Supplementary Information

Controlled inference of transcriptional bursting from single-cell RNA counts

Matthew D. Schwartz Martin Hemberg

The eight Supplementary Files collect the derivations, benchmarks, analysis plans, tables and supplementary figures that the main text points to, where they are cited under the same numbers as "Supplementary File Sn", "Supplementary Table Sn" and "Supplementary Fig. Sn". Every number printed here is produced from the deposited analysis outputs by the code described in Supplementary File S7, and the Supplementary Files carry their own reference list.

S1 Derivations and related work

The derivations collected here underlie the telegraph likelihood of the main text (§§S1.1–S1.3) and the two-step OFF model (§§S1.4–S1.7). The two closing subsections review earlier evaluations of the telegraph likelihood and earlier work on the separability of the two count distributions in more detail than the Introduction of the main text. Throughout, rates are in units of the mRNA degradation rate, $a := k_{\text{on}}$, $b := k_{\text{off}}$, $r := k_{\text{syn}}$, burst frequency is k_{on} and burst size is $k_{\text{syn}}/k_{\text{off}}$, and the synthesis rate of the two-step model is written ν.

S1.1 Closed forms and parameter conventions

The stationary count distribution of the telegraph model is a Poisson mixture over a Beta-distributed promoter activity,

$$P(n) \;=\; \int_0^1 \mathrm{Pois}(n; ru)\,\mathrm{Beta}(u; a, b)\,du \;=\; \frac{r^n}{n!}\,\frac{(a)_n}{(a+b)_n}\;{}_1F_1(a{+}n;\; a{+}b{+}n;\; -r)\,, \tag{S1}$$

with $(x)_n$ the rising factorial and ${}_1F_1$ Kummer's function (DLMF 13.2.2 [1]). Kummer's transformation ${}_1F_1(\alpha;\gamma;-r) = e^{-r}\,{}_1F_1(\gamma{-}\alpha;\gamma;r)$ (DLMF 13.2.39 [1]) gives the second form

$$P(n) \;=\; e^{-r}\,\frac{r^n}{n!}\,\frac{(a)_n}{(a+b)_n}\;{}_1F_1(b;\; a{+}b{+}n;\; r)\,. \tag{S2}$$

In Eq. (S2) the transformed numerator parameter $\gamma{-}\alpha = b$ does not depend on n, and for $a, b, r > 0$ every term of the series ${}_1F_1(b; c; r) = \sum_{k\ge 0} \frac{(b)_k}{(c)_k}\frac{r^k}{k!}$ with $c = a + b + n$ is positive. A sum of positive terms cannot lose digits to cancellation: the rounding error of the partial sum is a small multiple of the sum itself, whatever the size of the largest term. The alternating series in Eq. (S1) has no such protection, because its terms grow to a size of order e^r before they fall off. The cancellation then removes from the working precision as many digits as the base-ten logarithm of the ratio of the largest term to the result (the loss in double precision is quantified in Supplementary File S2, § S2.4). Therefore all exact evaluations in this article sum Eq. (S2).

We compared both closed forms with two computations that use no special-function identity. Desingularized tanh–sinh quadrature of the mixture integral reproduces Eq. (S1) at three random parameter points to at least 40 digits. At two parameter points the Poisson–Beta distribution matches the stationary distribution of the truncated chemical master equation to about 10^{-14}, with the master equation solved in double precision as a linear system in a finite-state-projection-style truncation [2].

In the literature the Poisson–Beta distribution appears under several parameter conventions: the two Beta shapes in either order, both signs of the Kummer argument, and the sum $a{+}b$ carried as a single symbol (in the closed form the ${}_1F_1$ denominator parameter is $a + b + n$, not $b + n$). Supplementary Table S1 maps each convention used here, and that of the published txburst code, onto (a, b, r). We confirmed each row numerically, and we took the txburst mapping from its source code (the map is the identity: `kon`, `koff`, `ksyn` are a, b, r).

Two cautions apply to any re-implementation. First, rates $k_{\text{on}}, k_{\text{off}}, k_{\text{syn}}$ from a fit reported in per-hour units must be divided by the degradation rate before they enter these formulas. Second, forming $a{+}b{+}n$ in double precision before calling a high-precision routine perturbs the denominator parameter at the level of the double-precision rounding unit, which caps the agreement between independent evaluations near 14 digits. All evaluations used here therefore form the sum $a + b + n$ at working precision.

Table S1: Parameter conventions and their map to the primary Beta-mixture parameters $(a, b, r) = (k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}})/\delta$. The last column names how each row was confirmed: the symbolic Pochhammer–Euler identity (residual zero); the closed form against desingularized tanh–sinh quadrature of the mixture integral (at least 40 digits at three random points); the Kummer transformation (at least 40 digits); and the closed form against the stationary distribution of the truncated master equation (agreement 10^{-14} in a double-precision linear solve). Two cautions: a fit reported in per-hour units must divide its three rates by the degradation rate before entering these formulas, and forming $a + b + n$ in double precision before calling a high-precision routine perturbs the denominator parameter at the 10^{-16} level and caps the agreement between evaluations near 14 digits, and all our evaluations therefore sum parameters at working precision.

source	its symbols	map to (a, b, r)	confirmed by
Beta mixture (primary)	a, b, r	identity	quadrature; master equation
Kummer function of the closed form	${}_1F_1(\alpha;\gamma;z)$	$\alpha = a+n$, $\gamma = a+b+n$, $z = -r$	quadrature
transformed (all-positive) form	${}_1F_1(\beta';\gamma;z')$	$\beta' = b$ (independent of n), $\gamma = a+b+n$, $z' = +r$	Kummer transformation
txburst [3], maximum-likelihood script	`kon`, `koff`, `ksyn` passed to the density as (`alpha`, `bet`, `lam`)	$a =$ `kon`, $b =$ `koff`, $r =$ `ksyn` (identity)	code read and executed; Gauss–Jacobi weight orientation checked
Peccoud–Ycart [4]	λ (OFF→ON), μ (ON→OFF), ν (synthesis), δ (degradation)	$a = \lambda/\delta$, $b = \mu/\delta$, $r = \nu/\delta$; generating function ${}_1F_1\big(\frac{\lambda}{\delta};\frac{\lambda+\mu}{\delta};\frac{\nu}{\delta}(z-1)\big)$	authors' manuscript; Taylor coefficients of the generating function; master equation

S1.2 The three-term recurrence and its stable direction

The count distribution satisfies a three-term contiguity recurrence, stated for the Poisson–Beta case by Willmot [5] (§S1.8),

$$n(n+1)\,P(n{+}1) \;=\; n\,(n+a+b+r-1)\,P(n) \;-\; r\,(n+a-1)\,P(n{-}1)\,, \qquad n \geq 1\,. \tag{S3}$$

Equivalently, if we write $P(n) = C_n M_n$ with $C_n := e^{-r}\frac{r^n}{n!}\frac{(a)_n}{(a+b)_n}$ and $M_n := {}_1F_1(b; g{+}n; r)$, $g := a + b$, the Kummer factor obeys the contiguity relation in its denominator parameter,

$$r\,(a+n)\,M_{n+1} \;=\; (g+n)(g+n+r-1)\,M_n \;-\; (g+n)(g+n-1)\,M_{n-1}\,. \tag{S4}$$

Equation (S3) can be derived in two ways. The first is integration by parts on the Euler integral: with $I_m := \int_0^1 u^{a+m-1}(1-u)^{b-1}e^{-ru}\,du$, integrating the pointwise identity

$$\frac{d}{du}\Big[u^{a+n}(1-u)^b e^{-ru}\Big] = e^{-ru}(1-u)^{b-1}\Big[(a+n)u^{a+n-1} - (a+b+n+r)\,u^{a+n} + r\,u^{a+n+1}\Big]$$

over the unit interval gives $rI_{n+1} = (a+b+n+r-1)I_n - (a+n-1)I_{n-1}$, because the integral of the left side is a boundary term that vanishes for $a + n > 0$, $b > 0$, and substituting $I_m = P(m)\,m!\,B(a,b)/r^m$ then yields Eq. (S3). The second is Gauss contiguity of ${}_1F_1$ (DLMF 13.3.2 [1]) specialized to $M(b, g{+}n, r)$, which gives Eq. (S4). Thus a full vector $P(0), \ldots, P(n_{\max})$ costs two seed evaluations of the series plus $O(n_{\max})$ arithmetic.

Table S2: Digits retained by the three-term recurrence of the two-state distribution run forward (upward in n) and backward, in six parameter regimes, at 60-digit working precision against per-n reference values (per-n series values evaluated in ball arithmetic, precision 400). Forward digits are $-\log_{10}$ of the relative error at $n = 10$, 50 and 200 (negative values mean the error exceeds the value itself); the predicted forward loss is the growth of the dominant solution relative to $P(n)$; the last column is the smallest number of correct digits of the backward run over all $n \le n_{\max}$.

(a, b, r)	$n_{\max}$	forward digits ($n = 10/50/200$)	predicted loss (10/50/200)	backward (min)
(1.3, 4.7, 5)	900	53.6 / 17.2 / −193.7	8.2 / 44.5 / 255.5	60.3
(1.3, 4.7, 50)	900	52.0 / 36.1 / −29.5	10.9 / 26.8 / 92.4	59.3
(1.3, 4.7, 500)	900	43.8 / −6.4 / −102.1	20.1 / 70.3 / 166.0	59.1
(0.05, 0.05, 1.5)	300	58.4 / 8.9 / −274.0	3.4 / 52.8 / 335.7	59.8
(0.05, 0.05, 2000)	1500	37.0 / −37.8 / −223.1	28.7 / 103.5 / 288.7	59.1
(20, 0.3, 2500)	1200	43.1 / −25.5 / −214.4	20.1 / 88.7 / 277.6	58.9

The second solution. The solution space of Eq. (S3) is two-dimensional, spanned by the wanted $P(n)$ and by a second ("parasitic") solution built from the second Kummer function U,

$$Q(n) \;=\; C_n \,\frac{\Gamma(a+b+n)}{\Gamma(a+n)}\, U(b;\, a{+}b{+}n;\, r)\,, \tag{S5}$$

where the Gamma ratio maps the contiguity relation of U onto Eq. (S4). We obtained Q by coefficient matching and confirmed it numerically: the relative residual of Eq. (S3) applied to Q is about 10^{-115}. Any numerical solution of the recurrence is $P(n) + \epsilon\, Q(n)$ for some rounding-induced ϵ, and the error at step n relative to the wanted solution is governed by the ratio $Q(n)/P(n)$. That ratio increases with n in all the regimes we examined. Run upward from $n = 0$, the recursion amplifies a seed error ϵ by $[Q(n)/P(n)]/[Q(0)/P(0)]$, which grows. Run downward from the top, $n_{\max}{+}1$ and $n_{\max}$, it multiplies the error by $[Q(n)/P(n)]/[Q(n_{\max})/P(n_{\max})] \le 1$, which shrinks. Hence backward recursion from two seeds at the top is stable at all n, and forward recursion from $P(0), P(1)$ is unstable at nearly all n. Forward recursion is not safe even while $n < r$, where the distribution might be expected to be Poisson-like: the Poisson–Beta distribution is flat, roughly power-law, over $[0, \sim r]$ because the Beta mixture puts mass at small u, while Q rises like the pure-Poisson bulk $r^n/n!$. Only the rate at which forward recursion loses digits depends on (a, b, r) (roughly the base-ten logarithm of $r^n/n!$ against the flat P, worse for large r).

Measurement. We measured both directions at 60-digit working precision against per-n series values evaluated in ball arithmetic at 400 bits, in six parameter regimes that include two large-synthesis-rate corners of the parameter range (Supplementary Table S2). Backward recursion retained at least 58.9 of the 60 digits at all n in each regime, over sweeps of up to 1,500 steps, a loss of at most about one digit; the worst case is the regime $(a, b, r) = (20, 0.3, 2500)$. Forward recursion at $(a, b, r) = (1.3, 4.7, 500)$ retained 43.7 digits at $n = 10$ and no correct digit by $n = 200$. The forward loss predicted from Q/P tracked the measured loss to within a few digits over hundreds of digits of dynamic range, and the predicted backward loss is non-positive at all tabulated n.

Ball radii. Backward recursion stabilizes the midpoints of the balls but not their radii: radius propagation takes absolute values of the coefficients, and the radii therefore grow like the dominant solution of the recurrence in the coefficient magnitudes even in the backward direction. We estimate this growth in cheap floating-point arithmetic and add it to the working precision. Each returned ball is then tested against the target $\mathrm{rad} < \mathrm{mid} \cdot 2^{-\mathrm{prec}}$ and the precision raised up to 4 times; if the target is still missed, no value is returned (main text, Methods).

Agreement between evaluations. At 20 random parameter points (a, b log-uniform in $[0.01, 100]$,

$r \in [1, 3000]$, $n_{\max} \in [200, 1500]$) the backward recurrence and direct per-n evaluation of Kummer's function, both at 90 digits, agreed to 87.5 digits at worst. At the two large-synthesis-rate corners they agreed to 77.9–78.2 digits, the accuracy of the 80-digit reference values used there. We sum the all-positive series of Eq. (S2) directly rather than call Arb's native ${}_1F_1$ routine, which returns NaN in one parameter region, $(b, c, r) = (0.05,\ c \gtrsim 1050,\ 2000)$, at every precision we tried up to 2,000 bits. A further test at parameter points outside the development set is in Supplementary File S2 (§ S2.8).

S1.3 Accuracy limit of an N-node Gauss–Jacobi rule

An N-node Gauss–Jacobi rule integrates polynomials of degree at most $2N-1$ exactly against the Beta weight (degree 99 at $N = 50$ nodes). The factor of the integrand in Eq. (S1) that is not a polynomial is the Poisson kernel $\mathrm{Pois}(n; ru)$ read as a function of u: it peaks at $u_* = n/r$ and has width $\sqrt{n}/r$ in u. Jacobi nodes cluster quadratically at the endpoints, $u_k \approx (k\pi/2N)^2$, and the local node spacing near u_* is therefore $\pi\sqrt{u_*}/N$. The ratio of the two lengths,

$$\frac{\text{kernel width}}{\text{node spacing at } u_*} \;=\; \frac{\sqrt{n}/r}{\pi\sqrt{n/r}\,/N} \;=\; \frac{N}{\pi\sqrt{r}}\,, \tag{S6}$$

does not depend on n. When this ratio falls below one, the rule has fewer than one node under the peak of the kernel for all n at once, and under-resolves the whole count vector as soon as

$$r \;\gtrsim\; (N/\pi)^2\,, \tag{S7}$$

that is, $r \approx 253$ for the 50-node rule, 162 for 40 nodes and 10.1 for 10. Conversely, a fixed rule that resolves the kernel needs $N \gtrsim \pi\sqrt{r}$ nodes, approximately 314 at $r = 10^4$, the upper bound on k_{syn} in the published txburst code. Equation (S7) predicts the ratio of onsets between node counts, $(N_1/N_2)^2$, and the scale, though not the exact value, of the absolute onset, which also depends on the Beta shape. In Supplementary File S2 (§ S2.3) both predictions are compared with the onsets measured for the three published rules.

S1.4 The two-step OFF distribution from the master equation

The minimal Markov form of a refractory silent period [6, 7] is the cycle ON $\xrightarrow{k_{\mathrm{off}}}$ OFF_2 $\xrightarrow{k_{12}}$ OFF_1 $\xrightarrow{k_{\mathrm{on}}}$ ON, with synthesis at rate ν while ON. With H the 3×3 generator of the promoter chain (columns summing to zero), $P_i(k)$ the joint probability of promoter state i and k transcripts, and $u = (\nu, 0, 0)$ the synthesis vector, the master equation is

$$\frac{dP_i(k)}{dt} \;=\; u_i\big[P_i(k{-}1) - P_i(k)\big] + (k{+}1)P_i(k{+}1) - kP_i(k) + (HP(k))_i\,.$$

Multiply by the falling factorial $(k)_m$ and sum over k. With M the transcript number and $f_m(i) := E[(M)_m\,\mathbf{1}_{\{\mathrm{state}=i\}}]$ the three terms give $m\,u_i\,f_{m-1}(i)$, $-m\,f_m(i)$ and $(Hf_m)_i$, and stationarity reads

$$(mI - H)\,f_m \;=\; m\,\nu\,f_{m-1}(1)\,e_1\,. \tag{S8}$$

Two linear-algebra facts reduce Eq. (S8) to a scalar recursion for $f_m(1)$. The first is Cramer's rule: $e_1^T(mI - H)^{-1}e_1 = \det(mI - H_{\{1\}})/\det(mI - H) = (m + a_1)(m + a_2)/[m(m + b_1)(m + b_2)]$, where $a_{1,2}$ are the eigenvalues of $-H_{\{1\}}$ (H with the ON row and column deleted) and $b_{1,2}$ are the nonzero eigenvalues of $-H$ (zero is an eigenvalue because the columns sum to zero). The second is

the column-sum fact $\mathbf{1}^T(mI - H) = m\,\mathbf{1}^T$, hence $\mathbf{1}^T(mI - H)^{-1}e_1 = 1/m$. Applying e_1^T and $\mathbf{1}^T$ to Eq. (S8) gives $f_m(1) = \nu f_{m-1}(1)\,(m + a_1)(m + a_2)/[(m + b_1)(m + b_2)]$ and $E[(M)_m] = \nu f_{m-1}(1)$. The $m \to 0$ limit of the Cramer fact gives the stationary ON probability $\pi_{\mathrm{ON}} = a_1a_2/(b_1b_2) = f_0(1)$, and the product telescopes to

$$E[(M)_m] \;=\; \nu^m\,\frac{(a_1)_m(a_2)_m}{(b_1)_m(b_2)_m}\,. \tag{S9}$$

Because M is a Poisson mixture with bounded mixing variable ($M \mid X \sim \mathrm{Pois}(\nu X_1)$, $X_1 \in [0,1]$ [8]), its generating function $g(z) = E[z^M] = E[e^{\nu(z-1)X_1}]$ is entire and equals its factorial-moment series, $g(z) = \sum_m E[(M)_m](z-1)^m/m! = {}_2F_2[a_1, a_2; b_1, b_2; \nu(z-1)]$. Extracting the coefficient of z^k (an absolutely convergent double sum, using $(x)_{k+j} = (x)_k(x+k)_j$) gives the stationary count distribution

$$P(k) \;=\; \frac{\nu^k\,(a_1)_k\,(a_2)_k}{k!\,(b_1)_k\,(b_2)_k}\;{}_2F_2(a_1{+}k,\,a_2{+}k;\;b_1{+}k,\,b_2{+}k;\;-\nu)\,, \tag{S10}$$

which recovers the distribution that Cao, Filatova, Oyarzún and Grima derived for this promoter [9] and Herbach's Corollary 5.5 [8]. Exact stationary distributions for multistate promoter models in generalized-hypergeometric form go back to Zhou and Zhang [10]. A direct comparison with the truncated master equation, which uses no special-function identity, is reported with the other accuracy tests of the two-step evaluation in Supplementary File S5 (§ S5.1).

The eigenvalue dictionary. For the refractory cycle the numerator parameters are exactly the two OFF exit rates, $\{a_1, a_2\} = \{k_{\mathrm{on}}, k_{12}\}$, but the b_i are the roots of $x^2 - E_1^b x + E_2^b$ with

$$E_1^b = b_1 + b_2 = k_{\mathrm{on}} + k_{12} + k_{\mathrm{off}}\,, \qquad E_2^b = b_1b_2 = k_{\mathrm{on}}k_{12} + k_{\mathrm{off}}(k_{\mathrm{on}} + k_{12})\,, \tag{S11}$$

and are not rates. Only their symmetric functions are simple, $b_1 + b_2 = (a_1 + a_2) + k_{\mathrm{off}}$ and $b_1b_2 = a_1a_2 + k_{\mathrm{off}}(a_1 + a_2)$, just as only the sum $a + b$ enters the denominator parameter of the two-state closed form (§S1.1). The naive guess $b = \{k_{\mathrm{on}}{+}k_{\mathrm{off}},\, k_{12}{+}k_{\mathrm{off}}\}$ already gives the wrong trace, counting k_{off} twice, and its distribution differs from the true one by up to tens of percent (Supplementary File S5, § S5.1). Complex-conjugate b is generic: the discriminant $(E_1^b)^2 - 4E_2^b$ is negative whenever the three rates are comparable (for example $k_{\mathrm{on}} = k_{12} = k_{\mathrm{off}} = 1$ gives $b = (3 \pm i\sqrt{3})/2$). The probabilities $P(k)$ stay real, and we avoid complex arithmetic by working throughout in the real symmetric functions $E_1^a = a_1 + a_2$, $E_2^a = a_1a_2$, E_1^b, E_2^b and ν, through

$$(a_1)_k(a_2)_k = \prod_{i<k} R_a(i), \quad R_a(x) = E_2^a{+}E_1^a x{+}x^2, \qquad (b_1)_k(b_2)_k = \prod_{i<k} R_b(i), \quad R_b(x) = E_2^b{+}E_1^b x{+}x^2, \tag{S12}$$

with $R_b(x) = |b_1 + x|^2$ when b is complex. No numerator parameter cancels against a denominator parameter at positive rates, $b_i \neq a_j$, because the characteristic polynomial of $-H$ evaluated at a_i equals k_{off} times the partner OFF rate, which is positive. The confluent two-state limit ($k_{12} \to \infty$ or $k_{\mathrm{off}} \to 0$) lies on the boundary of the positive-rate parameter space and can be reached exactly only in the eigenvalue parameterization.

Identifiability from stationary counts. Stationary snapshot data determine at most the eigenvalue tuple (a, b, ν) of Eq. (S10). Since $\{a_1, a_2\} = \{k_{\mathrm{on}}, k_{12}\}$ enters symmetrically, the map from rates to the distribution is invertible only up to the unordered pair $\{k_{\mathrm{on}}, k_{12}\}$. In particular, stationary counts do not identify which OFF stage is the slow one. Near the two-state boundary, where the mixing density becomes a Beta density, the model approaches the telegraph model. With N sequential OFF stages the same construction gives a ${}_NF_N$-type distribution and a recurrence of order $N{+}1$, a generalization not used in this article.

S1.5 Proof, minimality and stability of the order-three recurrence

The generating function g satisfies the ${}_2F_2$ differential equation in $w = \nu(z-1)$, $\vartheta_w(\vartheta_w+b_1-1)(\vartheta_w+b_2-1)F = w(\vartheta_w+a_1)(\vartheta_w+a_2)F$ with $\vartheta_w = w\,d/dw$. In terms of the operator $\eta = (z-1)\,d/dz$, and after division by the common factor $(z-1)$, this equation reads $D(\eta+b_1-1)(\eta+b_2-1)g = \nu(\eta+a_1)(\eta+a_2)g$ with $D = d/dz$. Collecting the coefficient of z^K in $g = \sum_k P(k)z^k$ then gives a four-term recurrence of order three whose coefficients again depend only on the real symmetric functions:

$$\begin{aligned}(K{+}1)(K{+}2)(K{+}3)\,P(K{+}3) \;=\; & (K{+}1)(K{+}2)\big(2K + E_1^b + \nu + 1\big)\,P(K{+}2) \\ & -(K{+}1)\big[K^2 + K(E_1^b + 2\nu) + \nu(E_1^a + 1) + E_2^b\big]\,P(K{+}1) \\ & +\nu\big(K^2 + E_1^a K + E_2^a\big)\,P(K), \qquad K \geq 0.\end{aligned} \tag{S13}$$

The two-state three-term analog, Eq. (S3), appears in the actuarial literature [5, 11], and we know of no previous statement of the three-state recurrence.

Proof. Substituting Eq. (S10) into Eq. (S13) and re-indexing $s = i + j$ turns the residual into $\sum_s v_s R_s(K)$, with v_s the ${}_2F_2$ term sequence and R_s a rational function of s, K and the parameters. In the two-state case the denominator-contiguity proof works term by term, but here R_s is not identically zero and a pointwise proof fails. The recurrence is proved instead by a Gosper–Wilf–Zeilberger telescoping identity: a computer-algebra search finds a polynomial $d(s)$ of degree four in s, with coefficients rational in K and the parameters, such that

$$R_s \;=\; d(s) - d(s{+}1)\,\rho(s)\,, \qquad \rho(s) = v_{s+1}/v_s\,, \qquad d(0) = 0\,, \tag{S14}$$

and $\sum_s v_s R_s$ then telescopes to $d(0)\,v_0 = 0$. The polynomial d of Eq. (S14) is included with the code (Supplementary File S7). As the confluent two-state limit requires, at $b_1 = a_1$ every solution of the two-state recurrence, Eq. (S3), also satisfies the order-three recurrence.

Minimality. $P(k)$ is transcendental in ν, and we therefore tested minimality on the exact formal-ν series over $\mathbb{F}_p[[\nu]]$ (100 orders in ν, $K \leq 60$, generic rational rates in the complex-b regime) at the two primes 2147483647 and 2147483629. No recurrence of order two exists with coefficient degrees at most 6 in k and at most 2 in ν: the null space is empty at both primes. Thus order three is minimal within these bounds, and at order three the null space is one-dimensional, spanned by the coefficient vector of Eq. (S13).

Stability. The other two solutions of Eq. (S13) are the coefficient sequences of the two solutions of the generating-function equation that are singular at $z = 1$. Their indicial structure gives a power-law fall-off $\sim k^{-\mathrm{Re}\,b_i}$ (oscillating when b is complex), against the super-exponential fall-off $\nu^k/k!$ of $P(k)$. The wanted solution is therefore the minimal one under upward recursion, and the conclusion of §S1.2 holds unchanged: backward recursion from three seeds at the top is stable at all k, and forward recursion is unstable nearly everywhere. We measured both directions at 60-digit working precision in four regimes, including complex b and ν up to 2,000, with $k_{\max}$ up to 1,800 (Supplementary Table S3). Backward recursion retained at least 55.1 digits at all k in each regime, a loss of at most about five digits including the seed error over the longest sweep. Forward recursion had lost every digit by $k = 50$–200 in each regime. To evaluate a full count vector we run the recurrence backward from three seeds, and we handle the radii of the ball recurrence, which grow like the dominant solution even backward, as in §S1.2.

S1.6 Series bounds for the ${}_2F_2$ evaluation

The two-state tail bound. For Eq. (S2), since $b < c := a + b + n$, the ratio of successive series terms obeys $t_{k+1}/t_k = (b+k)\,r/[(c+k)(k+1)] \leq r/(k+1)$. Once $k + 1 > r$, the tail beyond term K

Table S3: Digits retained by the order-three recurrence of the two-step-OFF distribution run forward and backward in four regimes (including complex-conjugate $b_{1,2}$ and $\nu = 2000$), at 60-digit working precision. Forward digits at $k = 10$, 100 and $k_{\max}$ (negative: error larger than the value); last column: smallest number of correct digits of the backward run over all $k \le k_{\max}$. The backward direction is stable in every regime (all four).

regime	$k_{\max}$	forward digits at $k = 10/100/k_{\max}$	backward digits (min)
complex b, small ν	400	58.1 / −31.2 / −532.4	59.9
complex b, $\nu = 300$	900	47.4 / −24.7 / −235.8	57.0
demo (Luo et al.-like), $\nu = 20$	400	57.7 / 22.5 / −296.7	58.0
difficult real b, $\nu = 2000$	1800	40.0 / −106.4 / −797.8	55.1

is bounded by the geometric sum $t_K/(1 - r/(K+1))$. We compute this bound T entirely in ball arithmetic and add it to the partial sum as the interval $T \cdot [0, 1]$.

The two-step series. A Kummer-type transformation that would make the ${}_2F_2(-\nu)$ series of Eq. (S10) sign-definite does not exist for generic ${}_2F_2$. The two-state trick ${}_1F_1(\alpha;\gamma;-r) = e^{-r}\,{}_1F_1(\gamma-\alpha;\gamma;r)$ has no ${}_2F_2$ analog, and $e^{\nu}\,{}_2F_2(-\nu)$ is not hypergeometric (its product coefficients are not sign-definite). In its place we use two proved properties of the series, the first of which is spectral dominance. From Eq. (S11),

$$R_b(x) - R_a(x) \;=\; k_{\text{off}}\,(E_1^a + x) \;\ge\; 0 \qquad \text{for all } x \ge 0, \tag{S15}$$

identically in the rates (at run time the implementation tests the two ball inequalities $E_1^b \ge E_1^a$, $E_2^b \ge E_2^a$, which are the trace and determinant relations of the dictionary in inequality form). Hence the ratio of successive terms of the ${}_2F_2$ series at count k obeys $|T_{j+1}/T_j| = \nu\, R_a(k+j)/[R_b(k+j)(j+1)] \le \nu/(j+1)$, a bound that is below one and strictly decreasing once $j+1 > \nu$. If the series is truncated at the first omitted term T_J with $\nu/(J+1) < \frac{1}{2}$, the tail is enclosed by $|T_J|/(1-\nu/(J+1))\cdot[-1, 1]$, again computed entirely in ball arithmetic. The second property is an a priori bound on the cancellation. Since $R_a \le R_b$, $\sum_j |T_j| \le e^{\nu}$; and the ${}_2F_2$ factor at count k equals $E[X^k e^{-\nu X}]/E[X^k] \ge e^{-\nu}$ for the mixing variable $X \in [0, 1]$. The alternating sum therefore loses at most $2\nu \log_2 e$ bits to cancellation. The working precision starts at the target precision plus $2\nu \log_2 e$ plus a margin, and the width of the result is then tested against the target as in §S1.2.

