Supplementary Information for

Inferring gene regulation from stochastic transcriptional variation across single cells at steady-state

Anika Gupta[a,b,1], Jorge D. Martin-Rufino[a,c,d,1], Thouis R. Jones[a], Vidya Subramanian[a], Xiaojie Qiu[e,f], Emanuelle I. Grody[a], Alex Bloemendal[a], Chen Weng[a,c,d,e], Sheng-Yong Niu[a], Kyung Hoi Min[e,g], Arnav Mehta[a,d,h], Kaite Zhang[a], Layla Siraj[a], Aziz Al' Khafaji[a], Vijay G. Sankaran[a,c,d], Soumya Raychaudhuri[a,b,i], Brian Cleary[a,2], Sharon Grossman[a,2], Eric S. Lander[a,j,k,2,*]

Email: agupta@broadinstitute.org and lander@broadinstitute.org

This PDF file includes:

- Supplementary Methods
- Figures S1 to S4
- Table S1
- SI References

Supplementary Methods

We set out to infer gene regulation from stochastic transcriptional variation across single cells at steady-state. To accomplish this, we first modeled the behavior of a pair of genes—one transcription factor and its target gene—undergoing transcriptional bursting and following the central dogma, as well as a regulatory link function between the two genes. In order to build ground truth scenarios, we simulated a population of cells that followed our underlying model of bursting, the central dogma, and gene regulation, tracking mRNA and protein abundances over time in order to estimate the regulatory signal present.

Burst switching function

We modeled a burst switching function for each gene in each cell, which determined whether transcriptional bursting switched from the OFF to ON state or vice versa, depending on which state that gene currently occupied in that cell (1-4). We made the burst switching parameters exponentially distributed to reflect the expected time in which the state would switch:

$$P(state\ switch\ at\ t) = p|k_{on}, k_{off} = \left\{\lambda e^{-(\lambda x)}\ if\ x < t;\ 0\ if\ x \geq t\right\} \quad (1)$$

where λ is $1/k_{on}$ or $1/k_{off}$ depending on whether bursting is switching ON or OFF, respectively, and x represents the time elapsed since in the current state.

Expected mRNA abundance

The expected mRNA abundance for a single gene *G* in a given cell at any moment in time ($G^{(RNA)}$), which, in our model, is a direct result of compounded transcriptional bursts, can be calculated as below. Specifically, the steady-state distribution of $G^{(RNA)}$ across cells follows a Beta-Poisson distribution:

$$E\left(s_{RNA}, p\right) = Poisson(sp)$$

$$p|k_{on}, k_{off} = Beta(k_{on}, k_{off})$$

such that

$$E(G^{(RNA)}|s, kon, koff) = s\ \frac{kon}{kon + koff} \quad (2)$$

where $G^{(RNA)}$ is the mRNA abundance for gene *G*, s_{RNA} is the transcription (i.e. mRNA synthesis) rate for gene *G*, and *p* is the state-switching probability, governed by the gene-specific parameters k_{on} and k_{off} as in Eqn. 1. We note that E(*p*) is the fraction of time spent in the ON state. Further, each of these parameters are in units of mRNA half-life and thus do not explicitly include mRNA decay in this formulation.

Modeling mRNA and protein decay

We modeled the decay of both mRNA and protein abundances for the two genes as Poisson processes, with rate parameters corresponding to the degradation rate:

$$G^{(RNA)}(t+1) = G^{(RNA)}(t) - Poisson(G^{(RNA)}(t) * \delta * step\_size)$$
$$G^{(P)}(t+1) = G^{(P)}(t) - Poisson(G^{(P)}(t) * \gamma * step\_size) \quad (3)$$

where δ is the mRNA degradation rate ($\ln 2/t_{1/2,\ RNA}$), γ is the protein degradation rate ($\ln 2/t_{1/2,\ Protein}$), t is the current time since the initial abundance is measured in minutes, and *step_size* is one time-step in minutes. To restrict future abundances such that they are less than or equal to the current mRNA or protein abundance for each cell, we update each abundance to be:

$$min(G^{(RNA)}(t),\ G^{(RNA)}(t+1))$$

or the protein equivalent, depending on the entity of interest.

mRNA variance decomposition

The variance of mRNA for each of the two genes can be calculated based on the key parameters that dictate mRNA abundance. Given the steady-state distribution of $G^{(RNA)}$ as in (Fig. 2*C*), the variance can be calculated as:

$$Var\left(G^{(RNA)}\right) = E[Var(G^{(RNA)}|s_{RNA}, p)] + Var[E\left(s_{RNA}, p\right)]$$

Since $G^{(RNA)}$ is Poisson-distributed,

$$E\left(G^{(RNA)}|s_{RNA}, p\right) = Var\left(G^{(RNA)}|s_{RNA}, p\right) = s_{RNA} * p$$

Thus,

$$= E\left(s_{RNA} * p\right) + s_{RNA}^2 * Var(p)$$

$$= b\frac{kon}{kon+koff} + b^2\frac{kon*koff}{(kon+koff)^2(kon+koff+1)} \quad (4)$$

where b is the gene's burst size in transcripts per burst, and k_{on} and k_{off} are gene-specific parameters. Thus, the burst size for each gene directly scales the mRNA variance for the gene across the cell population.

Relationship between mRNA and protein abundances in the limit of an infinite number of cells

To provide intuition for our stochastic simulations, we modeled the flow of information between TF mRNA to TF protein for the deterministic limit of an infinite number of cells with an ordinary differential equation that captures mRNA translation and decay of both mRNA and protein:

$$\frac{dp}{dt} = \alpha * m(t) - \gamma * p(t) \quad (5)$$
$$\frac{dm}{dt} = -\delta * m(t)$$

where α is the translation rate, γ is the protein degradation rate ($\ln 2/t_{1/2,\ prot}$), δ is the mRNA degradation rate ($\ln 2/t_{1/2,\ RNA}$), *p* is the protein abundance for the gene (tracked in thousands of molecules), and *m* is the mRNA abundance for the gene (in number of transcripts), both with respect to time. We note that the complete process of translation includes mRNA export, translation in the ribosome complex, and re-importation into the nucleus; however, the first and last steps are much faster, occurring on the order of minutes, and thus we did not explicitly model them.

The timing when the TF protein abundance best reflects the TF mRNA abundance at t=0 reflects approximately when the TF has its largest effect on its target genes' transcription. When solving for the time when $\frac{dp}{dt}$ = 0, which captures the time point when the protein abundance best reflects the mRNA abundance at t=0, we get:

$$\frac{log(\gamma) - log(\delta)}{\gamma - \delta} \quad (6)$$

Regulatory response function

We modeled the regulatory response function between the TF and its target gene as a Hill function, as previously experimentally measured and modeled (5-13). Specifically, by binding to the target gene, the $TF^{(P)}$ directly modulates the $Target^{(kon)}$, or the rate at which the target gene bursts. We can capture this as follows:

$$Target^{(kon\_adjusted)} = Target^{(kon)} * \frac{1}{1+(\frac{mean(TF^{(P)})}{(TF^{(P)})})^{n}} \quad (7)$$

where k_{on} is the gene-specific burst frequency and *n* is the Hill coefficient. We calculated the mean $TF^{(P)}$ across the entirety of simulated cells at one time point.

