Supplementary Figures and Tables

Two kinds of time in mammalian promoters: a refractory class that detects change faster per burst cycle

Matthew D. Schwartz and Martin Hemberg

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

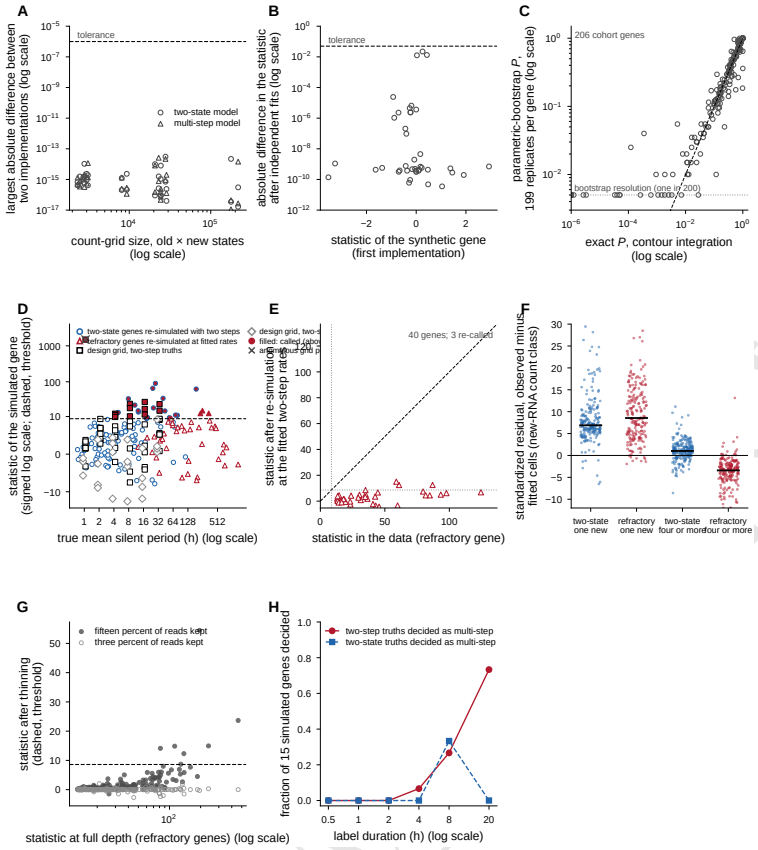

Figure S1. Numerical validation of the likelihood and the exact test, and detection power of a two-hour window at this depth, related to Figure 1 and Supplementary Methods. (A) Largest absolute difference between the values of the labeling-window likelihood computed by two independent implementations at 64 parameter points (count grids from 74 ×29 to 1,167 ×186 states; both promoter models). At every point the difference is below 3.1×10^{-14}, within the agreement threshold of 10^{-6} (Supplementary Methods, section 1.7). (B) Difference in the likelihood-ratio statistic between the two implementations, each fitting with its own optimizer, for 40 synthetic genes, each simulated with 120 cells. The largest difference is 0.023, within the agreement criterion of 0.05. (C) Exact upper-tail P by contour integration against the parametric-bootstrap P (199 re-simulations per gene at the fitted two-state rates) for the 206 genes of the calibration cohort with a complete bootstrap (Supplementary Methods, section 2.2); dotted, the bootstrap's resolution. (D) The detection map in full: the statistic of every simulated gene against its true mean silent period for 108 two-state genes re-simulated with two silent steps at their own rates (circles), 40 refractory genes re-simulated at their fitted two-step rates (triangles), the 48-point design grid (squares) and 24 genes simulated from the two-state model on the grid (diamonds); filled, detected (statistic above 8.6). No gene simulated from the two-state model is detected as multi-step; one grid point (silent period one hour, burst size twelve) has an anomalous statistic of 1,491 (cross; Supplementary Methods, section 4.1). (E) Statistic of the refractory genes re-simulated at their fitted two-step rates, compared with the data. On re-simulation 3 of 40 are re-detected and the median statistic is 2.0, below the observed values, indicating that the observed silences are more regular than two equal steps produce (Supplementary Methods, section 4.2). (F) Goodness of fit over 186 expression-matched pairs: standardized residual (observed minus fitted number of cells) at one new molecule and at four or more. Refractory genes have fewer cells with four or more new molecules than their class model predicts (median residual −3.4 versus 1.0 for the two-state partners under their model) and both classes have an excess of cells with one new molecule (8.6 and 6.9). (G) Depth thinning of the 186 refractory genes over 4,000 cells: 7 remain above threshold when each molecule is retained with probability 0.15 and 0 remain when the probability is 0.03 (median ratio of the thinned to the full statistic 0.024 at 0.15). (H) Label duration in an imaging-based design (15 simulated refractory-like genes, 19,603 cells): fraction of genes simulated with two silent steps that are classified as multi-step, and of genes simulated from the two-state model that are wrongly classified as multi-step, against label duration. Of the two-step genes, 73% are classified as multi-step with a 20-hour label, 27% with an eight-hour label and 0% with a two-hour label (Supplementary Methods, section 4.3).

