3'-5' crosstalk contributes to transcriptional bursting

Massimo Cavallaro,[1,2,3,*] Mark D. Walsh,[1] Matt Jones,[1] James Teahan,[4] Simone Tiberi,[5] Bärbel Finkenstädt,[2] and Daniel Hebenstreit[1,*]

[1] *School of Life Sciences, University of Warwick, Coventry, UK*
[2] *Department of Statistics, University of Warwick, Coventry, UK*
[3] *Mathematics Institute and Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research, University of Warwick, Coventry, UK*
[4] *Department of Chemistry, University of Warwick, Coventry, UK*
[5] *Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland*

Transcription in mammalian cells is a complex stochastic process involving shuttling of polymerase between genes and phase-separated liquid condensates. It occurs in bursts, which results in vastly different numbers of an mRNA species in isogenic cell populations. Several factors contributing to "transcriptional bursting" have been identified, usually classified as intrinsic, i.e., local to single genes, or extrinsic, relating to the macroscopic state of the cell. However, each factor only accounts partially for the observed phenomenon, and some possible contributors have not been explored yet. We focussed on processes at the 3' and 5' ends of a gene that enable reinitiation of transcription upon termination. Using Bayesian methodology, we measured the transcriptional bursting in inducible transgenes, showing that perturbation of polymerase shuttling typically reduces burst size, increases burst frequency, and thus limits transcriptional noise. Analysis based on paired-end tag sequencing (ChIA-PET) suggests that this effect is genome wide. The observed noise patterns are also reproduced by a generative model that captures major characteristics of the polymerase flux between a gene and a phase-separated compartment.

In many cellular systems, mRNAs appear to be produced in burst-like fashion. This is directly observed in real-time experimental studies [1–3] and also agrees with theoretical analyses of steady-state mRNA distributions among single cells [4, 5]. Such bursty dynamics are thought to be the signature of gene regulation and are often described in terms of transcriptional "noise" [5, 6]. Due to the central role of transcription in cellular functions, it is important to understand the mechanisms from where the bursting originates [7].

The microscopic dynamics underlying transcription are not yet well understood. Various factors have been found to influence transcriptional dynamics, mostly by modulating bursting parameters such as the size or frequency of bursts [3, 5]. These factors are often classified as either intrinsic or extrinsic, although this distinction is blurred in many cases. This classification originally derives from the observation that fluctuations in expression levels are partially correlated across multiple genes [8], thus suggesting common, extrinsic causes, while the remaining, independent fluctuations are intrinsic to each gene. Typical major extrinsic noise sources are the cell cycle [9–11] and cell-size fluctuations [12], the latter partially due to the former. Numerous additional factors such as neighbouring cells, cell morphology and others have been found to affect transcription to varying degrees [13]. Intrinsic factors include non-linear transcription factor interactions [4, 5, 8], changing chromatin status [14, 15], promoter architecture [3], transcription factor diffusion [16], and several others [17–19].

* To whom correspondence should be addressed. E-mail: d.hebenstreit@warwick.ac.uk and m.cavallaro@warwick.ac.uk

It is unclear how these phenomena relate to the local environment at transcribing genes. These are associated to clusters of RNA polymerase II (PolII), which have been interpreted as "transcription factories" [20] and suggested to modulate the temporal patterns of transcription [21, 22]. More recently, it has been found that, in proximity to active genes, the PolIIs are incorporated in membrane-less droplets, maintained by a liquid-liquid phase separation (LLPS) from the rest of the nucleus, with the net effect of locally increasing the concentration of the factors involved in initiation; when PolII is liberated from this domain, transcription can be initiated [23–28]. LLPS also provides an explanation for the hitherto enigmatic action-at-a-distance type of gene regulation by distal enhancers, as the nuclear condensates are indeed able to restructure the genome [29].

While a comprehensive description of the interactions between PolIIs, other factors, and the chromatin within these niches is missing, several observations suggest that termination is linked to reinitiation; these include the presence of factors at both ends of a gene, the reduction of initiation upon perturbation of 3' processes, and protein interactions that would juxtapose the promoter and the terminator DNA, forming a structure that has been referred to as a "gene loop" [30, 31]. One of the effects of such interactions is to favour the reinitiation, thus increasing the mean expression level of the gene, as demonstrated in [32]. LLPS is highly important in this regard, as PolII undergoes a sequence of post-translational modifications on its C-terminal domain during transcription, while integration into phase separated domains and reinitiation requires it to be unmodified [24]. In line with this, recent studies suggest that LLPS is also involved in 3'-end transcriptional processes [33]. It has been suggested that

a repetitive cycle of reinitiation and termination due to these mechanisms is likely to produce a rapid succession of mRNA creation events, thus potentially contributing to the transcriptional bursts [34] but, to the best of our knowledge, an experimental verification is as yet lacking.

In this paper, we investigate the interplay between bursty expression and 3'-5' interactions using an interdisciplinary approach. We first consider two integrated genes that permit studying transcription upon perturbation of their 3'-5' processes at different induction levels; we demonstrate that this interaction strikingly influences the transcription kinetics and typically elicits the transcriptional noise, by decreasing burst frequency and increasing burst size. We then focus on genome-wide 3'-5' interactions involved in transcription by means of PolII ChIA-PET sequencing data, showing that they are related to the gene-expression parameters similarly to the transgenes' results. This scenario is well described by a microscopic stochastic model of gene expression, where tuning a single parameter—corresponding to the probability of local polymerase recycling—naturally yields the observed expression patterns, without involving extrinsic-noise contributors.

RESULTS

Cell lines as model systems for PolII recycling. We utilized two HEK293 cell lines which contain on their genomes copies of the genes *β-globin* (HBB) [32] and a modified version of HIV-1-*env* [35], respectively, driven by inducible CMV promoters (Fig. 1 A and B). The first gene, HBB, is an example for long-range chromosomal interactions in its native genomic neighbourhood. Its expression involves spatial proximity between the promoter and a locus control region (LCR) over 50 Kbs away [36]. The LCR has been studied extensively in murine and human cells (see, e.g., references [37, 38]) and jointly regulates expression of several *β-globin* like genes at the locus, likely involving LLPS [39]. A recent study demonstrates burst-like expression of murine HBB and suggests that interactions between the LCR and the HBB promoter modulate the bursting parameters [9]. Our cell line features an ectopic insertion of human HBB under control of a tetracycline (Tet) responsive promoter. A previous study of this system has provided a substantial number of results suggesting that 3' mRNA processing contributes to reinitiation of transcription [32]. This notion is based on several findings relating to the introduction of a single point mutation in the SV40 late poly-adenylation (pA) site (Fig. 1 C). This includes decreased average mRNA expression levels, while "read-through" transcription downstream of the pA site is increased. Furthermore, the mutation leads to a decrease of PolII, TBP and TFIIB levels at the promoter shortly after gene induction, and to an accumulation at the "read-through" region instead. Reduced transcription initiation compared to wild-type (WT) cells was also supported by nuclear run on assays and by a changed profile of post-translational modifications of PolII. Noticeably, TFIIB has been demonstrated to be functionally involved in linking 3' and 5' transcriptional activities [40], while post-translational modifications of PolII are in part carried out by Ssu72, which is associated with gene-loop formation in yeast [41] and appears to have similar roles in vertebrates [42]. A further recent study that utilized the ectopic HBB system reports direct detection of gene looping based on a 3C assay in the WT cell line, but not the mutant [43].

