On the relation between input and output distributions of scRNA-seq experiments

Schwabe D., Falcke M.

Bioinformatics
2022 MAR 01 ; 38(5): 1336-1343
2021 DEC 15 (first published online: final publication)
doi: 10.1093/bioinformatics/btab841

Publisher: Oxford University Press

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

On the relation between input and output distributions of scRNA-seq experiments

Daniel Schwabe [1] and Martin Falcke [2,*]

[1] Mathematical Cell Physiology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[2] Department of Physics, Humboldt University Berlin, Newtonstr. 15, 12489 Berlin, Germany Country.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Single-cell RNA sequencing determines RNA copy numbers per cell for a given gene. However, technical noise poses the question how observed distributions (output) are connected to their cellular distributions (input).
Results: We model a single-cell RNA sequencing setup consisting of PCR amplification and sequencing, and derive probability distribution functions for the output distribution given an input distribution. We provide copy number distributions arising from single transcripts during PCR amplification with exact expressions for mean and variance. We prove that the coefficient of variation of the output of sequencing is always larger than that of the input distribution. Experimental data reveals the variance and mean of the input distribution to obey characteristic relations, which we specifically determine for a HeLa data set. We can calculate as many moments of the input distribution as are known of the output distribution (up to all). This, in principle, completely determines the input from the output distribution.
Contact: martin.falcke@mdc-berlin.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

When exploring cell populations, one might expect that cells of the same type and under the same environmental conditions should have the exact same gene expression profile. However, single-cell data has revealed that gene expression even among clonal cells is heterogeneous. This is called biological noise or cell-to-cell variability. While causes for this effect have been discussed extensively (Elowitz *et al.*, 2002; Roberfroid *et al.*, 2016; Sun and Zhang, 2020), the consequences or intent of it are more obscure (Roberfroid *et al.*, 2016; Osorio *et al.*, 2019; Mendenhall *et al.*, 2021). Nevertheless, it appears frequently in data sets such that quantification becomes more and more important.

Due to cell-to-cell variability (Tsimring, 2014), we face distributions of individual mRNA species across a cell population instead of a single abundance value for all of them. That true distribution of the abundance of a specific gene is the input distribution to single-cell RNA sequencing (scRNA-seq). The scRNA-seq experiment will then introduce significant technical noise (Tung *et al.*, 2017; Hicks *et al.*, 2018) and produce an observed or output distribution. The technical and biological noise are convoluted in the output distribution. We have to remove the technical noise from the data in order to estimate cell-to-cell variability. Many data-driven approaches have been established to deal with various aspects of technical noise. Among others, normalization (Butler *et al.*, 2018; Lun *et al.*, 2016; Bacher *et al.*, 2017; Hafemeister and Satija, 2019; Breda *et al.*, 2021), batch correction (Butler *et al.*, 2018; Haghverdi *et al.*, 2018; Tran *et al.*, 2020) and data imputation (van Dijk *et al.*, 2018; Huang *et al.*, 2018; Li and Li, 2018; Hou *et al.*, 2020) are now often part of scRNA-seq data analysis in an attempt to remove technical noise.

In contrast to such a data-driven approach, we model a simplified version of a scRNA-seq experiment and propose probability distributions for the data at each step. This results in an analytical formula for the output distribution for a given input distribution. Analysing this analytic expression reveals the moments of the input distribution, which is a first step towards quantifying cell-to-cell variability.

We simplify scRNA-seq experiments to the following essential steps:

1. Starting point is a cell population of n cells each containing m genes.

2. The mRNA is extracted and tagged with a cell barcode and a unique molecular identifier (UMI) sequence.
3. The mRNA material is amplified with l cycles of PCR.
4. The mRNA library is sequenced and reads containing the same UMI are collapsed.

2 The PCR distribution

Since each individual transcript is tagged by a cell barcode and a UMI, it can be uniquely identified amongst all other transcripts. The likelihood to produce the same cell barcode or UMI twice is small enough to justify ignoring these occurrences in our setting.

PCR is a method to amplify DNA material. At each cycle, every molecule has a chance p to be copied once. p is called the PCR efficiency. In an ideal scenario, p would be equal to 1. In reality, efficiency values vary greatly and the importance of having perfect efficiency depends entirely on the research question. Typical estimates put PCR efficiency between 0.9 and 1 (Svec *et al.*, 2015). While the PCR efficiency can be different from cycle to cycle (Booth *et al.*, 2010), we simplify our estimates and assume that the PCR efficiency p is constant across all cycles. We also note that we model only what is called the exponential phase where reactants are freely available.

We now want to know how many PCR copies each unique initial molecule produces after l cycles. The sequence of PCR cycles can be perceived as a branching process in terms of the theory of stochastic processes (Harris, 1964; Weiss and von Haeseler, 1995; Peccoud and Jacob, 1996; Stolovitzky and Cecchi, 1996). Let X_l be the random variable describing the number of copies after l cycles with $X_0 = 1$. The increase of copy numbers from X_{l-1} to X_l during cycle l follows a binomial distribution since each existing copy has chance p to be copied again. This means $(X_l - X_{l-1}|X_{l-1} = n) \sim \text{Binom}(n, p)$. Therefore, we can recursively derive an analytical formula for the probability distribution function (pdf) $\mathbb{P}[X_l = k]$ to obtain k copies after l cycles. This yields

$$\mathbb{P}[X_l = k] = \frac{p^{k-1}}{(1-p)^{k-2}} \sum_{i_{l-1}=\lceil \frac{k}{2} \rceil}^{\min\{k,2^{l-1}\}} \sum_{i_{l-2}=\left\lceil \frac{i_{l-1}}{2} \right\rceil}^{\min\{i_{l-1},2^{l-2}\}} \cdots \sum_{i_1=\left\lceil \frac{i_2}{2} \right\rceil}^{\min\{i_2,2\}} \binom{i_{l-1}}{k-i_{l-1}} \cdot \ldots \cdot \binom{i_1}{i_2-i_1} (1-p)^{\sum_{m=1}^{l-1} i_m}. \tag{1}$$

where

- the indices i_m for $m \in \{1, \ldots, l-1\}$ represent the number of copies after m cycles,
- the lower bound for index i_{m-1} for $m \in \{1, \ldots, l-1\}$ is equal to the rounded up integer of $\frac{i_m}{2}$ which is expressed by $\left\lceil \frac{i_m}{2} \right\rceil$,
- the upper bound for index i_{m-1} for $m \in \{1, \ldots, l-1\}$ is equal to the minimum of i_m and 2^{m-1} (as not more than 2^{m-1} copies can have been created after $m-1$ cycles).

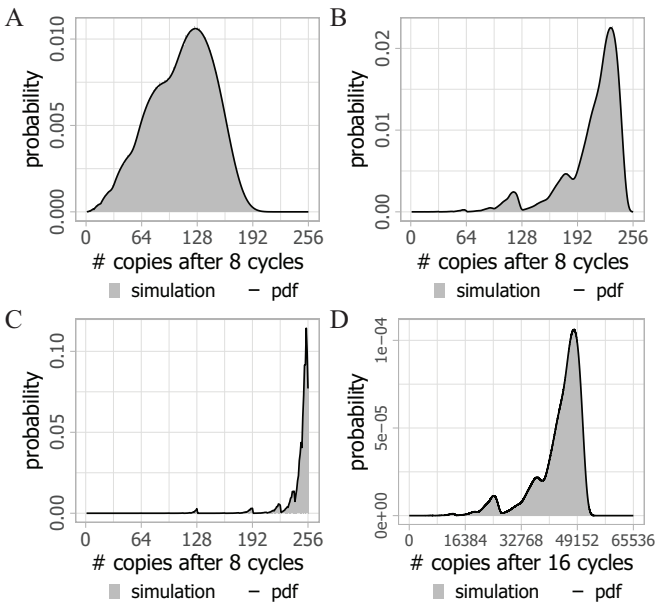


Fig. 1: The pdf of the PCR distribution matches Monte Carlo simulations. For each plot, 10^6 Monte Carlo simulations (blue) were performed to calculate the number of copies produced after (**A**)-(**C**) $l = 8$, (**D**) $l = 16$ cycles at PCR efficiency (**A**) $p = 0.8$, (**B**)$p = 0.95$, (**C**) $p = 0.99$, (**D**) $p = 0.95$. Additionally, the corresponding analytical pdf from Eq. (1) is plotted in black.

We hypothesise that the formula in Eq. (1) cannot be much simplified due to the multiplication of binomial coefficients. The same result can be obtained by formulating the problem as a Markov chain with an appropriate transition probability matrix (see Methods A.5). In Figure 1, we have run Monte Carlo simulations to generate draws from the PCR distribution. Comparing this simulation data to the corresponding pdf given by Eq. (1) yields very close agreement. However due to the nested sums, the pdf is very cumbersome to work with. We will therefore rather utilize the moments the PCR distribution generates.

Despite the rather complicated form of the pdf in Eq. (1), the expectation and variance can be written in much simpler terms:

$$\mathbb{E}[X_l] = (1+p)^l, \tag{2}$$

$$\text{Var}(X_l) = \frac{1-p}{1+p} \cdot \mathbb{E}[X_l] \cdot (\mathbb{E}[X_l] - 1)\,. \tag{3}$$

When the PCR efficiency p approaches 1, Eq. (2) equals a doubling of the copy numbers at each cycle. However since it is common practice to use large cycle numbers $l > 15$, small deviations of the PCR efficiency p get amplified quickly and result in noticeable deviations of the mean value from 2^l. The formula for the expectation has been used for many decades to approximate the PCR efficiency in amplification experiments (Saiki *et al.*, 1985; Li *et al.*, 1988; Keohavong and Thilly, 1989). Eqs. (2), (3) are also given in Weiss and von Haeseler, 1995 (with a typo in the expression for the variance). We provide the derivation of Eqs. (2), (3) in Methods A.2.

With the expectation and variance in hand, we can furthermore derive an equation for the coefficient of variation (CV) given by the ratio of the standard deviation to the expectation. The CV is an excellent measure for

variability as it is comparable between random variables.

$$\mathrm{CV}(X_l) = \sqrt{\frac{1-p}{1+p} \cdot \left(1 - \frac{1}{\mathbb{E}[X_l]}\right)} \tag{4}$$

$$\text{for large } l \approx \sqrt{\frac{1-p}{1+p}}. \tag{5}$$

The CV increases with the number l of PCR cycles due to the term $\left(1 - \frac{1}{\mathbb{E}[X_l]}\right)$. We note that this factor is between 0 and 1 and converges to 1 with increasing l. Therefore, the CV converges to the upper bound $\sqrt{\frac{1-p}{1+p}}$. This upper bound is $\mathrm{CV} = 0.229$ for $p = 0.9$ and $\mathrm{CV} = 0.071$ for $p = 0.99$. The speed of convergence depends on the PCR efficiency p. Performing more PCR cycles does not spoil the data by rising variabilty, but neither improves precision.

By setting a specific convergence threshold (see Methods A.3), we interestingly find that the difference between nearly perfect PCR efficiency ($p = 0.99$) and a PCR efficiency of 0.8 is only one cycle, with 7 or 8 cycles required respectively to reach convergence (Figure S2). This implies that experiments with more than 8 cycles of PCR amplification can generally be considered to have a fixed CV that only depends on the PCR efficiency.

3 Sequencing Probabilities

Let K denote the number of original mRNAs in the cell population. Each transcript $k \in \{1, \dots, K\}$ produces j_k copies during PCR where each j_k is a realization drawn from the PCR distribution X_l. After PCR amplification, all the PCR copies of all transcripts are sequenced together. Let R be the number of total PCR copies that is actually sequenced (i.e. the number of reads). Then each individual PCR copy has a chance $\hat{p}_l$ to be sequenced where $\hat{p}_l$ can be approximated by

$$\hat{p}_l := \frac{R}{\sum_{k=1}^{K} j_k} \approx \frac{R}{K\mathbb{E}[X_l]}. \tag{6}$$

Eq. (6) can be perceived as an estimate for $\hat{p}_l$ on the basis of the outcome of a specific experiment. For one particular original transcript k, the number of PCR copies actually sequenced out of the j_k copies produced is denoted by the random variable Y_{j_k}. It follows a Poisson distribution as we are continuously drawing items from a large pool (on average $\mathbb{E}[X_l]$ copies for each of the k molecules) with an extremely low probability. Such a binomial distribution with large repetition and small probability per trial converges to the Poisson distribution. The corresponding Poisson parameter λ_k is given by $\lambda_k = j_k \cdot \hat{p}_l$.

