

SOFTWARE

Open Access



BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells

Mariana Gómez-Schiavon^{1,2}, Liang-Fu Chen³, Anne E. West^{3*} and Nicolas E. Buchler^{4,5,6*} 

Abstract

Single-molecule RNA fluorescence in situ hybridization (smFISH) provides unparalleled resolution in the measurement of the abundance and localization of nascent and mature RNA transcripts in fixed, single cells. We developed a computational pipeline (BayFish) to infer the kinetic parameters of gene expression from smFISH data at multiple time points after gene induction. Given an underlying model of gene expression, BayFish uses a Monte Carlo method to estimate the Bayesian posterior probability of the model parameters and quantify the parameter uncertainty given the observed smFISH data. We tested BayFish on synthetic data and smFISH measurements of the neuronal activity-inducible gene *Npas4* in primary neurons.

Keywords: Gene expression, Stochastic process, Chemical master equation, Likelihood methods, Monte Carlo sampling, Bayesian posterior probability

Background

Cell-to-cell variation in gene expression across an isogenic population is a fact of life. The initiation of transcription involves a series of stochastic biochemical events, which includes chromatin accessibility, the binding of transcription factors, and the assembly of RNA polymerase at the promoter of a gene [1]. Distinct promoter states can often arise when one of these biochemical events is rate-limiting. The existence of multiple promoter states with different expression rates can generate transcriptional bursting, which are episodes of transcriptional activity followed by long periods of inactivity [2–4]. This phenomenon has been observed in bacteria [5, 6], yeast [7, 8], flies [9, 10], and mammals [11–16].

Cell-to-cell variability in gene expression is often studied using techniques that measure transcription in single

cells. One such technique, single-molecule RNA fluorescence in situ hybridization (smFISH) measures the abundance and localization of individual transcripts in a single cell [17, 18]. This method uses a cocktail of fluorescently labelled DNA oligos complementary to the target RNA and works in many organisms; see [19]. Each individual transcript is bound by fluorescent DNA probes and appears as a bright, diffraction-limited spot in a fluorescence microscope. When there are multiple transcripts (e.g., active transcription sites, TSs, at gene loci), the measured intensity can be significantly brighter. The smFISH technique is simple and has been rapidly adopted by other labs to address cell-to-cell variability in gene expression. This has been helped in part by software packages [20, 21] that facilitate image segmentation and spot analysis.

Gene expression is dynamic and the properties of transcriptional bursting must be inferred from smFISH data, which are static snapshots or distributions of mRNA and active TSs per cell sampled from a population. This inference is done using mathematical models of stochastic gene expression, whose predicted distributions of transcripts and active TSs in a population of cells are

*Correspondence: west@neuro.duke.edu; nicolas.buchler@duke.edu

³Department of Neurobiology, Duke University, Durham, NC, USA

⁴Department of Biology, Duke University, Durham, NC, USA

Full list of author information is available at the end of the article

then fitted to smFISH data to infer the model parameters and likely properties of transcriptional bursting [22]. The simplest model that generates such bursting is the two-state model, which presumes that a gene stochastically switches between two promoter states, a transcriptionally active state and an inactive state. The advantage of the two-state model is that the distributions have been solved analytically and model parameters are inferred by fitting moments (e.g., mean and variance) or the full distributions to the observed smFISH data using least-squares approaches. Despite the success of two-state models, more complicated models are often needed to explain the observed distributions properly [7, 15, 16, 23]. These complex models often do not have analytical solutions and one must resort to simplifying assumptions or computationally intensive numerical methods to calculate distributions [24]. This is especially true for genes that are not in a steady state, e.g., induced genes.

We used smFISH to measure transcripts of the neuronal activity-inducible gene *Npas4* in primary neurons after membrane depolarization with elevated extracellular potassium (Fig. 1). Our *Npas4* smFISH measurements showed a surprising amount of cell-to-cell variation in

both transcript levels and active TSs even when all neurons were exposed to a uniform external stimulus. Given prior studies of cell-to-cell variability in gene expression in other systems, this variability in the transcriptional response of activity-inducible genes is likely to arise from the probabilistic activation of transcriptional bursting at single alleles. We reasoned that we could use this single-cell transcriptional variability to build a model of activity-inducible *Npas4* induction that would inform our quantitative understanding of the transcriptional processes that drive dynamic changes in *Npas4* expression following neuronal activation.

Thus, we developed a computational pipeline (BayFish) that uses a Bayesian approach to infer the best model parameters from smFISH data and to quantify the uncertainty in those parameters rigorously. The user specifies any mathematical model of stochastic gene expression with an unknown set of parameters (θ) and provides smFISH data (Y) at different time points before and after induction. BayFish then uses a Monte Carlo method to estimate the Bayesian posterior probability $P(\theta|Y)$ of the model parameters, which elucidates the best-fitting parameters and quantifies their uncertainty given the current smFISH data. We first tested BayFish on synthetic

