Controlled inference of transcriptional bursting from single-cell RNA counts

Matthew D. Schwartz*
Department of Physics, Harvard University, Cambridge, MA 02138, USA

Martin Hemberg
The Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital, and Harvard Medical School, Boston, MA, USA

This manuscript is preliminary. The first draft was written by Claude (Anthropic), and the text and computations are still under review by the authors.

September 28, 2026

Abstract

Genes make RNA in bursts: a gene sits silent most of the time, switches on briefly, produces a run of RNA, and switches off. In the telegraph model each gene has a rate and a burst size and these numbers can be approximated by maximum-likelihood fits to single-cell RNA counts. Each likelihood is a hypergeometric function and a realistic genome-scale fit must multiply hundreds of them. Standard tools struggle with this multiplication and errors quickly grow out of control. We instead use ball arithmetic, a powerful mathematical tool used often in computer-assisted proofs, to produce fitted rates with sensible and trustworthy error bars. Applying this technique to the large allelic data set of Larsson et al., we determine the effect of regulatory DNA. We find that allelic differences in bursting are common, in roughly half of the genes tested, and that the differences we can resolve are in burst frequency. With error propagation under control, we also observe that no gene in this data set was measured in enough cells to reveal whether its promoter needs recovery time between bursts. More generally, the techniques we develop should allow for the extraction of much meaningful biology from noisy single-cell RNA datasets.

Keywords: transcriptional bursting; telegraph model; single-cell RNA-seq; maximum likelihood; ball arithmetic; allele-specific expression

Introduction

Genes make RNA in bursts: a promoter switches on, produces a run of transcripts, and switches off. Two numbers summarize each gene's regulation: how often bursts fire and how many transcripts a firing makes. Which of the two a regulatory difference moves is a mechanism question, because a gene can be turned up either way. Bursting was established by imaging transcription in living cells [1–3] and by counting single RNA molecules in fixed cells [4, 5], methods that follow a few chosen genes at a time. The question of how promoters and enhancers share control of burst frequency and burst size predates these measurements. One early proposal was that enhancers act by increasing the probability of transcription [6], and live imaging has since shown an enhancer

*Work done with Anthropic, but results and views are not endorsed by Anthropic.

controlling the burst frequency of its target gene [7]. Testing such proposals genome-wide requires burst kinetics inferred from single-cell RNA sequencing.

In hybrid mice from a cross of two inbred strains (here C57BL/6J and CAST/EiJ), single-cell sequencing counts the transcripts of each gene's two alleles separately [8]. The counts are small (zero in most cells, single digits when not, tens for a cell caught mid-burst), and burst kinetics are inferred from their distribution across cells with the two-state telegraph model [9, 10]. In the model a promoter switches from OFF to ON at rate k_{on} and back at rate k_{off} and transcribes at rate k_{syn} while ON, and transcripts are degraded at a constant rate. The stationary distribution of the count is a Poisson mixture over a Beta-distributed promoter activity and can be written in closed form as a confluent hypergeometric function [11, 12].

Kim and Marioni fitted the telegraph model to single-cell counts by Gibbs sampling over the Beta variable [13], and the differential-expression tools BPSC and D3E are built on the same Poisson–Beta distribution [14, 15]. Jiang et al. developed SCALE, which estimates allelic kinetics by the method of moments with a technical-noise model, and reported that allelic differences act predominantly on burst frequency [16]. Larsson et al. fitted the telegraph model transcriptome-wide to allele-resolved counts by maximum and profile likelihood with the txburst software, and reported that burst frequency is encoded primarily in enhancers and burst size in core promoters [8, 17]. Luo et al. generalized the model to non-exponential dwell times and inferred its parameters by approximate Bayesian computation [18]. Other methods use moments and special functions [19, 20], finite state projection of the master equation [21, 22], or a likelihood that includes transcript capture [23].

Existing software implementations evaluate the telegraph likelihood by Gauss–Jacobi quadrature of the mixture integral with a fixed number of nodes, by a truncated series, or by sampling, and none of them reports an error bound on the value it returns. Delmans and Hemberg wrote that the hypergeometric routines available to them "become unreliable" at large arguments [15], and a recent dissertation fitting finite-state-projection (FSP) likelihoods to one gene concluded that "direct likelihood maximization using the FSP is intractable" [24]. Likelihood-ratio tests, profile confidence intervals and model comparisons take the likelihood value as input, and any error in that value propagates into them. Extended-precision and log-space evaluations have been described [25, 26] without an error bound (Supplementary File S1), and to our knowledge the likelihood had not been evaluated with controlled error. A three-term recurrence for the Poisson–Beta distribution exists in the actuarial literature on mixed-Poisson counts [27, 28], and, as far as we could determine, no gene-expression tool uses it.

The accuracy of a single likelihood evaluation is only part of the problem, because a genome-scale fit multiplies hundreds of per-cell probabilities and searches the product across the parameter space. Therefore unmeasured error compounds once per cell at every step of the search, and controlling it means carrying each value's precision through the entire calculation. Techniques for such error control have been pioneered and optimized for computer-assisted proofs in mathematics and physics. One of them, ball arithmetic [29], computes with intervals (balls) instead of numbers, every quantity carried as a ball (a value plus a radius) proven to contain the true value and every operation updating both. The bookkeeping is automatic, and the result is a guarantee rather than a hope: every likelihood value computed this way provably lies inside its stated interval. The fitted rates themselves are still the output of an ordinary floating-point optimizer run on this likelihood surface, and the error bars are profile intervals whose endpoint likelihoods are evaluated in the same way. Hemberg and Barahona built the same kind of guarantee into stochastic simulation of the master equation, "removing a source of uncontrolled uncertainty" from comparisons of stochastic models [30, 31].

Here, we construct an exact evaluator for the stationary distribution of the two-state model

and of its extension with a two-step OFF state, using ball arithmetic to give each likelihood value a guaranteed error bound. We compare the quadrature rules of txburst and BPSC with this evaluator over the parameter region of the fits of Larsson et al. and re-evaluate each of those fits. The published point estimates change by less than their statistical uncertainty for nearly all genes. However, likelihood values change for genes with a large synthesis rate or with cells whose counts are improbable under the fitted model, and the endpoints of profile confidence intervals move for a wider set of genes. In addition, the chi-square cutoff in txburst's profile-likelihood routine gives confidence intervals whose coverage is below their nominal level.

With the exact likelihood we find, in deeply sequenced fibroblasts from the same mouse cross [32], that about half of the genes tested differ between alleles in burst kinetics, with the resolved differences in burst frequency or in both parameters. Finally, detecting a two-step OFF state in snapshot counts requires more cells than Larsson et al. measured for any gene, and we compute how many cells would be needed. The evaluator is available as a Python package, **[TODO: package name]**, at **[TODO: archive DOI]**, together with the corrected per-gene tables (Data availability).

Results

The telegraph likelihood and its numerical evaluation

The telegraph model and its stationary distribution are standard [4, 9–11], and we restate them to fix notation (Fig. 1a). A promoter switches from OFF to ON at rate k_{on} and back at rate k_{off}, and while it is ON it produces transcripts at rate k_{syn}, each of which is degraded at unit rate,

$$\mathrm{OFF} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} \mathrm{ON}, \qquad \mathrm{ON} \xrightarrow{k_{\mathrm{syn}}} \mathrm{ON} + \mathrm{mRNA}, \qquad \mathrm{mRNA} \xrightarrow{1} \varnothing. \tag{1}$$

All rates are then in units of the mRNA degradation rate, and following Larsson et al. [8] we call k_{on} the burst frequency and $k_{\mathrm{syn}}/k_{\mathrm{off}}$ the burst size throughout. At stationarity the transcript count n in a cell follows a Poisson distribution mixed over a Beta-distributed promoter activity u (the Poisson–Beta distribution),

$$P(n) = \int_0^1 \mathrm{Pois}(n; k_{\mathrm{syn}} u)\, \mathrm{Beta}(u; k_{\mathrm{on}}, k_{\mathrm{off}})\, du. \tag{2}$$

In this mixture $\mathrm{Pois}(n;\lambda)$ denotes the Poisson probability of n at mean λ and $\mathrm{Beta}(u; k_{\mathrm{on}}, k_{\mathrm{off}})$ the Beta density.

The Poisson–Beta integral has a closed form in the confluent hypergeometric function ${}_1F_1$ [12], and Kummer's transformation [33] gives a second form,

$$\begin{aligned} P(n) &= \frac{k_{\mathrm{syn}}^n}{n!}\, \frac{(k_{\mathrm{on}})_n}{(k_{\mathrm{on}}+k_{\mathrm{off}})_n}\, {}_1F_1(k_{\mathrm{on}}+n;\, k_{\mathrm{on}}+k_{\mathrm{off}}+n;\, -k_{\mathrm{syn}}) \\ &= e^{-k_{\mathrm{syn}}}\, \frac{k_{\mathrm{syn}}^n}{n!}\, \frac{(k_{\mathrm{on}})_n}{(k_{\mathrm{on}}+k_{\mathrm{off}})_n}\, {}_1F_1(k_{\mathrm{off}};\, k_{\mathrm{on}}+k_{\mathrm{off}}+n;\, k_{\mathrm{syn}})\,, \end{aligned} \tag{3}$$

where $(x)_n = x(x+1)\cdots(x+n-1)$ denotes the rising factorial. The two expressions are equal but behave differently in floating-point arithmetic: the series of the first alternates in sign, whereas the second has a positive argument and a positive numerator parameter k_{off} independent of n. Hence all terms of the second series are positive at any n and their sum incurs no cancellation error.

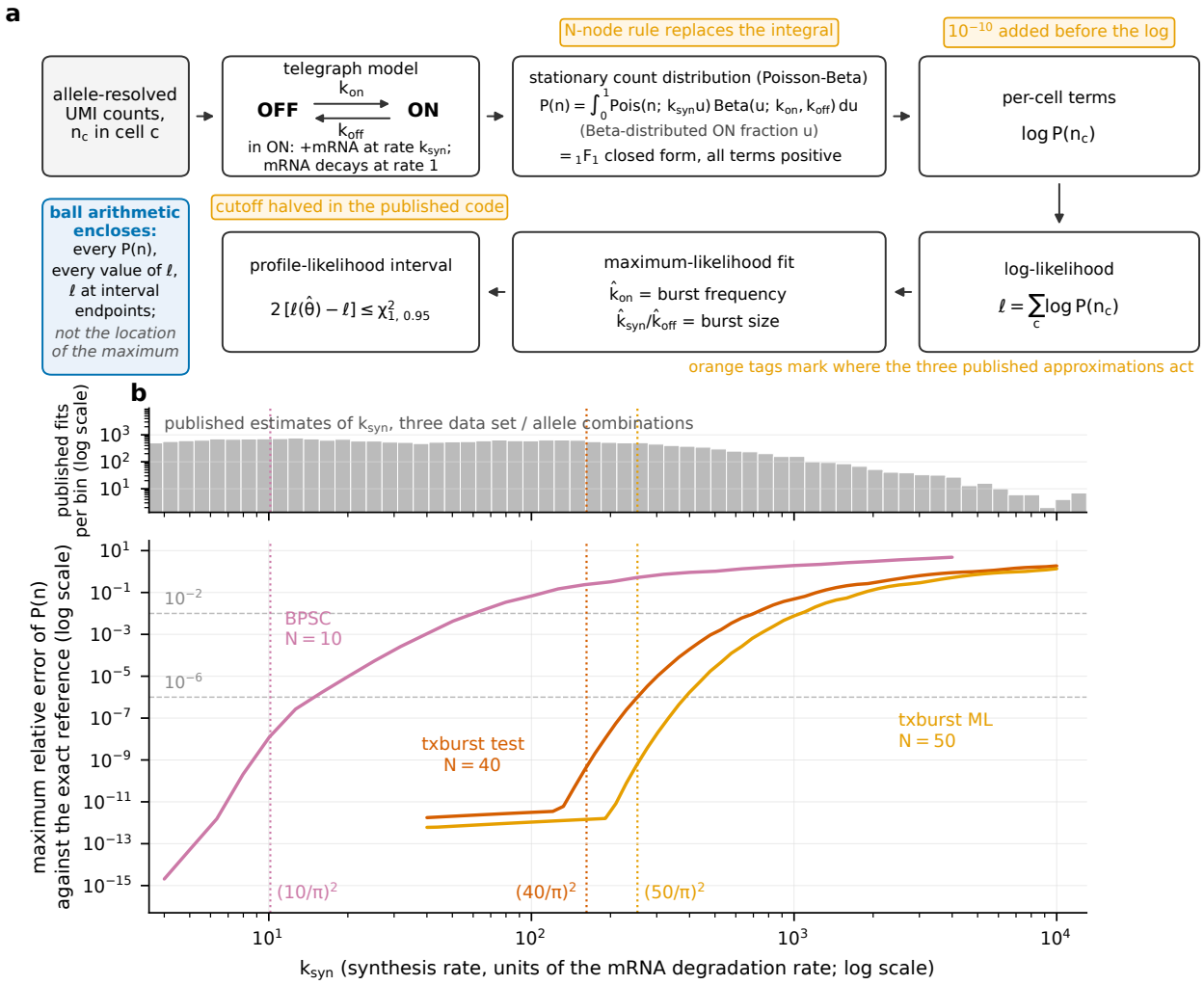

Figure 1: An N-node quadrature rule evaluates the telegraph count distribution accurately while $k_{\text{syn}} \lesssim (N/\pi)^2$, and most published fits lie below the onset of the 50-node rule. (a) The objects of the inference problem, from allele-resolved counts through the telegraph model, its Poisson–Beta distribution with the ${}_1F_1$ closed form, the per-cell logarithms and their sum ℓ to the fit and the profile interval. Orange tags mark where the three published approximations act (the N-node quadrature rule, the constant added before the logarithm and the halved interval cutoff); the box at lower left lists what evaluation in ball arithmetic encloses. (b) Top, published estimates of k_{syn} for the 32,376 fits of the three data set and allele combinations of Larsson et al. [8] (shown against burst size in Supplementary Fig. S1). Bottom, largest relative error of $P(n)$ against the exact reference at the median fibroblast shape (Table 1, row 2) for $N = 10$ (BPSC), 40 (txburst test) and 50 (txburst maximum likelihood). Dotted verticals mark $(N/\pi)^2 = 10.1$, 162 and 253; the errors first exceed 10^{-6} at $k_{\text{syn}} = 15.9$, 276 and 399. The onset is shape-dependent, and $(N/\pi)^2$ sets its scale.

