

Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics

David M. Suter,^{1*} Nacho Molina,^{2*} David Gatfield,^{1,3} Kim Schneider,¹ Ueli Schibler,^{1†‡} Felix Naef^{2†‡}

In prokaryotes and eukaryotes, most genes appear to be transcribed during short periods called transcriptional bursts, interspersed by silent intervals. We describe how such bursts generate gene-specific temporal patterns of messenger RNA (mRNA) synthesis in mammalian cells. To monitor transcription at high temporal resolution, we established various gene trap cell lines and transgenic cell lines expressing a short-lived luciferase protein from an unstable mRNA, and recorded bioluminescence in real time in single cells. Mathematical modeling identified gene-specific on- and off-switching rates in transcriptional activity and mean numbers of mRNAs produced during the bursts. Transcriptional kinetics were markedly altered by cis-regulatory DNA elements. Our analysis demonstrated that bursting kinetics are highly gene-specific, reflecting refractory periods during which genes stay inactive for a certain time before switching on again.

Polymerase II-mediated transcription of mammalian genes is a complex process consisting of several consecutive steps (1). Studies in prokaryotes (2–4), yeast (5–7), and higher eukaryotes (8–11) have suggested that genes are transcribed in a discontinuous fashion, resulting in stochastic production of RNA and protein molecules. Stochastic gene expression has been linked to phenotypic variability, for example, in the resistance to antibiotics in bacterial populations or in the control of developmental transitions in metazoans (8, 12–14). Transcriptional bursts can be abstracted in the random telegraph model of gene expression (2, 9, 15–17),

whereby transcription switches between “on” and “off” states (Fig. 1A). We monitored transcription kinetics by single-cell time-lapse bioluminescence imaging of mouse fibroblasts expressing a short-lived luciferase reporter gene controlled by endogenous loci, circadian regulatory sequences, or artificial promoters (Fig. 1B). Mathematical modeling allowed us to reconstruct temporal changes in mRNA and protein copy numbers as well as the gene activity state (“on” or “off”) at a resolution of 5 min over extended time periods.

Our strategy to monitor transcription is based on the following principle: If both the mRNA

transcribed from a gene and its protein product are short-lived, the fluctuations of protein expression should closely reflect the transcriptional bursting kinetics (Fig. 1A). Although this approach cannot capture abortive transcription events, it should reflect the production of mature mRNAs. We engineered a short-lived nuclear luciferase (NLS-luc) encoded by a short-lived mRNA, and fused it to blasticidin deaminase, separated by the 2A peptide of foot-and-mouth disease virus, to generate two polypeptides from one coding sequence (18) (Fig. 1B). This cassette was integrated into three types of constructs: (i) a gene trap (GT) lentivector, allowing transcription to be controlled by the endogenous locus at the insertion site (19) (the resulting cell lines were dubbed “GT:gene name”); (ii) vectors conferring circadian expression by the *Bmal1* promoter (20) (*Bmal1*/NLS-luc) or the *Dbp* (D site of albumin promoter binding protein) gene (*Dbp*/NLS-luc); and (iii) vectors carrying artificial promoters consisting of one or two copies of different CCAAT-box variants followed by a TATA box (21) (table S1). These constructs were used to generate

¹Department of Molecular Biology, Sciences III, University of Geneva, and National Centre of Competence in Research Frontiers in Genetics, 30 Quai Ernest Ansermet, 1211 Geneva, Switzerland.

²Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics, AAB 021 Station 15, CH-1015 Lausanne, Switzerland.

³Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland.

*These authors contributed equally to this work.

†These authors contributed equally to this work.

‡To whom correspondence should be addressed. E-mail: ueli.schibler@unige.ch (U.S.); felix.naef@epfl.ch (F.N.)

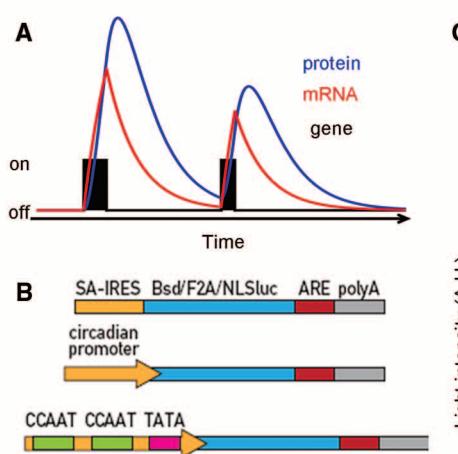
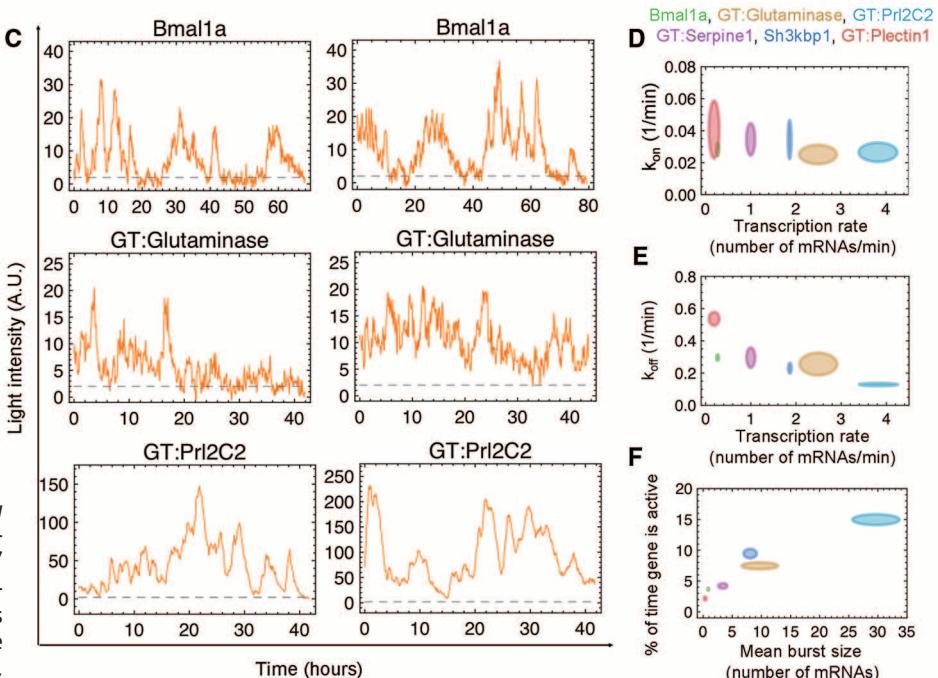


Fig. 1. Different transgene insertion sites show highly stereotyped luminescence profiles. (A) Schematic for a gene switching between “on” and “off” states and expressing short-lived mRNA and proteins. (B) Schematic representation of the vectors used to generate the different cell lines. SA, splice acceptor; IRES, internal ribosomal entry site; Bsd, blasticidin deaminase; F2A, foot-and-mouth virus peptide 2A; NLS-luc, destabilized nuclear luciferase; polyA, polyadenylation signal; ARE, AU-rich element. (C) Examples of single-cell traces. (D and E) Relationships between transcription rate and k_{on} (D) or k_{off} (E) confirm gene specificity of transcriptional kinetics. (F) Mean burst size versus percentage of time during which the gene is active. Ellipses represent means \pm 2SD. Color keys are shown on top of (D).



specificity of transcriptional kinetics. (F) Mean burst size versus percentage of time during which the gene is active. Ellipses represent means \pm 2SD. Color keys are shown on top of (D).

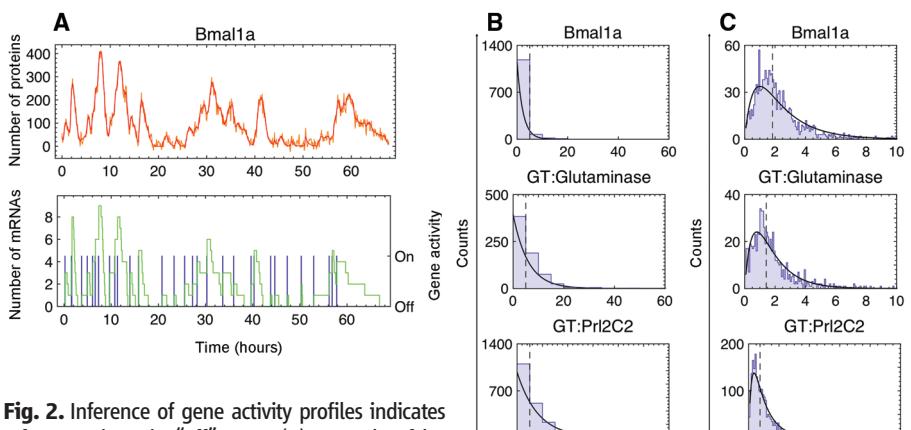


Fig. 2. Inference of gene activity profiles indicates refractory times in “off” state. **(A)** Example of luminescence trace (orange), protein copy number (red), mRNA copy number (green), and gene activity inferences (blue) for the Bmal1a cell line. **(B)** Distribution of “on” intervals; black lines show exponential fits. **(C)** Distribution of “off” intervals; black lines show best fits to “two-step” model (21). Black vertical dotted lines show medians of distributions.

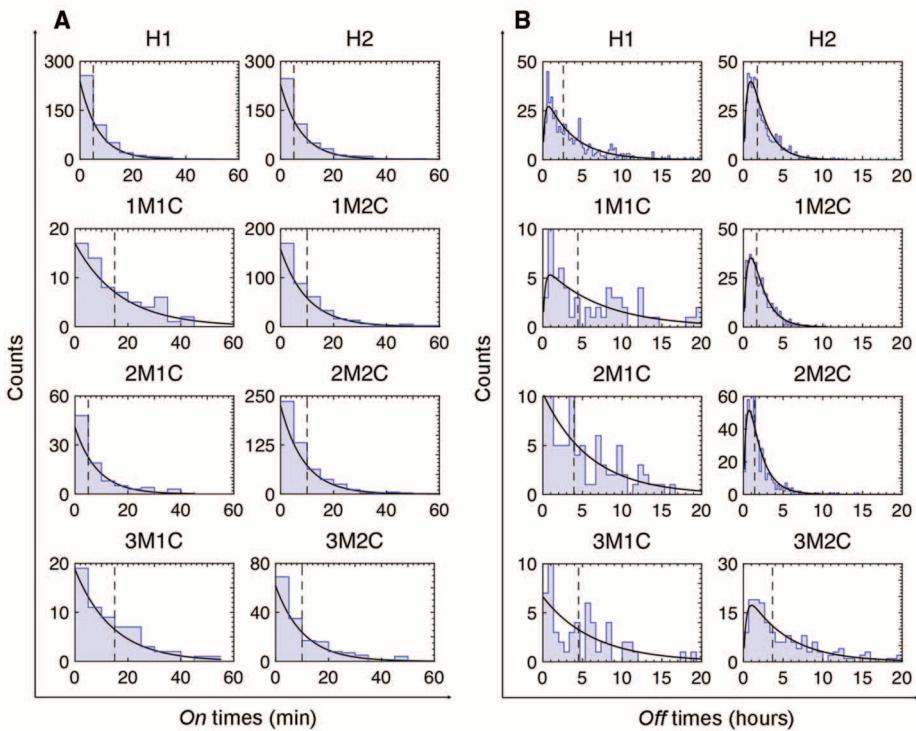


Fig. 3. Influence of promoter architecture on transcriptional bursting. **(A and B)** Cell lines carrying a single reporter gene containing either one (left panels) or two (right panels) CCAAT boxes with different affinities for NF-Y. “H” indicates high affinity for NF-Y, “M” the number of mutations surrounding the CCAAT sequence resulting in decreased affinity for NF-Y (21), and “C” the copy number of CCAAT boxes. **(A)** Distribution of “on” intervals; black lines show exponential fits. **(B)** Distribution of “off” intervals. Black lines show best fits to “two-step” model (21). Black vertical dotted lines show medians of distributions. **(C)** Mean burst sizes versus percentage of time during which the gene is active. Ellipses represent contours of mean burst sizes \pm 2SD. The numbers 0 to 3 refer to the numbers of mutations adjacent to the CCAAT box sequence.

monoclonal NIH-3T3 cell lines. We also isolated primary mouse fibroblasts from heterozygous and homozygous Per2::luciferase knock-in mice (22). Throughout the recordings, we used non-dividing cells to minimize confounding noise sources. All cell lines are listed in table S2.

Figure 1C and fig. S1A show examples of temporal bioluminescence profiles for the cell lines mentioned above. Examination of these data by wavelet analysis (21) revealed frequency spectra characteristic for each cell line (fig. S2). As translation is invariably initiated at the internal ribosomal entry site (IRES), discontinuous translation was unlikely to account for the observed differences in temporal bioluminescence profiles. To confirm the transcriptional origin of discontinuous luciferase expression, we blocked transcription in three different cell lines and observed a smooth decay of the luminescence signal in all cells (21) (fig. S3). Next, we generated a cell line containing a single transgene that specified a mRNA harboring 24 repeats of MS2-binding sites in its 3' untranslated region (21) whose transcription is driven by two CCAAT boxes. Transcriptional bursting was confirmed by discontinuous appearance of spots in the nucleus upon expression of a yellow fluorescent protein-tagged MS2 (fig. S4).

To further specify the kinetic model, we measured protein and mRNA stabilities by observing luminescence decays in cell populations after translation or transcription inhibition, respectively (21) (fig. S5). The photon emissions were calibrated to luciferase protein concentrations by correlating the mean numbers of photons generated by a cell to the mean number of luciferase molecules per cell. This was accomplished for five cell lines spanning a range of expression levels by comparing luminescence emission for a known number of cells with that from a recombinant luciferase having the same specific activity as NLS-luc (21) (fig. S6). Knowing the number of molecules per cell, we imaged single cells under conditions used for the time-lapse imaging experiments (fig. S6A) to calibrate photon emission versus protein copy numbers. We also calibrated the number of mRNA molecules using real-time quantitative polymerase chain reaction (fig. S7) (21). These calibrations enabled us to develop a stochastic model to quantitatively analyze the data. We computed the probability of the time traces according to the random telegraph model, and computed optimal model parameters using a maximum likelihood approach. We estimated that switching rates between the “on” and “off” gene activity states (k_{on} and k_{off}), the transcription (k_m) and translation (k_p) rates (fig. S8). To visualize the kinetic parameters of transcription for each gene, we plotted k_m against k_{on} and k_{off} . These parameters fell into clear clusters for each cell line, thereby confirming that the kinetic signatures were highly gene-specific (Fig. 1, D and E). Mean burst sizes (i.e., the mean number of transcripts produced per burst) (Fig. 1F and fig. S9B) were in the range of previous quan-

tifications (5, 9, 10, 23). To verify that the transcription rate of the tagged mRNA reflected that of the untagged allele, we measured the amount of *Ctgf* mRNA in the GT:ctgf cell line (21) and obtained a relative transcription rate of $88 \pm 3\%$ as compared to the tagged *Ctgf* allele. Hence, the insertion of a luciferase cassette did not markedly affect the *Ctgf* transcription rate.

Our model could also reconstruct the most probable temporal sequence of protein accumulation, mRNA accumulation, and gene activity states (Fig. 2A and fig. S9A). From this, we computed the distributions of time intervals during which the gene remained “on” and “off,” respectively. The “on” intervals followed exponential distributions, suggesting a first-order off-switching of gene transcription (Fig. 2B and fig. S9C). In contrast, the “off” intervals showed a local maximum that was best described by assuming two sequential exponential processes, indicating a refractory period in the “off” state before the gene can be switched on again (Fig. 2C and figs. S9D and S10A). The burst frequency and size of the circadian genes *Bmal1a/b*, *Dbp*, and *Per2::luc* oscillate during a circadian cycle (fig. S11), and we observed a clear phase advance of the former as compared to the latter in the *Bmal1a/b* and *Dbp* cell lines (fig. S11). Therefore, burst size and frequency can be uncoupled during up- and down-regulation of a gene.

We next investigated the role of promoter architecture, previously reported to affect on- and off-switching of gene activity in yeast (24). To keep the genomic environment invariable, we used NIH-3T3 Flp-In cells to engineer cell lines carrying a single copy of a different artificial construct at the same genomic locus (21). To drive NLS-luc expression, the transgenes contain one or two CCAAT boxes (25) of different affinities for the ubiquitously expressed NF-Y transcriptional

activator upstream of a TATA box, (21) (table S1). We generated and analyzed a total of eight different cell lines in which the overall strength of the artificial promoters correlated with both the number of CCAAT boxes and their affinity for NF-Y (fig. S12). Indeed, using two CCAAT boxes instead of one, or choosing CCAAT boxes of higher affinity for NF-Y, increased the mean burst sizes (Fig. 3C, x axis). This reflected mainly an increase in the number of transcripts produced during the “on” times, because the duration of “on” times was largely unaffected (Fig. 3A). In addition, the decrease in the duration of “off” intervals (Fig. 3B and fig. S10A) resulted in a higher percentage of activity time in the presence of two CCAAT boxes (fig. S10B).

Finally, we investigated the effect of histone acetylation and chromosomal location on transcription kinetics (Fig. 4), as chromatin environment plays a major role in gene activity (9, 12, 26). First, we blocked histone deacetylation in four cell lines with trichostatin A (TSA) and recorded single-cell luminescence traces. This treatment abolished the circadian component of the *Bmal1a* cell line, yet did not drastically affect general aspects of the bursting patterns (fig. S13). In addition, the values for k_{on} and k_{off} were preserved for *Bmal1a* as well as GT:NcKap1 (fig. S8). In contrast, GT:Prl2c2 and H1 had markedly altered transcription rates (fig. S8) and burst sizes (Fig. 4C) in the presence of TSA, but discontinuous gene activity persisted. Thus, although TSA leads to a spectrum of responses, the qualitative bursting characteristics were not strongly affected. Furthermore, the different chromosomal insertion sites of the transgene in the *Bmal1a* and *Bmal1b* cell lines did not markedly influence the temporal activity profiles (Fig. 1, C and F, and figs. S1A and S9B). These results suggest that the chromatin environment plays a secondary role in shaping bursting patterns.

Our findings show that the bursting kinetics of single-allele mammalian genes measured in individual cells at high temporal resolution are highly gene-specific. Cis-acting regulatory elements play a dominant role in shaping transcriptional kinetics.