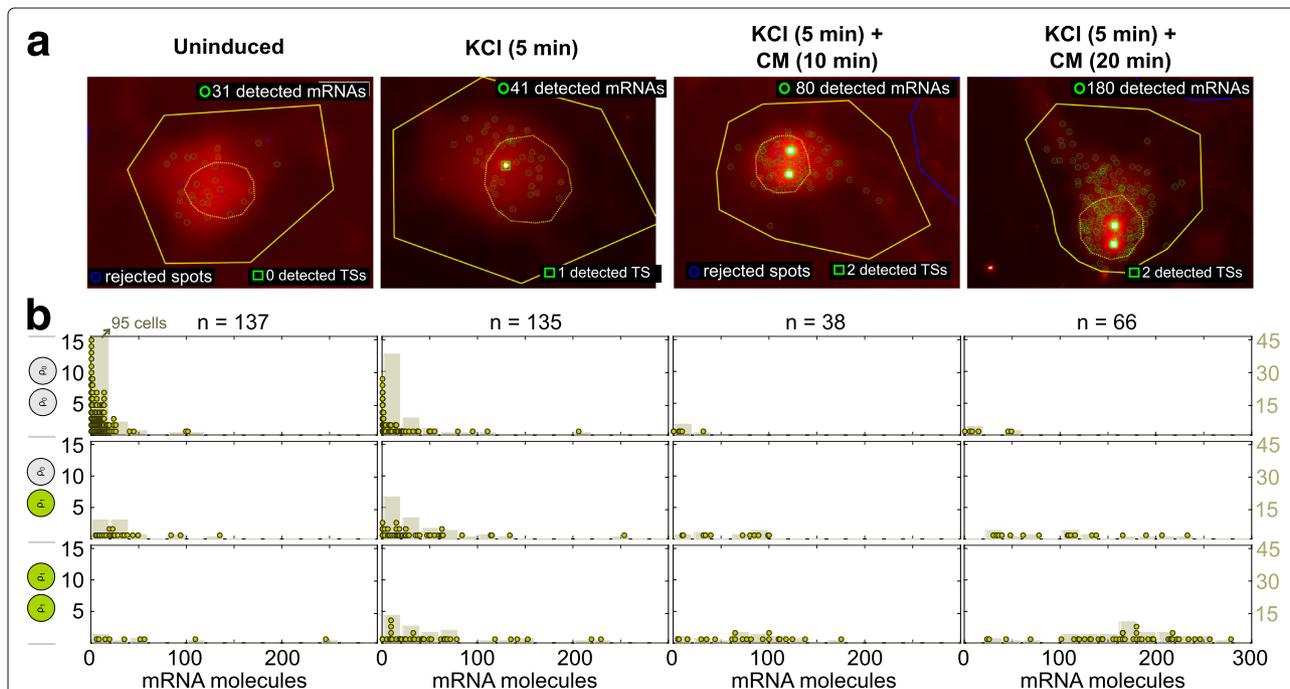


Fig. 1 Single-molecule fluorescence in situ hybridization data of *Npas4* mRNA in primary neurons after membrane depolarization. Measurements are shown before the stimulus (uninduced), 5 min after KCl exposure, and an additional 10 and 20 min later after cells were returned to the conditioned medium (CM); see “Methods”. **a** Example of an image-processed cell at each time point. We show detected mRNAs (green circles) and active transcription sites (TSs; green squares) within the cell contour (yellow line) and nucleus (dashed yellow line). Neurons are post-mitotic and, thus, we observed up to two active gene loci per diploid cell. **b** For each condition, a histogram of the number of mRNA molecules binned by the number of TSs (dots, left y-axis). Smoothed histogram with bins of 20 mRNAs (bars, right y-axis). The total number of cells (n) per time sample is listed at the top. CM conditioned medium, TS transcription site

data and demonstrate how to select the best model from multiple mathematical models by combining information criteria with the likelihood and Bayesian posterior calculated by BayFish. We then used BayFish on the *Npas4* smFISH data to infer the parameters of an underlying two-state model of gene expression that were likely affected by the stimulus. Our results show that a two-state promoter model can recapitulate *Npas4* dynamics after induction and we further inferred that the transition rate from the promoter off state to the on state is increased by the stimulus.

There is currently no software that allows a user to specify any model of stochastic gene expression, evaluate the time evolution of mRNA and active TS distributions after induction, and rigorously infer parameters and confidence intervals from smFISH data using the Bayesian posterior probability. We expect BayFish to fill an important gap that will facilitate the adoption of the smFISH technique by other laboratories that wish to address cell-to-cell variability in gene expression.

Results

BayFish is a software package that combines numerical methods with a Monte Carlo method to estimate the Bayesian posterior probability $P(\theta|Y)$ of model parameters (θ) given the observed smFISH data (Y) at different time points before and after induction. Bayes theorem states that $P(\theta|Y) = P(Y|\theta)P(\theta)/P(Y)$ where $P(Y|\theta)$ is the likelihood \mathcal{L} of the data given the parameters. $P(\theta)$ and $P(Y)$ are the prior probability distributions of the parameters and the data, respectively. Each iteration of the Monte Carlo method uses several numerical subroutines to calculate the time evolution of the mRNA and active TS distributions given a set of model parameters (θ), to evaluate the likelihood that the smFISH data (Y) were sampled from this distribution or $\mathcal{L} = P(Y|\theta)$, and to calculate the Bayesian posterior probability $\mathcal{P} = P(\theta|Y)$ given the likelihood and priors. The global program is based on the Metropolis random walk algorithm [25, 26]:

1. Specify a mathematical model of stochastic gene expression that has an unknown set of parameters θ .
2. Choose an initial θ and calculate the corresponding likelihood $\mathcal{L} = P(Y|\theta)$ and Bayesian posterior probability $\mathcal{P} = \mathcal{L}P(\theta)/P(Y)$ using several numerical subroutines.
3. Iterate over $t = \{1, 2, \dots, T\}$ as follows:
 - (a) Draw a random proposal $\phi \sim \theta_t + \mathcal{N}(0, \Sigma)$, where $\mathcal{N}(0, \Sigma)$ is a multivariate normal distribution with the same dimension as θ and with zero mean. Σ is the covariance matrix.

- (b) Evaluate the likelihood of the proposal $\mathcal{L}_\phi = P(Y|\phi)$ using several numerical subroutines.
- (c) Calculate the Bayesian posterior probability $\mathcal{P}_\phi = \mathcal{L}_\phi P(\phi)/P(Y)$.
- (d) Update parameters $\theta_{t+1} \leftarrow \phi$ and $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_\phi$ with probability $\min(\mathcal{P}_\phi/\mathcal{P}_t, 1)$; otherwise, $\theta_{t+1} \leftarrow \theta_t$ and $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_t$.

Over time, the algorithm will generate a Markov chain of θ_t whose distribution converges to the Bayesian posterior probability $P(\theta|Y)$. BayFish saves the likelihood \mathcal{L}_t and θ_t of each step. After discarding the early part of the chain (the burn-in phase), the remaining θ_t values were used to estimate the Bayesian posterior probability $P(\theta|Y)$; see “Methods.”

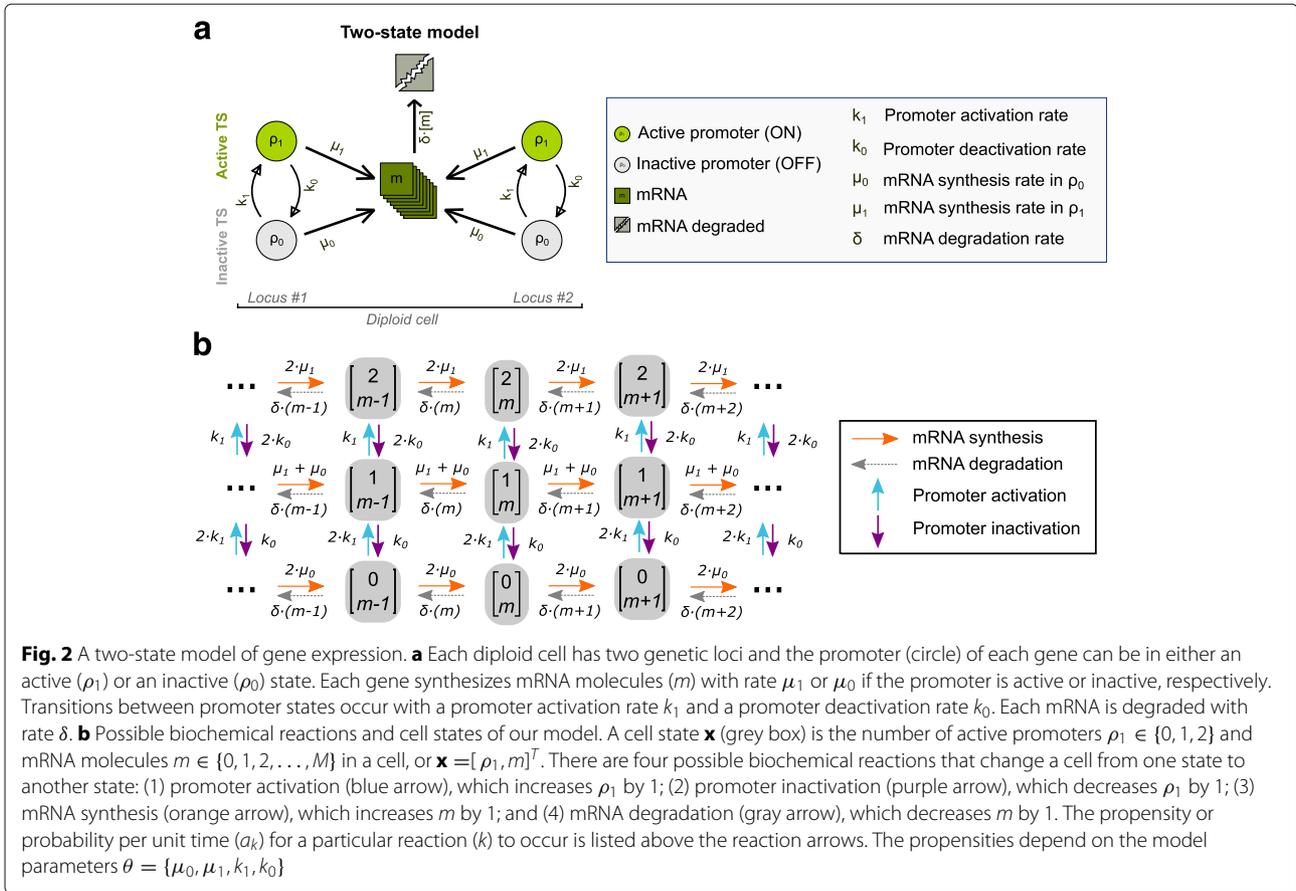
Mathematical model of stochastic gene expression