Another arrangement of the series, which we examined and rejected, starts from the fact that the mixing density of Eq. (S10) is itself a positive Beta mixture (Herbach [8], Sec. 5.4). Written with $\delta = E_1^b - E_1^a = k_{\text{off}}$ in the form $x^{a_{\max}-1}\,{}_2F_1[b_1-a_{\min},\, b_2-a_{\min};\, \delta;\, 1-x]$, it has ${}_2F_1$ coefficients $|(b - a_{\min})_m|^2 \ge 0$ even for complex b, and this form expresses $P(k)$ as a positive series of exact two-state kernels. However, its outer convergence is only polynomial ($\sim m^{-(a_{\min}+1)}$). A rigorous truncation also has to bound the inner ${}_1F_1$ from above, and the estimate that does so is loose by a factor of order e^{ν}. For evaluation this form is therefore strictly worse than the bounded alternating series.

An independent exact evaluation of the same distribution is Arb's rigorous ${}_pF_q$ [12, 13] called with the complex b_i. We require the imaginary part of the resulting complex ball to contain zero and take its real part as the value. The agreement of the series and this evaluation with each other and with two further multiprecision implementations, used for comparison only, is reported in Supplementary File S5 (§ S5.2).

S1.7 The boundary-calibrated likelihood-ratio test

Write $P_3(\theta)$ for the two-step distribution of Eq. (S10) and $P_2(\phi)$ for the telegraph distribution, and let

$$\mathrm{KL}(\theta) \;=\; \min_{\phi} \mathrm{KL}(P_3(\theta) \,\|\, P_2(\phi)) \tag{S16}$$

be the divergence from a true two-step distribution to the nearest telegraph distribution. Then $n\,\mathrm{KL}$ is the expected log-likelihood ratio that n independent stationary cells carry in favor of the true two-step distribution against its best two-state approximation (the Stein–Chernoff regime of a simple likelihood-ratio test). Translating $n\,\mathrm{KL}$ into the power of the *composite* test, in which both models are fitted, needs the null distribution of the likelihood-ratio statistic $\mathrm{LR} = 2[\sup_\theta \ell_3(\theta) - \sup_\phi \ell_2(\phi)]$ when the truth is two-state.

The telegraph family does not lie in the interior of the two-step family. In the dwell-split parameterization of Supplementary File S5, a two-step OFF period with split fraction $f \in (0, \frac{1}{2}]$ reduces to the telegraph model exactly at $f = 0$ (equivalently $k_{12} \to \infty$). Therefore the null value lies on the boundary of a single one-sided parameter, the case treated by Self and Liang [14]. Locally the statistic is the square of a signed root Z that is standard normal under the null, and the extra parameter can improve the fit in one direction only. With probability one half the unconstrained optimum falls outside the parameter space, the constrained fit coincides with the two-state fit, and $\mathrm{LR} = 0$. The null distribution is therefore the equal mixture

$$\mathrm{LR} \;\sim\; \tfrac{1}{2}\chi^2_0 + \tfrac{1}{2}\chi^2_1 \tag{S17}$$

rather than χ^2_1 [15], and the level-α critical value is the $(1-2\alpha)$ quantile of χ^2_1, $c_\alpha = z^2_{1-\alpha}$, smaller than the two-sided χ^2_1 quantile.

Under a two-step alternative at separation $n\,\mathrm{KL}$ the signed root is, to leading order, normal with unit variance and mean $\sqrt{2n\,\mathrm{KL}}$ (the noncentrality of the likelihood-ratio statistic is $2n\,\mathrm{KL}$), and the test rejects when $Z > z_{1-\alpha}$. The asymptotic power is then

$$\mathrm{power}(n\,\mathrm{KL}) \;\approx\; \Phi\big(\sqrt{2n\,\mathrm{KL}} - z_{1-\alpha}\big), \tag{S18}$$

with Φ the standard normal distribution function. Inverting it gives the separation needed for a target power $1-\beta_{\mathrm{err}}$, $n\,\mathrm{KL} = \frac{1}{2}(z_{1-\alpha} + z_{1-\beta_{\mathrm{err}}})^2$, and hence the cells-needed rule of the main text, $n = \mathrm{const}/\mathrm{KL}$. Using the two-sided χ^2_1 critical value in place of Eq. (S17) would understate the power. The numerical values (power at $n\,\mathrm{KL} = 2$ and 8, the separation at which power reaches 0.9) and their Monte-Carlo confirmation with exact likelihoods are in Supplementary File S5 (§ S5.5). Two steps of this argument are asymptotic: Eq. (S18) itself, and the identification of the composite test's noncentrality with the simple-test separation $2n\,\mathrm{KL}$. Both were confirmed by simulation only at the parameter sets reported there.

The contact between the two families is of high order. Writing $\epsilon = 1/k_{12}$ and expanding the factorial-moment ratio of the two-step law (Eq. (S9)) in ϵ, the first-order term is absorbed by the shift $k_{\mathrm{off}} \to k_{\mathrm{off}} + k_{\mathrm{on}}k_{\mathrm{off}}\,\epsilon$ of a telegraph law, and the second order by $k_{\mathrm{off}} \to k_{\mathrm{off}} + k_{\mathrm{on}}k_{\mathrm{off}}\,\epsilon + k_{\mathrm{on}}k_{\mathrm{off}}(k_{\mathrm{on}} + k_{\mathrm{off}})\,\epsilon^2$ together with $\nu \to \nu(1 + k_{\mathrm{on}}k_{\mathrm{off}}\,\epsilon^2)$, and the two-step law then agrees with a telegraph law through second order in ϵ. The third-order residual of the ratio, $-k_{\mathrm{on}}k_{\mathrm{off}}(k_{\mathrm{on}}k_{\mathrm{off}} + j^2)/(k_{\mathrm{on}} + k_{\mathrm{off}} + j)$ relative to the matched telegraph ratio at moment order j, grows with j and cannot be absorbed by any re-parametrization, because it lies outside the tangent space of the telegraph family. On $\log P(n)$ itself the first- and second-order departures project onto the three telegraph scores to rounding at every parameter point we examined, and the third order keeps a nonzero orthogonal component. The regular one-sided parameter of the boundary test is therefore $\delta = \epsilon^3$, and near the boundary $\mathrm{KL}(\theta)$ grows as ϵ^6 (as f^6 on the dwell-split family

of Supplementary File S5, where exact minimization gives local log-log slopes rising to six). Self and Liang's mixture holds in $\delta \geq 0$, but $\hat{\epsilon}$ converges at the rate $n^{-1/6}$, and the composite statistic therefore approaches the mixture slowly at the cell numbers of these data (Supplementary File S5, § S5.5). The per-gene and slice values of KL reported in the main text are computed directly from the two distributions and do not use this expansion.

S1.8 Related evaluations of the telegraph likelihood

The actuarial recurrence. Willmot and Panjer developed difference-equation methods for mixed-Poisson claim counts in the actuarial literature [11], and Willmot's survey states Eq. (S3) for the Poisson–Beta case, with ${}_1F_1$ boundary values [5]. We implemented his published formulas as printed (forward direction, negative-argument Kummer boundary values) and compared them with our backward recurrence. Over all $m \leq n_{\max}$ at 10 random points the two agree to at least 77.9 digits, close to the full working precision of the 80-digit comparison. Run in the published forward direction, the recurrence needed up to approximately 466 digits of extra working precision to remain accurate at $n_{\max}$, as the stability analysis of §S1.2 predicts. The actuarial papers state the recurrence in the forward direction only, with no stability analysis or error control.

No gene-expression tool that we know of uses the three-term recurrence. The published tools evaluate the count distribution by fixed Gauss–Jacobi quadrature (txburst [3], BPSC [16]), bypass its evaluation with Monte-Carlo sampling [17, 18], evaluate ${}_1F_1$ directly in floating point [19] or in extended precision [20], or compute the stationary distribution numerically from the master equation [21–23]. The nearest methodological precedent outside gene expression is Pessia and Tang's three-term recurrence for the ill-conditioned hypergeometric transition probabilities of the simple birth–death process [24], a recurrence whose accuracy they tested without error bounds.

Extended precision. The closest earlier high-precision implementation is the scModels R package [20], whose density function implements Eq. (S2), in our convention, in MPFR (the version examined is listed in Supplementary File S7). According to its source code, the series is summed at MPFR's default 53-bit mantissa (extended exponent range, no extra digits) and truncated when the last term's absolute value drops below 10^{-6}, with no tail bound or error policy. Against 60-digit references its relative error is 10^{-16}–10^{-11} across most of the parameter grid of Supplementary File S2. At one point, $(a, b, r, n) = (10, 0.1, 3000, 2900)$, it returns a value with relative error 4.6×10^{-6} and no warning, which is consistent with summing the series without a tail bound at large r and n.

Floating-point closed forms. Sukys and Grima found their log-space ${}_1F_1$ implementation "sufficiently numerically stable" [19]. In Supplementary File S2 we compare our re-implementation of that algorithm with ball-arithmetic references and find it accurate over the whole parameter grid, an accuracy that rests on the positivity of Eq. (S2). Gorin and Pachter note that closed-form approaches "presuppose the availability of efficient implementations" of the special functions [25]. For one widely used ${}_1F_1$ routine (the `scipy` row of the same benchmark) that assumption fails through underflow in part of the grid, although the routine is accurate wherever its result is finite. In Mottes, Zhu and Brenner's "exact stochastic kinetic models" the exactness is in the forward simulation: the fit minimizes a divergence between simulated and observed histograms by gradients through a relaxed sampler, and no likelihood is evaluated [26].

Truncated master equations. The finite state projection method evaluates truncated-master-equation likelihoods with a computable truncation bound [2] and has been used to fit multistate promoter models to single-molecule FISH data [27]. A recent method of this type evaluates truncated-master-equation telegraph likelihoods from single-cell RNA sequencing, dropping high-count states at a stated error tolerance of 10^{-4} [21, 22]. StochasticGene.jl fits promoters with any number

of states to stationary count histograms by Bayesian sampling [23]. In these methods the stated tolerance or projection bound controls the state-space truncation and does not bound the error of the floating-point evaluation of the stationary distribution.

Holonomic methods. The holonomic-systems community uses exact contiguity relations and Pfaffian systems for ${}_1F_1$ [28, 29]. We found no application of them to this likelihood, and the reference list of the holonomic gradient method maintained by that community (about 49 entries) contains no gene-expression application. To our knowledge, the telegraph likelihood had not been evaluated with controlled error before.

S1.9 Related work on the separability of two-step and two-state count distributions

Several studies have established that stationary snapshots discriminate promoter architectures poorly. Stinchcombe, Peskin and Tranchina showed for arbitrary dwell-time distributions that "empirical estimates of the steady-state mRNA probability mass function ... do not allow one to distinguish between underlying models with exponential and nonexponential dwell times in some relevant parameter regimes" [30]. Dattani, maximizing an exact multi-cycle likelihood on simulated snapshots designed after Zoller et al., found that "good fits to the empirical distribution are not necessarily indicative of the best parameter estimates" [31, 32]. Braichenko, Holehouse and Grima compared telegraph-like models with a mechanistic model of mammalian transcription and found a large parameter region where the models are "practically indistinguishable by comparison of their transcript numbers distributions" while still separable by the shape of their waiting-time distributions [33]. Miles's review [34] calls the phenomenon "model mimicry", and Jiao et al. demonstrated it for the model pair studied here [35].

Jiao et al. fitted the telegraph model to stochastically simulated refractory distributions at 625 grid points and found a divergence below 0.025 everywhere, with the small-sample Akaike criterion selecting the wrong model in over 90 % of parameter sets at 100 cells [35]. Cao et al. mapped the Hellinger distance between the refractory and telegraph distributions over a parameter plane and found that the refractory model "in general, is not well approximated by the telegraph model" [9]. That plane spans parameters at which the two distributions separate, whereas the per-gene calculation of the main text is made at the fitted, bursty shapes of the Larsson et al. data, where they nearly coincide. Our exact values corroborate this mimicry at the fitted shapes and make it quantitative per gene. The earlier studies computed the separation of the two distributions at chosen parameter points, with simulation and fitting noise in the resulting values. The calculation here evaluates the divergence of Eq. (S16) from the exact distributions at each gene's fitted parameters and multiplies it by the gene's actual cell count. The maxima of the separation over the parameter slices through two fitted shapes, and its distribution over the genes, are given in Supplementary File S5, and the error control of the calculation is described in § S5.4 of that file.

Gu et al. [21] gave a comparable per-gene analysis of *parameter* identifiability within the two-state model, a question that presupposes the model. Gorin et al. [36] introduced cells-needed curves for model discrimination through simulation-calibrated Bayes factors on a different model pair. They also reported that joint measurements of unspliced and spliced transcripts improved model distinguishability by roughly an order of magnitude, which suggests one way in which the presence of a second OFF step could become decidable gene by gene. The cells-needed curves of the main text are computed from the divergence of Eq. (S16), evaluated exactly, and converted to cell numbers through the asymptotic power relation of §S1.7; the error control of Supplementary File S5 applies to the divergence, while the power of the test at a finite number of cells is known only through that asymptotic relation.

S2 Evaluator benchmark

This Supplementary File reports the benchmark behind the accuracy results of the main text: the parameter grid and reference values, the six evaluation methods scored in Supplementary Table S4, and the measured onsets of quadrature error with their ratios between node counts. It also reports two failure modes other than quadrature and what the quadrature error implies for likelihood values and point estimates. The later subsections cover how the constant that txburst adds inside the logarithm changes the likelihood surface, the consequences of BPSC's own quadrature error for its fits and differential-expression tests, a test of the evaluator at parameter points outside its development set, the error control of its double-precision mode, and timings. All results concern the numerical evaluation of the stationary distribution of the telegraph model. The txburst parameters and BPSC's $(\alpha, \beta, \lambda_1)$ both map to (a, b, r) of Supplementary File S1 by the identity map (§S1.1 for txburst; both quadrature calls are quoted in Supplementary File S7, §S7.3), which excludes convention differences as a source of any error reported below. The three quadrature rules evaluated are those of the published tools at fixed versions (Supplementary File S7): the 50-node rule of txburst's maximum-likelihood and profile-likelihood scripts, the 40-node rule of its likelihood-ratio test script, and the 10-node rule of BPSC.

S2.1 Grid, reference and error metric

The benchmark grid has 228 cells: a logarithmic lattice with six values per axis, $k_{\text{on}} \in [0.03, 12.9]$, $k_{\text{off}} \in [0.3, 999]$, $r \in [2, 10^4]$ (216 points), plus 12 points taken from the published fits of Larsson et al. [3]. At each cell we evaluate the count distribution as a vector over n from zero to the model's 99.9 percent quantile (capped at $n = 3{,}000$). As the reference we use the ball-arithmetic evaluator of the main text at 150-bit precision, unmodified. Its largest relative ball radius over the entire grid is 1.9×10^{-60}, a factor of at least 4×10^{43} below the smallest deviation scored, and the failures counted here therefore do not depend on the reference's own uncertainty.

As the error metric we take the maximum over n of the relative error of $P(n)$ within a cell. A cell is *clean* when it has no hard failure and that maximum is at most 10^{-6}; a *hard failure* is a NaN, a $-\infty$ or zero, or a negative value at an n where the reference $P(n) \geq 10^{-300}$ (on this grid, everywhere) inside the quantile range. We located the onsets of quadrature error at a fixed shape $(k_{\text{on}}, k_{\text{off}})$ by scanning r in multiplicative steps of 1.096 (about 10% resolution) for the two txburst rules and in coarser logarithmic steps for BPSC's rule. Each onset below is the first scan point over the stated threshold; where two values are given, the smaller is the last point under it.

S2.2 The six evaluation methods

Supplementary Table S4 compares six methods with the reference. Rows 1 and 2 are the two Gauss–Jacobi rules of txburst as implemented in its own code, the 50-node rule of the fitting and profile scripts and the 40-node rule of the test script (Supplementary File S7, §S7.3, compares our copies with the published scripts). Row 3 is the closed form of Eq. (S1) with the ${}_1F_1$ factor from `scipy.special.hyp1f1` in double precision and a prefactor that we assemble in log space. Row 4a is a double-precision, fully log-space evaluation of the all-positive series of Eq. (S2) by per-term log-sum-exp. It is our re-implementation of the algorithm of Sukys and Grima [19], whose Julia code we could not run in our environment and whose concrete implementation may differ (for example in asymptotic switches).

Row 4b is the defining alternating series summed naively in double precision. No published tool evaluates it this way; we include it because a log-sum-exp evaluation of an alternating series would

have to fall back to it, and the comparison with row 4a shows what the Kummer transformation to an all-positive series contributes. Its prefactor is assembled in log space and the per-n arithmetic order is preserved under vectorization, both choices favoring its accuracy. Row 5 is the double-precision summation of this work (all-positive series, log-space prefactor, linear accumulation with per-element power-of-ten rescaling), the code also used for the maximizer comparison of §S2.5, and row R is the reference.

The clean-cell counts are 153 of 228 for the 50-node rule, 140 for the 40-node rule, 175 for the `scipy` closed form, 94 for the naive series, and 228 for both the log-space series and the double-precision summation of this work. The two txburst rules reach maximum relative errors of 2.90 and 4.78 at the grid corner $(0.03, 0.3, 10^4)$, and the 50-node rule exceeds 10^{-6} relative error in 75 of the 228 cells. Rows 4a and 5 evaluate the same all-positive series: per-term log-sum-exp takes 3.4 times as long as linear accumulation with power-of-ten rescaling and accumulates about 8 times more rounding error (1.5×10^{-10} against 1.9×10^{-11} worst case). The errors reported for rows 1 and 2 are those of the raw density returned by the txburst rules and exclude the effect of the constant 10^{-10} that the published likelihood adds inside the logarithm (§S2.6).

S2.3 Onsets at two shapes and their ratios

We measured the onsets against the reference at the two componentwise median shapes of the published fits, $(k_{\text{on}}, k_{\text{off}}) = (0.426, 9.45)$ (fibroblast) and $(0.433, 2.12)$ (mouse embryonic stem cells, mESC). Supplementary Fig. S1 shows the published fits with the onset lines. At the fibroblast shape the 50-node rule's maximum relative error is 1.1×10^{-12} at $r = 100$ and 1.7×10^{-8} at $r = 300$, just past the predicted onset $(50/\pi)^2 \approx 253$. On the scan it crosses 10^{-6} between $r = 364$ (error 4.4×10^{-7}) and 399 at the fibroblast shape, and at $r = 364$ at the mESC shape. The 40-node rule crosses between 252 (error 9.3×10^{-7}) and 276 (fibroblast) and at 230 (mESC), against a prediction of 162. BPSC's 10-node rule leaves its double-precision floor of about 1.1×10^{-14} between $r = 2$ and 10.99 (where its error is 2.0×10^{-12} on the grid) and crosses 10^{-6} at $r = 15.9$ at both shapes, against a prediction of 10.1.

Error at the percent level follows at roughly four to six times the 10^{-6} onset predicted by Eq. (S7): $r = 1{,}099$ and 1,002 for the 50-node rule (fibroblast and mESC shapes), 760 and 632 for 40 nodes, and 63.4 and 50.4 for 10 nodes. As multiples of the predicted onsets $(N/\pi)^2$ these values are 3.9–4.7 for the two txburst rules and 5.0–6.3 for the BPSC rule. The onset also depends on the Beta shape. At the grid corner with high burst frequency and low k_{off}, for example, the 50-node error is already 1.3×10^{-5} at an r below the median-shape onsets.

Equation (S7) predicts the ratio of onsets between node counts, and the measured ratios follow its N^2 scaling within the scan step. Within txburst the per-pair ratios are $276/399 = 0.69$ (fibroblast) and $230/364 = 0.63$ (mESC) against $(40/50)^2 = 0.64$. Across tools, languages and quadrature libraries (Python with `scipy` against R with `statmod`) they are $15.9/399 = 0.0399$ and $15.9/364 = 0.0437$ against $(10/50)^2 = 0.040$. The measured onsets are 1.4–1.7 times the predicted $(N/\pi)^2$; thus Eq. (S7) sets the scale of the onset and is not an equality.

At $(k_{\text{on}}, k_{\text{off}}, k_{\text{syn}}) = (0.43, 30, 2{,}000)$, the 99th-percentile burst-size point of the published fibroblast fits, the 40-node rule of the published likelihood-ratio test script carries 16% relative error in the count distribution (the 50-node rule 4.8%). The likelihood-ratio statistics of that script contain the rule's quadrature error wherever $r \gtrsim 230$, in addition to the effect of the added constant of §S2.6. We measured only the error of the evaluated distribution at that point and did not quantify its effect on the size of the test. Chen, Sheng, Xie and Jiao recommend maximum likelihood with Gauss–Jacobi quadrature as "the preferred choice for fitting experimental data" [37] on the basis of synthetic tests with synthesis rate $k_b \in [5, 50]$ that included no accuracy comparison

with a reference distribution. Their tested range lies below the 50-node onset $r \approx 253$, whereas the 99th-percentile burst-size point is at $r = 2{,}000$, and at such synthesis rates Eq. (S7) gives the number of nodes that a Gauss–Jacobi rule requires.

S2.4 Failure modes other than quadrature

Two of the failures in Supplementary Table S4 have causes other than quadrature error. The double-precision closed form through `scipy.special.hyp1f1` (row 3) fails because an intermediate factor leaves the representable range of double precision. The routine itself (at the version listed in Supplementary File S7) is accurate, with relative error of about 10^{-12} or better wherever its result is finite. But the Kummer factor $F = {}_1F_1(a{+}n; a{+}b{+}n; -r)$ carries the e^{-r} suppression inside the special function ($F = e^{-r}\,{}_1F_1(b; a{+}b{+}n; r)$ by the transformation of Supplementary File S1), and F therefore underflows double precision once $\log F < -745$ even though $P(n)$ itself is a representable probability.

The underflow of F produces $-\infty$ log-likelihood entries in 53 of 228 cells (92,313 entries over n). At the median-shape grid point with $r = 10^4$, the entries at 2,851 of the 3,001 values of n are $-\infty$. Hence a data set containing any one of those counts has infinite negative log-likelihood there. Near the underflow edge the failure is gradual: loss of subnormal precision yields finite but inaccurate values, up to 94% relative error. Because the underflow happens inside the library routine, assembling the prefactor in log space does not prevent it. Thus row 3 depends entirely on the range and accuracy of the library's ${}_1F_1$ routine, which is the limitation of closed-form evaluation that Gorin and Pachter point out (Supplementary File S1, §S1.8).

The defining alternating series in double precision (row 4b) fails by cancellation. Its terms grow to magnitudes of order e^{2r} relative to the result before they fall off. With the roughly sixteen significant digits of double precision the sum is therefore meaningless once the terms pass 10^{16} times the result, which happens already at $r = 40$, the smallest scan point (relative error 8.9×10^5). The series returns negative values or NaN in 132 of 228 cells where the true value is representable. It fails even at the median fitted point of the fibroblast data: at $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}}) = (0.426, 9.45, 36.7)$ its maximum relative error is 1.3×10^3 with 3 NaN entries. The median of this error over cells, 3.0×10^{-10}, is small only because half the lattice has $r \le 60$, and a spot test confined to small r would not reveal the failure.

S2.5 Likelihood values versus maximizers on synthetic data

The quadrature error is confined to the tail of the fitted parameter region. On all grid cells with $r \le 300$, a range containing about 90 % of the published fits, the 50-node rule has at most 3.1×10^{-9} relative error. The lattice has no cell at $r = 300$; on the finer scan of §S2.3 at the fibroblast shape the error there reaches 1.7×10^{-8}. The added constant changes the log-likelihood of data generated in this range by only about 10^{-9} nats per observation. The upper tail of the fits extends well beyond this range (95th and 99th percentiles of r among the fibroblast fits 871 and 2,382), although the error there tracks burst size more closely than r alone. For example, the fit at the 99th percentile of r, $(2.44, 592, 2382)$, has a small burst size and relative error 8.7×10^{-10}.

Per-gene point estimates change little even on synthetic data sets with large likelihood errors. On ten synthetic data sets of 200 cells drawn from the exact distribution, the data-set negative log-likelihood at the true parameters is in error by 0.21–0.56 nats at parameters within the range of the published fits and by 85.4–104.5 nats at the $r = 10^4$ bound. (Of the two sets at that bound, the one with the larger error had 11 of its 200 counts clamped at the sampler cap of 3,000 and is mildly off-model.) Yet on the four of those sets with the largest likelihood error, the shift between

the maximizer of the published likelihood and that of the exact likelihood is 0.00–0.23 conditional half-widths with a single start and 0.01–0.10 after multi-start polishing. We express these shifts in units of the conditional intervals of the exact likelihood; profile intervals are wider, and in those units the shifts would be smaller still.

In the two synthetic sets at the $r = 10^4$ bound both maximizers, that of the exact likelihood included, lie at $r \approx$ 2,800–3,800, far from the true value. The model is nearly non-identifiable along the (k_{off}, r) ridge there, a property of the likelihood geometry and not of the quadrature, and this near-degeneracy probably limits how far the quadrature error can displace a single-gene estimate. Near an optimum the quadrature error is nearly the same at neighboring parameter values and largely cancels in the comparisons an optimizer makes, but across parameter regions it varies systematically in sign. Hence likelihood values, and the likelihood-ratio statistics, profile intervals and cross-gene comparisons computed from them, carry that error in full, whereas single-gene point estimates shift only at the percent level.

S2.6 The added constant inside the logarithm

The published txburst likelihood adds 10^{-10} to the probability of each observed count inside the logarithm, which caps the negative log-likelihood contribution of one observation at $-\ln 10^{-10} =$ 23.026 nats. On the 228-cell grid the term is non-negligible in 39 cells, those in which the exact $P < 10^{-9}$ somewhere below the quantile cap. Where the model places almost no mass on the counts evaluated, the per-observation log-likelihood is in error by up to 76.4 nats: at $(k_{\text{on}}, k_{\text{off}}, r) = (12.9, 0.3, 10^4)$ the model mean is about 9,773, and the contributions of all counts $n \leq$ 3,000 are capped. For counts drawn from the model itself the contribution is rarely capped; the cap is reached for off-model outliers such as doublets and dropouts, and in the evaluations far from the optimum that an optimizer makes during a fit.

One slice along r through the likelihood surface of the synthetic 200-cell set with true parameters $(0.43, 5.7, 10^4)$ shows the effect of the added constant directly. Along this slice the synthetic set's negative log-likelihood rises steeply toward small r, and below $r \approx$ 2,800 the constant lowers it by up to 10,625 nats (exact NLL = 12,490 against 1,866 for the published likelihood at $r =$ 1,500). Each observation on which the model places almost no mass contributes 23.026 nats there instead of thousands (main text, Fig. 2a). Above $r \approx$ 3,000 on the same slice the published surface is jagged, with non-monotonic excursions of 10–100 nats between adjacent scan points, which is unfavorable to line searches. The two slice minima also differ: $r =$ 3,307 for the published likelihood against 3,579 for the exact one, about 0.6 conditional half-widths in burst size. The profile-likelihood and test scripts of txburst evaluate the same likelihood with the same constant, which therefore affects profile intervals and likelihood-ratio statistics as well. On the Larsson et al. data the most affected gene is *Ywhae* in mESC, all of whose cells have their log-likelihood contribution capped by the constant at the published estimate (main text and Supplementary File S3).

S2.7 Density error, bias and differential expression in BPSC

With $N = 10$, Eq. (S7) gives an onset at $r \approx 10.1$, and most of the lattice lies beyond it. In 50.9% of the lattice (110 of 216 cells) the maximum relative error exceeds 10^{-2}, in 43.5% (94 of 216) it exceeds ten percent, and the worst cell reaches 26.0. These fractions are not driven by the low-probability tail in n: 41.7% of cells have median-over-n error above 10^{-2} and 40.7% have mass-weighted error above it. Of the 12 points from the published fits, 7 exceed 10^{-2} and 5 exceed ten percent, while at the two medians the error is small ($r = 36.7$: 6.8×10^{-4}; $r = 9.4$: 4.2×10^{-8}). The lattice fraction is thus dominated by its high-r half, and among the published fits the tail from

the 95th percentile upward is affected while the median gene is not.

