Genome Biology

METHOD

Open Access

Synergising single-cell resolution and 4sU labelling boosts inference of transcriptional bursting

David M. Edwards[1*], Philip Davies[1] and Daniel Hebenstreit[1*]

*Correspondence: david.m.edwards@warwick.ac.uk; d.hebenstreit@warwick.ac.uk

[1] School of Life Sciences, University of Warwick, Coventry, UK

Abstract

Despite the recent rise of RNA-seq datasets combining single-cell (sc) resolution with 4-thiouridine (4sU) labelling, analytical methods exploiting their power to dissect transcriptional bursting are lacking. Here, we present a mathematical model and Bayesian inference implementation to facilitate genome-wide joint parameter estimation and confidence quantification (R package: burstMCMC). We demonstrate that, unlike conventional scRNA-seq, 4sU scRNA-seq resolves temporal parameters and furthermore boosts inference of dimensionless parameters via a synergy between single-cell resolution and 4sU labelling. We apply our method to published 4sU scRNA-seq data and linked with ChIP-seq data, we uncover previously obscured associations between different parameters and histone modifications.

Keywords: Transcription, Bursting, Dynamics, Inference, Time-resolved, 4sU, Single-cell, Genome-wide, Histone modification

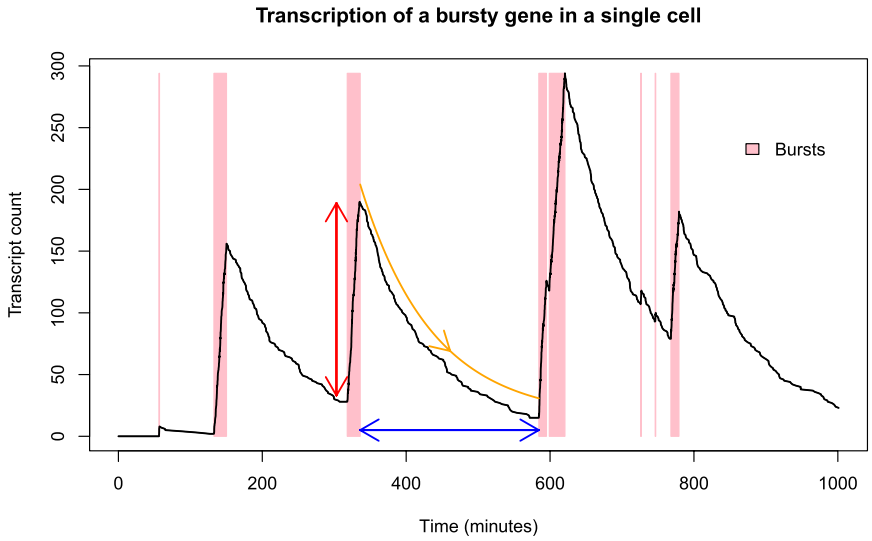
Background

The canonical understanding of transcription is that it consists of the steps of initiation, elongation and termination. During initiation of transcription in eukaryotes, RNA polymerase (RNAP) is recruited to the promoter via transcription factors (TF), followed by the synthesis of the first few bases of the new transcript [1]. Elongation succeeds initiation, in which RNAP processes along the gene, incorporating RNA nucleotides into the nascent transcript as it progresses [2]. Upon reaching the transcription end site (TES), termination occurs, in which the transcript and RNAP are released from the DNA [3]. Various processing steps take place at different points during transcription to allow for a mature transcript to be produced, including 5′ capping during initiation, splicing to remove intronic (non-coding) sequences during elongation of protein-coding genes, and polyadenylation and cleavage during termination [1–4].

Beyond the general mechanism outlined above, transcription is also a stochastic process subject to intrinsic noise through its fundamental dependence on probabilistic

collisions between molecules [5, 6], which are often present in relatively low numbers. Additionally, in many cases, transcription occurs only in short, intense bursts of activity followed by prolonged periods of inactivity, resulting in increased cell-cell variability in transcript counts [7, 8]. Indeed, studies have identified a broad spectrum of genes, from those that are transcribed in a Poissonian fashion, such as housekeeping genes, to those which are very bursty in nature and expressed only in relatively short windows [9, 10]. The transcriptional noise and cell-cell variance induced by bursting can be utilised to, for example, achieve alternative cell fates during differentiation of cell populations without requiring explicit control by genetic programming or external signals [11]. There are several different possible mechanisms thought to contribute to bursting, including the process of reinitiation, in which after transcribing a gene the RNAP is immediately recycled to the transcription start site (TSS) instead of simply terminating and disengaging [12]. This requires looping of the gene to bring the TSS and TES into physical proximity [13], and the link between TSS-TES interactions and bursting has been explored recently [14]. The chromatin state of a gene also plays an important role in governing transcriptional bursting dynamics, which in eukaryotes is dictated largely by histone modifications (HM). Different HMs may result in looser or tighter packing of the chromatin, respectively, with the chromatin density around the TSS being correlated with transcriptional noise [15]. Having active HMs at the TSS results in an increased probability of open chromatin, which facilitates initiation. This is proposed to reduce burstiness, possibly by reducing the duration between active periods [15, 16]. More recent studies have also reported genome-wide direct correlations between the presence of specific HMs at gene promoters and general transcriptional noise [17, 18], while further studies have even linked HMs with the underlying bursting dynamics, both at the individual gene level [19] and genome-wide [20]. Transcriptional bursting in bacteria can also result from supercoiling of the DNA [21]. The proposed mechanism is the accumulation of positive supercoiling caused by the RNAP proceeding through the gene, until it reduces the rate of elongation to the point that it prevents further transcription. Intermittent clearing of supercoiling followed by rapid transcription, and subsequent reaccumulation of supercoiling, results in bursty transcription. Studies have also observed the co-condensation of TFs with transcriptional coactivators such as p300, which mediates cooperative activation of genes by clusters of TFs [22]. This cooperative activation results in non-linear gene regulation and increased burst frequency and burst size for genes enriched in coactivators.

Transcriptional bursting may be understood in terms of several parameters (Fig. 1a), including the burst size (transcripts produced per burst, b), burst frequency (bursts per unit time, κ), decay rate (transcripts degraded per unit time, δ), transcript lifetime (average transcript survival time, $\gamma = 1/\delta$), burst rate (bursts per transcript lifetime, $a = \kappa/\delta$), and expression level (mean transcripts per cell, $\mu = b \times a$). Many studies make use of fluorescence microscopy-based approaches to interrogate transcriptional bursting dynamics. Single molecule fluorescence in situ hybridisation (smFISH) is a particularly popular approach here although the standard procedure offers only a snapshot of transcript counts across a cell population, with no time-variant information. Therefore, the timescales of bursting events may not be discerned [23], allowing estimation of μ, b and a but not κ or δ. Some smFISH-based experimental set-ups have progressed towards



(a)

Parameter	Bulk RNA-seq	scRNA-seq	Bulk 4sU RNA-seq	4sU scRNA-seq
Expression level				
Burst size				
Transcript lifetime				
Burst frequency				

(b)

Fig. 1 **a** Simulation demonstrating transcriptional bursting for a single gene in a single cell, indicating burst size (red), burst interval (blue, reciprocal of burst frequency), and decay rate (orange, reciprocal of transcript lifetime), while the thickness of the pink shaded regions indicate burst durations. **b** Table showing the parameters governing transcriptional dynamics that can theoretically be obtained using different RNA-seq approaches with no prior information. Dark blue and orange show if a data type does or does not inform a parameter, respectively

a level of understanding bursting timescales by using hybridisation specific to nascent transcripts [24, 25], although smFISH approaches generally suffer from scalability. While progress is being made towards multiplexing, it can still only analyse a handful of genes at a time compared with sequencing [26–28] or requires complex and labourious set-ups [29]. Sophisticated analysis methods [30] have been developed for time-lapse single-cell RNA imaging data [31] which allows dissection of transcriptional dynamics in great detail, however such approaches are even more limited scale-wise.

Single cell RNA-seq (scRNA-seq) experiments are widely used to analyse genome-wide bursting dynamics. However, scRNA-seq suffers from the same issue as standard smFISH regarding analysis of bursting timescales because it only provides a snapshot of the transcriptomes of a population of cells at a single point in time. Therefore, it has only been possible to obtain burst sizes (b) and burst rates (a), while burst frequencies (κ) may not be understood without making assumptions or using prior information on decay rates (δ) measured through separate experiments [10, 32–34]. On the other hand, bulk RNA-seq-based approaches have for several years made use of chemically labelled nucleotides, primarily 4-thiouridine (4sU) as in SLAM-seq, to understand RNA synthesis ($b \times \kappa$) and degradation (δ) rates [35, 36]. The cells are incubated in the presence of

4sU for a given duration, prior to RNA extraction. During this step, 4sU diffuses into the cell nucleus and becomes incorporated into nascently transcribed RNA. Labelled RNA can be bioinformatically distinguished from non-labelled RNA, previously residing in the cell, due to the higher rate of chemically induced cytosine conversion of 4sU relative to regular uracil. Using mathematical modelling, the ratio of labelled to unlabelled transcripts can be used to estimate the turnover rate [37]. However, since bulk RNA-seq neglects the cell-cell variability, it can not be used to study bursting dynamics. Recent advances combine scRNA-seq with 4sU and such datasets have the potential to fully characterise transcriptional bursting dynamics and their timescales (Fig. 1b). Thus far, they have been used for understanding dynamic changes in the transcriptome and/or RNA turnover/splicing rates that occur throughout the cell cycle and cell state transitions [38–42]. Studies with data of this type that have looked at bursting have only done so in a limited manner, using empirically derived statistics as a proxy for burstiness [43], while bursting timescales have remained uncharacterised in recent works [44]. This is despite previous modelling works having shown that degradation is expected to contribute significantly to transcriptional noise and therefore should be accounted for when investigating bursting dynamics [45].

Here, we construct mathematical models to relate observables from 4sU scRNA-seq data to the underlying bursting dynamics and develop an adaptive Markov chain Monte Carlo (MCMC) approach for Bayesian inference of the parameters governing those dynamics. We have produced an R package (https://github.com/hebenstreitLab/burstMCMC) from our method and applied this to published data from [38], demonstrating that we are able to characterise time-resolved transcriptional bursting dynamics for hundreds of genes in parallel. Our approach generates joint probability distributions of the parameters of interest from which estimates can be extracted and confidence in these quantified. This is the first method for joint inference of time-resolved bursting dynamics on a genome-wide scale and is generally applicable to 4sU scRNA-seq datasets. We also show that, even for the dimensionless parameters which can be obtained with conventional scRNA-seq, the accuracy and reliability of estimates can be improved by incorporating the additional information provided by 4sU scRNA-seq. Finally, we build on a previous study which interrogated correlations between bursting parameter estimates and HMs in a genome-wide manner, linking scRNA-seq with ChIP-seq data [20]. Our analysis reveals position-dependent associations between different parameters and HMs only apparent with 4sU scRNA-seq.

Results

Model comparison

We tested the advantages provided by 4sU scRNA-seq data coupled with our inference approach over conventional scRNA-seq by comparing our recovery of known bursting parameter values from a simulated dataset using different likelihood functions (Methods). The MCMC algorithm was run five times, using Eqs. 4, 15, 16, 19 and 20 as the likelihood functions, referred to as L1, L2, L1+L2, L3 and L1+L3, respectively.

- L1: The likelihood function of model 1, equivalent to scRNA-seq data without 4sU, relying solely on the UMI counts.

- L2: Equivalent to relying only on single cell T>C conversions, without fully incorporating the UMI counts.
- L1+L2: The likelihood function of model 2, equivalent to 4sU scRNA-seq data, incorporating all of the available information together.
- L3: Equivalent to bulk SLAM-seq data without spike-ins, ignoring UMI counts and using only cell-summed T>C conversions.
- L1+L3: The likelihood function of model 3, equivalent to combining bulk SLAM-seq data without spike-ins and scRNA-seq data.

Convergence to the target distribution is shown in Fig. 2 for each likelihood function, confirming that scRNA-seq data cannot resolve κ or δ, but does converge for the other parameters, while L2 and L1+L2 converge for all parameters, confirming that 4sU scRNA-seq data can time-resolve bursting. Unlike L2, L3 is unable to converge for any parameters other than δ, further demonstrating the advantage of cell-specific vs cell-summed T>C conversion data. Conversely, L1+L3 does converge for all parameters, with L1 informing burstiness while L3 informs timescales.

The resulting posteriors (Fig. 3) indicate that the accuracy and precision of estimates for a, b and μ are improved by incorporating the single-cell 4sU conversion data compared to relying solely on scRNA-seq or scRNA-seq with bulk SLAM-seq data, which is because the cell-cell variance in the T>C rate is a function of the transcriptional noise (burstiness) of the gene as well as turnover and, therefore, including such information makes the estimation more robust. Likewise, we see that while conventional scRNA-seq may not resolve κ or δ, including the UMI count information with the conversion data also results in more precise and accurate estimates of these parameters. This is because the set of T>C conversions is a function of a, b and δ, while the UMI counts are a function of a and b. Therefore, including the UMI data improves inference of a and b, which reduces the error associated with δ in our joint inference approach.

Overall, we see that L1+L2 outperforms all other likelihood functions for all parameters including L1+L3, demonstrating the benefits that a fully integrated analysis of time-resolved bursting dynamics using 4sU scRNA-seq data provides over more limited, separate treatments of subsets of the parameters by combining scRNA-seq (a and b) and bulk SLAM-seq (δ) information. This is apparent in this example of a gene with moderate expression, high transcriptional noise and a transcript lifetime similar to the 4sU pulse duration.