The second cell line, containing a Tet-inducible version of HIV-1-*env*, was studied previously in similar fashion to the HBB constructs. Results using a mutated version of the pA site (Fig. 1 D) mirrored those obtained with HBB, suggesting extensive 3'-5' crosstalk and recycling of factors including polymerase [32]. The *env* construct uses a BGH, not an SV40 pA site, which suggests that the findings are independent of the type of pA site. Notably, expression of the HIV-1 gene using its native long terminal repeat (LTR) promoter exhibits bursting dynamics [6].

We used these cell lines and their mutant versions as a model system for mammalian gene expression in presence and absence of 3'-5' crosstalk. We confirmed by total RNA-seq that HBB and *env* mRNAs are expressed inducibly in all cell lines (Fig. 1 A-D). At high Tet concentration (250 ng mL^{-1}), the fold changes over the uninduced samples were ≈ 16 and ≈ 26 for HBB and *env*, respectively. The mutants were expressed at lower levels and featured read-through transcription as described, with intact transcript sequences, i.e., not subject to splicing defects (Fig. 1 C-D). This indicated specificity of the pA site mutations.

In order to detect transcripts at the single molecule level, we designed probes for single molecule RNA-FISH (smFISH) and confirmed detection of large transcript numbers upon Tet stimulation of the cells, while the expression of a control gene, AKT1, remained constant (Fig. S3, *SI Appendix*). Microscopy-based smFISH is not ideal for HEK293 cells, since they tend to overlap and form aggregates when growing. We therefore decided to record the smFISH signal by adapting a flow-FISH technique based on flow cytometry [44]; this also resolves extrinsic-noise contributors such as cell size, morphology, and cycle, and, thanks to its high throughput, permits recording vast numbers of cells to analyse overall population structures (*SI Appendix*, sections S1 and S6).

While the flow-cytometer fluorescence signal from stained cells serves as a proxy for the mRNA abundance, it is returned in arbitrary units (a.u.) rather than in absolute counts. We thus used microscope imaging and nCounter® data to calibrate the flow-FISH fluorescence readings of HBB and *env* cells, respectively. Applying the clustering algorithm of [45] to the flow-FISH recordings allowed us to select single-cell readings against those from cell clumps, doublets, and debris (*SI Appendix*, section S1 and Fig. S1).

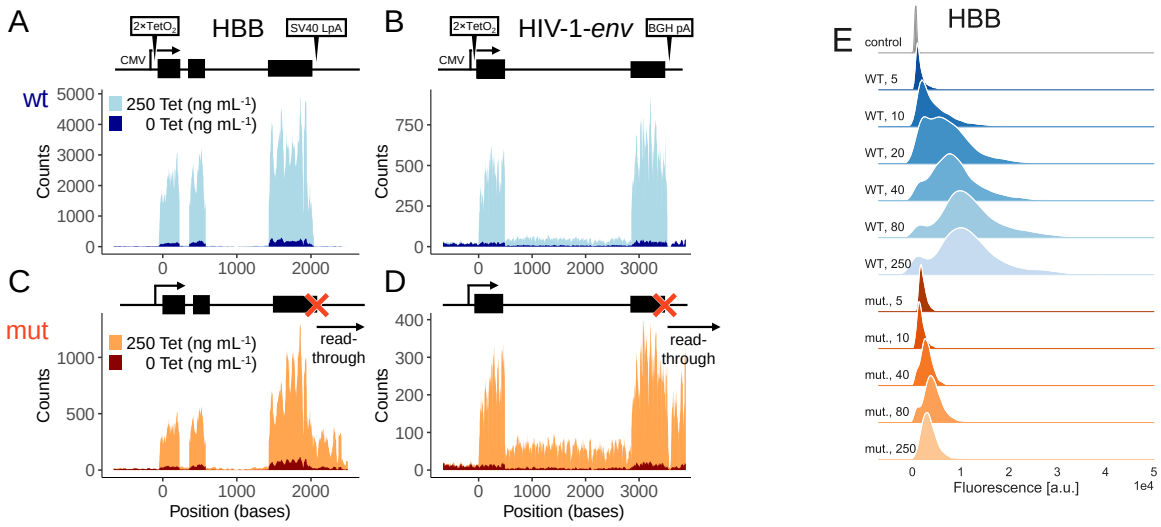

FIG. 1. Characteristics of transgenes used in this study. (A) schematic gene structure (top) of WT HBB including CMV promoter, Tet operator, pA signal, and exons (black blocks) as indicated. Total RNA-seq confirms Tet-inducible expression (bottom). (B) as (A), for *env*. (C)-(D) mutant versions of HBB and *env*, respectively. Point mutations in pA sites ('x') and read-through transcription are indicated. Positions are relative to TSS, blue and orange shades correspond to WT and mutant versions, respectively, and light and dark shades correspond to 250 and 0 ng mL^{-1} Tet, respectively. Coverages by sequencing reads are shown. (E) Kernel density estimates of the flow-FISH single-cell readings corresponding to the abundances of HBB transcripts, WT (blue), mutant (orange) variants, and control (gray) cells, from replicate $k = 1$, at different induction levels (Tet concentrations in unit of ng mL^{-1}, shades of colors, as indicated on the left). Gene expression increases and saturates upon increasing Tet concentration, mutant-cell expression is lower than the WT; a.u., arbitrary units; y-axes not to scale.

Flow-FISH data demonstrate Tet-dose dependent expression of HBB and *env*, indicating specific detection of transcripts above background noise. The stationary expression levels appeared to reach saturation at 80 ng mL^{-1} Tet (Fig. 1 E and Fig. S2, *SI Appendix*). Staining for the DNA content demonstrates a mild increase of HBB and *env* expression with increasing cell cycle stage. We found that the contribution to the total variability, measured as the squared coefficient of variation (CV^2) of the mRNA population, due to the cell cycle was minor (*SI Appendix*, section S6) and therefore focused on local genic mechanisms to investigate the observed noise pattern. The measured signal includes a background of unspecific staining and auto-fluorescence of the cells, which is subtracted from the total signal [46]. To gauge this background we deleted the *env* gene from its host cell line with Cas9 and performed the staining procedure as before. The resulting control cells had low fluorescence intensity that remained virtually unchanged upon maximal Tet stimulation, thus confirming specificity of our system and validating the use of this control to estimate the background (Table S1, *SI Appendix*).

Increased transcriptional bursting upon 3'-5' crosstalk. In order to gain insights into the transcriptional dynamics driving WT and mutant expression of HBB and *env*, we employed a Markov chain Monte Carlo (MCMC) sampling approach to fit statistical models to the flow-FISH data. Importantly, Bayesian modelling permitted using microscope and nCounter® data to estimate informative prior distributions that calibrate the absolute mRNA quantification, whilst retaining flexibility in this respect. We further incorporated the background signal in the Bayesian framework based on the estimates from the Tet-stimulated control cells (*Material and Methods* and *SI Appendix*, sections S2-S3).