Simulations setup

To model the behavior above, we wrote a simulation (Python v3.7) based on the widely accepted two-state stochastic bursting model of gene transcription and simulated two genes, one TF (regulator) and its target gene, across 20,000 cells. Computational state transitions involved updating each parameter according to the effect every other parameter had on it (if any). Each state represented a single minute, and to allow cells to reach steady-state, we let the simulator run for a burn-in time of 12 days.

The underlying stochastic model (see **Eqn. 1-2**) relies on the following transcriptional burst parameters, which we incorporated in a gene-specific manner: k_{on}, k_{off}, and burst size. Parameter values for each gene are based on TF- and non-TF ranges observed experimentally (see **Table 1**). We also incorporated splicing, translation, mRNA degradation, and protein degradation rates from the literature in a TF- and non-TF-specific manner (**Table 1**). Bursting state switches were modeled as exponential

with the rate parameter equal to k_{on} or k_{off} for turning bursting off or on, respectively (**Eqn. 1**). Transcription rate was calculated as the burst size * k_{off} for each gene, and the rate of mRNA synthesis was the product of whether bursting was on for that gene and the transcription rate. We used a Poisson decay model to enable discrete removal of mRNA transcripts (see **Eqn. 3**).

We modeled the $TF^{(P)}$:$Target^{(kon)}$ response function as a Hill function (see **Eqn. 7**), where the Hill coefficient and maximum possible effect were variable inputs. This choice of response function was grounded in prior modeling of experimental data to map TF concentration changes to target gene mRNA changes, where either a sigmoidal or Michaelis-Menten curve was predominantly concluded as best reflecting the dose-response nature of regulatory interactions (5-13). We only modeled activation in gene regulation, rather than including transcriptional repression. At each time step, we tracked $TF^{(\Delta RNA)}$,$TF^{(RNA)}$, $TF^{(P)}$, $Target^{(\Delta RNA)}$, and $Target^{(RNA)}$, and $Target^{(kon)}$, to reflect the abundance changes between each time point, where ∆RNA corresponds to the number of unspliced RNA transcripts at the time of sampling.

To assess regulatory signal, we calculated the following time-shifted RNA:RNA Spearman's rho correlations between: (1) TF_0 and $Target_T^{\Delta}$ (C_T^{Δ}) and (2) TF_0 and $Target_T$ (C_T), for each time point ranging from 0 hours after the burn-in time to up to 2 days after.

Gene-specific parameter sensitivity analyses

We measured the sensitivity of C_T and C_T^{Δ} to gene-specific parameters for a range of values from the literature. We varied each parameter to take on either the 25th, 50th, or 75th percentile value, taken from the literature (**Table 1**). To build confidence intervals around estimates of regulatory signal, we ran 25 independent simulations of 20,000 cells each. In each run, we estimated the regulatory effect using the interdecile ratio measure (ratio of the mean abundance of the top decile of cells to the mean of the bottom decile of cells) as well as Spearman's rho (C_T and C_T^{Δ}), as described in the main text.

We plugged in various values of mRNA and protein half-lives into **Eqn. 6** to analytically model each parameter's effect on protein production from a baseline mRNA abundance at t=0.

Simulation of state-based covariation

We simulated two cell "states" with no regulation between the TF and Target to test the gene pair covariation in the absence of direct regulation but with state-based structure. We first removed the link function between $TF^{(P)}$ and $Target^{(kon)}$ such that the target gene's burst frequency was not changed by $TF^{(P)}$. To simulate the cell states, we created one state with low mRNA abundances and another with high abundances for each gene—each state had 10,000 cells. Specifically, we varied the basal burst frequency for both the TF and Target to be either their first quartile or third quartile values (the "low" state had both the TF and Target bursting at their first-quartile frequencies, and the "high" state had the genes bursting at their third-quartile frequencies). At each time point, we then combined the 10,000 cells from each state to

yield a total of 20,000 cells that we tracked over time for the key entities described in the main text.

In order to check for state-based confounding, we tracked both the time-shifted Spearman's rho between the gene pair over time C_T^{Δ} and the role reversal Corr(Target$^{(RNA)}_0$, TF$^{(\Delta RNA)}_T$) Spearman's rho over time as well, as described in the text. We compared these values and the shapes of the response curves in each case to those generated by the simulated cells with regulation and only one cell state present.

Assessing the effect of sample size and sequencing inefficiencies

To mimic the combined effects of capture inefficiencies and read-depth in scRNA-seq (which we call "UMI detection efficiency"), we varied the sampling density of counts for each cell by introducing Poisson downsampling: 10-90% downsampling in intervals of 10%. Specifically, we sampled counts for each cell at each time point, with the Poisson rate parameter equal to the number of counts*capture_efficiency, and we chose the minimum between the raw and downsampled counts to ensure that we never overestimated the number of counts per cell.

To determine the effect of either changing the sample size (number of cells) or UMI detection efficiency on our ability to detect regulation-based covariation, we ran the simulator 25 times for each pair of capture efficiencies and sample sizes (10-100% in intervals of 10 for capture efficiencies, and 1e3, 2.5e3, 5e3, 10e3, 25e3, and 50e3 for number of cells), and we tracked the magnitude and variance in TF:Target Spearman's rho values across simulation runs at each time point for each pair of cell number and capture efficiency values. We also calculated the coefficient of variation (standard deviation/mean) of (1) C_T^{Δ} and (2) C_T^{Δ} - C_0^{Δ} across simulation runs as we varied the number of cells sampled, the UMI detection efficiency, and the TF and Target mRNA half-lives.

Pulse-chase experiment with 4-thiouridine

Low passage K562 erythroleukemia cells (ATCC, CCL-243) were cultured in RMPI 1640 + L-glutamine, with 10% FBS, 1% Penicillin / Streptomycin and 1% L-glutamine 200mM at 37ºC in a humidified atmosphere with 5% CO_2. For 4-thiouridine (4sU) experiments, cells were plated in 6-well plates ~12h before the start the experiment at a density of 5x 10^5 – 8 x 10^5 cells/mL in 5mL of fresh media. 1M 4sU stocks were prepared by dissolving 4sU powder (Sigma, T4509-25MG) in DMSO, and added to culture wells at a final concentration of 100µM. For the pulse phase, 4sU was added to the media for 24h, with renewal every 6-8h. Between the pulse and chase phases, the cells were washed twice with Dulbecco's Phosphate-Buffered Saline to remove any residual traces of 4sU. Subsequently, media with saturating concentrations of uridine (Sigma, U6381) at 10mM (a 100X excess compared to 4sU) was added. Cells were collected at 0h, 2h, 4h, 6h, 8h, and 10h.