For a gene with counts $n_1, \ldots, n_C$ in C cells, we write the log-likelihood of the three rates as

the sum of the per-cell log-probabilities,

$$\ell(k_{\text{on}}, k_{\text{off}}, k_{\text{syn}}) = \sum_{c=1}^{C} \log P(n_c). \tag{4}$$

The rate estimates are obtained by maximizing ℓ over the three rates, and profile-likelihood intervals by constrained maximization. Each probability in Eq. (4) is a hypergeometric function, a gene contributes hundreds of them (one per cell), and in a genome-scale analysis the sum is evaluated for thousands of genes, thousands of times each during optimization and profiling. The count distribution changes little when k_{off} and k_{syn} increase together at fixed burst size. Thus the likelihood has a ridge along which the two rates are poorly constrained individually [34]. We follow Larsson et al. in reporting burst frequency and burst size, which the counts determine more reliably [8].

Published software for fitting the telegraph model evaluates the log-likelihood of Eq. (4) by quadrature or replaces its evaluation by Monte Carlo sampling (Methods and Supplementary File S7). In txburst, the maximum-likelihood software of Larsson et al. [8], the mixture integral of Eq. (2) is evaluated with a fixed Gauss–Jacobi rule of 50 nodes (40 in the hypothesis-test module). The code adds 10^{-10} to each probability before taking the logarithm, a routine guard against $\log 0$ (Fig. 1a). BPSC, the beta-Poisson differential-expression package [14], uses the same quadrature with 10 nodes, and earlier sampling-based methods draw the Beta variable by Monte Carlo rather than evaluate $P(n)$ [13, 15]. A Gauss–Jacobi rule with N nodes is the natural quadrature for a Beta mixing density and replaces the integral by a weighted sum of N Poisson probabilities, one per node.

Textbook methods for evaluating the closed form of Eq. (3) are unreliable in part of the realistic parameter range. A general-purpose floating-point ${}_1F_1$ routine applied to the first form underflows there (Supplementary File S2), because the factor $e^{-k_{\text{syn}}}$ is absorbed into the special function, whose value falls below the floating-point range even where $P(n)$ is an ordinary probability. Summing the alternating series directly loses all accuracy to cancellation at moderate k_{syn}, because its terms grow far larger than their sum before they decay. Whether the quadrature error or the added constant matters depends on a gene's rates $(k_{\text{on}}, k_{\text{off}}, k_{\text{syn}})$ and on whether the likelihood value is used for a point estimate, a likelihood-ratio test or a confidence interval (Fig. 1a). To measure both dependences we first construct an exact reference.

Exact likelihoods with ball arithmetic

We evaluate the telegraph likelihood exactly through its all-positive series and a three-term recurrence run in the stable direction. For one gene (the vector of probabilities for 501 count values) the evaluation takes about 11.7 milliseconds (Table 1), at several-fold the cost of txburst's 50-node quadrature rule [8] and cheap enough to use for all reported likelihoods. Each probability is returned as a ball [29], a midpoint and a radius defining an interval that provably contains the true value. The resulting radii on the log-likelihood are far smaller than any of the likelihood differences reported below, and we call an evaluation exact when the returned radius is less than 10^{-25} of the value (Methods). We sum the all-positive form in Eq. (3) term by term in ball arithmetic without cancellation, and once the term index exceeds k_{syn} the terms decrease at least geometrically, which bounds the omitted tail (Methods).

Successive probabilities $P(n)$ in Eq. (3) satisfy a three-term recurrence,

$$n(n{+}1)\,P(n{+}1) \;=\; n\,(n + k_{\text{on}} + k_{\text{off}} + k_{\text{syn}} - 1)\,P(n) \;-\; k_{\text{syn}}\,(n + k_{\text{on}} - 1)\,P(n{-}1)\,, \qquad n \geq 1\,, \tag{5}$$

with the three rates in degradation units as in Eq. (1). Willmot stated this recurrence in the

Table 1: Exact reference values of the telegraph count distribution at eight parameter points, with the errors of the double-precision surrogate and of the three published quadrature rules and the cost of each method. Rows: (1) low expression; (2, 3) the median fibroblast and mouse embryonic stem cell (mESC) fits; (4) the onset region of the 50-node rule at large k_{off}; (5, 6) the 99th percentiles of burst size and of k_{syn} among the fibroblast fits; (7) the benchmark-grid corner of largest k_{on} and k_{syn} and smallest k_{off}, whose mass lies above the counts; (8) the grid's largest k_{syn}, the top of txburst's fitting range, at the median shape. Rates are in units of the degradation rate. $P(n)$, printed for the five rows that belong to the 25-point reference set (Supplementary File S7), is the exact value at the stated n truncated to twelve digits (relative ball radius below 10^{-60}; at most 1.9×10^{-60} over the whole 228-point grid). Error columns: largest relative error of $P(n)$ over the counts carrying all but 10^{-3} of the mass, two digits (Methods); "$-\infty$ (m)": the general-purpose ${}_1F_1$ routine returned zero probability at m attainable counts. Cost per 501-count vector, single core, grid median: exact 11.7 ms, surrogate 1.85 ms, 50-node rule 1.41 ms; the ratios vary with the point.

	$(k_{\text{on}}, k_{\text{off}}, k_{\text{syn}})$	n	exact $P(n)$	surrogate	N=50	N=40	N=10	${}_1F_1$ routine
1	$(0.03, 0.3, 2)$	2	$2.68418362465 \times 10^{-2}$	8.9×10^{-16}	1.9×10^{-10}	1.0×10^{-11}	1.7×10^{-15}	5.4×10^{-15}
2	$(0.426, 9.45, 36.7)$	10	$6.08326565079 \times 10^{-3}$	2.0×10^{-14}	6.0×10^{-13}	1.7×10^{-12}	6.8×10^{-4}	1.4×10^{-14}
3	$(0.433, 2.12, 9.39)$			1.1×10^{-14}	6.0×10^{-13}	3.2×10^{-13}	4.2×10^{-8}	9.8×10^{-15}
4	$(0.03, 999, 331)$	3	$1.60066732851 \times 10^{-4}$	6.5×10^{-13}	2.7×10^{-9}	5.4×10^{-9}	2.0×10^{-10}	6.7×10^{-13}
5	$(0.43, 30, 2000)$			7.0×10^{-13}	4.8×10^{-2}	1.6×10^{-1}	1.8	$-\infty$ (38)
6	$(2.44, 592, 2382)$			1.1×10^{-12}	8.7×10^{-10}	8.8×10^{-8}	4.7×10^{-2}	1.5×10^{-12}
7	$(12.9, 0.3, 10^4)$	1500	$1.32204786450 \times 10^{-14}$	1.4×10^{-11}	2.1	2.8	8.8	$-\infty$ (2,872)
8	$(0.426, 9.45, 10^4)$	1000	$1.90149378097 \times 10^{-4}$	1.2×10^{-11}	1.4	1.9	8.2	$-\infty$ (2,851)

actuarial literature [27, 28]. We run Eq. (5) downward in n from two series-evaluated values at the top of the count range, for a total cost of two series evaluations and one recurrence pass per vector.

Downward recursion is stable because the recurrence has a second solution that grows with n relative to $P(n)$, and errors in the two starting values are therefore damped in a downward pass and amplified in an upward one. Run downward for up to 1,500 steps, the recurrence lost at most about one decimal digit in each of the six parameter regimes we tested (Supplementary File S1). In contrast, when the recurrence was run upward from $P(0)$ and $P(1)$ in one of those regimes, no correct digits were left by $n \approx 200$. The working precision is chosen from an estimate of how fast the radii grow along the downward pass. If a returned radius nevertheless exceeds the required fraction of the value, the probability vector is re-evaluated at higher precision, and after a bounded number of failed attempts the evaluator returns no value rather than an unverified one (Methods).

On a single core, the exact evaluation takes 5.1 to 8.3 times as long as txburst's 50-node quadrature rule (1.41 ms per vector) and 5.5 to 6.4 times as long as a double-precision summation of the same all-positive series. Because a fit requires thousands of likelihood evaluations, we use this double-precision summation as a faster surrogate inside the optimizer. Each surrogate entry carries an a priori error bound, and the probability vector is recomputed in ball arithmetic if the bound on any entry exceeds 10^{-12} (Methods). Table 1 gives reference values of $P(n)$ for testing other implementations at representative parameter points inside and beyond the region of the published fits, with the errors of the double-precision surrogate and of the published quadrature rules. At 19 further parameter points the evaluator agrees with two independent high-precision computations to at least 57 digits (Methods and Supplementary File S2).

Ball arithmetic encloses the likelihood values but not the fitted rates that maximize them, which an ordinary floating-point optimizer finds on the exact likelihood surface. Each reported log-

likelihood is returned as a midpoint m and a radius r such that the true log-likelihood ℓ satisfies $|\ell - m| \le r$, and in all cases r is many orders of magnitude below one thousandth of a natural-log unit. In particular, this holds for the log-likelihoods at both endpoints of the profile-likelihood intervals, which are the error bars on the fitted rates. In the reanalysis of the published fits [8] we measure, for each gene, how far the original rate estimates move when the fit is repeated with the exact likelihood. The same series-and-recurrence evaluation in ball arithmetic extends to a promoter with a two-step OFF state, for which the recurrence has four terms.

Where fixed-order quadrature loses accuracy

An N-node Gauss–Jacobi rule evaluates the Poisson–Beta integral of Eq. (2) accurately for k_{syn} below a threshold near $(N/\pi)^2$, which we call the onset, and loses accuracy above it (Fig. 1b). The predicted onsets are $k_{\mathrm{syn}} \approx 253$, 162 and 10.1 for the 50 nodes of txburst's fitting routine [8], the 40 of its hypothesis-test module and the 10 of BPSC [14]. Near its peak the Poisson kernel in the integrand has width $\sqrt{n}/k_{\mathrm{syn}}$, and the Jacobi nodes there lie about $\pi\sqrt{n/k_{\mathrm{syn}}}/N$ apart. The ratio of kernel width to node spacing, $N/(\pi\sqrt{k_{\mathrm{syn}}})$, does not depend on the count n, and the rule therefore stops resolving the kernel for all counts at once (derivation in Supplementary File S1) when

$$k_{\mathrm{syn}} \gtrsim (N/\pi)^2 . \tag{6}$$

At the median Beta shapes $(k_{\mathrm{on}}, k_{\mathrm{off}})$ of the published fits, the measured onsets are 1.4–1.7 times the predicted ones ($k_{\mathrm{syn}} \approx 399$, 276 and 15.9 at 10^{-6} relative error, fibroblast shape), and Eq. (6) thus gives the scale of the onset up to a shape-dependent factor of order one. The three measured onsets scale as N^2. Hence, by Eq. (6), a fixed rule needs $N \gtrsim \pi\sqrt{k_{\mathrm{syn}}}$ nodes, about 314 at the top of txburst's fitting range ($k_{\mathrm{syn}} = 10^4$).

A fixed rule was nevertheless the reasonable choice for txburst, because about 90% of the fits of Larsson et al. [8] have $k_{\mathrm{syn}} \le 300$, where the 50-node rule is accurate to a relative error of 3.1×10^{-9}. Only the upper tail of the fitted synthesis rates lies beyond the onset (fibroblast ninety-fifth and ninety-ninth percentiles 871 and 2,382, Fig. 1b). BPSC [14] uses 10 nodes, and at about half of the regular benchmark-grid points (50.9%, Methods) its probabilities are in error by more than one percent. Figure 2 shows how the quadrature error, the added constant and the chi-square cutoff affect fits, tests and intervals on simulated data.

The constant 10^{-10} that txburst adds to each probability before taking the logarithm limits any cell's contribution to the negative log-likelihood to $-\ln 10^{-10} = 23.0$ natural-log units (nats). A cell whose count has almost no probability under the parameters being evaluated could otherwise contribute hundreds of nats. For each such cell the negative log-likelihood computed with the constant is therefore smaller than the exact value by the cell's true contribution minus 23.0 nats. We evaluated both likelihoods along k_{syn} for a synthetic 200-cell data set drawn from the telegraph distribution at a large synthesis rate (Fig. 2a). On this synthetic data set the negative log-likelihood with the constant lies below the exact one by up to 10,625 nats, while the two optima along k_{syn} differ by only about 0.6 conditional half-widths in burst size. When we re-fit such data sets, the joint maximizer moves by at most 0.23 conditional half-widths from a single start (Supplementary File S2), and because conditional intervals are narrower than profile intervals this number can only overstate the shift.

