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

by

David Michael Edwards

A thesis submitted for the degree of

Doctor of Philosophy

University of Warwick
School of Life Sciences

December 2022

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr Daniel Hebenstreit for his fantastic guidance and support throughout my PhD. I couldn't ask for a better supervisor, providing the perfect mix between quickly giving smart advice and ways to improve my work while allowing me the freedom to direct my work and be an independent researcher. I also thank my secondary supervisor, Dr Louise Dyson, for her much appreciated advice regarding the mathematical theory of my research. Next, Dr Philip Davies deserves huge credit, both for directly contributing towards the project and for being a brilliant person to work with in general where we could always bounce ideas back and forth. Specific thanks also goes to Francesca Mantellino for helpful input on appropriate data processing strategies. Thanks also to the BBSRC for funding my PhD through a MIBTP studentship.

I must also thank all of my group members, both past and present, as well as my friends for making my PhD such an enjoyable time.

Finally, I thank my parents for their huge and unending support throughout my PhD and life in general.

Declaration

I hereby declare that this thesis is an original work and has not been submitted for any other degree.

I carried out the entirety of the work presented other than the code for the simulation protocol outlined in section 2.7, which Dr Philip Davies also directly contributed towards.

A pre-print of part of the work has been published on bioRxiv:

David Michael Edwards, Philip Davies, and Daniel Hebenstreit. Synergising single-cell resolution and 4su labelling boosts inference of transcriptional bursting. bioRxiv, 2022.

Abbreviations

single-cell (sc), 4-thiouridine (4sU), RNA polymerase (RNAP), transcription factor (TF), transcription end site (TES), transcription start site (TSS), histone modification (HM), single molecule fluorescence in situ hybridisation (smFISH), Markov chain Monte Carlo (MCMC), Metropolis-adjusted Langevin algorithm (MALA), coefficient of variation (CV), unique molecular identifier (UMI), gene body (GB), partial differential equation (PDE), ordinary differential equation (ODE)

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. Presented here is a mathematical model and Bayesian inference implementation to facilitate joint estimation and confidence quantification of the parameters governing transcriptional bursting dynamics on a genome-wide scale. It is demonstrated 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. Accounting for various sources of both biological and technical noise, the observed cell-specific transcript turnovers and abundances are naturally integrated, thus reducing the error across all parameters of interest; both dimensionless and temporal. Applying the method to published 4sU scRNA-seq data indicated that large bursts are required for genes with very high expression levels, such as mitochondrial genes. Linking with published ChIP-seq data uncovered otherwise obscured associations between different parameters and histone modifications, agreeing with but advancing upon previously reported results. Evidence is provided for a link between histone modifications and modulation of bursting dynamics through, for example, effects on transcript stability, with these effects being dependent upon the location of the modification throughout the gene. Algorithm performance was validated using simulated datasets with ground truth target parameter values, both with a detailed analysis at a single-gene scale and with a high level analysis at a genome-wide scale.

1 Introduction

1.1 Transcription overview

Eukaryotic transcription is the process by which RNA polymerase (RNAP) enzymes generate RNA "transcripts" by processing along the DNA sequence of a gene. RNA nucleotides complementary to the template DNA sequence that the RNAP is passing over are added to a nascent transcript [1]. This process is known as elongation and is preceded and succeeded by initiation and termination, respectively. Initiation involves the recruitment of RNAP to the transcription start site (TSS) via the pre-initiation complex (PIC), which is made up of various transcription factors (TFs), including those which bind to regulatory DNA sequence motifs at the promoter [2]. The DNA double helix is then unwound to expose the template strand before the RNAP is loaded and the first few bases are synthesised, providing the DNA-RNA hybrid needed for elongation to proceed. Termination involves the release of both the nascent transcript and RNAP from the DNA followed by RNAP disassembly, usually upon reaching the transcription end site (TES) [3]. Various processing steps take place at different points during transcription to allow for a mature transcript to be produced. These include, in the case of protein-coding transcripts, 5' capping during initiation, splicing to remove intronic (non-coding) sequences during elongation, and polyadenylation and cleavage during termination [2, 1, 3, 4]. Mature messenger RNA (mRNA) is shuttled to the cytoplasm to undergo translation. Mature transcripts of RNA genes, on the other hand, may remain in the nucleus or be transported to the cytoplasm to fulfil their function without being translated. In eukaryotes, ribosomal RNA (rRNA) other than 5S rRNA is transcribed by RNAP I [5], the mRNA of protein-coding genes, along with various non-coding RNAs, are transcribed by RNAP II [2] and 5S rRNA, transport RNA (tRNA) and other small RNAs are transcribed by RNAP III [6].

1.2 Chromatin and nucleosomes

Since the process of transcription depends on recruitment of various proteins and factors to the DNA sequence which is being transcribed, it depends crucially upon the accessibility of the DNA to these factors. In eukaryotes, the structure of chromatin is governed primarily by the packaging of the DNA by nucleosomes [7]. The DNA is wound around nucleosomes, which are complexes composed of several protein subunits called histones, with the level of packaging of chromatin across different regions of the genome broadly classifying them into either heterochromatin or euchromatin if the DNA is more tightly or loosely packaged, respectively (figure 1). Each nucleosome is a histone octamer, with each of the core histone proteins (H2A, H2B, H3 and H4) being found with two copies, comprising a H3-H4 tetramer and two H2A-H2B dimers [8]. Each histone protein is made up of a globular domain, which resides in the nucleosome core and is able to non-specifically bind to DNA through electrostatic interactions, and an unstructured N-terminal tail domain which extends out from the nucleosome and is subject to various covalent post-translational modifications at many residues. These are known as histone marks or histone modifications (HMs) and modulate the interaction strength of the modified histone with the DNA [9, 10]. HMs can drastically alter the chromatin landscape and therefore, the transcriptional dynamics of proximal genes via the accessibility of the DNA for transcriptionally-relevant proteins and factors. If the DNA is too tightly packaged then transcription is prevented since the TFs required to recruit RNAP and the PIC in general are not able to interact with their DNA binding sites.

So-called pioneer TFs are able to open silent regions of chromatin at which active histone marks are not present and the DNA is too tightly wound around nucleosomes for other TFs to access

[11, 12, 13]. Unlike other TFs and transcriptionally-relevant DNA-binding proteins, pioneer TFs are capable of recognising, targeting and binding partially accessible nucleosomal DNA sequence motifs found in regions of otherwise inaccessible, non-active chromatin, although this is not possible at strongly repressed heterochromatin with repressive HMs, which is too tightly packaged for even pioneer TFs to access. One example is FOXA, which has been shown to displace linker histone proteins from the nucleosome it binds to [14], thereby directly disrupting the nucleosome structure and making the chromatin accessible to other TFs, histone modifiers and nucleosome remodelling complexes. Pioneer TFs may then recruit other TFs and even RNAP to the regulatory site. This mechanism can lead to the transformation of silent, unmarked chromatin into either active (enhancers and promoters) or repressive (silencers) regulatory sites by inducing euchromatin formation, with the outcome depending upon the site being targeted as well as the other factors, proteins and complexes available. Silent regions may also become repressed regions through the action of pioneer TFs by making chromatin accessible for the deposition of repressive HMs, which results in tighter packaging of nucleosomes and subsequent heterochromatin formation at the site. In this manner, pioneer TFs cause changes to gene regulatory networks which shift the transcriptomic landscape of the cell and ultimately result in, for example, cell fate determination during development [12, 13]. Mis-regulation of pioneer TFs, particularly their up-regulation, is also associated with cancer, such as SOX2 up-regulation in skin cancer, with SOX2 deletion being found to strongly reduce tumour formation [15].

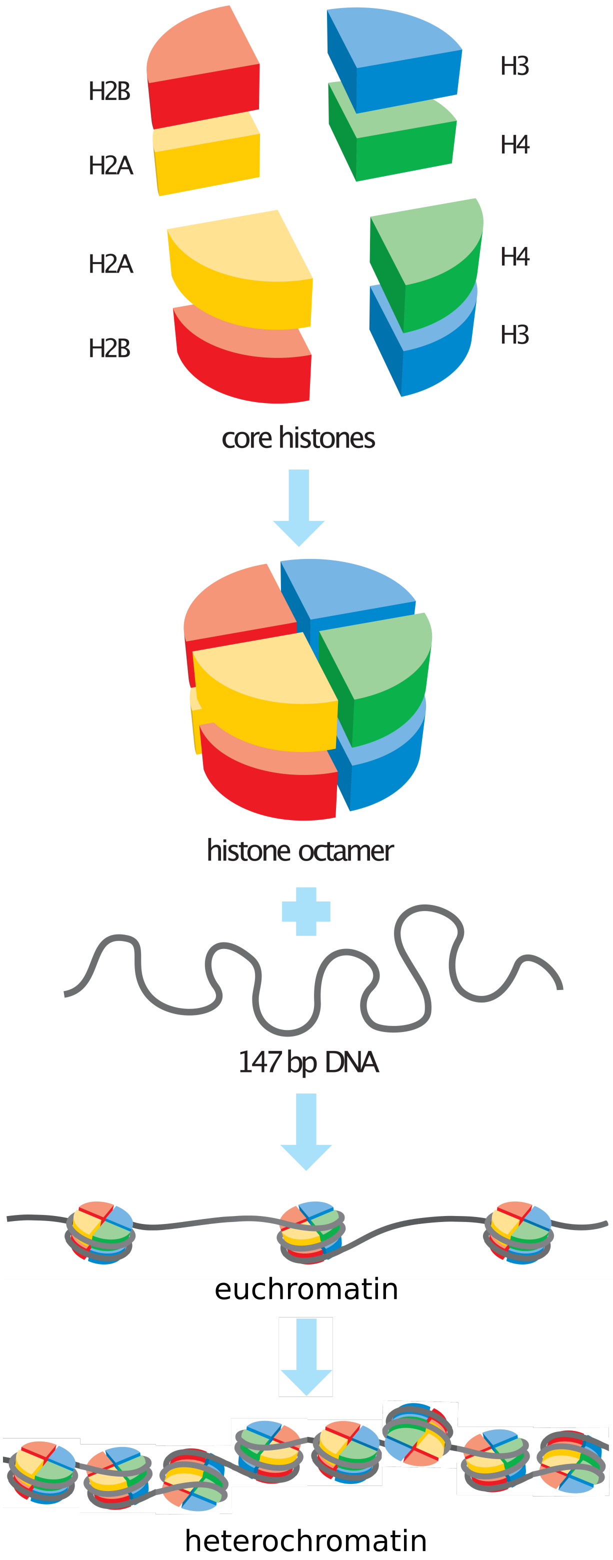


Figure 1: Schematic adapted from [16] illustrating the nucleosome-mediated assembly of chromatin by winding ∼ 147bp of DNA around each histone octamer, resulting in either loosely packaged euchromatin or tightly packaged heterochromatin.

1.3 Histone modifications

Two of the most important types of HM for transcriptional regulation are acetylation and methylation, although many others exist including phosphorylation, ubiquitylation and sumoylation [17]. Acetylation and deacetylation of histone tail residues is carried out by histone acetyl-transferases (HATs) and histone deacetylases (HDACs), respectively, while methylation and demethylation of residues relies on histone methyl-transferases (HMTs) and histone demethylases (KDMs, with the K referring to lysine, the methylated/demethylated residue although arginines may also be methylated). During acetylation, a single acetyl group will be transferred to the lysine residue, while a lysine may have one to three methyl groups as a result of methylation. Histone acetylation weakens the histone-DNA interaction due to the removal of positive histone charge with which to attract the negative DNA phosphate groups, which means that histone acetylation is generally associated with looser packaging of the DNA, greater accessibility and therefore active transcription [18]. The effects of histone methylation on DNA interaction strength vary based on the position of the residue being methylated and the number of methyl groups which are transferred, meaning that histone methylation may be linked with either active or repressed transcription. Regarding the issue of causality, there are actually suggestions that in some cases HMs simply reflect and are a result of changes in the chromatin rather than being the driving force. For example, HAT complexes may not be able to penetrate tightly closed chromatin without some initial remodelling to open it [19], although there is also clear evidence of acetylation driving chromatin remodelling based on the order of events, with acetylation occurring before any remodelling [20].

Out of the histone methylation marks associated with active transcription, some such as H3K4me3 are primarily localised to the TSS, while others like H3K36me3 and H3K79me2 are primarily localised to the gene body (GB) [21]. Others, such as H3K4me1, have been strongly linked to transcription via active enhancer in multiple studies [22, 23, 24, 25]. The active histone acetylation mark H3K27ac is often found at active enhancers and TSSs and is associated with enhancer-mediated gene activation [26]. However, an important study in mESCs in which the lysine at the 27th residue of the H3 protein was mutated to arginine, thereby strongly diminishing H3K27 acetylation, recently found only a modest effect on the binding of RNAP II and MED1 (a transcriptional coactivator) in mutant cells [27]. Indeed there were no major changes to core transcriptional regulator binding to chromatin or to transcription itself, indicating that H3K27ac is not essential for mESC transcription. Instead, it is hypothesised that acetylation of alternative lysine residues, for example, may compensate for its at H3K27 [27], suggesting that although H3K27ac is clearly an important active enhancer mark, it may not be crucial for transcription in many cases.

Further studies have characterised other marks and their relationships with transcription, many of which mark enhancers in addition to those found at gene promoters/GBs. For example, an AI-based approach was used in conjunction with ChIP-seq (and other) datasets to predict enhancers in different cell lines and found H3K18ac, H4K16ac and H3K79me3 to all be strongly enriched at enhancers, which were generally marked by high levels of H3K4me1 along with high levels of either H3K18ac or H4K16ac [22]. H3K4me2, H3K4me3, H3K79me2, and H3K9ac appeared to be associated with housekeeping enhancers whereas H3K4me1, H4K16ac and H3K27ac more so predicted developmental enhancers. One study investigated the Notch signalling pathway, which is very important for development in Drosophila, and found that Notch activation resulted in a rapid increase in H3K56 acetylation at Notch-regulated enhancers, causing elevated transcription levels, with this effect being conserved since it was observed both in Drosophila and mammalian systems [23]. H3K56ac directly alters chromatin accessibility by increasing the rate with which DNA unwraps from nucleosomes, thus increasing TF occupancy. The increase in H3K56ac occurs before transcription elongation, reinforcing the idea that it facilitates initiation by inducing euchromatin formation. Deposition of H3K56ac occurs primarily at enhancer loci already marked by H3K4me1, indicating that H3K4me1 is a pre-requisite for H3K56ac-mediated gene activation and may even facilitate the spread of H3K56ac, with H3K4me1 mutation resulting in reduced H3K56ac levels [28].

Another study on H3 acetylation of the globular domain (H3K64ac and H3K122ac) in mESCs in comparison to the more studied tail domain acetylation HMs, such as H3K27ac, showed that while H3K27ac has traditionally been used to identify enhancers, there is a class of enhancers marked by H3K122ac but lacking H3K27ac, in addition to the more well-known class of enhancers (marked by H3K64ac, H3K122ac and H3K27ac) [24]. Promoters were shown to be marked by both H3K64ac and H3K122ac, with further data from K562 cells indicating that H3K122ac was enriched at both active and poised promoters as well as strong enhancers. Globular acetylation HMs such as H3K64ac and H3K122ac exert an affect by perturbing the inter-nucleosomal interactions to alter nucleosome stability and turnover and also by directly facilitating activator binding [24]. Previous work on mESCs also investigated the lesser studied HMs found on H4, as opposed to H3, showing that loss of acetylation at H4K16 did not cause compaction of higher-order chromatin, although it likely involved smaller-scale chromatin remodelling since H4K16 facilitates neighbouring nucleosome interactions, with acetylation likely to disrupt this [25]. Again, this study showed that a subset of active enhancers are marked by H3K4me1 and H4K16ac but not H3K27ac, with expressed genes also having H4K16 around their TSS. Whereas H3K4me3, which is the the classic TSS-associated HM, occurs much more strongly on the +1 nucleosome, the presence of H4K16ac is equally strong on the +1 and -1 nucleosomes (which flank the TSS). Although it is not found around the TES, H4K16ac is also present in the GB, pointing to a potential role in elongation [25].

1.4 Initiation and pausing

Once heterochromatin is cleared from the promoter region of a gene by the modification of histones weakening interactions between DNA and nucleosomes, RNAP can be recruited to the TSS/promoter region. Initiation factors, such as the TATA box-binding protein, may bind in a sequence-specific manner at active promoters upstream of the TSS which can then recruit further proteins including TFs and RNAP to form the PIC, as well as HATs/HMTs for further chromatin remodelling. DNA-binding TFs often associate with specific sequence motifs at the promoter (or enhancer) region and may find their target site via 3d diffusion through the cytoplasm. However, TFs have been shown to locate targets up to two orders of magnitude faster than 3d diffusion alone would allow and that this is achieved through a combination of 1d diffusion, in which the TF slides along the DNA, intersegmental transfer, in which the sliding TF may "hop" between two proximal DNA strands, and standard 3d cytoplasmic/nucleoplasmic diffusion [29, 30] (figure 2). This occurs through the non-specific binding of the TF to segments of open euchromatin away from the target site through electrostatic interactions, which while tethered to the DNA filament, is bound weakly enough to enable free and rapid sliding of the TF along the DNA. The TF may hop between proximal DNA segments either by directly transferring between them during a brief period in which the TF is non-specifically associated with both segments, or alternatively by dissociating with the current segment, diffusing a short distance through the nucleoplasm to the new segment and associating with it [30]. Facilitated diffusion allows more rapid target finding, with the TF becoming strongly and specifically bound to the target sequence motif once it slides over it while diffusing in 1d along the DNA, subsequently contributing to transcriptional activation of the gene. Additionally, the TF may dissociate from the target site but remain tethered to the DNA, sliding back and forth and then pass over the target site to become specifically bound again before fully unbinding from the DNA, resulting in multiple gene activation events in quick succession [29]. In promoters which lack binding sequences on the DNA, initiation factors may instead anchor themselves directly to nucleosomes, as has been reported in the case of TFIID becoming anchored to TSS-flanking nucleosomes via the H3K4me3 modification, thus facilitating transcriptional activation [31]. The PIC is responsible for the opening of DNA at the promoter by denaturing it from the double-stranded to single-stranded state, which is achieved for the RNAP I and III PICs simply by binding which melts and breaks the hydrophobic base-base interactions, whereas the RNAP II PIC depends on the action of the XPB DNA translocase to unwind the DNA and make it accessible to RNAP [32, 33].

Turning specifically to RNAP II, its C-terminal domain (CTD) is central to the regulation of all phases of the transcription process. After unwinding of the DNA at the promoter, in order to initiate transcription the fifth and seventh serine (Ser5 and Ser7) residues of the RNAP II CTD is phosphorylated by CDK7 in a process facilitated by the Mediator coactivator (figure 3). This releases RNAP II from the PIC, allowing it to synthesise the first few bases of the nascent RNA transcript by proceeding over the template DNA strand [1, 34, 35]. Certain DNA sequences may interrupt the procession of RNAP II along the DNA, resulting in pausing which often occurs ~ 50 bps downstream of the TSS (figure 3), just before the $+1$ TSS-flanking downstream nucleosome. Pausing is stabilised by the DSIF and NELF factors, which bind to either end of the RNAP II, and this can lead to backtracking or premature termination. While paused, the phosphorylated Ser5 residue of the RNAP II CTD recruits factors responsible for 5' capping of the nascent transcript, protecting it from degradation. TFIIS to required for the release of promoter-proximal pausing into the productive elongation phase, but is inhibited from binding to RNAP II by NELF. Pausing release is automatically achieved by a TFIIS-like subunit of RNAP I and III but depends upon phosphorylation of NELF, DSIF and the second serine (Ser2) residue of the CTD by P-TEFb (CDK9). This causes disassociation of NEFL from the chromatin, allowing binding of TFIIS, and converts DSIF into a positive elongation factor, forming an important part of the elongation complex and playing a role in co-transcriptional processes. The Ser2-phosphorylated CTD not only recruits elongation factors but also other factors needed for co-transcriptional processes like modification of histones to continue clearing nucleosomes and RNA splicing to remove intronic sections of the nascent transcript and stitch together exonic regions ready for translation, while the CTD Ser5 becomes increasingly dephosphorylated by phosphatases as elongation progresses [1, 33, 35, 36].

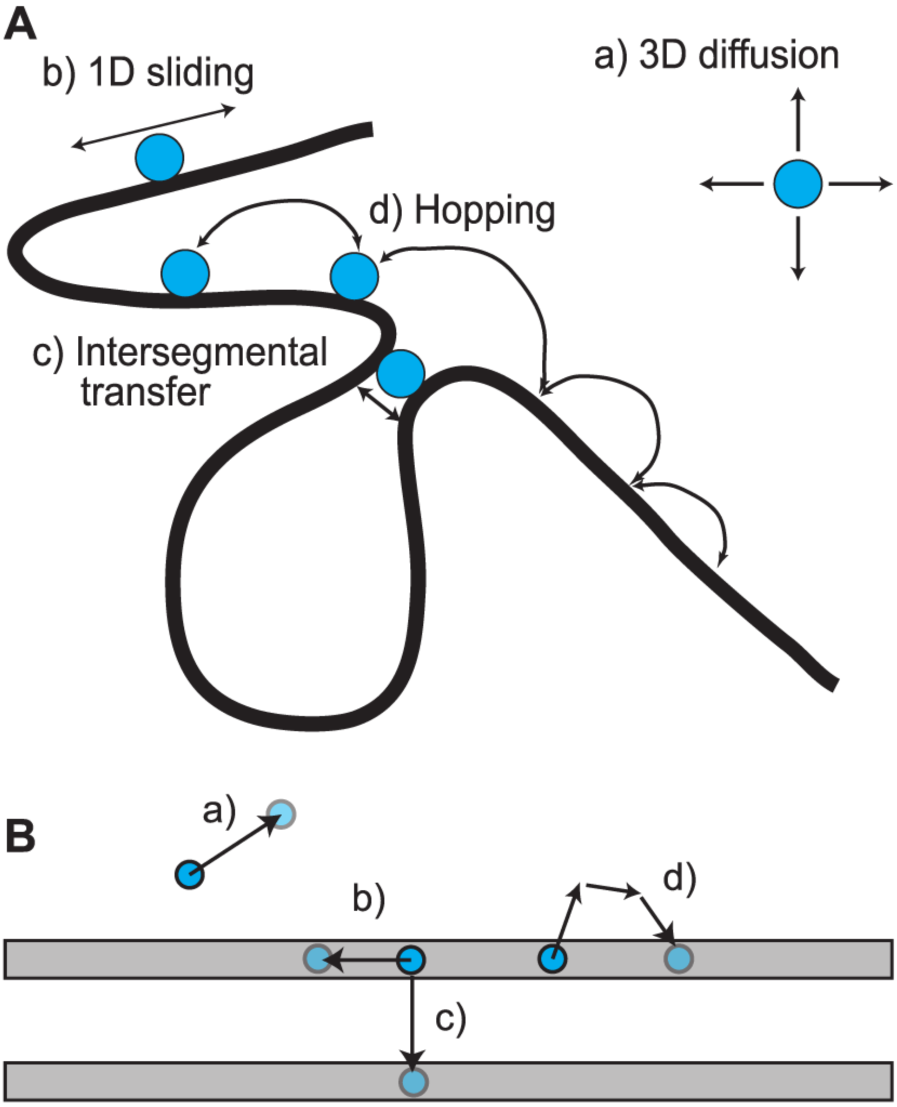


Figure 2: Figure taken from [30] showing how TFs move through the genome and nucleus towards their target sites. **A** Schematic showing how the TFs may be transported throughout a complex chromatin structure by 3d diffusion of the TF through the nucleoplasm (a), 1d sliding of the TF along the DNA segment when non-specifically bound (b), direct intersegmental transfer of the TF between proximal DNA segments without becoming fully untethered from the DNA (c) and hopping between segments or along the same segment, in which the TF alternates between 1d sliding and short excursions out away form the segment to diffuse in 3d before becoming non-specifically bound again (d). **B** Schematic showing the aforementioned modes of motion of a TF specifically in relation to two local DNA segments.

1.5 Elongation and splicing

There are several important co-transcriptional processes influencing the nascent transcript (like 5' capping during initiation/pausing) or chromatin which occur during the elongation phase of transcription or during the subsequent termination. This includes chromatin remodelling, RNA splicing, cleavage and polyadenylation (figure 3). During elongation the nucleosomes exhibit a dynamic turnover in which they are temporarily disassembled to facilitate movement of RNAP II through the chromatin and then reassembled in a process mediated by histone chaperones and chromatin remodellers [1]. The disassembly is induced by RNAP II and involves the removal of one of the H2A-H2B dimers from the octamer nucleosome, leaving a hexasome behind and facilitating passage of RNAP II. As previously mentioned, while some HMs are primarily localised to the promoter/TSS regions, such as H3K4me2, H3K4me3 and H3K27ac, others are found in the GB instead like H3K36me3 and H3K79me2. Although the exact action/regulation of H3K79me2 during elongation is unclear, H3K36me3 is transferred by the KMT SETD2, which is recruited to the chromatin during elongation by its association with the phosphorylated RNAP II CTD. This results in proper regulation of chromatin remodelling, such as deacetylation of coding regions to prevent initiation of intragenic transcription [37].

Splicing is another co-transcriptional process occurring during elongation in which the large ribonucleoprotein complex known as the spliceosome mediates removal of intronic sequences from the nascent transcript (or precursor mRNA, pre-mRNA) by cleaving the transcript at conserved splice sites. There is evidence that the elongation speed of RNAP II is lower at exons due to sequence and chromatin features such as increased H3K79me2 at introns as well as preferential binding of H3K36me3-modified nucleosomes at exons, resulting in reduced nucleosome density in introns compared to exons [38, 39, 40]. This, along with exon-binding proteins and the aforementioned exonic deposits of H3K36me3, may aid the spliceosome in distinguishing between exons and introns to facilitate proper splicing [2, 41]. Indeed, proximity to the chromatin has been shown to be important for splicing efficiency even when the transcript has undergone cleavage and polyadenylation [42]. Furthermore, there is strong evidence based on nucleoplasmic RNA measurements that constitutive introns (which are spliced in all isoforms of the gene) may be spliced almost entirely co-transcriptionally whereas alternative introns may frequently undergo post-transcriptional splicing [42]. H3K36me3 has also been shown to directly recruit splicing machinery to the DNA, such as providing a docking site for MRG15 which binds the splice factor PTB to recruit the nucleosome [43].

Additionally, there is evidence of a reverse interaction, with the intronic H3K36me3 density increasing further downstream in the GB specifically in a stepwise fashion after each subsequent exon, indicating that splicing also recruits SETD2 and enhances H3K36me3 deposition, especially considering that splicing inhibition has been directly shown to reduce SETD2 recruitment [38, 43]. Preferential assembly of nucleosomes and deposition of H3K36me3 at exons has also been linked to differential GC content which distinguishes exons and introns in addition to the presence of splicing sites [39]. H3K36me3, elongation and splicing are closely intertwined and the hypothesis that increased H3K36me3-modified nucleosome density in exons enhances splicing is supported by the finding that exons with weak splice sites, which are less efficiently targeted by the spliceosome than exons with strong splice sites, have higher levels of nucleosomes and H3K36me3 to compensate [39]. On the other hand, alternatively spliced exons have lower H3K36me3 levels than constitutively spliced exons, correlating H3K36me3 with the probability of the exon being spliced out of the flanking introns and included in the mature transcript [40], reinforcing its role as a splicing marker. The differences between introns and various types of exons in terms of both H3K36me3 levels and the nucleosome density is conserved, being previously observed in humans, mice, flies and worms [38, 39, 40].

Taken together, this evidence and previous work points to a mechanism whereby high exonic H3K36me3-modified nucleosome density is established by GC content, elongating RNAP and splicing factors, which then interacts with the spliceosome more strongly than introns, with their lower density of H3K36me3-modified nucleosomes, helping guide the splicing machinery towards the splice sites which border the exonic and intronic transcript sequences. While elongation in exons is maintained by H3K36me3-mediated recruitment of elongation factors, the higher nucleosome density reduces RNAP elongation speed, allowing time both for proper folding of the nascent transcript and for the spliceosome, which may be attached to the chromatin through splicing factors like PTB, to remain in proximity of the splice site for longer. This increases the chance of cleavage before further elongation may cause the splice site to become less closely tethered to the DNA/chromatin and, therefore, more distal to the spliceosome.

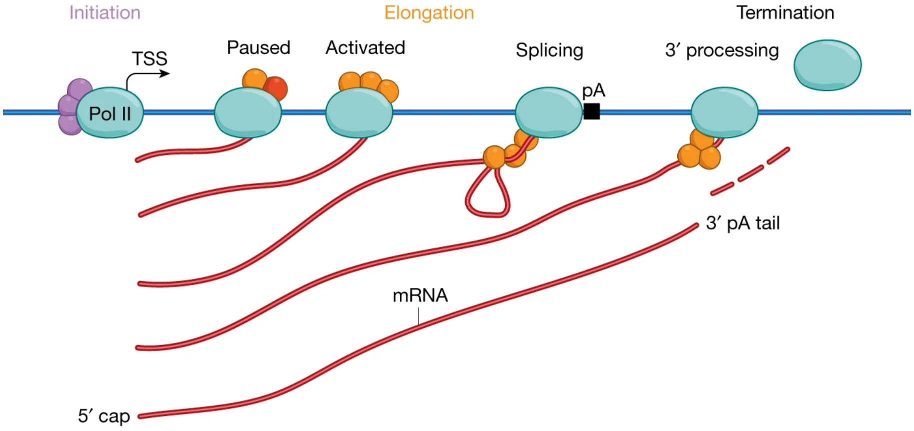


Figure 3: Diagram of the transcription cycle for RNAP II-mediated transcription taken from [33]. The different phases require the association of different sets of protein factors with RNAP II, as denoted by colour (purple for initiation and red/orange for elongation). Different sets of proteins are bound to RNAP II during pausing and elongation based on the phosphorylation state of the CTD. Splicing, polyadenylation and cleavage require transfer of proteins bound to the RNAP II directly onto the nascent pre-mRNA transcript.

1.6 Termination and transcript release

The final step of transcription sees RNAP II transition from the elongation to termination phase at the TES, although this is not necessarily the end of the line for RNAP II due to the phenomenon of transcription reinitiation through polymerase recycling. Cleavage and polyadenylation of the nascent transcript are key processes immediately preceding transcription termination, which occurs when RNAP II transcribes the polyadenylation signal sequence at the TES (figure 3). This triggers the cleavage and polyadenylation specificity factor (CPSF) to transfer from the RNAP II CTD to the polyadenylation signal site of the pre-mRNA, which along with other factors such as the Ser2-phosphorylated CTD-binding cleavage factor Pcf11, leads to cleavage of the 3' end of the transcript and subsequent polyadenylation catalysed by polyadenylate polymerase (PAP). Once PAP has extended the polyA tail of the transcript to about 200-250 nucleotides, its contact with CPSF weakens to breaking point, causing cessation of polyadenylation. As with 5' capping and the appropriate splicing of introns, the polyA tail is critical in determining the stability of the transcript, as well as exportation from the nucleus [1, 35, 44, 45, 46]. Even after cleaveage, polyadenylation and release of the transcript from the transcription complex and chromatin, RNAP II continues to transcribe for several thousand bp on average beyond the TES. There are two main mechanistic models explaining the eventual termination of transcription and disassociation of RNAP II; the allosteric model and the torpedo model. The allosteric model states that the loss of several important proteins from the transcription complex after the polyA signal results in a reduced efficiency of elongation due to a conformational change, which makes dissociation from the DNA at each subsequently transcribed base more likely, until eventually termination occurs. The alternative, 'torpedo' model instead posits that termination relies on degradation of the remaining uncapped RNA after cleavage by the XRN2 exonuclease, which degrades the RNA faster than RNAP II processes over the DNA until eventually XRN2 catches up and torpedos into RNAP II, thereby displacing it and resulting in termination [1, 2, 4, 35, 47].

1.7 Stochastic transcription, noise and bursting

Beyond the biochemical description outlined above, transcription is also a stochastic process subject to intrinsic noise through its fundamental dependence on probabilistic collisions between molecules [48, 49, 50]. These molecules are often present at relatively low counts, resulting in relatively larger fluctuations associated with each integer-valued change in copy number, which corresponds to poissonian noise and a poisson distribution of transcripts across cells in the regime of constant transcript synthesis and degradation. On top of this, in many cases, transcription occurs in bursts of higher activity followed by periods of lower activity or inactivity, which may greatly enhance the cell-cell variability in transcript counts and enable super-poissonian noise levels [51, 52]. 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, intense windows of activity [53, 54]. Variation in gene expression between cells may arise not only due to the intrinsic noise associated with each gene copy, but may also have a an extrinsic noise component, in which heterogeneity may arise due to to global variation between cells such as cell-cycle phase and varying numbers of cellular components (e.g. enzymes or metabolites) which affect gene copies within a cell in a correlated manner [55, 56, 57, 58]. Gene expression noise is not only restricted to transcription, but also to protein copy numbers, with translational noise having been a heavily-studied topic for many years, although the same fundamental principles underly noise intrinsic to transcription and translation; discrete molecule numbers magnifying fluctuations at low molecule numbers and the production of molecules in bursts due to switches between states with different production rates (gene on/off, mRNA present/absent). Large protein fluctuations are observed when transcription rates are slow and translation rates are fast, whereas small fluctuations are seen with slow transcription and fast translation, because the protein numbers are more or less sensitive to the instantaneous mRNA state of the cell, respectively [59]. This logic is also applicable to transcriptional noise since as proteins are produced from a mRNA copy during translation, transcripts are produced from a gene copy during transcription, with relatively slow and fast promoter (gene state) switching leading to large or small transcript fluctuations, respectively. The gene expression noise is generally viewed as the size of the fluctuations in copy numbers relative to the average copy number, such that although the absolute variance for a high expression gene may be larger than for a low expression gene, the fold-difference in copy numbers between any given pair of cells may be lower on average than for a low expression gene that experiences larger fluctuations as a proportion of its mean expression level. Therefore, transcriptional noise across a cell population is often quantified as the coefficient of variation (CV, standard deviation over mean), the Fano factor (variance over mean) or even the variance over the square of the mean [56, 57, 58].

Fluctuations in transcript levels can reduce the signal to noise ratio of information propagation through a gene regulatory network (GRN), which can be a harmful result for the cell [60]. On the other hand, gene expression noise and cell-cell variance can be utilised as a means of inducing diversity within a cell population. For example, in eukaryotic, a genetically identical stem cell population may achieve alternative cell fates during differentiation without requiring explicit control by genetic programming or external signals through the probabilistic switching in expression level of genes key for different cell fates [60, 61, 62]. It has also been hypothesised as an effective bet-hedging mechanism in prokaryotes to enable rapid adaption of a colony to stress and changing conditions by endowing different subpopulations of the colony with diverse phenotypes without requiring genetic mutations, such that at least one subpopulation may be able to survive and adapt to sudden environmental stresses. One example is that of persistent bacterial infections, in which the quickly growing bacteria are killed by antibiotic treatment but a handful of dormant, slow-growing bacteria, which exhibit this alternative phenotype due to gene expression noise, survive extended periods before later switching back into the fast-growth phenotype, resulting in the infection once again taking hold. Indeed gene-intrinsic noise is subject to natural selection in both prokaryotes and eukaryotes by modulating the level of bursting, while the observation of bursting across different domains of life demonstrates that transcriptional bursting and stochastic noise in gene expression are fundamental phenomena [60, 62]. Taking a classical view of transcriptional bursting, in which transcription occurs only during active periods which are separated by periods of inactivity, one may understand it in terms of several key parameters (figure 4) which govern the transcriptional dynamics. These include 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, $\mu = b \times a$).

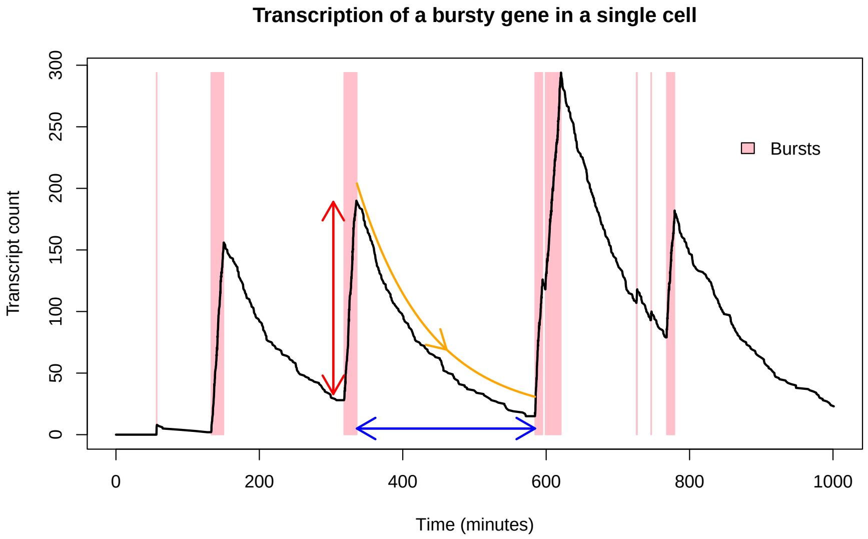


Figure 4: Simulation demonstrating transcriptional bursting for a single gene in a single cell, indicating burst size (red), burst interval (blue, reciprocal of burst frequency), and decay rate (orange, reciprocal of transcript lifetime), while the thickness of the pink shaded regions indicate burst durations.

1.7.1 Biological origins

As previously mentioned, intrinsic transcriptional noise arises as a result of random encounters of molecules which results in molecule-level noise since the biochemical processes themselves are inherently stochastic, as well as through the process of transcriptional bursting, which may push intrinsic noise to super-poissonian levels. Transcriptional bursting and the switching between gene states with differing activity levels is associated with a wide variety of underlying mechanisms, which may vary depending on the gene, cell type and organism, including TFs, chromatin, nucleosomes, enhancers, and chromosome topology [63]. Transcriptional bursts may be regulated at/before the time of initiation by factors which are proximal to the promoter or gene or by distal factors, and may be modulated by events occurring after initiation has taken place. Bursts are initiated through occupation of cis-regulatory elements (DNA sequences at/around the promoter) by TFs, which is primarily related to modulation of burst frequency. The capacity of TFs to bind at the promoter is crucial in determining the burst frequency, with the loss of binding motifs resulting in less frequent bursting, while depletion of nucleosomes from the promoter strongly increases burst frequency, since nucleosome packaging reduces promoter accessibility to TFs [64]. Single molecule fluorescence in situ hybridisation (smFISH) data on the c-Fos gene also showed that higher TF concentration resulted more frequent bursts, whereas the burst size is related to the lifetime of the TF on the promoter with longer TF occupation times resulting in longer bursts, while burst size is also increased for TFs with a stronger transactivation domain, instead through increased RNAP initiation rate which increases the number of transcripts produced during a give active window [65]. The c-Fos gene is enriched for paused RNAP at/near the TSS, which may strengthen the relationship between TF binding duration and burst size, since keeping the active state maintained for longer only produces a longer burst if there are enough RNAP to continue transcribing throughout the window. TF binding proximal to the TSS immediately preceding transcriptional bursts has also been observed, providing supporting evidence that TF binding initiates bursts. 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 [66]. This cooperative activation results in non-linear gene regulation and increased burst frequency and burst size for genes enriched in coactivators. The promoter-proximal sequence architecture also influences bursting, with increased burst frequency being achieved by having more and/or higher affinity cis-regulatory elements, although mutation of certain TF binding sites has also been shown to reduce burst size [63]. Recent works directly observing live-cell nascent RNA have found very short burst durations from seconds to minutes, with multiple rounds of transcriptional initiation. Large bursts observed with smFISH, for example, are actually likely a series of small bursts in quick succession, which links to ideas that TFs bind DNA both specifically (at the target site) and non-specifically, able to slide back and forth across the DNA, resulting in multiple tiny rapid bursts each time the TF passes over the target site, until the TF fully dissociates from the DNA, at which point no transcription occurs for a while [29]. From a physical perspective, these are in fact multiple successive small bursts during the tethered period in which the TF remains at least non-specifically bound with intermittent specific binding/unbinding events. However, since these sliding kinetics and small bursts occur on a much shorter timescale than transcript degradation or TF-DNA dissociation/association events they could be considered collectively as a single, large burst, with a duration corresponding to the total time the TF spends tethered to the DNA proximal to the target site before fully dissociating from the DNA back into the cytoplasm/nucleoplasm. Reducing TF binding affinity through mutation of the binding site may, therefore, result in reduced burst duration and burst size due to a lower fraction of tethered time being spent specifically bound at the target site and relatively more time spent non-specifically bound to the surrounding sequences, during which there is greater probability to become fully dissociated [63].