Our strategy requires a flexible model to represent the absolute mRNA abundance. We considered three stochastic models of gene expression to capture the phenomenology of the transcription process (*Materials and Methods*). According to the first model, the gene can stay in an "on" state, in which transcription occurs at rate $\tilde{\alpha}$, or in an "off" state, in which no transcription occurs. The gene switches from "off" to "on" and "on" to "off" at rates $\tilde{k}_{\text{on}}$ and $\tilde{k}_{\text{off}}$, respectively. Assuming that the mRNA degrades at constant rate $\tilde{d}$, this model corresponds to a Poisson-beta mixture distribution for the stationary per-cell mRNA population, which can be expressed in terms of the dimensionless rates α, k_{on}, and k_{off} (*SI Appendix*, section S2) [4, 47]. The second model is a simplified version of the former two-state model, where α and k_{off} approach infinity, whilst the ratio α/k_{off}, which is referred to as the average burst size [48] and incorporated as a single parameter, is held finite; this model gives rise to a negative binomial stationary mRNA distribution and allows much more efficient MCMC sampling than the Poisson-beta model (*SI Appendix*, sections S3-S4). The third model is the most naïve as it assumes that transcription events of individual mRNAs occur indepen-

dently at constant rate $\mu_X \cdot \tilde{d}$, where μ_X is the mean mRNA population, thus yielding a Poisson distributed mRNA population at equilibrium which is thought to characterise genes with unregulated expression [5]. Noise levels consistent with the Poisson model [49, 50] or higher [4, 13] have both been reported in the literature.

We obtained better fits for the Poisson-beta and the negative-binomial models than the Poisson model (*SI Appendix*, sections S4 and S7) for all the replicates. In the Poisson-beta case, the MCMC traces of the rates k_{off} and α had strong correlation; this revealed that most of the information about these two parameters is encoded in the ratio α/k_{off} (*SI Appendix*, section S6 and Fig. S12), which is more straightforwardly inferred by means of the negative-binomial model. In fact, for our data, these two models give consistent results in terms of CV^2, average burst size α/k_{off}, and burst frequency $\tilde{k}_{\text{on}}$. To study the transcriptional noise, we obtained the CV^2 of the mRNA abundance (which we refer to as CV^2_X) from the estimated parameters (*SI Appendix*, sections S2 and S7), and plotted it against the estimated mean expression levels μ_X (Fig. 2 A-C). These reveal a trend observed before in other systems [6, 51–53], i.e., the transcriptional noise decreases as μ_{X} increases, with the data of each experiment well fitted by a curve of the form $\text{CV}^2_X = A/\mu_{\text{X}} + B$, and seems to approach a lower limit beyond which it does not further decrease. Such a limit is known as the noise floor [54–58]. Strikingly, the presence of the mutation alters the noise trends, thus suggesting that PolII recycling indeed contributes to the noise. The transcriptional noise at intermediate expression levels is significantly higher in WT than mutant cells. For the HBB gene, this pattern extends throughout the range of all induction levels. *Env* shows less pronounced differences between WT and mutant cells for the highest expression levels but resembles HBB otherwise. In all these cases, the noise clearly appears higher than postulated by the Poisson prediction curve $\text{CV}^2_X = 1/\mu_X$ (solid lines in Fig. 2 A-C).

Using the DNA content as proxy of the cell-cycle progression, we heuristically selected populations corresponding to G1, S, and G2 phases from 40 ng mL^{-1} Tet-induced cells (Fig. 3 A), fitted the negative-binomial model to their mRNA-expression reads, and estimated kinetic parameters and noise at each of the three phases separately (Fig. 3 C-D). Based on this we found that the cell cycle, which is a major extrinsic noise contributor, only accounts for around 20% of total mRNA variability for the transgenes (Fig. 3 B), in contrast with [9, 10]; for further details see *SI Appendix*, section S6.

Modulation of rates. The overall rate estimates obtained from our fits are largely in agreement with previous findings from similar systems [3]. In fact, estimated values of $\tilde{k}_{\text{off}}$ ranged up to ≈ 2.5 events per minute, with $\tilde{k}_{\text{on}}$ roughly an order of magnitude lower. Increasing the Tet concentration boosts transcription by increasing the average burst size and the frequency $\tilde{k}_{\text{on}}$ (Fig. 2 D), thus shortening the average "off" state duration ($1/\tilde{k}_{\text{on}}$). Intriguingly, for the HBB gene, $\tilde{k}_{\text{on}}$ is higher in mutant than WT cells in all cases, while the average burst size is lower in mutant cells in all cases. These patterns are less definite for the *env* gene but appear to support the conclusions from the HBB gene (Fig. 2 E and *SI Appendix*, section S7). In other words, the 3'-5' crosstalk imposes a constraint on the transcriptional dynamics whose removal can cause bursts to be more frequent and smaller than in the WT gene.

PolII-mediated 3'-5' interactions by ChIA-PET. To jointly study the expression of a gene and its 3'-5' interactions we analysed publicly available datasets for the human cell line K562, obtained from chromatin-interaction analysis by paired-end tag sequencing (ChIA-PET) [59] and single-cell RNA-seq data (scRNAseq) [50]. We chose to use ChIA-PET against PolII to target chromatin interactions that are involved in transcription. We generated HiC-style interaction matrices (whose entries correspond to 2-Kb regions) from the ChIA-PET data using CHIA-PET2 [60]. We filtered the list of genes from the RefGene database with the hg19 reference genome to only contain those with unique gene symbols on chromosomes 1-22 and X, thus excluding alternatively spliced genes. As a proxy of the 3'-5' interaction of a gene, we first aggregated the reads corresponding to the interaction between the bins that include its transcription start site (TSS) and transcription end site (TES). The resulting metrics depend on the gene length, which we addressed by dividing the number of reads for each gene by the average read number from 10^4 genomic intervals of the same length as the gene, randomly sampled across the chromosome. We then applied the $\text{arcsinh}\sqrt{x+0.5}$ transformation to obtain a variance-stable interaction score [61]. We also discarded genes that are shorter than the resolution of our interaction matrices.

Fitting a negative binomial distribution to the scRNAseq UMI counts data of [50] allows us to estimate for expressed genes (sample UMI mean > 0.05) the noise CV^2_X, as well as the parameters k_{on} and α/k_{off} (*Materials and Methods*). These are plotted against the mean expression μ_X in Fig. 4 A-C. It is worth noting that burst frequency averaged over all the genes $\bar{k}_{\text{on}}$ seems to determine the average trends of CV^2_X and α/k_{off}. The noise trend appears to be explained by the curve $\text{CV}^2_X = 1/\mu_X + 1/\bar{k}_{\text{on}}$ (derived under the negative-binomial assumption, see *SI appendix*, section S2), which in fact separates the genes whose noise levels are higher than the mean predicts (blue and orange markers in Fig. 4) from those whose noise is lower than the prediction (yellow markers). As a measure of the deviation from this prediction, for each gene, we calculated the vertical distance ν of its expression noise to the curve $\text{CV}^2_X = 1/\mu_X + 1/\bar{k}_{\text{on}}$ in logarithmic scale, further separating noisy genes for which $\nu > \nu_1$ (blue makers in Fig. 4) from those for which $0 < \nu < \nu_1$ (orange makers). The interaction score of the high-noise genes is significantly higher than the score of the intermediate group, which in turn is higher than the low-noise genes' (Mann–Whitney U-test, $\text{P} < 2.2 \cdot 10^{16}$).

