68–83 *Nucleic Acids Research, 2023, Vol. 51, No. 1* *Published online 30 December 2022*
https://doi.org/10.1093/nar/gkac1204

Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics

Songhao Luo [1,2,†], Zihao Wang[1,2,†], Zhenquan Zhang[1,2], Tianshou Zhou[1,2,*] and Jiajun Zhang [1,2,*]

[1]Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou, 510275, P. R. China and [2]School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong Province, 510275, P. R. China

Received July 01, 2022; Revised November 06, 2022; Editorial Decision November 26, 2022; Accepted December 06, 2022

ABSTRACT

Gene expression in mammalian cells is highly variable and episodic, resulting in a series of discontinuous bursts of mRNAs. A challenge is to understand how static promoter architecture and dynamic feedback regulations dictate bursting on a genome-wide scale. Although single-cell RNA sequencing (scRNA-seq) provides an opportunity to address this challenge, effective analytical methods are scarce. We developed an interpretable and scalable inference framework, which combined experimental data with a mechanistic model to infer transcriptional burst kinetics (sizes and frequencies) and feedback regulations. Applying this framework to scRNA-seq data generated from embryonic mouse fibroblast cells, we found Simpson's paradoxes, i.e. genome-wide burst kinetics exhibit different characteristics in two cases without and with distinguishing feedback regulations. We also showed that feedbacks differently modulate burst frequencies and sizes and conceal the effects of transcription start site distributions on burst kinetics. Notably, only in the presence of positive feedback, TATA genes are expressed with high burst frequencies and enhancer–promoter interactions mainly modulate burst frequencies. The developed inference method provided a flexible and efficient way to investigate transcriptional burst kinetics and the obtained results would be helpful for understanding cell development and fate decision.

INTRODUCTION

The gene-expression variability resulting from programmed and stochastic processes has emerged as a central preoccupation for investigating gene regulation (1,2). Genes are stochastically transcribed often in a discontinuous bursting manner (3,4). Transcriptional bursting is regarded as a primary proxy of stochasticity in gene expression and contributes to cell-to-cell variability (5–7), but the molecular mechanisms governing transcriptional bursting kinetics still remain elusive. Many experimental studies have provided evidence for linking static promoter architecture and sequence to transcriptional bursting and, therefore, to the resulting variability in gene expression (5,8). This variability can propagate from mRNA to protein and further to the downstream target genes via a complex regulatory network (9,10). This raises important issues: On the genome-wide scale, how do static promoter regulatory sequences encode transcriptional burst kinetics, and how do dynamic gene regulatory networks shape burst kinetics?

An intuitive view is that there is an indispensable link of gene-expression variability to promoter architecture (11,12). This link is due to the fact that a basic step of RNA synthesis is to copy the genetic information from the gene promoter. Much effort has been devoted to rationalizing the promoter-architecture encoding of transcriptional burst kinetics on genome-wide scales. For example, genes with TATA boxes increase variability in expression levels, whereas the presence of CpG island significantly lowers the variability (13–15). The sharp distributions of transcription start sites (TSS) lead to higher gene-expression variability than the broad TSS distributions (13). A recent study (16) has revealed that the increases in burst sizes are dependent on the presence of TATA box and initiator elements (characteristics of the core promoter), and burst frequencies are regulated by enhancer–promoter (E–P) interactions. All these studies and others (17–21) indicate the importance of promoter architecture in modulating transcriptional burst kinetics.

Another viewpoint is that feedback regulations modulate transcriptional burst kinetics by creating a higher-level

*To whom correspondence should be addressed. Tel: +86 20 84111829; Email: zhjiajun@mail.sysu.edu.cn
Correspondence may also be addressed to Tianshou Zhou. Tel: +86 20 84134958; Email: mcszhtsh@mail.sysu.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

structure regulatory pattern (10). In fact, feedback regulations exist extensively in biological systems, and their functions may be reflected by the circuits of interacting genes and proteins (22). In particular, auto-regulatory feedback loops have been identified in various regulatory systems, where transcription factors directly or indirectly regulate the corresponding gene expression (23). In general, feedbacks can be categorized into positive and negative ones. Experimental investigations for a few genes or transcripts have demonstrated that different kinds of feedback played diverse roles (10). For example, negative feedback limits large expression variability and accelerates responses (24–26). Conversely, positive feedback amplifies expression variation, induces bimodal expression, and stimulates genes to 'active' states (27–30). In addition, negative feedback with a long delay loop can display increased variability (31). Theoretical analysis has also shown that different feedback mechanisms modulate burst kinetics in different manners (32,33). All these studies indicate the important roles of feedback regulations in mediating gene expression, including transcriptional bursts, but it is unclear whether the results obtained for case-by-case studies can hold on a genome-wide scale.

The above two viewpoints are not solitary but are complementary to each other. A challenging task is to investigate how static promoter architecture and dynamic feedback regulation coordinate transcriptional burst kinetics on a genome-wide scale. Previous studies of transcriptional bursting were limited to low-throughput experimental approaches, where observed experimental results could not be generalized across different genes or cell types (34–40). Recently, single-cell RNA sequencing (scRNA-seq) has enabled the in-depth measurement of expression levels within cell populations, providing an opportunity to study genome-wide transcriptional mechanisms (41). An important step toward this study is to develop mathematical models for the genome-wide inference of burst kinetics. The models for inference should satisfy some requirements. First, these models should be interpretable, i.e. they can capture essential gene-expression dynamics and convey kinetic information about transcriptional bursts (16,42–44) (https://doi.org/10.1101/2021.09.06.459173). Previous studies relied on inferring the direct correlations between features across molecular scales (13,45). However, these correlations are insufficient to uncover the mechanisms of gene expression. Second, the inference models should be tractable, i.e. they can effectively treat a large number of cells and genes. In general, a complex mechanistic model incorporating regulatory factors is difficult to analyze on the one hand (46), and a genome-wide inference needs expensive computational cost on the other hand. Therefore, an interpretable and tractable inference framework integrating experimental data and molecular mechanisms is strongly demanded.

Here we developed a statistical framework based on the model-driven and data-driven combination to perform a scalable genome-wide inference. This framework used the static snapshots of scRNA-seq data to infer the regulatory mechanisms underlying transcriptional burst kinetics. Specifically, it integrated the expected information on gene-expression variability, burst frequencies, burst sizes, and feedback regulation forms. Applying this inference method to the scRNA-seq data generated from embryonic mouse fibroblast cells (16), we showed that feedbacks differently modulate burst frequencies and sizes, TATA genes are expressed with high burst frequencies only in the presence of positive feedback, feedback regulations conceal the effects of TSS distribution on transcriptional burst kinetics, and E–P interactions mainly modulate burst frequencies only in the presence of positive feedbacks. Briefly, we found that characteristics of genome-wide transcriptional burst kinetics in the case without feedback regulations were different from those in the case with feedback regulations, implying Simpson's Paradox, an interesting statistical phenomenon.

MATERIALS AND METHODS

A mechanistic hierarchic model for statistical inference

The observed counts in a scRNA-seq experiment reflect a combination of the true expression level and the measurement level of each gene in each cell. We describe the observed counts by a two-level hierarchical model (See details in Supplementary Text 1.1, Figure 1D, and Supplementary Figure S1a–c):

$$P(Y=y)=\int_0^\infty P_{\text{meas}}(y\,|n)\,P_{\text{gene}}(n)dn, \tag{1}$$

where $P_{\text{meas}}(y|n)$ is for a measurement model and $P_{\text{gene}}(n)$ for a gene expression model.

The first level represents the measurement process for the observed count Y_{cg} conditional on the true expression level N_{cg} of gene g in cell c, with a conditional probability distribution (Supplementary Figure S1b):

$$Y_{cg}\big|\,N_{cg}\sim P_{\text{meas}}(y\,|n)\,. \tag{2}$$

The $P_{\text{meas}}(y|n)$ describes all aspects of the technical noise produced in the measurement process for a given true expression level N_{cg}, and is suggested as a Binomial distribution or Poisson distribution which is supported by empirical analyses and theoretical arguments in many existing methods (47,48). By adding an extra sampling probability λ_{cg} in the sequencing process, we characterize the sequencing depth and assume that intercellular molecules are independent of each other and only the proportional products are captured and sequenced using Binomial distribution

$$Y_{cg}\big|\,N_{cg}\sim \text{Binomial}\left(n,\lambda_{cg}\right). \tag{3}$$

In our calculations, we set the sampling probability $\lambda_{cg}=\lambda=0.5$ without loss of generality since the setting of λ_{cg} does not influence our qualitative results.

The second level is the true expression level of gene g across cells, which is assumed to follow a probability distribution

$$N_{cg}\sim P_{\text{gene}}(n)\,. \tag{4}$$

The underlying model describes the intrinsic dynamics of stochastic gene expression. How an appropriate gene-expression model $P_{\text{gene}}(n)$ is chosen is critical. In general, this choice needs to satisfy two basic requirements: (i) the model should capture the essential gene-expression dynamics of interest (e.g. transcriptional burst kinetics); and (ii) the inference based on the model should be effective and scalable to large numbers of cells and genes. As combining mechanistic models to infer the entire gene regulatory

network would lead to sophisticated models that become intractable, we simplified the gene regulatory network to a feedback loop, which is the most common form existing in gene expression systems. For these purposes, we adopt a model of stochastic gene expression (Supplementary Text 1.1, Figure 1B and Supplementary Figure S1c), which simultaneously characterizes transcriptional burst kinetics and auto-regulatory feedbacks with the below distribution

$$P_{\text{gene}}(n) = \mathcal{N}\int_0^{\infty} \text{Poisson}(n|x)\, x^{a(1+\varepsilon)-1} e^{-x/b}\left(1+(x/k)^h\right)^{-a/h} dx, \quad (5)$$

where $\mathcal{N}$ is a normalization factor. Note that the discrete gene expression distribution (Equation (5)) is a Poisson representation in form, i.e. $P_{\text{gene}}(n) = \int_0^{\infty} \text{Poisson}(n|x) f(x) dx$, where $f(x)$ is a kernel density function that has the same form as the continuous distribution of proteins in (49). As illustrated in Figure 1B, this kernel function $f(x)$ can extract several kinetic parameters, denoted by $\theta = (a, b, \varepsilon, k, h)$, of the steady-state gene-product distribution from a dynamic model with auto-regulatory feedback described by some meaningful kinetic parameters: switching rates between inactive and active state (k_{on}, k_{off}), mRNA transcription rates (k_{syn}) and degradation rates (k_{deg}). Here, a is the number of bursts per cell cycle (burst frequency), and b is the mean number of gene products generated per burst (burst size), and h is a vital parameter of capturing the feedback regulation dynamics of gene products, which is actually a Hill coefficient. Furthermore, this model can describe two most common feedback loops in gene expression: positive-feedback loop (i.e. $hlt; 0$) and negative feedback loop (i.e. $h > 0$). It should be noted that the auto-regulatory feedbacks involve gene products, which directly or indirectly regulate the corresponding target gene itself through feedback loops, resulting in a repressing or activating expression. The small leakiness proportion of the promoter ε contains the information on the baseline bursts in the absence of regulation, and k contains the information on the equilibrium binding constant (see details in Supplementary Text 1.1). Overall, Equation (5) is a mechanistic model, which can simultaneously describe the burst-production and feedback-regulation processes of gene expression. More importantly, as a special case of this model, $h = 0$ corresponds to the negative binomial distribution of gene expression $P_{\text{gene}}(n)$ (i.e. non-feedback).