Fits of real data with BPSC's own fitter (`estimateBP`) reach the affected region, where the error exceeds 10^{-2}. With this fitter we obtained, on 20 genes of the Larsson et al. mESC data, $\lambda_1 = 13.5$–74.9 for the ordinary genes (mean expression 4–18), values that straddle the percent-error line at $r \approx 50.4$–63.4, and $\lambda_1 = 337$ (*Malat1*) and 987 (*Abcb10*) well past it. The binned fitting objective of `estimateBP` is also affected, although the 10-node rule conserves total probability and only redistributes it across n: at $(k_{\text{on}}, k_{\text{off}}, r) = (0.426, 9.45, 300)$ with BPSC's own bins the per-bin relative errors run 1%–34% (for example +34% on the bin $(70, 81.7]$, a high-count bin).

Fits in the affected region can be biased by a factor of 2–3: at true parameters $(\lambda_1, \lambda_2) = (300, 1)$, λ_2 being BPSC's count-scale parameter, the published fitter returns $\lambda_1 \approx 90$–170 with $\lambda_2 \approx 1.1$–2.0, compensating along the ridge of constant mean. The quadrature error is confounded here with the fitting criterion, since `estimateBP`'s binned-deviance objective is not maximum likelihood even with an exact density. We therefore attribute the bias to fitting in the affected region and not to quadrature alone.

BPSC's differential-expression P values are nevertheless largely insensitive to the density error, because its test (`BPglm`) does not evaluate the density. The fitted parameters enter the test only through the quasi-likelihood variance shape $V(\mu) = \lambda_2\mu + \phi_2\mu^2$ with $\phi_2 = \beta/(\alpha(\alpha + \beta + 1))$ (file and lines in Supplementary File S7), λ_1 does not enter the generalized linear model, and the fit of that model re-estimates the quasi-dispersion, which absorbs most of the residual parameter error. We probed this mechanism, without attempting a systematic study, in 50 simulated experiments with 200 control and 200 treated cells each and counts drawn from the exact distribution. They comprise 20 null and 20 alternative experiments at $r = 300$ in the affected region, and 5 of each at $r = 8$ as an accurate-region control (25 null and 25 alternative in all).

When we replaced BPSC's fit with a maximum-likelihood fit under the exact likelihood, 0 of 50 calls changed at the 0.05 level: with either fit no null experiment was called significant and all alternative experiments were. The null P values changed little: in the 20 affected-region null experiments the median $|\Delta \log_{10} P|$ is 2.5×10^{-6} and the maximum 2.8×10^{-5}, and the largest across all 25 null experiments is 3.8×10^{-5}. In the 20 affected-region alternative experiments the P values moved by a median of 3.7×10^{-4} in $\log_{10}$. The largest change, 4.76, is between a Welch t-test P and a quasi-likelihood P, because the published fit did not converge in that experiment, in which case `BPglm` falls back to a t-test. Across the other 49 experiments the largest change among the alternative experiments is 7.7×10^{-3}. A further 10 experiments with new random draws gave the same result (0 of 60 calls changed in all). Because the exact-likelihood version changes both the density evaluation and the estimation procedure, this comparison measures their joint effect and does not isolate the effect of the quadrature.

We did not test the goodness-of-fit and model-selection outputs of `estimateBP` (its X^2, P value and Akaike criterion). They use the density directly, and the parameter bias above suggests but does not show that they are affected. We also did not test BPSC's Monte-Carlo null procedure, its zero-inflated five-parameter model, multi-group or batch designs, power and type-I error beyond these 60 experiments, or non-integer normalized inputs (FPKM or TPM). Such inputs are fitted with $\lambda_2 < 1$, which raises the effective λ_1 and may push real fits deeper into the affected region than integer counts suggest.

S2.8 Further accuracy tests and double-precision error control

We tested the evaluator on 19 parameter points not used during its development (Supplementary File S7): 12 random points in the benchmark lattice, four on its edges, two at $k_{\text{syn}} = 10^4$, and a point in the region where Arb's native ${}_1F_1$ returns NaN (Supplementary File S1). We compared the

evaluator's output with two independent evaluations, one in multiprecision arithmetic at 60 decimal digits and one in ball arithmetic at 220 bits, the latter by ball overlap. The minimum agreement is 57.5 digits against a requirement of 40, with the 60-digit precision of the multiprecision evaluation capping the measurable agreement near 60. All ball overlaps hold, and the largest relative radius of any ball is 6.3×10^{-72}.

For the double-precision summation we compute an a priori error bound for each entry and return no double-precision value under four conditions: a bound above 10^{-12}, non-convergence of the series, $\log P < -700$, or any non-finite value. The vector is then recomputed by the multiprecision implementation at two working precisions, whose results must agree at each n. We tested these conditions on the 228-cell benchmark grid, where 44,804 entries in 55 cells deviate from the reference balls by more than 10^{-12}. The number of those cells that meet none of the four conditions, and therefore escape recomputation, is zero. The a priori bound holds entry by entry (worst ratio of deviation to bound 0.652), and 96 of 228 cells (42%) are recomputed. Cells accepted in double precision have maximum deviation 2.15×10^{-13}, and the worst deviation anywhere is 8.96×10^{-12}. The four conditions are empirical safeguards, however, and do not prove that an accepted value is accurate: in an earlier version of the double-precision code, since corrected, a `log1p` argument near -1 at $a \ll b$ produced a deviation of 3.1×10^{-12} at $(k_{\mathrm{on}}, k_{\mathrm{off}}) = (0.03, 999)$ that met none of them.

S2.9 Timing

Wall times are single-core medians over the 216 lattice points for a full vector $n = 0, \ldots, 501-1$ (Supplementary Table S4). The ball reference costs 11.7 ms per 501-point vector, 6.4 times the double-precision summation of this work (1.85 ms) and 8.3 times the published 50-node rule (1.41 ms) in the run reported in the table. In a second run the medians changed by factors of 0.7–1.15 and these two ratios became 5.5 and 5.1, so the ratios should be read as single-digit multiples whose exact values vary between runs. Thus the exact evaluation is several times slower than the 50-node rule but remains fast enough to be run once for each gene of a data set (Supplementary File S3).

Table S4: Six ways of evaluating the telegraph count distribution against the exact reference over the 228-point parameter grid. The two published Gauss–Jacobi rules leave 10^{-6} relative error past $k_{\mathrm{syn}} \approx 364$–$399$ (50 nodes) and 252–276 (40 nodes) at the median fibroblast shape and reach relative errors of 2.90 and 4.78 at the grid corner. The double-precision closed form returns $-\infty$ in 53 cells where the probability is representable, and the naive alternating series fails from the smallest scan point. The all-positive series in double precision (rows 4a and 5) and the ball reference are clean on all 228 cells. Grid, reference, error metric and onset definition as in Methods. "ms" is the single-core median wall time over the grid for a 501-count vector; "largest" and "median error" are the maximum and the median over cells of the per-cell maximum relative error. The next two columns count cells whose error exceeds 10^{-6} and 10^{-2}; the onsets quoted above are the scan brackets of k_{syn} at the median fibroblast shape where the error first exceeds 10^{-6}. Timing is reproducible run to run within a factor below two, so the timing ratios are single-digit multiples, not constants. Hard failure: not-a-number, $-\infty$ or zero, or negative probability at a count where the exact $P(n) \geq 10^{-300}$.

#	method; provenance	ms	largest error	median error	$> 10^{-6}$	$> 10^{-2}$	hard failures
1	txburst maximum likelihood, 50-node Gauss–Jacobi; their code [a]	1.41	2.90	2.0×10^{-10}	75/228	58/228	0
2	txburst hypothesis test, 40-node Gauss–Jacobi; their code [b]	1.14	4.78	1.6×10^{-10}	88/228	62/228	0
3	closed form, general-purpose ${}_1F_1$ in float64; their routine, our assembly [c]	0.20	0.94 (finite)	4.6×10^{-13}	53/228	50/228	$-\infty$ where P is representable in 53 cells (92,313 counts)
4a	log-space Kummer-transformed ${}_1F_1$, float64, the method of [19]; our re-implementation [d]	6.20	1.5×10^{-10}	7.8×10^{-13}	0/228	0/228	0
4b	naive float64 alternating series; nobody's published code [e]	2.80	5.0×10^{270}	3.0×10^{-10}	63/157	62/157	132 cells not-a-number or negative; 71 total loss
5	this work's double-precision summation (positive series, rescaled accumulation); our method [f]	1.85	1.9×10^{-11}	5.2×10^{-13}	0/228	0/228	0
R	ball arithmetic at 150 bits, the reference; our method [g]	11.7	0 (radius $\leq 1.9 \times 10^{-60}$)				0

[a] The zero-change vendoring of the maximum-likelihood script of txburst at the fixed version of Supplementary File S7, confirmed identical to the published script on 20 genes of the repository data (no parameter differs). The table maps the raw quadrature density; the published likelihood also adds 10^{-10} inside the logarithm, whose effect is additional to the error shown here. [b] A line-for-line copy of the density function of the hypothesis-testing script of txburst (Supplementary File S7), its 40-node rule. [c] The special-function call is the double-precision ${}_1F_1$ of scipy (version in Supplementary File S7) on the defining closed form; the prefactor is assembled in logarithms, so every failure in this row is the routine's factor underflowing or losing digits. At the median shape the method is clean over the lower part of the scan and returns $-\infty$ above it (Supplementary File S2). [d] Sukys and Grima call a log-${}_1F_1$ implementation in Julia that we could not run in this environment; this row measures the algorithm (a fully log-space evaluation of the Kummer-transformed all-positive series) and not their binary. [e] Nobody's published code: the defining alternating series summed in double precision, included to quantify why the Kummer transformation is essential. [f] All-positive Kummer-transformed series, log-space prefactor, linear accumulation with per-element rescaling; identical on five comparison cells to the double-precision evaluator used for the error map. [g] The exact evaluator of the main text, used unmodified.

S3 Re-evaluation of the published fits

This Supplementary File reports in full the re-evaluation of the published txburst fits of Larsson et al. [3] in their three combinations of data set and allele (primary mouse fibroblasts with the C57 and with the CAST allele, and mouse embryonic stem cells, mESC, with the CAST allele). The analysis plan was specified before any statistic reported here had been computed, and §S3.1 summarizes it and tabulates the deviations from it. The later subsections report the likelihood values at the published parameters, the point-estimate shifts, the quality-call changes, an exploratory percentile comparison and the genes the published code cannot re-fit, followed by the profile-interval decomposition, the Monte-Carlo coverage of the published intervals and the corrected per-gene tables.

All results in this Supplementary File concern the numerics of the telegraph model as implemented in txburst on these three data sets. We did not test whether any biological conclusion of Larsson et al. changes with the corrected numbers. Throughout, "the quality filter" is the filter built into the published code and "filter-passing genes" are the 20,201 genes it retains (7,191+7,186+5,824 in the three data sets). The CAST-allele mESC count file in the txburst repository lists each of its 94 cells twice (188 columns). The published mESC fits were produced on that doubled matrix and are examined as published, whereas all exact evaluations, re-fits, intervals and per-gene entries reported here use each cell once (main text, Methods). On both sides of every mESC likelihood comparison the likelihood is evaluated on the 94 unique cells. The published parameter values and quality calls, which the authors obtained on the 188-column file, enter only as points at which both likelihoods are evaluated or as calls compared with calls, and they are invariant to the duplication up to optimizer tolerance. No likelihood value, statistic or interval computed on 188 columns is used.

S3.1 Pre-specified design

We tested whether replacing the published likelihood (the 50-node quadrature with 10^{-10} added to each per-cell probability) by the exact likelihood changes the likelihood values, profile intervals, point estimates or quality calls of the published fits. The re-evaluation covers the 32,376 finite published fits in the three combinations of data set and allele (224 fibroblast and 94 mESC cells) and, for re-fitting, the 20,201 filter-passing genes, at the txburst version fixed in Supplementary File S7. Both likelihoods are evaluated under identical parameter conventions (§S1.1). We fixed the analysis plan on 2026-07-08, before any statistic reported below had been computed on the data. Two computations had been run on the data before then, neither of which produced such a statistic: a timing measurement on twenty fibroblast genes whose likelihood values were not used and the comparison of optimizer versions described in Supplementary File S7 (§S7.2). Supplementary Table S5 lists all later changes to the plan.

The first primary endpoint was the likelihood value at the published estimate. For each fit g we defined $D(g) = \mathrm{NLL}_{\mathrm{pub}} - \mathrm{NLL}_{\mathrm{exact}}$, both terms evaluated at the published maximum-likelihood estimate, and called the fit discrepant when $|D|$ exceeds one nat. Calling a fit discrepant required in addition that the exact value had been evaluated in ball arithmetic, with relative radius below 10^{-25} and with a radius below a fixed fraction of $|D|$ (§S3.2). We made that evaluation for all discrepant fits, for all fits with $|D|$ between 0.5 and one nat, and for a random one percent of each data set. We predicted in the plan that discrepant fits would concentrate in the large-synthesis-rate region or among fits with cells on the added constant, with $|D|$ of order 0.5–13 nats. That region is $k_{\mathrm{syn}} \geq 500$, burst size at or above the ninety-fifth percentile of the data set, or a parameter at an optimizer bound (main text, Methods). In particular, we expected almost no discrepant

fits in the accurate region (fits with $k_{\mathrm{syn}} \leq 300$ and without cells on the constant), and specified that discrepant fits found there would trigger the stopping rule described below. We counted the prediction as confirmed if at least 10 discrepant fits existed and at least 80% of them lay in the large-synthesis-rate region or had cells on the constant.

The second primary endpoint was the set of profile intervals of the one data set for which intervals were published, fibroblast CAST (7,186 genes). We separated two effects: the likelihood effect, $\mathrm{CI}(L_{\mathrm{exact}}, t)$ against $\mathrm{CI}(L_{\mathrm{pub}}, t)$ at a fixed threshold t, and the cutoff effect, $\mathrm{CI}(L, 3.8415)$ against $\mathrm{CI}(L, 1.9207)$ under a fixed likelihood L (main text, Methods). All four intervals of a parameter are read off one stored profile curve per likelihood, with no re-fit between thresholds. We called an interval endpoint moved when it changed by more than 10%, and tallied moves above 50% separately. The criterion for the likelihood effect was that at least 5% of the large-synthesis-rate region's profiled genes had an endpoint moved by more than 10%. The criterion for the cutoff effect was that the measured half-width widening of the non-truncated majority agree with the factor 1.414 of the quadratic regime, apart from the deviations expected for ridge-shaped profiles. Because the published intervals had been computed under an older `scipy` version, we also re-ran the published profile code on 200 randomly chosen genes under the current library (§S3.7). Genes not reproduced within 20% were set aside, and if more than 50% had failed we would have replaced the published intervals in the comparison by recomputed ones (the interval-reproduction fallback).

In the first of two secondary analyses we re-fitted all filter-passing genes under both likelihoods with one fixed multi-start scheme and measured the shift between the two estimates in conditional half-widths of the exact likelihood (§S3.3). We expected shifts below one half-width for the large majority of genes, and set the threshold for asserting per-gene misestimation at more than one gene in twenty shifting by more than one half-width after per-gene re-evaluation and ridge separation. Second, we re-applied the published quality filter to the exact re-fits and counted the changed calls by direction and stratum. We specified that if the changed calls exceeded 1% of the filter-passing genes of a data set, each change would be re-evaluated and classified as numerics-driven (the published-likelihood optimum worse than the exact-likelihood optimum by more than 0.1 nats in exact negative log-likelihood), ridge-ambiguous, or non-identifiable with both interval scans truncated at the bounds. In addition, as an exploratory analysis without a decision rule, we compared the burst-size percentiles of the published fits with those of the exact re-fits, with gene-level bootstrap intervals.

Besides the interval-reproduction fallback we distinguished three possible outcomes and expected the default one. In the default outcome, point estimates move by less than their intervals, and the claims then concern likelihood values, tests and intervals, with no per-gene misestimation asserted. In the reduced outcome, which applied if fewer than 10 discrepant fits were found or the regional prediction failed, the statement about likelihood values is dropped and the conclusions rest on the interval results alone. The third outcome, no conclusion about the likelihood evaluation on these data, applied if in addition no profile endpoint in the large-synthesis-rate region moved. All outcomes were subject to a stopping rule with three conditions: a disagreement above 10^{-6} nats at a reported point between the ball value and the double-precision evaluation of Supplementary File S2 (§S2.8), which serves inside the optimizer as a surrogate for the ball value; a ball still failing the 10^{-25} criterion after a precision increase; or discrepant fits in the accurate region. When any of the three occurred, the analysis stopped until the cause was found, and our own code was examined first. Finally, we excluded in advance any statement on cell-to-cell heterogeneity, technical noise or dropout, and credited the three-term recurrence to Willmot [5].

Deviations from the pre-specified plan

The deviation log has eight entries and six clarifications, listed as nine rows in Supplementary Table S5. Seven of the rows list changes or clarifications that settle points of implementation and were made before any result had been computed. The first two of these rows describe the choices on which §S3.7 relies. One entry was made during the analysis before any value was reported, and one (row 8) after all results, when the mESC analysis was re-run with each cell counted once (main text, Methods). A tenth row records the entry made after all results during the review of the manuscript, the profile convention of §S3.9.

The entry made during the analysis (row 7) concerns the one time the stopping rule was triggered. At *Lyz2* in fibroblast C57 (220 cells, one outlier cell at $n = 156$) the ball value of the log-likelihood differed from the correct surrogate value by 1.22×10^{-3} nats. The cause was our own summation step, which converted each ball to double precision before taking the logarithm, and in double precision the outlier cell's probability of 1.5×10^{-322} is subnormal, with about two significant bits. We moved the summation into ball arithmetic, re-evaluated all values accepted until then (largest correction 2.3×10^{-13} nats), and changed the stopping rule: after a surrogate disagreement the likelihood is first recomputed with the multiprecision implementation, and the analysis stops only if that value also disagrees with the ball.

Table S5: Deviations from the pre-specified plan of the re-evaluation (§S3.1). One row per numbered entry of the archived deviation log, in the log's numbering, a row for the log's six clarifications and a last row for the entry made during the review of the manuscript. "Stage" places each entry relative to the endpoints.

No.	Date	Stage	Change	Reason	Effect on results
1	not dated in the log	before any endpoint	Profile scans continued past the published stopping point, which lies below $t = 3.8415$, and each endpoint left undefined on its own when its interpolation fails (the published code then voids both).	The plan's rule that both thresholds are read off one stored curve needs the longer scan.	None on the 200-gene reproduction test, which runs the published procedure without these changes. The endpoint rule applies identically under both likelihoods.
2	not dated (found in a test before the analysis)	before any endpoint	Inner k_{syn} bound of the exact-likelihood profile scans set to 10^4, the maximum-likelihood bound, in place of the profile code's 10^{10}, with scans approaching it marked for the multiprecision implementation.	The double-precision surrogate is tested up to $k_{\mathrm{syn}} = 10^4$ only and overflows far above it.	Published-likelihood scans unchanged. Six genes reached the bound (§S3.7), none affecting the criterion.
3	not dated in the log	before any endpoint	Our copy of the published profile script (Supplementary File S7) added to the input files listed with checksums.	The plan's file list omitted it.	None.

Table S5: Deviations from the pre-specified plan of the re-evaluation (continued).

No.	Date	Stage	Change	Reason	Effect on results
4	not dated in the log	before any endpoint	Scan centers fixed: published-likelihood curves centered at the published estimate, exact-likelihood curves at the exact re-fit.	The plan named no center.	None. Each curve profiles around its own optimum, as in the published procedure.
5	not dated in the log	before any endpoint	The exploratory likelihood-ratio measurement on the 40-node rule of the published test script was not carried out.	Exploratory, without a decision rule.	None. Supplementary File S2 (§S2.3) gives that rule's onset instead.
6	not dated in the log	before any endpoint	"Cell on the added constant" given an operational definition: published per-cell probability at or below 10^{-8}.	The plan used the term without a threshold.	Defines the count of 199 such fits (§S3.2).
7	2026-07-08	during the analysis: likelihood-value and re-fit results existed, no value had been reported	(a) Log-likelihood of the ball values summed in ball arithmetic, no longer after conversion to double precision. (b) After a surrogate disagreement above 10^{-6} nats the likelihood is first recomputed with the multiprecision implementation, and the analysis stops only if that value also disagrees.	The stopping rule was triggered at *Lyz2* (fibroblast C57) by the conversion, which lost accuracy on one subnormal probability. The surrogate and the ball evaluation were correct.	All values accepted until then re-evaluated, largest correction 2.3×10^{-13} nats. Thresholds and endpoints unchanged, no later stop, no precision increase.
8	2026-09-09	after all endpoints	The mESC CAST analysis re-run on a count file with each of the 94 cells once (the repository file lists each twice, 188 columns). The published fits, made on the doubled matrix, examined as published.	Duplicated columns found in the repository file; the repository's later correction covered the C57 file only.	The main text and the Supplementary Files print the de-duplicated values. Endpoints, thresholds, random draws and rules unchanged (main text, Methods).
9	not dated in the log	before any endpoint	Six points the plan left implicit, among them: region membership evaluated among filter-passing genes with each data set's ninety-fifth percentile, each re-fit under one likelihood started from the single-start optimum under the other, and the random sample drawn per data set.	Interpretation only, with no change of procedure.	None. The plan's fallback for the case that the profile scans proved too expensive to compute was not needed.

Table S5: Deviations from the pre-specified plan of the re-evaluation (continued).

No.	Date	Stage	Change	Reason	Effect on results
10	2026-09-26	after all endpoints, during the review of the manuscript	The profile at each fixed value of the profiled quantity taken over the closure of the telegraph family, the negative-binomial limit at $k_{\text{off}} \to \infty$ competing with the interior optimum over the nuisance parameters; endpoints attained at that limit noted, not marked (§S3.9).	The negative-binomial law is a member of the fitted family, and the plan's box on k_{off} is a numerical convenience.	The reference of 3,137 rows moved to the closure minimum, against which 12,512 crossing endpoints lie above the cutoff and are conservative (§S3.9). Endpoint values unchanged.

All pre-specified criteria were met (§S3.2 to §S3.8), the published intervals were reproduced under the current library, and point estimates moved by less than their intervals, which is the default outcome. The quality-call changes were almost all ridge or bound effects, and the exploratory percentile comparison showed no bias attributable to the likelihood evaluation. The verbatim plan, its deviation log and the plan-versus-outcome table are in the archive of Supplementary File S7 (§S7.5) as `S3_plan_verbatim.txt` (SHA-256 `122cca4de0c0…`), `S3_deviations_verbatim.txt` (`0feb0e704527…`) and `S3_plan_vs_outcome_verbatim.txt` (`905107e82fac…`).

Across the re-evaluation we made 76,115 exact evaluations, whose largest relative radius was 4.2×10^{-47} against the 10^{-25} criterion, and the largest deviation of the double-precision surrogate from a ball at a reported point was 1.2×10^{-9} nats against the 10^{-6} threshold. In addition, each gene with a reported likelihood discrepancy, quality-call change, shift above one half-width (including genes outside the filter) or interval-endpoint move was re-evaluated under both likelihoods, with ball evaluation at both optima and at the published point. These re-evaluations number 2,732 in all, with overlaps between the four sets removed.

S3.2 Likelihood values at the published parameters

For all 32,376 finite published fits we evaluated the exact negative log-likelihood of the telegraph model [4] at the published maximum-likelihood estimate itself. The 32,381 rows of the fit files also hold genes whose published fit is undefined (*Pigp* in fibroblast C57 and *Btg2*, *Fut8*, *Ppme1* and *Tgfb1* in fibroblast CAST), which are excluded from all counts. With $D(g) = \text{NLL}_{\text{pub}} - \text{NLL}_{\text{exact}}$ as in §S3.1, where NLL_{pub} is the published code's value at the same point, we call a fit *discrepant* when $|D|$ exceeds one nat with a ball radius below 10^{-3} of $|D|$. At the published parameters we find 205 discrepant fits (88, 86 and 31 in fibroblast C57, fibroblast CAST and mESC CAST; Supplementary Table S6), of which 199 have at least one cell on the added constant, a "capped" cell, operationally one whose published per-cell probability is at or below 10^{-8}.

The likelihood-value criterion of §S3.1 (at least 10 discrepant fits, at least 80% of them in the large-synthesis-rate region or with capped cells) is met: 100% of the discrepant fits lie in that region or have capped cells, or 97.6% under the strictest literal reading. The double-precision evaluation alone already finds 200 of the discrepant fits; the exceptions are five genes with published $k_{\text{syn}} = 2.5 \times 10^4$–$1.4 \times 10^5$, above the range over which that evaluation was tested, for which only the ball value exists.

In 200 of the 205 discrepant fits the published negative log-likelihood is the lower one ($D < 0$), and for the 199 of them with capped cells the mechanism is the added constant and not the

Table S6: Published fits whose log-likelihood at their own parameters differs from the exact value by more than one nat, by data set. $D = \ell_{\text{exact}} - \ell_{\text{published}}$ in negative-log-likelihood units, so $D < 0$ means the published code reports a better fit than the exact likelihood permits; a fit "has cells on the constant" when at least one cell's published probability is small enough for the added 10^{-10} to control its logarithm. In 200 of the 205 fits $D < 0$; 194 of them sit at published $k_{\text{syn}} < 300$ with cells on the constant, and the quadrature mechanism accounts for about six fits at large k_{syn} (largest +61.6 nats). These are statements about likelihood values at the published parameters, not about the estimates; mESC cells are counted once.

data set	fits with $\lvert D\rvert > 1$	with cells on constant	median $\lvert D\rvert$ (nats)	extreme D (nats)
fibroblast, C57 allele	88	85	14.0	−669.5 (*Lyz2*)
fibroblast, CAST allele	86	83	14.4	−201.8 (*Ap2a2*)
mESC, CAST allele	31	31	11.4	−4,826.9 (*Ywhae*)
total	205	199		

quadrature. The 10^{-10} added inside the logarithm replaces the large per-cell penalties that the exact likelihood assigns to counts the model cannot produce at those parameters. Hence the published code reports a log-likelihood above the exact value, by up to 4,826.9 nats (*Ywhae*, mESC, all 94 cells capped). The next largest are *Bub1b* ($D = -1{,}081.1$, 70 of 91 cells capped) and *Lyz2* (−669.5 from a single capped cell, the $n = 156$ outlier among 220). The median $|D|$ among discrepant fits is 14.0, 14.4 and 11.4 nats in the three data sets, whose largest discrepancies are those of *Lyz2*, *Ap2a2* ($D = -201.8$) and *Ywhae* respectively.

We had predicted quadrature under-resolution, with discrepant fits setting in near $k_{\text{syn}} \approx 800$ and $|D|$ of 0.5–13 nats, and the pre-specified criterion was met, but most discrepant fits arose by a different mechanism. The predicted population exists but is small: about six high-k_{syn} discrepant fits, none of them with capped cells, are likely attributable to the quadrature. Their $|D|$ lies between 2.4 and 61.6 nats (the largest for *Nsf*), and five of these fits, including *Cxcl14*, have $D > 0$. Instead, 94.6% of the discrepant fits at the published parameters (194 of 205) are low-k_{syn} genes (published $k_{\text{syn}} < 300$), all 194 with capped cells (Supplementary Fig. S1), and their largest discrepancies far exceed the predicted range. For 98.8% of fits the published likelihood value is accurate to better than 10^{-3} nats (median $|D|$ about 2×10^{-7}), consistent with the estimate of Supplementary File S2 (§S2.5) that about 90% of the published fits need no correction. Thus the discrepancies at the published parameters are confined to an identifiable tail of 205 genes, and any likelihood-ratio test or model comparison that uses the published likelihood values of those genes is affected by the same amounts. These discrepancies concern the likelihood value at the published estimate, whereas the estimates themselves move by less than their intervals (§S3.3).

S3.3 Point-estimate shifts and the *Krt10* case

We re-fitted all 20,201 filter-passing genes under both likelihoods with an identical multi-start procedure and a fixed `scipy` version (Supplementary File S7). In all, 12 genes (0.059%; four, six and two in the three data sets) move by more than one conditional half-width. The median shifts are 2.2×10^{-7}–2.6×10^{-7} half-widths in burst size and 9.7×10^{-7}–4.6×10^{-6} in k_{on}, that is, median burst-size shifts of 6×10^{-6}–7×10^{-6} percent. Because conditional intervals are narrower than profile intervals, this unit can only overstate a shift, and even so 99.94% of filter-passing genes shift by less than one half-width. Under the design of §S3.1 this is the default outcome, and no published parameter value is called misestimated.

The single extreme case is fibroblast-CAST *Krt10* (filter-passing, outside the large-synthesis-

rate region, 127 cells), which shifts k_{on} by 2,274.6 half-widths (Supplementary Fig. S2). The mechanism is a second optimum of the published likelihood at a negative value of the negative log-likelihood, which is impossible for a probability mass function and arises when the quadrature returns per-cell "probabilities" above one. The re-fit under the published likelihood converges to $(k_{\text{on}}, k_{\text{off}}, k_{\text{syn}}) = (75.98, 1000, 1.0)$ with published NLL $= -36.5$, a line-search trial value stored by the optimizer; direct re-evaluation of the published likelihood at that point gives -35.431, also negative. The exact NLL at the same point is 68.6, which is 28.8 nats worse than the optimum of the exact likelihood (NLL = 39.810 at (0.04754, 73.74, 208.86)), and the two estimates of k_{on} differ by a factor of about 1,600. The published quality filter would discard this re-fit, and we describe it only to illustrate the second optimum that the quadrature error creates.