There is a significant positive correlation between the

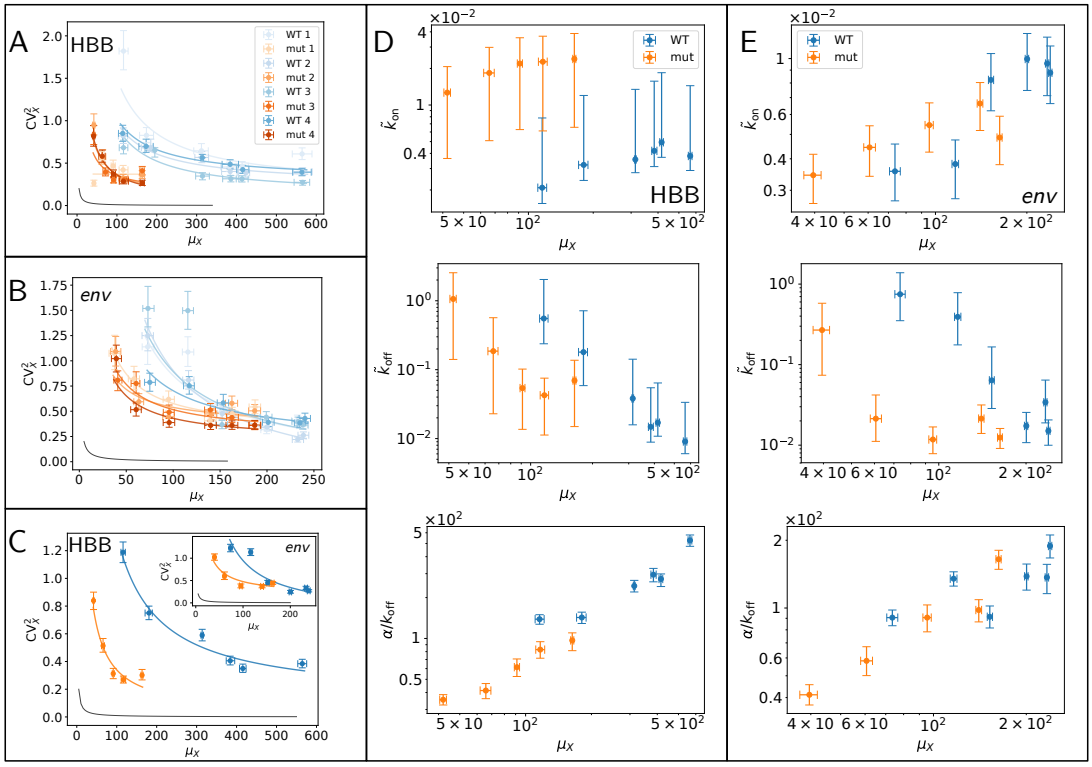

FIG. 2. Bayesian parameter estimates. Noise plots of HBB (A) and HIV (B) gene expressions, obtained from the Poisson-beta model for both WT (blue) and mutant (orange) gene variants. Different color intensities correspond to replicates. Mutation changes the balance between noise and average expression level. (C) Results from replicates are aggregated into consensus estimates (*SI Appendix*, section S4) for HBB and HIV (inset). Solid lines are orthogonal-distance regression curves $\mathrm{CV}_X^2 = A/\mu_X + B$. (D-E) Consensus estimates of Poisson-beta model parameters μ_X, $\tilde{k}_{\mathrm{on}}$, $\tilde{k}_{\mathrm{off}}$, and α/k_{off} for HBB (D) and HIV (E). WT (blue) and mutant (orange) show different patterns, with WT genes having highest average burst size and lower burst frequency than mutant at intermediate expression levels. Single-replicate estimates, and negative binomial and Poisson model results are in *SI Appendix*, section S7. Points and error bars correspond to medians and 90% HPD CIs of the posterior distributions.

distance ν and the interaction score ($\mathrm{P}< 2.2 \cdot 10^{-16}$, lm; Fig. 4 D), thus showing that the noise level of genes with high interaction score is typically higher than the mean predicts; we also observe a significant negative correlation between the interaction score and the burst frequency k_{on} ($\mathrm{P}< 2.2 \cdot 10^{-16}$, lm) and a significant positive correlation between the interaction score and the burst size ($\mathrm{P}< 2.2 \cdot 10^{-16}$, lm), consistent with the results on the transgenes. Filtering out zero-count genes, for which there is little statistical information, yields the scatter plots of Fig. 4 D-E and the boxplot of Fig. 4 F for the three groups. These results agree with those obtained from different ChIA-PET biological repeats and different bin resolutions (1 Kb and 7 Kb; *SI appendix*, Fig. S16).

Microscopic model. To shed further light on the biological mechanisms involved and test whether PolII shuttling can a priori alter the transcriptional noise as seen in the previous section, we constructed and simulated a more complex stochastic model that captures the most important features of our expression system, i.e., induction, polymerase flux between the LLPS droplet and the gene, transcription, and decay, whilst stripping away non-essential details (Fig. 5 and *Materials and Methods*). The model is designed around the idea that each PolII waits in a compartment until the transcription occurs [22], where the compartment represents an LLPS droplet (Fig. 5 A). This is immersed in its nuclear environment, which adds and removes PolIIs at rates γ and δ, respectively. In addition to this, by transcribing at rate β, the PolIIs leave the compartment with probability $1-l$ or are re-injected otherwise. This latter reaction represents the crosstalk between the 3'-end processing and the transcription initiation and helps to sustain the compartment population despite the presence of initiation, which in average contributes to depleting it. Consistently with the two genes integrated in our cell lines, the model encodes a Tet-repressor binding site downstream of the TSS which binds to the TetR factor, present in concentration n. Such a binding event interrupts the transcription, therefore tuning n allows us to control the blocking rate λ_{off} (*Materials and Methods*). The model parameters l and n are akin to the pA mutation and the Tet concentration, respectively, in the experimental settings. We assume that the pA mutation hinders but does not com-