By combining Equations (3) and (5) and substituting into Equation (1), the discrete probability distribution of Y_{cg} can be computed but is expressed in an integral form (see details in Supplementary Text 1.1):

$$\begin{aligned} P(Y=y;\cdot) &= \int_0^{\infty} P_{\text{meas}}(y|n)\, P_{\text{gene}}(n) dn \\ &= \int_0^{\infty} \text{Poisson}(Y=y|\lambda x) f(x)\, dx \\ &= \mathcal{N}\cdot \int_0^{\infty} \frac{(\lambda x)^y}{y!} e^{-\lambda x} x^{a(1+\varepsilon)-1} e^{-x/b}\left(1+(x/k)^h\right)^{-a/h} dx. \end{aligned} \quad (6)$$

For the case of non-feedback, we employ the negative binomial distribution, which is then given by

$$P(Y=y;\cdot) = \int_0^{\infty} \frac{(\lambda x)^y}{y!} e^{-\lambda x} \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b} dx. \quad (7)$$

where $\Gamma(\cdot)$ is the Gamma function.

Maximum likelihood estimation of parameters

Here we introduce a method of estimating the kinetic parameters in our hierarchical model using the expression data of each gene. For a given expression read of observed cells, the most common parameter estimation method is the maximum likelihood estimation, which can be formulated as the following optimization problem of five parameters $\theta = (a, b, \varepsilon, k, h)$ in our case

$$\arg\min_{\theta}(-L(y;\theta)) = \arg\min_{\theta} \sum_{y} -\ln(P(Y=y;\theta)), \quad (8)$$

where $P(Y=y;\theta)$ is described in Equation (6).

Because of the complex integral and unnormalized probability mass function in Equation (6), calculating the integral directly through the MCMC method (50) would be at a high cost of computation, and in particular, it is hard to use in the analysis of genome-wide data. Therefore, we apply the Generalized Gauss-Laguerre Quadrature Rules (51) to Equation (6) instead of the use of the MCMC method, realizing a rapid calculation in inference:

$$\begin{aligned} P(Y=y;\theta) &= \mathcal{N}\int_0^{\infty} \frac{(\lambda x)^y}{y!} e^{-\lambda x} x^{a(1+\varepsilon)-1} e^{-x/b}\left(1+(x/k)^h\right)^{-a/h} dx \\ &\approx \mathcal{N}\sum_{i=1}^{n} w_i f(x_i), \end{aligned} \quad (9)$$

where x_i, w_i can be determined by the generalized Laguerre polynomials.

In particular, a simple algebraic transformation in Equation (7) yields the explicit expression of the probability distribution in the case of non-feedback:

$$\begin{aligned} P(Y=y;\cdot) &= \int_0^{\infty} \frac{(\lambda x)^y}{y!} e^{-\lambda x} \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b} dx \\ &= \frac{(b\lambda)^y}{(a+y)\, B(a, y+1)\, (b\lambda+1)^{y+a}}, \end{aligned} \quad (10)$$

where $B(\cdot,\cdot)$ is the Beta function.

Optimization method and initial values setting

To realize a fast calculation for solving the optimization problem (Equation (8)) of parameter estimation, we use the *fmincon* function in the LBFGS method of MATLAB (https://www.mathworks.com/products/matlab.html), a nonlinear programming solver, to find the minimum of the optimization problem given a set of initial values and parameter intervals $a = (10^{-1}, 30)$, $b = (1, 20)$, $k = (1, 10^3)$, $h = (-10, -1)\text{or}(1, 10)$. For each gene and each case of positive, negative, and non-feedbacks, we repeatedly solve the optimization problem 30 times.

We restrict that the initial values of the optimization problem obey the following rules. First, we consider initial

points of a (burst frequency) and b (burst size). Here X is a random variable of the distribution in Equation (5). Since Gamma distribution is a special case of Equation (5), we assume that the initial points a and b follow

$$\mathrm{E}[X] = ab, \mathrm{Var}[X] = ab^2. \quad (11)$$

On the other hand, by considering the initial values of Y generated by our hierarchical model, we have

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y|X]] = \lambda \mathrm{E}[X] = \lambda ab, \quad (12)$$

$$\mathrm{Var}[Y] = \mathrm{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathrm{E}[Y|X]] = \lambda ab + \lambda^2 ab^2. \quad (13)$$

Given the expectation and variance of Y, we can in return estimate burst frequency a and burst size b by rearranging Equations (12) and (13), which are then taken as initial values

$$a = \frac{\mathrm{E}[Y]^2}{\mathrm{Var}[Y] - \mathrm{E}[Y]}, b = \left(\frac{\mathrm{Var}[Y]}{\mathrm{E}[Y]} - 1\right) \Big/ \lambda. \quad (14)$$

Usually, the transcriptional rate in the OFF state is much smaller than that in the ON state. And a small enough leaky rate does not affect the distribution shape of gene expression. Therefore, we fix ε at the constant of 0.05 during our inference. The h value is extracted from a uniform distribution between the integer -5 and -1 (positive feedback) or integers 1 and 5 (negative feedback). And the values of parameter k are extracted from the log-uniform (logarithm base 10) distribution on interval [0, 2].

Model selection

Using the above inference method, we obtain 90 inferred results for each case of positive, negative, and non-feedbacks. Then, we filter out the unreliable results on the inference boundary, which are possibly caused by the optimization program setting. On this basis, we compute the value of the corrected Akaike information criterion (AICc) (52) and select the best model corresponding to the smallest AICc,

$$\mathrm{AICc} = -2 \log L\left(\hat{\theta}\right) + 2k + \frac{2k(k+1)}{n-k-1}, \quad (15)$$

where the maximum likelihood $L(\hat{\theta})$ is the result during the inference run, k is the number of model parameters, and n is the sample size of observed data.

Validation on synthetic scRNA-seq data

In order to check whether the above statistical inference method can effectively infer burst frequency, burst size, and feedback form in our hierarchical model, we produce synthetic single-cell RNA data. Given a set of model parameter values $\theta = (a, b, \varepsilon, k, h)$, we first calculate the probability distributions of these parameters according to the method described in the above section and then carry out random samples according to the probability of each parameter value to obtain the input data for the inference process.

We show the precision regions for the inference of burst kinetics (burst frequencies and burst sizes) under different feedback strengths h and different equilibrium binding constants k (Supplementary Figure S4–S6). And the error between true parameters $\theta^{\mathrm{true}} = (bf^{\mathrm{true}}, bs^{\mathrm{true}})$ and estimated parameters $\theta^{\mathrm{est}} = (bf^{\mathrm{est}}, bs^{\mathrm{est}})$ is calculated according to:

$$\mathrm{Error}\left(\theta^{\mathrm{true}}, \theta^{\mathrm{est}}\right) = \left(\log\left(bf^{\mathrm{true}}\right) - \log\left(bf^{\mathrm{est}}\right)\right)^2 + \left(\log\left(bs^{\mathrm{true}}\right) - \log\left(bs^{\mathrm{est}}\right)\right)^2. \quad (16)$$

We show the robustness of the inference in the cases of positive, negative, and non-feedbacks, respectively (Supplementary Figure S4–S6). To explore the robustness of the cell numbers to the inference, we select different sampled cell numbers (200, 300, 500, 1000, 5000) to synthesize data 50 times, and at each time, set 30 different initial points for optimization in each case of feedback forms. The optimization process is the same as the inference process of real data. The same process is used to explore the effects of stochastic losses of mRNA molecules (sensitivity), missing randomly at a certain probability (0.1, 0.3, 0.5, 0.7 or 0.9) from sufficient samples (number = 2000). The results of inference robustness analysis are illustrated with two different distribution examples in the three cases of feedback forms (Supplementary Figure S4–S6).

Inference evaluation

To assess whether the observed data came from the distribution generated via the parameters inferred by our method, we use goodness-of-fit statistics that obey chi-square distribution of large samples:

$$\chi^2 = \sum_{k=0}^{\infty} \frac{(O_k - E_k)^2}{E_k}, \quad (17)$$

where O_k is the observed sample number whose mRNA number is k, and E_k is the expected sample number. Note that in some sequencing techniques, the cell samples of scRNA-seq data are not large enough, so it is needed to use the Monte Carlo method to generate the null distribution of chi-square goodness-of-fit test instead of the asymptotic distribution. For each gene, we first generate the same number of samples as that in the observed data from the probability of each point with the inferred parameters and then compute the χ^2_{sim} statistic according to Equation (17). After repeating the Monte Carlo simulation procedure for 1000 times, we judge whether the resulting inference is a good fit by comparing χ^2_{obs} with the resulting 1000 χ^2_{sim}. The criterion that an inferred parameter is a good fit is that the χ^2_{obs} is at least less than five percentage numbers of χ^2_{sim} (Supplementary Figure S7a).

Data analysis

scRNA-seq data processing. We utilize the processed scRNA-seq data for 10727 genes of transcriptomes from 224 individual mouse embryonic fibroblasts for each allele (C57 × CAST) (16). In that paper, the quantification of gene transcription is based on the Smart-seq2 scRNA-seq libraries, and UMI counts is used to reduce the amplification noise. To ensure that the inference process is not hin-

dered by low-quality elements of the data as far as possible, we carry out a certain degree of quality control of the original data (from the file: SS3_cast_UMIs_concat.csv and SS3_c57_UMIs_concat.csv). We filter out the genes expressed in less than 50 cells. Also, we filter out the cells expressed in <2000 genes. In addition, we filter out the genes whose overall average expression levels are <2. After these manipulations on each allelic data (C57 × CAST), the genes that meet the conditions are combined to facilitate inferences from more adequate samples and give a single-cell expression matrix composed of 2162 genes and 413 cells. This treatment is based on the assumption that the distributions of almost all genes for the CAST and c57 alleles have similar shapes and that the transcriptional dynamic behavior is consistent between alleles for most genes, which is also supported by previous studies (16,53). And, we removed the outlier data with the tail 5% of the distribution. In addition, our method can be also applied to any high-quality non-allelic scRNA-seq data.

Identification of promoter motif and TSS distribution. The recognition and coordinates of the promoter motifs (TATA box, Initiator, CCAAT box, GC box) are downloaded from 'the Select/Download Tool' of the EPD New database (54). In order to determine the TSS distribution of mouse embryonic fibroblasts, MEFs FANTOM5 Cap Analysis of gene expression data is retrieved through the CAGEr R package (55). After normalization and TSS clustering, TSS distribution is defined as 'sharp' if the promoter width is less than 15bp (this length is taken as the median of all genes), and as 'broad' otherwise.

Identification of enhancer–promoter intensity. The data about the interaction between enhancer and promoter is downloaded from (16). The dataset is used to compare the correlation between burst kinetics and enhancer activity of fibroblasts and mESCs. Enhancer activity is calculated according to the intensity of the H3K27ac peak measured in the defined EPUs region (which is considered that enhancer and promoter interactions occur more possibly) via ChIP-seq in a previous study (56). In our study, we only utilize the collated data that includes the peak of H3K27ac in EPUs of MEFs.

Statistical analysis

Gene expression variability. Gene expression variability is usually quantified by the square of the coefficient of variation (CV^2), which is defined as the ratio of the variance over the square of the mean. According to this definition, we calculate gene-expression variability in a given set of observed data Y for gene g. Similarly, we use the inferred θ^{est} to calculate the theoretical CV^2 of our hierarchical model for gene g, that is,

$$\mathrm{CV}_g^2 = \frac{\mathrm{Var}\left[Y;\theta_g^{est}\right]}{\mathrm{E}\left[Y;\theta_g^{est}\right]^2}. \tag{18}$$

When fitting CV^2 with a cubic spline, we find that there is a strong correlation between the mean expression level and CV^2 (Supplementary Figure S9a). Many studies have discussed the relationship between gene-expression variability and mean (57,58). Note that in the classical telegraph model, the total mRNA variability can be decomposed into two parts: the mRNA internal variability generated from transcription and the promoter variability due to the switching between active and inactive states. Inspired by (15), we adjust the variability by subtracting the inverse of the logarithmic mean (logarithm base 2), thus obtaining the residual squared coefficient of variation ($r\mathrm{CV}^2$). For example, for gene g, we have

$$r\mathrm{CV}_g^2 = \mathrm{CV}_g^2 - \frac{1}{\log_2\left(\mu_g\right)}, \tag{19}$$

where $\mu_g = \mathrm{E}[Y;\theta_g^{est}]$. As a result, the influence of the mean expression level on the expression variability is basically eliminated after performing a linear regression on $r\mathrm{CV}^2$ (Supplementary Figure S9b).