Light exposure to the samples was minimized between the addition of 4sU to the media until the completion of reverse transcription. All the samples from each pulse-chase experiment were processed the same day to minimize batch effects. To control for

genes induced by the long exposure to 4sU and to uridine, two control samples with chase only or DMSO only were included.

Conventional and temporally-resolved single-cell RNA sequencing

The same datasets were used to analyze stochastic transcriptional variation using correlations between steady-state total counts and gene-specific burst sizes and mRNA half-lives. Our scNT-seq protocol was adapted from previously published methods (14-15). This technique relies on the chemical conversion of 4sU into cytosine analogs that result in apparent T-to-C mutations in the cDNA following reverse transcription of labelled RNA. Briefly, following completion of the pulse-chase experiment, cells were collected and washed twice with 0.01% Bovine Serum Albumin (BSA, Sigma-Aldrich, A8806-5G) in DPBS, filtered through a 40µm cell strainer (Corning, 431750), counted, diluted to a concentration of 120 cells/uL in DPBS-0.01% BSA (200 cells/uL for the replicate with lentiviral barcoding for doublet detection –see below–, and 120 cells/uL for the replicate without lentiviral barcoding), and transferred to a 10mL Luer lock syringe with a magnet (V&P Scientific, 782N-6-150) for gentle stirring. Barcoded Oligo dT primer on beads for Drop-seq (26) (ChemGenes, MACOSKO-2011-10(V+)) were resuspended at a concentration of 130 beads/uL in 10mL of lysis buffer (4mL of RNAse free water, 3mL of 20% Ficoll® PM 400 (Sigma, 26873-85-8), 100uL of N-Lauroylsarcosine sodium salt solution 20% (Sigma, 137-16-6), 400uL of 0.5M EDTA (Invitrogen, 15575-038), 2mL of 1.0M Tris-HCl, pH 7.5, and 500uL of 1.0M DTT (Caiman Chemical, 700416) and transferred to a 10mL Luer lock syringe (BD, 300912) (with a magnet (V&P Scientific, VP 772DP-N42-5-2) for gentle stirring. Droplet generation oil (Biorad, 1863005) was loaded into a 30mL Luer lock syringe (BD, 302832). The experiment was performed using uFluidix Drop-seq chips with hydrophobic coating, using a Photron Fastcam SA5 camera, KD Scientific Syringe Pumps (KDS, 78-2910) and micromedical tubing (Scientific Commodities, BB31695-PE/2). Flow rates were set to 15,000 uL/h for oil, 4,000uL/h for cells and 4,000uL/h for beads, and a new collection was started every 15–20 minutes. Droplet emulsions were collected in 50mL conicals and kept in ice and protected from light until breakage.

Following oil removal, 30mL of 6X SSC (diluted in water from 20X SSC, Life Technologies, 15557044) and 1mL of 1H,1H,2H,2H-perfluoro-1-octanol 97% (Sigma-Aldrich, 370533) were added to the tubes, which were then manually shaken to break droplets. The tubes were spun at 1,000x*g* for 1 minute and the supernatant was discarded. 20mL of 6X SSC were added twice to resuspend beads, transferred to new tubes, and spun at 1,000x*g* for 2 minutes. Bead pellets were transferred to 2mL Lobind tubes and washed twice with 1mL of 6X SSC. This and all subsequent centrifugations were performed in a spinning bucket centrifuge at 1,000x*g* for 1 minute.

Chemical conversion was performed using TimeLapse-seq chemistry (14), which entails the oxidative-nucleophilic-aromatic substitution of 4sU into cytidine analogs, that yield apparent U-to-C mutations upon reverse transcription that can be detected using next generation sequencing. In preparation for chemical conversion, beads were washed once in 450uL of 3M sodium acetate pH 5.2. (Thermo Scientific, R1181**),** 4uL of 0.5M

EDTA pH 8.0 and 430 uL of water. For the chemical conversion reaction, beads were resuspended in a mix containing 8uL of 3M sodium acetate pH 5.2, 2uL of 0.5M EDTA pH 8.0, 214uL of H2O, 13uL of 2,2,2-Trifluoroethylamine (Sigma-Aldrich, 91692-5ML), and 13uL of 192mM freshly prepared NaIO4 (Sigma-Aldrich, 311448-5G) in water. The reaction was conducted at 45ºC for 1hr with rotation. Upon completion of the reaction, beads were washed once with 1mL of TE (Sigma-Aldrich, 93302-100ML), and incubated at 37ºC for 30 minutes in 500uL of reducing master mix containing 5uL of 1M Tris-HCl ph7.5, 5uL of 1M DTT, 10uL of 5M NaCl, 1uL of 0.5M EDTA, 10uL of RiboLock RNase inhibitor (Thermo Scientific, EO0381) and 469uL of water. Beads were subsequently washed with 1mL of Tris-HCl buffer (10mM, ph8.0) and 0.3mL of Maxima H Minus 5X RT buffer (Thermo Scientific, EP0751). For the reverse transcription reaction, beads were resuspended in 200uL of 80uL of water, 40uL of Maxima H Minus 5X RT buffer, 40uL of 20% Ficoll PM-400, 20ul of 10mM dNTPs (NEB, N0447L), 5uL of 100uM template switch oligo (AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG), 5uL of RiboLock RNase inhibitor, and 10uL of maxima H minus reverse transcriptase enzyme (Thermo Scientific, EP0751). Beads were incubated for 30 minutes at room temperature with rotation, and then for 120 minutes at 42ºC with rotation. Following completion of the reaction, beads were washed with 1mL of TE-SDS and twice with 1mL of TE-TW, and stored at 4ºC for 1-2 days. To determine the additional number of cycles for whole transcriptome amplification, we first performed PCR on an aliquot of 6,000 beads washed twice in water (by adding to beads 25uL KAPA HiFi HS Readymix (Roche, 07958935001), 0.4uL of 100μM TSO-PCR primer (AAGCAGTGGTATCAACGCAGAGT) and 24.6μL of water) using the following parameters: 95ºC for 3 min; 4 cycles of (98ºC for 20s, 65ºC for 45s and 72ºC for 3 min); 9 cycles of (98ºC for 20s, 67ºC for 20s and 72ºC for 3 min); 72ºC for 5min; and hold at 4ºC. The PCR product was then purified using one round of 0.7X AMPURE XP beads (Beckman Coulter, A63881). The additional number of cycles required for optimal amplification using qPCR (which we empirically determined as the number of cycles that coincided with three fourths of the exponential amplification stage in qPCR) by adding 1uL of purified cDNA with 4.5uL of KAPA HiFi HS Readymix spiked with SYBR Green Dye (Lonza, 12001-796), 0.07uL of 25μM TSO-PCR primer and 3.53uL of water and using the following parameters: 95ºC for 3min; 25 cycles of (95ºC for 15s, 63ºC for 30s, 72ºC for 30s). Subsequently, large-scale PCR amplification was performed on the rest of beads with the same parameters as below plus the additional number of cycles determined by qPCR. To ensure high diversity in our libraries, 2–5 tagmentations were performed for each of the time points' cDNA pool using the Nextera XT DNA Library Prep Kit (Illumina, FC-131-1096), using 550pg of purified cDNA as input, and amplified in a second round of PCR (15uL of Nextera PCR mix, 5uL of 2μM P5-TSO hybrid primer –AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*A*C–, and 5uL of 2μM Nextera N70X oligo from (97) using the following parameters: 95ºC for 30s; 12 cycles of (95ºC for 10s, 55ºC for 30s and 72ºC for 30s); 72ºC for 5 min; and hold at 4ºC. PCR products were purified using two rounds of 0.6X AMPURE XP beads and measured using the Agilent 2100 Bioanalyzer High Sensitivity DNA kit (Agilent Technologies, 5067-4626 and 5067-4627). Pooled libraries were quantified using the KAPA Library Quantification Kits (07960204001), and sequenced in an Illumina NovaSeq 6000 System using a S2 flow cell with a 20bp (Read 1), 72bp

