PNAS RESEARCH ARTICLE | SYSTEMS BIOLOGY OPEN ACCESS

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], and Eric S. Lander[a,j,k,2,3,4]

Contributed by Eric S. Lander; received May 9, 2022; accepted July 20, 2022; reviewed by Arjun Raj and Alexander van Oudenaarden

Regulatory relationships between transcription factors (TFs) and their target genes lie at the heart of cellular identity and function; however, uncovering these relationships is often labor-intensive and requires perturbations. Here, we propose a principled framework to systematically infer gene regulation for all TFs simultaneously in cells at steady state by leveraging the intrinsic variation in the transcriptional abundance across single cells. Through modeling and simulations, we characterize how transcriptional bursts of a TF gene are propagated to its target genes, including the expected ranges of time delay and magnitude of maximum covariation. We distinguish these temporal trends from the time-invariant covariation arising from cell states, and we delineate the experimental and technical requirements for leveraging these small but meaningful cofluctuations in the presence of measurement noise. While current technology does not yet allow adequate power for definitively detecting regulatory relationships for all TFs simultaneously in cells at steady state, we investigate a small-scale dataset to inform future experimental design. This study supports the potential value of mapping regulatory connections through stochastic variation, and it motivates further technological development to achieve its full potential.

transcriptional bursting | gene regulation | single-cell transcriptomics

Significance

Deciphering gene regulatory networks can help elucidate the molecular underpinnings that define cellular identity and disease processes. Current approaches to discover regulation compare different cell types or employ cellular perturbations. We show that, with enough data, it should be possible to identify regulatory relationships within a cell type without need for perturbation, by leveraging the intrinsic stochasticity in transcriptional bursting across individual cells at steady-state. Importantly, time-shifted correlations in RNA expression make it possible to distinguish covariation due to regulatory relationships within a cell state from covariation due to undetected cell states. Here, we present a theoretical framework for this approach and discuss future experimental design.

Systematically identifying gene regulatory networks in any specific cellular context has been a long-standing quest that remains unfulfilled. Existing approaches to link transcription factors (TFs) to their regulatory targets based on gene expression data include tracking gene coexpression across different cell states (1–4), perturbing TFs experimentally and measuring transcriptomic changes (5–8), and studying coordinated changes in genes' expression over dynamic biological processes (9–11). Single-cell RNA sequencing (scRNA-seq) technologies have made it possible to apply these approaches to study gene expression at unprecedented resolution (11–24).

However, systematic interrogation of human gene–gene regulatory interactions in steady-state cellular systems has been limited by the need for targeted experimental perturbation of specific genes (14, 25–27). The ability to learn gene regulation from unperturbed cells would provide a scalable approach applicable to any cell type and state, providing novel insights into the transcriptional programs that shape steady-state cellular identity.

Because all genes are transcribed in stochastic bursts (28–35), even isogenic cells in the same cellular state have substantial variability in gene expression across time (36–41). Cells thus carry out their own natural "perturbation" experiments, as the accumulation or depletion of recent bursts leads to varied messenger RNA (mRNA) abundances for each gene. These natural perturbations have the potential to help reveal gene regulation (42–45). Specifically, a transient increase in the transcription of a TF in a given cell generates a subsequent increase in the abundance of the TF protein, which in turn leads to an increase in mRNA for the target genes of the TF (provided the TF protein is localized to the nucleus).

In this paper, we demonstrate how stochastic fluctuations across individual cells in a single cell state can give rise to time-shifted covariation between the level of TFs and their target genes. Using this information, we lay out a theory for inferring regulatory relationships between TFs and target genes in cells at steady state, based on the time-shifted covariation between TF mRNA levels and target gene mRNA levels; the approach can be used to study simultaneously all TFs and potential targets.

Specifically, we use published ranges for gene-specific parameters to characterize the shape of the time-shifted correlation curve resulting from regulatory relationships. As we discuss, it is important to examine time-shifted correlations rather than just simultaneous correlations between the level of gene pairs at the same time point—because the

Competing interest statement: In advance of review, the authors noted that one of the authors (E.S.L.) and one of the reviewers (A.v.O.) were both among more than 80 coauthors on a community white paper describing plans for a Human Cell Atlas that was posted on eLife in December 2017 (https://elifesciences.org/articles/27041). PNAS determined that this connection was tangential and did not constitute recent scientific collaboration relevant to the review process.

[1]A.G. and J.D.M. contributed equally to this work.

[2]B.C., S.G., and E.S.L. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: lander@broadinstitute.org.

[4]Currently on leave from the Broad Institute of MIT and Harvard.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2207392119/-/DCSupplemental.

Published August 15, 2022.

latter can arise from the undetected presence of multiple cell states, whereas the former should not. Finally, we explore practical considerations in designing future experiments to detect gene regulation by measuring time-shifted correlations, including the sample sizes and improvements in mRNA detection efficiency that will be required.

Results

We first lay out the conceptual framework for the paper. Consider a transcription factor gene, TF, and a possible target gene, Target. Because gene expression occurs in bursts, the number of transcripts of TF and Target will fluctuate over time. (Throughout, $\text{Gene}^{(\text{RNA})}{}_{\text{T}}$ and $\text{Gene}^{(\Delta\text{RNA})}{}_{\text{T}}$ will denote the amount of a gene's total and nascent RNA at time t = T. Nascent RNA refers to unspliced, pre-mRNA.)

The key insight is that if TF regulates Target in cells of a given cell state, then the time-shifted correlation with either nascent or total Target RNA—that is, $C_T = \text{Corr}(\text{TF}^{(\text{RNA})}{}_0, \text{Target}^{\text{RNA}}{}_T)$ and $C_T^{\Delta} = \text{Corr}(\text{TF}^{(\text{RNA})}{}_0, \text{Target}^{\Delta\text{RNA}}{}_T)$—is expected to increase and then decrease, as T increases from 0 (Fig. 1 *A* and *B*). The reason is simple: Higher levels of $\text{TF}^{(\text{RNA})}$ at t = 0 give rise over time to higher levels of $\text{TF}^{(\text{Protein})}$ and then to higher levels of $\text{Target}^{(\Delta\text{RNA})}$ and $\text{Target}^{(\text{RNA})}$, resulting in higher maximum time-shifted correlations, until the relationship eventually breaks down as the initial $\text{TF}^{(\text{RNA})}$ degrades.

The temporal pattern is important because it allows one to distinguish correlation arising due to gene regulation from correlation arising from an important alternative possibility. Specifically, correlated gene expression across cells can also arise if the cells are an undetected mixture of different cell types. However, such correlations will not show the expected pattern of increase and decrease in time-shifted correlation C_T (Fig. 1 *C* and *D*).

We now turn to a more quantitative treatment, to study the properties of the time-shifted correlation and determine the feasibility of detecting it experimentally.

Transcriptional Bursting in Cells at Steady State Gives Rise to Temporal Covariation between TFs and Their Target Genes. We built a quantitative model that describes the processes linking the key quantities for a regulatory gene pair involving a TF and a target gene. The model builds on the traditional model of transcriptional bursting, by extending it to include translation of the TF mRNA and gene regulation of a target gene. The key quantities we track are total TF mRNA abundance [$\text{TF}^{(\text{RNA})}$], measured in number of transcripts; total TF protein abundance [$\text{TF}^{(\text{P})}$], measured in thousands of molecules; and target-gene mRNA abundance, assessed as either total RNA [$\text{Target}^{(\text{RNA})}$] or nascent RNA [$\text{Target}^{(\Delta\text{RNA})}$] (Fig. 2*A* and *Materials and Methods*). For each quantity, we can derive from our model the average abundance level, the distribution of abundance levels across cells, and the cross-correlation and auto-correlation between these quantities across cells and time. These values each depend on gene-specific kinetic parameters, which we describe below.

