model depicted in Fig. 1(b) can be roughly viewed as an approximation of the more realistic model with all undetectable microstates $(i, m, 0)$ combined as a single microstate $(3, 0)$.

Besides the ZINB and ZIP models discussed in the present work, many other statistical models have also been proposed to analyze scRNA-seq data. Some commonly used models include but not limited to the Gaussian mixture model [61], Poisson-negative binomial mixture model [17, 62], Poisson-gamma mixture model [63], Hurdle model [64], zero-inflated log-normal model [18], zero-inflated Gaussian mixture model [19], and Bayesian mixture model [65, 66]. We anticipate that the mesoscopic kinetic mechanisms for these models could be clarified. A deeper understanding of the connection between the kinetic approach and the statistical approach is expected.

Acknowledgements

The author acknowledges Michael Q. Zhang, Min Chen, and Cong Zhang at the University of Texas at Dallas, Yuxuan Liu at the University of Texas Southwestern Medical Center, Bochao Liu at Rutgers

Cancer Institute of New Jersey, and Xuegong Zhang and Jiaqi Li at Tsinghua University for stimulating discussions. The author is also grateful to the anonymous reviewers for their valuable comments and suggestions which help the author greatly in improving the quality of this paper.

Appendix

Here we provide the detailed derivation of the steady-state probability distribution of the reduced model depicted in Fig. 3(c). At the steady state, the steady-state probabilities of all microstates satisfy the following set of linear equations:

$$\begin{cases} 0 = vp_{2,1}^{ss} + (a_2 + a_3q)p_{3,0}^{ss} - (a_1p + b_2)p_{2,0}^{ss}, \\ 0 = b_2p_{2,0}^{ss} - (a_2 + a_3)p_{3,0}^{ss}, \\ 0 = \sum_{k=0}^{m-1} a_1p^{m-k}qp_{2,k}^{ss} + (m+1)vp_{2,m+1}^{ss} \\ \quad + a_3p^m qp_{3,0}^{ss} - (a_1p + mv)p_{2,m}^{ss}, \quad m \geq 1. \end{cases} \quad (7)$$

By the second equation in (7), we have

$$(a_2 + a_3)p_{3,0}^{ss} = b_2p_{2,0}^{ss}.$$

Inserting this equation into the first and third equations in (7) eliminates $p_{3,0}^{ss}$ and yields

$$\begin{cases} 0 = vp_{2,1}^{ss} - \tilde{a}_1pp_{2,0}^{ss}, \\ 0 = \tilde{a}_1p^m qp_{2,0}^{ss} + \sum_{k=1}^{m-1} a_1p^{m-k}qp_{2,k}^{ss} + (m+1)vp_{2,m+1}^{ss} - (a_1p + mv)p_{2,m}^{ss}, \quad m \geq 1, \end{cases} \quad (8)$$

where \tilde{a}_1 is the constant defined in (5). For convenience, set

$$w_0 = \frac{\tilde{a}_1}{a_1}p_{2,0}^{ss}, \quad w_m = p_{2,m}^{ss}, \quad m \geq 1.$$

Then the two equations in (8) can be rewritten in a unified way as

$$\sum_{k=0}^{m-1} a_1p^{m-k}qw_k + (m+1)vw_{m+1} - (a_1p + mv)w_m = 0, \quad m \geq 0. \quad (9)$$

To proceed, we introduce the generating function

$$F(z) = \sum_{m=1}^{\infty} w_m z^m.$$

Then the algebraic equation (9) can be converted into the ordinary differential equation

$$vF'(z) = \frac{a_1p}{1-pz}F(z),$$

whose solution is given by

$$F(z) = A(1-pz)^{-a_1/v},$$

where A is a constant. Therefore, w_m can be recovered from the generating function F as

$$w_m = \frac{F^{(m)}(0)}{m!} = A \cdot \frac{p^m (a_1/v)_m}{m!}.$$

This shows that

$$p_{2,0}^{ss} = A \cdot \frac{a_1}{\tilde{a}_1}, \quad p_{2,m}^{ss} = A \cdot \frac{p^m (a_1/v)_m}{m!}, \quad m \geq 1.$$

References

- [1] Paulsson, J. Models of stochastic gene expression. *Phys. Life Rev.* **2**, 157–175 (2005).
- [2] Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11**, 22 (2014).
- [3] Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat. Methods* **11**, 25 (2014).
- [4] Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
- [5] Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17**, 63 (2016).
- [6] Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053 (2014).
- [7] Landau, W. M. & Liu, P. Dispersion estimation and its effect on test performance in RNA-seq data analysis: a simulation-based comparison of methods. *PLoS one* **8**, e81415 (2013).
- [8] Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
- [9] Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* **5** (2016).
- [10] Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2017).
- [11] Jia, C. *et al.* Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.* **45**, 10978–10988 (2017).
- [12] Zhu, L., Lei, J., Devlin, B., Roeder, K. *et al.* A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.* **12**, 609–632 (2018).
- [13] Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220 (2018).
- [14] Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Mol. Cell* **58**, 598–609 (2015).
- [15] Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
- [16] McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
- [17] Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740 (2014).
- [18] McDavid, A. *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461–467 (2012).
- [19] Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200 (2013).
- [20] Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 241 (2015).
- [21] Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145 (2016).
- [22] Fang, R., Wagner, B., Harris, J. & Fillon, S. Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiology & Infection* **144**, 2447–2455 (2016).
- [23] Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14**, 565 (2017).
- [24] Gao, R. *et al.* Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat.*