Inference on data from Qiu

We next applied our method to 4sU scRNA-seq data published in 2020 by Qiu et al, which used human K562 cells [38]. Inference on the data from Qiu was carried out for all genes with at least one read and observed T>C conversion in both the 4sU and control datasets, running the MCMC algorithm in parallel on each to obtain a posterior from model 2 or model 3 if required (Methods). The final set of genes to be analysed was selected based on those with sufficient confidence in all parameter estimates. Therefore, a maximum CV value of 0.45 was imposed for all parameter estimates, so that only genes with no CV $>$ 0.45 would be included, leaving 584 genes as the final selected set.

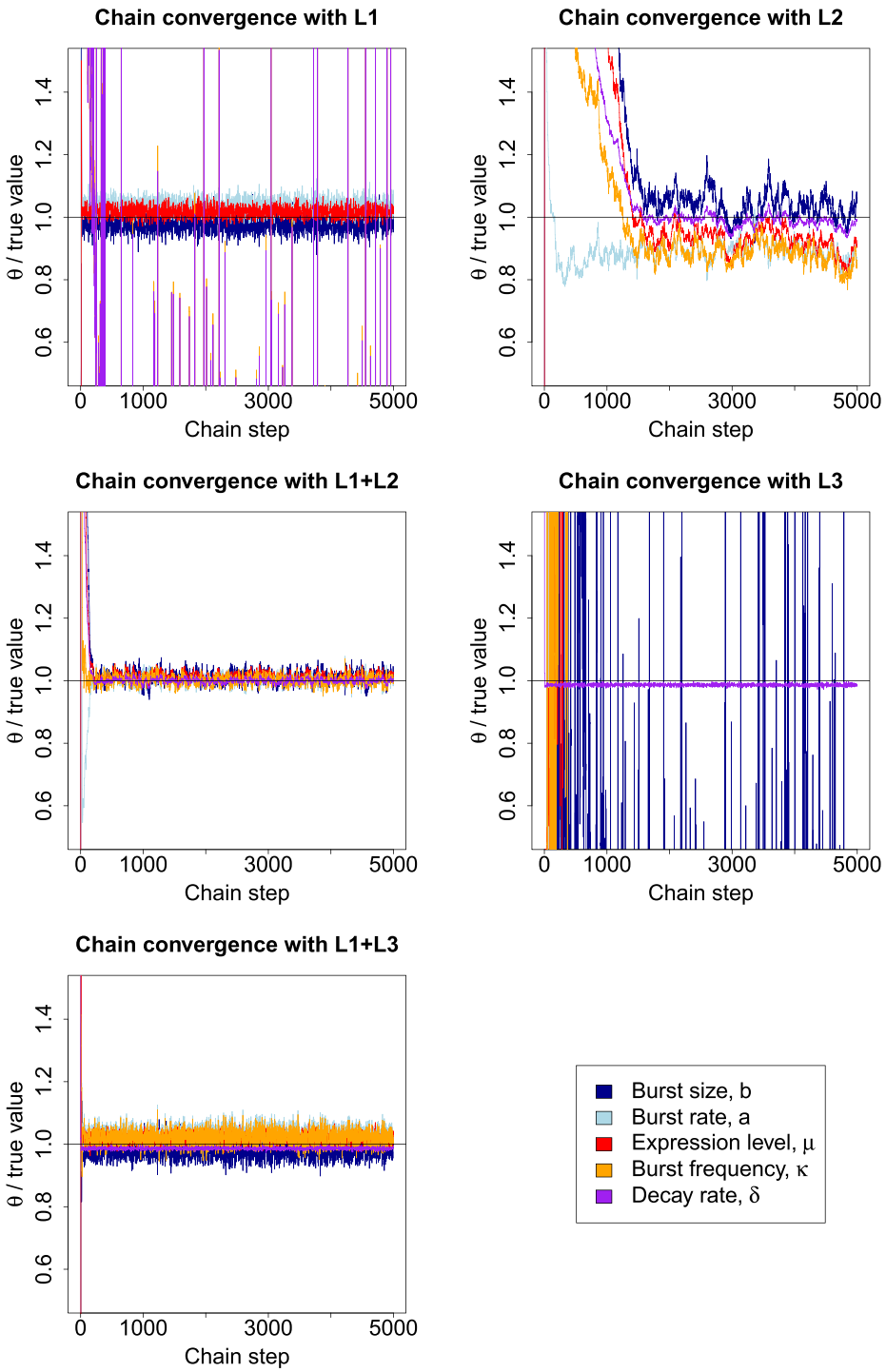


Fig. 2 Convergence of Markov chains to true parameter values with simulated data for three different likelihood functions. The parameter values, θ, in the chain are divided by the true value to allow for joint visualisation, with the black horizontal line representing the target value

For the selected genes we observe that the quality of our estimates depends upon the location of the gene within parameter space, as shown in Fig. 4, which depicts estimate vs CV for all parameters. $CV(\delta)$ has an optimal (minimum) value for δ

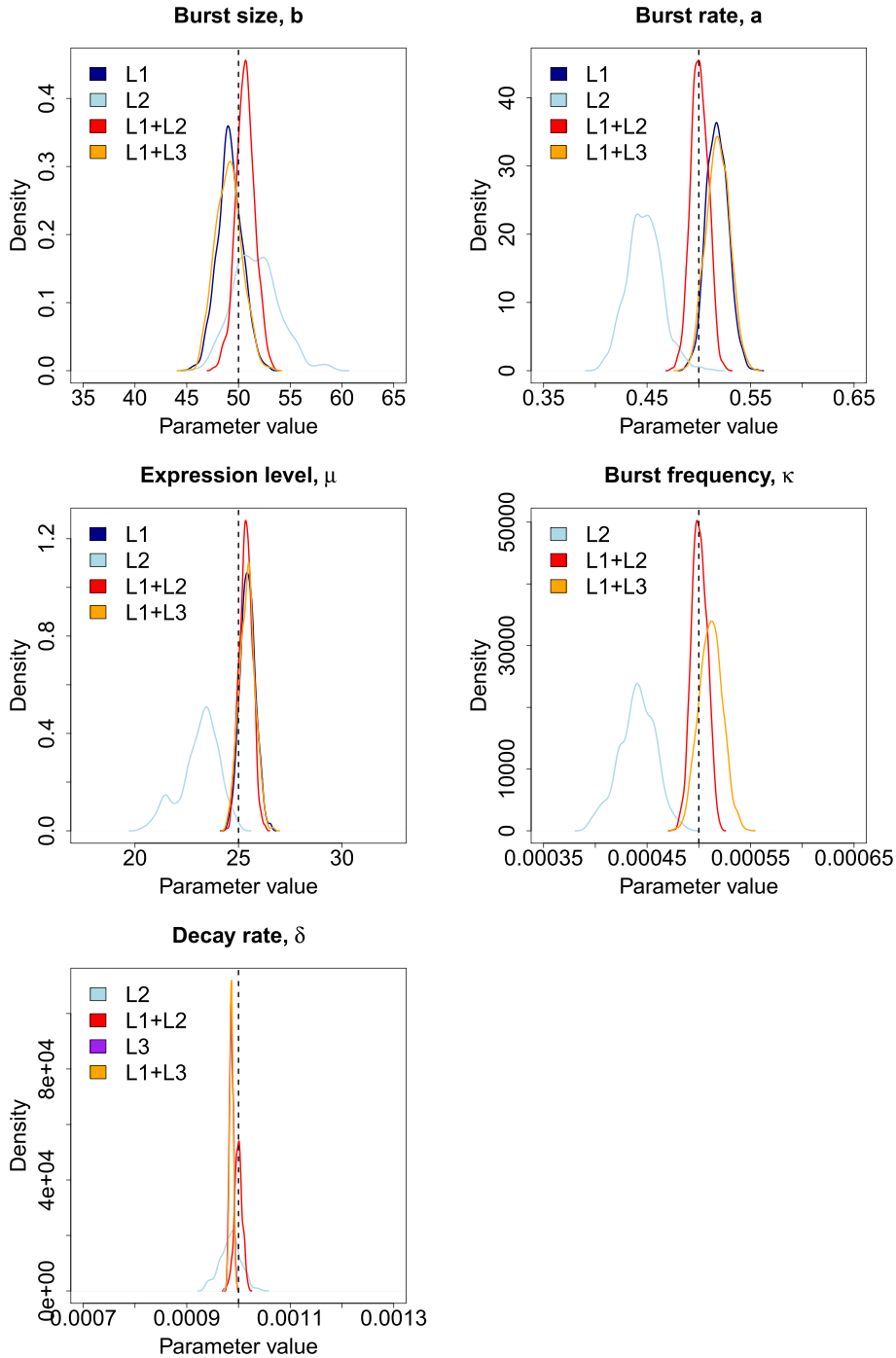


Fig. 3 Probability density functions of each parameter derived from posteriors obtained using different likelihood functions, with the dashed black lines representing the true parameter values that were used to simulate the dataset upon which inference was carried out. The densities for δ obtained with L3 and L1+L3 are difficult to distinguish because they almost perfectly overlap

corresponding to an average transcript lifetime equal to the 4sU pulse duration (4 hours), with confidence decreasing bidirectionally and outliers with very low $CV(\delta)$ corresponding to genes with $\mu \geq 1000$. We also have increased confidence in general

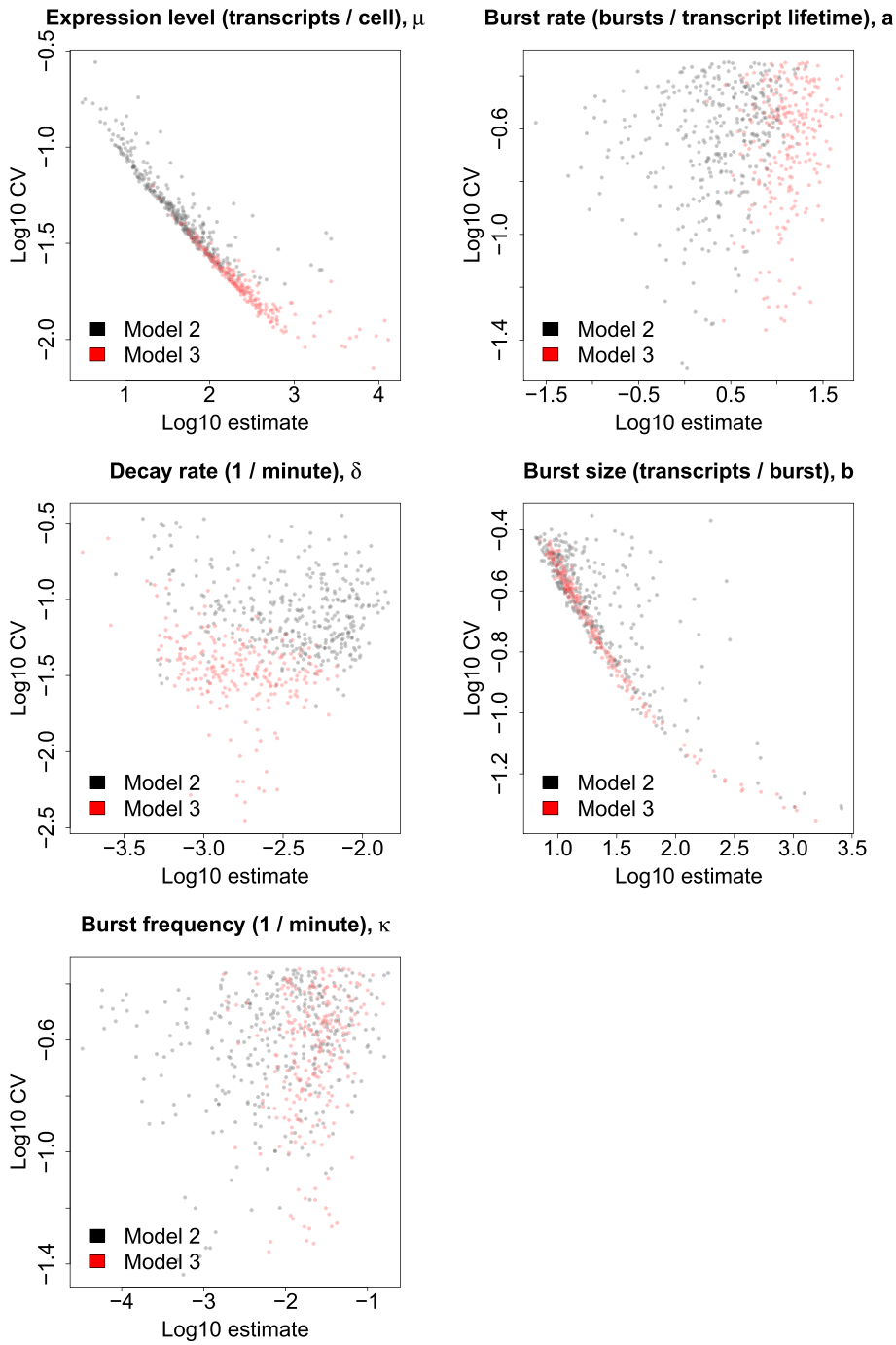


Fig. 4 Estimates vs CVs of all parameters derived from sampled posteriors for all 584 selected genes, with those obtained using models 2 or 3 displayed in black or red, respectively

for genes with higher μ since estimates for such genes are informed by a greater volume of data. Likewise, genes with greater b have greater confidence because, firstly, increased b results in higher μ. Secondly, for a given μ, having a higher b implies lower a, meaning that the transcriptional noise is higher, resulting in a more heavily skewed transcript count distribution (across cells) which may be more precisely

attributed to a region of parameter space. We do not see a visually obvious trend in confidence for a. This is because it is associated with higher expression level but lower transcriptional noise. Therefore, a gene with higher a has more data points with which to inform the estimate but a less skewed transcript count distribution, so that the effects on confidence tend to cancel each other out. The trend in confidence for κ is essentially dictated by the a and δ values for the gene.