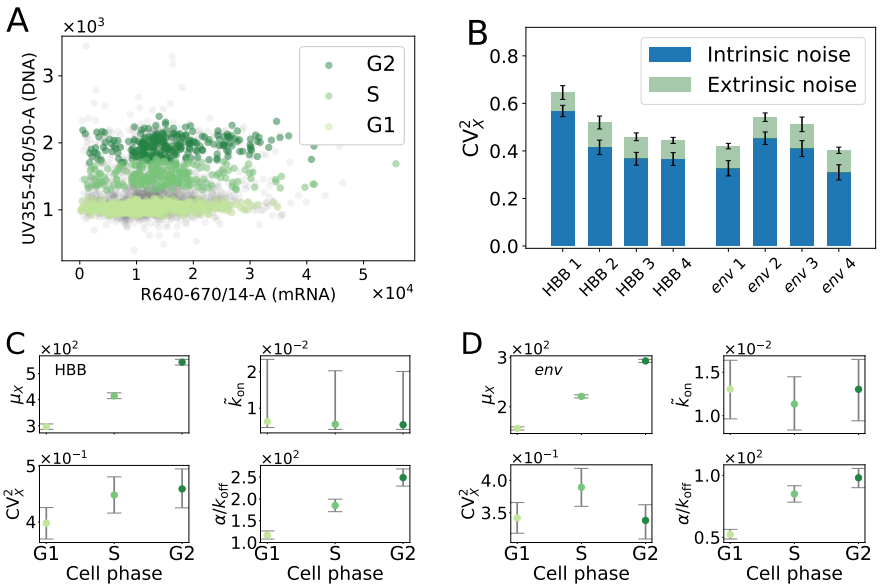

FIG. 3. Cell cycle analysis. (A) Scatter plot from flow-FISH signals (corresponding to mRNA vs DNA) for the HBB gene, replicate $k = 3$; cells from G1, S, and G2 phase highlighted with green-scale colors. (B) Extrinsic and intrinsic contributions to WT HBB and *env* genes expression noise, SE error bars obtained via bootstrap. (C) and (D) Consensus estimates of the negative-binomial model parameters for the same genes; points are medians, error bars comprise 90% HPD CIs.

pletely block PolII flux back to the compartment (which can also be facilitated by diffusion, for instance [16, 24]), therefore the parameter l is assumed to be small but still strictly positive even in the presence of pA mutation. Crucially, the transcription rate is proportional to the abundance of PolII in the compartment [22, 58], so that, when the blockade is released and the compartment is full, the transcription occurs repeatedly while the PolII population quickly drops. As simulation results demonstrate, the model is able to reproduce an increase of transcriptional bursting upon increasing the recycling probability l (Fig. 5). This behaviour is conserved under a broad range of different parameter settings, demonstrating that this is a generic result of our model. Fitting a negative binomial distribution with vague prior distributions to an ensemble of mRNA abundances, simulated from this microscopic model, shows patterns consistent with those obtained from the experimental data (Fig. 5 C and *SI Appendix*, section S8).

While actual transcriptional mechanisms are more complex than our idealised model, the latter provides a significant step towards a mechanistic explanation of our observations. In fact, it captures the essential features of the two gene constructs, and naturally reproduces the observed pattern by tuning only the shuttling probability l and the factor abundance n. Notably, our results demonstrate a minor role for extrinsic contributions to noise (Fig. 5 B); in fact, intrinsic factors suffice to yield the noise floor for a wide range of λ_{off} and μ_X, which contrasts with several other studies [54–58].

DISCUSSION

The wealth of existing results strongly suggests the occurrence of 3'-5' crosstalk in the WT variants of our transgene systems, involving physical interaction between factors at either gene end and shuttling of polymerases, which can be disrupted or strongly reduced upon a point mutation. Similarly, information of the interactions between the ends of genes involved in transcription can be accessed genome-wide by means of PolII ChIA-PET sequencing.

Based on both an in-depth analyses of the transgene systems (which provide a controlled experimental setting) and an observational study of ChIA-PET sequencing data (which provide a genome-wide view of chromatin interactions involved in transcription), we present results to suggest that PolII-mediated 3'-5' interactions are major contributors to transcriptional noise.

Building on standard phenomenological models, transcription parameters, such as average burst size and frequency, are consistently inferred across the different conditions using a Bayesian methodology, to demonstrate the presence of association between 3'-5' interactions and transcription kinetics. Modelling transcription requires abstraction and simplification due to the complexity of the molecular processes involved and the inadequacy of current experimental methodologies to dynamically resolve structural interactions at individual loci. Furthermore, the Bayesian estimates of the kinetic parameters reflect the incomplete quantitative information available on the experimental device. Nevertheless, our setting is sufficient to resolve specific patterns, which can be repro-

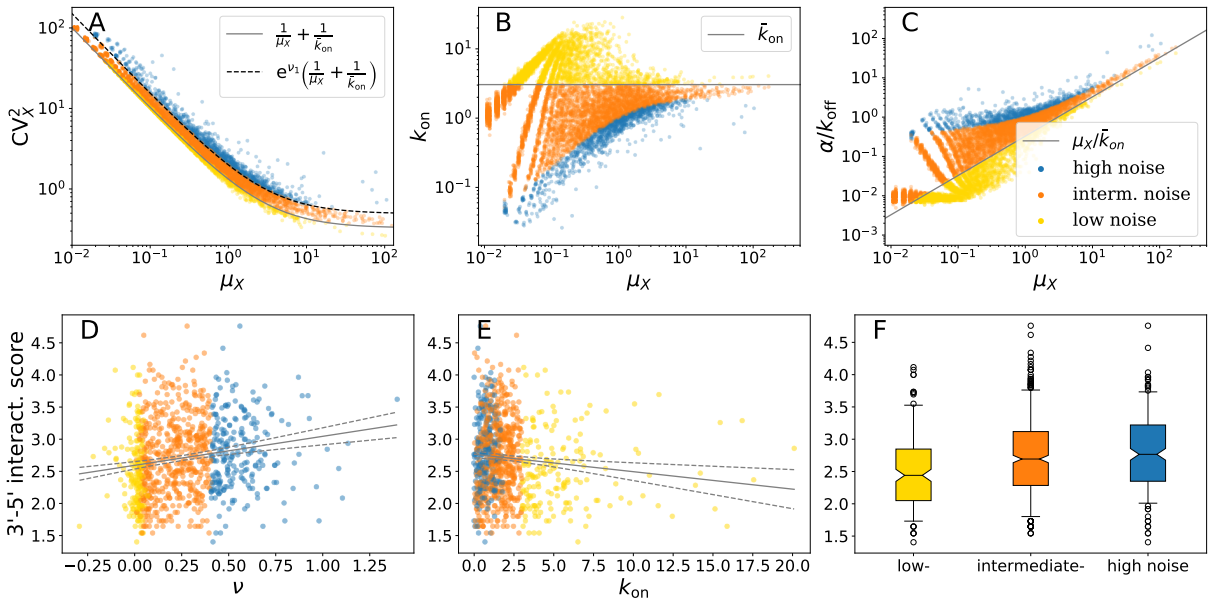

FIG. 4. Genome-wide estimates of transcription kinetics and 3'-5' interactions. (A)-(C) Scatter points correspond to genes, axes are medians of posterior distributions for expression parameters μ_X and CV^2_X, k_{on}, and α/k_{off}, respectively, obtained by Bayesian model fitting; solid lines correspond to the predictions obtained by assuming that all genes have burst frequency equal to the sample average $\overline{k}_{\mathrm{on}}$. Genes are divided into three groups corresponding to low-, intermediate-, and high-noise levels (yellow, orange, and blue markers, respectively). Dashed line is obtained by setting $\nu_1 = 4.5$ (equation inset in (A)) to separate intermediate- and high-noise genes. (D)-(E) 3'-5' interaction scores against expression noise (measured as distance ν from the solid-line prediction of A) and burst frequency k_{on}; (F) Partitioning the genes by ν shows that the interaction score is significantly higher in higher-noise genes than in lower-noise genes (Mann-Whitney U-test, $\mathrm{P} < 2.2 \cdot 10^{-16}$).

duced by an *ab-initio* mechanistic model, thus supporting our conclusions.