Transcriptional bursting can be abstracted into a two-state model, whereby transcription of a gene stochastically switches between transcriptional "on" and "off" states (46–50). A gene's burst frequency (k_{on}), burst size, and mRNA decay rate determine its cellular mRNA abundance, which—along with the translation and protein decay rates—determines the corresponding protein abundance over time (*SI Appendix*, Eqs. **1–3** and **5**). Following transcription from a burst, each nascent pre-mRNA transcript

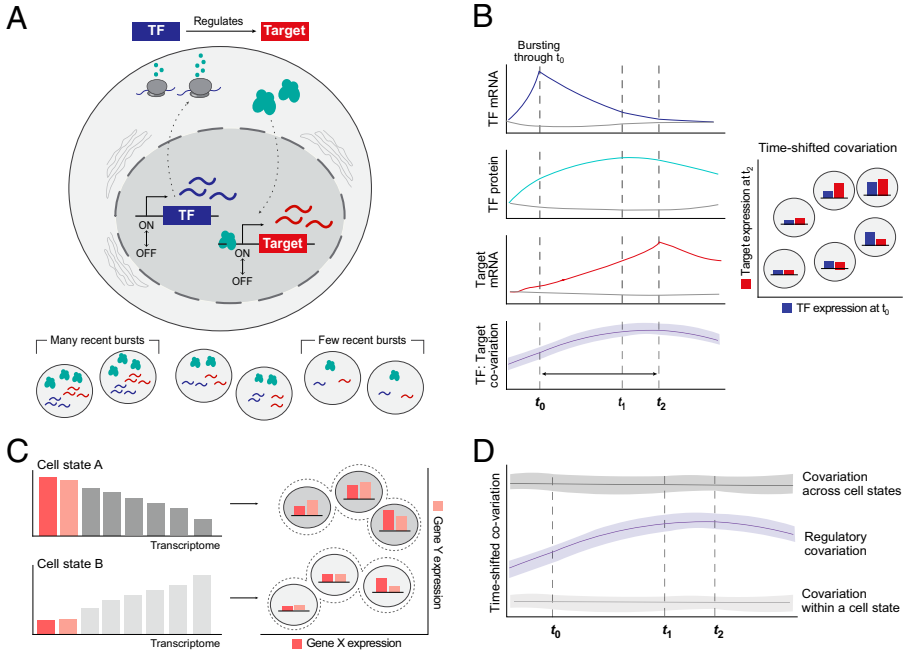


Fig. 1. Overview of the conceptual framework for inferring TF:Target gene regulation from single cells at steady state. (*A*) Transcriptional bursting leads to stochastic variation in the mRNA abundance of each gene, even within a population of isogenic cells at steady state. We invoke stochastic transcriptional bursting as a source of TF mRNA heterogeneity across steady-state cell populations. If a TF directly regulates a target gene, we hypothesize that their abundances will be correlated. (*B*) Idealized representation of the hypothesized time-shifted correlation between a TF and its target gene's mRNA abundances in the presence of regulation. Colored lines indicate the average behavior of cells that had at least one burst of the TF gene; gray lines represent those that did not have a burst. The time delay reflects the time required for TF mRNA translation into protein, translocation, and target site search in the nucleus. From left to right, dotted lines reflect the time of maximal TF mRNA (t_0), TF protein (t_1), or target mRNA abundance (t_2), respectively. (*C*) Subpopulations of cells (i.e., cell states, such as cells in different stages of the cell cycle) will also give rise to covariation—in this example, due to different baseline mRNA counts for genes in each state. Thus, correlation does not always imply regulation. (*D*) We can theoretically distinguish between regulation- and state-based covariation by looking at the shape over time: State-based covariation will tend to be more stable.

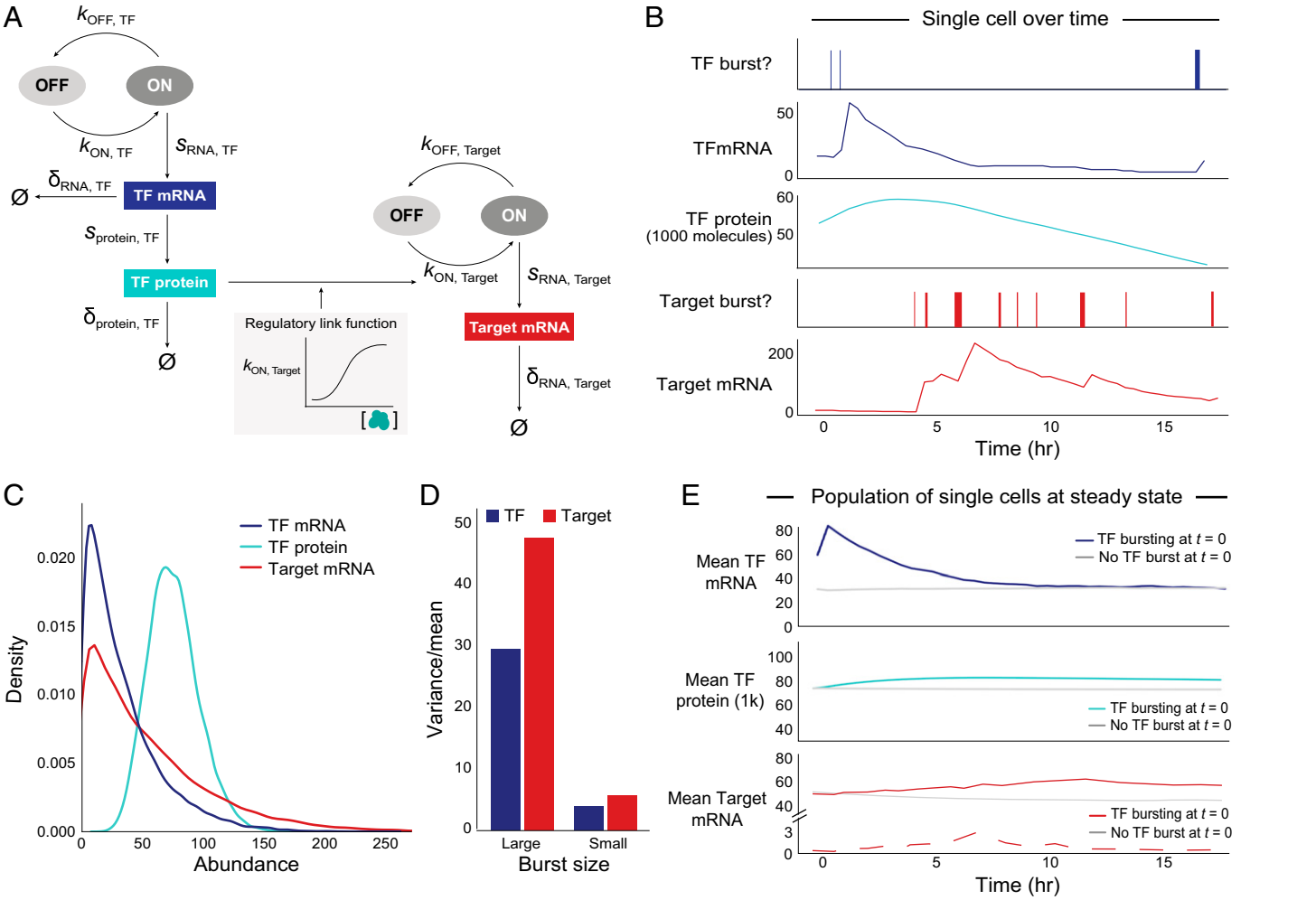


Fig. 2. Transcriptional bursting yields intrinsic variation for each gene across cells at steady state. (*A*) Two-state model of transcriptional bursting and regulation for one Regulator–Target pair. Variables: k_{ON} (burst frequency), k_{OFF} (1/burst duration), s_{RNA} (transcription = burst size*k_{OFF}), $s_{protein}$ (translation), δ (decay), Ø (no molecules left). Blue: TF, red: Target. (*B*) Simulation of $TF^{(RNA)}$, $TF^{(P)}$, and $Target^{(RNA)}$ bursting events and abundance for one cell in the presence of direct regulation between a pair of genes. (*C*) Abundance distributions of $TF^{(RNA)}$ (μ: 29, CV: 0.84), $TF^{(P)}$ in thousands (μ: 61, CV: 0.31), and $Target^{(RNA)}$ (μ: 50, CV: 0.93), for median values of burst parameters, across 20,000 simulated cells. (*D*) Overdispersion structure (variance/mean) of total mRNA for TF and Target for different burst sizes (large TF: 32 transcripts/burst, large Target: 40 transcripts/burst; small TF: 3, small Target: 4). (*E*) Mean $TF^{(RNA)}$ (*Top*), $TF^{(P)}$ (*Middle*), and $Target^{(RNA)}$ (*Bottom*) abundance over time, across cells that did have a burst of the TF gene (colored, n = 715) versus those that did not have a burst (gray, n = 715 randomly subsampled cells with no burst) at t = 0 (20,000 total cells). Solid red line indicates $TF^{(RNA)}$ and dashed red line indicates $TF^{(\Delta RNA)}$. Curves are based on data points at 30 min intervals. (TF mRNA at 30 min shows a sharp peak due to discrete sampling; the actual peak is smooth.)

encoding a TF is spliced into a mature transcript on the order of <10 min (51–53), after which it is translocated into the cytoplasm and translated, also on the order of minutes (54).