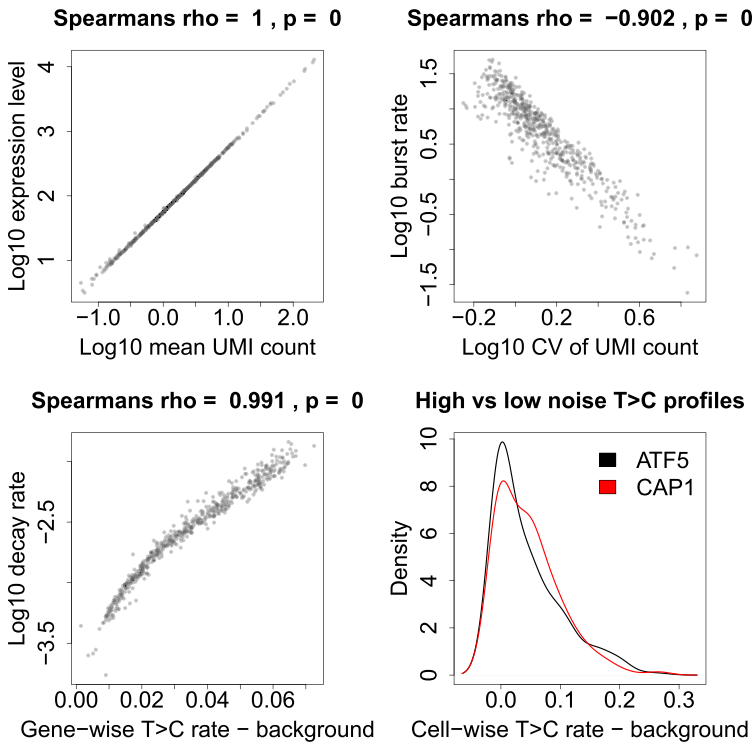
Instead of relying solely on model 2, for some genes we must switch to an alternative (model 3). This occurs when genes lie within a region of parameter space such that the solution to Eq. 9 becomes unstable. Figure 4 provides evidence supporting the reliability of our inference approach, since the model 2 and 3 genes generally occupy the same regions of the plot and exhibit the same relationships between confidence and estimate for each parameter. This also illustrates the increased probability for a gene to reside within unstable parameter space, and therefore require use of model 3, when μ and a are higher and when δ is lower.

We reinforce our results by demonstrating a strong positive correlation about the diagonal between our estimates of δ and cell-matched values calculated in [36] for the same genes (Additional file 1: Fig. S3). Further assessment of our parameter estimation and confidence quantification was provided by carrying out inference on simulated data. This simulation-based validation differs from the previously described model comparison analysis (Figs. 2 and 3) in that experimental settings, such as cell number, cell capture efficiency and sequencing depth, were equivalent to those in the Qiu dataset rather than being idealised, and the bursting parameter values estimated for each of the 12276 genes we analysed were used as the true values for a corresponding simulated gene. Strong, tight correlations about the diagonal between estimates and true parameter values confirmed the capacity for the algorithm to recover known parameter values (Additional file 1: Fig. S4).

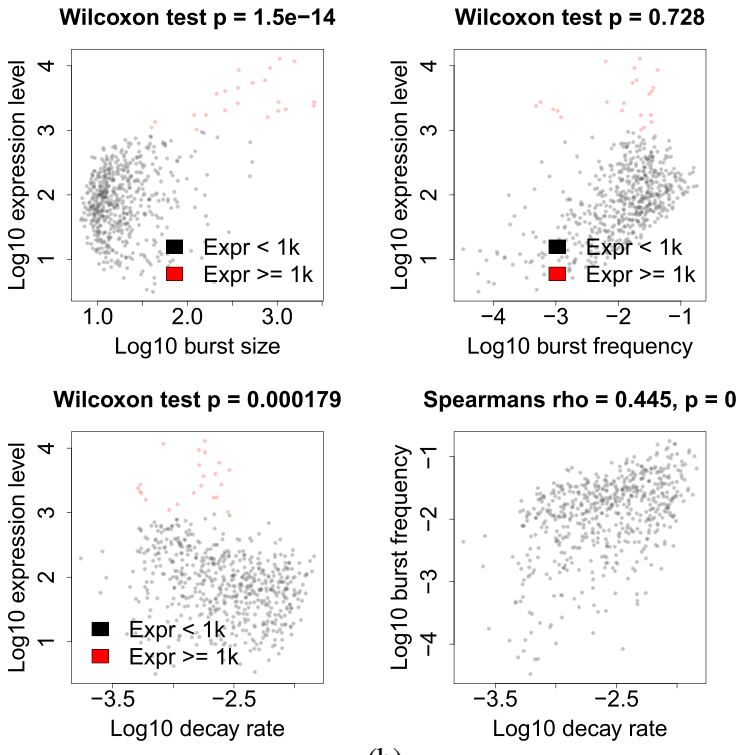
Now that we have estimates for all parameters of interest, it is possible to demonstrate how the different aspects of the data feed into informing the joint probability distribution. Figure 5a illustrates some expected correlations, showing that μ correlates very strongly with the mean UMI count and that δ correlates very strongly with the 4sU - control T>C rate, since these values reflect the overall activity and turnover of the gene, respectively. We see that a correlates strongly against the CV of the UMI count, which reflects the relationship between bursting and cell-cell variability. It is also possible to demonstrate the aforementioned complex relationship between burstiness and the shape of the single-cell T>C count data, but not in a genome-wide manner since the effect is masked by variation in μ and δ. Therefore, we instead compare a pair of genes (*ATF5* and *CAP1*) with very similar estimates for μ and δ but very different values of a (and

(See figure on next page.)

Fig. 5 **a** Correlations between statistics of the observable data and related bursting parameter estimates, with Spearman's rank correlation strength (rho) and statistical significance (p) displayed. Bottom right compares the cell-specific T>C rates minus gene-specific background for the *ATF5* and *CAP1* genes, which are expressed with high and low noise, respectively. **b** Estimates for different parameters plotted against each other. Statistical significance of difference in b, κ and δ for genes with very high ($\mu \geq 1000$) expression level vs other genes ($\mu < 1000$) is shown with the p-value calculated using the Wilcoxon test. Also shown in the bottom right is the Spearman's rank correlation strength (rho) and statistical significance (p) of κ against δ



(a)



(b)

Fig. 5 (See legend on previous page.)

Table 1 Parameter estimates for the *ATF5* and *CAP1* genes

	ATF5	CAP1
b	66.1	10.2
μ	62.5	66.1
κ	0.00426	0.0289
δ	0.00447	0.00409
a	0.957	7.07

Table 2 Parameter values for simulated high and low noise genes

	High noise	Low noise
b	250	25
μ	250	250
κ	0.001	0.01
δ	0.001	0.001
a	1	10

therefore also b and κ), with *ATF5* being expressed in a far more bursty fashion. The estimates for the different parameters of these genes are given by Table 1.

The density plot in Fig. 5a compares the distribution of cell-specific T>C rates (minus gene-specific background) across all reads in the cell for the aforementioned pair of genes. There is a clear difference in the shape of the distribution, with the bursty gene having a greater density at either extreme while the gene with less noisy expression has a greater intermediate density. This is because large, infrequent bursting has a binarising effect, meaning that most cells either have a low or high T>C rate. Those with a low rate correspond to those which have had no bursts occur during the 4sU pulse, resulting in their entire transcript population comprising those surviving from before the pulse. Those with a high rate correspond to those which have had at least one burst occur during the pulse. Since the bursts tend to be large, this results in the majority of the transcript pool being comprised of newly synthesised transcripts. On the other hand, smaller, more frequent bursts causes the surviving transcripts to gradually become replaced by new transcripts in a more uniform manner across cells. Similarly to how scRNA-seq reveals differences in cell-cell variation in transcript counts for two genes with otherwise equal expression levels, 4sU scRNA-seq also reveals differences in cell-cell variation in new transcript proportions for two genes with otherwise equal decay rates.

Despite controlling for μ and δ in this pairwise comparison of a high vs low noise gene, the effect of bursting on cell-specific T>C rates shown in Fig. 5a is still somewhat obscured by the variable cell-specific capture efficiencies, α, present in the data. Therefore, datasets were simulated in the same manner as for the model comparison analysis, except $\lambda_s = 0.001$, and $\alpha = 1$ to totally control for the effect of capture efficiencies. Datasets were simulated for a gene with high noise and another with low noise with parameter values set as shown in Table 2.

The differential transition from surviving to new transcript pool for high and low noise genes is demonstrated in Additional file 2, which shows the cell-specific T>C rate distributions for data simulated with different pulse durations. This illustrates the previously discussed effect of bursting on cell-cell turnover variation more clearly, visualising the bimodal vs unimodal transitions occurring under high vs low noise conditions with a video.

Biological findings

Correlating the parameter estimates against each other for our 584 genes also reveals that genes with extremely high expression levels, the majority of which are mitochondrial genes, are able to achieve these high levels primarily by having very large bursts, rather than very frequent bursts or very stable transcripts, although the decay rates do appear somewhat constrained (Fig. 5b). There may be biological upper limits on κ due to the various factors required to be in place to prime a gene for activity and, therefore, it may be preferable to instead increase burst duration (reduce k_{off}), and therefore burst size, for very high expression levels [32]. A similar phenomenon has been observed previously, in which *MYC* overexpression lead to increased expression in target genes through increased burst duration and size, rather than increased burst frequency [46, 47]. Estimates for κ and δ are also positively correlated, despite κ and δ varying across several orders of magnitude. This correlation may be the result of a selective pressure to limit the variability of a and, for example, prevent transcriptional noise levels from becoming excessively high. Alternatively, high burst frequencies would correspond to RNAP rapidly processing over the gene, allowing less time to pause for the nascent transcript to be folded/spliced appropriately than with lower burst frequencies [2], resulting in reduced transcript stability.

Histone modifications and bursting

We next explored the relationship between HMs and transcriptional bursting dynamics with a metagene analysis carried out using ChIP-seq data for eight previously analysed HMs [20] and two further HMs (Methods). In this analysis, we removed mitochondrial genes and genes for which we lacked HM data from our set with high confidence parameter estimates, with 505 genes ultimately being included. Of the eight previously studied HMs we analysed, the profiles generally fall into the two previously described categeories [20], being either predominantly promoter-localised (H3K4me2, H3K4me3, H3K9ac, H3K27ac, Fig. 6 and Additional file 1: Figs. S6-8) or gene body (GB)-localised (H3K4me1, H3K36me3, H3K79me2, H4K20me1, Additional file 1: Figs. S9-13). To better understand the association between HM profile and bursting parameters, the genes were split in half, sorted by parameter estimate for each of the five parameters. Metagene comparison reveals position-dependent associations for promoter-localised HMs, using H3K4me2 as an example (Fig. 6). It appears that HM presence at the promoter and through the GB is associated with increased μ and also a, while increased κ is specifically associated with promoter but not GB presence. Conversely, presence through the GB excluding the promoter region appears associated with increased b and reduced δ.

This analysis builds upon a previous scRNA-seq study which correlated bursting parameter estimates with HM localisation by averaging the ChIP-seq coverage from

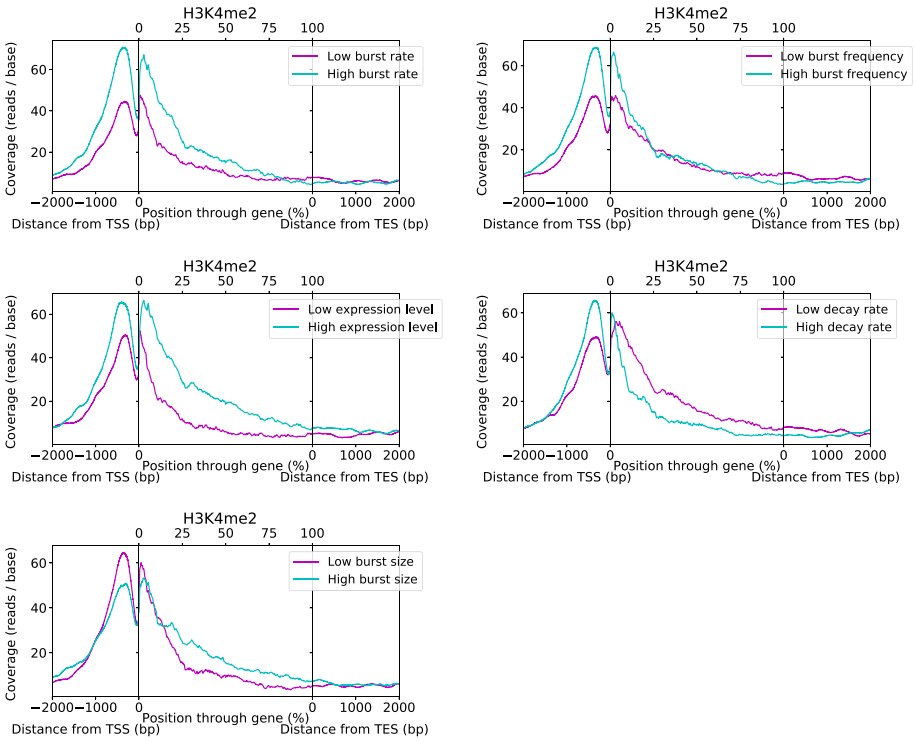


Fig. 6 Metagene plots of H3k4me2 coverage, comparing profiles for the top and bottom 50% of genes when split according to their estimates for each parameter, denoted by high and low, as indicated

2000 bp upstream of the TSS to the TES for each gene [20]. They were unable to obtain estimates of κ or δ due to a lack of published data on transcript turnover rates for the cell type (hESCs). Our results are in agreement with [20] despite having a different cell type, but additional complexities are revealed which are only apparent with our metagene analysis combined with the capacity to estimate κ and δ afforded by 4sU scRNA-seq. For promoter-localised HMs, they report positive associations between HM presence and both a and b, whilst we demonstrate that the association with b is specific to the GB. We confirm that the association with a holds throughout both the promoter and GB, but show that this is a result of a promoter-specific positive κ association and a GB-specific negative δ association, thereby further demonstrating the advantages of 4sU scRNA-seq inference.