The analysis suggests that recycling of the polymerase typically increases noise at a given expression level, while an alternative symmetric interpretation is possible, viz., that recycling permits higher expression at a given noise level. These relations are either a byproduct of the construction of the transcriptional machinery or were selected for. It will be interesting to further explore our findings from an evolutionary perspective. In particular, many studies show how selection of noisy expression can be critical by contributing to cell fate diversity [62, 63] and by favouring their long-term survival in adverse environments [64]. This could also have implications in synthetic biology, where the optimisation of gene expression and the control of its noise are desirable features [65, 66]. Our work provides an important contribution to the field of systems biology by identifying a single base, and thus a genetic determinant, that modulates the balance between the average expression level and its variation.

MATERIALS AND METHODS

Measurement equation and Monte Carlo estimation. We assume that the measured fluorescence Y_i of cell i is proportional to the true mRNA abundance X_i and therefore can be expressed as $Y_i^{(k)} = \epsilon_i^{(k)} + \kappa^{(k)} X_i^{(k)}$ where (k) indexes the replicate, $\kappa^{(k)}$ can be thought of as a scale and $\epsilon_i^{(k)}$ is the zero of such a scale, also corresponding to the background of unspecific staining and auto-fluorescence of the ith cell [46]. The background noise is measured, for each replicate k, by means of control cells whose gene of interest has been deleted. These are used to define informative priors for $\epsilon_i^{(k)}$. Our choice is $\epsilon_i^{(k)} \sim \mathrm{SN}(a^{(k)}, \mu_\epsilon^{(k)}, \sigma_\epsilon^{(k)})$, i.e., the control-cell fluorescence y is supposed to have Azzalini's skew-normal distribution

$$f_\epsilon(y|a^{(k)}, \mu_\epsilon^{(k)}, \sigma_\epsilon^{(k)}) = 2\Phi((y-\mu_\epsilon^{(k)})\,\sigma_\epsilon^{(k)}\,a^{(k)})\,\phi(y|\mu_\epsilon^{(k)}, \sigma_\epsilon^{(k)}),$$

where Φ and ϕ are the standard normal CDF and normal PDF, respectively, while the mean $\mu_\epsilon^{(k)}$, the standard deviation $\sigma_\epsilon^{(k)}$, and the skewness parameter $a^{(k)}$ are point estimates from the control data sets. Prior distributions for $\kappa^{(k)}$ are chosen based on the regression coefficients of gamma generalised linear model fits with identity link. For the remaining parameters we assume vague gamma priors with mean 1 and variance 10^3. Adaptive Metropolis–Hastings samplers for model fitting were implemented (*SI Appendix*, section S4).

Phenomenological two-state gene-expression models. The transcriptional bursting is fully characterised by the rates $\tilde{\alpha}, \tilde{k}_{\mathrm{on}}$, and $\tilde{k}_{\mathrm{off}}$ in units of min^{-1}. It is convenient to express the rates in units of the inverse of the mean mRNA life-time $\tilde{d}$, i.e., $\tilde{k}_{\mathrm{off}} = k_{\mathrm{off}}\,\tilde{d}$,

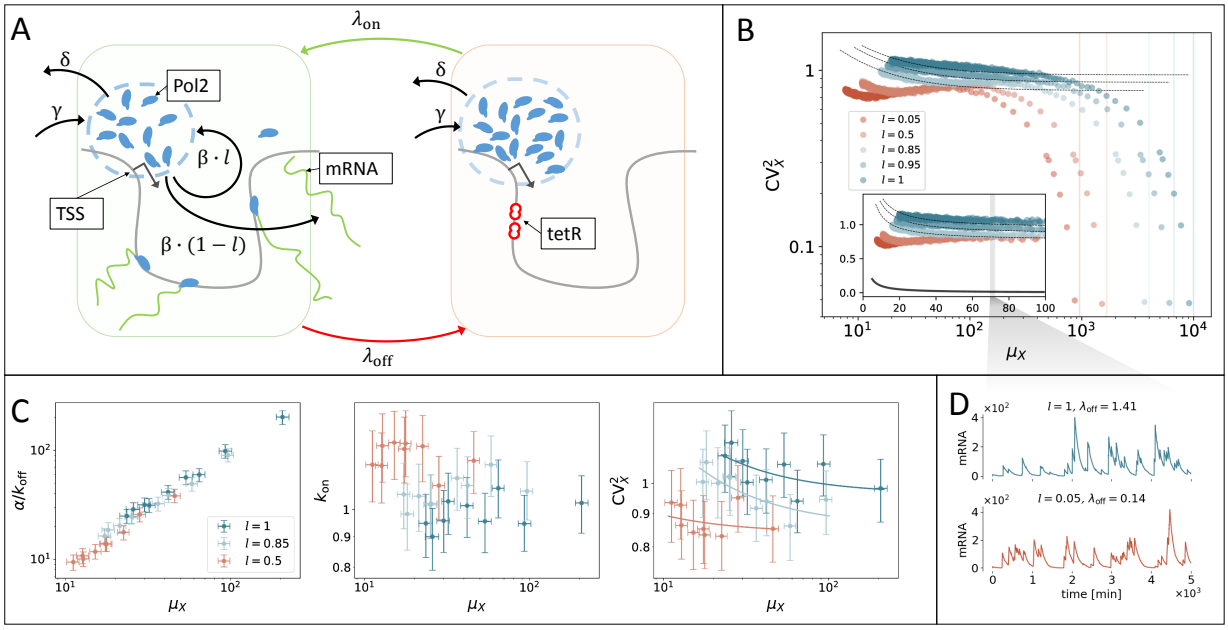

FIG. 5. Microscopic model of transcription in Tet-inducible genes. (A) PolIIs (blue) are stored in a compartment (dashed circle) in the proximity of the TSS. With rate β, each PolII leaves the compartment to transcribe mRNA and is re-injected with probability l. When TetR (tetracycline repressor) binds to the $TetO_2$ operator downstream of the TSS (this occurs at rate λ_{off}), transcription is interrupted and PolIIs accumulate in the compartment. At rate λ_{on}, TetR unbinds, thus releasing the large amount of PolIIs accumulated in the compartment to causes bursts, which can be phenomenologically described in terms of the rates $\tilde{\alpha}$, $\tilde{k}_{on}$, and $\tilde{k}_{off}$. The compartment also exchanges PolIIs with the nuclear environment (at rates δ and γ). The transcription rate is proportional to the abundance of PolIIs, which fluctuates in time and in turn elicits transcriptional noise. Similarly to our experimental system, here we can simulate different Tet concentrations and the recycling probability by tuning the "off"-switch rate λ_{off} and l, respectively. (B) Noise plots of simulated mRNA abundances. Setting $\lambda_{off} = nK_\lambda$ and $\lambda_{on} = K_\lambda$, we imitate the effect of different TetR concentration values by tuning n. As Tet presence prevents TetR-$TetO_2$ binding, small values of n correspond to high Tet-induction levels. For extremely small values of n, the gene can be thought of as being always in "on" state, CV^2 becomes very low, and the limiting value of μ_X can be analytically obtained (vertical lines, see also *SI Appendix*, section S8). n ranges from 0.1 to 100, values of the other parameters are $(\gamma, \beta, d, \delta, K_\lambda) = (10, 10, 0.01, 1, 0.01)$. Inset: Same scatter plot, axes in linear scale. At intermediate expression levels, CV_X^2 always increases with l. Dashed lines are orthogonal-distance regression curves $CV_X^2 = A/\mu_X + B$, solid line is Poisson-noise curve $CV_X^2 = 1/\mu_X$. (C) Negative-binomial model fit to 500 mRNA abundances simulated from the microscopic model with $\lambda_{off} = 0.5, 1, 1.5, 2, 2.5, 3, 2.5, 3.5, 4, 4.5$, values of other parameters as in B. (D) Simulated mRNA-population traces; the two parameter combinations yield almost identical average expressions (sample means of 71.3 ± 0.7 and 70.4 ± 0.6 over 10^4 realisations, respectively, SE obtained via bootstrap), but different biological noise (sample CV^2s of 0.78 ± 0.01 and 1.07 ± 0.02, respectively).