We considered a two-state model of gene expression (Fig. 2), where each promoter can be in an inactive off state ρ_0 with a basal transcription level (synthesis rate μ_0) or an active on state ρ_1 with a higher transcription level (synthesis rate μ_1). Transitions between promoter states occur with a promoter activation rate k_1 and a promoter deactivation rate k_0 . We chose a two-state model because it is the simplest model that can generate transcriptional bursting, a feature observed in our *Npas4* smFISH data (Fig. 1). Neurons are post-mitotic and, thus, our model does not include duplicated alleles (e.g., three or four active loci) that arise after DNA replication. Each promoter allele was assumed to be regulated independently, as shown previously [11, 15, 23, 27]. The two-state model parameter set, which determines the dynamics of mRNA and active promoters, is $\theta = \{\mu_0, \mu_1, k_1, k_0\}$. We fixed the mRNA degradation rate δ because it is a known quantity, but this parameter could be a free parameter in other models.

Our smFISH experiments measured gene expression both before and after the stimulus. We presumed that gene expression before the stimulus was at a steady state determined by one set of model parameters (θ^U , unstimulated parameter set). Upon induction, the stimulus changed one or more of the model parameters (θ^S , stimulated parameter set). Thus, the mRNA and active TS distribution will evolve towards a new steady state in response to the changed parameters. Below, we describe how we calculated the stationary mRNA and active TS distribution before the stimulus using θ^U and how we then calculated the time evolution of the distribution after the stimulus using θ^S .

Time evolution of the probability distribution

The chemical master equation (CME) is an infinite set of coupled differential equations that describe the dynamics



of the probability of the biochemical system being in a particular state \mathbf{x} at time t , $P(\mathbf{x}, t)$ [28, 29]. The probability flow into and out of each state \mathbf{x} is given by:

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} = \sum_k [a_k(\mathbf{x} - \mathbf{v}_k)P(\mathbf{x} - \mathbf{v}_k, t) - a_k(\mathbf{x})P(\mathbf{x}, t)]. \tag{1}$$

The summation is over all possible biochemical reactions k into and out of state \mathbf{x} :

$$\mathbf{x} \xrightarrow{a_k(\mathbf{x})} \mathbf{x} + \mathbf{v}_k \tag{2}$$

where $a_k(\mathbf{x}) \partial t$ is the probability that the biochemical reaction k will occur within the infinitesimal time interval ∂t given that the system is in state \mathbf{x} . The model parameters θ affect the propensities of different biochemical reactions (Fig. 2), and the stoichiometric vector (\mathbf{v}_k) of reaction k describes how the system state changes when the reaction k occurs. More generally, the CME is written in matrix form:

$$\frac{\partial \mathbf{P}(\mathbf{X}, t)}{\partial t} = \mathbf{A}(\theta) \cdot \mathbf{P}(\mathbf{X}, t) \tag{3}$$

where all possible cell states \mathbf{X} are enumerated as a vector $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^T$. $\mathbf{P}(\mathbf{X}, t)$ is the probability density state

vector $[P(\mathbf{x}_1, t), P(\mathbf{x}_2, t), \dots, P(\mathbf{x}_N, t)]^T$ of possible states organized identically to \mathbf{X} . The state reaction matrix $\mathbf{A}(\theta)$ has elements:

$$\mathbf{A}_{ij} = \begin{cases} -\sum_k a_k(\mathbf{x}_i), & \forall i = j, \\ a_k(\mathbf{x}_i), & \forall j \text{ such that } \mathbf{x}_j = \mathbf{x}_i + \mathbf{v}_k, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Pre-stimulus stationary distribution

We assumed that the pre-stimulus mRNA and active TS distribution $\mathbf{P}^*(\mathbf{X})$ is time-independent and stationary. We calculated the stationary distribution by setting Eq. 3 to zero and determined the nonzero eigenvector $\mathbf{V} \geq \mathbf{0}$ in the kernel of $\mathbf{A}(\theta^U)$ using the Arnoldi iteration algorithm [30] (eigs MATLAB function, or eig_gen Armadillo C++ library). Each element of \mathbf{P}^* is given by

$$P^*(\mathbf{x}_i) = \frac{V_i}{\sum_j V_j} \tag{5}$$

where V_i is the i th element in the vector $\mathbf{V} = [V_1, V_2, \dots, V_N]^T$ and $\sum_i P^*(\mathbf{x}_i) = 1$.

Post-stimulus distribution dynamics

Given an initial distribution $\mathbf{P}^*(\mathbf{X})$ at time zero and post-stimulus state reaction matrix $\mathbf{A}(\theta^S)$, the post-stimulus distribution $\mathbf{P}(\mathbf{X}, \tau)$ at time τ after stimulus is:

$$P(X, \tau) = \exp[A(\theta^S) \tau] P^*(X). \quad (6)$$

We calculated $P(X, \tau)$ after induction using the same MATLAB routines from the finite state projection method [24], or the equivalent functions in the Armadillo C++ library. We used finite state projection to verify that our estimated probability distributions were below the error threshold ($\epsilon \leq 10^{-12}$) for finite M ; see below.