In the setup described, we will observe an individual original molecule if at least one of its PCR copies is sequenced, meaning $Y_{j_k} > 0$. We call the probability to observe an original molecule p_s. Then, we can show (see Methods A.6) that

$$\begin{aligned} p_s &= \sum_{j_k=1}^{2^l} \mathbb{P}[X_l = j_k] \cdot \mathbb{P}[Y_{j_k} > 0] \\ &= 1 - \left(1 + \frac{1}{2}\hat{p}_l^2 \mathrm{Var}(X_l)\right) \cdot e^{-\hat{p}_l \mathbb{E}[X_l]}. \end{aligned} \tag{7}$$

Utilizing our formulas for the moments of the PCR distribution (Eqs. (2), (3)), we can express p_s in terms of the parameters of the experimental setup being the PCR efficiency p, the number of PCR cycles l and the probability to sequence a PCR copy $\hat{p}_l$:

$$\begin{aligned} p_s \overset{\text{Eqs. (2),(3)}}{=} 1 - \left(1 + \frac{1}{2}\hat{p}_l^2(1-p)(1+p)^{l-1} \cdot \left((1+p)^l - 1\right)\right) \\ \cdot e^{-\hat{p}_l(1+p)^l}. \end{aligned} \tag{8}$$

With the estimate from Eq. (6) for $\hat{p}_l$, we can additionally display p_s to be:

$$p_s \overset{\text{Eq. (6)}}{=} 1 - \left(1 + \frac{R^2}{2K^2}\mathrm{CV}(X_l)^2\right) \cdot e^{-\frac{R}{K}}. \tag{9}$$

The variable R represents the total number of observed reads during sequencing (# reads). K is the total number of original transcripts in the cell population which is roughly given by (# cells) $\cdot$ (# transcripts per cell). We also stated that Y_{j_k} signifies the number of PCR copies actually sequenced for a specific transcript $k \in \{1, \dots, K\}$. Then, we can conclude

$$\#\,\text{reads} = R = \sum_{k=1}^{K} Y_{j_k}. \tag{10}$$

The number of reads R is then correlated to the number of PCR cycles l through the variables Y_{j_k}

In order to obtain a simpler approximation for p_s, we neglect the term $\frac{R^2}{2K^2}\mathrm{CV}(X_l)^2$ in Eq. (9) (second order term, compare Methods A.6.1 and A.7). This simplifies the formula for p_s and we conclude

$$\begin{aligned} & p_s = 1 - e^{-\frac{R}{K}} \\ \Leftrightarrow \qquad & p_s = 1 - e^{-\frac{\#\,\text{reads}}{(\#\,\text{cells}) \cdot (\#\,\text{transcripts per cell})}}. \end{aligned} \tag{11}$$

Eq. (11) provides a clear guideline on how to estimate p_s from experimental data. With these considerations, we can also call p_s the sequencing depth of the data set.

4 Distribution of Observed Counts

So far, we have only considered what happens to one particular transcript of one specific gene that is originally present in one cell. We now investigate the distribution of the total number of transcripts for one particular gene G_1 across multiple cells. Let X_{in} describe the input distribution of the number of transcripts of G_1 found across a cell population and X_{out} be the corresponding output distribution, which represents the observed counts of G_1 across the cell population after the sequencing experiment.

We now aim to relate the pdf of X_{out} to the pdf of X_{in}. In the previous section, we have derived the probability p_s to sequence one particular transcript. We can utilize this result by considering the following: If in one particular cell there are i initial transcripts present of gene G_1 and we observe k of them after the scRNA-seq experiment, it implies that we sequenced k transcripts and failed to sequence $i - k$ transcripts. The number of possibilities to choose k out of these i transcripts is given by the binomial coefficient $\binom{i}{k}$. This leads to a binomial distribution with success probability p_s if the number of initial transcripts was known. Since this number is in fact not known, we need to sum over all probabilities for having i initial transcripts.

These considerations lead us to

$$\begin{aligned} \mathbb{P}[X_{\text{out}} = k] &= \sum_{i=k}^{\infty} \mathbb{P}[X_{\text{in}} = i] \cdot \mathbb{P}[\text{sequencing of } k \text{ out of } i \text{ transcripts}] \\ &= \sum_{i=k}^{\infty} \mathbb{P}[X_{\text{in}} = i] \cdot \binom{i}{k} p_s^k (1-p_s)^{i-k}. \end{aligned} \tag{12}$$

We can replace p_s by Eq. (8) to express the output distribution depending only on the input distribution, the PCR efficiency p, the number of PCR

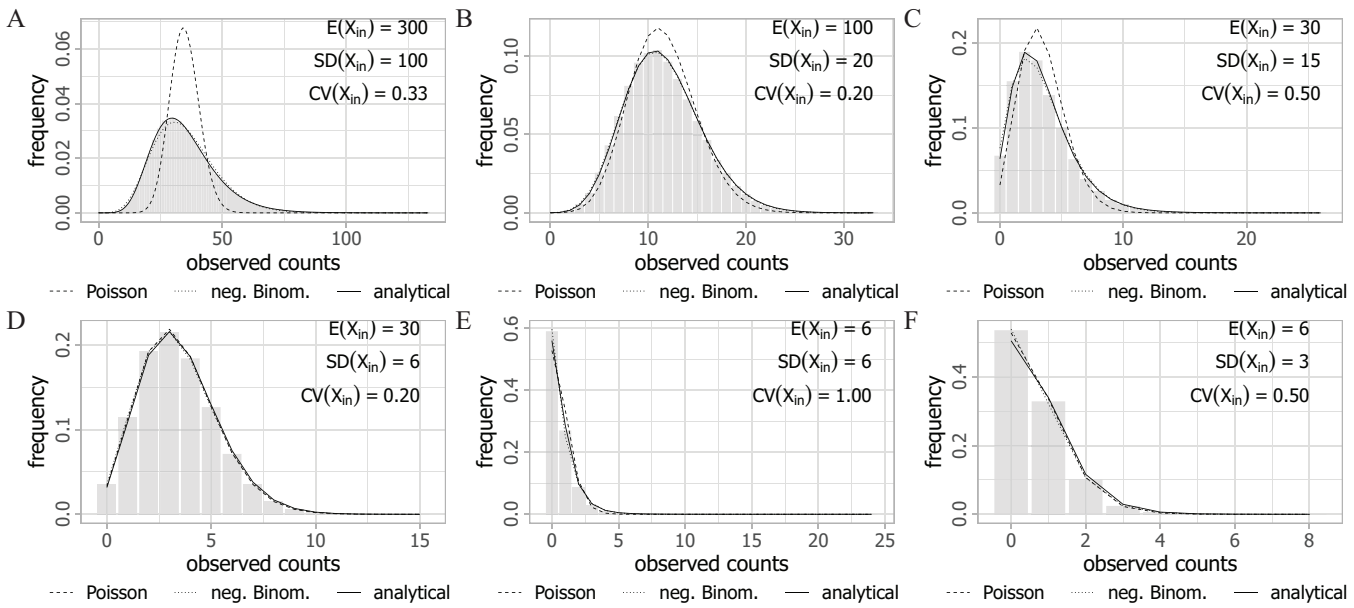


Fig. 2: The analytical output distribution matches Monte Carlo simulations. In all cases, a negative binomial (blue) is fit to the data, as well as a Poisson distribution. Different input parameters are noted inside the plots.

cycles l and the probability to sequence an individual PCR copy $\hat{p}_l$, which are the parameters of the experimental setup. Given a specific input distribution, Eq. (12) completely characterizes the output distribution. It is typically assumed that X_{in} follows a log-normal distribution, $X_{\text{in}} \sim \text{Log-normal}(\mu, \sigma^2)$ (Bengtsson *et al.*, 2005), which will therefore be our choice for simulations (Figure 2).

It is possible to derive formulas for the expectation, variance and CV of X_{out} (see Methods Eqs. (40), (43), (44)). However, in an experimental setting, we are rather interested in concluding characteristics of the input distribution when observing a certain output distribution. For these, we obtain:

$$\mathbb{E}[X_{\text{in}}] = \frac{1}{p_s}\mathbb{E}[X_{\text{out}}], \tag{13}$$

$$\text{Var}(X_{\text{in}}) = \frac{1}{p_s^2}\text{Var}(X_{\text{out}}) - \frac{1-p_s}{p_s^2}\mathbb{E}[X_{\text{out}}], \tag{14}$$

$$\text{CV}(X_{\text{in}}) = \sqrt{\text{CV}(X_{\text{out}})^2 - \frac{1-p_s}{\mathbb{E}[X_{\text{out}}]}}. \tag{15}$$

Since p_s is the probability to observe a single molecule, it is completely reasonable that the expected number of molecules of the output distribution $\mathbb{E}[X_{\text{out}}]$ is given by $p_s \cdot \mathbb{E}[X_{\text{in}}]$. These moment relations apply with all input distributions. If we assumed a certain distribution type such as the log-normal distribution for the input distribution, then it would be completely defined by calculating these moments.

We can show that the CV of the output data is always larger than the CV of the input data in feasible cases:

$$\begin{aligned} & \frac{\text{CV}(X_{\text{out}})^2}{\text{CV}(X_{\text{in}})^2} > 1 \qquad (16)\\ \Leftrightarrow \quad & \frac{1-p_s}{p_s \cdot \mathbb{E}[X_{\text{in}}] \cdot \text{CV}(X_{\text{in}})^2} > 0 \\ \Leftrightarrow \quad & 1 - p_s > 0, \end{aligned}$$

which is almost always true by definition. Only the case $p_s = 1$ does not satisfy this inequality. This would mean $X_{\text{out}} = X_{\text{in}}$ which is experimentally not feasible. For $p_s = 0$, the inequality is not defined as it would imply $X_{\text{out}} \equiv 0$ so that no CV can be defined.

We can derive even more general statements from Eq. (12). It determines all moments of the output distribution for given moments of the input distribution (see Methods A.8.4):

$$\mathbb{E}[X_{\text{out}}^n] = \sum_{k=0}^{n} \left\{ {n \atop k} \right\} p_s^k \, \mathbb{E}[X_{\text{in}}^{\underline{k}}], \tag{17}$$

where $\left\{ {n \atop k} \right\}$ is the Sterling factor of second kind and $X_{\text{in}}^{\underline{k}}$ is the k-th falling power of X_{in} (see Methods Eq. (48)). Vice versa, Eq. (17) can be rearranged to yield all moments of the input distribution from the output moments. We show this by an iterative approach which calculates the n-th moment of X_{in} from moments of X_{in} of order $n-1$ and smaller and moments of X_{out} with highest order n:

$$\mathbb{E}[X_{\text{in}}^n] = \frac{1}{p_s^n}\mathbb{E}[X_{\text{out}}^n] - \mathbb{E}[\underbrace{X_{\text{in}}^{\underline{n}} - X_{\text{in}}^n}_{\substack{\text{highest order} \\ \text{term is } n-1}}] - \sum_{k=0}^{n-1} \left\{ {n \atop k} \right\} p_s^{k-n} \, \mathbb{E}[X_{\text{in}}^{\underline{k}}]. \tag{18}$$

Knowledge of all moments implies a complete characterization of a distribution (Kampen, 2007). Hence, in principle we can recover the input distribution from scRNA-seq data, if the quality of the output data allows for calculation of higher moments.

5 Mean-variance relationship in input and output distributions

While traditionally count data is modelled using Poisson distributions, sequencing count data is typically modelled using negative binomial distributions (Anders and Huber, 2010; Robinson *et al.*, 2010; Hardcastle and Kelly, 2010). For scRNA-seq data, models relying on zero-inflated negative binomial distributions have become popular (Risso *et al.*, 2018; Lopez *et al.*, 2018; Eraslan *et al.*, 2019), although the presence of zero-inflation is still debated and doubtful (Svensson, 2020; Choi *et al.*, 2020; Sarkar and Stephens, 2021). Therefore, we will concentrate on the standard negative binomial distribution.

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

A Poisson distribution has the characteristic that its expectation μ and its variance σ^2 are equal. For a negative binomial distribution with parameters p and r, the following relation between mean and variance holds:

$$\sigma^2 = \mu + \frac{\mu^2}{r}.$$

We can investigate both of these relations empirically in scRNA-seq data of HeLa cells taken from Schwabe *et al.*, 2020.

In Figure 3A, every point represents an individual gene. Clearly visible is the so-called overdispersion which makes the scRNA-seq data fit rather a negative binomial than a Poisson distribution. The reason for this overdispersion is often attributed to cell-to-cell variability (biological noise) (Risso *et al.*, 2018; Lopez *et al.*, 2018; Love *et al.*, 2014).