$\tilde{k}_{on} = k_{on}\,\tilde{d}$, $\tilde{\alpha} = \alpha\,\tilde{d}$. It can be shown that the stationary mRNA abundance X for this model is Poisson beta with probability density function (PDF)

$$f_X(x|\alpha, k_{on}, k_{off}) = \int_0^1 f_{Poi}(x|\alpha p) f_{Be}(p|k_{on}, k_{off})\,\mathrm{d}p,$$

where $f_{Poi}(x|\alpha) = \alpha^x \mathrm{e}^{-\alpha}/x!$ and $f_{Be}(p|k_{on}, k_{off}) = p^{k_{on}-1}(1-p)^{k_{off}-1}\Gamma(k_{on}+k_{off})\,(\Gamma(k_{off})\Gamma(k_{on}))^{-1}$ are PDFs of Poisson and beta random variables (RVs), respectively. This expresses the hierarchy

$$X|\alpha, P \sim \mathrm{Poi}(\alpha P), \qquad P|k_{on}, k_{off} \sim \mathrm{Beta}(k_{on}, k_{off}).$$

It is convenient to reparametrise the Poisson-beta PDF in terms of its mean $\mu_X = \alpha k_{on}/(k_{off}+k_{on})$, to get

$$X|\mu_X, k_{on}, k_{off}, P \sim \mathrm{Poi}(\mu_X P\,(k_{off}+k_{on})/k_{on}),$$
$$f_X(x|\alpha, k_{on}, k_{off}) =: f'_X(x|\mu_X, k_{on}, k_{off}).$$

In fact, this allows us to exploit knowledge on μ_X in the form of informative priors and infer the dimensionless rates α, k_{off} and k_{on}. These are converted to min^{-1} by using $\tilde{d}$ estimated from data (*SI Appendix*, section S5). In the limit as $k_{off} \to \infty$, $\alpha \to \infty$, with their ratio α/k_{off} held finite, the population mean satisfies $\mu_X = k_{on}\alpha/k_{off}$, while the PDF of X approaches the negative binomial distribution

$$f''_X(x|k_{on}, k_{off}/\alpha) = \int_0^\infty f_{Poi}(x|\lambda) f_{Gamma}(\lambda|k_{on}, k_{off}/\alpha)\,\mathrm{d}\lambda,$$

where $f_{Gamma}(x|k_{on}, k_{off}/\alpha)$ is the density of a Gamma RV with mean μ_X and variance $\mu_X k_{off}/\alpha$; when this RV concentrates near the mean as $k_{on} \to \infty$ and $k_{off}/\alpha \to 0$, X is Poisson with PDF $f_{Poi}(x|\mu_X)$.

Microscopic model. The microscopic model is defined

by means of the following chemical reaction scheme:

$$\begin{aligned}
&\mathrm{DNA_{on}} + \mathrm{PolII} \xrightarrow{l\,\beta} \mathrm{mRNA} + \mathrm{DNA_{on}} + \mathrm{PolII},\\
&\mathrm{DNA_{on}} + \mathrm{PolII} \xrightarrow{(1-l)\,\beta} \mathrm{mRNA} + \mathrm{DNA_{on}},\\
&\mathrm{DNA_{on}} \xrightarrow{\lambda_{\mathrm{off}}} \mathrm{DNA_{off}}, \quad \mathrm{DNA_{off}} \xrightarrow{\lambda_{\mathrm{on}}} \mathrm{DNA_{on}},\\
&\mathrm{mRNA} \xrightarrow{d} \varnothing, \quad \varnothing \xrightarrow{\gamma} \mathrm{PolII}, \quad \mathrm{PolII} \xrightarrow{\delta} \varnothing.
\end{aligned}$$

By the law of mass action, $\lambda_{\mathrm{off}} = nK_\lambda$, $\lambda_{\mathrm{on}} = K_\lambda$, where K_λ and n represent the chemical affinity and concentration of TetR homodimers that bind to the $\mathrm{TetO_2}$ operators downstream of the TSS, respectively. When such a binding event occurs, the transcription is inhibited as elongation is impeded and the resulting locked DNA configuration is represented by the species $\mathrm{DNA_{off}}$. The switch to $\mathrm{DNA_{on}}$ corresponds to the release of the lock.

Data availability. Custom scripts have been made available at https://github.com/mcavallaro/gLoop. Data that support the findings of this study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) and are accessible through the GEO Series number GSE124682. All other relevant data are available on request.

ACKNOWLEDGMENTS

The research was supported by BBSRC grant BB/L006340/1, and utilised WISB computational and experimental facilities (grant ref: BB/M017982/1) funded under the UK Research Councils' Synthetic Biology for Growth programme. We thank Louise Dyson, Matt Moores, Lucy Ternent, and Jonathan Keith for valuable discussions, and Sharon Collier and Charlotte Petersen for minor contributions.