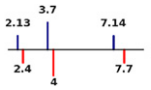
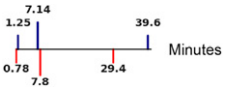
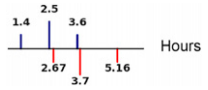
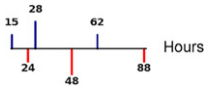
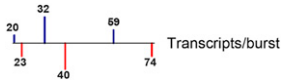
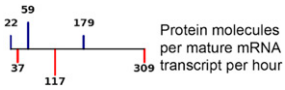
While modeling the processes of transcription and translation is straightforward, modeling the impact of changes in TF protein levels on the expression of target genes is more complicated. TF protein levels are generally considered to affect target genes by modulating their bursting frequency (although this is not strictly true in all cases; see refs. 55–64). We modeled the regulation of a target gene by a TF by adjusting the target gene's rate of burst initiation in a manner dependent on the concentration of the TF protein (*SI Appendix*, Eq. 7). Here, the derivative of the response function [i.e., the instantaneous $\Delta Target^{(kon)}/\Delta TF^{(P)}$] describes the proportional change in bursting frequency relative to the change in $TF^{(P)}$. We chose a Hill function to reflect the effect of changes in $TF^{(P)}$ on $Target^{(kon)}$, although the specific shape of the response curve can also be captured by other models (e.g., Michaelis–Menten).

We note that small fluctuations in $TF^{(P)}$ levels will only affect target gene transcription when $TF^{(P)}$ is neither extremely low nor near saturating concentrations. For instance, if a given target gene's promoter is already occupied with many molecules of a particular TF protein, then a small increase in that protein will likely have minimal to no effect on the target gene's transcription. In this work, we focus on regulatory relationships that are "meaningful" within a cell type, by which we mean that $TF^{(P)}$ is near the steep part of the response curve. Since the coefficient of variation for $TF^{(P)}$ across a population of cells at steady state is relatively small (~0.3; Fig. 2*C*), the TF protein abundance for any given cell is not likely to stray too far from the mean. In these cases, the relationship between $TF^{(P)}$ levels to $Target^{(kon)}$ is effectively linear, with a slope defined by the Hill coefficient n (*SI Appendix*, Fig. S2*C*). (The slope of the Hill coefficient roughly captures the degree of cooperativity in the effect of TF molecules on gene regulation.)

We simulated a TF regulating a target gene in a steady-state population of 20,000 cells, continuously tracking $TF^{(RNA)}$, $TF^{(P)}$, $Target^{(RNA)}$, and $Target^{(\Delta RNA)}$ for 20 h in each cell (*Materials and Methods*).

For parameters related to transcription and translation, we selected values based on published data in human and mouse cells (40, 55, 65–70) as reported in Table 1.

Table 1. Range of intrinsic parameter values for TFs and non-TF genes

Parameter	Interquartile ranges — TF — Non-TF	Refs.
Time between bursts ($1/k_{ON}$)		67, 70
Burst duration ($1/k_{OFF}$)		67, 70
mRNA half-life		65, 69, 81
Protein half-life		69
Burst size		55, 67, 74
Translation rate		65, 69

Derived from literature, where parameters were estimated and inferred from published experimental data. Lines represent first quartile, median, and third quartile values for TFs (blue) and non-TFs (red), respectively. Burst sizes used in simulations were based on previously published smFISH-based copy numbers, although we note that scRNA-seq data-based estimates are lower due to transcript capture inefficiencies (88–90). k_{on} = burst frequency; k_{off} = number of burst ends per hour.

For most gene-specific parameters, the estimates are consistent across studies. However, certain studies (e.g., refs. 71 and 72) suggest that mRNA half-lives may be longer than other estimates by two- to fourfold (65, 70). To be conservative, we used the shorter estimates of half-life in our simulations. We return to this point below, where we discuss the implications of longer half-lives for inferring gene regulation.

For the Hill coefficient of the regulatory response function (connecting changes in TF protein levels to target gene transcription), the value can range from zero (no regulatory effect) to a large positive number (with higher coefficients corresponding to a stronger effect). In our model, we chose a modest value of 2, representing the situation where doubling $TF^{(P)}$ results in a fourfold increase in the target gene's burst frequency [$Target^{(kon)}$], in bursts per hour (*SI Appendix*, Fig. S2*C*). With these parameters, the distributions of $TF^{(RNA)}$, $TF^{(P)}$, and $Target^{(RNA)}$ across cells are consistent with published, experimental measurements (*SI Appendix*, Fig. S1 *A* and *B*) (69, 73, 74).

While our simulations assume a single copy of the TF and target genes, we confirmed that simulations with two alleles at both genes, with each allele having half the bursting rate, produce essentially equivalent results (as expected, because bursting events are infrequent and thus the two alleles are rarely bursting at the same time) (*SI Appendix*, Fig. S2*B*). We also confirmed that our simulations produce equivalent results to using Gillespie stochastic simulations for our model (*SI Appendix*, Fig. S2*B*) (75).

Applying our model to follow a collection of single cells over time revealed the compounded effects of bursting and decay on each gene's mRNA abundance fluctuations (Fig. 2*B*). While much of the variation in $TF^{(RNA)}$ is buffered at the protein level due to longer protein half-lives, there is an approximately threefold difference in mean TF protein abundance between the top and bottom deciles of $TF^{(P)}$ (Fig. 2*C*), which is robust to changes in gene-specific parameter values over relevant ranges (*SI Appendix*, Fig. S1*C*). Importantly, the production of mRNA molecules in rapid succession due to bursts results in the overdispersion (variance/mean >1) of mRNA expression, with the amount of mRNA overdispersion directly proportional to the gene's burst size (Fig. 2*D* and *SI Appendix*, Eq. 4).

If we partition cells according to whether or not the TF gene was bursting at time t = 0, we find that cells that were bursting at t = 0 have substantially higher levels of $TF^{(RNA)}$, $TF^{(P)}$, and $Target^{(RNA)}$ over the subsequent hours than cells that were not bursting at t = 0; the difference eventually disappears as the levels converge back to the population mean (Fig. 2*E*). [As seen in the figure, we note that cells bursting at t = 0 already have higher levels of $TF^{(RNA)}$ at t = 0, because the burst will have typically begun before t = 0.] The temporal dynamics for each of the quantities is notably different:

1) $TF^{(RNA)}$ rises steeply for nearly 30 minutes, by which time the burst occurring at t = 0 will have ended and mRNA decay will have taken over.
2) $TF^{(P)}$ rises more gradually as the TF mRNA is steadily translated. In cells that were bursting at t = 0, $TF^{(P)}$ reaches a maximum increase of ∼12% increase at ∼6 h. Elevated levels of $TF^{(P)}$ persist for many hours, because $TF^{(P)}$ decays much more slowly than $TF^{(RNA)}$ (median TF protein half-life = 28 h).
3) $Target^{(\Delta RNA)}$ and $Target^{(RNA)}$ both rise at similar rates; however $Target^{(\Delta RNA)}$ peaks and drops off at ∼7 h, whereas $Target^{(RNA)}$ continues rising gradually for another > 5 h (approximately two mRNA half-lives). This difference reflects the short lifespan of unspliced transcripts. $Target^{(\Delta RNA)}$ thus provides a sensitive measure of the instantaneous regulatory effect of $TF^{(P)}$, while $Target^{(RNA)}$ reflects the accumulation of mRNA over time.

Propagation of Stochastic Variation from TF to Target Genes. We next analyzed in detail how the variance in a TF's mRNA level is propagated to a target gene's mRNA level through translation and then the TF protein's modulation of the target gene's transcriptional bursting (Fig. 3*A*).

The amount of $TF^{(P)}$ in a cell produced from $TF^{(RNA)}$ present at t = 0 can be expressed by a system of ordinary differential equations governed by translation and degradation rates (*SI Appendix*, Eq. **5**). $TF^{(P)}$ peaks after a time delay, because protein molecules continue to accumulate until the rate of new TF proteins produced by the remaining mRNA that was present at t = 0 is offset by the slow decay rate of the protein. For the parameters in our model, the $TF^{(RNA)}$:$TF^{(P)}$ correlation is initially ∼0.35 at t = 0 [consistent with experimental measurements in single cells (76)] and subsequently rises to a peak of 0.5 at around t = 6 h (*SI Appendix*, Fig. S2*A*), before gradually falling again over the subsequent 10+ h, based on the TF protein's half-life (Fig. 3*B*).

If we compare cells that were in the top decile (D10) and bottom decline (D1) of $TF^{(RNA)}$ at t = 0, those that were in the top decile have nearly double the level of $TF^{(P)}$ at the time of maximum correlation at ∼6 h (D10:D1 interdecile ratio [IR] = 1.7; Fig. 3*C*), showing how variance is propagated from mRNA to protein.