(Read 2) and 8bp (Index 1) configuration with a HPLC-purified custom read 1 primer (GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC).

Raw sequencing data demultiplexing, alignment, and temporal classification of reads

Sequencing files were demultiplexed using bcl2fastq v2.20 (Illumina). Fastq files from tagmentations belonging to the same time point were concatenated together. Using Dynast (https://github.com/aristoteleo/dynast-release), a wrapper around STARsolo (https://github.com/alexdobin/STAR/blob/master/docs/STARsolo.md), cell barcode and UMI sequencing errors were corrected and reads were aligned to the hg38v91 genome (16). Dynast count was used to parse the alignment BAM and quantify labeled/unlabelled/spliced/unspliced RNA and produce anndata h5ad files for each of the time points. We omitted background SNPs present in K562 by providing the genomic coordinates of SNPs computed on the non-chemically converted sample (which were defined as the genomic coordinates in which a mutation was present in >50% of reads).

To compute background mutation rates, we used two strategies: for the replicate with 0–6h chase time points, we used Dynast to compute the average mutation rate of non-T bases to any other base; for the replicate with 0–10h chase time points, we calculated the T-to-C mutation rate in a control sample treated with 4sU but without chemical conversion.

Lentiviral barcoding for homotypic doublet detection in single-cell RNA sequencing

Current doublet detection strategies for droplet-based single-cell transcriptomics are only able to distinguish heterotypic doublets (i.e. two different cell types in the same droplet), and are not well-suited for homotypic doublets (i.e. two cells from the same cell type in the same droplet). We devised a strategy that allowed us to increase the number of cells profiled by unit of time using Drop-seq by detecting doublets in homogeneous steady-state populations using expressed lentiviral degenerate barcodes. We also reasoned that removal of doublets would increase our chances of detecting stochastic differences in expression across single cells. Conceptually, the principle is to discard cell barcodes which have the top expressed lentiviral barcode and the second top lentiviral barcode expressed at a similar ratio. Briefly, we performed dial-out PCR on scRNA-seq libraries to selectively amplify the expressed lentiviral degenerate barcodes. Libraries were sequenced in an Illumina NextSeq System using a with a 20bp (Read 1) and a 20bp (Read 2) configuration with a HPLC-purified custom read 1 primer (GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC).

Sequencing files were demultiplexed using bcl2fastq v2.20 (Illumina). We used UMI-tools to error-correct cell barcodes, which we then linked to the corresponding expressed lentiviral barcode (https://genome.cshlp.org/content/early/2017/01/18/gr.209601.116.abstract). To generate a consensus list of expressed lentiviral barcodes, we used Starcode with Levenshtein distance of 3. We subsequently determined non-singlets using the ratio of

the highest expressed to the second highest expressed lentiviral barcode in each cell (please refer to this paper's Github https://github.com/sharigrossman/stochastic-regulation).

Gene-specific parameter estimation from real data (mRNA half-lives and burst sizes)

To estimate gene-specific half-lives from real data, the fraction of labeled counts for each gene at each time point was fitted to an exponential decay model (adapted from https://gist.github.com/johanvdw/443a820a7f4ffa7e9f8997481d7ca8b3). Genes in which the fraction of labeled counts increased overtime were removed from the analysis. Genes with a fitting $r^2 > 0.6$ were plotted for the comparison with half-lives from (14).

To estimate gene-specific burst sizes and frequencies, we used Dynamo's compute_bursting_properties function, which computes these parameters after fitting the distribution of total counts to a negative binomial distribution as previously described (17). Only genes with positive burst sizes were retained.

Identification of putative regulatory links using simultaneous correlation analysis

All AnnData .h5ad files were loaded into Scanpy to perform basic cell and gene quality filtering (18). In order to calculate simultaneous correlations, we first filtered out cells with low UMI and excluded genes expressed in only a small number of cells across replicates. We then scaled raw counts by total counts per cell to account for differences in library size. To determine the list of TFs and non-TF genes, we used a recently-published, curated list of human TFs (19).

Correlated genes for each TF were identified based on significant correlations in their expression profiles with the TF within the cells of each experimental time point, based on Spearman's rank correlation. Spearman's rho was used to quantify correlations based on ranked gene expression, computed with the correlatePairs function in scran (20). We assessed the significance of non-zero correlations using a permutation test, where the null hypothesis is that the ranking of normalized expression across cells is independent between genes. This allowed us to construct a null distribution by randomizing the ranks within each gene. The p-value for each gene pair is defined as the tail probability of this distribution at the observed correlation. We defined significantly correlated genes for each TF as the intersection of correlated genes with a p-value < 0.05 in all sampling time points.

Validation of *GATA1* putative targets using Perturb-seq data

To validate the putative targets of GATA1 identified using simultaneous correlation analysis, we leveraged a recently published dataset that combined CRISPR interference with a single-cell RNA sequencing readout (21). Processed count matrices were obtained from GSE132080 and loaded into Seurat v4.0.1 (22). Subsequently, the expression of each gene for each cell was normalized by the total expression, multiplied by a 10,000 scale factor, and log-transformed. A guide-cell barcode dictionary with the identity of the perturbation in each cell was obtained from GSE132080.

The *GATA1* knock-down (KD) perturbation-specific signature was identified in every cell using Mixscape (23). This model uses a mixture of two Gaussian distributions (one for cells in which the KD was effective, and another for cells that remained unperturbed, either because they received a non-targeting sgRNA or because the perturbation was not effective). A Wilcoxon Rank Sum test was used to identify differentially expressed genes between *GATA1* KD and non-perturbed cells. The set of differentially expressed genes upon *GATA1* KD was defined as those with Bonferroni-adjusted p-value < 0.001.

