Stochastic pausing at latent HIV-1 promoters generates transcriptional bursting

Katiana Tantale, Encarnation Garcia-Oliver, Adèle L'Hostis, Yueyuxio Yang, Marie- Cécile Robert, Thierry Gostan, Meenakshi Basu, Alja Kozulic-Pirher, Jean-Christophe Andrau, Florian Müller, et al.

Stochastic pausing at latent HIV-1 promoters generates transcriptional bursting

Katiana Tantale[1,2,*], Encarnation Garcia-Oliver[1,*], Adèle L'Hostis[3], Yueyuxio Yang[3], Marie-Cécile Robert[1,2], Thierry Gostan[1], Meenakshi Basu[1,2], Alja Kozulic-Pirher[1,2], Jean-Christophe Andrau[1], Florian Muller[4], Eugenia Basyuk[1,5,#]; Ovidiu Radulescu[3,#]; Edouard Bertrand[1,#]

[1]Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France

[2]Equipe labélisée Ligue Nationale Contre le Cancer, University of Montpellier, CNRS, Montpellier, France

[3]LPHI, UMR CNRS 5235, University of Montpellier, Montpellier, France

[4]Unité Imagerie et Modélisation, Institut Pasteur and CNRS UMR 3691, 28 rue du Docteur Roux, 75015 Paris; France

[5]Institut de Génétique Humaine, University of Montpellier, CNRS, Montpellier, France

[*]Equal contribution

[#]To whom correspondence should be addressed: Edouard.Bertrand@igmm.cnrs.fr; Ovidiu.Radulescu@umontpellier.fr; Eugenia.Basyuk@igh.cnrs.fr

Keywords: transcriptional noise, polymerase pausing, HIV-1 latency, bursting, single molecule imaging

Summary

Promoter-proximal polymerase pausing is a key process regulating gene expression. In latent HIV-1 cells, it prevents viral transcription and is essential for latency maintenance, while in acutely infected cells the viral factor Tat releases paused polymerase to induce viral expression. Pausing is fundamental for HIV-1, but how it contributes to bursting and stochastic viral reactivation is unclear. Here, we performed single molecule imaging of HIV-1 transcription, and we developed a quantitative analysis method that manages multiple time scales from seconds to days, and that rapidly fits many models of promoter dynamics. We found that RNA polymerases enter a long-lived pause at latent HIV-1 promoters (>20 minutes), thereby effectively limiting viral transcription. Surprisingly and in contrast to current models, pausing appears stochastic and not obligatory, with only a small fraction of the polymerases undergoing long-lived pausing in absence of Tat. One consequence of stochastic pausing is that HIV-1 transcription occurs in bursts in latent cells, thereby facilitating latency exit and providing a rationale for the stochasticity of viral rebounds.

Introduction

Transcription initiation is a complex process that comprises chromatin opening, assembly of a pre-initiation complex (PIC), polymerase recruitment and finally its maturation into an elongation-competent form (see [1] for review). In Drosophila and mammals, this last step is highly regulated and appears to be a key point in the control of gene expression ([2] for review). RNA polymerase II (RNAPII) is recruited by the PIC in a hypo-phosphorylated form and is then loaded on a short stretch of single stranded DNA, which is melted by TFIIH. The initiating polymerase starts elongating about a dozen of nucleotides and must undergo a number of modifications before leaving the promoter and entering productive elongation [3]. First, the TFIIH-associated CDK7 kinase phosphorylates the Serine 5 of the heptad repeats of the C-terminal domain (CTD) of RNAPII, thereby disrupting interaction with Mediator and facilitating promoter escape ([4,5] for reviews). The S5 phosphorylated CTD also recruits the RNA capping enzymes that access the RNA 5'-end when it emerges from the polymerase [6]. The polymerase then transcribes an additional 10-80 nucleotides and typically enters a paused state. Two factors appear particularly important to trigger pausing, in relation with TFIID [7]: DSIF (DRB sensitivity-inducing factor), which is composed of SPT4 and SPT5, and NELF (negative elongation factor), a four subunit complex that also interacts with the cap via the cap-binding complex [8] (CBC). A recent structure of the pausing complex indicates that the RNA-DNA hybrid adopts a tilted conformation within the polymerase that prevents further nucleotide addition [9]. This structure is stabilized by NELF and DSIF, which also prevent binding of TFIIS, a factor that can trigger cleavage of the RNA at the active site to restart backtracked polymerases [10]. Release from the paused state requires the positive transcription elongation factor b (P-TEFb), which is composed of Cyclin T1 or T2 associated with the kinase CDK9 [11], sometimes in association with the super-elongation complex

[12,13] (SEC). P-TEFb is activated by CDK7 [4,5,14] and it phosphorylates a number of components of the pausing complex to enable formation of an elongation-competent polymerase [9,15,16]. Phosphorylation of NELF triggers its dissociation from the polymerase, and this frees a binding site for PAF, an elongation factor that is required for transcription through chromatin. P-TEFb also phosphorylates the RNA polymerase CTD on its Serine 2, as well as the linker between the polymerase core and the CTD, creating a binding site for the elongation factor SPT6 [9]. DSIF functions both as a repressor and activator of elongation, and it is also phosphorylated by P-TEFb ([17] and ref therein). The structures of the paused and active elongation complex show that DSIF adopts different conformations in the two complexes. In particular, phosphorylated DSIF frees the nascent RNA and allows the polymerase to clamp around the DNA, promoting elongation while preventing release of the polymerase from DNA. Overall, P-TEFb mediated phosphorylation thus disrupts the pausing complex and triggers formation of an active elongation complex comprising the polymerase associated to DSIF, SPT6, and PAF.

While pausing is thought to be a key regulatory point for many cellular promoters in mammals and Drosophila, it is often revealed by a peak of RNAPII near the promoter that can in fact correspond to different molecular processes [18], such as slow elongation, polymerase arrest, or defective processivity (i.e. abortive initiation). Recent efforts have been made to clarify these mechanisms by measuring pausing duration. These studies indicated that pausing time vary from less than a minute up to an hour in Drosophila, depending on the promoter [19-23]. This revealed a surprising variability in pausing kinetics, with widely different regulatory potential.

Another major finding of the last 15 years is that transcription is a discontinuous process in vivo ([24] see [25,26] for reviews), with "active" genes going through active and inactive periods in a stochastic manner, a phenomenon also called transcriptional noise or gene bursting. In particular, recent evidences suggest that for many genes, expression levels are dynamically

encoded in the time domain by controlling the periods during which a gene is active, rather than by regulating the initiation rate [27-29]. Major efforts have been made to decipher the causes of gene bursting and in particular the molecular status of the postulated ON and OFF states. Indeed, the transitions between these states are kinetically rate limiting and therefore represent key regulatory checkpoints. However, despite these efforts and the importance of pausing in regulating gene expression, how pausing affects gene bursting remains not understood.

An important implication of gene bursting is that it creates cell-to-cell heterogeneity and this has multiple consequences on the phenotypes of single cells or multicellular organisms. For instance, stochasticity in the expression of Heat-Shock genes in yeasts is thought to help a fraction of the yeast population survive sublethal stresses [30], while in C. Elegans, mutations in a small gene regulatory network create a high expression variability, ultimately leading to variable phenotypic penetrance of the mutation [31]. In the case of HIV-1, transcriptional noise is thought to play a crucial role in the control of latency. Indeed, HIV-1 infection generates latent cells that can persist in the body for decades and can re-establish viral propagation when antiviral treatments are interrupted. Previous studies from the Siliciano and Weinberger labs have shown that latency exit is stochastic and linked to random fluctuations of viral transcription [32-34]. How the viral promoter creates bursts of gene expression in latent cells is not understood, but nevertheless fundamental as it is triggering latency exit. A better knowledge of mechanistic and quantitative aspects of the reactivation dynamics is indeed essential for the development of new strategies in combinatorial anti-retroviral therapies such as “shock and kill” and “block and lock”.

The ability of the virus to alternate between acute and latent forms lies in a positive transcriptional feedback loop established by the viral protein Tat ([32], see [35,36] for reviews). In latent cells, Tat levels are very low and viral transcription remains silent or also low. In acutely infected cells, Tat levels are elevated, strongly inducing viral transcription. It is well established

that in absence of Tat or when Tat levels are low, P-TEFb is limiting for viral transcription and the polymerases that initiate transcription enter a paused state after transcribing about 60 nucleotides and fail to enter productive elongation (reviewed in [35,36]; Figure 1A, left). Tat alleviates this block by binding both P-TEFb and the TAR stem-loop at the 5'-end of nascent HIV-1 RNAs, leading to the formation of a ternary complex that promotes elongation by recruiting P-TEFb and its associated super-elongation complex to paused polymerases [11-13] (Figure 1A, right). The HIV-1 promoter is thus strictly regulated at the level of pausing and P-TEFb recruitment, and these steps are controlled by Tat, which overall can activate viral transcription by more than 100 fold. These properties make HIV-1 an attractive model to decipher how pausing affects gene bursting, with direct relevance for HIV-1 latency and pathogenesis [37,38].

Here, we imaged HIV-1 transcription in live cells at the level of single polymerases. We characterized the effect of pausing on gene bursting by modulating the levels of Tat, which controls pausing at the HIV-1 promoter. We provide the first fully quantitative description of the stochastic activity of the HIV-1 promoter in basal and induced conditions, on timescales ranging from second to tens of hour. Surprisingly, we found that promoter-proximal pausing is a stochastic event that generates large viral bursts even in cells that do not express Tat. In HIV-1 latent cells with a functional but inactive Tat loop, stochastic pausing may be a key phenomenon that determines latency exit.

Results

Single molecule imaging of HIV-1 transcription with different levels of the pause release factor Tat

We previously developed an improved MS2 tagging system based on a 128xMS2 tag and designed for long term tracking of single RNAs [28]. To image viral transcription, we inserted this tag in the intron of an HIV-1 vector that had all the viral sequences responsible for transcription and RNA processing (Figure 1A-B). The corresponding pre-mRNA splices entirely post-transcriptionally, enabling imaging of transcription independently of splicing [28,39]. The high number of MS2 stem-loops present in this reporter allows for a 5-fold increase in signal as compared to our original 24xMS2 repeat [40]. This enables the use of a low illumination power to limit photo-bleaching, allowing to capture five times more images while still detecting single RNA molecules. By using the 128xMS2 tag and monitoring the brightness of the transcription site over time, it is possible to measure promoter activity with a temporal resolution in the second range and for hours.

It has been demonstrated by numerous studies that the HIV-1 promoter is regulated at the level of promoter proximal pausing (see [35,36] for reviews). Indeed, latent cells do not express a significant amount of Tat and in this case, polymerases that start transcribing are blocked ~60 nucleotides downstream the transcription start site and do not enter productive elongation. This block is relieved by Tat, which directly alleviates pausing by recruiting P-TEFb to the nascent viral RNAs and allowing polymerases to elongate throughout the entire viral genome. To characterize how pausing affects HIV-1 transcription, we therefore created isogenic cell lines expressing different levels of Tat. These lines all contained the 128xMS2 reporter integrated at the same chromosomal location. We previously generated a HeLa cell line that expressed in *trans*

a saturating amount of Tat (*High Tat* cells). In these cells, transcription was high and a further increase in the amount of Tat did not lead to more viral transcription [28]. We then created two new reporter cell lines with low levels of Tat to mimic the situation of latent cells where Tat is not expressed or only at very low levels [35,36]. The first cell line expresses Tat from the second cistron of a bicistronic vector (referred to as *Low Tat* cells), and Tat was not detected by Western blot although it promoted HIV-1 transcription by 2.7 fold (Figure 1C-E and Figure S1A). The second cell line entirely lacked Tat (referred to as *No Tat*). We first determined the expression levels of the HIV-1 reporter by performing smFISH experiments with probes binding the 128xMS2 repeat. We found that expression of the HIV-1 reporter depended on Tat as expected (Figure 1C-E), as the number of pre-mRNA molecules present in the nucleoplasm dropped from ~500 copies per nucleus in *High Tat* cells, to ~50 and ~20 in *Low Tat* and *No Tat* cells, respectively. This was mirrored by a similar decrease in the level of the nascent RNAs present at the transcription sites, with a mean of 32 copies for the *High Tat* cells, and only 5 and 1.8 for the *Low Tat* and *No Tat* cells, respectively (Figure 1C-E).

Next, we aimed at confirming that pausing was limiting viral transcription in *No Tat* cells. To this end, we overexpressed the two subunit of P-TEFb, Cdk9 and Cyclin T1, by transient transfection. We observed that this increased viral transcription as previously reported in other cellular systems (Figure S1B; [41]). Then, we fused CDK9 to a fluorescent catalytically inactive Cas9 variant (dCas9-tagBFP), and we transfected the resulting construct in *No Tat* cells together with vectors expressing three Cas9 guide RNAs targeting the HIV-1 promoter. By performing smFISH with probes against the 128xMS2 repeat, we found that expressing dCas9-CDK9-tagBFP alone increased HIV-1 RNA levels by 4 fold, while further targeting it to the HIV-1 promoter with three guide RNAs led to a 10-fold increase in expression (Figure S1C). Moreover, the basal HIV-1 transcriptional activity in *No Tat* cells was blocked when P-TEFb was

inactivated with KM05283, a drug that specifically inhibits CDK9 kinase activity (Figure S2A). This indicated that P-TEFb was both required for basal transcription and also limiting viral expression, providing functional indications that pausing was limiting in *No Tat* cells. Next, we tested whether the basal viral transcription observed without Tat was due to sporadic activation of the NF-κB pathway, as it is a well-known activator of the HIV-1 promoter that can recruit P-TEFb [42,43]. We treated cells with BAY11-7082, a drug that inhibits the IKK kinase and traps NF-κB subunits in the cytoplasm. No difference in HIV-1 expression was seen after 16h of treatment, indicating that the basal viral transcription was independent of NF-κB (Figure S2B-C). Taken together, these data indicate that in our cellular system, the basal HIV-1 transcription occurring in absence of Tat is P-TEFb dependent, and that the recruitment of this factor is a key step limiting viral transcription, as expected from a large body of previous studies.

The absence of Tat does not affect the formation of polymerase convoys but creates long inactive periods

When Tat is in excess, HIV-1 transcription occurs in the form of polymerase convoys, i. e. sets of closely spaced polymerases that transcribe the gene together (see schematic in Figure 2D; [28]). In average, the Tat-activated HIV-1 promoter produces convoys of 19 polymerases, each polymerase spaced every ~4 seconds, with a convoy being fired every ~2 minutes. In order to characterize how a limiting amount of Tat affects the viral transcriptional output, we performed live-cell imaging using MCP-GFP and monitored the brightness of transcription sites over time. The single molecules of unspliced pre-mRNA present in the nucleoplasm were used to calibrate the signal at the transcription site, which could then be expressed as an absolute number of RNA molecules (Figure 2). We previously showed that the Tat-activated HIV-1 promoter fluctuates on

time scales ranging from minutes to hours, and we therefore recorded two types of movies to cover the entire temporal range of transcriptional fluctuations [28]. 'Short movies' capture one image stack every 3 seconds for 15 to 20 minutes, and they allow a detailed characterization of rapid transcriptional fluctuations such as polymerase convoys. 'Long movies' last for 8 hours with a rate of one image stack every three minutes, and they allow to measure the frequency and duration of long inactive periods. Note that since a nascent RNA resides 2.8 minutes at the transcription site [28], this frame rate ensures that all the initiation events are detected in the long movies.

In the short movies, we observed transient increases in the brightness of transcription sites for all three cell lines: *High Tat*, *Low Tat* and *No Tat* (Figure 2A-C). They were in the minute range and quantification of the signals indicated that they corresponded to the synthesis of multiple RNA molecules (Figure 2A-C). Thus, viral transcription occurred in large bursts even in absence of Tat, resulting in the formation of polymerase convoys. To better characterize these rapid fluctuations, we focused on transcription cycles in which an inactive transcription site transiently turned on, and we fitted these data with a model of polymerase convoys ([28]; see schematic in Figure 2D). Surprisingly, the convoys formed in the *Low Tat* and *No Tat* cells were roughly similar to those formed when Tat was saturating (19 polymerases initiating every 4 seconds in *High Tat* cells, compared to 14 polymerases every 6 s in *Low Tat* cells and 12 polymerases every 8 s in absence of Tat; Figure 2E). This result was unexpected because decreasing Tat levels should increase pausing, which should increase in lag time between successive polymerases, possibly until convoys are no longer formed. It is also interesting to note that the differences observed at this rapid time-scale were small and could not account for the 30 fold difference in expression induced by Tat (Figure 1C-E).

Next, we analyzed fluctuations on slow time scales using long movies. The HIV-1 promoter was almost always active in cells expressing an excess of Tat (Figure 3A-B, left panels). In contrast, *No Tat* and *Low Tat* cells displayed long inactive periods that lasted for hours (Figure 3A-B, middle and right panels). In addition, active periods were brief and rare, yet yielded initiation of multiple polymerases in the form of convoys as for *High Tat* cells (see Figure 3A). The activity of the HIV-1 promoter in absence of Tat thus occurs mainly in the form of sparse, yet large bursts, with long inactive period explaining most of the difference in promoter activity with and without Tat.

Development of a novel analysis pipeline to characterize the fluctuations of promoter activity on multiple timescales

The fluctuations of promoter activity arise from stochastic transitions between active and inactive promoter states ([25,26]; Figure 4A). These transitions correspond to steps that are kinetically rate-limiting, and the characterization of these promoter states can thus yield important information on how promoters function and are regulated. To better understand how pausing and Tat control the activity of the HIV-1 promoter, we turned to machine learning and modeling approaches with the aim of elucidating how the promoter switches between active and inactive states. The analysis of the fluctuations of transcription sites brightness can be done by auto-correlation strategies [44,45]. This gives a direct measurement of the dwell time of the nascent RNAs and allows to estimate the elongation and 3'-end processing rates. However, there is currently no theoretical framework that can easily extend autocorrelation methods to models containing multiple promoter states besides a simple ON/OFF switch. In addition, correlation approaches are difficult to use when fluctuations are slow and approach the recording time of the movies. Other analysis strategies hypothesize a theoretical transition model and infer parameters using Bayesian or maximum

likelihood approaches [29,46-48]. These strategies rarely compare several models and do not directly characterize features such as polymerase convoys. To circumvent these difficulties, we turned to the direct analysis of polymerase waiting times, i.e. the lag time between two successive initiation events. Indeed, transcription can be modelled as a continuous time Markov chain in which a promoter stochastically switches between various non-productive states until it reaches an active state where it can initiate transcription (Figure 4A). In this case, waiting times between successive initiation events are interesting to consider because their distribution directly relates to transition rates of the Markov chain (see Supplemental Text). Moreover, we obtained for many different models the closed-form equations expressing the distribution of waiting times as a function of the model parameters (for full solutions to this direct problem, see Methods and Supplemental Text), as well as closed-form equations allowing to compute the model parameters directly from the distribution of waiting times, the so-called inverse problem (for full solutions, see Methods and Supplemental Text). In particular, if we consider a class of models containing several consecutive OFF states and one ON state that can initiate transcription (Figure 4A), the survival function of polymerase waiting times, which is one minus their cumulative distribution, is the sum of several exponentials with the number of exponentials corresponding to the number of promoter states (Figure 4A; see Supplemental Text). Thus, by fitting the survival function with various sums of exponentials, one can determine the number of states in the promoter model. In addition, the rates of promoter switching can be directly calculated from the coefficients of the fitted sum of exponentials (see Methods and Supplemental Text, inverse problem). Hence, if the distribution of waiting times can be extracted from the experimental data, it is straightforward to determine both the number of promoter states, as well as the rates of switching between these states.

Calculation of polymerase waiting times from short and long movies

We first reasoned that the inactive periods seen in the long movies correspond to long polymerase waiting times (Figure 3A-B). Since the frame rate is 3 minutes while a nascent RNA remains 2.8 minutes at the transcription site [28], these movies should detect all initiation events and thus to measure all the polymerase waiting times longer than 3 minutes. The short waiting times could be calculated from the short movies, which have a much higher frame rate (3 seconds). However, a difficulty is that the signal generated by a polymerase persists several minutes after it initiated, as the labelled nascent RNA leaves the transcription site only after it is transcribed to the end of the gene and 3'-end processed (see schematic in Figure 2D). Consequently, if the next polymerase appears before the nascent RNA disappears, the transcription site remains continuously fluorescent and it is not possible to directly calculate the polymerase waiting times. To circumvent this difficulty, we reasoned that the intensity of transcription sites over time is the result of the convolution of two functions: the signal produced by a single polymerase and the time sequence of firing events (see [25] and Figure 4B, left panels). The signal produced by a single polymerase depends on the polymerase elongation rate and the rate of 3'-end formation, which we determined previously for this HIV-1 reporter gene [28,39]. If we assume that all polymerases behave identically, it is thus possible to calculate the temporal position of polymerase initiation events by finding the best sequence of these events that reproduces the experimental transcriptional fluctuations (Figure 4B). It should thus be possible to extract polymerase waiting times from the short movies, keeping in mind that the waiting times longer than the movie will be truncated and require a correction (see Supplemental Text).