In addition, cells with high levels of $TF^{(P)}$ have substantially higher levels of nascent target-gene mRNA ($Target^{(\Delta RNA)}$), which closely reflects the instantaneous $Target^{(kon)}$. For our model (regulatory relationship with Hill coefficient = 2), comparison of cells in the top versus bottom decile of $TF^{(P)}$ (whose IR = 2.8) reveals that the former have approximately sixfold

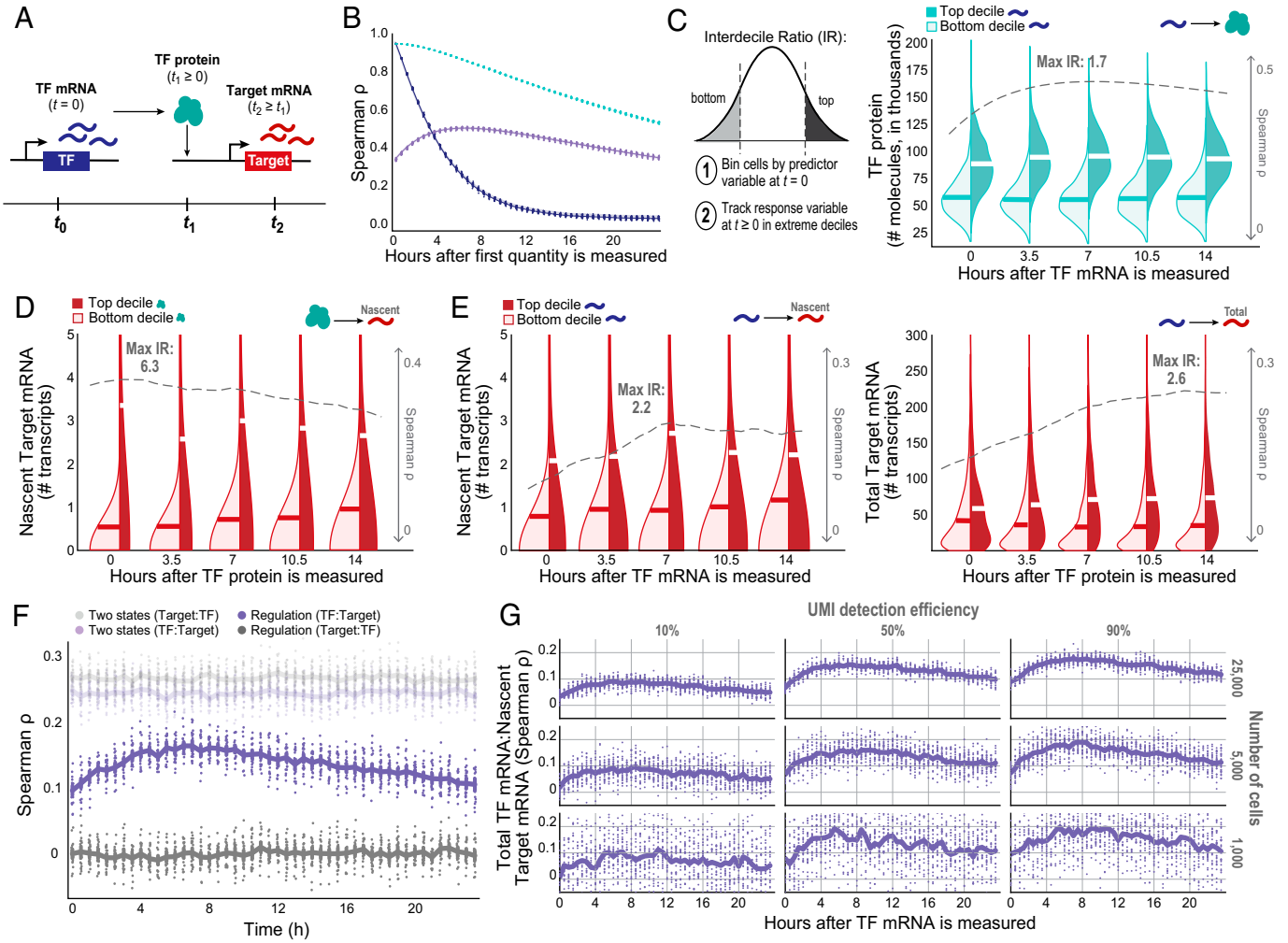


Fig. 3. Information flow between a TF and its target gene is time-dependent. (*A*) Schematic of the multistep inference question. (*B*) $TF^{(RNA)}$ and $TF^{(P)}$ Spearman's ρ autocorrelations and correlations between $TF^{(RNA)}$ and $TF^{(P)}$ over time, across 25 simulation runs (each included as its own dot). (*C*–*E*) Kernel density estimates comparing the extremes of distributions across cells at t ≥ 0, binned by the predictor molecule abundance at t = 0. Dashed gray line indicates Spearman's ρ, and "max IR" indicates maximum D10:D1 IR between top- and bottom-binned cells (alternative hypothesis: top decile of the dependent variable is greater than the bottom decile) value at each time point, with the highest magnitude effect listed. (*C*) $TF^{(RNA)}_0$:$TF^{(P)}_T$ correlation and D10:D1 IR of $TF^{(P)}$ distribution extremes at t ≥ 0 for cells binned by $TF^{(RNA)}$ at t = 0. (*D*) $TF^{(P)}_0$:$Target^{(kon)}_T$ correlation and D10:D1 IR of adjusted $Target^{(\Delta RNA)}$ distribution extremes for cells binned by $TF^{(P)}$ at t = 0, under the Hill function model of interaction. (*E*) Relying on mRNA only—C_T^{Δ} and $TF^{(RNA)}$-binned $Target^{(\Delta RNA)}$ D10:D1 IR (*Left*) and C_T and $Target^{(RNA)}$ D10:D1 IR (*Right*)—to infer regulation over time. (*F*) Correlation between a TF and its putative Target could be a result of cell-state-based structure; if so, the time-shifted correlation would have a stable magnitude over time. (*G*) Effect of down-sampling the number of cells and/or UMI detection efficiency on estimated TF:Target covariation trends over time (focusing on C_T^{Δ}). Estimates get both noisier and lower in magnitude due to these two technical considerations.

higher levels of $Target^{(\Delta RNA)}$ [maximum Spearman's ρ = 0.32 between $TF^{(P)}$ and $Target^{(\Delta RNA)}$; Fig. 3*D*].

In principle, to compare TF and Target RNA levels, we can look at either C_T or C_T^{Δ}; the former reaches higher magnitudes (e.g., a peak of 0.23 versus 0.18) but changes more gradually, while the latter peaks earlier and falls more sharply. We measured both of these time-shifted correlations for $0 \leq T \leq 15$ and focused on C_T^{Δ} for our analyses. The peak C_T^{Δ} occurs at around 7 h, not long after the peak $TF^{(RNA)}$ correlation with $TF^{(P)}$ (Fig. 3*E*, *Left*). The same is true for the D10:D1 IR of $Target^{(\Delta RNA)}$ (*SI Appendix*, Fig. S2*D*). This makes sense, because the level of nascent transcripts reflects current transcription rates and thus is closely related to TF protein levels. In contrast, C_T continues to rise for several more hours (Fig. 3*E*, *Right* and *SI Appendix*, Fig. S2*B*), because transcription continues to occur and total mRNA continues to accumulate.

Sensitivity of Time-Shifted Correlations to Gene-Intrinsic Parameters. We explored the sensitivity of the magnitude and timing of peak correlations to the choice of parameters in our model, by replacing them with the 25th and 75th percentile values (Table 1).

Broadly, the time-shifted correlation C_T^{Δ} is robust to variation in most variables, although the magnitude is more sensitive to changes in TF protein half-life and TF burst frequency (*SI Appendix*, Fig. S2*F*).

With respect to the time of maximum correlation, increasing the half-life of TF mRNA or TF protein from 25th to 75th percentile delays the time from 5 to 9 h (as expected from *SI Appendix*, Eq. **6**), whereas increasing $Target^{(kon)}$ shortens the time effect from 12 to 4 h (*SI Appendix*, Fig. S2 *D* and *F* and Table S1).

With respect to the magnitude of the correlation C_T^{Δ}, increasing the TF mRNA half-life increases $TF^{(1/koff)}$ (burst duration), which in turn increases the covariation by up to 25%, whereas increasing the TF protein half-life or $TF^{(kon)}$ decreases the magnitude by up to 60% (*SI Appendix*, Fig. S2 *E* and *F* and Table S1).

These results suggest that experiments to detect regulatory covariation should sample time points across the range of 0

to >12 h in order to capture the start, peak, and decay of the time-shifted correlation curves for most genes.

We also explored the sensitivity of the results to changing the Hill coefficient (*SI Appendix*, Fig. S2*C*). Whereas the Hill coefficient of 2 used in our model results in a maximum Spearman's ρ for $TF^{(RNA)}$:$Target^{(\Delta RNA)}$ of ∼0.2, larger values increase the magnitude of the covariation (e.g., 0.33 with a Hill coefficient of 4) while lower values decrease the magnitude of the covariation (0.13 with a Hill coefficient of 1, corresponding to no cooperativity).

Distinguishing Cell State-Based Structure from Regulatory Covariation. While we have focused on how TF:Target regulation gives rise to a characteristic time-shifted correlation C_T^{Δ}, it should be noted that such regulation also gives rise to simultaneous correlations—that is, when the quantities are measured at the same time point. Specifically. $TF^{(RNA)}$ and $TF^{(P)}$, as well as $TF^{(RNA)}$ and $Target^{(RNA\ or\ \Delta RNA)}$, show positive correlation even when the quantities are measured at t = 0 (Fig. 3 *B, C,* and *E*; Spearman's ρ = 0.35 and 0.1, respectively). This is due to the fact that TF mRNA resulting from earlier bursts of transcription persists long enough to overlap target gene expression resulting from those bursts.