- Commun.* **8**, 228 (2017).
- [25] Wallrapp, A. *et al.* The neuropeptide NMU amplifies ILC2-driven allergic lung inflammation. *Nature* **549**, 351 (2017).
- [26] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
- [27] Chen, W. *et al.* UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* **19**, 70 (2018).
- [28] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053 (2018).
- [29] Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **19**, 24 (2018).
- [30] Miao, Z., Deng, K., Wang, X. & Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **34**, 3223–3224 (2018).
- [31] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
- [32] Qian, H. Stochastic physics, complex systems and biology. *Quant. Biol.* **1**, 50–53 (2013).
- [33] Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
- [34] Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
- [35] Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E* **79**, 031911 (2009).
- [36] Mugler, A., Walczak, A. M. & Wiggins, C. H. Spectral solutions to stochastic models of gene expression with bursts and regulation. *Phys. Rev. E* **80**, 041921 (2009).
- [37] Chong, S., Chen, C., Ge, H. & Xie, X. S. Mechanism of transcriptional bursting in bacteria. *Cell* **158**, 314–326 (2014).
- [38] Kumar, N., Singh, A. & Kulkarni, R. V. Transcriptional bursting in gene expression: analytical results for general stochastic models. *PLoS Comput. Biol.* **11**, e1004292 (2015).
- [39] Jia, C., Zhang, M. Q. & Qian, H. Emergent Lévy behavior in single-cell stochastic gene expression. *Phys. Rev. E* **96**, 040402 (2017).
- [40] Klindziuk, A. & Kolomeisky, A. B. Theoretical Investigation of Transcriptional Bursting: A Multistate Approach. *J. Phys. Chem. B* **122**, 11969–11977 (2018).
- [41] Livesey, F. Strategies for microarray analysis of limiting amounts of RNA. *Briefings in Functional Genomics* **2**, 31–36 (2003).
- [42] Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
- [43] Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
- [44] Harper, C. V. *et al.* Dynamic analysis of stochastic transcription cycles. *PLoS Biol.* **9**, e1000607 (2011).
- [45] Rieckh, G. & Tkačik, G. Noise and information transmission in promoters with multiple internal states. *Biophys. J.* **106**, 1194–1204 (2014).
- [46] Corrigan, A. M., Tunnacliffe, E., Cannon, D. & Chubb, J. R. A continuum model of transcriptional bursting. *Elife* **5**, e13051 (2016).
- [47] Bintu, L. *et al.* Dynamics of epigenetic regulation at the single-cell level. *Science* **351**, 720–724 (2016).
- [48] Cai, L., Dalal, C. K. & Elowitz, M. B. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* **455**, 485 (2008).
- [49] Moran, M. A. *et al.* Sizing up metatranscriptomics. *The ISME journal* **7**, 237 (2012).

- [50] Pigolotti, S. & Vulpiani, A. Coarse graining of master equations with fast and slow states. *J. Chem. Phys.* **128**, 154114 (2008).
- [51] Cappelletti, D., Wiuf, C. *et al.* Elimination of intermediate species in multiscale stochastic reaction networks. *Ann. Appl. Probab.* **26**, 2915–2958 (2016).
- [52] Jia, C. Reduction of Markov chains with two-time-scale state transitions. *Stochastics* **88**, 73–105 (2016).
- [53] Jia, C. Simplification of irreversible Markov chains by removal of states with fast leaving rates. *J. Theor. Biol.* **400**, 129–137 (2016).
- [54] Bo, S. & Celani, A. Multiple-scale stochastic processes: decimation, averaging and beyond. *Phys. Rep.* **670**, 1–59 (2016).
- [55] Jia, C. Simplification of Markov chains with infinite state space and the mathematical theory of random gene expression bursts. *Phys. Rev. E* **96**, 032402 (2017).
- [56] Jia, C., Li, Y. & Qian, M. A general analysis of single IP₃ receptors modulated by cytosolic Ca²⁺ and IP₃. In *The Third International Symposium on Optimization and Systems Biology*, 89–101 (Zhangjiajie, China, 2009).
- [57] Jia, C., Jiang, D. & Qian, M. An allosteric model of the inositol trisphosphate receptor with nonequilibrium binding. *Phys. Biol.* **11**, 056001 (2014).
- [58] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
- [59] Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. *bioRxiv* 574574 (2019).
- [60] Svensson, V. Droplet scRNA-seq is not zero-inflated. *bioRxiv* 582064 (2019).
- [61] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495 (2015).
- [62] Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241 (2016).
- [63] Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nature methods* **15**, 539 (2018).
- [64] Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
- [65] Prabhakaran, S., Azizi, E., Carr, A. & Peer, D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, 1070–1079 (2016).
- [66] Sun, Z. *et al.* A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. *Nat. Commun.* **10**, 1649 (2019).