S3.4 Quality-call changes and their decomposition

Re-applying the published quality filter to the exact re-fits changes the call for 5.03%, 4.81% and 6.89% of filter-passing genes in the three data sets (362, 346 and 380 genes, 1,088 in all; for mESC the denominator is the set the committed rule keeps). The published mESC calls were made with an earlier version of the filter, without its $k_{\text{syn}} > 1$ criterion, and differ from the committed rule at 312 genes; the mESC count above applies the committed rule on both sides, whereas the count against the calls as published is 676 (1,384 changes in all). All three fractions exceed the pre-specified 1% reporting threshold (§S3.1). We re-evaluated and classified all 1,384 genes of this as-run set, and Supplementary Table S7 gives their decomposition; the like-for-like mESC set is not decomposed separately.

Of the changes, 772 (55.8%) are ridge-ambiguous: the signed difference ΔN, the exact negative log-likelihood at the published-likelihood optimum minus that at the exact-likelihood optimum, is at most 0.1 nats. Over all changed calls the median $|\Delta N|$ is 7.9×10^{-11} and the ninety-fifth percentile 7.6×10^{-7}. In this class the two optima have, for almost all genes, the same likelihood to numerical accuracy, which suggests that the call changes because small moves along the flat $(k_{\text{off}}, k_{\text{syn}})$ ridge cross the filter's thresholds. A further 606 (43.8%) are practically non-identifiable in the profile-likelihood sense [38], with both interval scans truncated at the optimizer bounds, where the usual chi-square calibration of the interval does not hold [14].

Only 6 genes (0.4% of changes, 0.03% of filter-passing genes), all fibroblast, are numerics-driven, that is, have ΔN above 0.1 nats: *Serpina3g* ($\Delta N = 78.23$ nats, capped cells), *Lrp6* (12.56, capped), *Cd276* (8.94, capped), *Rpl13a* (0.107), *Zc3h6* (0.168) and *Prdm15* (0.106). The three large differences belong to the genes with capped cells and are therefore produced by the added constant (§S3.2). Altogether 17 changes have $|\Delta N| > 0.05$ nats, 10 of them with $\Delta N < 0$, meaning that the exact-likelihood optimum is the worse one and the change does not reflect an error of the published likelihood. In the mESC data set the change with the largest $|\Delta N|$, *Bmp2k* at 0.613 nats, has $\Delta N < 0$ and is classed ridge-ambiguous by the signed rule $\Delta N \leq 0.1$. As a pre-specified descriptive statistic we also computed the fraction of the published fits at the upper k_{off} bound that move more than 0.1% off the bound under the exact likelihood. That fraction is 78.1%, 77.8% and 72.3% in the three data sets.

S3.5 Exploratory percentile comparison

In the like-for-like comparison, with both likelihoods re-fitted under the same fixed `scipy` version, the burst-size percentile differences between the two re-fits are consistent with zero at all percentiles in all three data sets (the gene-level bootstrap intervals include zero). An offset between the published fits themselves and the exact re-fits remains in the mESC tail (ninety-ninth-

Table S7: Quality-filter calls that change when the published quality filter is re-applied to the exact-likelihood re-fit, over the 20,201 genes passing the original filter, and their decomposition with the thresholds of the analysis plan (Supplementary File S3). A change is numerics-driven when the exactly evaluated log-likelihood difference between the two optima exceeds the tolerance of the plan, non-identifiable when the interval scans of both optima truncate at the parameter bounds, and ridge-ambiguous otherwise. All but six of the 1,384 changes are ridge-ambiguous or non-identifiable; six genes change for likelihood reasons. The raw fractions are meaningful only together with this decomposition. The mESC row counts the published calls, which were made with an earlier version of the filter, against the committed rule on the re-fit; with the committed rule applied on both sides the mESC count is 380 (6.89% of the genes that rule keeps) and the pooled count 1,088 (§ S3.4), whose decomposition is not re-derived.

data set	changed	% of passing genes	ridge-ambiguous	non-identifiable	numerics-driven
fibroblast, C57 allele	362	5.03	275	83	4
fibroblast, CAST allele	346	4.81	270	74	2
mESC, CAST allele	676	11.61	227	449	0
total	1,384		772	606	6

percentile difference +9.71 [+3.40, +33.37]), whereas the fibroblast ninety-ninth-percentile offsets, −1.81 [−6.12, +1.55] and −1.14 [−2.65, +4.15], include zero. Because the like-for-like comparison is null, this mESC offset is most likely due to the older optimizer version under which the published fits were made and not to the likelihood evaluation.

S3.6 Genes the published code cannot re-fit

Under the fixed `scipy` version the published code cannot re-fit 46 genes (3, 3 and 40 in the three data sets). When the optimizer reaches large $(k_{\text{on}}, k_{\text{off}})$ inside the parameter bounds, the banded eigenvalue routine that constructs the Gauss–Jacobi nodes overflows and raises an error. The limitation lies in the published code at that library version and does not affect the exact evaluation. None of the 46 passes the quality filter. Therefore the filter-passing statistics are unaffected. Supplementary Table S8 lists these genes, which are excluded from all denominators.

Table S8: The 46 genes the published code cannot re-fit under the fixed library version (§S3.6). None passes the published quality filter.

data set	genes
fibroblast, C57 allele	*Atp11c*, *Chkb*, *Fli1*
fibroblast, CAST allele	*Dvl3*, *Fli1*, *Lipo1*
mESC, CAST allele	*Ado*, *Aoah*, *Arl15*, *Atxn1*, *B4galt7*, *B9d2*, *Bex2*, *Ccdc163*, *Cnnm4*, *Dmd*, *Dnajc6*, *Dpysl2*, *Frat1*, *Gm5779*, *Hhex*, *Hps3*, *Itpripl1*, *Lpcat1*, *Pgrmc1*, *Phospho2*, *Plk3*, *Polg2*, *Praf2*, *Qtrtd1*, *Rfwd2*, *Rnf122*, *Rnf165*, *Rtkn2*, *Sigirr*, *Tmsb15b1*, *Tmsb15b2*, *Tsix*, *Ublcp1*, *Ypel1*, *Zfp280c*, *Zfp317*, *Zfp335*, *Zfp58*, *Zfp617*, *Zfp948*

S3.7 Profile-interval decomposition and its stability

The likelihood effect on the profile intervals is measured on the one Larsson et al. data set with published profile-likelihood output, the 7,186 fibroblast-CAST genes of the published profile table.

For each gene and each of the two profiled parameters we recomputed the profile curve with the published scan procedure (constrained two-parameter re-fits along the parameter grid, cubic or linear endpoint interpolation as published), once under the published likelihood L_{pub} and once under the exact likelihood L_{exact}, each curve centered on its own likelihood's optimum. Each curve is stored once and read at both thresholds $t \in \{1.921, 3.841\}$. The likelihood effect is $\text{CI}(L_{\text{exact}}, t)$ against $\text{CI}(L_{\text{pub}}, t)$ at fixed t. The cutoff effect, $\text{CI}(L, 3.841)$ against $\text{CI}(L, 1.921)$ at fixed L, involves no re-fit between thresholds and therefore has no re-fitting error once the curve is stored. Because the decomposition is in principle path-dependent, we give the likelihood effect at both thresholds and the cutoff effect under both likelihoods below.

Two implementation choices, the first two rows of Supplementary Table S5, enter the profile-interval decomposition. The published scan stops before $t = 3.841$ can be read, and we therefore ran the published step procedure with its stopping criterion evaluated at the higher threshold. In the exact-likelihood scans we bounded the inner k_{syn} search at 10^4, the maximum-likelihood bound (the published scans allow 10^{10}), because the double-precision surrogate for L_{exact} is tested only up to that bound. Six genes reached it (*Akr1c18*, *Cyp1b1*, *Frmd4a*, *Ifit2*, *Mrpl36*, *Pabpc1*), and none of them affects whether the pre-specified criterion below is met. To control for the library version we also re-ran the published profile code on a fixed 200-gene random subsample under the current `scipy`. Of these genes, nine (4.5%) were not reproduced within 20% and are excluded as version-confounded from all later counts (four of the nine would otherwise count as moved). This failure rate is far below the 50% threshold of the pre-specified interval-reproduction fallback (§S3.1).

At the correct cutoff the exact-likelihood endpoints are those of the corrected tables (§S3.9), whose re-certification over the closure of the family moved 926 endpoints of the three tables inward and 42,628 outward, and the published-likelihood endpoints are read off the stored scans as planned. At the correct cutoff $t = 3.841$, 2,066 of 7,177 non-confounded genes (28.8%) have a burst-frequency (k_{on}) or burst-size endpoint moved by more than 10% between the two likelihoods (1,142 genes for k_{on}, 1,574 for burst size), and 309 genes move an endpoint by more than 50%. In the large-synthesis-rate region (§S3.1; the ninety-fifth-percentile burst size is 18.4 in this data set), 413 of 1,030 profiled genes, 40.1%, have a moved endpoint (176 a k_{on} endpoint, 314 a burst-size endpoint). The share thus meets the pre-specified criterion of at least 5%. It also does so under the two alternative treatments of the version-confounded genes, which give 40.3% (413/1,025) when those genes are excluded from the denominator and 40.5% (417/1,030) when their moved genes are counted.

However, only 413 of the 2,066 genes with a moved endpoint at the correct cutoff lie in the large-synthesis-rate region. Thus most moved endpoints belong to genes whose fitted parameters are in the range where the quadrature is accurate. The added constant acting on cells with improbable counts (§S3.2) may account for many of these moves, but we did not classify them by mechanism. On the stored scans alone, with both likelihoods read off the same scan, 1,253 genes move an endpoint at $t = 3.841$ and 1,033 at the published threshold $t = 1.921$ (k_{on} 677, burst size 477), and the decomposition is therefore stable against the threshold.

The cutoff effect measured on the stored profile curves matches the chi-square prediction. The median half-width widening from $t = 1.921$ to $t = 3.841$ under L_{exact} is 1.4388 (interquartile range [1.4138, 1.4691], n = 6,482 k_{on} curves) and 1.4510 ([1.4336, 1.4997], n = 5,874 burst-size curves); under L_{pub} it is 1.4369 (n = 6,204) and 1.4521 (n = 5,780). The prediction in the quadratic regime is $\sqrt{3.8415/1.9207} = 1.414$. The medians are 1.6%–2.7% above it with positive skew, as expected for ridge-shaped profiles, and the interquartile ranges bracket or nearly bracket the prediction. Hence the pre-specified criterion for the cutoff effect (§S3.1) is met. A user cannot recover the corrected intervals by rescaling the published ones with a single factor, because the chi-square widening must be applied to the exact-likelihood curve and for approximately one gene in six that curve itself differs from the published one.

Table S9: Monte-Carlo coverage of nominal 95 % profile-likelihood intervals of burst frequency (k_{on}) and burst size ($k_{\mathrm{syn}}/k_{\mathrm{off}}$) computed by the published interval code, 400 synthetic telegraph data sets of 200 cells at two parameter sets, $(1.0, 10.0, 100.0)$ and $(0.4, 5.0, 200.0)$, inside the region where the 50-node rule is accurate. "Corrected cutoff" is the identical code called with the significance argument that makes its cutoff formula return 3.8415 instead of 1.9207. Data sets for which the code returned an interval endpoint as not-a-number are excluded from the denominator and counted in their own column; that happens when the profile scan reaches the parameter bounds before crossing the cutoff. P is an exact two-sided binomial test against 95 %. Coverage in the large-synthesis-rate region and for bound-hitting fits is not measured here; the burst-size shortfall under the corrected cutoff comes from the 200-cell ridge and from the scan's inner optimizer (Supplementary Table S10, § S3.8).

set	target	cutoff	covered/n	coverage [Wilson 95 %]	no interval	P vs. nominal
A	burst frequency	published	160/197	0.812 [0.752, 0.861]	3	
A	burst frequency	corrected	189/200	0.945 [0.904, 0.969]	0	
A	burst size	published	154/199	0.774 [0.711, 0.827]	1	
A	burst size	corrected	178/199	0.895 [0.844, 0.930]	1	
B	burst frequency	published	135/161	0.839 [0.774, 0.887]	39	
B	burst frequency	corrected	188/200	0.940 [0.898, 0.965]	0	
B	burst size	published	156/200	0.780 [0.718, 0.832]	0	
B	burst size	corrected	181/200	0.905 [0.856, 0.938]	0	
pooled	burst frequency	published	295/358	0.824 [0.781, 0.860]	42	4.6×10^{-18}
pooled	burst frequency	corrected	377/400	0.943 [0.915, 0.961]	0	0.49
pooled	burst size	published	310/399	0.777 [0.734, 0.815]	1	1.1×10^{-32}
pooled	burst size	corrected	359/399	0.900 [0.866, 0.926]	1	4.2×10^{-5}

S3.8 Monte-Carlo coverage per parameter set

To measure coverage, we ran our copy of the published profile-likelihood code (Supplementary File S7, §§S7.2 and S7.3) on 400 synthetic telegraph data sets of 200 cells at two parameter sets where the 50-node rule is accurate, $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}}) = (1.0, 10.0, 100.0)$ (burst size 10) and $(0.4, 5.0, 200.0)$ (burst size 40), 200 replicates each with fixed random draws. The maximum-likelihood start is taken from the published estimator as in the published procedure, and all 400 fits pass the quality filter. Each fit is profiled twice with the same code, at the default level (nominal 95% intervals, cutoff 1.9207) and at $\alpha = 0.005574596680784416$, the value at which the published cutoff formula gives 3.8415, and the two conditions therefore differ only in the cutoff constant. Supplementary Table S9 gives the pooled and per-set coverages with Wilson intervals, and Supplementary Fig. S3 plots them.

Pooled over both parameter sets, the burst-frequency coverage of the published nominal 95% intervals is 0.824 at 200 cells. This coverage is below the nominal 95% ($P = 4.6 \times 10^{-18}$, exact two-sided binomial test) but indistinguishable from 0.8342, the coverage predicted for nominal 95% intervals under the halved cutoff ($P = 0.62$), and correcting only the cutoff restores it to nominal (0.943, $P = 0.49$). The median k_{on} interval-width ratio, published to corrected, is 0.711 against the Wald prediction $\sqrt{1.9207/3.8415} = 0.707$, which suggests that the halving accounts for the shortfall.

The halved cutoff does not account for the whole coverage shortfall of the burst-size intervals. Published burst-size coverage of the nominal 95% intervals, 0.777, is below the halved-cutoff prediction as well ($P = 3.0 \times 10^{-3}$ against 0.8342), and corrected-cutoff coverage is 0.900 [0.866, 0.926],

still below the nominal 95% ($P = 4.2 \times 10^{-5}$). That residual shortfall is a finite-sample effect of the $(k_{\text{off}}, k_{\text{syn}})$ ridge at 200 cells together with the scan's chained inner optimizer, which under-converges along the ridge (below), and comes from neither the cutoff nor the quadrature. The true parameters lie in the accurate region by construction. In 8.3% of the fits the estimate moved along the ridge to fitted $k_{\text{syn}} > 500$, where the 50-node quadrature degrades, but the same fits enter both conditions, and these ridge moves therefore cannot produce a coverage difference between the conditions. Thus the halved cutoff removes roughly 11–17 points of coverage at 200 cells. The simulations test the interval construction only where the quadrature is accurate. At large k_{syn} and at bound-hitting genes, where the quadrature and the added constant also change the profile curve, coverage is likely to be worse, and for that regime we rely on the real-data decomposition of §S3.7 instead.

Under the halved cutoff the published script also stops its scan early, often with fewer than the 4 points that cubic interpolation needs. Our copy then returns undefined (NaN) k_{on} endpoints for 42 of 400 data sets (10.5%, with 3 and 39 in the two parameter sets and none in the corrected condition). Intervals with an undefined endpoint are excluded from the coverage denominators and counted separately in the table. The script as published instead raises an exception at such an early stop under current `numpy` releases, which our copy avoids by a one-line change (Supplementary File S7, §S7.2).

We repeated the study at 94 and 682 cells at the same two parameter points (400 data sets per point and cell number; Supplementary Table S10), adding parametric-bootstrap percentile and basic intervals and a bootstrap-calibrated profile cutoff (200 refits per data set). At 94 cells the corrected-cutoff intervals, as computed by txburst's scan, cover burst frequency in a fraction 0.938 and burst size in 0.909 of cases; the percentile bootstrap covers burst size in 0.936. At 682 cells the scan's burst-frequency interval is undefined for 153 and 398 of 400 data sets at the two points, because the profile crosses the cutoff within two or three of its five-percent steps and the cubic interpolation needs four; its burst-size coverage is 0.880, against 0.958 for the percentile bootstrap. On 199 of the same data sets an exact profile at the same cutoff, computed without the scan, covers the true burst size in 16 cases where the scan's interval misses it and misses it in 1 case where the scan covers it. The remaining burst-size shortfall of the corrected arm at 682 cells therefore comes from the scan's chained inner optimizer, which under-converges along the $(k_{\text{off}}, k_{\text{syn}})$ ridge; only at 94 cells is there a finite-sample shortfall of the profile itself. The undefined intervals are excluded from the coverage denominators and counted separately in the table, which also gives every coverage with undefined intervals counted as misses.

S3.9 The corrected per-gene tables

Corrected per-gene tables for the three combinations of data set and allele are deposited with the code (Supplementary File S7), one table per data set in two formats. The CSV file is for spreadsheet use, and the JSONL file has the same rows plus the per-endpoint numerical detail and a description of how each value was produced. Each row is one filter-passing gene, and together the three tables cover all 20,201 of them. The filter is the one applied to the original fits, held fixed to keep the gene set identical to that of the published tables (the quality-call changes of §S3.4 are a measurement and are not applied).

The columns give, in order, the gene symbol as in the source count matrix, the number of cells used, the exact-likelihood maximum-likelihood estimates of k_{on}, k_{off} and k_{syn} in units of the mRNA degradation rate, and the burst size $k_{\text{syn}}/k_{\text{off}}$ and burst frequency k_{on} at that estimate. Next come the nominal 95% profile-likelihood intervals for burst size and for burst frequency, computed under the exact likelihood with the other two parameters re-optimized at each fixed value of the profiled

quantity. These intervals are read at $2\Delta \log L = 3.841$ in place of the published 1.921, which also removes the cutoff error, and a column naming the interval method follows them. The last columns hold the quality marks (defined in the next paragraph; empty means none) and, in the fibroblast-CAST table only, the intervals published with the original code for comparison. The published point estimate and its quality call, and the exact $D(g)$ at the published estimate, are included in each row of the JSONL file.

A row is unmarked when it carries none of the five marks defined in this paragraph (bound, scan-truncated, non-identifiable ridge, disconnected confidence set and endpoint nuisance-bound); the note nb-limit endpoint of the next paragraph is not a mark. Most rows, 14,067 (69.6%), carry no mark, and the marked rows are left as computed (Supplementary Table S11). The mark *bound* (384 genes) denotes an estimate within 0.1% of an edge of the original code's parameter bounds. The mark *scan-truncated* (6,030 genes) denotes a profile side that did not cross the threshold within the scan range, in which case the endpoint is reported at the last scanned value and the true interval extends at least that far. The mark *non-identifiable ridge* (2,593 genes) denotes both endpoints of the same parameter truncated, with the likelihood flat along the $(k_{\text{off}}, k_{\text{syn}})$ ridge and the parameter not identified at the nominal level. Finally, *disconnected confidence set* (12 genes, 0.06%) denotes a profile-likelihood confidence set that is disconnected at scan resolution, in which case the reported interval is the connected component containing the estimate and the remaining components are listed in the JSONL row. The mark *endpoint nuisance-bound* (698 genes, 707 endpoints) denotes a reported crossing endpoint at which the optimum over the two nuisance parameters, taken inside the optimizer box, sits at an edge of that box. For bound and ridge rows the Wilks calibration behind the nominal 95% does not apply, so those intervals describe the likelihood surface and are not calibrated confidence sets.

The profile at each fixed value of the profiled quantity is taken over the closure of the telegraph family (a deviation from the analysis plan, Supplementary Table S5): the negative-binomial law reached as $k_{\text{off}} \to \infty$ at fixed burst size competes with the interior optimum over the two nuisance parameters, and the lower negative log-likelihood is the profile value. A box on k_{off} is a numerical convenience, and the limit law is a member of the family the data are fitted to. Rows with a reported crossing endpoint attained at that limit carry the note *nb-limit endpoint* (16,615 genes, 31,750 endpoints, 530 of the genes also marked endpoint nuisance-bound). The note records that the data do not bound k_{off} from above at that endpoint; it describes the model there, leaves the interval valid and is not counted as a mark.

Because the profile statistic is not monotone for ridge genes, we located each interval endpoint by outward bisection from the estimate instead of interpolation, which can return a crossing on the far branch. Every crossing endpoint of the tables as first deposited was then re-certified from bracketing-scan and estimate-scaled nuisance starts, with the profile taken over the closure of the family as stated above, and re-located by outward bisection where the certified drop fell short of the cutoff. Of 68,657 endpoints, 25,103 kept their value, 42,628 moved outward, 926 moved inward and 396 reached the scan limit, with 5,582 moved by more than 10% and 830 by more than 50% (4,317 rows). The inward moves, by up to a factor of 16.2 for the lower burst-frequency endpoint of *Timm17b* on the CAST allele, show that the chained scan had placed those endpoints at the wrong location and not only certified them with the wrong likelihood drop: at the deposited value the profile lies above the cutoff, and the interval had been reported wider than the data allow. After the repair every endpoint not at the scan limit satisfied $|2\Delta\ell - 3.8415| \leq 0.02$ against the interior scan minimum of its row. For 3,137 rows the negative-binomial maximum lies more than 0.02 nat below that minimum, and the row's reference is the closure minimum instead; against it 12,512 crossing endpoints (4,519 by more than 0.1 nat) lie above the cutoff, and these intervals, 3,076 of them on unmarked rows, are therefore conservative at the nominal level. Each reported

endpoint has its negative log-likelihood evaluated in ball arithmetic (49,054 endpoints) or, at the negative-binomial limit, in closed form at 50 digits (31,750 endpoints), and no endpoint needed the multiprecision implementation. The ball radii are at most 1.5×10^{-34} nats for the table endpoints and at most 8.8×10^{-35} nats for the likelihood differences of the mESC data set evaluated on the de-duplicated cells, both negligible against the one-nat criterion of §S3.2.

Table S10: Coverage of the true burst frequency and burst size at 94 and 682 cells, 400 synthetic data sets at each of the two parameter sets of Supplementary Table S9, by interval method: txburst's profile scan at the corrected cutoff, parametric-bootstrap percentile and basic intervals and the profile at a bootstrap-calibrated cutoff (200 refits per data set). Coverage is given with the data sets whose interval is undefined excluded from the denominator and, in parentheses, with them counted as misses; the last column counts them. The scan's burst-frequency interval is undefined for most data sets at 682 cells because the profile crosses the cutoff within two or three of its five-percent steps and the cubic interpolation needs four. The burst-size shortfall of the corrected-cutoff scan at 682 cells comes from the scan's inner optimizer, which under-converges along the $(k_{\mathrm{off}}, k_{\mathrm{syn}})$ ridge (§ S3.8).

cells	set	target	method	coverage (undefined as misses)	undefined
94	A	burst frequency	profile, corrected cutoff	0.942 (0.938)	0
94	A	burst frequency	bootstrap percentile	0.962 (0.958)	0
94	A	burst frequency	bootstrap basic	0.920 (0.915)	0
94	A	burst frequency	profile, bootstrap-calibrated cutoff	0.950 (0.945)	0
94	A	burst size	profile, corrected cutoff	0.929 (0.918)	3
94	A	burst size	bootstrap percentile	0.950 (0.945)	0
94	A	burst size	bootstrap basic	0.867 (0.863)	0
94	A	burst size	profile, bootstrap-calibrated cutoff	0.937 (0.925)	3
94	B	burst frequency	profile, corrected cutoff	0.935 (0.930)	2
94	B	burst frequency	bootstrap percentile	0.938 (0.938)	0
94	B	burst frequency	bootstrap basic	0.895 (0.895)	0
94	B	burst frequency	profile, bootstrap-calibrated cutoff	0.935 (0.930)	2
94	B	burst size	profile, corrected cutoff	0.890 (0.890)	0
94	B	burst size	bootstrap percentile	0.923 (0.923)	0
94	B	burst size	bootstrap basic	0.903 (0.903)	0
94	B	burst size	profile, bootstrap-calibrated cutoff	0.895 (0.895)	0
682	A	burst frequency	profile, corrected cutoff	0.968 (0.598)	153
682	A	burst frequency	bootstrap percentile	0.943 (0.943)	0
682	A	burst frequency	bootstrap basic	0.880 (0.880)	0
682	A	burst frequency	profile, bootstrap-calibrated cutoff	0.972 (0.613)	148
682	A	burst size	profile, corrected cutoff	0.845 (0.845)	0
682	A	burst size	bootstrap percentile	0.955 (0.955)	0
682	A	burst size	bootstrap basic	0.865 (0.865)	0
682	A	burst size	profile, bootstrap-calibrated cutoff	0.875 (0.875)	0
682	B	burst frequency	profile, corrected cutoff	1.000 (0.005)	398
682	B	burst frequency	bootstrap percentile	0.953 (0.953)	0
682	B	burst frequency	bootstrap basic	0.938 (0.938)	0
682	B	burst frequency	profile, bootstrap-calibrated cutoff	1.000 (0.033)	387
682	B	burst size	profile, corrected cutoff	0.915 (0.915)	0
682	B	burst size	bootstrap percentile	0.960 (0.960)	0
682	B	burst size	bootstrap basic	0.950 (0.950)	0
682	B	burst size	profile, bootstrap-calibrated cutoff	0.943 (0.943)	0

Table S11: Most rows of the corrected per-gene tables, 14,067 of 20,201 (69.6%), carry no quality mark. Rows per data set carrying each mark of §S3.9; a row is unmarked when it carries none of the five marks, and the nb-limit note (a reported crossing endpoint attained at the negative-binomial limit $k_{\text{off}} \to \infty$, where the data do not bound k_{off} from above) is listed but not counted as a mark. A row can carry several marks.

mark	fibroblast C57	fibroblast CAST	mESC CAST	all
rows	7,191	7,186	5,824	20,201
unmarked	5,736	5,722	2,609	14,067
bound	58	56	270	384
scan-truncated	1,405	1,424	3,201	6,030
non-identifiable ridge	609	590	1,394	2,593
disconnected confidence set	10	2	0	12
endpoint nuisance-bound	129	159	410	698
nb-limit endpoint (note, not a mark)	6,026	5,999	4,590	16,615
nb-limit and endpoint nuisance-bound	113	139	278	530

S4 The allele-resolved analysis

This Supplementary File documents the allele-resolved analysis of the main text. It summarizes the analysis plan specified before any test statistic had been computed and the deviations from it, and reports the pilot on the Larsson et al. matrices and the resulting amendment, which changed only the data set. The later subsections cover the test statistics and calibration criteria with their outcomes, the strata of the significant genes and the class counts under five stratifications, the genome-wide allelic balance, the exclusion of chromosome X and the per-gene results (Supplementary Data S4).

We analyzed the allele-resolved fibroblast data of Johnsson et al. [39], ArrayExpress accession E-MTAB-11054, which comprise 682 Smart-seq3 primary fibroblasts from one female F1 animal of a C57BL/6J × CAST/EiJ cross. Each allelic difference reported here is a contrast of stationary telegraph-model parameters between the two alleles of one gene, in one cell type of one animal, and burst size and burst frequency are quantities of that model. Such a difference is either a strain (cis) effect or a parent-of-origin effect, and the two cannot be distinguished in a cross of a single direction. Residual mapping bias cannot be excluded for any individual gene, and we assign no mechanism to any gene named here.

S4.1 Pre-specified design

For each gene we ask whether the two alleles, each fitted with the exact likelihood, differ in the parameters of the stationary telegraph model. The alleles share the same cells and hence the same trans environment, and a difference may lie in burst size ($k_{\mathrm{syn}}/k_{\mathrm{off}}$), in burst frequency ($k_{\mathrm{on}}$) or in both. We wrote the first plan for the allele-resolved matrices of Larsson et al. [3] (224 fibroblasts, and the 94 embryonic stem cells present in both allele files), after confirming that the two allele files contain the same cells and each comparison is therefore paired within a cell (§S4.2).

We fixed the plan on 2026-07-09, before any likelihood-ratio statistic, P value or class had been computed on real data. Only count-level quantities had been computed by then, namely the pairing comparison, the per-cell chromosome-X allele fractions, the raw allelic fractions of the candidate imprinted genes, and the expression filters with the resulting family sizes. The raw allelic fractions of the imprinted genes carry some information on the direction later tested in the positive controls. However, we took the control list itself from the literature, with the expression filter as the only selection applied to it. We also fixed the order of work: null calibration and power first, since they need only shared-parameter fits and simulations, then the imprinted controls, and the autosomal discovery family last.