ChIP-seq enrichment

We obtained K562 ChIP data from the ENCODE consortium (24). TF targets using ChIP-seq data were defined as those genes whose promoter or enhancer regions had high-quality (ENCODE peak calling q-val<0.01) ChIP binding signal for a given TF (enhancers were linked to genes using the Activity-By-Contact model (25)).

For the enrichment analysis, we only considered TFs that had at least 15 non-TF putative targets (defined as the intersection of correlated genes with p-value < 0.05 in all time points), since those with too few putative targets yield noisy enrichment estimates. We then used Fisher's Exact Test for count data (R stats package) to test the null of independence of rows and columns in a contingency table that contained putative targets from both methods, from only one of them, or for none, and computed odds ratios and confidence intervals for each TF. An Odds Ratio > 1 indicates that the ratio of correlated genes with the TF bound to the TF unbound is higher than for uncorrelated genes.

Supplementary Figures

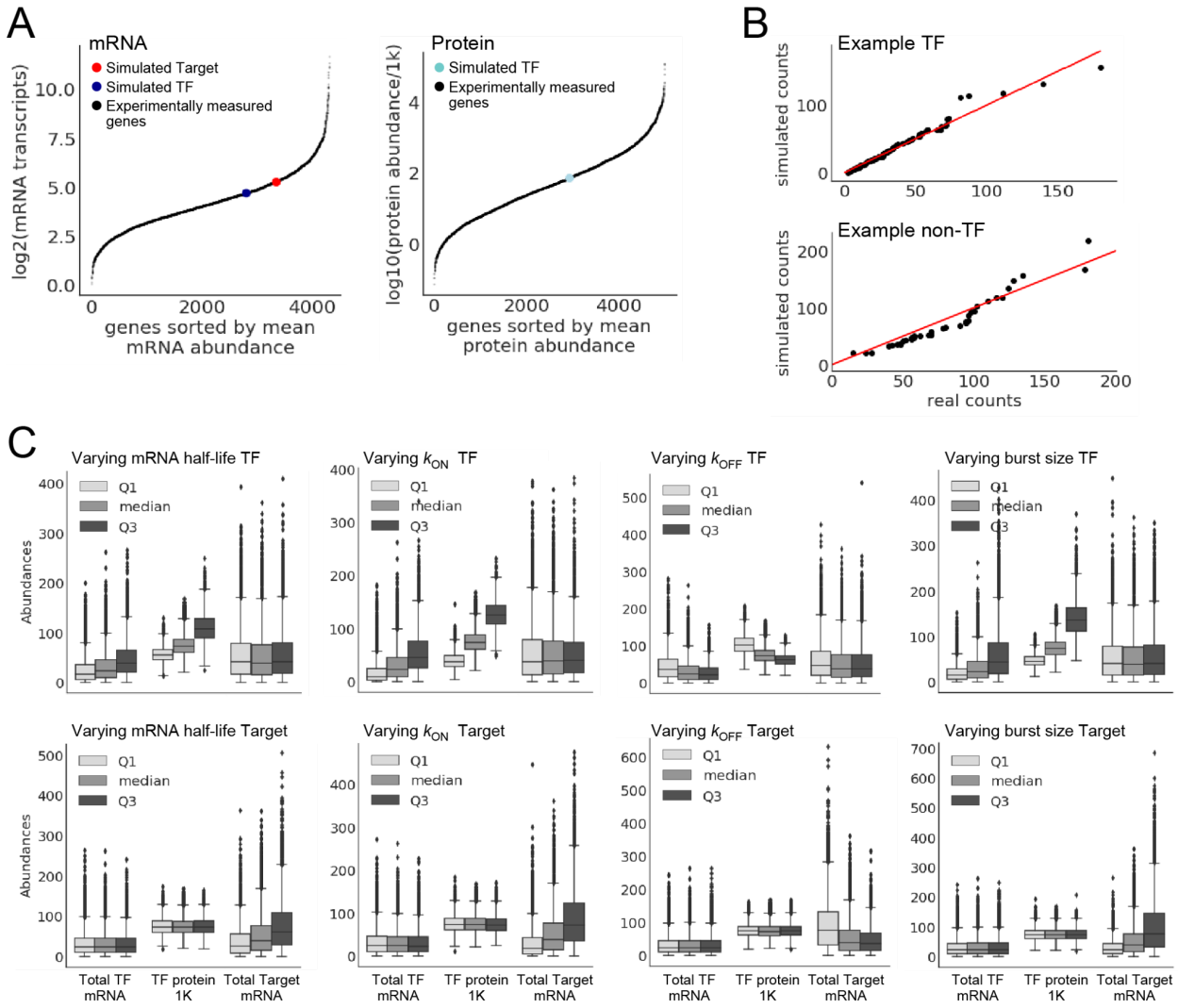

Figure S1. Checks for simulated mRNA and protein steady-state distributions. (*A*) Mean (left) mRNA abundances of the simulated TF and Target genes and (right) simulated TF protein abundance, in comparison to the range of mean abundances for 4,309 genes and 5,000 proteins, respectively, measured experimentally by Schwanhausser et al., *Science*, 2011. Experimentally measured genes are sorted by their increasing abundances, which are reported in $\log_2$(transcripts) or $\log_2$(protein molecules/1000), respectively. (*B*) Sorted mRNA copy numbers taken from published experimental data (*NANOG*, a sample TF taken from smFISH data in Ochiai et al., *Scientific Reports*, 2014 and *TNFR1*, a sample non-TF gene taken from Lin et al., *Nat Comm*, 2019). Each dot indicates the copy number for that gene in a single cell. Red line indicates where the ranked cells would fall if they had the same mRNA copy numbers between simulated and real cells. (*C*) Abundance distributions of key entities (total TF mRNA, total TF protein in 1,000 molecules, and total Target mRNA) are shown across a population of 20,000 simulated cells at steady-state, as each of the intrinsic parameters is varied (separately for TF and non-TF). Parameter quartiles are derived from literature (see Table 1).