Altogether, the long movies give access to waiting times longer than the frame rate (waiting times in the 3 min-10h range), and the short movies provide waiting times shorter than the movie length (in the 3s-20min range). The combination of these movies thus allows to

297 reconstruct and estimate the distribution of polymerase waiting times over 4 logarithmic decades,
298 i.e. 3s-10h (see Supplemental Text for the reconstruction procedure). This analysis pipeline has
299 three advantages. First, by determining the number of exponentials required to fit the survival
300 function, one can directly determine the number of promoter states in the model. Second, given
301 that equations describing the distribution of waiting times can be obtained for many models, it is
302 straightforward to fit these models and to estimate which model best fits the experimental data.
303 Finally, this pipeline enables to combine data acquired at multiple time-scales, from seconds to
304 ten hours, and therefore provides an ideal framework to quantify transcriptional dynamics in live
305 cells.
306
307 **Validation of the analysis pipeline by simulations**
308 To evaluate the precision and reliability of the analysis pipeline, we first tested the performance
309 of the deconvolution algorithm on simulated datasets. The initiation times of several polymerases
310 were simulated and the signal of an imaginary transcription site was calculated using
311 experimentally measured elongation and 3'-end processing rates [28,39]. We then added a realistic
312 amount of noise and tested the ability of the deconvolution algorithm to reconstruct the proper
313 initiation timing from the noisy signal (Figure 5A and Supplemental Text). The algorithm is
314 composed of two parts: a genetic algorithm to obtain the rough position of initiation events, and
315 by a local optimization to refine the position of initiation events. In both presence and absence of
316 noise, the combination of the two steps allowed an accurate positioning of the initiation events.
317 Next, we validated the entire analysis pipeline by simulating a three state branched
318 promoter model with the Gillespie algorithm, using several realistic sets of parameters (i.e.
319 corresponding to values obtained with our cell lines, see below). We computed the brightness of
320 many statistically equivalent imaginary transcription sites as above, and added different amounts

of noise (1x, 2x and 4x), with the 1x condition corresponding to the noise observed in our experimental data (Figure 5B, upper panels; see Supplemental Text). The intensities of the simulated transcription sites were then resampled to create artificial short and long movies, which were treated exactly as real data. Simulated short movies were deconvolved and the distribution of waiting times was computed separately for the short and long movies. These distributions were then combined to reconstruct the entire distribution of waiting times (Figure 5B, middle panels), which was fitted to a sum of three exponentials to calculate the parameters of the promoter model. In absence of noise, all the model parameters were recovered accurately and with high precision (i.e. a small confidence interval), for the three sets of parameter value used to generate the artificial data (Figure 5B, lower panels). With the 1x and 2x amount of noise, parameter recovery was still accurate, while for the 4x noise condition, some parameters were recovered with a low precision, in particular those corresponding to rapid transition rates. Overall, these simulations indicated that our analysis pipeline worked well, even with complex promoter models, and was robust with respect to noise.

Modeling indicates that pausing is stochastic and that pauses are long-lived

We analyzed the movies produced from cells expressing different amounts of Tat and created several models describing how the HIV-1 promoter may operate. The simplest model has two promoter states, ON and OFF as shown in Figure 6A, and assumes that once initiated, RNAPII enters directly into productive elongation without a pausing step. This would likely be the case when expression of Tat is high and pausing not rate-limiting, but not when Tat is limiting or absent. We thus created a model that included a pausing step. It consisted of the same simple model with two promoter states (OFF and ON), but with initiating polymerases undergoing an obligatory pause (PAUSE), before either progressing into elongation or aborting (Figure 6A

345 middle; model M3). Note that once the polymerase exits the pause or aborts, the promoter goes
346 immediately back in the ON state. A large body of work indicates that Tat is promoting
347 elongation by recruiting P-TEFb and in agreement, P-TEFb is limiting for HIV-1 transcription in
348 the *Low Tat* and *No Tat* cells used here (Figure S1B-C). Therefore, we expected to have a high
349 abortion rate (k_{abort}) and/or a low rate of pause release ($k_{release}$) in absence of Tat, and the opposite
350 when Tat is abundant. Conversely, the rates of switching between the ON and OFF state should
351 not be much affected by the amount of Tat.

352 For the obligatory pausing model, the symbolic solution describing the distribution of
353 polymerase waiting times is the sum of three exponentials, but with one of the five parameter
354 being constrained and expressed as a function of the others (see Supplemental Text section 4.6).
355 After fitting the experimental distributions of polymerase waiting times using this symbolic
356 solution, we estimated the quality of the fit with three criteria: (i) the sum of squared residuals,
357 evaluated from the function minimized during the fit (i.e. the objective function, with the inverse
358 of its minimal value giving the fit score); (ii) the certainty of the value of the fitted model
359 parameters, evaluated by their confidence intervals; (iii) the realistic nature of the parameter
360 values, and in particular the pausing times and the effects of Tat. According to these
361 considerations, the fit of the 3 state model with an obligatory pause was poor, and this was the
362 case of all the Tat cell lines. First, the model scores were low and not better than the simple 2
363 state model without pause, even in the *Low Tat* / *No Tat* cell lines were P-TEFb recruitment
364 limits viral transcription (Figure 6B-C). Second, the uncertainty in some parameter was high, as
365 shown by the large confidence intervals of the parameters of the fitted exponentials (see Table 4
366 of Supplemental Text). Third, the pausing time, estimated from the rates of pause release and
367 transcription abortion, was short (Figure 6D; less than 10 seconds whether Tat was present or
368 not), while most of the regulation induced by Tat occurred at the transition between the ON and

OFF state and not pausing (see Figure 17 of the Supplemental Text). It is also interesting to note that the fitted abortion rate was found to be > 100 fold faster than the rate of pause release (Figure 17 of Supplemental Text). Because the promoter goes directly to the ON state upon abortion or pause release, a high abortion or release rate creates a collapse between the ON and PAUSE states and therefore simplifies the 3 state model with pause into a simple 2 state ON/OFF model without pause. This explains why these two models have identical scores and fitted survival functions (Figure 6B, compare curves with '+' and 'x'). In order to try improving the model with obligatory pause, we made a four state model having two successive OFF states, one ON state and an obligatory pause (Model M4; see Figure 18 of Supplemental Text). This model fitted the data better and had a better overall score (see value of the objective function in Table 5 of the Supplemental text). However, it suffered from similar flaws as the previous model (see Figure 18 of Supplemental Text): (i) short pausing time whether Tat was present or not (<10 s); and (ii) high abortion rates, which similarly collapsed the 4 state model with an obligatory pause into a 3 state model without pause. Overall, increasing the number of OFF states in the model with an obligatory pause still yields short pausing times not regulated by Tat. Thus, an obligatory pause does not provide a benefit over a model without pause, with most of the effect of Tat occurring at the level of transitions between OFF and ON states. It is important to realize that given the high degree of bursting without Tat, with polymerases rapidly succeeding one another to form convoys during periods of gene activity (Figure 2), an obligatory pause necessarily means that pausing is short. In addition, since Tat mainly affects long inactive periods (Figure 3B), short pauses mean that the regulation by Tat cannot be on pausing, but rather on other steps able to produce long OFF periods. Hence, the occurrence of polymerase convoys in absence of Tat implies that an obligatory pause cannot be the step regulated by Tat to increase transcription.

This questioned the validity of the model and we thus sought for alternatives. In the previous models, pausing is an obligatory step, but it could be imagined that pausing is a facultative step, for instance if entry into the pause is stochastic. In this case, initiating polymerases have the choice of either directly progressing into productive elongation or entering a paused state, from which they can exit by either aborting or entering elongation (Figure 6A right, model M2+). To test this model, we first used a simplified variant of model M2+, in which polymerases systematically abort when exiting a facultative pause (model M2, see Figure 1 of the Supplemental Text). This model could fit the data from all the three cell lines, *High Tat*, *Low Tat* and *No Tat* (Figure 6B), with scores higher than the 3- or 4-state models with an obligatory pause (Figure 6C; Table 5 of Supplemental Text). Moreover, all parameters had a high precision with small confidence intervals (see Table 2 of Supplemental Text), and the model correctly predicted the number of pre-mRNA per cell (Figure 6E), with only a slight under-estimation for the *High Tat* cells. The fitted parameters indicate that pausing is infrequent, even in cells lacking Tat (Figure 6D). This implies that the fate of the paused polymerase will only marginally affect the promoter output, indicating that models in which the paused polymerase enters productive elongation would give similar results. Because the simplified model M2 is symmetrical, it is not possible to determine with certainty which parameters correspond to the ON-OFF transition, and which correspond to the ON-facultative pause. Nevertheless, both possibilities indicate a long pausing time from 15 minutes to 3h in *No Tat* cells, which is regulated by Tat as it decreases to either 1 or 15 minutes in *High Tat* cells. Pausing is also always predicted to be infrequent, varying from one every 20 to 180 polymerases in *No Tat* cells, down to one every 40 to 3900 in *High Tat* cells (Figure 6D).

Measurement of pausing duration by biochemical approaches.

To further assess models with obligatory or stochastic pausing, we attempted to test their most discriminative prediction. Obligatory pausing predicts a pausing time in the second range, while facultative pausing predicts a duration in the hour or sub-hour range (Figure 6D). Pausing duration can be estimated by measuring RNAPII residency time, and this can be achieved by performing chromatin immunoprecipitation (ChIP) during a time-course with Triptolide, a drug that inhibits TFIIH and prevents loading new polymerases without removing the ones that already initiated. We treated *High Tat* and *No Tat* cells with Triptolide for up to an hour and performed an RNAPII ChIP experiment. We analyzed the HIV-1 promoter as well as the GAPDH promoter, as a constitutively active control gene (Figure 7A). In the *High Tat* cells, similar levels of RNAPII were found on both the GAPDH and the viral promoters, while about 6-fold less polymerases was found on the HIV-1 promoter in absence of Tat, consistent with previous results (Figure S3; [49]). Most importantly, treatment with Triptolide led to the rapid disappearance of RNAPII at the GAPDH promoter, with only ~20% of the signal remaining after 10 min of treatment (Figure 7A). Interestingly, the kinetics observed at the HIV-1 promoter were dependent on Tat. In *High Tat* cells, the RNAPII signal also decreased rapidly and this was consistent with the rapid succession of polymerase firing that we measured in live cells (on every 4-6 seconds; [28]). In contrast, the polymerases remained associated a much longer time with the viral promoter in absence of Tat, with 88% of the signal remaining after 10 minutes of treatment (Figure 7A). Extrapolation of the half-life of the promoter-associated polymerases indicated 10 minutes for the GAPDH promoter and for the HIV-1 promoter when Tat levels are high. However, this half-life raised to 38 minutes for the HIV-1 promoter when Tat was absent, consistent with a long pause. These long values may moreover be underestimated as hour-long treatment with Triptolide were shown to cause degradation of RNA polymerase II in human cells [50]. Altogether, these data verify a key discriminative prediction of the facultative pausing model, namely that paused polymerases

exhibit a half-life in the sub hour range and not in the second range as expected from an obligatory pausing scenario.

Next, we wished to determine whether long pausing time requires a specific feature of the HIV-1 promoter or could be induced at any promoter by depleting P-TEFb. We thus repeated the GAPDH RNAPII ChIP time course, but pretreated cell with the Cdk9 inhibitor KM05382 for 2h before performing the Triptolide time course. The residency time of RNAPII at the GAPDH promoter was similarly short whether cells were pretreated with KM05382 or not (Figure 7B), indicating that the lack of P-TEFb activity is not sufficient in itself to induce long pauses. This suggests that the HIV-1 promoter likely has additional features that specify this property.

Discussion

Cells latently infected with HIV-1 prevent patients from clearing the virus, as the stochastic activation of these cells can re-establish viral propagation [32,34]. Latent cells do not express the viral genome and pausing of RNA polymerases at the viral promoter is a key block that prevents HIV-1 transcription [35-38]. Pausing thus plays a fundamental role in HIV-1 biology, and yet, how it contributes to bursting and stochastic reactivation of the virus is not known. Here, we harnessed the power of single molecule transcriptional imaging and modeling to study how pausing affects HIV-1 transcription in single cells. We find that pausing is a stochastic process, and modeling as well as biochemical experiments indicate that it is long lived inhibitory state that impacts only a small fraction of the initiating polymerases. Stochastic pausing therefore generates viral transcriptional bursts in absence of Tat, which may cause viral reactivation, latency exit and viral rebounds in patients.

A frequentist approach accurately and robustly model transcriptional fluctuations

Single molecule transcription imaging is a powerful technique that becomes indispensable for understanding transcriptional regulation in vivo. However, the signal produced by this technique integrates processes with widely distributed timescales, and not directly accessible by simple data processing. Hence, new modeling methods are needed to cope with the multiscale nature of transcription. To this end, we developed a new machine learning and modeling method. Using numerical deconvolution, this approach generates a time map of transcriptional initiation events indicating, for each transcription site, when RNAPII molecules start producing an mRNA. This feature is unique in our analysis pipeline and not available in other approaches that directly fit a particular transcription model to experimental data [29,45-48]. Our method generates a multiscale

473 cumulative distribution function of polymerase waiting times, which separate successive
474 transcription initiation events. This distribution function has the unique advantage of integrating
475 temporal information on transcriptional processes with an unprecedented dynamic range from
476 seconds to days. Moreover, we have analytically solved the inverse problem consisting in
477 computing the model parameters as a function of the waiting time distribution, for a large number
478 of models. By allowing easy and quick comparison of many different models of promoter
479 dynamics, this method removes a bottleneck essential for hypothesis testing in gene regulation
480 studies.
481
482 **Polymerase pausing generates long-lived inactive states that limit HIV-1 transcription**
483 P-TEFb is an essential elongation factor that is required for both the basal and Tat-induced
484 activity of the HIV-1 promoter [11,35,36]. By default, the HIV-1 promoter leads to pausing and
485 inefficient elongation, and Tat functions as a promoter specific elongation factor by recruiting P-
486 TEFb to the nascent viral RNAs. When Tat is present in saturating concentrations, we observe
487 that polymerases initiate rapidly, one after another (every 4-6s in average; [28]). This indicates that
488 the maturation of initiating polymerases into a processive complex is rapid, in agreement with the
489 fact that P-TEFb recruitment is not rate-limiting when Tat is abundant. When Tat is limiting or
490 absent, we observe a biphasic behavior. HIV-1 promoters are mostly inactive, and yet sometimes
491 transcribe the viral genome in brief pulses containing tens of polymerases. These polymerases are
492 fired in rapid succession (one every 7-15s), and they form convoys resembling the ones observed
493 when Tat is saturating. Modelling the imaging data in *Low Tat* and *No Tat* cells confirms this
494 biphasic behavior and further indicates that in average, 20-35 polymerases initiate during active
495 periods of 5 minutes, followed by inactive periods of 20 minutes or 3h. Given that pausing is
496 limiting transcription in the absence of Tat, these long inactive periods are likely caused by long-

497 lived pauses at the HIV-1 promoter. Indeed, direct measurements of RNAPII residency time at
498 the HIV-1 promoter indicate that the absence of Tat generates long pauses in the sub-hour range,
499 which are therefore responsible for at least some of the long periods without viral transcription.
500 Recent genome-wide data obtained in Drosophila with Triptolide time-course experiments
501 indicate that the half-life of polymerases at cellular promoters varies from less than a minute to
502 about an hour [19-23]. Analysis of a series of promoter variants further indicates that an initiator
503 element with a G at position +2 is a key determinant of long pausing time (>40 minutes; [51]). It is
504 not known whether this rule also applies to vertebrates, but it is worth noting that the HIV-1
505 promoter has an unusual initiator element required for Tat activation that contains a G at +2 [52].
506 Moreover, inhibiting P-TEFb does not generate long pauses at the GAPDH promoter, suggesting
507 that some promoter specific features exist. In the future, it will be interesting to determine
508 whether long-lived and short-lived paused polymerases have a similar 3D structure. Indeed,
509 recent data in NELF KO cells suggests that polymerases can have several pausing sites and states
510 [53]. Because of their half-life, long-lived paused polymerase may display additional features such
511 as backtracking or other stabilizing properties, and backtracking was indeed shown to occur at the
512 pause site in the case of HIV-1 [54]. Long-lived paused polymerases are especially interesting
513 because of their properties, which effectively limit transcription but maintain the promoter in an
514 open state [22].
515
516 **Stochastic pausing generates transcriptional bursting**
517 In the traditional model of transcription initiation, polymerase pausing is an obligatory step
518 during the formation of the elongation complex [2,55]. In contrast, our live cell data on HIV-1
519 transcription suggest that pausing is a stochastic event that occurs rarely: 1 every 20-180
520 polymerase in absence of Tat and down to 1 every 40-3900 when Tat is abundant. A model

reconciling these views would be the existence of two fates during pausing: a pause could lead to either rapid enzyme maturation, or to a long-lived inactive state that would inhibit further transcription. In this scenario, the polymerases initiating at the HIV-1 promoter would mature into a processive elongation complex but would have a low probability of entering a long-lived paused state (Figure 7C). A key feature of this model is that long-lived pauses are stochastic, and this changes the nature of this process as long-lived pauses would not be a step required for proper polymerase maturation but an inhibitory state preventing transcription. In essence, stochastic long-lived pauses are analogous to an inactive promoter state (Figure 7C). In the case of HIV-1, long-lived pauses would be a key regulatory step in transcriptional regulation, and by ensuring an efficient recruitment of P-TEFb, Tat would drastically reduce the probability of long inhibitory pauses (see model in Figure 7C). This is consistent with the fact that the HIV-1 promoter is fully occupied in a model of latent cells [56], even if in some cases Tat can slightly enhance PIC occupancy [49]. It is also consistent with the known function of Tat as a P-TEFb/SEC recruiting factor, with a major function in reducing pausing.

The basal activity of the HIV-1 promoter requires P-TEFb and it is surprising that the factors responsible for P-TEFb recruitment in absence of Tat allow for the firing of a series of polymerases before switching back to a long inactive state. Indeed, the HIV-1 promoter is active for periods of ~ 5 minutes in absence of Tat, firing 20 polymerases on average. A possibility to explain this behavior would be a switching mechanism, in which P-TEFb would be present and active for several minutes at the HIV-1 promoter, and then leave for long time periods. Our data show that NF-κB is not involved in the basal transcriptional activity of HIV-1 in our cellular system, and we can thus rule out sporadic activation of this pathway as a cause of transcriptional bursts in absence of Tat. Another possibility would involve the diffusion dynamics of P-TEFb.

Indeed, it has been shown that P-TEFb is a local explorer that repetitively visit the same location [57], and recent data further suggest that P-TEFb undergoes transient liquid-liquid phase transitions [58]. FRAP studies showed that the residency time of P-TEFb is 11 seconds at the HIV-1 promoter in absence of Tat [59], and 55 seconds at the transcription site of a CMV-based reporter [58]. While this is too short to explain the 5 minutes active periods without Tat, single particle tracking of P-TEFb subunits indicate a wide range of binding times [58]. Moreover, P-TEFb also might exchange rapidly from longer-lasting liquid condensates. It is also possible that other phenomena are responsible for P-TEFb recruitment, or that long pauses arise from an inherently stochastic and inefficient process.

The stochastic nature of long-lived polymerase pausing and their low probability has important consequences for HIV-1 pathogenesis. There are evidences that the stochastic activation of the viral promoter is responsible for the stochasticity of latency exit, at least in part [32-34,37]. Moreover, latent viruses do not express Tat or at very low levels [35,36], and we show that in these conditions the spontaneous release of a long-lived pause leads to the synthesis of a large series of viral RNAs. In some cases, this may be sufficient to activate the viral promoter and to initiate the Tat positive feedback loop, leading to acute viral replication. The stochastic nature of long-lived pausing may thus be an important feature of HIV-1 regulation that favorizes spontaneous latency exit [34,37,38]. It is also possible that even if the viral RNAs produced do not initiate the Tat-feedback loop, they may still produce a small amount of viral particles, which may infect naive cells and could thus participate in the viral rebounds or viremia blips seen in patients. It is also important to note that quiescent memory T cells have a low P-TEFb activity [35,36], possibly leading to very long periods without HIV-1 transcription. Finally, stochastic pausing has also been reported in developing Drosophila embryos, where it may finely tune gene

567 expression after zygotic genome activation (see accompanying paper). Stochastic pausing may be

568 a general property of cellular promoters important for gene regulation.

569

Methods

Cell culture and drug treatments

HeLa Flp-in H9 cells (a kind gift of S. Emiliani) were maintained in DMEM supplemented with 10% fetal bovine serum, penicillin/streptomycin (10 U/ml) and glutamin (2.9 mg/ml), in a humidified CO_2 incubator at 37°C. Cells were transfected with the indicated plasmids with JetPrime (Polyplus), following manufacturer recommendations. Drugs were used at the following concentrations: Triptolide, 1 µM; KM05382 100 µM; BAY11-7082, 2 µM.

Stable expression of MCP-GFP was achieved by retroviral-mediated integration of a self-inactivating vector containing an internal ubiquitin promoter (as described in [28]). The MCP used dimerizes in solution and contained the deltaFG deletion, the V29I mutation, and an SV40 NLS. MCP-GFP expressing cells were grown as pool of clones and FACS-sorted to select cells expressing low levels of fluorescence. Isogenic stable cell lines expressing the 128xMS2 HIV-1 reporter gene were created using the Flp-In system and a HeLa H9 strain expressing various levels of Tat (see below) and MCP-GFP. Flp-In integrants were selected on hygromycin (150 µg/ml). For each construct, several individual clones were picked and analyzed by in situ hybridization.

No Tat cells expressed the 128xMS2 HIV-1 reporter gene but did not express any Tat protein. To obtain low level of Tat expression, a Tat-Flag fused to an Auxin-inducible degron (AID) and cloned as a second cistron after auxin receptor F-box protein AFB2 and instead of GFP in a previously described vector AAV-CAGGS-eGFP [60]. The resulting vector was integrated in genomic AAVS1 site using CRISPR-Cas9 and clones were selected using puromycin as described [60]. Cells were not treated with Auxin.