From our pdf for the output distribution in Eq. (12), we can already conclude that this is an accurate characterization. The presence of cell-to-cell variability is expressed by the fact that X_{in} is a distribution rather than a constant value. If we assumed the opposite, that $X_{\text{in}} \equiv n \geq k$, then:

$$\begin{aligned}\mathbb{P}[X_{\text{out}} = k] &= \sum_{i=k}^{\infty} \mathbb{P}[X_{\text{in}} = i] \cdot \binom{i}{k} p_s^k (1-p_s)^{i-k} \\ &= \binom{n}{k} p_s^k (1-p_s)^{n-k}.\end{aligned}$$

This is a binomial distribution with small success probability p_s. We know that such a binomial can be approximated by a Poisson distribution. Hence in the absence of cell-to-cell variability, the sequencing count data would indeed follow a Poisson distribution.

For numerical simulations in Figure 2, we choose a log-normal distribution as our input distribution. For multiple parameter choices of the log-normal distribution, we generate observed counts as described in Figure S4. We also set $p_s = 0.116$ as we have estimated this for a particular scRNA-seq experiment involving HeLa cells (see Methods A.7). We observe that the analytical output distribution matches the simulation data well (Figure 2), which confirms Eq. (12). In addition to the analytical pdf, we also fit a Poisson and a negative binomial distribution. In all examples, it is possible to fit a negative binomial distribution very close to the analytical distribution, justifying its usage in numerical simulation. The Poisson fit on the other hand differs noticeably from the numerical simulation for high expression genes. The larger the CV of the input data, the larger the differences become. This is in line with our previous assertion that a constant input distribution would provide a Poisson output distribution. For lowly expressed genes on the other hand, the Poisson distribution provides a good fit to the data despite a large CV of the input data. The same fact can be derived from the mean-variance relation in Figure 3A.

The input distribution types fitting experimental results well - negative binomial and log-normal - both have 2 independent parameters (Risso *et al.*, 2018; Lopez *et al.*, 2018; Eraslan *et al.*, 2019). A priori, we expect each gene to have its specific value for these parameters independently of the values for other genes, since transcription might be coordinated within groups of genes but not across all genes. That is, we do not expect a priori a systematic and fixed relation between the distribution parameters across all genes. If contrary to this expectation such a relation exists, it attests to a process involved in transcription fixing the relation between mean and variance across genes. Such a process might be established by transcription regulation or by very basic properties of the transcription mechanism. It is well known and also illustrated by Figure 3, that a relation between mean and variance exists in the output of scRNA-seq experiments and consequently also in the input distribution. We cannot identify the process establishing it, but we can turn the mean-variance relation of the output into a mean-variance relation of the input distribution with our results. That provides a quantitative property against which hypotheses on the nature on the process fixing the mean-variance relation across genes can be verified.

We express the mean-variance relation for the input of the data in Figure 3B by its Taylor expansion up to third order

$$\text{Var}(X_{\text{in}}) = \sum_{i=1}^{3} a_i \mathbb{E}[X_{\text{in}}]^i. \tag{19}$$

The resulting mean-variance relation fits the data reasonably well confirming our choice of terms of the Taylor expansion. If we furthermore assume the input distributions for the data to be log-normal (Bengtsson *et al.*, 2005), a relation between the two parameters μ and σ^2 of a log-normal distribution can be established (see Methods A.9). This means that if the output data shown in Figure 3 do stem from log-normal distributions,

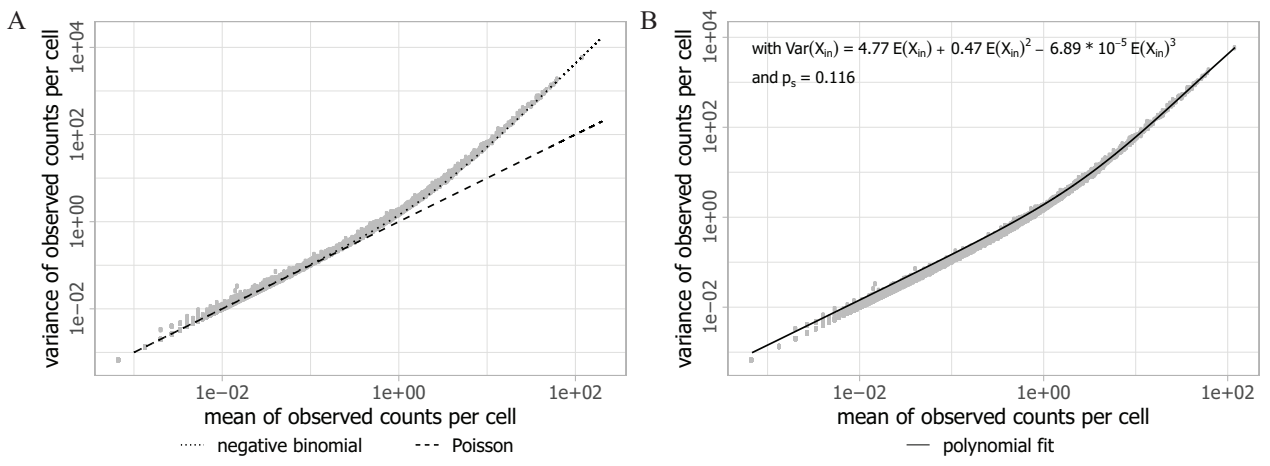


Fig. 3: The mean-variance relation of scRNA-seq data for individual genes of a HeLa data set (Schwabe *et al.*, 2020) satisfies a negative binomial rather than a Poisson distribution and establishes a relation between mean and variance of the input distribution. Each point represents an individual gene. Only non-variable genes with dispersion smaller than 0 are shown (see Methods of Schwabe *et al.*, 2020). (**A**) A Poisson distribution (green) and negative binomial distribution (blue) are fitted to the data cloud. (**B**) By incorporating a cubic relation between the mean and variance of the input distribution, we find a good fit to the data. All three parameters of the cubic function are found to be significant for the fit.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

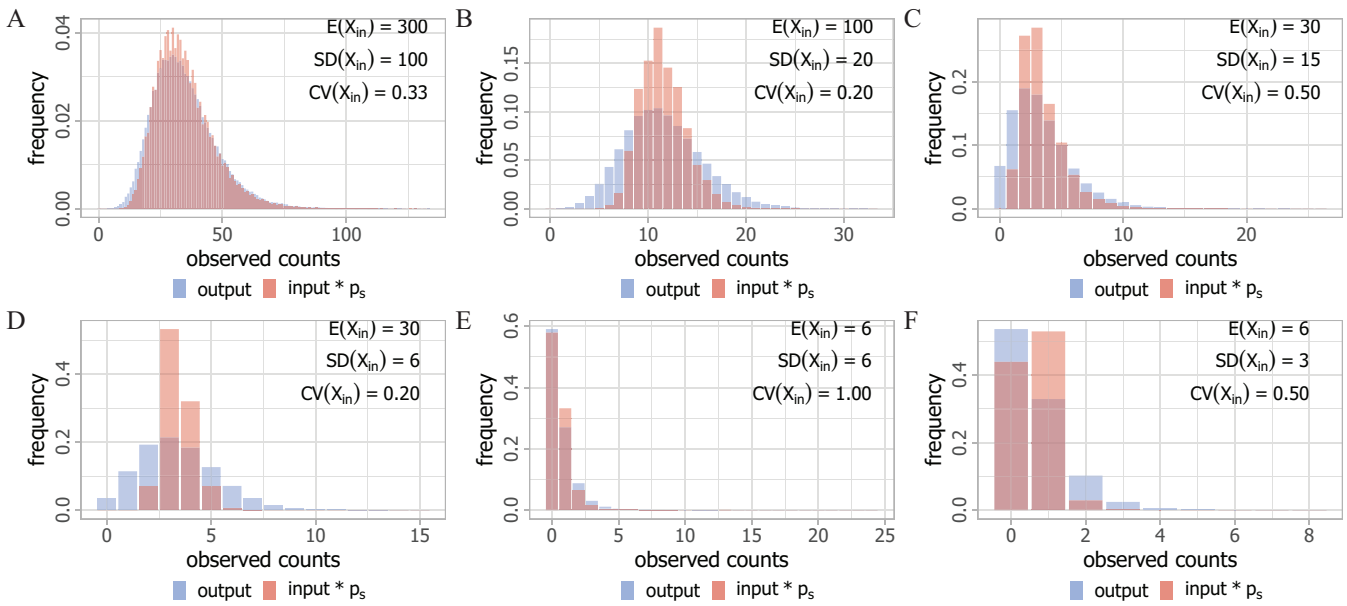


Fig. 4: Visualization of the differences between input and output distribution. The histogram of the output distribution X_{out} (blue) of 10^5 cells is plotted together with the histogram of the re-scaled input values $p_s \cdot X_{\text{in}}$ (red) for $p_s = 0.116$ (see Methods A.7). It illustrates that the CV of the output is always larger than the one of the input according to Eq. (16).

the parameters of these distributions obey:

$$\mu = \ln\left(\frac{\mathrm{e}^{\sigma^2} - a_2 - 1 \pm \sqrt{\left(\mathrm{e}^{\sigma^2} - a_2 - 1\right)^2 - 4a_1a_3}}{2a_3}\right) - \frac{\sigma^2}{2}. \quad (20)$$

The parameters a_1, a_2 and a_3 are fitted from the data and are displayed for the example data utilized here in Figure 3B. The quality of the fit suggests that we can specify the general assumption of a log-normal distribution for cellular transcripts to a log-normal distribution with a relation between the parameters like Eq. (20). The remaining parameter σ is fixed by the average transcript number. We obtain with the values from Figure 3B for a_1, a_2 and a_3 for our HeLa data set

$$\mathrm{Var}(X_{\text{in}}) = 4.77\,\mathbb{E}[X_{\text{in}}] + 0.47\,\mathbb{E}[X_{\text{in}}]^2 - 6.89\,10^{-5}\,\mathbb{E}[X_{\text{in}}]^3 \quad (21)$$

$$\mu = \ln\left(\left|\mathrm{e}^{\sigma^2} - 1.47 - \sqrt{\left(\mathrm{e}^{\sigma^2} - 1.47\right)^2 + 1.315\,10^{-3}}\right|\right) + 8.89 - \frac{\sigma^2}{2}. \quad (22)$$

Note, this equation is a relation between average μ and σ. The averaging is over all genes with the same average copy number.

Another question of interest is to investigate how the shape of the input distribution compares to the output distribution. In order to compare these two distributions within one plot, we scale the input distribution by the factor p_s due to the relation of the expectations from Eq. (2): $\mathbb{E}[X_{\text{out}}] = p_s \cdot \mathbb{E}[X_{\text{in}}]$. We observe in Figure 4 that the output distributions have fatter tails, while the input distributions are rather weighted towards its mean. This confirms the conclusion in Eq. (16) that the CV of the output distribution is larger than the input distribution.

6 Relation between input and output CV

From Eq. (15), we know that

$$\mathrm{CV}(X_{\text{out}}) = \sqrt{\frac{1 - p_s}{p_s \cdot \mathbb{E}[X_{\text{in}}]} + \mathrm{CV}(X_{\text{in}})^2}. \quad (23)$$

We investigate the dependence of $\mathrm{CV}(X_{\text{out}})$ on $\mathrm{CV}(X_{\text{in}})$ and the sequencing probability p_s (which we can also call the sequencing depth) in Figure 5. Most noticeably, when $\mathrm{CV}(X_{\text{in}})$ attains large values and thus dominates all other terms, the relation approaches a linear dependency. This is obvious also from Eq. (23). Considering the fold change of $\mathrm{CV}(X_{\text{out}})$ to $\mathrm{CV}(X_{\text{in}})$, we observe that small $\mathrm{CV}(X_{\text{in}})$ cause huge amplification of the variability whereas the fold change approaches 1 for larger values of $\mathrm{CV}(X_{\text{in}})$.

The effects of the sequencing depth p_s on $\mathrm{CV}(X_{\text{out}})$ are altogether not surprising as well. The more transcripts are sequenced, the smaller the increase of $\mathrm{CV}(X_{\text{out}})$. However, we do note that for the choice of $\mathrm{CV}(X_{\text{in}}) = 0.2$ and $\mathbb{E}[X_{\text{in}}] = 300$ displayed here, the sequencing depth $p_s = 0.116$ is still located in an area of the hyperbola branch with increased growth. Therefore, increases in the sequencing depth p_s would still benefit noise reduction noticeably. Interestingly, these improvements start to drop off quickly (consider the slope of the graph). That means that improvements past a sequencing depth of around $p_s = 0.5$ would have far smaller effects on $\mathrm{CV}(X_{\text{out}})$. The same results are plotted in Figure S5 and Figure S6 in three dimensions for further illustration.