Likelihood of smFISH data from probability distributions

The smFISH data are for a finite sample of cells at several time points $\{0, \tau_1, \tau_2, \dots, \tau_S\}$ after induction. Each cell was in a state, i.e., number of active TSs and mRNA molecules, contained within $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^T$. The size (N) of the vector and matrix is determined by $N = p(M + 1)$, where p is the number of distinct promoter states per cell ($p = 3$ for a two-state model and two alleles per cell, i.e., a cell can have zero, one, or two active TSs). M is the maximum number of mRNA molecules a cell can display, which could, in principle, be infinite. For practical purposes, we chose $M = 500$ because it is finite and larger than the expected mRNA levels in our smFISH data. The smFISH data vector \mathbf{Y}^t for sample t is a count of observed cell states, where $[n_1, n_2, \dots, n_N]^T$. The likelihood of having sampled the observed data given the calculated distributions $P(X, \tau)$ for model parameters θ is a product of multinomial distributions:

$$\mathcal{L} = P(Y|\theta) = \prod_{t=0}^S \left[\left(\frac{(\sum_j Y_j^t)!}{\prod_k Y_k^t!} \right) \prod_{i=1}^N [P(\mathbf{x}_i, \tau_t)]^{Y_i^t} \right]. \quad (7)$$

Calculating the Bayesian posterior probability

The Bayesian posterior probability is the likelihood \mathcal{L} multiplied by $P(\theta)$ and divided by $P(Y)$, which are the prior probability distributions of the parameters and data. These priors are often unknown and $P(\theta)$ and $P(Y)$ are presumed flat and constant, i.e., any parameter set and data set are equally likely. BayFish assumes flat priors unless specified otherwise. We implemented a Heaviside step function for $P(\theta)$, where the prior was zero for non-physiological parameters, but otherwise flat and constant. Non-physiological parameters include negative numbers (i.e., below 10^{-8}) or a maximum transcription rate (i.e., 12–18 mRNAs per minute; see [31]).

Validating BayFish with synthetic smFISH data

To test the ability of BayFish to infer parameters correctly, we generated synthetic smFISH data from a two-state model with known parameters. Our first model was a k_1 -stimulus model, where k_1 changed from k_1^U to k_1^S upon induction and all other parameters stayed constant; see

“Methods”. We created three technical replicates of synthetic smFISH data with a similar sampling density and number of time points as our *Npas4* data. Each technical replicate (Fig. 3a) is different from the others only because of sampling error. We then ran BayFish using an underlying k_1 -stimulus model to infer the free parameters of each technical replicate. The mRNA degradation rate was not a free parameter in these BayFish runs and was fixed to its known value to mimic our situation for *Npas4*.

If the synthetic smFISH data were too sparse to constrain the model, then we would expect the Bayesian posterior distributions to be flat. However, each BayFish run converged to well-defined Bayesian posterior distributions of model parameters and the technical replicates had posterior distributions that were relatively close to one another and overlapped the true underlying parameters (Fig. 3b). This demonstrates that sparsely sampled smFISH data at multiple time points already constrain the parameters of the underlying model. We then created a synthetic smFISH data set using the same k_1 -stimulus model, but varied the sampling density at each time point ($n = 30, 100, 300, 1000$ cells). As expected, increasing the sampling density better constrained the Bayesian posterior distribution and more accurately estimated the underlying model parameters (Fig. 3c).

Model selection using BayFish and information criteria

Previously, we initialized BayFish with the correct underlying model (k_1 -stimulus model). However, one does not usually know the correct model and it has to be inferred along with the unknown parameters. It is well known that models with more parameters have a higher likelihood of fitting the data. Thus, we combined BayFish with several likelihood-based metrics to evaluate different underlying models and penalize those with more free parameters (see “Methods”). These metrics are the Bayesian information criterion (BIC) [32] and the Akaike information criterion (AIC) [33], which are based on the maximum likelihood calculated by BayFish. The deviance information criterion (DIC) [34] uses both the likelihood and the Bayesian posterior distribution calculated by BayFish.

To test the ability of BayFish and information criteria to select the correct model, we generated two synthetic smFISH data sets from different parameter-stimulus models. The first set was generated using a k_1 -stimulus model, whereas the second set was generated using a more complex (k_1, k_0, μ_1) -stimulus model; see “Methods.” We then systematically ran BayFish using multiple underlying parameter-stimulus models, where different combinations of parameters were affected by the stimulus: k_1 -, k_0 -, μ_1 -, (k_1, μ_1) -, (k_0, μ_1) -, and (k_1, k_0, μ_1) -stimulus models. The one-parameter-stimulus models had five free parameters and the three-parameter-stimulus model had seven free parameters to be inferred. As before, the mRNA

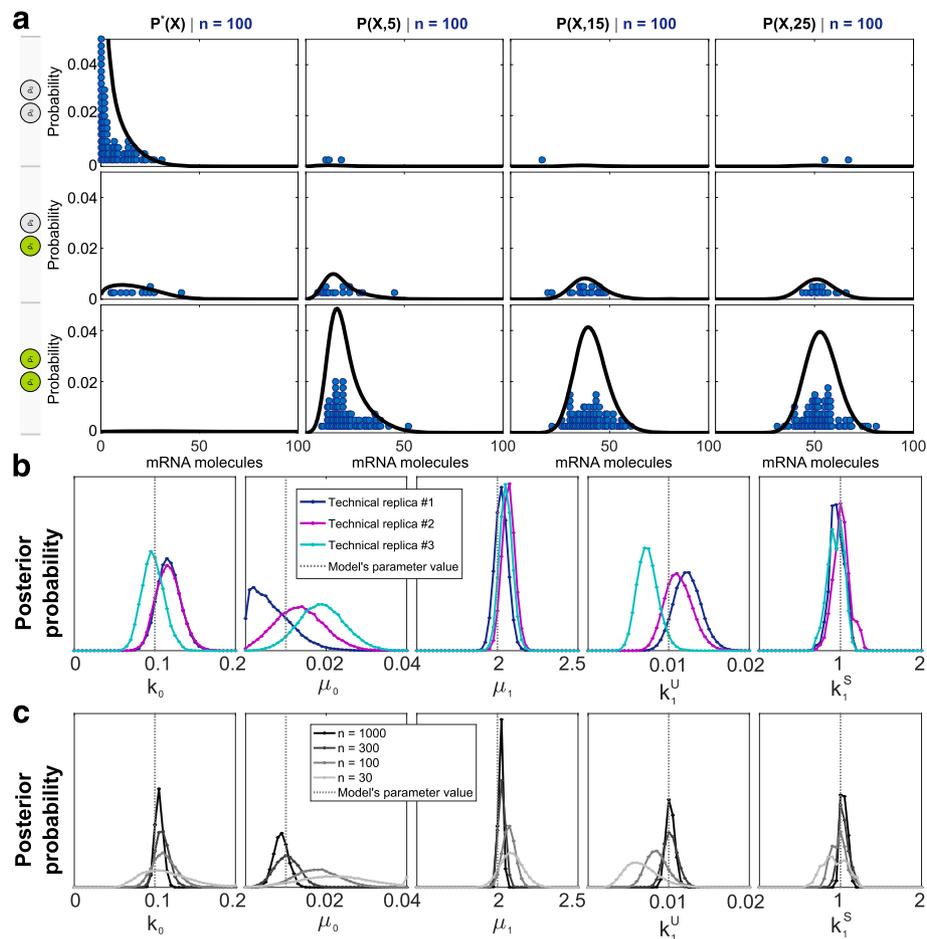


Fig. 3 Validating BayFish on synthetic smFISH data. **a** Example of a technical replicate of synthetic smFISH data generated for an underlying k_1 -stimulus model, where $n = 100$ cells per time point were sampled from the mRNA and active transcription site distribution (solid lines). Our synthetic data were sampled at $t = 0, 5, 15, 25$ min after induction. **b** Marginal Bayesian posterior distributions of parameters estimated by BayFish on three technical replicates (different colors). **c** Marginal Bayesian posterior distributions of parameters estimated by BayFish for different sampling densities ($n = 30, 100, 300, 1000$ cells per time point) of the same technical replicate. Vertical dashed lines are the true parameters of the k_1 -stimulus model used to generate the synthetic data

degradation rate was fixed to its known value. We ran three replicas of BayFish with random initial parameters for $T = 10^5$ iterations for each underlying parameter-stimulus model. We then plotted the different information metrics obtained from each BayFish run on the k_1 -stimulus synthetic data set (Fig. 4a) and the (k_1, k_0, μ_1) -stimulus synthetic data set (Fig. 4b). A lower information criterion score indicates that the underlying model had a better fit. Our results with synthetic data demonstrate that BayFish and the different information criteria select the correct underlying model.