592 *High Tat* cells [28] were created using the plasmid pSpoII-Tat. In this plasmid, the CMV
593 promoter transcribes a Tat-Flag cDNA followed by an IRES-Neo selectable marker. Following
594 Neomycin selection (400 µg/ml), expression levels of individual clones were verified by western
595 blotting and by immunofluorescence to ensure homogeneity both between clones and between
596 cells of a clone.
597
598 **Plasmids**
599 Sequences of the plasmids are available upon request. The 128xMS2 HIV-1 reporter and High
600 Tat expression vector were described previously [28]. AAV-CAGGS-eGFP vector, used to obtain
601 low Tat cells, Cas9 encoding vector and AAVS1-site targeting RNA-guides were obtained from
602 Dr. G. M. Church [60]. pcDNA3-CDK9-GFP and pcDNA3-CyclinT1-GFP plasmids were obtained
603 by Gateway technology, CDK9 and Cyclin-T1 were amplified by PCR from the vectors provided
604 by Dr. L. Lania [61]. pHR-SFFV-dCas9-BFP plasmid used for CDK9 cloning is #46911 from
605 Addgene. The RNA guides were cloned in a home-made U6 expression vector with an optimized
606 guide RNA scaffold [62].
607
608 **dCas9 tethering and pTEFb overexpression**
609 For P-TEFb overexpression, Hela 128xMS2 HIV-1 No Tat cells without MCP-GFP were plated
610 on coverslips and the next day transfected with CDK9-GFP, Cyclin-T1-GFP, or both, using
611 jetprime (polyplus). pBluescript was used as a negative control and GFP-Tat as a positive control.
612 24-hours after transfection cells were fixed and the reporter RNA was detected by smFISH with
613 Cy3-labeled fluorescent probes against MS2 repeats, the RNA expression was scored in
614 transfected GFP-positive cells.

For CDK9 tethering, RfB gateway cassette was cloned in pHR-SFFV-dCas9-BFP between dCas9 and BFP. CDK9 was next introduced by LR recombination. The resulting plasmid pHR-SFFV-dCas9-CDK9-BFP was transfected in Hela *No Tat* cells without MCP-GFP together with 3 RNA guides encoding plasmids as described above. pHR-SFFV-dCas9-CDK9-BFP without guides and pHR-SFFV-dCas9-BFP were used as controls. 24 hours after transfection cells were fixed and subjected to smFISH with probes against 128xMS2. The numbers of RNA molecules in BFP-positive cells were counted using FISH-QUANT [63,64]. The sequences of RNA guides were as follows CCGCCTAGCATTTCATCACG, CCACGTGATGAAATGCTAGG, TGCTACAAGGGACTTTCCGC.

SmFISH and RNA quantification

SmFISH was performed as previously described [28], with a mix of 10 fluorescent oligos hybridizing against the MS2x32 repeat, each oligo containing four molecules of Cy3. Since each oligo bound four times across the 128xMS2 repeat, each molecule of pre-mRNA hybridized with 40 oligos, thereby providing excellent single molecule detection and signal-to-noise ratios.

To obtain the number of nascent and released pre-mRNAs per cell and the distribution of this parameter in the cell population, cells processed for smFISH were imaged on a ZEISS Axioimager Z1 wide-field microscope (63X~, NA 1.4; 40X~, NA 1.3), equipped with an sCMOs Zyla 4 .2 camera (Andor) and controlled by MetaMorph (Universal Imaging). 3D image stacks were collected with a Z-spacing of 0.3 µm. Figures were prepared with Image J, Photoshop and Illustrator (Adobe), and graphs were generated with R or MatLab.

Raw, 3D smFISH images were analyzed to count the number of pre-mRNA per nuclei, using populations of >300 cells per experiment. Briefly, nuclei were segmented using the DAPI

638 signal with Imjoy [65], and transcription sites (TS) were identified manually. Isolated pre-mRNA
639 molecules located in the nucleoplasm were then detected with *FISH-quant* [63,64], after manual
640 thresholding of Laplacian on Gaussian filtered image. This defined the PSF and the total light
641 intensity of single molecules, which were averaged to obtain an average PSF. The average PSF of
642 single RNA molecule was used to determine the number of nascent pre-mRNA molecules at the
643 TS.
644
645 **Live cell imaging**
646 Cells were plated on 25 mm diameter coverslips (0.17 mm thick) in non-fluorescent media
647 (DMEM gfp-2 with rutin; Evrogen). Coverslips were mounted in a temperature-controlled
648 chamber with CO_2 and imaged on an inverted OMXv3 Deltavision microscope in time-lapse
649 mode. A 100x, NA 1.4 objective was used, with an intermediate 2X lens and an Evolve 512x512
650 EMCCD camera (Photometrics). Stacks of 11 planes with a z-spacing of 0.6 µm were acquired.
651 This spacing still allowed accurate PSF determination without excessive oversampling.
652 Illuminating light and exposure time were set to the lowest values that still allowed visualization
653 of single molecules of pre-mRNAs (laser at 1% of full power, exposure of 15 ms per plane). This
654 minimizes bleaching and maximizes the number of frames that can be collected. Yet, it
655 guarantees that transcription can be detected early on, when one or a few nascent chains are in the
656 process of being transcribed. For short movies, one stack was recorded every 3 seconds for 15 to
657 20 minutes. For long movies, one stack was recorded every three minutes for 8 hours.
658
659 **Quantification of short movies**

Extract the TS signal in the short movies was done as previously described [28]. We manually defined the nuclear outline and the region within which the TS is visible. The stack was corrected for photobleaching by measuring the fluorescence loss of the entire nucleus and fitting this curve with a sum of three exponentials. This fitted curve was then used to renormalize each time-point such that its nuclear intensity was equal to the intensity of the first time-point. We then filtered the image with a 2-state Gaussian filter. First, the image was convolved with a larger kernel to obtain a background image, which was then subtracted from the original image before the quantification is performed. Second, the background-subtracted image was smoothened with a smaller Kernel, which enhances the SNR of single particles to facilitate spot pre-detection.

We then pre-detected the position of the TS in each frame of the filtered image by determining in the user-specified region the brightest pixel above a user-defined threshold. If no pixel was above the threshold, the last known TS position was used. Pre-detected position was manually inspected and corrected. Then the TS signal was fitted with a 3D Gaussian estimating its standard deviation σ_{xz} and σ_z, amplitude, background, and position. We performed two rounds of fitting: in the first round all fitting parameters were unconstrained. In the second round, the allowed range was restricted for some parameters, to reduce large fluctuations in the estimates especially for the frames with a dim or no detectable TS. More specifically, the σ_{xz} and σ_z were restricted to the estimated median value +/- standard deviation from the frames where the TS could be pre-detected, and the background was restricted to the median value. The TS intensity was finally quantified by estimating the integrated intensity above background expressed in arbitrary intensity units.

With the live cell acquisition settings, the illumination power was low and we could not reliably detect all individual molecules. We therefore collected right after the end of the movie one 3D stack – termed calibration stack - with increased laser intensity (50% of max intensity,

compared to 1% for the movie), which allowed reliable detection of individual RNA molecules. We also collected slices with a smaller z-spacing for a better quantification accuracy (21 slices every 300 nm). Quantification of TS site intensity in the calibration stack was done with *FISH-quant* as follows: (a) when calculating the averaged image of single RNA molecules, we subtracted the estimated background from each cell to minimize the impact of the different backgrounds; (b) when quantifying the TS in a given cell, we rescaled the average image of single RNA molecules such that it had the same integrated intensity as the molecules detected in the analyzed cell.

To calibrate the TS intensities in the entire movie, i.e. to express the TS intensity as a number of equivalent full-length transcripts, we used the fact that the last movie frame was acquired at the same time as the calibration stack. We then normalized the extracted TS intensity in the movies, I_{MS2}, to get the nascent counts $N_{nasc;calib}$:

$$N_{nasc;calib}(t) = I_{MS2}(t) * (N_{nasc,final} / I_{final}),$$

where $N_{nasc,final}$ stands for the estimated number of nascent transcripts in the calibration stack and I_{final} for the averaged intensity of the last 4 frames. Note that the approach was limited to movies where the TS was active at the movie end since otherwise its intensity could not be quantified. More than 100 cells were used in each condition.

Quantification of long movies

To quantify the long movies acquired at low frames rate (one 3D stack per 3 minutes), we used *ON-quant* [28], a rapid analysis tool that identified the ON and OFF periods and measured their length. This did not require an absolute quantification of the number of nascent pre-mRNAs and we therefore defined an intensity threshold, based on the mean intensity of single molecules, under which a TS is considered to be silent, and above which a TS is considered to be active.

This threshold corresponded to the intensity of 1.5 pre-mRNA. For each cell line between 100 and 150 cells were analyzed.

Mathematical modelling

A detailed description of the algorithm can be found in the Supplemental Text, and the software algorithms are in the Supplemental file.

Deconvolution and RNAPII Positioning

The RNAPII positions were found by combining a genetic algorithm with a local optimisation procedure. Before initiation of the analysis algorithm, several key parameters were established. The RNAPII elongation speed was fixed at 67 bp/s [28]. The reporter construct transcript was divided into three sections consisting of the pre-MS2 fragment (PRE=700 bp), 128xMS2 loops (SEQ=2900 bp), and post-MS2 fragment (POST=1600 bp). An extra time P_{poly}=100s was added to POST, corresponding to the polyadenylation signal (during this time the polymerase has finished transcription and waits on the transcription site). The temporal resolution of short movies was 3 s/frame. This frame rate is sufficient to detect processes that occur on the order of seconds.

The possible polymerase positions were discretized using a step of 30 bp. This step was chosen as it is smaller than the minimum polymerase spacing and large enough to have a reasonable computation time. For a movie of 20 min length this choice corresponds to a maximum number of 2680 positions. The deconvolution algorithm was implemented in Matlab R2020a using Global Optimization and Parallel Computing Toolboxes for optimizing RNAPII positions in parallel for all nuclei in a collection of movies. The resulting positions are stored for

731 analysis in the further steps of our computational pipeline. The deconvolution step is common to
732 all of the MS2 data analysis pipelines.
733
734 *Long movies waiting time distribution*
735 For long movies, the low resolution (3min) does not allow RNAPII positioning. In this case we
736 binarize the signal by considering that the transcription site is active or inactive if the measured
737 intensity is above or below a threshold level, respectively. The inactive intervals indicate long
738 waiting times between successive polymerases. The active intervals are used to estimate the
739 probability that waiting times are larger than the movie resolution (see Supplemental Text).
740
741 *Multi-exponential regression fitting of the survival function and model reverse engineering using*
742 *the survival function*
743 Data from several short movies corresponding to the same phenotypes was first pooled together.
744 Waiting times were extracted as differences between successive RNAPII positions from all the
745 resulting traces and the corresponding data was used to estimate the nonparametric cumulative
746 short movie distribution function by the Meyer-Kaplan method. Data from long movies and the
747 same phenotype are also pooled and generate the nonparametric cumulative long movie
748 distribution function. The two conditional distribution functions are fitted together into a
749 multiscale cumulative distribution function using the total probability theorem and estimates of
750 two parameters p_l and p_s, representing the probabilities that waiting times are longer than the long
751 movie resolution, and longer than the length of the short movie, respectively (see Figure 4 and
752 Supplemental Text for details).
753 Then, a multi-exponential regression fitting of the multiscale distribution function
754 produced a set of 2N-1 distribution parameters, where N is the number of exponentials in the

regression procedure (3 for N=2 and 5 for N=3). The regression procedure was initiated with multiple log-uniformly distributed initial guesses and followed by local gradient optimisation. It resulted in a best-fit solution with additional suboptimal solutions (local optima with objective function value larger than the best fit).

The 2N-1 distribution parameters can be computed from the 2N-1 kinetic parameters of a N state transcriptional bursting model. Conversely, a symbolic solution for the inverse problem was obtained, allowing computation of the kinetic parameters from the distribution parameters and reverse engineering of the transcriptional bursting model. In particular, it is possible to know exactly when the inverse problem is well-posed, i.e. there is a unique solution in terms of kinetic parameters for any given distribution parameters in a domain.

The transcriptional bursting models used in this paper are as following:

For N=2, there were 3 distribution parameters and 3 kinetic parameters.

The distribution parameters are $A_1, \lambda_1, \lambda_2$, defining the survival function

$$S(t) = A_1 e^{\lambda_1 t} + (1 - A_1) e^{\lambda_2 t}.$$

The solution of the inverse problem for the ON-OFF telegraph model (Figure 6A) is

$$k_2 = -S_1, k_1^- = S_1 - \frac{S_2}{S_1}, k_1^+ = \frac{S_3 S_1 - S_2^2}{S_1 (S_1^2 - S_2)},$$

$$S_1 = A_1\lambda_1 + A_2\lambda_2, S_2 = A_1\lambda_1^2 + A_2\lambda_2^2, S_3 = A_1\lambda_1^3 + A_2\lambda_2^3, A_2 = 1 - A_1,$$

where k_2, k_1^+, k_1^- are the initiation rate, the OFF to ON and ON to OFF transition rates, respectively.

For N=3, there were 5 distribution parameters and 5 kinetic parameters.

The distribution parameters are $A_1, A_2, \lambda_1, \lambda_2, \lambda_3$, defining the survival function

776 $S(t) = A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} + (1 - A_1 - A_2) e^{\lambda_3 t}$.

777 The inverse problem has a unique solution for the 3 state model (stochastic, facultative pause)
778 with one OFF state, one PAUSE state and one ON state (Figure 6A, model M2 of Supplemental
779 Text). Note that the kinetic parameter of Figure 6A (model M2+) are noted below as follow for
780 model M2: $k_{ini} = k3$; $k_{pause} = k2^-$; $k_{abort} = k2^+$; $k_{release} = 0$.

781 $$k_3 = -S_1, k_2^+ = \frac{1}{2}\left[-L_1 + \frac{S_2}{S_1} - \frac{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}{S_1}\right], k_2^- = \frac{1}{2}\left[S_1 - \frac{S_2}{S_1} + \frac{-S_1^2L_1 + S_1S_2 + S_1L_2 - L_3 + \frac{S_2^2}{S_1} - S_3}{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}\right],$$

782 $$k_1^+ = \frac{1}{2}[-L_1 + \frac{S_2}{S_1} + \frac{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}{S_1}], k_1^- = \frac{1}{2}[S_1 - \frac{S_2}{S_1} - \frac{-S_1^2L_1 + S_1S_2 + S_1L_2 - L_3 + \frac{S_2^2}{S_1} - S_3}{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}],$$

783 where

784 $S_1 = A_1\lambda_1 + A_2\lambda_2 + A_3\lambda_3, S_2 = A_1\lambda_1^2 + A_2\lambda_2^2 + A_3\lambda_3^2, S_3 = A_1\lambda_1^3 + A_2\lambda_2^3 + A_3\lambda_3^3, A_3 =$

785 $1 - A_1 - A_2$,

786 $L_1 = \lambda_1 + \lambda_2 + \lambda_3, L_2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2, L_3 = \lambda_1^3 + \lambda_2^3 + \lambda_3^3$,

787 and $k_3, k_2^+, k_2^-, k_1^+, k_1^-$ are the transcription initiation, OFF to ON, ON to OFF, PAUSE to ON, and
788 ON to PAUSE rates, respectively.

789 Duration of the ON, OFF, and PAUSE states can be calculated thusly:

$$T(PAUSE) = \frac{1}{k_{1+}}, T(OFF) = \frac{1}{k_{2+}}, T(ON) = \frac{1}{k_{1-} + k_{2-}}$$

790 For this model, the steady state probability to be in a given promoter state is

$$p_{PAUSE} = \frac{k_1^- k_2^+}{k_1^+ k_2^{\mp} + k_1^- k_2^{\mp} + k_1^+ k_2^-}, p_{OFF} = \frac{k_1^+ k_2^-}{k_1^+ k_2^+ + k_1^- k_2^+ + k_1^+ k_2^-}, p_{ON}$$

$$= \frac{k_1^+ k_2^+}{k_1^+ k_2^+ + k_1^- k_2^+ + k_1^+ k_2^-}.$$

791 The alternative 3 state model with obligatory pause (Figure 6A, model M3) satisfies the
792 following relation among distribution parameters (see Supplemental Text for a proof):

$$A_1\lambda_1 + A_2\lambda_2 + (1 - A_1 - A_2)\lambda_3 = 0.$$

This means that only 4 and not 5 distribution parameters are free, which further constrains the three exponential fitting. In order to infer this model, a constrained fitting was performed but the bad quality of fitting recommended rejection of the model (Figure 6B-C; see results).

Testing the method with artificial data

The entire computational pipeline was tested using artificial data. Artificial traces were generated by simulating the model using the Gillespie algorithm with parameter sets similar to those identified from data. The simulations generated artificial polymerase positions, from which a first version of the signal was computed by convolution. The results are provided in Figure 5 and Supplemental Text.

Error intervals

Distribution parameters result from multi-exponential regression fitting using gradient methods with multiple initial data. These optimization methods provide a best fit (global optimum) but also suboptimal parameter values. Using an overflow ratio (a number larger than one, in our case 2) to restrict the number of suboptimal solutions, we define boundaries of the error interval as the minimum and maximum parameter value compatible with an objective function less than the best fit times the overflow.

mRNA levels

Steady state mRNA levels can be computed from the parameters of the multi-exponential fit. We showed in the Supplemental Text that:

$$mRNA = -\frac{T_{mRNA}}{\sum_{i=1}^{N}\frac{A_i}{\lambda_i}},$$

where T_{mRNA} is the mean lifetime of the mRNA. The formula is valid for all N and we have used T_{mRNA} = 45 min [28].

Chromatin immunoprecipitation

High Tat and *No Tat* HeLa cells were treated with 1 µM of triptolide at 0, 10, 30 and 60 minutes. *High Tat* HeLa cells were treated with 100 µM of KM05382 during 1 hour followed by 1 µM of triptolide at 0, 10, 20 and 30 minutes. Cells were cross-linked by adding crosslinking solution (11% formaldehyde, 100 mM NaCl, 1 mM EDTA pH 8, 0.5 mM EGTA pH 8, 50 mM Hepes pH 7.8) directly to cultures (1% final) and incubated for 10 min at room temperature. Then, 250 mM final glycine was added, and cultures were incubated for 5 min at room temperature. Cells were then washed four times with cold PBS, scraped in cold PBS with Protease Inhibitor cocktail and centrifuged at 1350×g for 10 min.Crude nuclei were prepared by hypotonic lysis. The pellet was resuspended in 5 mL of BufferA (50 mM Hepes pH 8.0, 85 mM KCl, 0.5% Triton-X-100, Protease Inhibitor cocktail, 1 mM PMSF), incubated on ice for 10 min and centrifuged at 1350xg for 10 min. Then, the pellet was resuspended in 5 mL of BufferA' (50 mM Hepes pH 8.0, 85 mM KCl, Protease Inhibitor cocktail, 1 mM PMSF) and centrifuged at 1350xg for 10 min. Finally, the pellet was resuspend in 0.9 mL of Buffer B (50 mM Tris-HCl pH 8, 1% SDS, Protease Inhibitor cocktail, 1 mM PMSF), incubated on ice for 10 min and then stored at the -80°. Pellets were sonicated at 4ºC using a Bioruptor (Diagenode) to shear the chromatin to a mean length of 300 bp by repeated cycles (16 cycles of 30 s ON and 30 s OFF). After sonication cellular debris was removed by centrifugation at 20000×g for 10 min. The chromatin solution was diluted 10-fold in FA/SDS Like buffer (50 mM Hepes KOH pH 7.5, 150 mM NaCl, 1% Triton-X-100, 0.1% Na

deoxycholate, Protease Inhibitor cocktail, 1 mM PMSF) and precleared for 1 hour at 4°C with 25 µl of protein G Dynabeads (Invitrogen). The precleared chromatin solution (1.5 × 106 cells) was incubated overnight with 50 µL of BSA-blocked protein G Dynabeads (previously bound with 3 ug of the corresponding antibody, POLII F-12 sc-55492 Lot K1516 Santacruz, during 1 hour at 4ºC). Samples were washed once with FA/SDS buffer (50 mM Hepes KOH pH 7.5, 150 mM NaCl, 1% Triton-X-100, 0.1% Na deoxycholate, 1 mM EDTA, 0.1% SDS, Protease Inhibitor cocktail, 1 mM PMSF), three times with FA/SDS Buffer supplemented with 300mM NaCl, once with washing Buffer (10 mM Tris-HCl pH 8, 0.25 M LiCl, 1 mM EDTA, 0.5% NP40, 0.5% Na deoxycholate) and once with TE Buffer. Elution was performed adding 125 µl of Elution Buffer (25 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.5% SDS) and incubating at 65°C for 25 min. The eluates were digested with 50 µg/mL of RNase A at 37°C for 30 min and with 50 µg/ml of proteinase K at 50°C for 1 h. Then, they were incubated at 65°C overnight to reverse cross-links. DNA was recovered by phenol extraction followed by a Qiaquick purification (PCR purification columns, Qiagen, Germany). Specific sequences in the immunoprecipitates were quantified by real-time PCR using the primers listed below. The signal of each sample was normalized with the average signal obtained from the input of the same sample with each pair of primers used. Each experiment was done analysing two independent biological replicates.

Primers used:

GAPDH promoter F: 5' AAAGGCACTCCTGGAAACCT

GAPDH promoter R: 5' GGATGGAATGAAAGGCACAC

GAPDH negative control F: 5' CTAGCCTCCCGGGTTTCTCT

GAPDH negative control R: 5' ACAGTCAGCCGCATCTTCTT

TSS HIV1 +92 F: 5' GCTTCAAGTAGTGTGTGCCC

861 TSS HIV1 +92 R: 5' GCTTTCAAGTCCCTGTTCGG

862

References

1 Schier, A. & Taatjes, D. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev.* **34**, 465-488, doi:10.1101/gad.335679.119 (2020).

2 Jonkers, I. & Lis, J. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol.* **16**, 167-177, doi:10.1038/nrm3953 (2015).