References and Notes

- N. J. Fuda, M. B. Ardehali, J. T. Lis, *Nature* **461**, 186 (2009).
- I. Golding, J. Paulsson, S. M. Zawilski, E. C. Cox, *Cell* **123**, 1025 (2005).
- L. Cai, N. Friedman, X. S. Xie, *Nature* **440**, 358 (2006).
- Y. Taniguchi *et al.*, *Science* **329**, 533 (2010).
- D. Zenklusen, D. R. Larson, R. H. Singer, *Nat. Struct. Mol. Biol.* **15**, 1263 (2008).
- R. Z. Tan, A. van Oudenaarden, *Mol. Syst. Biol.* **6**, 358 (2010).
- L. Cai, C. K. Dalal, M. B. Elowitz, *Nature* **455**, 485 (2008).
- A. Raj, S. A. Rifkin, E. Andersen, A. van Oudenaarden, *Nature* **463**, 913 (2010).
- A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, *PLoS Biol.* **4**, e309 (2006).
- J. R. Chubb, T. Trcek, S. M. Shenoy, R. H. Singer, *Curr. Biol.* **16**, 1018 (2006).
- A. Paré *et al.*, *Curr. Biol.* **19**, 2037 (2009).
- A. Raj, A. van Oudenaarden, *Annu Rev Biophys* **38**, 255 (2009).
- A. N. Boettiger, M. Levine, *Science* **325**, 471 (2009).
- M. F. Wernet *et al.*, *Nature* **440**, 174 (2006).
- J. M. Pedraza, J. Paulsson, *Science* **319**, 339 (2008).
- D. R. Larson, R. H. Singer, D. Zenklusen, *Trends Cell Biol.* **19**, 630 (2009).
- J. Peccoud, B. Ycart, *Theor. Popul. Biol.* **48**, 222 (1995).
- M. D. Ryan, J. Drew, *EMBO J.* **13**, 928 (1994).
- W. L. Stanford, J. B. Cohn, S. P. Cordes, *Nat. Rev. Genet.* **2**, 756 (2001).
- E. Nagoshi *et al.*, *Cell* **119**, 693 (2004).
- See supporting material on *Science* Online.
- S. H. Yoo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 5339 (2004).
- S. Yunger, L. Rosenfeld, Y. Garini, Y. Shav-Tal, *Nat. Methods* **7**, 631 (2010).
- L. M. Octavio, K. Gedeon, N. Maheshri, *PLoS Genet.* **5**, e1000673 (2009).
- R. Mantovani, *Gene* **239**, 15 (1999).
- X. Darzacq *et al.*, *Annu. Rev. Biophys.* **38**, 173 (2009).
- We thank A. Liani and Y.-A. Poget for building a bioluminescence reader, C. Bauer and J. Bosset for help with time-lapse imaging, P. Salmon and D. Trono for providing lentiviral constructs, A. Oates for critical reading of the manuscript, and N. Roggli for the artwork. The computations were performed at Vital-IT (www.vital-it.ch). The U.S.’s laboratory was supported by the Swiss National Science Foundation (SNF 31-113565, SNF 31-128656/1, and the NCCR program grant Frontiers in Genetics), the European Research Council (ERC- 2009-AdG 20090506), the State of Geneva, and the Louis Jeantet Foundation of Medicine. F.N.’s laboratory was supported by the Swiss National Science Foundation (SNF grants 3100A0-113617 and 31-130714) and the Ecole Polytechnique Fédérale de Lausanne (EPFL).

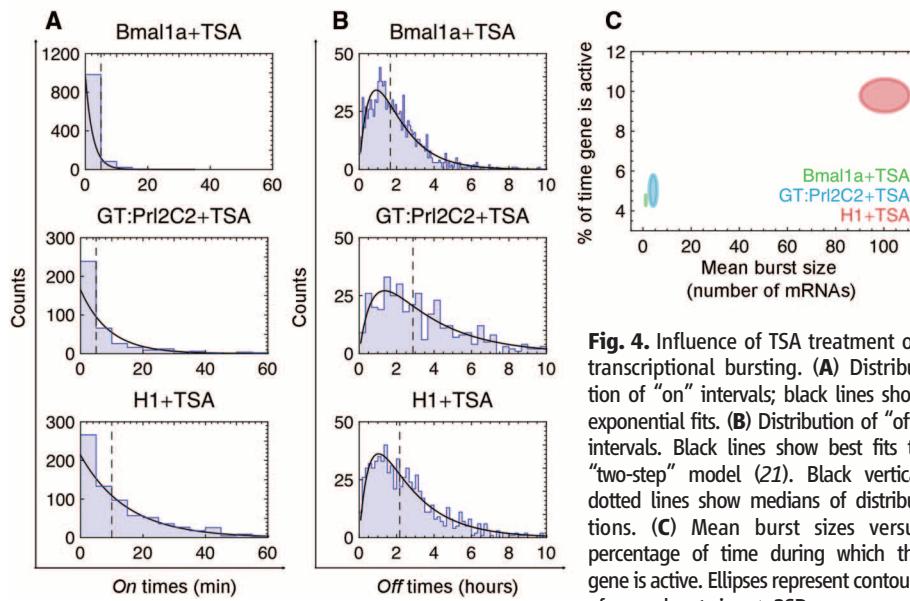


Fig. 4. Influence of TSA treatment on transcriptional bursting. (A) Distribution of “on” intervals; black lines show exponential fits. (B) Distribution of “off” intervals. Black lines show best fits to “two-step” model (21). Black vertical dotted lines show medians of distributions. (C) Mean burst sizes versus percentage of time during which the gene is active. Ellipses represent contours of mean burst sizes \pm 2SD.

Supporting Online Material

- www.sciencemag.org/cgi/content/full/science.1198817/DC1
Materials and Methods
Tables S1 to S3
Figs. S1 to S20
Movies S1 to S3
References

10 December 2010; accepted 4 March 2011
Published online 17 March 2011;
10.1126/science.1198817

Contents

1 Experimental methods	2
1.1 DNA constructs	2
1.2 Lentivector production and titration	3
1.3 Cell culture and transduction	3
1.4 Identification of insertion sites by modified 5'RACE and inverse PCR	4
1.5 Luminescence microscopy	5
1.6 Fluorescence time-lapse microscopy	5
1.7 Quantification of proteins	5
1.8 Quantification of mRNA	6
1.9 Correlation of protein and mRNA copy number with luminescence	7
1.10 Homogeneity of luciferase expression in cell populations	7
1.11 Supplementary movies	7
2 Computational methods	8
2.1 The telegraph model of gene expression	8
2.2 Protein and mRNA half-lives	10
2.3 Calibration and noise model	11
2.4 Parameter estimation and deconvolution	11
2.5 On-times and off-times histograms	13
2.6 Validation of the algorithm and performance of the inference	13
2.7 Comparison of the inference using a three vs. a two state model	14
2.8 Wavelet analysis	15
2.9 Burst frequency and size during a circadian cycle	15
3 Supplementary tables and figures	15
References	42

1 Experimental methods

1.1 DNA constructs

Lentivectors

To generate a gene trap lentivector construct, the following elements were assembled from 5' to 3': the adenovirus major late transcript splice acceptor [1]; the EMCV internal ribosomal entry site (IRES) [2]; STOP codons in all three reading frames; an open reading frame coding for three different coding sequences: the blasticidin deaminase, the 2A peptide from the foot and mouth disease virus [3] and a destabilized version of firefly luciferase (pgl4.24, Promega Corp.); a polyadenylation site. This construct was cloned in reverse orientation with respect to the 5' and 3'LTR of the p53rpapgkGFP lentivector (kindly provided by Patrick Salmon, University of Geneva Medical School). We then attempted to use this vector to generate gene trap cell lines, but were unable to obtain clones after blasticidin selection. As this might have been caused by insufficient expression of blasticidin resistance, we aimed at controlling mRNA stability by including one or several AU-rich element(s) in our construct. AU-rich elements are known to destabilize mRNAs, but under certain conditions, such as treatment with growth factors, they can have a stabilizing effect [4]. We thus integrated one or two AU-rich elements 5' to the polyadenylation site, transduced NIH-3T3 cells with the different lentivectors and monitored luminescence decay after actinomycin D treatment. Cells transduced with the lentivector carrying a single AU-rich element showed the fastest luminescence decay (Fig. S15A). Then, we monitored the luminescence signal of confluent NIH-3T3 cell populations transduced with the different lentivectors in the presence of basic fibroblast growth factor (bFGF) for several hours. We observed an increase of luminescence emission (Fig. S15B), suggesting that bFGF increased the stability and/or expression of the transgenic mRNA. These results encouraged us to pursue our experiments with the lentivector containing one AU-rich element. Finally, we integrated a nuclear localization signal to increase the local concentration of luciferase and thus the local bioluminescence signal strength. In addition, the nuclear localization reduced the half-life of luciferase from about 60 minutes to 22 minutes, which allows for a higher temporal resolution in measuring gene expression in individual cells. The final resulting gene trap lentivector is referred as to pGTNLS-luc throughout the text. To generate pBmal1/NLS-luc, the splice acceptor and IRES sequences of pGTNLS-luc were replaced by the Bmal1 promoter [5].

Flp-In constructs

To generate the pFRT-DBPNLS-luc construct, a DNA fragment containing an EMCV internal ribosomal entry site (IRES) followed by the firefly luciferase coding sequence was removed from pFRT-DBP-IRES-luc [6] and replaced by a fragment containing the EMCV IRES followed by bsdF2ANLS-luc, an AU-rich element and a polyadenylation signal. To generate the different CCAAT box reporter constructs, a DNA fragment containing bsdF2ANLS-luc, an AU-rich element and a polyA signal was cloned into pcDNA5/FRT. Two oligonucleotides encoding the top and bottom strands of a miniCMV promoter (pgl4.24, Promega corp.) flanked by cohesive ends were annealed and inserted 5' to the sequence described above. The different CCAAT boxes were designed according to previous work measuring in vitro binding affinities of NF-Y to different CCAAT box sequences [7]. The high affinity and low affinity CCAAT box sequences were directly chosen from the study of Bi et al. [7], whereas the 1M, 2M and 3M sequences were designed by insertion of one, two or three contiguous mismatches in the high affinity sequence. All oligonucleotides

were flanked by cohesive ends to MluI-digested DNA. Double stranded DNA fragments encoding the different CCAAT box elements were generated by annealing top and bottom oligonucleotides corresponding to each sequence, and DNA fragments containing one of the CCAAT boxes were cloned as one or two copies 5' to the miniCMV promoter, in antisense orientation. In the final constructs, the CCAAT box and the TATA box of the miniCMV promoter are spaced by 63 base pairs (six integral turns of DNA), and when two CCAAT boxes are present, they are spaced by 42 base pairs (four integral turns of DNA). We generated one additional cell line carrying a single reporter gene, with a promoter containing two high affinity CCAAT boxes as described above, with an additional cassette consisting of 24 tandem repeats of MS2-binding sites downstream of bsdF2ANLS-luc. We will refer to this cell line as H2-MS2.

1.2 Lentivector production and titration

The lentivectors were produced by transient transfection of 20 μ g of the lentivector construct, 15 μ g of the PAX2 and 5 μ g of the MD2G plasmids (kindly provided by Didier Trono, EPFL, Lausanne) using calcium phosphate, as previously described [8]. Titration of the lentivector was performed by transduction of HeLa cells followed by real time PCR of the integrated construct into genomic DNA, as described in [9].

1.3 Cell culture and transduction

NIH-3T3 cells were maintained in high glucose DMEM supplemented with 10% FBS and 1% penicilline/streptomycine/glutamine (Gibco). To generate the different gene trap NIH-3T3 cell lines, the lentivector was added on freshly trypsinized NIH-3T3 at a multiplicity of infection of 0.5-1. Approximately 1% of the cell population was expressing luciferase after inspection by luminescence microscopy, consistent with single active insertions. Three days after transduction, cells were treated with blasticidin at a final concentration of 1 μ g/ml and recombinant bFGF (R&D systems) at a final concentration of 100 μ g/ml for four days. bFGF allowed to transiently increase expression of bsdF2ANLS-luc through increased stability of the mRNA, which was mediated by the AU-rich element ([4]). This increased blasticidin resistance during the treatment by bFGF. Blasticidin selection without further bFGF treatment was continued for five additional days. Next, cells were diluted to a density of approximately 50 cells per 10cm dish. 9 to 13 days later, single clones were picked and replated into individual wells of a 24-well plate. One week later, the clones were split into black-walled 24 well plates (Genetix) and time-lapse luminescence recordings were performed in a home-made luminescence reader (André Liani and Yves-Alain Poget, unpublished). To generate the BmalNLS-luc NIH-3T3 cell lines, cells were transduced with pBmalNLS-luc and subsequently diluted to generate single clones. Each clone was analyzed by time-lapse luminescence recording, and three of them were selected for the mapping of their insertion sites. To generate the different Flp-In cell lines, NIH-3T3 cells carrying one single FRT site were purchased from Invitrogen. These cells were cotransfected with pOG44 and the different Flp-In constructs as recommended by the manufacturer. Cells were subsequently cultured for 48 hours at 30°C, and split in 10cm dishes. 24 hours later, the medium was supplemented with hygromycin B at a final concentration of 150 μ g/ml. Two weeks later, individual clones were pooled and maintained in medium containing hygromycin B at a final concentration of 50 μ g/ml. To prepare primary mouse tail fibroblasts, six month old male litter mate mice carrying one or two alleles of the mPer2Luciferase gene (Yoo et al., 2004) were anesthetized and a small piece of the tail tip (ca. 3 mm) was removed using a pair of scissors. The tail tip was cut into smaller pieces with a scalpel and

digested overnight in 2 ml culture medium (DMEM, 15% FCS, 1% PSG, 1% non-essential amino acids, 1 mM sodium pyruvate, 87 mM b-mercaptoethanol, 18 mM HEPES pH 7.0) containing 1 mg/ml collagenase type 1A (Sigma) and 1% Amphotericin B. The next morning, digested tail fragments and liberated cells were washed using 10 ml culture medium, collected by centrifugation, seeded to a 10 cm culture dish in 10 ml culture medium, and kept in a low-oxygen incubator (5% CO₂, 3% O₂). The medium was replaced every 5 days for the time that fibroblast colonies appeared and became confluent (typically after ca. 2 weeks). From then, cells were split every 5 days 1:3 and used for the described imaging experiments. To inhibit histone deacetylation, cells were pretreated for 24 hours and during time-lapse imaging with trichostatin A at a final concentration of 2µM. The effect of TSA treatment on histone acetylation was confirmed by immunofluorescence staining (Fig.S16).

1.4 Identification of insertion sites by modified 5'RACE and inverse PCR

To identify insertion sites in the gene trap cell lines, we developed a modified 5' rapid amplification of cDNA ends (5'RACE) protocol. RNA was extracted using TRI reagent (Sigma), reverse transcribed with superscript II (Invitrogen) with a primer specific to the IRES sequence (GSP1), in reverse orientation. The resulting cDNA was then purified using the QIAquick PCR purification kit and subsequently A-tailed using terminal transferase (New England Biolabs). These resulting DNA fragments were then amplified by PCR using a nested IRES-specific biotinylated primer (GSP2) and an anchor primer annealing on the polyA tail. An aliquot of PCR product was loaded on a gel and visible bands were cut, DNA was extracted, purified and sequenced. When no PCR band was visible, the PCR product was purified using streptavidin-coated magnetic beads (Dynabeads, Invitrogen) and digested with BfuCI and NlaIII. These two enzymes do not cut in the vector sequence which is part of the PCR product. Therefore, after digestion, a product containing both a fragment of the vector sequence and of unknown transcript sequence remains attached to the beads. Beads were subsequently washed and digested with ApaI, cutting into the vector sequence and thus releases the RACE fragment. Subsequently, A-tailing with terminal transferase was performed on the supernatant and the resulting product was digested with BsrGI, which cuts the vector sequence upstream of the splice acceptor. This last digestion step allows to cleave PCR products resulting from mRNAs in which the transgene-encoded RNA moiety was not correctly spliced. Finally, the sample was amplified by PCR with the anchor primer and GSP3, loaded on an agarose gel, and DNA was extracted from bands and sequenced. As an alternative approach, the bead-purified PCR product generated with the GSP2 and anchor primers was digested with BfaI and MseI. These enzymes generate compatible ends and cut within the amplified vector sequence, thereby releasing the fragments from the beads. The fragments contained in the supernatant were then circularized by T4 DNA ligase and subsequently redigested with the AclI restriction enzyme. AclI cleaves within the remaining vector sequence, resulting in the linearization of the ligated product, with both ends containing known vector sequences. This fragment was then amplified by PCR with two primers specific for vector sequences present at both ends of the template (GSP4 and GSP5). The resulting product was gel-purified and sequenced. To identify (the) genomic site(s) of integration in the clones transduced with pBmalNLS-luc, 500ng of genomic DNA were digested overnight with MspI. After heat inactivation of the enzyme, digestion products were diluted at 2ng/µl and self-ligated overnight at 16°C using T4 DNA ligase. The next day, the reaction was heat-inactivated and digested with EcoNI, which cuts within the vector sequence to linearize the products containing the insertion site. PCR was then performed using two vector specific primers (BmalGTIDF

and BmalGTIDR), and the PCR product(s) was (were) sequenced.

1.5 Luminescence microscopy

All time-lapse imaging experiments were performed in a LV200 luminoview microscope (Olympus) equipped with an EM-CCD cooled camera (Hamamatsu photonics, EM-CCD C9100-13). One day before starting the recordings, the luminescent cell lines were diluted 1:50 to 1:100 into non-luminescent 3T3 cells at 4×10^5 cells/cm² in 2.2cm Wilco dishes. Next day, the medium was replaced with phenol red-free high glucose DMEM supplemented with 10% FBS, 1% penicilline/streptomycine/glutamine (Gibco), and 275 μ g/ml of luciferin (nanolight). For cells carrying a circadian reporter gene, cells were first synchronized by replacing the culture medium by a mixture of DMEM and 50% of horse serum, and after 2 hours the medium was replaced as described above. The microscope environment was maintained at 37°C and 5% of CO₂, and the cells were monitored for 48 to 72 hours, using a 20-fold magnification objective. Photon counts were acquired as five minute time intervals, and a binning of 4x4 pixels was used to maximize the signal-to-noise ratio. Time lapse acquisitions were analyzed using the Image J software. Each cell was tracked manually using an analysis area of 20 pixels. Background measurements were performed in regions without detectable cellular signals with the same measurement area. For some experiments, two luminescent cell lines were analyzed simultaneously. One cell line was transduced with a lentivector driving the expression of a red fluorescent protein (DsRed) and was checked for homogeneous DsRed expression. Then, this cell line was mixed with another luminescent cell line (non-fluorescent) and non-luminescent 3T3, in ratios of 1:1:98, respectively. Our Lumiview microscope is also equipped with a 565nm light-Emitting Diode (Coolled, www.coolled.com), allowing us to excite and record DsRed. We first identified the red cells to differentiate the two cell lines, and then performed time-lapse acquisitions as described. At the end of the acquisition, another fluorescence picture was taken to check again the identity of each luminescent cell.

1.6 Fluorescence time-lapse microscopy

To follow transcriptional activity by live-imaging, we transfected the H2-MS2 cell line with a plasmid encoding an MS2-YFP fusion protein (kindly provided by Dr. Robert Singer, Albert Einstein College of Medicine, New York) by nucleofection, using an Amaxa nucleofector apparatus (Lonza group Ltd). 24 hours after transfection, cells were imaged at five-minute intervals during 16-20 hours in a Leica AFX6000 wide-field microscope using a 100-fold magnification objective. At each time point, stacks of 13-16 pictures were acquired. Subsequently, images were deconvoluted using the Leica LAS software, and Image J was used to build average projections of each stack.

1.7 Quantification of proteins

To quantify the average number of luciferase molecules per cell, each cell line was split at a density of 4×10^5 cells/cm² into two wells of a 24-well plate. The next day, medium was replaced by white medium (as described above). 40-60 minutes later, the medium was removed and cells were scraped into 75 μ l of ice-cold PBS and transferred to a well of a 96-well plate. Serial dilutions of recombinant luciferase (QuantiLum Recombinant Luciferase, Promega) were performed in a cell suspension of luciferase-negative NIH-3T3 cells. Subsequently, 75 μ l of lysis buffer containing luciferin (Dual Glo assay kit, Promega) was added to each well, and the plate was placed on a horizontal shaker for 15 minutes.