Running BayFish on *Npas4* smFISH data

We then used BayFish to infer parameters and select an underlying parameter-stimulus model for the *Npas4* smFISH data. We used the same approach as above, but

the *Npas4* mRNA degradation rate constant was fixed to $\delta = 0.0559 \text{ min}^{-1}$ [35]. Our results demonstrate that the best underlying model with the fewest parameters is the (k_1, k_0) -stimulus model (Fig. 5). The inferred mRNA and active TS distribution and Bayesian posterior distribution of the (k_1, k_0) -stimulus model are shown in Fig. 6 and summarized in Table 1. Model selection using BayFish and information criteria also showed that not all parameters are equivalent. Regulation of k_1 by the stimulus consistently gave a better fit to the observed data than regulation by k_0 or μ_1 alone or in combination.

Discussion

Like any inference approach, BayFish is limited by the information content of the data and the underlying model assumptions. For example, our smFISH data measured the

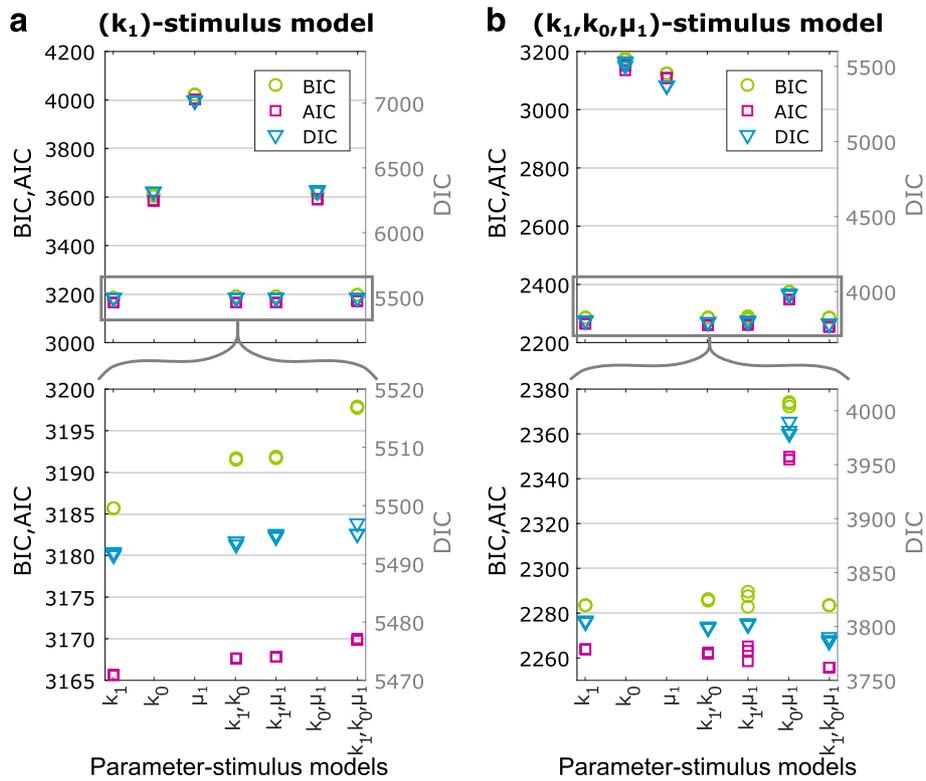


Fig. 4 Model selection using BayFish and information criteria. We applied the Bayesian information criterion (BIC), the Akaike information criterion (AIC), and the deviance information criterion (DIC) metrics to the BayFish results obtained with the different parameter-stimulus models listed on the x-axis. All models were run on the same synthetic smFISH data. The maximum likelihood observed in each BayFish run was used for BIC and AIC metrics, and the full likelihood and Bayesian posterior distribution, excluding the burn-in period, were used for DIC. Models with the lowest BIC and AIC scores (left, y-axis) and DIC (right, y-axis) are the most informative models with the fewest parameters. **a** BayFish results for synthetic smFISH data ($n = 100$ cells per time point) generated for an underlying k_1 -stimulus model. **b** BayFish results for synthetic smFISH data ($n = 100$ cells per time point) generated for an underlying (k_1, k_0, μ_1) -stimulus model. AIC Akaike information criterion, BIC Bayesian information criterion, DIC Deviance information criterion

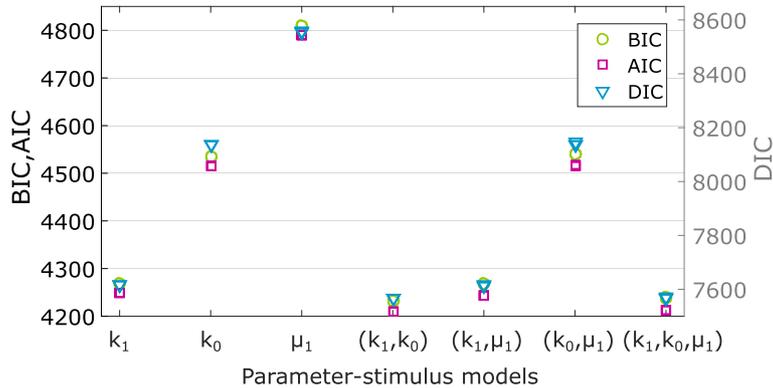


Fig. 5 Comparing different stimulus models for *Npas4* smFISH data. We applied the BIC, AIC, and DIC metrics to the *Npas4* BayFish results obtained with the different parameter-stimulus models listed on the x-axis. For each parameter-stimulus model, three replicas of BayFish were run with different initial conditions. The Bayesian posterior distributions for each parameter-stimulus model are shown in Additional file 1: Figure S1. AIC Akaike information criterion, BIC Bayesian information criterion, DIC Deviance information criterion