7 Discussion

Simplified models do not capture every detail of an experiment and the precision of their predictions is limited. However given cell-to-cell variability, capturing all details limits the scope of models (and precision beyond the relative variance of noise is meaningless anyway). The advantage of simplified models is that they provide equations which can be used for a large class of experiments. In that spirit, we investigate

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

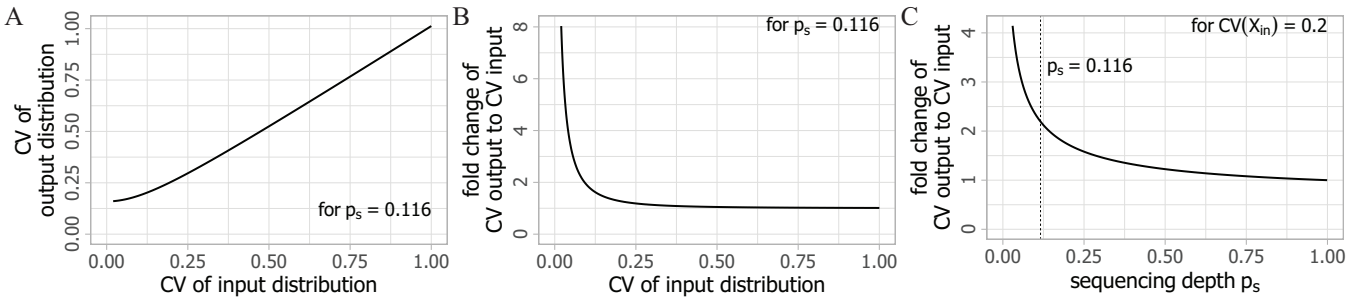


Fig. 5: The CV of X_{out} shows characteristic dependencies on CV(X_{in}) and p_s providing guidelines for obtaining desired noise levels. Effects of changes in input CV and success probability p_s on the output CV.

the processes involved in scRNA-seq with a simplified model. It enables us to derive a set of equations useful for experiment design and analysis of results, and sheds light on some of the causes for and implications of various noise terms arising during scRNA-seq experiments.

Within the validity of our model setup, Eqs. (1), (2) , (3), (8), (12) provide a complete characterization of a scRNA-seq experiment in terms of the experimental parameters PCR efficiency p, PCR cycle number l and the probability to sequence a PCR copy $\hat{p}_l$. An experiment can be designed on that basis to provide a desired precision. The output distribution of an individual gene obeys Eq. (12) for a given input distribution. Knowledge of the moments of the input distribution independent of the distribution type suffice if we want to determine only moments of the output distribution. That is especially interesting in light of the finding that the input distribution obeys moment relations fixing its variance to the mean (Fig. 3, Eq. (19)). Hence, both mean and variance of the output are determined by the mean of the input (see Eqs. (40), (43)).

The purpose of scRNA-seq experiments usually is to determine the input distribution from the experiment's output. We solve this inverse problem by calculating all moments of the input distribution from the moments of the output distribution by Eq. (18). Since in principle a distribution is completely determined if all its moments are known (Kampen, 2007), this recovers the complete information of the input from the output distribution.

In practice, the output distributions are often not of sufficient quality allowing for the calculation of all moments. However, Eq. (18) enables us to determine as many input moments as we know output moments and thus guarantees that we can use all the information hidden in a measured output distribution. It is generally assumed that cellular RNA copy numbers obey a log-normal distribution (Bengtsson *et al.*, 2005). That assumption can be verified on the basis of Eq. (18) if output distributions providing three or more moments are available.

We note that the assumptions we have made in our example simulations for the input distribution are limited to homogeneous cell populations. Our formulas can in principal be applied to heterogeneous cell populations with bi- or multi-modal distributions just as well. The number of moments to be determined from the output distribution to recover the input distribution needs to be as large as the number of unknown parameters of the input distribution. If this is possible, appropriate ansatzes for the input distribution may also allow for calculating its parameters and thus to determine bi-modal input as well.

We apply our results to a HeLa data set (Fig. 3). Two of the fundamental characteristics of single-cell data are that they exhibit relations between mean and variance and show higher variability than is typically expected for count data. The latter is called overdispersion (in comparison to a Poisson distribution). We determine the moment relation of the HeLa data set on the basis of Eqs. (40), (43) and obtain very good agreement with a third order polynomial (Fig. 3B). Assuming that the type of input distribution is the same for most of the genes (as Fig. 3A suggests to be the case) and that type to be log-normal, we determined the relation between the parameters of the cellular distribution characterising the HeLa data set. That is the complete characterization of the scRNA-seq input data which is possible within the validity of the assumptions since for each average number of RNA copies we can calculate the copy number distribution. The moment relation reproduces of course also the overdispersion in the HeLa data set. Relating it to Eq. (12) confirms overdispersion in general to be due to cell-to-cell variability, as has been assumed before.

The square root of the coefficient a_2 of the moment relation of the input distribution (Eq. (19)) is the coefficient of variation for highly expressed genes. The fit in Figure 3B shows this CV to be 0.686 for the HeLa data set. That means among the cells with transcript numbers within the range mean$\pm$SD we find a variability up to factor 5.4 ($\frac{\text{mean+SD}}{\text{mean−SD}} = \frac{1+\text{CV}}{1-\text{CV}}$) in transcript copy numbers, which we feel is surprisingly large for highly expressed genes. The CV for lowly expressed genes is even larger.

This large cell-to-cell variability within a single cell type raises the question how a cell type or a cell state can be defined. Is it a qualitative concept only, which denotes a specific list of expressed genes (essentially the DNA and its epigenetic state) as the state of a cell? Or, is it also possible to define quantitative criteria? That question does not only appear with respect to gene expression profiles but also arises from functional data exhibiting large cell-to-cell variability (Thurley *et al.*, 2014). Cell-to-cell variability renders simply adding copy numbers to the list of genes as a quantitative definition of cell state meaningless. However, data describing cell variability like Eq. (21) might provide a quantitative description of cell state. It is valid for all genes translated in a given state. Ideally, it should not depend on the experimental methods used to establish it. Of course, we cannot guarantee that for the very parameter values of Eq. (21), since p_s enters the calculation (Eq. (56)). But it will be very interesting to compare Eq. (21) to future results from other labs and conditions to learn about its degree of universality.

We note an interesting concurrent observation with functional and translational data. With Ca^{2+} spiking data reported in Thurley *et al.*, 2014, it was the relation between the variance (or standard deviation) and the average of interspike intervals which applied to all individual cells in a given state, while the average interspike interval exhibited large cell-to-cell variability. We observe the same here. Individual copy numbers of transcripts scatter by a factor 5, but the moment relation Eq. (21) is universal for all individual cells. In both cases, cell-to-cell variability appears to prevent quantitative definitions of cell state. In both cases, experimental results show that surprisingly a moment relation exists (with regard to Ca^{2+} spiking, Skupin and Falcke, 2009 show that spike generating systems in general do not exhibit a unique moment relation) and these relations quantifying cell variability are the universal relations.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

This suggests a definition of cell state, where the DNA and its epigenetic state provide for the qualitative description and the moment relations of cell variability for a quantitative description.

The results of our study can be used in several ways. They help to optimize the experimental parameters and they can be used to obtain as much information on the input distribution moments as desired and supported by the quality of the output data. It is a theory study which we hope will be picked up and checked by experimental labs.

Acknowledgements

We would like to thank Laleh Haghverdi for very helpful comments and advice on the manuscript.

Funding

DS was a member of the Computational Systems Biology graduate school (GRK1772), which was funded by Deutsche Forschungsgemeinschaft (DFG).

Conflict of interest

The authors declare no competing interests.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**(10), R106.

Bacher, R., Chu, L. F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., and Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods*, **14**(6), 584–586.

Bengtsson, M., Stahlberg, A., Rorsman, P., and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res*, **15**(10), 1388–92.

Booth, C. S., Pienaar, E., Termaat, J. R., Whitney, S. E., Louw, T. M., and Viljoen, H. J. (2010). Efficiency of the polymerase chain reaction. *Chem Eng Sci*, **65**(17), 4996–5006.

Breda, J., Zavolan, M., and van Nimwegen, E. (2021). Bayesian inference of gene expression states from single-cell RNA-seq data. *Nat Biotechnol*, **39**(8), 1008–1016.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*, **36**(5), 411–420.

Choi, K., Chen, Y., Skelly, D. A., and Churchill, G. A. (2020). Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol*, **21**(1), 183.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, **297**(5584), 1183–6.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun*, **10**(1), 390.

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

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol*, **36**(5), 421–427.

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Harris, T. E. (1964). *The theory of branching process*. Rand Corporation Rand report. Rand Corp., Santa Monica, Calif.,.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, **19**(4), 562–578.

Hou, W., Ji, Z., Ji, H., and Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol*, **21**(1), 218.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods*, **15**(7), 539–542.

Kampen, N. G. v. (2007). *Stochastic processes in physics and chemistry*. North-Holland personal library. Elsevier, Amsterdam ; New York, 3rd edition.

Keohavong, P. and Thilly, W. G. (1989). Fidelity of DNA polymerases in DNA amplification. *Proc Natl Acad Sci U S A*, **86**(23), 9253–7.

Li, H. H., Gyllensten, U. B., Cui, X. F., Saiki, R. K., Erlich, H. A., and Arnheim, N. (1988). Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature*, **335**(6189), 414–7.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell RNA-seq data. *Nat Commun*, **9**(1), 997.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat Methods*, **15**(12), 1053–1058.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, **15**(12), 550.

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol*, **17**, 75.

Mendenhall, A. R., Martin, G. M., Kaeberlein, M., and Anderson, R. M. (2021). Cell-to-cell variation in gene expression and the aging process. *Geroscience*, **43**(1), 181–196.

Osorio, D., Yu, X., Zhong, Y., Li, G., Yu, P., Serpedin, E., Huang, J. Z., and Cai, J. J. (2019). Single-cell expression variability implies cell function. *Cells*, **9**(1).

Peccoud, J. and Jacob, C. (1996). Theoretical uncertainty of measurements using quantitative polymerase chain reaction. *Biophys J*, **71**(1), 101–8.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun*, **9**(1), 284.

Roberfroid, S., Vanderleyden, J., and Steenackers, H. (2016). Gene expression variability in clonal populations: Causes and consequences. *Crit Rev Microbiol*, **42**(6), 969–84.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–40.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**(4732), 1350–4.

Sarkar, A. and Stephens, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat Genet*, **53**(6), 770–777.

Schwabe, D., Formichetti, S., Junker, J. P., Falcke, M., and Rajewsky, N. (2020). The transcriptome dynamics of single cells during the cell cycle. *Mol Syst Biol*, **16**(11), e9946.

Skupin, A. and Falcke, M. (2009). From puffs to global ca2+ signals: how molecular properties shape global signals. *Chaos*, **19**(3), 037111.

Stolovitzky, G. and Cecchi, G. (1996). Efficiency of dna replication in the polymerase chain reaction. *Proc Natl Acad Sci U S A*, **93**(23), 12947–52.

Sun, M. and Zhang, J. (2020). Allele-specific single-cell RNA sequencing reveals different architectures of intrinsic and extrinsic gene expression noises. *Nucleic Acids Res*, **48**(2), 533–547.

Svec, D., Tichopad, A., Novosadova, V., Pfaffl, M. W., and Kubista, M. (2015). How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments. *Biomol Detect Quantif*, **3**, 9–16.

Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol*, **38**(2), 147–150.

Thurley, K., Tovey, S. C., Moenke, G., Prince, V. L., Meena, A., Thomas, A. P., Skupin, A., Taylor, C. W., and Falcke, M. (2014). Reliable encoding of stimulus intensities within random sequences of intracellular ca2+ spikes. *Sci Signal*, **7**(331), ra59.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*, **21**(1), 12.

Tsimring, L. S. (2014). Noise in biology. *Rep Prog Phys*, **77**(2), 026601.