Beyond the onset of the 10-node rule, BPSC's estimate of the synthesis rate is biased low on simulated data (Fig. 2c). For data simulated with $k_{\mathrm{syn}} = 300$ at the median fibroblast shape, BPSC's fitting routine returns synthesis rates low by a factor of 2–3, with the other parameters shifted to keep the mean count unchanged. Because its binned fitting objective differs from maxi-

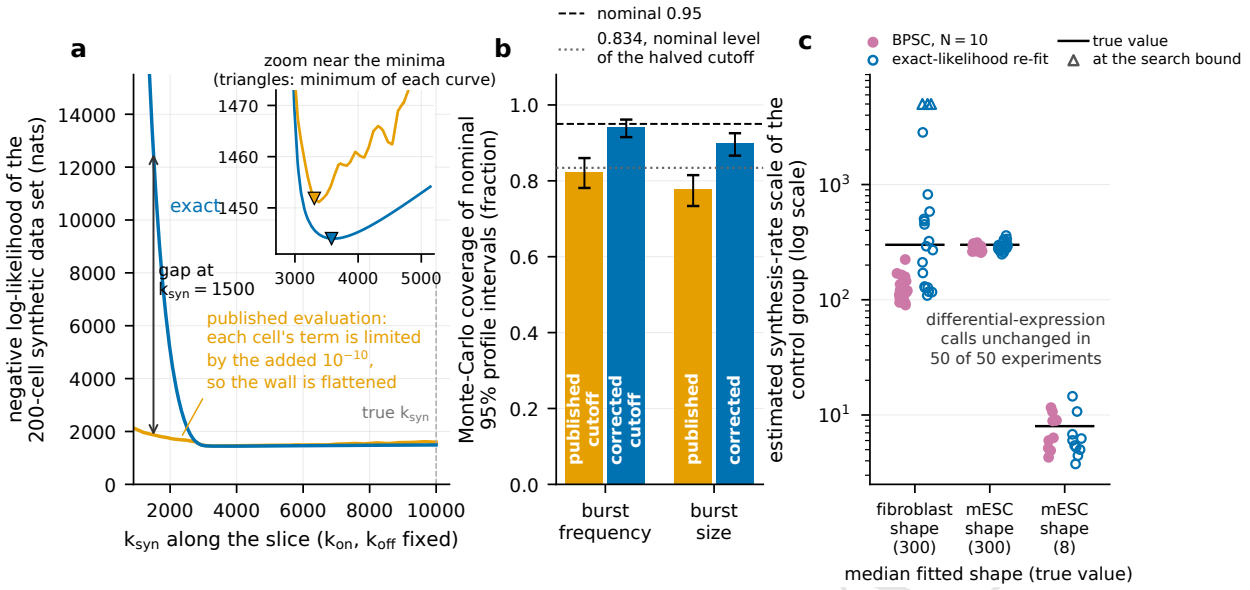

Figure 2: On synthetic data with known parameters the published evaluation changes likelihood values and interval coverage but hardly moves txburst's rate estimates or alters BPSC's test decisions. (a) Negative log-likelihood along k_{syn} of a synthetic 200-cell data set with known parameters $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}}) = (0.43, 5.7, 10^4)$: the added 10^{-10} limits each cell's term to 23.0 nats, so at $k_{\mathrm{syn}} = 1{,}500$ the published value lies 10,625 nats below the exact one, yet the two minima (inset) differ by 0.6 conditional half-widths. (b) Monte Carlo coverage of nominal 95% profile intervals, 400 sets of 200 cells at two parameter points where the quadrature is accurate: burst frequency and size 0.824 and 0.777 at the published cutoff, 0.943 and 0.900 at the corrected cutoff, the burst-size residual being a 200-cell ridge effect (per-point coverage in Supplementary Fig. S3). Halving the cutoff alone makes nominal 95% intervals 0.834 intervals (dotted line); large k_{syn} is not measured here. (c) BPSC's 10-node synthesis-rate estimate beside an exact-likelihood re-fit of the same cells, 50 simulated two-group experiments (a limited probe): null estimates at the median fibroblast shape span 90–170 for a true value of 300; 0 calls changed.

mum likelihood even for an error-free density, this bias is likely due mainly to the fitting procedure in that region and only partly to the quadrature error. However, BPSC's differential-expression test uses only the variance function of the fitted distribution and does not evaluate the density. In 60 simulated two-group experiments (the 50 of Fig. 2c and 10 more; a limited probe, Methods), replacing BPSC's fit by an exact-likelihood fit changed none of the 60 differential-expression calls at the 0.05 level. We did not test its goodness-of-fit and model-selection outputs, which do use the density.

Separately, txburst's profile-likelihood routine places each interval endpoint where twice the drop in log-likelihood from the maximum reaches 1.9207, half of the 95% chi-square quantile 3.8415 given by Wilks' theorem [35]. The halving is consistent with the textbook cutoff for $\Delta \log L$ applied to a statistic that is already $2\Delta \log L$. The implied asymptotic coverage of the nominal 95% intervals is

$$P(\chi_1^2 \le 1.9207) \;=\; \mathrm{erf}\,\sqrt{1.9207/2} \;=\; 0.8342, \tag{7}$$

where χ_1^2 denotes a chi-square variable with one degree of freedom. The nominal 95% intervals are therefore 83.4% intervals. Error bars labeled to miss the truth one time in twenty in fact miss it

about one time in six.

We also measured the coverage of the nominal 95% intervals directly, running txburst's profile-likelihood routine on 400 synthetic data sets at two parameter points where the quadrature is accurate (Methods). In these 200-cell data sets the intervals covered the true burst frequency in a fraction 0.824 of cases (0.777 for burst size, Fig. 2b). With only the cutoff corrected, the two coverages rose to 0.943 and 0.900. The remaining burst-size shortfall comes from the $(k_{\mathrm{off}}, k_{\mathrm{syn}})$ ridge [34] at 200 cells and from the under-converged inner optimizer of txburst's profile scan; the cutoff and the quadrature contribute none of it (Supplementary File S3).

Thus, quantities that depend only on where the likelihood attains its maximum, such as txburst's rate estimates and BPSC's differential-expression statistic, are nearly unaffected by the quadrature error and the added constant. Any quantity computed from the likelihood value itself is affected by the full error in that value. We measure both effects on the published fits of Larsson et al. (Fig. 3).

Reanalysis of the published fits

Of the 32,376 txburst fits with finite parameter values published by Larsson et al. [8], 205 have a log-likelihood at their own parameters that differs from the exact value by more than one nat (by up to 4,827 nats for one gene, Fig. 3a). Apart from this small group, the published values agree with ours to better than 10^{-3} nats for about 99% of the fits. The fits cover three allele-specific count tables, from primary mouse fibroblasts (C57 and CAST alleles) and mouse embryonic stem cells (mESCs; CAST allele). We compared four quantities, each against a threshold fixed before the re-evaluation (Methods).

In 200 of the fits differing by more than one nat, the published log-likelihood is the higher one, and txburst thus reports a better fit than the exact likelihood permits. Few of them lie beyond the quadrature onset (Eq. (6)). In 194 of them, instead, the fitted k_{syn} is below 300 and at least one cell carries a count to which the model assigns almost no probability. For such cells the added constant raises the computed log-likelihood above the exact value (Fig. 2a). The largest difference, for *Ywhae* in the embryonic stem cells, arises because all 94 of its cells carry such a count (color scale of Fig. 3a). Over all fits, by contrast, we find a median absolute difference of 2×10^{-7} nats, as expected from Fig. 1b for the bulk of the fitted region (Supplementary File S3).

When we re-fit on the exact likelihood surface with txburst's optimizer settings [8], 12 of the 20,201 genes whose published fits pass the original quality filter (0.059%) have burst-frequency or burst-size estimates that move by more than one conditional half-width (Fig. 3b). The published point estimates are in effect unchanged. The extreme case among these shifts is *Krt10* on the fibroblast CAST allele (Supplementary Fig. S2). For this gene the published likelihood surface has a second optimum at a positive log-likelihood, which requires per-cell probabilities above one, and this optimum is absent from the exact likelihood surface. Applied to the re-fits, the original quality filter gives a different call (in either direction) for 1,088 genes when one filter rule is applied on both sides (Supplementary File S3). Of the 1,384 changes counted against the calls as published, the change in likelihood value accounts for only six. The rest lie on the non-identifiable $(k_{\mathrm{off}}, k_{\mathrm{syn}})$ ridge [34] or at the edge of the optimizer's range, where the call is unstable under either likelihood (Supplementary File S3).

Profile intervals were published [8] for the 7,186 fibroblast genes on the CAST allele, of which we set aside nine whose endpoints were not reproduced in a partial re-run under the current SciPy (Methods). At the correct cutoff, 2,066 of the remaining 7,177 genes (28.8%) have a burst-frequency or burst-size endpoint that moves by more than 10% between the published likelihood and the exact likelihood, whose endpoints are those of the corrected per-gene tables described below. Of these genes, 309 move an endpoint by more than half (Fig. 3c).

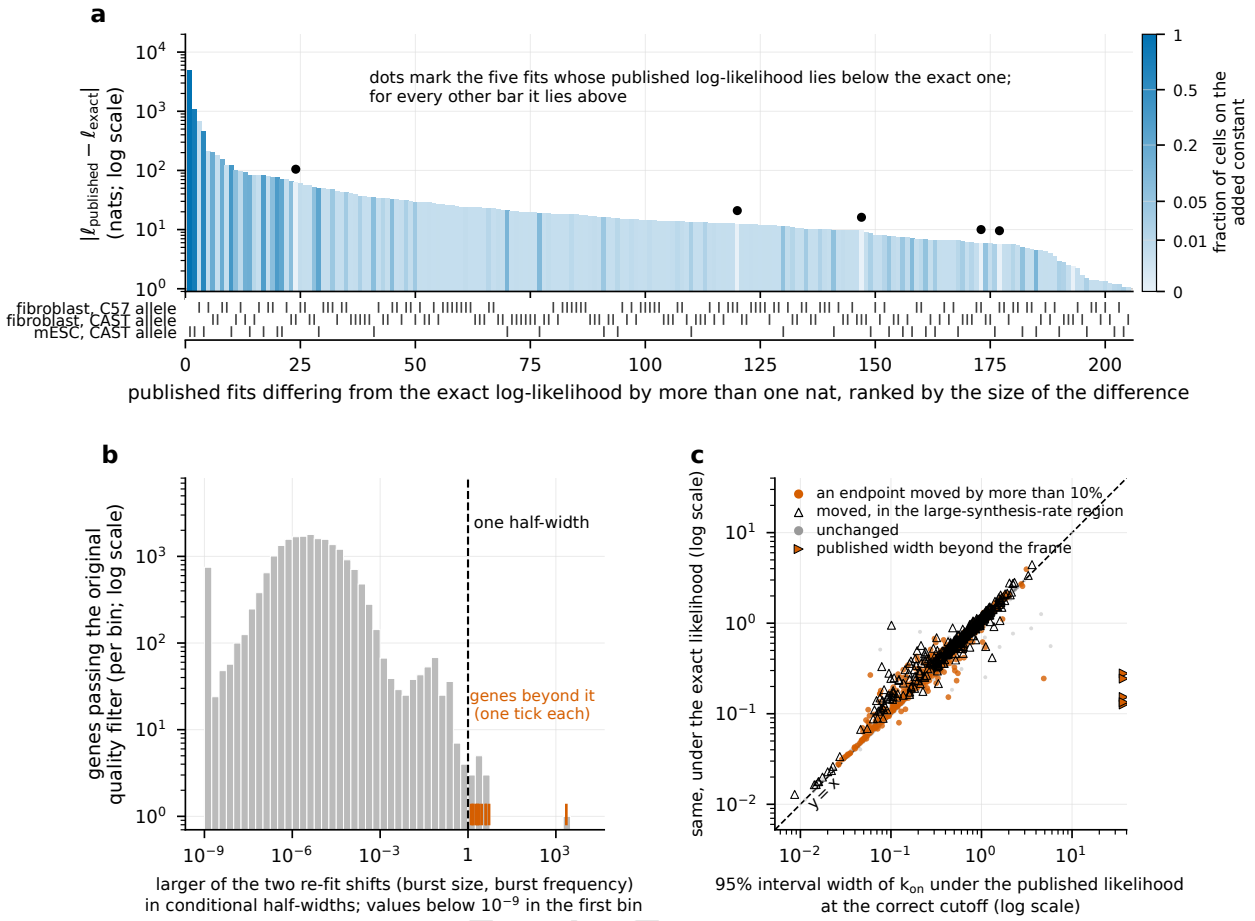

Figure 3: Of the 32,376 published fits of the three data set and allele combinations of Larsson et al. [8], 205 differ from the exact log-likelihood by more than one nat at their own parameters, while only 12 of 20,201 re-fitted estimates move by more than one conditional half-width. (a) The 205 fits ranked by the difference, evaluated at the published parameter values, up to 4,827 nats, each embryonic stem cell counted once. Color: fraction of cells whose published probability sits on the added 10^{-10} (nonzero in 199 fits); dots: the five fits whose published log-likelihood lies below the exact one. (b) Larger of the burst-size and burst-frequency re-fit shifts, published versus exact likelihood, for the 20,201 genes whose published fits pass the original quality filter, in conditional half-widths. (c) Width of the 95% interval of k_{on} under the published likelihood and under the exact likelihood with the endpoints of the corrected tables, for the fibroblast CAST set: 2,066 of 7,177 genes have a burst-frequency or burst-size endpoint moved by more than 10% at the correct cutoff (colored where drawn; genes with a truncated or undefined width on either side are not drawn). Triangles: moved genes inside the large-synthesis-rate region; most moved genes lie outside it.