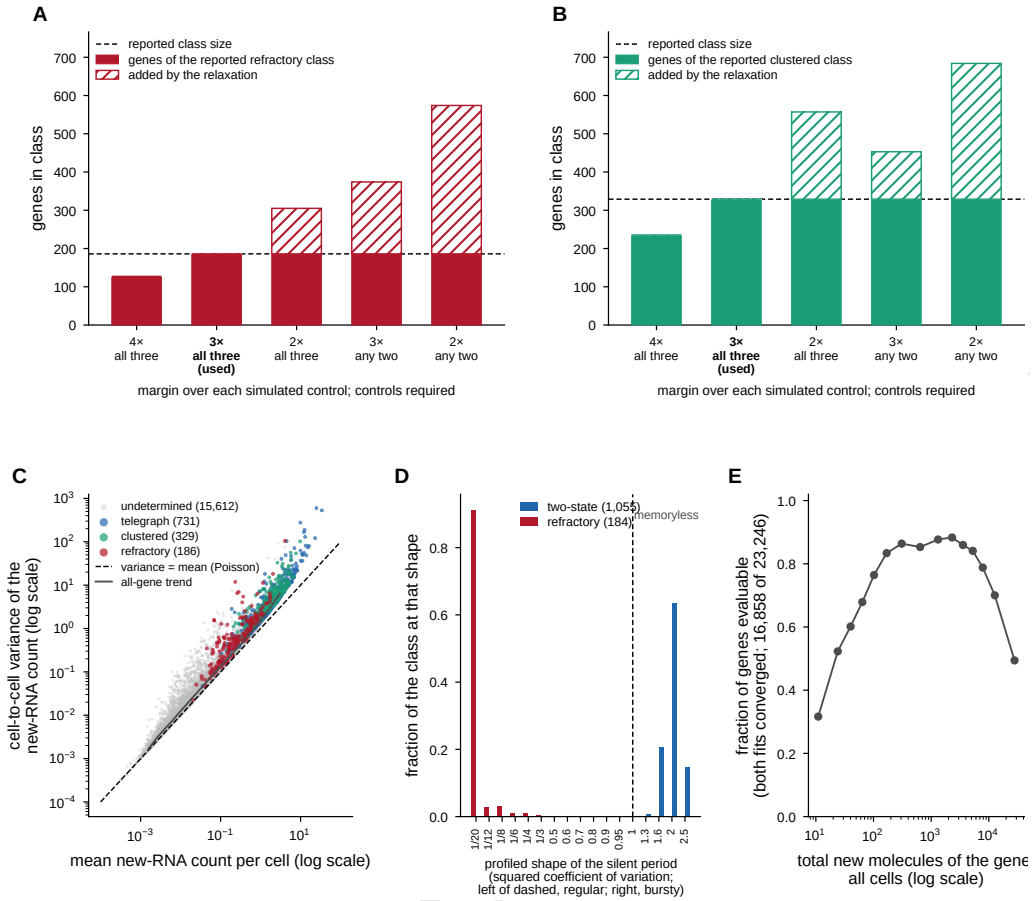

Figure S2. Class sizes under alternative control criteria and the shape of the silent period, related to Figure 1. (A) Refractory and (B) clustered class sizes recomputed with the margin over both gene-matched simulated controls set to four, three (the value used) or two times with all three controls (the two simulated controls and the G1 refit) required, and set to three or two times with any two of the three required. Filled, genes of the reported class; hatched, genes added by the relaxation; dashed, the reported size. Every relaxation contains the reported class in full (with the margin lowered to two times, 305 refractory and 557 clustered genes; with any two of the three controls required, 374 and 453; with both relaxations combined, 574 and 684), and the stricter four-times margin keeps 127 of the 186 refractory and 235 of the 329 clustered genes. Without controls, 1,231 genes exceed the refractory threshold and 3,737 are significant on the exact test alone; 744 of the 1,060 two-state genes exceed the clustered threshold. The median margin of a gene of the reported class over the larger of its two control statistics is 5.1 times in the refractory class and 5.2 times in the clustered class. (C) Cell-to-cell variance against mean of the new-RNA count for all 16,858 evaluable genes (raw counts over 8,916 cells); dashed, Poisson; line, the all-gene trend. Every classified gene lies above the Poisson line; refractory and clustered genes lie above the trend at matched mean (1.51- and 1.37-fold at the median; 91% and 92% of genes above), telegraph genes on it (1.06-fold; 68%). (D) The profiled shape of the silent period (its squared coefficient of variation, one coordinate running from many equal steps through the memoryless value to two-timescale mixtures; Supplementary Methods, section 3) for the 184 refractory and 1,055 two-state genes with a complete profile (of 1,658 profiled, 2,000 selected). The likelihood intervals of all 184 refractory genes lie on the regular side, and for 168 of them the likelihood is highest at the smallest value profiled (one over 20), where the profiled value is only an upper bound on the true one. The intervals of 1,047 two-state genes lie on the bursty side and those of 4 include the memoryless value. (E) Fraction of the 23,246 fitted genes that are evaluable (both fits converged) against the total number of new molecules of the gene. The median is 928 molecules for evaluable genes and 53 for genes where neither fit converged. The evaluable fraction falls at the highest counts, where the count grids are largest and fits converge less often.