Given that TF:Target regulation results in simultaneous correlation in between TF and Target mRNA levels at any given time t, one might ask why bother looking at time-shifted correlation. After all, simultaneous correlation is easier to study, as it can be measured in a simple scRNA-seq experiment. The reason is that, while simultaneous correlation can arise as a consequence of gene regulation, it can also arise from the undetected presence of multiple stable cell states. In drawing inferences about gene regulation, it is important to rule out the latter possibility.

To illustrate the issue, we compared the time-shifted correlation C_T that arises from gene regulation versus a mixture of two stable cell "states" (*Materials and Methods* and *SI Appendix*, Fig. S3*A*). Moreover, we examined both the standard "forward" time-shifted correlation studied above $C_T^{\Delta} = Corr(TF_{t=0}^{(RNA)}: Target_{t \geq 0}^{(\Delta RNA)})$ and, as a negative control, a "reverse" time-shifted correlation $Corr(Target^{(\Delta RNA)}{}_0:TF^{(RNA)}{}_T)$, which assesses whether the Target regulates the transcription of the TF gene (*SI Appendix*, Fig. S3*B*).

As expected, the mixture of two stable cell states gives rise to a time-shifted correlation that is roughly constant over time and shows similar results in the forward and reverse directions—in sharp contrast to the results seen for time-shifted correlation arising from gene regulation (Fig. 3*F* and *SI Appendix*, Fig. S3*B*).

Notably, it is easier to distinguish the time-shifted correlation arising from gene regulation from the time-shifted correlation arising from a mixture of cell states when focusing on C_T^{Δ} rather than C_T. This is because C_T^{Δ} shows an earlier, sharper peak that is more readily distinguished from the stability of state-based correlations. In contrast, C_T changes more gradually, since the preexisting target-gene mRNA dilutes the acute regulatory effect of $TF^{(P)}$ at t = 0.

Finally, we note that "stable cell states," as used above, refers to states that persist for the period over which time-shifted correlations are measured—e.g., at least one cell division. [Stable states thus include both permanent states and meso-stable states (77, 78)]. States that persist for much shorter periods of a few hours are best regarded as "transient states"; they might arise from rapid processes, such as cell cycle progression or temporary activation of neuronal or immune cells. To distinguish transient states from regulation, it will be useful to know gene expression patterns associated with these rapid processes (as are available for cell cycle progression).

From Theory to Experimental Design. The results above show it should be possible to systematically learn gene regulatory connections from stochastic variation in steady-state cells, provided one can measure transcriptome-wide mRNA levels at pairs of time points in individual cells with sufficient sensitivity and power. We now turn to how one might implement the theory in practice.

Several high-throughput technologies exist for using pulse–chase labeling to distinguish mRNAs synthesized at different time points (68, 71, 72, 79–81). One approach involves 1) pulse-labeling cells with 4-thiouridine (4sU, a uridine analog that incorporates into RNA) to mark transcripts synthesized during an initial interval (t_1, $t_1 + \delta$), 2) chasing the label with uridine, and 3) harvesting and profiling the cells at a later time t_3 using single-cell RNA-seq approaches. By performing chemical conversion of 4sU into cytosine analogs before performing scRNA-seq, one can use the presence of U-to-C substitutions to identify transcripts synthesized during (t_1, $t_1 + \delta$). This approach can be used to study both total mRNA and nascent RNA, via intronic sequences.

Suppose we wish to detect an increase δ in the time-shifted correlation—for example, a doubling from t = 0 to t = 7 (Fig. 3*G*). In this case, $\delta = C_7^{\Delta} - C_0^{\Delta}$. The power to detect this increase depends on the number of cells studied and the sensitivity to detect transcripts.

First, consider the case of infinitely many cells and perfect transcript detection α = 1 (that is, 100% sensitivity, such that every transcript is represented by a unique molecular identifier [UMI]). There will be no variance in independent correlation estimates. Thus, δ can be estimated perfectly: The coefficient of variation (CV_δ, defined as SD divided by the mean of δ) is zero.

Next, consider the case of a finite number *n* of cells with perfect transcript detection α = 1. In this case, one must overcome only biological variation across cells. Suppose the true mean abundances are E(X) = 29 and E(Y) = 50 transcripts (Fig. 2*C*). Simulations show that the time-shifted correlation doubles, from $C_0^{\Delta} = 0.09$ to $C_7^{\Delta} = 0.18$ (*SI Appendix*, Fig. S3*C*), with δ = 0.09 and $CV_\delta \approx 0.36/\sqrt{(n/1{,}000)}$, corresponding to $CV_\delta \approx 0.17$ for n = 2,500 and 0.04 for n = 50,000 (*SI Appendix*, Fig. S3*D*). To estimate δ with an SD equal to 25% of the mean (a reasonable level of precision) requires roughly 5,000 cells (*SI Appendix*, Fig. S3*E*).

Now, consider the case of a finite number of cells with incomplete transcript detection α < 1 (Fig. 3*G*). In this case, one must overcome both biological variation and measurement noise. For example, current scRNA-seq methods detect only a small fraction (∼10 to 15%) of the number of mRNA transcripts present in each cell (82), owing to incomplete capture of transcripts(73, 83) and incomplete sequencing of captured transcripts (83–91). If we can detect only 10% of transcripts in the preceding example, we have E(X) = 2.9 and E(Y) = 5.0. Simulations show that the time-shifted correlations are now smaller (roughly, $C_0^{\Delta} = 0.04$ and $C_7^{\Delta} = 0.08$) (*SI Appendix*, Fig. S3*C*) and the CV_δ is larger (by 2.24-fold) (*SI Appendix*, Fig. S3*F*). To estimate δ with $CV_\delta = 0.25$ requires many more cells—roughly 50,000 cells (*SI Appendix*, Fig. S3*G*).

Finally, we noted above that some papers have suggested mRNA half-lives may be two- to fourfold longer than previously reported values, which were used in our simulations. Increasing the half-life (while maintaining the mean level) of both the TF and Target mRNA by two- or fourfold increases

the value and time of the maximum time-shifted correlation C_T, but it has minimal effect on δ or on the number of cells required (*SI Appendix*, Fig. S3*H*).

In summary, to reliably detect relevant increases in time-shifted correlation, it will be necessary to improve current methods for time-resolved labeling by increasing 1) transcript detection efficiency, 2) the number of cells that it is feasible to study, or 3) both. For example, a detection efficiency of 50% would allow the use of 17,500 cells per time point for TF and Target mRNA half-lives of 2.5 and 3.6 h, respectively. Such improvements seem achievable but will require focused efforts.

Leveraging Simultaneous Correlations in a Single-Cell Dataset to Enrich for Gene Regulation. We noted earlier that regulatory relationships give rise to two types of correlation between $TF^{(RNA)}$ and $Target^{(RNA)}$: 1) a positive correlation C_0 between the quantities when measured simultaneously [because $TF^{(RNA)}$ from past bursts persists long enough in cells to overlap $Target^{(RNA)}$ due to those bursts] and 2) an increase δ in the time-shifted correlation after an appropriate time interval. Simultaneous correlation has the disadvantage that it could also arise from undetected cell types, whereas the time-shifted correlation approach controls for this possibility.

Because the time-shifted correlation approach is not yet feasible with current labeling methods, we undertook an experiment to examine genes that show simultaneous correlation with a TF, to look for other evidence that some of the correlations may be due to regulation. Using a droplet-based method, scNT-sEq. (71), we generated a temporally resolved scRNA-seq dataset with K562 cells labeled with 4sU for a 24-h pulse phase, followed by a 10-h chase phase with uridine, during which we sampled cells every 2 h (Fig. 4*A*, *SI Appendix*, Fig. S4 *A–D*, and *Materials and Methods*). We collected ∼1,000 to 4,000 cells at each of six time points (total of 13,679 cells) (*SI Appendix*, Fig. S4*B*). We confirmed that burst sizes (*SI Appendix*, Fig. S4*E*) and mRNA half-lives (*SI Appendix*, Fig. S4*F*) estimated from the K562 data are consistent with recent, published estimates from scRNA-seq data (55, 71, 72, 80), including the longer estimates of mRNA half-lives.