However, only 413 of these genes lie in the large-synthesis-rate region defined for this comparison (Methods). Correcting the cutoff alone widens the intervals by median factors of 1.437 to 1.452 (1.6% to 2.7% above $\sqrt{2} \approx 1.414$, the quadratic-profile value), and this small excess has the sign expected for ridge-shaped profiles. Thus the published intervals cannot be corrected by a single rescaling, because the correct cutoff must be applied to the exact-likelihood profile, which itself differs from the published profile for the 2,066 genes of Fig. 3c.

The per-gene tables (Data availability) list the published estimate and quality call, the exact

log-likelihood at the published and re-fitted parameters, and the re-fitted k_{on}, k_{off} and k_{syn} with burst frequency and size. Each row also gives profile intervals for both under the exact likelihood at the correct cutoff, with endpoint likelihoods evaluated in ball arithmetic. When we re-evaluated the 68,657 endpoints of these tables as first computed against the exact profile, 926 moved inward, by up to a factor of 16.2 (Supplementary File S3). For those the scan had placed the endpoint too far out. Rows are marked where the nominal calibration does not apply: a parameter at an optimizer bound, an endpoint beyond the scanned range, a fit on the non-identifiable ridge, a disconnected confidence set or a nuisance optimum at the edge of the optimizer box (Supplementary File S3).

In all, 69.6% of the rows are unmarked. Whether any biological conclusion of Larsson et al. [8] is affected by these numerical differences was not tested.

The effect of regulatory DNA

Among the 77 clean hits in the fibroblast data of Johnsson et al. [32], the one-degree tests resolve 19: 17 are frequency-resolved calls, 2 are resolved on both axes, and none is a size-resolved call (Fig. 4a; Table 2). These counts are outcomes of tests with unequal power, 0.74 for a 2-fold difference in burst frequency against 0.36 for the same difference in burst size, and they do not measure how often each mechanism occurs (Methods). Jiang et al. concluded from SCALE analyses of mouse blastocyst and human fibroblast cells that "cis control in gene expression overwhelmingly manifests as differences in burst frequency" [16]. Larsson et al. attributed control of burst frequency mainly to enhancers and of burst size to core promoter elements [8]. If so, enhancer sequence differences between the strains would appear as allelic differences in burst frequency. The allele-resolved studies by Larsson et al. of the C57BL/6J × CAST/EiJ cross that we also analyze [8, 17] are the closest prior work to ours. The calls reported here do not test these statements.

The data set of Johnsson et al. [32] (ArrayExpress E-MTAB-11054) contains 682 Smart-seq3 fibroblasts from one female F1 animal of that cross. In these cells we tested a stratified random sample of 198 of the 13,123 autosomal genes that met the coverage criteria, a family size fixed in the analysis plan before any fitting (Methods). We adopted this data set after a pilot in which the same tests on the allelic matrices of Larsson et al. (224 and 94 cells) did not meet our calibration criteria (Methods and Supplementary File S4).

For each gene the primary test is a joint likelihood-ratio contrast of the three telegraph parameters between the alleles (three degrees of freedom), computed with exact likelihoods. One-degree-of-freedom contrasts of burst size and of burst frequency are secondary tests, and the false discovery rate is controlled at 5% separately for each of the three tests (Benjamini–Hochberg [36]). We call a gene that is significant on the joint test size-resolved if only the burst-size contrast is also significant, frequency-resolved if only the burst-frequency contrast is, resolved on both if both are, and unresolved if neither is. The call is an outcome of tests whose power depends on the gene's expression.

The analysis plan, including three calibration criteria, was specified before the allelic test results were unblinded (Methods). All 20 imprinted genes used as positive controls were recovered in the expected direction. The P values of the three statistics were uniform on 500 data sets simulated under the null (Kolmogorov–Smirnov test, Fig. 4b). Median power at observed expression for a 2-fold difference is 0.74 in burst frequency and 0.36 in burst size (Fig. 4c). Therefore the burst-size contrast fails the criterion of 0.5. Its first implementation met the criterion at 0.87 because the constrained fit had stopped at its starting point for most genes, and the repaired statistic is the one reported (Methods and Supplementary File S4).

By the joint test, 104 of the 198 genes differ in burst kinetics between the alleles (Table 2; per-gene catalog in Supplementary File S4). The clean set comprises the 77 of these genes with

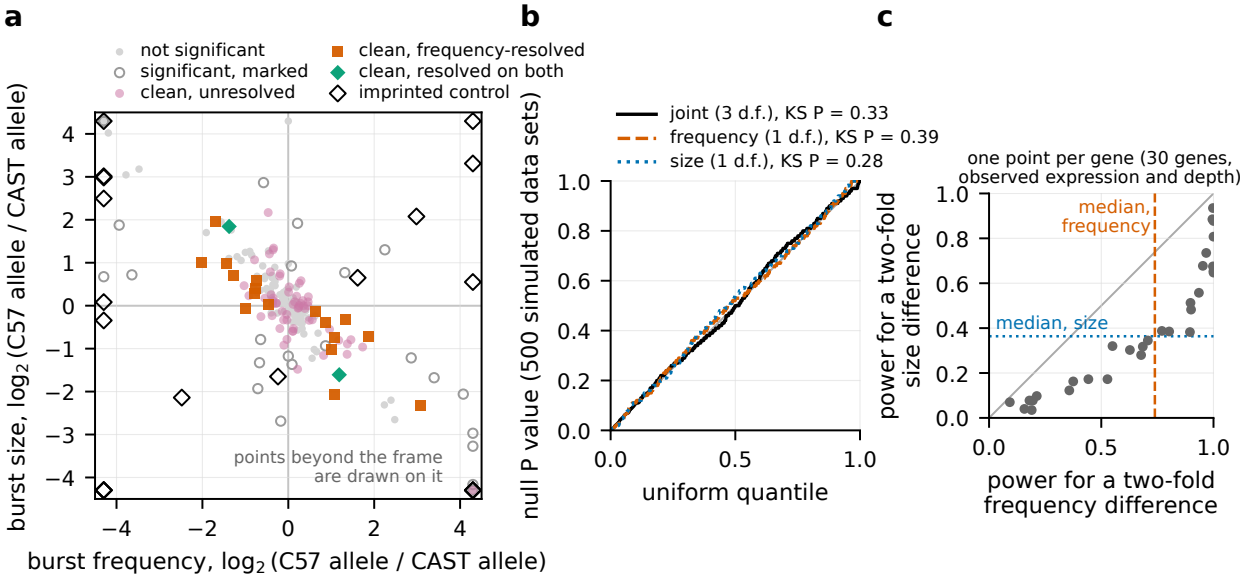

Figure 4: Among the 19 clean genes the one-degree tests resolve, 17 differ between alleles in burst frequency, 2 in both parameters and none in burst size alone; the burst-size contrast has less than half the power of the burst-frequency contrast, and the split reflects that difference in power. (a) Allelic $\log_2$ ratio (C57 over CAST allele) of burst size against that of burst frequency for the 198 autosomal genes tested in 682 fibroblasts of Johnsson et al. [32]. The 77 clean significant genes (no artifact mark, balanced counts; Methods) are colored by the call the one-degree tests return (size-resolved, frequency-resolved, both, unresolved), 58 of them unresolved and none size-resolved; marked and non-significant genes gray; the 20 imprinted controls, all recovered, diamonds; points beyond the axis range are drawn at the frame edge. Each point contrasts telegraph parameters between the alleles of one single-direction cross; strain and parent-of-origin effects are not distinguished. (b) Null P values of the joint, frequency and size statistics on 500 data sets simulated with equal alleles against uniform quantiles (Kolmogorov–Smirnov $P = 0.33$, 0.39 and 0.28). (c) Simulated power per gene for a 2-fold difference at observed expression; medians 0.74 (frequency) and 0.36 (size).

none of the four artifact marks (boundary fit, low expression, possible mapping dropout, or a likelihood ridge) and with total-count allelic balance inside the accepted band (Methods). We added this balance criterion post hoc, after re-fits showed that the largest apparent allelic effects were artifacts of read mapping or of an allele nearly absent from the data (Supplementary File S4). Among the resolved clean genes the allelic effects are modest, with a median unsigned ratio between the alleles of about 2.1-fold in the affected parameter (Fig. 4a; $|\log_2|$ ratio 1.07, per-class medians in Table 2). The 58 clean genes that remain unresolved have small or mixed allelic effects, or effects the one-degree tests lack the power to attribute.

The split between frequency-resolved and size-resolved calls says little about the balance of mechanisms. In a simulation of the whole procedure with planted allelic differences (Methods and Supplementary File S4), a planted burst-frequency difference is returned as a frequency-resolved clean call in a fraction 0.142 of cases and a planted burst-size difference as a size-resolved clean call in 0.034. An equal mixture of the two yields a split at least as uneven as ours in a fraction 0.79 of experiments (Wilson 95% interval 0.52 to 0.92). A difference in k_{off} alone, which changes burst size and burst frequency together, is returned as size-resolved in none of 390 planted cases. Hence we

Table 2: Counts and calibration of the allele-resolved analysis on 198 autosomal genes in 682 fibroblasts. A gene is significant when the joint three-degree-of-freedom likelihood-ratio test of equal kinetics on the two alleles rejects at a Benjamini–Hochberg level of 5%; it is clean when its fit carries no artifact mark and its total-count allelic balance lies in $[0.3, 0.7]$, a criterion adopted after unblinding (Methods). The one-degree-of-freedom tests return, for each clean difference, a size-resolved, frequency-resolved, both or unresolved call; these are outcomes of tests with unequal power (frequency 0.74, size 0.36 for a 2-fold difference), and the expected calls under an equal planted mixture come from a simulation of the whole procedure (Supplementary File S4). Effects are unsigned allelic ratios of model parameters between the two alleles of one single-direction C57BL/6J×CAST/EiJ cross. Every likelihood entering a test was evaluated exactly; zero evaluations failed their error bound, and at identical parameters the double-precision search statistics agree with the exact statistics to 1.3×10^{-7} nats. 13 nested statistics were negative by up to 0.98 because the free per-allele fits stopped at least 0.55 nats short of their optimum, and were set to zero (all with $P = 1$).

Genes	
tested (autosomal, 682 cells)	198
significant, joint test	104
clean	77
marked, plausible	15
marked, likely artifact	12
clean, resolved by the one-degree tests	19
size-resolved	0
frequency-resolved	17
resolved on both	2
clean, unresolved	58
expected clean calls under an equal planted mixture, size / frequency	0.33 / 6.67
median $\lvert \log_2 \rvert$ effect, clean resolved (about 2.1-fold)	1.07
size / frequency / both	– / 1.07 / 1.73
unresolved (larger of the two ratios)	0.53
Calibration	
imprinted positive controls recovered	20 of 20
null P values against the uniform on 500 simulated sets, Kolmogorov–Smirnov P (joint / frequency / size)	0.33 / 0.39 / 0.28
power at a 2-fold difference, burst frequency / burst size	0.74 / 0.36

make no statement about whether regulatory differences act more often on burst size or on burst frequency.

In Larsson et al.'s own transcriptome-wide comparison of the C57 and CAST alleles in fibroblasts, similar numbers of genes differed in burst frequency (307) and in burst size (276) [8]. That analysis used the published txburst implementation examined in Fig. 3, so every gene call in that comparison carried unmeasured error. The disagreement between the fibroblast results and the SCALE result may reflect differences in species, cell type, platform and statistical test, and we do not attempt to resolve it here. An allele-specific capture or mapping ratio ρ multiplies the inferred burst-size ratio by ρ exactly (binomial thinning, Methods), and neither of the 2 clean genes significant in the burst-size family keeps its call at a ratio 10% above or below unity (Supplementary File S4).

Each entry in the resulting catalog (Table 2, Supplementary File S4) is an allelic difference in parameters of the stationary telegraph model, inferred from the published count matrices rather

than measured directly. With a single-direction cross, an allelic difference may be a cis effect of sequence differences between the strains or a parent-of-origin effect such as imprinting, and these data cannot distinguish the two. Genes with signs of mapping dropout are left out of the clean set, but residual per-gene mapping bias, including reference bias at SNP positions, cannot be excluded (genome-wide allelic balance in Supplementary File S4). The bootstrap treats the two alleles of a cell as independent. A within-cell label swap of the joint contrast, exact when the two alleles are exchangeable within a cell, keeps 90 hits at the same false-discovery level and drops 14, 12 of them clean and one of them resolved on both axes (Supplementary File S4). For these reasons we do not assign a specific mechanism to any individual gene. Finally, with cells from one animal and one cell type, the analysis has no biological replicate.