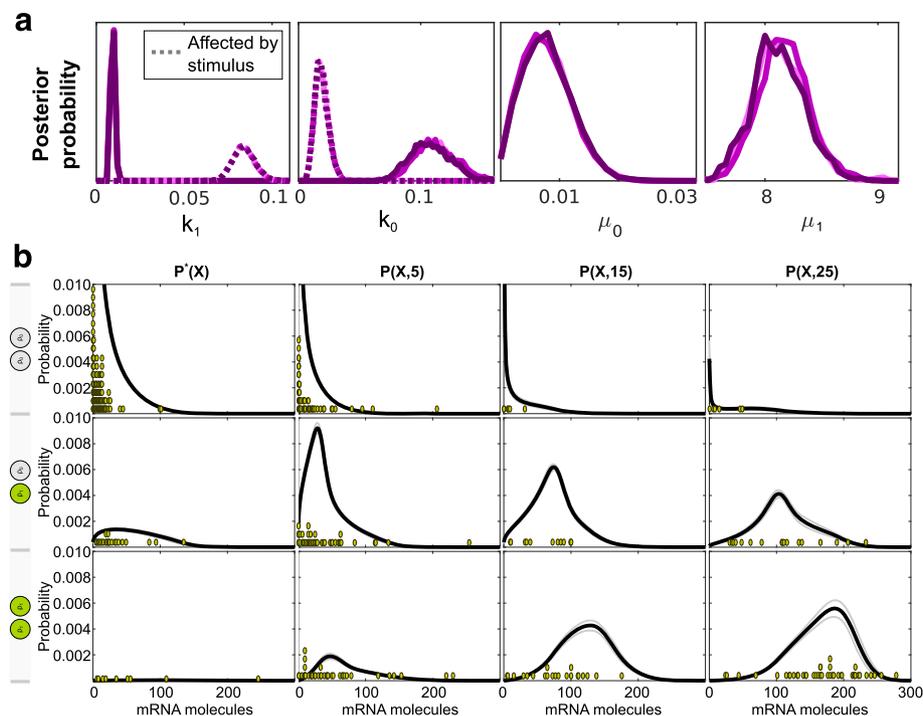


Fig. 6 Bayesian posterior distribution for (k_1, k_0) -stimulus model run on *Npas4* smFISH data. **a** Marginal posterior distributions of parameters for BayFish replicas (i.e., different colors correspond to distinct random number generator seeds and initial conditions). There are two distributions for k_0 and k_1 , one of which is the pre-stimulus parameter (continuous lines) and the other is the post-stimulus parameter (dotted lines). **b** The mean distribution of mRNA and active transcription site ($P(X, \tau)$) as inferred from the Bayesian posterior distribution of parameters. The standard deviation (σ_p) is shown in gray. A histogram of experimental data is shown for comparison (green dots)

mRNA and active TS counts per cell, but one could also measure the brightness of TS sites from the smFISH data to estimate the number of nascent mRNAs and, hence, the transcription rate (μ_1) [20]. This additional information could further constrain the underlying model, as has been done by others [15]. We did not include or fit TS intensity in our *Npas4* model and, thus, this provides an independent test of the μ_1 parameter inferred by BayFish. We estimated nascent mRNAs and the transcription rate from *Npas4* active TSs (Additional file 1: Figure S2). The estimated transcription rate has a strong mode between 7–10 mRNA min^{-1} , which is consistent with our inferred μ_1 of 8 mRNA min^{-1} ; see Table 1.

However, our analysis also shows that there are caveats with estimating the transcription rate from the integrated intensity of active TSs. First, the estimate depends on the choice of transcription elongation rate, which can vary across genes and organisms [31]. Second, some active TSs have more nascent mRNAs and a higher transcription rate than theoretically possible (grey area in Additional file 1: Figure S2). The simplest explanation for these unusually bright spots is that mRNAs continue to be associated with chromatin at active TSs after transcription until further processing [36]. Thus, the integrated intensity of an active TS cannot be assumed to represent only nascent mRNAs in the process of transcription.

Table 1 Estimated parameters for *Npas4* (k_1, k_0) -stimulus model

	Parameter	Mean	Standard deviation	Units
k_1^U	Activation rate	0.0093	0.0010	min^{-1}
k_1^S	Activation rate after stimulus	0.0839	0.0063	min^{-1}
k_0^U	Deactivation rate	0.1108	0.0172	min^{-1}
k_0^S	Deactivation rate after stimulus	0.0189	0.0056	min^{-1}
μ_0	ρ_0 synthesis rate	0.0078	0.0042	mRNA min^{-1}
μ_1	ρ_1 synthesis rate	8.14	0.2305	mRNA min^{-1}
δ	mRNA degradation rate	0.0559	–	min^{-1}

Our mathematical model of stochastic gene expression also assumed that each promoter allele was regulated independently [11, 15, 23, 27]. However, previous work has also shown that genes can exhibit strongly correlated gene expression, particularly when integrated adjacent to one another on the same chromosome [14, 37]. If one *Npas4* allele is independent of the other, then we expect the active TS to exhibit a binomial distribution of zero, one, or two active TSs with probability $(1-p)^2$, $2p(1-p)$, or p^2 where $p = k_1/(k_0 + k_1)$, i.e., the probability of an active allele or burst fraction. Although our results show a statistically significant difference between the measured and expected fractions for independent alleles (Additional file 1: Figure S3), the data are closer to independent alleles than perfectly correlated alleles (i.e., there are no cells with one active TS). Modeling the weak correlations between alleles and the post-transcriptional processing of mRNAs at active TSs is beyond the scope of our current software package, but these could be potentially informative extensions of BayFish.

Conclusions

We developed a suite of MATLAB programs (BayFish), and an alternative C++ version, that use Bayesian inference to estimate model parameters robustly from smFISH data. We expect this software package to be useful for other labs because it fills a critical gap in the downstream analysis of population snapshots of smFISH in single cells. The user specifies any mathematical model of stochastic gene expression with an unknown set of parameters (θ) and provides smFISH data (Y) of mRNA and active TS counts in a population of cells at different time points before and after induction. BayFish uses a Monte Carlo method to estimate the Bayesian posterior probability $P(\theta|Y)$ of the model parameters, which elucidates the best-fitting parameters and quantifies their uncertainty. Based on the confidence intervals of inferred parameters from a current data set, BayFish permits labs to design the next set of experiments and collect additional smFISH data (e.g., different times or more cells) that is maximally informative.

We generated synthetic data to validate the ability of BayFish to infer the correct parameters and tested its performance on smFISH data sets with sampling error. We further demonstrated how BayFish can be combined with information criteria to select the most informative underlying model. Finally, we used BayFish to extract meaningful biological information from *Npas4* gene expression in single neurons (Fig. 1). Our results favor a two-state model where the stimulus increases k_1 and decreases k_0 . Both parameters modulate the *Npas4* burst fraction, e.g., fraction of time that a promoter spends in the active, on state, without changing the transcription rate of the on or off state. Modulation of the burst fraction upon induction

is consistent with previous observations for other genes [13, 15, 38], although modulation of the transcription rate (μ_1) upon induction has also been documented [14]. Future experiments will address mechanisms of activation and cell-to-cell variability in *Npas4* and other immediate-early genes of primary neurons. This can be done by combining genetic and pharmacological perturbations of gene expression with downstream BayFish analysis of multi-color smFISH distributions of several immediate-early genes.