Measurements were performed on a Microwin plate reader in triplicates. A standard curve was derived from the serial dilutions of recombinant luciferase to calibrate the system. We then calculated the total number of luciferase molecules per cell lysate, and using the cell count results we determined the number of luciferase molecules per cell. We then measured the specific activity of the recombinant luciferase, in comparison to the unstable nuclear luciferase we engineered. For this purpose, we first generated a construct to express high levels of bsdF2ANLS-luc. We replaced eGFP from pEGFP-C1 by the bsdF2ANLS-luc coding sequence and transfected 20 μ g of this construct in parallel with 2 μ g of CMV-luc construct expressing the unmodified firefly luciferase in 293T cells. 48 hours after transfection, cells were trypsinized, and one third of the cells were lysed and assayed for luciferase activity in a plate reader at the same time as the standard described above. Whole cell protein extracts were prepared from the remaining two third of cells. For each protein sample, an amount corresponding to a given luminescence signal in the plate reader for each sample was completed with luciferase-free NIH-3T3 protein extracts, to reach a total amount of 200 μ g of protein. These protein extracts were then loaded on a polyacrylamide gel alongside standards prepared using serial dilutions of recombinant luciferase diluted into luciferase-free protein extracts. This gel was then used to perform western blot assays with an anti-luciferase antibody, and band intensities corresponding to luciferase were quantified using the ImageJ software.

1.8 Quantification of mRNA

Quantification of mRNA from gene trap clones

We first measured the average RNA quantity per cell. For this purpose, cells were plated at 4x10⁵ cells/cm² in 2 wells of a 24-well plate. The next day, one replicate well was trypsinized and cells were counted. RNA was extracted, quantified, and the mean amount of RNA per cell was calculated to be about 22pg/cell. To prepare a standard for RNA quantification, the coding sequence for bsdF2ANLS-luc was subcloned into pKS Bluescript downstream of the T3 polymerase promoter. This construct was used as a template for in vitro transcription. This transcript was serially diluted into RNA extracted from wild type NIH-3T3 cells. Subsequently, reverse transcription was performed simultaneously on RNA extracted from the different cell lines and RNA containing different dilutions of in vitro transcribed RNA. Real-time PCR was performed with two gene specific primers and SYBR green reagent (Roche), in a Lightcycler 480 II (Roche). cDNA was amplified using primers for luciferase (QlucF and QlucR), and primers for cyclophilin B (QcycF and QcycR) were used for normalization. The amount of transgene-encoding RNA per μ g of RNA used for reverse transcription was calculated, from which the number of transgene-encoding RNA molecules per cell could be deduced. To control for genomic DNA contamination, real-time PCR was performed on the corresponding amounts of RNA before reverse transcription. Contribution of residual genomic DNA was found to be negligible.

In order to examine whether the wild-type *Ctgf* allele and the *Ctgf* allele tagged with a luciferase cassette were transcribed at similar rates, we quantified wild-type *Ctgf* mRNA and *Ctgf*-luciferase mRNA in the GT:*ctgf* cell line by real-time PCR. To this end, two different pairs of primers complementary to sequences downstream of the luciferase reporter insertion site and a primer pair complementary to luciferase sequences were used for wild-type and luciferase reporter mRNA, respectively. We calculated the PCR efficiencies for all primer pairs, and normalized the CT values accordingly. At equal synthesis rates the steady state concentrations of mRNAs should be proportional to the mRNA half-lives. Hence, we normalized the values found for the wild-type allele using a published mRNA

half-life of 92.7 minutes [10] compared to the somewhat longer half-life (126.5 minutes) of the mRNA issued from the luciferase-tagged allele. The transcription rate of the wild-type *Ctgf* allele was remarkably similar (88%) to that of the tagged *Ctgf* allele.

Quantification of mRNA from Bmal clones and Flp-In clones

The Bmalla and Bmallb clones had one insertion into an active gene in the sense and antisense orientation, respectively. In principle, the Flp-In clones could also be transcribed in the sense or antisense orientation from a spurious promoter. This raised the possibility that another transcript might contain the coding sequence for NLS-luc and therefore be amplified by real-time PCR. To make sure that we were only amplifying sense transcripts, we performed reverse transcription with a gene-specific primer (instead of random hexamers) annealing just before the polyA site to the sense strand of the transgene (RTluc). We then performed real-time PCR with the primers Qluc2F and Qluc2R, amplifying the end of the luciferase coding sequence to make sure that the cDNA we were measuring contained a full length luciferase coding sequence. In parallel, we also performed a reverse transcription reaction using random hexamers in order to quantify cyclophilin cDNA, which was used to normalize the gene-specific reverse transcripts.

1.9 Correlation of protein and mRNA copy number with luminescence

For each cell line analyzed, intensities of at least 30 cells from at least 3 independent time-lapse experiments were measured at the corresponding time point of protein and mRNA extraction. These measurements were averaged and correlated to protein and mRNA copy numbers as outlined in the previous sections (1.6 and 1.7).

1.10 Homogeneity of luciferase expression in cell populations

In principle, we could be overestimating mean cell luminescence calculated from single-cell measurements from the luminoview microscope by missing cells that never express detectable amounts of luciferase over 48 hours. To exclude this possibility, all gene trap cell lines and the transgene cell lines were plated either at high density without non-luminescent cells, or transduced with a lentivector expressing DsRed and mixed with non-luminescent/non-fluorescent NIH-3T3 cells. The large majority of cells (> 95%) within each cell line exhibited a detectable luminescence signal in the luminoview microscope in a 24 hours time-lapse experiment. As this was confirmed in the H1, H2, 1M2C, and 3M2C cell lines, we assumed that this also held true for the remaining CCAAT box cell lines (1M1C, 2M1C, 2M2C and 3M1C) that share the same genomic insertion site.

1.11 Supplementary movies

- Supplementary Movie 1: 48 hours-movie of the GT:Prl2c2 cell line, with one frame every five minutes, using 20-fold magnification in the luminoview microscope.
- Supplementary Movie 2: 72 hours-movie of the Bmalla cell line, with one frame every five minutes, using 20-fold magnification in the luminoview microscope.
- Supplementary Movie 3: 48 hours-movie of the cell line carrying a single high-affinity CCAAT box (H1), with one frame every five minutes, using 20-fold magnification in the luminoview microscope.

2 Computational methods

2.1 The telegraph model of gene expression

Our kinetic model of gene expression considers three basic processes leading to protein expression: activation of the gene, transcription of the DNA sequence into a mRNA molecule and translation of the mRNA into protein. Transcription was assumed to behave according to a random telegraph model representing the active and inactive states of the gene. This model has been widely used to study fluctuations in mRNA and protein copy numbers [11, 12, 13, 14, 15, 16, 17, 18, 19]. The hypotheses of the model are the following:

1. Activation and inactivation of the gene are described by a random telegraph signal with constant rates of switching k_{on} and k_{off} .
2. When the gene is active or 'on', transcription occurs with constant rate k_m . In other words, 'on' defines a state that is competent for the synthesis of new mRNAs. In the inactive state, there is no transcription.
3. Translation of mRNA into protein occurs with constant rate k_p per molecule of mRNA .
4. Degradation of protein and mRNA is described by constant rates γ_p and γ_m per molecule.

Note that while the model is rich enough to analyse the data presented here (and make predictions related to the distributions of 'on' and 'off' times), we should keep it mind that it makes a number of assumptions related to the process of transcription and translation. The rate of transcription in the 'on' state is an effective rate that summarizes a series of processes including the synthesis, degradation and splicing of the precursor mRNA, as well as mRNA export from the nucleus. Thus, transcription rate is an effective rate representing the number of mature mRNA molecules that reach the cytoplasm per unit of time during which the gene is in the 'on' state. One consequence could be that we are underestimating the 'bare' transcription rate, defined as the number of pre-mRNAs made per unit time during activity periods. Because the transcripts generated by the different vectors are short (less than 10 kb), it is likely that few pre-mRNAs will be degraded during elongation and processing into mature mRNA; thus, the modeled transcription rate will be close to the true transcription initiation rate. Similarly the process of translation is highly simplified as we do not model explicitly translation initiation, peptide elongation and termination. In our model the translation rate is proportional to the number of mRNA molecules, which is well supported by recent ribosome profiling experiments [20].

Note that the kinetic rates could explicitly depend on time, e.g. for clones with circadianly regulated promoters. However, the goal of our stochastic model is not to describe explicitly all the biochemical processes that take place during the recordings, but rather to provide a minimal model (with a small number of parameters) to deconvolve the data (see below).

Several properties of the model are important to consider. First, the state of the system at any time t is determined by three integers pmg that represent the number of proteins p , the number of mRNA molecules m , and the state of the gene g . The first two can, in principle, take any integer value, while the gene state is defined to be 0 if the gene is 'off' and 1 if the gene is 'on'. Second, the stochastic model has only six parameters: the six constant rates introduced above. And third, the model is Markovian, meaning that the

state of the system in a particular instant depends only on the state at the previous time point and not on the full history.

The transition probability (or propagator) of finding the system in the state pmg at time t (given that it was in the state $\tilde{p}\tilde{m}\tilde{g}$ at time $t = 0$), is obtained by solving the following master equation:

$$\begin{aligned} \frac{P_{p,m,g}}{dt} = & k_p m P_{p-1,m,g} + \gamma_p(p+1) P_{p+1,m,g} \\ & + k_m g P_{p,m-1,g} + \gamma_m(m+1) P_{p,m+1,g} \\ & - k_p m P_{p,m,g} - \gamma_p p P_{p,m,g} \\ & - k_m g P_{p,m,g} - \gamma_m m P_{p,m,g} \\ & + k_g P_{p,m,\bar{g}} - k_{\bar{g}} P_{p,m,g} \end{aligned} \quad (1)$$

subject to the initial condition and the normalization constraint,

$$\begin{aligned} P_{p,m,g}(0) &= \delta_{pmg|\tilde{p}\tilde{m}\tilde{g}} \\ \sum_{pmg} P_{p,m,g} &= 1 \end{aligned}$$

where the bar on top of the gene state in equation (1) represents the negation operator: $\bar{0} \equiv 1$ and $\bar{1} \equiv 0$.

Although some analytical solutions under different approximations exist [14, 18, 17], solving the full differential equation is not possible. As we are interested in computing the propagator over a time interval of $t = 5$ minutes (corresponding to the sampling rate in our luminescence recordings), we introduced an approximation that assumes constant gene activity and number of mRNA molecules in the interval, which allows us to factorize the probability distribution as:

$$P(pm|g|\tilde{p}\tilde{m}\tilde{g}) \simeq P(p|\tilde{p}\tilde{m})P(m|\tilde{m}\tilde{g})P(g|\tilde{g}). \quad (2)$$

The first factor accounts for the change in the amount of protein given a constant number \tilde{m} of mRNA molecules, the second for the change of the mRNA molecules given a constant gene state \tilde{g} , and the last for the change in the gene state. The advantage of this approximation is that the three factors can be computed analytically, as we will see below.

The transition probability of the gene state is described by the following differential equation:

$$\frac{dP_g}{dt} = k_g P_{\bar{g}} - k_{\bar{g}} P_g$$

which can be integrated directly using the normalization constrain $\sum_g P_g = 1$ and the initial condition $P_g(0) = \delta_{g\bar{g}}$ leading to the solution,

$$P(g|\tilde{g}) = e^{-kt} \delta_{g\bar{g}} + (1 - e^{-kt}) \frac{k_g}{k} \quad (3)$$

where $k = \sum_g k_g$.

The transition probabilities of both the mRNA and protein are described by a birth-death process with constant rate k of particle production and constant rate γ of particle degradation. The differential equation for this process,

$$\frac{dP_n}{dt} = k P_{n-1} + \gamma(n+1) P_{n+1} - k P_n - \gamma n P_n$$

plus the normalization constrain $\sum_n P_n = 1$ and the initial condition $P_n(0) = \delta_{n\tilde{n}}$ can be solved using the generating function $G(z, t) = \sum_n P_n(t)z^n$ [14, 18]. The solution is

$$P(n|\tilde{n}) = \sum_{q=0}^{\tilde{n}} \binom{\tilde{n}}{q} P_o(n-q|k\gamma t)(e^{-\gamma t})^q(1-e^{-\gamma t})^{\tilde{n}-q} \quad (4)$$

where the binomial part quantifies the probability to keep q particles out of the initial \tilde{n} particles, whereas the term

$$P_o(n|k\gamma t) = \frac{\langle n(t) \rangle^n}{n!} e^{-\langle n(t) \rangle}$$

represents the probability of production of n particles, with $\langle n(t) \rangle = \frac{k}{\gamma}(1 - e^{-\gamma t})$ being the mean number.

Finally, setting the rates $k = mk_p$ and $\gamma = \gamma_p$ we obtain the solution for the transition probability between protein states and, similarly, we obtain the equivalent expression for mRNA states by substituting the rates $k = gk_m$ and $\gamma = \gamma_m$.

2.2 Protein and mRNA half-lives

To estimate the half-life of the nuclear luciferase, we performed time-lapse luminescence recordings of different gene trap cell lines in a lumicycle apparatus (Actimetrics). After stabilization of the signal, cells were treated with 25µg/ml of cycloheximide, and the decay of luminescence signal was recorded. To infer the half-life of the transgenic mRNA in different cell lines, we performed time-lapse luminescence recordings of cell populations as described above, and treated cells with 5µg/ml of actinomycin D.

Since the experiments were carried out at the population level we used a deterministic model to describe the processes of transcription and translation. The concentrations of protein and mRNA were determined by the following system of differential equations:

$$\begin{aligned} \dot{p}(t) &= k_p m(t) - \gamma_p p(t) \\ \dot{m}(t) &= k_{\text{eff}} - \gamma_m m(t) \end{aligned} \quad (5)$$

where the effective rate of transcription $k_{\text{eff}} = k_m(k_1/(k_0+k_1))$ is a function of the switching rates of the gene.

When translation is blocked ($k_p = 0$) the solution is straightforward since equation (5) becomes a pure decay process with the solution $p(t) = p_0 e^{-\gamma_p t}$. An exponential function was fitted to the experimental data to obtain the degradation rate of the reporter γ_p (see Fig. S5A and S5C). When transcription is inhibited, ($k_{\text{eff}} = 0$) the solution of the equation (5) is

$$p(t) = \left(p_0 - \frac{k}{\gamma_p - \gamma_m} \right) e^{-\gamma_p t} + \left(\frac{k}{\gamma_p - \gamma_m} \right) e^{-\gamma_m t} \quad (6)$$

where $k = m_0 k_p$. We fitted this function to the experimental data of each clone assuming a Gaussian noise model. We used a Monte Carlo sampling method [21] to estimate the degradation rate of the mRNA γ_m knowing the degradation rate of the protein γ_p (See Fig. S5B and S5D).

We wished to examine the reliability of the method used to estimate mRNA half-lives (by inference from bioluminescence traces, Fig. S5B) by a more direct method. To this end, we collected RNA every 30 minutes after the addition of actinomycin D to H2 cells, converted it into cDNA, and performed real-time PCR experiments with primers against luciferase mRNA and 18S ribosomal RNA (for normalization). As indicated in Figure S5F, the half-life thus determined (65.4 +/- 6 min), is very close to that inferred from luminescence traces (54.9 +/- 16 minutes).

2.3 Calibration and noise model

To connect the stochastic model introduced above with the experimental data, it is imperative to find the relationship between the amount of emitted light and protein copy numbers. As described above, we quantified both the amount of protein and the corresponding emitted light for five clones (Fig. S6A). We then fitted a linear function $s = \alpha p$ to the calibration data and obtained a proportionality constant $\alpha = 7.63$ ($r^2 = 0.8$) between measured gray levels (s) and protein copies (p). We independently confirmed the linear behavior of the camera down to the lowest intensities by following individual cells after translation inhibition by cycloheximide (Fig. S6A). As it has been routinely performed and empirically supported [22], we modeled the noise by a combination of two sources: the background fluctuations plus a Poissonian fluctuation that is proportional to the amount of light: $\sigma^2 = \sigma_b^2 + \beta p$. The probability of measuring the signal s given p proteins is then,

$$P_e(s|p) = \frac{1}{\sqrt{2\pi(\sigma_b^2 + \beta p)}} e^{-\frac{(s-\alpha p)^2}{2(\sigma_b^2 + \beta p)}} \quad (7)$$

To estimate σ_b , we measured the standard deviation of the gray levels in a region of identical measurement size but devoid of cells. This was done independently for each movie. Finally, we fit β to best explain the fluctuations observed in single cell recordings where translation was blocked with cycloheximide (Fig. S6A). The estimated parameters were thus $\alpha = 7.63$, $\beta = 54$ and $\sigma_b \simeq 100$.

2.4 Parameter estimation and deconvolution

Having defined the emission and transition probabilities (3, 4 and 7), it is now possible to write the probability of observing a time series of light measurements $D = \{s_1, \dots, s_T\}$ given a set of parameters $\Theta = \{k_p, \gamma_p, k_m, \gamma_m, k_0, k_1\}$ as the probability of measuring the data given a particular state path $\Lambda = \{(p_1 m_1 g_1), \dots, (p_T m_T g_T)\}$, times the probability of that particular state path, and then summing over all possible state paths, i.e:

$$P(D|\Theta) = \sum_{\Lambda} P(D|\Lambda\Theta)P(\Lambda|\Theta) \quad (8)$$

To keep the notation simple we took only one time series of D measurements but the formalism can be easily extended to multiple independent time series $\{D_n\}$ (i.e. multiple cells in our case) by taking the product in (8) over the different time series. We then rewrite the likelihood of the data $P(D|\Theta)$ applying the Markovian property of the stochastic model as,

$$P(D|\Theta) = \sum_{\Lambda} \prod_i P_e(s_i|p_i)P_t(p_i|m_i g_i|p_{i-1}m_{i-1}g_{i-1})$$

which, introducing the approximation for the transition probabilities (2), becomes:

$$P(D|\Theta) = \sum_{\Lambda} \prod_i P_e(s_i|p_i)P_t(p_i|p_{i-1}m_{i-1})P_t(m_i|m_{i-1}g_{i-1})P_t(g_i|g_{i-1}) \quad (9)$$

where we have used notation of HMM parlance, i.e. we used the index e for 'emission' and t for 'transition' [23, 24].

The analysis of the experimental data led to two inference problems which can be posed as follows:

1. Given a time series of light measurements D , what is the set of parameters Θ that maximizes the likelihood of the data $P(D|\Theta)$.
2. Given a time series of light measurements D and a set of parameters Θ , what is the state path Λ , i.e. the time series of protein, mRNA and gene states, that maximizes the probability $P(D|\Lambda\Theta)P(\Lambda|\Theta)$.

Solving the first problem is required to compute the sum in equation (9). However, note that the number of possible paths grows exponentially with the total time of recording, meaning that the sum becomes infeasible even for short recordings. To overcome this problem we applied a standard technique used in Hidden Markov Models [23, 24]. The so-called forward score $F_{i,pmg}$ is defined as the probability of the data up to time i and being, at that time, in the state pmg . This quantity can be computed recursively using the following expression:

$$F_{i,pmg} = \sum_{\tilde{p}\tilde{m}\tilde{g}} P_e(s_i|p)P_t(p|\tilde{p}\tilde{m})P_t(m|\tilde{m}\tilde{g})P_t(g|\tilde{g})F_{i-1,\tilde{p}\tilde{m}\tilde{g}} \quad (10)$$

and the (forward) probability of the data is then,

$$P(D|\Theta) = \sum_{pmg} F_{T,pmg}$$

where T is the last time point of the time series.