3 Harlen, K. & Churchman, L. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat Rev Mol Cell Biol.* **18**, 263-273 (2017).

4 Fisher, R. Cdk7: a kinase at the core of transcription and in the crosshairs of cancer drug discovery. *Transcription* **10**, 47-56 (2019).

5 Rimel, J. & Taatjes, D. The essential and multifunctional TFIIH complex. *Protein Sci* **27**, 1018-1037 (2018).

6 Ghosh, A., Shuman, S. & Lima, C. Structural insights to how mammalian capping enzyme reads the CTD code. *Mol Cell.* **43**, 299-310 (2011).

7 Fant, C. *et al.* TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. *Mol Cell.* **78**, 785-793 (2020).

8 Narita, T. *et al.* NELF interacts with CBC and participates in 3' end processing of replication-dependent histone mRNAs. *Mol Cell.* **26**, 349-365 (2007).

9 Vos, S., Lucas Farnung, L., Henning Urlaub, H. & Patrick Cramer, P. Structure of paused transcription complex Pol II-DSIF-NELF. *Nature* **560**, 601-606 (2018).

10 Cheung, A. & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**, 249-253, doi:10.1038/nature09785 (2011).

11 Wei, P., Garber, M., Fang, S., Fischer, W. & Jones, K. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92**, 451-462 (1998).

12 He, N. *et al.* HIV-1 Tat and host AFF4 recruit two transcription elongation factors into a bifunctional complex for coordinated activation of HIV-1 transcription. *Mol Cell.* **38**, 428-438 (2010).

13 Sobhian, B. *et al.* HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. *Mol Cell.* **38**, 439-451 (2010).

14 Nilson, K. *et al.* THZ1 Reveals Roles for Cdk7 in Co-transcriptional Capping and Pausing. *Mol Cell.* **59**, 576-587 (2015).

15 Vos, S. *et al.* Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. *Nature* **560**, 607-612 (2018).

16 Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D. & Handa, H. Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *EMBO J.* **17**, 7395-7403 (1998).

17 Yamada, T. *et al.* P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Mol Cell.* **21**, 227-237 (2006).

18 Ehrensberger, A. H., Kelly, G. P. & Svejstrup, J. Q. Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. *Cell* **154**, 713-715 (2013).

19 Henriques, T. *et al.* Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol Cell.* **52**, 517-528 (2013).

20 Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407 (2014).

21 Buckley, M. S., Kwak, H., Zipfel, W. R. & Lis, J. T. Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev.* **28**, 14-19 (2014).

22 Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet.* **49**, 1045-1051 (2017).

23 Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell.* **67**, 411-422 (2017).

24 Chubb, J., Trcek, T., Shenoy, S. & Singer, R. Transcriptional pulsing of a developmental gene. *Curr Biol.* **16**, 1018-1025 (2006).

25 Pichon, X., Lagha, M., Mueller, F. & Bertrand, E. A Growing Toolbox to Image Gene Expression in Single Cells: Sensitive Approaches for Demanding Challenges. *Mol Cell.* **71**, 468-480, doi:10.1016/j.molcel.2018.07.022 (2018).

26 Rodriguez, J. & Larson, D. Transcription in Living Cells: Molecular Mechanisms of Bursting. *Annu Rev Biochem.* **89**, 189-212 (2020).

27 Lionnet, T. *et al.* A transgenic mouse for in vivo detection of endogenous labeled mRNA. *Nat Methods* **8**, 165-170 (2011).

28 Tantale, K. *et al.* A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nat Commun.* **7**, 12248 (2016).

924 29 Rodriguez, J. *et al.* Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression
925 Heterogeneity. *Cell* **176**, 213-226 (2019).
926 30 Blake, W. *et al.* Phenotypic consequences of promoter-mediated transcriptional noise. *Mol*
927 *Cell.* **24**, 853-865, doi:10.1016/j.molcel.2006.11.003 (2006).
928 31 Raj, A., Rifkin, S., Andersen, E. & A., v. O. Variability in gene expression underlies
929 incomplete penetrance. *Nature* **463**, 913-918 (2010).
930 32 Weinberger, L., Burnett, J., Toettcher, J., Arkin, A. & Schaffer, D. Stochastic gene
931 expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic
932 diversity. *Cell* **122**, 168-192 (2005).
933 33 Ho, Y. *et al.* Replication-competent noninduced proviruses in the latent reservoir increase
934 barrier to HIV-1 cure. *Cell* **155**, 540-551 (2013).
935 34 Rouzine, I., Razooky, B. & Weinberger, L. Stochastic variability in HIV affects viral
936 eradication. *Proc Natl Acad Sci U S A.* **111**, 13261-13262 (2014).
937 35 Mbonye, U. & Jonathan Karn, J. The Molecular Basis for Human Immunodeficiency Virus
938 Latency. *Annu Rev Virol* **4**, 261-285 (2017).
939 36 Shukla, A., Ramirez, N. & D'Orso, I. HIV-1 Proviral Transcription and Latency in the New
940 Era. *Viruses* **12**, 555, doi:10.3390/v12050555 (2020).
941 37 Tyagi, M., Pearson, R. & Karn, J. Establishment of HIV latency in primary CD4+ cells is
942 due to epigenetic transcriptional silencing and P-TEFb restriction. *J. Virol.* **84**, 6425-6437
943 (2010).
944 38 Jiang, G. *et al.* Synergistic Reactivation of Latent HIV Expression by Ingenol-3-Angelate,
945 PEP005, Targeted NF-kB Signaling in Combination with JQ1 Induced p-TEFb Activation.
946 *PLoS Pathog.* **11**, e1005066 (2015).
947 39 Boireau, S. *et al.* The transcriptional cycle of HIV-1 in real-time and live cells. *J Cell Biol*
948 **179**, 291-304 (2007).
949 40 Fusco, D. *et al.* Single mRNA molecules demonstrate probabilistic movement in living
950 mammalian cells. *Curr Biol.* **13**, 161-167 (2003).
951 41 Yedavalli, V. S., Benkirane, M. & Jeang, K. Tat and trans-activation-responsive (TAR)
952 RNA-independent induction of HIV-1 long terminal repeat by human and murine cyclin T1
953 requires Sp1. *J Biol Chem* **278**, 6404-6410, doi:10.1074/jbc.M209162200 (2003).

954 42 Barboric, M., Nissen, R., Kanazawa, S., Jabrane-Ferrat, N. & Peterlin, B. NF-kappaB binds
955 P-TEFb to stimulate transcriptional elongation by RNA polymerase II. *Mol Cell.* **8**, 327-337
956 (2001).
957 43 West, M., Lowe, A. & Karn, J. Activation of human immunodeficiency virus transcription in
958 T cells revisited: NF-kappaB p65 stimulates transcriptional elongation. *J. Virol.* **75**, 8524-
959 8537 (2001).
960 44 Larson, D., Zenklusen, D., Wu, B., Chao, J. & RH., S. Real-time observation of transcription
961 initiation and elongation on an endogenous yeast gene. *Science* **332**, 475-478 (2011).
962 45 Desponds, J. *et al.* Precision of Readout at the hunchback Gene: Analyzing Short
963 Transcription Time Traces in Living Fly Embryos. *PLoS Comput Biol.* **12**, e1005256,
964 doi:10.1371/journal.pcbi.1005256 (2016).
965 46 Coulon, A. & Larson, D. Fluctuation Analysis: Dissecting Transcriptional Kinetics with
966 Signal Theory. *Methods Enzymol.* **572**, 159-191 (2016).
967 47 Corrigan, A., Tunnacliffe, E., Cannon, D. & Chubb, J. A continuum model of transcriptional
968 bursting. *Elife* **5**, e13051 (2016).
969 48 Lammers, N. *et al.* Multimodal transcriptional control of pattern formation in embryonic
970 development. *Proc Natl Acad Sci U S A.* **117**, 836-847 (2020).
971 49 D'Orso, I. & Frankel, A. D. RNA-mediated displacement of an inhibitory snRNP complex
972 activates transcription elongation. *Nat Struct Mol Biol* **17**, 815-821 (2010).
973 50 Vispé , S. *et al.* Triptolide is an inhibitor of RNA polymerase I and II-dependent
974 transcription leading predominantly to down-regulation of short-lived mRNA. *Mol Cancer*
975 *Ther* **8**, 2780-2790 (2009).
976 51 Shao, W., Alcantara, S. & Zeitlinger, J. Reporter-ChIP-nexus reveals strong contribution of
977 the Drosophila initiator sequence to RNA polymerase pausing. *Elife* **8**, e41461,
978 doi:10.7554/eLife.41461 (2019).
979 52 Rittner , K., Churcher, H. J., Gait, M. J. & Karn, J. The human immunodeficiency virus long
980 terminal repeat includes a specialised initiator element which is required for Tat-responsive
981 transcription. *J Mol Biol* **248**, 562-580 (1995).
982 53 Aoi, Y. *et al.* NELF Regulates a Promoter-Proximal Step Distinct from RNA Pol II Pause-
983 Release *Mol Cell.* **78**, 261-274 (2020).

54 Palangat , M. & Landick, R. Roles of RNA:DNA hybrid stability, RNA structure, and active site conformation in pausing by human RNA polymerase II. *J Mol Biol* **311**, 265-282 (2001).

55 Wissink, E. M., Ihervaara, A., Tippens, N. D. & T., L. J. Nascent RNA Analyses: Tracking Transcription and Its Regulation. *Nat Rev Genet.* **20**, 705-723 (2019).

56 Demarchi, F., D'Agaro, P., Falaschi, A. & Giacca, M. In vivo footprinting analysis of constitutive and inducible protein-DNA interactions at the long terminal repeat of human immunodeficiency virus type 1. *J Virol.* **67**, 7450-7460 (1992).

57 Izeddin, I. *et al.* Single-molecule tracking in live cells reveals distinct target-search strategies of transcription factors in the nucleus. **Elife**, e02230 (2014).

58 Lu, H. *et al.* Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318-323 (2018).

59 Molle, D. *et al.* A real-time view of the TAR:Tat:P-TEFb complex at HIV-1 transcription sites. *Retrovirology* **4**, 36 (2007).

60 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826 (2013).

61 Majello, B., Napolitano, G., Giordano, A. & Lania, L. Transcriptional regulation by targeted recruitment of cyclin-dependent CDK9 kinase in vivo. *Oncogene* **18**, 4598-4605 (1999).

62 Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479-1491 (2013).

63 Tsanov, N. *et al.* smiFISH and FISH-quant - a flexible single RNA detection approach with super-resolution capability. *Nucleic Acids Res.* **44**, e165 (2016).

64 Mueller, F. *et al.* FISH-quant: automatic counting of transcripts in 3D FISH images. *Nat Methods.* **10**, 277-278 (2013).

65 Ouyang, W., Mueller, F., Hjelmare, M., Lundberg, E. & Zimmer, C. ImJoy: an open-source computational platform for the deep learning era. *Nat Methods* **16**, 1199-1200 (2019).

Figure Legends

Figure 1. Single cell characterization of HIV-1 gene expression, with and without Tat.

A-Schematic of HIV-1 transcriptional regulation. Left: in absence of Tat, pTEFb is not recruited and polymerases binds NELF and DSIF and pause near the promoter. Right: in presence of Tat, pTEFb, composed of Cyclin T1 and Cdk9 associated to the super elongation complex, is recruited to the nascent TAR RNA. Cdk9 phosphorylates NELF, DSIF and RNA polymerase II, thereby triggering pausing exit and processive elongation.

B-Schematic of the HIV-1 reporter construct. SD1: major HIV-1 splice site donor; SA7: last HIV-1 splice site acceptor; ψ: packaging signal; RRE: Rev-responsive element; LTR: long terminal repeat.

C- Expression of the 128xMS2 HIV-1 tagged reporter in cells expressing high levels of Tat. Left panel: microscopy images of High Tat HeLa cells where the unspliced HIV-1 pre-mRNA is detected by smFISH with probes against the 128xMS2 tag. Cells bear a single copy of the reporter gene integrated with the Flp-in system. The bright spots in the nuclei correspond to nascent RNA at their transcription sites, while the dimmer spots correspond to single pre-mRNA molecules. Scale bar : 10 μm. Middle panel: distribution of the number of released HIV-1 pre-mRNAs per cell, in High Tat cells. Experimental RNA distribution are from smFISH data. X-axis: number of HIV-1 pre-mRNA molecules per cell; y-axis: number of cells; inset: mean number of HIV-1 pre-mRNAs per cell. Right panel: distribution of the number of nascent HIV-1 pre-mRNAs per transcription site, in High Tat cells. Experimental RNA distribution are from smFISH data. X-axis: number of nascent HIV-1 pre-mRNA molecules per transcription site; y-axis: number of transcription sites; inset: mean number of nascent HIV-1 pre-mRNAs per cell.

D- Expression of the 128xMS2 HIV-1 tagged reporter in cells expressing low levels of Tat. Legend as in C, except that experiments are from Low Tat cells.

E- Expression of the 128xMS2 HIV-1 tagged reporter in cells not expressing Tat. Legend as in C, except that experiments are from No Tat cells. Image contrast adjustment is identical for panels C, D and E.

Figure 2. Fluctuation of HIV-1 transcription over short time periods, with and without Tat.

A-C Fluctuations of HIV-1 transcription over 15-20 minute periods, with one image stack recorded every 3 seconds. Left: each graph is a single transcription site; the x-axis represents the time (in minutes) and y-axis represents the intensity of transcription sites, expressed in equivalent numbers of full-length pre-mRNA molecules. Right: each line is a cell and the transcription site intensity is color-coded (scale on the right). A: High-Tat cells; B: Low-Tat cells; C: No-Tat cells.

D-Schematic of a polymerase convoy. Top: a polymerase convoy, with polymerases in orange and the gene represented as a black horizontal arrow. N_{pol} : number of polymerases; t_{space} : spacing between successive RNA polymerases (in seconds); v_{el} : elongation rate. Bottom: schematics describing the different phases of a transcription cycle (left) and the position of the polymerase convoy on the MS2 tagged gene (right; the green box is theMS2 tag).

E-Box-plots representing the parameters values of the best-fit models, measured for a set of isolated transcription cycles in each cell line. t_{proc} is the 3'-end RNA processing time; *Npol* is the number of polymerases in the convoy; V_{el} is the elongation rate (in kb/min); t_{space} is the spacing between successive polymerase (in seconds). The bottom line displays the first quartile, the box corresponds to the second and third quartile, the top line to the last quartile, and the double circle is the median. Small circles are outliers (1.5 times the inter-quartile range above or below the upper and lower quartile, respectively).

Figure 3. Fluctuation of HIV-1 transcription over long time periods, with and without Tat.

A-Fluctuations of HIV-1 transcription over 8 hours, with one image stack recorded every 3 minutes. The x-axis represents the time (in hours) and y-axis represents the intensity of transcription sites, expressed in arbitrary units. Periods of HIV-1 promoter activity are colored in green, and periods of inactivity in red.

B-Active and inactive periods of the HIV-1 promoter, for the indicated cell lines. Each line is a cell and the activity of the HIV-1 promoter is color-coded (green: active; red: inactive), using the threshold shown in panel A. x-axis: time in hours.

Figure 4. Analysis and modeling strategy for the live cell transcriptional data.

A- Determination of models for transcription initiation. Left: example of a complex promoter models describing the different steps leading to transcription initiation and their kinetic relationship. OFF: inactive promoter state; ON: active promoter state; orange ball: RNA polymerase. Right: the survival function (equal to one minus the cumulative function) describes the distribution of polymerase waiting times (delay between two successive initiation events). For linear models such as the one depicted on the left, the survival function can be fitted by a sum of exponentials, with the number of exponentials being equal to the number of promoter states. Branched models also lead to sums of exponentials (see text).

B- Experimental and machine learning strategy to determine the survival function of polymerase waiting times. Left: signals of short movies made at high temporal resolution result from the convolution of the signal from a single polymerase and the sequence of temporal positions of initiation events. The sequence of initiation events can thus be reconstructed by a deconvolution numerical method, provided that the signal of a single polymerase is known. This allows to

estimate the distribution of waiting times for waiting times shorter than the movie duration (i.e. a conditional distribution). Right: long movies made with a lower temporal resolution, in the order of the residency time of RNA polymerase on the gene (3 minutes), allow to estimate the distribution of polymerase waiting times for waiting times greater than the temporal resolution. The two conditional survival functions, short and long, can then be combined to reconstitute the complete survival function, with the constraint that waiting times of short movies, smaller than the frame rate of the long movie, must fill the active periods in the long movie. Finally, the complete survival function is fitted with a sum of exponentials to determine the number of promoter state, the kinetics of transitions between them, and the initiation rate. Multiple models can be easily fitted to the same survival function and the most appropriate one is selected based on parsimony, parametric indeterminacy and consistency with complementary experiments.

Figure 5. Accuracy and robustness of the analysis and modeling pipeline.

A- Fidelity and robustness of the deconvolution method. Left panels: simulation of short movies for an artificial set of polymerase initiation events, with noise added (bottom), or without (top). x-axis is time in minutes; y-axis is the intensity of transcription sites (expressed in number of RNA molecules). Right panels: positions of the transcription initiation events (vertical bars), for the original artificial data (black; bottom lines), the reconstructed data from the simulated short movies after the genetic algorithm (GA, red, middle lines), or the final reconstruction after both the GA and the local optimization (blue; top lines). x-axis is time in minutes.

B- Fidelity and robustness of the overall analysis pipeline. Top schematic: the linear three state promoter model used for Monte Carlo simulations. Top graphs: examples of artificial short movies (black lines), with various levels of noise added (red lines). Note that the noise level measured experimentally corresponds to the 1x condition. x-axis is time in seconds; y-axis is the

intensity of transcription sites expressed in number of RNA molecules. Middle graphs: survival functions reconstructed from artificial short and long movies (red and green circles, respectively), and fitted to a sum of three exponentials (black line). The theoretical survival function obtained with the model parameters used for the simulation is shown for comparison (blue line). x-axis: time intervals between successive initiations events, in seconds and in $\log_{10}$ scale. y-axis: probability of $\Delta t > x$ ($\log_{10}$ scale).

C- Accuracy of determining the model parameters. Graphs plot the parameters used to generate the artificial data (x-axis), against the parameter measured by the deconvolution and fitting procedure (y-axis). Vertical bars: confidence intervals. Three parameter sets were used, corresponding to the values obtained with the experimental data from the High Tat cells (circles), Low Tat cells (crosses), and No Tat cells (triangles).

Figure 6. A facultative pausing model reproduces the live cell transcription data and predicts a long-lived pause.

A- Schematics of the different models used to fit the live cell HIV-1 transcriptional data. Left: a two-state ON/OFF promoter mode; middle: a three state promoter model including an obligatory pause as traditionally represented (model M3); right: a three state promoter model with a facultative pause (model M2+). Polymerases are represented by small orange balls.

B-Fits of the experimental survival functions. Graphs represent the survival functions reconstructed from the live cell data for the High Tat, Low Tat and No Tat conditions, with the part deriving from the short and long movies in red and green, respectively. Blue line: fit of the 3-state model with a facultative pause; "+": fit of the 3-state model with an obligatory pause; "x":

fit with a facultative pause. x-axis: time intervals between successive initiations events, in seconds and in $\log_{10}$ scale. y-axis: probability of $\Delta t > x$ ($\log_{10}$ scale).

C- Model scores. The graph depicts the score of each model (inverse of the minimal value of the fitted Objective Function), for each of the model and cell line.

D- Pausing characteristics predicted by the models. Top: predicted pausing times, for the relevant models and cell lines (see text for details). Bottom: predicted pausing frequencies (in %), for the indicated cell line and model. For the model with the facultative pause, the two indicated values come from the two branches of the model that could each correspond to the paused state (see model M2 in the Supplemental Text).

E- Features of the model with the facultative pause. Left: the graphs represent the number of mRNA per cell measured by smFISH experiments (violet bars), or predicted from the model parameters (blue bars). Error bars are the standard deviation for the smFISH data (estimated from replicate measurements) and the confidence intervals for the prediction from the model. Middle: initiation rate (in s^{-1}), for the three cell lines. Error bars are confidence intervals. Right: fraction of the cells with the promoter in the ON state (in %), for the three cell lines. Error bars are confidence intervals.

Figure 7. Biochemical measurements indicate a long-lived paused state at the HIV-1 promoter.

A- Residency time of RNA polymerase II at the HIV-1 promoter. The graph depicts the RNA polymerase II ChIP signals at the HIV-1 and GAPDH promoters during a Triptolide time course experiment, for the High Tat and No Tat cell lines. GAPDH TSS: transcription start site of the human GAPDH gene; HIV-1 TSS: transcription start site of the HIV-1 promoter; Control DNA: a

non-transcribed genomic locus. ChIP signals were measure by qPCR and values are expressed as percent of input and normalized to the zero time point. For the control genomic regions (Control DNA), values are normalized to that of GAPDH TSS at time zero.

B- Effect of pTEFb inhibition on the residency time of RNA polymerase II at the GAPDH promoter. Legend as in panel A, except that the KM sample was pretreated with the Cdk9 inhibitor KM05382 for 2h before triptolide addition.

C- Model depicting the dynamics of the HIV-1 promoter and highlighting the positive and negative effects of Tat. The numbers are from the facultative pausing model fitted to the High Tat and No Tat data (see Figure 6C and supplemental text, Table 3). The model with facultative pausing has two symmetrical branches (see model M2 in the Supplemental Text), and each branch of the model could correspond to the paused state. The values indicated attribute the pause state to the branch that is most affected by the presence of Tat.

Supplemental Figure S1. Transcriptional activation of HIV-1 128xMS2 reporter in Hela Flp-in cells is P-TEFb dependent.