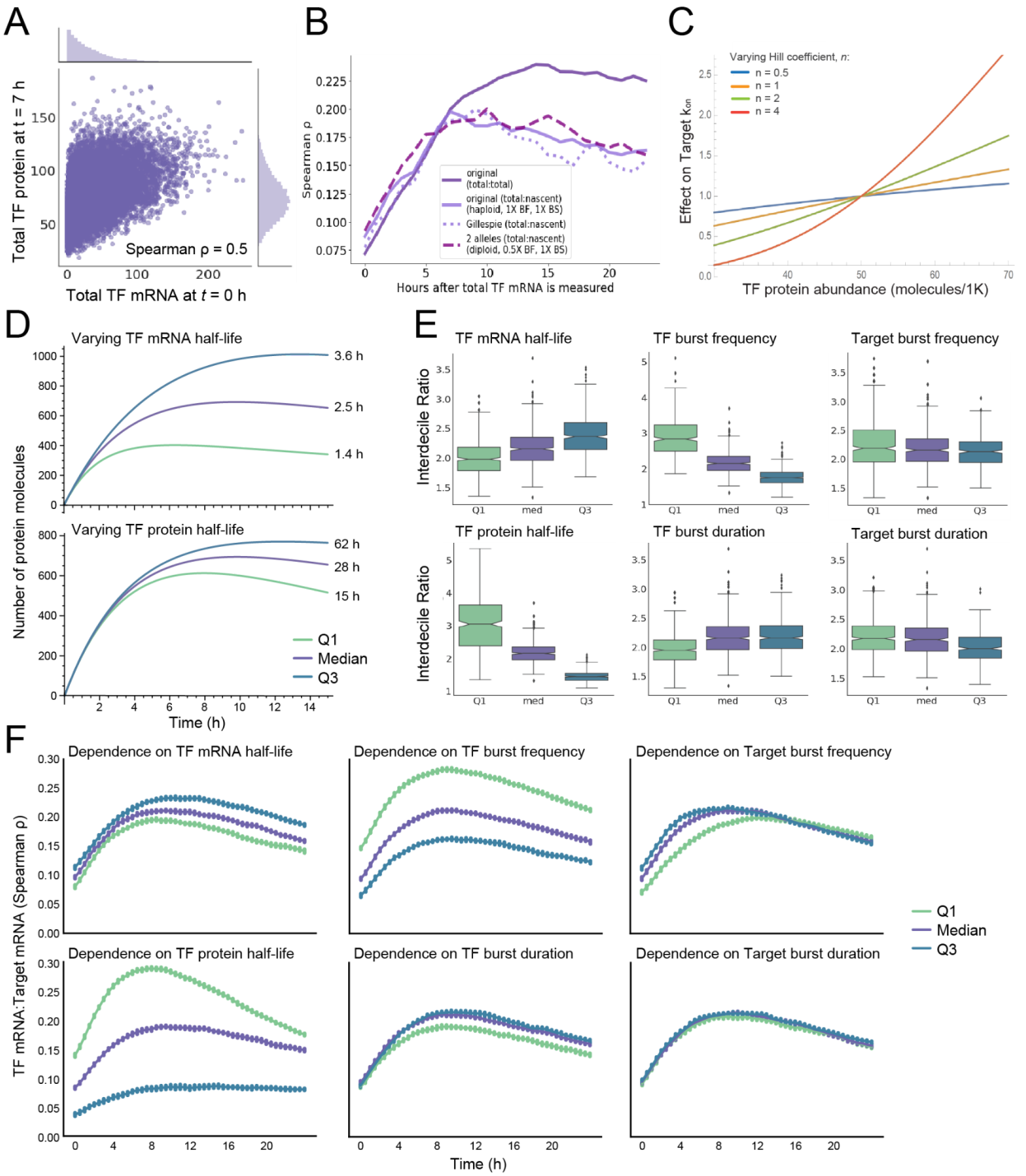

Figure S2. Time-dependency of TF:Target regulatory effect. (*A*) Joint distribution of $TF^{(RNA)}_0$ and $TF^{(P)}_6$ (in 1K molecules). (*B*) Spearman's ρ of C_T (the time-shifted $TF^{(RNA)}_0$:$Target^{(RNA)}_T$) and C_T^{Δ} ($TF^{(RNA)}_0$:$Target^{(\Delta RNA)}_T$). Curves for our simulations for the diploid case with two alleles, as well as Gillespie-based stochastic simulations, are also included. (*C*) Effect of varying the Hill coefficient on how $Target^{(kon)}$ changes as $TF^{(P)}$ changes. (*D*) Effect of varying mRNA (top) and protein (bottom) half-lives (in hours) on the time when protein abundance best reflects mRNA abundance at t=0. (*E*) Effect of varying individual burst parameters (first quartile, median, and third quartile) on the maximum estimated, time-shifted $TF^{(RNA)}_0$:$Target^{(\Delta RNA)}_T$ D10:D1 IR (across all sampling time points over 20 hours). Twenty five simulations were run for each burst parameter value, each with 20k cells. (*F*) Effect of varying six TF- and Target-specific parameter values (first quartile, median, and third quartile) on the shape of the regulatory response curve, as estimated by the time-shifted Spearman's rank correlation C_T^{Δ} between TF mRNA and Target nascent RNA

abundances (across all sampling time points over 24 hours). Twenty five simulations were run for each burst parameter value, each with 20k cells (mean +/- 1 SD is plotted per time point).