It will be convenient to introduce also a backward score $B_{i,pmg}$, which is defined as the probability of the data from time i to the end of the recording (given that at time i the state of the system was pmg). The recursive expression is then,

$$B_{i,pmg} = \sum_{\tilde{p}\tilde{m}\tilde{g}} P_e(s_{i+1}|\tilde{p})P_t(\tilde{p}|pm)P_t(\tilde{m}|ng)P_t(\tilde{g}|g)B_{i+1,\tilde{p}\tilde{m}\tilde{g}} \quad (11)$$

To find the set of parameters Θ^* that maximizes the likelihood of the data (9), we used the Baum-Welch algorithm [23, 24] which is an Expectation-Maximization (EM) iterative algorithm. Given a set of parameters $\Theta^{(n)}$ at the n th iteration, the expectation step (E) of the algorithm consists of obtaining the expected number of times $\langle N_{pmg,\tilde{p}\tilde{m}\tilde{g}} \rangle$ that the states pmg and $\tilde{p}\tilde{m}\tilde{g}$ occur next to each other. These numbers can be computed using the forward and backward scores. The expression reads,

$$\langle N_{pmg,\tilde{p}\tilde{m}\tilde{g}} \rangle = \frac{\sum_i F_{i,pmg} P_e(s_{i+1}|\tilde{p})P_t(\tilde{p}|pm)P_t(\tilde{m}|mg)P_t(\tilde{g}|g)B_{i+1,\tilde{p}\tilde{m}\tilde{g}}}{\sum_{pmg} F_{T,pmg}} \quad (12)$$

Then, in the maximization step (M) the parameters were re-estimated maximizing the likelihood of observing these numbers, i.e. finding Θ^* that maximizes,

$$P(\{\langle N_{pmg,\tilde{p}\tilde{m}\tilde{g}} \rangle\}|\Theta) = \prod_{pmg,\tilde{p}\tilde{m}\tilde{g}} [P_t(\tilde{p}|pm)P_t(\tilde{m}|mg)P_t(\tilde{g}|g)Q(pm)]^{\langle N_{pmg,\tilde{p}\tilde{m}\tilde{g}} \rangle}$$

where the prior probability $Q(pm)$ was taken as constant.

Last, to close the iterative loop, the new parameters $\Theta^{(n+1)} = \Theta^*$ were used to recompute the forward and backward score as well as the counts (12). The iterative process was stopped when the relative difference between the old likelihood $P(D|\Theta^{(n)})$ and the new one $P(D|\Theta^{(n+1)})$ was less than 5%. The search was seeded with values for the parameters that were consistent with the measured protein and mRNA half lives, as well as their copy

numbers (Fig. S7). Note that EM algorithms guarantee convergence to local minima; the global parameter landscape may be further explored using Monte Carlo sampling methods. The latter would also inform on how the identifiability of model parameters depends on the amount of data and regime in parameter space.

We estimated the maximum likelihood parameters from each movie independently and, for each clone, we reported the mean over all the movies of that clone as well as the standard deviation. In all parameter estimations in the manuscript, we have held the mRNA and protein half-lives fixed as these were independently measured experimentally.

To solve the second inference problem, i.e. to obtain the path of states Λ^* that maximize the likelihood of the data we use a Viterbi algorithm [23, 24]. The maximum likelihood $M_{i,pmg}$ that is possible to achieve up to time i and being at this time in state pmg is obtained, recursively, with the following expression:

$$M_{i,pmg} = \max_{\tilde{p}\tilde{m}\tilde{g}} P_e(s_i|p)P_t(p|\tilde{p}\tilde{m})P_t(m|\tilde{m}\tilde{g})P_t(g|\tilde{g})M_{i-1,\tilde{p}\tilde{m}\tilde{g}}$$

The maximum path is recovered by traversing backwards the matrix M . Using the parameters that were inferred previously we obtained the maximum paths for all analyzed cells. These inferred traces were used to compute the burst size distributions as well as the 'on' and 'off' times distributions (see below).

2.5 On-times and off-times histograms

All the histograms shown in the main text and in the supplementary material were obtained using the maximum path solutions. The on(off)-times histograms were directly calculated from the observed frequency of number of consecutive on(off)-states that occur across all the cell traces of a given clone.

Exponential distributions $P(t|\tau) \propto e^{-t/\tau}$, with a characteristic time scale τ , were fitted to the on-times histograms. In contrast, the off-times frequencies were modeled as the combination of a two-step reaction process with a distribution of times that reads,

$$P(t|\tau_1\tau_2) = \int_0^t P(t_1|\tau_1)P(t-t_1|\tau_2)dt_1 = \frac{(e^{-t/\tau_2} - e^{-t/\tau_1})}{\tau_2 - \tau_1}$$

which is the result of the convolution of two exponential processes of characteristic time scales τ_1 and τ_2 . This means that, after an active gene period, a refractory period of mean time τ_1 is followed by an *inactive* period of mean time τ_2 before the gene is switched on again. The mean total off-time is then $\bar{\tau} = \tau_1 + \tau_2$. In contrast to single exponential distributions, these type of distributions show a maximum at a time different than zero. The values of the refractory time τ_1 ranged from 15 to 40 minutes and those of the *inactive* time τ_2 from 100 to 300 minutes.

The observed refractory times suggests that we may extend our gene model to a three state model inserting a refractory state in between the current 'on' and 'off' states. This extended model can in principle be treated in the same inference framework. In Fig. SS20, we show the result obtained for the three-state model for the Bmal1a clone. In this case, the model parameters were calibrated by sampling the forward probability using the Metropolis-Hastings algorithm.

2.6 Validation of the algorithm and performance of the inference

We applied our algorithm to simulated data in order to examine its ability of inferring the transcriptional kinetic parameters. To this end, we produced artificial luminescence

recordings by simulating temporal traces of protein, mRNA and gene activity using the Gillespie algorithm [25]. To evaluate how the performance depends on the parameter values we built 25 different sets of parameters (our "in silico" cell lines) and for each set we simulated 25 cells recorded for two days (similar to the real data). The distribution of parameter values is as follows: 6 clones with different k_{on} and the same k_{off} and k_m (Fig. S17A); 11 clones with different k_{off} and the same k_{on} and k_m (Fig. S17B); 8 simulated clones with different transcription rates and the same k_{on} and k_{off} (Fig. S17C). Comparing the inferred parameters to the real ones showed a quite satisfactory performance given the complexity of the inference (Fig. S17). It appears that there is a tendency to underestimate the transcription rate k_m as well as the off-switching rate k_{off} when the real k_{off} become of the order of the inverse sampling rate (Fig. S17B). This is a consequence of the approximation used (2) in which the gene is considered to be on for the entire five minute interval, lowering the transcription rate to compensate for the extra time the gene is kept on. Therefore, both the inferred transcription rate and the off-switching rate should be considered as lower bounds. Importantly, the mean burst size $b = k_m/k_{off}$ was always recovered accurately, and similarly for the on-switching rate k_{on} .

In Fig. SS18, we have assessed how imprecision in the measurement of the mRNA half-live (cf. Fig. S5B) affects the estimation of the other model parameters. For this we simulated data with a given mRNA half-live and estimated parameters by holding fixed a modified value with a given relative error. We find that the relative errors in the predicted k_{on} and k_{off} , were less or comparable than the half-live error, while the error on k_m did not vary much with the error on the mRNA half-life. Thus, as we would expect, errors in measured parameters do affect estimation, but such errors are certainly not amplified and thus accurate measurements will lead to similarly accurate parameter estimation.

To show that the maxima in the inferred distribution of off-times are a property of the data and not an artifact related to the reconstruction method, we produced artificial data using the same parameters as the ones inferred from the clones Bmal1a, GT:Prl2C2, GT:Glutaminase and H2. For each set of parameters, we simulated 1000 cells recorded for two days. We obtained the maximum path solution of each temporal trace and computed the on-times and off-times distributions. In Figure S19 we compare the distributions obtained from real and artificial data. The on-times distributions are similar whereas there are clear differences between the off-times distributions. This indicates that the characteristic shapes of real off-times distributions are not an artifact of the computational method.

2.7 Comparison of the inference using a three vs. a two state model

The peaked shape of the predicted off-time distributions indicates that the gene dynamics in the off state reflects at least two exponential steps, so that in sum the minimal gene model would be a three state switching process. We investigated the consequences of estimating parameters using a two state model, as we have done throughout, when the underlying model is instead a three-state model. We generated data with same parameters as in Fig. SS20 except that the 'off' state was split into two states of equal lengths, 'off-1' and 'off-2', with effective on rate $k_{on} = 1/(k_{on,1}^{-1} + k_{on,2}^{-1})$. The predicted parameters are shown in Fig. SS20A and show good agreement with the real ones: we find a slight overestimation for k_{on} and a slightly underestimation of k_{off} , both in the range of $\lesssim 20\%$. Meanwhile the predicted k_m is very accurate and the burst size is slightly underestimated as consequence of k_{off} .

To verify that prediction for the off time distribution are insensitive to the assumed model, we analyzed the data for the Bmal1a clone using an extension of our algorithm that

incorporates a three-state model for the gene activity. In this case, since the expectation maximization (EM) is much more involved we sampled the posterior distribution on the model parameters using the Metropolis-Hastings method. For this, the new three-state propagator describing the dynamics of the gene needed in Eq. 2 is calculated numerically using matrix exponentials. In Fig. SS20B we show that parameter values inferred for the two and three state models show consistent values. Most importantly, Fig. SS20C shows that the predicted ‘on’ and ‘off’ time distributions are very similar for the two models, indicating that the data is informative enough and that the peaked off distribution are a robust feature of the data.

2.8 Wavelet analysis

Continuous wavelet transforms [26] used Morlet wavelets and were implemented using the wmtsa package in R (<http://cran.r-project.org/>). Wavelet transforms for each individual cell were performed considering the first two days of recording (zero padding was used in case of shorter traces). To represent clone-specific profiles, the wavelet spectra for all cells of one clone were averaged (Fig. S2). The power spectra (global wavelet spectrum) represent the time averaged wavelet profiles [27].

2.9 Burst frequency and size during a circadian cycle

To obtain profiles of burst frequency and mean burst size during a circadian cycle (Fig. S11) we proceeded as follows: first, we smoothed the single cell luminescence traces to identify peaks of circadian activity. We then aligned all peaks taking the maximum luminescence activity as a reference plus 12 hours both before and after the maximum. We computed the burst frequency and the mean burst size within a window of four hours which was slid within the circadian cycle. The burst frequency is defined as the number of on-periods and the burst size is approximated as the net change of mRNA molecules taking place during the on-periods. Finally, at each time point we averaged the burst frequency and the mean burst size across all maxima in our data.

3 Supplementary tables and figures

Oligo name	Oligo sequence
GSP1	TACGCTTGAGGAGAGCCATT
GSP2	AGGAAC TGCTTCCTCACGA
GSP3	CCAAAAGACGGCAATATGGT
GSP4	GAGAGGGCGGTCTTATT
GSP5	TGGAATAAGGCCGGTGTG
Anchor	GGTTGTGAGCTCTTAGATGGTTTTTTTTTTTTTT
QlucF	GCAATT CACGAATCCCAACT
QlucR	AGGTGCTTCTCGATCTGCAT
QcycF	GGAGATGGCACAGGAGGAA
QcycR	GCCC GTAGTGCTTCAGCTT
RTluc	GCCC GTAGTGCTTCAGCTT
Qluc2F	GTC CACGAACACAACACCAC
Qluc2R	AACCATGACCGAGAAGGAGA
Qlucpgl3F	TCAAAGAGGCGAAGTGTGTG
Qlucpgl3R	CGCTTCCGGATTGTTACAT
BmalGTIDF	TCAAAATTTATCGATCACGAGAC
BmalGTIDR	TTCCC ATCGCGATCTAACCT
miniCMVtop	CGCGTAAGATCTGGCCTCGGGGGCCAAGCTTAGACACTAGAGGGTATATAATGGAAGCTCGACTTCCAGT
miniCMVbottom	CGCGACTGGAAGTCGAGCTTCCATTATACCCCTCTAGTGTCTAAGCTTGGCCCGAGGCCAGATCTTA
High affinity top	CGCGTAGCGCGTACAAGAGATTAACCAATCACGTACGGTCTA
High affinity bottom	CGCGTAGACCGTACGTGATTGGTTAATCTCTTGACCGCTA
Low affinity top	CGCGTAGCGCGTACAACCGTACAACCAATTAAAGTCACTCTA
Low affinity bottom	CGCGTAGAGTGACTTAATTGGTTACGGTTACGGCTA
1Mtop	CGCGTAGCGCGTACAAGAGATAACCAATCACGTACGGTCTA
1Mbottom	CGCGTAGACCGTACGTGATTGGTTATCTCTTGACCGCTA
2Mtop	CGCGTAGCGCGTACAAGAGATAACCAATCACGTACGGTCTA
2Mbottom	CGCGTAGACCGTACGTGATTGGTTATCTCTTGACCGCTA
3Mtop	CGCGTAGCGCGTACAAGAGATAACCCAAATCACGTACGGTCTA
3Mbottom	CGCGTAGACCGTACGTGATTGGGTATCTCTTGACCGCTA
18SF	AGTCCCTGCCCTTGACACA
18SR	GATCCGAGGGCCTCACTAAC
ctgf11F	CAGTGGGAATTGTGACCTGA
ctgf1R	GCTTTATCACCTGCACAGCA
ctgf2F	TGGGAATTGTGACCTGAGTG
ctgf2R	TCACCTGCACAGCATTGTT

Table S1: Oligo sequences (5' to 3')

Cell line	mRNA Accession number	Gene name	Insertion site
GT:Nckap1	NM_016965	Mus musculus NCK-associated protein 1 (Nckap1)	after exon 17
GT:glutaminase	NM_001081081	Mus musculus glutaminase	after exon 7
GT:Serpine1	NM_025814.2	Serpine 1/PAI-1 gene	in exon 6
GT:Prl2c2	NM_031191	Mus musculus prolactin family 2, subfamily c, member 2 (Prl2c2),	after exon 4
GT:Sh3kbp1	NM_021389.4	Mus musculus SH3-domain kinase binding protein 1 (Sh3kbp1),	in exon 7
GT:Plectin1	NM_011117	Mus musculus plectin 1 (Plec1), transcript variant 1, mRNA.	after exon 1
GT:Hmga2	NM_010441.2	Mus musculus high mobility group AT-hook 2 (Hmga2), mRNA	after exon 3
GT:Ctgf	NM_010217	Mus musculus connective tissue growth factor (Ctgf), mRNA	in exon 5
Bmal1a	NM_148945	ribosomal protein S6 kinase polypeptide 3	after exon 15, sense
Bmal1b	NM_173189	Microcephalin	in exon 6, antisense
DBP-NLS-luc	NM_016974.3	D site of albumin promoter (albumin D-box) binding protein	FRT site of NIH-3T3 Flp-In
Per2luc heterozygous	NM_011066.3	Period homolog 2	Exon 23
Per2luc homozygous	NM_011066.3	Period homolog 2	Exon 23
H1			FRT site of NIH-3T3 Flp-In
H2			FRT site of NIH-3T3 Flp-In
1M1C			FRT site of NIH-3T3 Flp-In
1M2C			FRT site of NIH-3T3 Flp-In
2M1C			FRT site of NIH-3T3 Flp-In
2M2C			FRT site of NIH-3T3 Flp-In
3M1C			FRT site of NIH-3T3 Flp-In
3M2C			FRT site of NIH-3T3 Flp-In
H2-MS2			FRT site of NIH-3T3 Flp-In

Table S2: mRNA accession numbers, gene names and insertion sites

Clone ID	Number of mRNA per cell
GT:Nckap1	64.5 ± 23.2
GT:Glutaminase	30.9 ± 15.13
GT:Serpine1	23.0 ± 5.3
GT:Prl2c2	31.3 ± 13.1
GT:Sh3kbp1	38.9 ± 3.0
GT:Plectin1	17.7 ± 3.6
GT:Hmga2	9.8 ± 1.0
GT:Ctgf	312.7 ± 139.2
Bmal1a	2.5 ± 0.5
Bmal1b	0.5 ± 0.4
DBP	12.9 ± 7.8
H1	63.8 ± 24.1
H2	90.7 ± 22.7
1M2C	63.6 ± 28.3
Per2luc (homozygous)	9.8 ± 1.4