A- Western blot of the extracts of HIV-1 128xMS2 Hela cell lines with no, low and high Tat expression. Tat-Flag was detected with anti-Flag antibodies; loading control is tubulin.

B- CDK9-GFP and cyclinT1-GFP activate transcription of the HIV-1 reporter. Fluorescent microscopy images of Hela Flp-in cells with the HIV-1 128xMS2 reporter, not expression Tat nor MCP-GFP, and co-transfected with plasmids encoding for CDK9-GFP and cyclinT1-GFP (24h after transfection). First row from the left: RNA of HIV-1 reporter detected by smFISH with Cy3 probes against 128xMS2 tag; second row: GFP signal corresponding to the cells transfected with CDK9-GFP and cyclinT1-GFP; third row: nuclear staining with dapi; last row: merge. Top panel:

cells transfected with CDK9-GFP and cyclinT1-GFP. Bottom panel: control transfection with pBluescript. The scale bar is 10 μm.

C- Tethering of CDK9 to the HIV-promoter using dCas9 leads to transcriptional activation. The histogram shows the results of mRNA counting on smFISH images 24h after transfection the Hela Flp-in HIV-1 128xMS2 no Tat cells (without MCP-GFP) with dCas9-CDK9-BFP fusion and 3 RNA guides targeting the CDK9 fusion specifically to the HIV-1 promoter (middle bar); dCas9-CDK9-BFP fusion without guides (right bar) or dCas9-BFP alone (left bar) were transfected in control experiments. On y axis is the mRNA number. Error bars are standard errors of the mean.

Supplemental Figure S2. Transcriptional activation of HIV-1 reporter in the absence of Tat depends on enzymatic activity of CDK9 and is independent of NF-κB pathway.

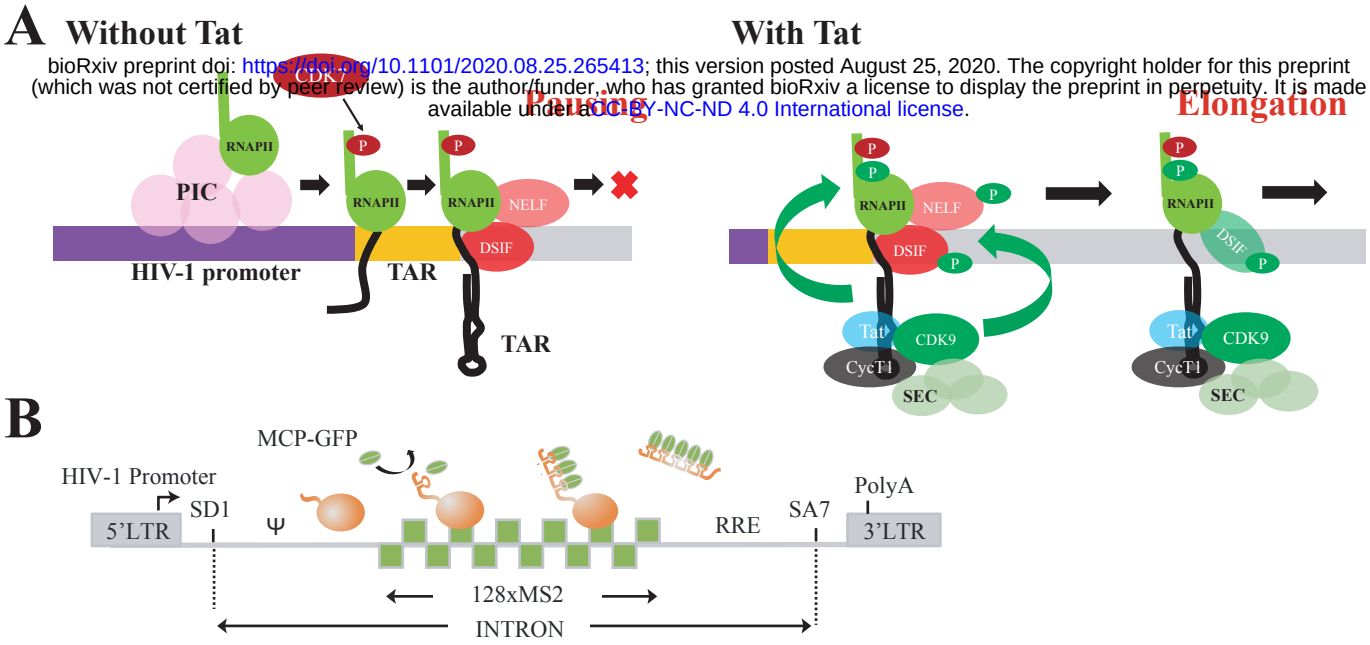
A- CDK9 inhibitor KM05283 inhibits HIV-1 transcription. Images of Hela Flp-in HIV-1 128xMS2 MCP-GFP no Tat cells treated with 100 μM KM05382 for 4 h, using GFP filter. Left – non-treated control; right – 4 hours of KM05382 treatment. The scale bar is 10 μM.

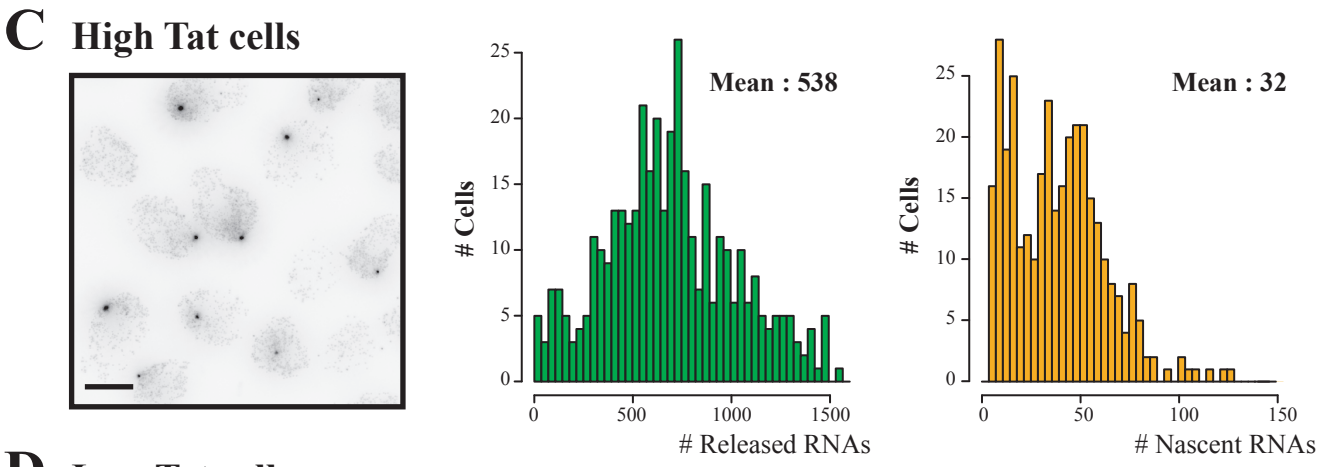
B- NF-κB inhibitor BAY11-7082 does not affect HIV-1 reporter transcription. Left panel: Images of smFISH with Cy3 labeled probes of the cells Hela Flp-in HIV-1 128xMS2 MCP-GFP no Tat. Left - non-treated control; right -16h treatment with 2 μM BAY11-7082.

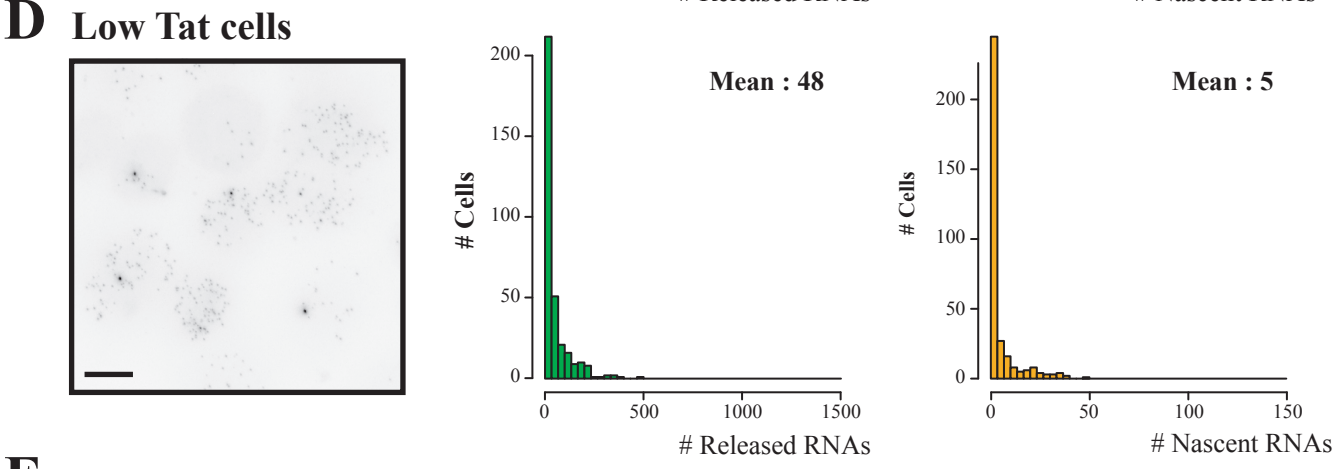
C- Histogram showing the quantification of mature and nascent RNA number on the smFISH images after 16h inhibition of NF-kB with 2 μM BAY11-7082. On y axis is the RNA number. Error bars are standard deviations.

Supplemental Figure S3. RNA Pol II ChIP in presence and absence of Tat.

1196 The graph depicts the RNA polymerase II ChIP signals at HIV-1 and GAPDH loci for the High
1197 Tat and No Tat cell lines. GAPDH TSS: transcription start site of the human GAPDH gene; HIV-
1198 1 TSS: transcription start site of the HIV-1 promoter; Control DNA: a non-transcribed genomic
1199 locus. ChIP signals were measure by qPCR and values are expressed as percent of input (y axis).
1200 The scale bar is 10 μM.
1201