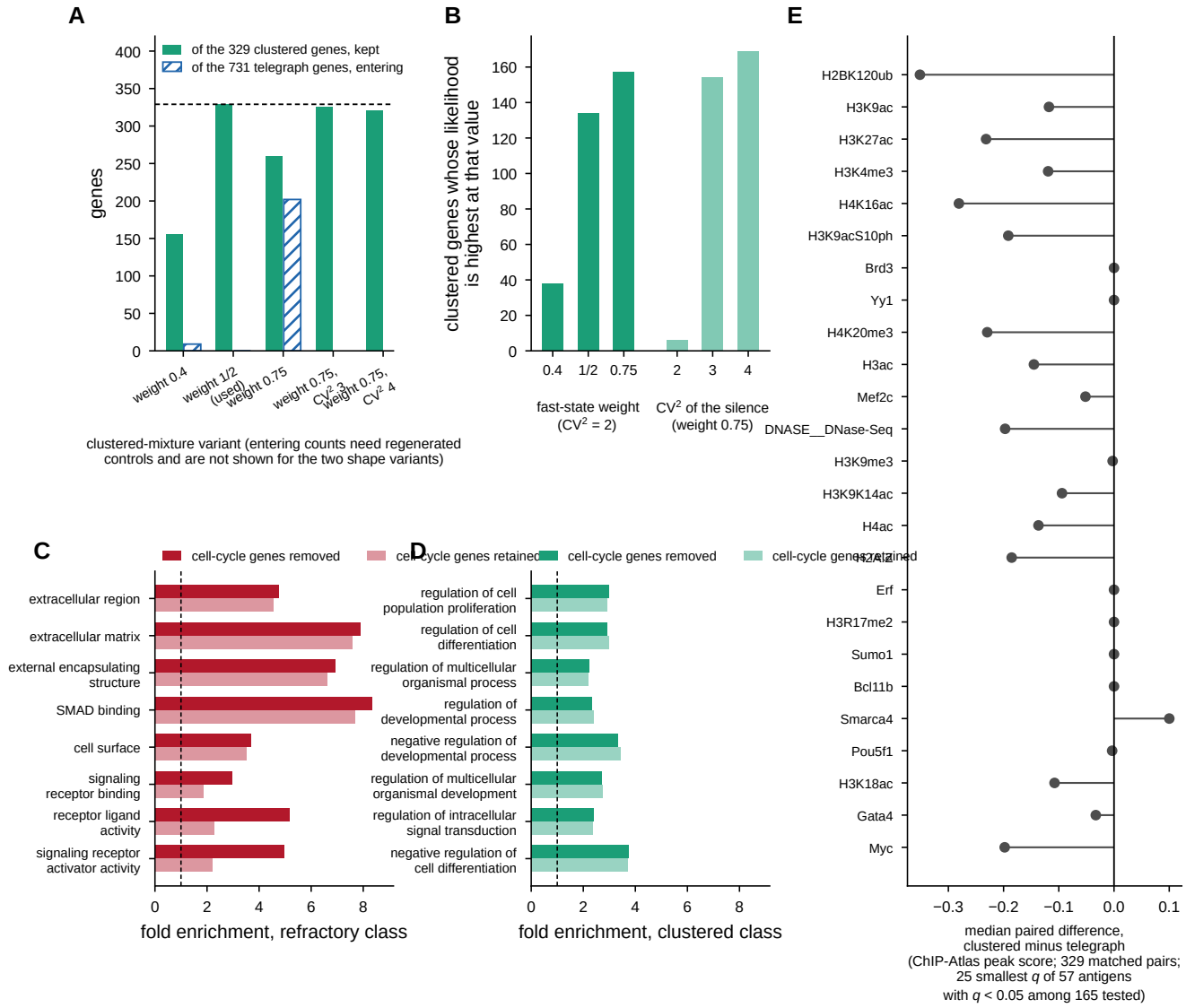

Figure S3. The clustered class under other mixture conventions, and class enrichments with and without cell-cycle genes, related to Figures 2 and 3. (A) Numbers of the 329 clustered genes that remain in the class, and of the 731 telegraph genes that enter it, when the fast-state weight of the mixture is moved from one half to 0.4 (156 remain, 9 enter) or to 0.75 (260 remain, 202 enter; 118 remain under all three weights). At weight 0.75, with the squared coefficient of variation raised from 2 to three or four, 326 and 321 remain (the telegraph genes entering at these shapes are not counted, because counting them would require controls regenerated at each shape). At each setting 1,033 two-state genes were refitted (27 lie above the count-grid limit). (B) The setting at which each clustered gene's likelihood is highest: fast-state weight 0.75 (157), one half (134) or 0.4 (38); squared coefficient of variation four (169), three (154) or two (6). Class membership therefore depends on the stated conventions, and a weight or shape per gene is not identifiable from one two-hour window at this depth (Supplementary Methods, section 3.5). (C) Gene Ontology fold enrichments of the refractory class with cell-cycle-annotated genes removed (164 annotated genes) or retained (171), with the 12,243 annotated evaluable genes as background: extracellular region 4.8-fold and 4.6-fold, respectively ($P = 1.3 \times 10^{-22}$ with cell-cycle genes retained), extracellular matrix 7.9-fold and 7.6-fold. (D) The same for the clustered class (299 annotated genes after removal). (E) The ChIP-Atlas promoter antigens with the smallest q values for a difference between clustered genes and their expression-matched telegraph partners (329 pairs). Active marks are modestly lower at clustered promoters, whereas promoter methylation (difference 1.5 percentage points, $P = 0.72$) and CpG-island overlap (78% versus 77%, $P = 0.85$) do not differ.