Table S3: Number of NLSluc mRNA copies per cell

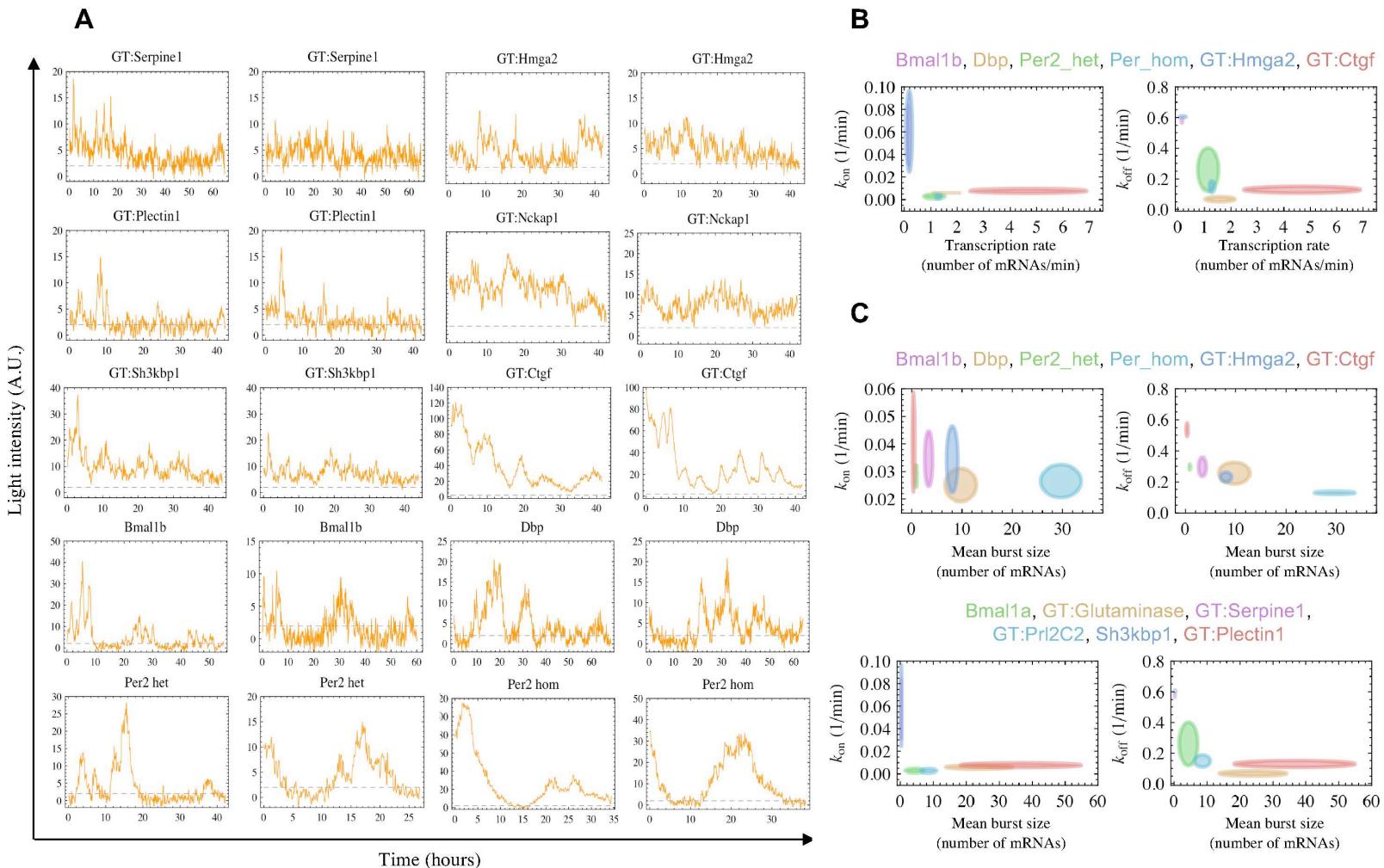
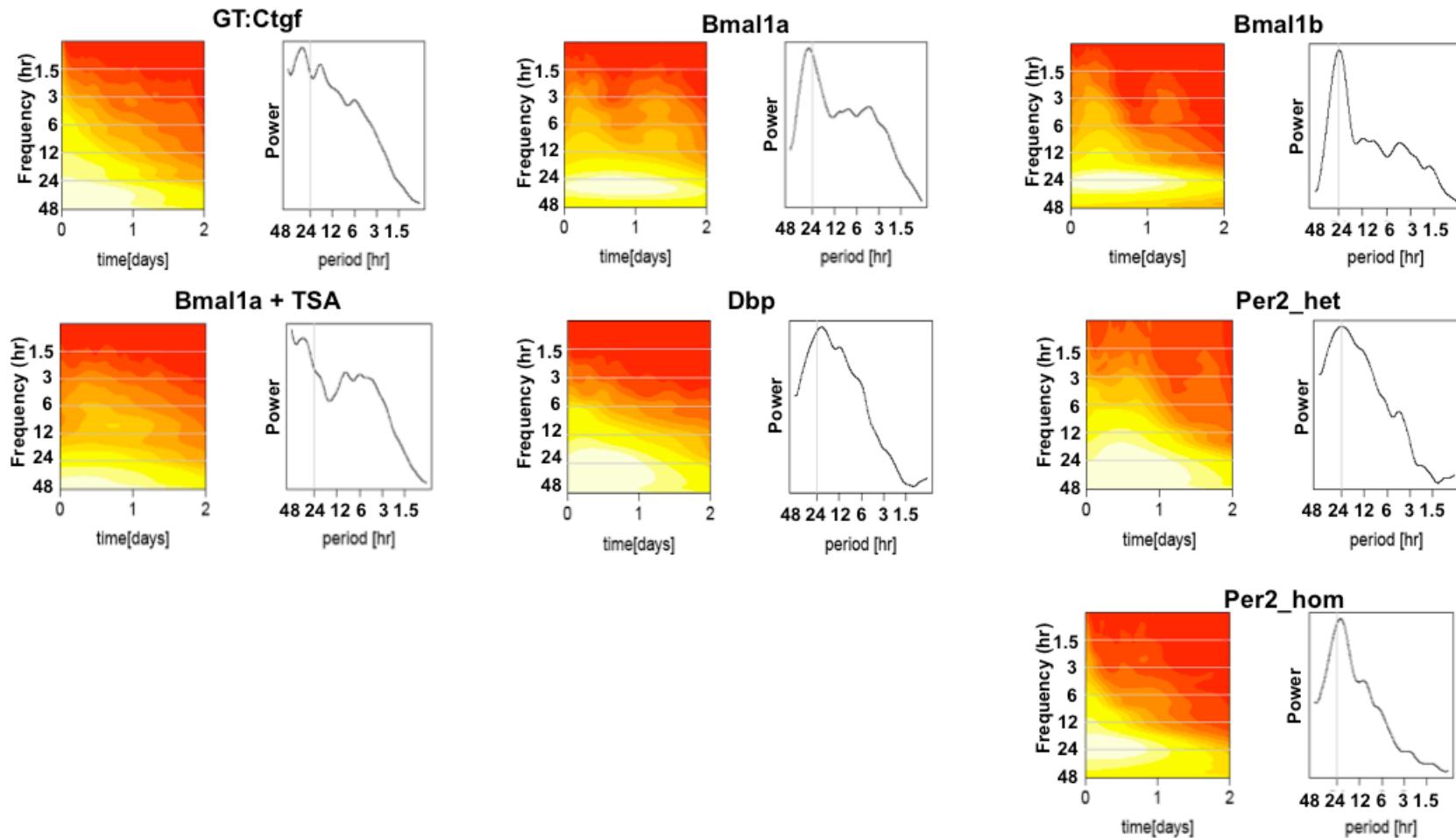
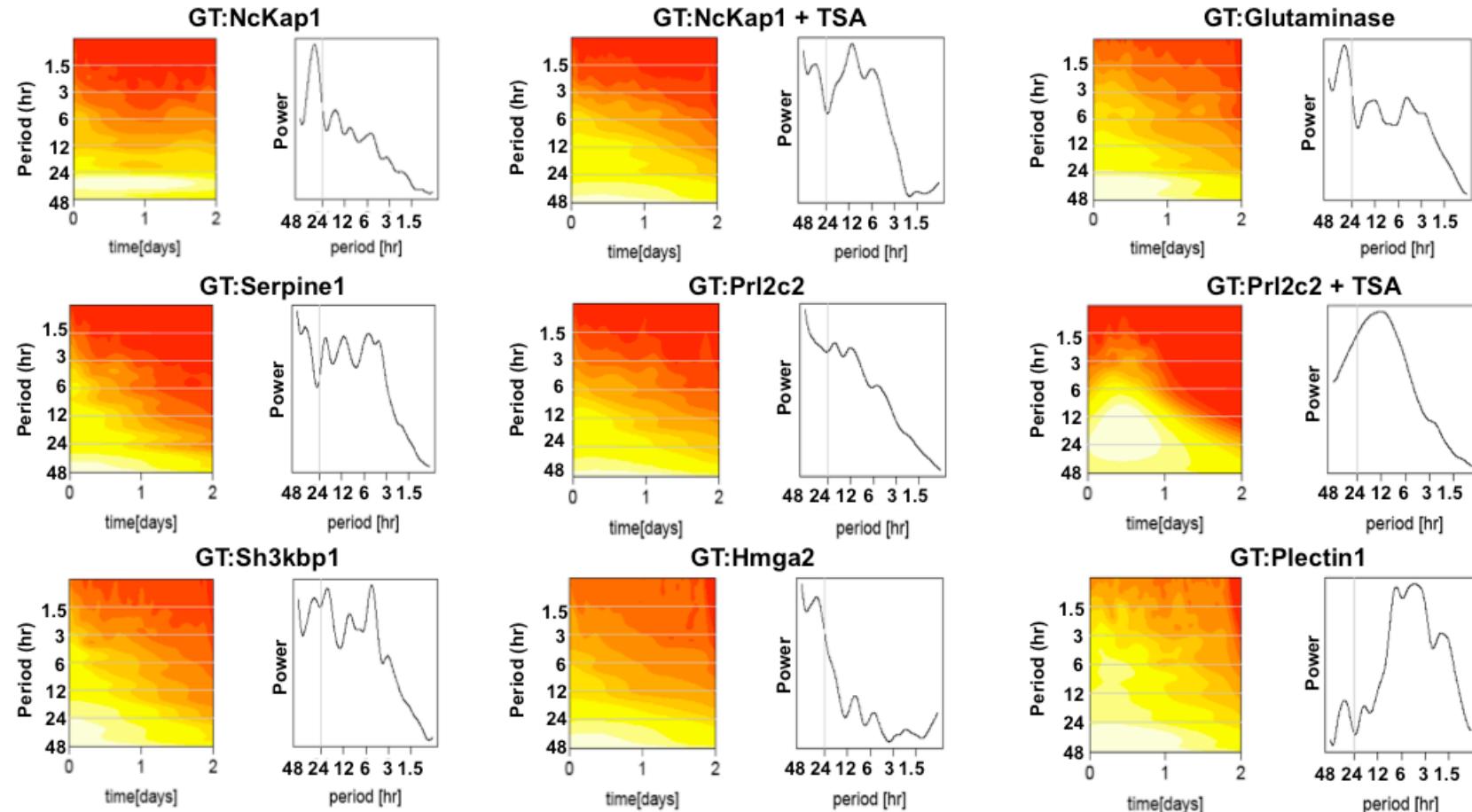


Figure S1: (A) Individual cell traces of GT:Serpine1, GT:Hmga2, GT:Plectin1, GT:Nckap1, GT:Sh3kbp1, GT:Ctgf, Bmal1b, DBP, Per2_heterozygous, and Per2_homozygous cell lines. (B, C) Relationship between transcription rate k_m (B) and burst size (C) with the switching rates k_{on} (left) and k_{off} (right). Ellipses represent means \pm two standard deviations. A.U.: arbitrary units. Color legends are shown on top of each diagram.





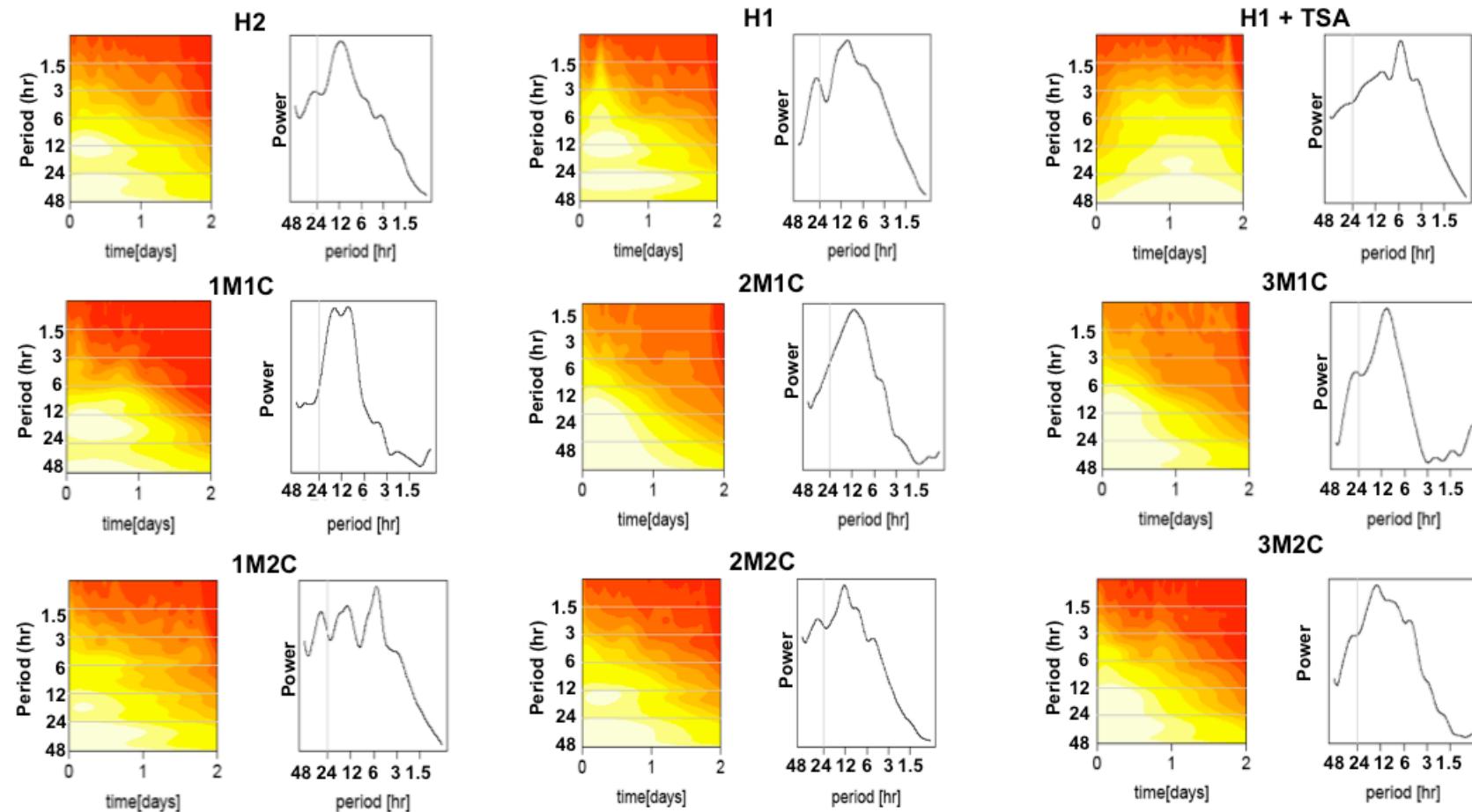


Figure S2: Wavelet spectra of the different cell lines. Morlet wavelets were used, and all cells from one clone were averaged (cf. supplemental methods). Left panels: heat map representations in function of time, logarithmically spaced periods are indicated. Right panels: power spectra (global wavelet spectra) represent the square root of the time-averaged spectra in the left panels.

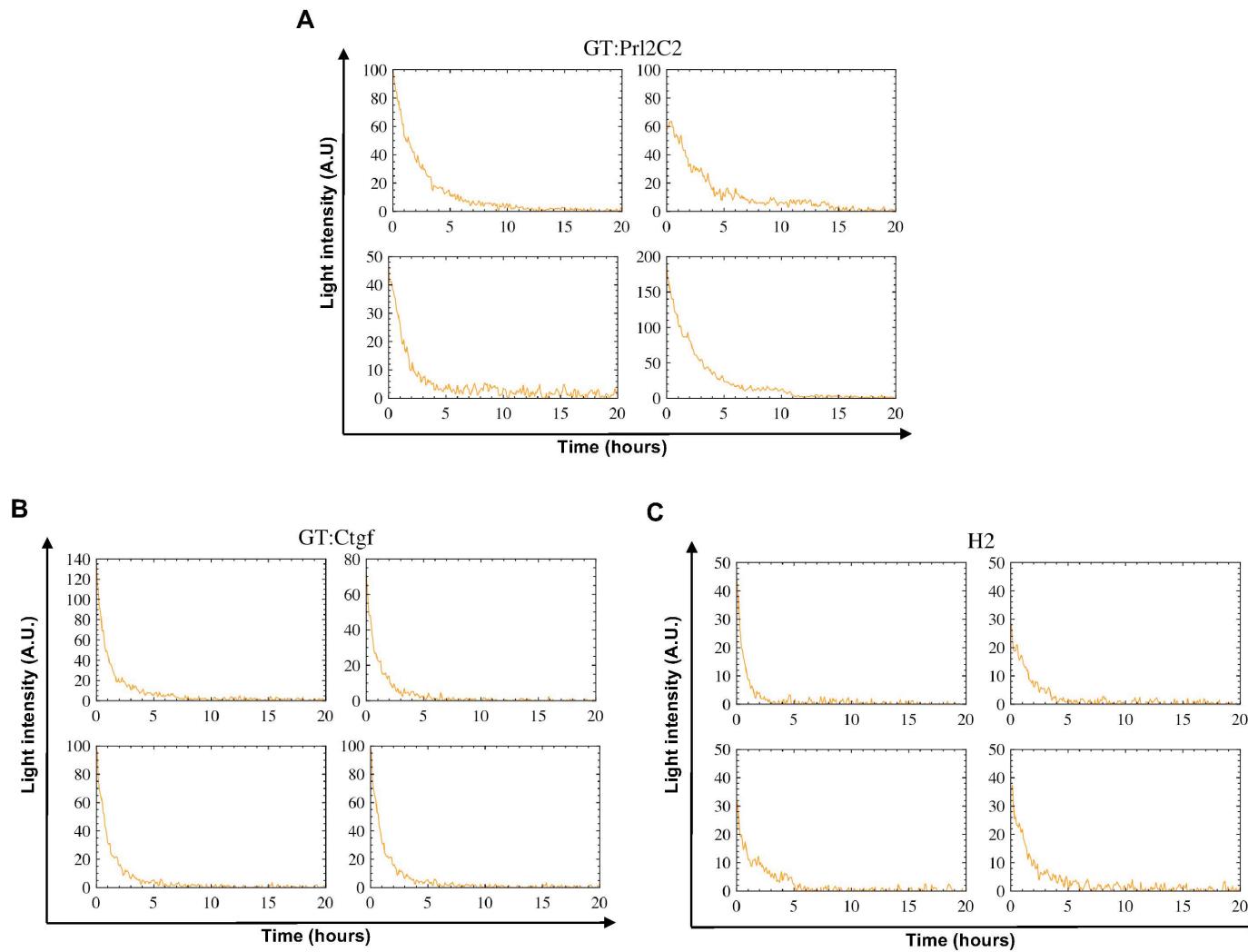


Figure S3: The GT:Prl2C2 (A), GT:Ctgf (B) and H2 (C) cell lines were treated with Actinomycin D and luminescence signal from individual cells were tracked for 20 hours.

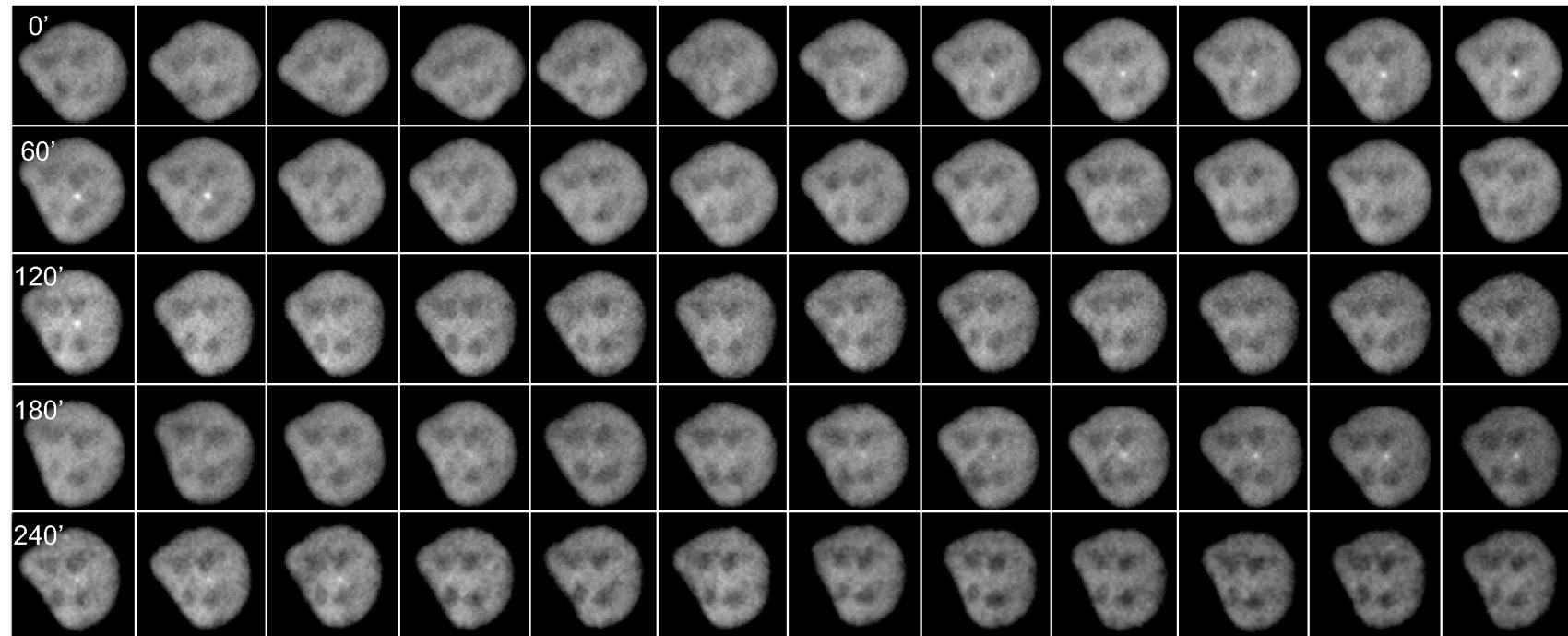


Figure S4: Example of time-lapse imaging of YFP-tagged MS2 in the H2-MS2 cell line. Image stacks were acquired every five minutes, deconvoluted and projected on a single plane. Numbers on the first column indicate time points (minutes) at which images were acquired. Note the presence of three episodes of appearance and disappearance of a spot in the center of the nucleus.