The pilot on the Larsson et al. matrices did not meet the calibration criteria (§S4.2), and on the same day we amended the plan by changing only the data set (to the Johnsson et al. fibroblasts). Under the amendment we mapped the Ensembl identifiers of the new matrices to gene symbols by a fixed rule that does not use the counts, dropping genes with no symbol or an ambiguous one, and re-applied the expression filter at the original thresholds (§S4.3). The discovery family is 198 genes drawn from the 13,123 expressed autosomal genes by the sampling scheme of the first plan (at random within strata of total molecule count), with seed and draw order written down before any fitting. In the amendment we also retained the rule excluding chromosome X from the discovery family (§S4.7) and noted that the new data lack a second cell type and a reciprocal cross (§S4.2).

The primary test is the joint likelihood-ratio statistic of §S4.3, which contrasts a six-parameter model with both alleles free against a three-parameter model with all parameters shared. We calibrated it for each gene by parametric bootstrap at the constrained estimate of that gene's own null and controlled the false discovery rate at 5% within each family [40]. Three criteria had to hold before the discovery family was unblinded: (a) at least four fifths of the expressed

imprinted controls, here 16 of 20, significant in the expected direction; (b) null P values uniform by a Kolmogorov–Smirnov test, with the fraction at or below 0.05 inside $[0.02, 0.08]$; and (c) median power of at least 0.5 for a 2-fold burst-size ratio at observed expression. If a criterion failed, the analysis of that cell type ended there, with no provision for re-testing the criterion after changes to the procedure.

We treat the burst-size and burst-frequency statistics, each of which constrains one quantity across the alleles, as secondary. They are calibrated in the same way as the joint statistic and define the classification families (§S4.3), with effect directions read at the free-model estimate. For each fit we noted which parameters lay at their bounds, and a gene whose free fits put k_{on} at the upper bound on both alleles is labeled uninformative for frequency and kept. The plan also designated three gene sets as descriptive only: a second panel of imprinted genes whose imprinting is tissue-restricted, the ten most highly expressed autosomal genes as a stress set for the boundary handling, and the chromosome-X genes.

We handled the known confounds of an allelic contrast as follows. Mapping bias toward the reference genome is small genome-wide and per chromosome at the count level, and the residual bias of individual genes is discussed in §S4.6. Allele-specific capture roughly halves the counts per allele, and the resulting loss of power enters criterion (c) directly, because the power in that criterion is computed at the observed per-allele expression. We took the expected direction of each imprinted control from the literature and set criterion (a) on the whole panel, because a strain effect could mimic the direction of any single control. Transient monoallelic expression in single cells is expected when the two alleles burst independently and is therefore consistent with the model. Cross-allele dependence from extrinsic noise lies outside the model, as do cell-cycle effects shared by the two alleles, and both limit the interpretation of the P values (§S4.6). The caveat marks (§S4.4) label a result and are not grounds for excluding it.

Deviations from the pre-specified plan

The deviation logs of the two plans hold fifteen entries (Supplementary Table S12). Twelve precede all endpoints, one is a scheduling note written afterwards with no effect on any number, one, entered after the significant set was known, gives the stratification adopted in the main text, and the last, entered during the review of the manuscript, records the repair of the equal-burst-size fit (§S4.3). Of the eight entries in the first log, five shrink the computation to a feasible size or substitute a numerically bounded implementation, one settles a detail the plan left open, one is an observational note and one brings the replicate code into line with the plan; none of them changes a threshold, a statistic, the calibration principle, the control lists or a decision rule. The two plans and their logs are in the archive of Supplementary File S7 (§S7.5): the first plan `S4_plan_verbatim.txt` (SHA-256 `1acb5e246579`…) with its deviation log `S4_deviations_verbatim.txt` (`5ff91dd6f71d`…), and the amendment `S4_amendment_verbatim.txt` (`5db18c3966ac`…) with the deviation log of the amended plan `S4_deviations_v2_verbatim.txt` (`310ebcc4b203`…).

Table S12: Deviations from the pre-specified plan of the allele-resolved analysis, one row per entry of the two deviation logs, in log order: the first eight entries are from the log of the first plan, the remaining seven from the log of the amended plan. Date: the date the entry carries. Stage: when the change was made relative to the endpoints, where an endpoint is any of the three calibration criteria or any real-data statistic of the discovery family. Effect on results: where the change enters the numbers of the main text or of this Supplementary File.

No.	Date	Stage	Change	Reason	Effect on results
1	2026-07-09	before any endpoint	Optimizer and bootstrap-replicate fits use a double-precision backward recurrence seeded by two all-positive series sums. Final observed-data likelihoods are recomputed in the evaluator's double-precision mode, with doubtful cases passed to the multiprecision implementation; reported statistics in ball arithmetic, as planned	at the measured cost per call, the planned evaluation mode made the bootstrap's more than a million re-fits infeasible	replicate statistics carry an evaluation error many orders of magnitude below the bootstrap's Monte-Carlo resolution; criterion (b) was evaluated with this substitution in place
2	2026-07-09	before any endpoint	Multi-start scheme replaced by fixed mean-scaled, seeded random and chained starts, with the best few polished, identical for observed data and every replicate; statistics clipped at zero; convergence tolerance loosened but kept far below the resolution the statistics need	computing budget	any effect on the null calibration is measured by criterion (b)
3	2026-07-09	before any endpoint	Pilot discovery family reduced by about one third, to 198 genes per cell type, with the same seed and stratified scheme and the gene lists written before fitting; the ten most highly expressed autosomal genes added as a descriptive stress set	computing budget	fixes the family size later used in the main text; the stress set never enters the discovery family
4	2026-07-09	before any endpoint	Calibration set reduced to a seeded subset of genes across expression quintiles, each simulated under the null repeatedly, with a two-stage inner bootstrap; power computed on a seeded subset of genes for the burst-size (criterion) and burst-frequency (descriptive) alternatives	computing budget; the add-one P value is super-uniform at any replicate number, and the bootstrap is still tested at the fitted null on the reduced set	the tail fraction of criterion (b) is estimated from the reduced number of null data sets, with the corresponding binomial sampling error

continued on the next page

Table S12 (continued)

No.	Date	Stage	Change	Reason	Effect on results
5	2026-07-09	before any endpoint	Third bootstrap stage dropped; replicates capped at the second stage, with a resulting floor on the bootstrap P values	computing budget	Benjamini–Hochberg remains decidable except for a single isolated floor-P gene in a family, which would be reported if it occurred
6	2026-07-09	before any endpoint	Imprinted-control family defined as the primary-panel genes only; the second panel descriptive, with raw P values	a detail left open in the plan	none on criterion (a)
7	2026-07-09	observational note	In the pilot fibroblasts, the reciprocal-cross strata assigned to cells by a leave-one-gene-out vote of the imprinted panel differ slightly in cell count from an earlier count-level tally (tie handling at the panel thresholds)	no action needed; no endpoint involved	none; the per-gene assignment defined in the plan is used
8	2026-07-09	before any endpoint; found during calibration by an independent boundary-null re-run	Replicate fits made to follow exactly the observed-data fit chain (pooled null fit, chained starts, constrained fits). The earlier replicate code started fits at the generating parameters, which made the burst-size null conservative on ridge genes	the plan specified identical optimizer, bounds and starts for observed and simulated data	calibration results computed before the correction discarded and recomputed; no observed-data statistic affected; boundary-null uniformity confirmed on fresh seeds
9	2026-07-09	before any endpoint	Second imprinted panel corrected by the expression filter (*Phlda2* out, *Gpr1* in; still four genes)	the amended plan's text mislisted two genes; the filter is the membership rule	none (descriptive panel)
10	2026-07-09	before any endpoint	Imprinted controls computed before calibration and power had finished; all three criteria judged together afterwards	scheduling; calibration is the longest computation	none; no autosomal real-data statistic existed before all three criteria were met with the statistics as then implemented (the burst-size criterion is not met by the repaired statistic; last entry of this table)
11	2026-07-09	before any fitting	Seed streams and the draw order for the discovery, calibration and power samples fixed and written down before fitting	implementation detail	none

continued on the next page

Table S12 (continued)

No.	Date	Stage	Change	Reason	Effect on results
12	2026-07-10	during the computation, before any criterion was evaluated	Inner recurrence loop compiled and the log-gamma table cached, with the floating-point operation sequence unchanged	the fits of high-count genes ran far slower than budgeted	none; the fits already completed were retained, and the two code paths give identical results on them
13	2026-07-10	written after the endpoints (scheduling note)	Discovery family computed with more parallel workers on a second computer	computing time	none; random seeds are keyed by gene label and independent of the worker count
14	2026-07-10	after the primary endpoint (significant set known)	The four pre-specified caveat marks together with a total-count balance band adopted as the primary stratification of the results (clean, marked but plausible, artifact; §S4.4), after independent spot re-fits had shown that the largest raw effects were artifacts of read mapping or of a missing allele	the plan used marks as labels only, and over the whole significant set the marks alone did not separate the artifacts (§S4.4)	nothing removed: all 198 genes and the family count of 104 are reported, and the class counts under the stratification variants are given in Supplementary Table S13
15	2026-09-24	after all endpoints, during the review of the manuscript	The equal-burst-size fit re-implemented in fiber coordinates that map a box onto the feasible set of the shared burst size and the two k_{syn} values; the burst-size statistic, its bootstrap P values, the burst-size legs of the calibration criteria and the class counts recomputed	the implementation used until then stopped at its starting point for 135 of 198 genes, because a penalty outside the k_{syn} bounds zeroed the finite-difference gradient on the first step (§S4.3)	the burst-size family and the class counts of the main text (31 of 104 hits change class); the joint and burst-frequency statistics are byte-identical

S4.2 The pilot and the change of data set

We ran the pilot on the allele-resolved matrices of Larsson et al. [3] for fibroblasts and for mouse embryonic stem cells. The pilot did not meet the calibration criteria in either cell type, and we stopped it without computing any statistic on its discovery family. In the stem cells 5 of the 12 imprinted positive controls were recovered. In fibroblasts the null P values of the burst-size statistic failed the uniformity test (Kolmogorov–Smirnov $P = 0.0043$), with too few mid-range P values, that is, in the conservative direction. The median power for a 2-fold burst-size effect at observed expression was 0.30 in the stem cells and 0.38 in fibroblasts, below the required one half.

The Johnsson et al. fibroblasts analyzed instead have 3.85 times the per-cell, per-allele depth of the pilot fibroblasts. They have the paired same-cell design of the pilot data, with identical cell columns and identical missing-entry masks across the two allele files, and the power reached on them is reported in §S4.3. The change of data set removed the second cell type and the reciprocal cross from the original design. The second cell type has no counterpart in the new data, so the cross-cell-type replication was dropped. Thus all results concern fibroblasts alone. In addition, the new data come from one animal of one cross direction, without the reciprocal cross of the first plan, and a strain (cis) effect therefore cannot be distinguished from a parent-of-origin effect.

S4.3 Filters, statistics, calibration criteria and their outcomes

A gene enters testing when it has at least 25 cells with allele-informative entries and at least 50 summed molecules over both alleles, and chromosome X is excluded (§S4.7). The discovery family is the stratified random sample of 198 autosomal genes from the expressed pool described in §S4.1. For each gene we fit the two alleles' counts on the non-missing paired cells under four telegraph models: both alleles free (six parameters), all parameters shared (three), equal burst size $k_{\mathrm{syn}}/k_{\mathrm{off}}$ with the rest free (five), and equal burst frequency k_{on} with the rest free (five). The primary statistic is the joint likelihood-ratio contrast of the free model against the shared model. The burst-size and burst-frequency statistics contrast the free model against the two five-parameter models and define the two classification families.

In place of chi-square asymptotics, we calibrate each statistic per gene by parametric bootstrap at the constrained maximum-likelihood estimate of its own null, with the add-one P value. The bootstrap remains valid for fits at a parameter bound or on a likelihood ridge, where the chi-square approximation does not hold. Replicates are drawn in two stages, the second of which is run only for genes with a small P value after the first. All three families are Benjamini–Hochberg corrected at false discovery rate 5% within family [40]. A gene significant in the joint statistic is called size-resolved when only the burst-size statistic is also significant, frequency-resolved when only the burst-frequency statistic is, resolved on both when both are, and unresolved when neither is. The bootstrap draws the two alleles independently and the cells independently from the constrained fit, and it therefore checks the numerics and the tests under that model class; shared cell state is addressed by the within-cell label-swap test of §S4.9. The equal-burst-size fit is optimized in fiber coordinates that map a box onto the feasible set of the shared burst size and the two k_{syn} values (the last entry of Supplementary Table S12): the implementation used before the review stopped at its starting point for 135 of 198 genes because a penalty outside the k_{syn} bounds zeroed the finite-difference gradient on the first step. Every burst-size statistic, its bootstrap P value, the burst-size legs of the calibration criteria and the class counts were recomputed with the repaired fit (31 of the 104 hits change class). The joint and burst-frequency statistics, their fits and their bootstrap draws are identical under both implementations.

We computed the discovery-family statistics only after the three calibration criteria had been met on this data set with the statistics as then implemented. All 20 of the 20 imprinted positive-control genes of the primary panel were significant in the expected direction (the criterion was at least 16 of 20). On 500 simulated-null data sets the P values of all three statistics were uniform by the Kolmogorov–Smirnov test ($P = 0.33$, 0.39 and 0.28 for the joint, burst-frequency and burst-size statistics), and the fraction of null P values at or below 0.05 lay between 0.048 and 0.054 (required band $[0.02, 0.08]$). The median power over 30 genes for a 2-fold burst-frequency effect at observed expression was 0.74, against a criterion of 0.5; for the matching burst-size effect it is 0.36 with the repaired statistic (0.87 with the defective one, whose null quantiles were too large). The burst-size test is therefore the less powerful of the two and does not meet the criterion. The null P values of the repaired burst-size statistic on the same 500 simulated sets are uniform, with a fraction 0.050 at or below 0.05, and 4 of the 198 genes are significant in the repaired burst-size family (33 with the defective statistic). The observed-data likelihoods entering the three statistics were evaluated in ball arithmetic, with zero failed evaluations among the 198 genes of the family. At identical parameters the double-precision values used during the search agree with the ball-evaluated statistics to 1.3×10^{-7} nats (1.7×10^{-10} for the evaluator's guarded route); 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 $P = 1$).

S4.4 Stratification of the significant genes and artifact diagnosis

By the joint statistic, 104 of 198 genes are significant at false discovery rate 5%. This is a family-level count, and the conclusions of the main text rest on the *clean* stratum, whose defining rule we state here in full. Its inputs are the four pre-specified caveat marks (boundary fit, low expression, possible mapping dropout at an allelic mean-expression ratio above 50, and ridge) together with the band $[0.3, 0.7]$ on the total-count C57 fraction. A hit is *clean* when it carries none of the four marks and its balance lies inside the band. It is an *artifact* when it carries the mapping-dropout mark, or when it lies outside the band with a boundary mark, a low-expression mark or $|\log_2$ burst-size effect$| > 2$. The remaining hits form the *marked but plausible* stratum of Table 2 of the main text.

We defined the stratification rule after the significant set was known, once spot re-fits had shown that the largest raw effects in the family were artifacts of read mapping or of the near-absence of one allele (the last entry of Supplementary Table S12), and applied it uniformly to all 104 hits. An earlier version of its artifact clause used the maximum $|\log_2|$ effect over both parameters in place of the burst-size effect. The clean count is 77 under either version. All 198 genes and all 104 hits are reported with their stratum (Supplementary Data S4), and, as specified in advance, marked genes are not presented individually as primary results.

The strata are 77 clean, 15 marked but plausible and 12 artifact (main text, Table 2). Among the 77 clean hits the one-degree tests resolve 19, 17 as frequency-resolved, 2 on both axes and none as size-resolved, and 58 clean hits are unresolved. Typical clean allelic effects are modest: the median classified effect is $|\log_2| = 1.07$, an unsigned allelic ratio of about twofold (2.1) in the classified model quantity, with per-class medians 1.07 (frequency-resolved) and 1.73 (both). The 58 clean unresolved hits have small or mixed effects (median $\max|\log_2| = 0.53$), consistent with the limited power of the one-degree tests for smaller effects at this sequencing depth (the median powers given in §S4.3 are for a twofold effect). Because these genes carry no mark, their unresolved status is more likely due to small effects than to a fitting artifact. Across all 104 hits, 55 are unresolved, of which 58 are clean and 33 lie below $|\log_2| = 1$ (effect sizes by stratum in Supplementary Fig. S4).

Count-imbalance artifacts. The genes with the largest raw effects in the family ($|\log_2|$ up to 16.4) are pseudogenes and predicted genes in which one allele is almost absent from the data, and the maximum-likelihood fit represents that absence as extreme kinetics. For one of these genes, *Gm26563*, neither of the two marks intended to detect such artifacts (mapping dropout and boundary) is set. This gene has 6 of its 149 molecules on CAST, an allelic mean ratio of 23.9, below the dropout threshold of 50 in that ratio, and a fitted $\log_2$ burst-size ratio of +14.9, a value that lies on a flat, degenerate ridge. An independent spot re-fit moved the CAST-allele fit by 0.08 in $\log_{10} k_{\text{on}}$ and 0.30 in $\log_{10} k_{\text{off}}$ and placed k_{syn} exactly at its lower bound, at a likelihood within 1.4×10^{-5} nats of the reported fit, while the burst-size ratio moved by about one $\log_2$ unit between the two equally good optima. Hence whether the boundary mark is set for such a gene depends on where along that ridge the optimizer stops: the reported fit lay 0.048 (in $\log_{10} k_{\text{syn}}$) above the lower bound, and the mark was not set.

The low-expression mark of *Gm26563* is set and by itself keeps the gene out of the clean stratum, independently of the two artifact-diagnostic marks that failed here. However, the gene lies outside the total-count balance band, as does any gene with one nearly absent allele, and this observation led us to add the band to the stratification post hoc. *Rps6-ps1*, the second case, has 36 of its 10,897 molecules on C57, an allelic mean ratio of 301, above the dropout threshold, and its C57 fit places k_{off} at the upper bound and k_{on} at $10^{-2.16}$. Its nominal frequency label most likely reflects allelic mapping dropout expressed as a boundary-k_{on} fit. Therefore we do not use the six resolved labels of the artifact stratum when comparing burst-size and burst-frequency modulation.

Table S13: Class counts of the significant genes under four stratification variants with the repaired burst-size statistic (frequency-resolved f, size-resolved s, both b, unresolved u; the counts with the statistic as first implemented in parentheses). Under every variant the frequency-resolved count exceeds the size-resolved count; the stratification changes the size of the quoted effects, which the marked strata inflate (Supplementary Fig. S4). The a priori exclusion by pseudogene or predicted-gene name was not recomputed with the repaired statistic. What the split can say about the mixture of mechanisms is quantified in § S4.10.

stratification	n	f	s	b	u
V1: all significant genes, no stratification	104	23 (16)	1 (23)	3 (10)	77 (55)
V2: genes with a fit mark removed	86	19 (12)	0 (19)	3 (10)	64 (45)
V3: count-balance outliers removed	83	18 (12)	0 (20)	2 (8)	63 (43)
V4: clean stratum (primary; both removed)	77	17 (11)	0 (19)	2 (8)	58 (39)

The compensating ridge. Zranb3 has the largest effect in the clean stratum, with a $\log_2$ burst-size ratio of -5.9 against a compensating $+5.5$ in $\log_2 k_{\text{on}}$ (both C57 relative to CAST), balanced counts (C57 fraction 0.495) and no marks. Neither one-degree test survives the Benjamini–Hochberg correction within its family, and under the pre-specified classification rule the gene is therefore unresolved. Because the burst-size and burst-frequency estimates trade off against each other along the (k_{on}, burst size) ridge, either one quoted alone may overstate the allelic difference.

S4.5 Class counts under alternative stratifications

We repeated the split under the stratification variants of Supplementary Table S13 with the repaired burst-size statistic. With no stratification the 104-gene family splits into 23 frequency-resolved, 1 size-resolved, 3 both and 77 unresolved; the pre-specified marks alone give 19 / 0 / 3 / 64 (of 86 hits), the balance band alone 18 / 0 / 2 / 63 (of 83), and the clean set 17 / 0 / 2 / 58. Under every recomputed variant the frequency-resolved count exceeds the size-resolved count; the a priori exclusion by pseudogene name was not recomputed with the repaired statistic. Excluding the marked fits changes the quoted effect magnitudes, which the artifact strata inflate. The table concerns the choice of stratification only; the planted simulation of §S4.10 quantifies what the split can say about the mixture of mechanisms.

S4.6 Allelic balance reference and mapping bias

The total-count C57 fractions of the 77 clean hits lie inside $[0.3, 0.7]$ by definition and fall at genome-wide percentiles 0.079–0.921, none in the extreme two-percent tails. The reference distribution is the per-gene total-count C57 fraction over the 13,168 autosomal genes with at least 50 molecules (median 0.4993, with 13.4% outside the band). Over the whole data set we measure a C57 fraction of allele-informative molecules of 0.508 (autosomes only 0.495), and global reference bias is therefore small compared with the effects reported. The per-gene spread of this fraction mixes true cis effects, imprinting and residual mapping bias, which these data cannot separate gene by gene (§S4.2). Hence residual mapping bias, including reference bias against the non-reference allele at variant positions, may affect any individual call; we did not attempt a per-gene correction, and genes marked possible mapping dropout are not reported as primary results. Cross-allele dependence from extrinsic noise, which neither the independent-alleles likelihood nor its parametric bootstrap accounts for, is small but present: the genome-wide median cross-allele Spearman correlation is $+0.012$ over 10,890 autosomal genes. Among the 77 clean hits the partial Spearman correlation

of the two alleles' counts given sequencing depth has median −0.042 and is significant at the 0.05 level for a fraction 0.66 of them under a within-decile permutation null, with both signs; adjusting also for the first expression principal component and plate leaves the median at −0.050. The label-swap test of §S4.9, exact under within-cell exchangeability, checks the bootstrap's independence assumption, and the 14 hits it removes are reported with that qualification.

S4.7 The X chromosome

We had planned to extend the analysis to X-linked genes that escape X inactivation, with burst kinetics conditioned on each cell's inactive X. This extension was dropped before any fitting, because all 682 cells are C57-X-monoallelic (median per-cell chromosome-X C57 fraction 0.975) and *Xist*-negative: the population has no X-inactivation mosaic, and there is no inactive-X state to condition on. For the same reason chromosome X was excluded from the discovery family by rule: X-linked allelic imbalance in these cells is a whole-chromosome state, not gene-level regulation, and with no CAST-X-active cells the two cannot be separated here. The only allele-resolved fibroblast data set available to us with an X-inactivation mosaic is the 224-cell set of Larsson et al. [3] re-evaluated in Supplementary File S3.

S4.8 Per-gene results (Supplementary Data S4)

Supplementary Data S4 is a flat table with one row per gene of the discovery family (198 rows), deposited with the code (Supplementary File S7). Its columns are the gene symbol, the number of paired cells used, and the per-allele maximum-likelihood estimates of k_{on}, k_{off}, k_{syn}, burst size and burst frequency, followed by the three likelihood-ratio statistics with their bootstrap P values, replicate counts and Benjamini–Hochberg adjusted P values. The remaining columns give the class, the four caveat marks, the total-count C57 fraction and the stratum. Supplementary Table S14 lists the 77 clean genes with their class and effect.

S4.9 Within-cell label swap, mapping-ratio sensitivity and goodness of fit

The bootstrap of §S4.3 simulates the two alleles of a cell independently. As a check that makes no distributional assumption about shared cell state, we permuted the allele labels within each cell (2,000 swaps per gene) and recomputed the joint statistic, a test that is exact when the two alleles of a cell are exchangeable under the null. The joint statistic was reproduced for all 104 hits. With Benjamini–Hochberg control at 5% applied to the label-swap P values, 90 of the 104 hits remain significant and 14 lose significance (12 of them clean; Supplementary Table S15); 4 hits have a label-swap P above 0.05 (3 clean). The two calibrations disagree on a fraction 0.13 of the hits, below the fraction at which we had undertaken to recalibrate with a paired generative model, and the Spearman correlation of their $\log P$ values is 0.77. Reported both ways, the clean split is 17 frequency-resolved, 0 size-resolved, 2 both and 58 unresolved with the bootstrap calibration; removing the 3 clean genes whose label-swap P exceeds 0.05 gives 16 / 1 / 57 (of 74), and removing the 12 clean genes that lose significance gives 15 / 1 / 49 (of 65). *Dcaf10*, one of the two genes resolved on both axes, is in both removed sets.

An allele-specific capture or mapping efficiency ratio $\rho = \beta_{\mathrm{C}}/\beta_{\mathrm{A}}$ enters the likelihood as binomial thinning of one allele and multiplies the inferred burst-size ratio by ρ exactly; a burst-size call whose effect is smaller than $|\log_2 \rho|$ therefore cannot be separated from mapping bias. For the 2 clean genes significant in the repaired burst-size family (*Dcaf10* and *Slc39a14*) we recomputed the burst-size statistic with one allele thinned by 10% and by 25% in either direction, against the null distribution regenerated at $\rho = 1$ with the seed of the main analysis and taken as invariant in ρ: zero of the

Table S14: The 77 clean significant genes of the allele-resolved analysis (no artifact mark; total-count C57 fraction inside $[0.3, 0.7]$), grouped by the call the one-degree tests return with the repaired burst-size statistic (17 frequency-resolved, 2 resolved on both axes, 0 size-resolved, 58 unresolved). Effects are $\log_2$ ratios (C57 over CAST allele) of the fitted burst size $k_{\mathrm{syn}}/k_{\mathrm{off}}$ and burst frequency k_{on}; the last column is the C57 share of allele-informative molecules. Each entry contrasts telegraph-model parameters between the alleles of one single-direction cross; strain and parent-of-origin effects are not distinguished and no mechanism is assigned.

gene	call	$\log_2$ size	$\log_2$ freq.	C57 frac.	gene	call	$\log_2$ size	$\log_2$ freq.	C57 frac.
Ampd3	frequency	−0.06	−1.00	0.33	*En1*	unresolved	+1.34	−0.34	0.63
Ankra2	frequency	+0.71	−1.27	0.38	*Ephx1*	unresolved	+0.18	−0.35	0.48
C1galt1	frequency	+0.98	−1.43	0.38	*Ghdc*	unresolved	+0.12	+0.34	0.58
Ccn5	frequency	+0.03	−0.45	0.44	*Glrx2*	unresolved	+0.17	−0.55	0.44
Dhx40	frequency	−1.01	+1.00	0.62	*Gm16033*	unresolved	+0.74	−0.19	0.60
Emb	frequency	−0.71	+1.87	0.66	*Gnb1l*	unresolved	+0.47	−1.03	0.38
Erlin2	frequency	−2.08	+1.07	0.48	*Hnrnpul1*	unresolved	−1.48	+0.22	0.41
Gemin6	frequency	−0.14	+0.64	0.59	*Ibtk*	unresolved	+0.05	+0.29	0.56
Gmpr2	frequency	+0.59	−0.73	0.44	*Jam2*	unresolved	−0.79	−0.03	0.38
Lekr1	frequency	−2.33	+3.09	0.66	*Kifc5b*	unresolved	−0.53	−0.20	0.37
Lrrc4c	frequency	−0.33	+1.34	0.67	*Lrrc8d*	unresolved	−0.60	+0.11	0.45
Mgmt	frequency	+0.39	−0.76	0.43	*Man1b1*	unresolved	+1.07	−0.79	0.46
Nle1	frequency	+0.27	−0.79	0.42	*Nbr1*	unresolved	+0.79	−0.42	0.54
Rexo1	frequency	+1.95	−1.70	0.44	*Ndufc2*	unresolved	+0.31	+0.01	0.55
Slc41a3	frequency	+1.02	−2.03	0.33	*Ndufv1*	unresolved	+0.03	+0.11	0.52
Thsd1	frequency	−0.74	+1.07	0.56	*Nob1*	unresolved	−0.06	+0.23	0.53
Thyn1	frequency	−0.39	+0.87	0.61	*Numb*	unresolved	+0.05	−0.28	0.45
Dcaf10	both	+1.84	−1.38	0.47	*Pcbp4*	unresolved	+0.06	−0.28	0.45
Slc39a14	both	−1.61	+1.19	0.53	*Phlpp1*	unresolved	−1.13	+1.37	0.58
3110056K07Rik	unresolved	−0.96	+1.73	0.65	*Pigs*	unresolved	−0.59	+0.10	0.45
Adamts5	unresolved	+0.00	+0.25	0.55	*Pqlc3*	unresolved	−0.07	+0.39	0.53
Afg1l	unresolved	−0.79	+1.46	0.65	*Psmc1*	unresolved	−0.20	−0.04	0.47
Aggf1	unresolved	−0.81	+1.17	0.59	*Rbfa*	unresolved	+0.58	+0.30	0.59
Alad	unresolved	+0.25	−0.40	0.44	*Rps6kc1*	unresolved	−0.00	+0.46	0.58
Aqr	unresolved	−0.33	+0.56	0.53	*Runx1t1*	unresolved	+0.52	+0.26	0.61
Arl2bp	unresolved	−0.53	+0.12	0.50	*Scarf2*	unresolved	−0.25	−0.64	0.35
Atp13a3	unresolved	−0.33	−0.11	0.47	*Sdhaf3*	unresolved	+0.23	−0.92	0.38
Bbx	unresolved	−0.00	+0.35	0.56	*Sem1*	unresolved	+0.09	+0.15	0.54
Bcdin3d	unresolved	+0.21	+0.47	0.60	*Sgcb*	unresolved	−0.23	−0.38	0.44
Castor2	unresolved	+1.30	−0.36	0.63	*Smim14*	unresolved	+0.44	+0.02	0.56
Cbr4	unresolved	+0.22	+0.19	0.56	*Syt12*	unresolved	−0.92	−0.10	0.31
Ccdc115	unresolved	−0.55	+0.96	0.58	*Tha1*	unresolved	−0.89	+1.37	0.60
Chd4	unresolved	−0.33	−0.03	0.48	*Tmsb10*	unresolved	−0.13	−0.14	0.45
Commd7	unresolved	+1.21	−0.44	0.54	*Tst*	unresolved	−0.34	+0.49	0.55
Cotl1	unresolved	−0.04	+0.24	0.54	*Usp7*	unresolved	+0.04	−0.37	0.46
Crls1	unresolved	+0.32	−0.73	0.42	*Wdyhv1*	unresolved	+2.17	−0.47	0.65
Cyp20a1	unresolved	−1.16	+0.82	0.53	*Xrn2*	unresolved	+0.69	−0.20	0.47
Dpp8	unresolved	+0.11	−0.36	0.45	*Zranb3*	unresolved	−5.90	+5.51	0.50
Dtnbp1	unresolved	+0.09	+0.46	0.57					

Table S15: The 14 joint-test hits that lose Benjamini–Hochberg significance at 5% when the bootstrap P value is replaced by the within-cell label-swap P value (2,000 swaps per gene; § S4.9), with the joint statistic, both P values, the stratum and the call under the repaired burst-size statistic. The remaining 90 hits keep their significance under both calibrations.

gene	T_3	bootstrap P	label-swap P	stratum	call
Ephx1	13.4	0.0064	0.0235	clean	unresolved
Pcdhga8	8.6	0.0164	0.0270	marked, plausible	unresolved
Usp7	10.8	0.0141	0.0300	clean	unresolved
Bcdin3d	14.0	0.0027	0.0300	clean	unresolved
Rexo1	19.4	9.1×10^{-4}	0.0310	clean	frequency-resolved
Arl2bp	9.1	0.0241	0.0340	clean	unresolved
Zranb3	10.2	0.0191	0.0350	clean	unresolved
Aqr	12.4	0.0064	0.0415	clean	unresolved
Man1b1	16.4	4.5×10^{-4}	0.0430	clean	unresolved
Tst	9.3	0.0236	0.0450	clean	unresolved
Dcaf10	18.2	4.5×10^{-4}	0.0650	clean	both
Lrrc8d	10.9	0.0127	0.0775	clean	unresolved
Myg1	17.6	0.0014	0.0835	marked, plausible	unresolved
Erlin2	12.2	0.0068	0.1364	clean	frequency-resolved

two calls survive at either ratio, and both genes are reported as within mapping-ratio uncertainty. The two calls are marginal at $\rho = 1$ (*Dcaf10* sits at the family cut), and the losses at 10% amount to one or two exceedances on a 2,198-draw null; the result holds at Monte-Carlo resolution only.