Linear regression model in promoter motif analysis. After having obtained the promoter motifs of each gene from the EPD database and its burst kinetics ($r\mathrm{CV}^2$, burst frequency, burst size) by inference, we conduct multivariate linear regression with interaction terms to find the correlations between quantities of interest in cases of positive, negative, and non-feedbacks. Specifically, we perform multivariate linear regression according to

$$\begin{aligned} r\mathrm{CV}^2 &\sim (TATA * Inr + CCAAT * GC) \times feedback, \\ \log_{10}(bf) &\sim (TATA * Inr + CCAAT * GC) \times feedback, \\ \log_{10}(bs) &\sim (TATA * Inr + CCAAT * GC) \times feedback. \end{aligned} \tag{20}$$

In Figure 4B–D and Supplementary Figure S10c–e, we show the t-value in the regression results. The absolute of t-value is larger in the test of the linear regression coefficient, indicating that the resulting correlations are significant.

RESULTS

An integrated statistical framework for learning promoter-dependent yet feedback-constrained transcriptional burst kinetics on a genome-wide scale

Cell-to-cell heterogeneity in gene expression is primarily attributed to transcriptional bursting (12,59,60), which is represented by a vector $\boldsymbol{y}$ of components including burst frequency, burst size, expression variability, etc. (Figure 1C). Transcriptional bursts result from complex molecular processes on multilayered sources (1), which are represented by a vector $\boldsymbol{x}$ of components including static DNA sequences, epigenetic modifications (61), transcription, translation, dynamic network regulations, etc. (2). Then, the question of how these molecular processes coordinate transcriptional busting can be mathematically described as $\boldsymbol{y} = f(\boldsymbol{x})$, where f is a vector function describing the correlation of $\boldsymbol{x}$ to $\boldsymbol{y}$.

Static promoter architecture is an essential DNA sequence for binding transcription factors during mRNA synthesis. Specifically, promoter motifs (such as initiator, TATA-box, CCAAT-box, GC-box), broad and sharp TSS

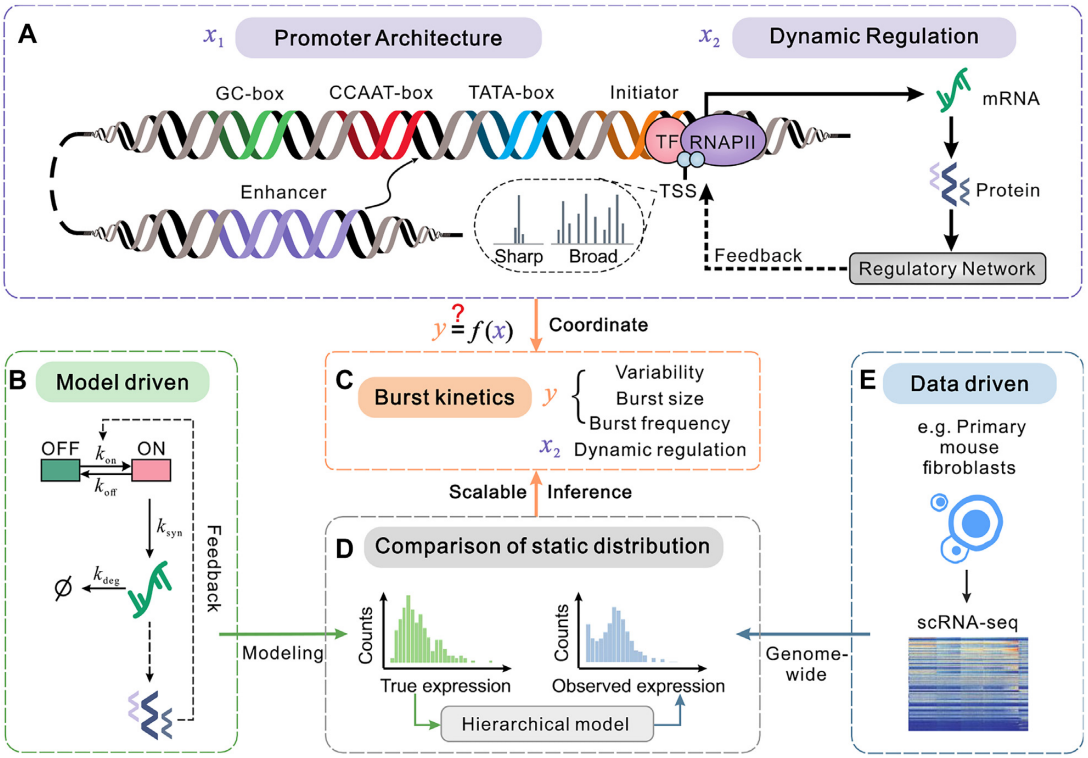

Figure 1. Overview of a scalable genome-wide inference method. (**A**) Schematic for important ingredients in gene expression process, including static promoter architecture information and dynamic regulation. Promoter architecture (represented by $\boldsymbol{x}_1$) consists of promoter motifs (Initiator, TATA-box, CCAAT-box, and GC-box), TSS distributions ('sharp' and 'broad') and enhancer–promoter interactions. Dynamic regulation (represented by $\boldsymbol{x}_2$) is referred to as a series of processes, such as transcription and translation as well as feedback loops, in which the gene product (as a transcription factor) regulates its own expression, possibly via a complex regulatory network. (**B**) Model-driven: schematic for a mechanistic model of stochastic gene expression, which considers an active (ON) state and an inactive (OFF) state of the promoter and auto-regulatory feedback. Here k_{on} is the switching rate from OFF to ON and k_{off} from ON to OFF; k_{syn} is the transcription rate when the gene is in ON state and k_{deg} is the degradation rate of mRNAs. (**C**) Kinetic parameters to be inferred, which include expression variability, burst frequency, burst size, and dynamic regulation. (**D**) Comparison between two static distributions (the left panel is for 'true' mRNA levels in the mechanistic gene model and the right panel for 'observed' mRNA counts in a given set of scRNA-seq data) by a hierarchical model can determine the values of the kinetic parameters in (C) via a scalable genome-wide inference method. (**E**) Data-driven: genome-wide scRNA-seq data of mouse embryonic fibroblasts gives an expression matrix that further gives the observed static distribution in (D).

distributions, and enhancer–promoter interactions are essential features of eukaryotic promoter architecture (Figure 1A, left). Meanwhile, variability in gene expression can propagate from mRNA to protein and further to target genes, possibly through a dynamic and complex gene regulatory network. A common form of dynamic regulation is auto-regulation which directly or indirectly regulates the corresponding target gene itself through a feedback loop, resulting in a repressing or activating expression (Figure 1A, right). For clarity, we let vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ represent static promoter architecture and dynamic feedback regulation, respectively (Figure 1A). The information on promoter architecture ($\boldsymbol{x}_1$) can be recovered from public bioinformatics databases such as the EPD database (54), Bioconductor (62), and UCSC Genome Browser (63). In general, the mechanisms of dynamic regulation ($\boldsymbol{x}_2$) and burst kinetics ($\boldsymbol{y}$) are not directly measurable but hidden in data sets. Unlike some imaging-based technologies such as MS2 system (64) that were limited to a few genes and could not be extended to the whole genome, single-cell sequencing technologies made it possible to recover the information on dynamic regulations ($\boldsymbol{x}_2$) and burst kinetics ($\boldsymbol{y}$) from static snapshots (Figure 1E). Figure 1B–E summarizes the genome-wide inference procedure proposed here. This procedure used a statistical framework of the model-driven (Figure 1B) and data-driven (Figure 1E) combination to infer dynamic feedback regulations and transcriptional burst kinetics from static scRNA-seq data (Figure 1C, D) under the assumption that the abundances of mRNA and protein were highly dependent (65).

Specifically, our statistical inference framework used a mechanistic model of gene expression (Figure 1B), which simultaneously considered transcriptional burst kinetics ($\boldsymbol{y}$) and feedback regulations ($\boldsymbol{x}_2$), to obtain 'true' gene expression distributions (Figure 1D, left). On the other hand, the known scRNA-seq data gave 'observed' gene expression distributions, implying possible errors in the sequencing technologies (66,67). A hierarchical statistical model (see 'Materials and Methods') was proposed to link 'true' gene-expression levels (Figure 1D, left) and 'observed' mRNA counts (Figure 1D, right), thus determining key kinetic parameters (expression variability, burst size, burst frequency, and dynamic regulation) (Figure 1C). We emphasized that the proposed framework was a scalable genome-wide inference, which was particularly useful in revealing how both static promoter architecture and

dynamic feedback regulation coordinate transcriptional bursting.

A hierarchical model provides the genome-wide inference of transcriptional burst kinetics and feedback regulations from single-cell snapshots

The hierarchical statistical model developed here can give a mechanistic interpretation for Unique Molecular Identifiers (UMIs) based on scRNA-seq data. In fact, this model not only captured the characteristics of transcriptional burst kinetics and feedback regulations, but also described the measured noise of UMIs data (see 'Materials and Methods'). Then, we used the maximum likelihood method to determine burst kinetics and feedback forms (positive-, negative-, non-feedback) within biologically reasonable ranges of model parameters. Note that the inferences with traditional MCMC methods (50) would need huge and even unaffordable computational costs since the static mRNA distribution was expressed in a high-order integral that is difficult to solve. To overcome this difficulty, we developed a fast algorithm for computing this distribution based on generalized Gauss-Laguerre quadrature rules, thus realizing a scalable genome-wide inference (51) (see 'Materials and Methods').

To evaluate the validity of our inference method, we first explored the sensitivity of distribution shapes to changes in model parameters. We found that the genes with high expression levels were more sensitive to model parameters than the other genes (Supplementary Figures S2 and S3). Then, to test the reliability of the method in inferring kinetic parameters, we generated synthetic single-cell RNA data by stochastic sampling from the distribution for the hierarchical model with known parameter values. Through inference using the synthetic data, we can give robust estimates of burst frequencies, burst sizes, and feedback forms from the corresponding static mRNA distributions (Supplementary Figure S4–S6). Besides, we also assessed the robustness of our inference method to different cell numbers and stochastic losses of mRNA molecules (mimicking the incomplete mRNA detection in scRNA-seq protocols) (Supplementary Figures S4b–c, S5b–c and S6b–c). Overall, we provided a mechanistic model and an effective, robust and scalable inference method for learning dynamic burst kinetics and feedback forms from static snapshot data, which can be conveniently used in the analysis of scRNA-seq data.