We started by considering *GATA1*, a well-known erythroid regulator, in these steady-state cells (Fig. 4*B*) and looking for genes whose mRNA levels showed significant correlation with *GATA1* mRNA levels across cells (Spearman's ρ $P < 0.05$; *Materials and Methods*). Specifically, we analyzed each time point separately and looked only at those genes that showed significant correlation with *GATA1* at all six time points. A total of 36 genes met these criteria. Based on two lines of evidence, many of the correlations seem likely to reflect regulation by *GATA1*: 1) The genes are enriched (3.8-fold) for genes differentially expressed (92) in an independent *GATA1* Perturb-seq knockdown experiment (Fig. 4*C*, *SI Appendix*, Fig. S4*G*, and *Materials and Methods*) and 2) 22 of the genes are well-known *GATA1* targets (see refs. 93–97 for a few prominent examples), of which 19 are significantly down-regulated and 3 significantly up-regulated upon *GATA1* knockdown (*SI Appendix*, Fig. S4*H*).

We next calculated mRNA correlations for 342 TFs, finding a total of 17,152 TF:non-TF pairs that showed significant correlations at all six time points (mean number of significant links per TF = 50). We focused on 56 TFs with at least 15 significantly correlated genes whose genome-wide binding sites had been measured in K562 cells using chromatin immunoprecipitation and sequencing (ChIP-seq) in the ENCODE project (98). For each such TF, we identified putative binding sites

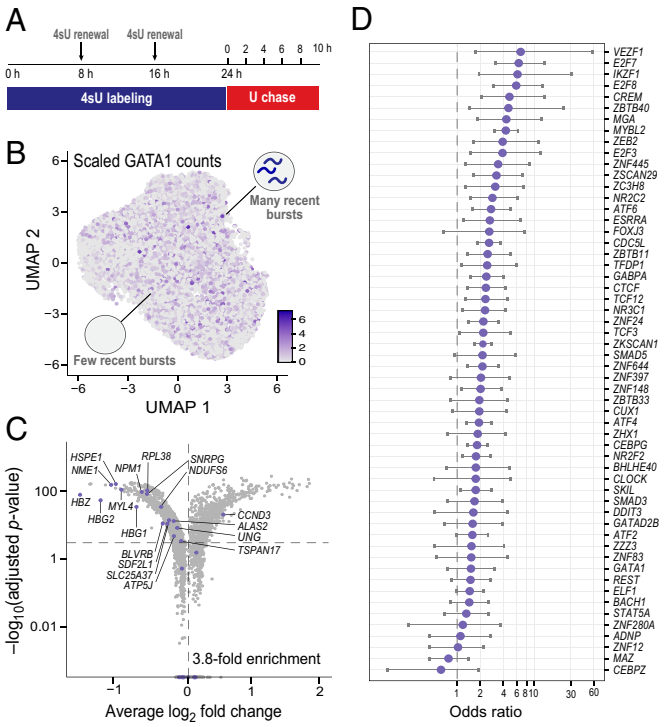


Fig. 4. Enrichment of gene regulatory signal from simultaneous correlations in scRNA-seq data of K562 cells at steady state. (*A*) Schematic of the experimental design of a pulse–chase metabolic labeling experiment to capture two temporally resolved snapshots of RNA abundance in the same single cells. U, uridine. (*B*) UMAP of 13,679 unperturbed K562 single cells across six time points (∼1,000 to 4,000 cells per time point), colored by *GATA1* scaled counts. (*C*) Differentially expressed genes upon *GATA1* knockdown, inferred from an orthogonal *GATA1* knockdown Perturb-seq experiment in K562 cells (92). The horizontal dotted line represents a Bonferroni-adjusted *P* value threshold of 0.001, and the vertical one a $\log_2$ fold change of 0. Purple dots denote the set of correlated genes with $P < 0.05$ at all time points, which have a 3.8-fold enrichment. (*D*) Enrichment of predicted TF binding from simultaneous $Corr(TF^{(RNA)}_T:non\text{-}TF^{(RNA)}_T)$ correlations for ChIP-seq binding signal across 56 TFs with ENCODE ChIP-seq data that have at least 15 significantly correlated genes from the K562 data.

based on stringent ChIP-seq peaks; we then assigned these sites to a specific gene based on the Activity-By-Contact model (ref. 99 and *Materials and Methods*). We found that 53/56 (95%) of the TFs were enriched for binding in enhancers and promoters assigned to the significantly correlated genes relative to all other genes expressed in K562 cells (Fig. 4*D*).

These results suggest that many of the genes that show simultaneous correlations are likely to be regulatory targets, although time-resolved experiments will be necessary to rule out alternative possibilities. If confirmed, this framework would provide a method for identifying regulatory connections in steady-state cells.

Discussion

The ability to infer gene regulation from observations of cells at steady state, without experimental perturbations, would vastly expand the study of regulatory networks in any cell type—providing insight into previously inaccessible realms of biology. Here, we introduce a framework to do so that uses the stochasticity of transcription to identify, in a principled way, all pairs of covarying TF and target genes in cells at steady state.

Our model, which is grounded in experimentally-derived parameters, captures a distinctive, time-shifted correlation curve between the abundance of the mRNA of a TF and nascent RNA of a target gene, which rises and then falls, at time scales

reflecting the half-lives of the TF mRNA and TF protein. While we focus on C_T^{Δ}, we show that C_T (using total Target mRNA) results in a similar curve that increases and decreases more gradually. Importantly, these temporal relationships distinguish correlation due to gene regulation from correlation due to the undetected presence of multiple cell states.

Simulations indicate that identifying regulatory pairs should be feasible, in principle, for genes within typical gene-specific parameter ranges. The key challenge is that current experimental methods lack adequate power. For example, given current transcription detection efficiencies of ∼10%, detecting a two-fold increase in the time-shifted correlation would require ∼50,000 cells sampled per time point, which is not currently practical. We quantify the effects of improving transcript detection and sample sizes on the ability to infer gene regulation.

While simultaneous correlation does not allow us to conclusively distinguish between gene regulation and undetected cell states, we investigated simultaneous correlation between TFs and other genes in K562 cells at steady state. Independent results from Perturb-seq and ChIP-seq strongly suggest that many of the significantly correlated pairs represent true gene regulation. These results support the idea that pairwise correlations of gene mRNA abundances in cells at steady state can highlight potential instances of gene regulation, which can then be tested via patterns of time-shifted correlation.

A reviewer raised an interesting question that is worth addressing: Given that the variation in TF protein levels across cells should cause a positive correlation in bursting between the two alleles of a gene within a cell, why do studies often find no such allelic correlation (100, 101)? The answer is that there is indeed an allelic correlation, but its magnitude is tiny (typically, in the range of 0.003) given the variation of TF protein levels and the regulatory response in our model and thus would be difficult to detect.

Our results have various limitations. Estimates of kinetic parameters from scRNA-seq data remain imperfect. Our simulations assume a regulatory effect size (Hill coefficient) of 2; extremely weak regulatory relationships in a cell type [i.e., corresponding to very high or very low concentrations of $TF^{(P)}$] will not be detected but likely constitute less meaningful biology in the cell type. Our model does not currently include such factors as nuclear buffering of mRNA (102), TF autoregulation (76, 103, 104), and posttranslational modifications (105). Such factors could be incorporated, but we believe that the current model provides sufficient guidance for initial experimental designs.

While our work focuses on scRNA-seq–based transcriptomic correlations, technologies to profile complementary aspects of cells broaden the spectrum of informative approaches. Joint protein and mRNA measurements in single cells (106) can enable simultaneous measurement of $TF^{(P)}$ with the $Target^{(\Delta RNA)}$. Additionally, pairing intronic gene abundances with spatial information (107) could leverage the spatial organization of the nascent transcriptome to yield more nuanced regulatory insights.

Given the critical roles of gene regulatory networks in cell types, we anticipate that the ideas presented here—coupled with improvements in high-throughput single-cell technologies—could provide powerful approaches for understanding biology.

Materials and Methods

Modeling Transcriptional Bursting and Gene Regulation. The underlying stochastic model (*SI Appendix*, Eqs. **1** and **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 (Table 1). We used a Poisson decay model to enable discrete removal of mRNA transcripts and protein molecules (*SI Appendix*, Eq. **3**). We also incorporated splicing and translation from the literature in a TF- and non-TF-specific manner (Table 1 and *SI Appendix*, Eqs. **4**–**6**). Transcription rates were 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. As a check, we confirmed that the resulting mRNA and protein abundances match those estimated by deterministic ordinary differential equations (*SI Appendix*, Eq. **5**). We modeled the $TF^{(P)}$:$Target^{(kon)}$ response function as a Hill function (*SI Appendix*, Eq. **7**), with a Hill coefficient of 2.

Simulating Two-Gene Regulation in a Population of Cells at Steady State. We wrote a simulation (Python v3.7) based on the two-state stochastic bursting model of gene transcription above and simulated two genes, one TF (regulator) and its target gene, across 20,000 cells. At each time step, we tracked $TF^{(\Delta RNA)}$, $TF^{(RNA)}$, $TF^{(P)}$, $Target^{(\Delta RNA)}$, $Target^{(RNA)}$, and $Target^{(kon)}$, to reflect the abundance changes between each time point. To assess regulatory signal, we calculated the following time-shifted RNA:RNA Spearman's ρ 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 h after the burn-in time to up to 2 d after. We have made our simulator code, including code to run the simulator with various parameters, and a notebook with the key analyses included in this paper, available on GitHub (https://github.com/agupta-landerlab/stochastic-regulation-code).