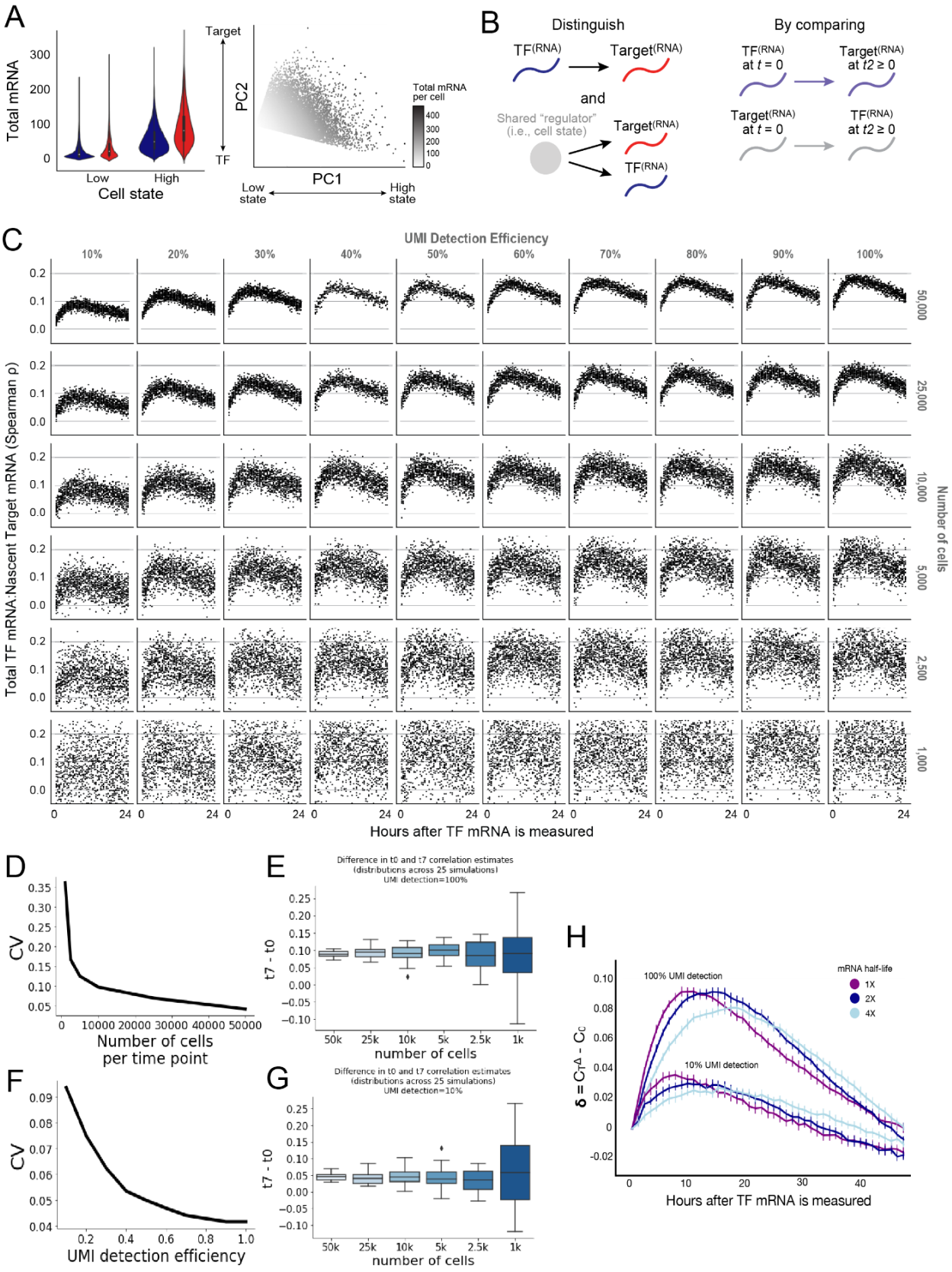
A
Total mRNA
Low
High
Cell state
Target
TF
PC2
PC1
Low state
High state
Total mRNA per cell
B
Distinguish
By comparing
and
Shared "regulator" (i.e., cell state)
C
UMI Detection Efficiency
Total TF mRNA:Nascent Target mRNA (Spearman ρ)
Number of cells
Hours after TF mRNA is measured
D
CV
Number of cells per time point
E
Difference in t0 and t7 correlation estimates (distributions across 25 simulations)
UMI detection=100%
t7 - t0
number of cells
F
CV
UMI detection efficiency
G
Difference in t0 and t7 correlation estimates (distributions across 25 simulations)
UMI detection=10%
t7 - t0
number of cells
H
100% UMI detection
10% UMI detection
mRNA half-life
1X
2X
4X
Hours after TF mRNA is measured

Figure S3. Effect of downsampling and cell state on detected, time-shifted TF:Target covariation. (*A*) Two simulated cell states with varying mRNA abundances for the two genes. (*B*) We can distinguish cell state-based covariation from that due to gene regulation by comparing (1) the time-shifted $TF^{(RNA)}$:$Target^{(RNA)}$ Spearman's rho and (2) the reverse scenario of time-shifted $Target^{(RNA)}$:$TF^{(RNA)}$ Spearman's rho. In a situation with cell states, both will be constant and high over time; in the case of gene regulation, the forward will peak then decrease, and the reverse will be closer to zero. (*C*) Spearman's rho between a TF and its target gene across time as we vary UMI detection efficiencies (simulated via Poisson downsampling) and sample sizes (number of cells per time point). 25 simulations were run for each combination of detection efficiency, sample size, and time point. Comparison made at each time point is C_T^{Δ} ($TF^{(RNA)}_0$:$Target^{(\Delta RNA)}_T$). (*D*) Coefficient of variation (standard deviation/mean) of the time-shifted $TF^{(RNA)}$:$Target^{(\Delta RNA)}$ Spearman's rho C_T^{Δ} across the 25 simulation runs, as the number of cells sampled changes. (*E*) Difference in C_7^{Δ} and C_0^{Δ} (the maximum difference for genes with median gene-specific parameter values) across simulation runs, as the number of cells sampled values. UMI detection efficiency = 100%. Error bars indicate 95% CIs. (*F*) Coefficient of variation (standard deviation/mean) of the time-shifted $TF^{(RNA)}$:$Target^{(\Delta RNA)}$ Spearman's rho C_T^{Δ} across the 25 simulation runs, as the UMI detection efficiency changes. We note that the lower threshold of 0.04 indicates the lowest CV possible with 50,000 cells. (*G*) Same as (S4*E*), but UMI detection efficiency = 10%. (*H*) δ (C_T^{Δ}-C_0^{Δ}) over time, as the TF and Target mRNA half-lives are doubled and quadrupled, for both 100% and 10% UMI detection efficiencies. 25 simulations were run, each with 20,000 cells, to obtain these curves (error bars indicate +/- 1 standard deviation).