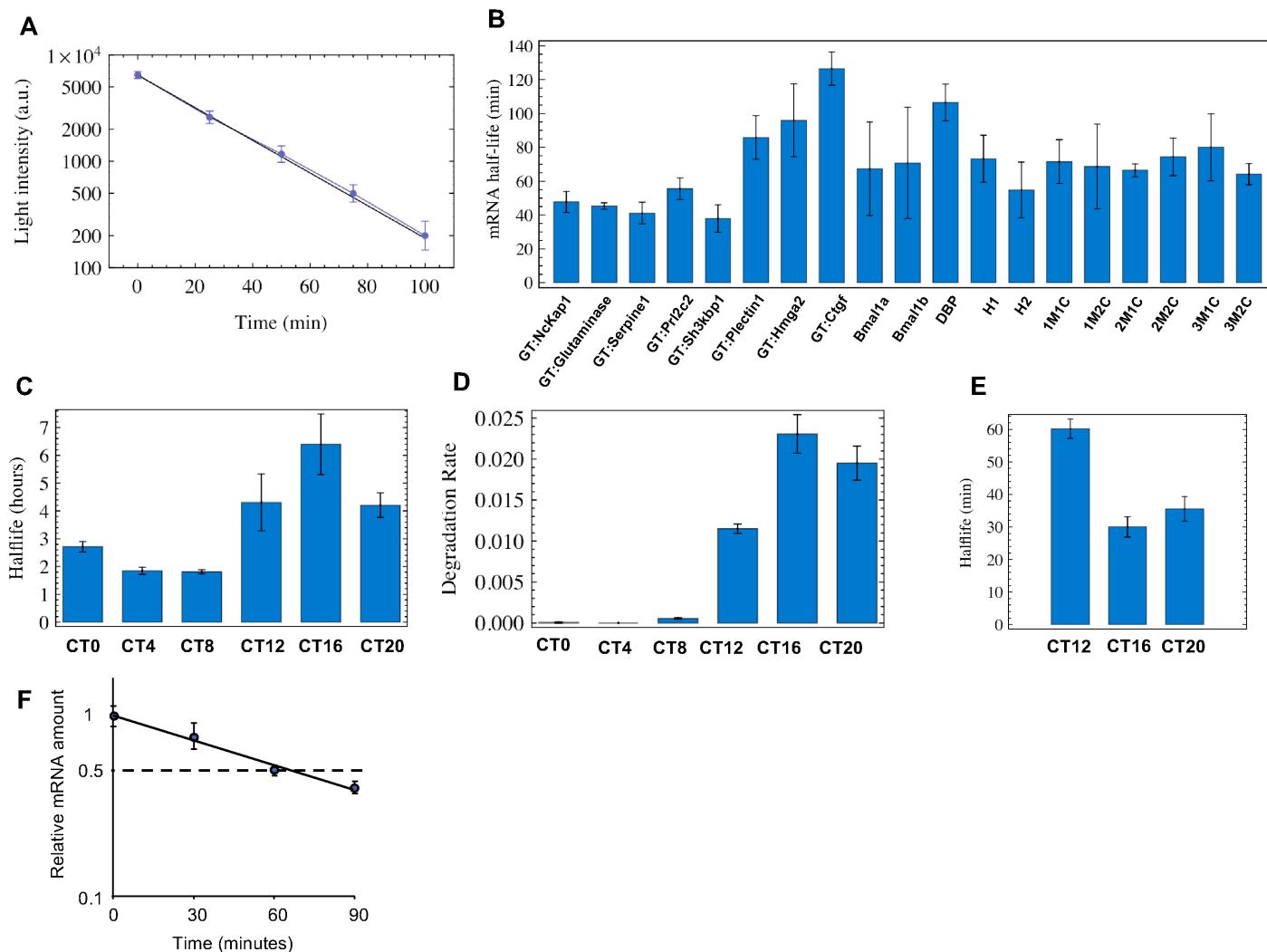


Figure S5: Luminescence decay was measured in the whole cell population after addition of cycloheximide or actinomycin D. **(A)** mean of luminescence signal decay in three different clones after cycloheximide treatment. **(B)** half-lives of mRNAs, calculated from luminescence signal decay after actinomycin D treatment. **(C-E)** protein (C) and mRNA (D, E) stability of Per2luc around the clock after serum shock. **(F)** real-time PCR measurement of bsdF2ANLS-luc decay after actinomycin D treatment of the H2 cell line. Errors bars: standard error of the mean.

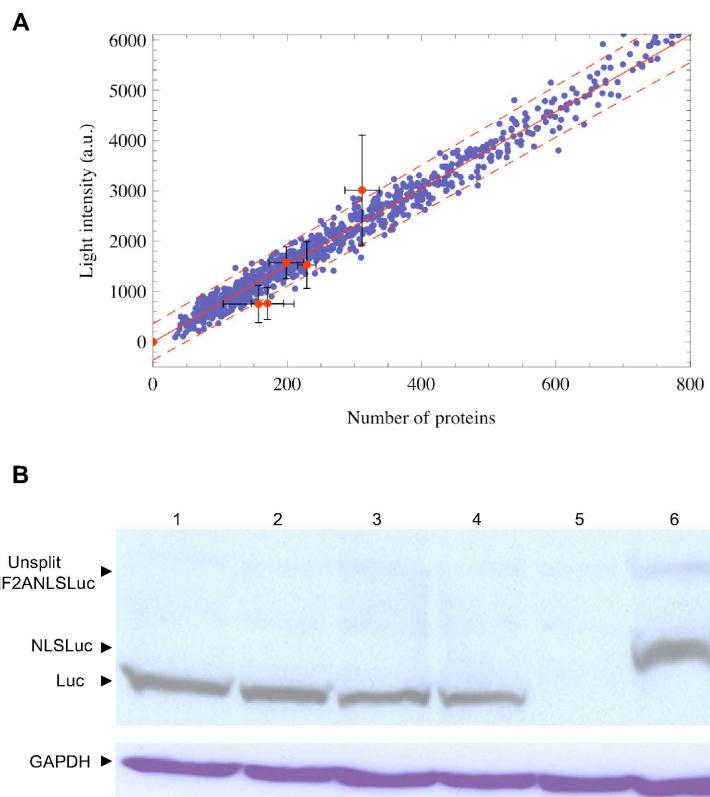


Figure S6: (A) linear relationship between average gray levels and average number of luciferase molecules per cell for five different clones (GTNcKap1, GTglutaminase, GTSerpine1, GTPrl2c2 and GTSh3kbp1). To verify the linearity of photon counting of the camera, we blocked translation of the cell lines described above (blue dots). The calibration used quantification of measured intensity against number of luciferase molecules (red dots, see Methods). (B) Western blotting for luciferase. Lane 1-5: wt NIH-3T3 extract with 30ng (1), 20ng (2), 15ng (3), 10ng (4) and 0ng (5) of recombinant luciferase. Lane 6: Extract from 293T cells transfected with a vector expressing NLSLuc, mixed with extract from wt NIH-3T3 cells. The amount of 293T cell extract loaded in lane 6 is equivalent to 20ng of recombinant luciferase for the same specific activity. The ratio of the specific activity of NLSLuc compared to recombinant luciferase was equal to 1.002 ± 0.187 .

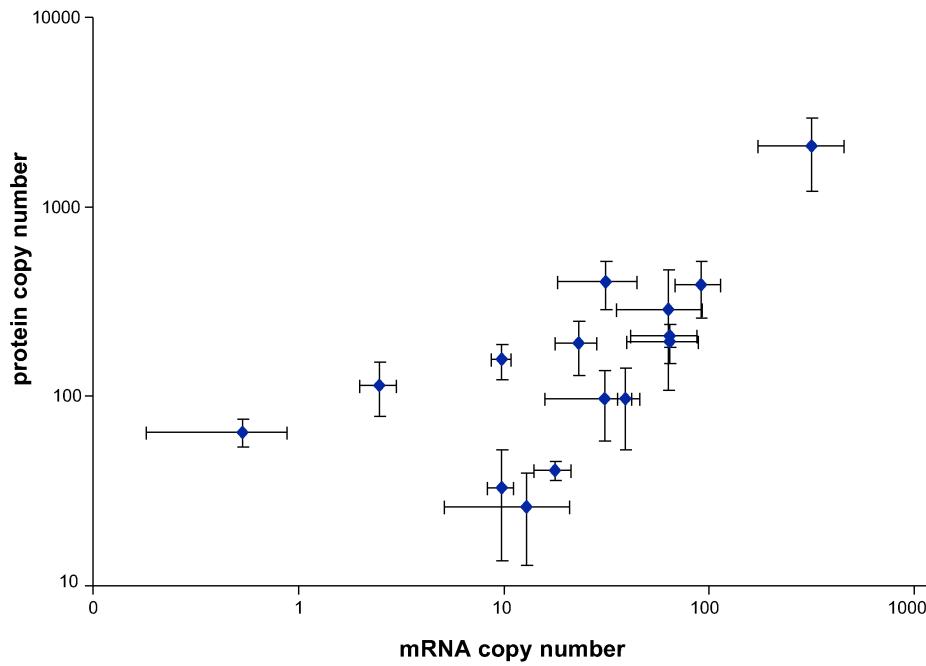


Figure S7: Correlation between measured mRNA copy numbers and protein copy numbers inferred from luminescence of individual cells (see Methods).

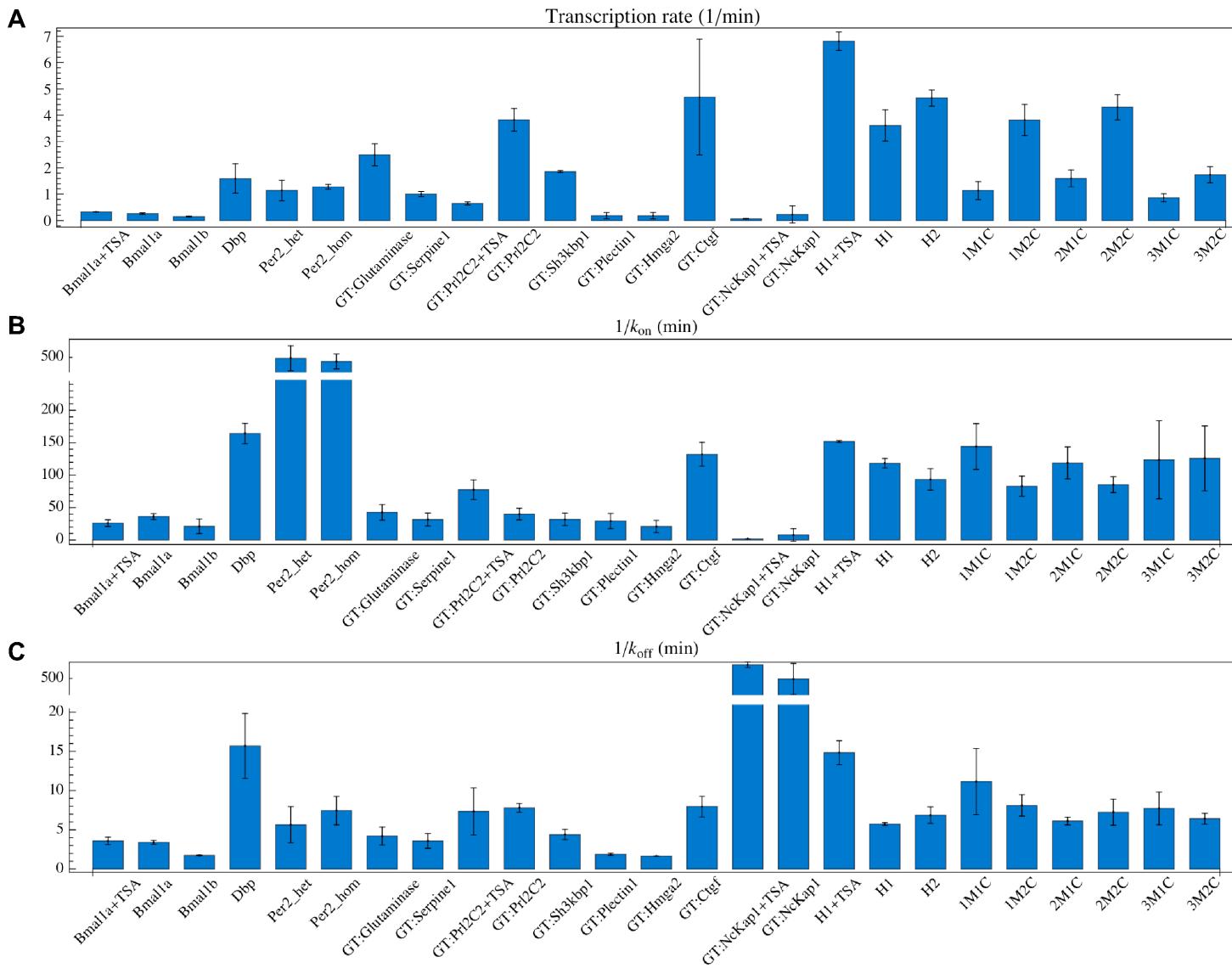


Figure S8: Inferred parameters of the gene expression model: (A) transcription rates k_m ; (B) mean off-time ($1/k_{on}$); (C) mean on-time ($1/k_{off}$)

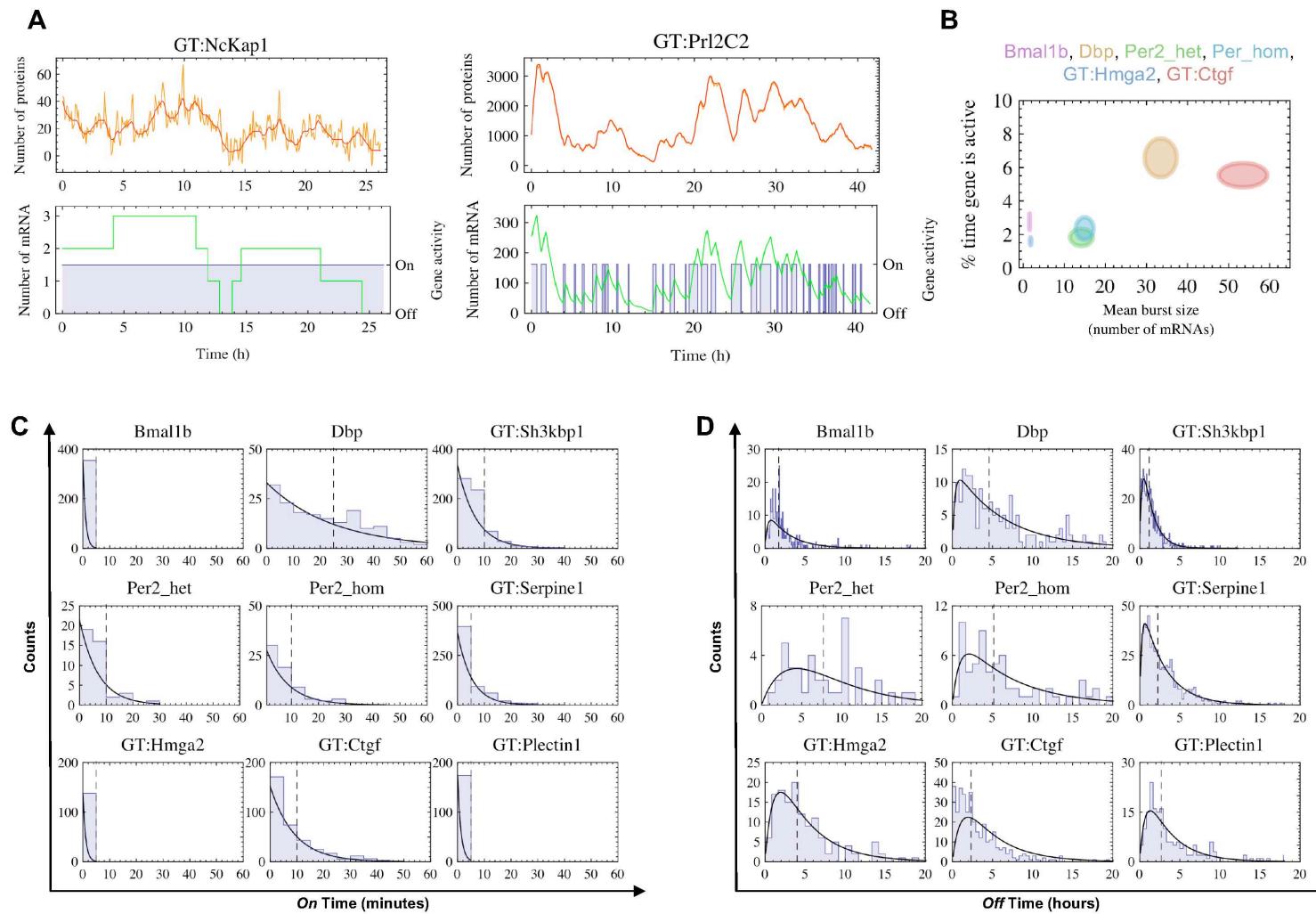


Figure S9: Inference of gene activity profiles and computation of burst sizes, on-times and off-times distributions of Bmal1b, DBP, GT:Sh3kbp1, Per2_heterozygous, Per2_homozygous, GT:Serpine1, GT:Hmga2, GT:Ctgf, and GT:Plectin1. **(A)** Other examples of protein copy number (red), mRNA copy number (green), and gene activity (blue) inferences from a luminescence trace (orange). **(B)** 2D diagram of mean burst sizes versus percentage of time the gene is active (color legend is shown on top of the diagram). Ellipses represent means \pm two standard deviations. **(C)** Distributions of ‘on’ intervals; black lines show exponential fits. **(D)** Distributions of ‘off’ intervals; black lines show best fits to ‘two-step’ model (cf. Methods). Black dotted vertical lines show medians of distributions.