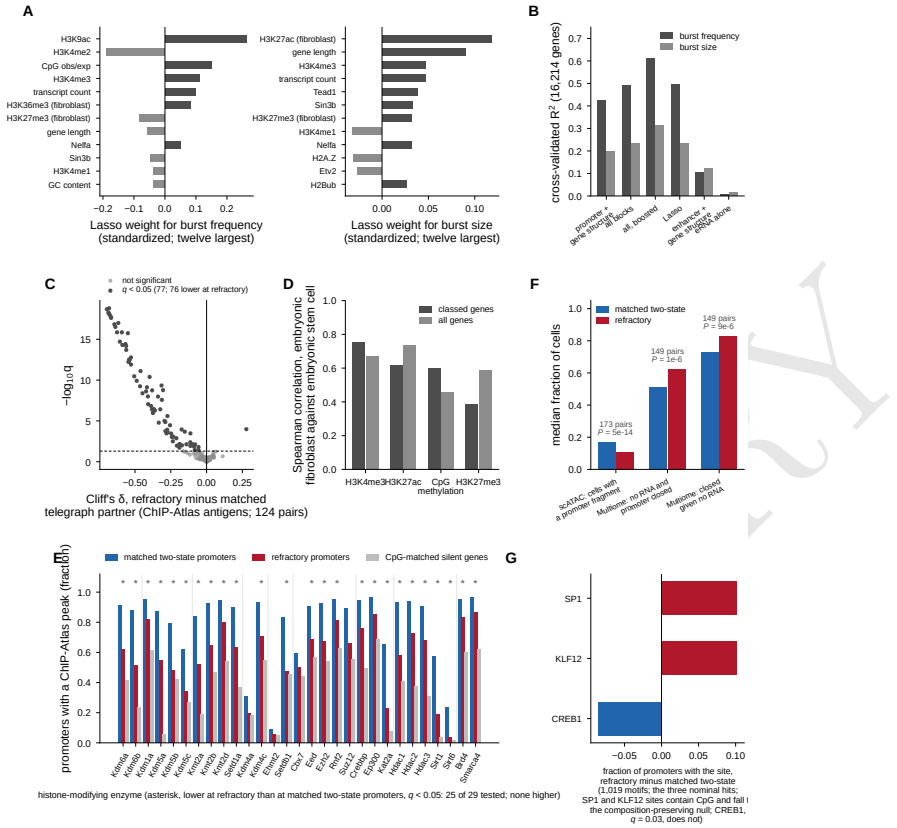

Figure S4. Chromatin regressions and promoter machinery, related to Figure 4. (A) The twelve largest Lasso weights for burst frequency (140 features kept, cross-validated $R^2 = 0.49$) and burst size (121 kept, $R^2 = 0.23$) over 16,214 genes with 220 standardized features (16 promoter and gene-structure features and, from embryonic-fibroblast ChIP-Atlas, 38 histone marks, 163 DNA-binding antigens, Pol II and accessibility). No transcription-factor weight exceeds 0.05 for frequency or 0.04 for size, and the largest transcription-factor weight for size is that of TEAD1 (+0.04). (B) Cross-validated R^2 for burst frequency and burst size by model: the promoter block and all blocks under ridge regression, all blocks under boosted trees (0.61 and 0.32, respectively), the Lasso, the enhancer block (linked candidate elements within 100 kb plus gene structure; 0.08 and 0.04) and enhancer RNA alone (0.006 and 0.014; increments over the full promoter model −0.0002 and +0.006). (C) Antigens differing between refractory promoters and expression-matched telegraph partners (124 pairs): 77 of the tested antigens differ at $q < 0.05$, 76 of them lower at refractory promoters. (D) Agreement of promoter marks between embryonic fibroblasts and embryonic stem cells over the classified genes and over all genes (Spearman; H3K4me3 0.76, H3K27ac 0.62, methylation 0.60, H3K27me3 0.39 over the classified genes), the basis for pooling experiments across cell types in Figure 4I. (E) Fraction of promoters with a ChIP-Atlas peak for 29 histone-modifying enzymes with data (of 33 surveyed) at matched two-state, refractory and CpG-matched silent promoters. Of these enzymes 25 are lower at refractory than at matched two-state promoters and none is higher ($q < 0.05$). Compared with CpG-matched silent genes (175 pairs), refractory promoters have more p300 and HDAC1, less H3K27me3 and H3K9me3 (silent over refractory 2.3- and 2.1-fold) and methylation that is not significantly different (35% versus 45%, $P = 0.70$), and they do not differ in any Polycomb component (0 of 5). (F) Three measures of promoter chromatin state in matched pairs. The first is the fraction of untreated fibroblasts with a single-cell ATAC-seq fragment at the promoter (173 pairs; the classes do not differ in the zero-inflation index, $P = 0.27$). The other two are computed from joint RNA and ATAC profiles of 1,935 cells: the fraction of cells with no RNA of the gene and a closed promoter (refractory 0.62 versus two-state 0.51 over 149 pairs re-matched on the RNA data, $P = 1 \times 10^{-6}$) and, among cells with no RNA of the gene, the fraction with a closed promoter. (G) Nominally significant motifs from a scan of 1,019 vertebrate motifs in refractory versus matched two-state promoters (186 pairs). The two motifs nominally more frequent in refractory promoters contain CpG and are not significant under a composition-preserving null (930 draws). The only significant motif is a CREB-type site more frequent in the two-state class ($q = 0.034$; 8.6% of matched two-state promoters), and the power of the scan is limited by the 186 matched pairs.