Methods

Npas4 smFISH measurements in single neurons

Neuron-enriched cultures were generated from the cortex of male and female E16.5 CD1 mouse embryos (Charles River Laboratories Inc., Wilmington, MA, USA) and cultured as previously described [39]. Neurons were treated with 1 μ M tetrodotoxin (TTX) (Tocris Cookson, Ballwin, MO, USA), a sodium channel inhibitor, at DIV6 and depolarized by elevating the extracellular potassium concentration to 55 mM with an isotonic KCl solution at DIV7 [40], which activates L-type voltage-gated calcium channel dependent transcription of *Npas4* [41]. Cells were fixed at four time points: no KCl, 5 min KCl treatment, 5 min KCl treatment plus 10 min condition medium, and 5 min KCl treatment plus 20 min condition medium as indicated in Fig. 1.

Neurons were fixed in 4% Paraformaldehyde (PFA) at room temperature for 10 min after sampling and permeabilized by 70% (v/v) EtOH at 4 °C overnight. The mouse *Npas4* mRNAs were hybridized with the Quasar[®] 570 Stellaris RNA FISH Probe set following the manufacturer's instructions, which are available online. Custom Stellaris[®] FISH Probes were designed against mouse *Npas4* mRNA by utilizing the Stellaris[®] RNA FISH Probe Designer (Biosearch Technologies, Inc., Petaluma, CA, USA), which is available online. We hybridized probes to samples in a hybridization buffer (10% formamide, 10% 20 \times SSC, 10% dextran sulfate, 1 mg mL⁻¹ *Escherichia coli* tRNA, 2 mM vanadyl ribonucleoside complex, and 20 μ g mL⁻¹ Bovine serum albumin (BSA)) at 37 °C for 4 hours followed by Hoechst staining. Z-stack images were captured on a wide-field microscope (DMI4000, Leica) equipped with a CCD camera (DFC365 FX, Leica) and controlled by MetaMorph (Molecular Devices). An objective with NA 1.4 and 63 \times magnification yielded an xy pixel-size of 146 nm. Then, 35–45 Z-slices were recorded with a 200 nm step-size and 1 second exposure time.

We used FISH-quant [21] to identify and count absolute mRNA numbers and active TSs in single cells (Fig. 1). The active TSs can be detected because nascent mRNAs are transiently attached to the elongating RNA Polymerase II in the gene, accumulating fluorescent probes around

active sites, and then appear as highly intense dots (one or two, as there are two copies of the gene) in the nucleus of the diploid cell. We and others have confirmed that these nuclear spots mark the active TSs because they colocalize in two-color smFISH with probes specific for the gene introns, which are present only in nascent RNAs (data not shown and [42]).

Monte Carlo sampling and burn-in

The number of iterations (T), covariance matrix (Σ), and burn-in period were determined by monitoring the acceptance rate of proposals and the distribution of parameters and likelihood in the stationary phase of the Monte Carlo algorithm. The rate at which the Markov chain approaches stationarity (i.e., the region with higher likelihood) depends on the covariance matrix Σ used to draw new proposals. We defined the burn-in as the initial period where the log-likelihood was increasing and less than 99.5% of the maximum. The burn-in period is sensitive to the initial parameters and the parameter-stimulus model. Given our experimental data, we verified that $T = 10^5$ iterations and our covariance matrix Σ were sufficient for BayFish to achieve stationarity and adequately sample the Bayesian posterior distribution after discarding the burn-in. The final covariance matrix Σ was diagonal with 10^{-5} for k_0, k_1, μ_0 and 10^{-3} for μ_1 proposals.

Generating synthetic smFISH data

For a given stimulus model and known parameter set (truth), we calculated the mRNA and active TS distributions using the algorithms described in the main text. The pre-stimulus stationary distribution at $t = 0$ min was generated using the unstimulated parameters, whereas the post-stimulus distributions at $t = 5, 15, 25$ min were generated using the stimulated parameters. From these distributions, we created a technical replicate by randomly sampling n cells from each mRNA and active TS distribution calculated at each time point. We generated technical replicates of synthetic smFISH data for two parameter-stimulus models: a k_1 -stimulus and a (k_1, k_0, μ_1) -stimulus model. The k_1 -stimulus model had the following six parameters: $k_1^U = 0.01 \text{ min}^{-1}$, $k_1^S = 1 \text{ min}^{-1}$, $k_0 = 0.1 \text{ min}^{-1}$, $\mu_1 = 2 \text{ mRNA min}^{-1}$, $\mu_0 = 0.01 \text{ mRNA min}^{-1}$, and $\delta = 0.05 \text{ min}^{-1}$. The (k_1, k_0, μ_1) -stimulus model had the following eight parameters: $k_1^U = 0.01 \text{ min}^{-1}$, $k_1^S = 1 \text{ min}^{-1}$, $k_0^U = 1 \text{ min}^{-1}$, $k_0^S = 0.01 \text{ min}^{-1}$, $\mu_1^U = 0.2 \text{ mRNA min}^{-1}$, $\mu_1^S = 2 \text{ mRNA min}^{-1}$, $\mu_0 = 0.01 \text{ mRNA min}^{-1}$, and $\delta = 0.05 \text{ min}^{-1}$.

Information criterion and model fitting

We used several information criteria, such as BIC [32], AIC [33], and DIC [34], to evaluate the likelihood of different models and to penalize model over-fitting:

- Bayesian information criterion:

$$\text{BIC} = -2 \ln(\hat{\mathcal{L}}) + m \ln(n). \quad (8)$$

- Akaike information criterion:

$$\text{AIC} = -2 \ln(\hat{\mathcal{L}}) - 2m + \frac{2m(m+1)}{n-m-1}. \quad (9)$$

The maximum likelihood $\hat{\mathcal{L}} = P(Y|\hat{\theta})$ is the maximum value of \mathcal{L} obtained during the BayFish run, m is the number of free parameters that were fit, and n is the total sample size. These metrics do not take full advantage of the Bayesian posterior probability estimated by BayFish. Thus, we also used:

- Deviance information criterion:

$$\text{DIC} = 2\bar{D} - D(\bar{\theta}). \quad (10)$$

The deviance is

$$D(\theta) = -2 \ln P(Y|\theta) = -2 \ln \mathcal{L}. \quad (11)$$

$\bar{D} = E[D(\theta)]$ is the mean of the deviance $D(\theta)$ calculated from the Bayesian posterior probability, whereas $D(\bar{\theta}) = D(E[\theta])$ is the deviance of the mean of θ calculated from the Bayesian posterior probability.

Availability and requirements

Project name: BayFish

Project homepage: <https://github.com/mgschiavon/BayFish>; <http://doi.org/10.5281/zenodo.830056>

Operating system: Platform independent

Programming language: MATLAB or C++

Other requirements: See README file in the project homepage.

License: GNU General Public License v3.0

Additionally, the datasets analyzed during the current study are also available in the GitHub repository, <https://github.com/mgschiavon/BayFish/tree/master/DATA>.

Additional file

Additional file 1: Supplementary figures S1–S3. (PDF 233 kb)

Abbreviations

AIC: Akaike information criterion; BIC: Bayesian information criterion; CME: Chemical master equation; DIC: Deviance information criterion; FISH: Fluorescence in situ hybridization; RNA: Ribonucleic acid; smFISH: Single-molecule RNA FISH; TS: transcription site

Acknowledgments

We are grateful to Sayan Mukherjee and Stefano Di Talia for advice and feedback.