Enhancers, although potentially very distal to their target genes, are key for transcriptional activation in many cases and have the capacity to recruit TFs and chromatin remodelling factors and may either silence or activate target genes. Enhancers which are proximal to the promoter in 3d space may increase local TF concentration by having multiple TF binding sites, which in turn recruit RNAP and other cofactors. This proximity may be achieved through the transient looping of chromatin to bring enhancer and promoter close together and deposit the recruited factors from the enhancer to the TSS. If the enhancer interacting with the promoter enables transcription, then longer, more stable interactions increase the duration of the active transcription window, thereby increasing the burst size [63]. Loss of enhancers has also been shown to result in the reduction of burst frequency as well as a weaker reduction in burst size, which is hypothesised to occur through transcriptional activation instead being mediated by alternative/redundant enhancers which have less frequent and lower stability interactions with the promoter [63]. Increased looping of the chromatin between enhancer and promoter has been shown to increase burst frequency through higher probability of the enhancer-promoter interaction occurring. Topological regulation in general is also crucial for bursting regulation. The accessibility for factors binding to trans-regulatory elements, such as enhancers, or to cis-regulatory elements at the promoter, as well as recruitment of factors to the GB, depends upon the structure of the DNA helix. This is strongly linked to the level of supercoiling of the DNA, with supercoiling being caused by transcription in both prokaryotes and eukaryotes [63]. As RNAP processes through the gene, it leaves positive super coiling behind it and induces negative supercoiling in front of it. Successive RNAPs result in a greater accumulation of supercoiling, which can eventually cause structural inhibition of further transcription, thereby potentially acting as an inherent constraint on burst size and leading to a refractory period. Supercoiling has indeed been shown to negatively impact transcription mediated by both RNAP I and II. Topoisomerase enzymes can relieve supercoiling in a concentration-dependent fashion, with bursting having been shown to rely upon expression levels of such supercoiling-relieving enzymes in prokaryotes. 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 re-accumulation of supercoiling, results in bursty prokaryotic transcription [67]. In eukaryotes, supercoiling has been shown to facilitate spacing between RNAPs during a transcriptional burst to avoid them crashing into each other. The pausing of RNAP after initiation may also be important in regulating transcriptional bursting [63]. Genome-wide enrichment of RNAP II 50-100 bps downstream of the TSS indicates polymerase pausing proximal to the TSS, with the transition from pausing to elongation involving phosphorylation of the C-terminal domain of RNAP II, as previously mentioned. Pause times of roughly 40 seconds have previously been inferred mathematically, with up to 90% of RNAP II undergoing abortive transcription rather than entering productive elongation. Cells may be able to modulate the number of RNAP II released from pausing during an active window to regulate the burst size as an alternative to modulating the total number of RNAP II recruited and initiated. Promoter-proximal pausing dynamics could also influence the frequency of bursts in conjunction with initiation rate [68]. RNAP may also undergo the process of reinitiation, in which after transcribing a gene the RNAP is immediately recycled to the TSS instead of simply terminating and disengaging [69]. This requires looping of the gene to bring the TSS and TES into physical proximity [70], and the link between TSS-TES interactions and bursting has been explored recently, pointing to increased burst size by facilitating a larger number of total RNAP initiation (and reinitiation) events during a given active period [71].

In eukaryotes, the nucleosome packaging of chromatin plays a central role in transcriptional bursting, with peak nucleosome turnover in flies is found to be located immediately downstream of the TSS, coinciding with the locus of maximal RNAP II enrichment [63]. Nucleosome turnover also occurs at enhancers, and expression level of a gene is directly proportional to the level of nucleosome turnover at its promoter, while histone turnover time is similar to the time between transcriptional bursts. Different HMs may result in looser or tighter packing of the chromatin, with the chromatin density around the TSS being correlated with transcriptional noise [72]. 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 [72, 73]. Histone acetylation is important for controlling burst frequency over circadian rhythms, and in mouse cells is known to modulate both burst frequency and size [74]. H3K4me3 has also been indicated to play a role in determination of burst duration. Histone acetylation turnover occurs on short timescales similar to burst durations, but histone methylation turnover occurs over much longer timescales, noting that daughter cells have been shown to have inherited burst frequencies close to the mother cells, with the inheritance being dependent upon H3K4 methylation, a known active transcription histone mark [63]. Histone acetylation may be more important for regulating transcriptional bursting dynamics occurring on short timescales, whereas histone methylation may play a more important role in determining transcriptional bursting dynamics over long timescales. Strongly repressed chromatin, which can also be achieved through histone methylations like H3K27me3, in which transcription is inhibited for large periods of time, is associated with increased transcriptional noise, where occasional bursts cause a small handful of cells to express the gene while the rest of the population is silent [63]. Proximity of nucleosomes to TF binding sites have also been shown to increase the dissociation rate by up to 1000-fold, resulting in reduced occupation time, burst duration and burst size. Although TF binding is the rate-limiting step in transcriptional bursting, it can only occur when permitted by brief periods of nucleosome remodelling, when nucleosomes are cleared from the TSS area to facilitate TF binding, leading to a transcriptional burst. Nucleosomes with different HMs may have different propensities to be cleared and for different durations. Recent studies have also reported genome-wide direct correlations between the presence of specific HMs at gene promoters and general transcriptional noise [75, 76], while further studies have even linked HMs with the underlying bursting dynamics, both at the individual gene level [74] and genome-wide [77].

The higher order chromosomal architecture, such as that measured by HiC, also has an impact on transcriptional bursting. Eukaryotic genomes are organised in 3d into topologically associated domains (TADs), which may form via loop extrusion. Enhancers and promoters within the same TAD are assumed to interact more frequently and strongly than those located in different TADs, which would mean that TADs are important for facilitating or preventing regulation of target genes by enhancers. This would result in increased burst frequency and sizes for genes located within the same TAD as their regulating enhancer compared with those located in separate TADs [63]. However, the difference is not extreme, with those within the same TAD twice as likely to interact as those in adjacent TADs. Different chromosomal configurations have been observed across a cell population, indicating that the TAD organisation is dynamic, changing over time, such that genes could interact with different sets of enhancers at different times depending on the TADs they inhabit, resulting in shifts in their transcriptional bursting dynamics and changes in transcriptional noise. CTCF as well as cohesin are important for TAD establishment, with enhancer-proximal CTCF sites acting to strengthen enhancer-promoter interactions. Deletion of such sites resulted in a minimal change in expression level despite reduced enhancer-promoter interaction frequency, whereas noise levels were increased, indicating that burst frequency was reduced but was compensated by increased burst size [63].

Extrinsic noise is also crucial for determining the total transcriptional noise alongside intrinsic noise, and has a wide array of biological sources. Such extrinsic noise sources may be entirely stochastic and unpredictable or can have a deterministic component, in which they vary periodically, and are therefore more predictable, with stochastic fluctuations about the deterministic mean, such as the cell cycle phase and circadian rhythm [55]. Fully stochastic extrinsic noise sources may include historical events in the cells history, such as the activation of a gene whose product is related to transcription, such as a TF, with a positive feedback loop magnifying the effect, resulting in a long-term accumulation of the TF in that cell, causing a global increase in the burst frequency in that cell until TF levels become depleted to levels more comparable with its neighbour cells [55]. Asynchronised cell populations also have a non-uniform distribution of cells over the cell cycle because new daughter cells are more abundant than those on the cusp of division since each dividing cell produces two new ones.

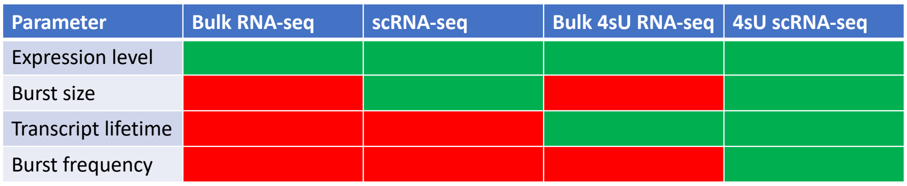
1.7.2 Data generation

Early studies of stochastic gene expression focussed on translational noise due to the greater ease of quantifying protein than transcript abundances with previously available technology, since protein copy numbers are generally higher and more easily detectable. A common theme was the decomposition of noise into intrinsic vs extrinsic and the quantification of each component. Such analyses relied on dual reporter systems, with two differently coloured fluorescent reporter proteins controlled by identical regulatory sequences or promoters, which in previous studies have been inserted into both prokaryotic and eukaryotic genomes, specifically at the same locus in homologous chromosomes in the eukaryotic case [78, 56]. This aimed to ensure that the two reporters were under identical intrinsic noise conditions, with measurement of total noise levels being achieved through quantification of the fluorescent intensities observed in each cell across a population upon imaging. The proportion of noise originating from extrinsic sources was obtained by measuring the correlation between the two reporter levels across the cell population, with the remaining noise belonging to intrinsic, gene copy-specific fluctuations. The total noise was quantified as the CV, with stronger correlations indicating a larger extrinsic noise contribution. One eukaryotic study of this nature found the majority of measured expression noise to be composed of extrinsic noise sources [56]. However, control of factors such as cell size, shape and cell-cycle phase through the use of fluorescence-activated cell sorting (FACS) reduced but did not eliminate extrinsic noise despite separate analyses of subpopulations homogenous in these factors, indicating other sources of heterogeneity within the cells which are not apparent under the microscope, such as variations in cellular components.

Due to the importance of small molecule mRNA fluctuations in stochastic noise, the development of fluorescence imaging technologies with the capacity to distinguish individual molecules from background levels was a major breakthrough in directly quantifying transcript copy numbers and thus transcriptional noise, with transcript levels being more sensitive to the gene state than proteins, which are longer-lived . An early study of this type directly showed that transcriptional bursting occurs in prokaryotes by using the MS2-GFP fusion protein to tag transcripts, causing them to fluoresce and facilitating their integer-valued detection at single-molecule resolution [79]. This was contrary to prevailing ideas that poisson noise dominates bacterial transcription as opposed to accepted idea of super-poissonian bursty eukaryotic transcription, with observed super-poissonian protein noise in prokaryotes arising due to translational not transcriptional bursting. Transcriptional bursts were directly observed in E. coli using time-series data, with measurements on burst size and wait times conforming to mathematical predictions made under the assumption of transcriptional bursting. Observing cells throughout cell division has also indicated binomial partitioning of transcripts into daughter cells, contributing to transcriptional noise extrinsically, with higher total transcript numbers resulting in a smaller injection of noise upon partitioning [79].

Many modern studies also make use of fluorescence microscopy-based approaches to interrogate transcriptional bursting dynamics by imaging a cell population and quantifying transcript levels within individual cells based on the continuous-valued fluorescent intensity observed within each cell. Single molecule fluorescence in situ hybridisation (smFISH) is a particularly popular approach in recent years although the standard procedure, which uses fluorescently tagged RNA probes to target specific mRNA and emit a signal, 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 [80], allowing estimation of expression level, burst size and burst rate (μ, b and a) but not burst frequency or decay rate (κ or δ). Some smFISH-based experimental set-ups have progressed towards a level of understanding bursting timescales by using hybridisation specific to nascent transcripts [81, 82], 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 [83, 84, 85] or requires complex and labourious set-ups [86]. Sophisticated analysis methods [87] have been developed for time-lapse single-cell RNA imaging data [88] which allows dissection of transcription dynamics in great detail, however such approaches are even more limited scale-wise.

Unlike fluorescence-based techniques, RNA sequencing (RNA-seq) technologies have long been used to detect transcripts present in a sample across the entire genome [89]. The procedure begins with the extraction of the RNA content from a sample of cells of interest, which is often achieved by lysing the cells and then precipitating any RNA with a polyA tail out of the lysis by running it over a column or magnetic beads with oligo-d(T) molecules attached, which selects for polyA RNA through the A-T complimentary interaction [90, 91]. This approach, in which the RNA content across the entire population of cells is immediately pooled without regard for which the cell from which each transcript originated, is known as bulk RNA-seq. Purified RNAs are then fragmented to a certain size using either chemical (alkaline solutions or solutions with cations), biological (RNA cleavage enzymes) or physical (sonication) methods before carrying out reverse transcription of RNA fragments to cDNA using random hexamer primers [91]. During the preparation of the cDNA library after reverse transcription, cDNAs produced from the fragmented RNAs are ligated to flanking adapter sequences. These adapters act as primers for the subsequent PCR amplification of the cDNA, which is carried out to ensure enough material is present for sequencing, and the adapters also allow the cDNA to bind to the sequencing flow cell [91]. The sequencing of cDNA may then be carried out using a variety of technologies, such as sequence-by-synthesis, in which complementary strand synthesis is carried out and each subsequently incorporated nucleotide is detected and identified through, for example, pH changes [92] or via fluorescent signal when using nucleotides modified with fluorophores [93], such that the original RNA sequence can be discerned and aligned with a reference genome sequence to find its corresponding gene. Such bulk RNA-seq methods may be used to analyse the overall expression levels of genes averaged across the cell population (figure 5) and to look at relative expression levels between different genes within a sample or between different samples for the same gene (differential gene expression analysis) [94].



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

Figure 5: Table showing the parameters governing transcription dynamics that can theoretically be obtained using different RNA-seq approaches with no prior information. Green and red show if a data type does or does not inform a parameter, respectively.

While bulk RNA-seq neglects the cell-cell variation in transcript numbers, single cell RNA-seq (scRNA-seq) experiments tag sequencing read with cell-unique barcode IDs, and have therefore been 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 [54, 95, 96, 97]. On the other hand, bulk RNA-seq-based approaches have for several years made use of chemically labelled nucleotides, primarily 4-thiouridine (4sU), to understand RNA synthesis ($b \times \kappa$) and degradation (δ) rates, with SLAM-seq being the most well-known example [98, 99]. 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 [100]. 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 (figure 5). 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 [101, 102, 103, 104, 105]. 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 [106], while bursting timescales have remained uncharacterised in recent works [107]. Figure 6 provides a step-by-step diagram as an example of the 4sU scRNA-seq experimental protocol.

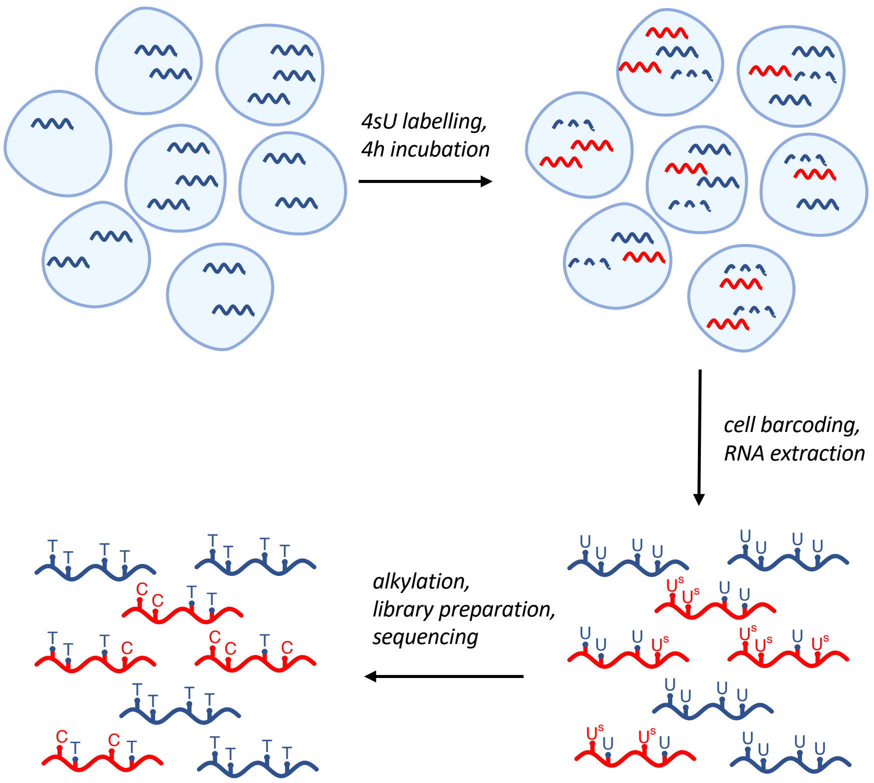


Figure 6: Cells with variable mRNA content (blue lines) shown for an example gene. Cells are incubated in the presence of the uracil analogue 4sU for a set amount of time, which is 4 hours in this example but can vary depending on the desired outcome of the experiment, such as using a shorter incubation (pulse) to analyse unstable transcripts. 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.

1.7.3 Modelling and analysis

Effective modelling of gene expression in biological systems demands the use of stochastic modelling approaches in conjunction with single-cell measurements of molecule numbers, as opposed to deterministic models with bulk assays, in order to quantify noise and how it is linked with cell-cell variability [108]. Deterministic models simply work with continuous-value concentrations whereas stochastic models explicitly account for the discrete nature of individual molecules within the biological system, along with their random synthesis, degradation and association events. Deterministic results for gene expression dynamics only align with stochastic models and experimental data when there are very high molecule numbers (large system size) and where the promoter kinetics (the switching on and off rates) are much faster than the other reactions that take place in the system (e.g. degradation rate) [60]. In this often unrealistic scenario, gene state lifetimes must be much shorter than molecule lifetimes, such that the only thing about the promoter kinetics that influences the molecule dynamics is the fraction of the time the gene spends in each state. If promoter kinetics are slow, the molecule numbers become more sensitive to the gene state. Rather than being unimodally distributed around a mean with poisson noise, they can become highly skewed, or even bimodally distributed with high and low expression state cells being observed throughout a population. Promoter state transitions, especially in eukaryotes, can be slow enough to influence the molecule dynamics, with the activation rate in particular often being comparable to the timescales upon which molecule degrade, for example [60]. This is due to factors such as nucleosome packaging of the DNA around promoters making them inaccessible to transcriptional machinery for extended periods. This increases the heterogeneity observed in a cell population, potentially even leading to a binary scenario in the extreme, in which cells either have high expression distributed around a value or none, corresponding to cells with the gene in the on or off states, respectively, with molecule levels responding rapidly to changes in gene state. The reliance of deterministic models on large populations which approximate concentrations is also broken by the prevalence of low copy numbers inside the cell of genes, transcripts, proteins and other components. The gene expression noise level, often quantified as the CV as previously mentioned, is larger with small molecule numbers, with CV scaling as the reciprocal of the square root of the number of molecules such that lower copy numbers leads to greater heterogeneity in a population of cells, which is called the finite-number effect. Noise may sometimes be alternatively quantified as the Fano factor for convenience since then it scales with the reciprocal of the number of molecules. These reasons have cemented stochastic modelling as the foremost analytical method for understanding transcriptional dynamics [60].

The simplest, and most analytically tractable, stochastic models of gene expression consider expression of a single gene copy, with the most simple model being the constitutively active gene which transcribes one RNA at a time at a constant rate, in which transcription events are randomly (uniformly) distributed throughout time independent of each other and each RNA is degraded at a constant rate with exponential survival times, leading to a poisson distribution of molecule counts throughout time or across cells and therefore poisson noise levels. The abundant observations of super-poissonian gene expression noise have lead to many microscopic mathematical models of transcription (and translation) capable of generating additional noise, including a very early study modelling the binomial partitioning of transcripts during cell division as an additional noise source [109]. In order to account for the observed phenomenon of transcriptional (and translational) bursting, models of gene expression with multiple states were required, capable of generating super-poissonian intrinsic noise. This led to the use of the two-state model, or random telegraph model, of gene expression (figure 7) as the most common and simple model of bursting in gene expression, in which the gene can switch between a repressed/off state with no transcription and an active/on state with a higher transcription rate, such that there are four possible reactions; activation of the gene when repressed, repression of the gene when active, RNA synthesis (transcription) when the gene is active and RNA degradation/decay at all times. The two-state model and variations of it have been used for decades to model stochastic transcription and bursting, often in conjunction with computer simulations, to explore how transcriptional dynamics and noise change with different parameter settings, such as TF binding affinity as the determinant of the on and off rates of the gene. One early study modelled the gene as being on or off depending on whether the transcription complex (of TFs and RNAP etc) is bound or not, with the switching between states being determined by the transcription complex association and dissociation rates, which are controlled by the stability of the complex on the regulatory region of the gene [110]. The model used did not account for degradation of transcripts, instead allowing product to accumulate over a given time interval. Two different scenarios with higher or lower transcription complex dissociation rates were examined across a range of association rates, with the transcription rate being fixed. Therefore, the on state half-life was variable between being either long and short, corresponding to scenarios with larger or smaller bursts, depending on whether the binding is more or less stable, respectively. The total accumulation of transcripts over a given time across different association rates, which correspond to burst frequencies, was simulated for the two dissociation rates (burst durations/sizes) for multiple independent gene copes to obtain simulated distributions. Gene induction, defined as the total accumulation of transcripts across the given timescale, was broadly found to be homogenous or heterogeneous across gene copies, with higher affinity resulting in more homogenous gene induction in both scenarios, while higher stability resulted in more heterogenous gene induction for a given affinity level. This is one of the earliest examples of theoretical modelling of stochastic transcription indicating that larger and less frequent bursts results in noisier transcription [110]. However, the approach suffers due to transcript degradation being neglected, with more recent studies showing that mRNA decay has a high contribution to noise in gene expression [111]. The switching time between high and low (or zero) expression states is controlled by degradation rate such that there is a trade-off between high and low degradation in terms of the gene expression signal-to-noise ratio, with a more clear signal in cases where the the switching time is fast, although this corresponds to greater sensitivity to the gene state and therefore higher total variability and noise in transcript levels [111]. Faster degradation also corresponds to lower expression levels and lower copy numbers, such that the transcript count becomes subject to stronger relative fluctuations [111].

The chemical master equation (CME) of the full two-state model has been used by previous works to derive discrete-value solutions to the steady state transcript count distribution across independent gene copies (such as those across a cell population), which takes the form of a poisson-beta compound distribution [112, 113]. This becomes a negative binomial distribution in the limit of instantaneous bursting (with vanishingly small burst durations and transcription rates approaching infinity, but finite burst size as the product of the two), which is valid under the assumption that burst durations are much shorter than transcript lifetimes, with bursts arriving in a poisson fashion with exponential wait times (based on the on rate, or burst frequency) and geometric burst size distributions (determined by the ratio of the transcription rate and off rate) [112, 51, 114, 113]. The analysis reported in [112] focussed on counting transcript numbers in individual cells using FISH, finding large variation between cells due to transcription occurring in short, intense bursts of activity, followed by longer periods of inactivity. Additionally, although the gene state transitions which cause bursting are themselves intrinsic, bursting in genes within a nearby genomic region were found to be correlated, pointing to the action of GRNs and/or local chromatin state as influences on the underlying bursting rate parameters [112]. Transcripts may provide a better read-out of gene expression noise and the gene state than proteins since they are a direct product of the active gene and are comparatively short-lived, meaning that noise generated by transcriptional bursts may be buffered downstream by slow protein degradation rates, with protein levels being more stable than transcript levels. However, the two-state model in the limit of instantaneous bursts, or the geometric burst model, has also been applied to the stochastic modelling of protein bursts, in which the on/off gene state dichotomy maps to the presence/absence of an individual mRNA copy, while the mRNA count distribution maps to a protein count distribution, such that the model structure remains unchanged, with activation, repression, transcription and mRNA decay mapping to (constitutive) transcription, mRNA decay, translation (from a single mRNA) and protein decay, respectively. Translational bursts are assumed to be instant, relying on transcript lifetimes being much shorter than protein lifetimes, as with the transcriptional burst model assuming that burst durations are much shorter than transcript lifetimes. Additionally, only a single mRNA copy may exist at any one instant, resulting in random, uncorrelated translational bursts, as with the transcriptional burst model, which only allows a single active gene copy at any instant. Since the two models are mathematically identical, the wealth of theoretical work on translational bursting is also relevant to understanding transcriptional bursting dynamics [115, 58, 114, 51]. Studies modelling translational bursting with this approach have previously treated protein abundances as continuous-valued concentrations, as opposed to the aforementioned realistic discrete-value transcriptional bursting approach. This was more immediately applicable to the continuous-value fluorescence imaging data which was used to quantify protein levels for which discretisation, such as that carried out in [112], was non-trivial due to higher technical noise levels at the time and because protein levels may often be higher than transcript levels, meaning that the proportional increase in fluorescence associated with one additional protein is less than for one additional transcript [115, 58]. These studies derived a steady state solution from the continuous-value version of the CME, again resulting in poisson arrival of bursts but exponential rather than geometric burst sizes, which is the continuous analogue of the geometric distribution, and a gamma distribution of molecule copy numbers rather than a negative binomial, which again is the continuous analogue [115, 58]. [58] used this gamma distribution solution to model stochastic translation in E. coli, finding that it fits well to experimental fluorescence-based steady state protein concentration measurements, even being robust at low expression levels, with an appropriately inflated peak at zero.

A recent study compared the steady state distributions of different mathematical models of transcription for a single gene copy, including thermodynamic and discrete-value kinetic models, based on statistical mechanics and chemical kinetics, respectively, which include constitutive and bursty promoters with/without repression [113]. Coarse-graining over different molecular processes of transcription led to indistinguishable results in terms of mean expression levels but differing results in terms of the higher order moments. It was found that thermodynamic models are only sufficient for describing mean expression levels at steady state, whereas kinetic models can go further in predicting higher order moments of the distributions, albeit with greater numbers of parameters to potentially complicate inference. Quantifying transcriptional noise with the Fano factor, it was shown to be < 1 for kinetic models with constitutive, non-bursty promoters, far lower than for experimental E.

coli data, with Fano factors > 1 nearly always being observed. Out of all models considered, only the two bursty promoter kinetic models of transcription fit experimental data in terms of noise levels, one of which is the standard two-state model (figure 7) and one of which is the previously mentioned instant geometric burst model. It is suggested that since both models are capable of explaining experimental data, the mathematically simpler version with instant bursts is preferable in practice, which has fewer parameters (3 instead of 4) [113]. Analytical progress has also been made towards discrete-value time-dependent solutions to the molecule copy number distribution of the two-state model of transcription (figure 7). An approximate time-dependent solution of translational bursting (which also maps to transcriptional bursting) was presented in [114], which is valid for timescales longer than several molecule lifetimes and relies on the assumption that proteins are much more stable than transcripts or that transcripts are much more stable than the active gene state depending on whether translational or transcriptional bursts are being modelling, eliminating the fast variable from the master equation which does not influence the dynamics of the stochastic system. Therefore, bursts are treated as instantaneous with a geometric size distribution, such that the time-dependent solution becomes the previously mentioned negative binomial steady steady solution in the limit as time tends towards infinity. A solution is described for zero initial molecules as well as for arbitrary initial conditions, also defining a three-stage model which specifically includes both transcriptional and translational bursting simultaneously [114].

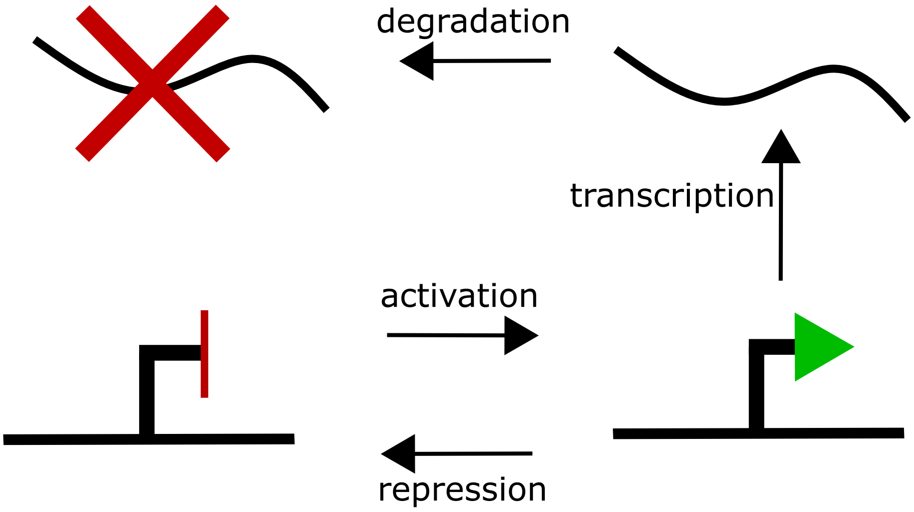


Figure 7: Schematic representation of the two state model, with the four reactions (activation, repression, transcription and degradation) acting on the three species (repressed gene, active gene and transcript). In this model no transcription occurs when the gene is repressed, while transcript degradation occurs independent of the gene state.

Further work has also explored how different types of gene activation influence transcriptional bursting dynamics and noise, including microscopic modelling of TF dynamics [29]. Two models of gene activation were examined, the first of which has TFs diffusing through the cytoplasm in 3d in a random walk process, binding specifically to the target site (at the promoter) when proximal with an association rate and unbinding from the DNA with a dissociation rate to re-enter the 3d diffusion mode, with transcription only occurring when the TF is bound, which corresponds to the standard two-state model of transcription. The other model allows for facilitated diffusion in which, as previously described, TFs diffuse through the cytoplasm in 3d but may also non-specifically bind/unbind to any DNA sequence with an association/dissociation rate, remaining loosely tethered to the DNA. This enables it to rapidly slide back and forth along the strand in a 1d diffusion process, becoming specifically bound upon passing over the target site to enable transcription, before returning to the non-specifically tethered state, leading to a more complex model of activation and transcription. Although facilitated TF diffusion is known to occur in biological systems, DNA-binding proteins have been shown to slide rapidly along the DNA, with the modelling carried out in [29] demonstrating that the kinetics of facilitated diffusion are fast enough that they do not influence the noise and transcriptional dynamics beyond what can be captured with the two-state model, serving only to modulate the burst frequency and/or size rather than changing the shape of the transcript count distributions. Additional modelling work on gene activation also investigated the role of promoter leakage on transcriptional noise and bursting [116], in which the standard two-state model is modified such that the off state with zero transcription is replaced with a low activity state with a non-zero transcription rate lower than the high activity (on) state [117]. Analytical solutions to the steady state transcript count distribution were derived for the two-state promoter leakage model of stochastic transcription. It was found in all cases that greater leakage (smaller relative difference between the high and low activity states) results in reduced noise for a given mean expression level, weakening the biomodality exhibited by genes with similar promoter state switching rates until the distribution becomes unimodal as the lower efficiency transcription rate approaches the higher efficiency transcription rate. Higher leakage also corresponds to shorter bursts for a given expression level. Therefore, promoter leakage may act as a mechanism of dampening gene expression noise, with the results holding for both fast and slow TF binding/unbinding rates [117].

Analyses have also investigated the possibility of avoiding certain assumptions made by the two-state model (or random telegraph model), which is that bursts are geometric and arrive in a poisson fashion with exponential wait times [64]. The paper focussed on generalising the model to allow for scenarios with non-poissonian burst arrival as well as non-geometric bursting by mapping ideas from queuing theory to analytically describe steady state stochastic gene expression distributions from which noise levels can be quantified. Stochastic models with a generic burst arrival process were represented by multi-step activation process to represent different initiation factors being recruited to the promoter or alternatively through the use of an arbitrary wait time distribution function, as opposed to single-step poissonian arrival, with generic bursting also being represented by an arbitrary burst size distribution. By mapping the multi-step gene activation process to multi-person queue times, queuing theory was exploited to derive analytical solutions to the moments of transcripts counts at steady state, including higher-order moments such as skewness and kurtosis, from which bursting parameters can be inferred, even for models with non-poissonian burst arrival and/or non-geometric burst size. Since in some cases, burst waiting time has been indicated to deviate from an exponential distribution, the capacity to infer whether this is the case for a given set of steady state measurements of transcript counts is a desirable goal while confirmation that burst sizes are geometric would also be an advantage. It was found that the first three measured steady-state transcript count moments are sufficient to determine whether there is a deviation from poissonian burst arrival or from geometric burst size [64].

When modelling gene expression noise it is also important to consider the decomposition of total noise into the intrinsic and extrinsic components, which has been the subject of several previous works [55, 112, 58, 57]. Importantly, discrete-valued low molecule number reactions and bursting are associated with intrinsic but not extrinsic noise, which arises due to cell-specific factors, as previously mentioned, with extrinsic noise being reported to play a less important role in eukaryotic systems compared to prokaryotic systems [112, 58]. One study on protein levels in E. coli quantified noise as being the variance over the square of the mean, which is the inverse to the number of protein bursts occurring per protein lifetime, with the proportion of that noise arising from extrinsic sources corresponding to the correlation strength in two fluorescent protein signals across cells, as previously mentioned [58]. At higher expression levels, genes exhibited much higher noise than the theoretical intrinsic minimum (poissonian) noise level, with a noise floor being observed below which genes do not drop regardless of expression level, indicating that extrinsic noise sources play a stronger role in shaping distributions at higher expression levels in prokaryotes. This stronger extrinsic protein noise component for higher expression level genes is hypothesised to be related to cell-cell variation in copy number of things like ribosomes, enzymes and metabolites, which high expression genes are inherently more sensitive to as they become rate-limiting steps in gene expression [58]. Many studies have carried out the statistical decomposition of noise and quantification of extrinsic noise based on the experimental solution of embedding identical independent systems within the same environment with observed correlations in the variation reflecting the influence of their common environment (extrinsic noise), whereas uncorrelated variation corresponds to intrinsic noise arising from within the system. However, this method only accounts for static extrinsic noise sources, which is often broken in dynamic biological systems by time-varying environmental heterogeneity [55, 57].

Such dynamic extrinsic noise sources may exhibit entirely stochastic fluctuations over time and be unpredictable, or may have a deterministic component that varies periodically and thus is more predictable, with stochastic fluctuations of varying magnitude about the deterministic mean. These types of noise sources, which are sometimes also referred to as upstream drivers of gene expression, can be categorised as periodic (with a repeating deterministic component), entrained (periodic and synchronised across the cell population) or random (entirely stochastic), with examples including the cell-cycle (periodic), circadian rhythms (entrained) and historical events in the cells life (random) such as a sudden accumulation of a TF due to the action of a GRN with a positive feedback loop, which could cause a global increase in the burst frequency in that cell until TF levels become depleted to levels more comparable with its neighbour cells. Some approaches aim to subtract either the intrinsic or dynamic extrinsic noise away from the total noise to allow separate modelling of the remaining component, but it was found to be easier to model the extrinsic noise by subtracting out the intrinsic noise than the reverse [55]. Subtracting extrinsic noise is difficult because of various complications, like asynchronised cell populations having a non-uniform distribution of cells over the cell cycle which is skewed towards newer cells. Another factor is that intrinsic noise levels depend upon extrinsic upstream drivers in a multiplicative manner rather than being additive, with an example being that the rates of biochemical reactions of transcription, like degradation of transcripts, depends on the product of the stoichiometries of enzymes and substrates, as well as rate constants, with the enzyme number being the source of extrinsic noise in this case. In a microscopic model of transcription this is generally not accounted for, instead collapsing the rate constant and enzyme number/concentration into one parameter as, for example, the degradation rate, which is multiplied by the transcript count (the substrate stoichiometry), making it difficult to statistically subtract away the extrinsic noise and decompose total noise into intrinsic noise. The impact of environmental variation may also vary strongly on a gene-by-gene bases. Microscopic models may resort to taking the average of rate values that fluctuate due to extrinsic noise, which therefore preserve the average molecule numbers in the system, however, this may not preserve the intrinsic noise and can lead to incorrect inference of parameter values in the scenario with transcriptional bursting [55]. More recent studies have developed approaches for modelling both the intrinsic noise sources (microscopically) and the time-varying extrinsic upstream drivers simultaneously, representing the dynamic extrinsic noise as changes in the parameter values of the microscopic model of transcription, with the specific modelling approaches depending upon the upstream driver being modelled. [57] presented a modelling framework for cells with time-varying rate parameters using both the constitutive one-state and bursty two-state models of transcription. This was an advancement on previous attempts to model dynamic upstream drives which used rate parameters that varied entirely deterministic, instead modelling parameters as time-dependent functions of stochastic variables. Several different models were developed, which included both an upstream driver extrinsic noise component (account for periodicity, entrainment or randomness) and a downstream intrinsic noise component (either poisson or poisson-beta), allowing for CMEs which do not rely on the assumption that gene expression is uncorrelated between cells. The solution to the CMEs are composed of a discrete transcriptional noise component and a continuous component corresponding to the time-evolution of the parameters governing transcriptional dynamics which varies depending on the upstream variation which is being modelled. This generalises the classical microscopic representations of transcriptional dynamics to systems with time-variant stochastic rates, thus allowing for both cell-cell variability in transcript levels but also in the upstream driver, even if the driver is correlated across cells. For example, the two-state model was implemented with a time-varying promoter strength depending on cyclical upstream drivers (such as cell-cycle), facilitating stochastically time-variable gene activation/repression rates with variable synchronicity, thus enabling modulation of burst frequencies and/or burst sizes by extrinsic noise sources [57]. One issue with this kind of approach is the large numbers of degrees of freedom, such that the posterior distribution of the parameters, which represents the probability density of parameters and hyper-parameters given observed data, may be strongly multi-modal, with several different combinations of values (points in parameter space) potentially explaining the given dataset. This could hinder the capacity to make biologically relevant inferences for all but the most detailed and information-rich datasets.

In the field of systems/synthetic biology, the theory of GRNs has been explored using various mathematical approaches from stochastic modelling, including the CME and the chemical Langevin equation, as well as numerical methods like Gillespie's stochastic simulation algorithm [118], as opposed to previously utilised deterministic frameworks like reaction rate equation models [119]. This explicitly accounts for both the interactions between different genes within a network as well as the stochastic fluctuations, which, especially at small molecule numbers, result in the propagation of noise through the network. This noise may be magnified in systems with positive feedback and feedforward control loops, while other GRNs may be robust against noise through their design, exhibiting negative feedback/forward control loops which dampen fluctuations. Thus, understanding of transcriptional noise is crucial for the inference of GRNs from data and for the effective construction and implementation of artificial gene circuits [119]. Observed correlations in molecule numbers with single-cell measurements across a population at steady stead in conjunction with such stochastic models can be used to infer GRNs and which genes interact and regulate each other based on the perturbations to the GRN caused by large fluctuations in copy numbers due to bursts [108]. A burst of one gene induces bursts in genes it regulates downstream, for example, resulting in a positive correlation across cells, demonstrating the importance of modelling bursting explicitly for such analyses. This approach may reveal clear gene groups with strongly correlated variation which often are involved in the same biological function, such as stress response genes. However, correlations based on steady state measurements only indicate genes with static relationships and may miss instances where one gene regulates another but there is a lag-time such that the two genes are not expressed simultaneously. To understand this, time-series analyses of gene expression are required in conjunction with cross-correlation statistics, which quantify the correlation between two gene product numbers at a given time-separation across the time-series data (such as time-lapse fluorescence microscopy with dual reporter genes) [108]. This can indicate both activating and repressive regulatory relationships depending on whether the cross-correlation score is positive of negative, respectively.

More recent work has been carried out in the context of a GRN with two mutually repressive genes which explored alternative modelling approaches and the links between models of gene expression noise at the microscopic and mesoscopic scales [51]. Microscopic models describe the processes underlying intrinsic noise in more detail, whereas mesoscopic models make assumptions about the finer details to facilitate computational scalability and can be more easily embedded within models of GRNs, for example. The microscopic model for the aforementioned GRN model explicitly models genes as being one-state with mRNA transcription and degradation, protein translation and degradation, and protein-mediated transcriptional repression as a Hill function of the protein number, modelling the state changes associated with each reaction on the individual molecule level. This maps to a two-state model of transcription in which the transcripts are mutually repressive agents. Mesoscopic models are also defined which eliminate the fast variable (the transcript in this case, but also mappable to be the gene state for the two state model), which model translational bursts, rather than translation of individual molecules, with either a constant size or with a geometrically distributed size governed by a mean parameter, with a frequency corresponding to the transcription rate. Only degradation is modelled for each protein explicitly, while mRNA copy numbers are not explicitly modelled. This maps to the two-state model of transcription in the limit of instantaneous bursts (geometric burst model of transcription) in which the transcripts are mutually repressive agents. Comparing simulated molecule count distributions, the geometric burst mesoscopic model matched well the microscopic model distributions, but the constant burst model did not. A diffusion approximation was then used to solve the master equation for the geometric burst model, in which the discrete-molecule stochastic process is mapped to an approximate continuous-concentration gaussian process described by a Fokker-Planck equation, facilitating an analytical solution for the time-dependent copy number distribution. This was also applied to the microscopic model, but only works when the molecule numbers in the system are high enough that a concentration approximation is appropriate, and thus fails to capture the intrinsic noise statistics in the microscopic model due to the low mRNA numbers. While the diffusion approximation of the geometric burst model performed better than the microscopic model, it was still not faithful to the simulated statistics and distributions and failed to capture the intrinsic noise arising from bursting dynamics. Therefore, an alternative modelling approach which can generate analytical solutions was proposed called the piecewise deterministic Markov process (PDMP), which was shown to outperform the diffusion approximation-based solution of the geometric burst model for biologically relevant regions of parameter space. The PDMP is a mesoscopic, coarse-grained model describing the time-evolution of the two mutually repressive protein numbers using different pairs of deterministic ordinary differential equations (ODEs). There is a pair corresponding to each of three defined states, in which there is no mRNA, or one mRNA of either gene (but not both simultaneously), with stochastic switching between states, going from no mRNA to one copy of either based on the Hill function-defined transcription rate and switching back to no mRNA at the decay rate for the gene. Molecule abundances are treated as a continuous concentration, modelling bursts as exponentially distributed, which is the analogue of discrete-molecule geometric bursting, as previously mentioned. This approximation matched the simulation data better than the geometric burst diffusion approximation for this GRN, which has a pair of mutually repressive genes, in terms of the steady state distribution and noise statistics when the mean burst size is ≥ 5. This is because no gaussian noise is introduced into the system by the PDMP, unlike with diffusion approximations, but will not perform any better with low burst sizes since the stochastic fluctuations associated with the degradation of individual molecules is neglected [51]. This paper demonstrates the development of methods for analytical steady state solutions of GRNs with bursting dynamics which can capture the intrinsic noise faithfully.