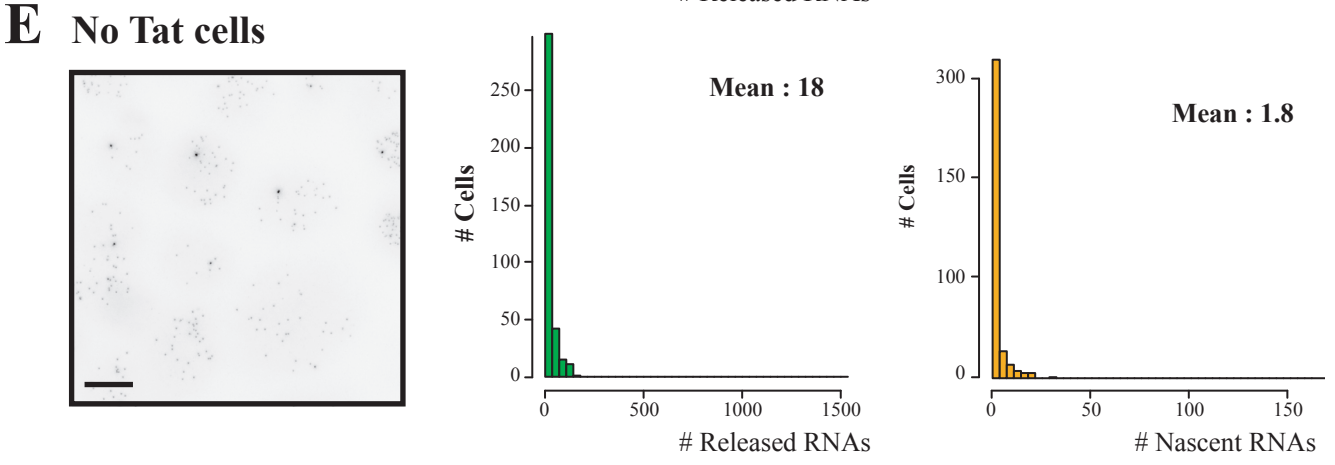


FIGURE 1

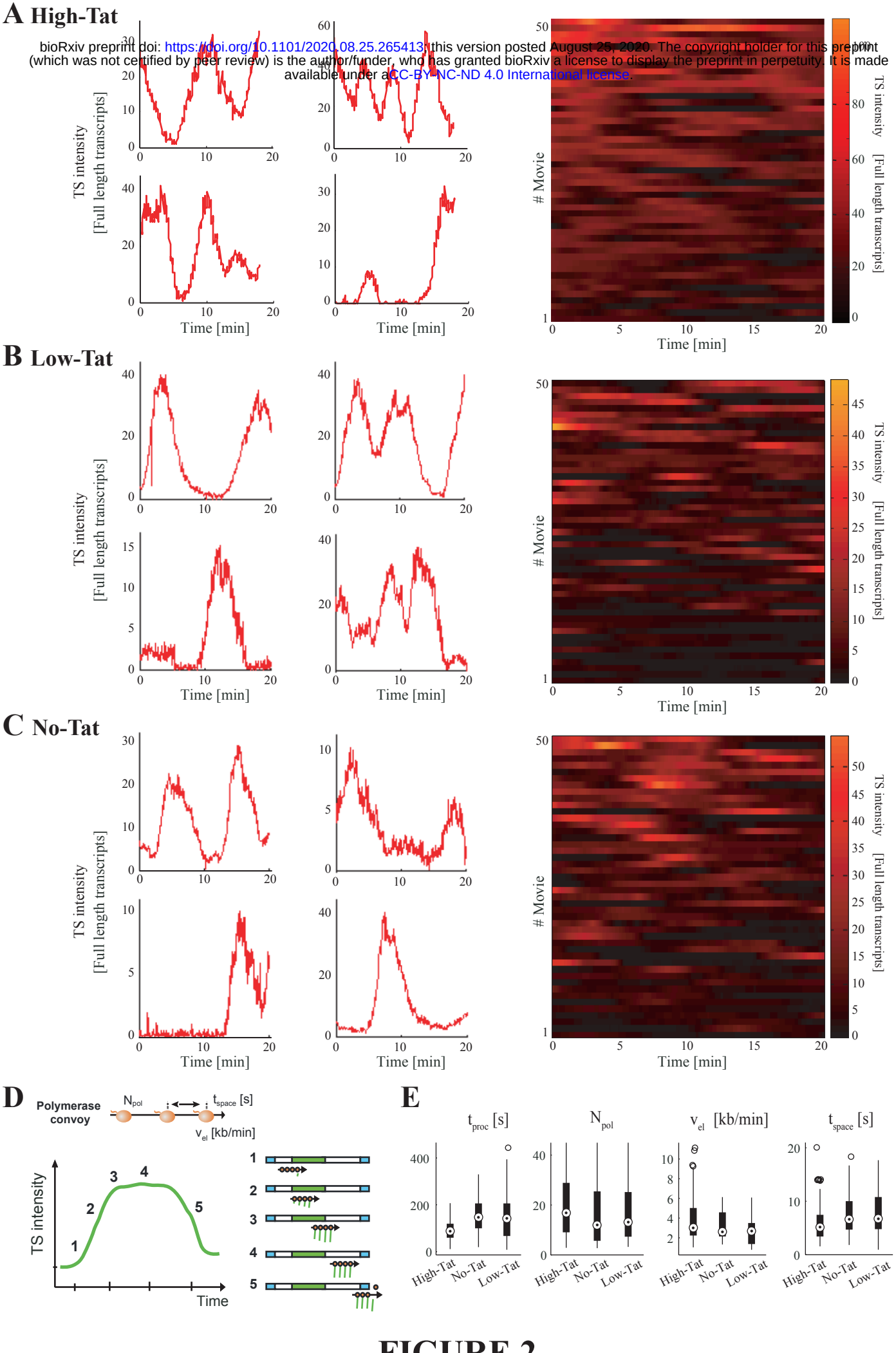


FIGURE 2

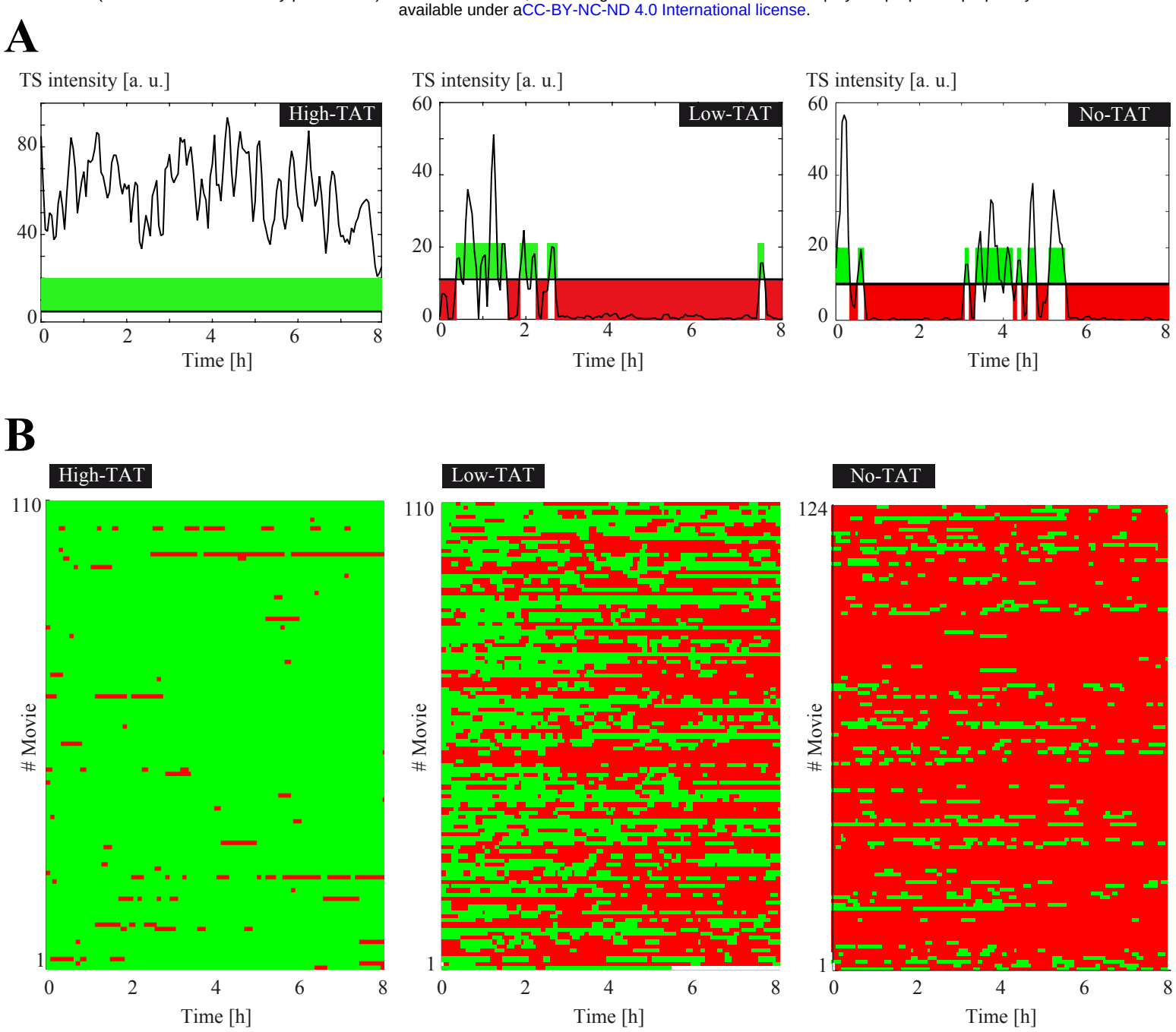


FIGURE 3

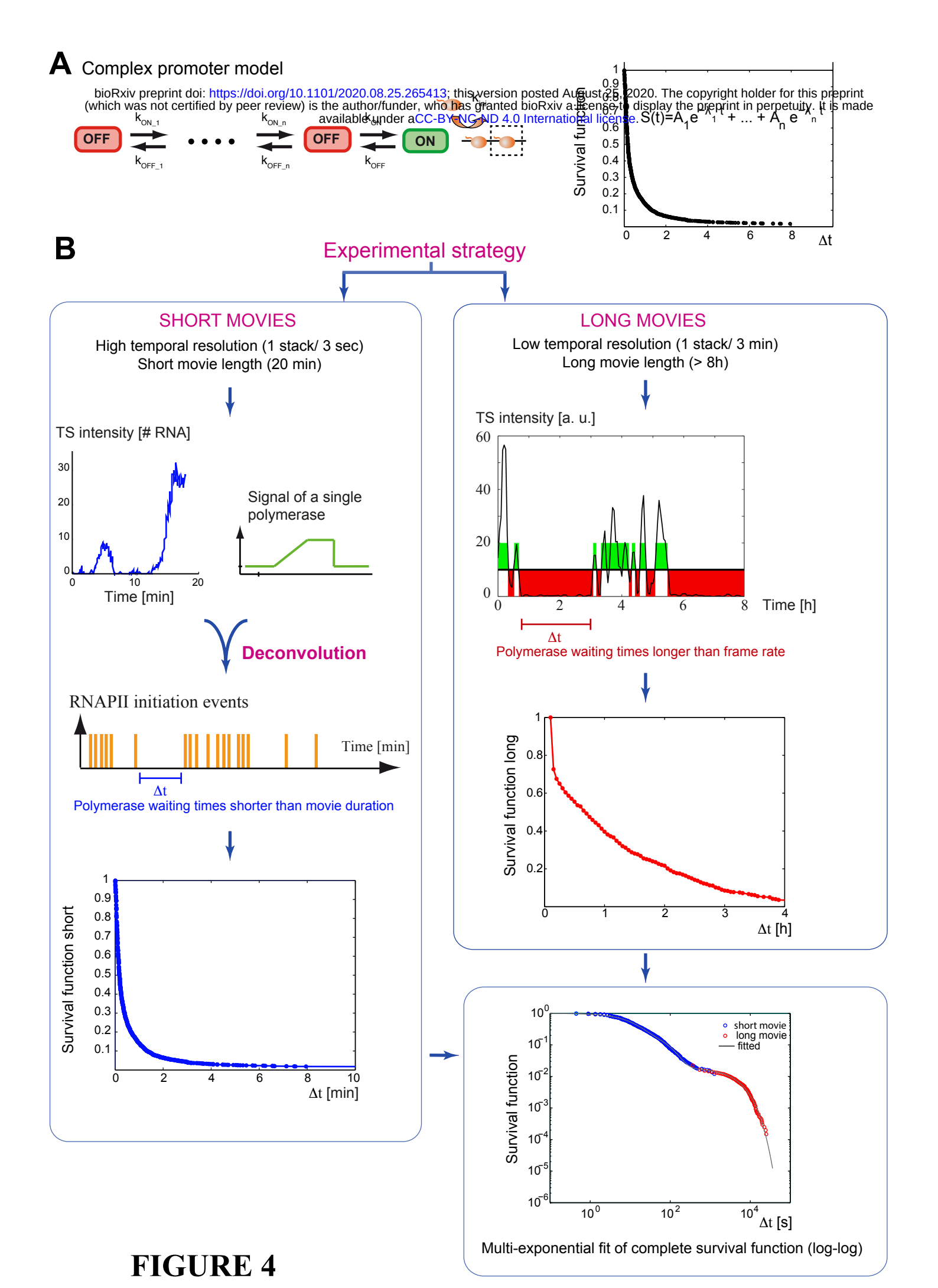
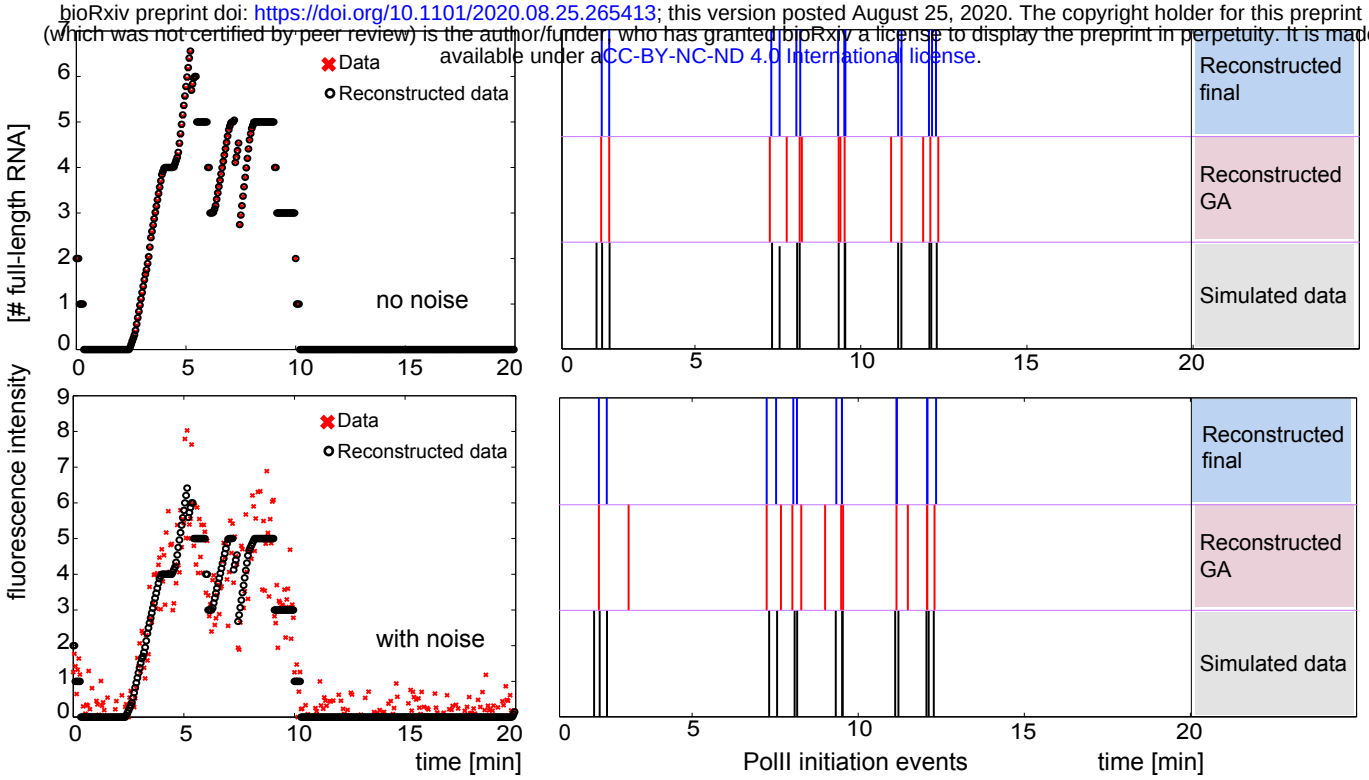


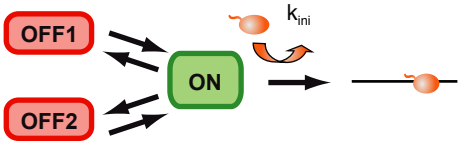
FIGURE 4

A Accuracy and robustness of the two-step deconvolution method: 1-genetic algorithm; 2-local optimisation

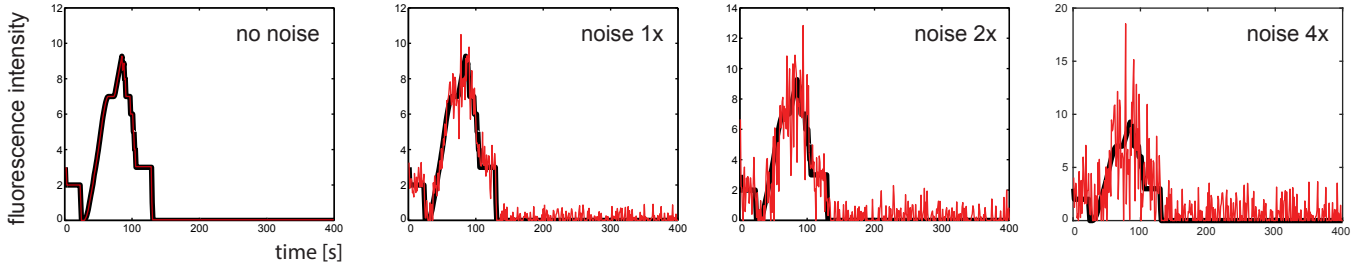


B Accuracy and robustness of the overal analysis and modeling strategy

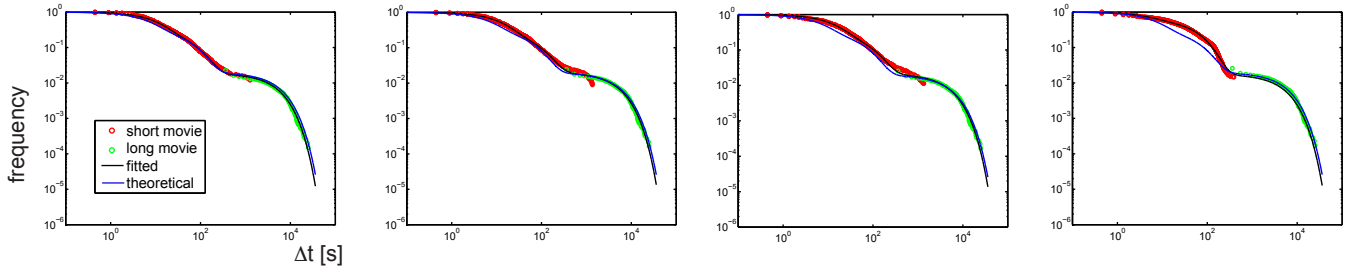
Simulated model :



Simulated short movies



Reconstructed and fitted survival functions



Accuracy of estimated model parameters, for a range of simulated values

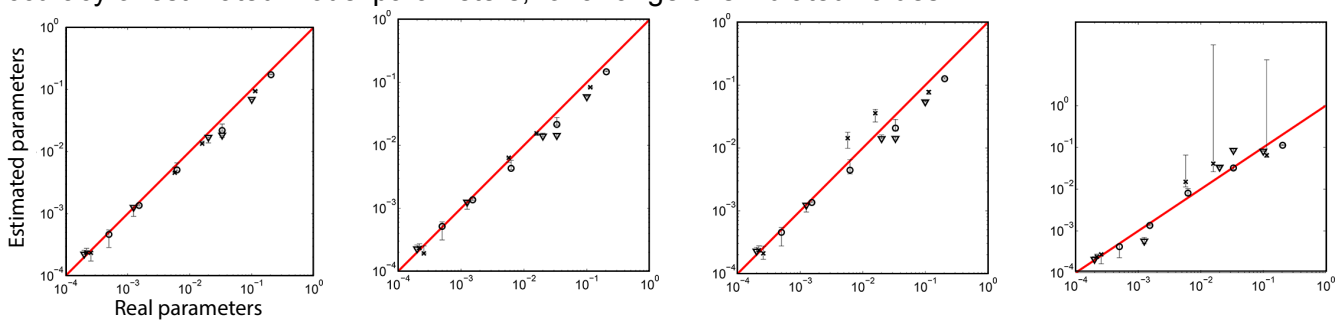
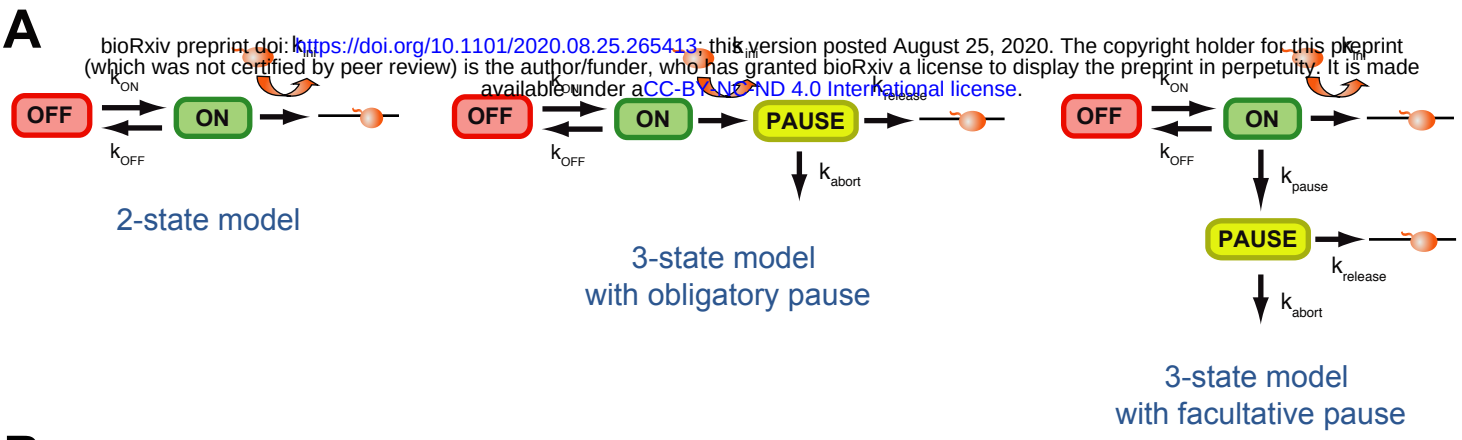
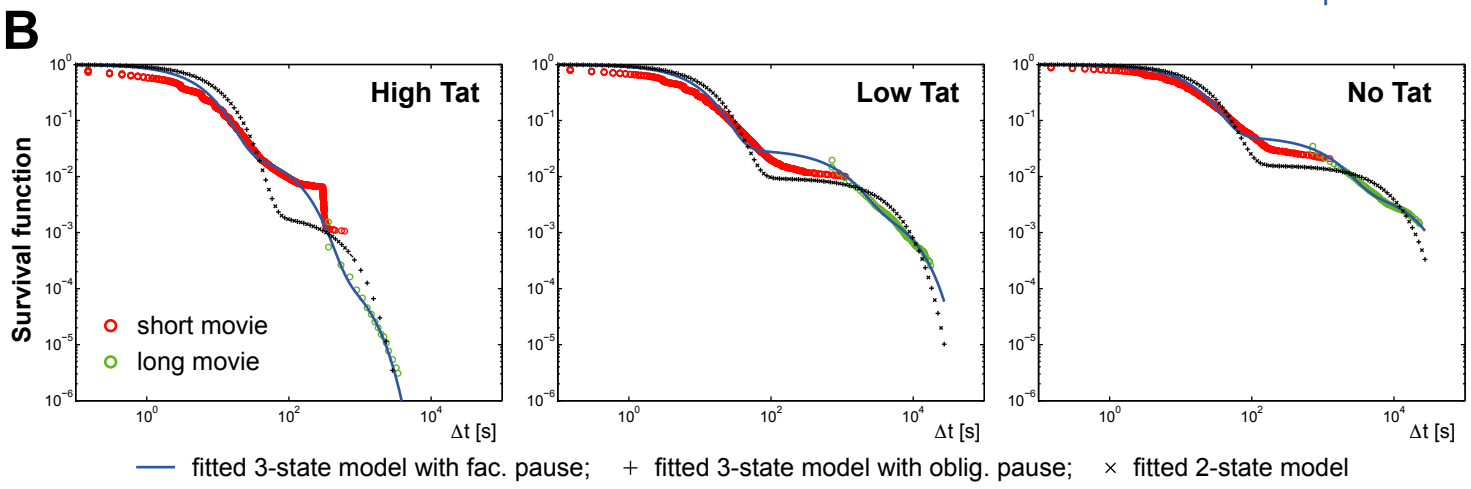
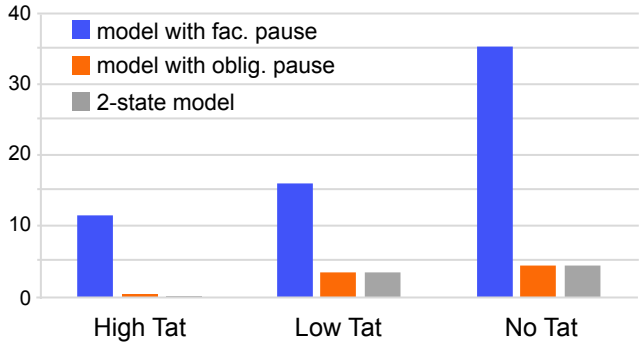


FIGURE 5







D

Predicted pausing time

	Fac. pause	Oblig. pause
High Tat	1-15 min	< 10 seconds
Low Tat	14 min - 2h	< 10 seconds
Not Tat	18 min - 3h	< 10 seconds

Predicted pausing frequencies

High Tat	0.03 - 2.5%	100%
Low Tat	0.3 - 2.7 %	100 %
Not Tat	0.5 - 5.0 %	100 %

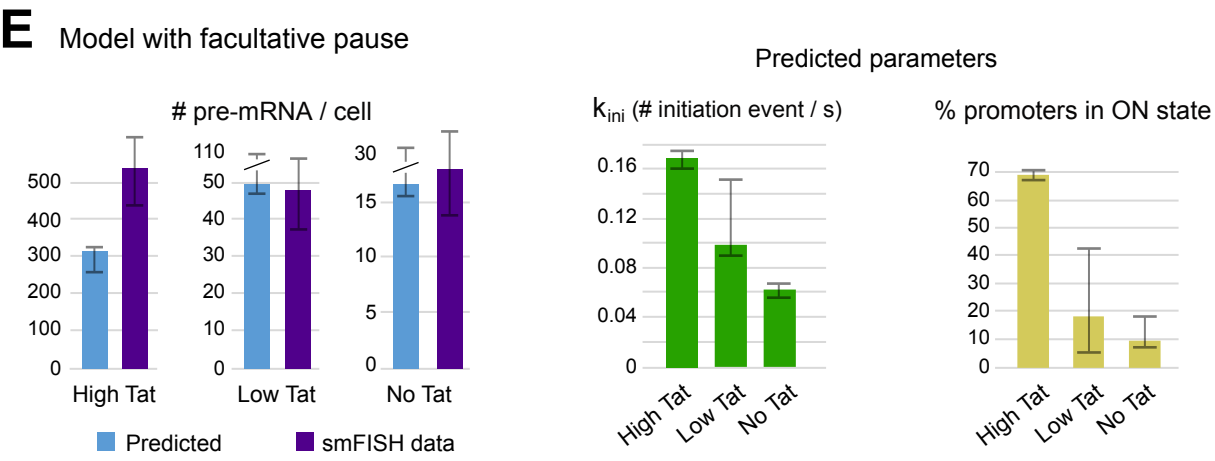
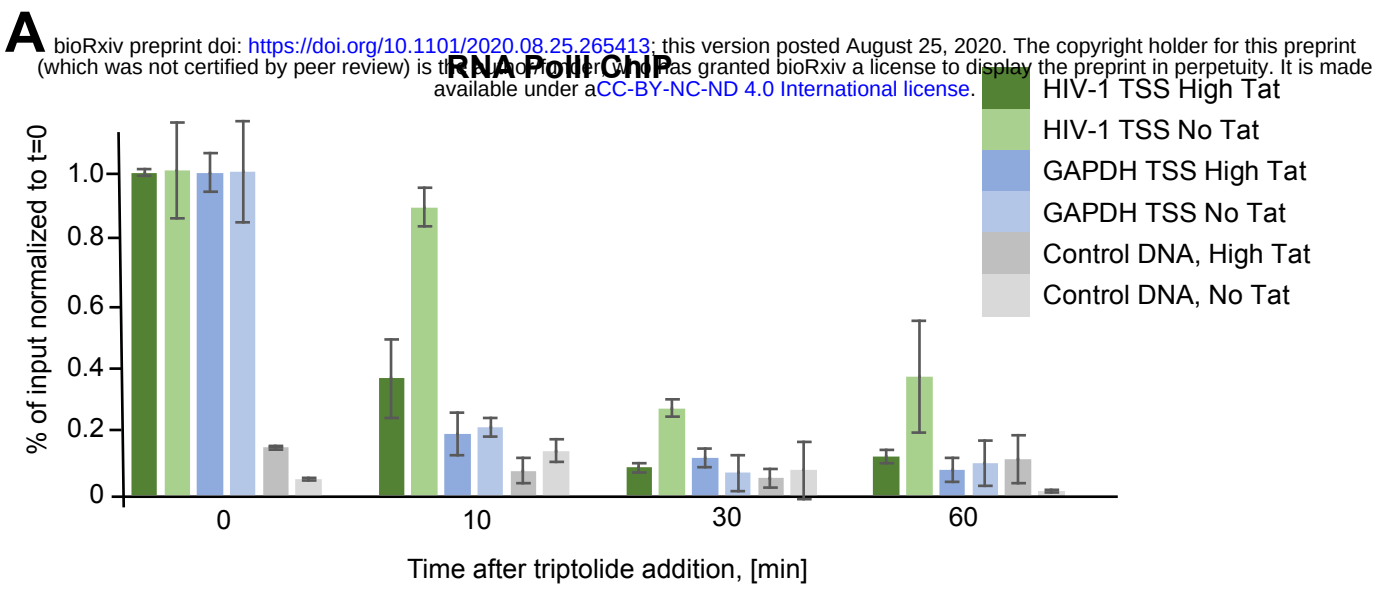
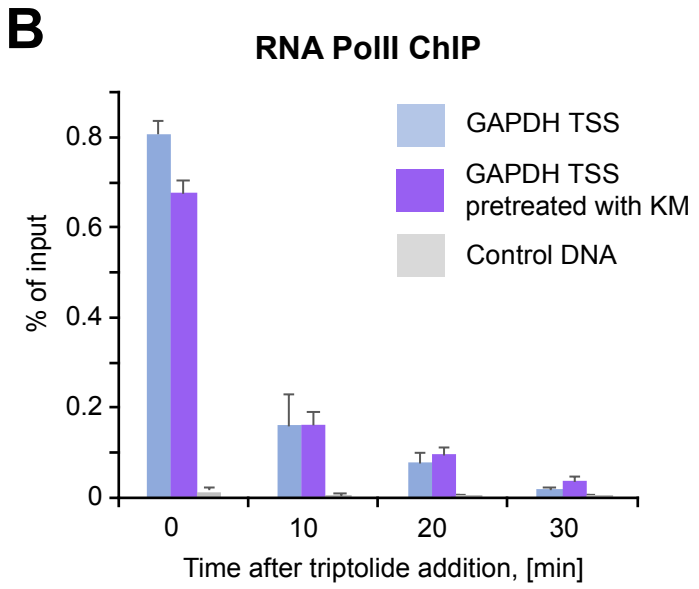