Exact likelihoods for refractory promoters

The stationary count distribution of a promoter with a two-step OFF period is a generalized hypergeometric function (Eq. (8)), and we evaluate its log-likelihood for 200 cells in ball arithmetic to within 5.1×10^{-44} nats in 0.004 s. In live-cell recordings many promoters are refractory, passing through more than one rate-limiting step before firing again [3, 37]. Hypergeometric stationary solutions for multistate promoters go back to Zhou and Zhang [38]. Cao et al. derived the stationary distribution for two sequential OFF states as a ${}_2F_2$ function and used it to compare count distributions between promoter models [39]. For their generalized telegraph model, Luo et al. found approximating the count distribution "computationally prohibitive, making it impossible to evaluate the likelihood function directly" and used approximate Bayesian computation [18]. Among inference tools for sequencing counts and works citing these solutions, we found no likelihood evaluation with an error bound for more than one OFF state (Supplementary File S1).

The three-state model we evaluate has one active state and a Markovian two-step OFF state, $\mathrm{ON} \xrightarrow{k_{\mathrm{off}}} \mathrm{OFF}_2 \xrightarrow{k_{12}} \mathrm{OFF}_1 \xrightarrow{k_{\mathrm{on}}} \mathrm{ON}$, in which the OFF dwell is the sum of two exponential stages (Fig. 5a). Synthesis proceeds at rate ν in the ON state, with all rates in units of the mRNA degradation rate. The model's stationary count distribution $P(k)$, rederived from the master equation (Supplementary File S1), is

$$P(k) \;=\; \frac{\nu^k\,(a_1)_k\,(a_2)_k}{k!\,(b_1)_k\,(b_2)_k}\;{}_2F_2(a_1{+}k,\; a_2{+}k;\; b_1{+}k,\; b_2{+}k;\; -\nu)\,, \tag{8}$$

where $(x)_k$ is the rising factorial, $a_{1,2} = \{k_{12}, k_{\mathrm{on}}\}$ are the OFF-state exit rates, and $b_{1,2}$, possibly complex conjugate, are fixed by $b_1{+}b_2 = k_{\mathrm{on}}{+}k_{12}{+}k_{\mathrm{off}}$ and $b_1 b_2 = k_{\mathrm{on}}k_{12}{+}k_{\mathrm{off}}(k_{\mathrm{on}}{+}k_{12})$ (Methods).

The probabilities $P(k)$ obey a recurrence of order three,

$$\begin{aligned}(K{+}1)(K{+}2)(K{+}3)\,P(K{+}3) \;=\;& (K{+}1)(K{+}2)(2K + E_1^b + \nu + 1)\,P(K{+}2)\\ &- (K{+}1)\big[K^2 + K(E_1^b + 2\nu) + \nu(E_1^a + 1) + E_2^b\big]\,P(K{+}1)\\ &+ \nu\big(K^2 + E_1^a K + E_2^a\big)\,P(K),\end{aligned} \tag{9}$$

where $E_{1,2}^a$ and $E_{1,2}^b$ denote the elementary symmetric functions of the a's and b's. We prove the recurrence by a telescoping identity and show that no lower-order recurrence with polynomial coefficients exists up to the degrees we searched (Supplementary File S1). Like Eq. (5), the recurrence is stable only when run downward in K (Supplementary File S1). Luo et al. report about 75 s per gene for approximate Bayesian inference of this model class on their hardware [18]. In that time the exact 200-cell likelihood at their demonstration parameters can be evaluated about 2×10^4 times on our hardware (0.004 s each).

When counts are drawn from a three-state distribution, the expected log-likelihood ratio in favor of the extra state is $n\,\mathrm{KL}$, the number of cells n times the Kullback–Leibler divergence from that distribution to the nearest two-state distribution. At the demonstration parameters of Luo et al.'s first example in Markov form (Methods), 200 cells carry only $n\,\mathrm{KL} \approx 0.1$ (Fig. 5a), which is too little for any method to recover the second OFF step from stationary counts. Indeed, the exact three-state maximum-likelihood fit to 200 simulated cells converges to the two-state boundary, where the OFF_2 stage is effectively instantaneous, and the likelihood-based model comparison favors two states ($\Delta\mathrm{AIC} = +2$). However, with 20,000 cells at an identifiable parameter set, $n\,\mathrm{KL} \approx 48$ (Fig. 5a) and direct maximum likelihood recovers the second OFF step ($\Delta\mathrm{AIC} = -64.4$). With the exact likelihood we compute $n\,\mathrm{KL}$ for each gene in the data of Larsson et al. [8] at its own cell count and fitted parameters (Fig. 5b).

The limits of snapshot data

Across the 20,216 genes evaluated in the data of Larsson et al. [8], the largest separation between a two-step and a single-step OFF state at a gene's own cell count is $n\,\mathrm{KL} = 0.58$, for *Timp1* with 224 cells (Fig. 5b). The corresponding asymptotic detection power is at most 0.285, and the power simulated at the critical value of its own design is 0.222 (Methods). Stationary snapshots are known to discriminate promoter architectures poorly [41, 42]; Braichenko et al. found telegraph-like and mechanistic models "practically indistinguishable by comparison of their transcript numbers distributions" in a large parameter region [43].

For three-state parameters θ we measure the distance from the two-state family by

$$\mathrm{KL}(\theta) = \min_{\phi} \mathrm{KL}(P_3(\theta) \,\|\, P_2(\phi))\,, \tag{10}$$

where $P_3(\theta)$ and $P_2(\phi)$ are the count distributions of Eqs. (8) and (2) and the minimizing ϕ defines the nearest two-state distribution. Because the two-state family lies on the boundary of the three-state family, the likelihood-ratio null is asymptotically an even mixture of χ^2_0 and χ^2_1 [44] (Methods), and the asymptotic detection power is 0.64 at $n\,\mathrm{KL} = 2$ and 0.99 at $n\,\mathrm{KL} = 8$. At these cell numbers the mixture is approached slowly: the 5% critical value simulated at nine two-state designs lies between 0.59 and 3.03 against the asymptotic 2.706 (Supplementary File S5), and the power values quoted here are asymptotic.

We computed $n\,\mathrm{KL}$ for each of the 20,217 genes whose exact re-fit passes the original quality filter (median 206 cells per gene), except one that could not be evaluated (Supplementary File S5). We divided each gene's fitted mean OFF dwell into two exponential stages in varying proportions; the equal division gives the largest separation, and all evaluated genes stay below $n\,\mathrm{KL} = 2$ at their own cell counts (Fig. 5b). The median separation is $n\,\mathrm{KL} = 0.0069$, at which the power to detect a second OFF step is close to the test's false-positive rate.

Over the 95% likelihood region of the *Timp1* two-state fit the separation ranges from $n\,\mathrm{KL} = 0.45$ to 0.70. No test of any form at the 5% level has power above 0.64 anywhere in that region (Pinsker's inequality, Methods). Thus, at the fitted parameters and across their confidence region, no gene in this data set was measured in enough cells to reveal whether its promoter needs recovery time between bursts. Even at the ninetieth-percentile fibroblast parameters (Methods) a gene would need more than 20,000 cells to reach $n\,\mathrm{KL} = 8$ (Supplementary Fig. S5).

From Eq. (10), the number of cells N needed for $n\,\mathrm{KL} = 8$ is

$$N \simeq \frac{8}{\mathrm{KL}(\theta)}, \tag{11}$$

and $N \simeq 4.3/\mathrm{KL}$ gives power 0.9. Capture at a fixed efficiency β enters exactly as binomial thinning, $\nu \to \beta\nu$ (Methods); the fitted rates are rates of captured molecules, $\nu_{\mathrm{fit}} = \beta_0\nu$, with β_0 the capture efficiency of the experiment, which is not known for these libraries. At the experiment's own capture and the most favorable two-step division, a gene with the median fibroblast parameters needs $N = 1.2 \times 10^5$ cells (Fig. 5c). With every transcript counted, the same gene would need between 1.0×10^4 and 4.3×10^4 cells for β_0 between 0.1 and 0.5. Thinning the observed counts to 0.3 and to the droplet-like 0.06 of the experiment's capture raises the requirement to 1.8×10^6 and 3.9×10^8 cells (Supplementary Fig. S6). A droplet experiment at the absolute mean efficiency 0.06 of Klein et al. [40] and Tang et al. [23] would need between 3.2×10^5 and 3.2×10^7 cells over the same bracket. A gene with the identifiable parameter set of Fig. 5a would reach $n\,\mathrm{KL} = 2$ at 832 cells and power 0.9 near 1,800 cells. Zhu et al. [45] estimated that about 10^5 cells of stationary counts separate a three-state model with polymerase dynamics from the two-state model, the same order of magnitude as our best case; their third state subdivides the active period and ours the OFF period, and the cell numbers do not transfer.

Our cell numbers bound the two-step case only; OFF periods with more stages, whose dwell-time distributions are more sharply peaked (5–7 rate-limiting steps have been inferred from live-cell traces [37]), are not bounded. Jiao et al. [46] showed model mimicry [47] between the two- and three-state models on 625 simulated parameter sets, and Cao et al. [39] mapped a parameter plane on which the two distributions do separate.

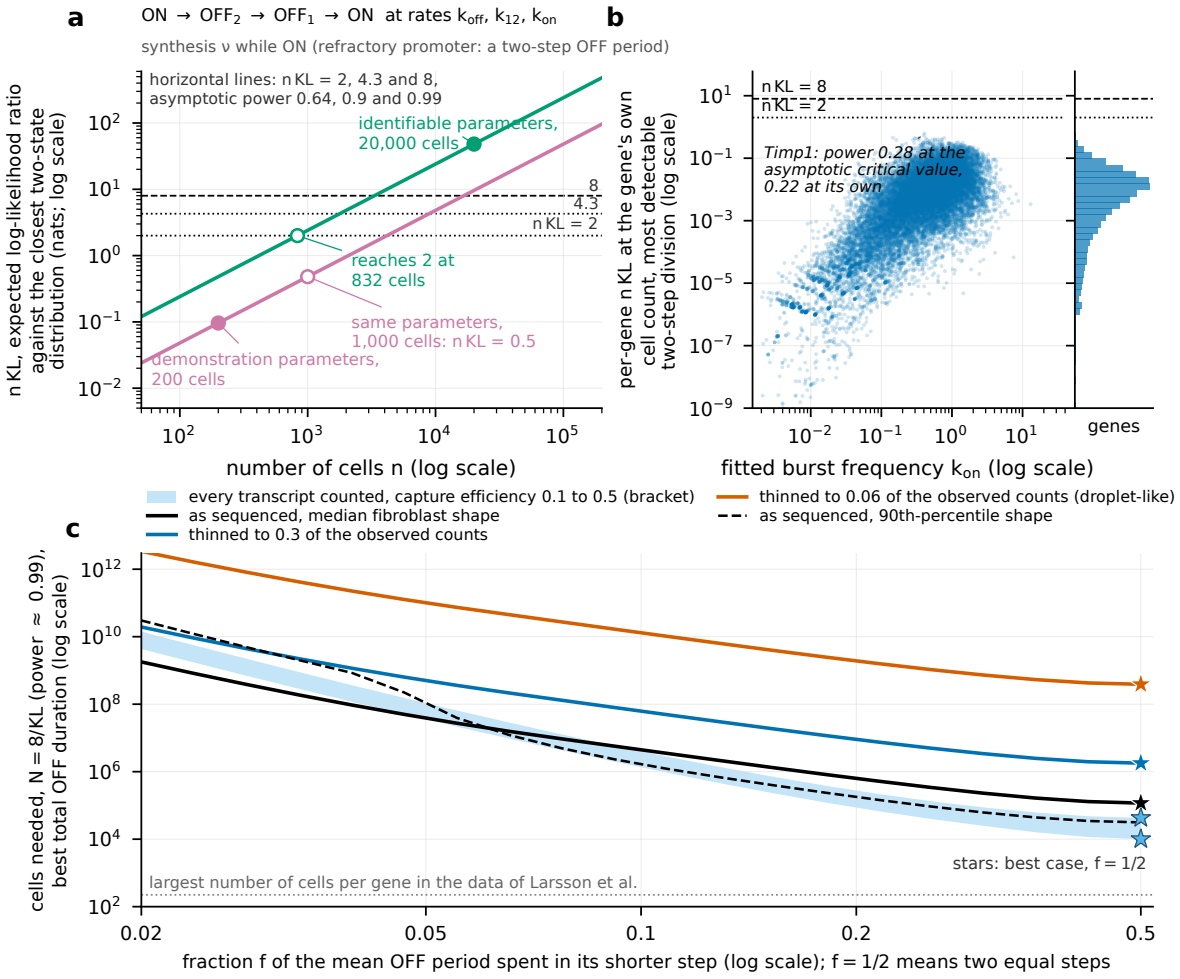

Figure 5: At the cell numbers of Larsson et al. [8] no gene reaches the separation needed to detect a two-step OFF period, and a gene with the median fitted parameters would need 1.2×10^5 cells at the capture efficiency of the experiment and of order 10^4 (1.0×10^4 to 4.3×10^4) if every transcript were counted. (a) Expected log-likelihood ratio n KL of counts from a two-step model against the closest two-state distribution: demonstration parameters (0.1 at 200 cells; open circle, 0.5 at 1,000 cells) and identifiable parameter set (48 at 20,000; open circle, n KL = 2 at 832 cells). Horizontal lines: separations n KL = 2, 4.3 and 8, at which the asymptotic boundary-calibrated power is 0.64, 0.9 and 0.99 (Methods); the 5% critical values simulated at the designs of Supplementary File S5 range from 0.59 to 3.03 against the asymptotic 2.706, and the power values are asymptotic. (b) Largest n KL over two-step divisions of the fitted OFF period at each gene's own cell count, 20,216 fits; maximum 0.58 (*Timp1*, 224 cells; asymptotic power 0.285; simulated 0.283 at the asymptotic critical value and 0.222 at the critical value simulated for its design). Values are conditional on each gene's fitted two-state distribution (equal split of the fitted OFF period). (c) Cells needed to detect a two-step OFF period at the median fibroblast shape, $N = 8/\mathrm{KL}$, against the OFF-period division f, at the capture of the experiment (as sequenced) and after further thinning to 0.3 and to the droplet-like 0.06 of it [40]; the shaded band is the same gene with every transcript counted, for an experimental capture efficiency between 0.1 and 0.5. Values condition on the fitted two-state distribution with capture as binomial thinning only and bound the two-step case only; deeper OFF cascades are not bounded, and real requirements are larger.