Subsequent work has also yielded analytical solutions to time-dependent transcript count distributions with transcriptional bursting using this PDMP approach, in addition to just stationary distributions, and this was applied not only to the two-state model but also to models with > 2 states [120]. Modelling these scenarios with PDMP involves each individual state being modelled as a poisson process, with fluctuations about the state-specific mean, stochastically switching between (gene) states with a PDMP mixing kernel to result in a dynamic poisson mixture model of transcription. This approximate approach allowed for much more efficient calculation of time-dependent mRNA count distributions than standard methods using numerical forward integration of the CME. This modelling approach was embedded within the BayFISH framework, which is a computational pipeline for Bayesian inference of time-resolved transcriptional bursting kinetic parameters designed for smFISH data with time-dependent gene inductions, which utilises a Markov chain Monte Carlo (MCMC) algorithm with a Metropolis-Hastings step to generate posterior distributions to quantify uncertainty in the parameter estimates [121]. This was applied to simulated smFISH data with gene induction at different time points, obtaining parameter estimates for models with different numbers of states before carrying out model selection based on comparison of observed and theoretically predicted mRNA distributions generated by the parameter estimates for the given model. The Kullbeck-Leibler divergence between observed and predicted distributions was used to select optimal models for different datasets, using either the Bayesian or Akaike information criterion to penalise more complex models, which may achieve a lower divergence from the data than a simpler model due to overfitting rather than being a more appropriate model [120, 121]. The development of such modelling and inference approaches have enabled estimation of the parameters governing time-resolved transcriptional bursting dynamics for an individual gene based on smFISH data generated following gene induction at a specific time-point. However, inference and modelling approaches to obtain genome-wide time-resolved bursting dynamics from the wealth of currently available sequencing data are lacking.

1.8 Aims

Here we construct mathematical models to relate observables from 4sU scRNA-seq data to the underlying bursting dynamics and develop a MCMC approach for Bayesian inference of the parameters governing those dynamics. Applying this method to published data from [101] demonstrates that we are able to characterise time-resolved transcriptional bursting dynamics for hundreds of genes in parallel. The 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 [77]. This analysis reveals position-dependent associations between different parameters and HMs only apparent with 4sU scRNA-seq.

2 Methods

2.1 Data processing and analysis

2.1.1 4sU scRNA-seq

The main datasets that were used for parameter inference in this study were produced in Qiu et al 2020 [101], 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 hours before chemical treatment, with GEO sample IDs GSM4512696 and GSM4512697, respectively. K562 cells are a human cancer cell line originally derived from a patient with chronic myeloid leukimia in the 1970s [122], which in [101] were cultured under non-stress conditions permissive to indefinite cell growth. K562 cells have been shown to have the capacity for spontaneous development of features associated with blood cells such as erythrocytes, monocytes and granulocytes, including the production of proteins related to oxygen transport such as haemoglobin [123, 124]. Their robust nature and capacity to indefinitely undergo cell division makes them popular cells lines to work with in experimental biology. Another key feature of these cells in their massive overexpression of Aurora kinases, which in healthy cells are required for mitosis, leading to uncontrolled cell division in these cancerous cells [125]. The aforementioned Qiu datasets are Drop-seq datasets [126] 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 [127] 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 (based on detecting at least six consecutive As in the transcript sequence read). Barcode reads with a missing last base in the cell ID sequence were identified by $\geq 80\%$

of UMI sequences corresponding to the given cell ID having a T as the last base due to the polyT segment, and were repaired by inserting an N at the end of the cell ID before the UMI sequence and deleting the T at the end of the barcode. Then any cell IDs with the same base present at the same position in $\geq 80\%$ of the corresponding UMI sequences were discarded. This deals with the potential occurrence of known drop-seq barcode synthesis errors [126].

The trimmed/repaired reads were then aligned to the primary human genome assembly (GRCh38.p13), the fasta file for which was obtained from gencode [128] (https://www.gencodegenes.org/human/), using bwa to build the genome index and for the actual alignment [129]. 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, taking only reads which overlap with exonic regions, 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. The gene-specific, cell-specific UMI counts were calculated after collapsing all UMI barcodes corresponding to the given cell and gene with a Hamming distance [130] of 1 to account for sequencing errors in the barcodes, recursively finding the barcode with the most other barcodes within 1 Hamming distance and collapsing those barcodes to it, then repeating until every pair of barcodes has a Hamming distance ≥ 2.

Cell selection was then carried out to exclude those cell IDs corresponding to empty droplets by ordering the cell IDs by descending total aligned read pairs and then selecting the top 400 or 795 IDs for the control and 4sU dataset, respectively, as specified in [101] after confirming that these values correspond to the elbow of the cumulative distribution function, thus indicating empty droplets beyond (figures 8 and 9). Cell IDs with < 25 total UMIs were automatically discarded. 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. The code used for pre-processing this dataset as described is available on GitHub (https://github.com/hebenstreitLab/burstMCMCpreprocessing).

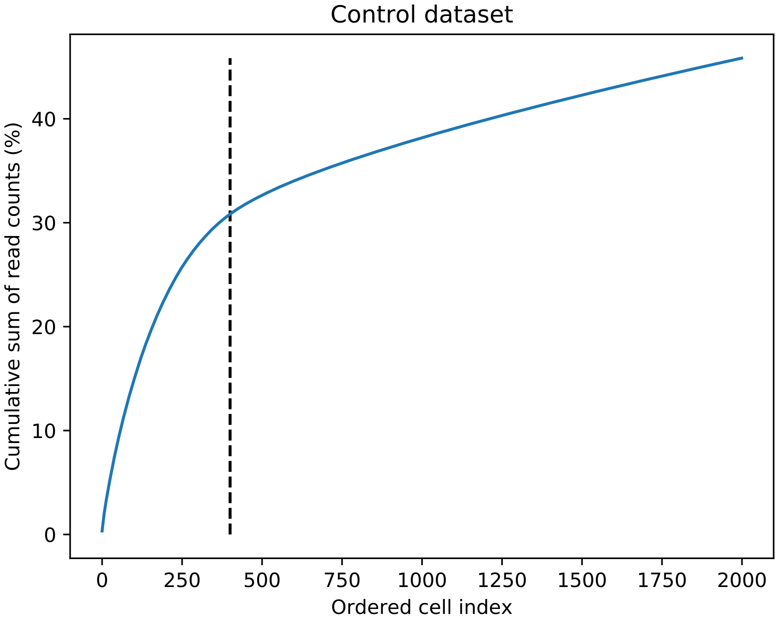


Figure 8: Cumulative sum of total read count of cells in descending order in the Qiu control dataset (no 4sU), with dashed line indicating the number of non-empty, cell-containing droplets (400). Only the top 2000 cell IDs are shown.

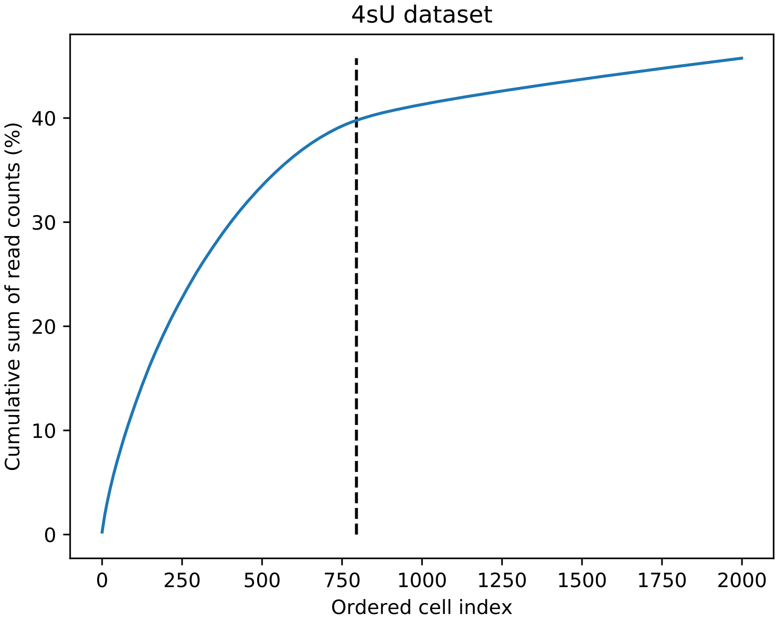


Figure 9: Cumulative sum of total read count of cells in descending order in the Qiu 4sU dataset, with dashed line indicating the number of non-empty, cell-containing droplets (795). Only the top 2000 cell IDs are shown.

2.1.2 ChIP-seq

Publicly available ChIP-seq datasets for eight 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 already been processed with alignment to the hg19 human genome build [131]. 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 [132]. 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 gtf downloaded from Gencode [133]. 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 position through/around the gene across all specified genes, similarly to the metagene analysis described in [134].

The ChIP-seq datasets were aligned to the hg19 reference genome as opposed to the hg38 reference genome, which the 4sU scRNA-seq datasets were aligned to. The main difference between the two references is that hg38 contains alternative sequences for a greater number of specific genomic regions which exhibit strong variation across the human population. Other significant differences include updates to centromeres and the mitochondrial genome. Our analyses focus on genic regions across the genome, which was not the primary focus of sequence updates, meaning that in general the sequences of genes will be a close match from hg19 to hg38, with the main difference being that the coordinates of the gene within the two reference genomes is different. Therefore, although a given read may align to a different position on the two reference genomes, the shift in coordinates of the corresponding gene it belongs to between the hg19 and hg38 gtf files will match that of the read, such that the read maps to the appropriate gene regardless of which reference is used. Even if there is a mismatch (or a small number) between the aligned read and the reference sequence due to the hg19 sequence being outdated, the mapq score will still be high given that the rest of the sequence matches, resulting the correct mapping being retained with high confidence. This is the case for a substitution mismatch, although an insertion or deletion mismatch would prevent proper alignment and mapping. Mitochondrial genes were excluded from analyses involving ChIP-seq data due the strongly differing HM landscape in the mitochondria compared to the nuclear genome. Therefore, the significant updates to the mitochondrial genome in the hg38 reference will not impact the analyses presented here in any way. The main disadvantage of using hg19-aligned ChIP-seq data here instead of hg38-aligned is that there will inevitably be certain outdated genic sequences which deviate from the read sequences corresponding to that loci to such a degree that alignment will fail, thus reducing the sequencing depth of the dataset. However, this is not expected to occur with a high frequency since alignment rates to hg19 and hg38 were found not to be significantly different across multiple alignment tools [135].

2.2 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 [53, 54, 71]. 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

$$
\begin{gathered}
G_{off} \xrightarrow{k_{on}} G_{on} \\
G_{on} \xrightarrow{k_{off}} G_{off} \\
G_{on} \xrightarrow{\beta} G_{on} + RNA \\
RNA \xrightarrow{\delta} \emptyset
\end{gathered}
$$

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 the introduction (section 1.7.3, figure 7). In our scheme no transcription may occur when the gene is in the inactive state, although so-called leakage (low level transcription in the inactive state) may in fact occur, which was shown to result in reduced transcriptional noise, with lower noise being achieved by higher leakage levels for a given overall expression level as previously mentioned (section 1.7) [116, 117]. However, leakage is neglected here in order to make mathematical progress, with the modelling becoming analytically intractable if leakage is included.

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 [51, 114], and when $k_{on} << k_{off}$, which is supported by the parameter estimates reported in [95]. This model simplifies $\kappa = \lim\limits_{k_{off}\to\infty} \frac{1}{(1/k_{on})+(1/k_{off})} = k_{on}$ while b remains finite with $b = \lim\limits_{\beta, k_{off}\to\infty} \frac{\beta}{k_{off}}$ [113].

As outlined in section 1.7.3, previous studies have shown that the gamma distribution may be used to effectively model stochastic gene expression and recover bursting kinetic parameters since it has both a scale and shape parameter, enabling it to capture the skewness introduced by transcriptional and translational bursts in transcript and protein amounts, respectively [58, 115]. The gamma distribution models continuous random variables, making it appropriate to model transcripts and proteins as concentrations for dealing with fluorescence microscopy and flow cytometry data, which produces continuous-valued data. However, our discrete transcript molecule scheme, while analogous and also capable of fully capturing the skew introduced by bursting, is better suited than the gamma distribution for dealing with sequencing data, which produces exact discrete UMI and read count values rather than continuous fluorescence values.

2.2.1 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 (see section 2.4) of that cell, α, in a similar fashion to the technical noise model outlined in [136]. 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 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 a Poisson-Beta distribution, which under instantaneous bursting becomes a negative binomial distribution [54, 114, 112, 113]

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

where

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

The derivation of equation 1 is shown in section 2.6. 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 [136]

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

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 equations 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 equation 1, which avoids summing to ∞, achieving a finite state projection (FSP) [137, 138] with an error of 0.0001. This leads us to the likelihood function of model 1 by taking the product of equation 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 $L = (l_1, \dots, l_k)$, 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

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

which we achieve through MCMC sampling.

2.2.2 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. 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 $1 - F_{Exp}(X \leq t|\delta) = f_{Pois}(0|\tau)$ since it corresponds to having a degradation wait time longer than the pulse duration, which is an exponential random variable. 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 (equation 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

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

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 [114] 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, defined by a power series which terminates in our case since we have a non-positive integer, $-n$, as the first parameter reducing the series to a polynomial generally defined as

$${}_2F_1\left(-n, x, y; z\right) = \sum_{k=0}^{n} (-1)^k \frac{\Gamma(n+1)}{\Gamma(n-k+1)} \frac{(x)_k}{(y)_k} \frac{z^k}{k!}$$

where $(x)_k$ represents the Pochhammer symbol for the rising factorial such that $(x)_k = \Gamma(x+k)/\Gamma(x)$. The derivation of equation 9 is shown in section 2.6. 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 equations 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 [99] and compounding with equation 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 (see section 2.5). We may now model the cell-specific T>C conversion rate for the given gene by compounding equations 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 equation 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 equation 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 equations 4 and 15

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

As in equation 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 equation 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.

2.2.3 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

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

This is similar to equation 12 but is independent of the total transcript count, m, and is therefore not cell specific. We can apply equation 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 equations 14 and 15. We define the final likelihood function of model 3 as the product of equations 4 and 19.

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

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

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

2.3 Markov chain Monte Carlo algorithm

MCMC was employed in order to sample from the posterior distributions outlined in equations 5, 17 or 21 using a Metropolis-adjusted Langevin algorithm (MALA) within a Gibbs sampler, which simulates a Markov chain using Langevin dynamics [139] and corrects the Euler-Maruyama integration error with an accept-reject step as with the Metropolis-Hastings algorithm [140]. The code for carrying out bursting parameter inference with this algorithm is available on GitHub (https://github.com/hebenstreitLab/burstMCMC) which can be downloaded and installed as an R package (called "burstMCMC"). The Markov 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 (see section 2.5), 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 $\theta^{(1)}$ with the above sampling approach 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

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

where

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

with support $[y, z]$. This approach exploits the easily accessed prior information to give rough parameter estimates where possible and initialise the Markov chain closer to the target density than with random sampling. The expression level estimate, $\hat{\mu}$, is very robust, which is why it is a fixed value for initialisation. Conversely, the estimate for transcript lifetime, $\hat{\gamma}$, is less robust, particularly for very high/low values (hence it is truncated at either extreme), so a normal distribution is chosen to to avoid fixing the initialisation at implausible values (which would violate the prior distributions). We have limited prior information with which to estimate burst rate, a, so the log-uniform distribution from 1 to 10 is chosen for initialisation to provide unbiased samples across an order of magnitude which ranges from high burstiness (but not extreme) to moderate/lower burstiness (but still with super-poissonian noise). If the burst rate, a, is initialised at too high a value, the likelihood surface will become flat and the Markov chain will struggle to converge down to the target value. Extremely low (close to zero) burst rates correspond to extremely high noise and burst size, making very high transcript counts possible and causing the state space being summed over in the likelihood function calculations to explode, potentially slowing down the algorithm significantly. That is why only burst rate, a, values between 1 and 10 are accepted for initialisation, with the Markov chain able to move outside of that range after the first step.

The Markov chain proceeds through three dimensional parameter space with $\theta = (\mu, a, \gamma)$. This parameterisation was chosen for Markov chain progression to minimise correlations between parameters and because they generally do not have values $<< 1$, being > 1 in most cases, which helps avoid potential numerical issues. At each step, j, in the chain, the next step is sampled by proposing jumps to new positions in parameter space from the current position, so choosing a parametrisation which avoids values $<< 1$ prevents frequent proposals to negative (unsupported) values. The classic Metropolis-Hastings algorithm [140] 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 [141]. 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 [139]. 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 Wiener process (standard Brownian motion) diffusion term, W, and a drift term determined by the vector gradient, ∇, of the log-density of the posterior, $\pi(\theta) \propto P(X|\theta)P(\theta)$, with respect to θ evaluated at θ_t. The Fokker-Planck equation for the time-evolution of the probability density is

$$\frac{\partial}{\partial t}[p(\theta, t)] = -\frac{\partial}{\partial \theta}[p(\theta, t)\nabla \log \pi(\theta)] + \frac{\partial^2}{\partial \theta^2}[p(\theta, t)] \tag{23}$$

with the diffusion coefficient being independent of θ. Since we do not have an analytical solution for $\nabla \log \pi(\theta)$ in equation 22, 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 [142, 143], 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 equation 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

$$A = \min\left(1, \frac{\pi(\theta_1^{(j+1)}, \ldots, \theta_{k-1}^{(j+1)}, \theta_k^{(*)}, \theta_{k+1}^{(j)}, \ldots, \theta_d^{(j)})}{\pi(\theta_1^{(j+1)}, \ldots, \theta_{k-1}^{(j+1)}, \theta_k^{(j)}, \ldots, \theta_d^{(j)})}\right) \tag{24}$$

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 equations 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 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 [141]

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

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 [139]. At step 1, we initialise $\nabla = 0$, $S = \theta^{(1)}/100$ and $\eta = 0.1$.

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 level each step takes longer but the stronger evidence means that less 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. This allows for a discarded "burn-in" period of 500 or 2500 steps, which is demonstrated to be sufficient to allow the Markov chain to reach equilibrium and achieve convergence in the vast majority of cases. We carried out (model 2) inference on simulated data for thousands of genes to validate the performance of the algorithm, as described in section 2.9. By showing the Markov chain traces for a random sample of 100 simulated genes we demonstrate that the aforementioned burn-in periods allow convergence to equilibrium for all three of the chosen parameters (μ, a and γ) that our Markov chains move in (figure 10). This is also demonstrated for 23 genes simulated with $\mu \geq 1000$ (figure 11), which all have the shorter burn-in period (500 steps). Finally, this was shown for 100 randomly sampled genes for which we have relatively high confidence in their parameter estimates, corresponding to those genes with coefficient of variation (CV) < 0.45 associated with the posteriors generated for each of the three parameters, which have a mixture of the longer and shorter burn-in periods (figure 12).

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 equation 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. The inference algorithm was computationally implemented using custom scripts written in the Python [127] and R [144] programming languages.

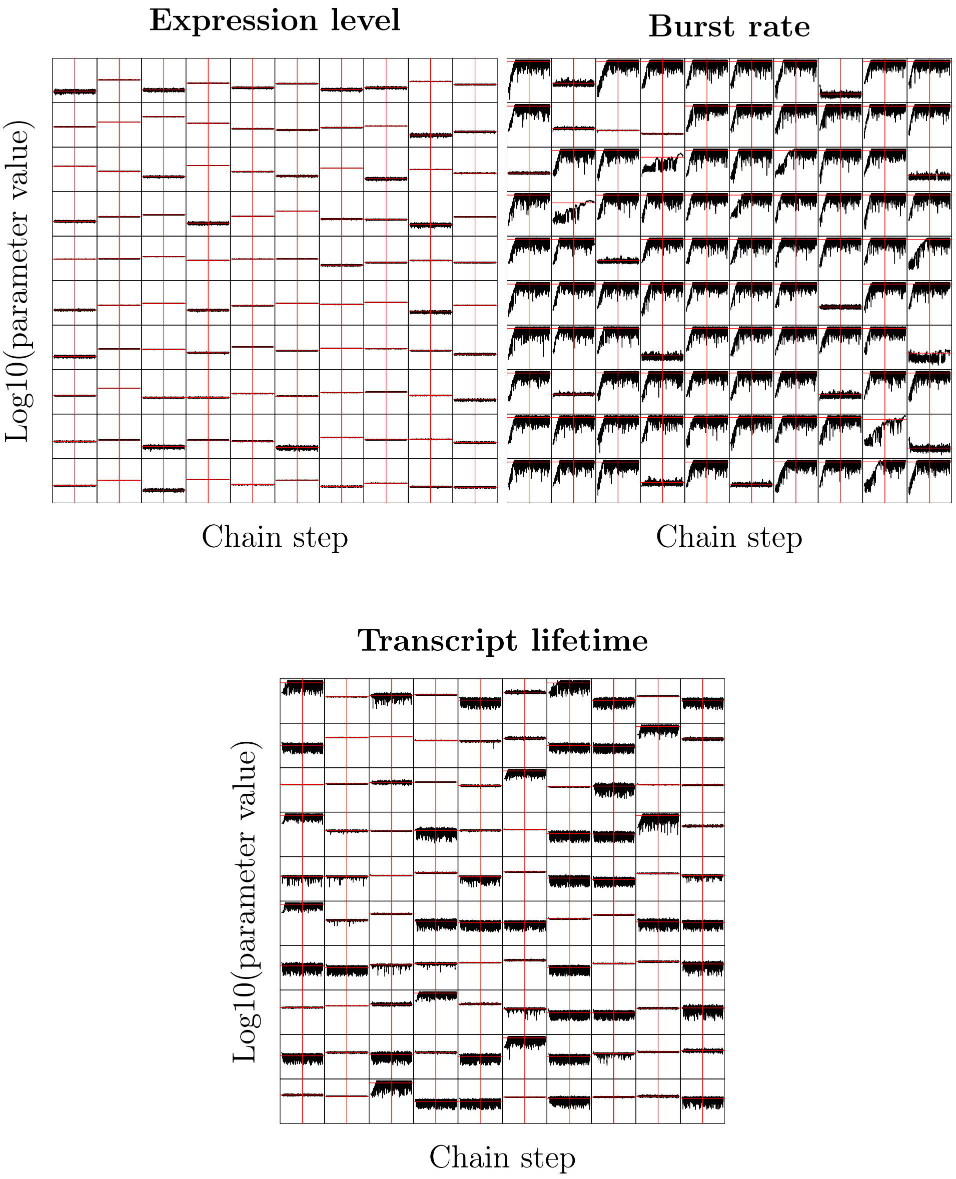


Figure 10: Markov chain traces for 100 randomly sampled genes with data simulated using known parameter values (red horizontal lines), with inference being carried out using model 2 (or 3 if required). Convergence of the Markov chains to equilibrium is shown to be achieved by the time the end of burn-in (the initial steps discarded to allow convergence, shown as the red vertical line) is reached for all three parameters (expression level, burst rate and transcript lifetime). The Markov chain is run for 1500 total steps with a 500 step burn-in or 5000 total steps with a 2500 step burn-in for genes estimated to have an expression level $\geq$ 1000 or other genes, respectively, with all shown belonging to the latter. The y-axes limits are from -2 to 5 (prior upper bound).

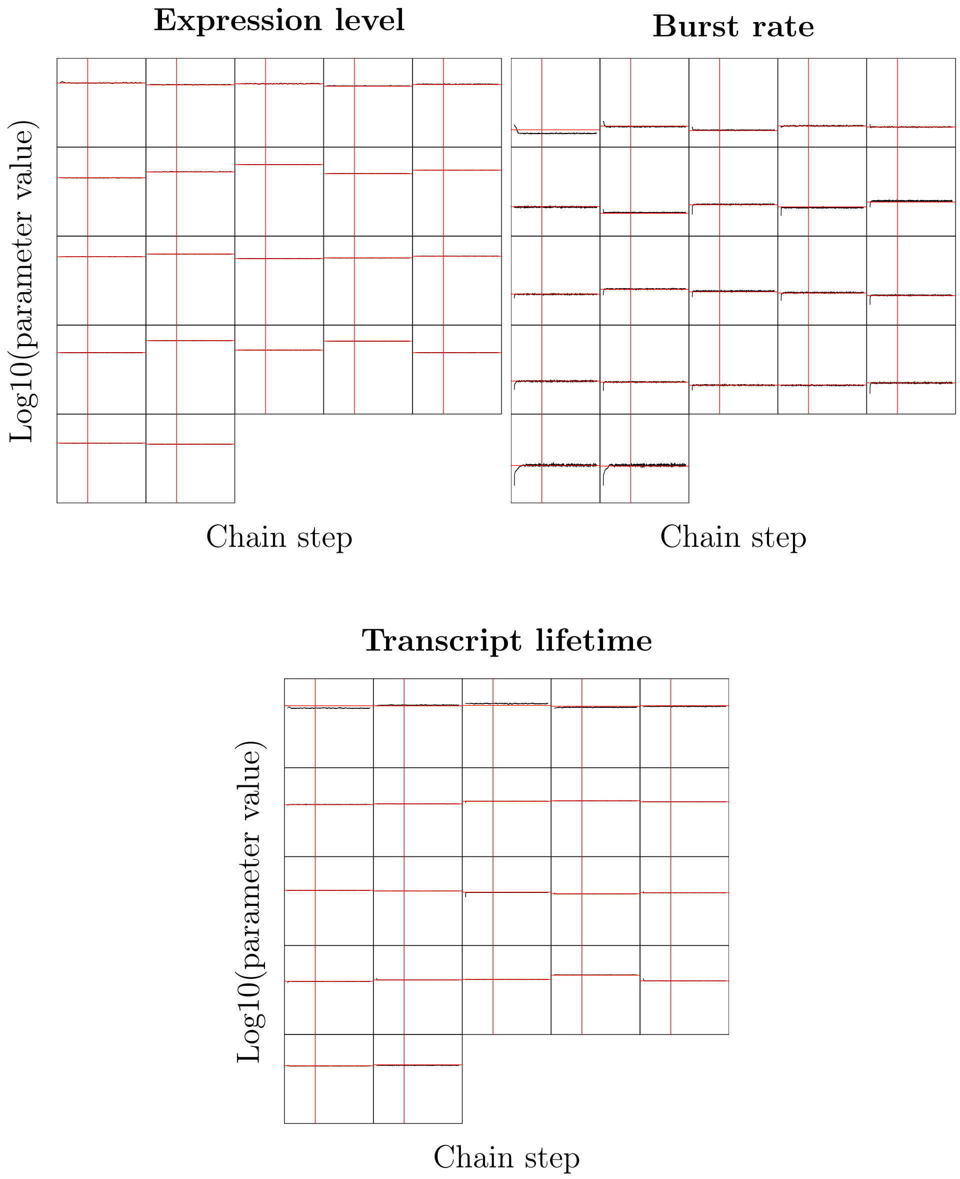


Figure 11: Markov chain traces for 23 genes simulated with high expression levels ($\mu \geq 1000$) using known parameter values (red horizontal lines), with inference being carried out using model 2 (or 3 if required). Convergence of the Markov chains to equilibrium is shown to be achieved by the time the end of burn-in (the initial steps discarded to allow convergence, shown as the red vertical line) is reached for all three parameters (expression level, burst rate and transcript lifetime). The Markov chain is run for 1500 total steps with a 500 step burn-in or 5000 total steps with a 2500 step burn-in for genes estimated to have an expression level $\geq$ 1000 or other genes, respectively, with all shown belonging to the former. The y-axes limits are from -1 to 5 (prior upper bound).

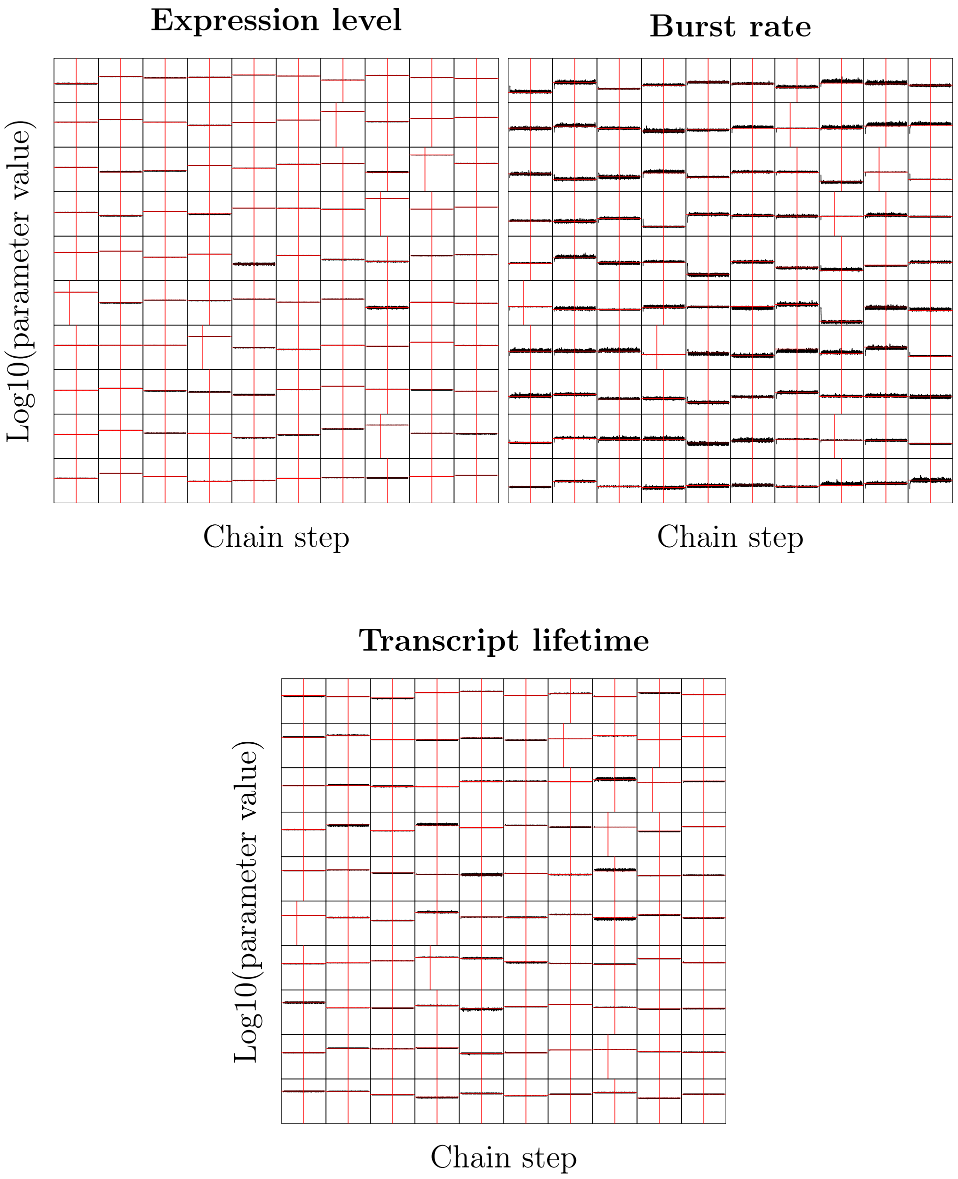


Figure 12: Markov chain traces for 100 randomly sampled genes with high confidence parameter estimates ($CV < 0.45$) with data simulated using known parameter values (red horizontal lines), with inference being carried out using model 2 (or 3 if required). Convergence of the Markov chains to equilibrium is shown to be achieved by the time the end of burn-in (the initial steps discarded to allow convergence, shown as the red vertical line) is reached for all three parameters (expression level, burst rate and transcript lifetime). The Markov chain is run for 1500 total steps with a 500 step burn-in or 5000 total steps with a 2500 step burn-in for genes estimated to have an expression level $\geq$ 1000 or other genes, respectively, with a variety of the former and latter shown. The y-axes limits are from -2 to 5 (prior upper bound).

2.4 Cell-specific capture efficiencies

Our models require the capture efficiency, α, (proportion of transcripts from each cell with at least 1 corresponding read) of each cell to be known. This necessitates the use of RNA spike-in probes, in which a known quantity of material is added to each cell and the proportion of molecules detected in the sequencing gives α. Spike-ins were not used in the Qiu datasets, but capture efficiencies may be inferred by using data from Klein et al 2015 [145], which has cell-matched (K562) scRNA-seq data (with GEO sample ID GSM1599501) which does contain ERCC spike-in probes. We construct a Bayesian mathematical model to obtain the probability distribution of the capture efficiencies in the 4sU Qiu dataset, α_q based on the Klein data, under the assumption that since both datasets were produced with K562 cells, the underlying probability distribution of the total transcript count in each cell, m, is the same for both datasets. According to [136] the true number of spike-in molecules loaded to each cell, x, may be modelled as a poisson random variable with rate based on the expected number of molecules loaded per cell, λ, (12467.64 in the case of the Klein dataset)

$$x \sim Pois(\lambda)$$

The capture efficiency of each of the 953 cells in the Klein dataset, α_k, is then

$$\alpha_k \sim Beta(y, x - y)$$

where y represents the total number of spike-in molecules detected in the given cell. The total number of transcripts present in each cell in Klein is modelled as

$$m - l_k \sim NBin(l_k, 1 - \alpha_k)$$

where l_k represents the total UMI counts across all genes in Klein for the given cell, using the negative binomial parametrisation defined for equation 1. The capture efficiency of each cell in Qiu, α_q, may then be obtained using the total number of UMIs in the given cell across all genes, l_q,

$$\alpha_q \sim Beta(l_q, m - l_q)$$

The probability density function for each cell in Qiu is then solved by numerically integrating the above distributions through random number generation to obtain

$$\begin{gathered} P(\alpha_q|\lambda, y, l_k, l_q) = \\ \sum_{m=l_q}^{\infty} f_{Beta}(\alpha_q|l_q, m - l_q) \frac{1}{N} \sum_{i=1}^{N} \int f_{NBin}(m - l_{k,i}|l_{k,i}, 1 - \alpha_{k,i}) \\ \sum_{x=y_i}^{\infty} [f_{Beta}(\alpha_{k,i}|y_i, x - y_i) f_{Pois}(x|\lambda)] \, d\alpha_{k,i} \end{gathered} \tag{25}$$

where N is the number of cells in Klein, i refers to the ith cell of Klein and

$$f_{beta}(\alpha_q|l_q, m - l_q) = \frac{\alpha_q^{l_q}(1 - \alpha_q)^{m - l_q}}{B(l_q, m - l_q)}$$

with

$$B(l_q, m - l_q) = \frac{\Gamma(l_q)\Gamma(m - l_q)}{\Gamma(m)}$$

Estimates may be derived from $P(\alpha_q|\lambda, y, l_k, l_q)$ with $E[\alpha_q]$ and confidence may be quantified with $E[\alpha_q^2]$. Figure 13 indicates high confidence in our estimates through the low CV, while the estimated capture efficiencies for Qiu are lower than for Klein, at around 0.02 on average. A quick, simple method for calculating α_q estimates without quantifying confidence is as follows

$$\alpha_q = l_q/\hat{m}$$

where

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} l_{k,i}\lambda/y_i$$

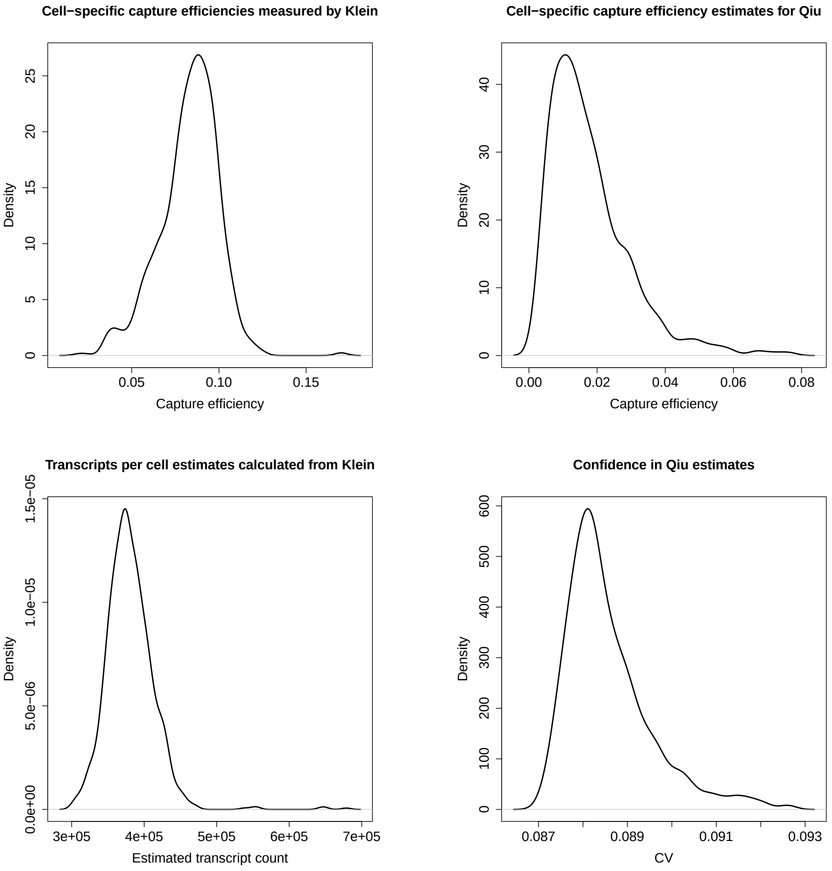


Figure 13: Density plots of the capture efficiencies measured for the Klein dataset, the total transcript content per cell estimated from Klein based on measured capture efficiencies and total UMI counts per cell, the capture efficiencies estimated for the Qiu dataset based on total UMI counts per cell and total transcript content estimated from Klein, and the confidence in those estimates for Qiu as represented by the CVs obtained with our Bayesian capture efficiency model.

2.5 Conversion rates

Models 2 and 3 also require the gene-specific background and gene-invariant 4sU-mediated T>C conversion rates to be known, λ_s and λ_n, respectively. As previously mentioned (in section 2.2.2), λ_s is defined as the proportion of genomic Ts in all reads and all selected (see section 2.1.1) cells that appeared as Cs in the control dataset for the given gene. Therefore, the conversion rate observed in the 4sU dataset corresponds to $\lambda_s + \lambda_n$. The conversion rates of all genes for which we have high confidence in the rate estimate in both datasets are shown in figure 14, which was 6259 genes. Confidence is obtained by modelling the T>C rate, λ, as

$$\lambda \sim Beta(C, T - C)$$

classing those with a resulting CV $< 10^{-0.5}$ in both datasets as having high confidence. We expect the rate in the 4sU dataset to be at least as large as in the control, hence genes tend to appear on the diagonal or above it. Genes with higher turnover are expected to appear further above the diagonal while those with low turnover are expected to appear closer to it. The curve along the top of the plot represents λ_s (x-axis) added to our estimate of λ_n. λ_n is estimated by first assuming that all reads correspond to new transcripts (synthesised during the pulse) and then calculating

$$p = 1 - F_{Bin}(C - 1 | T, \lambda_s + \lambda_n) = F_{Bin}(T - C | T, 1 - \lambda_s - \lambda_n)$$

for each gene so that $\mathbf{p} = (p_1, \ldots, p_{6259})$ where

$$F_{Bin}(C | T, \lambda_s + \lambda_n) = \sum_{i=0}^{\lfloor C \rfloor} \binom{T}{i} (\lambda_s + \lambda_n)^i (1 - \lambda_s - \lambda_n)^{T-i}$$

The estimate for λ_n is then the minimum value for which $\sum_g [p_g < 10/6259] < 10$. With this approach, $\lambda_n \approx 0.07547$.

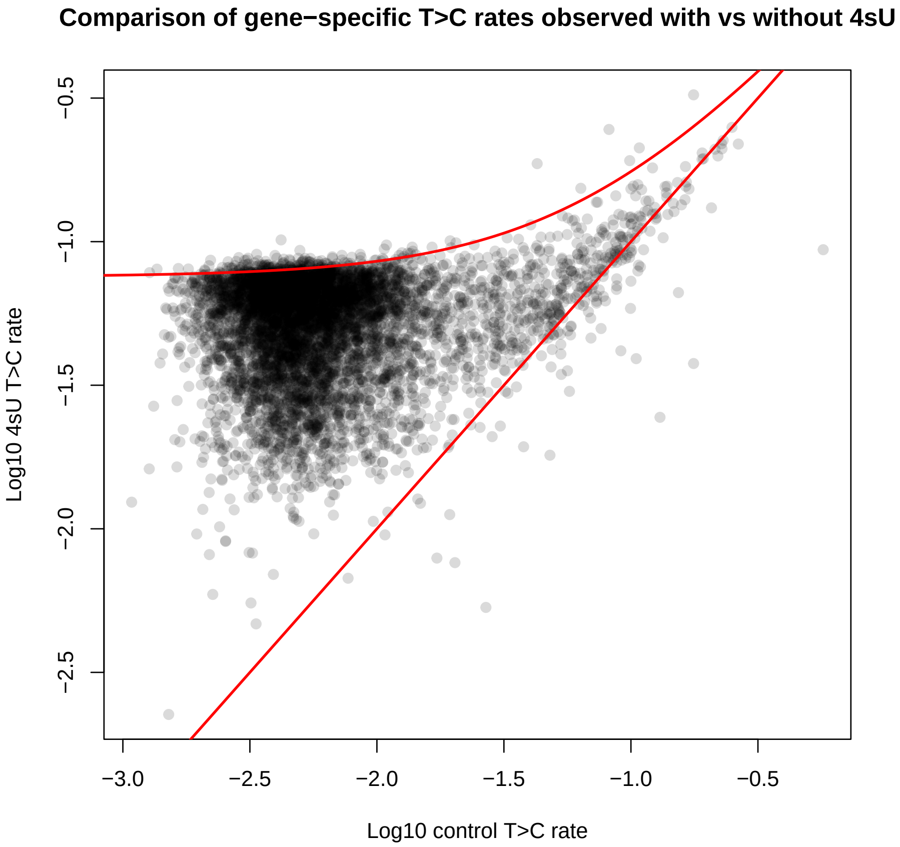


Figure 14: Gene-specific T>C conversion rates in the control vs 4sU Qiu datasets for 6259 genes for which we have high confidence in the observed T>C rate in both datasets. The lower red line represents the background T>C mutation rate, λ_s, while the upper one represents the maximum possible T>C rate in the 4sU dataset, $\lambda_s + \lambda_n$, which is the sum of the observed gene-specific background rate and the estimated global 4sU-mediated T>C rate.