In order to statistically test these apparent associations, the average HM coverage values around the promoter and through the GB excluding the promoter were obtained for each HM (Methods), taking the average value from 2000 bp upstream of the TSS to 5% through the GB (-2000:5%) and from 5% through the GB to the TES (5%:100%), respectively. Spearman's rank correlation of the mean value for each promoter-localised HM against each parameter across our 505 genes confirmed the direction and quantified the strength (Fig. 7a), as well as confirmed the statistical significance of the suspected associations (Fig. 7b).

The association between promoter-localised HM presence and reduced decay rate is consistent with previous reports of a link between HMs and pre-RNA processing. The RNAP elongation speed may be modulated by HMs or they may be responsible for

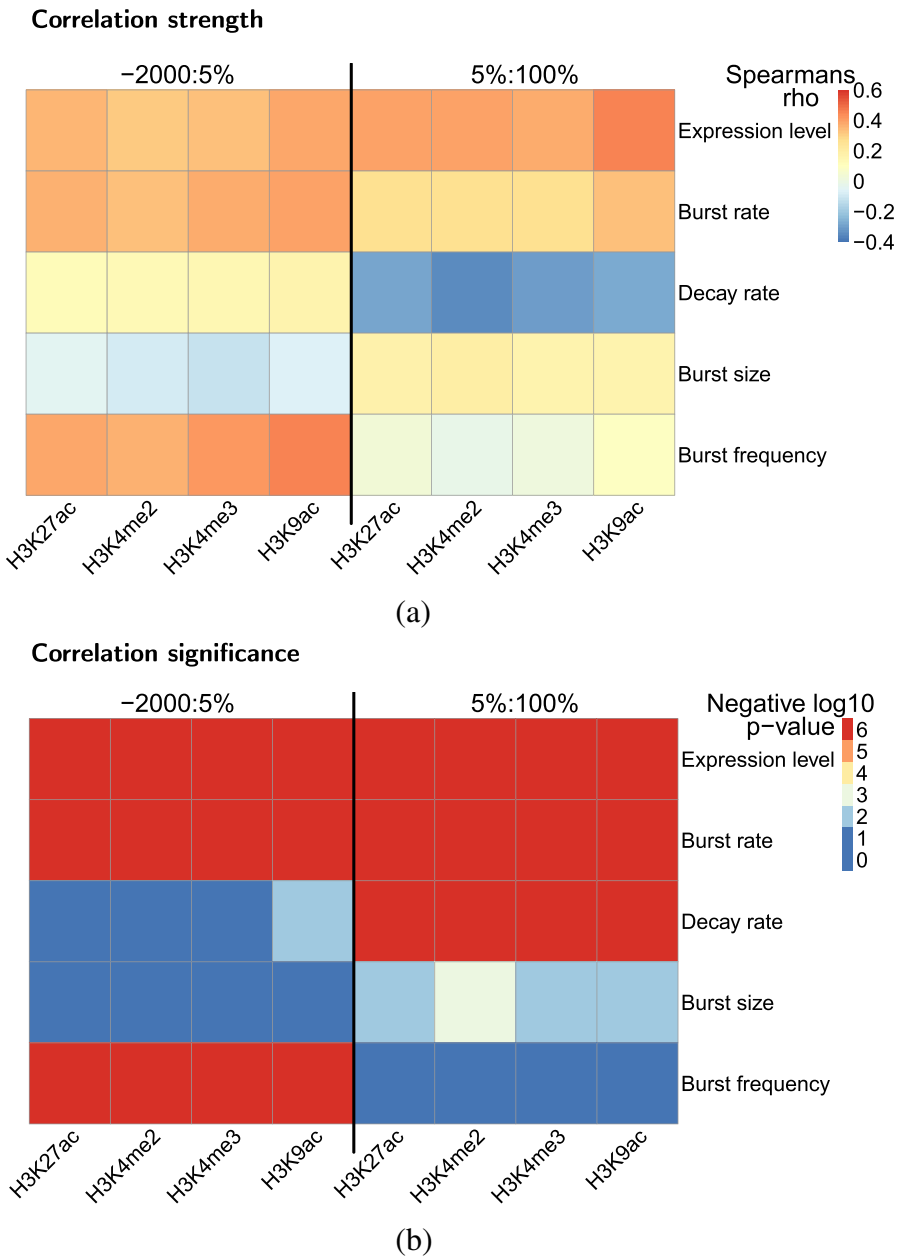


Fig. 7 **a** Heatmap showing the Spearman's rank rho as the heat intensity value for the correlations between bursting parameter estimates and the mean promoter-localised HM coverage values across the -2000:5% and 5%:100% regions. More intense red or blue colouration indicates a stronger positive or negative correlation, respectively, while neutral indicates no/weak correlation. **b** Heatmap showing the Spearman's rank *p*-value (adjusted for multiple hypothesis testing) as the heat intensity value for the correlations between bursting parameter estimates and the mean promoter-localised HM coverage values across the -2000:5% and 5%:100% regions. The heat values are discretised, corresponding to negative log10 *p*-value thresholds. For example, the most intense blue indicates that, for the given correlation, $10^{-2} < p$, meaning no statistical significance, the neutral colour indicates that $10^{-4} < p \leq 10^{-3}$, while the most intense red indicates that $p \leq 10^{-6}$

the recruitment of splicing factors [48, 49]. This could result in more stable RNA by ensuring correct splicing and/or polyadenylation. GB presence of promoter-localised HMs could also result in increased burst size by facilitating TSS-TES contact through

the maintenance of the open chromatin state around the TES. Coupled with the free movement of RNAP through the GB, this may increase the burst size by allowing RNAPs to quickly and repeatedly generate multiple transcripts by promoting polymerase recycling [14]. Another hypothesis is the presence alternative TSSs in the GB, which may transcribe simultaneously upon gene activation, causing increased burst size, with 483 out of the 505 analysed genes exhibiting alternative TSSs according to the gtf. Metagenes for the rest of the aforementioned HMs along with the correlation/statistical analysis of the GB-localised HMs can be found in Additional file 1, as well as analysis of two additional HMs (H3K18ac and H4K16ac) not analysed in [20] (Additional file 1: Figs. S14-16), but which were shown to be strongly linked to active enhancer regions [50].

Discussion

With the inference approach presented here, we demonstrate the capacity to obtain genome-wide estimates of the parameters governing transcriptional bursting dynamics and the timescales upon which they occur from a single dataset with no prior knowledge. By sampling from the full joint probability distributions of the parameter values given the data, we are able to quantify confidence in our estimates and take into account the complex interdependencies between the different parameters and 4sU scRNA-seq data, revealing the regions of parameter space for which we have the most accurate and precise estimates. We show that the distribution of 4sU-induced T>C conversions across cells is shaped not only by the turnover rate and expression level of the gene but also by the transcriptional noise and that this information can therefore be used to improve estimates of burst rate (a) and burst size (b) beyond the level obtainable with conventional scRNA-seq. In this way, combining metabolic labelling and single cell resolution has an effect greater than the sum of their parts on inference power. Previous analysis of transcriptional bursting using 4sU scRNA-seq data has tapped into this idea by estimating the proportion of new transcripts (based on T>C conversions) in each cell for a particular gene and then using the standard deviation of this new to total ratio as a proxy for burstiness [43]. However, as clearly demonstrated by the video in Additional file 2, this distribution, and therefore its standard deviation, is shaped not only by transcriptional noise but also by RNA turnover and may be skewed by technical noise such as variation in capture efficiency. Therefore, along with the overall expression level, this needs to be explicitly accounted for in order to accurately quantify burstiness, as is naturally achieved with our mathematical model.

Having genome-wide estimates of the parameters governing transcriptional dynamics means that it is possible to use the variation which naturally exists between genes to examine the relationships between the different parameters and other features, such as HMs, instead of having to rely on experiments which artificially perturb the cells to gain insight via a single gene system. In agreement with previous reports [42], we find that the genes with very high expression levels are primarily mitochondrial genes. Going beyond this, we show that such activity levels are achieved by having large burst sizes rather than increased RNA stability or burst frequency, which we hypothesise could be due to biological constraints on the rate of switching between active and inactive states [32], potentially making it favourable to instead increase the duration of bursts,

and therefore the burst size, as has similarly been observed for *MYC*-driven transcription [46, 47]. Whereas some studies have found the variation in decay rates (in mESCs) across genes to be an order of magnitude lower than for the other parameters, and therefore negligible [32], we found significant variation in K562 cells which was important to account for in order to properly estimate burst frequencies. This is in line with previous predictions that transcript stability plays an important role in modulating gene expression noise [45]. Indeed, our analysis revealed an unexpected positive correlation between burst frequency and decay rate, resulting in the burst rate, and therefore transcriptional noise, being constrained. One may speculate that only noise levels within a certain range are tolerated, with extreme values resulting in too few cells expressing the gene for a given function to be achieved, such as the appropriate proportion of cells in an isogenic population undergoing differentiation [11, 51], manifesting as the observed correlation. A mechanistic, rather than evolutionary, explanation is that high burst frequencies result in rapid flux of RNAP through the gene, such that less time is allowed for pausing, during which appropriate folding and/or splicing of the nascent transcript is facilitated [2]. This would would reduce transcript stability and cause the observed correlation.

Examining the relationship between bursting parameters and HMs genome-wide produced results consistent with but advancing upon previous work [20]. Combining our metagene analysis with the additional information provided by 4sU scRNA-seq over inference on conventional scRNA-seq reveals intricacies that were not previously apparent. The presence of GB-localised HMs throughout the gene is generally associated with increased burst rate (bursts per transcript lifetime) via increased burst frequency (bursts per minute), while promoter-localised HMs are only associated with increased burst frequency when found around the TSS. Their presence further downstream remains associated with increased burst rate, and therefore reduced transcriptional noise, but through reduced decay rate rather than increased burst frequency. The association with reduced decay rate may be related to the previously documented influence of HMs on pre-RNA processing, which is achieved, for example, by modulating RNAP elongation speed and/or by recruiting splicing factors [48, 49]. This may increase RNA stability by reducing the probability of incorrect splicing or polyadenylation. Presence of promoter-localised HMs throughout the GB but not at the TSS is also associated with increased burst size. Downstream presence could facilitate interactions between the TSS and the TES by maintaining the open chromatin state around the TES. This, along with maintaining the free movement of RNAP through the GB, could promote polymerase recycling and therefore increased burst size by allowing RNAPs to quickly and repeatedly fire off multiple transcripts during an active period [14]. Another possible explanation for the association between promoter-localised HM presence in the GB with burst size is that there are multiple, alternative TSSs found within genes. These may initiate transcription in a non-independent manner, leading to increased burst size when there are more/stronger alternative TSSs, as signalled by HM presence.

The inference approach described here is generally applicable to 4sU scRNA-seq datasets which have RNA spike-ins and UMIs for any organism or cell type. Furthermore, the model could easily be expanded to integrate an arbitrarily large number of repeat experiments by extending the Markov chain according to the product of the likelihood functions of each dataset. Indeed, such a scheme which utilised datasets with different

4sU pulse durations could theoretically characterise the transcriptional dynamics of all genes genome-wide. For example, inference carried out using two datasets with long and short pulse durations would facilitate estimates for genes with long and short transcript lifetimes, respectively, along with everything in between. A caveat of our analysis is the asynchronisation of the cell cycle phase across the population. This may confound the results in two ways, firstly because different phases have a different cellular environment, influencing the global transcriptional dynamics and causing variation in the underlying parameter values for the same gene between cells in different phases. Secondly, there is variation in the copy number of genes throughout the cell cycle, with an unknown proportion of cells having one or two copies of each nuclear gene. Confounding effects on the inference could be resolved by separation of the different subpopulations of cells by cell cycle phase using, for example, fluorescence-activated cell sorting prior to sequencing [33], and/or by using allele-specific/sensitive scRNA-seq approaches combined with metabolic labelling [17, 34]. As 4sU scRNA-seq data becomes more common place and there are improvements in capture efficiencies, sequencing depths and cell numbers, it will be possible to robustly infer time-resolved transcriptional bursting dynamics for a far greater number of genes from a single experimental set up. Our findings on burst dynamics and their associations with HMs could be a valuable starting point to inform future experimental work investigating this area, while further application of our method beyond what is presented here might hint at other, novel mechanistic relations.