Discussion

We have shown that the telegraph likelihood can be evaluated with a guaranteed error bound, fast enough for the exact value to serve as the reference in all fits, tests and intervals (Table 1). Against that reference, a fixed quadrature rule loses accuracy at large synthesis rates, and the constant added inside the logarithm flattens the likelihood for cells whose counts are improbable under the fitted distribution (Fig. 2a). On re-evaluation nearly all point estimates of Larsson et al. [8] change by less than their uncertainty, whereas likelihood values change for a small minority of genes with large synthesis rates or such improbable counts (Fig. 3). Interval endpoints move for more genes than that, and txburst's halved cutoff makes all published intervals too narrow (Fig. 2b).

In deeply sequenced fibroblasts of the same C57BL/6J × CAST/EiJ cross [32], per-gene likelihood-ratio tests give a catalog of allelic differences in burst kinetics (Fig. 4 and Table 2). Where the one-degree tests resolve a clean hit, the difference lies in burst frequency or in both parameters and never in burst size alone. The tests cannot say whether regulatory DNA changes burst size as often as burst frequency: the burst-size contrast has less than half the power of the burst-frequency contrast, and an equal mixture of the two yields a split at least as uneven as this one in a fraction 0.79 of simulated experiments. Finally, no gene in the data of Larsson et al. has enough cells to detect a two-step OFF state, and the number needed follows from each gene's fitted parameters (Fig. 5).

Which published results should be recomputed depends on whether an analysis uses the likelihood value or only the location of its maximum. Point estimates from txburst and gene rankings built on them (Fig. 3b), and the differential-expression calls of BPSC [14], which use only the fitted variance function (Fig. 2c), can be taken as published. Likelihood-ratio tests, AIC-type model comparisons and goodness-of-fit statistics use the value itself. We recommend recomputing them with the exact likelihood for genes with k_{syn} beyond the quadrature onset or with a cell whose log-likelihood term is set by the added constant.

A developer can choose the number of nodes from the onset relation, Eq. (6), since a fixed N-node rule is adequate while $k_{\mathrm{syn}} \lesssim (N/\pi)^2$. Chen et al. [48] recommended Gauss–Jacobi quadrature after synthetic tests at synthesis rates in this range. A tool that fits with such a rule could state that range and, outside it, decline to fit or add nodes until $N \gtrsim \pi\sqrt{k_{\mathrm{syn}}}$. For profile intervals the full cutoff, a one-line change to txburst's profile-likelihood routine, should be applied to the exact-likelihood profile, as in our corrected per-gene tables. The inner optimizer of the scan must also converge along the $(k_{\mathrm{off}}, k_{\mathrm{syn}})$ ridge, where txburst's chained scan returns burst-size intervals narrower than the exact profile (Supplementary File S3).

Our approach differs from earlier inference methods in three ways. First, we evaluate the likelihood value with a guaranteed error bound, whereas txburst [8] and BPSC [14] use a fixed quadrature rule and Kim and Marioni [13] used Gibbs sampling. Direct two-state evaluation without such a bound is available elsewhere [19, 25, 26]. Second, SCALE [16] estimates the kinetics from moments after removing technical noise with a spike-in model, whereas our likelihood has no technical-noise component. Third, for promoters with a two-step OFF state we evaluate the likelihood directly, whereas Luo et al. [18] used approximate Bayesian computation. They pooled the two fibroblast alleles of Larsson et al. into one quality-filtered matrix of 413 cells and fitted their generalized model to it genome-wide. At their own demonstration parameters, that many cells cannot constrain the second OFF step by any method, since the expected log-likelihood ratio in its favor is below half a nat (Fig. 5a).

Our analysis has four limitations. First, like the implementations we re-evaluate, the model has no measurement layer: counts enter as molecules, with no capture efficiency, ambient RNA or depth variation, so every comparison is like-for-like under the tools' own model, and absolute

rates inherit that model's realism [23, 49, 50]. Uniform capture at a known efficiency only rescales the synthesis rate (Methods). Modeling cell-specific efficiencies remains future work. Second, cells are treated as independent draws from one stationary distribution, with no cell-cycle, cell-size or other extrinsic variation, and the misspecification bias that Grima and Esmenjaud [51] quantified for such variation persists under exact evaluation. In the allelic tests the two alleles of a cell are treated as independent as well. Their depth-adjusted within-cell correlation among the clean hits is small (median partial Spearman -0.042) but significant for a fraction 0.66 of them, and the label-swap test, exact under exchangeability, removes 14 hits (Supplementary File S4). Third, rates are known only relative to mRNA degradation. Fourth, the detectability bound for refractory promoters covers a two-step OFF state, and deeper cascades [37] are not bounded by it. Live-cell imaging [3, 37] or snapshot samples of at least the size computed here (Fig. 5c) could detect a second OFF step in individual genes. Our re-evaluation concerns the numerics and statistics of the telegraph model as implemented in txburst and BPSC at the versions given in Methods. Whether any biological conclusion of Larsson et al. changes was not tested.

We also applied the per-gene likelihood-ratio tests to published cell-state comparisons in budding yeast, *Arabidopsis* and *Hydra* (Supplementary File S6). In yeast, stress appears to change mainly burst frequency, whereas in *Arabidopsis* and *Hydra* neither size nor frequency dominates. Because low capture inflates the estimated share of frequency effects in these two droplet data sets, we draw no cross-kingdom conclusion.

We regard this paper as the first step of a broader program rather than a finished reanalysis: with the likelihood evaluated exactly, the expected log-likelihood ratio between competing models gives the number of cells an experiment needs to distinguish them. In particular, the per-gene separations of Fig. 5b indicate for which genes a second OFF step could be detected with the fewest additional cells. Higher-order recurrences of the same form, not implemented here, may extend the evaluation to deeper OFF cascades, and finite-sample power calculations can replace the asymptotic relation we used. For the catalog, a reciprocal cross would separate strain from parent-of-origin effects and a second tissue would provide replication. The corrected tables cover only the data of Larsson et al., and profile intervals from txburst's routine in other studies need the same recomputation.

Conclusions

Exact evaluation of the telegraph likelihood is fast enough to be the default in analyses that use likelihood values, tests or intervals. Indeed, for the published fits of Larsson et al. [8] our reanalysis confirms the point estimates, finds different likelihood values for a small subset of genes and corrects all profile intervals. In deeply sequenced fibroblasts of the same cross, 104 of the genes tested differ between alleles in burst kinetics. The 19 resolved clean differences lie in burst frequency (17) or in both parameters (2), never in burst size alone. Detecting a two-step OFF state in a typical gene from snapshot counts requires on the order of 10^5 cells as sequenced and 10^4 with every transcript counted, more than any data set analyzed here contains. The evaluator accompanies the article, with reference values and corrected per-gene tables, and computes the telegraph likelihood with a guaranteed error bound for other single-cell count data.

Methods

Data sets and published estimates

The allele-resolved UMI count tables of Larsson et al. [8] and the published per-gene fits (point estimates and quality calls, with profile intervals for the fibroblast CAST table) were taken from the txburst repository (Supplementary File S7).

We re-evaluated three combinations of data set and allele: primary fibroblasts on the C57 and on the CAST allele (224 cells and 10,727 genes in each table) and embryonic stem cells on the CAST allele (94 cells, 10,927 genes). The CAST-allele embryonic-stem-cell count file in the repository lists each of its 94 cells twice (188 columns); a later commit corrected only the C57 file, and the published embryonic-stem-cell fits were produced on the doubled matrix. We examined those fits as published; all exact evaluations, re-fits, intervals and per-gene separations reported here use each cell once.

The tests for allelic differences used the count matrices of Johnsson et al. [32] (ArrayExpress E-MTAB-11054): 682 Smart-seq3 fibroblasts from one female offspring of a C57BL/6J × CAST/EiJ cross.

Model conventions and closed forms

Rates are expressed in units of the mRNA degradation rate δ, and burst frequency and burst size denote k_{on} and $k_{\mathrm{syn}}/k_{\mathrm{off}}$ as in Larsson et al. [8]. The derivations and the software use a Beta-mixture shorthand, related to the rates and to the Kummer arguments of Eq. (3) at count n by

$$(a,\, b,\, r) \;=\; \frac{1}{\delta}\,(k_{\mathrm{on}},\, k_{\mathrm{off}},\, k_{\mathrm{syn}}), \qquad \alpha = a + n, \qquad \gamma = a + b + n. \tag{12}$$

The txburst routines take (a, b, r) in this order and in these units, and the activation, inactivation and synthesis rates of Peccoud and Ycart [10], divided by δ, map onto (a, b, r) in that order; Supplementary Table S1 lists the conventions of other implementations.

Kummer's transformation [33] relates the two forms of Eq. (3),

$${}_1F_1(\alpha;\gamma;-r) \;=\; e^{-r}\,{}_1F_1(\gamma-\alpha;\gamma;r), \qquad \gamma-\alpha = b, \tag{13}$$

and yields the second form, summed in all exact evaluations reported here, in which the numerator parameter b does not depend on n and all terms are positive for $a, b, r > 0$. The closed form agrees with high-precision quadrature of the mixture integral to at least 40 digits and with the stationary solution of the truncated master equation [21] to about 10^{-14} at two parameter points (Supplementary File S1).

Series, recurrence and ball-arithmetic evaluation

We call an evaluation exact when it is carried out in ball arithmetic [29] and the returned radius is less than 10^{-25} of the value. All reported likelihood values, including those at interval endpoints, meet this criterion.

For each gene the all-positive series of Eq. (13) is summed in ball arithmetic at the two largest counts required, $n_{\max}$ and $n_{\max}+1$. Because $b < \gamma = a + b + n$, consecutive terms t_k obey a geometric bound, which encloses all terms from index K onward once $K + 1 > r$,

$$\frac{t_{k+1}}{t_k} \;=\; \frac{(b+k)\,r}{(\gamma+k)(k+1)} \;\le\; \frac{r}{k+1}, \qquad \sum_{k\ge K} t_k \;\le\; \frac{t_K}{1-q}, \quad q = \frac{r}{K+1} < 1. \tag{14}$$

The tail bound is added to the partial sum as a ball. The remaining probabilities follow from the recurrence Eq. (5), run downward in n from these two starting values; the second solution of the recurrence, built from the second Kummer function, grows relative to $P(n)$ as n increases. Thus, the probabilities are the minimal solution, which the downward pass computes stably (Supplementary File S1).

Ball radii grow along the recurrence because radius propagation takes absolute values; the working precision is set beforehand from a floating-point estimate of that growth, and a ball that fails the 10^{-25} criterion is re-evaluated at higher precision up to 4 times, after which no value is returned. A second, multiprecision implementation runs the same recurrence at two working precisions and requires agreement at every n.

We also tested the evaluator at 19 parameter points (random points inside txburst's parameter bounds, the edges and top of those bounds, and points where a general-purpose ${}_1F_1$ routine returns no value) against a 60-digit multiprecision value and an independent high-precision ball evaluation, with agreement to at least 57 digits (Supplementary File S2). The implementation uses Arb through python-flint, with mpmath and SymPy (versions in Supplementary File S7).

Floating-point surrogate and precision policy

During optimization and profile scans the all-positive series is summed in double precision, with the prefactor in log space and per-element rescaling, and each probability carries an a priori error bound. A probability vector is recomputed in ball arithmetic if that bound exceeds 10^{-12}, the series does not converge, a log-probability underflows the double-precision range, or any entry is non-finite.

Published implementations evaluated, benchmark grid and error metrics

We evaluated txburst [8] at commit 6b6525f and BPSC [14] at commit 5a512e5. The maximum-likelihood routine of txburst (`txburstML.py`) evaluates Eq. (2) with a 50-node Gauss–Jacobi rule and adds 10^{-10} to each probability inside the logarithm. Its profile-likelihood routine uses the same rule and constant and places interval endpoints where the profile statistic $2(\ell_{\max} - \ell)$ crosses `cutoff = chi2.ppf(1-alph, 1)/2`, half the χ^2_1 quantile [35]. Its hypothesis-test module uses 40 nodes and the same constant. The general-purpose routine of Table 1 is `hyp1f1` of SciPy 1.17.1 applied to the first form of Eq. (3) with a log-space prefactor. Each routine was copied without algorithmic change and reproduces the original output on the repository's example data (Supplementary File S7).