2.6 Derivations

In this section our mathematical model of transcription will be formally defined with its chemical master equation (CME), from which we will derive the partial differential equation (PDE) for the corresponding probability generating function (PGF). This PDE will the be solved in order to obtain both the steady state and time-dependent transcript count PMFs of equations 1 and 9, respectively.

2.6.1 Chemical master equation

Our biological model corresponds to the infinitesimal limit of the standard two-state model, converting it into a one-state model with poissonian arrival of geometric bursts, which is described in [51, 113]. In this scheme, only two events are possible; the exponential decay of a single transcript at rate δ, or the generation of a geometrically distributed number of transcripts through an instantaneous transcriptional burst at rate κ with average size b. In this section we derive the CME by following the approach of [113], defining the probability that the next event during a burst is either the synthesis of a transcript or the end of the burst as ξ and Θ, respectively, such that $\xi = \frac{b}{1+b}$, $\Theta = \frac{1}{1+b}$ and $b = \xi/\Theta$. For convenience now we write the geometric PMF as $g(m) = \xi^m\Theta = f_{Geom}(m|\Theta)$. The probability of having m transcripts at time $t+\Delta t$, $P_m(t+\Delta t)$ is then described by the probability of flowing into state m from other states or of remaining in state m minus the probability of flowing out of state m

$$\begin{aligned}P_m(t+\Delta t) =& P_m(t) + \delta(m+1)\Delta t P_{m+1}(t) - \delta m \Delta t P_m(t)\\ &+\kappa\Delta t\sum_{n=0}^{m-1} g(m-n)P_n(t) - \kappa\Delta t\sum_{n=m+1}^{\infty} g(n-m)P_m(t)\end{aligned} \tag{26}$$

From this we obtain the chemical master equation (CME) describing the time-evolution of the probability distribution of the state of the system by subtracting $P_m(t)$ from both sides, dividing by Δt and taking $\lim\limits_{\Delta t\to 0}$

$$\begin{aligned}\frac{dP_m(t)}{dt} =& \delta(m+1)P_{m+1}(t) + \kappa\sum_{n=0}^{m-1} g(m-n)P_n(t)\\ &-\delta m P_m(t) - \kappa\sum_{n=m+1}^{\infty} g(n-m)P_m(t) \qquad (27)\end{aligned}$$

We can then normalise the timescale of our CME into units of transcript lifetimes by dividing both sides by δ, since $a = \kappa/\delta$ and $\tau = t\delta$

$$\begin{aligned}\frac{dP_m(\tau)}{d\tau} =& (m+1)P_{m+1}(\tau) + a\sum_{n=0}^{m-1} g(m-n)P_n(\tau)\\ &-mP_m(\tau) - a\sum_{n=m+1}^{\infty} g(n-m)P_m(\tau) \qquad (28)\end{aligned}$$

Notice that bursts with size of zero are excluded from the burst terms in equation 28 as they do not affect the rate of change of $P_m(\tau)$. We can reindex both terms more conveniently to include burst sizes of zero since the zero instances from the two terms cancel each other out, in which $m = n$. Then we have

$$a\sum_{n=0}^{m-1} g(m-n)P_n(\tau) \to a\sum_{n=0}^{m} g(m-n)P_n(\tau)$$

and

$$a\sum_{n=m+1}^{\infty} g(n-m)P_m(\tau) \to a\sum_{n=m}^{\infty} g(n-m)P_m(\tau) = aP_m(\tau)$$

since we sum across all possible burst sizes in the second instance and $\sum_{m=0}^{\infty} g(m) = 1$. This leads us to the final form of our CME

$$\frac{dP_m(\tau)}{d\tau} = (m+1)P_{m+1}(\tau) - mP_m(\tau) - aP_m(\tau) + a\sum_{n=0}^{m} g(m-n)P_n(\tau) \qquad (29)$$

2.6.2 Probability generating function

In order to solve the CME of equation 29 we must first derive the corresponding probability generating function (PGF), which we achieve in this section by following the approach of [113]. The PGF is defined as

$$G(\tau, z) = \sum_{m=0}^{\infty} z^m P_m(\tau) \tag{30}$$

such that by multiplying equation 29 by z^m and summing over m with $\sum_m \equiv \sum_{m=0}^{\infty}$ we have

$$\begin{aligned}\frac{dG(\tau, z)}{d\tau} =& \sum_m z^m (m+1) P_{m+1}(\tau) - \sum_m z^m m P_m(\tau) \\ &+ a \sum_m z^m \sum_{n=0}^{m} g(m-n) P_n(\tau) - aG(\tau, z)\end{aligned} \tag{31}$$

The aim now is to write this entirely in terms of $G(\tau, z)$ and z rather than $P_m(\tau)$ and m. Beginning with the second term, notice that

$$z^m m = z \frac{dz^m}{dz}$$

and since $P_m(\tau)$ is independent of z we have

$$\sum_m z^m m P_m(\tau) = z \frac{d}{dz} \left(\sum_m z^m P_m(\tau) \right) = z \frac{dG(\tau, z)}{dz} \tag{32}$$

Turning to the first term, we define $l = m + 1$ and thus

$$\sum_m z^m (m+1) P_{m+1}(\tau) = z^{-1} \sum_{l=1}^{\infty} z^l l P_l(\tau) = z^{-1} \sum_{l=0}^{\infty} z^l l P_l(\tau)$$

because we multiply by zero when $l = 0$. Therefore, due to the relation shown in equation 32 we have

$$\sum_m z^m (m+1) P_{m+1}(\tau) = \frac{dG(\tau, z)}{dz} \tag{33}$$

Finally coming to the third term, we first re-order the summations with

$$\sum_{m=0}^{\infty} \sum_{n=0}^{m} = \sum_{n=0}^{\infty} \sum_{m=n}^{\infty}$$

and substitute $\xi^m \Theta$ for $g(m)$ so that

$$a \sum_m z^m \sum_{n=0}^{m} g(m-n) P_n(\tau) = a \sum_{n=0}^{\infty} \sum_{m=n}^{\infty} z^m \xi^{m-n} \Theta P_n(\tau)$$

which can be rearranged to

$$a\Theta \sum_{n=0}^{\infty} \xi^{-n} P_n(\tau) \sum_{m=n}^{\infty} (z\xi)^m$$

Then redefining $l = m - n$ we re-write the last sum

$$\sum_{m=n}^{\infty} (z\xi)^m = \sum_{l=0}^{\infty} (z\xi)^{l+n} = (z\xi)^n \sum_{l=0}^{\infty} (z\xi)^l$$

When solving $G(\tau, z)$ to find the PMF of the transcript count it will be evaluated at $z = 0$ or $z = 1$, while $0 < \xi < 1$ by definition, meaning that $|z\xi| \leq 1$ so the resulting sum is a terminating geometric series and therefore has a closed form solution

$$(z\xi)^n \sum_{l=0}^{\infty} (z\xi)^l = \frac{(z\xi)^n}{1 - z\xi}$$

which upon substituting back in gives

$$a \sum_{m} z^m \sum_{n=0}^{m} g(m-n) P_n(\tau) = \frac{a\Theta}{1-z\xi} \sum_{n=0}^{\infty} z^n P_n(\tau) = \frac{a\Theta}{1-z\xi} G(\tau, z) \tag{34}$$

Substituting equations 32, 33 and 34 back into equation 31 results in a PDE for the PGF

$$\frac{\partial G(\tau, z)}{\partial \tau} = \frac{\partial G(\tau, z)}{\partial z} - z\frac{\partial G(\tau, z)}{\partial z} + \frac{a\Theta}{1-z\xi} G(\tau, z) - aG(\tau, z)$$

which we can rearrange to

$$\frac{(z-1)\xi}{1-z\xi} aG(\tau, z) = \frac{\partial G(\tau, z)}{\partial \tau} + (z-1)\frac{\partial G(\tau, z)}{\partial z} \tag{35}$$

since

$$\frac{\Theta}{1-z\xi} - 1 = \frac{\Theta - 1 + z\xi}{1-z\xi} = \frac{(z-1)\xi}{1-z\xi}$$

2.6.3 Steady state solution

Now that we have derived the PDE for the PGF in equation 35, we can solve it to complete our derivation of the steady state transcript count PMF described by equation 1 from the CME. We accomplish this in this section by following the approach of [113]. Under steady state, we may substitute

$$\frac{\partial G(\tau, z)}{\partial \tau} = 0$$

into our PDE, which reduces it to the solvable ordinary differential equation (ODE) which is no longer time-dependent

$$\frac{\xi}{1-z\xi} aG(z) = \frac{dG(z)}{dz}$$

which can be directly integrated

$$\int \frac{dG(z)}{G(z)} = \int \frac{a\xi dz}{1-z\xi}$$

giving

$$\log G(z) = -a\log(1-z\xi) + c$$

and finally

$$G(z) = c(1-z\xi)^{-a} \tag{36}$$

Referring back of the definition in equation 30 a normalisation constraint is apparent

$$G(z=1) = \sum_m P_m = 1$$

Therefore we may substitute $G(z=1) = z = 1$ into equation 36 to find c

$$c = (1-\xi)^a$$

which leads to our final steady state solution

$$G(z) = \left(\frac{1-\xi}{1-z\xi}\right)^a \tag{37}$$

The PGF has the property that the PMF the system state is given by the mth coefficient of its Taylor expansion about $z=0$, such that

$$P_m = \frac{1}{m!}\frac{d^m G(z)}{dz^m}\bigg|_{z=0} \tag{38}$$

because taking the mth derivative of the PGF at $z=0$ results in

$$\frac{d^m G(z)}{dz^m}\bigg|_{z=0} = \sum_n \frac{n!}{(n-m)!} z^{n-m} P_n\bigg|_{z=0}$$

which is zero for every term except where $m = n$ since $0^0 = 0! = 1$ so that we have

$$\left.\frac{d^m G(z)}{dz^m}\right|_{z=0} = n! P_n$$

Differentiating equation 37 in this manner results in

$$\left.\frac{d^m G(z)}{dz^m}\right|_{z=0} = \frac{\Gamma(m+a)}{\Gamma(a)} \Theta^a \xi^m$$

and substituting this into the Taylor expansion coefficient described in equation 38 and expressing in terms of b leads to

$$P_m = \frac{\Gamma(m+a)}{\Gamma(m+1)\Gamma(a)} \left(\frac{1}{1+b}\right)^a \left(\frac{b}{1+b}\right)^m \tag{39}$$

finally arriving at our negative binomial PMF for the steady state transcript count used in equation 1.

2.6.4 Time-dependent solution

Returning to equation 35, we can solve for the time-dependent transcript count solution of equation 9 as in [114, 146] by keeping the time-derivative term. We initially make progress towards this solution by following the approach of [146]. First of all we transform $z \to z + 1$ by redefining a PGF as

$$G(\tau, z) = \sum_{m=0}^{\infty} (z+1)^m P_m(\tau) \tag{40}$$

which also allows us to conveniently express the PDE in terms of b rather than ξ, remembering that $b = \xi/\Theta$, as

$$\frac{abz}{1-bz} G(\tau, z) = \frac{\partial G(\tau, z)}{\partial \tau} + z \frac{\partial G(\tau, z)}{\partial z} \tag{41}$$

since with our transformation

$$z - 1 \to z$$

and

$$\frac{(z-1)\xi}{1-z\xi} \to \frac{\Theta^{-1}}{\Theta^{-1}}\frac{z\xi}{\Theta - z\xi} = \frac{bz}{1-bz}$$

Now we define

$$F(\tau, z) = (1-bz)^{a} G(\tau, z) \tag{42}$$

which still satisfies the form of equation 41 [146]

$$F(\tau, z) = \frac{1-bz}{abz}\left(\frac{\partial F(\tau, z)}{\partial \tau} + z\frac{\partial F(\tau, z)}{\partial z}\right) \tag{43}$$

With a change of variables [146, 147] where $\omega = \log(z) - \tau$ and $\zeta = z$, equation 43 is converted into a solvable ODE

$$F(\omega, \zeta) = \frac{1-b\zeta}{ab\zeta}\left(\zeta\frac{\partial F(\omega, \zeta)}{\partial \zeta}\right) = \frac{1-b\zeta}{ab}\frac{\partial F(\omega, \zeta)}{\partial \zeta} \tag{44}$$

which can be directly integrated

$$\int \frac{dF(\omega, \zeta)}{F(\omega, \zeta)} = \int \frac{abd\zeta}{1-b\zeta}$$

leading to

$$c + \log(F(\omega, \zeta)) = -a\log(1-b\zeta)$$

and finally

$$cF(\omega, \zeta) = (1-b\zeta)^{-a} \tag{45}$$

where c represents an initial condition at $\tau = 0$, with $P_{m=0}(\tau = 0) = 1$, which is set as $c = (1-bz_0)^{-a}$ in order to also convert back to our original function $F \to G$ based on equation 42, with z_0 representing the value of z when $\tau = 0$. Before substituting back in, z_0 is transformed to be expressed in terms of our new variables subject to initial conditions. When $\tau = 0$, $\omega = \log(z) + 0 = \log(z)$, so that $z_0 = e^{\omega} = ze^{-\tau}$. Finally, we can substitute c back into equation 45 and write the solution in terms of our original variables,

τ and z, as

$$G(\tau, z) = \left(\frac{1 - bze^{-\tau}}{1 - bz}\right)^a \tag{46}$$

Remembering that we previously made the transformation $z \to z + 1$ using the PGF defined in equation 40, we can now apply the reverse transformation $z \to z - 1$ to convert the solution back in terms of our original PGF defined in equation 30

$$G(\tau, z) = \left(\frac{1 - b(z-1)e^{-\tau}}{1 - bz + b}\right)^a \tag{47}$$

Notice that our choice of c has satisfied both our normalisation and initial condition constraints, $F(\tau, z = 1) = 1$ and $F(\tau = 0, z = 0) = 1$, respectively. We reach our solution by following the approach of [114] throughout the rest of this section, re-writing equation 47 as

$$G(\tau, z) = \left(\frac{1 + be^{-\tau}}{1 + b}\right)^a \times \frac{\left(1 - \frac{b}{1+b}z\right)^{-a}}{\left(1 - \frac{b}{e^{\tau}+b}z\right)^{-a}} \tag{48}$$

since

$$1 - \frac{b}{1+b}z = \frac{1 + b - bz}{1 + b} = \frac{1 - bz + b}{1 + b}$$

with the numerator and denominator corresponding to the denominators of equations 47 and 48, respectively, while

$$1 - \frac{b}{e^{\tau} + b}z = \frac{e^{-\tau}}{e^{-\tau}}\frac{e^{\tau} + b - bz}{e^{\tau} + b} = \frac{1 - b(z-1)e^{-\tau}}{1 + be^{-\tau}}$$

with the numerator and denominator corresponding to the numerators of equations 47 and 48, respectively. At this point, it is possible to apply the Taylor expansion described in equation 38 to our solved PGF from equation 48 to generate the time-dependent transcript count PMF with zero initial transcripts using the identities

$$\frac{\partial^m}{\partial z^m}[1-qz]^{-a}\bigg|_{z=0} = \frac{\Gamma(a+m)}{\Gamma(a)}q^m \tag{49}$$

and

$$\frac{\partial^m}{\partial z^m}\frac{x(z)}{y(z)} = m!\sum_{k=0}^{m}\frac{\partial^{m-k}}{\partial z^{m-k}}x(z)\sum_{j=0}^{k}\frac{(-1)^j(k+1)y(z)^{-j-1}}{(j+1)!(m-k)!(k-j)!}\frac{\partial^k}{\partial z^k}y(z)^j \tag{50}$$

wherein we substitute the numerator and denominator of equation 48 for $x(z)$ and $y(z)$ in equation 50, respectively, so that q in equation 49 corresponds to either $\frac{b}{1+b}$ or $\frac{b}{e^\tau+b}$. According to equations 38, 49 and 50 we obtain

$$\begin{aligned}P_m(\tau) = &\left(\frac{1+be^{-\tau}}{1+b}\right)^a\sum_{k=0}^{m}\frac{\Gamma(a+m-k)}{\Gamma(a)}\left(\frac{b}{1+b}\right)^{m-k}\\ &\times\sum_{j=0}^{k}\frac{(-1)^j(k+1)}{(j+1)!(m-k)!(k-j)!}\frac{\Gamma(aj+k)}{\Gamma(aj)}\left(\frac{b}{e^\tau+b}\right)^k\end{aligned} \tag{51}$$

which can be simplified since

$$\sum_{j=1}^{k}\frac{(-1)^j\Gamma(aj+k)}{\Gamma(aj)(j+1)!(k-j)!} = \frac{(-1)^k\Gamma(a+1)}{\Gamma(a-k+1)(k+1)!} \tag{52}$$

so that we arrive at

$$\begin{aligned}P_m(\tau) = &\left(\frac{b}{1+b}\right)^m\left(\frac{1+be^{-\tau}}{1+b}\right)^a\\ &\times\sum_{k=0}^{m}\frac{(-1)^k}{k!}\frac{\Gamma(a-k+m)}{\Gamma(m-k+1)\Gamma(a-k+1)}\left(\frac{1+b}{e^\tau+b}\right)^k\end{aligned} \tag{53}$$

Finally, we can write our solution using the hypergeometric function, ${}_2F_1$, which terminates when $-m$ is a non-positive integer such that

$$ {}_2F_1\left(-m, x, y; z\right) = \sum_{k=0}^{m} (-1)^k \frac{\Gamma(m+1)}{\Gamma(m-k+1)} \frac{(x)_k}{(y)_k} \frac{z^k}{k!} $$

where $(x)_k$ represents the Pochhammer symbol for the rising factorial defined as $(x)_k = \Gamma(x+k)/\Gamma(x)$ and which satisfies

$$ \Gamma(a+1) = (-1)^k (-a)_k \Gamma(a-k+1) $$

Writing

$$ \Gamma(a-k+m) = \Gamma(a+m-1-k+1) $$

leads us to the final form of our solution for the time-dependent transcript count PMF given zero transcripts initially

$$ P_m(\tau) = \frac{\Gamma(a+m)}{\Gamma(m+1)\Gamma(a)} \left(\frac{b}{1+b}\right)^m \left(\frac{1+be^{-\tau}}{1+b}\right)^a {}_2F_1\left(-m, -a, 1-a-m; \frac{1+b}{e^\tau+b}\right) \tag{54} $$

which we used in equation 9.

2.7 Simulations for model comparison

The performance of inference using different likelihood functions was tested on simulated data. Gillespie's exact algorithm (stochastic simulation algorithm) [118] was used to simulate data according to the reactant matrix shown in table 1 and the product matrix shown in table 2, with the stoichiometry matrix shown in table 3, using the following chemical reaction scheme which allows for simulation of the pool of transcripts in the cell that was synthesised before (RNA_0) and after (RNA_1) the 4sU pulse started.

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

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

$$G_{on} \xrightarrow{\beta_0} G_{on} + RNA_0$$

$$G_{on} \xrightarrow{\beta_1} G_{on} + RNA_1$$

$$RNA_0 \xrightarrow{\delta_0} \emptyset$$

$$RNA_1 \xrightarrow{\delta_1} \emptyset$$

	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 1: Reactant 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 2: Product 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

Table 3: Stoichiometry matrix for new and surviving transcript count Gillespie algorithm simulations.

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 a bursty gene with moderate 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 $t_1 = 1000$ minutes was used here so that the average transcript lifetime matches the 4sU pulse duration, ensuring a roughly split between new and surviving transcripts. This gives the final state of the system X_{t_1}, and importantly gives the counts for RNA_0 and RNA_1 in the cell. The biological parameters used results in an expression level of 25 transcripts per cell, a burst size of 50 and a burst interval and transcript lifetime of roughly 1000 minutes. The values all fall within the previously observed ranges [53, 95, 96] and are therefore biophysically reasonable. Although the given transcript lifetime corresponds to a biologically plausible value, the exact number used in the simulations is irrelevant since identical results could be produced by scaling all parameters with the transcript lifetime, only their values relative to each other matters from a simulation/modelling perspective rather than the actual timescales. The simulation was repeated to simulate $N = 10000$ cells and in-silico sequencing data was then generated based on these simulated transcript count values. Cell-specific capture efficiencies were drawn

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

to enable comparison of model performance with highly variable capture efficiencies, 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 [148]. 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, which is set to ensure sufficient T>C count data to enable clear comparison of model performance. If the reads are too short there will be fewer T>Cs per read which will impede the capacity for models which harness the T>C count data to do so, making the effects/differences on inference performance weaker. 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 $\lambda_n = 0.075$, both of which are within the range of values observed in [99], which used the same chemistry for converting incorporated 4sU as the previously mentioned 4sU scRNA-seq dataset that we sourced (section 2.1.1). The λ_n value is also close to the one inferred in section 2.5. 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 (see section 2.3) to sample posterior distributions.

2.8 Simulations for validating model predictions

Gillespie algorithm simulations [118] were carried out to generate in-silico data for a single gene to validate the predictions made by our mathematical models (see section 2.2). This was performed using the method described for the model comparison simulations (see section 2.7) but using different values for some parameters to allow for the most clear possible visualisation of the various distributions, whereas in section 2.7 the aim was just to simulate data for a gene with relatively high transcriptional noise to demonstrate inference capacities. Biological parameters were set to $\theta = (\beta_0 = 25, \beta_1 = 0, \delta_0 = 0.001, \delta_1 = 0.001, k_{on} = 0.002, k_{off} = 1)$ when running to steady state and then $\theta = (\beta_0 = 0, \beta_1 = 25, \delta_0 = 0.001, \delta_1 = 0.001, k_{on} = 0.002, k_{off} = 1)$ for the post-4sU simulation, also setting $N = 100000$ cells and $\alpha = 1$ for all cells. Analytical distributions from our models were produced with $\theta = (\mu = 50, a = 2, \gamma = 1000)$ to match the parameter values used in the simulations, corresponding to a gene with bursty transcription. The biological parameter values all fall within previously observed ranges [53, 95, 96]. These settings were used to directly visualise the agreement between different analytical distributions and their simulated counterparts. This agreement/difference was quantified as $\sum_x |P_{ana}(x) - P_{sim}(x)|$, the sum of the absolute value of the difference in probability mass between the two distributions across the state space, x. For several distributions, how the divergence varies across multiple points in parameter space was tested, simulating with $N = 10000$ cells and fixing all parameters at the aforementioned values other than the on being varied. β_0 and β_1 were varied together to achieve a range of expression levels (μ), k_{on} was varied to achieve a range of burst rates (a), and k_{on}, δ_0 and δ_1 were varied together to achieve a range of transcript lifetimes (γ) while maintaining a constant burst rate ($a = 2$).

Beginning with the steady state transcript count PMF (equation 1), we confirm our analytical results by directly comparing the distributions generated with our model analytically or via simulations in the aforementioned manner, seeing a close agreement (figure 15). Figure 16 also shows how the previously described probability difference value for the steady state distribution changes when varying one parameter at a time. The patterns observed in this case, and indeed in all subsequent cases, can be explained by how concentrated or dispersed the probability mass is across state space at different parameter values, with a more highly concentrated distribution resulting in a lower difference value. This is because there is a greater degree of certainty associated with these distributions, whereas distributions dispersed over a larger state space will require more cells, N, for the simulated distribution to match the analytical distribution to the same degree. Therefore, increasing the expression level (and burst size) while fixing burst rate and transcript lifetime results in the observed increase in probability difference due to the distribution being scaled over greater state space. Conversely, increasing burst rate (and burst frequency) while fixing the other two parameters resulted in reduced probability difference, since at lower burst rates the distribution is more skewed and therefore more dispersed over more state space, whereas at high burst rate the distribution converges to a poisson distribution, becoming concentrated around the mean value. As expected, varying the transcript lifetime has no effect on the difference since the steady state distribution is governed purely by dimensionless parameters. Since the patterns observed are explained simply by state space variation rather than biases, this result provides validation of our model.

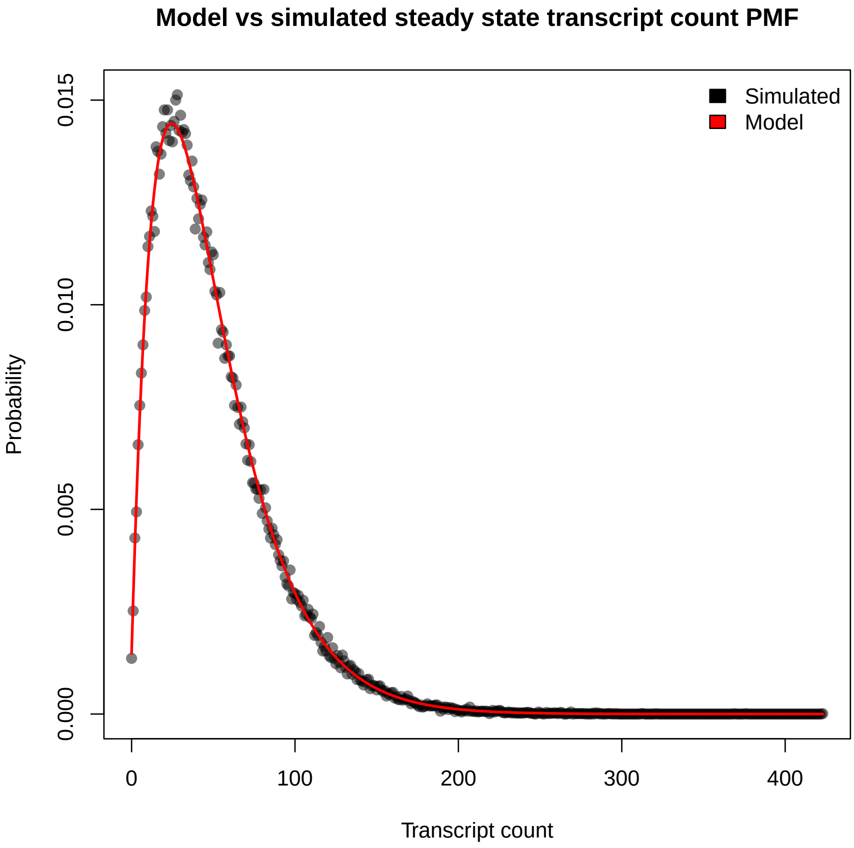


Figure 15: The steady state transcript count PMF generated with simulations or analytically using matching parameter values.

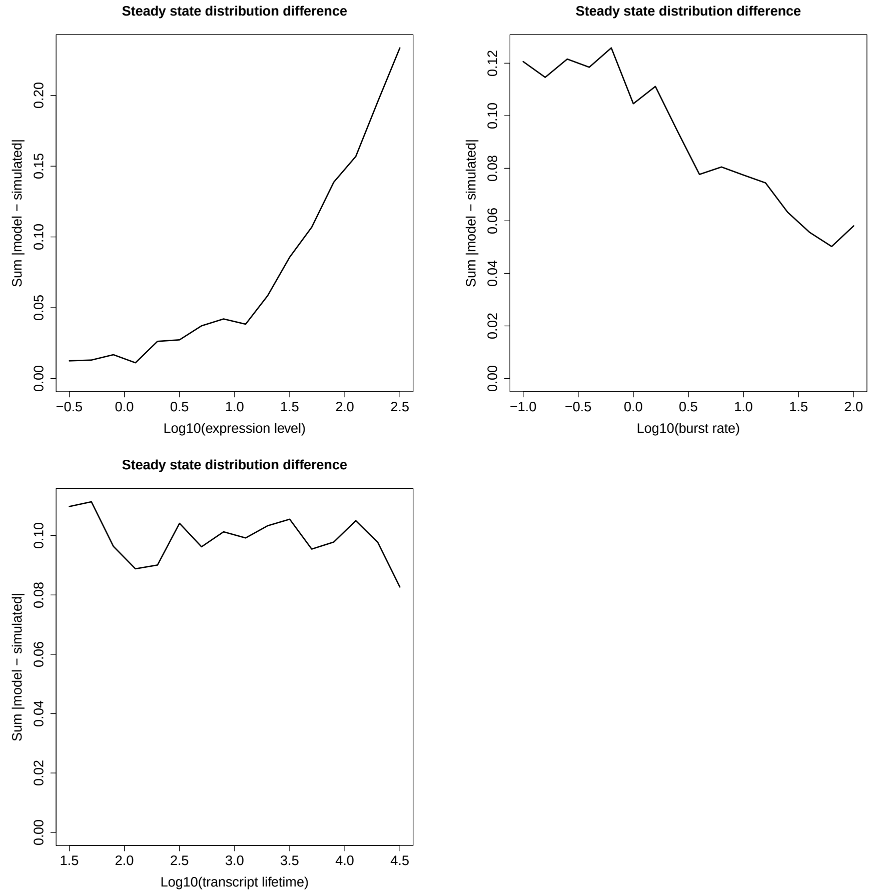


Figure 16: The sum of the absolute value of the difference in probability across state space for the steady state transcript count PMF when generated analytically or with simulations using matching parameter values, with expression level, burst rate and transcript lifetime being varied one at a time. Lower y-axis values indicates closer agreement between the analytical distribution of the model and the simulated distribution for the given parameter value.

Moving on to our approximation of the surviving transcript count PMF (equation 8), its accuracy is supported by direct comparison to a simulated PMF (figure 17), observing a close agreement as before. Again, figure 18 shows how the probability difference varies when changing each of our three parameters. Identical patterns can be observed for expression level and burst rate as in figure 16 due to increased expression level and reduced burst rate causing increased state space. Unlike figure 16 however, the probability difference now varies with transcript lifetime with longer transcript lifetimes resulting in a larger probability difference, whereas very short transcript lifetimes have almost zero difference. This is again explained by variation in the concentration of the mass over state space at different transcript lifetimes. At very short lifetimes, there will be few/no transcripts surviving from before the 4sU pulse, such that $\lim_{\gamma \to 0} P(s = 0) = 1$, which corresponds to the maximum concentration level for the distribution and therefore the minimum probability difference. Conversely, at very long lifetimes, there will be few/no degradation events during the 4sU pulse, such that $\lim_{\gamma \to \infty} P(s) = P(m)$, which corresponds to maximum dispersion level for the surviving distribution (the steady state distribution) and therefore the maximum probability difference.

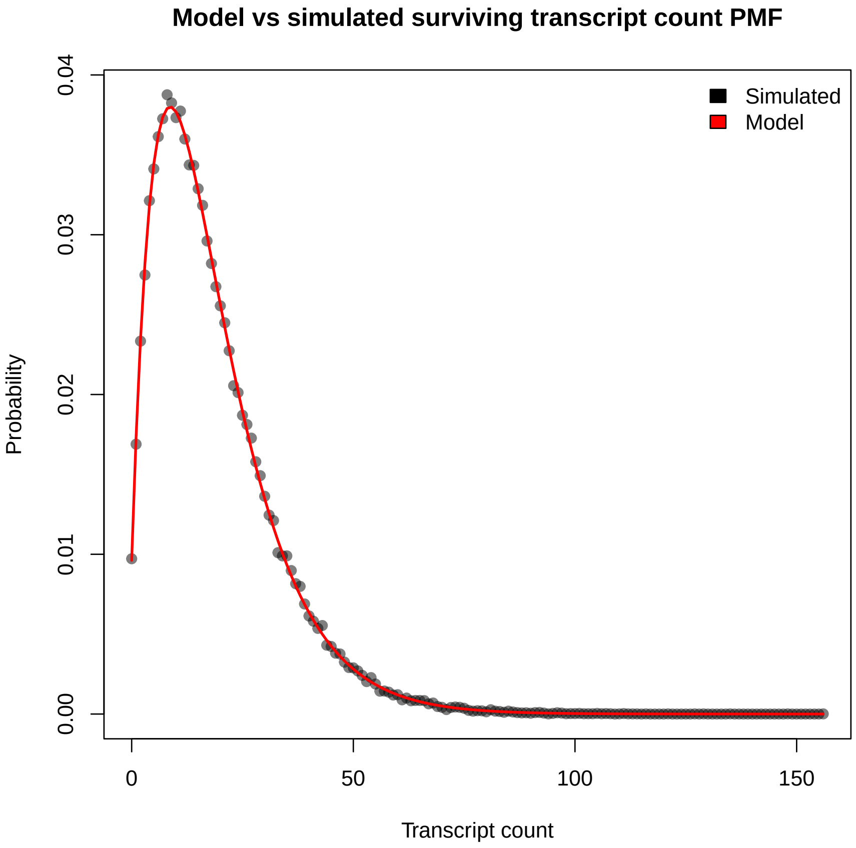


Figure 17: The surviving transcript count PMF generated with simulations or analytically using matching parameter values.

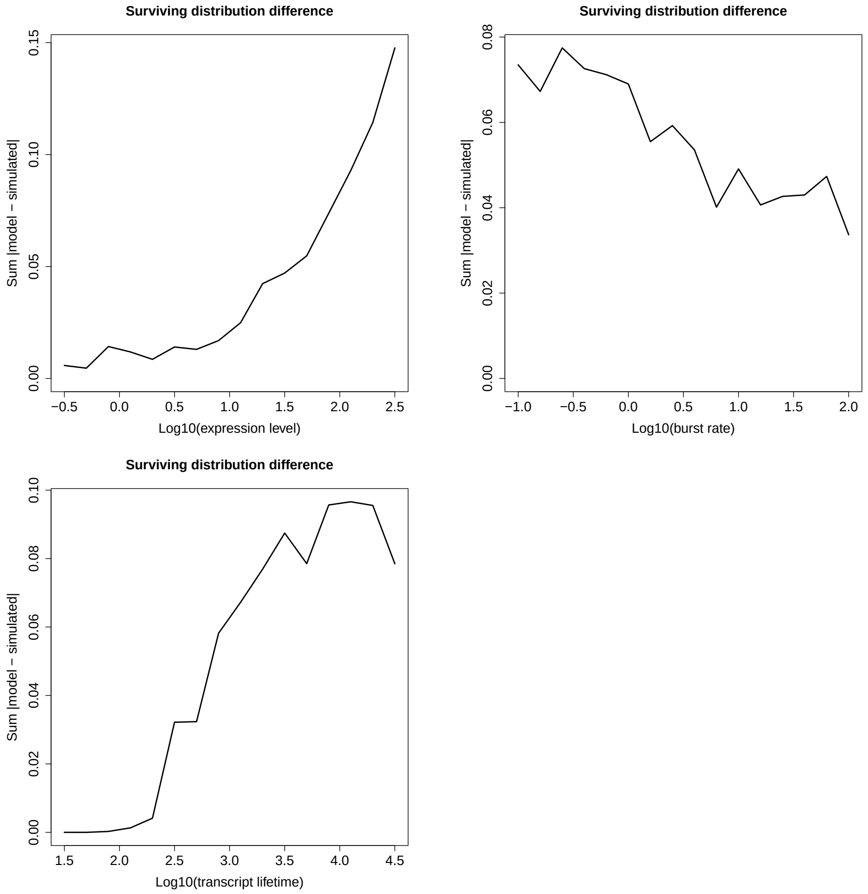


Figure 18: The sum of the absolute value of the difference in probability across state space for the surviving transcript count PMF when generated analytically or with simulations using matching parameter values, with expression level, burst rate and transcript lifetime being varied one at a time. Lower y-axis values indicates closer agreement between the analytical distribution of the model and the simulated distribution for the given parameter value.

Now turning to our approximation of the new (non-equilibrium) transcript count PMF (equation 9), we again validate it by directly comparing to its corresponding simulated PMF (figure 19), observing a good agreement. Looking once more at the variation in probability difference associated with the three parameters (figure 20), we note the previously observed pattern of increased probability difference with increase expression level due to increase dispersion of the distribution over the state space. The reverse relationship between transcript lifetime and probability difference is observed compared to figure 18. Since expression level and burst rate are fixed when varying transcript lifetime, we increase transcript lifetime by reducing both the decay rate and burst frequency together. Therefore, very long transcript lifetimes correspond to few/no new transcripts being produced during the 4sU pulse, such that $\lim_{\gamma \to \infty} P(n = 0) = 1$, which corresponds to the maximum concentration level for the distribution and therefore the minimum probability difference. Conversely very short transcript lifetimes correspond to many new transcripts being produced (and degraded) during the 4sU pulse, such that $\lim_{\gamma \to 0} P(n) = P(m)$, which corresponds to maximum dispersion level for the new distribution as it settles into equilibrium (steady state) and therefore the maximum probability difference. For the new transcript distribution we observe a slightly different relationship between probability difference and burst rate compared with figures 16 and 18, although this can again be explained by variation in the distribution of probability mass over the state space. We see an initial increase in probability difference with increasing burst rate up to a certain point, followed by a decrease beyond that point. Lower burst rate does result in increased skew in the distribution, as with the previous cases, and therefore, one would expect a greater probability difference with increasingly low burst rate due to the greater dispersion over state space. However, figure 19 shows that the new transcript count PMF is heavily zero-inflated to account for instances in which no bursts occur during the 4sU pulse (or a small burst occurs but all resulting new transcripts are degraded). Therefore, the lower the burst rate, the more heavily concentrated the distribution becomes at zero, leading to reduced probability difference. At high burst rates the distribution approaches a poisson distribution and becomes more concentrated about the mean, leading to reduced probability difference. At intermediate burst rates the distribution is not heavily concentrated at zero but still retains super-poissonian noise, skew and therefore dispersion, leading to maximal probability difference. Again, being able to explain the change in agreement between our analytical and simulation results purely through the dispersion of the distribution over state space provides confidence that our model is not biased over these regions on parameter space.

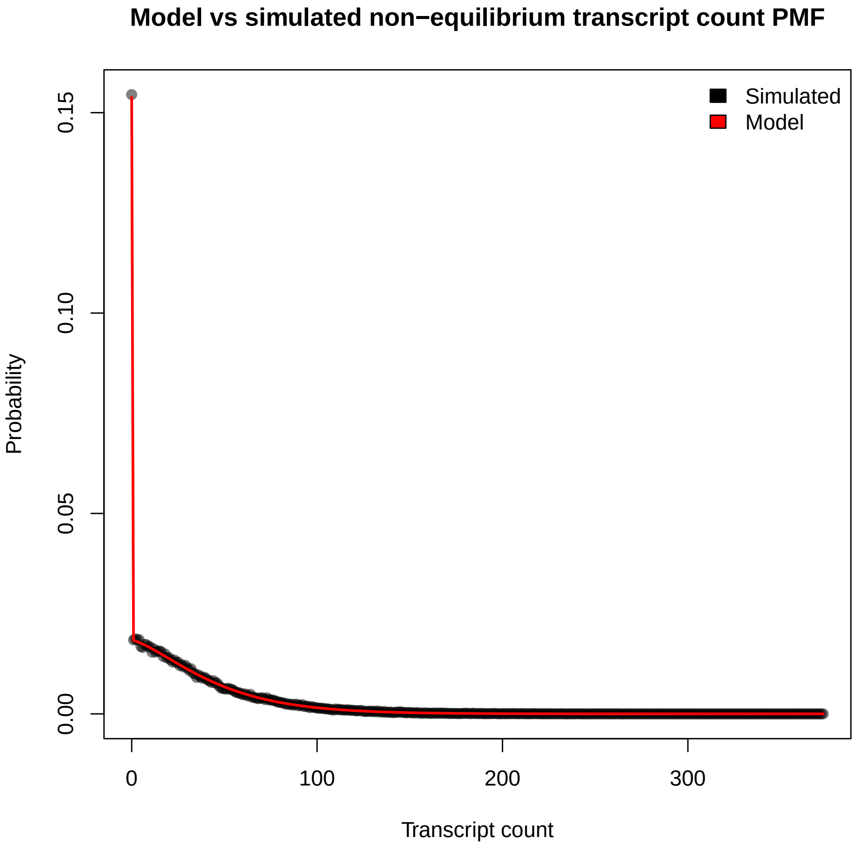


Figure 19: The non-equilibrium transcript count PMF generated with simulations or analytically using matching parameter values.

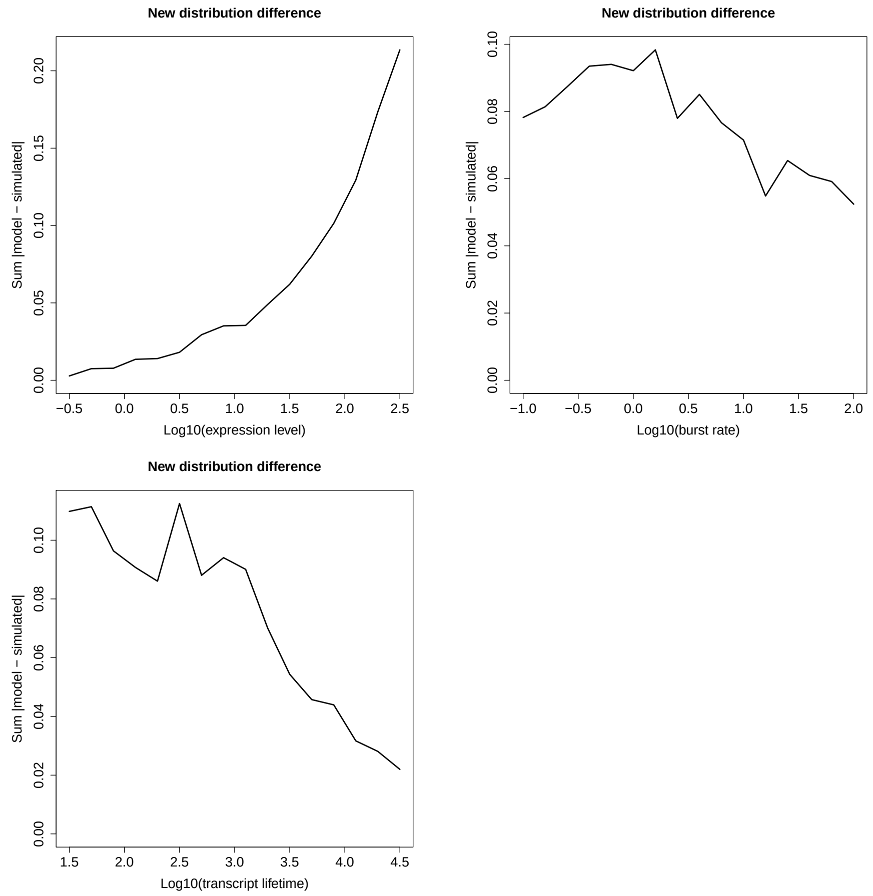


Figure 20: The sum of the absolute value of the difference in probability across state space for the new (non-equilibrium) transcript count PMF when generated analytically or with simulations using matching parameter values, with expression level, burst rate and transcript lifetime being varied one at a time. Lower y-axis values indicates closer agreement between the analytical distribution of the model and the simulated distribution for the given parameter value.