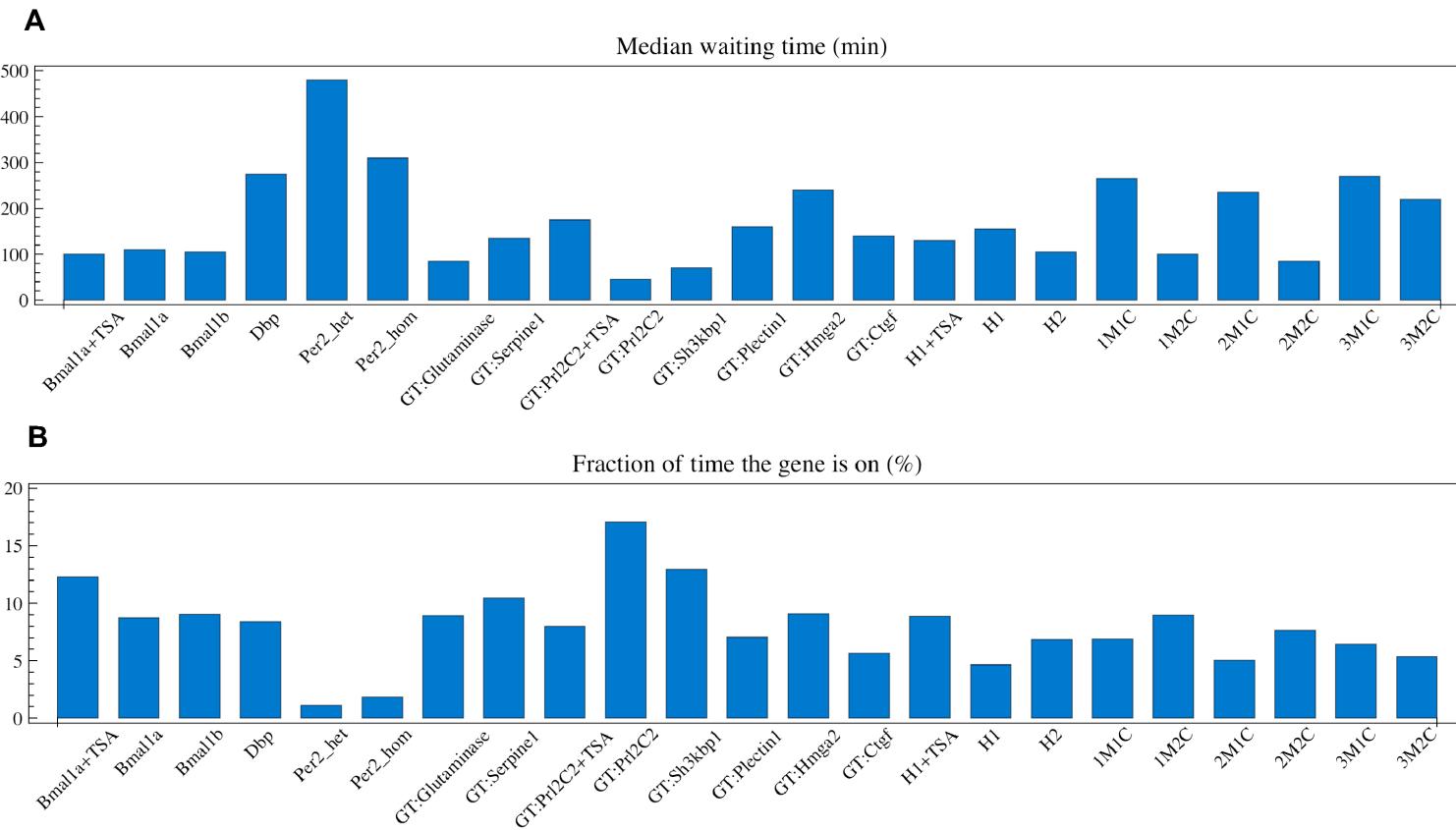


Figure S10: (A) Medians of waiting time distributions from off to on state. (B) percentage of time the gene is on.

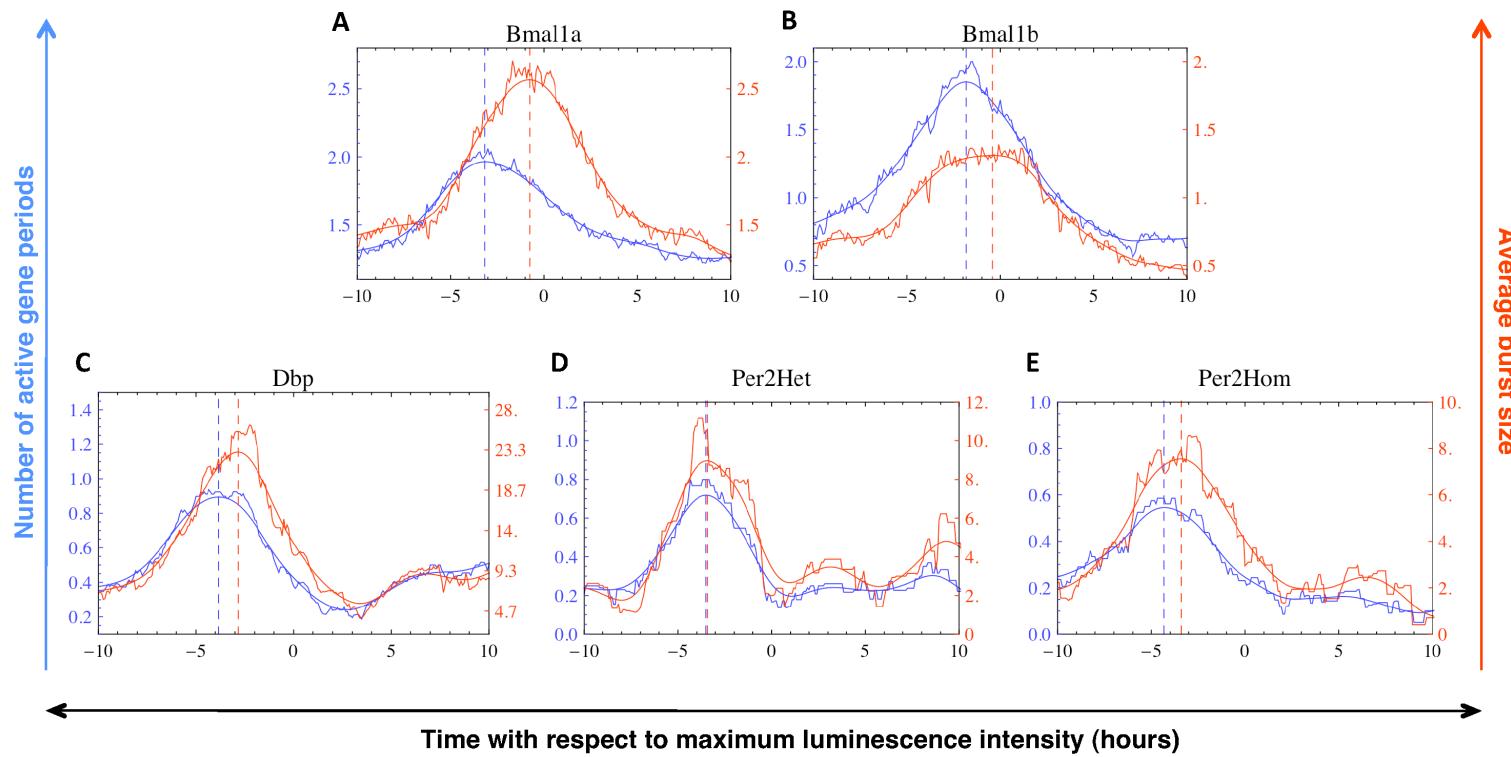


Figure S11: Variations of burst frequencies and sizes during a circadian cycle. Number of active periods (blue) and mean burst sizes (red) were computed using time windows of four hours. Smoothed curves were obtained using a Gaussian filter. Vertical dotted lines indicate time at which the number of active periods (blue) and mean burst sizes (red) reach their peak. **(A)** Bmal1a; **(B)** Bmal1b; **(C)** DBP; **(D)** Per2luc heterozygous; **(E)** Per2luc homozygous.

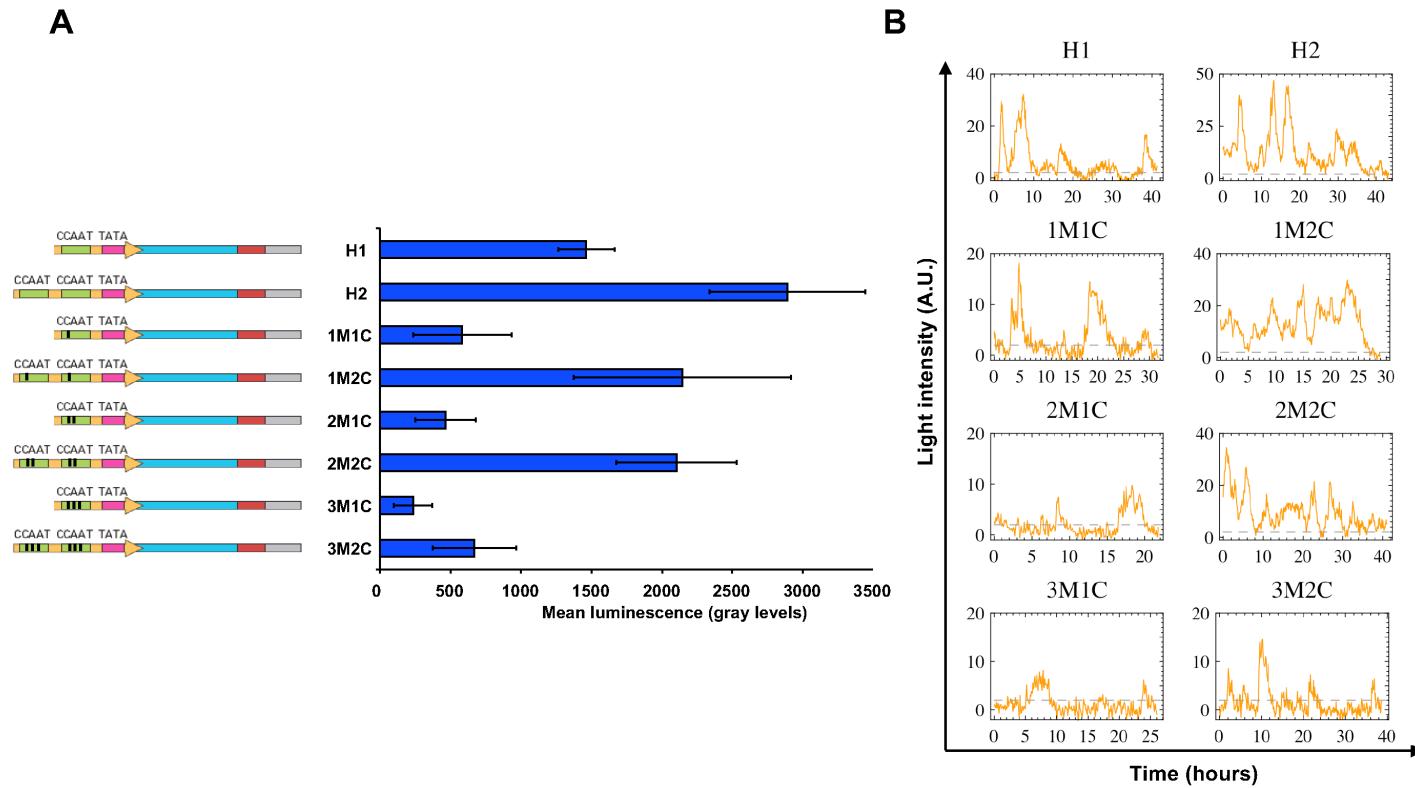


Figure S12: (A) Average luminescence intensity of CCAAT box cell lines from single-cell measurements in the lumiview microscope. (B) Single cell traces from cell lines carrying a single-copy of artificial promoter constructs containing one or two CCAAT box(es).

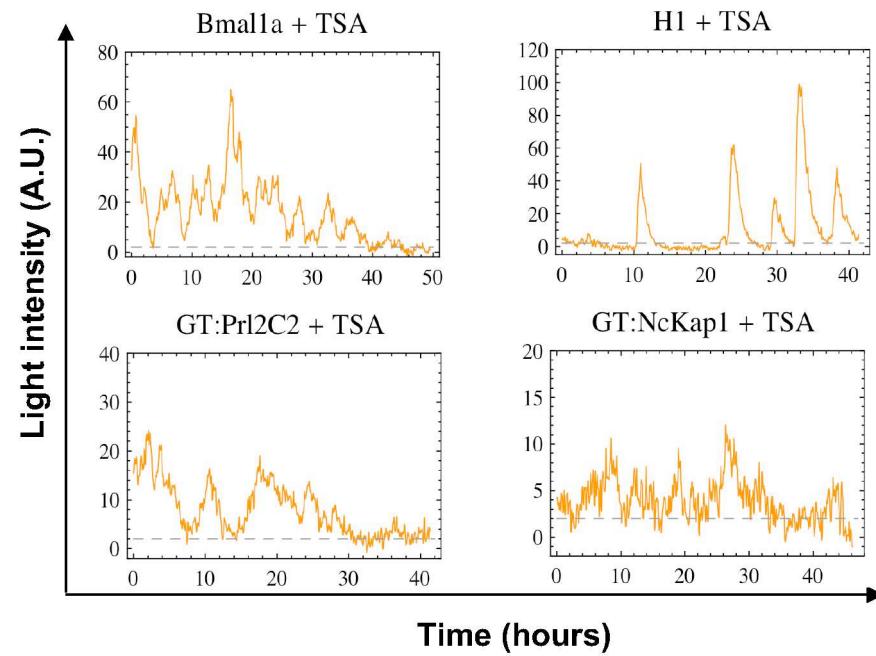


Figure S13: Single-cell traces from cell lines treated with TSA.

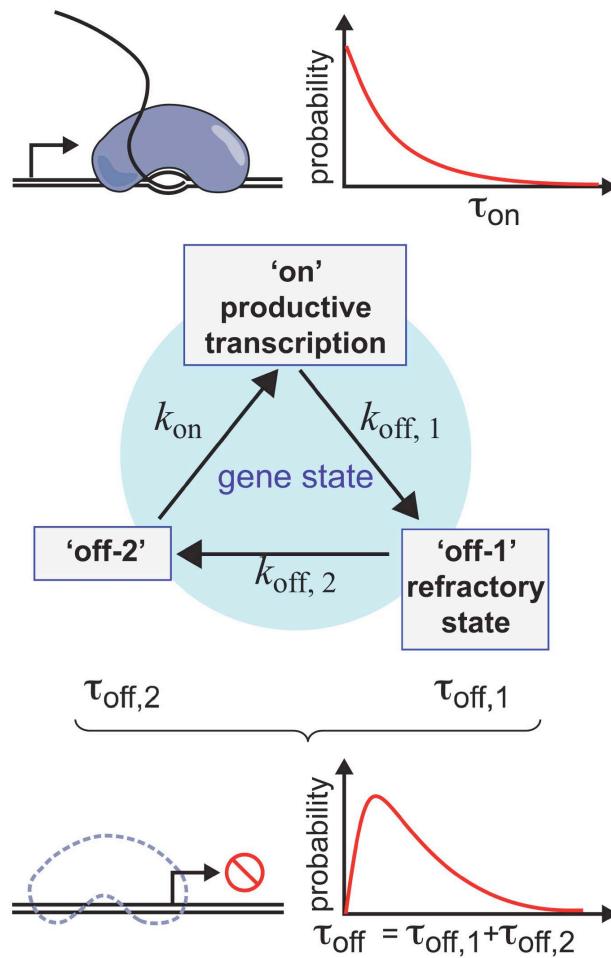


Figure S14: Three-state model of gene expression constrained by refractory periods of inactivity: the gene activity proceeds sequentially through the 'on' state (only here can mature mRNAs be produced), the 'off-1' (or refractory state), the 'off-2' state, and returns to 'on', with the indicated rates. Thus the 'on' times (τ_{on}) are exponentially distributed while the net 'off' times ($\tau_{\text{off}} = \tau_{\text{off},1} + \tau_{\text{off},2}$) show a peaked distribution. In the 'off' states, polymerases could either be absent from the promoter or present in a paused or otherwise inactive state. The times τ_{on} , $\tau_{\text{off},1}$ and $\tau_{\text{off},2}$ are defined as $1/k_{\text{off},1}$, $1/k_{\text{off},2}$ and $1/k_{\text{on}}$, respectively.

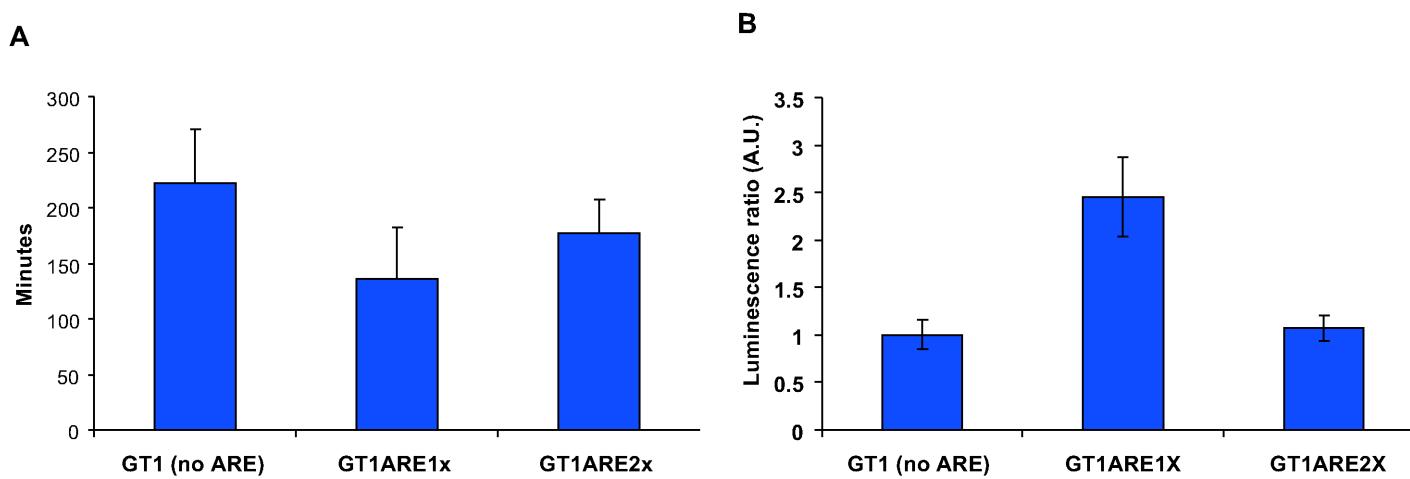


Figure S15: NIH-3T3 were transduced with a gene trap lentivector containing either none, one or two AU-rich element(s) (ARE). **(A)** Half-luminescence decay time. **(B)** Luminescence emission ratio between cells treated with 100ng/ml of bFGF for five hours and non-treated cells. A.U.: arbitrary units.