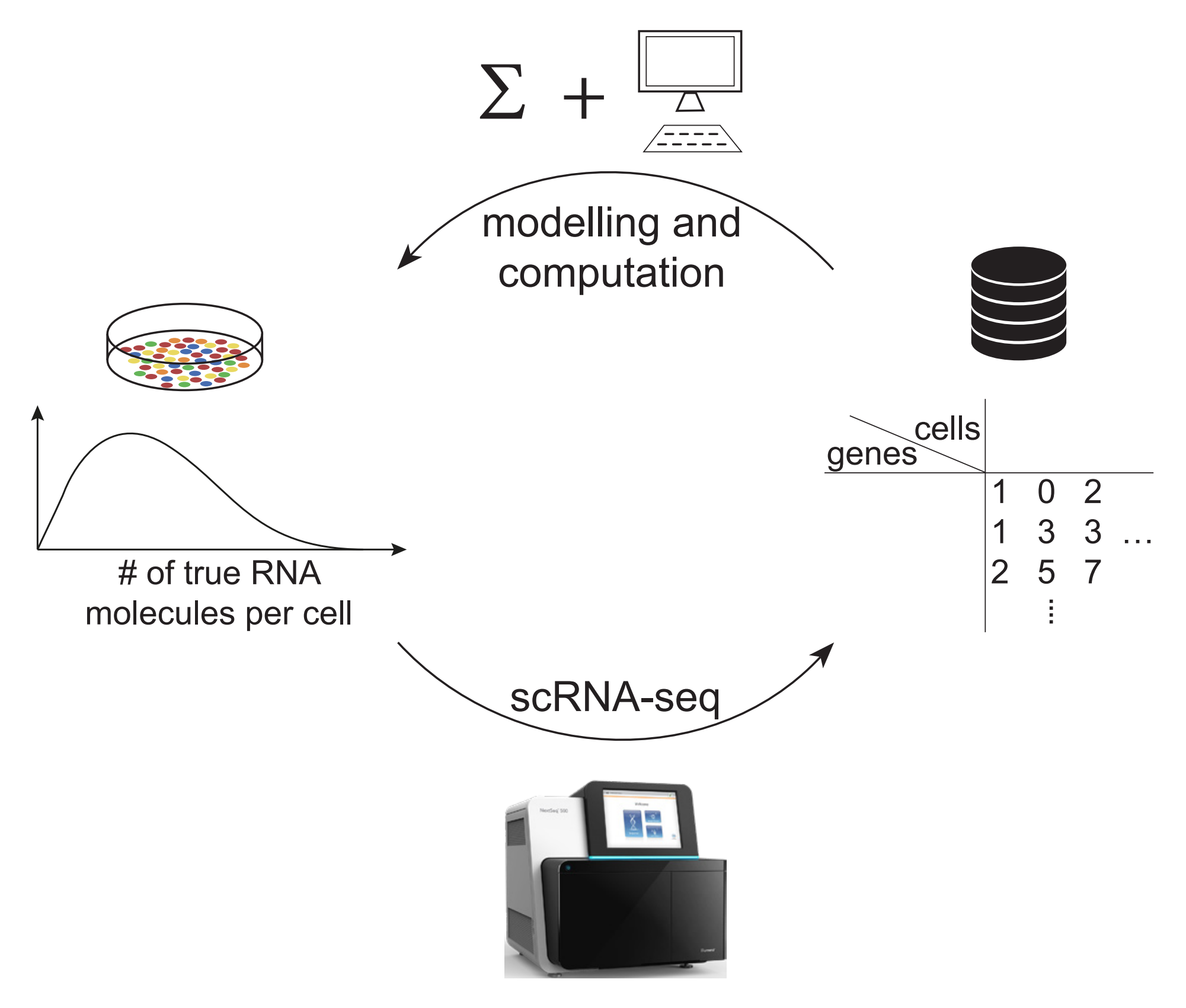
Tung, P. Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci Rep*, **7**, 39921.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**(3), 716–729 e27.

Weiss, G. and von Haeseler, A. (1995). Modeling the polymerase chain reaction. *J Comput Biol*, **2**(1), 49–61.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture


Σ +
modelling and computation
cells
genes
1 0 2
1 3 3 …
2 5 7
# of true RNA molecules per cell
scRNA-seq

Supplementary data for the manuscript "On the relation between input and output distributions of scRNA-seq experiments"

Daniel Schwabe[1] and Martin Falcke[1,2,*]

[1]Mathematical Cell Physiology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[2]Department of Physics, Humboldt University Berlin, Newtonstr. 15, 12489 Berlin, Germany
[*]Correspondence to: martin.falcke@mdc-berlin.de

A Methods

A.1 The probability distribution function caused by PCR

In order to derive a pdf for the PCR distribution, we start at the final copy number k and consider the possible copy numbers i_{l-1} at the previous $(l-1)$ cycle that could have produced k final copies. This subsequently traces the experiment backwards until we arrive at the first PCR cycle. Each individual cycle represents a draw of the binomial distribution. We calculate

$$\begin{aligned}
\mathbb{P}[X_l = k] &= \sum_{i_{l-1}=\lceil \frac{k}{2} \rceil}^{k} \underbrace{\mathbb{P}[X_l = k \mid X_{l-1} = i_{l-1}]}_{\text{Binomial distrib.}} \cdot \mathbb{P}[X_{l-1} = i_{l-1}] \\
\xrightarrow{\substack{\text{plugging in binomial distribution} \\ \text{to draw } k - i_{l-1} \text{ out of } i_{l-1}}} &= \frac{p^k}{(1-p)^k} \sum_{i_{l-1}=\lceil \frac{k}{2} \rceil}^{k} \binom{i_{l-1}}{k - i_{l-1}} p^{-i_{l-1}} (1-p)^{2 i_{l-1}} \cdot \mathbb{P}[X_{l-1} = i_{l-1}] \\
\xrightarrow{\substack{\text{plugging in binomial distribution} \\ \text{for each subsequent cycle}}} &= \frac{p^k}{(1-p)^k} \sum_{i_{l-1}=\lceil \frac{k}{2} \rceil}^{k} \sum_{i_{l-2}=\lceil \frac{i_{l-1}}{2} \rceil}^{i_{l-1}} \cdots \\
&\quad \sum_{i_1=\lceil \frac{i_1}{2} \rceil}^{i_2} \binom{i_{l-1}}{k - i_{l-1}} \cdot \ldots \cdot \binom{i_1}{i_2 - i_1} (1-p)^{\sum_{m=1}^{l-1} i_m} (1-p)^{i_1} p^{-i_1} \cdot \mathbb{P}[X_1 = i_1] \\
\xrightarrow{\substack{(1-p)^{i_1} p^{-i_1} \cdot \mathbb{P}[X_1 = i_1] = (1-p)^2 p^{-1} \\ \text{for both } i_1 = 1 \text{ and } i_1 = 2}} &= \frac{p^{k-1}}{(1-p)^{k-2}} \sum_{\substack{i_{l-1}=\lceil \frac{k}{2} \rceil \\ i_{l-1} \le 2^{l-1}}}^{k} \sum_{i_{l-1}=\lceil \frac{k}{2} \rceil}^{\min\{k, 2^{l-1}\}} \sum_{i_{l-2}=\lceil \frac{i_{l-1}}{2} \rceil}^{\min\{i_{l-1}, 2^{l-2}\}} \cdots \\
&\quad \sum_{i_1=\lceil \frac{i_2}{2} \rceil}^{\min\{i_2, 2\}} \binom{i_{l-1}}{k - i_{l-1}} \cdot \ldots \cdot \binom{i_1}{i_2 - i_1} (1-p)^{\sum_{m=1}^{l-1} i_m}. \qquad (24)
\end{aligned}$$

A.2 Proof for the moments of the PCR distribution

We now prove the Eqs. (2), (3) with the help of the induction principle.

A.2.1 Expectation

We stated previously that the increase in number of copies from X_{l-1} to X_l follows a binomial distribution $(X_l - X_{l-1} | X_{l-1} = n) \sim \text{Binom}(n, p)$. We also note that for any binomial distribution $Z \sim \text{Binom}(n, p)$, it holds that $\mathbb{E}[Z] = n \cdot p$. Hence,

$$\mathbb{E}[X_l - X_{l-1} | X_{l-1} = n] = n \cdot p \qquad (25)$$

and for the conditional expectation it holds

$$\mathbb{E}[X_l - X_{l-1}|X_{l-1}] = X_{l-1} \cdot p. \tag{26}$$

We now perform a proof via induction.
Base case: $l = 1$

$$\begin{aligned}\mathbb{E}[X_1] &= \sum_{k=1}^{2} k \cdot \mathbb{P}[X_1 = k] \\ &= 1 \cdot \mathbb{P}[X_1 = 1] + 2 \cdot \mathbb{P}[X_1 = 2] \\ &= 1 + p.\end{aligned}$$

Induction hypothesis: $\mathbb{E}[X_{l-1}] = (1+p)^{l-1}$.
Induction step:

$$\begin{aligned}\mathbb{E}[X_l] &= \mathbb{E}[X_l - X_{l-1}] + \mathbb{E}[X_{l-1}] \\ \text{induction hypothesis} \rightarrow \quad &= (1+p)^{l-1} + \mathbb{E}[X_l - X_{l-1}] \\ \text{law of total expectation} \rightarrow \quad &= (1+p)^{l-1} + \mathbb{E}[\mathbb{E}[X_l - X_{l-1}|X_{l-1}]] \\ \text{Eq. (26)} \rightarrow \quad &= (1+p)^{l-1} + p\mathbb{E}[X_{l-1}] \\ \text{induction hypothesis} \rightarrow \quad &= (1+p)^{l-1} + p(1+p)^{l-1} \\ &= (1+p)^{l}.\end{aligned}$$

□

Hence, the expectation is calculated with the help of the expectation of the output distribution.

A.2.2 Variance

From $(X_l - X_{l-1}|X_{l-1} = n) \sim \text{Binom}(n, p)$, we can also conclude that $\text{Var}(X_l - X_{l-1}|X_{l-1} = n) = p \cdot (1-p) \cdot n$ and furthermore

$$\text{Var}(X_l - X_{l-1}|X_{l-1}) = p \cdot (1-p) \cdot X_{l-1}. \tag{27}$$

The law of total variance states for two random variables X and Y where X has finite variance that

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}\left(\mathbb{E}[X|Y]\right). \tag{28}$$

For the proof of the variance formula in Eq. (3), we need to assess the variance of the increase of copy numbers between two cycles $\text{Var}(X_l - X_{l-1})$:

$$\begin{aligned}\text{Var}(X_l - X_{l-1}) &\overset{\text{Eq. (28)}}{=} \mathbb{E}\left[\text{Var}(X_l - X_{l-1}|X_{l-1})\right] + \text{Var}\left(\mathbb{E}[X_l - X_{l-1}|X_{l-1}]\right) \\ &\overset{\text{Eq. (27)}}{=} p(1-p)\mathbb{E}[X_{l-1}] + \text{Var}\left(\mathbb{E}[X_l - X_{l-1}|X_{l-1}]\right) \\ &\overset{\text{Eq. (26)}}{=} p(1-p)\mathbb{E}[X_{l-1}] + p^2\text{Var}\left(X_{l-1}\right).\end{aligned} \tag{29}$$

We also require an expression for the covariance of $X_l - X_{l-1}$ and X_{l-1}:

$$\begin{aligned}&\text{Cov}(X_l - X_{l-1}, X_{l-1}) \\ &= \sum_{i=1}^{2^{l-1}} \sum_{j=0}^{i} \left(i - \mathbb{E}[X_{l-1}]\right)\left(j - \mathbb{E}[X_l - X_{l-1}]\right)\mathbb{P}[X_l - X_{l-1} = j, X_{l-1} = i] \\ \text{Bayes} \rightarrow \quad &= \sum_{i=1}^{2^{l-1}} \left(i - \mathbb{E}[X_{l-1}]\right)\mathbb{P}[X_{l-1} = i] \sum_{j=0}^{i} \left(j - \mathbb{E}[X_l - X_{l-1}]\right)\mathbb{P}[X_l - X_{l-1} = j|X_{l-1} = i] \\ \text{law of total expectation and Eq. (26)} \rightarrow \quad &= \sum_{i=1}^{2^{l-1}} \left(i - \mathbb{E}[X_{l-1}]\right)\mathbb{P}[X_{l-1} = i] \sum_{j=0}^{i} \left(j - p\mathbb{E}[X_{l-1}]\right)\mathbb{P}[X_l - X_{l-1} = j|X_{l-1} = i] \\ \text{conditional expectation and total probability} \rightarrow \quad &= \sum_{i=1}^{2^{l-1}} \left(i - \mathbb{E}[X_{l-1}]\right)\mathbb{P}[X_{l-1} = i]\left(\mathbb{E}[X_l - X_{l-1}|X_{l-1} = i] - p\mathbb{E}[X_{l-1}]\right)\end{aligned} \tag{30}$$

$$\begin{aligned}\text{Eq. (25)} \rightarrow \quad &= p \sum_{i=1}^{2^{l-1}} \left(i - \mathbb{E}[X_{l-1}]\right)^2 \mathbb{P}[X_{l-1} = i] \\ &= p\text{Var}(X_{l-1}). \end{aligned} \tag{31}$$

Now, we can proof Eq. (3) via induction:
Base case: $l = 1$

$$\begin{aligned}\text{Var}(X_1) &= \sum_{k=1}^{2} (k - \mathbb{E}[X_1])^2 \mathbb{P}[X_1 = k] \\ &= (1 - (1+p))^2 \mathbb{P}[X_1 = 1] + (2 - (1+p))^2 \mathbb{P}[X_1 = 2] \\ &= p^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p \cdot (1-p) \\ &= \frac{1-p}{1+p}(1+p)(1+p-1) \\ &= \frac{1-p}{1+p}\mathbb{E}[X_1](\mathbb{E}[X_1] - 1).\end{aligned}$$

Induction hypothesis: $\text{Var}(X_{l-1}) = \frac{1-p}{1+p} \cdot \mathbb{E}[X_{l-1}] \cdot (\mathbb{E}[X_{l-1}] - 1)$.
Induction step:

$$\begin{aligned}\text{Var}(X_l) &= \text{Var}(X_l - X_{l-1}) + \text{Var}(X_{l-1}) + 2\text{Cov}(X_l - X_{l-1}, X_{l-1}) \\ \text{Eqs. (29), (31)} \rightarrow \quad &= p(1-p)\mathbb{E}[X_{l-1}] + (1+p)^2 \text{Var}\left(X_{l-1}\right) \\ \text{induction hypothesis} \rightarrow \quad &= (1+p)(1-p) \cdot \mathbb{E}[X_{l-1}]^2 - (1-p)\mathbb{E}[X_{l-1}] \\ \mathbb{E}[X_l]=(1+p)^l \rightarrow \quad &= \frac{1-p}{1+p} \cdot \mathbb{E}[X_l] \cdot (\mathbb{E}[X_l] - 1). \end{aligned}$$

□

A.3 Convergence of the coefficient of variation of the PCR distribution

We define that the term $\left(1 - \frac{1}{\mathbb{E}[X_l]}\right)$ from Eq. (4) is negligible when $\frac{1}{\mathbb{E}[X_l]} < 0.01$ since then $0.994 < \sqrt{1 - \frac{1}{\mathbb{E}[X_l]}} < 1$ meaning the actual CV deviates less than 0.6% from the approximation. This essentially removes the dependency of the CV on the number of cycles l. We call the term $\frac{1}{\mathbb{E}[X_l]}$ the CV convergence factor which should be large when l is small and approach 0 when l is large.