Verification of our solution to the PMF for the new transcript count conditional on total transcript count (equation 11) is achieved by comparing with simulation results for a given value of m (figure 21). Further validation is provided by comparing analytical and simulation-based expected new transcript counts conditional on total transcripts across different values of m (figure 22).

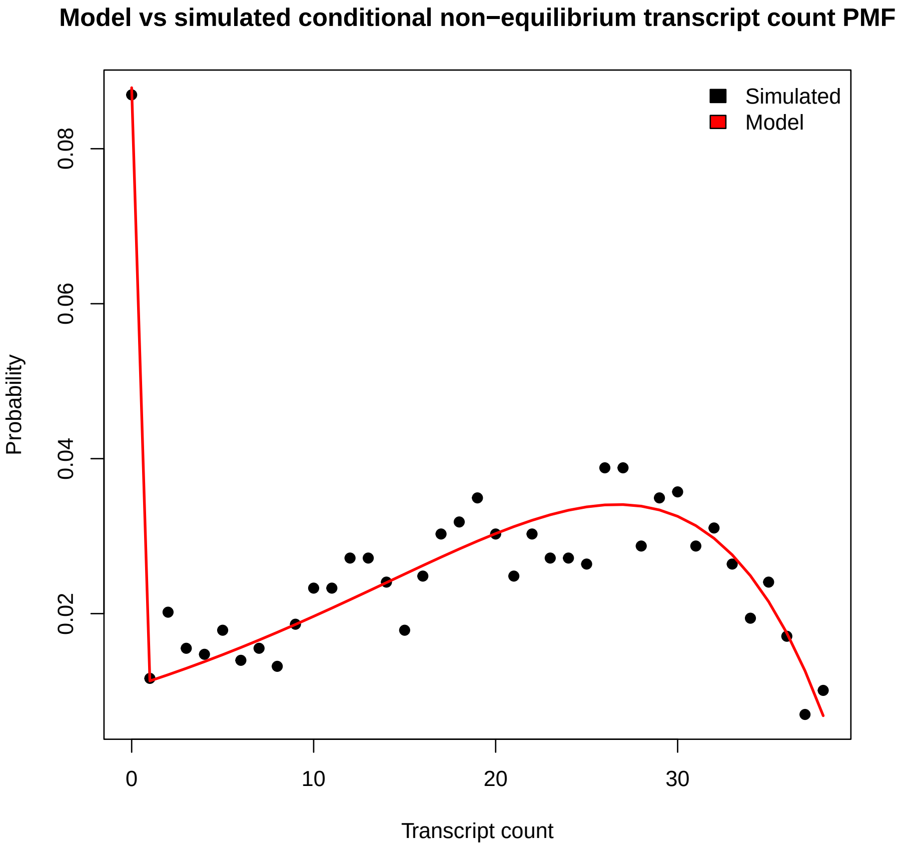


Figure 21: The non-equilibrium transcript count PMF conditional on total transcripts, in which the total transcript count $m = 38$, generated with simulations or analytically using matching parameter values.

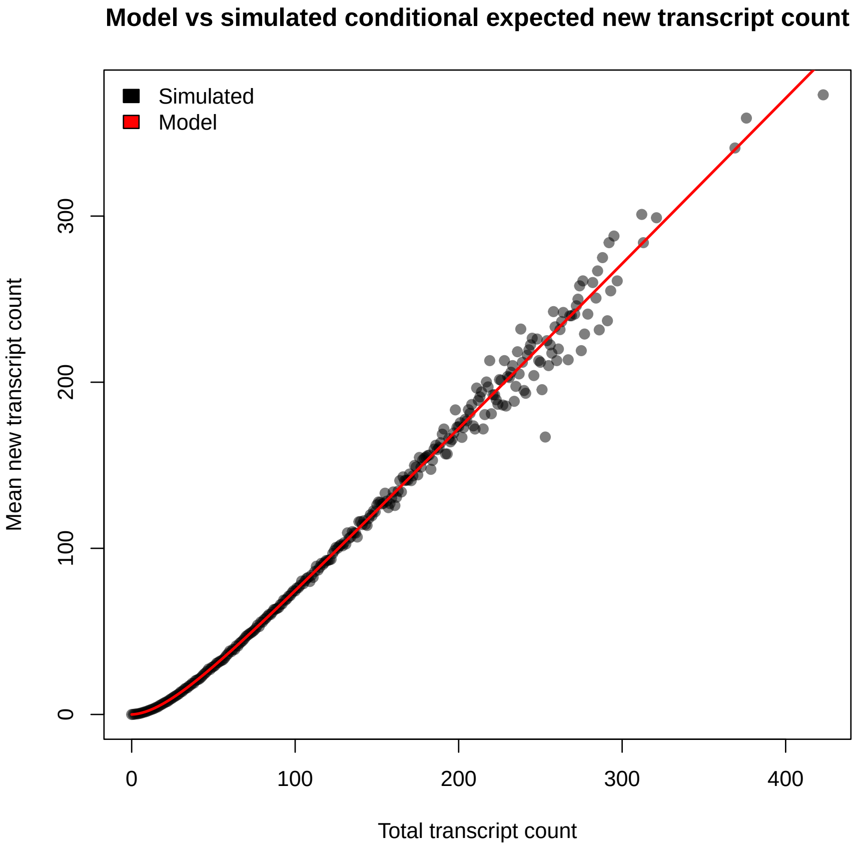


Figure 22: The expected new transcript count conditional on total transcript count across different values of m generated with simulations or analytically using matching parameter values.

We can also demonstrate the alignment of our model with simulation results for the full joint PMF of the new and surviving transcript counts (figures 23 and 24). Here we show the joint distribution for $s \in [0, 70]$ and $n \in [1, 170]$, truncating $n = 0$ for visualisation purposes due to the inflated value of $P(n = 0)$ for the chosen parameter values (figures 19 and 21), with the agreement between the two demonstrated by the probability difference heatmap (figure 25). Additionally, we examine the dependence on the probability difference between the analytical and simulated joint distribution with each of three individually varied parameters (figure 26). As in previous cases, the probability difference depends on how concentrated the probability mass is in state space at each parameter value, with higher concentration resulting in reduced probability difference. As in figures 16, 18 and 20, we again observe an increase in probability difference with increased expression levels due to increased dispersion over state space. This is a much stronger effect over the same parameter values than previously observed because the state space is now 2-dimensional rather than 1-dimensional, so the total state space covered by the distribution scales with the square relationship to the expression level rather than linearly. Therefore, the total state space over which the probability mass is distributed becomes vast, meaning that a much larger number of simulated cells, N, would be required to accurately populate this space, with a very large number of tiny probability values. The relationship of the probability difference with burst rate is similar to that observed in figure 20 but the effect is much more clear and strong. This is because at low burst rates the more heavily skewed individual distributions, and especially the zero-inflated new transcript distribution, ensures that a significant portion of the probability mass is concentrated at a small set of joint values at the lower end of both axes close to zero, restricting the probability difference. At high burst rates, each individual distribution approaches poisson, resulting in a concentration of mass at the two mean values, thereby reducing the probability difference. At intermediate burst rates, the distributions are less heavily skewed but still with super-poissonian skew and with reduced zero-inflation in the new distribution, resulting in both the new and surviving distributions having their probability mass dispersed over a large region of state space, which increases in a square fashion as previously mentioned, maximising the probability difference and the observed sharp increase over these burst rates. Finally coming to the transcript lifetime, the state space, and therefore the probability difference is maximised with intermediate transcript lifetimes because neither the surviving or new distribution is concentrated towards zero at these values. Either extreme of the transcript lifetime axis corresponds to the maximum dispersion of one distribution but the minimum of the other, such that the total dispersion is maximised when the transcriptome exhibits a roughly even split between new and surviving transcripts. Observing that the variation in probability difference for the joint PMF follows the patterns expected based on the distribution of probability mass over the state space provides confidence in our model, demonstrating that varying different parameters does not introduce biases into our model predictions that would be missed when directly comparing analytical vs simulated distributions at fixed values.

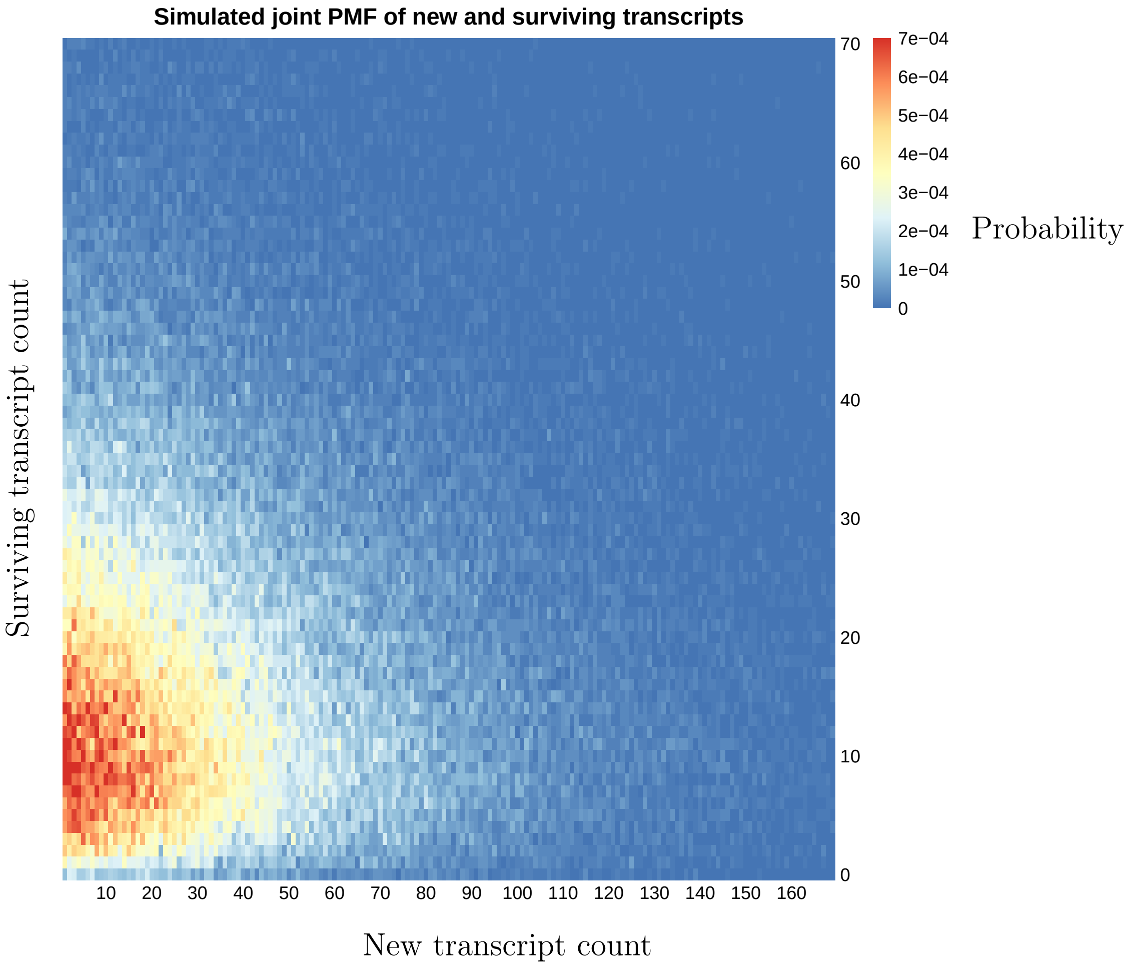


Figure 23: The joint new and surviving transcript count PMF as generated by simulations for surviving transcript count values $s \in [0, 70]$ and new transcript count values $n \in [1, 170]$.

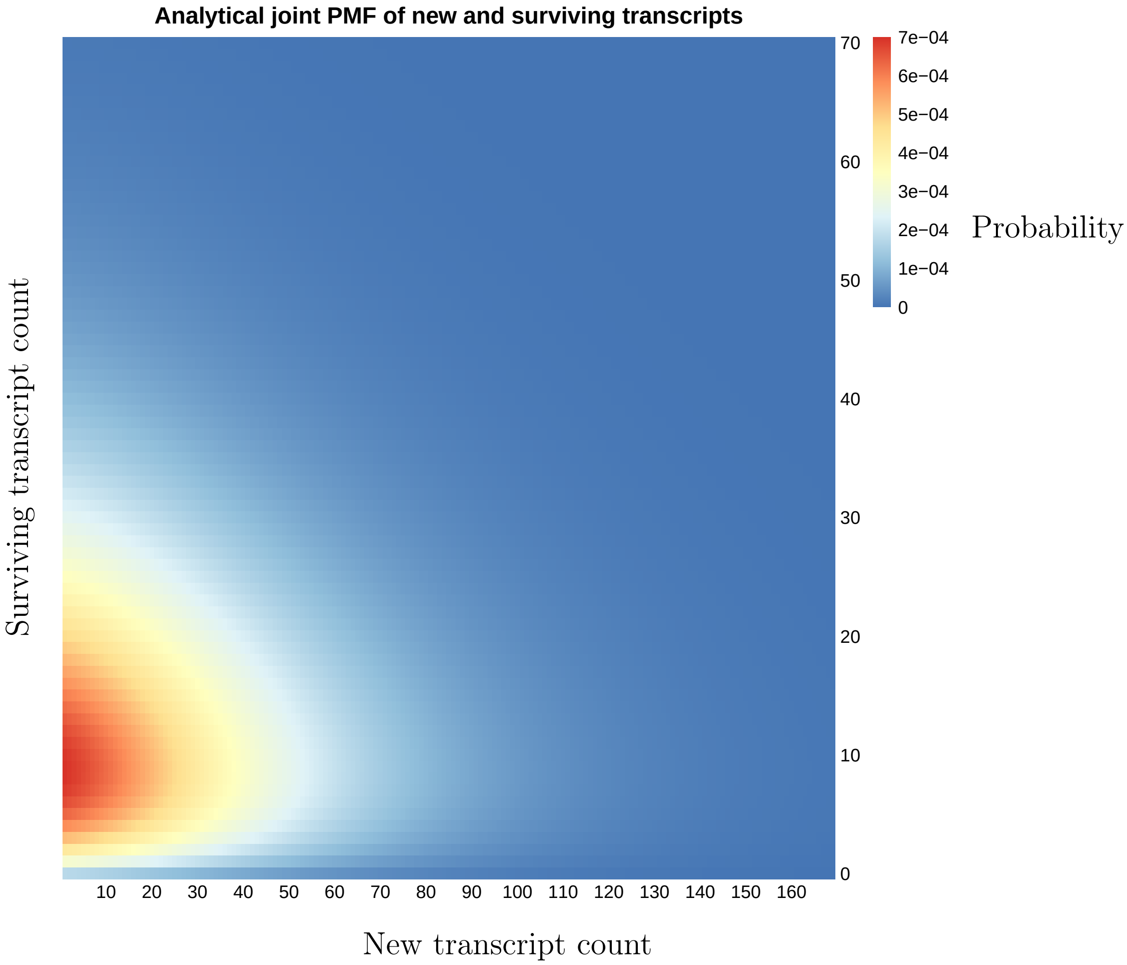


Figure 24: The analytical joint new and surviving transcript count PMF for surviving transcript count values $s \in [0, 70]$ and new transcript count values $n \in [1, 170]$.

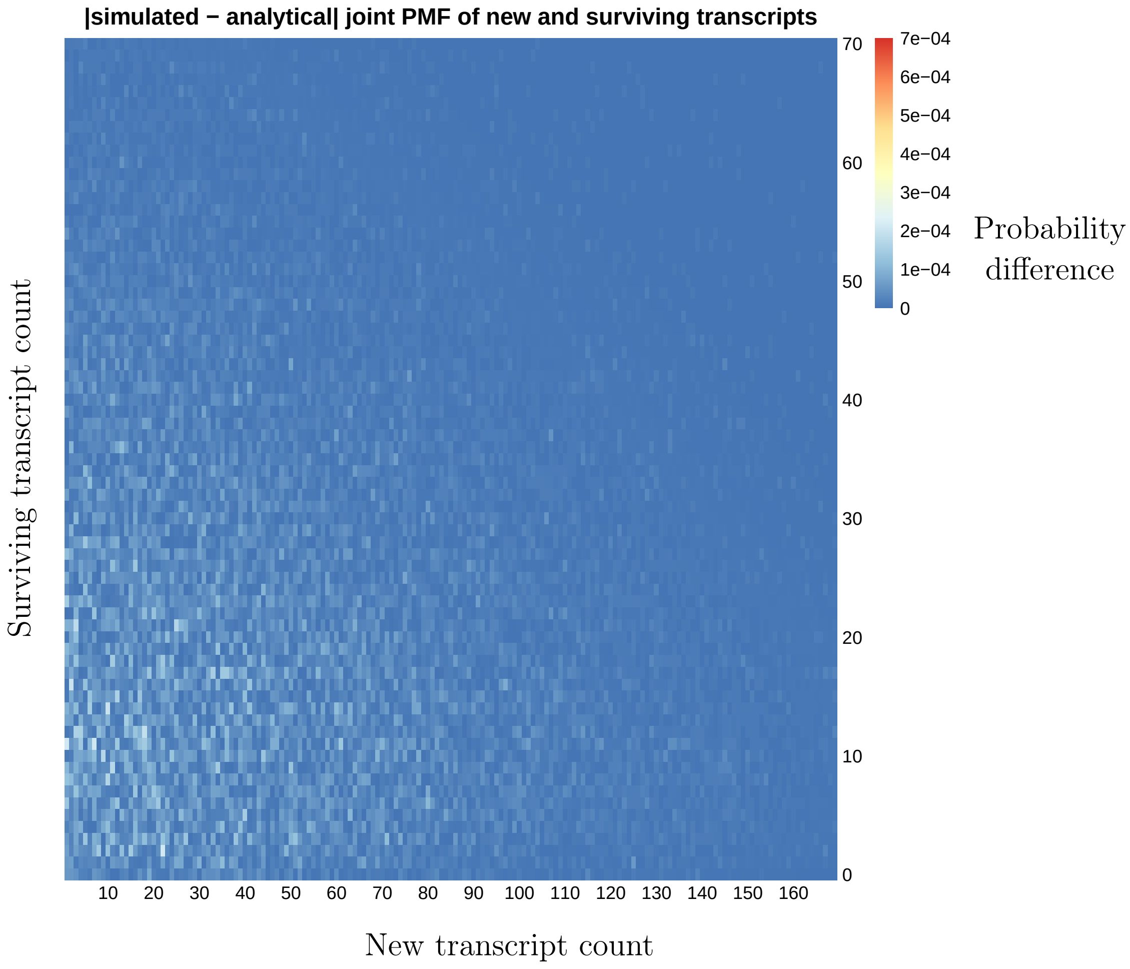


Figure 25: The absolute value of the difference in probability between the analytically generated and simulation-based joint new and surviving transcript count PMFs for surviving transcript count values $s \in [0, 70]$ and new transcript count values $n \in [1, 170]$ produced with matching parameter values.

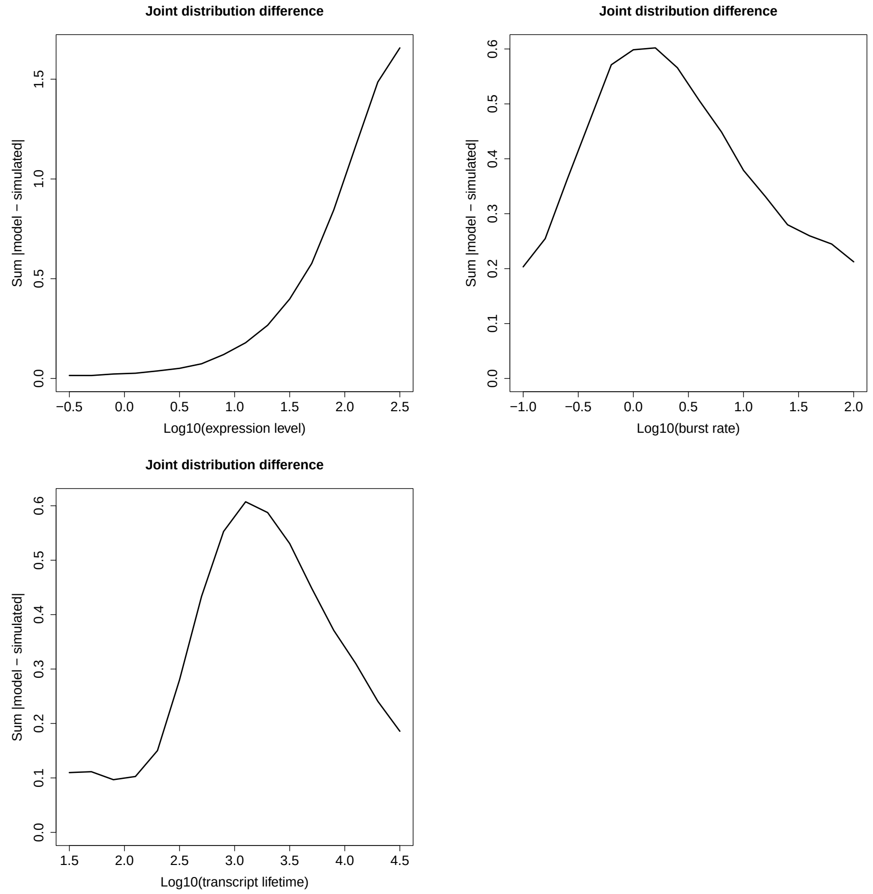


Figure 26: The sum of the absolute value of the difference in probability across state space for the joint surviving and new (non-equilibrium) transcript count PMF when generated analytically or with simulations using matching parameter values, with expression level, burst rate and transcript lifetime being varied one at a time. Lower y-axis values indicates closer agreement between the analytical distribution of the model and the simulated distribution for the given parameter value.

The accuracy of our solution for the T>C count PMF conditional on total transcript count (equation 12) is also demonstrated by comparing with simulations for a given total transcript count, m (figure 27). This distribution corresponds to the total T>C count for any given read corresponding to either new or surviving transcripts for a cell with the given total transcript count, m, and is already normalised by the summation over the empirically observed total uracils per read distribution, $P(u)$, in equation 12. Further verification is once more provided through comparison of analytical and simulation-based expected T$>$C counts conditional on total transcripts across different values of m (figure 28).

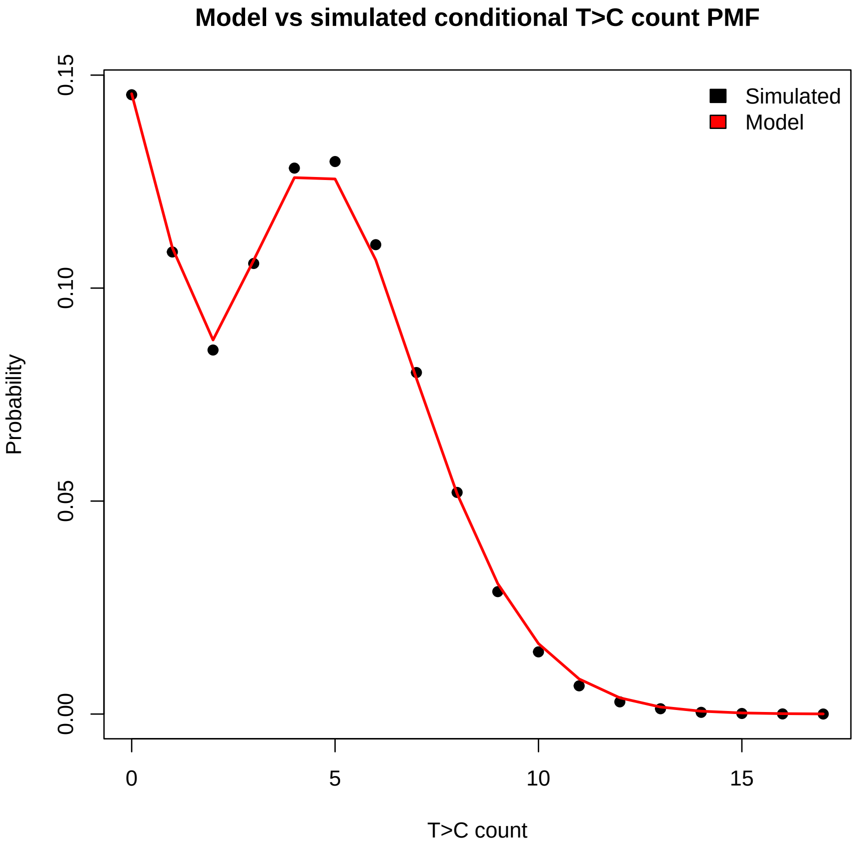


Figure 27: The T$>$C count PMF conditional on total transcripts, in which the total transcript count $m = 100$, generated with simulations or analytically using matching parameter values.

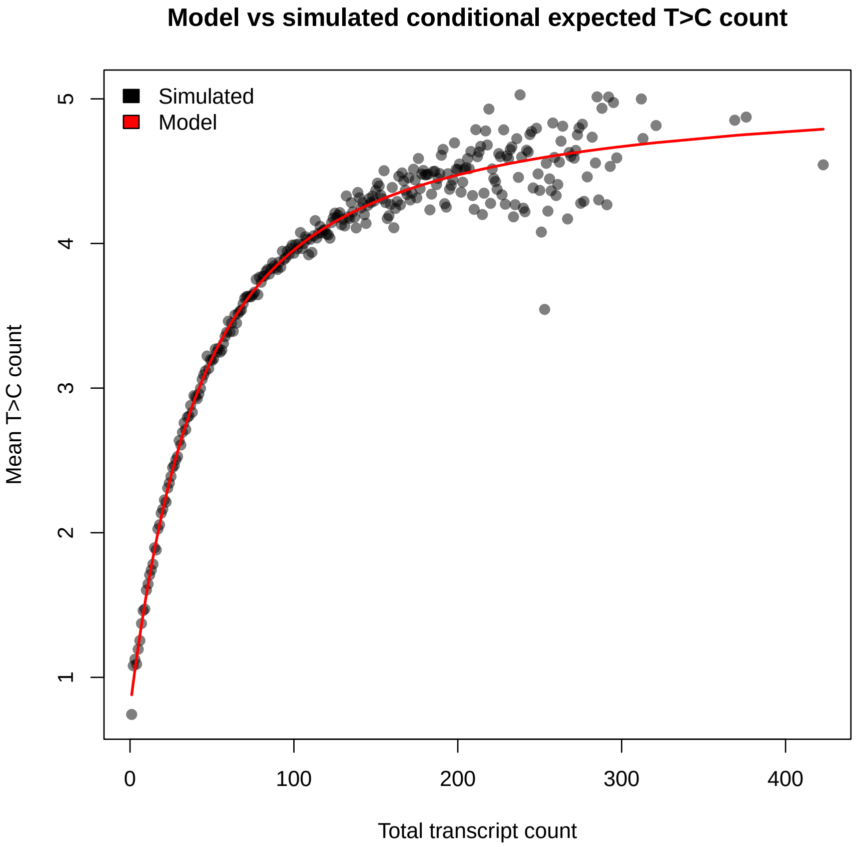


Figure 28: The expected T>C count conditional on total transcript count across different values of m generated with simulations or analytically using matching parameter values.

Finally, simulation-based verification of our solution to the T>C count PMF independent of total transcript count (equation 18) is demonstrated (figure 29). Like figure 27, this distribution represents the total T>C count for any given read corresponding to either new or surviving transcripts but in any given cell regardless of total transcript count, again normalised by the summation over the empirically observed total uracils per read distribution, $P(u)$, in equation 12.

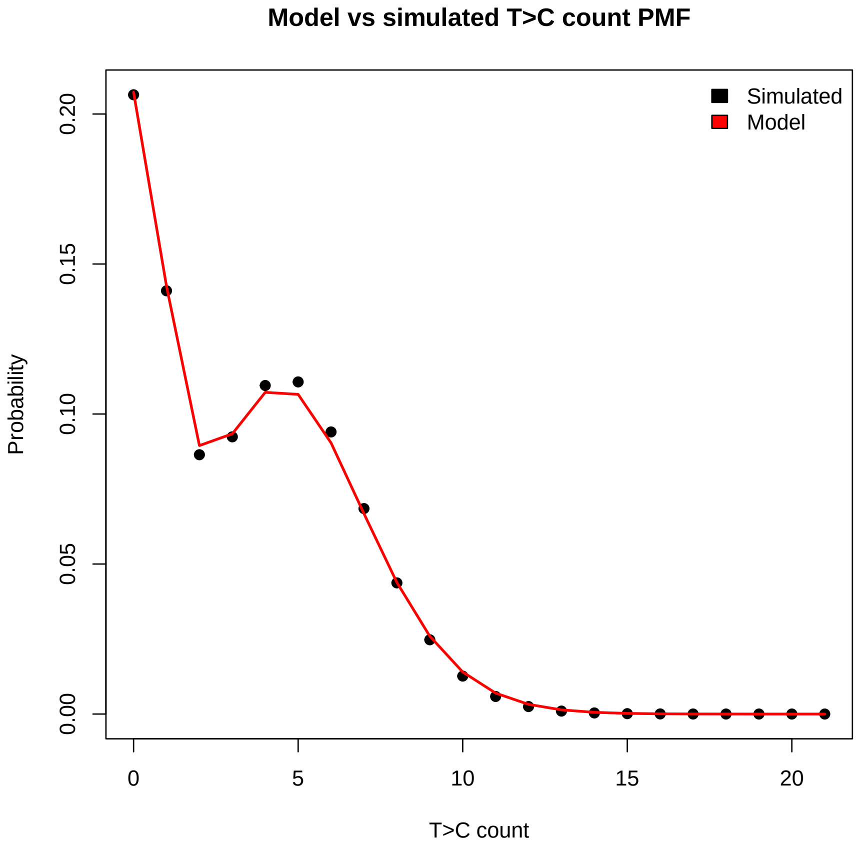


Figure 29: The cell-invariant T>C count PMF, which is not conditional on total transcripts, generated with simulations or analytically using matching parameter values.

2.9 Simulations for genome-wide inference validation

In order to assess the performance of the inference algorithm, a dataset was simulated for each of the genes from the real Qiu dataset whose bursting dynamics were inferred. The θ estimates obtained from Qiu were used as the "ground truth" parameter values to simulate new datasets. A dataset for each gene is simulated by drawing the steady state transcript counts for $N = 795$ cells, matching the real data, at the start of the pulse

$$m_0 \sim NBin\left(a, \frac{b}{1+b}\right)$$

the surviving transcript counts at the end of the pulse

$$s \sim Bin\left(m_0, e^{-\tau}\right)$$

the number of bursts occurring during the pulse

$$\beta \sim Pois(\kappa t)$$

how long before the end of the pulse each burst occurs at

$$T \sim Unif(0, 240)$$

the size of each burst

$$\sigma \sim Geom\left(\frac{1}{1+b}\right)$$

where

$$f_{Geom}\left(\sigma | \frac{1}{1+b}\right) = \left(\frac{b}{1+b}\right)^{\sigma}\left(\frac{1}{1+b}\right)$$

and the number of newly synthesised transcripts from each burst which survive to the end of the pulse

$$n \sim Bin\left(\sigma, e^{-\delta T}\right)$$

Then the total transcripts in each simulated cell at the end of the pulse is $m = s + \hat{n}$, summing the number of transcripts surviving from each burst occurring during the pulse as $\hat{n} = \sum n$. The number of UMIs corresponding to surviving and new transcripts in each cell is then drawn as

$$l_s \sim Bin(s, \alpha)$$

and

$$l_n \sim Bin(\hat{n}, \alpha)$$

where α is sampled without replacement from the set of estimated capture efficiencies (see section 2.4). Then we have the total UMI count for each cell $l = l_s + l_n$ and L, where L_c is the UMI count of cell c. For each UMI, we draw the number of corresponding reads, r, as

$$r \sim ZTPois(\rho)$$

where ρ is the maximum likelihood estimate given the observed ratio of reads to UMIs, R, across all cells for the given gene in the real data. ρ is obtained by minimising

$$\left| R - \frac{\rho}{1 - e^{-\rho}} \right|$$

From this we have the total number of reads corresponding to surviving, r_s and new, r_n transcripts for each cell. The number of uracils, u, in each read is then drawn from the gene-specific empirical probability mass function, $P(u)$, from equation 12. Finally, the number of uracils in each read which undergo a T>C conversion, i, is then drawn for surviving reads as

$$i \sim Bin(u, \lambda_s)$$

and for new reads as

$$i \sim Bin(u, \lambda_s + \lambda_n)$$

Now we have y, where y_i represents the number of reads with i total conversions in the given cell and and Y, where Y_c is the the vector y for cell c. We could then carry out inference with L and Y as previously described (see section 2.3).

3 Inferring transcriptional dynamics

Here we outline the results of applying our mathematical models and Bayesian inference algorithm to both real and simulated datasets. The improvement in estimation of the parameters governing transcriptional bursting offered by 4sU scRNA-seq is explored. This is followed by the extraction of genome-wide joint estimates for time-resolved bursting dynamics and subsequent simulation-based validation of algorithm performance.

3.1 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 (see sections 2.2 and 2.7 for details about the likelihood functions and simulation approach, respectively). The MCMC algorithm was run five times, using equations 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 (figure 30) for each likelihood function, confirming that scRNA-seq data cannot resolve burst frequency or decay rate (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 decay rate, δ, further demonstrating the advantage of cell-specific vs cell-summed T>C conversion data. As expected, L1+L3 converges for all parameters, with L1 informing burstiness while L3 informs timescales.

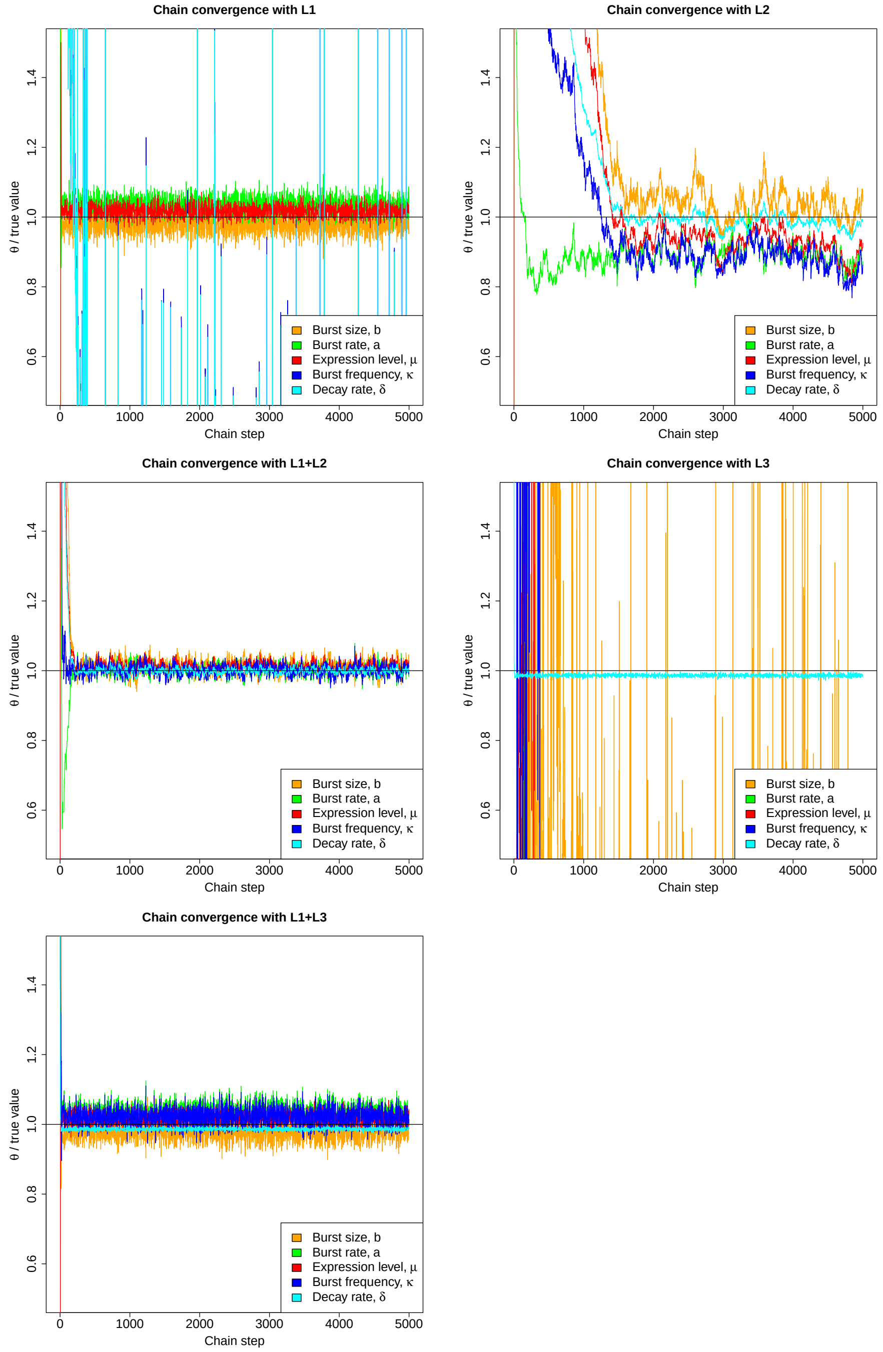


Figure 30: Convergence of Markov chains to true parameter values with simulated data for five different likelihood functions. The parameter values, θ, in the chain are divided by the true value to allow for joint visualisation, with the black horizontal line representing the target value. The different likelihood functions use different parts of the simulated 4sU scRNA-seq data and so Markov chains are able to converge to the target values for different subsets of parameters depending on the likelihood function used.

The resulting posteriors (figure 31) indicate that the accuracy and precision of estimates for burst rate, burst size and expression level (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 burst frequency or decay rate (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 burst rate, burst size and decay rate (a, b and δ), while the UMI counts are a function of burst rate and burst size (a and b). Therefore, including the UMI data improves inference of burst rate and burst size (a and b), which reduces the uncertainty associated with decay rate (δ) in our joint inference approach.

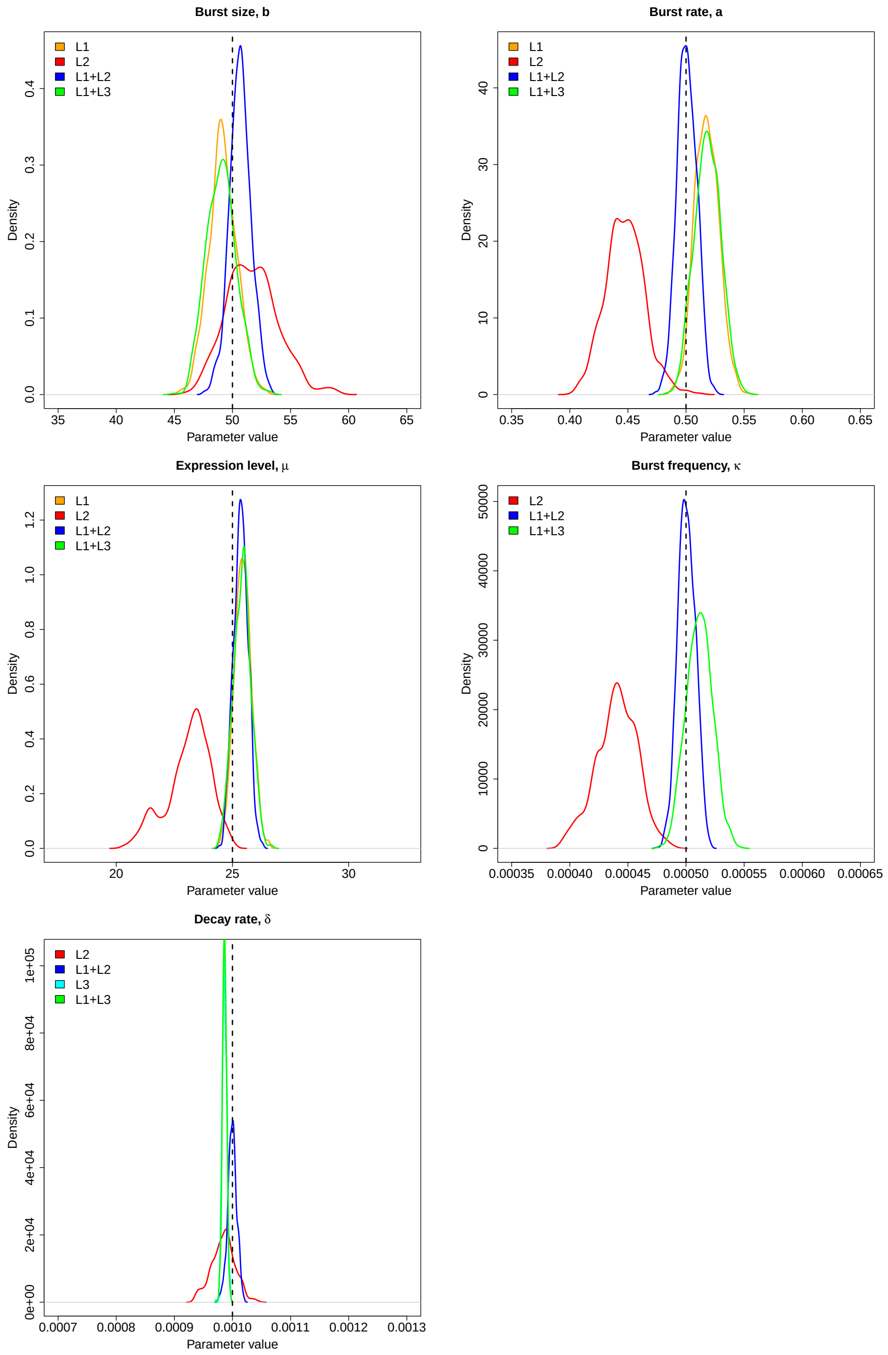


Figure 31: Posterior densities of each parameter obtained using different likelihood functions, with the dashed black lines representing the true parameter values that were used to simulate the dataset upon which inference was carried out. The densities for δ obtained with L3 and L1+L3 are difficult to distinguish because they almost perfectly overlap. The different likelihood functions use different parts of the simulated 4sU scRNA-seq data and so Markov chains are able to generate samples centred about the target value for different subsets of parameters depending on the likelihood function used.