Next, we applied our hierarchical model and inference approach to the scRNA-seq data of primary mouse fibroblasts (16). From the original UMIs data containing 10727 genes and 224 cells, we selected 2162 highly expressed genes using a quality control method and then merged two allelic expression data into a matrix to infer burst kinetics and feedback forms. We observed that these selected genes were transcribed with widely different burst kinetics (68), and in particular, those genes with the same average expression level exhibited diverse burst kinetics, implying that the expressions of different genes were regulated possibly by different molecular mechanisms (Figure 2A). To check the validity of these inferred results, we performed a goodness-of-fit test (see 'Materials and Methods'). We found that the distributions from the dataset were consistent with those obtained using the inferred parameters (Supplementary Figure S7a), and confirmed that the mRNA mean and variability in the mechanistic model matched those in the data (Supplementary Figure S7b). All the good-fit genes can be classified into three categories: 626 positive-feedback genes, 625 negative-feedback genes, and 840 non-feedback genes. The inferred results for example genes: Mbnl2, Prr13, Ralb, and Plod1 were demonstrated in Figure 2a_1-a_4, showing that these genes had different feedback forms and followed different distributions. Interestingly, our hierarchical model can particularly recover bimodal distributions from static data, which however were fitted as unimodal distributions via the telegraph model without feedback (69) (e.g. the distribution of the Mbnl2 gene as shown in Figure 2a_1 and more genes as shown in Supplementary Figure S8). In addition, we compared the inferred results between our hierarchical model and the telegraph model, finding that both models captured almost the same gene-expression variability (CV^2, Figure 2B) while keeping high correlations between burst frequencies and burst sizes (Figure 2C, D, P-value $< 2.2 \times 10^{-16}$). Notably, we found that the forms of dynamic feedback regulations can lead to different burst kinetics on a genome-wide scale but cannot be inferred by previous methods (Figure 2C, D) (16,43).

Feedbacks modulate burst frequencies and sizes differently

Having inferred each gene's burst kinetics and feedback forms, we next investigated how feedback regulations affected expression variability (CV^2) and transcriptional burst kinetics on a genome-wide scale. Interestingly, we found the statistical phenomenon of Simpson's paradox. First, we observed from Figure 3A that there were no significant differences in variability distributions between the positive-feedback and the negative-feedback genes, but the non-feedback genes exhibited higher expression variability. The latter result seemed inconsistent with the previous conclusions that positive feedback amplified variability and negative feedback attenuated variability (70). This can be interpreted by the fact that the expression level and the expression variability were negatively correlated (57,58,71) (Supplementary Figure S9a). To show this point, we introduced the average expressed variable by dividing all the selected genes into five equal boxes based on average expression levels and tracked the expression-variability changes when the average gene-expression levels were increased. Then, we found that the expression variability was indeed negatively correlated with the average expression levels, regardless of feedback forms (Figure 3D). Furthermore, the positive-feedback genes showed relatively higher expression variability than the negative-feedback genes at the same expression levels (Figure 3D), consistent with the results obtained in previous studies (70,72).

Next, we checked the genome-wide effects of feedback regulations on transcriptional burst frequencies and burst sizes. Interestingly, we found that positive and negative feedback differently modulated burst frequencies and sizes (Figures 3B, C, and Supplementary Figures S9c, d). Specifically, the burst frequencies of positive-feedback genes were significantly higher than those of negative-feedback genes on the whole genome (Figure 3B) and at the same expression level (Figure 3E). By contrast, the burst sizes of positive-feedback genes were smaller than those of

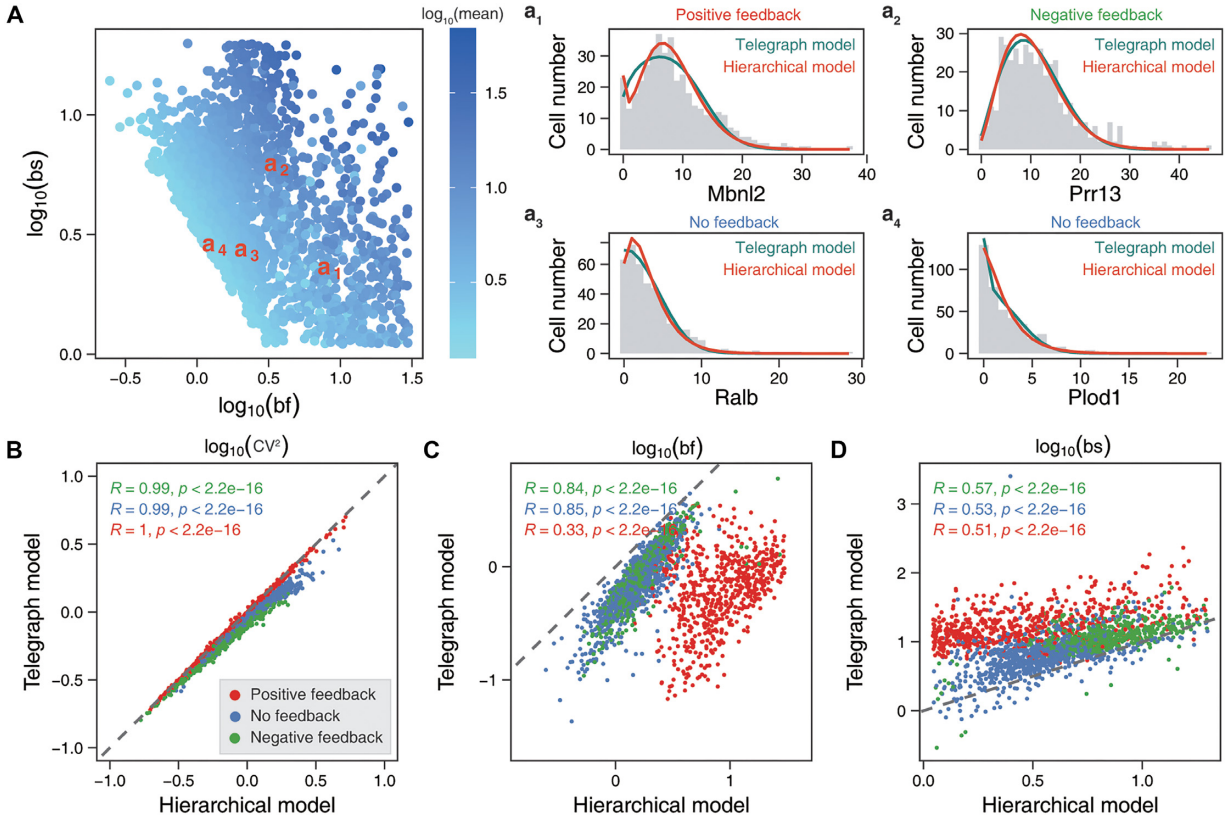

Figure 2. Genome-wide characteristics of transcriptional burst kinetics inferred from the scRNA-seq data of primary mouse fibroblasts. (**A**) Scatter plots of burst frequencies (bf) and burst sizes (bs), where the colored points represent mean expression levels. $\mathbf{a_1}$–$\mathbf{a_4}$ Examples for comparison of the inferred distributions between our hierarchical model (orange line) and the telegraph model (green line), where the gray histograms represent the distributions of mRNA counts. (**B–D**) Scatter plots of the expression variability (CV^2, B), burst frequencies (C) and burst sizes (D), which are correlated in the sense of Pearson correlation test (see the indicated values of *R* and *P*-value). The values of these kinetic parameters are obtained via the hierarchical model and the telegraph model, respectively. Red dots correspond to positive feedback, blue dots to non-feedback, and green dots to negative feedback. The slope of dashed lines equals 1.

negative-feedback genes (Figure 3C, F). In addition, the effects of negative feedback and non-feedback on burst frequencies were difficult to distinguish (Figure 3B, E), but there was a significant difference in burst sizes (Figure 3C, F). This observation suggested that burst size could be a distinguishable characteristic between negative-feedback and non-feedback genes.

Finally, in this subsection, we point out that an unexplored issue is how promoter architecture affects transcriptional burst kinetics in the presence of feedback regulation on a genome-wide scale. Below, we address this issue from three aspects: promoter motifs, TSS distributions, and enhancer–promoter interactions in the following.

TATA genes are expressed with high burst frequencies only in the presence of positive feedback

It was reported that promoter motifs such as TATA box and initiator regulated transcriptional bursting directly (13,14,16,57,73). On the other hand, we showed in the previous section that different feedback regulations led to different burst kinetics. This raised an unexplored question: how do promoter motifs modulate transcriptional burst kinetics in the presence of feedback regulation on the genome-wide scale?

We first identified promoter motifs (TATA box, initiator, GC-box, and CCAAT-box) of each gene from the EPD database (54) (see 'Materials and Methods') (Figure 4A). Then, we found that both the TATA box and initiator positively regulated mean transcriptional levels, in line with the results obtained in previous studies (74) (Supplementary Figure S10a). Besides, we verified that the TATA genes with positive feedback had higher proportions than those genes with negative feedback or without feedback, whereas the other promoter motifs were uncorrelated to feedback forms (Supplementary Figure S10b). These results implied that the TATA box was a critical promoter motif for the regulation of transcription by a positive feedback mechanism, which might be supported by the following experimental observation: TATA boxes were enriched in the promoters of genes with fewer transcriptional pauses (75), and the TATA box sequence was specifically bound by the TATA-binding proteins that acted as general transcription factors to facilitate the localization of RNA polymerase II and transcription (76,77).

To investigate the genome-wide effects of promoter motifs on burst kinetics in the presence of feedback regu-

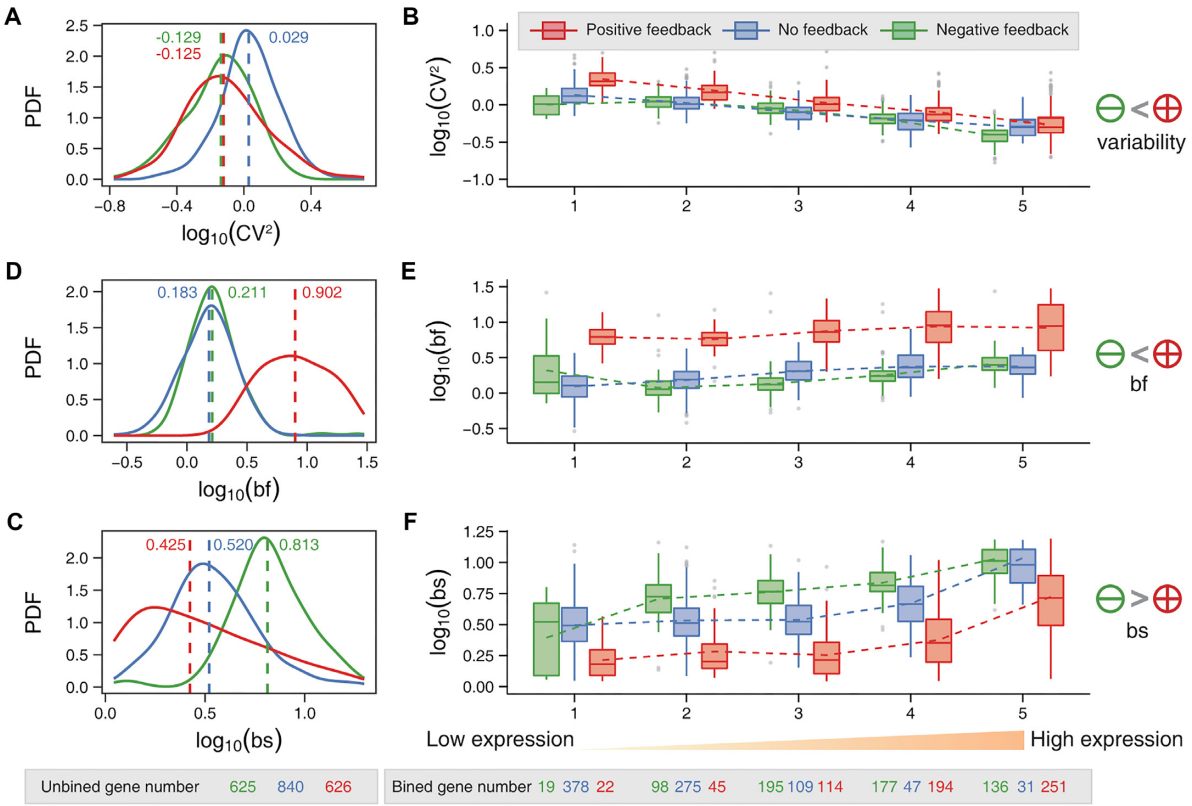

Figure 3. Genome-wide comparison of transcriptional burst kinetics in three cases of feedback regulation. (**A–C**). Three probability density functions (PDF) of expression variability (CV^2, A), burst frequencies (bf, B), and burst sizes (bs, C) for positive-feedback genes (red), non-feedback genes (blue) and negative-feedback genes (green), where dashed lines represent the medians. (**D–F**) Boxplots of expression variability (D), burst frequency (E) and burst size (F). The genes are divided into five boxes with an equal number of genes, and the gene-expression level increases from left to right, where the dashed line connects the mean expression levels in each box. The number of good-fit genes per feedback type is shown at the bottom of the figure.

lations, we performed multivariate statistical analysis using linear regression models (Figure 4B–D, see 'Materials and Methods'). We also observed the Simpson's paradox that the effect of promoter motifs on variability and burst kinetics is different between distinguishing feedback regulation and without distinguishing feedback regulation.