[1] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell **123**, 1025 (2005).
[2] J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer, Current Biology **16**, 1018 (2006).
[3] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, Science **332**, 472 (2011).
[4] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, PLoS Biology **4**, e309 (2006).
[5] B. Munsky, G. Neuert, and A. van Oudenaarden, Science **336**, 183 (2012).
[6] A. Singh, B. Razooky, C. D. Cox, M. L. Simpson, and L. S. Weinberger, Biophysical Journal **98**, L32 (2010).
[7] A. J. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg, Nature **565**, 251 (2019).
[8] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 1183 (2002).
[9] C. J. Zopf, K. Quinn, J. Zeidman, and N. Maheshri, PLoS Computational Biology **9**, e1003161 (2013).
[10] M. S. Sherman, K. Lorenz, M. H. Lanier, and B. A. Cohen, Cell Systems **1**, 315 (2015).
[11] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, eLife **5** (2016), 10.7554/eLife.12175.
[12] O. Padovan-Merhar, G. Nair, A. Biaesch, A. Mayer, S. Scarfone, S. Foley, A. Wu, L. Churchman, A. Singh, and A. Raj, Molecular Cell **58**, 339 (2015).
[13] N. Battich, T. Stoeger, and L. Pelkmans, Cell **163**, 1596 (2015).
[14] J. M. Raser and E. K. O'Shea, Science **304**, 1811 (2004).
[15] L. Weinberger, Y. Voichek, I. Tirosh, G. Hornung, I. Amit, and N. Barkai, Molecular Cell **47**, 193 (2012).
[16] J. S. van Zon, M. J. Morelli, S. Tanase-Nicola, and P. R. ten Wolde, Biophysical Journal **91**, 4350 (2006).
[17] S. Chong, C. Chen, H. Ge, and X. S. Xie, Cell **158**, 314 (2014).
[18] T. Fukaya, B. Lim, and M. Levine, Cell **166**, 358 (2016).
[19] C. Bartman, S. Hsu, C.-S. Hsiung, A. Raj, and G. Blobel, Molecular Cell **62**, 237 (2016).
[20] A. Papantonis and P. R. Cook, Chemical Reviews **113**, 8683 (2013).
[21] I. I. Cisse, I. Izeddin, S. Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, and X. Darzacq, Science **341**, 664 (2013).
[22] W. K. Cho, N. Jayanth, B. P. English, T. Inoue, J. O. Andrews, W. Conway, J. B. Grimm, J. H. Spille, L. D. Lavis, T. Lionnet, and I. I. Cisse, eLife **5** (2016), 10.7554/eLife.13617.
[23] A. J. Plys and R. E. Kingston, Science **361**, 329 (2018).
[24] M. Boehning, C. Dugast-Darzacq, M. Rankovic, A. S. Hansen, T. Yu, H. Marie-Nelly, D. T. McSwiggen, G. Kokic, G. M. Dailey, P. Cramer, X. Darzacq, and M. Zweckstetter, Nature Structural and Molecular Biology **25**, 833 (2018).
[25] S. Chong, C. Dugast-Darzacq, Z. Liu, P. Dong, G. M. Dailey, C. Cattoglio, A. Heckert, S. Banala, L. Lavis, X. Darzacq, and R. Tjian, Science **361**, eaar2555 (2018).
[26] B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty, and R. A. Young, Science **361**, eaar3958 (2018).
[27] W. K. Cho, J. H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, Science **361**, 412 (2018).
[28] P. Cramer, Nature (2019), 10.1038/s41586-019-1517-4.
[29] Y. Shin, Y. C. Chang, D. S. Lee, J. Berry, D. W. Sanders, P. Ronceray, N. S. Wingreen, M. Haataja, and C. P. Brangwynne, Cell **175**, 1481 (2018).
[30] M. J. Moore and N. J. Proudfoot, Cell **136**, 688 (2009).
[31] J. N. Kuehner, E. L. Pearson, and C. Moore, Nature Reviews Molecular Cell Biology **12**, 283 (2011).
[32] C. K. Mapendano, S. Lykke-Andersen, J. Kjems, E. Bertrand, and T. H. Jensen, Molecular Cell **40**, 410

(2010).
[33] X. Fang, L. Wang, R. Ishikawa, Y. Li, M. Fiedler, F. Liu, G. Calder, B. Rowan, D. Weigel, P. Li, and C. Dean, Nature **569**, 265 (2019).
[34] D. Hebenstreit, Trends in Genetics **29**, 333 (2013).
[35] C. K. Damgaard, S. Kahns, S. Lykke-Andersen, A. L. Nielsen, T. H. Jensen, and J. Kjems, Molecular Cell **29**, 271 (2008).
[36] D. Carter, L. Chakalova, C. S. Osborne, Y.-f. Dai, and P. Fraser, Nature Genetics **32**, 623 (2002).
[37] A. Reik, A. Telling, G. Zitnik, D. Cimbora, E. Epner, and M. Groudine, Molecular and cellular biology **18**, 5992 (1998).
[38] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat, Molecular cell **10**, 1453 (2002).
[39] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp, Cell **169**, 13 (2017).
[40] Y. Wang, J. A. Fairley, and S. G. E. Roberts, Current biology : CB **20**, 548 (2010).
[41] A. Ansari and M. Hampsey, Genes and Development **19**, 2969 (2005).
[42] S. Wani, M. Yuda, Y. Fujiwara, M. Yamamoto, F. Harada, Y. Ohkuma, and Y. Hirose, PLoS ONE **9**, e106040 (2014).
[43] S. M. Tan-Wong, J. B. Zaugg, J. Camblong, Z. Xu, D. W. Zhang, H. E. Mischo, A. Z. Ansari, N. M. Luscombe, L. M. Steinmetz, and N. J. Proudfoot, Science **338**, 671 (2012).
[44] H. M. Shapiro, *Practical Flow Cytometry* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2003).
[45] K. Lo, R. R. Brinkman, and R. Gottardo, Cytometry Part A **73A**, 321 (2008).
[46] S. Tiberi, M. Walsh, M. Cavallaro, D. Hebenstreit, and B. Finkenstädt, Bioinformatics **34**, i647 (2018).
[47] J. Kim and J. C. Marioni, Genome Biology **14**, R7 (2013).
[48] M. Dobrzynski and F. J. Bruggeman, Proceedings of the National Academy of Sciences **106**, 2583 (2009).
[49] D. Zenklusen, D. R. Larson, and R. H. Singer, Nature Structural and Molecular Biology **15**, 1263 (2008).
[50] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, Cell **161**, 1187 (2015).
[51] R. D. Dar, S. M. Shaffer, A. Singh, B. S. Razooky, M. L. Simpson, A. Raj, and L. S. Weinberger, PLoS ONE **11**, e0158298 (2016).
[52] R. D. Dar, B. S. Razooky, L. S. Weinberger, C. D. Cox, and M. L. Simpson, PLoS ONE **10**, e0140969 (2015).
[53] M. Soltani, C. A. Vargas-Garcia, D. Antunes, and A. Singh, PLoS Computational Biology **12**, e1004972 (2016).
[54] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, and N. Barkai, Nature Genetics **38**, 636 (2006).
[55] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, Nature **441**, 840 (2006).
[56] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, Science **329**, 533 (2010).
[57] J. Stewart-Ornstein, J. S. Weissman, and H. El-Samad, Molecular cell **45**, 483 (2012).
[58] S. Yang, S. Kim, Y. Rim Lim, C. Kim, H. J. An, J.-H. Kim, J. Sung, and N. K. Lee, Nature Communications **5**, 4761 (2014).
[59] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C.-L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W.-K. Sung, M. Snyder, and Y. Ruan, Cell **148**, 84 (2012).
[60] G. Li, Y. Chen, M. P. Snyder, and M. Q. Zhang, Nucleic Acids Research **45**, e4 (2017).
[61] M. S. Bartlett, Biometrics **3**, 39 (1947).
[62] E. Kussell and S. Leibler, Science **309**, 2075 (2005).
[63] A. Eldar and M. B. Elowitz, Nature **467**, 167 (2010).
[64] H. J. E. Beaumont, J. Gallie, C. Kost, G. C. Ferguson, and P. B. Rainey, Nature **462**, 90 (2009).
[65] K. F. Murphy, R. M. Adams, X. Wang, G. Balázsi, and J. J. Collins, Nucleic Acids Research **38**, 2712 (2010).
[66] L. Bandiera, S. Furini, and E. Giordano, Frontiers in Microbiology **7**, 479 (2016).
[67] A. Patil, D. Huard, and C. Fonnesbeck, Journal of Statistical Software **35**, 1 (2010).