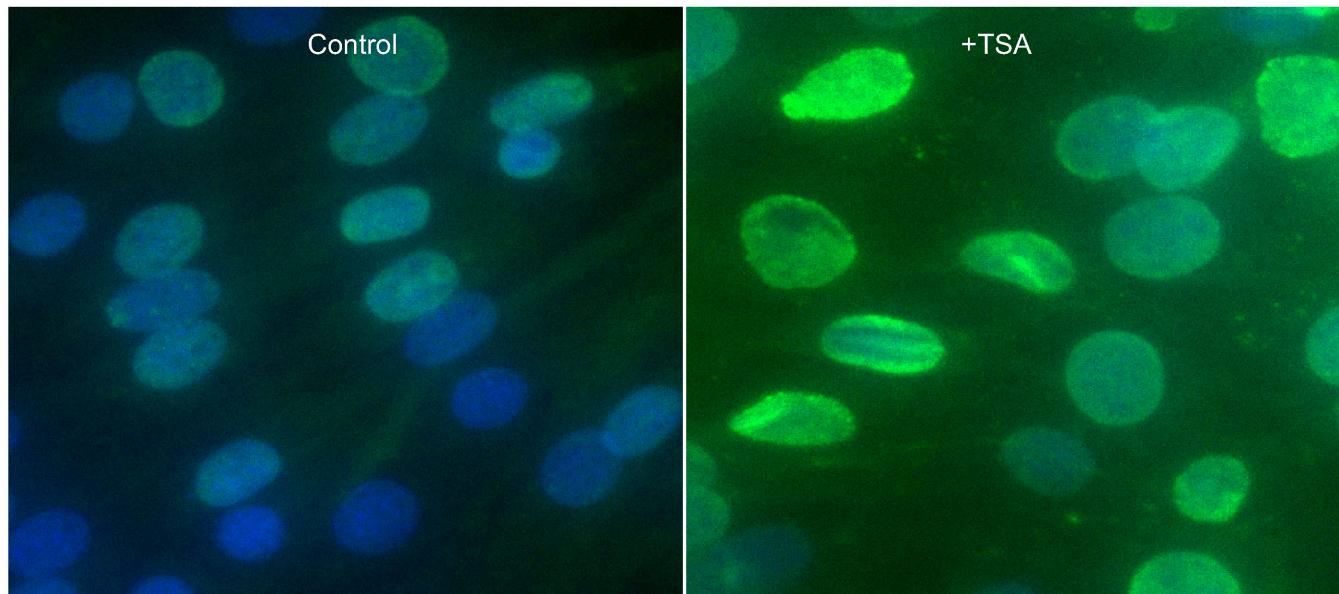


Figure S16: Immunofluorescence using an antibody for acetylated histone H4 (green), with or without a 24 hours $2\mu\text{M}$ trichostatin A treatment. Nuclei were stained with DAPI (blue).

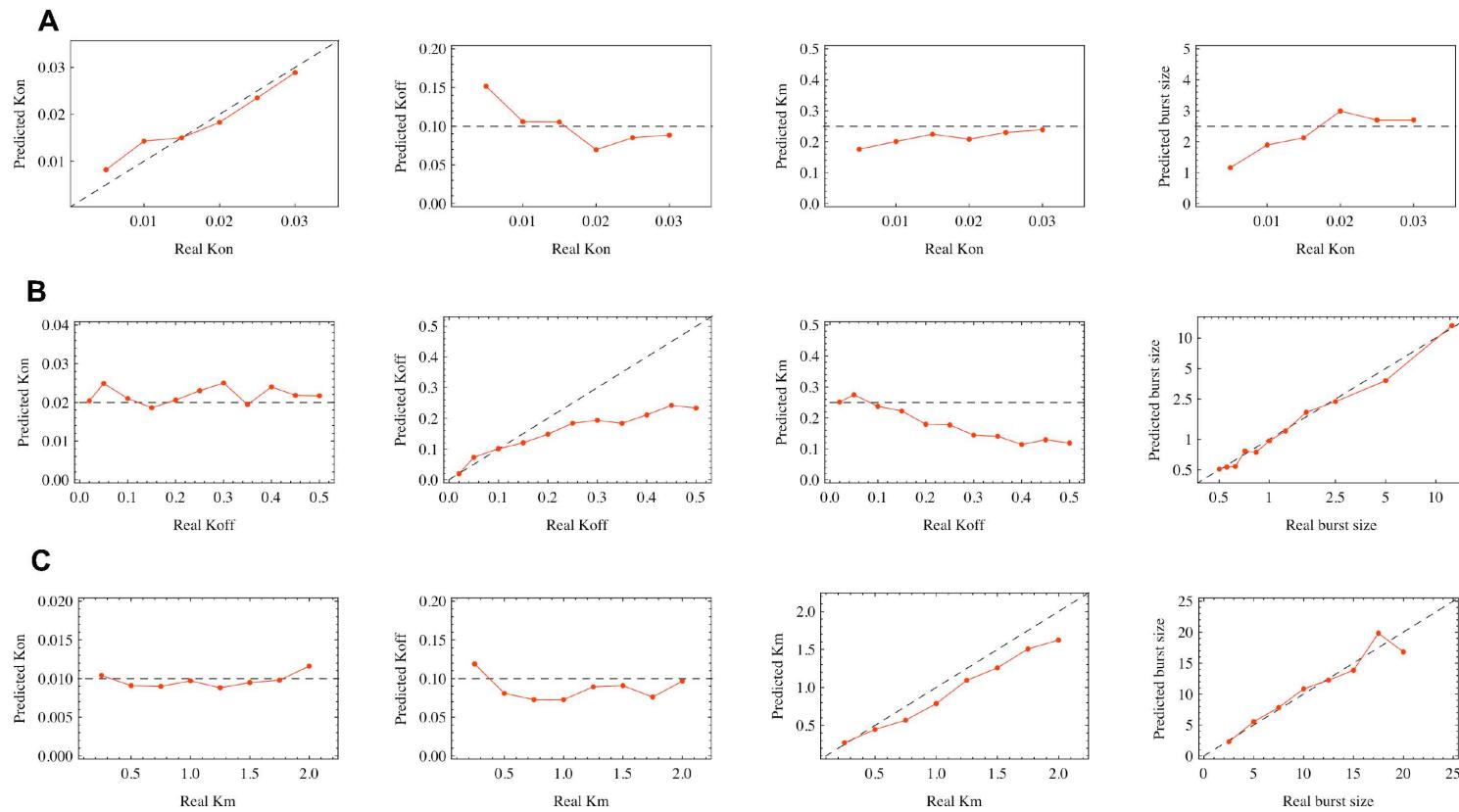


Figure S17: Validation of the parameter inference. **(A)** predicted versus inferred rates of six simulated clones with different k_{on} 's and $k_{\text{off}} = 0.1 \text{ min}^{-1}$ and $k_{\text{m}} = 0.25 \text{ min}^{-1}$. **(B)** Similar for 11 simulated clones with different k_{off} 's and $k_{\text{on}} = 0.02 \text{ m}^{-1}$ and $k_{\text{m}} = 0.25 \text{ min}^{-1}$. **(C)** Similar for 8 simulated clones with different transcription rates and $k_{\text{on}} = 0.01 \text{ min}^{-1}$ and $k_{\text{off}} = 0.1 \text{ min}^{-1}$.

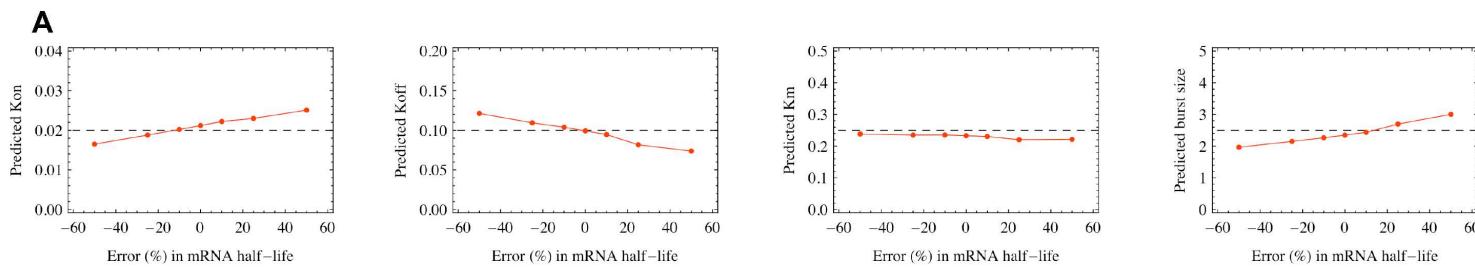


Figure S18: Effects of the estimated mRNA half-lives on parameter estimation. Here we simulated data with parameters as in Fig. S17A and estimated parameters by holding fixed a modified value of the mRNA half life to mimic a biased estimation. We find that the relative errors in the predicted k_{on} and k_{off} were comparable or less than the errors in half-lives, while the error on k_m did not vary much with the error on the mRNA half-life. The error on the burst sizes, given by the ratio k_m/k_{off} mainly reflect k_{off} .

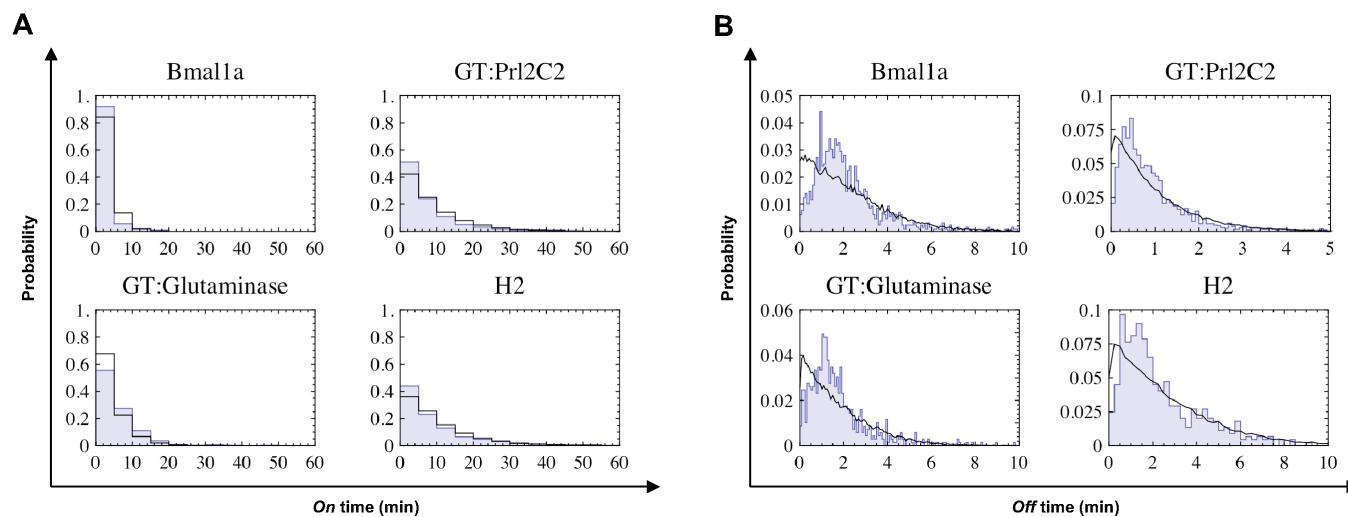


Figure S19: Distributions from four different clones (blue) and four simulated clones (black) with the same kinetic parameters. **(A)** 'on' time distributions. **(B)** 'off' time distributions.

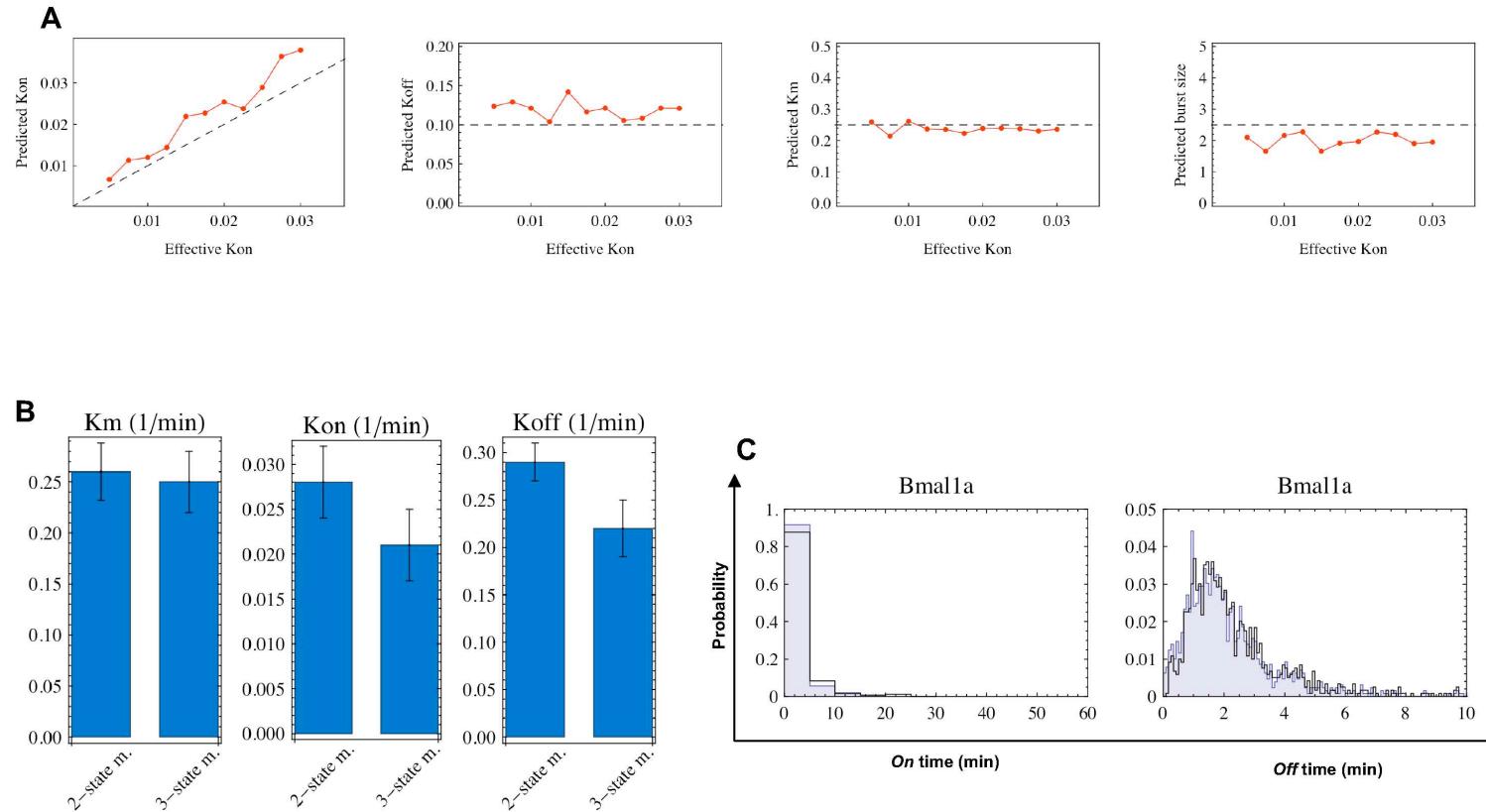


Figure S20: Three vs. two state model. **(A)** Data were generated with a three state model ('on', 'off-1', 'off-2') and analyzed with the two state model (cf. Section 2.7 in the supplement). Simulated parameters are as in Fig. S17A except that the 'off' state has been split into two equal length 'off-1' and 'off-2' states. The 'effective' $k_{on} = 1/(\tau_{off,1} + \tau_{off,2})$ is varied and predicted parameters are shown. **(B, C)** The data for the Bmal1a clone is analyzed with the two (blue) or three state (black) models (for the latter, parameter estimation is done via sampling, cf. section 2.7 in supplement). **(B)** Parameter values inferred for the two and three state models show consistent values. **(C)** The predicted 'on' and 'off' time distributions are very similar.

References

- [1] G Friedrich and P Soriano. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev*, 5(9):1513–23, Sep 1991.
- [2] S K Jang and E Wimmer. Cap-independent translation of encephalomyocarditis virus rna: structural elements of the internal ribosomal entry site and involvement of a cellular 57-kd rna-binding protein. *Genes Dev*, 4(9):1560–72, Sep 1990.
- [3] M D Ryan and J Drew. Foot-and-mouth disease virus 2a oligopeptide mediated cleavage of an artificial polyprotein. *The EMBO Journal*, 13(4):928–33, Feb 1994.
- [4] R Winzen, M Kracht, B Ritter, A Wilhelm, C Y Chen, A B Shyu, M Müller, M Gaestel, K Resch, and H Holtmann. The p38 map kinase pathway signals for cytokine-induced mrna stabilization via map kinase-activated protein kinase 2 and an au-rich region-targeted mechanism. *The EMBO Journal*, 18(18):4969–80, Sep 1999.
- [5] Emi Nagoshi, Camille Saini, Christoph Bauer, Thierry Laroche, Felix Naef, and Ueli Schibler. Circadian gene expression in individual fibroblasts: cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell*, 119(5):693–705, Nov 2004.
- [6] Jürgen A Ripperger and Ueli Schibler. Rhythmic clock-bmal1 binding to multiple e-box motifs drives circadian dbp transcription and chromatin transitions. *Nat Genet*, 38(3):369–74, Mar 2006.
- [7] W Bi, L Wu, F Country, B de Crombrugghe, and S N Maity. Dna binding specificity of the ccaat-binding factor cbf/nf-y. *J Biol Chem*, 272(42):26562–72, Oct 1997.
- [8] T Dull, R Zufferey, M Kelly, R J Mandel, M Nguyen, D Trono, and L Naldini. A third-generation lentivirus vector with a conditional packaging system. *J Virol*, 72(11):8463–71, Nov 1998.
- [9] <http://medweb2.unige.ch/salmon/lentilab/qpcritration.html>.
- [10] Caroline C. Friedel, Lars DÄ¶lken, Zsolt Ruzsics, Ulrich H. Koszinowski, and Ralf Zimmer. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Research*, 37(17):e115, 2009.
- [11] J Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, Jan 2005.
- [12] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mrna synthesis in mammalian cells. *Plos Biol*, 4(10):e309, Oct 2006.
- [13] Nir Friedman, Long Cai, and X Sunney Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.*, 97(16):168302, Oct 2006.
- [14] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA*, 105(45):17256–61, Nov 2008.
- [15] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-rna counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, 15(12):1263–71, Dec 2008.

- [16] Jeroen S van Zon, Marco J Morelli, Sorin Tănase-Nicola, and Pieter Rein ten Wolde. Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophysical Journal*, 91(12):4350–4367, Nov 2008.
- [17] S Iyer-Biswas, F Hayot, and C Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E*, Jan 2009.
- [18] Aleksandra M Walczak, Andrew Mugler, and Chris H WIggins. Analytic methods for modeling stochastic regulatory networks. *arXiv*, q-bio.MN, May 2010.
- [19] Rui Zhen Tan and Alexander van Oudenaarden. Transcript counting in single cells reveals dynamics of rdna transcription. *Mol Syst Biol*, 6:1–7, Jan 2010.
- [20] Huili Guo, Nicholas T Ingolia, Jonathan S Weissman, and David P Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–40, Aug 2010.
- [21] H Haario, M Laine, A Mira, and E Saksman. Dram: efficient adaptive mcmc. *Statistics and Computing*, Jan 2006.
- [22] <http://learn.hamamatsu.com/articles/ccdsnr.html>.
- [23] LR Rabiner and BH Juang. A tutorial on hidden markov models. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [24] Richard Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.
- [25] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annual review of physical chemistry*, 58:35–55, Jan 2007.
- [26] 19. D B Percival and A T Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, 2003.
- [27] C Torrence and GP Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 1998.