First, we studied gene-expression variability. We characterized this variability with the residual squared coefficient of variation (rCV^2) (see 'Materials and Methods') since this coefficient can disentangle the correlation of the CV^2 and the average expression levels across cells (Supplementary Figure S9b). Therefore, we focus on rCV^2 instead of CV^2. By performing the linear regression of rCV^2 (see 'Materials and Methods'), we found the synergy between positive feedback and the TATA box (or initiator or CCAAT box) can amplify the expression variability (Figure 4B). This result was actually an extension of the previous result that the TATA box enlarged the gene-expression variability when feedback regulations were not distinguished (Supplementary Figure S10c) (13,78). As an additional evaluation, we used the rCV^2 rank to predict the presence of the TATA-box and showed that the area under the ROC (receiver operating characteristic) curve, denoted by AUC, was larger in the case of positive feedback than in the case of negative feedback or non-feedback (Figure 4E), indicating that TATA boxes led to the larger gene-expression variability in the former case.

Next, we assessed burst frequencies and sizes. Similar to the case of expression variability, we also performed multivariate linear regression analyses on them. When feedbacks were not distinguished, we showed that TATA boxes significantly boosted burst frequencies of the genes (Supplementary Figure S10d). However, when considering different feedback forms, we observed that only TATA genes with positive feedback increased burst frequencies (Figure 4C). In addition, we observed that other promoter motifs had different degrees of effect on burst frequency, depending on feedback forms. These results were masked without distinguishing feedback forms (Supplementary Figure S10d). For burst sizes, it was reported that the genes with TATA box or initiator had larger burst sizes than those without TATA box or without initiator (16). We reproduced similar results (Supplementary Figure S10e), but observed that the TATA genes were expressed with larger burst sizes, independently of feedback regulation, and the genes with initiator had larger burst sizes only in the case of negative feedback (Figure 4D). GC-box and CCAAT-box on the distal promoter had opposite effects on burst sizes in the cases of positive and negative feedback (Figure 4D). In particular, no difference was found for all the genes if feedback forms were not distinguished (Supplementary Figure S10e).

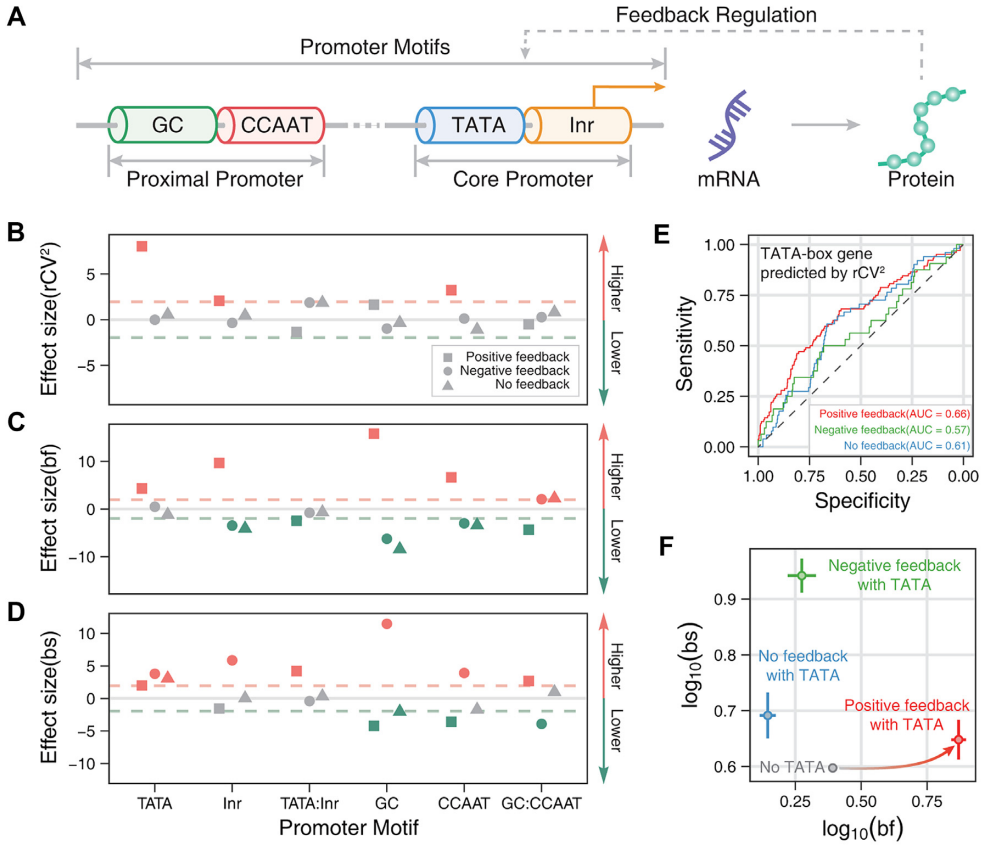

Figure 4. Genome-wide effects of promoter motifs on transcriptional burst kinetics in three cases of feedback regulation. (**A**) Schematic for a gene model that considers feedback regulation and promoter motifs (such as initiator, TATA-box, CCAAT-box, and GC-box). (**B–D**) Dependences of variability (rCV^2, B), burst frequencies (bf, C) and burst sizes (bs, D) on promoter motifs for different feedback regulations, obtained through linear regressions. Each symbol shows the *t*-value in a multivariate linear regression model, which is used to judge whether to reject the null hypothesis (i.e. the feature does not correlate with the dependent variable). Color: significantly higher (red symbol), significantly lower (green symbol), and no apparent effect (gray symbol). Different symbols stand for different feedbacks: square for positive feedback, circle for negative feedback, and triangele for non-feedback. (**E**) ROC curves are used to distinguish the genes with TATA boxes according to the relative rCV^2 rank. AUC is the area under the ROC curves. (**F**) Scatter plots of mean burst frequencies and mean burst sizes among the genes without TATA (gray), with positive feedback and TATA (red), with negative feedback and TATA (green), and without feedback but with TATA (blue). The solid lines near the scatter are error bars.

Briefly, the above results indicated that the TATA box played a pivotal role in transcriptional bursting. It worked as a static promoter element to up-regulate burst sizes and simultaneously utilized a dynamic positive feedback regulation mechanism to increase burst frequencies (Figure 4F).

Feedback regulations concealed the effects of TSS distribution on transcriptional burst kinetics

TSS can be divided into two classes according to its distribution: single TSS (sharp promoter) and multiple TSSs (broad promoter), both being important for gene expression (79,80). It was reported that the shapes of TSS distribution correlated with the category of genes, such as housekeeping genes and cell-type-specific genes, both exhibiting different transcriptional burst patterns (81). On the other hand, some experimental results indicated that feedback can regulate transcriptional initiation (23,82,83). A question naturally arose: how do the shapes of TSS distribution affect transcription burst kinetics in the presence of feedback regulation?

To address this question, we used the R package CAGEr (55) to read CAGE data of FANTOM5 MEF cell (see 'Materials and Methods') and classified the promoters into 'broad' and 'sharp' ones (79) according to the median (15bp) of the widths of all sampled promoters as depicted in Figure 5A. Similarly, the influence of the TSS distribution on variability and burst kinetics was subject to Simpson's paradox in the case of with and without distinguishing feedback regulations.

We showed that the impacts of different TSS distributions on the mean expression level did not exhibit apparent differences in three cases of feedback regulation and all genes (Supplementary Figure S11a,b). This property can avoid possible errors in evaluating the expression variability (rCV^2). Consistent with the observations in previous experimental studies (13), sharp promoters resulted in a significantly higher expression variability than broad promoters, independent of feedback forms (Supplementary Figure S11c, d). The rCV^2 declined with increasing the width (< 15bp) of 'sharp' promoters but was almost unchanged with increasing the width of 'broad' promoters (Figure 5B). No-

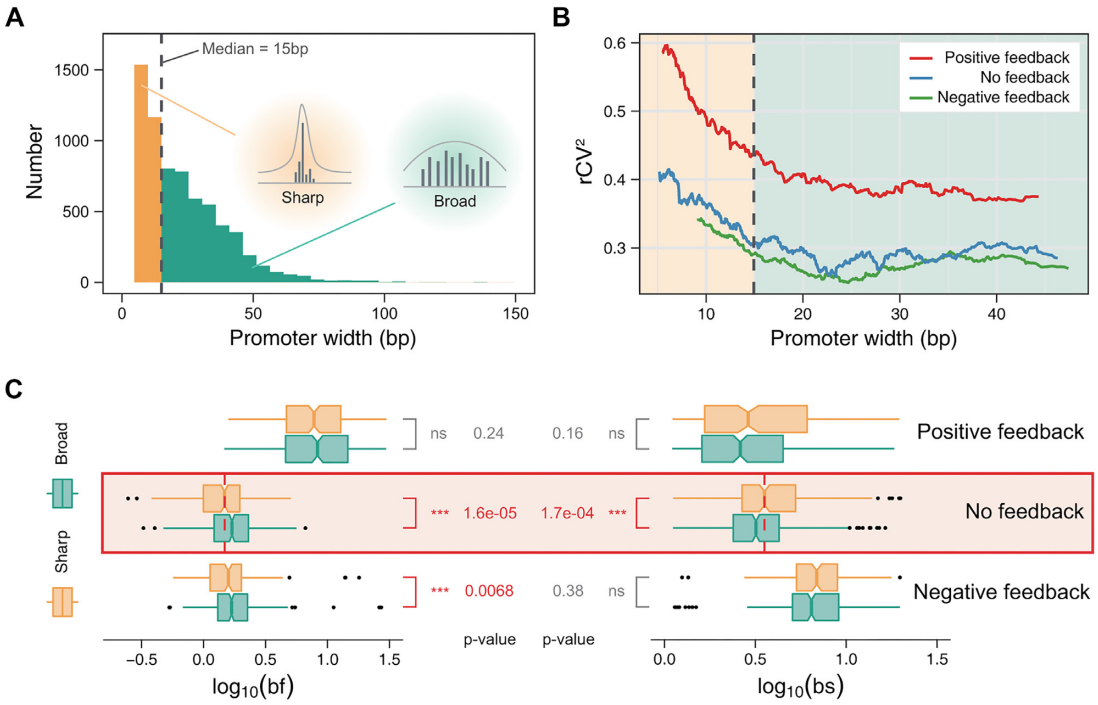

Figure 5. Genome-wide effects of TSS distributions on transcriptional burst kinetics in three cases of feedback regulation. (**A**) Histogram of genes, which are divided into two groups (sharp and broad) based on the median (dashed line) of promoter widths. (**B**) Changing trends of variability (rCV^2) as a function of promoter width in three cases of feedback regulation: positive (red), negative (green) and non-feedbacks (blue). The left-hand side of the dashed line stands for sharp promoters (yellow region) and the right-hand side for broad promoters (green region). (**C**) Boxplots of burst frequencies (left) and burst sizes (right), where yellow squares stands for sharp promoters and green squares for broad promoters. *P*-values are indicated, and ns is the abbreviation of no significance.

tably, the curve of rCV^2 vs. promoter width for the positive-feedback genes was always above that for the genes with negative feedback or non-feedback (Figure 5B).