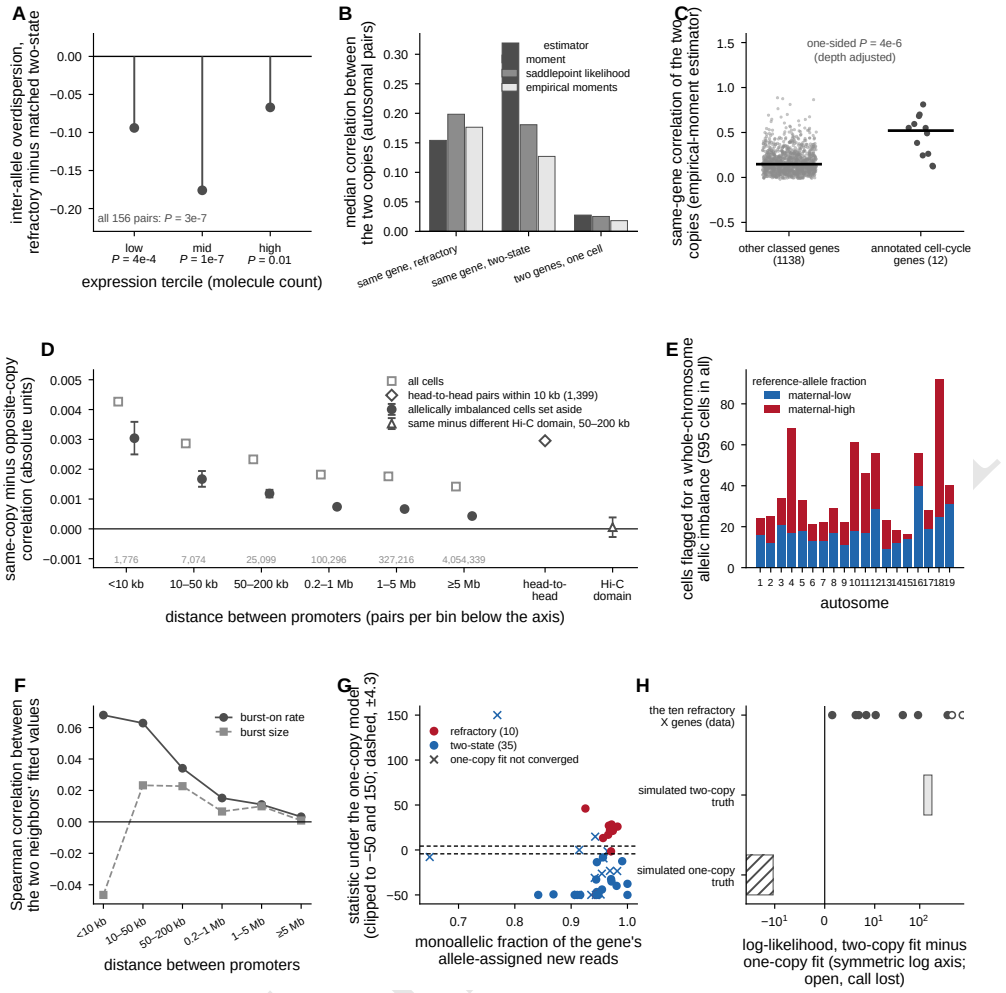

Figure S5. Allele statistics and their controls, related to Figure 5. (A) The difference in inter-allele overdispersion, refractory gene minus its matched two-state partner, within each expression tercile of the 156 matched pairs (low -0.094, $P = 4 \times 10^{-4}$; mid -0.176, $P = 1 \times 10^{-7}$; high -0.067, $P = 1 \times 10^{-2}$; all pairs $P = 3 \times 10^{-7}$). Because refractory genes have 4.22 allele-count units per new molecule compared with 1.67 for two-state genes, the contrast at matched expression, not the level, is interpreted. (B) Median correlation between the two copies under the three estimators of the bivariate-mixing model (Supplementary Methods, section 5.2) over the 149 autosomal pairs: same refractory gene 0.14, 0.19 and 0.18 (moment, likelihood, empirical); same two-state gene 0.32, 0.18 and 0.13; two different genes in one cell 0.03 (empirical 0.02). Under all three estimators the two copies of one gene are correlated in either class, but which class has the larger correlation depends on the estimator. (C) Positive control: the 12 annotated cell-cycle genes among the classified genes have a same-gene correlation of 0.52, compared with 0.15 for the rest (one-sided $P = 4 \times 10^{-6}$, adjusted for depth). (D) The neighbor statistic (same-copy minus opposite-copy correlation, absolute units) by distance, computed with all cells and again after setting aside the 595 cells (6.7%) that carry a chromosome with more than 80% of allele-assigned reads from one parent. Setting these cells aside removes most of the distance-independent component (0.0014 to 0.0004) and leaves the short-range excess (0.0028 to 0.0026; 0.0030 within ten kilobases). Head-to-head pairs within ten kilobases, which cannot share a read-through transcript, give 0.0030 (1,399 pairs). Pairs in the same Hi-C insulation domain exceed pairs in different domains at matched distance by 0.0001 (-0.0003 to 0.0004), compared with a distance effect of 0.0023 in the same distance bin, and pairs on different chromosomes give -0.00007 (s.e. 0.00007, 20,000 pairs). (E) Cells set aside for allelic imbalance, per autosome (16 to 92), split by the direction of the imbalance. (F) Spearman correlation between neighbors' fitted burst-on rates (0.07 within ten kilobases, 0.06 at ten to fifty, 0.003 beyond five megabases) and burst sizes (at most 0.05 in absolute value). (G) The 45 confidently classified X-linked genes refitted under one copy: monoallelic fraction (refractory genes 93 to 98%) against the one-copy statistic. One-copy fits converged for 68% of these genes and two-copy fits for 96%, and two genes of the two-state class acquire a nominally multi-step statistic, both with unconverged one-copy fits. (H) The simulated log-likelihood differences that define the shaded band of Figure 5H: a two-copy gene fitted with one copy loses 123 to 187 log-likelihood units and a one-copy gene gains 11 to 43, both simulated at the rates of the two genes reclassified under one copy; the observed differences of the ten refractory X-linked genes are overlaid.