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 (burst rate, a, and burst size, b) and bulk SLAM-seq (decay rate, δ) information. This is apparent in this example of a gene with moderate expression, high transcriptional noise and a transcript lifetime equal to the 4sU pulse duration.

3.2 Inference on real data

We next applied our method to 4sU scRNA-seq data published in 2020 by Qiu et al, which used human K562 cells [101]. 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, if required, model 3 (see section 2.3). 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. A CV of 0.45 was chosen in order to maximise the number of selected genes available for downstream analysis. The trade-off for increasing the CV (and number of genes) is that the minimum level of precision in our estimates is also reduced. Therefore, the aim was to increase the CV but not beyond the point that some estimates would be completely uninformative due to a lack of precision. The exact CV value was also informed by the diminishing returns in terms of number of genes obtained by increasing the CV over the informative range. Figure 32 shows 0.45 to coincide with the point at which the relative number of extra genes obtained with each increase in CV drops off, making it less worthwhile to increase the CV beyond 0.45.

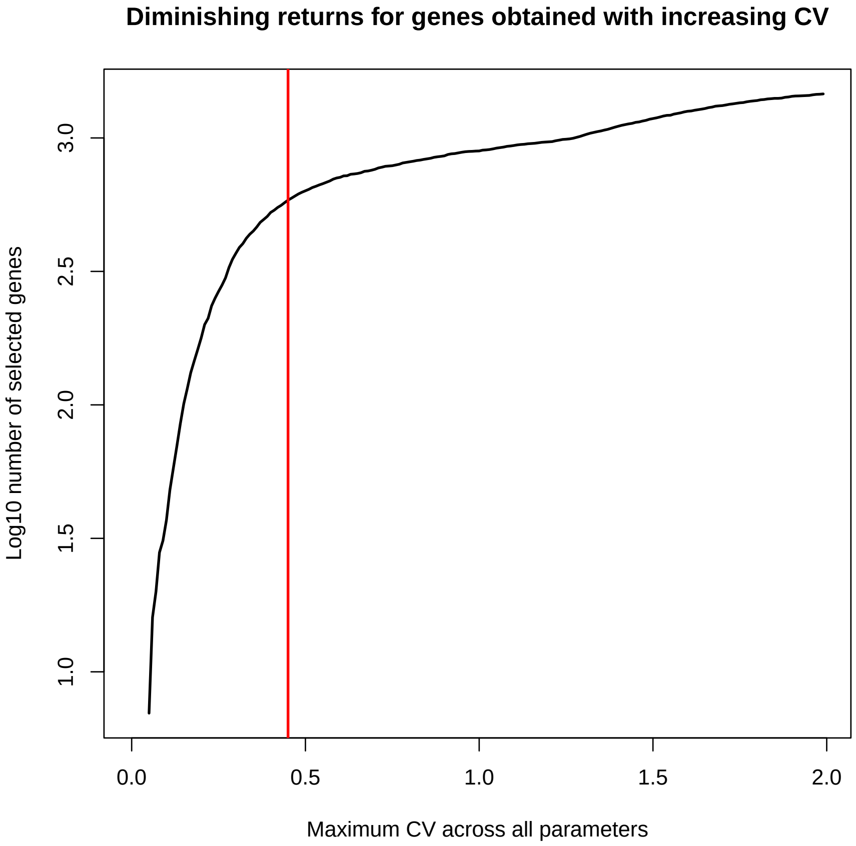


Figure 32: The number of genes selected for different CV values based on allowing no CV $> X$ for any parameter estimates inferred from the Qiu data for that gene. The chosen CV cutoff of 0.45 is shown as the red line, beyond which there are diminishing returns.

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 figure 33, which depicts estimate vs CV for all parameters. $CV(\delta)$ has an optimal (minimum) value for decay rate, δ, 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 expression level $\mu \geq 1000$. We also have increased confidence in general for genes with higher expression level, μ, since estimates for such genes are informed by a greater volume of data. Likewise, genes with greater burst size, b, have greater confidence because, firstly, increased burst size, b, results in higher expression level, μ. Secondly, for a given expression level, μ, having a higher burst size, b, implies lower burst rate, 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 burst rate, a. This is because it is associated with higher expression level but lower transcriptional noise. Therefore, a gene with higher burst rate, 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 burst frequency, κ, is essentially dictated by the burst rate and decay rate (a and δ) values for the gene.

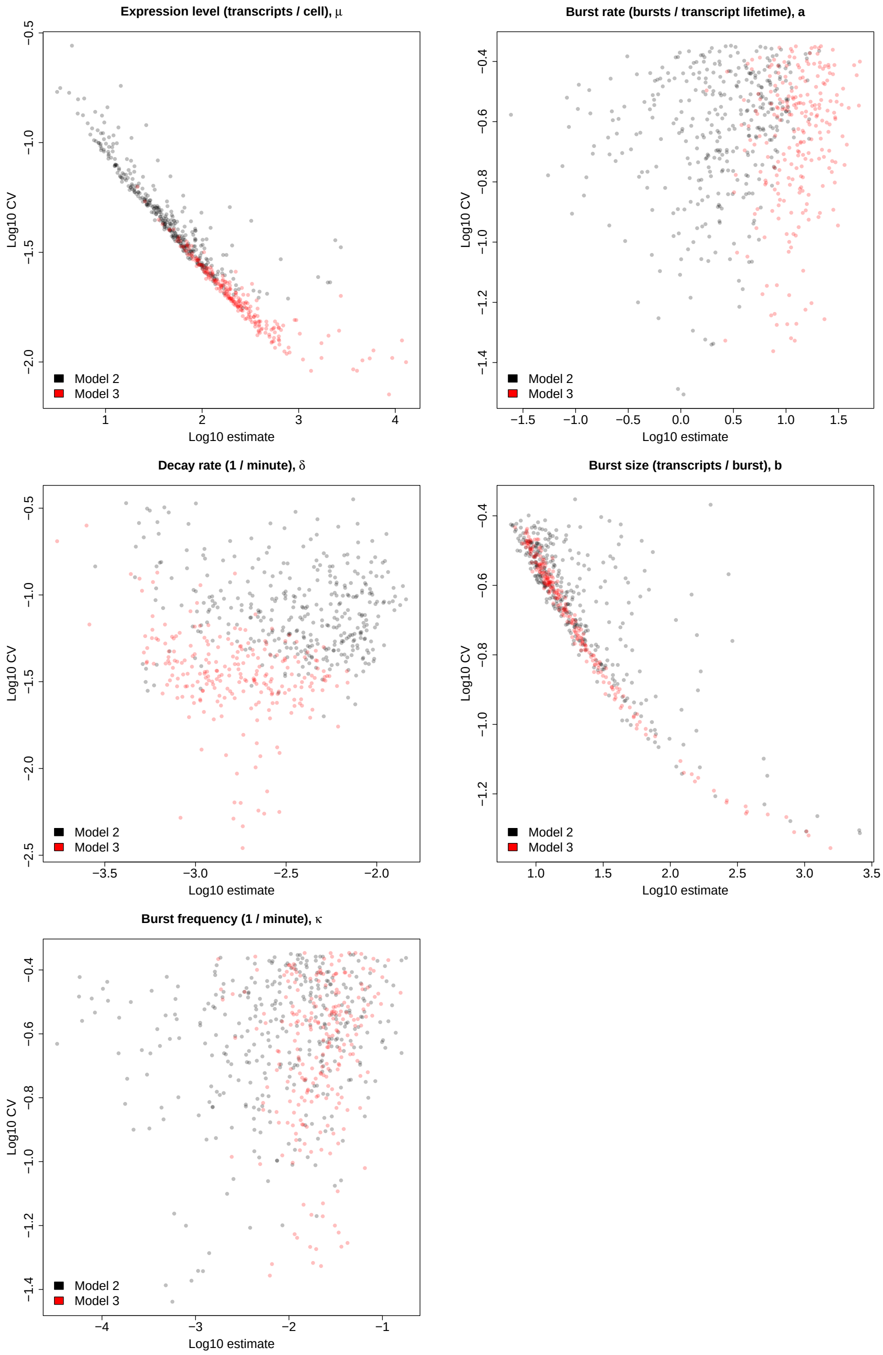


Figure 33: Mean value estimates vs corresponding CVs of all parameters derived from posteriors generated by inference on the Qiu data for all 584 selected genes (which had $CV < 0.45$ for all parameter estimates), with those obtained using models 2 or 3 displayed in black or red, respectively. Lower y-axis values indicates lower CV and, therefore, higher confidence and precision.

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 equation 9 becomes unstable. Figure 33 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 expression level, μ, and burst rate, a, are higher and when decay rate, δ, is lower. Additional confidence in our results is provided by showing the strong correlation about the diagonal between the decay rate estimates we obtained from Qiu for our selected high confidence genes and those previously calculated in [99] for the same cell type (figure 34).

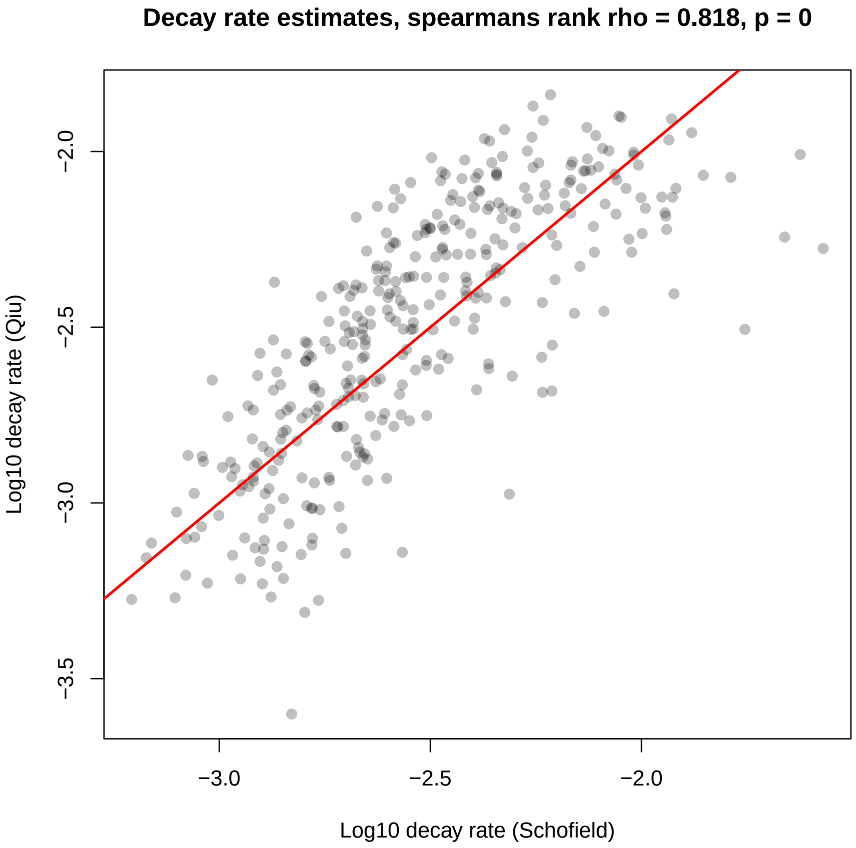


Figure 34: Decay rate (δ) estimates we obtained from Qiu for our high confidence gene set (584 genes which had $CV < 0.45$ for all parameter estimates) vs those calculated by Scofield for the same genes, with the Spearman's rank correlation statistics shown and the diagonal displayed (red line).

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 35 illustrates some expected correlations, showing that expression level, μ, correlates very strongly with the mean UMI count and that decay rate, 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 expression level and decay rate (μ and δ). Therefore, we instead compare a pair of genes (*ATF5* and *CAP1*) with very similar estimates for expression level and decay rate (μ and δ) but very different values of burst rate, a, (and therefore also burst size, b, and burst frequency, κ), with *ATF5* being expressed in a far more bursty fashion. Only two genes are analysed due to the scarcity of pairs in our 584 selected genes which are appropriate for comparison (matching expression level, μ, and decay rate, δ, but strongly differing burst rate, a). The estimates for the different parameters of these genes are given by table 4.

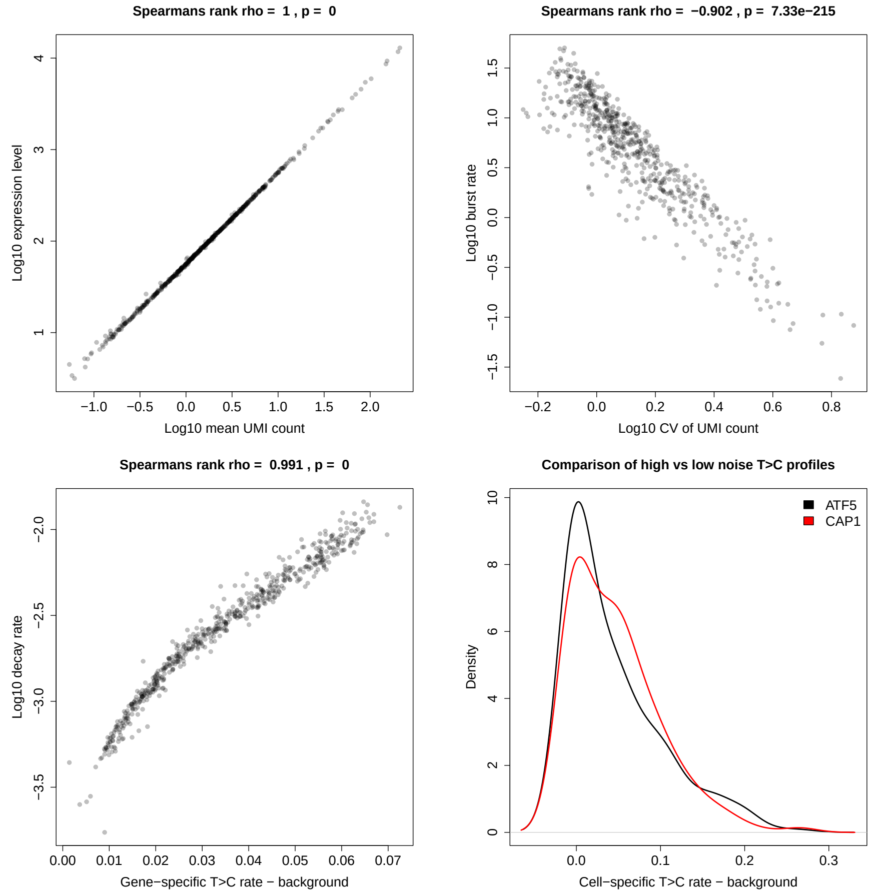


Figure 35: Correlations between statistics immediately observable in the Qiu dataset and inferred bursting parameter estimates are shown in the first three subplots, with Spearman's rank correlation strength (rho) and statistical significance (p) displayed. Each point corresponds to a gene, with the parameter estimate shown on the y-axis and the x-axis showing the statistic (mean of the UMI counts across cells, CV of the UMI counts across cells, or the mean T>C rate across cells in the 4sU dataset minus that in the control dataset, each for the given gene). 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 (due to having very different burst rates), respectively, but have closely matching expression levels and decay rates.

	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 4: Parameter estimates for the *ATF5* and *CAP1* genes.

The density plot in figure 35 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 expression level and decay rate (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 figure 35 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 (see section 3.1), except $\lambda_s = 0.001$, and $\alpha = 1$ to totally control for the effect of background T>C mutation and capture efficiencies, respectively. Datasets were simulated for a gene with high noise and another with low noise with parameter values set as shown in table 5.

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

Table 5: Parameter values for simulated high and low noise genes.

The differential flow from the surviving to new transcript pool for high and low noise genes is demonstrated in figures 36 and 37, 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 and unimodal transitions occurring under high and low noise conditions, respectively, and emphasising the key phenomenon which empowers 4sU scRNA-seq inference.

Differential transition from surviving to new transcripts

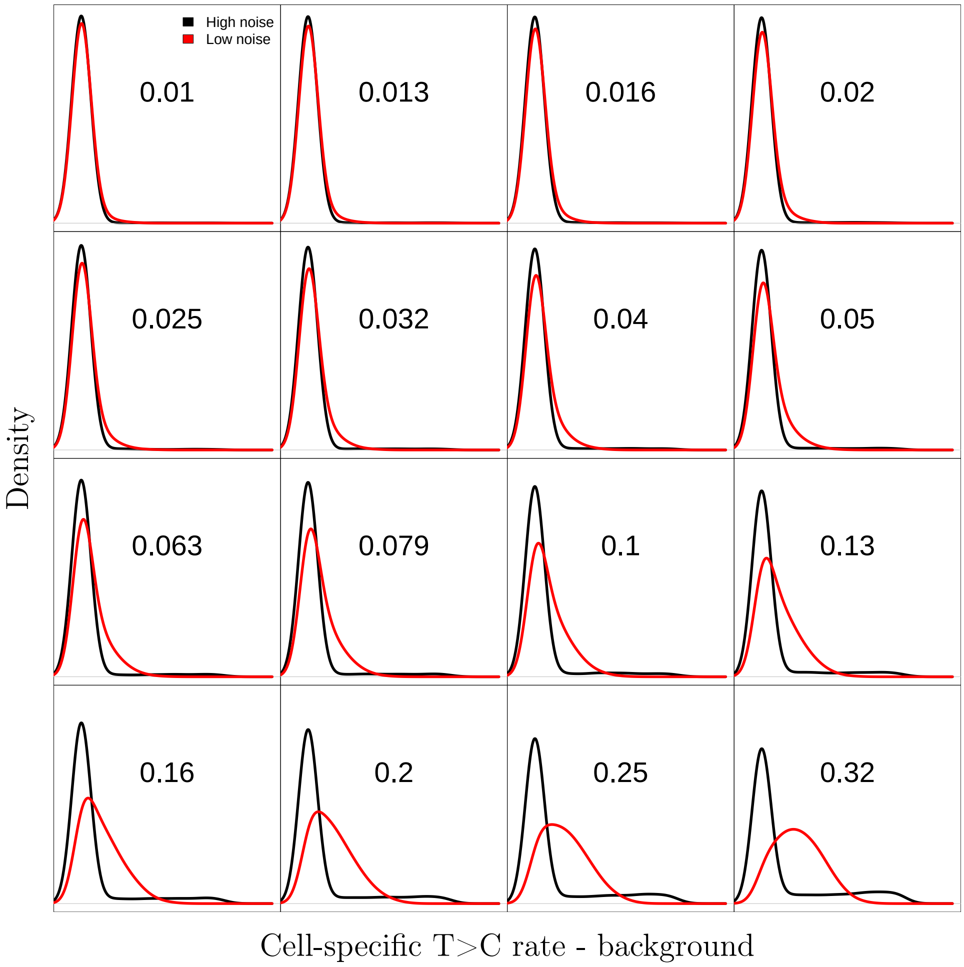


Figure 36: The differential transition from surviving to new transcript pool for high and low noise genes through the cell-specific T>C rate distributions for data simulated with different pulse durations, which are normalised to be in units of transcript lifetimes and displayed in the centre of each subplot. The x-axis values correspond to the T>C rate in the simulated cell minus the background T>C rate for the simulated gene, with values further on the left or right indicating a lower or higher proportion of the transcripts in the cell being new (synthesised after the 4sU pulse started), respectively. This figure shows the simulated transition across lower pulse durations.

Differential transition from surviving to new transcripts

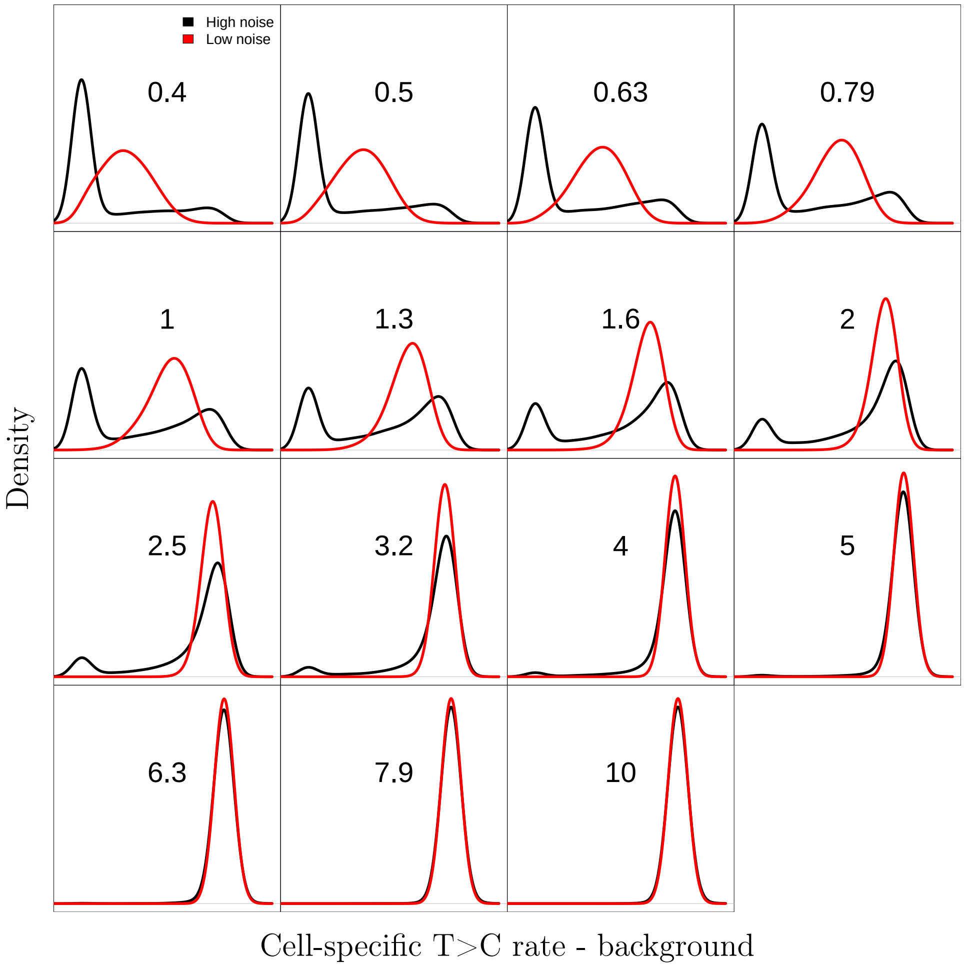


Figure 37: The differential transition from surviving to new transcript pool for high and low noise genes through the cell-specific T>C rate distributions for data simulated with different pulse durations, which are normalised to be in units of transcript lifetimes and displayed in the centre of each subplot. The x-axis values correspond to the T>C rate in the simulated cell minus the background T>C rate for the simulated gene, with values further on the left or right indicating a lower or higher proportion of the transcripts in the cell being new (synthesised after the 4sU pulse started), respectively. This figure shows the simulated transition across higher pulse durations.

3.3 Inference on simulated data

To validate the performance of our parameter estimation and confidence quantification and facilitate comparisons between our models on a genome-wide level, in-silico data was generated for each of the 12276 genes from the real Qiu dataset whose bursting dynamics were inferred (in section 3.2). The θ estimates obtained from Qiu were used as the "ground truth" parameter values for these simulations (see section 2.9), with inference being carried out both using model 1 and model 2 (or alternatively model 3). This simulation-based validation differs from the previously described model comparison analysis (figures 30 and 31) in that experimental settings, such as cell number, cell capture efficiency and sequencing depth, matched 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 each corresponding simulated gene.

First we turn to our inference of simulated with model 2 in order to confirm algorithm performance. The same selection of genes that was used for the real data was applied, selecting based on a maximum CV of 0.45 across all parameters. The correlations between the ground truth values and the parameter estimates derived from our sampled posteriors for the 422 selected simulated genes are shown in figure 38. The strong, tight correlations about the diagonal between estimates and true parameter values demonstrate the successful recovery of ground truths for all parameters. The error increases for genes with very low burst size or decay rate (b or δ), which reflects the increased CV for such estimates shown in figure 33. Higher error for burst rate, a, estimates primarily corresponds to those genes with very low, error-prone burst size, b.

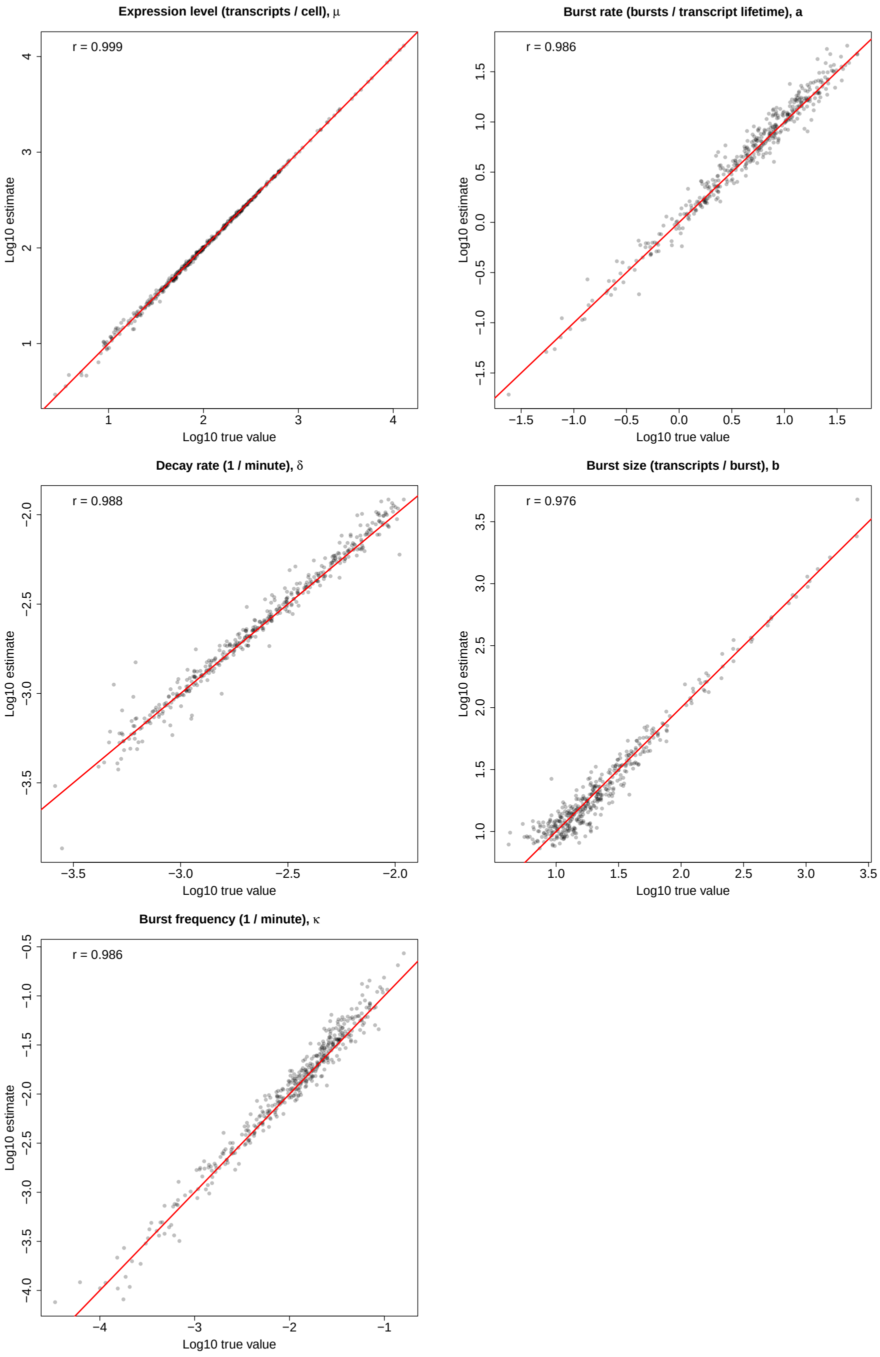


Figure 38: Correlations between "true" parameter values inferred from the Qiu dataset for real genes used as ground truth parameters with which to simulate data, and the estimates obtained from sampled posteriors for those simulated genes. 422 simulated genes are shown, selecting only those with $CV < 0.45$ across all five parameters shown. Red lines represent diagonals and the Pearson correlation coefficient (r) is indicated.

Having validated our model and algorithm performance, we can carry out a genome-wide comparison of models 1 and 2 under realistic conditions to compliment our idealised, single-gene demonstration (section 3.1). As before, we begin by selecting genes with $CV < 0.45$ but only considering dimensionless parameters, (expression level, burst rate and burst size), in order to compare model 1 and model 2/3, since model 1 does not converge on temporal parameters (figure 30). This leads us to select 549 and 584 genes based on our model 1 and 2 inference, respectively, for which we have sufficient confidence. We observe strong, tight correlations about the diagonal when plotting estimates vs true values for dimensionless parameters for both models 1 and 2 (figure 39), with both models performing equally well.

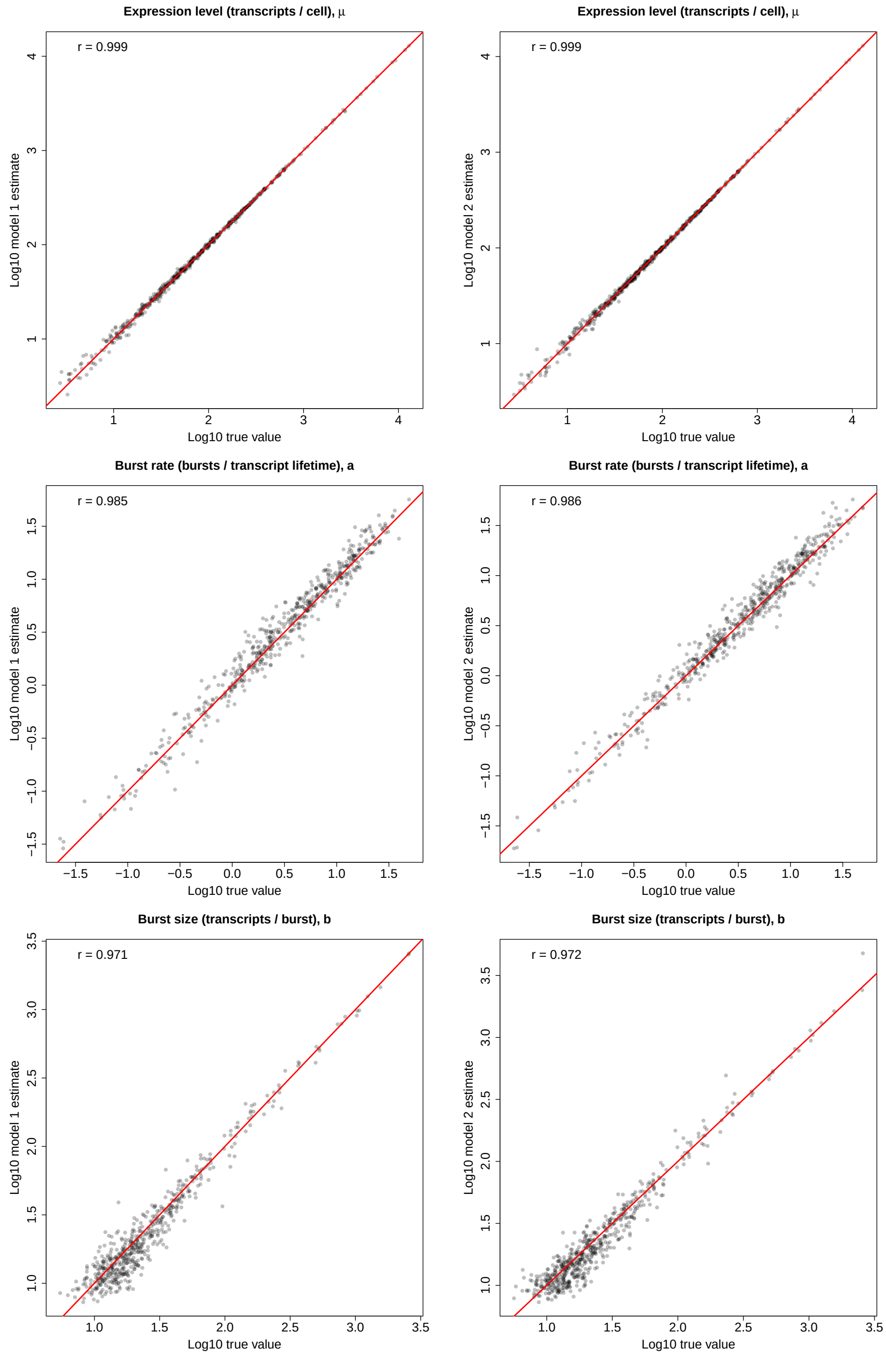


Figure 39: Correlations between "true" parameter values inferred from the Qiu dataset for real genes used as ground truth parameters with which to simulate data, and the estimates obtained from sampled posteriors for those simulated genes. Estimates for each simulated gene were inferred using model 1 (left) and then model 2 (right), showing 549 and 584 simulated genes, respectively, selecting only those with $CV < 0.45$ across the three (dimensionless) parameters shown. Red lines represent diagonals and the Pearson correlation coefficient (r) is indicated.

Taking a measure of the error our selected gene sets as the absolute value of the log fold-change between the true parameter value and our estimate, we again see no visual or statistical difference between models 1 and 2 for any of our dimensionless parameters (figure 40), thus further reinforcing that both approaches perform equally well for a given CV threshold.

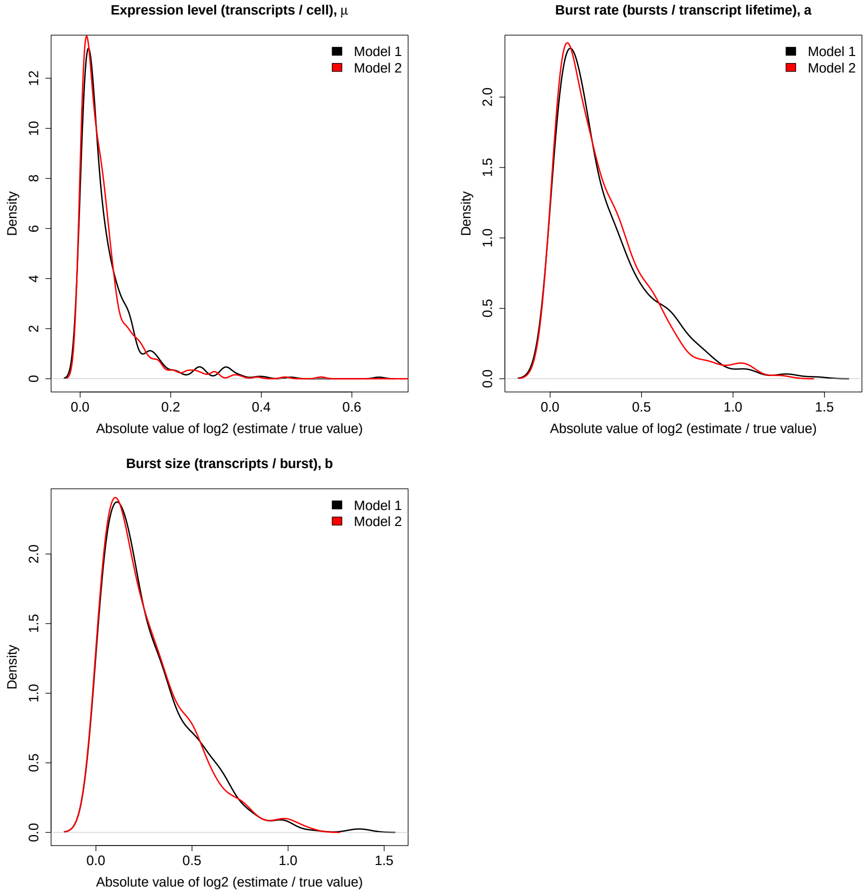


Figure 40: Comparison of the distribution of absolute value of the log fold-change between "true" parameter values inferred from the Qiu dataset for real genes used as ground truth parameters with which to simulate data, and the estimates obtained from sampled posteriors for those simulated genes. Estimates for each simulated gene were inferred using model 1 (left) and then model 2 (right), showing 549 and 584 simulated genes, respectively, selecting only those with $CV < 0.45$ across the three (dimensionless) parameters shown. There is no statistical difference between any of the three pairs of distributions based on Wilcoxon test.

This is confirmed by plotting the relationship between CV and the log fold-change between true parameter value and estimate (figure 41), indicating that the our quantification of confidence reflects true certainty in our estimate to the same degree for both models.

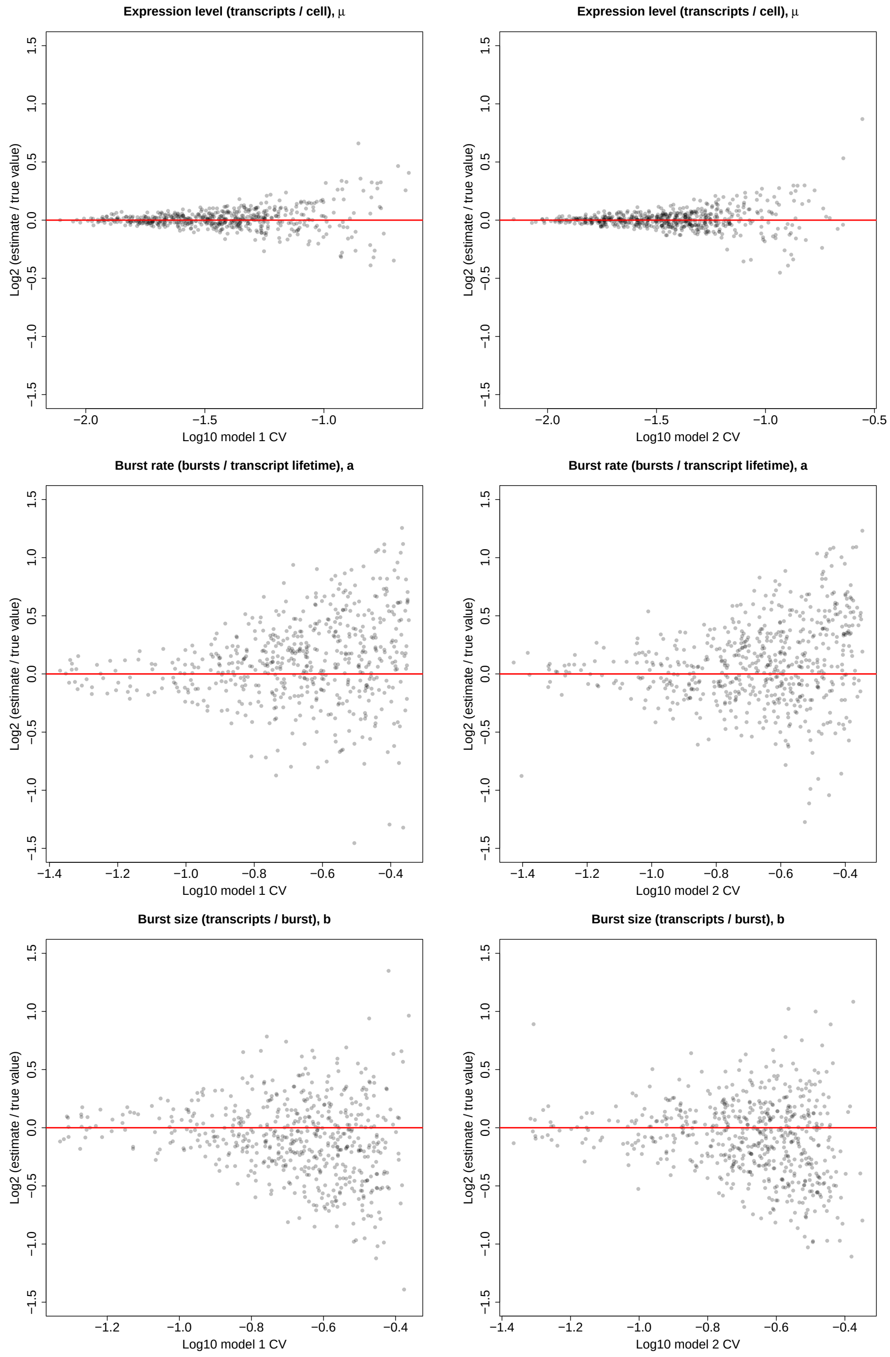


Figure 41: The log fold-change between "true" parameter values inferred from the Qiu dataset for real genes used as ground truth parameters with which to simulate data, and the estimates obtained from sampled posteriors for those simulated genes is shown on the y-axis. The CV obtained from sampled posteriors for those simulated genes is shown on the x-axis. Estimates and CVs for each simulated gene were inferred using model 1 (left) and then model 2 (right), showing 549 and 584 simulated genes, respectively, selecting only those with $CV < 0.45$ across the three (dimensionless) parameters shown.