Conclusions

In conclusion, we have developed a mathematical model to maximally exploit the power of 4sU scRNA-seq datasets to examine transcriptional bursting, tapping into the synergy between single-cell resolution and 4sU labelling which manifests in the cell-specific T>C rate distributions. The advantages over conventional scRNA-seq were demonstrated in detail using small-scale simulations and performance of the algorithm across parameter space was validated with large-scale simulations. We applied our inference approach to published 4sU scRNA-seq data to obtain genome-wide joint parameter estimates and confidence quantifications, finding an unexpected correlation between burst frequency and decay rate, and that genes with extremely high expression levels achieve this primarily through increased burst size. Finally, we linked our estimates with published ChIP-seq data, revealing position-dependent associations between different histone modifications and parameter estimates which only become apparent with 4sU scRNA-seq as opposed to conventional scRNA-seq.

Methods

Data processing and analysis

4sU scRNA-seq

The main datasets that were used for parameter inference in this study were produced in Qiu et al 2020 [38], downloaded from the GEO series GSE141851. Two datasets from this series were used, both using K562 cells; a negative control dataset with TFEA chemical conversion treatment but with no 4sU added, and another dataset which had 4sU added 4 h before chemical treatment, with GEO sample IDs GSM4512696 and GSM4512697, respectively. These are Drop-seq datasets and thus

were processed according to the "Drop-seq alignment cookbook" (https://mccarrolllab.org/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf). A custom Python script was used to carry out trimming of read pairs with any base with phred quality ≤ 10, and to clip adaptor and polyA tail sequences. The trimmed reads were then aligned to the primary human genome assembly (GRCh38.p13), the fasta file for which was obtained from gencode (https://www.gencodegenes.org/human/), using bwa to build the genome index and for the actual alignment [52]. Custom Python scripts were then used to map the aligned reads with mapq score ≥ 10 to their genes according to the gencode.v36 primary human genome assembly annotation gtf file, before extracting cell-specific (using the cell ID part of the read 1 barcode) UMI counts and total read counts for each gene, along with gene-specific, cell-specific information for each read about the number of genomic T bases (found in the fasta sequence across the aligned read positions) and the number of those which were converted to C bases in the read sequence. Cell selection was then carried out to exclude those cell IDs corresponding to empty droplets by ordering the cell IDs based on the total number of corresponding read pairs and then selecting the top 400 or 795 IDs for the control and 4sU dataset, respectively, as specified in [38]. The control dataset was then used to derive the gene-specific background T>C conversion rates, λ_s, based on the proportion of genomic Ts which were converted to Cs across all reads across all selected cells for the given gene. Figure 8a provides a schematic overview of how the T>C conversion data arises from the 4sU scRNA-seq experimental protocol.

ChIP-seq

Publicly available ChIP-seq datasets for ten active HMs produced with K562 cells were downloaded for our analysis. A H3K4me3 ChIP-seq dataset was obtained from the GEO series GSE108323 with sample ID GSM2895356, which had been processed with alignment to the hg19 human genome build [53]. Seven more ChIP-seq datasets, which had also been processed with alignment to the hg19 human genome build, were obtained from the GEO series GSE29611 with sample IDs GSM733778, GSM733651, GSM733653, GSM733656, GSM733675, GSM733692 and GSM733714, corresponding to H3K9ac, H3K4me2, H3K79me2, H3K27ac, H4K20me1, H3K4me1 and H3K36me3, respectively [54]. Two additional ChIP-seq datasets for H3K18ac (aligned to hg19) and H4K16ac (aligned to hg38) were obtained from the series GSE106964 and GSE158736 with sample IDs GSM2862934 and GSM4809274, respectively [55, 56]. The position and read count information from these datasets was used to obtain the single-base resolution coverage values for each HM. These values were associated with their corresponding genes using the information from the comprehensive gene annotation hg19 (or hg38 for H4K16ac) gtf downloaded from Gencode. Analysis of the correlations between bursting parameter estimates and HM coverage at different sections of the gene was carried out by taking the average coverage value for all bases across the specified section (e.g. from 2k bp upstream of the TSS to the TES) for each gene, so a single value is obtained per gene per HM. Metagene plots were produced by averaging the coverage values for each

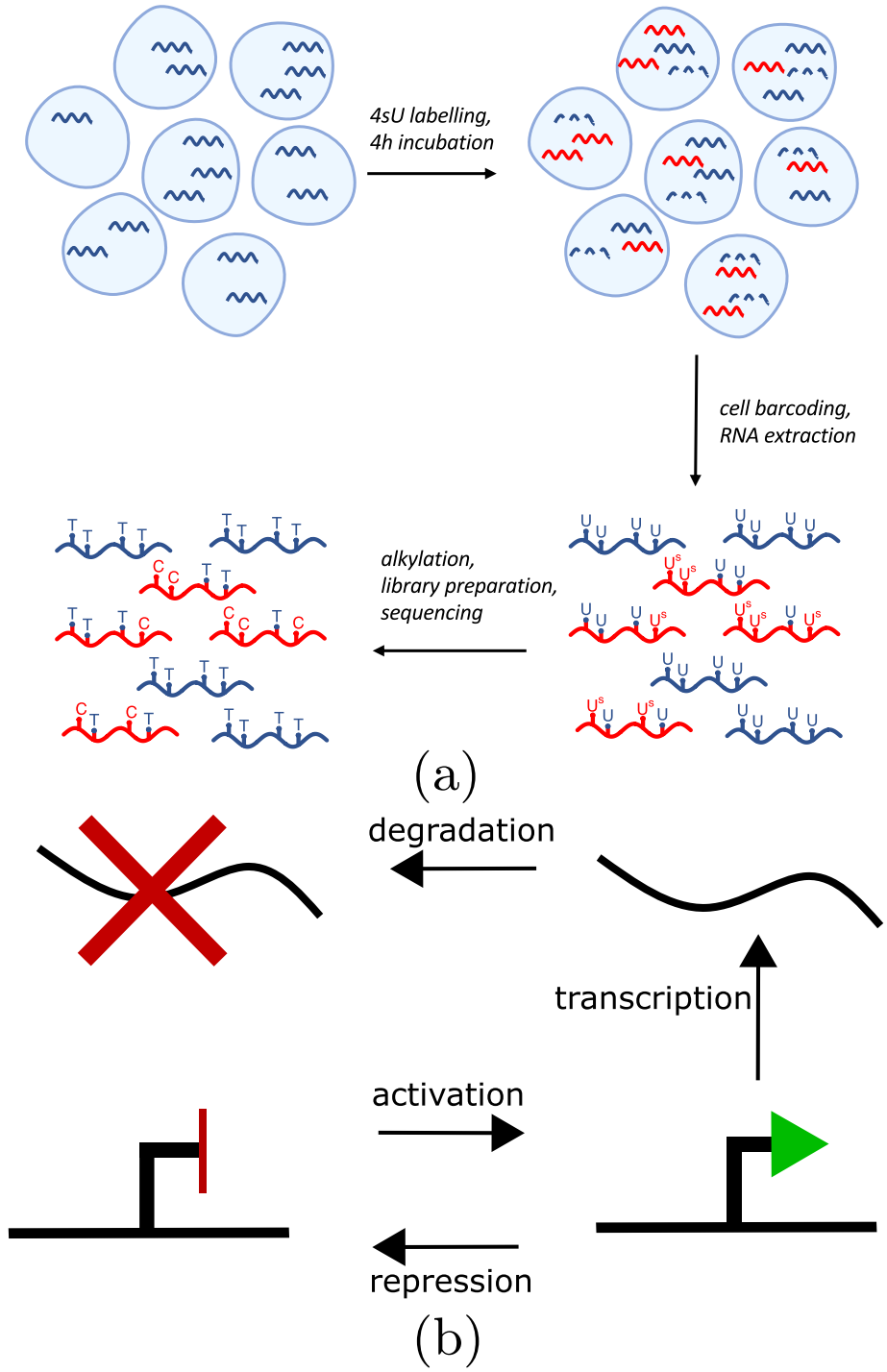


Fig. 8 **a** Cells with variable mRNA content (blue lines) shown for an example gene. Cells are incubated in the presence of the uracil analogue 4sU for a set amount of time (4h). Transcripts that are produced during that period (red) become labelled with 4sU, which is incorporated instead of uracil. During the incubation period, natural mRNA decay also takes place (dashed lines). Following cell barcoding and RNA extraction, the RNA is chemically treated, resulting in the modification (alkylation) of the 4sU moieties incorporated into all labelled transcripts (U^S). In turn, this introduces T-to-C base flip mutations at the points of 4sU incorporation during the first stage of cDNA library preparation (reverse transcription), which are subsequently detected by sequencing. **b** Schematic representation of the two state model, with the four reactions (activation, repression, transcription and degradation) acting on the three species (repressed gene, active gene and transcript)

position through/around the gene across all specified genes, similarly to the meta-gene analysis described in [57].

Mathematical modelling

In general, we model bursty transcription as a stochastic process closely related to the standard two-state model, as many previous works have [9, 10, 14]. The two-state model has four possible processes of gene activation, gene repression, transcription and degradation, where transcription may only occur with the gene in an active state while degradation acts continuously. This is represented by the following chemical reaction scheme

$$G_{off} \xrightarrow{k_{on}} G_{on}$$

$$G_{on} \xrightarrow{k_{off}} G_{off}$$

$$G_{on} \xrightarrow{\beta} G_{on} + RNA$$

$$RNA \xrightarrow{\delta} \emptyset$$

in which k_{on}, k_{off}, β and δ represent the rate constants for gene activation, gene repression, transcription and RNA degradation, respectively, while G_{off}, G_{on} and RNA represent the different species of repressed gene, active gene and transcript, respectively. A schematic representation of the system is shown in Fig. 8b. Gene activation is known to involve facilitated diffusion, in which TFs not only diffuse through the cytoplasm but can also slide along the DNA after becoming associated, to find the target/activation site. This makes the process more complicated than the simple on/off switch of our model, which reflects TF association/disassociation events without DNA sliding. However, previous modelling work has shown that the two-state model accurately captures gene activation due to the high speeds at which DNA-binding proteins slide along the DNA observed in biological systems, which ensures that they do not influence the transcriptional bursting dynamics [58].

With this model, we have burst frequency, $\kappa = \frac{1}{(1/k_{on})+(1/k_{off})}$ and burst size, $b = \frac{\beta}{k_{off}}$, and we recall the burst rate, $a = \frac{\kappa}{\delta}$. Aiming to understand bursting and its timescales specifically, we make the assumption that bursts occur instantaneously, arrive according to a Poisson process and burst in a geometric fashion, which is valid when $\delta << k_{off}$ since a transcript produced in a given burst is unlikely to have degraded before the burst is over [7, 59], and when $k_{on} << k_{off}$, which is supported by the parameter estimates reported in [32]. This model simplifies $\kappa = \lim_{k_{off}\to\infty} \frac{1}{(1/k_{on})+(1/k_{off})} = k_{on}$ while b remains finite with $b = \lim_{\beta,k_{off}\to\infty} \frac{\beta}{k_{off}}$ [60].

Model 1

The first model aims to model the observed unique molecular identifier (UMI) counts of a given cell, l, from the estimated capture efficiency (Additional file 1: Fig. S1) of

that cell, α, in a similar fashion to the technical noise model outlined in [61]. The capture efficiency, α, represents the transcript detection rate for that cell (probability of at least one read corresponding to a particular transcript). Based on the instantaneous bursting version of the two-state model described above, the steady state distribution of the transcript count, m, can be derived directly from the master equation and corresponds to the negative binomial distribution [10, 59, 60, 62]

$$P(m) = f_{N\ Bin}\left(m|a, \frac{b}{1+b}\right) \tag{1}$$

which is illustrated by the schematic in Fig. 9, where

$$f_{N\ Bin}\left(m|a, \frac{b}{1+b}\right) = \frac{\Gamma(m+a)}{\Gamma(m+1)\Gamma(a)}\left(\frac{1}{1+b}\right)^a\left(\frac{b}{1+b}\right)^m$$

The full derivation is available in [60]. We may then model the probability distribution of observing l UMIs given m transcripts in the cell with a capture efficiency of α, as a poisson approximation of the true binomial process

$$P(l|m,\alpha) = f_{Pois}(l|m\alpha) \tag{2}$$

where

$$f_{Pois}(l|m\alpha) = \frac{(m\alpha)^l e^{-m\alpha}}{l!}$$

which is valid when α is small. We model the observed data, linked by the unobserved steady state transcript distribution by compounding Eqs. 1 and 2 across the state space of m and marginalise

$$P(l|\alpha) = \sum_{m=0}^{M} P(l|m,\alpha)P(m) \tag{3}$$

where M is an upper bound corresponding to the 0.9999 quantile of Eq. 1, which avoids summing to ∞, achieving a finite state projection (FSP) [63, 64] with an error of 0.0001. The resulting distribution and its dependence on $P(m)$ is shown in Fig. 9. The approximation in Eq. 2 permits non-zero probability values when $m < l$, which allows our MCMC algorithm to more efficiently escape regions of parameter space for which $M < l$. This leads us to the likelihood function of model 1 by taking the product of Eq. 3 across all cells in the data

$$P(L|\theta) = \prod_c P(l_c|\alpha_c) \tag{4}$$

where l_c and α_c represent the observed UMI count (for the given gene) and capture efficiency for cell c, respectively, and $L = (l_1, \ldots, l_k)$, with k cells in total in the data and $\theta = (\mu, a, \gamma)$. Since we wish to infer the values of θ for each gene from the data using this model, we aim to obtain the posterior