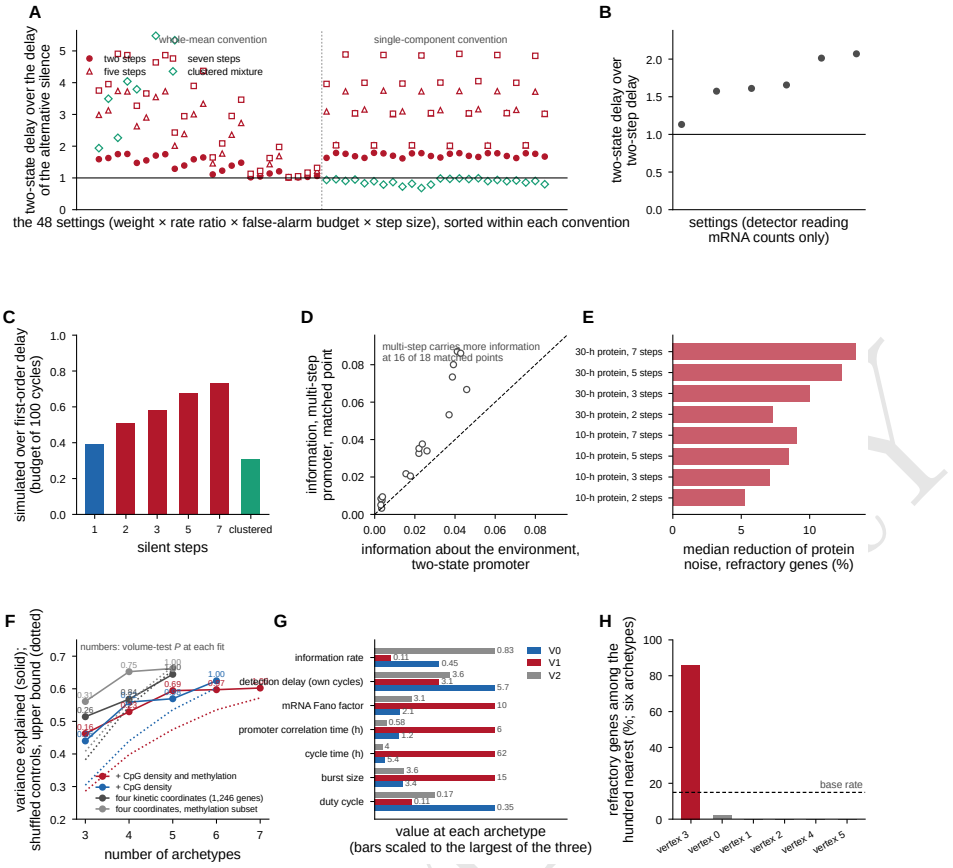

Figure S6. Detection, noise and information calculations and the archetype analysis, related to Figure 6. (A) Ratio of the detection delay of the two-state promoter to that of a promoter with a two-, five- or seven-step silence and to that of the clustered promoter, at equal mean interval, over the 48 settings (40,000 realizations each; Supplementary Methods, section 6.3). When the change to be detected scales the mean of every component of the silent period (whole-mean convention; 2.8- to 24-fold changes) the five-step ratio ranges from 2.0 to 3.8 (median 3.1) and the two-step ratio from 1.6 to 1.8. With only the fast component of the clustered promoter's silent period scaled (single-component convention; the mean interval then changes 1.01- to 2.7-fold), the corresponding ranges are 1.02 to 3.7 and 1.01 to 1.75. Detection with the clustered promoter is never faster than with the two-state promoter. (B) The same comparison with a detector that observes only the mRNA counts and not the promoter state (44 settings of the *TFF1* grid recomputed at the count level; Supplementary Methods, section 6.2): the change is detected sooner with the two-step silence at every setting. (C) Finite-budget factors, simulated delay over the first-order delay at a false-alarm budget of 100 cycles: 0.39 (one step), 0.51 (two), 0.58 (three) and 0.31 (clustered). (D) Information that a population of 150 readout cells carries about a switching environment, for the multi-step versus the two-state promoter at 18 matched combinations of duty cycle, mean and switching rate. When the environment modulates the burst-on rate, this information is higher with the multi-step promoter by more than two standard errors at 16 of these combinations (median ratio 1.6; the two-standard-error criterion was chosen post hoc). The low-state Fano factor of the multi-step promoter is the lower of the two at all 18 matched combinations. Across all 36 combinations (the matched combinations under both architectures), modulation of the synthesis rate (the burst-size channel) is the most informative at 31 of them and modulation of the burst-on rate (the burst-frequency channel) at five. Each information value is a lower bound on the information about the full environment path (Supplementary Methods, section 6.5). (E) Median reduction of protein noise over the 186 refractory genes when the silent period has two to seven steps at fixed rates, for ten-hour and thirty-hour proteins (5 to 13% of the coefficient of variation). (F) Variance explained by principal convex hull fits compared with the upper bound of column-shuffled controls at every number of archetypes, for the four kinetic coordinates (51% versus 38% at three; 29% versus 17% at two) and with CpG density and methylation added (612 genes with methylation; CpG density and methylation correlate with the burst-on rate at 0.48 and −0.40). By the volume test, none of these fits is a significant polytope ($P \geq 0.16$ for each). (G) Quantities at the three archetypes (V0, V1 and V2, where V1 is the refractory vertex): silent period 3.5, 55 and 3.3 h; duty cycle 0.35, 0.11, 0.17; Fano factor 2.1, 10, 3.1; detection delay 5.7, 3.1 and 3.6 own cycles (31, 189 and 14 h); information rate 0.45, 0.11, 0.83; the synthesis rate at V2 is 2.8 times that at V0. (H) Refractory genes among the hundred genes nearest each archetype of the six-archetype fit: at one archetype 86% of them are refractory, compared with a base rate of 15% ($P = 4 \times 10^{-50}$, Fisher's exact test); archetype positions are stable over 200 bootstrap resamples.

Supplementary tables