In Figure S2, we observe a sharp drop of the CV convergence factor between 1 and 7 cycles. We have coloured all values $\frac{1}{\mathbb{E}[X_l]} < 0.01$ white so as to investigate how many PCR cycles are necessary to saturate the CV at different PCR efficiencies.

A.4 Simulation data for the PCR distribution matches formulas for moments

In addition to the visual confirmation of our calculations in Figure 1, we can also investigate the mean μ, variance σ^2 and CV of the simulation data and compare them to our analytical formulas in Eqs. (2), (3), (5) for X_l:

	μ	$\mathbb{E}[X_l]$	σ	$\text{SD}(X_l)$	σ^2	$\text{Var}(X_l)$	CV	$\text{CV}(X_l)$
$l = 8$, $p = 0.8$	110.18	110.20	36.55	36.57	1336	1337	0.332	0.332
$l = 8$, $p = 0.95$	209.12	209.06	33.37	33.40	1114	1115	0.160	0.160
$l = 8$, $p = 0.99$	245.92	249.94	17.43	17.40	304	303	0.071	0.071
$l = 16$, $p = 0.95$	43713	43707	6995	6999	$48 \cdot 10^6$	$48 \cdot 10^6$	0.160	0.160

The data are within a very reasonable error margin and therefore confirm our analytical conclusions.

A.5 Markov chain formulation of the PCR distribution

The problem of PCR amplification can also be posed as a Markov chain illustrated for $l = 2$ in Figure S3. We note that for larger l the stationary probabilities for state 3 and 4 will have to be adjusted. These are set to 1 for X_2 in order to adhere to the rule that the rows of transition probabilities have to sum to 1. We can state the corresponding

probability transition matrix P_{X_2} which is given by

$$P_{X_2} = \begin{pmatrix} (1-p) & p & 0 & 0 \\ 0 & (1-p)^2 & 2p(1-p) & p^2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The pattern that emerges is that this is an upper triangle matrix. Within each row i, all entries (i,j) with $j < i$ equals 0. The diagonal element (i,i) is the starting point of writing down all elements of a binomial distribution for success p and failure $(1-p)$ followed by zeros until the end of the row is reached. This is only true for the first 2^{l-1} rows. The bottom half of the matrix is the identity matrix. In more general terms, we have the following pattern:

- $(i,j) = 0,$ if $j < i$,
- $(i,i) = (1-p)^i,$ if $i \leq 2^{l-1}$,
- $(i,i) = 1,$ if $i > 2^{l-1}$,
- $(i, i+k) = \binom{i}{k} p^k (1-p)^{i-k},$ if $0 < k < 2i$ and $i \leq 2^{l-1}$,
- $(i, 2i) = p^i,$ if $i \leq 2^{l-1}$,
- $(i,j) = 0,$ if $j > 2i$,
- $(i,j) = 0,$ if $j > i$ and $i > 2^{l-1}$.

A general probability transition matrix for X_l takes the form:

$$P_{X_l} = \left(\begin{array}{c|c|c|c|c|c|c|c} (1-p) & p & 0 & \dots & & & & 0 \\ 0 & (1-p)^2 & 2p(1-p) & p^2 & 0 & \dots & & 0 \\ \vdots & & \ddots & & & & & \vdots \\ 0 & \dots & 0 & (1-p)^{2^{l-1}} & \dots & \binom{2^{l-1}}{k} p^k (1-p)^{2^{l-1}-k} & \dots & p^{2^{l-1}} \\ 0 & \dots & & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & \ddots & & \vdots \\ \vdots & & & & & & 1 & 0 \\ 0 & \dots & & & & \dots & 0 & 1 \end{array}\right).$$

From the theory of Markov chains, we know that if we consider l steps along the Markov chain, the transition probabilities would be provided by the l-th power of P_{X_l}. Furthermore, the index (i,j) of that power yields the probability to transition from state i to state j in l steps. Since we are only interested in the transition probabilities from starting state 1, we can conclude that

$$\mathbb{P}[X_l = k] = (P_{X_l}^l)_{1k}. \tag{32}$$

This means that the probability to obtain k copies after l PCR cycles is given by the entry $(1,k)$ of the matrix obtained by raising P_{X_l} to the l-th power.

A.6 Probability to observe a molecule

$$
\begin{aligned}
p_s &:= \mathbb{P}[\text{"observation of single transcript"}] \\
&= \sum_{i=1}^{2^l} \mathbb{P}[\text{"}i\text{ copies produced by PCR"}] \\
&\qquad \cdot \mathbb{P}[\text{"at least one of the } i \text{ copies is sequenced"}] \\
&= \sum_{i=1}^{2^l} \mathbb{P}[X_l = i] \cdot (1 - \mathbb{P}[Y_i = 0]) \\
\text{Poisson distribution} \rightarrow \quad &= \sum_{i=1}^{2^l} \mathbb{P}[X_l = i] \cdot \left(1 - e^{-i\hat{p}_l}\right) \\
&= 1 - \mathbb{E}[e^{-\hat{p}_l X_l}] \\
&\approx 1 - e^{-\hat{p}_l \mathbb{E}[X_l]} - \frac{1}{2}\hat{p}_l^2 \cdot e^{-\hat{p}_l \mathbb{E}[X_l]} \cdot \mathrm{Var}(X_l) \qquad (33) \\
&= 1 - \left(1 + \frac{1}{2}\hat{p}_l^2 \mathrm{Var}(X_l)\right) \cdot e^{-\hat{p}_l \mathbb{E}[X_l]} \qquad (34)
\end{aligned}
$$

For Eq. (7) to hold, we need to prove the approximation in Eq. (33).

A.6.1 Proof of approximation

For a general infinitely differentiable function $f(X)$ of a random variable X, we can consider the Taylor expansion around the mean:

$$
\begin{aligned}
f(X) = f(\mathbb{E}[X]) &+ f'(\mathbb{E}[X])(X - \mathbb{E}[X]) + \frac{f''(\mathbb{E}[X])}{2!}(X - \mathbb{E}[X])^2 \\
&+ \frac{f'''(\mathbb{E}[X])}{3!}(X - \mathbb{E}[X])^3 + \dots
\end{aligned}
$$

$$
\overset{\mathbb{E}[.] \text{ of both sides}}{\Rightarrow} \qquad \mathbb{E}[f(X)] = f(\mathbb{E}[X]) + \frac{f''(\mathbb{E}[X])}{2}\mathrm{Var}(X) + \sum_{k=3}^{\infty} \frac{f^k(\mathbb{E}[X])}{k!}\mathbb{E}[(X - \mathbb{E}[X])^k].
$$

We now choose $f(X_l) = e^{-\hat{p}_l X_l}$ and note

$$
f''(X_l) = \hat{p}_l^2 e^{-\hat{p}_l X_l}, \qquad (35)
$$

$$
f^k(X_l) = (-1)^k \hat{p}_l^k e^{-\hat{p}_l X_l}. \qquad (36)
$$

Hence,

$$
\begin{aligned}
\mathbb{E}[e^{-\hat{p}_l X_l}] &= e^{-\hat{p}_l \mathbb{E}[X_l]} + \frac{1}{2}\hat{p}_l^2 e^{-\hat{p}_l \mathbb{E}[X_l]}\mathrm{Var}(X_l) + \sum_{k=3}^{\infty} \frac{(-1)^k}{k!}\hat{p}_l^k e^{-\hat{p}_l \mathbb{E}[X_l]}\mathbb{E}[(X_l - \mathbb{E}[X_l])^k] \\
&= e^{-\hat{p}_l \mathbb{E}[X_l]} \left(1 + \frac{1}{2}\mathrm{Var}(\hat{p}_l X_l) + \sum_{k=3}^{\infty} \frac{(-1)^k}{k!}\mathbb{E}[(\hat{p}_l X_l - \mathbb{E}[\hat{p}_l X_l])^k]\right).
\end{aligned}
$$

The elements of the sum are declining rapidly if $\hat{p}_l$ is small. Since the series is moreover alternating, it is reasonable to ignore terms for $k \geq 3$.

A.7 Estimating parameter values for $\hat{p}_{25}$ and p_s

We want to roughly estimate $\hat{p}_l$ from experimental parameters. We have

$$
\begin{aligned}
\hat{p}_l &= \frac{R}{K\mathbb{E}[X_l]} \\
&= \frac{\#\text{ reads}}{(\#\text{ cells}) \cdot (\#\text{ transcripts per cell}) \cdot (\text{average }\#\text{ copies produced by PCR})} \qquad (37)
\end{aligned}
$$

where # reads is the total number of observed reads during sequencing which we previously denoted by R, K is the average total number of original transcripts in the cell population and $\mathbb{E}[X_l]$ is the average number of PCR copies

produced per transcript. We also stated that Y_{j_k} signifies the number of PCR copies actually sequenced for a specific transcript $k \in \{1, \dots, K\}$. Then, we can conclude

$$\#\text{ reads} = R = \sum_{k=1}^{K} Y_{j_k}. \tag{38}$$

The number of reads depends on the number of PCR cycles l.

It is estimated that a single cell contains about 1 million mRNA transcripts. In an experiment involving HeLa data [2], we perform around $l = 25$ PCR cycles at an estimated $p = 0.95$ PCR efficiency. The sequencing run for the HeLa data yields roughly $2 \cdot 10^8$ mapped reads for 1624 cells (before filtering). We therefore estimate the sequencing efficiency $\hat{p}_{25}$ to be

$$\begin{aligned}
\hat{p}_{25} &\approx \frac{2 \cdot 10^8}{1624 \cdot (10^6) \cdot (1 + 0.95)^{25}} \\
&\approx \frac{2 \cdot 10^8}{1624 \cdot (10^6) \cdot (1.8 \cdot 10^7)} \\
&\approx 6.8 \cdot 10^{-9}.
\end{aligned}$$

For this particular data set, we can then also estimate the probability p_s to detect a specific transcript with Eq. (11):

$$\begin{aligned}
p_s &= 1 - \mathrm{e}^{-\frac{R}{K}} \\
&= 1 - \mathrm{e}^{-\frac{2 \cdot 10^8}{1624 \cdot 10^6}} \\
&\approx 0.116.
\end{aligned}$$

A.8 Moments of Output Distribution

In order to characterize the output distribution in Eq. (12), we investigate the first and second moments as well as the CV.