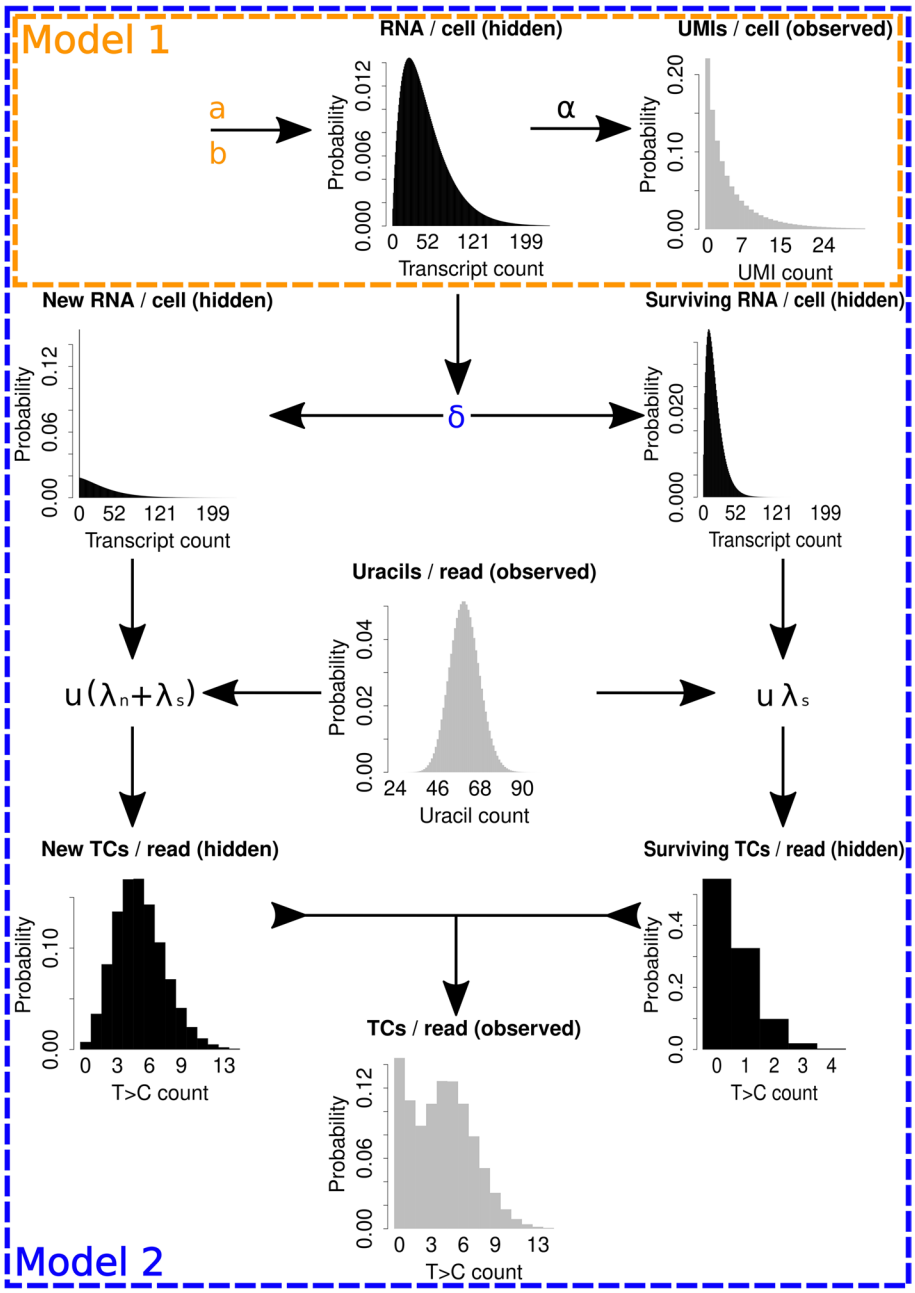


Fig. 9 Schematic showing several of the hidden (black) and observed (grey) data we model and their governing parameters. For this illustration, values were set as $a = 2, b = 25$ and $\delta = 0.001$ for the biological parameters and $t = 1000, u \sim Pois(60)$, $\lambda_n = 0.075$, $\lambda_s = 0.01$ and $\alpha \sim Beta(1, 9)$ for the technical parameters. The encompassing boxes indicate the information used during parameter inference by model 1 (a and b) and 2 (a, b and δ). The direction of the arrows indicate how the distributions feed into each other as dictated by the accompanying parameters. For example, a and b determine the steady state distribution, which determines the new and surviving transcript count distribution as dictated by δ for given t, while the new and surviving T>C count distributions combine to form the observed T>C count distribution, which is conditional upon the cell's transcript count, m, with $m = 100$ shown here. More information on estimating α and λ_n specifically for the Qiu dataset is found in Additional file 1: Figs. S1 and S2, respectively

$$P(\theta|L) = \frac{P(L|\theta)P(\theta)}{\int_\theta P(L|\theta)P(\theta)d\theta} \tag{5}$$

which we achieve through MCMC sampling.

Model 2

We will now construct a model which unifies the UMI and T>C conversion aspects of the data with the aim of understanding both bursting dynamics and the timescale upon which they occur. Figure 8a illustrates how the T>C conversion data arises from the experimental protocol. First of all, we define $\tau = t\delta$ where t is the time before sequencing at which the 4sU nucleotides were added to the cells, otherwise known as the pulse duration. τ therefore represents unitless time in terms of transcript lifetimes. Next, we must obtain the probability mass function of the number of transcripts surviving to the sequencing point which were produced before the 4sU was added, otherwise known as the surviving transcripts, s. This distribution, $P(s)$, may be understood as the time-decay of the steady state distribution, $P(m)$, where we have $\lim_{t\to\infty} P(s = 0) = 1$ and $P(s|t = 0) = P(m)$ when $\delta > 0$. Degradation acts upon each individual transcript molecule with rate δ, and therefore the probability of a given transcript produced before 4sU was added surviving is $1 - F_{Exp}(X \leq t|\delta) = f_{Pois}(0|\tau)$. Therefore, the probability of having s transcripts surviving given m originally is

$$P(s|m) = f_{Bin}(s|m, f_{Pois}(0|\tau)) \tag{6}$$

where

$$f_{Bin}(s|m, f_{Pois}(0|\tau)) = \binom{m}{s} f_{Pois}(0|\tau)^s F_{Exp}(X \leq t|\delta)^{m-s}$$

and

$$F_{Exp}(X \leq t|\delta) = 1 - e^{-\tau}$$

giving the conditional distribution of s. Compounding this with the steady state distribution (Eq. 1) we obtain the marginal

$$P(s) = \sum_{m=0}^{M} P(s|m)P(m) \tag{7}$$

We compute this distribution efficiently by using the approximation

$$P(s) = f_{N\ Bin}\left(m|a, \frac{f_{Pois}(0|\tau)b}{1 + f_{Pois}(0|\tau)b}\right) \tag{8}$$

Next, we obtain the probability mass function of the newly synthesised transcript count, $P(n)$, for those transcripts that were produced after the 4sU was added and therefore have a higher T>C conversion rate than the background. This may be understood in reverse to $P(s)$, as it describes the convergence of the newly synthesised transcript count from a point mass at zero to the steady state distribution where we have $P(n = 0|t = 0) = 1$ and $\lim_{t\to\infty} P(n) = P(m)$ when $a, b, \delta > 0$. An approximate solution

to such a distribution was derived as a model of translation in [59] though the assumed relationships apply here. The solution is

$$P(n) = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{1+b}\right)^{n}\left(\frac{1+be^{-\tau}}{1+b}\right)^{a}{}_2F_1\left(-n,-a,1-a-n;\frac{1+b}{e^{\tau}+b}\right) \tag{9}$$

which is valid when $k_{off} >> \delta$ and $\tau >> \delta/k_{off}$, where ${}_2F_1$ refers to the hypergeometric function. The general dependency of the surviving and new transcript distributions on $P(m)$, as dictated by δ, is illustrated by Fig. 9. Next, we obtain the probability distribution of transcripts at steady state conditional on our observed cell-specific capture efficiency, α, and UMI count, l, by using Eqs. 1 and 2

$$P(m|l,\alpha) = \frac{P(l|m,\alpha)P(m)}{\sum_m P(l|m,\alpha)P(m)} \tag{10}$$

Now, we describe the probability distribution of n conditional on m as the joint distribution of n and s

$$P(n|m) = \frac{P(n)P(s=m-n)}{\sum_{n=0}^{m} P(n)P(s=m-n)} \tag{11}$$

with the convolution $\sum_{n=0}^{m} P(n)P(s=m-n) \approx P(m)$ being used as a normalising value in place of $P(m)$ due to the approximate nature of $P(n)$, ensuring that $\sum_{n=0}^{m} P(n|m) = 1$. It is now possible to model the number of T>C conversions observed in a given read conditional on m, where we have expanded and built upon the poisson mixture model of conversions described in [36] and compounding with Eq. 11

$$P(i|m) = \sum_{n=0}^{m}\sum_{u} P(u)\left(\frac{n}{m}f_{Pois}(i|u(\lambda_n+\lambda_s)) + \left(1-\frac{n}{m}\right)f_{Pois}(i|u\lambda_s)\right)P(n|m) \tag{12}$$

where $P(u)$ is the gene-specific empirical probability mass function of observing u uracils across the fasta sequence corresponding to a given read's mapping position. λ_s is the gene-specific background conversion rate observed in the control dataset (without the addition of 4sU) which represents conversion due to random mutations or other sources outside of chemical conversion. λ_n is the gene-invariant conversion rate due to 4sU incorporation and conversion which was estimated from the data (Additional file 1: Fig. S2). $P(u)$ and the conditional T>C count distribution are shown in Fig. 9, along with the dependence of $P(i|m)$ on $P(u)$, $P(s)$ and $P(n)$ as dictated by λ_n and λ_s. We may now model the cell-specific T>C conversion rate for the given gene by compounding Eqs. 10 and 12

$$P(i|l,\alpha) = \sum_{m=0}^{M} P(i|m)P(m|l,\alpha) \tag{13}$$

where M is an upper bound corresponding to the 0.9999 quantile of Eq. 1, again giving a FSP with error 0.0001. We are finally in a position to complete the model and link all our observables together. The observed counts of conversions in each cell may be

represented by y, where y_i is the number of reads that have i conversions. Therefore, the cell-specific observed distribution of conversions per read may be understood as a multinomial distribution with a probability vector determined by Eq. 13

$$P(y|l,\alpha) = \frac{(\sum_i y_i)!}{\prod_i y_i!} \prod_i P(i|l,\alpha)^{y_i} \tag{14}$$

enabling us to model the conversion data conditional on the UMI data. A likelihood function may now be obtained with

$$P(Y|L,\theta) = \prod_c P(y_c|l_c,\alpha_c) \tag{15}$$

where y_c is the conversions per read distribution observed in cell c and $Y = (y_1,\ldots,y_k)$ where $y_{c,i}$ is the number of reads with i conversions in cell c for the given gene. The final likelihood function of model 2 is now defined as the product of Eqs. 4 and 15

$$P(Y,L|\theta) = P(Y|L,\theta)P(L|\theta) \tag{16}$$

As in Eq. 5, MCMC sampling was used to obtain

$$P(\theta|Y,L) = \frac{P(Y,L|\theta)P(\theta)}{\int_\theta P(Y,L|\theta)P(\theta)d\theta} \tag{17}$$

One thing to note about model 2 is that Eq. 9 is an approximate solution and breaks down in certain regions of parameter space. When a and/or b become too large and/or τ becomes too small, the function will oscillate around the true probability distribution function, with these oscillations quickly becoming more extreme to the point that the approximate solution gives negative probability values. The solution can be said to become unstable in these regions of parameter space, and therefore such regions will be referred to as unstable parameter space. If a gene is found to reside within an unstable region of parameter space then an alternative to model 2 must be used.

Model 3

Our final model acts as an alternative to model 2 when a gene resides within an unstable region of parameter space. Unlike model 2, this model ignores the cell-specific T>C information in favour of simply pooling the conversions across all cells. We define the probability distribution of observing i conversions for a given read

$$P(i) = \sum_u P(u)\left[F_{Exp}(X \le t|\delta) f_{Pois}(i|u(\lambda_n + \lambda_s)) + f_{Pois}(0|\tau) f_{Pois}(i|u\lambda_s)\right] \tag{18}$$

This is similar to Eq. 12 but is independent of the total transcript count, m, and is therefore not cell specific. We can apply Eq. 18 to the full set of observed conversions across cells, Y, again using the multinomial distribution to obtain a likelihood function

$$P(Y|\theta) = \frac{(\sum_i y_i)!}{\prod_i y_i!} \prod_i P(i)^{y_i} \tag{19}$$

where y_i represents the number of reads with i conversions summed across all cells rather than being a cell-specific value as in Eqs. 14 and 15. We define the final likelihood function of model 3 as the product of Eqs. 4 and 19.

$$P(L, Y|\theta) = P(L|\theta)P(Y|\theta) \tag{20}$$

As in Eqs. 5 and 17, MCMC sampling was used to obtain

$$P(\theta|L, Y) = \frac{P(L, Y|\theta)P(\theta)}{\int_\theta P(L, Y|\theta)P(\theta)d\theta} \tag{21}$$