Noting that despite no differences in performance between our models for a given CV threshold, there is a greater number of genes with estimates meeting our confidence criteria with model 2 inference than with model 1. Therefore, instead of selecting based on CV, we selected simulated genes based on their ground truth values, accepting those with expression level $\mu > 60$ and burst rate $a < 4$. Then comparing the CVs across the dimensionless parameters, those obtained with model 1 are clearly and statistically significantly greater than those obtained with model 2 (figure 42), indicating higher confidence in our model 2 estimates. This demonstrates that while a given CV threshold reflects equivalent performance in model 1 and 2, we obtain accurate and precise estimates for a greater number of genes when using model 2 in the analysis of moderately active genes with bursty transcription. Therefore, we further confirm the advantages of complimenting UMI counts with single-cell T>C data when inferring bursting dynamics with a genome-wide analysis using realistic conditions.

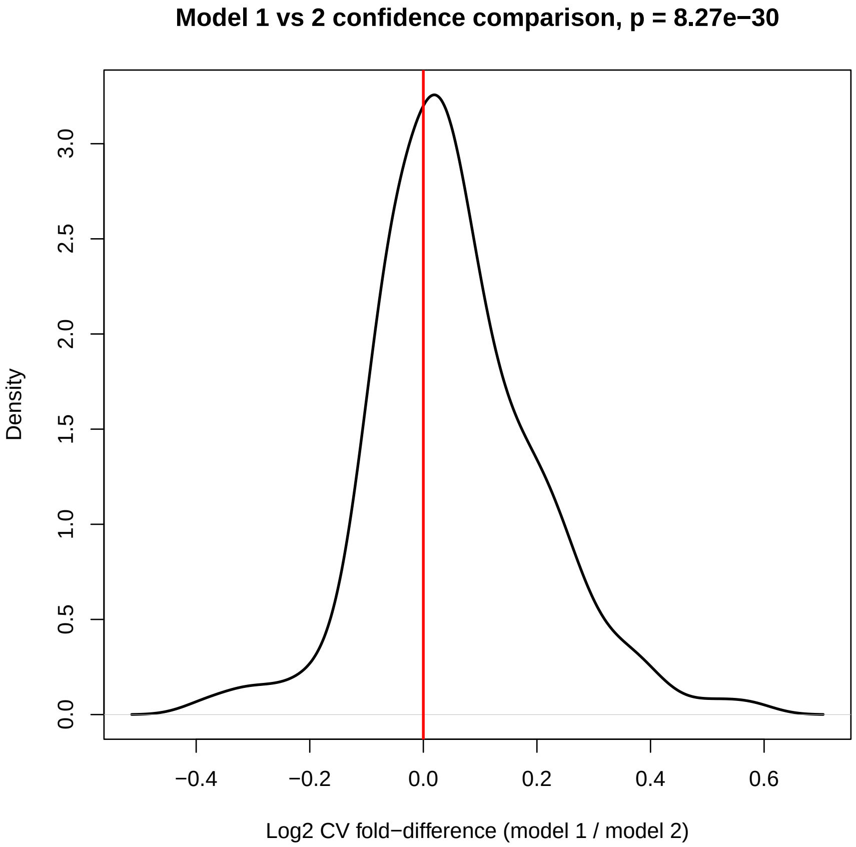


Figure 42: The log-ratio of the CVs obtained from sampled posteriors generated with models 1 and 2 for the three dimensionless parameters (expression level, burst rate and burst size) for genes simulated using "true" parameter values that were inferred from the Qiu dataset for real genes. Only simulated genes with a "true" expression level $\mu > 60$ and burst rate $a < 4$ are shown. The significant Wilcoxon test p-value (p) indicates asymmetry about zero (vertical red line).

3.4 Acceptance rates

Density plots of the MCMC acceptance rates (after burn-in removal) for all 12276 genes analysed in the Qiu dataset and the corresponding simulated genes are shown (figure 43). This diagnostic indicates that overall the desired mixing behaviour is achieved in all cases for each of the three parameters in our chosen parametrisation, with the ideal acceptance rate being roughly 0.574 [139].

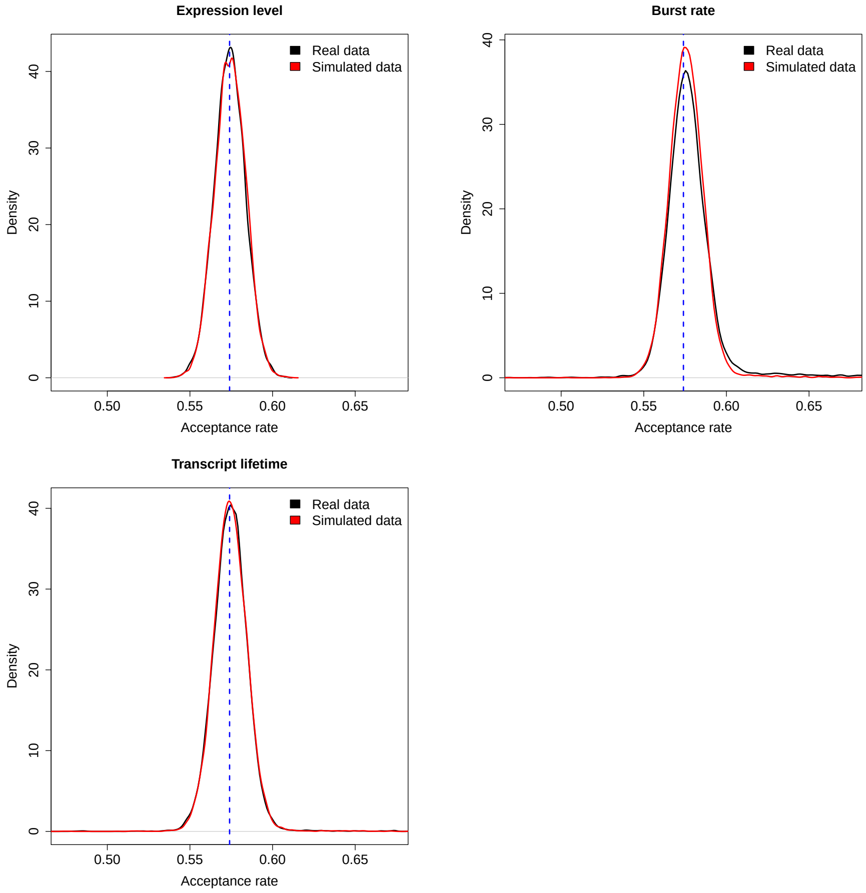


Figure 43: Markov chain acceptance rates (excluding burn-in) for inference on the Qiu dataset and the corresponding dataset simulated using ground truth values inferred from Qiu, for each of the three parameters in our chosen parametrisation, with the vertical dashed lines indicating the optimal acceptance rate (0.574, the proportion of proposals which accepted).

3.5 Discussion

Thus far we have demonstrated that a fully integrated inference approach, which maximally exploits the information available in 4sU scRNA-seq data, results in optimal estimation for all parameters governing the transcriptional bursting dynamics (figures 30 and 31). This stems from the cell-specific T>C counts being shaped by the transcriptional noise in addition to turnover rate and expression level (figures 36 and 37), building on previous results which substituted a measure of the cell-cell variation in T>C rate as a proxy of burstiness [106]. Running the algorithm in a genome-wide manner on published data showed that we have highest confidence for noisy, highly expressed genes with transcript lifetimes similar to the 4sU pulse duration (figure 33). Genome-wide inference on simulated data demonstrated the capacity of the algorithm to robustly recover known parameter values with realistic biological and technical conditions (figure 38). Comparison of inference with models 1 and 2 on realistic simulated data demonstrated that a higher number of genes with high confidence parameter estimates in dimensionless parameters may be recovered using 4sU scRNA-seq data than with conventional scRNA-seq data (figure 42), despite 4sU labelling not being required for convergence on these parameters, confirming the advantages provided by synergising 4sU labelling and single-cell resolution. The improvement in inference with model 2 over model 1 would likely be stronger for datasets with deeper sequencing, more cells and higher capture efficiencies.

An important point to note is that the role of extrinsic noise is neglected when estimating parameters. Our model assumes that the conditions of the cells are identical, although variation in cellular conditions may be caused by asynchronised cell-cycles. The expression of certain genes is necessarily highly sensitive to extrinsic sources of noise, such as those involved in cell-cycle regulation, which are required to be more or less expressed at specific stages, resulting in our parameter estimates for these genes likely being strongly influenced by extrinsic noise [49, 75]. A related issue is that of copy number variation of (nuclear) genes between one and two throughout the cell cycle. Therefore, an unknown portion of the cells have double the copy number of each gene. Even if the true parameter values associated with each copy of a gene remain constant throughout the cell-cycle, the inferred burst frequency would be increased relative to the true value since the model assumes a single copy per cell. Finally, the size of cells varies depending on cell-cycle phase. If the overall concentration of transcripts is to remain constant, higher expression levels would be required in larger cells, leading to variability in the underlying parameter values for a given gene between cells in order to achieve differing expression levels. A solution would be to separate the cells by cell-cycle phase with fluorescence-activated cell sorting before sequencing [96], or to make use of allele-specific/sensitive scRNA-seq in conjunction with metabolic labelling [75, 97]. Other sources of extrinsic noise have also been shown to impact the chromatin landscape and transcriptional bursting, such as cyclical changes in the levels of the H3K27ac HM at promoters of genes related to the circadian rhythm in mouse cells, which directly translated to changes in the burst frequencies of those genes [74].

4 Biological findings

4.1 High expression level genes

Now we return to the estimates obtained from the real dataset on which inference was carried out with model 2 (see section 3.2). Correlating the parameter estimates against each other for our 584 genes 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 are also somewhat constrained (figure 44). There may be biological upper limits on burst frequency, 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 [95]. 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 [149, 150]. Estimates for burst frequency and decay rate (κ and δ) are also positively correlated, perhaps because the cells are only able to tolerate a certain degree of noise in the expression of any given gene, otherwise too small a proportion may express the gene for the function to be fulfilled. This may manifest as a correlation between these two parameters to stabilise burst rate, a, and thus the transcriptional noise.

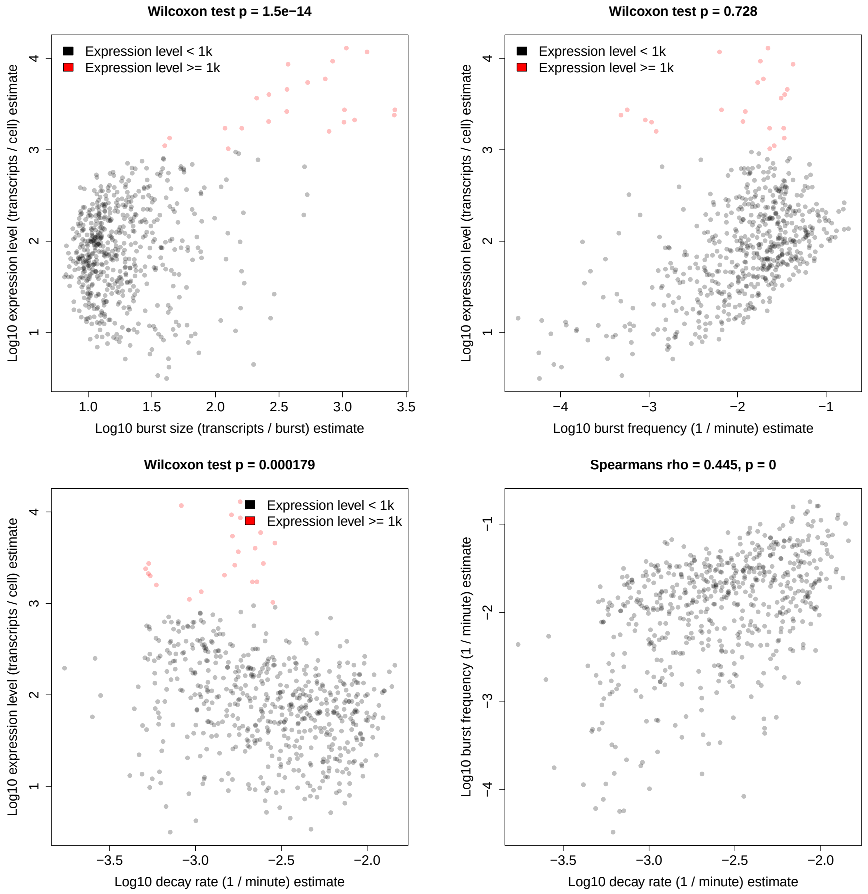


Figure 44: Estimates obtained via inference on the Qiu dataset using model 2 (or 3) for different parameters plotted against each other. Only our high confidence gene set is shown (584 genes which had $CV < 0.45$ for all parameter estimates). Statistical significance of difference in burst size, burst frequency and decay rate (b, κ and δ) for genes with very high expression level ($\mu \geq 1000$) 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 burst frequency against decay rate (κ against δ).

4.2 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 HMs (see section 2.1.2). 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 active HMs analysed, the profiles generally fall into the two previously described categeories [77], being either predominantly promoter-localised (H3K4me2, H3K4me3, H3K9ac, H3K27ac) or gene body (GB)-localised (H3K4me1, H3K36me3, H3K79me2, H4K20me1). 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.

4.2.1 Promoter-localised histone modifications

Metagene comparison reveals position-dependent associations for promoter-localised HMs, using H3K4me2 as an example (figure 45). It appears that HM presence at the promoter and through the GB is associated with increased expression level, μ, and also burst rate, a, while increased burst frequency, κ, is specifically associated with promoter but not GB presence. Conversely, presence through the GB excluding the promoter region appears associated with increased burst size, b, and reduced decay rate, δ.

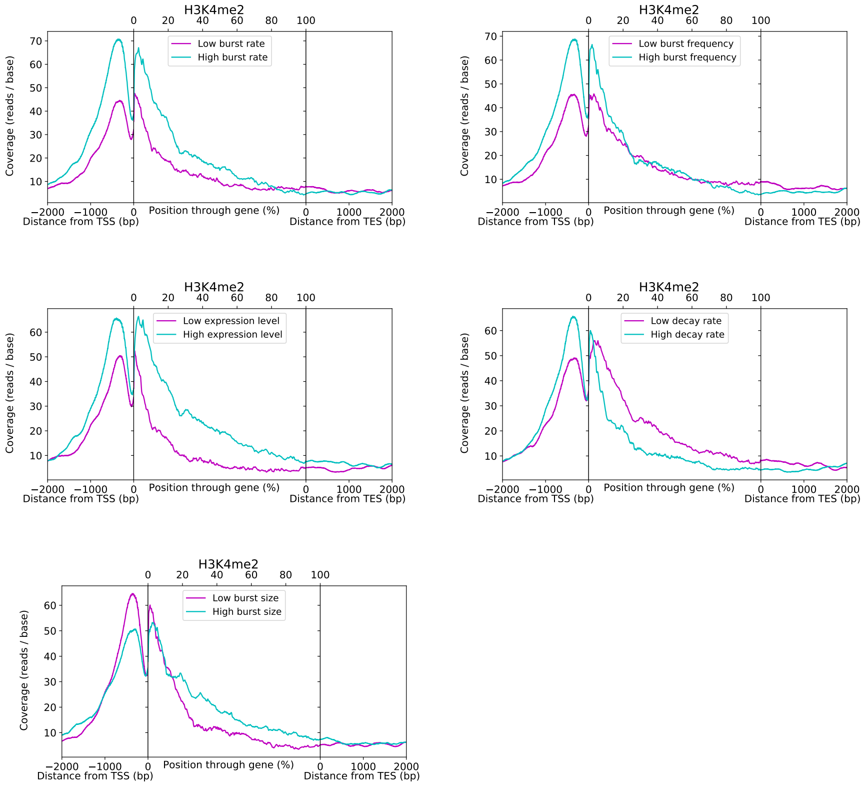


Figure 45: Metagene plots of H3k4me2 coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

This analysis builds upon a previous scRNA-seq study which correlated bursting parameter estimates with HM localisation by averaging the ChIP-seq coverage from 2000 bp upstream of the TSS to the TES for each gene [77]. They were unable to obtain estimates of burst frequency or decay rate (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 [77] 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 burst frequency and decay rate (κ and δ) afforded by 4sU scRNA-seq. For promoter-localised HMs, they report positive associations between HM presence and both burst rate and burst size (a and b), whilst we demonstrate that the association with burst size, b, is specific to the GB. We confirm that the association with burst rate, a, holds throughout both the promoter and GB, but show that this is a result of a promoter-specific positive burst frequency, κ, association and a GB-specific negative decay rate, 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 (see section 2.1.2), taking the average value from 2000 bp upstream of the TSS to 5% through the GB (-2000:5%) and from 5% through the gene body 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 (figure 46), as well as confirmed the statistical significance of the suspected associations (figure 47). Figure 47 also confirms that there is no statistical evidence of an association between decay rate or burst size and HM presence around the TSS or between burst frequency and HM presence throughout the GB downstream of the TSS, which aligns with our expectations based on the metagene profile (figure 45).

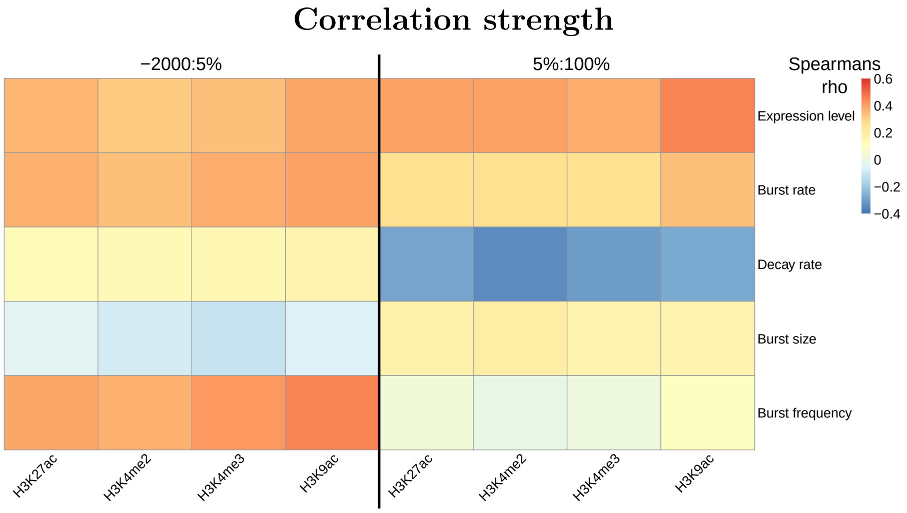


Figure 46: Heatmap showing the Spearman's rank rho as the heat intensity value for the correlations between bursting parameter estimates inferred from Qiu and the mean promoter-localised HM coverage values across the -2000:5% and 5%:100% regions for selected genes which had $CV < 0.45$ for all parameter estimates. More intense red or blue colouration indicates a stronger positive or negative correlation, respectively, while neutral indicates no/weak correlation.

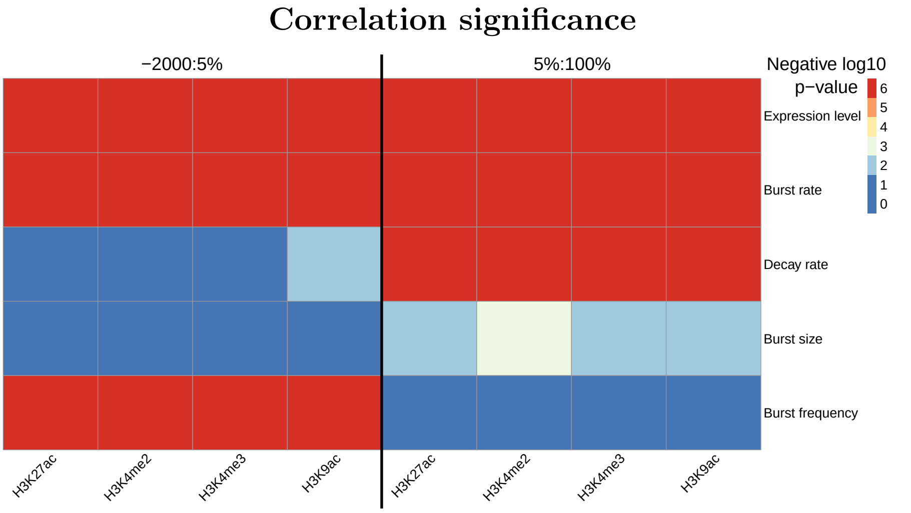


Figure 47: Heatmap showing the Spearman's rank p-value (adjusted for multiple hypothesis testing) as the heat intensity value for the correlations between bursting parameter estimates inferred from Qiu and the mean promoter-localised HM coverage values across the -2000:5% and 5%:100% regions for selected genes which had $CV < 0.45$ for all parameter estimates. 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 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 the recruitment of splicing factors [151, 152]. 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 [71].

Metagene analyses of the other promoter-localised HMs that were represented by H3K4me2 show the similar profiles of H3K4me3 (figure 48), H3K9ac (figure 49) and H3K27ac (figure 50).

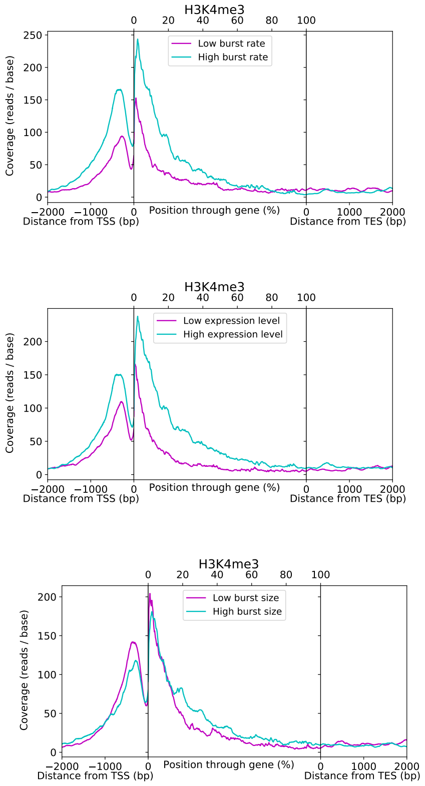


Figure 48: Metagene plots of H3K4me3 coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

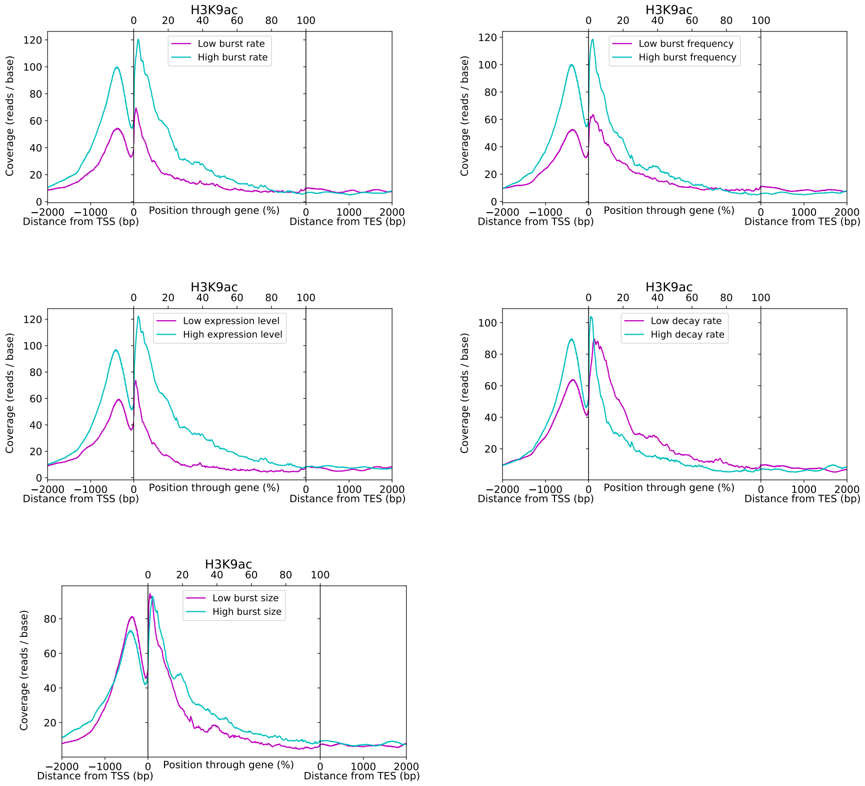


Figure 49: Metagene plots of H3K9ac coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

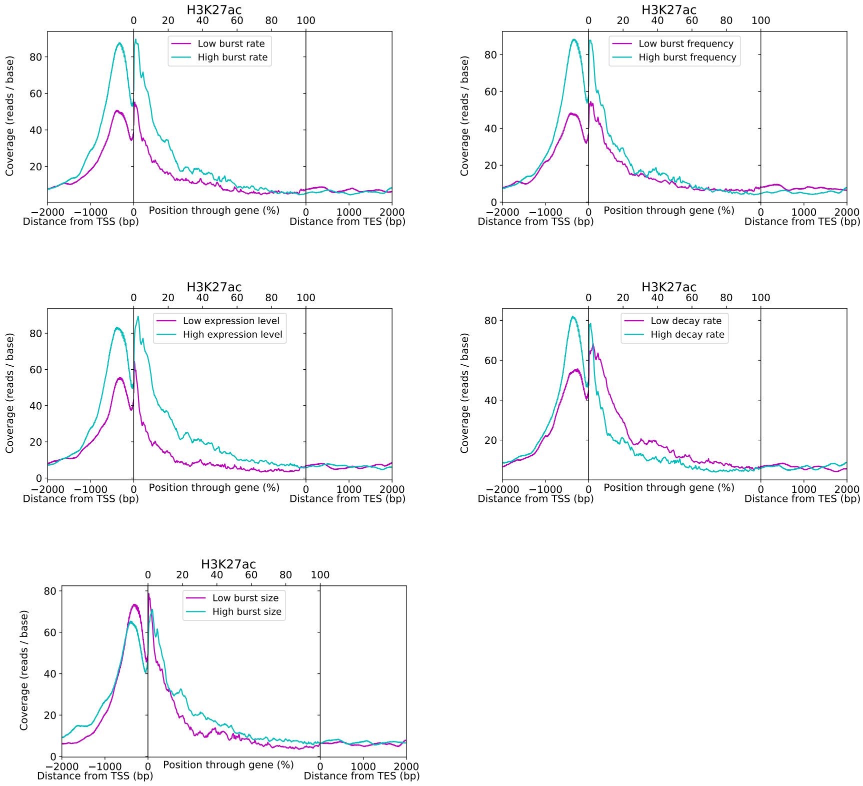


Figure 50: Metagene plots of H3K27ac coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

4.2.2 Gene body-localised histone modifications

A similar analysis of the GB-localised HMs was also carried out, where we use H3K36me3 as a representative example, although their metagene profiles and bursting associations are somewhat more diverse than with the four promoter-localised HMs. H3K4me1 was categeorised as being primarily promoter associated in [77] but we find its connections to transcriptional dynamics instead to be contingent upon its presence throughout the GB, and have therefore reclassified it for this context. The profiles of H3K36me3 halved by the different bursting parameter estimates as before (figure 51) indicate that presence throughout the GB and around the TES seems to be associated with increased expression level, burst rate and burst frequency (μ, a and κ) in a uniform manner. No association with burst size or decay rate (b or δ) is apparent, suggesting that this HM is associated with increased expression level, μ, purely through increased burst frequency, κ. In this case, we are able to support the previously reported correlation with burst rate, a, [77], and confirm that the inability of scRNA-seq data to distinguish burst rate and burst frequency (a and κ) did not skew the final conclusions by quantifying the strength (figure 52) and statistical significance (figure 53) of H3K36me3 and the other GB-localised HMs which it represents (H3K79me2 and H4K20me1). It should be noted, however, that based on the metagene analysis, while both H3K79me2 and H4K20me1 appear to be primarily associated with increased expression level, burst rate and burst frequency (μ, a and κ), along with H3K4me1 they look to have a positive and negative association with burst size and decay rate (b and δ), respectively, when found throughout the 20%:100% region. However, this is statistically significant only for H3K4me1 (figure 53), which also has no association with burst frequency, κ, and no significant correlation with burst rate, a. Therefore, for the GB-localised HMs, H3K36me3, H3K79me2 and H4K20me1 can be regarded as similar in their associations with bursting dynamics, while H3K4me1 is an outlier. The regions of association for the GB-localised HMs vary to a degree, as dictated by the metagene analysis, with the values used for the correlation analysis shown in figures 52 and 53 being averaged across 0%:2000 (TSS to 2000bp downstream of the TES) for H3K36me3 and H4K20me1, -2000:100% for H3K79me2 and 20%:100% for H3K4me1. The metagene analyses of H3K79me2 (figure 54), H4K20me1 (figure 55) and H3K4me1 (figure 56) are shown.

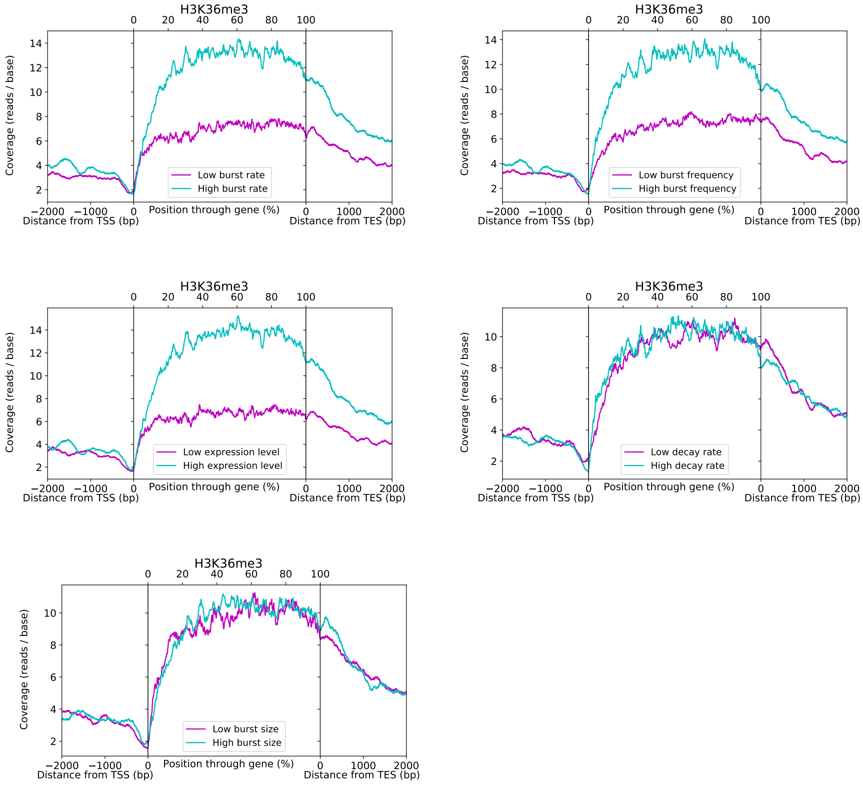


Figure 51: Metagene plots of H3K36me3 coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

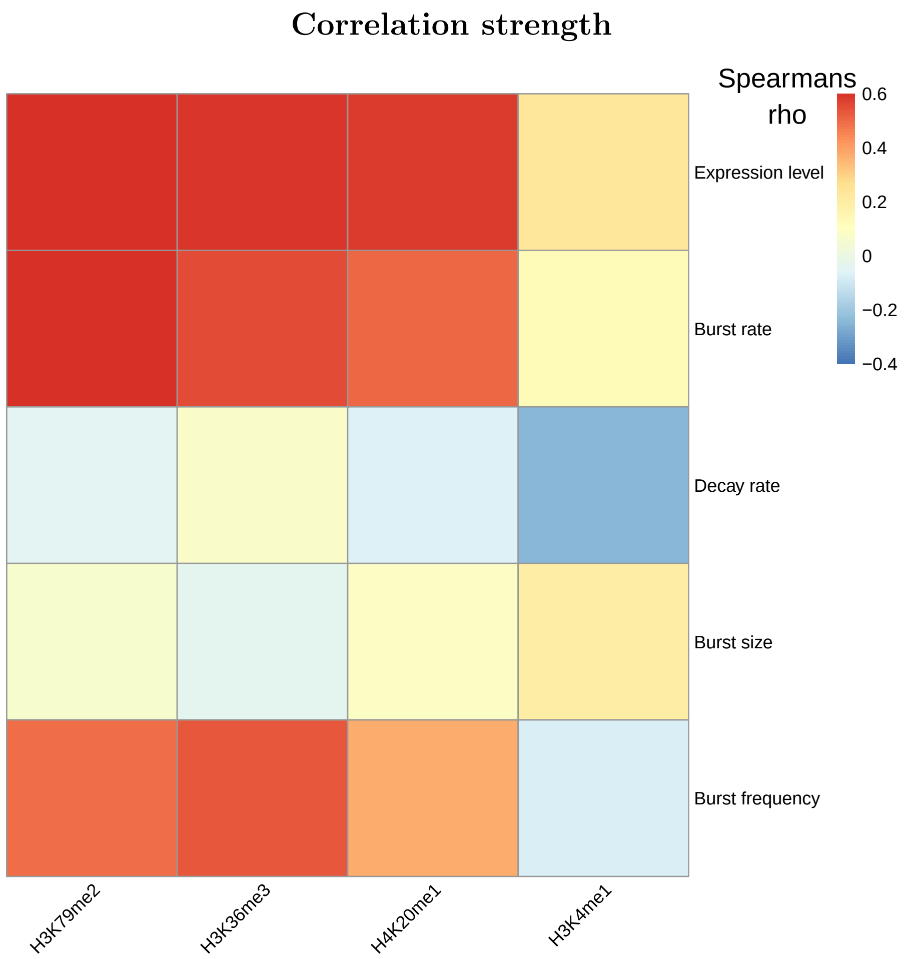


Figure 52: Heatmap showing the Spearman's rank rho as the heat intensity value for the correlations between bursting parameter estimates inferred from Qiu and the mean GB-localised HM coverage values across 0%:2000 for H3K36me3 and H4K20me1, -2000:100% for H3K79me2 and 20%:100% for H3K4me1 for selected genes which had $CV < 0.45$ for all parameter estimates. More intense red or blue colouration indicates a stronger positive or negative correlation, respectively, while neutral indicates no/weak correlation.

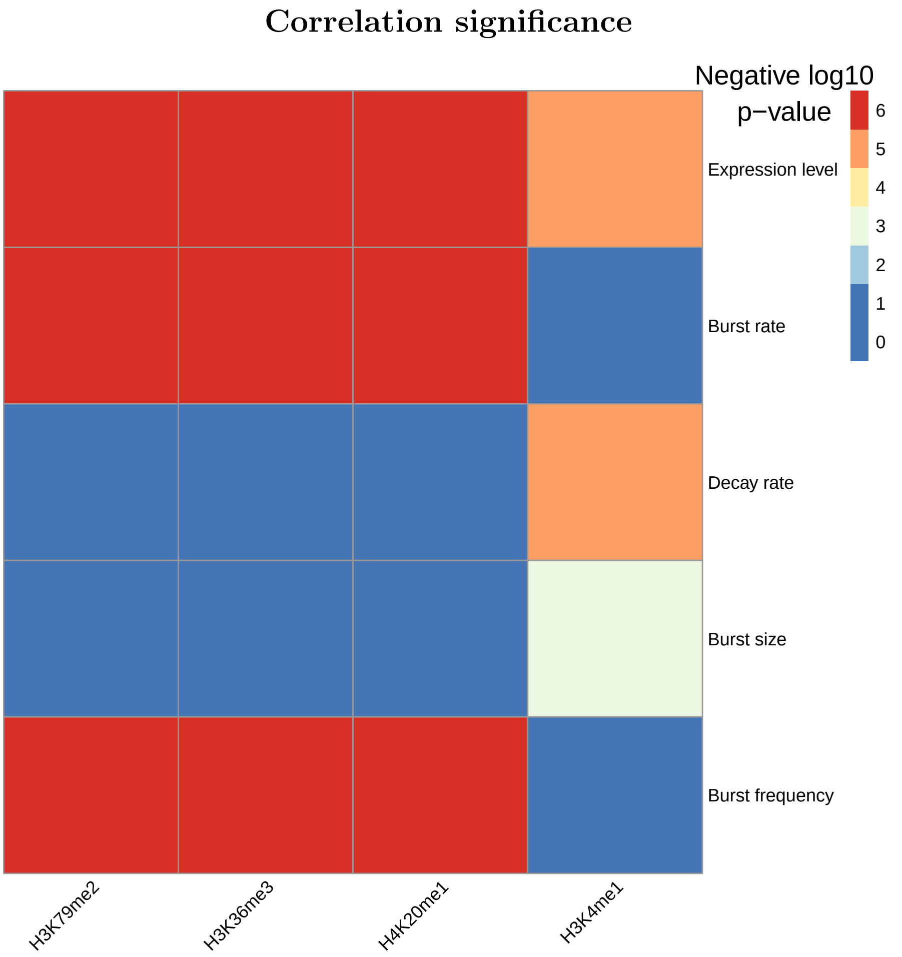


Figure 53: Heatmap showing the Spearman's rank p-value (adjusted for multiple hypothesis testing) as the heat intensity value for the correlations between bursting parameter estimates inferred from Qiu and the mean GB-localised HM coverage values across 0%:2000 for H3K36me3 and H4K20me1, -2000:100% for H3K79me2 and 20%:100% for H3K4me1 for selected genes which had $CV < 0.45$ for all parameter estimates. 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 $p \leq 10^{-6}$.

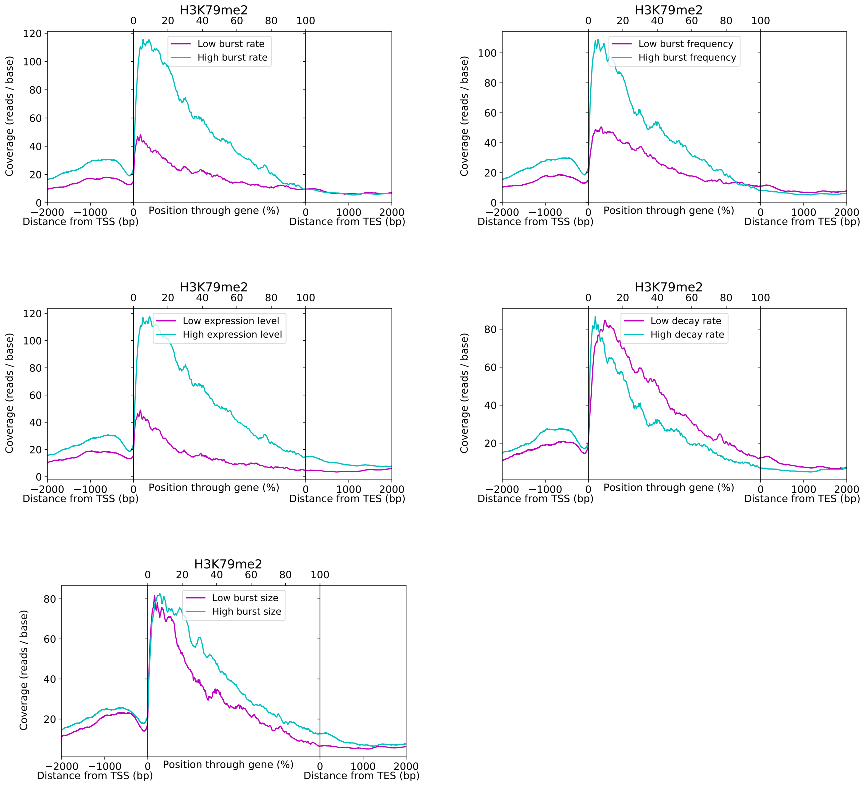


Figure 54: Metagene plots of H3K79me2 coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

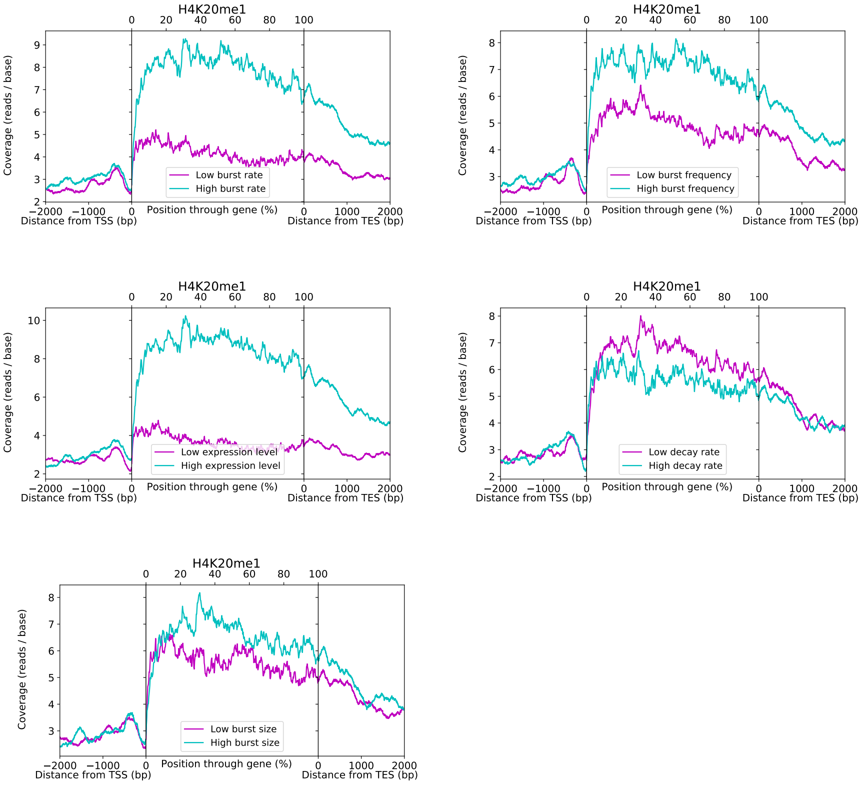


Figure 55: Metagene plots of H4K20me1 coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

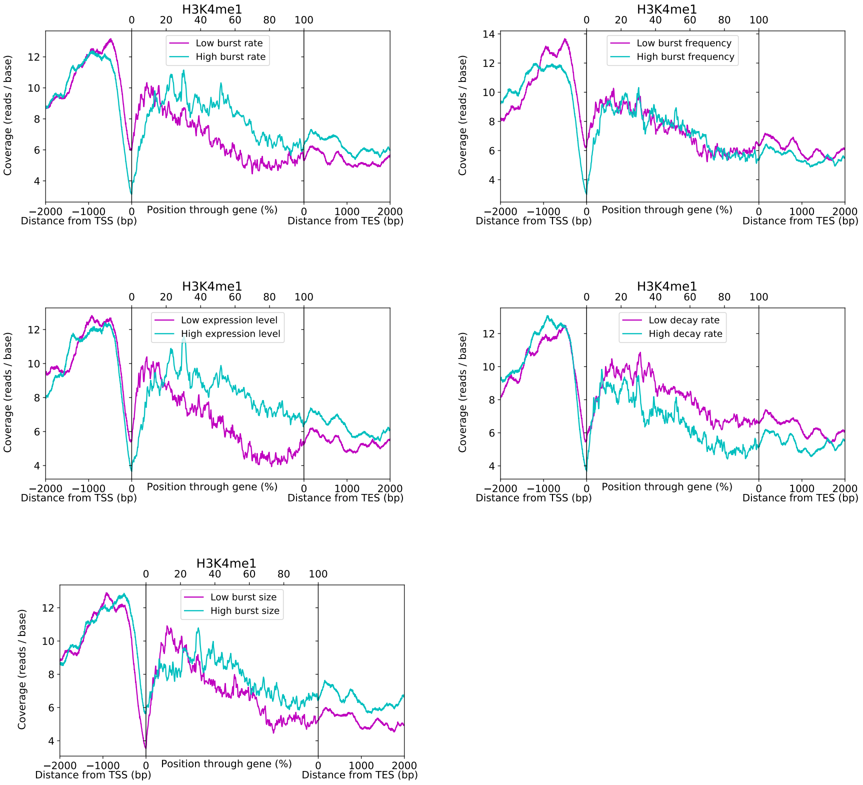


Figure 56: Metagene plots of H3K4me1 coverage, comparing profiles for the top and bottom 50% of selected genes, which had $CV < 0.45$ for all parameter estimates inferred from Qiu, when split according to their estimates for each parameter, denoted by high and low, as indicated.