Gene-Specific Parameter Sensitivity Analyses. We measured the sensitivity of C_T^{Δ} to gene-specific parameters, with each taking on either the 25th, 50th, or 75th percentile value, taken from the literature (Table 1). We ran 25 independent simulations of 20,000 cells each. In each run, we estimated the regulatory effect using the IR measure (ratio of the mean abundance of the top decile of cells to the mean of the bottom decile of cells) as well as C_T^{Δ}. We plugged in various values of mRNA and protein half-lives into *SI Appendix*, Eq. **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 by removing the link function between $TF^{(P)}$ and $Target^{(kon)}$ and creating one "state" with low mRNA abundances and another with high abundances for each gene. Specifically, we varied the basal burst frequency for both the TF and Target to be either their first quartile or third quartile values, depending on the state. 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.

Assessing the Effect of Sequencing Inefficiencies and Sample Size. 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 down-sampling, with the Poisson rate parameter equal to the number of counts*capture_efficiency. We chose the minimum between the raw and down-sampled 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 simulated 25 runs for each pair of capture efficiencies and sample sizes.

Pulse–Chase Experiment with 4sU. K562 erythroleukemia cells (ATCC, CCL-243) were cultured in RMPI 1640 + L-glutamine, with 10% fetal bovine serum, 1% penicillin/streptomycin, and 1% L-glutamine 200 mM. For 4sU experiments, cells were plated in six-well plates ∼12 h before the start the experiment at a density of 5×10^5 to 8×10^5 cells per mL in 5 mL of fresh media. The 4sU (Sigma, T4509-25MG) in dimethyl sulfoxide (DMSO) was added to culture wells at a final concentration of 100 μM for 24 h, with renewal every 6 to 8 h. Between the pulse and chase phases, the cells were washed twice to remove any residual traces of 4sU. Subsequently, media with saturating concentrations of uridine (Sigma, U6381) at 10 mM was added. Cells were collected at 0, 2, 4, 6, 8, and 10 h. 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 and to calculate gene-specific burst sizes and mRNA half-lives. Our scNT-seq protocol was adapted from previously published methods (71–73), with modifications described in *SI Appendix*. Details on raw sequencing data demultiplexing, alignment, and doublet detection can also be found in *SI Appendix*. These scRNA-seq data were deposited to Gene Expression Omnibus (GEO), with accession number GSE202292. We have made code related to the analysis of these real data available on GitHub (https://github.com/agupta-landerlab/stochastic-regulation-code).

Gene-Specific Parameter Estimation from Real Data (mRNA Half-Lives and Burst Sizes). To estimate gene-specific half-lives from real data, we fit the fraction of labeled counts for each gene at each time point to an exponential decay model (*SI Appendix*). 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 ref. 71.

Identification of Significantly Correlated TF:non-TF Pairs Using Simultaneous Correlation Analysis. To determine the list of TFs and non-TF genes, we used a recently published, curated list of human TFs (66). We filtered out cells with low UMI and genes expressed in few cells across time points. We scaled raw counts to account for differences in library size. Correlated genes for each TF were identified based on genes whose expression profiles had a Spearman's rank correlation with the TF's expression that fell in the right tail of the null distribution (defined by a permutation test to randomize ranks). 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.

Enrichment of *GATA1* Correlated Genes in Perturb-seq Data. We compared the genes significantly correlated with *GATA1* with a recently published dataset that combined CRISPR interference on *GATA1* with a single-cell RNA sequencing readout (92). A guide-cell barcode dictionary with the identity of the perturbation in each cell was obtained from GSE132080. We used a Wilcoxon rank sum test to identify differentially expressed genes between *GATA1* KD and nonperturbed cells (Bonferroni-adjusted $P < 0.001$).

ChIP-seq Enrichment for 56 TFs. We obtained K562 ChIP data from the ENCODE consortium (98). We linked enhancers to genes using the Activity-By-Contact model (99). For 56 TFs with at least 15 significantly correlated genes, we used Fisher's exact test to check for enrichment of TF binding: 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.

Data, Materials, and Software Availability. scRNA-seq data have been deposited in Gene Expression Omnibus (GSE202292) (108). Code related to the analysis of these data is available on GitHub (https://github.com/agupta-landerlab/stochastic-regulation-code).

ACKNOWLEDGMENTS. We thank Galit Lahav, Aviv Regev, Rajet Vatsa, Yakir Reshef, Dylan Kotliar, Peter Kharchenko, Jesse Engreitz, Dig Bijay Mahat, and Caroline Uhler for critical discussions regarding this work and Leslie Gaffney for help with graphics. This work was supported by funds from the Broad Institute. Jorge Diego Martin Rufino is supported by a fellowship from La Caixa Foundation (ID 100010434).

Author affiliations: [a]Broad Institute of MIT and Harvard, Cambridge, MA 02142; [b]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115; [c]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA 02115; [d]Dana-Farber Cancer Institute, Boston, MA 02215; [e]Whitehead Institute for Biomedical Research, Cambridge, MA 02142; [f]HHMI, Massachusetts Institute of Technology, Cambridge, MA 02139; [g]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; [h]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114; [i]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA 02115; [j]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [k]Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Author contributions: A.G., J.D.M., T.R.J., A.B., V.G.S., B.C., S.G., and E.S.L. designed research; A.G., J.D.M., T.R.J., E.I.G., A.M., K.Z., V.G.S., and S.G. performed research; X.Q., C.W., K.H.M., and A.A.K. contributed new reagents/analytic tools; A.G., J.D.M., T.R.J., V.S., S.N., L.S., V.G.S., S.R., B.C., S.G., and E.S.L. analyzed data; A.G., J.D.M., B.C., S.G., and E.S.L. wrote the paper; A.G. performed modeling, cowrote simulations, and ran simulated and real data analyses; and J.D.M. performed the K562 experiment and ran real data analyses.

Reviewers: A.R., University of Pennsylvania; and A.v.O., Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences.

1. A. J. Butte, I. S. Kohane, Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* **2000**, 418–429 (2000).
2. S. L. Carter, C. M. Brechbühler, M. Griffin, A. T. Bond, Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**, 2242–2250 (2004).
3. J. M. Stuart, E. Segal, D. Koller, S. K. Kim, A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
4. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17 (2005).
5. Z. Hu, P. J. Killion, V. R. Iyer, Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.* **39**, 683–687 (2007).
6. P. Kemmeren *et al.*, Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**, 740–752 (2014).
7. T. L. Lenstra *et al.*, The specificity and topology of chromatin interaction pathways in yeast. *Mol. Cell* **42**, 536–549 (2011).
8. S. van Wageningen *et al.*, Functional overlap and regulatory links shape genetic interactions between signaling pathways. *Cell* **143**, 991–1004 (2010).
9. Z. Bar-Joseph, A. Gitter, I. Simon, Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552–564 (2012).
10. J. D. Buenrostro *et al.*, Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548.e1516 (2018).
11. S. R. Hackett *et al.*, Learning causal networks using inducible transcription factors and transcriptome-wide time series. *Mol. Syst. Biol.* **16**, e9174 (2020).
12. S. Aibar *et al.*, SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
13. T. E. Chan, M. P. H. Stumpf, A. C. Babtie, Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* **5**, 251–267.e3 (2017).
14. A. Dixit *et al.*, Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866; e1817 (2016).
15. R. Oelen *et al.*; single-cell eQTLGen consortium; BIOS Consortium, Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat. Commun.* **13**, 3267 (2022).
16. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
17. D. Kotliar *et al.*, Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019).
18. S. Ma *et al.*, Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
19. T. M. Norman *et al.*, Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).
20. A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, T. M. Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
21. X. Qiu *et al.*, Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Syst.* **10**, 265–274.e11 (2020).
22. A. J. Rubin *et al.*, Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell* **176**, 361–376.e17 (2019).
23. D. van Dijk *et al.*, Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
24. X. Zheng, Y. Huang, X. Zou, scPADGRN: A preconditioned ADMM approach for reconstructing dynamic gene regulatory network using single-cell RNA sequencing data. *PLOS Comput. Biol.* **16**, e1007471 (2020).
25. B. Adamson *et al.*, A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 (2016).
26. M. Gasperini *et al.*, A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 1516 (2019).
27. D. Schraivogel *et al.*, Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).
28. S. Chong, C. Chen, H. Ge, X. S. Xie, Mechanism of transcriptional bursting in bacteria. *Cell* **158**, 314–326 (2014).
29. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
30. S. Itzkovitz *et al.*, Single-molecule transcript counting of stem-cell markers in the mouse intestine. *Nat. Cell Biol.* **14**, 106–114 (2011).
31. E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, A. van Oudenaarden, Regulation of noise in the expression of a single gene. *Nat. Genet.* **31**, 69–73 (2002).
32. A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, S. Tyagi, Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
33. P. S. Swain, M. B. Elowitz, E. D. Siggia, Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12795–12800 (2002).
34. D. Zenklusen, D. R. Larson, R. H. Singer, Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* **15**, 1263–1271 (2008).
35. S. C. Little, M. Tikhonov, T. Gregor, Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell* **154**, 789–800 (2013).
36. K. Bahar Halpern *et al.*, Bursty gene expression in the intact mammalian liver. *Mol. Cell* **58**, 147–156 (2015).
37. W. J. Blake, M. KAErn, C. R. Cantor, J. J. Collins, Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
38. I. Golding, J. Paulsson, S. M. Zawilski, E. C. Cox, Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–1036 (2005).
39. R. Losick, C. Desplan, Stochasticity and cell fate. *Science* **320**, 65–68 (2008).