Markov chain Monte Carlo algorithm

MCMC was employed in order to sample from the posterior distributions outlined in Eqs. 5, 17 or 21 using a Metropolis-adjusted Langevin algorithm (MALA) within a Gibbs sampler, which simulates a Markov chain using Langevin dynamics [65] and corrects the Euler-Maruyama integration error with an accept-reject step as with the Metropolis-Hastings algorithm [66]. The chain is initialised semi-randomly, setting $\theta^{(1)}$ in a manner which takes advantage of the information immediately available from the data to start the chain relatively close to the target density. We calculate empirical estimates of the expression level, μ, and transcript lifetime, γ, as

$$\hat{\mu} = \frac{1}{N} \sum_{c=1}^{N} l_c/\alpha_c$$

where $N = 795$ is the number of cells in the dataset, and as

$$\hat{\gamma} = -t/\log(\max[0.1, \min\{0.9, (1 - ((\lambda - \lambda_s)/\lambda_n)\}])$$

where λ is the observed conversion rate for the given gene across all reads, while λ_s and λ_n represent the background conversion rate measured in the control dataset and the estimated 4sU-mediated conversion rate, respectively. We then set $\mu = \hat{\mu}$ and draw

$$a \sim LUnif\,(1, 10)$$

and

$$\gamma \sim \mathcal{N}(\hat{\gamma}, \hat{\gamma}/5)$$

where

$$f_{LUnif}\,(x|y, z) = \frac{1}{x \ln(z/y)}$$

with support $[y, z]$ for $y > 0$ and

$$f_{\mathcal{N}}(x|M,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-M}{\sigma}\right)^2}$$

We repeatedly draw $\theta^{(1)}$ in this way until $P(X|\theta^{(1)})P(\theta^{(1)}) > 0$ where X is the dataset and $P(\theta)$ represents the prior distribution, which in this case is defined to be an uninformative multivariate uniform distribution such that

$$P(\theta = (\mu, a, \gamma)) = f_{UInif}(\mu|0, 100000) f_{UInif}(a|0, 100000) f_{UInif}(\gamma|1, 100000)$$

where

$$f_{UInif}(x|y,z) = \frac{1}{z-y}$$

with support $[y, z]$. At each step, j, in the Markov chain, the next step is sampled by proposing jumps to new positions in parameter space from the current position, proceeding through three dimensional parameter space with $\theta = (\mu, a, \gamma)$. This parameterisation was chosen for Markov chain progression to minimise correlations between parameters and proposals to negative (unsupported) values. The classic Metropolis-Hastings algorithm [66] corresponds to a random walk through parameter space, which converges relatively slowly to the target density, and which samples from the posterior inefficiently due to slow mixing of the chain, with the optimal acceptance rate (proportion of accepted proposals) being only 0.234 [67]. Therefore, we make use of the MALA as a superior alternative, which converges much more efficiently, requiring only $O(d^{1/3})$ steps, where d is the dimension of the target density, whereas the random walk requires $O(d)$ steps, while the higher optimal acceptance rate of 0.574 allows for faster mixing and reduced dependence between samples [65]. The Markov chain is treated as an itô diffusion and behaves according to Langevin dynamics with stochastic differential equation

$$d\theta_t = \nabla \log \pi(\theta_t) + \sqrt{2} dW_t \tag{22}$$

evolving θ in imaginary time with a standard Brownian motion diffusion term, W, and a drift term determined by the vector gradient, ∇, of the logarithm of the posterior density, $\pi(\theta) \propto P(X|\theta)P(\theta)$, with respect to θ evaluated at θ_t. However, we do not have an analytical solution for $\nabla \log \pi(\theta)$ which means we must estimate this numerically using the change in likelihood observed between the current step, j, and the previous one when generating a proposal. This leads to an additional complication, wherein we may not propose a new sample for all parameters simultaneously since then the observed change in likelihood would be the combined effect of the change in each parameter, making the individual gradients impossible to estimate. Therefore, we must sequentially update each parameter conditional on the current value of all other parameters, which are treated as fixed constants. This corresponds to embedding our MALA within a Gibbs sampler [68, 69], meaning that d sub-steps are required to move from step j to $j+1$. At step j, we cycle through each parameter, k, from 1 to d, and draw a new proposal for parameter k from a proposal distribution as determined by Eq. 22

$$\theta_k^{(*)} = \theta_k^{(j)} + S_k \nabla_k \log \pi(\theta) + \sqrt{2S_k}\xi$$

where ξ is a standard normal random variable and S is an adaptive scaling constant such that the proposal is drawn from

$$\theta_k^{(*)} \sim \mathcal{N}\left(\theta_k^{(j)} + S_k \nabla_k \log \pi(\theta), \sqrt{2S_k}\right)$$

This is accepted with a probability given by the likelihood ratio at the proposed and current value

$$A = \min\left(1, \frac{\pi(\theta_1^{(j+1)}, \ldots, \theta_{k-1}^{(j+1)}, \theta_k^{(*)}, \theta_{k+1}^{(j)}, \ldots, \theta_d^{(j)})}{\pi(\theta_1^{(j+1)}, \ldots, \theta_{k-1}^{(j+1)}, \theta_k^{(j)}, \ldots, \theta_d^{(j)})}\right) \tag{23}$$

where substituting $\pi(\theta)$ for $P(X|\theta)P(\theta)$ gives an equivalent ratio due to the proportionality, which allows us to refer directly to the target density, π. Note that the intractable integrals in the denominators of Eqs. 5, 17 and 21 cancel out to allow the acceptance probability to be calculated with only the likelihood function and the prior density. In our special case with uniform priors, these also cancel, only serving to reject proposals outside of the plausible ranges of parameter space as defined by the prior. With probability A we set $\theta_k^{(j+1)} = \theta_k^{(*)}$, otherwise $\theta_k^{(j+1)} = \theta_k^{(j)}$ and since we treat parameters other than θ_k as constants, we iteratively draw θ from the conditional rather than joint densities as

$$\theta_k^{(j+1)} \sim P(\theta_k^{(j+1)} | \theta_1^{(j+1)}, \ldots, \theta_{k-1}^{(j+1)}, \theta_{k+1}^{(j)}, \ldots, \theta_d^{(j)})$$

If the proposal is accepted, we update our estimate of the local gradient for the parameter k as

$$\nabla_k = \frac{\log \pi(\theta_k^{(j+1)}) - \log \pi(\theta_k^{(j)})}{\theta_k^{(j+1)} - \theta_k^{(j)}}$$

otherwise we set $\nabla_k = 0$. We also recursively update the adaptive scaling constant associated with parameter k in the manner described for the Adaptive Scaling Metropolis algorithm of [67]

$$S_k = e^{(\log(S_k) + \eta(A - 0.574))}$$

with a recursively updated decay term

$$\eta = 0.999\eta$$

which in the long-term results in the MALA mixing close to the optimal parameter-specific acceptance rate of 0.574 [65]. At step 1, we initialise $\nabla = 0$, $S = \theta^{(1)}/100$ and $\eta = 0.1$. Density plots indicate that in the vast majority of cases acceptance rates close to 0.574 were achieved (Additional file 1: Fig. S5).

The process repeats until 5000 steps have been completed ($j = 5000$) if $\hat{\mu} < 1000$ or 1500 if $\hat{\mu} \geq 1000$, since for these genes with very high expression levels each step takes longer but the stronger evidence means that fewer steps are required. Therefore, the Markov chain converges to the posterior distribution according to its gradient. Posteriors were produced from the sampled chain using the last 1000 or 2500 steps for high

expression or other genes, respectively, with a thinning factor of 2, where only every 2nd point in the chain is used in order to reduce dependency between points, resulting in smoother posterior densities and sample sizes of 500 or 1250. When using model 2, for each step, we check if the proposal for any sub-step was rejected because of negative probability values appearing in Eq. 9 due to the approximate non-equilibrium solution failing for an unstable point in parameter space. We set a rolling window size, w, equal to 100 or 500 for high expression or other genes, respectively. We then check at each step, j, if the number of steps with a rejection of this nature is $\geq w/20$ for steps $[\max((w/2)+1, j-w+1), j]$ and if this condition is met then the Markov chain is restarted using model 3 instead of model 2.

Simulations for model comparison

The performance of inference using different likelihood functions was tested on simulated data. Gillespie's exact algorithm (stochastic simulation algorithm) [70] was used to simulate data according to the reactant matrix shown in Table 3 and the product matrix shown in Table 4, with the stoichiometry matrix shown in Table 5.

Table 3 Reactant matrix for new and surviving transcript count Gillespie algorithm simulations

	RNA_0	RNA_1	G_{on}	G_{off}
β_0	0	0	1	0
β_1	0	0	1	0
δ_0	1	0	0	0
δ_1	0	1	0	0
k_{on}	0	0	0	1
k_{off}	0	0	1	0

Table 4 Product matrix for new and surviving transcript count Gillespie algorithm simulations

	RNA_0	RNA_1	G_{on}	G_{off}
β_0	1	0	1	0
β_1	0	1	1	0
δ_0	0	0	0	0
δ_1	0	0	0	0
k_{on}	0	0	1	0
k_{off}	0	0	0	1

Table 5 Stoichiometry matrix for new and surviving transcript count Gillespie algorithm simulations

	RNA_0	RNA_1	G_{on}	G_{off}
β_0	1	0	0	0
β_1	0	1	0	0
δ_0	-1	0	0	0
δ_1	0	-1	0	0
k_{on}	0	0	1	-1
k_{off}	0	0	-1	1

This allows simulation of the pool of transcripts in the cell that was synthesised before (RNA_0) and after (RNA_1) the 4sU pulse started. The simulation is run with initial conditions $X_0 = (0, 0, 0, 1)$ where $X = (RNA_0, RNA_1, G_{on}, G_{off})$ and rate constant values are set for bursty expression $\theta = (\beta_0 = 50, \beta_1 = 0, \delta_0 = 0.001, \delta_1 = 0.001, k_{on} = 0.0005, k_{off} = 1)$, running until $t_0 = 200000$ to bring the system to steady state. The system state at the end of this run, X_{t_0}, is then used as the initial condition for a second run, where we now set $\theta = (\beta_0 = 0, \beta_1 = 50, \delta_0 = 0.001, \delta_1 = 0.001, k_{on} = 0.0005, k_{off} = 1)$ to simulate the newly synthesised transcripts produced during the 4sU pulse along with decay of pre-existing transcripts. A pulse duration of $t_1 = 1000$ min was used here, giving the final state of the system X_{t_1}, and importantly giving the counts for RNA_0 and RNA_1 in the cell. This was repeated to simulate $N = 10000$ cells. In-silico sequencing data was then generated based on these simulated transcript count values. Cell-specific capture efficiencies were drawn

$$\alpha \sim Beta(1, 9)$$

before drawing the cell-specific UMI counts, l, corresponding to the two pools of transcripts as

$$l_k \sim Bin(RNA_k, \alpha)$$

for $k = 0$ and $k = 1$, so that the total UMI count for the given cell is $l = l_0 + l_1$. The cell-specific total number of reads corresponding to each UMI in the two pools is then drawn

$$r_{k,j} \sim ZTPois(\nu)$$

where $\nu = 5$ represents sequencing depth and reads per UMI is a zero-truncated Poisson random variable with

$$f_{ZTPois}(r|\nu, r > 0) = \frac{\nu^r}{(e^\nu - 1)r!}$$

using the same logic of Poisson assignment of reads to UMIs as in [71]. Then, the cell-specific total number of reads of the given pool is

$$r_k = \sum_{j=1}^{l_k} r_{k,j}$$

The number of uracils across the sequenced part of the transcript is then drawn for each read

$$u_{k,j} \sim Pois(\hat{u})$$

where $\hat{u} = 60$ is the average number of uracils per read. The number of conversions in each read in the cell is then drawn for the two pools of transcripts as

$$i_{0,j} \sim Bin(u_{0,j}, \lambda_s)$$

and

$$i_{1,j} \sim Bin(u_{1,j}, \lambda_s + \lambda_n)$$

where we set $\lambda_s = 0.01$ and $\lambda_n = 0.075$. The overall conversion data across all reads in the cell is then $i = (i_0, i_1)$, where $i_0 = (i_{0,1}, \ldots, i_{0,r_0})$ and $i_1 = (i_{1,1}, \ldots, i_{1,r_1})$, from which we obtain y, where y_i is the number of reads with i conversions in the given cell. Now we have our simulated dataset which we can use to demonstrate our capacity to recover known parameter values. MCMC was carried out with different likelihood functions in the previously described manner to sample posterior distributions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02977-y.

Additional file 1. Supplementary information: Contains methodological information on estimating capture efficiencies and 4sU-mediated TC rates and for simulating data for validating algorithm performance. Also contains results of metagene and correlation analyses for various HMs not shown in the main text, as well as a correlation between our estimated transcript decay rates and previously published cell-matched decay rates.

Additional file 2. High vs low noise cell-specific T>C rate distribution transition: Video gif showing the differential transition from surviving to new transcript pool for high and low noise genes through the cell-specific T>C rate distributions for data simulated with different pulse durations.

Additional file 3. Review history.

Acknowledgements
We thank Louise Dyson for useful discussions regarding the mathematical theory relevant to the study and Francesca Mantellino for valuable input on the data processing strategy.

Review history
The review history is available as Additional File 3.

Peer review information
Veronique van den Berghe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
DME, PD and DH conceived the study, while DME and DH designed the work. DME and PD acquired the data which DME analysed and DME, PD and DH interpreted. DME and PD created the software. DME wrote the manuscript with input from PD and DH. All authors approve the submission of the manuscript.

Funding
The research was funded by the Biotechnology and Biological Sciences Research Council (BB/M01116X/1) and by the Engineering Physical Sciences Research Council (EP/T002794/1).