Next, we investigated whether different TSS distributions affected burst frequencies and burst sizes differently. Although genes with 'sharp' promoters led to a higher expression variability than those with 'broad' promoters for arbitrary feedback forms, burst frequencies and sizes regulated by TSS distributions can exhibit significant discrepancy only in the absence of feedback (Figure 5C, Supplementary Figure S11e, f). Broad promoters led to higher burst frequencies and smaller burst sizes than sharp promoters (Figure 5C, red box), in agreement with the experimental observation that broad promoters tended to occur in the case of low RNA polymerase II pause, whereas sharp promoters tended to occur in the case of high RNA polymerase II pause (84–86). These results implied that on the genome-wide scale, feedback regulations significantly weakened the impacts of TSS distributions on transcriptional burst kinetics.

E–P interactions mainly modulate burst frequencies only in the presence of positive feedbacks

Enhancers, DNA sequences located upstream of the promoter, are important regulators of eukaryotic development (87). Several lines of experimental evidence supported that E–P interactions (Figure 6A) may facilitate gene transcription (88–91) and can regulate transcriptional burst kinetics (14,16,92–95). In addition, some studies showed that enhancer and promoter activations might require positive and negative feedback regulations, each contributing the elements of the protein complement required for activation of other genes (96). These results raise important questions: does the genome-wide control of burst kinetics by E–P interactions involves feedback regulations? If so, how do feedbacks affect burst kinetics?

To address these questions, we first recovered the intensities of E–P interactions from (16) and performed LOESS regression. With the involvement of feedback regulations, the modulation of variability and transcriptional burst kinetics by E–P interactions also presents Simpson's paradox. We then showed that, for all genes, increasing the E–P intensities led to the rise of mean gene-expression levels (Supplementary Figure S12d) but to the decline of variability (rCV^2) (Supplementary Figure S12e), indicating that stronger enhancers raised the expressions levels but lowered cell-to-cell variability in contrast to weaker enhancers (97). However, when distinguishing genes by feedback types, this pattern only appears in the case of positive feedback (Figure 6B, Supplementary Figure S12a), implying the important role of positive feedback in E–P interactions.

Next, we focused on burst frequencies and sizes. Previous molecular experiments and genome-wide inferences from scRNA-seq data showed that burst frequencies and sizes increased with promoting E–P interactions (93,94), and that enhancers mainly controlled burst frequencies (14,16,92–95,98–100). The same conclusion was obtained when we

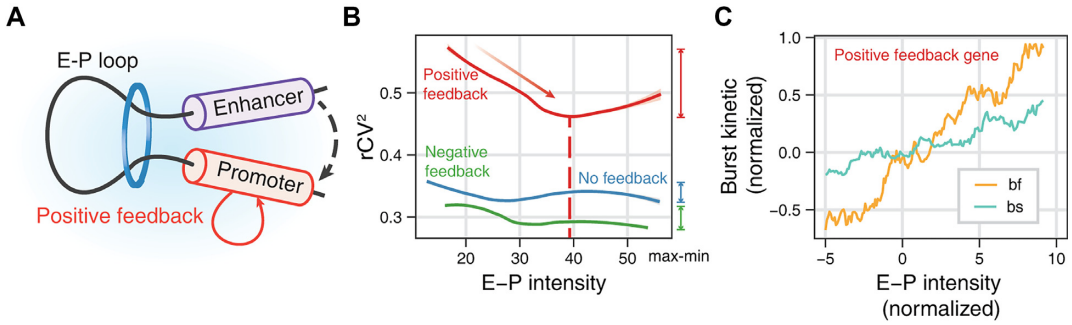

Figure 6. Genome-wide effects of enhancer–promoter interactions on transcriptional burst kinetics in three cases of feedback regulation. (**A**) Illustration of the E–P interaction with a positive feedback loop. (**B**) Dependence of noise (rCV^2) on E–P interaction intensity for different feedback forms, where the dashed line represents the valley in case of positive feedback and the line segment on the right-hand side of the picture represents the maximum minus the minimum, that is, the amplitude of affecting the variability. Color: positive (red line), negative (green line), and non-feedbacks (blue line). (**C**) Dependences of normalized burst frequencies and sizes on E–P interaction intensity in the case of positive feedback.

performed analysis without distinguishing feedback types (Supplementary Figure S12f, g). Notably, when the feedback types was considered, we found that as the E–P intensity increased, changes in burst frequencies and sizes were most apparent in the case of positive feedback (Supplementary Figure S12b,c). Moreover, the slope of the line for the dependence of burst frequencies on E–P intensity was larger than that for the dependence of burst sizes on E–P intensity (Figure 6C). In addition, we observed that this regulation effect of enhancers was saturated when the E–P interaction intensity exceeded a threshold (∼40) (Figure 6B and Supplementary Figure S12a-c). This result indicated that the function of the enhancer was not unlimited, in agreement with the theoretical prediction in our previous study (https://doi.org/10.1101/2022.01.24.477520).

The above genome-wide results provided direct support for the fact that the control of burst kinetics by E–P interactions was constrained by positive feedback regulations, in accordance with previous experimental results for a small number of genes (30,101).

DISCUSSION

As the core process of life, gene transcription occurs stochastically, leading to variability in the mRNA and further protein abundances. This variability is believed to be mainly attributed to transcriptional bursting, a phenomenon that occurs commonly in both prokaryotes and eukaryotes. From the viewpoint of biophysics, the sources of transcriptional bursting are multilevel and multiscale (1). In this study, we have developed a statistical framework of the model-driven and data-driven integration to infer dynamic feedback regulations and transcriptional bursting kinetics from static scRNA-seq data, using a mechanistic mathematical model as the connecting thread.

The mechanistic model used in our inference framework was interpretable. It captured the scRNA-seq measurement process and the molecular mechanisms of transcriptional bursting processes. We showed that not only burst frequencies and sizes as well as expression variability but also feedback forms can be effectively and robustly inferred to explain biophysical phenomena, which were masked in the scRNA-seq data. Meanwhile, our inference method made the interpretable model tractable. We utilized the Gauss-Laguerre Quadrature Rules instead of the classical MCMC method to compute mRNA distribution with a high-order integral that is difficult to solve, thus making our scalable inference applicable on genome-wide scales. Our statistical inference framework laid a solid foundation for exploring the molecular mechanisms of stochastic gene expression based on single-cell data.

Our inference method provided a powerful tool for analyzing the joint effects of feedback regulation and promoter architecture and for revealing the genome-wide mechanisms of transcriptional burst kinetics. First, we found that at the same gene-expression levels, positive-feedback genes exhibited significantly higher gene-expression variability and higher burst frequencies as well as smaller burst sizes than negative-feedback genes on genome-wide scales. This finding indicated that different regulatory networks played distinct roles in modulating transcriptional burst kinetics (10). Second, we revealed that the TATA box, apart from being indicatives of enlarging the expression variability and raising burst sizes as suggested in previous studies (13,16), can utilize a positive feedback mechanism to increase burst frequencies. This result may explain the phenomenon that the RNA polymerase II on the TATA box gene had better localization and fewer transcriptional pauses (75,76). Third, broad promoters with multiple TSSs led to higher burst frequencies and smaller burst sizes, which were concealed by the feedback regulations. Finally, we showed that enhancer–promoter interactions modulated burst kinetics and primarily controlled burst frequency in the presence of positive feedback. All these results were obtained under the hidden hypothesis that the intrinsic behaviors of the different gene were statistically identical. Overall, these genome-wide evidences indicated that transcriptional burst kinetics was not only encoded by static promoter architectures but also constrained by dynamic gene regulatory networks.

Our inference framework based on the model-driven and data-driven combination was an extensible one for studying the general principles of transcriptional bursting. First, gene expression variability caused by transcriptional bursts

comes not only from technical noise and feedback regulation as described in our hierarchical model, but also from many other potentially complex mechanisms, such as RNA polymerase II recruitment and pause release (102–105), alternative splicing (106,107), post-transcriptional regulations via mRNA degradation (108) and nuclear retention (109), chromatin movement (110), etc. (111–114), which all may affect burst kinetics. Second, promoter architecture can be described by a multi-state model since a transcription process would involve many molecular steps (115,116). It is unclear whether the multi-state architecture is more descriptive than the two-state model. Determining the number of gene states and studying the effect on burst kinetics is a long-term effort. Third, our hierarchical model only considered self-regulatory feedback (117), the simplest feedback form. More complex regulatory forms may exist in gene-expression systems (118). However, since they reflect high-level structure regulation (10), more complex yet reasonable mathematical models and more powerful inference methods need to be developed for better studying transcriptional burst kinetics. Fourth, most of the traditional models of gene expression were based on the Markov hypothesis (69,119). In organisms, however, the processes of molecular synthesis may be non-Markovian, and increasing time-resolved data have verified the extensive existence of molecular memory (120,121). Therefore, it is necessary to extend Markov models to non-Markov ones (122–124). But this is a great challenge to numerical solutions and statistical inferences. Finally, we point out that choosing a suitable model involves trade-off problems since more complex models would bring less consensus on general principles of transcriptional bursting (4).

Finally, studying transcriptional burst kinetics may start with a data-driven approach as done in our statistical inference framework. Our predictions of burst kinetics using scRNA-seq data were based on the assumption that the abundances of mRNA and protein were highly dependent (65). Recently, more and more studies of sequencing methods have paid attention to measuring the profiles of multitype molecules in single-cell levels, such as simultaneous quantification of intracellular mRNA and protein (125), which can better describe cell states (126). For feedback loop types our method predicted, we found that many genes have been confirmed by biological experiments (Supplementary Table S1). Moreover, the identification of feedback loops can be more convincing by using multimodal data combined with scRNA-seq such as ENCODE (127) and some automated packages (128). In addition, time-resolved data can provide more information compared to static data. We believe that with the continuous progress in measurement technologies, time-resolved single-cell data will be primary means to study the transcription burst kinetics in the future (https://doi.org/10.1101/2022.06.19.496754). Meanwhile, spatial transcriptome multimodal data (129–132) and chromatin structural data (133) provided good opportunities for in-depth studies of burst kinetics. Analysis of those multimodal single-cell data or integrated data can help us discover more credible biological knowledge but would also bring challenges for developing statistical methods to infer dynamic molecular mechanisms masked in static single-cell data.

DATA AVAILABILITY

All the analysis results and inference code that support the findings of this study are provided through https://github.com/cellfate/BurstFeedback or https://zenodo.org/record/7371318 (DOI: 10.5281/zenodo.7371318).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: J.Z. conceived of the study. S.L., Z.W. and J.Z. implemented the method, performed the analysis, and interpreted the results. Z.Z. helped with data analysis. J.Z. and T. Z. supervised the study. S.L., J.Z. and T.Z. drafted the manuscript with input from all the authors. All authors read and approved the final manuscript.