Goodness of fit of the telegraph model to each allele of the 77 clean genes was assessed by parametric bootstrap (200 replicates per gene and allele, scored at the observed-data estimate, which makes the marks conservative) on six statistics; Pearson's chi-square marks 10 of the 77 genes at a false-discovery level of 0.10, and the other five statistics mark none.

S4.10 Planted-mixture simulation of the full procedure

To learn what the split of the clean calls can say about the mixture of mechanisms, we ran the whole procedure, from the per-gene fits through the three bootstrap families, the artifact marks, the balance band, the class rule and the strata, on synthetic experiments of 198 genes with planted allelic differences. Each experiment plants differences in 104 genes, with effect sizes resampled from the observed clean effects (capped at their ninety-fifth percentile, with a floor), under four compositions: all frequency, all size, an equal mixture, and a mixture in which a quarter of the differences are in k_{off} alone (which changes burst size and burst frequency together) and the rest in both parameters. Each composition was run for 15 experiments with 199 first-stage replicates per statistic, without the unconditional second stage, and floor genes were refined where a family's count depended on them. Supplementary Table S16 gives, for each composition, the planted truth against the returned call. 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 (0.24 of the frequency recall). An equal mixture yields on average 0.33 size-resolved and 6.67 frequency-resolved clean calls per experiment (Table 2 of the main text), and the observed split, 0 size-resolved against 17 frequency-resolved (Wilson 95% interval on the size share 0.00 to 0.18), is at least as uneven as the split of a fraction 0.79 of the equal-mixture experiments (Wilson 95% interval 0.52 to 0.92). A k_{off}-only difference is returned as unresolved in 98 and as not significant in 283 of 390 planted cases and as size-resolved in zero. The balance band, a cut on the sum of the

Table S16: Planted truth against returned call in the planted-mixture simulation of the full procedure (§ S4.10): counts of gene-experiments over 15 experiments per composition. Calls are the clean-stratum calls (size-resolved, frequency-resolved, both, unresolved), 'n.s.' is not significant on the joint test, and 'marked' is significant but removed from the clean stratum by an artifact mark or the balance band. A planted size difference is returned as a clean size-resolved call in a fraction 0.034 of cases and a planted frequency difference as a frequency-resolved call in 0.142; a k_{off}-only difference, which changes burst size and burst frequency together, is returned as size-resolved in zero of 390 cases.

composition	planted	size	frequency	both	unresolved	n.s.	marked
all frequency	no difference	1	1	0	15	1390	3
	frequency	0	222	3	577	599	159
all size	no difference	0	0	0	14	1392	4
	size	53	0	0	723	554	230
equal mixture	no difference	0	0	0	16	1389	5
	frequency	0	99	0	297	307	77
	size	5	1	0	379	270	125
k_{off} only (a quarter), both (the rest)	no difference	0	0	1	16	1392	1
	both	18	71	17	464	280	320
	k_{off} only	0	2	0	98	283	7

two $\log_2$ ratios, also removes large same-sign changes, most often for mixed effects. The split of the clean calls is therefore an outcome of the procedure's unequal recall of the two kinds of difference and does not measure their prevalence.

S5 Accuracy and separability of the two-step OFF model

The two-step (refractory, three-state) promoter model of the main text, whose stationary distribution and likelihood-ratio test are derived in Supplementary File S1 (§§ S1.4–S1.7), is the cycle ON $\to$ OFF$_2$ $\to$ OFF$_1$ $\to$ ON with rates k_{off}, k_{12}, k_{on} and synthesis rate ν, all in units of the mRNA degradation rate. This Supplementary File reports the accuracy of the three-state likelihood evaluation (Supplementary Table S17), the two reference parameter sets, the computation of the separation n KL with the Monte-Carlo calibration of the boundary likelihood-ratio test, and the separation on slices through two fitted shapes (Supplementary Fig. S5). It then gives the separation for each gene of the Larsson et al. data and the number of cells needed at each capture efficiency and OFF division (Supplementary Fig. S6), and compares the cost of exact evaluation with that of simulation-based inference. All bounds given here cover a two-step OFF period only, with no measurement noise beyond binomial thinning where stated.

S5.1 Eigenvalue dictionary and the master-equation comparison

The stationary distribution (Supplementary File S1, Eq. (S10)) is parameterized by the eigenvalues $a_{1,2} = \{k_{\text{on}}, k_{12}\}$ of the generator with the ON state deleted and the nonzero eigenvalues $b_{1,2}$ of the full generator, whose symmetric functions are $b_1+b_2 = k_{\text{on}}+k_{12}+k_{\text{off}}$ and $b_1 b_2 = k_{\text{on}}k_{12}+k_{\text{off}}(k_{\text{on}}+k_{12})$. At five random rate points (log-uniform rates, ν in a moderate range) we compared the closed form with the stationary solution of the truncated master equation, obtained directly by sparse banded elimination at 70 digits with two truncation levels agreeing to at least 47 digits. Over the count range examined the closed form agrees with this solution to 45.2 digits at worst. The naive dictionary $b = \{k_{\text{on}}+k_{\text{off}}, k_{12}+k_{\text{off}}\}$, which counts k_{off} twice, deviates from the master-equation distribution at the same points by a maximum relative error of 0.09–0.57.

S5.2 Accuracy of the two-step evaluation

Supplementary Table S17 collects the accuracy tests, all against a target of at least 40 digits, beginning with the master-equation comparison just described. We measured the agreement of four independent evaluations at 20 points: the ball series with its proved tail and cancellation bounds, Arb's rigorous ${}_pF_q$ at complex parameters, a multiprecision real series, and an adaptive multiprecision ${}_2F_2$. Of the points, 12 are random and eight are chosen to be difficult (ν up to 2,000 with k up to 1,500 for real and for complex b, small rates at large ν, nearly degenerate $b_1 - a_1$, nearly double b, exactly double b, widely spread a). The worst pairwise agreement is 60.2 digits, the balls of the series and of ${}_pF_q$ overlap at all points, the multiprecision values lie inside the balls, and the largest relative ball radius is 2×10^{-60}.

In the confluent limit the three-state distribution reduces to the telegraph distribution, and with eigenvalue input $b_1 = a_1$ exactly, the three-state evaluation reproduces the two-state evaluation of the main text to 57.8 digits with overlapping balls. The three evaluations of the full count vector (backward ball recurrence, per-k series, two-precision multiprecision) agree to 54.4 digits at worst, with $|\sum_k P(k) - 1|$ of order 9×10^{-54}. The table's remaining rows concern the proof, minimality and stability scan of the recurrence, given in Supplementary File S1 (§ S1.5), and the demonstration fits of §S5.3.

S5.3 The two reference parameter sets

The demonstration parameter set is the minimal Markov embedding of Example 1 of Luo et al. [41], matched in mean dwells and synthesis rate: $k_{\text{on}} = 2/5$, $k_{12} = 2/7$, $k_{\text{off}} = 1/4$, $\nu = 20$, evaluated

Table S17: Accuracy checks of the two-step-OFF evaluator (all passed; digit targets at least 40). "Digits" is the number of agreeing significant digits between independent evaluations; relative ball radii are bounds returned by the ball arithmetic. An independent re-derivation of the eigenvalue dictionary and of the master-equation solver, with the ball-arithmetic evaluations repeated at fresh points and two precisions, reproduced every row (Supplementary File S5).

check	content	result
master-equation comparison	Eq. (8) of the main text against the stationary distribution of the truncated master equation, 5 points	worst 45.2 digits; the naive parameter dictionary is off by 0.09–0.57
agreement of four evaluations	four independent evaluations, 20 points (ν to 2000, k to 1500, doubled b)	worst pair 60.2 digits; relative radii $\leq 2 \times 10^{-60}$
recurrence proof and minimality	telescoping identity proving the order-3 recurrence; minimality modulo two primes	residual identically zero; no order-2 recurrence with degrees ≤ 6 in k, ≤ 2 in ν
confluent limit	$b_1 = a_1$ against the two-state evaluator	57.8 digits, balls overlap
vector evaluations	backward recurrence in balls, per-k series, two-precision multiprecision	worst 54.4 digits; $\lvert\sum_k P(k) - 1\rvert = 9.4 \times 10^{-54}$
stability	4 regimes to $k = 1800$	backward stable at all k (min 55.1 digits); forward not
demonstration fits	simple rejection sampling versus direct maximum likelihood on the two reference parameter sets	Supplementary File S5

at 200 cells, with a complex-conjugate b pair. Their example uses Gamma-distributed dwells and 1,000 cells, whereas ours has a hypo-exponential OFF period and an exponential ON period, and with the means matched we treat it as representative of theirs. On a data set of 200 cells drawn from this parameter set, the exact log-likelihood is $-585.601817533880382\ldots$ nats with error at most 5.1×10^{-44} nats, obtained in 0.004–0.005 s. The cost is per histogram and not per cell. A histogram of 20,000 cells takes 0.012 s. At these parameters and this cell number the refractory structure is not identifiable from stationary counts. The divergence from the three-state truth to the nearest two-state distribution is $\mathrm{KL} = 4.8 \times 10^{-4}$ nats. The separation is therefore $n\,\mathrm{KL} \approx 0.1$ at $n = 200$ and ≈ 0.5 at the $n = 1{,}000$ of Luo et al.'s own example, and at these values even the optimal likelihood-ratio test has little power to detect the structure.

The three-state maximum-likelihood fit to the demonstration data set converges to the two-state boundary ($k_{12} \approx 21$, OFF_2 effectively instantaneous) with Δloglik $= 0$ and $\Delta\mathrm{AIC} = +2$ in favor of two states. Because the null lies on a boundary (Supplementary File S1, § S1.7) we give the Akaike difference in place of a chi-square P value. A simple rejection-sampling fit to the same data (approximate Bayesian computation over the four rates, the demonstration fit listed in Supplementary Table S17) yields a pseudo-posterior but no model-comparison statistic, and hence no sign that the structure is unidentifiable at this sample size. The exact likelihood gives both a model-comparison statistic (the Akaike difference above) and the expected log-likelihood ratio $n\,\mathrm{KL}$ of the previous paragraph.

The identifiable reference parameter set is $k_{\mathrm{on}} = 4/5$, $k_{12} = 3/5$, $k_{\mathrm{off}} = 3/10$, $\nu = 100$ at 20,000 cells, with $\mathrm{KL} = 2.40 \times 10^{-3}$ and $n\,\mathrm{KL} \approx 48$. Direct maximum likelihood recovers the structure: Δloglik $= +33.2$ against the best two-state fit, $\Delta\mathrm{AIC} = -64.4$, with k_{off} estimated to within 5.5%, ν to 0.2%, and the slower and the faster of the two OFF rates to 11% and 33%. The likelihood is flattest in the plane of the two OFF rates, and because they enter the distribution symmetrically (Supplementary File S1, § S1.4), only the unordered pair is determined. A gene with these four

rates reaches $n\,\mathrm{KL} = 2$ at $n = 832$ cells (asymptotic power 0.64, measured 0.61 there for the simple likelihood-ratio test, §S5.5) and power 0.9 near $n \approx 1{,}800$.

S5.4 Computation and accuracy of the separation

For a three-state truth θ the separation per cell is $\mathrm{KL}(\theta) = \min_\phi \mathrm{KL}(P_3(\theta) \| P_2(\phi))$ (Supplementary File S1, Eq. (S16)). We carry out the inner minimization over $\phi = (a, b, r)$ by a simplex search in log-parameters, initialized at a few warm starts along each parameter sweep (§S5.6, §S5.7). Its objective is evaluated on double-precision log-probability vectors built from the backward recurrences (order three for the three-state distribution, three-term for the telegraph distribution) and seeded by single-point evaluations in ball arithmetic. We compared these double-precision vectors with two-precision multiprecision vectors for the three-state distribution (at least 13.1 digits) and with balls for the two-state distribution (at least 12.8 digits).

Each of the 1,360 KL values reported on the slices (§S5.6) was re-evaluated at the double-precision argmin with ball-evaluated probability vectors for both distributions and a ball KL sum; the largest resulting ball radius is 6.9×10^{-36}. Optimizer error enters the re-evaluated KL only at second order, by the envelope theorem. For the genome-wide per-gene calculation we used the double-precision summation, with the ball re-evaluations described in §S5.7.

The KL sum is truncated at an adaptive $k_{\max}$ with ball-evaluated residual mass $1 - \sum P_3$ below 10^{-12}, and at each re-evaluation we also computed the drift $\mathrm{KL}(k_{\max}{+}\mathrm{pad}) - \mathrm{KL}(k_{\max})$ (largest |drift| on the slices 4×10^{-12}). The omitted tail is bounded rigorously: both laws are Poisson mixtures, so $p_3(k{+}1)/p_3(k) \le \nu/(k{+}1)$, and the local Poisson rate of the telegraph law is non-decreasing, so $\log(p_3/q)$ grows at most linearly beyond the cutoff with the slope read there. The remainder is enclosed two-sidedly in ball arithmetic and is at most 1.6×10^{-6} of KL at the cutoff used for the reported values on the median-shape point, at most 3.3×10^{-13} at five other points spanning the printed cell numbers, and at most 7.4×10^{-17} at twice the cutoff. The result is independent of the working precision (agreement to at least 27.2 digits between 30 and 60 working digits), and six interior starts of the minimizer land in one basin at every point (relative spread at most 4.5×10^{-9}). The negative-binomial limit is the minimizer only at the ninetieth-percentile point, where the reported value exceeds the infimum by 0.78% and the cells needed at the infimum are 3.2×10^4 against the reported 3.1×10^4; no other printed cell number changes in two significant figures. The inner minimizer is a heuristic simplex search. Restarted repeatedly at six parameter points (the two reference parameter sets of §S5.3 and two points on each slice panel), it returns the same argmin to within 10^{-6} (relative) at the four points with an interior minimum. A second minimizer using a separate three-state evaluation reproduces twenty per-gene KL values of §S5.7 to within 1.3%.

On the 90th-percentile panel of §S5.6, and on 710 rows of a stratified sample of 2,001 genes evaluated before the genome-wide calculation, the inner argmin over ϕ lies on the boundary of the telegraph family, its negative-binomial (Poisson–Gamma) closure $b \to \infty$. There the reported value is a tight upper bound on the infimum. Minimizing separately on the negative-binomial boundary itself, we found the reported KL above the true infimum by at most 1.3% near every contour and by at most 2.7% across the boundary rows of that sample, reaching 54% only where $\mathrm{KL} \lesssim 10^{-6}$, two orders of magnitude below the lowest contour. Where this error occurs, it inflates the reported separation. Thus the cell numbers derived from the separation remain lower bounds.

Table S18: Null distribution of the composite two-step-OFF likelihood-ratio statistic at nine two-state designs (three fitted shapes at 94, 224 and 682 cells; 500 replicates each, exact likelihoods, both models re-fitted in every replicate). Rejection at the asymptotic critical values 2.706 (5%) and 5.412 (1%) with Wilson 95 % intervals, the mass at zero, and the simulated 5% and 1% critical values, against the mixture $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ (mass at zero one half, critical values 2.706 and 5.412). Replicates whose loose fit had a statistic above one were refitted at tight tolerances and the corrected rates are printed. The power labels of the main text are asymptotic; *Timp1* is quoted at its own simulated critical value (§ S5.7).

design	cells	rejection at 5% [Wilson]	at 1%	mass at zero	c_{95}	c_{99}
Timp1 fit	94	0.022 [0.012, 0.039]	0.000	0.296	2.19	3.21
Timp1 fit	224	0.066 [0.047, 0.091]	0.002	0.364	3.03	4.51
Timp1 fit	682	0.042 [0.028, 0.063]	0.008	0.464	2.57	5.11
median fibroblast shape	94	0.000 [0.000, 0.008]	0.000	0.336	0.59	0.78
median fibroblast shape	224	0.000 [0.000, 0.008]	0.000	0.370	0.85	1.31
median fibroblast shape	682	0.000 [0.000, 0.008]	0.000	0.436	1.18	1.64
ninetieth-percentile shape	94	0.000 [0.000, 0.008]	0.000	0.282	0.84	1.38
ninetieth-percentile shape	224	0.000 [0.000, 0.008]	0.000	0.278	1.09	1.80
ninetieth-percentile shape	682	0.012 [0.006, 0.026]	0.000	0.280	1.84	2.72

S5.5 Monte-Carlo calibration of the boundary test

Under the boundary null of Supplementary File S1 (Eq. (S17)) the critical value of the composite likelihood-ratio statistic at level 0.05 is 2.706, below the two-sided χ^2_1 quantile. Equation (S18) then gives asymptotic power 0.64 at $n\,\mathrm{KL} = 2$ and 0.99 at $n\,\mathrm{KL} = 8$, and power 0.9 at $n\,\mathrm{KL} \approx 4.3$. We confirmed the simple likelihood-ratio form of Eq. (S18) (Neyman–Pearson test of the three-state truth against its nearest two-state distribution) by Monte Carlo with exact likelihoods at four parameter points and two target separations ($n\,\mathrm{KL} = 2$ and $n\,\mathrm{KL} = 8$), 2,000 replicates each. The parameter points are the median shape of §S5.6 at two OFF divisions (the Erlang split and an unequal one), the 90th-percentile shape at the Erlang split, and the identifiable reference parameter set. Measured power is 0.61–0.64 at the cell numbers where $n\,\mathrm{KL} = 2$ and at least 0.988 where $n\,\mathrm{KL} = 8$, within Monte-Carlo error of Eq. (S18) in all eight runs.

Under the composite test, in which both models are re-fitted in each replicate, the boundary mixture is approached slowly. At nine two-state designs (the *Timp1* fit, the median fibroblast shape and the ninetieth-percentile shape, each at 94, 224 and 682 cells; 500 replicates each, exact likelihoods, the two-state and three-state fits from the same starts as the per-gene calculation) the rejection rate at the asymptotic critical value 2.706 lies between 0.000 and 0.066, the mass at zero between 0.28 and 0.46, and the simulated 0.05 critical value between 0.59 and 3.03 (Supplementary Table S18). The median and ninetieth-percentile shapes at every cell number reject far less often than the mixture predicts; on the ninetieth-percentile shape up to 41% of the two-state null fits end near the shared parameter box, where the two-state law sits on the (k_{off}, ν) ridge and the third rate is barely identified. The power labels of the main text are therefore asymptotic, and for *Timp1* we quote the power at its own simulated critical value 3.03 (bootstrap 95% interval 2.61–3.35 over resamples of the null statistics; §S5.7). The earlier null run at the identifiable parameter set ($n = 832$, 200 replicates) is consistent with the mixture: $P(\mathrm{LR} > 2.706) = 0.060 \pm 0.017$. We measured the composite test's power under the alternative by Monte Carlo at one parameter set only, the per-gene maximum of §S5.7, and elsewhere its power is taken from Eq. (S18).

S5.6 Slice panels through two fitted shapes

The slice plane through a fitted two-state distribution (a_0, b_0, r_0) has two axes, each a ratio of mean OFF dwell times. The abscissa $f \in (0, \frac{1}{2}]$ is the OFF-dwell split, the mean of OFF_2 over the total mean OFF dwell: $f \to 0$ is the telegraph limit and $f = \frac{1}{2}$ is the symmetric Erlang split, which has the narrowest OFF-time distribution and is the most detectable division. The ordinate $s \in [\frac{1}{4}, 4]$ is the total OFF dwell relative to the fitted one, with the ON dwell and the synthesis rate held at the fitted values ($s = 1$ preserves the fitted mean OFF dwell exactly). We computed two panels on a 20×17 grid, with each point re-evaluated in ball arithmetic (Supplementary Fig. S5). The first passes through the componentwise median of the exact fibroblast (C57) re-fits, $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}}) = (0.421, 11.0, 39.5)$ in units of the degradation rate, which differs slightly from the median of the published fits used for the onset scans of Supplementary File S2. The second passes through the 90th-percentile shape of the same re-fits, $(1.23, 206, 534)$.

At $n = 200$ cells, the scale of the Larsson et al. data, neither panel reaches the $n\,\mathrm{KL} = 2$ contour anywhere. The slice maxima, 6.8×10^{-5} nats (median shape) and 2.55×10^{-4} nats (90th percentile), both at the Erlang split, fall short of that contour by factors of 147 and 39. Even at $n = 20{,}000$ the median panel does not reach $n\,\mathrm{KL} = 2$ (maximum 1.36). On the 90th-percentile panel the $n\,\mathrm{KL} = 2$ region covers 15.3% of the slice after the boundary-infimum correction of §S5.4, and $n\,\mathrm{KL} = 8$ is not reached (maximum 5.06). Therefore, reaching the $n\,\mathrm{KL} = 8$ target with the 90th-percentile shape would take more than 20,000 cells, about 3.1×10^4 at its best split.

S5.7 The genome-wide per-gene calculation

The per-gene calculation covers all 20,217 exact two-state re-fits that pass the quality filter in the three combinations of data set and allele (a net 16 more than under the published fits, since genes outside the published filter-passing set were re-fitted as well; median cell count 206, maximum 224). For each gene we sought the smallest two-step OFF split with $n\,\mathrm{KL} \geq 8$ by evaluating the separation along the dwell-preserving family $f \in (0, \frac{1}{2}]$ through the gene's fit at its own cell count, and obtained the separation for 20,216 of the 20,217 genes. For all evaluated genes the separation lies below $n\,\mathrm{KL} = 8$ and below $n\,\mathrm{KL} = 2$ even at $f = \frac{1}{2}$. Hence none of them has a minimal detectable split. The median $n\,\mathrm{KL}(f{=}\frac{1}{2})$ is 0.0069, three orders of magnitude under the $n\,\mathrm{KL} = 8$ threshold and small enough that the boundary-calibrated power is close to the false-positive rate. The maximum is 0.581 (*Timp1*, fibroblast CAST, $n = 224$), for which the asymptotic power for a two-step OFF period is 0.285.

The genes whose double-precision truncation residual exceeded the 10^{-12} threshold (residuals 1.0×10^{-12} to 2.7×10^{-10}) received no value from the double-precision summation and were re-evaluated in ball arithmetic. For these genes a relaxed truncation served only to locate the argmin, and the reported value is the ball re-evaluation at that argmin, under the same residual and drift criteria as for the randomly sampled genes below. This re-evaluation gave a value for all but one of them, *Sel1l3* (mESC), which failed the drift criterion. For six genes it gave only an upper bound, because their true KL is at or below the double-precision noise floor (all at $n\,\mathrm{KL} < 2.1 \times 10^{-5}$, orders of magnitude below the detection thresholds). To bound the double-precision error of the per-gene values, we re-evaluated 60 randomly chosen genes with ball vectors. The relative deviation of the double-precision KL on that sample is at most 1.3×10^{-4} (median 2.3×10^{-8}), and on a further 60 mESC genes (duplicate-corrected counts, Supplementary File S3) it is at most 7.0×10^{-4} (median 1.0×10^{-8}).

A finite-n composite Monte Carlo at the *Timp1* parameters and cell number, with full two- and three-state maximum likelihood in each replicate, gives 564 rejections at the 2.706 critical value in

1,996 completed replicates (of 2,000 started). The estimated power at the asymptotic critical value is 0.283, with Wilson interval $[0.263, 0.303]$, which contains the asymptotic two-step power 0.285; at the critical value simulated for this design under the null (3.03, §S5.5) the same replicates give 0.222 $[0.205, 0.241]$, and 0.185 to 0.302 over the bootstrap interval of that critical value. Over the 95% likelihood region of the two-state fit ($2\Delta\mathrm{NLL} \leq 3.841$; 219 certified parameter vectors on 60 ray directions and directed searches for the extremes) KL ranges from 2.0×10^{-3} to 3.1×10^{-3}, $n\,\mathrm{KL}$ from 0.45 to 0.70, the cells needed for $n\,\mathrm{KL} = 8$ from 2,586 to 3,935 (3,085 at the plug-in fit), and the asymptotic power from 0.24 to 0.33; the equal split $f = \frac{1}{2}$ is the most detectable division at every point. For any level-0.05 test the power is at most $\alpha + \sqrt{n\,\mathrm{KL}/2} = 0.59$ at the plug-in fit (Pinsker's inequality), 0.52 to 0.64 over the region. Deeper OFF cascades with 5–7 rate-limiting steps [7] have a narrower OFF-time distribution than the two-stage Erlang case used here and may therefore be easier to detect; the present numbers, which cover two OFF stages only, do not bound the cells required to detect them.

S5.8 Cell numbers by capture efficiency and OFF division

From the separation computed on the slices of §S5.6, the number of cells an experiment needs follows as $N = 8/\mathrm{KL}$ for the detection power 0.99 reached at $n\,\mathrm{KL} = 8$ (§S5.5; power 0.9 already at $N \approx 4.3/\mathrm{KL}$). The fitted synthesis rate is a rate of captured molecules, $\nu_{\mathrm{fit}} = \beta_0\nu$, so $\beta = 1$ on the slices is the capture of the experiment and smaller β is further thinning of the observed law. The best case on the whole median-shape slice, the Erlang split, then needs $N \approx 1.2 \times 10^5$ cells as sequenced, 1.8×10^6 after thinning to 0.3 of the experiment's capture and 3.9×10^8 after thinning to 0.06. The same gene with every transcript counted needs $N(1/\beta_0)$: no primary estimate of β_0 exists for these libraries (Larsson et al. state that inferred burst size scales linearly with sensitivity and used no spike-ins [3]), and over the bracket $\beta_0 \in [0.1, 0.5]$ the requirement is 1.0×10^4 to 4.3×10^4 cells (2.4×10^4 at $\beta_0 = 0.3$); a droplet experiment at the absolute mean efficiency 0.06 measured by Klein et al. [42], as adopted by Tang et al. [43], needs 3.2×10^5 to 3.2×10^7 cells over the same bracket (6.0×10^6 at $\beta_0 = 0.3$). Supplementary Fig. S6 extends the as-sequenced and thinned numbers over the whole median-shape slice. Binomial thinning at a fixed efficiency acts on a mixed-Poisson distribution exactly as $\nu \to \beta\nu$, and the exact likelihood therefore applies unchanged to the thinned counts.