Availability of data and materials
The inference algorithm has been made available as a GitHub-distributed R package (https://github.com/hebenstreitLab/burstMCMC) [72]. The pre-processing scripts are similarly available in a GitHub repository (https://github.com/hebenstreitLab/burstMCMCpreprocessing) [73]. Both repositories are licensed under the MIT license. The scripts along with the processed data used in the analysis are available in Zenodo (https://doi.org/10.5281/zenodo.7707970) under the MIT license [74]. The raw 4sU scRNA-seq data used can be found under the GEO sample IDs GSM4512696 and GSM4512697 [38, 75]. The processed scRNA-seq dataset used for calculating capture efficiencies can be found under the GEO sample ID GSM1599501 [76, 77]. The processed ChIP-seq datasets used for metagene analysis can be found under GEO sample IDs GSM2895356, GSM733651, GSM733653, GSM733656, GSM733675, GSM733692 GSM733714, GSM733778, GSM2862934 and GSM4809274 [53–56, 78–81].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 6 September 2022 Accepted: 25 May 2023
Published online: 16 June 2023

References

1. Wissink EM, Vihervaara A, Tippens ND, Lis JT. Nascent RNA analyses: tracking transcription and its regulation. Nat Rev Genet. 2019;20(12):705–23.
2. Chen FX, Smith ER, Shilatifard A. Born to run: control of transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol. 2018;19(7):464–78.
3. Kuehner JN, Pearson EL, Moore C. Unravelling the means to an end: RNA polymerase II transcription termination. Nat Rev Mol Cell Biol. 2011;12(5):283–94.
4. Proudfoot NJ. Transcriptional termination in mammals: stopping the RNA polymerase II juggernaut. Science. 2016;352(6291):aad9926.
5. Paulsson J. Models of stochastic gene expression. Phys Life Rev. 2005;2(2):157–75.
6. Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci. 2002;99(20):12795–800.
7. Lin YT, Galla T. Bursting noise in gene expression dynamics: linking microscopic and mesoscopic models. J R Soc Interface. 2016;13(114):20150772.
8. Fukaya T, Lim B, Levine M. Enhancer control of transcriptional bursting. Cell. 2016;166(2):358–68.
9. Dar RD, Razooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, et al. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc Natl Acad Sci. 2012;109(43):17454–9.
10. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. 2013;14(1):R7.
11. Losick R, Desplan C. Stochasticity and cell fate. Science. 2008;320(5872):65–8.
12. Dieci G, Bosio MC, Fermi B, Ferrari R. Transcription reinitiation by RNA polymerase III. Biochim Biophys Acta (BBA)-Gene Regul Mech. 2013;1829(3-4):331–41.
13. Shandilya J, Roberts SG. The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. Biochim Biophys Acta (BBA)-Gene Regul Mech. 2012;1819(5):391–400.
14. Cavallaro M, Walsh MD, Jones M, Teahan J, Tiberi S, Finkenstädt B, et al. 3′-5′ crosstalk contributes to transcriptional bursting. Genome Biol. 2021;22(1):1–20.
15. Dey SS, Foley JE, Limsirichai P, Schaffer DV, Arkin AP. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. Mol Syst Biol. 2015;11(5):806.
16. Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet. 2020;21(2):71–87.
17. Sun M, Zhang J. Allele-specific single-cell RNA sequencing reveals different architectures of intrinsic and extrinsic gene expression noises. Nucleic Acids Res. 2020;48(2):533–47.
18. Benayoun BA, Pollina EA, Ucar D, Mahmoudi S, Karra K, Wong ED, et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell. 2014;158(3):673–88.
19. Nicolas D, Zoller B, Suter DM, Naef F. Modulation of transcriptional burst frequency by histone acetylation. Proc Natl Acad Sci. 2018;115(27):7153–8.
20. Wu S, Li K, Li Y, Zhao T, Li T, Yang YF, et al. Independent regulation of gene expression level and noise by histone modifications. PLoS Comput Biol. 2017;13(6):e1005585.
21. Chong S, Chen C, Ge H, Xie XS. Mechanism of transcriptional bursting in bacteria. Cell. 2014;158(2):314–26.
22. Ma L, Gao Z, Wu J, Zhong B, Xie Y, Huang W, et al. Co-condensation between transcription factor and coactivator p300 modulates transcriptional bursting kinetics. Mol Cell. 2021;81(8):1682–97.
23. Engl C, Jovanovic G, Brackston RD, Kotta-Loizou I, Buck M. The route to transcription initiation determines the mode of transcriptional bursting in E. coli. Nat Commun. 2020;11(1):1–11.
24. Dobrinić P, Szczurek AT, Klose RJ. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. Nat Struct Mol Biol. 2021;28(10):1–14.
25. Popp AP, Hettich J, Gebhardt JCM. Altering transcription factor binding reveals comprehensive transcriptional kinetics of a basic gene. Nucleic Acids Res. 2021;49(11):6249–66.
26. Maynard KR, Tippani M, Takahashi Y, Phan BN, Hyde TM, Jaffe AE, et al. dotdotdot: an automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smFISH) images in complex tissues. Nucleic Acids Res. 2020;48(11):e66.
27. Li G, Neuert G. Multiplex RNA single molecule FISH of inducible mRNAs in single yeast cells. Sci Data. 2019;6(1):1–9.
28. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc Natl Acad Sci. 2016;113(39):11046–51.
29. Shah S, Takei Y, Zhou W, Lubeck E, Yun J, Eng CHL, et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. Cell. 2018;174(2):363–76.
30. Gorin G, Wang M, Golding I, Xu H. Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics. PLoS ONE. 2020;15(3):e0230736.
31. Tunnacliffe E, Corrigan AM, Chubb JR. Promoter-mediated diversification of transcriptional bursting dynamics following gene duplication. Proc Natl Acad Sci. 2018;115(33):8364–9.
32. Larsson AJ, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR, Reinius B, et al. Genomic encoding of transcriptional burst kinetics. Nature. 2019;565(7738):251–4.
33. Ochiai H, Hayashi T, Umeda M, Yoshimura M, Harada A, Shimizu Y, et al. Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells. Sci Adv. 2020;6(25):eaaz6699.
34. Johnsson P, Ziegenhain C, Hartmanis L, Hendriks GJ, Hagemann-Jensen M, Reinius B, et al. Transcriptional kinetics and molecular functions of long noncoding RNAs. Nat Genet. 2022;54(3):1–12.
35. Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, et al. Thiol-linked alkylation of RNA to assess expression dynamics. Nat Methods. 2017;14(12):1198.
36. Schofield JA, Duffy EE, Kiefer L, Sullivan MC, Simon MD. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. Nat Methods. 2018;15(3):221.
37. Jürges C, Dölken L, Erhard F. Dissecting newly transcribed and old RNA using GRAND-SLAM. Bioinformatics. 2018;34(13):i218–26.

38. Qiu Q, Hu P, Qiu X, Govek KW, Cámara PG, Wu H. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. Nat Methods. 2020;17(10):991–1001.
39. Cao J, Zhou W, Steemers F, Trapnell C, Shendure J. Sci-fate characterizes the dynamics of gene expression in single cells. Nat Biotechnol. 2020;38(8):1–9.
40. Battich N, Beumer J, de Barbanson B, Krenning L, Baron CS, Tanenbaum ME, et al. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. Science. 2020;367(6482):1151–6.
41. Hendriks GJ, Jung LA, Larsson AJ, Lidschreiber M, Forsman OA, Lidschreiber K, et al. NASC-seq monitors RNA synthesis in single cells. Nat Commun. 2019;10(1):1–9.
42. Qiu X, Zhang Y, Martin-Rufino JD, Weng C, Hosseinzadeh S, Yang D, et al. Mapping transcriptomic vector fields of single cells. Cell. 2022;185(4):690–711.
43. Erhard F, Baptista MA, Krammer T, Hennig T, Lange M, Arampatzi P, et al. scSLAM-seq reveals core features of transcription dynamics in single cells. Nature. 2019;571(7765):419–23.
44. Boileau E, Altmüller J, Naarmann-de Vries IS, Dieterich C. A comparison of metabolic labeling and statistical methods to infer genome-wide dynamics of RNA turnover. Brief Bioinform. 2021;22(6):bbab219.
45. Zabet NR, Chu DF. Computational limits to binary genes. J R Soc Interface. 2010;7(47):945–54.
46. Patange S, Ball DA, Wan Y, Karpova TS, Girvan M, Levens D, et al. MYC amplifies gene expression through global changes in transcription factor dynamics. Cell Rep. 2022;38(4):110292.
47. Lu D, Jambhekar A, Lahav G. Louder for longer: MYC amplifies gene expression by extended transcriptional bursting. Cell Rep. 2022;38(9):110470.
48. Jimeno-González S, Reyes JC. Chromatin structure and pre-mRNA processing work together. Transcription. 2016;7(3):63–8.
49. Rahhal R, Seto E. Emerging roles of histone modifications and HDACs in RNA splicing. Nucleic Acids Res. 2019;47(10):4911–26.
50. Wolfe JC, Mikheeva LA, Hagras H, Zabet NR. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila. Genome Biol. 2021;22(1):1–23.
51. Arias AM, Brickman JM. Gene expression heterogeneities in embryonic stem cell populations: origin and function. Curr Opin Cell Biol. 2011;23(6):650–6.
52. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
53. Mchaourab ZF, Perreault AA, Venters BJ. ChIP-seq and ChIP-exo profiling of Pol II, H2A. Z, and H3K4me3 in human K562 cells. Sci Data. 2018;5(1):1–8.
54. Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.
55. Vazquez BN, Thackray JK, Simonet NG, Chahar S, Kane-Goldsmith N, Newkirk SJ, et al. SIRT7 mediates L1 elements transcriptional repression and their association with the nuclear lamina. Nucleic Acids Res. 2019;47(15):7870–85.
56. Radzisheuskaya A, Shliaha PV, Grinev VV, Shlyueva D, Damhofer H, Koche R, et al. Complex-dependent histone acetyltransferase activity of KAT8 determines its role in transcription and cellular homeostasis. Mol Cell. 2021;81(8):1749–65.
57. De S, Edwards DM, Dwivedi V, Wang J, Varsally W, Dixon HL, et al. Genome-wide chromosomal association of Upf1 is linked to Pol II transcription in Schizosaccharomyces pombe. Nucleic Acids Res. 2022;50(1):350–67.
58. Schoech AP, Zabet NR. Facilitated diffusion buffers noise in gene expression. Phys Rev E. 2014;90(3):032701.
59. Shahrezaei V, Swain PS. Analytical distributions for stochastic gene expression. Proc Natl Acad Sci. 2008;105(45):17256–61.
60. Morrison M, Razo-Mejia M, Phillips R. Reconciling kinetic and thermodynamic models of bacterial transcription. PLoS Comput Biol. 2021;17(1):e1008572.
61. Wang J, Huang M, Torre E, Dueck H, Shaffer S, Murray J, et al. Gene expression distribution deconvolution in single-cell RNA sequencing. Proc Natl Acad Sci. 2018;115(28):E6437–46.
62. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006;4(10):e309.
63. Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation. J Chem Phys. 2006;124(4):044104.
64. Gupta A, Mikelson J, Khammash M. A finite state projection algorithm for the stationary solution of the chemical master equation. J Chem Phys. 2017;147(15):154101.
65. Roberts GO, Rosenthal JS. Optimal scaling of discrete approximations to Langevin diffusions. J R Stat Soc B Stat Methodol. 1998;60(1):255–68.
66. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97-109.
67. Vihola M. On the stability and ergodicity of adaptive scaling Metropolis algorithms. Stoch Process Appl. 2011;121(12):2839–60.
68. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell. 1984;6:721–41.
69. Gilks WR, Best NG, Tan KK. Adaptive rejection Metropolis sampling within Gibbs sampling. J R Stat Soc Ser C Appl Stat. 1995;44(4):455–72.
70. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81(25):2340–61.
71. Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, et al. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proc Natl Acad Sci. 2014;111(5):1891–6.
72. Edwards DM, Davies P, Hebenstreit D. burstMCMC. GitHub. 2023. https://github.com/hebenstreitLab/burstMCMC. Accessed 30 Mar 2023.

73. Edwards DM, Davies P, Hebenstreit D. burstMCMCpreprocessing. GitHub. 2023. https://github.com/hebenstreitLab/burstMCMCpreprocessing. Accessed 30 Mar 2023.
74. Edwards DM, Davies P, Hebenstreit D. Synergising single-cell resolution and 4sU labelling boosts inference of transcriptional bursting. Zenodo. 2023. https://doi.org/10.5281/zenodo.7707970. Accessed 30 Mar 2023.
75. Qiu Q, Hu P, Qiu X, Govek KW, Cámara PG, Wu H. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. Gene Expression Omnibus. 2020. https://identifiers.org/geo:GSE141851. Accessed 13 Jan 2021.
76. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.
77. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Gene Expression Omnibus. 2015. https://identifiers.org/geo:GSE65525. Accessed 25 Oct 2020.
78. Mchaourab ZF, Perreault AA, Venters BJ. ChIP-seq and ChIP-exo profiling of Pol II, H2A. Z, and H3K4me3 in human K562 cells. Gene Expression Omnibus. 2018. https://identifiers.org/geo:GSE108323 (2017). Accessed 16 Mar 2022.
79. Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Gene Expression Omnibus. 2012. https://identifiers.org/geo:GSE29611 (2011). Accessed 14 Mar 2022.
80. Vazquez BN, Thackray JK, Simonet NG, Chahar S, Kane-Goldsmith N, Newkirk SJ, et al. SIRT7 mediates L1 elements transcriptional repression and their association with the nuclear lamina. Gene Expression Omnibus. 2019. https://identifiers.org/geo:GSE106964. Accessed 3 Mar 2023.
81. Radzisheuskaya A, Shliaha PV, Grinev VV, Shlyueva D, Damhofer H, Koche R, et al. Complex-dependent histone acetyltransferase activity of KAT8 determines its role in transcription and cellular homeostasis. Gene Expression Omnibus. 2021. https://identifiers.org/geo:GSE158736. Accessed 3 Mar 2023.

Publisher's Note