The benchmark grid has 228 points: a $6 \times 6 \times 6$ logarithmic grid over $k_{\text{on}} \in [0.03, 12.9]$, $k_{\text{off}} \in [0.3, 999]$ and $k_{\text{syn}} \in [2, 10^4]$ (216 points), plus 12 points drawn from the published fits. The error of a routine at a grid point is the maximum over n of the relative error of $P(n)$ against the exact value, whose radius is far smaller than any error reported (Table 1). The onset of an N-node rule is the first k_{syn}, on a geometric scan at fixed $(k_{\text{on}}, k_{\text{off}})$, at which that error exceeds 10^{-6}. Under the current SciPy, txburst cannot re-fit 46 genes, all failing the original quality filter, because its node construction overflows at their large published $(k_{\text{on}}, k_{\text{off}})$; they are excluded from the counts reported here (Supplementary File S3).

Re-evaluation design

The four quantities compared in the re-evaluation and their thresholds were specified before the data were analyzed (Supplementary File S3). The first is the difference between the published and the exact negative log-likelihood at the published parameters, called large when it exceeds one nat in magnitude. The second is the shift of the re-fitted burst frequency and burst size in conditional

half-widths of the exact likelihood, called large when it exceeds one. Conditional intervals (other parameters held fixed) are narrower than profile intervals. Hence, shifts in this unit can only be overstated. The third is a change of the published quality call on the exact re-fit, split into ridge-ambiguous (two optima within a tenth of a nat), range-truncated (both interval scans reaching the parameter bounds) and likelihood-driven cases.

The fourth quantity is the position of the profile-interval endpoints: the profile curve is computed under each likelihood L and read at both thresholds t,

$$\begin{aligned} \mathrm{CI}(L,t) \;&=\; \{\theta :\; 2\,[\ell_{\max}(L) - \ell_{\mathrm{prof}}(\theta; L)] \le t\}, \\ L \in \{L_{\mathrm{pub}},\, L_{\mathrm{exact}}\}, &\qquad t \in \{1.9207,\, 3.8415\}, \end{aligned} \tag{15}$$

where ℓ_{prof} is the profile log-likelihood under L and $\ell_{\max}(L)$ its maximum; fixing t isolates the likelihood effect and fixing L the cutoff effect, and an endpoint is called moved when it changes by more than 10%.

Re-fits used txburst's optimizer settings and the same multi-start scheme under both likelihoods, and the exact-likelihood profile scans stopped at $k_{\mathrm{syn}} = 10^4$, the top of the surrogate's tested range. We also re-ran the txburst profile routine on 200 genes under the current SciPy and excluded the nine whose published endpoints were not reproduced. The large-synthesis-rate region used in the Results is $k_{\mathrm{syn}} \ge 500$, burst size at or above the ninety-fifth percentile of the published fits in the gene's data set, or any parameter at an optimizer bound. Genes whose surrogate and exact log-likelihoods differed by more than 10^{-6} nats were also evaluated with the multiprecision implementation. The one amendment to the plan (Supplementary File S3) moved the summation of per-cell log-probabilities, which had lost precision on one subnormal probability, into ball arithmetic, and all reported values postdate that change.

Coverage and estimator simulations

Profile-interval coverage was measured on 400 synthetic 200-cell data sets, 200 at each of two parameter points where the quadrature is accurate, $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}}) = (1.0, 10.0, 100.0)$ and $(0.4, 5.0, 200.0)$. The txburst routine fitted and profiled each set at its default α and at the α for which its formula gives the χ^2_1 quantile, with undefined endpoints excluded and counted (Supplementary File S3). Estimator shifts in conditional half-widths were obtained by re-fitting, under both likelihoods, the four of the ten synthetic 200-cell sets whose published and exact log-likelihoods differed by more than a fifth of a nat (Supplementary File S2).

The BPSC comparison, a limited probe, used 50 two-group experiments of 200 cells per group, 25 without and 25 with a simulated difference, 20 of each at $k_{\mathrm{syn}} = 300$ beyond the onset of BPSC's rule and the rest at $k_{\mathrm{syn}} = 8$. Each was analyzed with BPSC's fit and with an exact-likelihood fit, and 10 more were simulated and analyzed with an independent implementation. Power at the gene with the largest separation was estimated from 2,000 data sets simulated at the *Timp1* fit (1,996 completed; Supplementary File S5), each fitted by two- and three-state maximum likelihood and tested at the boundary-calibrated critical value and at the critical value simulated at its own design.

Allele-resolved tests, calibration and class rule

The test family comprises 198 autosomal genes sampled at random, within strata of total UMI count, from those with at least 25 cells carrying non-missing allelic entries and 50 UMIs over both alleles. The X chromosome was excluded because all cells in these data express only the C57 X

(Supplementary File S4). Each allele was fitted separately by multi-start maximization of the exact telegraph likelihood within fixed parameter bounds.

Writing ℓ_{C} and ℓ_{A} for the C57 and CAST allelic log-likelihoods, the joint statistic is

$$T_3 \;=\; 2\Big[\max_{\theta,\theta'}\{\ell_{\mathrm{C}}(\theta)+\ell_{\mathrm{A}}(\theta')\} - \max_{\theta}\{\ell_{\mathrm{C}}(\theta)+\ell_{\mathrm{A}}(\theta)\}\Big], \tag{16}$$

and the one-degree statistics constrain one derived quantity q across alleles,

$$\begin{gathered} T_q \;=\; 2\Big[\max_{\theta,\theta'}\{\ell_{\mathrm{C}}(\theta)+\ell_{\mathrm{A}}(\theta')\} - \max_{q(\theta)=q(\theta')}\{\ell_{\mathrm{C}}(\theta)+\ell_{\mathrm{A}}(\theta')\}\Big], \\ q \in \{k_{\mathrm{syn}}/k_{\mathrm{off}},\; k_{\mathrm{on}}\}, \end{gathered} \tag{17}$$

namely $q = k_{\mathrm{syn}}/k_{\mathrm{off}}$ for T_{bs} and $q = k_{\mathrm{on}}$ for T_{bf}.

For nearly silent alleles the fitted k_{on} lies at the lower parameter bound, where the χ^2_3 and χ^2_1 references do not apply [44]. Therefore, P values were obtained by parametric bootstrap at the constrained fit of each null, re-fitting each simulated data set (Supplementary File S4). The bootstrap draws the two alleles independently, with cells drawn independently from the constrained fit, and it checks the numerics and the tests under the fitted model class only. A within-cell label-swap permutation of T_3 (2,000 swaps per gene), exact when the two alleles of a cell are exchangeable, is reported beside it as the check that makes no assumption about shared cell state (Supplementary File S4). The constrained fit of T_{bs} is optimized in coordinates that keep the shared burst size and each allele's k_{syn} inside their bounds by construction. An earlier implementation with a penalty outside the bounds stopped at its starting point for most genes and was replaced, with the bootstrap rerun, before the results reported here (Supplementary File S4). A T_3-significant gene is called size-resolved if only T_{bs} is also significant, frequency-resolved if only T_{bf} is, resolved on both if both are, and unresolved otherwise.

The allelic analysis plan, specified before the data were analyzed (Supplementary File S4), defined four artifact marks: boundary fit, low expression, possible mapping dropout (allelic mean-expression ratio exceeding 50) and likelihood ridge. A further criterion, total-count balance with C57 fraction inside $[0.3, 0.7]$, was adopted after unblinding, when re-fits showed that the largest apparent effects came from read mapping or a missing allele, and is reported as a deviation (Supplementary File S4). A significant gene is clean if it has none of the four marks and balance inside the band. It is a likely artifact if it has the dropout mark, or balance outside the band together with a boundary mark, a low-expression mark or an absolute $\log_2$ burst-size effect exceeding 2. The remaining significant genes form the "marked, plausible" stratum of Table 2.

The allelic analysis plan also set three calibration criteria. First, at least 16 of 20 imprinted positive-control genes were recovered in the expected direction. Second, null P values of all three statistics were uniform by a Kolmogorov–Smirnov test on 500 simulated null data sets, with the fraction at or below 0.05 inside the band $[0.02, 0.08]$ (observed 0.048–0.054). Third, median power over 30 genes for a 2-fold effect at observed expression was to be at least 0.5. An earlier pilot of the identical plan on the allelic matrices of Larsson et al. did not meet these three criteria, after which only the data set was changed (Supplementary File S4).

Two-step OFF model and detection power

For the two-step model ON $\xrightarrow{k_{\mathrm{off}}}$ OFF_2 $\xrightarrow{k_{12}}$ OFF_1 $\xrightarrow{k_{\mathrm{on}}}$ ON (synthesis rate ν while ON) the parameters of Eq. (8) are the OFF exit rates and the roots of a quadratic,

$$a_{1,2} = \{k_{12},\, k_{\mathrm{on}}\}, \quad b_1 + b_2 = k_{\mathrm{on}} + k_{12} + k_{\mathrm{off}}, \quad b_1 b_2 = k_{\mathrm{on}} k_{12} + k_{\mathrm{off}}(k_{\mathrm{on}} + k_{12}), \tag{18}$$

where $b_{1,2}$ may be complex conjugates and only their real symmetric functions $E^a_{1,2}$, $E^b_{1,2}$ of Eq. (9) are formed. Equation (9) is proved by a telescoping identity, is of minimal order for its coefficient degrees, and is stable when run downward (Supplementary File S1). The alternating ${}_2F_2$ series is summed in ball arithmetic with a geometric tail bound from $(x+b_1)(x+b_2) - (x+a_1)(x+a_2) = k_{\text{off}}(E^a_1 + x) \geq 0$ and extra working precision to absorb cancellation, and agrees with Arb's built-in ${}_pF_q$ (Supplementary File S5).

We minimized $\text{KL}(\theta)$ of Eq. (10) over the two-state family by a simplex search using double-precision probability vectors, truncating the sum over n at bounded remaining mass, with a rigorous bound on the omitted tail (Supplementary File S5).

The two-state family is the $k_{12} \to \infty$ boundary of the three-state family. The first two orders of the departure in $1/k_{12}$ are absorbed by shifts of k_{off} and ν and the first identifiable order is the third (Supplementary File S1). With $\delta = k_{12}^{-3} \geq 0$ as the one-sided parameter the likelihood-ratio null is asymptotically the Self and Liang mixture $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ [44], with critical value $c = 2.706$ at $\alpha = 0.05$. We checked this null by simulating the composite test, with both models re-fitted in every replicate, at nine two-state designs (three fitted shapes at 94, 224 and 682 cells, 500 replicates each; Supplementary File S5). The asymptotic power at separation $n\,\text{KL}$ is

$$\text{power} \;\simeq\; \Phi\big(\sqrt{2\,n\,\text{KL}} - \sqrt{c}\,\big), \qquad \sqrt{c} = \Phi^{-1}(1-\alpha), \tag{19}$$

where Φ is the standard normal distribution function, giving 0.64 at $n\,\text{KL} = 2$, 0.9 at $n\,\text{KL} = 4.3$ and 0.99 at $n\,\text{KL} = 8$. For any level-α test the power is at most $\alpha + \sqrt{n\,\text{KL}/2}$ (Pinsker's inequality applied to the n-cell product distributions). Exact-likelihood simulations at four parameter sets (2,000 replicates each) gave power 0.61 to 0.64 at $n\,\text{KL} = 2$ and at least 0.988 at $n\,\text{KL} = 8$. Binomial capture at fixed efficiency β leaves the switching rates unchanged and rescales the synthesis rate to $\beta\nu$; the fitted ν of Fig. 5 already contains the experiment's efficiency β_0, and β there is capture relative to the experiment. The cells needed are then

$$\sum_{k \geq m} \binom{k}{m} \beta^m (1-\beta)^{k-m}\, P(k;\nu) \;=\; P(m;\beta\nu), \qquad N(\beta) \;\simeq\; \frac{8}{\text{KL}(\theta;\beta\nu)}, \tag{20}$$

for power near 0.99 (4.3 replaces 8 for power 0.9).

The demonstration parameter set of Fig. 5a (a Markov embedding of Luo et al.'s first example [18], matched in mean dwell times and synthesis rate) has $k_{\text{on}} = 2/5$, $k_{12} = 2/7$, $k_{\text{off}} = 1/4$, $\nu = 20$ and the identifiable set $k_{\text{on}} = 4/5$, $k_{12} = 3/5$, $k_{\text{off}} = 3/10$, $\nu = 100$. The median and ninetieth-percentile fibroblast parameter sets (componentwise, over the exact C57 re-fits) are $(k_{\text{on}}, k_{\text{off}}, k_{\text{syn}}) = (0.421, 11.0, 39.5)$ and $(1.23, 206, 534)$. The calculations assume stationary, independent, identically distributed cells, one active state with a two-step OFF period only, and no measurement layer beyond fixed-efficiency thinning. For *Timp1* we propagated the fit's uncertainty by taking the range of KL over the 95% likelihood region of the two-state fit (Supplementary File S5), and elsewhere the values are at the point estimates.

Analysis plans and statistical reporting

Both analysis plans appear with their deviations in Supplementary Files S3 and S4. All reported values were recomputed, and the significant sets and classifications reproduced, with a second, separately written set of scripts. Tests other than likelihood-ratio tests are two-sided, false-discovery control is Benjamini–Hochberg [36] at 0.05 within each of the three allelic statistics, intervals are profile-likelihood intervals at the cutoff reported with them, and intervals on simulated proportions are Wilson intervals. Seeds and versions are in Supplementary File S7.