Because all components of technical noise other than fixed-efficiency capture are omitted (per-cell efficiency variation, ambient RNA, errors in molecule counting, cell-size covariates), real requirements are likely larger and the cell numbers just given are lower bounds. In addition, the calculation assumes stationary snapshot counts, and the relation between $n\,\mathrm{KL}$ and power is asymptotic, supported by the Monte Carlo of §S5.5 but without exact finite-n bounds. Finally, the per-gene results condition on each fitted two-state distribution without propagating its uncertainty, and those distributions were fitted to counts uncorrected for capture, whereas capture-modified likelihoods give systematically different burst parameters [43].

S5.9 Simulation cost

Luo et al. [41] infer the generalized telegraph model by sequential Monte-Carlo approximate Bayesian computation with 10,000 first-generation simulations, five generations and 1,000 particles. They report approximately 75 s per gene on their hardware, the time of roughly 2×10^4 of the exact evaluations of §S5.3. We did not run their code, and with our own stochastic simulator this run-time ratio would be larger: their lower bound of at least 1.4×10^4 simulations per inference times our measured 0.10 s per simulated data set gives at least 1,400 s per inference, the time of 10^5–10^6 exact

evaluations. However, their simulator and summary statistics are likely much faster to evaluate than our reference simulator, and the main text therefore quotes their reported time in place of this estimate.

S6 A cross-kingdom map of burst modulation

The allele-resolved analysis of the main text compares the two alleles of one gene within one cell type by per-gene likelihood-ratio tests. Here we apply the same per-gene comparison to one pair of conditions or cell states in each of three species from three kingdoms, with the calls decided by the rule of §S6.2.

S6.1 Question, species and data

For each of the three species we ask whether the change in expression between its two conditions is carried by burst frequency or by burst size, and we treat the answer (frequency-dominant, size-dominant or mixed) as a phylogenetic character of the species. We then ask whether the three characters determine the ancestral state at the eukaryote root. The pairs of conditions were chosen in advance, and the per-gene calls on them give the character matrix (§§S6.2 and S6.3); where a second pair was specified for a species, that pair is used only as a concordance check. In yeast the stress contrast appears to be carried by burst frequency, whereas the plant and cnidarian contrasts are mixed.

The yeast data are read counts, analyzed under the approximation that reads are proportional to molecules, whereas the plant and cnidarian data are molecule counts at low capture efficiency. In that regime the per-gene test detects frequency shifts more readily than size shifts, so that the plant and cnidarian frequency shares (frequency calls as a fraction of the frequency and size calls) are upper bounds (§S6.3). The reconstruction of §S6.5 leaves the ancestral state at the eukaryote root unresolved. We have not evaluated the decision rule of the analysis plan; with two non-metazoan species in the matrix the rule's outcome would be that the map cannot support a cross-kingdom conclusion, and we draw none (§S6.7). Independently of the species data, we also show that promoters with up to four states can change mean expression through burst size at nearly constant burst frequency (§S6.6).

Three species enter the character matrix: budding yeast (*Saccharomyces cerevisiae*, read counts, GEO series GSE102475), *Arabidopsis thaliana* (droplet molecule counts, GSE121619) and the cnidarian *Hydra* (Drop-seq molecule counts, GSE121617). Two further species were planned but do not enter the matrix. The counts of the bacterial data set (microSPLiT, GSE151940) include fractional multi-mapper values as deposited and have yet to be re-quantified from the raw reads. The malaria parasite (*Plasmodium berghei*, Smart-seq2 read counts) is excluded as ceiling-limited under the pre-specified rules, because most of its genes fit at the upper bound on burst size and too few remain to classify (§S6.4). We wrote the analysis plan before any species data were fitted; §S6.7 summarizes its decision rule and tabulates the deviations from it.

S6.2 Per-gene calls and genes without a call

We assign each gene one of five calls (frequency, size, mixed, not modulated, ambiguous), decided from its exact per-gene likelihoods under the two conditions. A gene receives no call when any of six exclusions applies: low expression, a fitted mode at one of the parameter bounds set in the plan (bound-limited), an expression contrast below the floor of $\kappa = 1.5$, saturation, failure of the per-gene separability screen, or model ambiguity. A gene is model-ambiguous when the separation between its frequency and size descriptions does not exceed, by the pre-specified margin, the separation that unresolved promoter structure could produce. Model ambiguity is distinct from the ambiguous call, which a gene clear of all six exclusions receives when none of the frequency, size and mixed descriptions is favored by the required factor. Genes whose exact evaluation did not finish within

the per-gene time limit of 600 s also receive no call and are counted separately as evaluation-limited. These are the most highly expressed genes, which are therefore under-represented among the calls: in the *Arabidopsis* data set 461 of 37,336 genes (1.23%) are evaluation-limited.

We later re-ran the 461 evaluation-limited *Arabidopsis* genes with a two-state evaluation built for large counts. This re-run called 3 of them, all frequency, and ended with 8 still unfinished. The other 450 reached the same 600 s limit again, because they need a three-state evaluation for large counts that we have not built. Therefore the most highly expressed genes are still under-represented among the calls. In addition, because genes within a species share regulons and chemistry, their calls may be correlated. In the plant and *Hydra* counts, genes meeting one of the six exclusions receive no call; an earlier version of the calling code had labeled them ambiguous (Supplementary Table S20, row 7).

S6.3 The character matrix and its calibration caveats

Supplementary Table S19 gives the character matrix. For each species, F_{std} is the frequency share among single-axis calls (frequency or size), standardized over pooled activity quintiles, and x is the mixed share among genes called frequency, size or mixed. A species is frequency-dominant when $F_{\mathrm{std}} \geq \frac{2}{3}$ and $x < \frac{1}{2}$, size-dominant when $F_{\mathrm{std}} \leq \frac{1}{3}$ and $x < \frac{1}{2}$, and mixed otherwise. By this rule the yeast contrast is frequency-dominant and the plant and cnidarian contrasts are mixed, and all three characters are stable under the pre-specified ± 0.05 perturbation of the thresholds.

Table S19: Character matrix of the cross-kingdom analysis: three species, one pair of conditions or cell states each, chosen in advance (the bacterial data set awaits re-quantification of its counts; the malaria data set is excluded as ceiling-limited because most of its genes fit at the upper bound on burst size, §S6.4). Columns: per-gene counts of frequency, size, mixed, not-modulated and ambiguous calls; F_{std}, the frequency share among single-axis calls standardized over pooled activity quintiles; x, the mixed share among genes called frequency, size or mixed; the character by the rule of §S6.3. In words, the yeast stress contrast appears to be carried by burst frequency, and the plant and cnidarian contrasts are mixed. The plant and *Hydra* frequency shares are upper bounds (calibration sensitivity 0.200 to injected frequency shifts against 0.000 to injected size shifts at the contrast floor), and the *Hydra* share is computed without the bound-limited genes, among which a counterfactual call would have favored frequency over size by 53 to 12. The yeast calibration is asymmetric in the same direction (sensitivity 0.115 to size shifts against 0.280 to frequency shifts at a twofold contrast), and the yeast share therefore also over-counts frequency. The three characters leave the eukaryote root unresolved (§S6.5), and no conclusion across kingdoms is drawn from them (§S6.7).

species (platform)	F/S/M/NM/AMB	F_{std}	x	character
S. cerevisiae (reads)	216/53/27/268/167	0.8080	0.0912	frequency-dominant
A. thaliana (droplet molecule counts)	559/785/578/239/668	0.3874	0.3007	mixed
Hydra (Drop-seq molecule counts)	98/124/27/35/62	0.4027	0.1084	mixed

The frequency shares of the plant and *Hydra* data sets are upper bounds, because at low capture efficiency the per-gene test detects frequency shifts more readily than size shifts of the same magnitude. Both data sets lie in the low-capture regime of our synthetic calibration, with a molecule-capture efficiency near 0.06. In that regime, at the smallest expression contrast at which a call is made ($\kappa = 1.5$), an injected frequency shift was detected in 40 of 200 simulated genes (sensitivity 0.200), whereas an injected size shift of the same ratio was detected in 0 of 200

(sensitivity 0.000).

A further caveat applies to the *Hydra* share, which is computed only on the genes not excluded as bound-limited (§S6.2). A counterfactual call on the excluded genes would have favored frequency over size by approximately 53 to 12 (48 against 11 above the contrast floor). A second *Hydra* cell-state pair (endodermal) agrees with the first in giving a mixed character, but only narrowly. Its share $F = 0.3375$ lies 0.0042 above the size-dominant threshold of $\frac{1}{3}$, and six frequency-to-size changes among its 1,354 single-axis calls would move it below that threshold. The +0.05 perturbation of the thresholds moves the second pair to size-dominant, whereas the primary pair is stable under all perturbations. Thus the character of the second pair is sensitive to small changes in the data.

In the yeast read-count regime of the synthetic calibration the per-gene test is likewise more sensitive to frequency shifts than to size shifts. At a twofold contrast an injected size shift was detected in 23 of 200 genes (sensitivity 0.115) and an injected frequency shift in 56 of 200 (sensitivity 0.280), a size-to-frequency sensitivity ratio of 0.41. Hence the yeast frequency share also over-counts frequency relative to size. We fitted the yeast data set on read counts under the approximation that reads are proportional to molecules, an approximation supported by the concordance of molecule- and RPKM-based burst parameters reported by Larsson et al. [3]. PCR amplification adds dispersion that the model omits and remains a possible confound for all yeast and malaria estimates.

S6.4 The malaria read-count data set

In the malaria data set (*P. berghei*, Smart-seq2 read counts) only one gene remains after all the per-gene exclusions of §S6.2, fewer than the pre-specified minimum for a character-matrix entry, and the data set is therefore classed as ceiling-limited. Of the 363 genes passing the low-expression filter, 305 (a fraction of 0.8402) were excluded as bound-limited at the upper bound on burst size, the largest such exclusion in the analysis. The exclusion most likely reflects the count scale of the platform: Smart-seq2 read counts run 10 to 100 times above molecule scale, whereas the upper bound on burst size was set from molecule-scale fits. Of the 305 excluded genes, 229 have their fitted burst size at the bound itself, and the excluded fraction rises monotonically with sequencing depth.

The bound-limited exclusion removed the 84% of filter-passing malaria genes with the highest counts, and only low-count genes remained eligible for a call. In the synthetic calibration, fits at a bound occurred only at the lower bound on burst size and only in the low-capture regime of the bacterial, plant and *Hydra* data sets. However, we did not foresee read-scale counts reaching the upper bound, and for a future read-count data set the upper bound on burst size should be scaled to read counts when the analysis is planned.

S6.5 Reconstruction to the root

We reconstructed the three-species character to the most recent common ancestor of eukaryotes under an equal-rates Markov model of character evolution. The reconstruction does not resolve the root: the posterior is uniform ($\frac{1}{3}$ per state) against a call threshold of 0.80. We also ran the pre-specified sensitivity analyses. Fitch parsimony places the mixed character at the root. Dropping the yeast data set gives a mixed root call, whereas dropping either of the other two species leaves the root unresolved. Collapsing the character to two states (frequency against size) gives a frequency-dominant root call at posterior 0.9995. Unit branch lengths, a stationary root prior, and an added *Escherichia coli* leaf on the bacterial branch with its character fixed in advance

from the literature as frequency-dominant all leave the root unresolved. Two planned analyses were not run: the symmetric-rates variant, because the pre-specified grid covers a single rate parameter only, and the variant adding the mouse data set, because no species-level mouse character was fixed in advance. Because bacterial and eukaryotic transcription machineries are not homologous, a character shared across the two domains could reflect only convergence under physical constraint, and character homology in the strict sense applies within the eukaryote subtree alone.

S6.6 Burst-size modulation at constant burst frequency

Separately from the species data, we asked whether a promoter with at most four states, with every transition rate inside the physiological ranges fixed in the plan, can raise its mean expression by a factor of 1.5 through an increase in mean burst size of at least that factor while its burst frequency stays constant to within a prescribed factor. We found concrete rate assignments that do so at the tightest frequency constraint examined, a factor of 21/20 (a frequency drift of at most 5%), both with the synthesis rate free and with it held fixed. In each of them the burst frequency, evaluated in rational arithmetic, is exactly the same before and after the change in expression. In particular, such assignments exist for the two-state telegraph promoter and for the three-state refractory topology of the main text.

To find such rate assignments we used the exact factorization of mean expression into burst frequency, mean burst size and the fraction of time spent OFF. This factorization holds in exact rational arithmetic at 10,000 random parameter draws on each of the 298 distinct promoter topologies with up to four states. In a lattice search with 16 points per rate axis over the same rate ranges we found pairs of lattice points realizing such a size-carried change within the 5% drift allowance on each topology examined: 4.3×10^5 pairs on a two-state telegraph promoter, 7.8×10^7 on a three-state refractory promoter and 1.7×10^{10} on a four-state refractory cycle. The largest achievable changes in mean expression lie at the edges of the rate ranges.

The nearest published theorem, by Zoller et al. [44], constrains a different object: reproducing an observed invariance of the promoter switching correlation time requires regulatory architectures with at least four internal states and non-equilibrium dynamics that break detailed balance. Their result is a necessary condition on promoter architecture for reproducing that observable, whereas ours shows that promoters with up to four states can realize a particular mode of modulation, a change in mean expression carried by burst size at bounded drift in burst frequency. Neither result implies the other, and their minimum of four states coincides with the largest number of states considered here.

S6.7 Pre-specified decision rule, deviations and scope

Before any species data were fitted, we fixed in a written plan the species and the pair of conditions or cell states for each, the per-gene call rule of §S6.2 with its contrast floor $\kappa = 1.5$, and the character thresholds of §S6.3 with a ± 0.05 perturbation of them as a sensitivity analysis only. We also specified a positive control on mouse data, built on TATA-box and enhancer-interaction gene lists whose membership we fixed in a later amendment (Supplementary Table S20, row 10), and a go/no-go criterion on the per-species counts of separable genes. For the root we specified the equal-rates reconstruction of §S6.5 with a call at posterior at least 0.80. Finally, we fixed an ordered decision rule that selects one conclusion from the per-species results, as follows.

The branches of the rule are examined in a fixed order, and the first branch whose condition holds gives the conclusion. First, if the go/no-go criterion fails, the analysis stops with only the per-species counts of separable genes, the detection ceilings and the fitted parameters, which are

reported whatever the outcome of the rule. Second, the analysis stops likewise if the mouse positive control fails with adequate power, whereas if the control is under-powered the evaluation continues and any later conclusion is stated as provisional. Third, it stops if fewer than three non-metazoan species have character-matrix entries after ceiling-limited data sets are excluded, because the map then cannot support a cross-kingdom conclusion.

If none of the three stops applies, four conclusions remain and are tested in order. The first, that the differences between species reflect the activity levels of their genes, is reached when the spread of the raw frequency share across species is at least $\frac{1}{3}$ and the spread of the activity-standardized share is below $\frac{1}{3}$. The second, that frequency modulation is universal among the sampled species, requires all species to be frequency-dominant. The third, that the modulation axis differs between lineages, requires at least one size-dominant species, a concordant call at the second condition pair of each size-dominant species that has one, and at least one size-dominant species free of the caveat that a capture-efficiency change between its conditions could mimic size modulation. Otherwise the rule selects the fourth conclusion, that the axis is set at the level of conditions or regulons, with no phylogenetic signal.

We have not evaluated the decision rule, because two of its inputs, the mouse positive control and the go/no-go criterion, have not been computed. The outcome of the rule on the present matrix nevertheless follows directly unless one of the two earlier stops applies first: the third stop requires at least three non-metazoan species with character-matrix entries, whereas only yeast and *Arabidopsis* qualify (*Hydra* is a metazoan), and the rule would therefore end at that stop.

The results of this analysis are the per-species characters of Supplementary Table S19, subject to the calibration caveats of §S6.3, and the demonstration in §S6.6 of burst-size modulation at constant burst frequency. With three species we reach no conclusion across kingdoms and leave the root unresolved. Each call is a comparison of fitted burst parameters between two conditions in public data and does not by itself identify a regulatory mechanism. Four further pre-specified items remain to be done: a test that would apply only to a data set containing model-ambiguous genes, a second mathematical question that we have deferred, the re-quantification of the bacterial counts (§S6.1), and a three-state evaluation for large counts (§S6.2).

Deviations from the pre-specified plan

The deviation log has twelve entries, all at stages that precede the decision rule, and three of them carry a date in their title; entries six to eight are merged into one row here. Supplementary Table S20 summarizes each entry from its title, with the species names in place of the plan's data-set codes. The decision-rule section of the plan and the titles of the logged deviations are in the archive of Supplementary File S7 (§S7.5) as `S6_decision_rule_verbatim.txt` (SHA-256 `adfeae7bfee5…`) and `S6_deviation_titles_verbatim.txt` (`1b690fdf96b0…`), together with the full plan and the full deviation texts.

Table S20: Deviations from the pre-specified plan of the cross-kingdom analysis, one row per entry of the deviation log (entries six to eight merged), numbered in log order with the log identifier in parentheses. Date: the date in the entry title, if any. Stage: when the change was made relative to the decision rule. Change and Reason: from the entry title. Effect on results: where the change enters the numbers of this Supplementary File.

No.	Date	Stage	Change	Reason	Effect on results
1	not dated in the title	data preparation, before the decision rule	(CA-D-1) Counts of the yeast data set (GSE102475) re-quantified from the raw reads deposited at the European Nucleotide Archive	the public deposit holds normalized values only	yeast enters the matrix with re-quantified read counts (§S6.3)
2	not dated in the title	data loading, before the decision rule	(CA-D-2) Loader for the bacterial microSPLiT matrix (GSE151940) repaired, the deposited matrix having no header or index, and the integer-count guard extended	the deposit carries fractional multi-mapper counts	led to entry 3
3	not dated in the title	matrix assembly, before the decision rule	(CA-D-3) The bacterial data set removed from the primary character matrix pending re-quantification of its counts	consequence of entry 2	one fewer non-metazoan species in the matrix, which bears on the third stop of the rule (§§S6.1, S6.7)
4	not dated in the title	per-gene evaluation, before the decision rule	(CA-D-4) Per-gene disposition "evaluation-limited" defined for genes whose exact evaluation needs more series terms than the fixed schedule of term limits allows; counted as a separate exclusion class	genes with large counts exceed the schedule	the class is counted per species (§S6.2)
5	not dated in the title	per-gene evaluation, before the decision rule	(CA-D-5) Per-gene time limit of 600 s, after which a gene is classed evaluation-limited; the correlation of this exclusion with expression to be reported; an evaluation for large counts deferred; partial per-gene results saved on interruption	as entry 4	§S6.2 states the limit, the under-representation of highly expressed genes and the later re-run of the evaluation-limited genes
6	not dated in the titles	data loading, before the decision rule	(CA-D-6 to CA-D-8) The *Arabidopsis* data set, deposited as an R data object, read through R into matrix and table files for the loader; condition labels mapped (`whole_root_Control_2` to control, `whole_root_Heatshock` to heat shock) by a stated evidence rule	the deposit is an R object, and the plan had not specified which cells belong to each of the two conditions	fixes which cells form the pre-specified pair; thresholds untouched

continued on the next page

Table S20 (continued)

No.	Date	Stage	Change	Reason	Effect on results
7	not dated in the title	data-set disposition, before the decision rule	(CA-D-9) Handling of the *Hydra* data set decided: the defect in the calling code (§S6.2) logged and its correction scheduled; the data set, held back while the defect was uncorrected, re-admitted conditionally	the defect was found while this data set was processed	the plant and *Hydra* rows of Supplementary Table S19 are the corrected counts (*Hydra*: 98/124/27/35/62; $F_{\mathrm{std}} = 0.4027$)
8	2026-07-23	data loading, before the decision rule	(CA-D-10) The malaria annotation file read as Latin-1 text	the file contains bytes that are not valid UTF-8	none
9	2026-07-23	data-set disposition, before the decision rule	(CA-D-11) The malaria data set classed as ceiling-limited and kept out of the primary matrix by the pre-specified rules, its second condition pair thereby moot; a re-run with re-scaled bounds noted as future work and not run	read-scale counts reach the upper bound on burst size	§S6.4 reports the data set descriptively
10	2026-07-23	before the positive control, which has not been computed	(CA-D-12) The gene lists of the mouse positive control fixed by an amendment to the plan's operative definitions	the identity of the lists had been left open	none yet; the control has not been computed (§S6.7)

S7 Software, versions and reproducibility

This Supplementary File documents the article's software and its deposited archive, beginning with the evaluator of the telegraph likelihood and the fixed software versions under which all numbers in the article were produced. It also gives the source lines of the published implementations we examined, compares our copies of their code with the originals, describes the contents of the archive (Supplementary Table S22), and states how the analysis plans were archived. The evaluator is distributed as **[TODO: package name]**, and the archive is available at **[TODO: archive DOI]**.

S7.1 The evaluator

The evaluator is a single Python file with pip-installable dependencies only (numpy, mpmath, sympy, and python-flint, which provides Arb [12]). It evaluates the telegraph count distribution and log-likelihood of Supplementary File S1 in three modes. The ball-arithmetic mode sums the all-positive series in Arb ball arithmetic with the geometric tail bound of Supplementary File S1 (§S1.6) and obtains vectors of probabilities from the backward ball recurrence with two starting values. If a returned ball is wider than the requested precision, the mode raises the working precision and repeats the evaluation, up to 4 times. If the ball is still wider than requested, the mode returns no value, and otherwise the radius of the returned ball is a rigorous bound on the log-likelihood error. The self-checking multiprecision mode runs the backward recurrence at two working precisions and accepts the result only if the two values agree to the requested number of digits at each n. The double-precision mode computes the per-entry a priori error bound and tests the four conditions of Supplementary File S2 (§S2.8), and by default recomputes a vector that meets one of them in the multiprecision mode.

A self-test computes $P(n) = A + Be^{-r}$ with exact rational A, B (from sympy) for integer (a, b) and rational r in five cases and requires each computed ball to contain the exact value. It reports success only if containment also fails when the dominant coefficient is perturbed by a relative change of 10^{-30}. Thus the self-test detects errors of that size. A file of reference values contains 25 parameter points with $P(n)$ to 50 digits (relative ball radius below 10^{-55} per value, generator seed 271,828) for regression tests. The 19 further test points of Supplementary File S2 (§S2.8) were drawn with a different generator seed, 31415926. The evaluator's JSON outputs are deterministic (sorted keys, no timestamps). The program does not access the network. Example command-line usage:

```
python3 telegraph_evaluate.py loglik --a 0.4 --b 9.5 --r 300 \
    --counts 0,3,7,2,9 --mode certified
python3 telegraph_evaluate.py --selftest
```

The `-mode` value in the first command selects the ball-arithmetic mode described above.

A second module evaluates the count distribution and log-likelihood of the two-step (three-state) model, with the accuracy reported in Supplementary File S5 (§S5.2). It offers three evaluations (the ball series with the spectral-dominance tail bound, Arb's rigorous ${}_pF_q$, two-precision multiprecision) and the backward order-three ball recurrence for vectors of probabilities. The polynomial d in the telescoping proof of the order-three recurrence (Supplementary File S1, Eq. (S14)) is deposited with the module.

S7.2 Dependencies and fixed versions

All numbers in the main text and in the Supplementary Files were produced under Python 3.12.3, numpy 2.4.1, scipy 1.17.1, pandas 3.0.3, mpmath 1.3.0, python-flint 0.8.0 (Arb) and sympy 1.14.0,

with single-threaded linear algebra. We examined the published implementations at fixed versions: txburst at commit `6b6525f` of its public repository (the maximum-likelihood, profile and test scripts `txburstML.py`, `txburstPL.py`, `txburstTEST.py`), BPSC at commit `5a512e5` (`R/pBP.R`, `R/pBPi.R`, `R/BPfam.R`), and scModels at version 1.0.5.

Two version differences affect the comparisons in Supplementary File S3. First, the published txburst fit files appear to have been produced with an earlier scipy release than ours. When we re-ran the published maximum-likelihood script on a fixed 200-gene sample under our scipy release, the optimizer stopped at a different point from the published fit for 28 of the genes (median $|\Delta\mathrm{NLL}|$ 0.17 nats), which suggests a change in the optimizer between releases. Therefore we compare point estimates only between re-fits made under one version and exclude from the profile comparison the genes whose published intervals are not reproduced under our version (§S3.7). Second, the published profile script names a numpy exception alias that numpy 2 removed. Under numpy 2 the early-stopped scans of Supplementary File S3 (§S3.8) raise an exception instead of returning NaN. We restored the alias with one added line and left the rest of the script unchanged.

S7.3 Published implementations examined

The txburst maximum-likelihood script evaluates the count distribution and the log-likelihood in two lines of `txburstML.py` at commit `6b6525f`, quoted verbatim:

```
41    x,w = j_roots(50,alpha = bet - 1, beta = alpha - 1)
49    return(-np.sum(np.log( dBP(vals,kon,koff,ksyn) + 1e-10) ) )
```

Line 41 is the fixed-node Gauss–Jacobi discretization of the mixture integral, mathematically correct up to truncation of the rule at that node count, with no adaptivity and no error estimate. Line 49 adds the constant inside the logarithm (a routine guard against $\log 0$). The profile script `txburstPL.py` uses the same rule (line 28) and the same constant (lines 35 and 40), and sets the interval cutoff at line 59 (identically at line 130):

```
cutoff = chi2.ppf(1-alph, 1)/2
```

and interpolates the endpoints where the statistic $2(\mathrm{NLL}(\theta) - \mathrm{NLL}_{\min})$ crosses `cutoff`. The likelihood-ratio test script `txburstTEST.py` builds the same kind of rule with 40 nodes at line 27 and adds the same constant at lines 39 and 48. BPSC at commit `5a512e5` uses the same construction in `R/pBP.R`, quoted verbatim:

```
26    w = gauss.quad(10,'jacobi', alpha=bet-1, beta=alp-1)
```

with the same rule in the interval version `R/pBPi.R` (line 30), which is the version called by its fitting routine. To confirm the node count from the function's output, we compared the published `pBP` with a `scipy` implementation of the same rule, and the two functions match to 1.2×10^{-13} at $r = 1{,}000$. The quasi-likelihood variance function used by BPSC's differential-expression test (Supplementary File S2, §S2.7) is defined at lines 49–52 of `R/BPfam.R`.

Our copies of the published code used in Supplementary Files S2 and S3 give the same output as the unmodified scripts. We copied the txburst maximum-likelihood code without algorithmic change and compared it with the unmodified script on 20 genes of the repository data, on which the two returned the same fitted parameters. Our 40-node density function is a line-for-line copy of the nested function in the test script (lines 18–30), with its imports unchanged. We copied the profile functions exactly from `txburstPL.py` and compared them with the original (the numpy alias restored in both) on 12 genes of the repository's mESC count file, 4 of which pass the published quality filter. The 8 profile intervals returned by the copy are identical to the original's, including the NaN entries. Supplementary Table S21 lists the files with their commits and checksum prefixes.

Table S21: Published source files examined, with the repository commit, the checksum prefix of the file as examined, and the lines quoted or cited in this article.

file	commit	checksum prefix	lines
`txburstML.py`	`6b6525f`	`82260daa...5257`	41 (rule), 49 (added constant), 105 (count rounding)
`txburstPL.py`	`6b6525f`	`507ceb7a...409d`	28 (rule), 35, 40 (constant), 59, 130 (cutoff)
`txburstTEST.py`	`6b6525f`	`565be1ab...57ed`	18–30 (density), 27 (rule), 39, 48 (constant)
`R/pBP.R`	`5a512e5`	(repository)	26 (rule)
`R/pBPi.R`	`5a512e5`	(repository)	30 (rule)
`R/BPfam.R`	`5a512e5`	(repository)	49–52 (variance shape)

S7.4 File map of the deposited archive

Supplementary Table S22 gives, for the principal results of the main text and of the Supplementary Files, the archive file from which each is computed, with names relative to the archive root. The archive contains the code and the reference values, the corrected per-gene tables of Supplementary File S3 (CSV and JSONL), the per-gene results of the allele-resolved analysis (Supplementary Data S4), the per-gene separability results of Supplementary File S5, and the benchmark and double-precision error-control results of Supplementary File S2. It also contains the analysis plans of Supplementary Files S3, S4 and S6, with the S4 amendment, the deviation logs and the S3 plan-versus-outcome table (`archive/analysis_plans/`, §S7.5). The archive also holds a script that loads the per-gene tables and a file manifest.