FUNDING

National Key R&D Program of China [2021YFA1302500]; Natural Science Foundation of P. R. China [12171494, 11931019, 11775314]; Guangdong Basic and Applied Basic Research Foundation [2022A1515011540]; Key-Area Research and Development Program of Guangzhou, P. R. China [2019B110233002, 202007030004]; Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University [2020B1212060032]. Funding for open access charge: National Key R&D Program of China [2021YFA1302500]; Natural Science Foundation of P. R. China [12171494, 11931019, 11775314]; Guangdong Basic and Applied Basic Research Foundation [2022A1515011540]; Key-Area Research and Development Program of Guangzhou, P. R. China [2019B110233002, 202007030004]; Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University [2020B1212060032].
Conflict of interest statement. None declared.

REFERENCES

1. Eling,N., Morgan,M.D. and Marioni,J.C. (2019) Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.*, **20**, 536–548.
2. Raj,A. and Van Oudenaarden,A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**, 216–226.
3. Rodriguez,J. and Larson,D.R. (2020) Transcription in living cells: molecular mechanisms of bursting. *Annu. Rev. Biochem.*, **89**, 189–212.
4. Tunnacliffe,E. and Chubb,J.R. (2020) What is a transcriptional burst?*Trends. Genet.*, **36**, 288–297.
5. Dar,R.D., Razooky,B.S., Singh,A., Trimeloni,T.V., McCollum,J.M., Cox,C.D., Simpson,M.L. and Weinberger,L.S. (2012) Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 17454–17459.
6. Phillips,R., Kondev,J. and Theriot,J. (2009) In: *Physical Biology of the Cell*. 2nd edn. Garland Science, NY.
7. Zenklusen,D., Larson,D.R. and Singer,R.H. (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.*, **15**, 1263–1271.
8. Jones,D.L., Brewster,R.C. and Phillips,R. (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, **346**, 1533–1536.

9. Pedraza,J.M. and Van Oudenaarden,A. (2005) Noise propagation in gene networks. *Science*, **307**, 1965–1969.
10. Chalancon,G., Ravarani,C.N., Balaji,S., Martinez-Arias,A., Aravind,L., Jothi,R. and Babu,M.M. (2012) Interplay between gene expression noise and regulatory network architecture. *Trends. Genet.*, **28**, 221–232.
11. Silander,O.K., Nikolic,N., Zaslaver,A., Bren,A., Kikoin,I., Alon,U. and Ackermann,M. (2012) A genome-wide analysis of promoter-mediated phenotypic noise in Escherichia coli. *PLoS Genet.*, **8**, e1002443.
12. Sanchez,A. and Golding,I. (2013) Genetic determinants and cellular constraints in noisy gene expression. *Science*, **342**, 1188–1193.
13. Faure,A.J., Schmiedel,J.M. and Lehner,B. (2017) Systematic analysis of the determinants of gene expression noise in embryonic stem cells. *Cell Syst.*, **5**, 471–484.
14. Ochiai,H., Hayashi,T., Umeda,M., Yoshimura,M., Harada,A., Shimizu,Y., Nakano,K., Saitoh,N., Liu,Z., Yamamoto,T. *et al.* (2020) Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells. *Sci. Adv.*, **6**, eaaz6699.
15. Morgan,M.D. and Marioni,J.C. (2018) CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. *Genome Biol.*, **19**, 81.
16. Larsson,A.J., Johnsson,P., Hagemann-Jensen,M., Hartmanis,L., Faridani,O.R., Reinius,B., Segerstolpe,Å., Rivera,C.M., Ren,B. and Sandberg,R. (2019) Genomic encoding of transcriptional burst kinetics. *Nature*, **565**, 251–254.
17. Friedrich,D., Friedel,L., Finzel,A., Herrmann,A., Preibisch,S. and Loewer,A. (2019) Stochastic transcription in the p53-mediated response to DNA damage is modulated by burst frequency. *Mol. Syst. Biol.*, **15**, e9068.
18. Skupsky,R., Burnett,J.C., Foley,J.E., Schaffer,D.V. and Arkin,A.P. (2010) HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS Comput. Biol.*, **6**, e1000952.
19. Hendy,O., Campbell Jr,J., Weissman,J.D., Larson,D.R. and Singer,D.S. (2017) Differential context-specific impact of individual core promoter elements on transcriptional dynamics. *Mol. Biol. Cell.*, **28**, 3360–3370.
20. Tunnacliffe,E., Corrigan,A.M. and Chubb,J.R. (2018) Promoter-mediated diversification of transcriptional bursting dynamics following gene duplication. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 8364–8369.
21. Sanchez,A., Garcia,H.G., Jones,D., Phillips,R. and Kondev,J. (2011) Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput. Biol.*, **7**, e1001100.
22. Davidson,E.H. (2010) Emerging properties of animal gene regulatory networks. *Nature*, **468**, 911–920.
23. Crews,S.T. and Pearson,J.C. (2009) Transcriptional autoregulation in development. *Curr. Biol.*, **19**, R241–R246.
24. Becskei,A. and Serrano,L. (2000) Engineering stability in gene networks by autoregulation. *Nature*, **405**, 590–593.
25. Rosenfeld,N., Elowitz,M.B. and Alon,U. (2002) Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.*, **323**, 785–793.
26. Austin,D., Allen,M., McCollum,J., Dar,R., Wilgus,J., Sayler,G., Samatova,N., Cox,C. and Simpson,M. (2006) Gene network shaping of inherent noise spectra. *Nature*, **439**, 608–611.
27. Alon,U. (2007) Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, **8**, 450–461.
28. To,T.-L. and Maheshri,N. (2010) Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science*, **327**, 1142–1145.
29. Venturelli,O.S., El-Samad,H. and Murray,R.M. (2012) Synergistic dual positive feedback loops established by molecular sequestration generate robust bimodal response. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E3324–E3333.
30. Becskei,A., Séraphin,B. and Serrano,L. (2001) Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.*, **20**, 2528–2535.
31. Pigolotti,S., Krishna,S. and Jensen,M.H. (2007) Oscillation patterns in negative feedback loops. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 6533–6537.
32. Bokes,P. and Singh,A. (2017) Gene expression noise is affected differentially by feedback in burst frequency and burst size. *J. Math. Biol.*, **74**, 1483–1509.
33. Bokes,P. (2022) Exact and WKB-approximate distributions in a gene expression model with feedback in burst frequency, burst size, and protein stability. *Discrete Cont. Dyn. B*, **27**, 2129–2145.
34. Bartman,C.R., Hamagami,N., Keller,C.A., Giardine,B., Hardison,R.C., Blobel,G.A. and Raj,A. (2019) Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Mol. Cell*, **73**, 519–532.
35. Chubb,J.R., Trcek,T., Shenoy,S.M. and Singer,R.H. (2006) Transcriptional pulsing of a developmental gene. *Curr. Biol.*, **16**, 1018–1025.
36. Golding,I., Paulsson,J., Zawilski,S.M. and Cox,E.C. (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.
37. Zoller,B., Little,S.C. and Gregor,T. (2018) Diverse spatial expression patterns emerge from unified kinetics of transcriptional bursting. *Cell*, **175**, 835–847.
38. Senecal,A., Munsky,B., Proux,F., Ly,N., Braye,F.E., Zimmer,C., Mueller,F. and Darzacq,X. (2014) Transcription factors modulate c-Fos transcriptional bursts. *Cell Rep.*, **8**, 75–83.
39. Donovan,B.T., Huynh,A., Ball,D.A., Patel,H.P., Poirier,M.G., Larson,D.R., Ferguson,M.L. and Lenstra,T.L. (2019) Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. *EMBO J.*, **38**, e100809.
40. Chong,S., Chen,C., Ge,H. and Xie,X.S. (2014) Mechanism of transcriptional bursting in bacteria. *Cell*, **158**, 314–326.
41. Tanay,A. and Regev,A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331–338.
42. Munsky,B., Li,G., Fox,Z.R., Shepherd,D.P. and Neuert,G. (2018) Distribution shapes govern the discovery of predictive models for gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 7533–7538.
43. Kim,J.K. and Marioni,J.C. (2013) Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.*, **14**, R7.
44. Jiang,Y., Zhang,N.R. and Li,M. (2017) SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol.*, **18**, 74.
45. Wu,S., Li,K., Li,Y., Zhao,T. and Qian,W. (2017) Independent regulation of gene expression level and noise by histone modifications. *PLoS Comput. Biol.*, **13**, e1005585.
46. Cao,Z. and Grima,R. (2018) Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.*, **9**, 3305.
47. Sarkar,A. and Stephens,M. (2021) Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.*, **53**, 770–777.
48. Wang,J., Huang,M., Torre,E., Dueck,H., Shaffer,S., Murray,J., Raj,A., Li,M. and Zhang,N.R. (2018) Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E6437–E6446.
49. Friedman,N., Cai,L. and Xie,X.S. (2006) Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.*, **97**, 168302.
50. Karandikar,R.L. (2006) On the markov chain monte carlo (MCMC) method. *Sadhana*, **31**, 81–104.
51. Abromowitz,M. and Stegun,I.A. (1972) In: *Handbook of Mathematical Functions*. Dover, NY.
52. Cavanaugh,J.E. and Neath,A.A. (2019) The Akaike information criterion: background, derivation, properties, application, interpretation, and refinements. *Wires Comput. Stat.*, **11**, e1460.
53. Deng,Q., Ramsköld,D., Reinius,B. and Sandberg,R. (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.
54. Dreos,R., Ambrosini,G., Groux,R., Cavin Périer,R. and Bucher,P. (2017) The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.*, **45**, D51–D55.
55. Haberle,V., Forrest,A.R., Hayashizaki,Y., Carninci,P. and Lenhard,B. (2015) CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.*, **43**, e51.
56. Shen,Y., Yue,F., McCleary,D.F., Ye,Z., Edsall,L., Kuan,S., Wagner,U., Dixon,J., Lee,L., Lobanenkov,V.V. *et al.* (2012) A map