Tables S1–S6 are provided as separate Excel files. The first sheet of each file, named Legend, repeats the title and legend given here and defines every column of the data sheets.

Table S1. Fitted rates, test statistics, controls and class of every fitted gene, related to Figure 1. Each row of the first sheet is one fitted gene (23,246 genes). The columns give the fixed inputs (mRNA degradation rate and count grid), an indicator of whether each fit converged, the fitted rates of the two-state, multi-step and clustered models, and the likelihood-ratio statistic with its exact P. Further columns give the statistics of the capture-jitter and copy-number controls, the outcome of the G1 refit, the class, and membership in the refractory and clustered classes under every variant of the class rule (Figures S2A and S2B). The last columns give the mean and variance of the new-RNA count per cell and the fraction of cells with a new molecule, the profiled shape of the silent period (Figure S2D), and the mean silent and active periods under the model of the gene's class. Rates are per hour, and the degradation rate of every fitted gene is the per-gene estimate published in the source data of the NASC-seq2 study (Methods), entered as a fixed input. The 39 genes whose count range exceeded the largest feasible grid are listed on the second sheet. Sheets: fitted genes; not fitted.

Table S2. Gene Ontology enrichments and gene lists of the refractory, clustered and telegraph classes, related to Figures 2 and 3. Each class sheet gives the fold enrichment and Benjamini–Hochberg-adjusted P of Gene Ontology terms in that class against a background of the evaluable genes with a Gene Ontology annotation, computed once with cell-cycle-annotated genes removed and once with them retained (Figures 2A, 3F, S3C and S3D). The last sheet lists every classified gene with its class. Sheets: refractory; clustered; telegraph; class gene lists.

Table S3. Chromatin features, regressions of burst kinetics on chromatin, and matched promoter contrasts between the classes, related to Figure 4. The first three sheets define every promoter, gene-structure, histone-mark, DNA-binding-factor, polymerase and accessibility feature and give the cross-validated R^2 of the regression of each burst parameter on successive feature blocks, together with the Lasso coefficient of each standardized feature (Figures 4D, S4A and S4B). The next four compare ChIP-Atlas antigens at refractory and at clustered promoters with those at expression-matched telegraph partners (Figures S4C and S3E), CpG-binding enzymes at refractory promoters with those at matched two-state partners (Figure 4I), and histone-modifying enzymes at refractory promoters with those at matched two-state and at CpG-matched silent promoters (Figure S4E). The eighth gives the agreement of promoter marks between embryonic fibroblasts and embryonic stem cells (Spearman correlation over genes; Figure S4D), the basis for pooling ChIP-seq experiments across cell types. The remaining four list the predictions stated before the enzyme analyses with their outcomes, the motif scan (Figure S4G), single-cell promoter accessibility (Figures 4H and S4F) and PRO-seq pausing indices in two datasets (Figure 4C). Contrasts use expression-matched pairs unless stated, occupancy is the ChIP-Atlas peak score, and q denotes the Benjamini–Hochberg-adjusted P. Sheets: feature definitions; cross-validated R^2; Lasso weights; refractory vs telegraph antigen; clustered vs telegraph antigens; CpG-binding enzymes; histone enzymes; fibroblast vs stem cell marks; pre-specified predictions; motif scan; single-cell accessibility; PRO-seq pausing.