A.8.1 Expectation

The expectation is calculated as

$$\begin{aligned}
\mathbb{E}[X_{\text{out}}] &= \sum_{k=0}^{\infty} k \cdot \mathbb{P}[X_{\text{out}} = k] \\
\text{Eq. (12)} \rightarrow \quad &= \sum_{k=0}^{\infty} \sum_{i=k}^{\infty} k \mathbb{P}[X_{\text{in}} = i] \binom{i}{k} p_s^k (1 - p_s)^{i-k} \\
\text{swap order of summation} \rightarrow \quad &= \sum_{i=0}^{\infty} \mathbb{P}[X_{\text{in}} = i] \sum_{k=0}^{i} k \binom{i}{k} p_s^k (1 - p_s)^{i-k} \\
\text{expectation of a binomial distribution} \rightarrow \quad &= \sum_{i=0}^{\infty} \mathbb{P}[X_{\text{in}} = i] \cdot i \cdot p_s \qquad (39) \\
&= p_s \cdot \mathbb{E}[X_{\text{in}}]. \qquad (40)
\end{aligned}$$

A.8.2 Variance

We first consider the second moment

$$\begin{aligned}
\mathbb{E}[X_{\text{out}}^2] &= \sum_{k=0}^{\infty} k^2 \cdot \mathbb{P}[X_{\text{out}} = k] \\
\text{Eq. (12)} \rightarrow \quad &= \sum_{k=0}^{\infty} \sum_{i=k}^{\infty} k^2 \mathbb{P}[X_{\text{in}} = i] \cdot \binom{i}{k} p_s^k (1 - p_s)^{i-k} \\
\text{swap order of summation} \rightarrow \quad &= \sum_{i=0}^{\infty} \mathbb{P}[X_{\text{in}} = i] \sum_{k=0}^{i} k^2 \cdot \binom{i}{k} p_s^k (1 - p_s)^{i-k} \\
\text{second moment of a binomial distribution} \rightarrow \quad &= \sum_{i=0}^{\infty} \mathbb{P}[X_{\text{in}} = i] \left(i \cdot p_s \cdot (1 - p_s) + i^2 \cdot p_s^2 \right) \qquad (41) \\
&= p_s \cdot (1 - p_s) \cdot \mathbb{E}[X_{\text{in}}] + p_s^2 \cdot \mathbb{E}[X_{\text{in}}^2]. \qquad (42)
\end{aligned}$$

Now, the variance is given by

$$\begin{aligned}\mathrm{Var}(X_{\mathrm{out}}) &= \mathbb{E}[X_{\mathrm{out}}^2] - \mathbb{E}[X_{\mathrm{out}}]^2 \\ \text{Eqs. (40), (42)} \rightarrow \quad &= p_s \cdot (1-p_s) \cdot \mathbb{E}[X_{\mathrm{in}}] + p_s^2 \cdot \mathrm{Var}(X_{\mathrm{in}}). \end{aligned} \tag{43}$$

A.8.3 Coefficient of Variation

The coefficient of variation is obtained by considering

$$\begin{aligned}\mathrm{CV}(X_{\mathrm{out}}) &= \frac{\sqrt{\mathrm{Var}(X_{\mathrm{out}})}}{\mathbb{E}[X_{\mathrm{out}}]} \\ &= \sqrt{\frac{p_s \cdot (1-p_s) \cdot \mathbb{E}[X_{\mathrm{in}}] + p_s^2 \cdot \mathrm{Var}(X_{\mathrm{in}})}{p_s^2 \cdot \mathbb{E}[X_{\mathrm{in}}]^2}} \\ &= \sqrt{\frac{1-p_s}{p_s \cdot \mathbb{E}[X_{\mathrm{in}}]} + \mathrm{CV}(X_{\mathrm{in}})^2}. \end{aligned} \tag{44}$$

A.8.4 General moments

We can generalize the previous considerations and write down an equation for the n-th moment of X_{out}. Then,

$$\begin{aligned}\mathbb{E}[X_{\mathrm{out}}^n] &= \sum_{k=0}^{\infty} k^n \cdot \mathbb{P}[X_{\mathrm{out}} = k] \\ \text{Eq. (12)} \rightarrow \quad &= \sum_{k=0}^{\infty} \sum_{i=k}^{\infty} k^n \mathbb{P}[X_{\mathrm{in}} = i] \cdot \binom{i}{k} p_s^k (1-p_s)^{i-k} \\ \text{swap order of summation} \rightarrow \quad &= \sum_{i=0}^{\infty} \mathbb{P}[X_{\mathrm{in}} = i] \sum_{k=0}^{i} k^n \cdot \binom{i}{k} p_s^k (1-p_s)^{i-k} \end{aligned} \tag{45}$$

The second sum corresponds to the n-th moment of a binomial distribution $\mathrm{Binom}(i, p_s)$ as given by [1]:

$$\mathbb{E}[k^n](i) = \sum_{k=0}^{n} \left\{ {n \atop k} \right\} i^{\underline{k}} \, p_s^k, \tag{46}$$

where $\left\{ {n \atop k} \right\}$ is the Stirling number of the second kind given by

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^n, \tag{47}$$

and $i^{\underline{k}}$ is the k-th falling power (also called falling factorial) of i which is calculated as

$$i^{\underline{k}} = i \cdot (i-1) \cdot \ldots \cdot (i-k+1). \tag{48}$$

Then from Eq. (45), we conclude

$$\mathbb{E}[X_{\mathrm{out}}^n] = \sum_{i=0}^{\infty} \sum_{k=0}^{n} \mathbb{P}[X_{\mathrm{in}} = i] \cdot \left\{ {n \atop k} \right\} i^{\underline{k}} \, p_s^k \tag{49}$$

$$\text{swap order of summation} \rightarrow \quad = \sum_{k=0}^{n} \left\{ {n \atop k} \right\} p_s^k \sum_{i=0}^{\infty} \mathbb{P}[X_{\mathrm{in}} = i] \cdot i^{\underline{k}} \tag{50}$$

$$= \sum_{k=0}^{n} \left\{ {n \atop k} \right\} p_s^k \; \mathbb{E}[X_{\mathrm{in}}^{\underline{k}}]. \tag{51}$$

This provides a straightforward equation for calculating the n-th moment of the output distribution which depends only on the first n moments of the input distribution. We can check our previous calculations from Eqs. (40), (42):

$$\begin{aligned}\mathbb{E}[X_{\mathrm{out}}^1] &= \left\{ {1 \atop 0} \right\} + \left\{ {1 \atop 1} \right\} p_s^1 \; \mathbb{E}[X_{\mathrm{in}}^{\underline{1}}] \\ &= p_s \; \mathbb{E}[X_{\mathrm{in}}], \end{aligned} \tag{52}$$

and

$$\begin{aligned}\mathbb{E}[X_{\text{out}}^2] &= \left\{ {2 \atop 0} \right\} + \left\{ {2 \atop 1} \right\} p_s^1 \, \mathbb{E}[X_{\text{in}}^{\underline{1}}] + \left\{ {2 \atop 2} \right\} p_s^2 \, \mathbb{E}[X_{\text{in}}^{\underline{2}}] \\ &= p_s \, \mathbb{E}[X_{\text{in}}] + p_s^2 \mathbb{E}[X_{\text{in}}^2 - X_{\text{in}}] \\ &= (p_s - p_s^2) \, \mathbb{E}[X_{\text{in}}] + p_s^2 \mathbb{E}[X_{\text{in}}^2]. \end{aligned} \tag{53}$$

The third moment is obtained by considering

$$\begin{aligned}\mathbb{E}[X_{\text{out}}^3] &= \left\{ {3 \atop 0} \right\} + \left\{ {3 \atop 1} \right\} p_s^1 \, \mathbb{E}[X_{\text{in}}^{\underline{1}}] + \left\{ {3 \atop 2} \right\} p_s^2 \, \mathbb{E}[X_{\text{in}}^{\underline{2}}] + \left\{ {3 \atop 3} \right\} p_s^3 \, \mathbb{E}[X_{\text{in}}^{\underline{3}}] \\ &= p_s \, \mathbb{E}[X_{\text{in}}] + 3p_s^2 \mathbb{E}[X_{\text{in}}^2 - X_{\text{in}}] + p_s^3 \mathbb{E}[X_{\text{in}}^3 - 3X_{\text{in}}^2 + 2X_{\text{in}}] \\ &= p_s^3 \mathbb{E}[X_{\text{in}}^3] + 3p_s^2(1 - p_s)\mathbb{E}[X_{\text{in}}^2] + p_s(2p_s^2 - 3p_s + 1)\mathbb{E}[X_{\text{in}}]. \end{aligned} \tag{54}$$

We note that the highest order moment of X_{in} required to calculate the n-th moment of X_{out} is also the n-th moment. This enables an iterative strategy for calculating the moments for X_{in} given the moments of X_{out}. If all moments of X_{out} are known, the distribution of X_{in} including its distribution type is fully defined.

Eq. (51) can be rearranged to obtain an expression for $\mathbb{E}[X_{\text{in}}^n]$ that only depends on the first n moments of X_{out} and the first $n-1$ moments of X_{in}:

$$\begin{aligned}\mathbb{E}[X_{\text{out}}^n] &= \sum_{k=0}^{n} \left\{ {n \atop k} \right\} p_s^k \, \mathbb{E}[X_{\text{in}}^{\underline{k}}] \\ &= p_s^n \cdot \mathbb{E}[X_{\text{in}}^{\underline{n}}] + \sum_{k=0}^{n-1} \left\{ {n \atop k} \right\} p_s^k \, \mathbb{E}[X_{\text{in}}^{\underline{k}}] \\ &= p_s^n \cdot \mathbb{E}[X_{\text{in}}^n] + p_s^n \cdot \mathbb{E}[\underbrace{X_{\text{in}}^{\underline{n}} - X_{\text{in}}^n}_{\substack{\text{highest order} \\ \text{term is } n-1}}] + \sum_{k=0}^{n-1} \left\{ {n \atop k} \right\} p_s^k \, \mathbb{E}[X_{\text{in}}^{\underline{k}}] \end{aligned}$$

$$\Leftrightarrow \qquad \mathbb{E}[X_{\text{in}}^n] = \frac{1}{p_s^n} \mathbb{E}[X_{\text{out}}^n] - \mathbb{E}[X_{\text{in}}^{\underline{n}} - X_{\text{in}}^n] - \sum_{k=0}^{n-1} \left\{ {n \atop k} \right\} p_s^{k-n} \, \mathbb{E}[X_{\text{in}}^{\underline{k}}]. \tag{55}$$

All terms on the right hand-side can be iteratively expressed by moments of the output distribution. If all moments of X_{out} exist, so do the moments of X_{in} which means that the probability distribution of X_{out} completely defines the probability distribution of X_{in}.

A.9 Relations of the parameters of a log-normal distribution required to satisfy the mean-variance relation of experimental scRNA-seq data

In Figure 3A, there is a clear dependency of the variance on the expectation of the output data of a scRNA-seq experiment. We have fitted them with a Poisson and a negative binomial distribution. We now utilize the output distribution we have derived and state requirements on the parameters of the log-normal distributions which is the distribution type typically chosen to model input distributions.

From Eqs. (40), (43), we have

$$\text{Var}(X_{\text{out}}) = (1 - p_s) \cdot \mathbb{E}[X_{\text{out}}] + p_s^2 \cdot \text{Var}(X_{\text{in}}).$$

Let X_{in} satisfy the following relation:

$$\text{Var}(X_{\text{in}}) = a_1 \mathbb{E}[X_{\text{in}}] + a_2 \mathbb{E}[X_{\text{in}}]^2 + a_3 \mathbb{E}[X_{\text{in}}]^3, \tag{56}$$

where a_1, a_2 and a_3 are parameters. Hence, we assume a polynomial relation between the expectation of the input distribution and its variance. This yields

$$\text{Var}(X_{\text{out}}) = (1 - p_s + p_s a_1) \cdot \mathbb{E}[X_{\text{out}}] + a_2 \mathbb{E}[X_{\text{out}}]^2 + \frac{a_3}{p_s} \mathbb{E}[X_{\text{out}}]^3. \tag{57}$$

We can fit the parameters and obtain the graph shown in Figure 3B. We find all three parameters to be significant, meaning all three orders of the polynomial are essential for the fit.