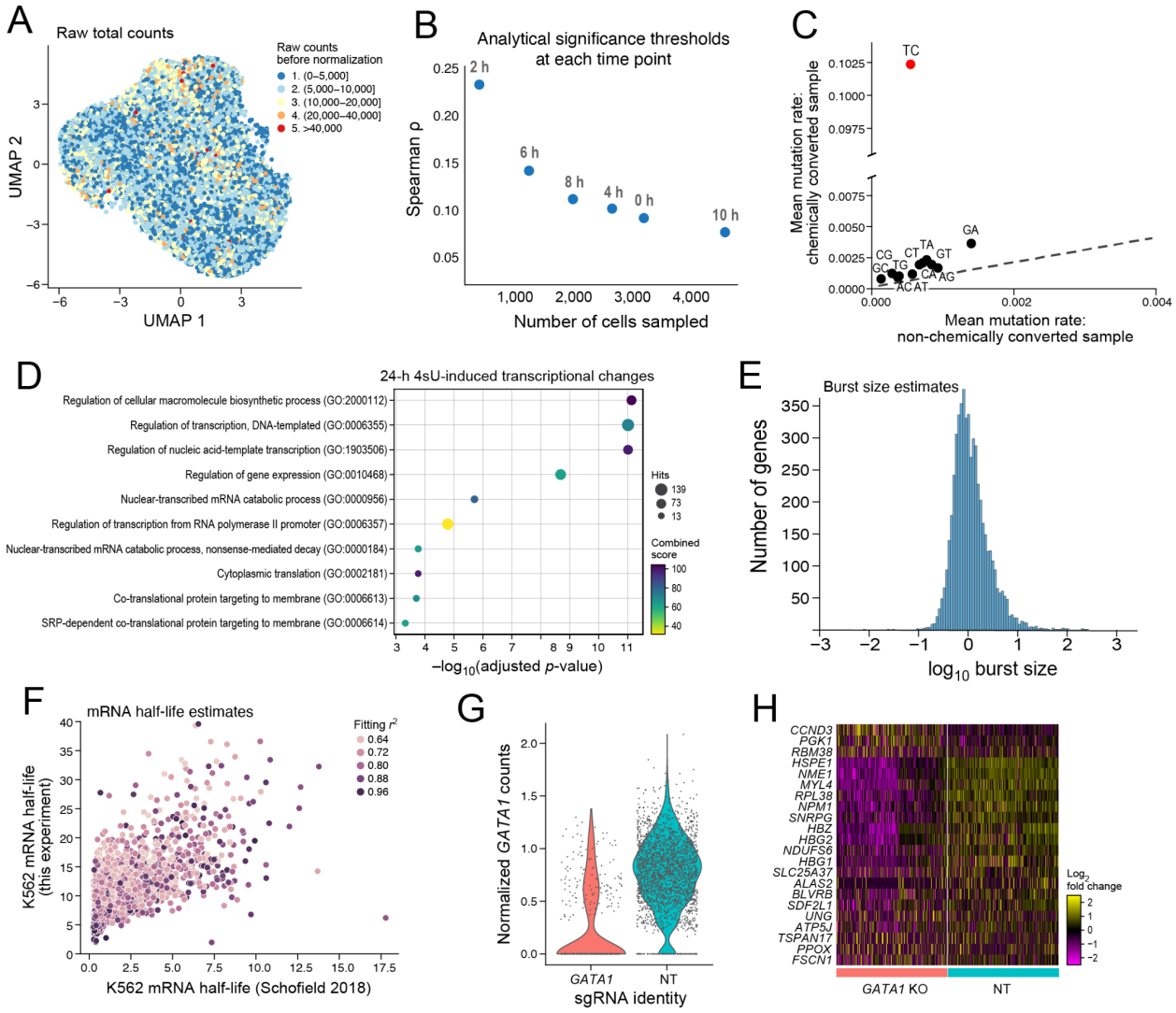

Figure S4. Learnings from a time-resolved experiment in K562 cells at steady-state. (*A*) UMAP of 14,544 unperturbed K562 single cells in one of the replicate experiments, colored by raw total counts. (*B*) Number of cells sampled at each time point, and the resulting analytical significance threshold for Spearman's correlation. (*C*) 40:1 signal to noise ratio in the T to C mutation rate of the chemically converted samples relative to the non-converted controls as a result of the addition of 4-thiouridine (4sU) (*Materials and Methods*). (*D*) Gene Ontology terms enriched upon 24h of 4sU induction. (*E*) Estimated burst sizes from the time-resolved data computed over the integral of gene ΔRNA over the chase phase (i.e. unlabeled counts) (*Materials and Methods*). (*F*) Estimated mRNA half-lives (in hours) in our data (y-axis) and prior estimates using bulk time-resolved RNA-seq (Schofield, Nature Methods, 2018). The color scale shows the r^2 fitting of the fraction of labeled transcripts for each gene to an exponential decay model (*Materials and Methods*). (*G*) Normalized *GATA1* expression in cells with the *GATA1* targeting sgRNA and with non-targeting (NT) sgRNAs from a published Perturb-seq dataset (Replogle et al., 2020). (*H*) Heatmap of the expression levels in *GATA1* knock-down and non-targeting (NT) cells of the genes with the highest correlation coefficients from stochastic regulation. The columns in the heatmap display expression levels in 300 individual cells.

Table S1. Effect of changing gene-intrinsic parameter values on the timing and magnitude of maximum time-shifted TF:Target Spearman's rho C_T^{Δ}.

	Timing of maximum covariation (hours)		**Maximum Spearman's rho**		**Maximum D10:D1 Interdecile Ratio**	
	25th percentile	75th percentile	25th percentile	75th percentile	25th percentile	75th percentile
TF mRNA half-life	5	9	0.18	0.25	2	2.5
TF protein half-life	5	9	0.3	0.07	3	1.5
TF burst frequency	7	7	0.28	0.15	3	1.8
TF burst duration	7	7	0.16	0.21	2	2.3
Target burst frequency	12	4	0.18	0.2	2.2	2.1
Target burst duration	7	7	0.19	0.19	2.2	2

Supplementary References

1. K. Fujita, M. Iwaki, T. Yanagida, Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA. *Nature Communications* **7** (2016).
2. T. Lionnet, R. H. Singer, Transcription goes digital. *EMBO Rep* **13**, 313-321 (2012).
3. B. Munsky, G. Neuert, A. van Oudenaarden, Using gene expression noise to understand gene regulation. *Science* **336**, 183-187 (2012).
4. A. Sanchez, S. Choubey, J. Kondev, Stochastic models of transcription: from single molecules to single cells. *Methods* **62**, 13-25 (2013).
5. R. D. Dar *et al.*, Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc Natl Acad Sci U S A* **109**, 17454-17459 (2012).
6. 51. J. Estrada, F. Wong, A. DePace, J. Gunawardena, Information Integration and Energy Expenditure in Gene Regulation. *Cell* **166**, 234-244 (2016).
7. 52. A. Hafner *et al.*, Quantifying the Central Dogma in the p53 Pathway in Live Single Cells. *Cell Systems* **10**, 495-505.e494 (2020).
8. 53. S. K. Hortsch, A. Kremling, Characterization of noise in multistable genetic circuits reveals ways to modulate heterogeneity. *PLoS One* **13**, e0194779 (2018).
9. 54. Y. Jiang, N. R. Zhang, M. Li, SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biology* **18** (2017).
10. 55. C. Li, F. Cesbron, M. Oehler, M. Brunner, T. Hofer, Frequency Modulation of Transcriptional Bursting Enables Sensitive and Rapid Gene Regulation. *Cell Syst* **6**, 409-423 e411 (2018).
11. 56. M. Razo-Mejia *et al.*, Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. *Cell Systems* **6**, 456-469.e410 (2018).
12. 57. A. Senecal *et al.*, Transcription Factors Modulate c-Fos Transcriptional Bursts. *Cell Reports* **8**, 75-83 (2014).
13. 58. Y. Wang, T. Ni, W. Wang, F. Liu, Gene transcription in bursting: a unified mode for realizing accuracy and stochasticity. *Biol Rev Camb Philos Soc* (2018).
14. J. A. Schofield, E. E. Duffy, L. Kiefer, M. C. Sullivan, M. D. Simon, TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nature Methods* **15**, 221-225 (2018).
15. Q. Qiu *et al.*, Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nat Methods* **17**, 991-1001 (2020).
16. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
17. A. J. M. Larsson *et al.*, Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251-254 (2019).
18. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
19. S. A. Lambert *et al.*, The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
20. A. T. Lun, D. J. McCarthy, J. C. Marioni, A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).
21. J. M. Replogle *et al.*, Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol* **38**, 954-961 (2020).
22. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
23. E. Papalexi *et al.*, Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet* **53**, 322-331 (2021).
24. C. A. Sloan *et al.*, ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**, D726-732 (2016).
25. C. P. Fulco *et al.*, Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664-1669 (2019).
26. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).