Table S4. Allele-resolved statistics and one-copy refits of X-linked genes, related to Figure 5. The first sheet gives the median co-bursting statistic of refractory, cross-class and two-state gene pairs and the difference between the refractory and cross-class medians with its permutation null (Figures 5C and 5D). The second gives the inter-allele overdispersion of refractory genes and their expression-matched two-state partners (Figures 5A, 5B and S5A). The third gives the per-gene estimates of the

between-copy correlation under the bivariate-mixing model from the moment, likelihood and empirical estimators (Supplementary Methods, section 5.2; Figures 5E and S5B). The fourth gives the same-copy minus opposite-copy correlation of neighboring genes by distance, computed once over all cells, once with the allelically imbalanced cells set aside, and once over G1 cells without the imbalanced cells (Figures 5F and S5D). The next two list the most extreme neighbor pairs, which include the imprinted pairs, and the number of cells set aside for allelic imbalance on each autosome, split by the direction of the imbalance (Figure S5E). The last two give the 45 X-linked genes refitted under one active copy, with their rates and statistics under one and two copies (Figures 5G, 5H, S5G and S5H). The remaining classified X-linked gene, the two-state gene *Tmsb4x*, was not refitted because its counts exceed the feasible grid. In the column labels, R and T denote refractory and two-state genes, respectively, or same-class pairs of such genes on the co-bursting sheet, and X denotes cross-class pairs of one refractory and one two-state gene. A same-gene correlation is between the two parental copies of one gene and a cross-gene correlation between copies of two different genes in the same cell. Lower inter-allele overdispersion means more strongly correlated alleles. Sheets: co-bursting; inter-allele overdispersion; bivariate mixing per gene; neighbor bins; extreme neighbor pairs; imbalance per chromosome; X-linked genes; X-linked fits.

Table S5. Detection-delay, noise and information calculations and the archetype analysis, related to Figure 6. The first sheet gives the detection delays of the optimal sequential detector by the number of steps in the silent period, at the *TFF1* operating point and its variants (Figures 6A–6C and S6B). The second gives the same delays at the 48 settings that vary the interval distribution, the false-alarm budget and the size of the change (Figure S6A). The third gives, for every classified gene at its fitted rates, the closed-form Fano factor, the information per burst cycle and per hour about a step change in its input, and the first-order and finite-budget detection delays (Figures 6D, 6E and S6C). The fourth and fifth give the information that a population of readout cells carries about a switching environment under multi-step and two-state promoters, over the full grid of settings and at the settings matched in duty cycle, mean and switching rate (Figure S6D; Supplementary Methods, section 6.5). The sixth gives the worked example of the burst-train information rate (Supplementary Methods, section 6.6). The last three give the rates of the three archetypes, the kinetic and functional quantities at each vertex, and the comparison of the hundred genes nearest archetype V0 with the hundred nearest archetype V2, the two vertices occupied by two-state genes (Figures 6F–6H and S6F–S6H). Archetype coordinates and neighborhoods are descriptive only, because the classified genes do not form a significant polytope by the hull-volume test at any number of archetypes for which the test is defined (Figure S6F). Sheets: TFF1 detection grid; 48 settings; closed forms per gene; environment encoding; encoding, matched points; information rate example; archetypes (3); vertex quantities; V0 vs V2 nearest 100.

Table S6. Detection power of the design and numerical checks of the likelihood and the exact test, related to Figure 1 and Supplementary Methods. The first three sheets give every simulated gene of the detection map with its true rates, statistic and call (Figures 1F and S1D), the statistics of the refractory genes after depth thinning (Figure S1G), and the fraction of simulated genes classified as multi-step by label duration (Figure S1H). The next two compare the two implementations of the likelihood, by their values at 64 parameter points (Figure S1A) and by the likelihood-ratio statistics of 40 synthetic genes fitted with each of them (Figure S1B). The sixth gives the calibration cohort of genes on which the exact test was compared with the parametric bootstrap, with both P values per gene (Figure S1C; Supplementary Methods, section 2.2). The last two give the likelihood profiles over the number of silent steps for the genes of Figure 1G and the profiles of the shape of the silent period (Figure S2D). For simulated genes the rates listed are those used in the simulation, and every statistic is $2(\log L_{\text{multi-step}} - \log L_{\text{two-state}})$ unless stated otherwise in the column label. Sheets: detection map; depth thinning; label duration; two implementations, 64 points; synthetic genes; bootstrap cohort; depth profiles (32 genes); shape profiles.