40. N. Molina *et al.*, Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20563–20568 (2013).
41. A. Raj, A. van Oudenaarden, Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
42. N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, M. B. Elowitz, Gene regulation at the single-cell level. *Science* **307**, 1962–1965 (2005).
43. J. M. Pedraza, A. van Oudenaarden, Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).
44. J. Stewart-Ornstein, J. S. Weissman, H. El-Samad, Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae*. *Mol. Cell* **45**, 483–493 (2012).
45. O. Padovan-Merhar, A. Raj, Using variability in gene expression as a tool for studying gene regulation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **5**, 751–759 (2013).
46. K. Fujita, M. Iwaki, T. Yanagida, Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA. *Nat. Commun.* **7**, 13788 (2016).
47. T. Lionnet, R. H. Singer, Transcription goes digital. *EMBO Rep.* **13**, 313–321 (2012).
48. B. Munsky, G. Neuert, A. van Oudenaarden, Using gene expression noise to understand gene regulation. *Science* **336**, 183–187 (2012).
49. A. Sanchez, S. Choubey, J. Kondev, Stochastic models of transcription: From single molecules to single cells. *Methods* **62**, 13–25 (2013).
50. J. Peccoud, B. Ycart, Markovian modeling of gene product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
51. T. Alpert, L. Herzel, K. M. Neugebauer, Perfect timing: Splicing and transcription rates in living cells. *Wiley Interdiscip. Rev. RNA* **8**, 10.1002/wrna.1401 (2017).
52. M. Carmo-Fonseca, T. Kirchhausen, The timing of pre-mRNA splicing visualized in real-time. *Nucleus* **5**, 11–14 (2014).
53. M. Shamir, Y. Bar-On, R. Phillips, R. Milo, SnapShot: Timescales in cell biology. *Cell* **164**, 1302–1302.e1 (2016).
54. B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland Science, 2004).
55. 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).
56. J. Estrada, F. Wong, A. DePace, J. Gunawardena, Information integration and energy expenditure in gene regulation. *Cell* **166**, 234–244 (2016).
57. A. Hafner *et al.*, Quantifying the central dogma in the p53 pathway in live single cells. *Cell Syst.* **10**, 495–505.e4 (2020).
58. S. K. Hortsch, A. Kremling, Characterization of noise in multistable genetic circuits reveals ways to modulate heterogeneity. *PLoS One* **13**, e0194779 (2018).
59. Y. Jiang, N. R. Zhang, M. Li, SCALE: Modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol.* **18**, 74 (2017).
60. C. Li, F. Cesbron, M. Oehler, M. Brunner, T. Höfer, Frequency modulation of transcriptional bursting enables sensitive and rapid gene regulation. *Cell Syst.* **6**, 409–423.e11 (2018).
61. M. Razo-Mejia *et al.*, Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction. *Cell Syst.* **6**, 456–469.e10 (2018).
62. A. Senecal *et al.*, Transcription factors modulate c-Fos transcriptional bursts. *Cell Rep.* **8**, 75–83 (2014).
63. 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).
64. C. R. Bartman *et al.*, Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Mol. Cell* **73**, 519–532.e4 (2019).
65. V. A. Herzog *et al.*, Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).
66. S. A. Lambert *et al.*, The human transcription factors. *Cell* **172**, 650–665 (2018).
67. A. J. M. Larsson *et al.*, Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
68. 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. *Nat. Methods* **15**, 221–225 (2018).
69. B. Schwanhäusser *et al.*, Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
70. D. M. Suter *et al.*, Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
71. Q. Qiu *et al.*, Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nat. Methods* **17**, 991–1001 (2020).
72. N. Battich *et al.*, Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science* **367**, 1151–1156 (2020).
73. J. Lin *et al.*, Ultra-sensitive digital quantification of proteins and mRNA in single cells. *Nat. Commun.* **10**, 3544 (2019).
74. H. Ochiai, T. Sugawara, T. Sakuma, T. Yamamoto, Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. *Sci. Rep.* **4**, 7125 (2014).
75. D. Gillespie, Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
76. J. Reimegård *et al.*, A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun. Biol.* **4**, 624 (2021).
77. S. M. Shaffer *et al.*, Memory sequencing reveals heritable single-cell gene expression programs associated with distinct cellular behaviors. *Cell* **182**, 947–959.e17 (2020).
78. G. S. Kinker *et al.*, Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
79. J. Cao, W. Zhou, F. Steemers, C. Trapnell, J. Shendure, Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.* **38**, 980–988 (2020).
80. G. J. Hendriks *et al.*, NASC-seq monitors RNA synthesis in single cells. *Nat. Commun.* **10**, 3138 (2019).
81. L. Kiefer, J. A. Schofield, M. D. Simon, Expanding the nucleoside recoding toolkit: Revealing RNA population dynamics with 6-thioguanosine. *J. Am. Chem. Soc.* **140**, 14567–14570 (2018).
82. M. Zhang *et al.*, Highly parallel and efficient single cell mRNA sequencing with paired picoliter chambers. *Nat. Commun.* **11**, 2118 (2020).
83. A. Wagner, A. Regev, N. Yosef, Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
84. M. Hagemann-Jensen *et al.*, Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
85. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
86. G. Kar *et al.*, Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. *Nat. Commun.* **8**, 36 (2017).
87. C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, J. C. Marioni, Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **14**, 565–571 (2017).
88. X. Zhang *et al.*, Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-Seq systems. *Mol. Cell* **73**, 130–142.e5 (2019).
89. X. Zhang, C. Xu, N. Yosef, Simulating multiple faceted variability in single cell RNA sequencing. *Nat. Commun.* **10**, 2611 (2019).
90. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
91. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
92. J. M. Replogle *et al.*, Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.* **38**, 954–961 (2020).
93. P. A. Doerfler *et al.*, Activation of γ-globin gene expression by GATA1 and NF-Y in hereditary persistence of fetal hemoglobin. *Nat. Genet.* **53**, 1177–1186 (2021).
94. S. C. Hsu *et al.*, The BET protein BRD2 cooperates with CTCF to enforce transcriptional and architectural boundaries. *Mol. Cell* **66**, 102–116.e7 (2017).
95. K. Kaneko *et al.*, Identification of a novel erythroid-specific enhancer for the ALAS2 gene and its loss-of-function mutation which is associated with congenital sideroblastic anemia. *Haematologica* **99**, 252–261 (2014).
96. C. R. Scherzer *et al.*, GATA transcription factors directly regulate the Parkinson's disease-linked gene alpha-synuclein. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10907–10912 (2008).
97. M. Yu *et al.*, Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell* **36**, 682–695 (2009).
98. C. A. Sloan *et al.*, ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44** (D1), D726–D732 (2016).
99. C. P. Fulco *et al.*, Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
100. K. Shah, S. Tyagi, Barriers to transmission of transcriptional noise in a c-fos c-jun pathway. *Mol. Sys. Bio.* **9**, 687 (2013).
101. M. J. Levesque, A. Raj, Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods* **10**, 246–248 (2013).
102. N. Battich, T. Stoeger, L. Pelkmans, Control of transcript variability in single mammalian cells. *Cell* **163**, 1596–1610 (2015).
103. M. Müller-McNicoll, O. Rossbach, J. Hui, J. Medenbach, Auto-regulatory feedback by RNA-binding proteins. *J. Mol. Cell Biol.* **11**, 930–939 (2019).
104. A. Y. Mitrophanov, E. A. Groisman, Positive feedback in cellular control systems. *BioEssays* **30**, 542–555 (2008).
105. C. T. Walsh, S. Garneau-Tsodikova, G. J. Gatto Jr., Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew. Chem. Int. Ed. Engl.* **44**, 7342–7372 (2005).
106. H. Chung *et al.*, Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat. Methods* **18**, 1204–1212 (2021).
107. S. Shah *et al.*, Dynamics and spatial genomics of the nascent transcriptome by intron seq-FISH. *Cell* **174**, 363–376.e16 (2018).
108. A. Gupta *et al.*, Inferring gene regulation from stochastic transcriptional variation across single cells at steady state. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE202292. Deposited 5 May 2022.