FIGURE 6





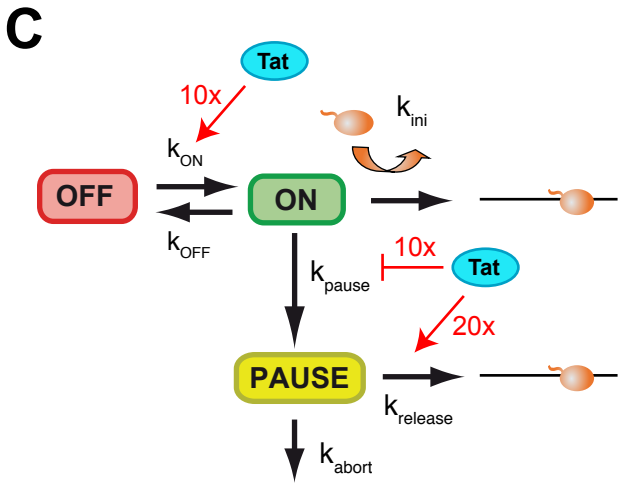
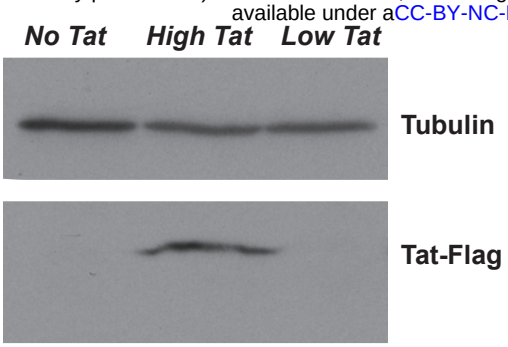
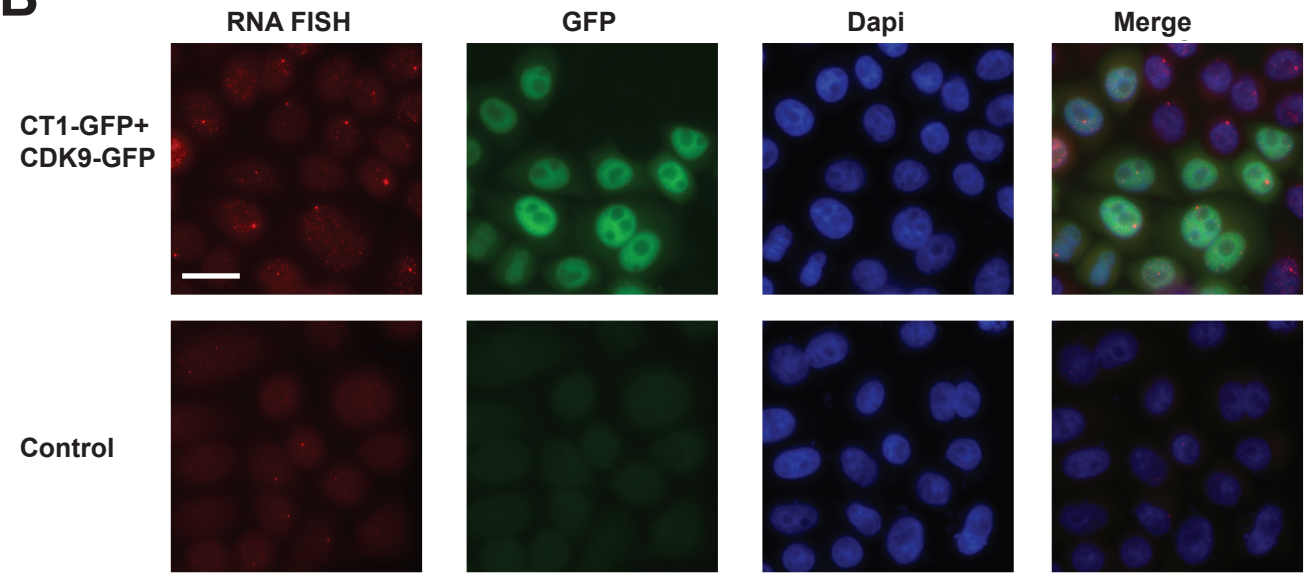


FIGURE 7

A



B



C

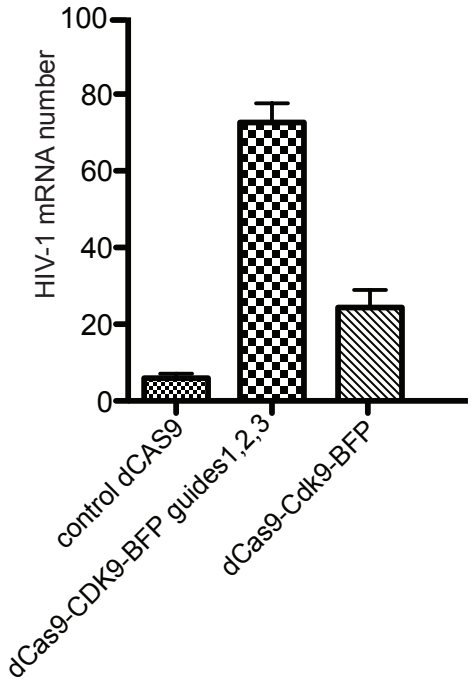
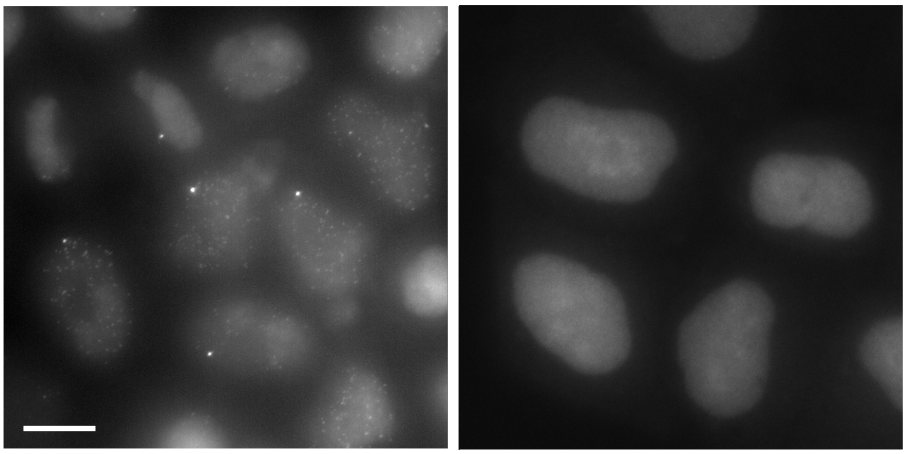


Figure S1

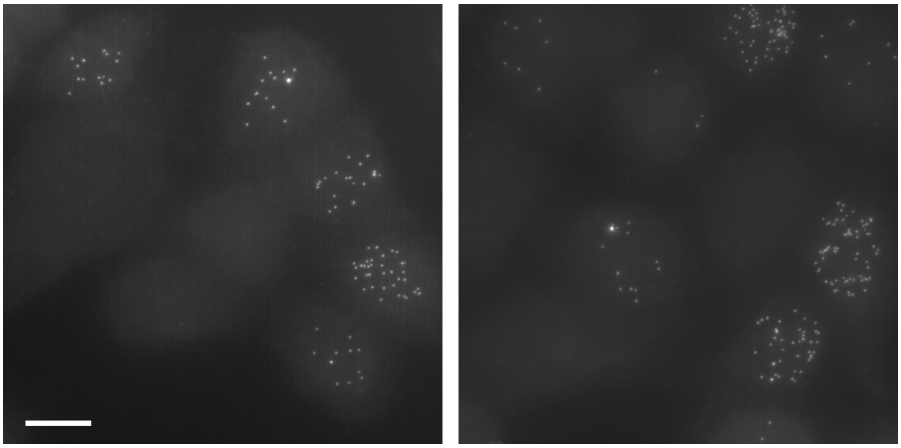
A



No Tat, -KM

No Tat,+KM, 4h

B

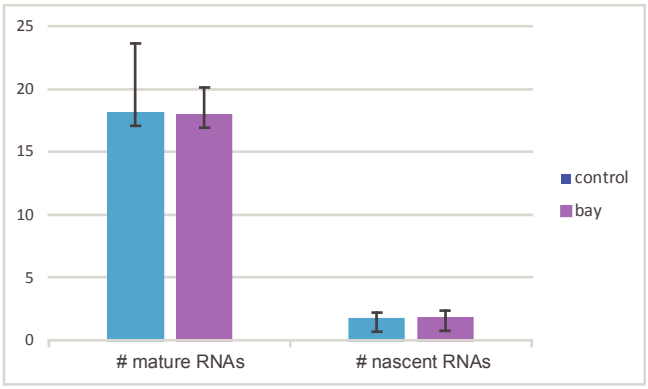


No Tat, control

No Tat, +Bay, 16h

NF-kB inhibition in no Tat cells

C



NF-kB inhibition in no Tat cells

Figure S2

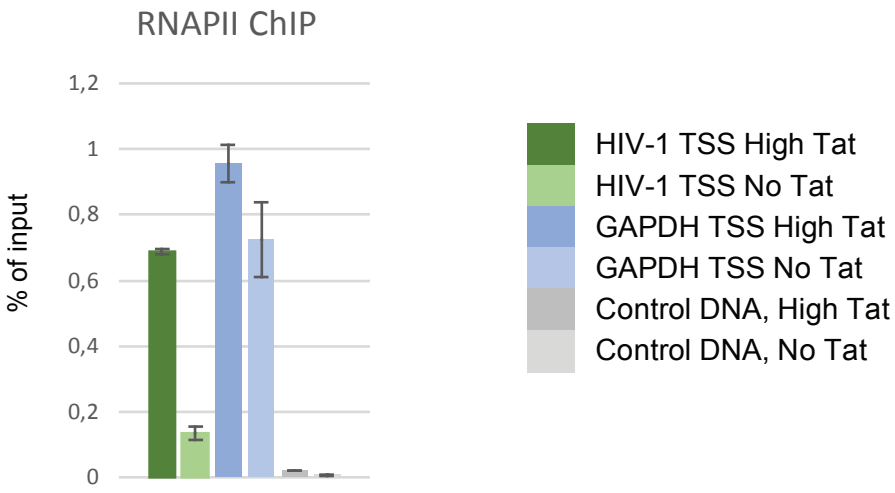


Figure S3

Hybrid symbolic/numeric method for reverse engineering of transcriptional bursting processes

Supplemental text to accompany: Stochastic pausing at latent HIV-1 promoters generates transcriptional bursting

Tantale K., Garcia-Oliver E., L'Hostis A., Yang Y., Robert MC., Gostan T., Basu M., Kozulic-Pihrer A., Andrau JC., Muller F., Basyuk E., Radulescu O., Bertrand E.

August 6, 2020

1 Introduction

1.1 Summary of the method

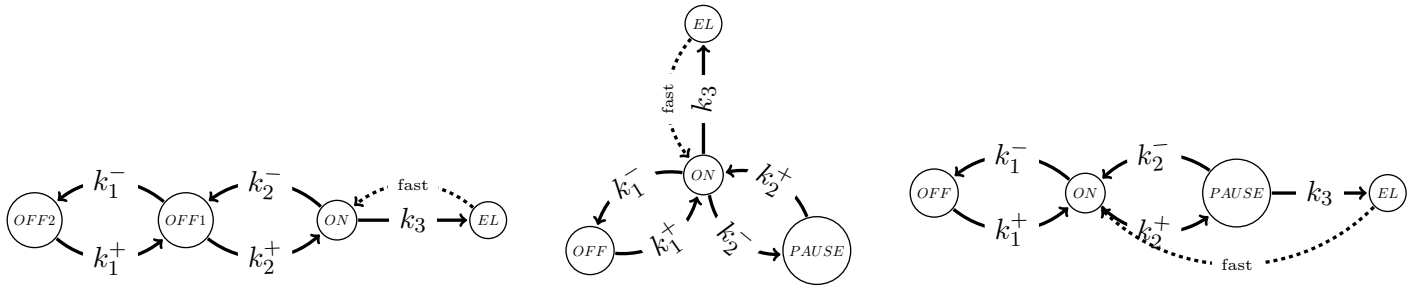
We use machine learning to derive characteristics of single cell transcription activity from MS2 data. The output of the machine learning procedure is threefold. Using a **deconvolution method** and high resolution movies, we generate a **time map of transcription events** indicating, for each cell, the moments when different RNAP molecules start producing mRNA. This direct readout of transcriptional events in a cell population, represents a unique feature of our method, not available in other methods that fit directly a particular transcription model to the MS2 data such as methods based on the autocorrelation function [2, 5, 4], or maximal likelihood estimate [1] or on Bayesian inference [7, 6]. The map can be used for direct characterisation of transcription features, such as polymerase convoys and various statistics of inter-event times. A second output of the approach is a **multiscale cumulative distribution function** of the waiting time separating succesive transcription events (or the complementary function, called survival function). We provide both non-parametric Kaplan-Meyer and parametric multi-exponential estimates of the multiscale distribution function. This distribution, obtained by combining short, high-resolution and long, low-resolution movies, covers timescales from second to 10 hours. The dynamical range of our method supersedes those of other extant methods that are based on much smaller

sampling rates and/or much shorter movie lengths. The waiting time distribution is model-free, but can be further used to identify various models of transcription dynamics. The third output of our method is the **model parameter identification**, simultaneously for several models that fit the data equally well. Although we focus on discrete transcription models based on Markovian transitions among hidden promoter states (for different number of states and for a rich collection of transition graph topologies), our method can be extended to the identification of more general models, including continuous or hybrid ones. Contrary to other methods that need separate fitting procedures for different models, in our method a single parametric fit of the multiscale waiting time distribution function is enough for identifying simultaneously a large collection of models that are all compatible with data and perform equally well. Another novelty with respect to other model fitting methods is the use of exact symbolic solutions, relating the parameters of the multiscale distributions to kinetic parameters of the model. For several models there is one-to-one relation between parameters of the distribution and kinetic parameters of the model. In this situation, the model kinetic parameters can be obtained analytically from the parameters of the multiscale distribution. Our method also leads to uncertainty estimates of the model parameters, based on optimal and close-to-optimal parametric fits of the multiscale distributions. As a matter of fact, models that fit equally well, can differ in their parametric uncertainties. Therefore, parametric uncertainty can be used as a model selection criterion that favor sure and reject uncertain models. The symbolic part of our method also identifies situations when parametric uncertainty results from redundancy, more precisely when there are manifolds of parameters that lead all to exactly the same goodness of fit. This is typically the situation when the relation between parameters of the multiscale distribution and the model parameters is one to many. Model and/or parameter uncertainty can be ultimately lifted by direct measurements of one or several kinetic parameters by alternative methods.

1.2 Discrete Markovian models for transcription dynamics

A Markovian model of transcription dynamics includes stochastic transitions between several ON and OFF promoter states (Figure 1). Rather generally there is a ON state and several OFF states. The promoter transcribes only in the state ON when it can trigger several departures of RNAP molecules along DNA. The departure of one RNAP is when the model reaches the state EL. It is considered that immediately after departure the operator site becomes free (the transition from EL to ON is instantenous). The transitions define a continuous time Markov chain characterized by a set of positive parameters k_{ij} representing the transition probability per unit time (or equivalently the inverse mean transition time) from state i to state j. Given the number of states N,

the structure of the Markov chain is defined by the directed graph $G = \{(i,j)|1 \leq i \leq N, 1 \leq j \leq N, k_{i,j} \neq 0\}$; several possible structures with $N = 3$ are shown in Figure 1. We show here how the parameters k_{ij} of a model can be adjusted to reproduce the transcriptional bursting and RNA synthesis observed in the live cells experiments. The parameter estimates are performed simultaneously for several possible model structures.



a) M1 (no pause, two non-productive states)

b) M2 (eventual pause, systematic abortion)

c) M3 (systematic pause, eventual abortion)

Figure 1: Three state, three exponential models of transcription dynamics. Transcription dynamics is represented as transitions between two OFF and one ON promoter states. ON represents the productive state. EL represents the elongation state which immediately liberates the promoter (from EL there is always fast return to ON). ON may not lead systematically to EL, for instance if there is transcription pausing. In these models, the pausing state is represented as one of the OFF state. If the pause leads systematically to transcription abortion the pause state leads to ON (model M2); otherwise it leads both to ON and to EL, with different probabilities (model M3). In model M1 the inactive state OFF1 can lead to another inactive state OFF2. The represented topologies differ by the connections between different states. The constants k_i are inverses of transition times, as such a) M1: k_3 is the initiation rate; b) M2: k_3 is the initiation rate, k_2^+ is the abortion rate, k_2^- is the pause enter rate; c) M3: k_3 is the pause exit rate and k_2^- is the abortion rate. All the represented models have 3 states (excepting the final elongation state), 5 kinetic parameters and their stochastic transcription activity can be described by a three exponential survival function.

1.3 The machine learning procedure

This procedure was initially designed for MS2 data obtained from human cell cultures but it has also been applied to in vivo study of Drosophila embryo development [3].

The machine learning procedure has several steps:

a) The first step is the **numerical deconvolution of the signal and Pol II Positioning**. The signal from each cell is a convolution between the contribution of a single polymerase and the point process (set of time points) describing all the start transcription events. For each cell, we reconstruct the start events by a least square optimization method performed by a genetic algorithm. We prefer optimization to Fourier transform based deconvolution in order to avoid the Gibbs phenomenon (the signal produced by a single polymerase is discontinuous).

b) The second step is the **non-parametric estimate of the survival function**. The data resulting at a) (positions of transcription start events) is used to estimate the survival function which is the complementary cumulative distribution function of the inter-event (waiting) times. Further complexity is brought in at this step by the utilization of two types of movies with short and long time resolution. Only short movies data undergoes deconvolution, the long movies are used to obtain long waiting times directly. This results into two distribution functions that are joined together (by affine transformations corresponding to the law of total probability) to cover many decades of timescales (second to ten hours). In certain applications the hours scale is not reachable because of biological constraints. For instance, in developmental biology, the studied developmental stage may be too short. In this case, we use only short movies and the joining step is not needed.

c) The third step is the **multi-exponential regression of the survival function**, performed by gradient optimization with random starting guesses, uniformly distributed in logarithmic scale (this choice is dictated by the multiscale nature of the signal). Usually, two or three exponentials (i.e. two or three time scales) are enough to describe our data. Chosing more than three exponentials is justified when this improves the fit without increasing parameter uncertainty. Conversely, choosing less exponentials is justified if this does not diminish the fit while decreasing parameter uncertainty. The first three steps of our procedure are model-free because they make no assumption about the dynamics of the transcription regulation.

d) The last step of the procedure is the **symbolic reverse engineering of transcription models** from the survival function. We consider that the transcription machinery has several discrete states among which only one is productive. Then, the waiting time between successive transcription start events is the first return time to the productive state. The distribution of this waiting time satisfies a system of ODEs whose solution can be expressed as a sum of exponentials. The inverse problem consists in computing model's kinetic parameters from the parameters of the multi-exponential regression. We have developed

a symbolic solution to perform this step. Our symbolic solution also tackles the ill-posed character of the inverse problem. Indeed, although the same distribution function can be produced by several models with different structures, the significance and the value of each parameter are different in different models. Moreover, we know precisely how to pass from one model to another by changing the parameter values. It is therefore enough to perform a direct independent experimental measurement of a single parameter in order to discriminate between different models. In the case of redundant parameters (parameters not influencing independently the observed distribution function) and parameter uncertainty, some parameters may remain independent and can be used for model discrimination.

2 Numerical deconvolution of short movies

2.1 Description of the problem

The experimental data obtained from short movies is shown in the Figure 2 for the HIV-1 promoter. The signal intensity from the mRNA MS2 reporter is represented as a function of time for each active transcription site. We are interested in reconstructing from this signal the sequence of waiting times between successive transcription start events (see Figure 3), for each transcription site.

Transcription events can not be straightforwardly detected from local features of the intensity signal because at a given time and for the same transcription site, more than one polymerase transcribe simultaneously. Furthermore, the signal from one polymerase does not appear immediately after initiation (see below).

One should thus consider that experimental data is a a convolution between the sequence of start events $\{t_i, 1 \leq i \leq N_{pol}\}$ and the signal $h(t)$ from a polymerase molecule:

$$S(t) = \sum_{i=1}^{N_{pol}} h(t - t_i), \tag{1}$$

where N_{pol} is the number of polymerases contributing to the signal. N_{pol} is not known and will be determined by the optimization procedure (see below). The parameters t_i are the initial polymerase positions on the DNA, indicating the transcription start events.

The polymerase signal $h(t)$ can be described as follows (see Figure 4):

i) Transcription begins when the RNA polymerase II leaves the promoter. However, no signal will be generated yet.

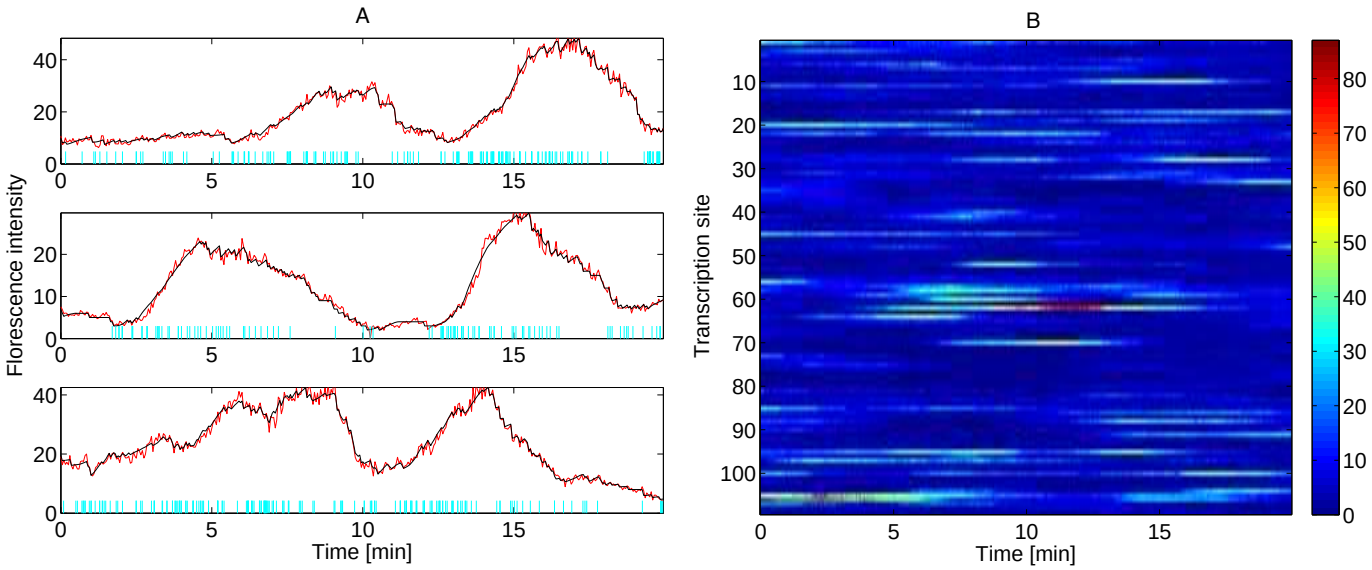


Figure 2: Short movie data for a HIV-1 promoter (no tat condition). A) fluorescence intensity vs. time for several transcription sites (red line); the reconstructed polymerase positions and signal are indicated as vertical cyan bars and black line, respectively. B) colormap of intensity for all transcription sites in a short movie.