Acknowledgements

[**TODO-for-Matt: acknowledgments**]

Funding

[**TODO-for-Matt: funding statement**]

Author contributions

This manuscript is preliminary. The first draft was written by Claude (Anthropic), and the text and computations are still under review by the authors. [**TODO-for-Matt: contributions of M.D.S. and M.H. (CRediT)**]

Competing interests

[**TODO-for-Matt: competing interests**]

Data availability

The sequencing data analyzed here are available from ArrayExpress under accessions E-MTAB-6362, E-MTAB-6385, and E-MTAB-7098 (Larsson et al. [8]; the published fits re-evaluated and the cells-needed calculation) and E-MTAB-11054 (Johnsson et al. [32]; the 682 Smart-seq3 fibroblasts of the allele-resolved analysis). The published per-gene estimates re-evaluated here are those of the txburst repository at the commit given in Methods (Supplementary File S7). The materials produced here are deposited as one archive at [**TODO: archive DOI**]. The archive holds the corrected per-gene tables for the three data set and allele combinations as CSV files, one row per gene passing the original quality filter, with 69.6% of rows carrying no caveat mark (column dictionary in Supplementary File S3). It also holds the exact evaluator with reference probabilities at 25 parameter points to 50 digits each, the per-gene results of the allele-resolved analysis (Supplementary File S4), and a file manifest. No sequencing data are redistributed.

Code availability

The evaluator is released as the Python package [**TODO: package name**]. It evaluates the stationary count distributions of the two-state and two-step-OFF models exactly in ball arithmetic, with a self-checking multiprecision mode and a fast double-precision mode, and includes a self-test. It is distributed under an OSI-approved license [**TODO-for-Matt: license name**] in the same archive and in a public repository [**TODO-for-Matt: repository URL**], with the fixed dependency versions listed in Supplementary File S7. No R interface is provided.

Ethics statement

Not applicable.

Supplementary information

The supplementary material is a single PDF, Supplementary Files S1–S8, with its own tables, figures and reference list; the main text cites its parts as Supplementary File Sn, Supplementary Table Sn and Supplementary Fig. Sn. Supplementary File S1: derivations. Closed forms and

conventions, the three-term recurrence and its stable direction, the quadrature onset, the two-step OFF distribution and its order-three recurrence, series tail bounds, the boundary likelihood-ratio null, and related evaluations in the literature. Supplementary File S2: evaluator benchmark. Evaluation methods compared, failure of general-purpose routines, measured onsets, estimator shifts under the added constant, the BPSC probe, the double-precision policy and held-out agreement. Supplementary File S3: re-evaluation of the published fits. Analysis plan and deviations, likelihood differences at the published parameters, re-fit shifts, quality-call changes, excluded genes, profile-interval decomposition, coverage simulations and the column dictionary of the corrected per-gene tables. Supplementary File S4: the allele-resolved analysis. Analysis plans and deviations, the pilot, test statistics and calibration, stratification and artifact checks, class counts under four stratification variants, genome-wide allelic balance, the X chromosome and the per-gene catalog. Supplementary File S5: the two-step OFF model. Accuracy checks, the demonstration and identifiable parameter sets, the separability calculation, finite-sample power, per-gene separations and cells needed under incomplete capture. Supplementary File S6: application of the per-gene tests to published cell-state comparisons in budding yeast, *Arabidopsis* and *Hydra*. Supplementary File S7: software and reproduction. The evaluator and its self-test, dependency versions, repository files and commits of the implementations evaluated, reference values, random seeds and the archive file manifest. Supplementary File S8: Supplementary Figs. S1–S6 with legends.

Word counts Main text 7,252 words (words of the main text from the Introduction through the Conclusions, excluding abstract, display items, Methods, end matter and references); Methods 2,490; abstract 202; 5 figures; 2 tables; 51 references.

References

[1] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005. doi: 10.1016/j.cell.2005.09.031.

[2] Jonathan R. Chubb, Tatjana Trcek, Shailesh M. Shenoy, and Robert H. Singer. Transcriptional pulsing of a developmental gene. *Current Biology*, 16(10):1018–1025, 2006. doi: 10.1016/j.cub.2006.03.092.

[3] David M. Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011. doi: 10.1126/science.1198817.

[4] Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006. doi: 10.1371/journal.pbio.0040309.

[5] Daniel Zenklusen, Daniel R. Larson, and Robert H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271, 2008. doi: 10.1038/nsmb.1514.

[6] M. C. Walters, S. Fiering, J. Eidemiller, W. Magis, M. Groudine, and D. I. Martin. Enhancers increase the probability but not the level of gene expression. *Proceedings of the National Academy of Sciences*, 92(15):7125–7129, 1995. doi: 10.1073/pnas.92.15.7125.

[7] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer control of transcriptional bursting. *Cell*, 166(2):358–368, 2016. doi: 10.1016/j.cell.2016.05.025.

[8] Anton J. M. Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R. Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg.

Genomic encoding of transcriptional burst kinetics. *Nature*, 565:251–254, 2019. doi: 10.1038/s41586-018-0836-1.

[9] Minoru S. H. Ko. A stochastic model for gene induction. *Journal of Theoretical Biology*, 153 (2):181–194, 1991. doi: 10.1016/S0022-5193(05)80421-7.

[10] Jean Peccoud and Bernard Ycart. Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995. doi: 10.1006/tpbi.1995.1027.

[11] Vahid Shahrezaei and Peter S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008. doi: 10.1073/pnas.0803850105.

[12] Srividya Iyer-Biswas, F. Hayot, and C. Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Physical Review E*, 79:031911, 2009. doi: 10.1103/PhysRevE.79.031911.

[13] Jong Kyoung Kim and John C. Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14(1):R7, 2013. doi: 10.1186/gb-2013-14-1-r7.

[14] Trung Nghia Vu, Quin F. Wills, Krishna R. Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, 32(14):2128–2135, 2016. doi: 10.1093/bioinformatics/btw202.

[15] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (D3E) — a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 17:110, 2016. doi: 10.1186/s12859-016-0944-6.

[16] Yuchao Jiang, Nancy R. Zhang, and Mingyao Li. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biology*, 18:74, 2017. doi: 10.1186/s13059-017-1200-8.

[17] Anton J. M. Larsson, Christoph Ziegenhain, Michael Hagemann-Jensen, Björn Reinius, Tina Jacob, Tim Dalessandri, Gert-Jan Hendriks, Maria Kasper, and Rickard Sandberg. Transcriptional bursts explain autosomal random monoallelic expression and affect allelic imbalance. *PLoS Computational Biology*, 17(3):e1008772, 2021. doi: 10.1371/journal.pcbi.1008772.

[18] Songhao Luo, Zhenquan Zhang, Zihao Wang, Xiyan Yang, Xiaoxuan Chen, Tianshou Zhou, and Jiajun Zhang. Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model. *Royal Society Open Science*, 10(4):221057, 2023. doi: 10.1098/rsos.221057.

[19] Gennady Gorin and Lior Pachter. Special function methods for bursty models of transcription. *Physical Review E*, 102:022409, 2020. doi: 10.1103/PhysRevE.102.022409.

[20] Gennady Gorin, John J. Vastola, Meichen Fang, and Lior Pachter. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nature Communications*, 13:7620, 2022. doi: 10.1038/s41467-022-34857-7.

[21] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124:044104, 2006. doi: 10.1063/1.2145882.

[22] Junhao Gu, Nandor Laszik, Christopher E. Miles, Jun Allard, Timothy L. Downing, and Elizabeth L. Read. Scalable inference and identifiability of kinetic parameters for transcriptional bursting from single cell data. *Bioinformatics*, 41(11):btaf581, 2025. doi: 10.1093/bioinformatics/btaf581.

[23] Wenhao Tang, Andreas Christ Sølvsten Jørgensen, Samuel Marguerat, Philipp Thomas, and

Vahid Shahrezaei. Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics. *Bioinformatics*, 39(7):btad395, 2023. doi: 10.1093/bioinformatics/btad395.

[24] E. M. Ron. *Single-cell quantification and stochastic modeling of glucocorticoid receptor transport and DUSP1 mRNA dynamics*. PhD thesis, Colorado State University, 2026. PhD thesis, advisor B. Munsky; hdl.handle.net/10217/245508.

[25] Lisa Amrhein, Kumar Harsha, and Christiane Fuchs. *scModels: Fitting Discrete Distribution Models to Count Data*, 2026. R package version 1.0.5, CRAN.

[26] Augustinas Sukys and Ramon Grima. Cell-cycle dependence of bursty gene expression: insights from fitting mechanistic models to single-cell RNA-seq data. *Nucleic Acids Research*, 53(7): gkaf295, 2025. doi: 10.1093/nar/gkaf295.

[27] Gordon E. Willmot and Harry H. Panjer. Difference equation approaches in evaluation of compound distributions. *Insurance: Mathematics and Economics*, 6:43–56, 1987. doi: 10.1016/0167-6687(87)90006-0.

[28] Gordon E. Willmot. On recursive evaluation of mixed Poisson probabilities and related quantities. *Scandinavian Actuarial Journal*, 1993(2):114–133, 1993.

[29] Fredrik Johansson. Arb: efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66(8):1281–1292, 2017. doi: 10.1109/TC.2017.2690633.

[30] Martin Hemberg and Mauricio Barahona. Perfect sampling of the master equation for gene regulatory networks. *Biophysical Journal*, 93(2):401–410, 2007. doi: 10.1529/biophysj.106.099390.

[31] Martin Hemberg and Mauricio Barahona. A Dominated Coupling From The Past algorithm for the stochastic simulation of networks of biochemical reactions. *BMC Systems Biology*, 2 (1):42, 2008. doi: 10.1186/1752-0509-2-42.

[32] Per Johnsson, Christoph Ziegenhain, Leonard Hartmanis, Gert-Jan Hendriks, Michael Hagemann-Jensen, Björn Reinius, and Rickard Sandberg. Transcriptional kinetics and molecular functions of long noncoding RNAs. *Nature Genetics*, 54(3):306–317, 2022. doi: 10.1038/s41588-022-01014-1.

[33] DLMF. NIST digital library of mathematical functions. https://dlmf.nist.gov/. F. W. J. Olver et al., eds.

[34] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009. doi: 10.1093/bioinformatics/btp358.

[35] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9(1):60–62, 1938. doi: 10.1214/aoms/1177732360.

[36] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57: 289–300, 1995.

[37] Benjamin Zoller, Damien Nicolas, Nacho Molina, and Felix Naef. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Molecular Systems Biology*, 11 (7):823, 2015. doi: 10.15252/msb.20156257.

[38] Tianshou Zhou and Jiajun Zhang. Analytical results for a multistate gene model. *SIAM Journal on Applied Mathematics*, 72(3):789–818, 2012. doi: 10.1137/110852887.

[39] Zhixing Cao, Tatiana Filatova, Diego A. Oyarzún, and Ramon Grima. A stochastic model of

gene expression with polymerase recruitment and pause release. *Biophysical Journal*, 119(5): 1002–1014, 2020. doi: 10.1016/j.bpj.2020.07.020.

[40] Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015. doi: 10.1016/j.cell.2015.04.044.

[41] A. R. Stinchcombe, C. S. Peskin, and D. Tranchina. Population density approach for discrete mRNA distributions in generalized switching models for stochastic gene expression. *Physical Review E*, 85(6):061919, 2012. doi: 10.1103/PhysRevE.85.061919.

[42] J. Dattani. *Exact solutions of master equations for the analysis of gene transcription models*. PhD thesis, Imperial College London, Department of Mathematics, 2016. PhD thesis (dated November 2015, degree awarded May 2016), supervisor M. Barahona; hdl.handle.net/10044/1/45656.

[43] Svitlana Braichenko, James Holehouse, and Ramon Grima. Distinguishing between models of mammalian gene expression: telegraph-like models versus mechanistic models. *Journal of the Royal Society Interface*, 18(183):20210510, 2021. doi: 10.1098/rsif.2021.0510.

[44] Steven G. Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987. doi: 10.1080/01621459.1987.10478472.

[45] Chunjuan Zhu, Liang Chen, Zhishan Qiu, Jiaxin Chen, Feng Jiao, and Jianshe Yu. Correlation and distinction between stochastic gene transcription models with and without polymerase dynamics. *Physical Review Research*, 7(2):023050, 2025. doi: 10.1103/PhysRevResearch.7.023050.

[46] Feng Jiao, Jing Li, Ting Liu, Yifeng Zhu, Wenhao Che, Leonidas Bleris, and Chen Jia. What can we learn when fitting a simple telegraph model to a complex gene expression model? *PLoS Computational Biology*, 20(5):e1012118, 2024. doi: 10.1371/journal.pcbi.1012118.

[47] Christopher E. Miles. Mechanistic inference of stochastic gene expression from structured single-cell data. *Current Opinion in Systems Biology*, 42:100555, 2025. doi: 10.1016/j.coisb.2025.100555.

[48] Liang Chen, Ying Sheng, Zhihui Xie, and Feng Jiao. Systematic comparison and optimization of parameter estimation methods for stochastic gene expression kinetics. *Electronic Research Archive*, 34(4):2261–2278, 2026. doi: 10.3934/era.2026102.

[49] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, 2014. doi: 10.1038/nmeth.2930.

[50] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics*, 53(6):770–777, 2021. doi: 10.1038/s41588-021-00873-4.

[51] Ramon Grima and Pierre-Marie Esmenjaud. Quantifying and correcting bias in transcriptional parameter inference from single-cell data. *Biophysical Journal*, 123(1):4–30, 2024. doi: 10.1016/j.bpj.2023.10.021.