If we assume $X_{\text{in}} \sim \mathcal{LN}(\mu, \sigma^2)$, then we can establish a relation between the two parameters μ and σ^2 of the input distribution with the help of Eq. (56). For a log-normally distributed variable Z, we know

$$\mathbb{E}[Z] = \mathrm{e}^{\mu + \frac{\sigma^2}{2}}, \tag{58}$$

$$\text{Var}(Z) = \mathrm{e}^{2\mu + \sigma^2} \left(\mathrm{e}^{\sigma^2} - 1 \right). \tag{59}$$

Then, we conclude from Eq. (56) that

$$\mathrm{e}^{2\mu+\sigma^2}\left(\mathrm{e}^{\sigma^2}-1\right)=a_1\mathrm{e}^{\mu+\frac{\sigma^2}{2}}+a_2\mathrm{e}^{2\mu+\sigma^2}+a_3\mathrm{e}^{3\mu+\frac{3\sigma^2}{2}} \tag{60}$$

$$\Leftrightarrow \quad 0=\mathrm{e}^{2\mu+\sigma^2}-\frac{\mathrm{e}^{\sigma^2}-a_2-1}{a_3}\mathrm{e}^{\mu+\frac{\sigma^2}{2}}+\frac{a_1}{a_3} \tag{61}$$

$$\Leftrightarrow \quad \mathrm{e}^{\mu+\frac{\sigma^2}{2}}=\frac{\mathrm{e}^{\sigma^2}-a_2-1\pm\sqrt{\left(\mathrm{e}^{\sigma^2}-a_2-1\right)^2-4a_1a_3}}{2a_3} \tag{62}$$

$$\Leftrightarrow \quad \mu=\ln\left(\mathrm{e}^{\sigma^2}-a_2-1\pm\sqrt{\left(\mathrm{e}^{\sigma^2}-a_2-1\right)^2-4a_1a_3}\right)-\ln\left(2a_3\right)-\frac{\sigma^2}{2}. \tag{63}$$

Thus, the log-normal distributions which we assume to generate the output distributions shown in Figure 3A and Figure 3B should satisfy Eq. (63) with the parameters a_1, a_2 and a_3 fitted in Figure 3B.

B Supplementary figures

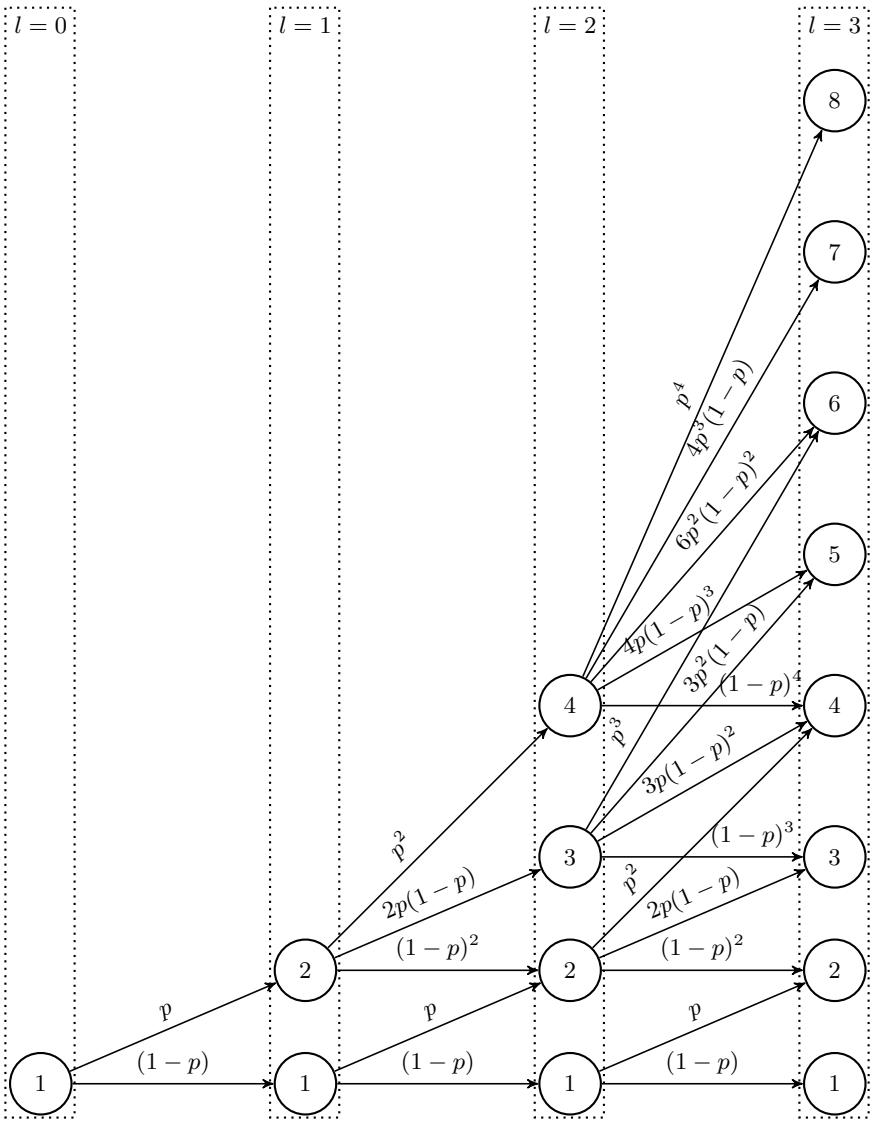


Fig. S1. Probability transitions for the first three cycles of PCR. Each node represents the number of PCR copies produced after l cycles where l is represented by the columns.

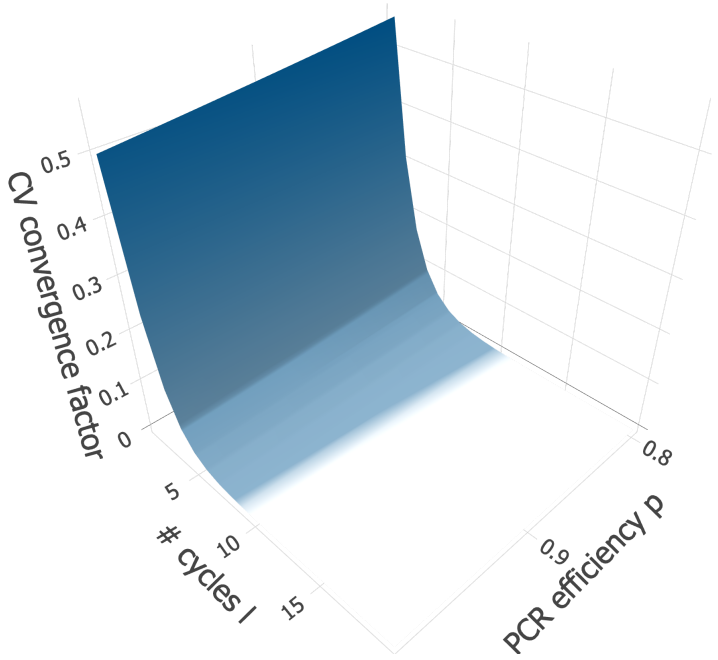


Fig. S2. The CV convergence factor converges fast, reaching the convergence threshold at around 7 and 8 cycles. The CV convergence factor is calculated in dependence on parameters l and p.

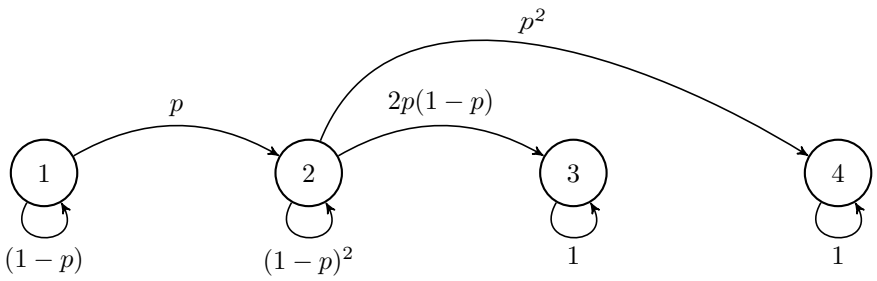


Fig. S3. The PCR distribution as a Markov chain.

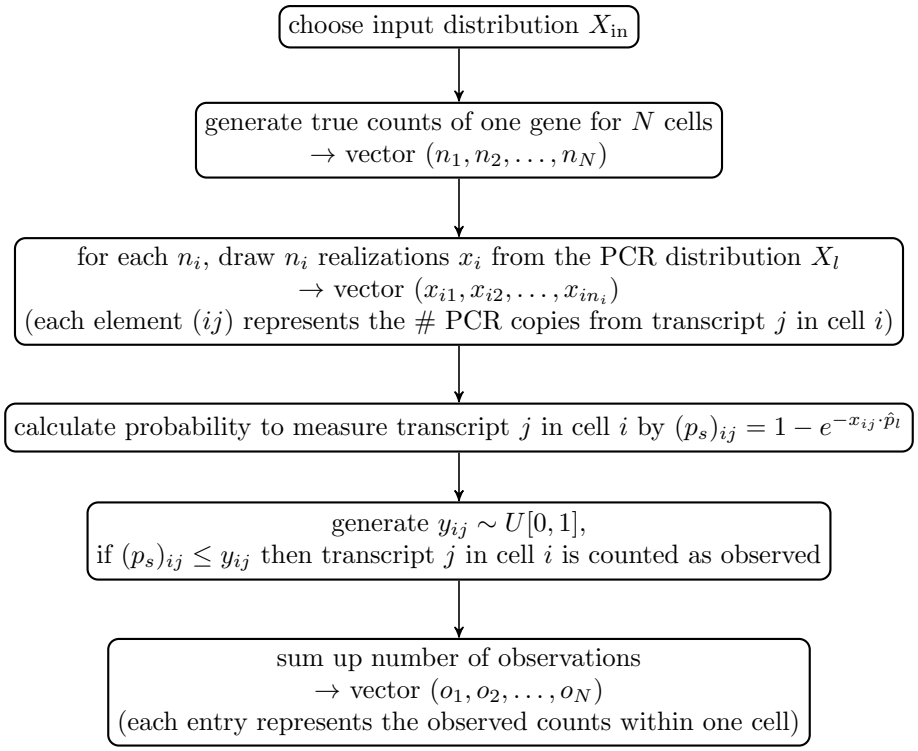


Fig. S4. Principle steps of a numerical simulation of a simplified scRNA-seq experiment to obtain observed counts given an input distribution.

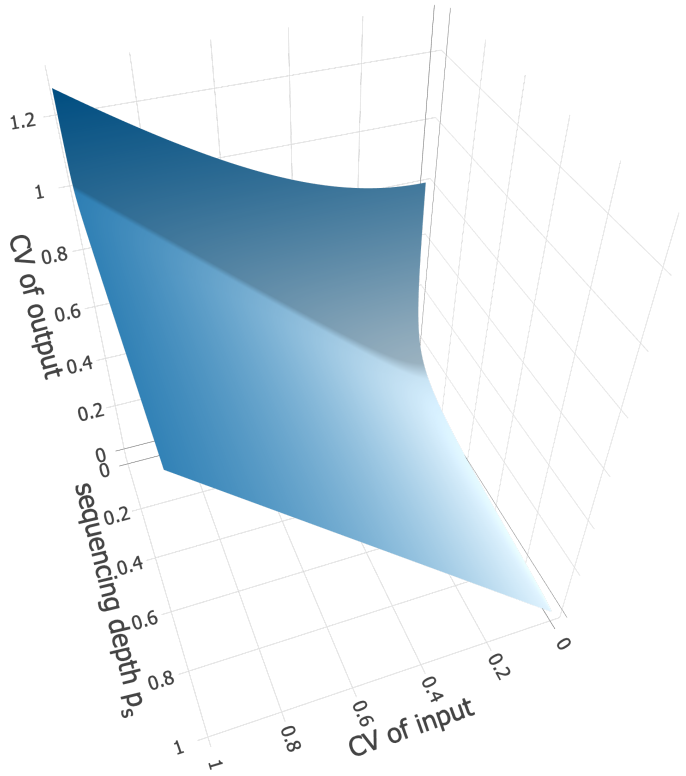


Fig. S5. Effects of changes in input CV and success probability p_s on the output CV in three dimensions.

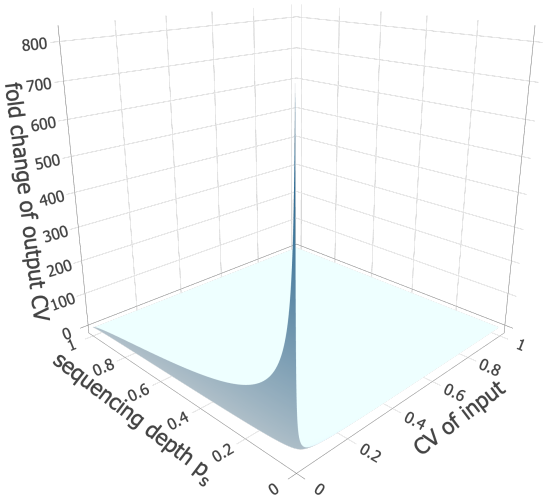


Fig. S6. Effects of changes in input CV and success probability p_s on the fold change of the output CV compared to input CV in three dimensions.

References

[1] Knoblauch A (2008). Closed-Form Expressions for the Moments of the Binomial Probability Distribution. SIAM Journal on Applied Mathematics, 69(1), 197–204.

[2] Schwabe D, Formichetti S, Junker JP, Falcke M, Rajewsky N (2020). The transcriptome dynamics of single cells during the cell cycle. Mol Syst Biol., Nov;16(11):e9946. doi: 10.15252/msb.20209946. PMID: 33205894.