4.3 Discussion

Using the genome-wide transcriptional bursting parameter estimates obtained in section 3, genes with very high expression levels are revealed to be primarily mitochondrial, as was previously reported [105]. Furthermore, such genes were shown to achieve high expression levels through increased burst size rather than burst frequency (figure 44), potentially due to biophysical limitations on the rates of switching between active and inactive states, which would make longer bursts the more favourable option, as was seen with *MYC*-

driven transcription [149, 150]. As opposed to previous reports of mESC cells [95], we found significant variation in transcript decay rates in K562 cells, emphasising the importance of accounting for this explicitly, which lead to the perhaps surprising finding of a positive correlation between burst frequency and decay rate (figure 44). We offer two hypothesis to explain this observation; one evolutionary and one mechanistic. Perhaps there are selection pressures acting to constrain the transcriptional noise levels, which is determined by burst rate as the ratio of burst frequency and decay rate, with extreme high or low values being unfavourable in, for example, ensuring the correct proportion of cells express a gene stochastically to induce their differentiation [61, 153]. 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 [1], resulting in reduced transcript stability.

Genome-wide analysis of HMs and transcriptional bursting was guided by a previous study [77], but revealed many additional complexities by combining metagene analyses with 4sU scRNA-seq inference results rather than conventional scRNA-seq. GB-localised HMs were generally associated with more frequent bursts (figure 51), thereby reducing transcriptional noise, which aligns with previous results regarding the HM-mediated recruitment of elongation factors to the GB leading to increased burst frequency [96]. On the other hand, promoter-localised HMs only exhibit an association with burst frequency when present around the TSS, instead becoming associated with reduced decay rate when present further downstream, maintaining the association with burst rate across the whole gene (figure 45). Since HMs have been shown to be important for pre-RNA processing via recruitment of splicing factors and modulation of RNAP elongation rate [151, 152], this may offer an explanation to the observed association with reduced decay rate and promoter-localised HM presence through the GB by enhancing the likelihood of correct co-transcriptional 5' capping, splicing, polyadenylation and transcript folding to increase transcript stability.

Another key observation is the association between promoter-localised HMs and increased burst size when found throughout the GB but not the TSS region (figures 46 and 47). One possible explanation is that their presence further downstream may facilitate TSS-TES cross-talk, allowing the same RNAP to repeatedly transcribe during the active period by being immediately recycled and reinitiating after termination [71]. Another hypothesis is that downstream presence of promoter-localised HMs reflects that locations of alternative TSSs, which are capable of transcribing simultaneously when the chromatin landscape of the gene facilitates active transcription, resulting in increased burst size since bursts from alternative TSSs would be non-independent. 483 out of the 505 genes used in the metagene analyses had alternative TSSs at multiple positions based on inspection of the gtf, which reinforces this as a plausible explanation. An association with increased burst size is also observed with H3K4me1 presence in the GB (figure 56), although the profiles observed point to this HM being something of an outlier in the sense that it's set of associations don't conform to those observed for either the promoter-localised or it's fellow GB-localised HMs, which are otherwise consistent within their own groups. H3K4me1 is known to be strongly associated with enhancers [154], indicating that its presence throughout the portion of the GB downstream of the TSS may signify intronic enhancers [155], which could result in larger bursts by maintaining the active state of the gene they are contained within for longer periods. Indeed, this is a limitation of the metagene analyses presented here, that no distinction between the exonic and intronic regions of the gene are made, otherwise we would expect to see the association of H3K4me1 with burst size being restricted to intronic not exonic regions. It would be beneficial for future analyses to dissect the different regions of genes into separate metagene analysis to check whether other observed associations are specific to exons or introns, especially considering that exons and introns are known to exhibit differing densities of HMs, such as H3K36me3 being relatively enriched in exons [40].

One caveat of the analysis presented here is that the relatively small number of genes available for statistical analysis of their association with HMs prevented further subdivision of the gene set. For example, burst size (b) has been shown in previous papers to negatively correlate with gene length [95]. Therefore, the position-specific associations between promoter-localised HMs and burst size may be stronger for short genes than long genes. Additionally, there are possible issues introduced by the selection of genes with high confidence parameter estimates, which biases the selection towards genes which tend to be more highly expressed in K562 cells. From a gene ontology perspective, this results in genes with related functions, like haemoglobin and myoglobin synthesis genes and others related to oxygen transport in the blood being preferentially selected due to K562 cells being bone marrow cancer cells capable of exhibiting characteristics associated with erythrocytes and other blood cell types [123, 124]. Therefore, the bursting-HM associations observed may be specific to genes with certain functions which may have similar regulatory mechanisms rather than being genome-wide features and also might not universally apply to alternative cell types and organisms.

5 Discussion

5.1 Exploiting the synergy

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 (figure 33). 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 dimensionless parameters beyond the level obtainable with conventional scRNA-seq and also improves inference of temporal parameters compared to combined analysis of scRNA-seq and bulk SLAM-seq data (figures 30 and 31). 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 [106]. However, as clearly demonstrated by simulations showing the flow from the surviving to new transcript pool (figures 36 and 37), 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. These benefits are confirmed by comparison of genome-wide parameter inference with model 1 vs 2 on data simulated under realistic conditions, with the results reflecting a reduction in uncertainty achieved when incorporating all of the available information in the data (figure 42).

5.2 Genome-wide dynamics

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 [105], we find that the genes with very high expression levels are primarily mitochondrial genes (figure 44). 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 [95], 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 [149, 150]. 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 [95], we found significant variation in K562 cells which was important to account for in order to properly estimate burst frequencies. 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 (figure 44). 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 [61, 153], 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 [1]. This would would reduce transcript stability and cause the observed correlation.

5.3 Histone modifications

Examining the relationship between bursting parameters and HMs genome-wide produced results consistent with but advancing upon previous work [77]. 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 (figures 51 and 45, respectively). 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 (figures 46 and 47). 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 [151, 152]. This may increase RNA stability by reducing the probability of incorrect splicing or polyadenylation, and by allowing time for the nascent transcript to fold properly at the appropriate times by polymerase pausing, for example, which is also crucial for transcript stability by facilitating 5' capping [1].

Presence of promoter-localised HMs throughout the GB but not at the TSS is also associated with increased burst size (figures 46 and 47). 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 [71]. 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, all capable of initiating transcription. Indeed, examination of the gtf indicates that 483 out of the 505 genes used in the metagene analyses do have $1 <$ TSS, which supports this hypothesis. This would result in an association with burst size rather than burst frequency if bursts from the different TSSs of a given gene are not independent of each other, as would be expected since if, for example, heterochromatin is cleared from across a gene then would permit transcription at all TSSs simultaneously. If transcription is able to occur from one TSS at a given moment then it is likely to also be possible at the other TSSs. This mechanism would allow more total initiation events to occur in a given time window, until the gene is repressed again by heterochromatin formation, resulting in increased burst size.

5.4 Future directions and caveats

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 longer and shorter pulse durations would facilitate estimates for genes with long and short transcript lifetimes, respectively, along with everything in between. Data from control datasets without 4sU could also easily be folded in to improve inference. Taking the Qiu datasets we analysed as an example, we only used the negative control scRNA-seq dataset for calculating gene-specific background T>C rates (figure 14) in order to streamline the analysis and demonstrate what can be accomplished using a single dataset/experiment. However, the UMI count information could have been included via likelihood function 1 relatively straightforwardly to make the inference more robust, under the assumption that there are no underlying differences between the control and 4sU dataset other than the 4sU pulse. In a similar manner, one could also include other auxiliary datasets within our framework, such as incorporating bulk SLAM-seq T>C count data to aid estimation via likelihood function 3. 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.

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 [96], and/or by using allele-specific/sensitive scRNA-seq approaches combined with metabolic labelling [75, 97]. Another point worth noting is the assumption of instantaneous bursting, which enables analytical progress towards exploitation of the 4sU-sc synergy via construction of model 2. This assumes that burst duration has a negligible effect on the transcriptional dynamics and is support by previous experimental results across the vast majority of genes analysed for which the gene spends much more time in the inactive than active state, $k_{off} >> k_{on}$ [95]. This assumption corresponds to a three parameter model of transcriptional bursting, unlike the standard four parameter model (figure 7), collapsing transcription rate (β) and gene inactivation rate (k_{off}) into the burst size parameter (b). However, there is likely a small subset of genes which spend a similar amount of time in the active and inactive states, where $k_{off} \approx k_{on}$, in which case the burst duration will strongly influence the transcript count distribution, dynamics and noise. In these cases, it will be appropriate to switch to an alternative treatment to analyse corresponding 4sU scRNA-seq data, replacing the negative binomial distribution in equation 1 with a Poisson-Beta compound distribution [54] which can account for variable time spent in inactive and active gene states

$$P(m) = \int f_{Pois}\left(m|\beta p/\delta\right) f_{Beta}(p|k_{on}/\delta, k_{off}/\delta)dp$$

This would alter likelihood function 1, which could then be combined with likelihood function 3 (not likelihood function 2 since it also assumes instantaneous bursts) essentially giving a modified version of model 3 capable of estimating transcription rate, decay rate, activation rate and repression rate (β, δ, k_{on} and k_{off}) or some subset depending on the location of the gene within parameter space.

One must also consider both the type of cells and the conditions in which they were cultured, and how these factors influence the results presented and the generalisability of conclusions drawn. Being a cancer cell line derived from a patient with chronic myeloid leukimia [122], K562 cells are capable of exhibiting characteristics associated with blood cells such as erythrocytes, monocytes and granulocytes, including the production of proteins related to oxygen transport such as haemoglobin [123, 124]. They also massively overexpress Aurora kinases relative to non-cancerous cells, which are required for mitosis in healthy cells, leading to uncontrolled cell division in K562 cells [125]. Overall, the transcriptome in K562 cells is highly skewed towards genes involved in blood cell functions as well as those required for cancerous cell divisions, which may have distinct regulatory programmes, resulting in parameter estimate profiles which may be biased towards genes associated with these biological functions, with the regulatory patterns for cell division especially deviating considerably from that of non-cancerous cells. An additional complication regarding uncontrolled cell division is that transcripts of mitosis-related genes must be degraded at specific points to allow progress from one cell-cycle stage to another, which under a rapid, cancerous cell-cycle regime could alter our decay rate estimates associated with these genes compared to in healthy cells [156]. Overall, this means that caution should be used when generalising to non-blood and especially non-cancer cell types the biological conclusions drawn regarding correlations of parameter estimates with each other (figure 44) and associations between HMs and bursting (figures 46, 47, 52 and 53). On the other hand the more mathematical/fundamental results reported in section 3 would not be influenced by these factors.

Regarding the influence of culture conditions, starvation of eukaryotic cells has been shown to cause different responses in terms of transcriptional bursting dynamics across different genes, even when belonging to closely related families, with both increases and decreases in expression level being observed through variations in burst size and frequency [88], making it difficult to predict the effects of altered culture conditions on genome-wide transcriptional bursting dynamics. Different culture conditions can also be used to induce differentiation of K562 cells into erythroid cells, macrophages and megakaryocytes [157], which would result in a massive shift in the transcriptome away from rapid cell division and towards the specific functions/processes associated with those cell types. Therefore, the biological results and conclusions drawn depend upon culture conditions that facilitate indefinite growth and do not stress the cells, or induce differentiation, for example, which is not the case in many biologically relevant scenarios, such as in the mature and developing tissues of animals. Again this makes the generalisability of the results reported in section 4 to other cell types and environmental scenarios more tenuous, although culture conditions would not be a confounding factor for the more fundamental insights presented in section 3.

5.5 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.

6 References

References

[1] Fei Xavier Chen, Edwin R Smith, and Ali Shilatifard. Born to run: control of transcription elongation by rna polymerase ii. *Nature reviews Molecular cell biology*, 19(7):464–478, 2018.

[2] Erin M Wissink, Anniina Vihervaara, Nathaniel D Tippens, and John T Lis. Nascent rna analyses: tracking transcription and its regulation. *Nature Reviews Genetics*, 20(12):705–723, 2019.

[3] Jason N Kuehner, Erika L Pearson, and Claire Moore. Unravelling the means to an end: Rna polymerase ii transcription termination. *Nature reviews Molecular cell biology*, 12 (5):283–294, 2011.

[4] Nick J Proudfoot. Transcriptional termination in mammals: Stopping the rna polymerase ii juggernaut. *Science*, 352 (6291), 2016.

[5] Stefan GE Roberts, Robert OJ Weinzierl, Robert J White, Jackie Russell, and Joost CBM Zomerdijk. The rna polymerase i transcription machinery. In *Biochemical Society Symposia*, volume 73, pages 203–216. Portland Press, 2006.

[6] Giorgio Dieci, Gloria Fiorino, Manuele Castelnuovo, Martin Teichmann, and Aldo Pagano. The expanding rna polymerase iii transcriptome. *TRENDS in Genetics*, 23(12):614–622, 2007.

[7] Gary Felsenfeld and Mark Groudine. Controlling the double helix. *Nature*, 421(6921):448–453, 2003.

[8] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, 389 (6648):251–260, 1997.

[9] Brian D Strahl and C David Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, 2000.

[10] Thomas Jenuwein and C David Allis. Translating the histone code. *Science*, 293(5532):1074–1080, 2001.

[11] Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, 25(21):2227–2241, 2011.

[12] Makiko Iwafuchi-Doi and Kenneth S Zaret. Cell fate control by pioneer transcription factors. *Development*, 143(11):1833–1837, 2016.

[13] Kenneth S Zaret. Pioneer transcription factors initiating gene network changes. *Annual review of genetics*, 54:367–385, 2020.

[14] Makiko Iwafuchi-Doi, Greg Donahue, Akshay Kakumanu, Jason A Watts, Shaun Mahony, B Franklin Pugh, Dolim Lee, Klaus H Kaestner, and Kenneth S Zaret. The pioneer transcription factor foxa maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation. *Molecular cell*, 62(1):79–91, 2016.

[15] Soufiane Boumahdi, Gregory Driessens, Gaelle Lapouge, Sandrine Rorive, Dany Nassar, Marie Le Mercier, Benjamin Delatte, Amelie Caauwe, Sandrine Lenglez, Erwin Nkusi, et al. Sox2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature*, 511(7508):246–250, 2014.

[16] Meredith E Crosby. Cell cycle: principles of control. *The Yale journal of biology and medicine*, 80(3):141, 2007.

[17] Craig L Peterson and Marc-André Laniel. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, 2004.

[18] Gabriel E Zentner and Steven Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3):259–266, 2013.

[19] Aharon Novogrodski and Nir Friedman. *Histone Modification in Transcriptional Regulation*. Citeseer, 2010.

[20] Theodora Agalioti, Stavros Lomvardas, Bhavin Parekh, Junming Yie, Tom Maniatis, and Dimitris Thanos. Ordered recruitment of chromatin modifying and general transcription factors to the ifn-β promoter. *Cell*, 103(4):667–678, 2000.

[21] Ashwini Jambhekar, Abhinav Dhall, and Yang Shi. Roles and regulation of histone methylation in animal development. *Nature reviews Molecular cell biology*, 20(10):625–641, 2019.

[22] Jareth C Wolfe, Liudmila A Mikheeva, Hani Hagras, and Nicolae Radu Zabet. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in drosophila. *Genome Biology*, 22(1):1–23, 2021.

[23] Lenka Skalska, Robert Stojnic, Jinghua Li, Bettina Fischer, Gustavo Cerda-Moya, Hiroshi Sakai, Shahragim Tajbakhsh, Steven Russell, Boris Adryan, and Sarah J Bray. Chromatin signatures at notch-regulated enhancers reveal large-scale changes in h3k56ac upon activation. *The EMBO journal*, 34(14):1889–1904, 2015.

[24] Madapura M Pradeepa, Graeme R Grimes, Yatendra Kumar, Gabrielle Olley, Gillian CA Taylor, Robert Schneider, and Wendy A Bickmore. Histone h3 globular domain acetylation identifies a new class of enhancers. *Nature genetics*, 48(6): 681–686, 2016.

[25] Gillian CA Taylor, Ragnhild Eskeland, Betül Hekimoglu-Balkan, Madapura M Pradeepa, and Wendy A Bickmore. H4k16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome research*, 23(12):2053–2065, 2013.

[26] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010.

[27] Aditya Sankar, Faizaan Mohammad, Arun Kumar Sundaramurthy, Hua Wang, Mads Lerdrup, Tulin Tatar, and Kristian Helin. Histone editing elucidates the functional roles of h3k27 methylation and acetylation in mammals. *Nature Genetics*, 54(6):754–760, 2022.

[28] Xiaoyan Guan, Neha Rastogi, Mark R Parthun, and Michael A Freitas. Discovery of histone modification crosstalk networks by stable isotope labeling of amino acids in cell culture mass spectrometry (silac ms). *Molecular & Cellular Proteomics*, 12(8):2048–2059, 2013.

[29] Armin P Schoech and Nicolae Radu Zabet. Facilitated diffusion buffers noise in gene expression. *Physical Review E*, 90 (3):032701, 2014.

[30] Hugo G Schmidt, Sven Sewitz, Steven S Andrews, and Karen Lipkow. An integrated model of transcription factor diffusion shows the importance of intersegmental transfer and quaternary protein structure for target site finding. *PLOS one*, 9 (10):e108575, 2014.

[31] Michiel Vermeulen, Klaas W Mulder, Sergei Denissov, WWM Pim Pijnappel, Frederik MA van Schaik, Radhika A Varier, Marijke PA Baltissen, Henk G Stunnenberg, Matthias Mann, and H Th Marc Timmers. Selective anchoring of tfiid to nucleosomes by trimethylation of histone h3 lysine 4. *Cell*, 131(1):58–69, 2007.

[32] Jean-Marc Egly and Frédéric Coin. A history of tfiih: two decades of molecular biology on a pivotal transcription/repair factor. *DNA repair*, 10(7):714–721, 2011.

[33] Patrick Cramer. Organization and regulation of gene transcription. *Nature*, 573(7772):45–54, 2019.

[34] Roger D Kornberg. Mediator and the mechanism of transcriptional activation. *Trends in biochemical sciences*, 30(5): 235–239, 2005.

[35] Jing-Ping Hsin and James L Manley. The rna polymerase ii ctd coordinates transcription and rna processing. *Genes & development*, 26(19):2119–2137, 2012.

[36] Leighton Core and Karen Adelman. Promoter-proximal pausing of rna polymerase ii: a nexus of gene regulation. *Genes & development*, 33(15-16):960–982, 2019.

[37] Michael J Carrozza, Bing Li, Laurence Florens, Tamaki Suganuma, Selene K Swanson, Kenneth K Lee, Wei-Jong Shia, Scott Anderson, John Yates, Michael P Washburn, et al. Histone h3 methylation by set2 directs deacetylation of coding

regions by rpd3s to suppress spurious intragenic transcription. *Cell*, 123(4):581–592, 2005.

[38] Hagen Tilgner, Christoforos Nikolaou, Sonja Althammer, Michael Sammeth, Miguel Beato, Juan Valcárcel, and Roderic Guigó. Nucleosome positioning as a determinant of exon recognition. *Nature structural & molecular biology*, 16 (9):996–1001, 2009.

[39] Schraga Schwartz, Eran Meshorer, and Gil Ast. Chromatin organization marks exon-intron structure. *Nature structural & molecular biology*, 16(9):990–995, 2009.

[40] Paulina Kolasinska-Zwierz, Thomas Down, Isabel Latorre, Tao Liu, X Shirley Liu, and Julie Ahringer. Differential chromatin marking of introns and expressed exons by h3k36me3. *Nature genetics*, 41(3):376–381, 2009.

[41] Iris Jonkers, Hojoong Kwak, and John T Lis. Genome-wide dynamics of pol ii elongation and its interplay with promoter proximal pausing, chromatin, and exons. *elife*, 3:e02407, 2014.

[42] Amy Pandya-Jones, Dev M Bhatt, Chia-Ho Lin, Ann-Jay Tong, Stephen T Smale, and Douglas L Black. Splicing kinetics and transcript release from the chromatin compartment limit the rate of lipid a-induced gene expression. *Rna*, 19(6): 811–827, 2013.

[43] Sérgio Fernandes de Almeida and Maria Carmo-Fonseca. Reciprocal regulatory links between cotranscriptional splicing and chromatin. In *Seminars in cell & developmental biology*, volume 32, pages 2–10. Elsevier, 2014.

[44] Corey R Mandel, Syuzo Kaneko, Hailong Zhang, Damara Gebauer, Vasupradha Vethantham, James L Manley, and Liang Tong. Polyadenylation factor cpsf-73 is the pre-mrna

3'-end-processing endonuclease. *Nature*, 444(7121):953–956, 2006.

[45] Paul B Balbo and Andrew Bohm. Mechanism of poly (a) polymerase: structure of the enzyme-mgatp-rna ternary complex and kinetic analysis. *Structure*, 15(9):1117–1131, 2007.

[46] Donny D Licatalosi, Gabrielle Geiger, Michelle Minet, Stephanie Schroeder, Kate Cilli, J Bryan McNeil, and David L Bentley. Functional interaction of yeast pre-mrna 3 end processing factors with rna polymerase ii. *Molecular cell*, 9(5): 1101–1111, 2002.

[47] Emanuel Rosonina, Syuzo Kaneko, and James L Manley. Terminating the transcript: breaking up is hard to do. *Genes & development*, 20(9):1050–1056, 2006.

[48] Johan Paulsson. Models of stochastic gene expression. *Physics of life reviews*, 2(2):157–175, 2005.

[49] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[50] Jonathan M Raser and Erin K O'shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743): 2010–2013, 2005.

[51] Yen Ting Lin and Tobias Galla. Bursting noise in gene expression dynamics: linking microscopic and mesoscopic models. *Journal of The Royal Society Interface*, 13(114):20150772, 2016.

[52] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer control of transcriptional bursting. *Cell*, 166(2):358–368, 2016.

[53] Roy D Dar, Brandon S Razooky, Abhyudai Singh, Thomas V Trimeloni, James M McCollum, Chris D Cox, Michael L Simpson, and Leor S Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, 2012.

[54] Jong Kyoung Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. *Genome biology*, 14(1):R7, 2013.

[55] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29): 12167–12172, 2011.

[56] Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. *science*, 304(5678):1811–1814, 2004.

[57] Justine Dattani and Mauricio Barahona. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. *Journal of The Royal Society Interface*, 14(126):20160833, 2017.

[58] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *science*, 329(5991): 533–538, 2010.

[59] Arjun Raj and Alexander Van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[60] Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

[61] Richard Losick and Claude Desplan. Stochasticity and cell fate. *science*, 320(5872):65–68, 2008.

[62] Avigdor Eldar and Michael B Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.

[63] Joseph Rodriguez and Daniel R Larson. Transcription in living cells: molecular mechanisms of bursting. *Annual review of biochemistry*, 89:189–212, 2020.

[64] Niraj Kumar, Abhyudai Singh, and Rahul V Kulkarni. Transcriptional bursting in gene expression: analytical results for general stochastic models. *PLoS computational biology*, 11 (10):e1004292, 2015.

[65] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription factors modulate c-fos transcriptional bursts. *Cell reports*, 8(1):75–83, 2014.

[66] Liang Ma, Zeyue Gao, Jiegen Wu, Bijunyao Zhong, Yuchen Xie, Wen Huang, and Yihan Lin. Co-condensation between transcription factor and coactivator p300 modulates transcriptional bursting kinetics. *Molecular cell*, 81(8):1682–1697, 2021.

[67] Shasha Chong, Chongyi Chen, Hao Ge, and X Sunney Xie. Mechanism of transcriptional bursting in bacteria. *Cell*, 158 (2):314–326, 2014.

[68] Caroline R Bartman, Nicole Hamagami, Cheryl A Keller, Belinda Giardine, Ross C Hardison, Gerd A Blobel, and Arjun

Raj. Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Molecular cell*, 73(3):519–532, 2019.

[69] Giorgio Dieci, Maria Cristina Bosio, Beatrice Fermi, and Roberto Ferrari. Transcription reinitiation by rna polymerase iii. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1829(3-4):331–341, 2013.

[70] Jayasha Shandilya and Stefan GE Roberts. The transcription cycle in eukaryotes: from productive initiation to rna polymerase ii recycling. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(5):391–400, 2012.

[71] Massimo Cavallaro, Mark D Walsh, Matt Jones, James Teahan, Simone Tiberi, Bärbel Finkenstädt, and Daniel Hebenstreit. 3-5 crosstalk contributes to transcriptional bursting. *Genome biology*, 22(1):1–20, 2021.

[72] Siddharth S Dey, Jonathan E Foley, Prajit Limsirichai, David V Schaffer, and Adam P Arkin. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Molecular systems biology*, 11(5):806, 2015.

[73] Robin Andersson and Albin Sandelin. Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, 21(2):71–87, 2020.

[74] Damien Nicolas, Benjamin Zoller, David M Suter, and Felix Naef. Modulation of transcriptional burst frequency by histone acetylation. *Proceedings of the National Academy of Sciences*, 115(27):7153–7158, 2018.

[75] Mengyi Sun and Jianzhi Zhang. Allele-specific single-cell rna sequencing reveals different architectures of intrinsic and ex-

trinsic gene expression noises. *Nucleic acids research*, 48(2): 533–547, 2020.

[76] Bérénice A Benayoun, Elizabeth A Pollina, Duygu Ucar, Salah Mahmoudi, Kalpana Karra, Edith D Wong, Keerthana Devarajan, Aaron C Daugherty, Anshul B Kundaje, Elena Mancini, et al. H3k4me3 breadth is linked to cell identity and transcriptional consistency. *Cell*, 158(3):673–688, 2014.

[77] Shaohuan Wu, Ke Li, Yingshu Li, Tong Zhao, Ting Li, Yu-Fei Yang, and Wenfeng Qian. Independent regulation of gene expression level and noise by histone modifications. *PLoS computational biology*, 13(6):e1005585, 2017.

[78] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[79] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

[80] Christoph Engl, Goran Jovanovic, Rowan D Brackston, Ioly Kotta-Loizou, and Martin Buck. The route to transcription initiation determines the mode of transcriptional bursting in e. coli. *Nature communications*, 11(1):1–11, 2020.

[81] Paula Dobrinić, Aleksander T Szczurek, and Robert J Klose. Prc1 drives polycomb-mediated gene repression by controlling transcription initiation and burst frequency. *Nature Structural & Molecular Biology*, pages 1–14, 2021.

[82] Achim P Popp, Johannes Hettich, and J Christof M Gebhardt. Altering transcription factor binding reveals comprehensive transcriptional kinetics of a basic gene. *Nucleic Acids Research*, 49(11):6249–6266, 2021.

[83] Kristen R Maynard, Madhavi Tippani, Yoichiro Takahashi, BaDoi N Phan, Thomas M Hyde, Andrew E Jaffe, and Keri Martinowich. dotdotdot: an automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smfish) images in complex tissues. *Nucleic acids research*, 48 (11):e66–e66, 2020.

[84] Guoliang Li and Gregor Neuert. Multiplex rna single molecule fish of inducible mrnas in single yeast cells. *Scientific data*, 6 (1):1–9, 2019.

[85] Jeffrey R Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39):11046–11051, 2016.

[86] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, Christopher Cronin, Christoph Karp, Eric J Liaw, et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqfish. *Cell*, 174(2):363–376, 2018.

[87] Gennady Gorin, Mengyu Wang, Ido Golding, and Heng Xu. Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics. *Plos one*, 15(3):e0230736, 2020.

[88] Edward Tunnacliffe, Adam M Corrigan, and Jonathan R Chubb. Promoter-mediated diversification of transcriptional bursting dynamics following gene duplication. *Proceedings of the National Academy of Sciences*, 115(33):8364–8369, 2018.

[89] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[90] Marc Sultan, Vyacheslav Amstislavskiy, Thomas Risch, Moritz Schuette, Simon Dökel, Meryem Ralser, Daniela Balzereit, Hans Lehrach, and Marie-Laure Yaspo. Influence of rna extraction methods and library selection schemes on rna-seq data. *BMC genomics*, 15(1):1–13, 2014.

[91] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. Rna-seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1364, 2017.

[92] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

[93] Jingyue Ju, Dae Hyun Kim, Lanrong Bi, Qinglin Meng, Xiaopeng Bai, Zengmin Li, Xiaoxu Li, Mong Sano Marma, Shundi Shi, Jian Wu, et al. Four-color dna sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences*, 103(52):19635–19640, 2006.

[94] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PloS one*, 12(12):e0190152, 2017.

[95] Anton JM Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, 2019.

[96] Hiroshi Ochiai, Tetsutaro Hayashi, Mana Umeda, Mika Yoshimura, Akihito Harada, Yukiko Shimizu, Kenta Nakano,

Noriko Saitoh, Zhe Liu, Takashi Yamamoto, et al. Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells. *Science advances*, 6(25):eaaz6699, 2020.

[97] Per Johnsson, Christoph Ziegenhain, Leonard Hartmanis, Gert-Jan Hendriks, Michael Hagemann-Jensen, Björn Reinius, and Rickard Sandberg. Transcriptional kinetics and molecular functions of long noncoding rnas. *Nature Genetics*, pages 1–12, 2022.

[98] Veronika A Herzog, Brian Reichholf, Tobias Neumann, Philipp Rescheneder, Pooja Bhat, Thomas R Burkard, Wiebke Wlotzka, Arndt von Haeseler, Johannes Zuber, and Stefan L Ameres. Thiol-linked alkylation of rna to assess expression dynamics. *Nature methods*, 14(12):1198, 2017.

[99] Jeremy A Schofield, Erin E Duffy, Lea Kiefer, Meaghan C Sullivan, and Matthew D Simon. Timelapse-seq: adding a temporal dimension to rna sequencing through nucleoside recoding. *Nature methods*, 15(3):221, 2018.

[100] Christopher Jürges, Lars Dölken, and Florian Erhard. Dissecting newly transcribed and old rna using grand-slam. *Bioinformatics*, 34(13):i218–i226, 2018.

[101] Qi Qiu, Peng Hu, Xiaojie Qiu, Kiya W Govek, Pablo G Cámara, and Hao Wu. Massively parallel and time-resolved rna sequencing in single cells with scnt-seq. *Nature methods*, 17 (10):991–1001, 2020.

[102] Junyue Cao, Wei Zhou, Frank Steemers, Cole Trapnell, and Jay Shendure. Sci-fate characterizes the dynamics of gene expression in single cells. *Nature Biotechnology*, pages 1–9, 2020.

[103] Nico Battich, Joep Beumer, Buys de Barbanson, Lenno Krenning, Chloé S Baron, Marvin E Tanenbaum, Hans Clevers, and Alexander van Oudenaarden. Sequencing metabolically labeled transcripts in single cells reveals mrna turnover strategies. *Science*, 367(6482):1151–1156, 2020.

[104] Gert-Jan Hendriks, Lisa A Jung, Anton JM Larsson, Michael Lidschreiber, Oscar Andersson Forsman, Katja Lidschreiber, Patrick Cramer, and Rickard Sandberg. Nasc-seq monitors rna synthesis in single cells. *Nature communications*, 10(1): 1–9, 2019.

[105] Xiaojie Qiu, Yan Zhang, Jorge D Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N Pogson, Marco Y Hein, Kyung Hoi Joseph Min, Li Wang, et al. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711, 2022.

[106] Florian Erhard, Marisa AP Baptista, Tobias Krammer, Thomas Hennig, Marius Lange, Panagiota Arampatzi, Christopher S Jürges, Fabian J Theis, Antoine-Emmanuel Saliba, and Lars Dölken. scslam-seq reveals core features of transcription dynamics in single cells. *Nature*, 571(7765):419–423, 2019.

[107] Etienne Boileau, Janine Altmüller, Isabel S Naarmann-de Vries, and Christoph Dieterich. A comparison of metabolic labeling and statistical methods to infer genome-wide dynamics of rna turnover. *Briefings in Bioinformatics*, page bbab219, 2021.

[108] Brian Munsky, Gregor Neuert, and Alexander Van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[109] Otto G Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of theoretical biology*, 71(4):587–603, 1978.

[110] Minoru SH Ko. A stochastic model for gene induction. *Journal of theoretical biology*, 153(2):181–194, 1991.

[111] Nicolae Radu Zabet and Dominique F Chu. Computational limits to binary genes. *Journal of the Royal Society Interface*, 7(47):945–954, 2010.

[112] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006.

[113] Muir Morrison, Manuel Razo-Mejia, and Rob Phillips. Reconciling kinetic and thermodynamic models of bacterial transcription. *PLoS computational biology*, 17(1):e1008572, 2021.

[114] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[115] Nir Friedman, Long Cai, and X Sunney Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical review letters*, 97(16): 168302, 2006.

[116] Hsin-Ho Huang, Christian Seeger, U Helena Danielson, and Peter Lindblad. Analysis of the leakage of gene repression by an artificial tetr-regulated promoter in cyanobacteria. *BMC Research Notes*, 8(1):1–8, 2015.

[117] Lifang Huang, Zhanjiang Yuan, Peijiang Liu, and Tianshou Zhou. Effects of promoter leakage on dynamics of gene expression. *BMC systems biology*, 9:1–12, 2015.

[118] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25): 2340–2361, 1977.

[119] Hana El Samad, Mustafa Khammash, Linda Petzold, and Dan Gillespie. Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, 15(15):691–711, 2005.

[120] Yen Ting Lin and Nicolas E Buchler. Exact and efficient hybrid monte carlo algorithm for accelerated bayesian inference of gene expression models from snapshots of single-cell transcripts. *The Journal of chemical physics*, 151(2):024106, 2019.

[121] Mariana Gómez-Schiavon, Liang-Fu Chen, Anne E West, and Nicolas E Buchler. Bayfish: Bayesian inference of transcription dynamics from population snapshots of single-molecule rna fish in single cells. *Genome biology*, 18:1–12, 2017.

[122] CB Lozzio and BB Lozzio. Human chronic myelogenous leukemia cell-line with positive philadelphia. *Blood*, 45(3): 321, 1975.

[123] C Michael Fordis, Nicholas P Anagnou, Ann Dean, Arthur W Nienhuis, and Alan N Schechter. A beta-globin gene, inactive in the k562 leukemic cell, functions normally in a heterologous expression system. *Proceedings of the National Academy of Sciences*, 81(14):4485–4489, 1984.

[124] Bismarck B Lozzio, Carmen B Lozzio, Elena G Bamberger, and Aurora S Feliu. A multipotential leukemia cell line (k-562) of human origin. *Proceedings of the Society for Experimental Biology and Medicine*, 166(4):546–550, 1981.

[125] Yanhua Fan, Hongyuan Lu, Li An, Changli Wang, Zhipeng Zhou, Fan Feng, Hongda Ma, Yongnan Xu, and Qingchun Zhao. Effect of active fraction of eriocaulon sieboldianum on human leukemia k562 cells via proliferation inhibition, cell cycle arrest and apoptosis induction. *Environmental Toxicology and Pharmacology*, 43:13–20, 2016.

[126] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[127] Guido Van Rossum and Fred L Drake. *Python 3 reference manual*. CreateSpace, 2009.

[128] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[129] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[130] Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

[131] Zenab F Mchaourab, Andrea A Perreault, and Bryan J Venters. Chip-seq and chip-exo profiling of pol ii, h2a. z, and h3k4me3 in human k562 cells. *Scientific data*, 5(1):1–8, 2018.

[132] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489 (7414):57, 2012.

[133] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[134] Sandip De, David M Edwards, Vibha Dwivedi, Jianming Wang, Wazeer Varsally, Hannah L Dixon, Anand K Singh, Precious O Owuamalam, Matthew T Wright, Reece P Summers, et al. Genome-wide chromosomal association of upf1 is linked to pol ii transcription in schizosaccharomyces pombe. *Nucleic acids research*, 50(1):350–367, 2022.

[135] Bohu Pan, Rebecca Kusko, Wenming Xiao, Yuanting Zheng, Zhichao Liu, Chunlin Xiao, Sugunadevi Sakkiah, Wenjing Guo, Ping Gong, Chaoyang Zhang, et al. Similarities and differences between variants called with human reference genome hg19 or hg38. *BMC bioinformatics*, 20(2):17–29, 2019.

[136] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences*, 115(28):E6437–E6446, 2018.

[137] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006.

[138] Ankit Gupta, Jan Mikelson, and Mustafa Khammash. A finite state projection algorithm for the stationary solution of the chemical master equation. *The Journal of chemical physics*, 147(15):154101, 2017.

[139] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[140] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 1970.

[141] Matti Vihola. On the stability and ergodicity of adaptive scaling metropolis algorithms. *Stochastic processes and their applications*, 121(12):2839–2860, 2011.

[142] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[143] Wally R Gilks, Nicky G Best, and Keith KC Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(4):455–472, 1995.

[144] RFfS Computing et al. R: A language and environment for statistical computing. *Vienna: R Core Team*, 2013.

[145] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161 (5):1187–1201, 2015.

[146] Chen Jia and Youming Li. Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms. *BioRxiv*, 2022.

[147] Zhixing Cao and Ramon Grima. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nature communications*, 9(1):1–15, 2018.

[148] Glenn K Fu, Weihong Xu, Julie Wilhelmy, Michael N Mindrinos, Ronald W Davis, Wenzhong Xiao, and Stephen PA Fodor. Molecular indexing enables quantitative targeted rna sequencing and reveals poor efficiencies in standard library preparations. *Proceedings of the National Academy of Sciences*, 111(5):1891–1896, 2014.

[149] Simona Patange, David A Ball, Yihan Wan, Tatiana S Karpova, Michelle Girvan, David Levens, and Daniel R Larson. Myc amplifies gene expression through global changes in transcription factor dynamics. *Cell reports*, 38(4):110292, 2022.

[150] Dan Lu, Ashwini Jambhekar, and Galit Lahav. Louder for longer: Myc amplifies gene expression by extended transcriptional bursting. *Cell Reports*, 38(9):110470, 2022.

[151] Silvia Jimeno-González and José C Reyes. Chromatin structure and pre-mrna processing work together. *Transcription*, 7(3):63–68, 2016.

[152] Raneen Rahhal and Edward Seto. Emerging roles of histone modifications and hdacs in rna splicing. *Nucleic acids research*, 47(10):4911–4926, 2019.

[153] Alfonso Martinez Arias and Joshua M Brickman. Gene expression heterogeneities in embryonic stem cell populations: origin and function. *Current opinion in cell biology*, 23(6): 650–656, 2011.

[154] Alvaro Rada-Iglesias. Is h3k4me1 at enhancers correlative or causative? *Nature genetics*, 50(1):4–5, 2018.

[155] Beatrice Borsari, Pablo Villegas-Mirón, Sílvia Pérez-Lluch, Isabel Turpin, Hafid Laayouni, Alba Segarra-Casas, Jaume Bertranpetit, Roderic Guigó, and Sandra Acosta. Enhancers with tissue-specific activity are enriched in intronic regions. *Genome research*, 31(8):1325–1336, 2021.

[156] Lenno Krenning, Stijn Sonneveld, and Marvin E Tanenbaum. Time-resolved single-cell sequencing identifies multiple waves of mrna decay during the mitosis-to-g1 phase transition. *Elife*, 11:e71356, 2022.

[157] Jai-Sing Yang, Chao-Ying Lee, Hsin-Chung Cho, Chi-Cheng Lu, Sheng-Chu Kuo, Yen-Fang Wen, Fuu-Jen Tsai, Miau-Rong Lee, and Shih-Chang Tsai. Itr-284 modulates cell differentiation in human chronic myelogenous leukemia k562 cells. *Oncology reports*, 39(1):383–391, 2018.