The raw sequencing data are not part of the archive and are available from the original repositories. The Larsson et al. data are at ArrayExpress under accessions E-MTAB-6362, E-MTAB-6385 and E-MTAB-7098, and the processed allelic count tables and fit files are in the txburst repository at the commit above. The Johnsson et al. matrices are at ArrayExpress under E-MTAB-11054, and the data for the three species of Supplementary File S6 are GEO series GSE102475, GSE121619 and GSE121617.

S7.5 The analysis plans and their archive

The analysis plans summarized in Supplementary Files S3, S4 and S6 were committed to the project's version-control repository, without an external registry or independent timestamp, so the order of plan and analysis could be checked only against the repository's commit history. Each plan lists the computations already made on the data when it was written, and each later change is a dated or numbered entry in its deviation log, tabulated in the corresponding Supplementary File. The verbatim texts named in Supplementary Files S3, S4 and S6, transliterated to ASCII with local directory paths shortened, are in the archive (`archive/analysis_plans/`, with the SHA-256 manifest `MANIFEST.sha256`). We archived the full plan and the full deviation log of Supplementary File S6 there as unmodified copies of the original files, `S6_plan_verbatim.txt` (SHA-256 `02915657eee5...`) and `S6_deviations_verbatim.txt` (`f1d6b38497f2...`).

Table S22: File map of the deposited archive: where each claim is reproduced. Paths are relative to the archive root; the raw re-evaluation outputs under `phase2/driver/out/` and `out_dedup_mesc/` are read-only.

claim	file
exact evaluator; held-out agreement to 57 digits (required ≥ 40)	`evaluator/telegraph_evaluate.py`; `evaluator/GATE_EVALUATOR.json`
25 reference values, $P(n)$ to 50 digits	`evaluator/REFERENCE_VALUES.json`
double-precision error model; 228-cell validation, zero uncaught	`evaluator/FLOAT_POLICY_VALIDATION.json`
$(N/\pi)^2$ quadrature onset; the added-constant mechanism	`pilot/gate/GATE_REPORT.md`
published-method benchmark against the exact reference	`benchmark/BENCHMARK_RESULTS.json`
likelihood values at the published parameters (Fig. 3a)	`phase2/driver/out/p1_nll.json`, `p1_nll.jsonl` (fibroblasts); `phase2/driver/out_dedup_mesc/` (mESC, each cell once)
interval endpoint moves; cutoff widening (Fig. 3c)	`phase2/driver/out/p2_ci.json`, `p2_ci.jsonl`
point-estimate negative control (Fig. 3b)	`phase2/driver/out/p3_shifts.json`, `p3_shifts.jsonl` (fibroblasts); `out_dedup_mesc/` (mESC)
quality-call changes and their decomposition (Supplementary Table S7)	`phase2/driver/out/p4_flags.json`, `recheck.jsonl` (fibroblasts); `out_dedup_mesc/` (mESC)
coverage Monte Carlo for the interval cutoff (Fig. 2b)	`checklist/pl_ci_mc_summary.json`; `checklist/PL_CI_VERIFY.md`
BPSC density errors; differential-expression probe (Fig. 2c)	`bpsc/BPSC_AUDIT.json`; `bpsc/de_probe.json`
two-step-OFF evaluator; maximum-likelihood and rejection-sampling demonstration	`multistate/multistate_core.py`; `multistate/abc_mle_demo.json`
per-gene separability calculation and its cell-number maps (Fig. 5b, c); independent re-derivation	`atlas/atlas_pilot.json`; `atlas/out/pergene_genomewide.json` (fibroblasts); `atlas/out_dedup_mesc/` (mESC); `atlas/VERIFY_ADVERSARIAL.md`
verbatim analysis plans and deviation logs summarized in Supplementary Files S3, S4 and S6 (SHA-256 manifest) and an independent re-reading of the raw re-evaluation results	`archive/analysis_plans/MANIFEST.sha256`; `phase2/PHASE2_PREREG.md`; `phase2/BLIND_READJUDICATION.md`
allele-resolved analysis: per-gene results, primary result, independent recomputation	`cis_v2/discovery/RESULTS_autosomal_V2.jsonl`; `cis_v2/discovery/CIS_V2_RESULT.md`; `cis_v2/VERIFY_ADVERSARIAL.md`
application across species (Supplementary File S6): character matrix, root reconstruction, accompanying theorem	`controlaxis/RESULT.md`; `controlaxis/e1_assembly/e1_out/`; `controlaxis/e2_x0/E2_RECEIPT.md`

S8 Supplementary figures

Supplementary Fig. S1 shows the published estimates of burst size and synthesis rate, with the onset lines $k_{\mathrm{syn}} \approx (N/\pi)^2$ overlaid, and marks the fits in which some cell's published probability reduces to the added constant (Supplementary Files S2, § S2.3, and S3, § S3.2). Supplementary Fig. S2 shows the negative log-likelihood profiles of *Krt10* under the published and the exact evaluation (Supplementary File S3, § S3.3), and Supplementary Fig. S3 the Monte-Carlo coverage per parameter set (Supplementary File S3, § S3.8). Supplementary Fig. S4 shows the allelic effect sizes by stratum and the class counts under four stratification variants (Supplementary File S4, §§ S4.4–S4.5). Supplementary Fig. S5 shows the separation $n\,\mathrm{KL}$ on slices through two fitted shapes (Supplementary File S5, § S5.6), and Supplementary Fig. S6 the number of cells needed at each capture efficiency and OFF division (Supplementary File S5, § S5.8).

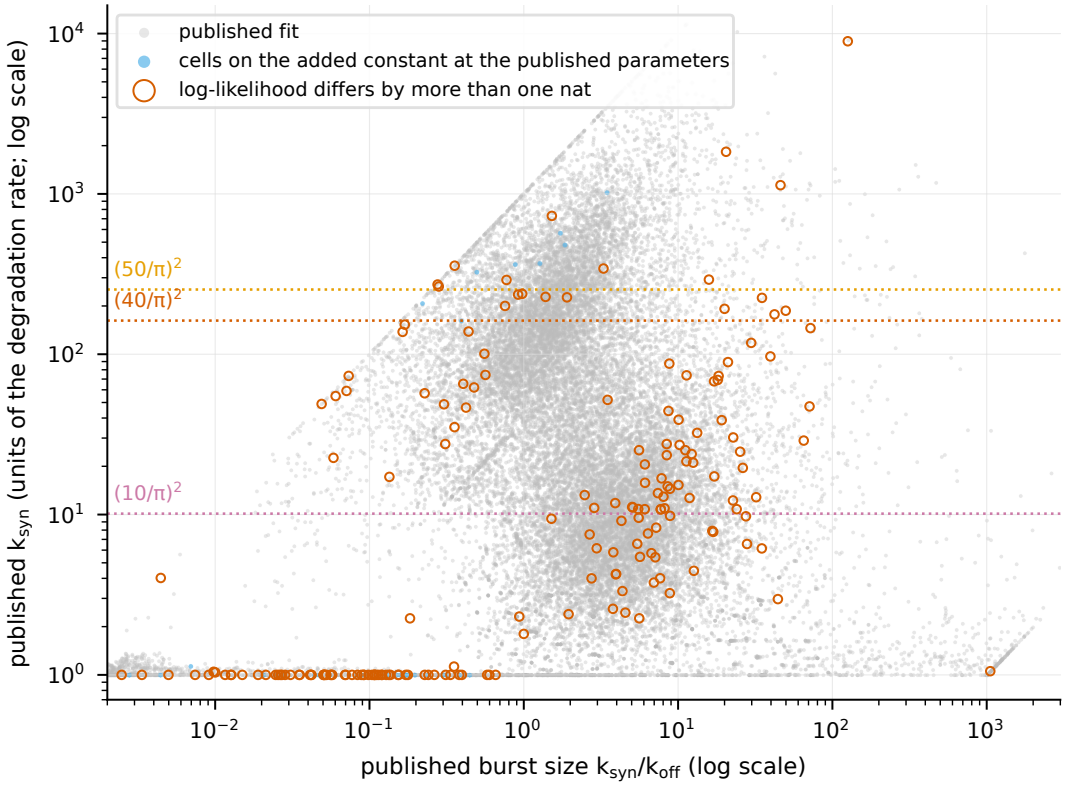

Figure S1: The published fits fall mostly where the 50-node rule is accurate, and nearly all fits whose log-likelihood error exceeds one nat lie at small k_{syn}, where the added constant dominates some cells' published probabilities, rather than beyond the quadrature onsets. Published estimates of burst size $k_{\text{syn}}/k_{\text{off}}$ and of k_{syn} (in units of the degradation rate) for the 32,376 fits of Larsson et al. [3] in the three combinations of data set and allele (mESC cells counted once). Dotted horizontal lines mark the onsets $(N/\pi)^2$ of the 10-, 40- and 50-node rules. Blue points: fits with at least one cell whose published probability reduces to the added 10^{-10} at the published parameters. Open circles: the 205 fits whose published log-likelihood differs from the exact value by more than one nat (Fig. 3a of the main text); most lie at the lower bound on k_{syn} of the published code, where the published distribution puts almost no mass on the observed nonzero counts.

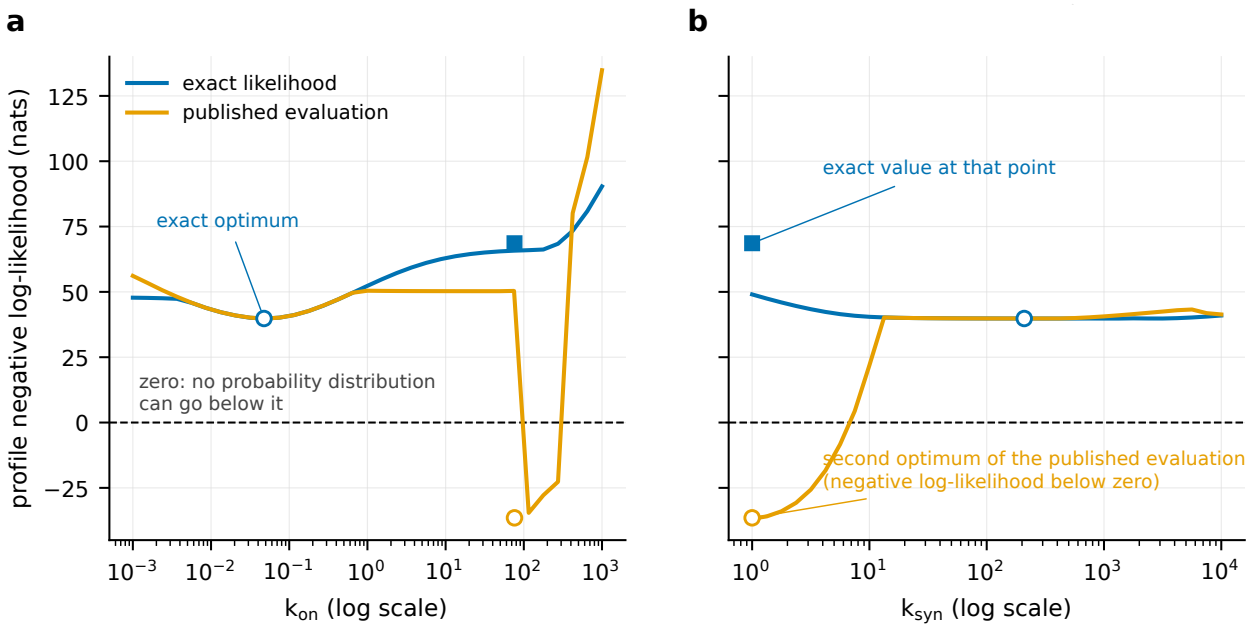

Figure S2: For *Krt10* (fibroblasts, CAST allele, 127 cells) the published evaluation has a second optimum at negative log-likelihood below zero, which no probability distribution can produce, and the exact negative log-likelihood at that point is 68.6 nats, 28.8 nats above the exact optimum. One-dimensional profiles of the negative log-likelihood over k_{on} (a) and k_{syn} (b) within the parameter bounds of the published code, the other two parameters optimized at each grid value by deterministic multi-start local search, under the published evaluation (orange) and the exact evaluation (blue). Open circles: the optimum of each curve (published, −36.5 nats); filled square: the exact value at the optimum of the published evaluation. At that optimum k_{on} is larger than the exact estimate by a factor of about 1,600 and the 50-node rule returns probabilities above one for most cells. *Krt10* is the single extreme case among the re-fitted genes, and the published quality filter would discard this re-fit; we show it only to illustrate the second optimum that the quadrature error creates (Supplementary File S3).

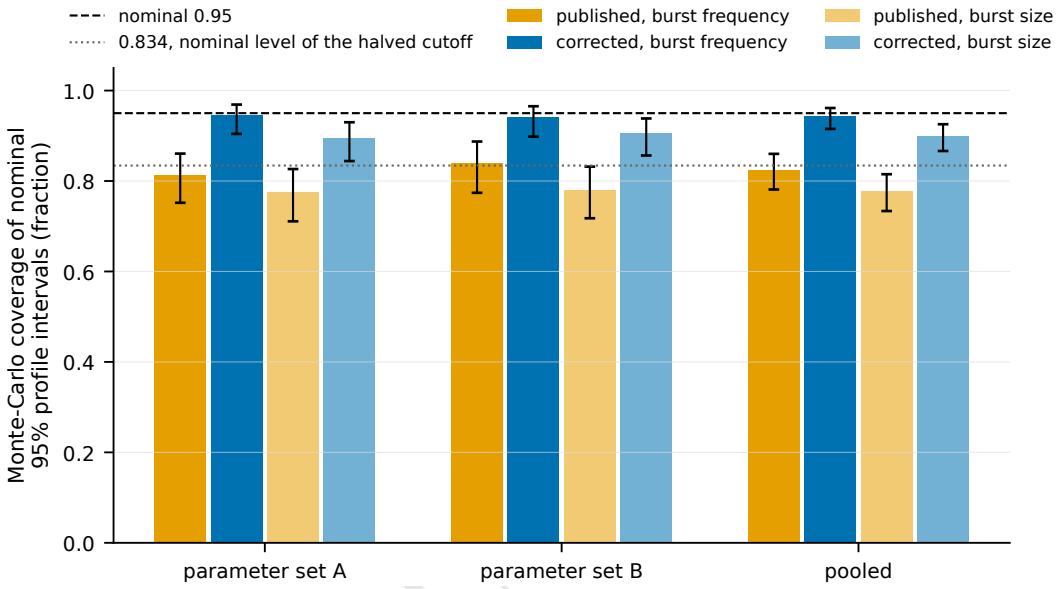

Figure S3: The coverage shortfall under the published cutoff and the coverage restored by the corrected cutoff differ little between the two parameter sets, where the quadrature is accurate. Monte-Carlo coverage of nominal 95 % profile-likelihood intervals for burst frequency and burst size at parameter set A $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}}) = (1.0, 10.0, 100.0)$ and set B $(0.4, 5.0, 200.0)$, each with half of the 400 simulated 200-cell data sets, and pooled (Fig. 2b of the main text); bars are Wilson 95 % intervals. Dashed line: nominal 95 %; dotted line: 0.834, the coverage predicted for nominal 95 % intervals under the halved chi-square cutoff. Pooled values: published cutoff 0.824 (burst frequency) and 0.777 (burst size), corrected cutoff 0.943 and 0.900; the burst-size shortfall under the corrected cutoff comes from the likelihood ridge at 200 cells and from the inner optimizer of the profile scan (§ S3.8); data sets for which the published code returned no interval are excluded from the denominators (Supplementary Table S9).

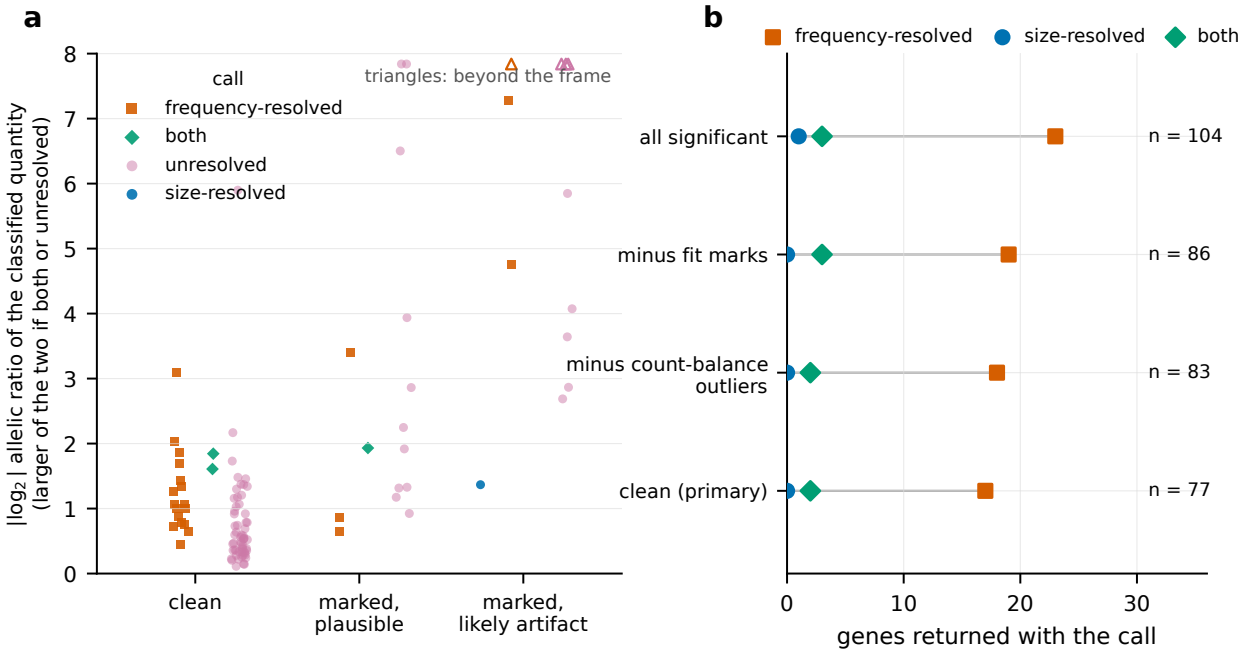

Figure S4: Under the repaired burst-size statistic the frequency-resolved count exceeds the size-resolved count under every recomputed stratification, and the large allelic effects concentrate in the marked strata. (a) Unsigned allelic $\log_2$ ratio of the classified quantity (burst size for size-resolved genes, burst frequency for frequency-resolved genes, the larger of the two otherwise) for the 104 significant genes, by stratum (clean, 77; marked but plausible, 15; artifact, 12) and call; triangles mark values beyond the frame. The large ratios concentrate in the marked strata, as expected if artifacts inflate the estimates of the marked genes. (b) Genes returned as frequency-resolved, size-resolved or both under each stratification (Supplementary File S4): all 104 significant genes (23 frequency-resolved, 1 size-resolved, 3 both, 77 unresolved), significant genes without a fit mark, those without a count-balance outlier, and the clean stratum of the main text (17, 0, 2); the exclusion by pseudogene name is not recomputed with the repaired statistic; n is the number of genes in each stratification.

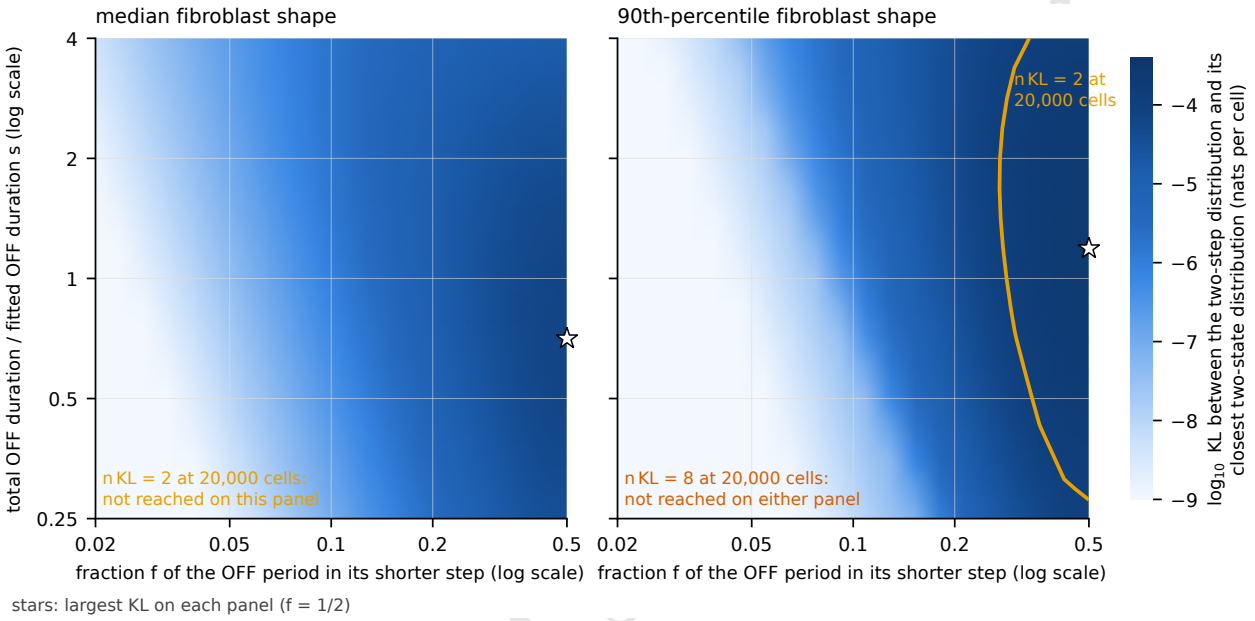

Figure S5: Along the slice through the median fibroblast shape, no two-step division of the OFF period reaches the separation detectable with 20,000 cells, and along the 90th-percentile shape only nearly equal divisions do. Kullback–Leibler divergence per cell between the stationary distribution of a two-step OFF model and its closest two-state distribution, over the fraction f of the mean OFF period spent in the shorter step and the total OFF duration s relative to the fitted one, with the ON dwell and the synthesis rate held at the fitted values (Supplementary File S5); left, the median fibroblast shape $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{syn}}) = (0.421, 11.0, 39.5)$; right, the 90th-percentile shape $(1.23, 206, 534)$. Stars: the largest divergence on each panel, 6.8×10^{-5} and 2.55×10^{-4} nats per cell, both at $f = 1/2$. Orange contour: $n\,\mathrm{KL} = 2$ at $n = 20{,}000$ cells (boundary-calibrated power 0.64); the $n\,\mathrm{KL} = 8$ level (power 0.99) is not reached on either panel, and even the 90th-percentile shape needs 3.1×10^4 cells in the best case.

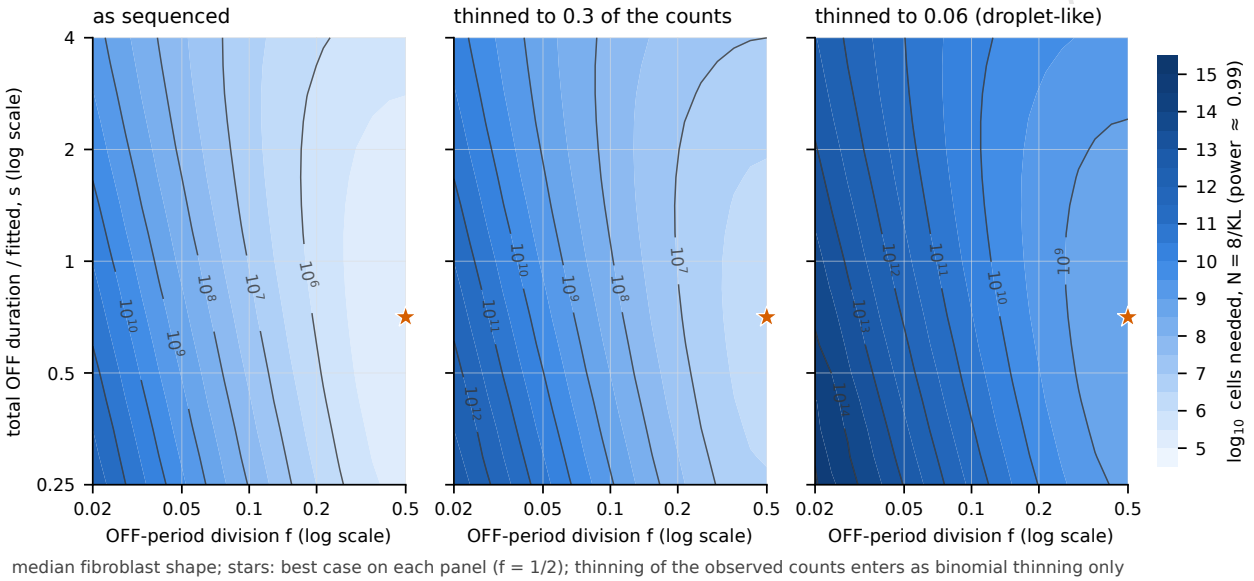

Figure S6: Over the whole median-shape slice, the number of cells needed to detect a two-step OFF period (Fig. 5c of the main text) grows by orders of magnitude away from the equal division and with further thinning of the observed counts. Cells needed with power about 0.99, $N = 8/\mathrm{KL}$, at the median fibroblast shape $(0.421, 11.0, 39.5)$ over the OFF-period division f and the relative total OFF duration s, at the capture of the experiment (as sequenced) and after further thinning of the observed counts to 0.3 and to the droplet-like 0.06 of it [42]; contours every decade. Stars: the best case on each panel, 1.2×10^5, 1.8×10^6 and 3.9×10^8 cells, all at $f = 1/2$; with every transcript counted the best case is 1.0×10^4 to 4.3×10^4 cells for an experimental capture efficiency between 0.1 and 0.5 (§ S5.8). Capture enters as binomial thinning of the synthesis rate; per-cell efficiency variation, ambient RNA and amplification noise are not modeled, so real requirements are likely larger and the plotted cell numbers are lower bounds.

References

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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] Ulysse Herbach. Stochastic gene expression with a multistate promoter: breaking down exact distributions. *SIAM Journal on Applied Mathematics*, 79(3):1007–1029, 2019. doi: 10.1137/18M1181006.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Fredrik Johansson. Computing hypergeometric functions rigorously. *ACM Transactions on Mathematical Software*, 45(3):1–26, 2019. doi: 10.1145/3328732.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] Read-Lab-UCI. TelegraphLikelihoodInfer. Zenodo software deposit, version 1.0, doi:10.5281/zenodo.16915450, 2025.

[23] Carson C. Chow. StochasticGene.jl: Julia package to fit and analyze stochastic gene transcription models. Julia package, version 2.0.4, Julia General registry; `https://github.com/nih-niddk-mbs/StochasticGene.jl`, 2026. With the laboratory of D. R. Larson (NIH).

[24] Alberto Pessia and Jing Tang. A three-term recurrence relation for accurate evaluation of transition probabilities of the simple birth-and-death process. *BIT Numerical Mathematics*, 61:561–585, 2021. doi: 10.1007/s10543-020-00836-x.

[25] 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.

[26] Francesco Mottes, Qian-Ze Zhu, and Michael P. Brenner. Gradient-based optimization of exact stochastic kinetic models. *arXiv preprint*, arXiv:2601.14183, 2026. doi: 10.48550/arXiv.2601.14183. v2, 2026-03-05.

[27] Gregor Neuert, Brian Munsky, Rui Zhen Tan, Leonid Teytelman, Mustafa Khammash, and Alexander van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119):584–587, 2013. doi: 10.1126/science.1231456.

[28] Hiromasa Nakayama, Kenta Nishiyama, Masayuki Noro, Katsuyoshi Ohara, Tomonari Sei, Nobuki Takayama, and Akimichi Takemura. Holonomic gradient descent and its application to the Fisher–Bingham integral. *Advances in Applied Mathematics*, 47(3):639–658, 2011. doi: 10.1016/j.aam.2011.03.001.

[29] Hiroki Hashiguchi, Yasuhide Numata, Nobuki Takayama, and Akimichi Takemura. The holonomic gradient method for the distribution function of the largest root of a Wishart matrix. *Journal of Multivariate Analysis*, 117:296–312, 2013. doi: 10.1016/j.jmva.2013.03.011.

[30] 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.

[31] 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.

[32] Jiajun Zhang, Luonan Chen, and Tianshou Zhou. Analytical distribution and tunability of noise in a model of promoter progress. *Biophysical Journal*, 102(6):1247–1257, 2012. doi: 10.1016/j.bpj.2012.02.001.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Benjamin Zoller, Alexis Bénichou, Thomas Gregor, and Gašper Tkačik. Invariant nonequilibrium dynamics in gene regulation optimize information flow. *Proceedings of the National Academy of Sciences*, 123(28), July 2026. ISSN 1091-6490. doi: 10.1073/pnas.2524855123. URL `http://dx.doi.org/10.1073/pnas.2524855123`.