of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.
57. Hornung,G., Bar-Ziv,R., Rosin,D., Tokuriki,N., Tawfik,D.S., Oren,M. and Barkai,N. (2012) Noise–mean relationship in mutated promoters. *Genome Res.*, **22**, 2409–2417.
58. Newman,J.R., Ghaemmaghami,S., Ihmels,J., Breslow,D.K., Noble,M., DeRisi,J.L. and Weissman,J.S. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, **441**, 840–846.
59. Chubb,J.R. and Liverpool,T.B. (2010) Bursts and pulses: insights from single cell studies into transcriptional mechanisms. *Curr. Opin. Genet. Dev.*, **20**, 478–484.
60. Yu,J., Xiao,J., Ren,X., Lao,K. and Xie,X.S. (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.
61. Nicolas,D., Zoller,B., Suter,D.M. and Naef,F. (2018) Modulation of transcriptional burst frequency by histone acetylation. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 7153–7158.
62. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L. and Girke,T. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.
63. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
64. Tutucci,E., Vera,M., Biswas,J., Garcia,J., Parker,R. and Singer,R.H. (2018) An improved MS2 system for accurate reporting of the mRNA life cycle. *Nat. Methods*, **15**, 81–89.
65. Liu,Y., Beyer,A. and Aebersold,R. (2016) On the dependency of cellular protein levels on mRNA abundance. *Cell*, **165**, 535–550.
66. Stegle,O., Teichmann,S.A. and Marioni,J.C. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, **16**, 133–145.
67. Qiu,P. (2020) Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.*, **11**, 1169.
68. Suter,D.M., Molina,N., Gatfield,D., Schneider,K., Schibler,U. and Naef,F. (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science*, **332**, 472–474.
69. Peccoud,J. and Ycart,B. (1995) Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.*, **48**, 222–234.
70. Müller-McNicoll,M., Rossbach,O., Hui,J. and Medenbach,J. (2019) Auto-regulatory feedback by RNA-binding proteins. *J. Mol. Cell Biol.*, **11**, 930–939.
71. Carey,L.B., Van Dijk,D., Sloot,P.M., Kaandorp,J.A. and Segal,E. (2013) Promoter sequence determines the relationship between expression level and noise. *PLoS Biol.*, **11**, e1001528.
72. Dublanche,Y., Michalodimitrakis,K., Kümmerer,N., Foglierini,M. and Serrano,L. (2006) Noise in transcription negative feedback loops: simulation and experimental analysis. *Mol. Syst. Biol.*, **2**, 41.
73. Pimmett,V.L., Dejean,M., Fernandez,C., Trullo,A., Bertrand,E., Radulescu,O. and Lagha,M. (2021) Quantitative imaging of transcription in living Drosophila embryos reveals the impact of core promoter motifs on promoter state dynamics. *Nat. Commun.*, **12**, 4504.
74. Deng,W. and Roberts,S.G. (2005) A core promoter element downstream of the TATA box that is recognized by TFIIB. *Gene Dev.*, **19**, 2418–2423.
75. Ramalingam,V., Natarajan,M., Johnston,J. and Zeitlinger,J. (2021) TATA and paused promoters active in differentiated tissues have distinct expression characteristics. *Mol. Syst. Biol.*, **17**, e9866.
76. Lee,T.I. and Young,R.A. (2000) Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.*, **34**, 77–137.
77. Tantale,K., Mueller,F., Kozulic-Pirher,A., Lesne,A., Victor,J.-M., Robert,M.-C., Capozi,S., Chouaib,R., Bäcker,V. and Mateos-Langerak,J. (2016) A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nat. Commun.*, **7**, 12248.
78. Miller-Jensen,K., Skupsky,R., Shah,P.S., Arkin,A.P. and Schaffer,D.V. (2013) Genetic selection for context-dependent stochastic phenotypes: sp1 and TATA mutations increase phenotypic noise in HIV-1 gene expression. *PLoS Comput. Biol.*, **9**, e1003135.
79. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engström,P.G. and Frith,M.C. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
80. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T. and Morishita,S. (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.
81. Haberle,V. and Stark,A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.*, **19**, 621–637.
82. Ngondo,R.P. and Carbon,P. (2014) Transcription factor abundance controlled by an auto-regulatory mechanism involving a transcription start site switch. *Nucleic Acids Res.*, **42**, 2171–2184.
83. Kiełbasa,S.M. and Vingron,M. (2008) Transcriptional autoregulatory loops are highly conserved in vertebrate evolution. *PLoS One*, **3**, e3210.
84. Meers,M.P., Adelman,K., Duronio,R.J., Strahl,B.D., McKay,D.J. and Matera,A.G. (2018) Transcription start site profiling uncovers divergent transcription and enhancer-associated RNAs in Drosophila melanogaster. *BMC Genomics*, **19**, 157.
85. Nechaev,S., Fargo,D.C., Santos,G., Liu,L., Gao,Y. and Adelman,K. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, **327**, 335–338.
86. Rach,E.A., Winter,D.R., Benjamin,A.M., Corcoran,D.L., Ni,T., Zhu,J. and Ohler,U. (2011) Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet.*, **7**, e1001274.
87. Peng,Y. and Zhang,Y. (2018) Enhancer and super-enhancer: positive regulators in gene transcription. *Anim. Model Exp. Med.*, **1**, 169–179.
88. Zuin,J., Roth,G., Zhan,Y., Cramard,J., Redolfi,J., Piskadlo,E., Mach,P., Kryzhanovska,M., Tihanyi,G. and Kohler,H. (2022) Nonlinear control of transcription through enhancer–promoter interactions. *Nature*, **604**, 571–577.
89. Xiao,J.Y., Hafner,A. and Boettiger,A.N. (2021) How subtle changes in 3D structure can create large changes in transcription. *Elife*, **10**, e64320.
90. Li,J., Hsu,A., Hua,Y., Wang,G., Cheng,L., Ochiai,H., Yamamoto,T. and Pertsinidis,A. (2020) Single-gene imaging links genome topology, promoter–enhancer communication and transcription control. *Nat. Struct. Mol. Biol.*, **27**, 1032–1040.
91. Chen,H., Levo,M., Barinov,L., Fujioka,M., Jaynes,J.B. and Gregor,T. (2018) Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.*, **50**, 1296–1303.
92. Walters,M.C., Fiering,S., Eidemiller,J., Magis,W., Groudine,M. and Martin,D. (1995) Enhancers increase the probability but not the level of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 7125–7129.
93. Yokoshi,M., Segawa,K. and Fukaya,T. (2020) Visualizing the role of boundary elements in enhancer–promoter communication. *Mol. Cell*, **78**, 224–235.
94. Fukaya,T., Lim,B. and Levine,M. (2016) Enhancer control of transcriptional bursting. *Cell*, **166**, 358–368.
95. Bartman,C.R., Hsu,S.C., Hsiung,C.C.-S., Raj,A. and Blobel,G.A. (2016) Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol. Cell*, **62**, 237–247.
96. Kim,T.-K., Hemberg,M. and Gray,J.M. (2015) Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *CSH Perspect. Biol.*, **7**, a018622.
97. Urban,E.A. and Jr,Johnston, R.J. (2018) Buffering and amplifying transcriptional noise during cell fate specification. *Front. Genet.*, **9**, 591.
98. Li,C., Cesbron,F., Oehler,M., Brunner,M. and Höfer,T. (2018) Frequency modulation of transcriptional bursting enables sensitive and rapid gene regulation. *Cell Syst.*, **6**, 409–423.
99. Rodriguez,J., Ren,G., Day,C.R., Zhao,K., Chow,C.C. and Larson,D.R. (2019) Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*, **176**, 213–226.
100. Larson,D.R., Fritzsch,C., Sun,L., Meng,X., Lawrence,D.S. and Singer,R.H. (2013) Direct observation of frequency modulated transcription in single cells using light activation. *Elife*, **2**, e00750.
101. Brown,J.C. (2020) Involvement of promoter/enhancers in a feedback loop to regulate human gene expression. *Heliyon*, **6**, e04934.

102. Sun,X.-M., Bowman,A., Priestman,M., Bertaux,F., Martinez-Segura,A., Tang,W., Whilding,C., Dormann,D., Shahrezaei,V. and Marguerat,S. (2020) Size-dependent increase in RNA Polymerase II initiation rates mediates gene expression scaling with cell size. *Curr. Biol.*, **30**, 1217–1230.
103. Fujita,K., Iwaki,M. and Yanagida,T. (2016) Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA. *Nat. Commun.*, **7**, 13788.
104. Engl,C., Jovanovic,G., Brackston,R.D., Kotta-Loizou,I. and Buck,M. (2020) The route to transcription initiation determines the mode of transcriptional bursting in E. coli. *Nat. Commun.*, **11**, 2422.
105. Tantale,K., Garcia-Oliver,E., Robert,M.-C., L'Hostis,A., Yang,Y., Tsanov,N., Topno,R., Gostan,T., Kozulic-Pirher,A. and Basu-Shrivastava,M. (2021) Stochastic pausing at latent HIV-1 promoters generates transcriptional bursting. *Nat. Commun.*, **12**, 4503.
106. Wan,Y., Anastasakis,D.G., Rodriguez,J., Palangat,M., Gudla,P., Zaki,G., Tandon,M., Pegoraro,G., Chow,C.C. and Hafner,M. (2021) Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell*, **184**, 2878–2895.
107. Gorin,G. and Pachter,L. (2022) Modeling bursty transcription and splicing with the chemical master equation. *Biophys. J.*, **121**, 1056–1069.
108. Nordick,B., Yu,P.Y., Liao,G. and Hong,T. (2022) Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression. *Nucleic Acids Res.*, **50**, 3693–3708.
109. Wang,Q. and Zhou,T. (2015) Dynamical analysis of mCAT2 gene models with CTN-RNA nuclear retention. *Phys. Biol.*, **12**, 016010.
110. Liu,T., Zhang,J. and Zhou,T. (2016) Effect of interaction between chromatin loops on cell-to-cell variability in gene expression. *PLoS Comput. Biol.*, **12**, e1004917.
111. Cao,Z. and Grima,R. (2020) Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 4682–4692.
112. Skinner,S.O., Xu,H., Nagarkar-Jaiswal,S., Freire,P.R., Zwaka,T.P. and Golding,I. (2016) Single-cell analysis of transcription kinetics across the cell cycle. *Elife*, **5**, e12175.
113. Peterson,J.R., Cole,J.A., Fei,J., Ha,T. and Luthey-Schulten,Z.A. (2015) Effects of DNA replication on mRNA noise. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 15886–15891.
114. Beentjes,C.H., Perez-Carrasco,R. and Grima,R. (2020) Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. *Phys. Rev. E*, **101**, 032403.
115. Klindziuk,A. and Kolomeisky,A.B. (2018) Theoretical investigation of transcriptional bursting: a multistate approach. *J. Phys. Chem. B*, **122**, 11969–11977.
116. Neuert,G., Munsky,B., Tan,R.Z., Teytelman,L., Khammash,M. and Van Oudenaarden,A. (2013) Systematic identification of signal-activated stochastic gene regulation. *Science*, **339**, 584–587.
117. Holehouse,J., Cao,Z. and Grima,R. (2020) Stochastic modeling of autoregulatory genetic feedback loops: a review and comparative study. *Biophys. J.*, **118**, 1517–1525.
118. Öcal,K., Gutmann,M.U., Sanguinetti,G. and Grima,R. (2022) Inference and uncertainty quantification of stochastic gene expression via synthetic models. *J. R. Soc. Interface*, **19**, 20220153.
119. Shahrezaei,V. and Swain,P.S. (2008) Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 17256–17261.
120. Jia,T. and Kulkarni,R.V. (2011) Intrinsic noise in stochastic models of gene expression with molecular memory and bursting. *Phys. Rev. Lett.*, **106**, 058102.
121. Pedraza,J.M. and Paulsson,J. (2008) Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, **319**, 339–343.
122. Zhang,J. and Zhou,T. (2019) Markovian approaches to modeling intracellular reaction processes with molecular memory. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 23542–23550.
123. Zoller,B., Nicolas,D., Molina,N. and Naef,F. (2015) Structure of silent transcription intervals and noise characteristics of mammalian genes. *Mol. Syst. Biol.*, **11**, 823.
124. Fritzsch,C., Baumgärtner,S., Kuban,M., Steinshorn,D., Reid,G. and Legewie,S. (2018) Estrogen-dependent control and cell-to-cell variability of transcriptional bursting. *Mol. Syst. Biol.*, **14**, e7678.
125. Darmanis,S., Gallant,C.J., Marinescu,V.D., Niklasson,M., Segerman,A., Flamourakis,G., Fredriksson,S., Assarsson,E., Lundberg,M. and Nelander,S. (2016) Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep.*, **14**, 380–389.
126. Stuart,T. and Satija,R. (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*, **20**, 257–272.
127. ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
128. Nordick,B. and Hong,T. (2021) Identification, visualization, statistical analysis and mathematical modeling of high-feedback loops in gene regulatory networks. *BMC Bioinf.*, **22**, 481.
129. Burgess,D.J. (2019) Spatial transcriptomics coming of age. *Nat. Rev. Genet.*, **20**, 317.
130. Maynard,K., Jaffe,A. and Martinowich,K. (2020) Spatial transcriptomics: putting genome-wide expression on the map. *Neuropsychopharmacol.*, **45**, 232–233.
131. Chen,K., Boettiger,A., Moffitt,J., Wang,S. and Zhuang,X. (2015) RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science***348**, aaa6090.
132. Eng,C.-H.L., Lawson,M., Zhu,Q., Dries,R., Koulena,N., Takei,Y., Yun,J., Cronin,C., Karp,C. and Yuan,G.-C. (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, **568**, 235–239.
133. Kempfer,R. and Pombo,A. (2020) Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.*, **21**, 207–226.