Funding

This work was supported by a CONACYT graduate fellowship (MGS), the National Institutes of Health Exploratory/Developmental Research Grant Award R21DA041878 (AEW), the National Institutes of Health Director's New Innovator Award DP2 OD008654-01 (NEB), the Burroughs Wellcome Fund

CASI Award BWF 1005769.01 (NEB), and seed funding from the Duke Center for Genomic & Computational Biology (AEW and NEB).

Authors' contributions

MGS developed the software and analyzed the data. LFC performed the smFISH experiments and analyzed the data. AEW and NEB conceived the project and supervised the research. MGS and NEB wrote the manuscript. All authors read, edited, and approved the final manuscript.

Ethics approval

All experiments were conducted in accordance with an animal protocol approved by the Duke University Institutional Animal Care and Use Committee.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Program in Computational Biology & Bioinformatics, Duke University, Durham, NC, USA. ²Present address: Department of Biochemistry & Biophysics, University of California, San Francisco, CA, USA. ³Department of Neurobiology, Duke University, Durham, NC, USA. ⁴Department of Biology, Duke University, Durham, NC, USA. ⁵Department of Physics, Duke University, Durham, NC, USA. ⁶Center for Genomic & Computational Biology, Duke University, Durham, NC, USA.

Received: 23 April 2017 Accepted: 10 August 2017

Published online: 04 September 2017

References

- Lenstra TL, Rodriguez J, Chen H, Larson DR. Transcription dynamics in living cells. *Annu Rev Biophys*. 2016;45(1):25–47.
- Kaufmann BB, van Oudenaarden A. Stochastic gene expression: from single molecules to the proteome. *Curr Opin Genet Dev*. 2007;17(2):107–12.
- Sanchez A, Golding I. Genetic determinants and cellular constraints in noisy gene expression. *Science*. 2013;342(6163):1188–93.
- Suter DM, Molina N, Naef F, Schibler U. Origins and consequences of transcriptional discontinuity. *Curr Opin Cell Biol*. 2011;23(6):657–62.
- Golding I, Paulsson J, Zawilski SM, Cox EC. Real-time kinetics of gene activity in individual bacteria. *Cell*. 2005;123(6):1025–36.
- Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*. 2010;329(5991):533–8.
- Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. Systematic identification of signal-activated stochastic gene regulation. *Science*. 2013;339(6119):584–7.
- Zenkhusen D, Larson DR, Singer RH. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*. 2008;15(12):1263–71.
- Bothma JP, Garcia HG, Esposito E, Schlissel G, Gregor T, Levine M. Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living *Drosophila* embryos. *Proc Natl Acad Sci*. 2014;111(29):10598–603.
- Fukaya T, Lim B, Levine M. Enhancer control of transcriptional bursting. *Cell*. 2016;166(2):358–68.
- Bahar Halpern K, Tanami S, Landen S, Chapal M, Szlak L, Hutzler A, et al. Bursty gene expression in the intact mammalian liver. *Mol Cell*. 2015;58(1):147–56.
- Battich N, Stoeger T, Pelkmans L. Control of transcript variability in single mammalian cells. *Cell*. 2015;163(7):1596–610.
- Dar RD, Razoooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, et al. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc Natl Acad Sci*. 2012;109(43):17454–9.
- Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol*. 2006;4(10):e309.
- Senecal A, Munsky B, Proux F, Ly N, Braye FE, Zimmer C, et al. Transcription factors modulate c-Fos transcriptional bursts. *Cell Rep*. 2014;8(1):75–83.
- Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. *Science*. 2011;332(6028):472–4.
- Femino AM, Fay FS, Fogarty K, Singer RH. Visualization of single RNA transcripts in situ. *Science*. 1998;280(5363):585–90.
- Levsky JM, Shenoy SM, Pezo RC, Singer RH. Single-cell gene expression profiling. *Science*. 2002;297(5582):836–40.
- Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods*. 2008;5(10):877–9.
- Bahar Halpern K, Itzkovitz S. Single molecule approaches for quantifying transcription and degradation rates in intact mammalian tissues. *Methods*. 2016;98:134–42.
- Mueller F, Senecal A, Tantale K, Marie-Nelly H, Ly N, Collin O, et al. FISH-quant: automatic counting of transcripts in 3D FISH images. *Nat Methods*. 2013;10(4):277–8.
- Munsky B, Fox Z, Neuert G. Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. *Methods*. 2015;85:12–21.
- Sepulveda LA, Xu H, Zhang J, Wang M, Golding I. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science*. 2016;351(6278):1218–22.
- Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys*. 2006;124(4):044104.
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953;21:1087–92.
- Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. 1970;57:97–109.
- Skinner SO, Xu H, Nagarkar-Jaiswal S, Freire PR, Zwaka TP, Golding I. Single-cell analysis of transcription kinetics across the cell cycle. *eLife*. 2016;5(12):7250–7.
- McQuarrie DA. Stochastic approach to chemical kinetics. *J Appl Probab*. 1967;4:413–78.
- Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*. 1977;81(25):2340–61.
- Lehoucq RB, Sorensen DC. Deflation/Techniques for an implicitly re-started Arnoldi iteration. *SIAM J Matrix Anal Appl*. 1996;17:789–821.
- Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol*. 2007;14(2):103–5.
- Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461–4.
- Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, editors. *Selected papers of, Hirotugu Akaike*. New York: Springer New York; 1998. p. 199–213.
- Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol*. 2002;64(4):583–639.
- Speckmann T, Sabatini PV, Nian C, Smith RG, Lynn FC. Npas4 transcription factor expression is regulated by calcium signaling pathways and prevents tacrolimus-induced cytotoxicity in pancreatic beta cells. *J Biol Chem*. 2016;291(6):2682–95.
- Bhatt DM, Pandya-Jones A, Tong AJ, Barozzi I, Lissner MM, Natoli G, et al. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*. 2012;150(2):279–90.
- Becskei A, Kaufmann BB, van Oudenaarden A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat Genet*. 2005;37(9):937–44.
- Larson DR, Fritsch C, Sun L, Meng X, Lawrence DS, Singer RH. Direct observation of frequency modulated transcription in single cells using light activation. *eLife*. 2013;2(2):1–20.
- McDowell KA, Hutchinson AN, Wong-Goodrich SJ, Presby MM, Su D, Rodriguiz RM, et al. Reduced cortical BDNF expression and aberrant memory in Carf knock-out mice. *J Neurosci*. 2010;30(22):7453–65.

40. Lyons MR, Chen LF, Deng JV, Finn C, Pfenning AR, Sabhlok A, et al. The transcription factor calcium-response factor limits NMDA receptor-dependent transcription in the developing brain. *J Neurochem*. 2016;137(2):164–76.
41. Lin Y, Bloodgood BL, Hauser JL, Lapan AD, Koon AC, Kim TK, et al. Activity-dependent regulation of inhibitory synapse development by Npas4. *Nature*. 2008;455(7217):1198–204.
42. Levesque MJ, Raj A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat Methods*. 2013;10(3):246–8.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