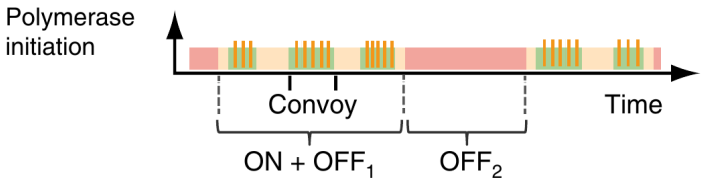


Figure 3: Dynamics of the promoter changing states. Start events are represented as red bars. During a ON period several polymerases start, forming a convoy. For this signal, two types of non-productive states, short (OFF1) and long (OFF2) can be obseved.

ii) The fluorescence signal is generated as soon as the polymerase reaches the MS2 sequence. During the transcription of the MS2 sequence the signal can be represented as a linear ramp-up.

iii) The signal will stay constant from the end of the MS2 sequence until when the polymerase leaves the transcription site, when the signal falls abruptly.

In order to compute the times corresponding to the three stages we use the length (expressed in base pairs) of the three sequences PRE, SEQ and POST (before MS2, MS2 and post MS2). These lengths depend on the MS2 construction (the values in our HIV-1 experiments are PRE=700bp , SEQ=2900bp, POST=1600bp). The

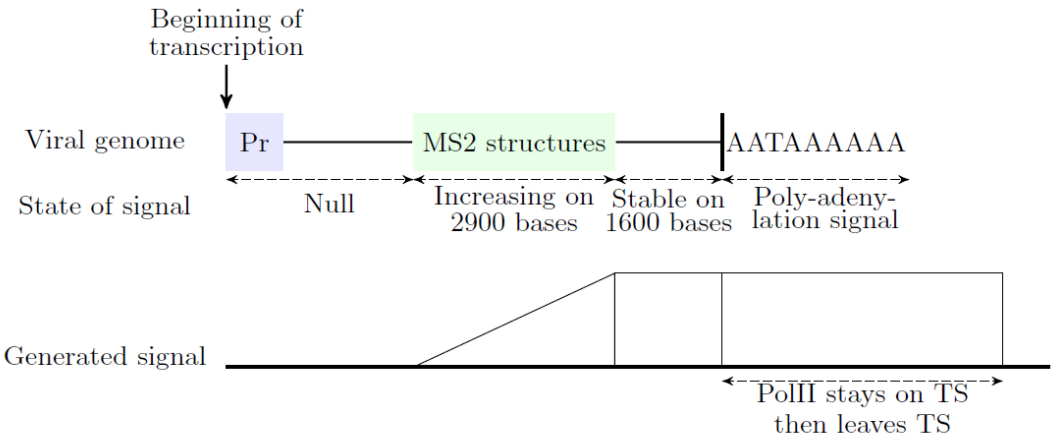


Figure 4: Representation of the signal from one polymerase for the HIV1-promoter. The parameters are indicative and can change for other applications.

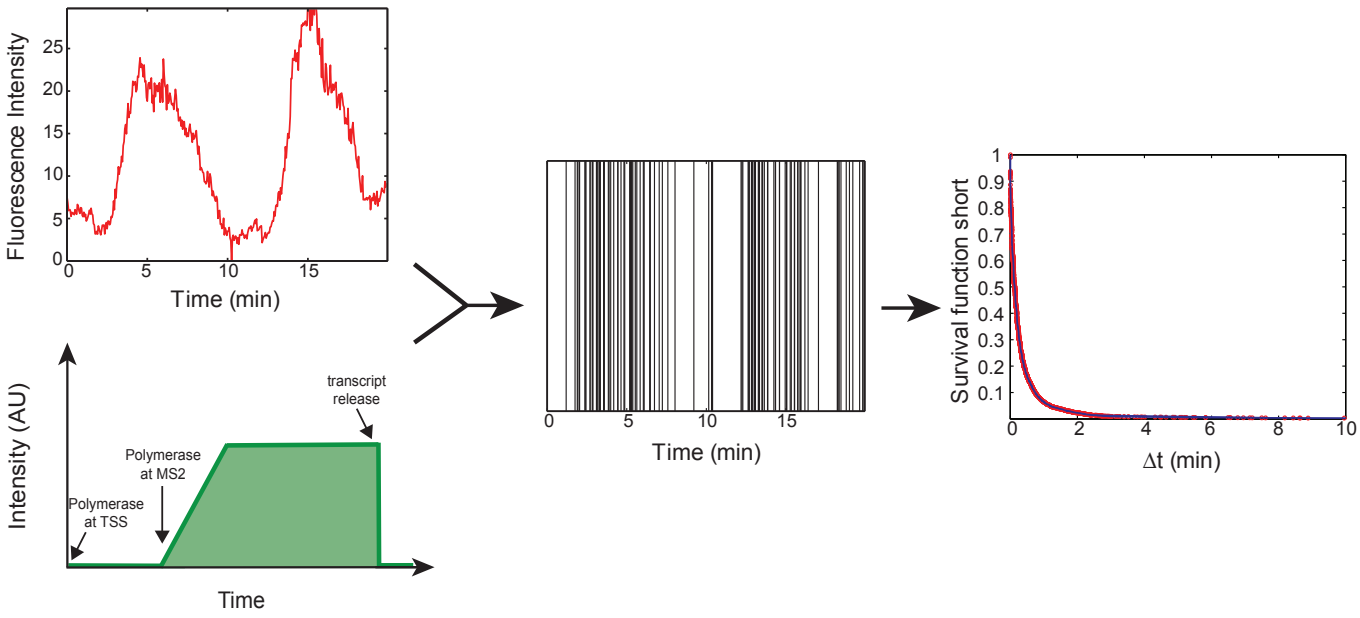


Figure 5: Distribution of transcription initiation events can be reconstructed by deconvolution.

sequence lengths are divided by the polymerase speed V_{pol} to be transformed into times. For our HIV-1 promoter we have used $V_{pol} = 67bp/s$ (see main text). An extra time $P_{poly} = 100s$ is added to POST, corresponding to the polyadenylation signal (during this time the polymerase has finished transcription and waits on the transcription site).

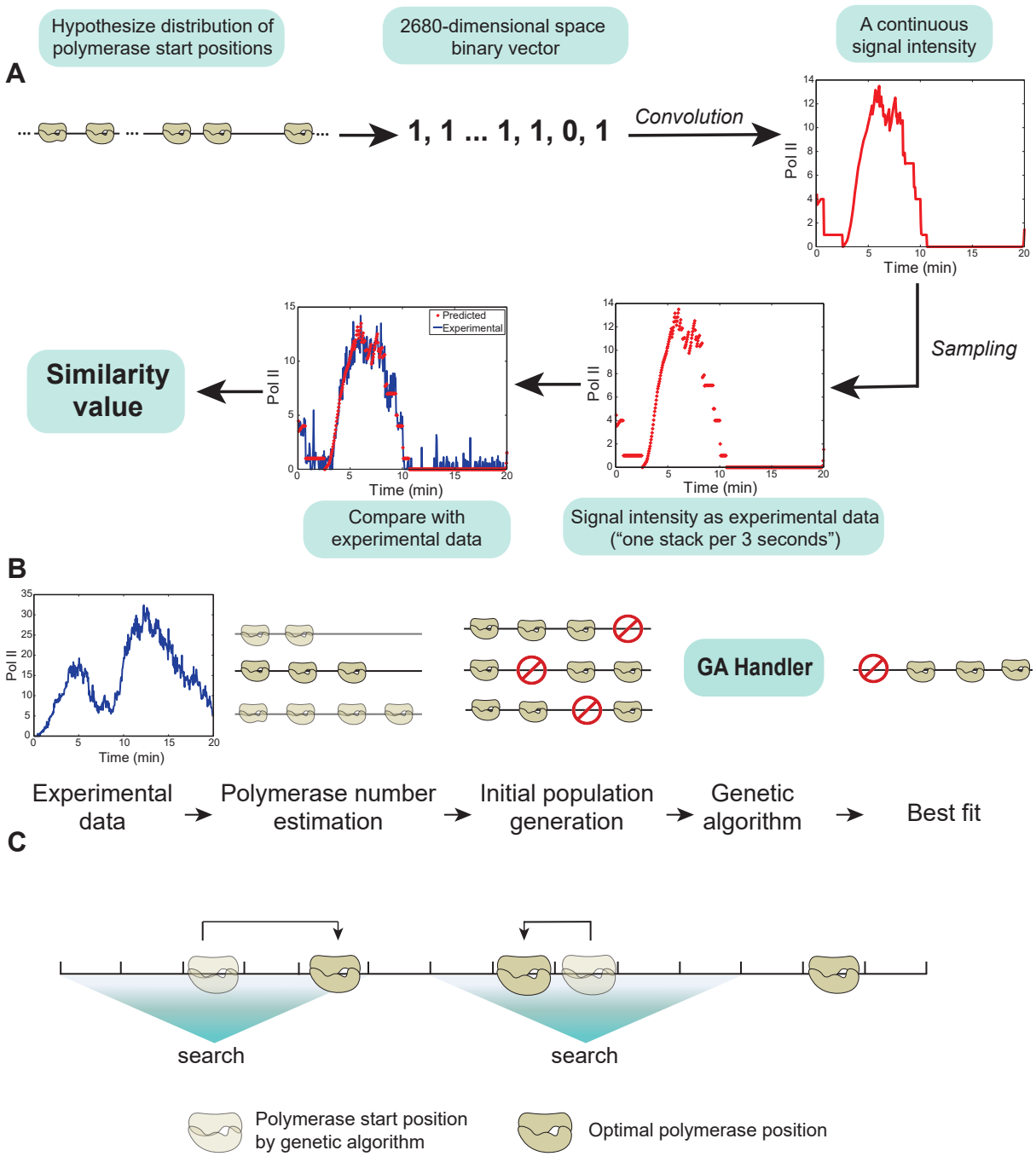


Figure 6: A. Flowchart of the numerical deconvolution method. B. Genetic algorithm step. C. Local optimization step.

After signal calibration the unit of fluorescence represents the amplitude of the signal from one polymerase.

Using this model we want to reconstruct the sequence of initiation events by deconvolution (see Figure 5). More precisely, we will determine N_{pol} and $t_i, 1 \leq i \leq N_{pol}$ that minimize the following objective function:

$$\mathcal{O}_1 = \sum_{k=1}^{N_{exp}} [S(k\delta) - S_{exp}(k\delta)]^2, \tag{2}$$

where δ is the time step (inverse frame rate), N_{exp} is the number of frames, S is described by (1) and S_{exp} is the experimental signal. For a short movie, the frame rate is $1/3s^{-1}$, thus $\delta = 3s$ and $N_{exp} = 400$ for a movie length of $P_{max} = 20$min.

2.2 Discretization of the optimization problem

It is useful to use a dual representation of the polymerase positions t_i in terms of seconds and base pairs on the DNA sequence. Although the polymerase positions are in principle continuous variables, for computation reasons we discretize them. In this dual representation, it is natural to consider that possible polymerase positions are multiples of the minimum distance d_{min} between two polymerases ($d_{min} = 30$bp in our program). The precise value of d_{min} is not needed. Generally, d_{min} should be chosen as small as possible to guarantee precision of the polymerase positions. It should be smaller than the real minimum distance between polymerases and larger than a value dictated by the computation costs (the computation costs increase when d_{min} decreases).

Using this discretization, polymerase positions are coded as a binary vector. Every possible polymerase position will have either 1 or 0 value, which represents if there is a polymerase in the current position or not. Considering that the polymerase speed is constant, the polymerase positions are all we need to determine the signal. For a movie of length $P_{max} = 20$min and for $V_{poly} = 67$bp/s, the polymerase positions are represented as a binary vector of length $N = P_{max}V_{poly}/d_{min} = 2680$. Considering discretization steps larger than d_{min} is also possible. In this case binary vectors are shorter and computation is faster, however the precision may be reduced.

2.3 Solve the deconvolution problem by a genetic algorithm followed by local optimization

Every binary vector of dimension N represent the polymerase start positions and determine a value of the objective function (2). The opposite of the objective function is the *fitness*. The deconvolution problem represents finding the minimum of this objective function (maximum of fitness) in the N dimensional binary space (Fig-

ure 6). In order to solve this hard combinatorial problem, we apply first a global optimization genetic algorithm (GA).

As shown in Figure 6 B, GA follows three steps: estimating the amount of polymerases, generating an initial population and applying genetic algorithm. We estimate the number of polymerases N_{pol} from the signal integral intensity, as the ratio of integral intensities of the experimental signal and of the single polymerase signal. The resulting amount is not an accurate number, and it is a rough estimation which can be used to accelerate next steps. Then we prepare an initial population according to the estimation of polymerase amount. Starting with a vector with N '0's, we randomly pick N_{pol} positions and change them into '1's. After the preparation of initial population, we use the genetic algorithm implemented in the GA solver provided by Matlab global optimization toolbox. Mutation, crossover and selection are processed by the MATLAB built-in function ga (MATLAB, version (R2013b), Natick, Massachusetts: The MathWorks Inc.). At each step, the genetic algorithm solver selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution.

In order to verify this method, we implemented a test using an artificial experimental signal. We deconvolved the artificial signal, for which we know exactly the polymerase start positions. The simulation of genetic algorithm, as in the example of Figure 7, shows that the genetic algorithm can approximately reconstruct the signal. However, the global minimum is not precisely reached and the polymerase start positions of simulation are not exactly the same as the artificial ones (Figure 7).

There are various reasons why polymerases were not exactly placed into right positions as follows: the limitation of the maximum number of iterations, the limitation of population size, the initial error in the estimated number of polymerases, the noise generated by the algorithm, etc.. Although GA can not give a precise result (or it is time consuming to get a precise result), it provides a solution not far from the optimal result.

With this in mind, we use a local exhaustive search to accomplish local optimization. The idea is to "move" a polymerase left or right relative to the GA found position to see if this improves the fitness function (Figure 6 C). For every polymerase we find the best position which has the highest fitness value and we update the best positions for all of them. The local optimization result is shown in Figure 7. By this method, practically all the polymerases were arranged into the correct positions. The local optimisation method has limitations, for instance it does not allow correction of the total number of polymerases; we suppose that this number has already been found by the GA.

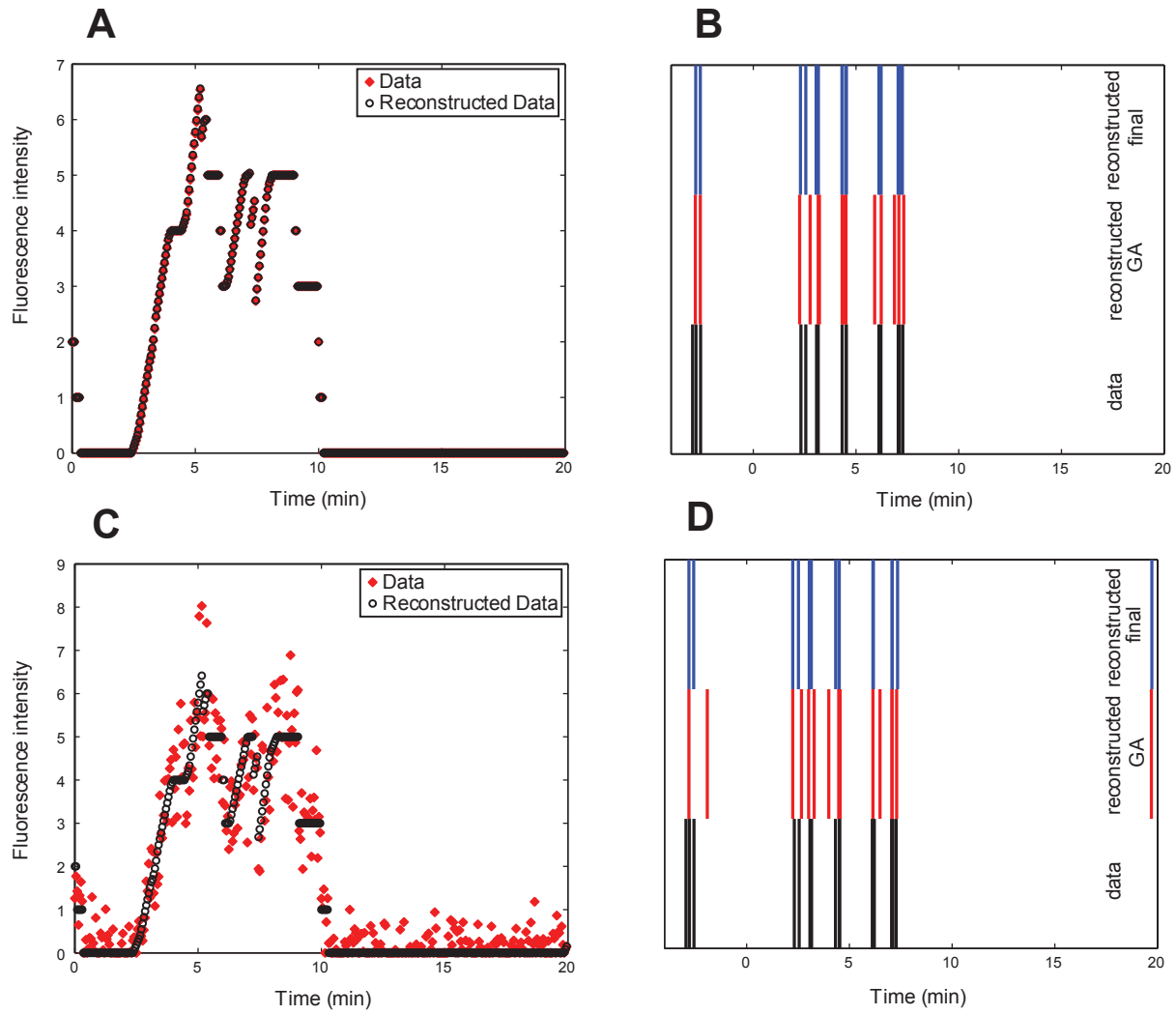


Figure 7: Result of the deconvolution step on artifical data. A) Signal generated artificially (no noise). B) Original polymerase positions compared to positions resulting from the genetic algorithm step and to the final positions corrected by local optimisation (no noise). C) Signal generated artificially (noise added according to the procedure described in Section 8). D) Original polymerase positions compared to positions resulting from the genetic algorithm step and to the final positions corrected by local optimisation (noise added).

3 Multi-exponential regression of the distribution function

From the numerical deconvolution step, we obtain the series of intial polymerase positions, for each active transcription site detected in the short movies.

From each transcription site we compute waiting times defined as time interval between successive positions.

When the position is the last one in the movie, the waiting time is defined as the distance to the end of the movie. Considering that all transcription sites are statistically equivalent, we gather the waiting times from all sites that are active in the same short movie.

Long movies also provide waiting times by a different method, without numerical deconvolution (see below).

We consider that the transcription events form a renewal process with independent, identically distributed waiting times Δ. This property is valid at stationarity, but not only. For instance, in Markovian models, the property is true if after every transcription event the system allways returns to the same state. Given that non-Markovian models can be made Markovian by adding hidden states, we believe that the property is quite general. All the models from Figure 1 satisfy this property, because from the EL state one can only go to the ON state.

We want to estimate the *complementary cumulative distribution function* (also called *survival function* in survival analysis) of the waiting times defined as:

$$S(t) = \mathbb{P}[\Delta > t]\,. \tag{3}$$

3.1 Waiting times from short movies: Δ_s

Outliers handling

Several observed transcription sites had abnormal behaviour (too many or too few events). The decision was made to take them off the data set as follows

1. Compute the amount of events of transcription that happened during the movie (Ev).

2. Compute the 1st and the 3rd quartile (respectively Q1 and Q3) of the distribution of Ev .

3. Only consider the transcription sites where

$$Q1 - 2.5(Q3 - Q1) < Ev < Q3 + 2.5(Q3 - Q1)$$

Affine transformation, parameter p_s

The Δ_s are the waiting times deduced from the short movies, therefore they all satisfy the condition $\Delta_s < P_{max}$, where P_{max} is the movie length. Therefore, this data does not reconstruct the full survival function, but the conditional survival function $S_{<P_{max}}(t) = \mathbb{P}[\Delta > t | \Delta < P_{max}]$.

In order to compute the relation between $S(t)$ and $S_{<P_{max}}(t)$ we use the total probability theorem:

$$\mathbb{P}[\Delta > t] = \mathbb{P}[\Delta > t | \Delta < P_{max}] \left(1 - \mathbb{P}[\Delta > P_{max}]\right) + \mathbb{P}[\Delta > t | \Delta > P_{max}] \, \mathbb{P}[\Delta > P_{max}] \, . \tag{4}$$

Let us note that for $t < P_{max}$, one has $\mathbb{P}[\Delta > t | \Delta > P_{max}] = 1$. Hence, from (4) it follows that

$$S_s(t) = (1 - p_s) S_{<P_{max}}(t) + p_s, \text{for } t < P_{max}, \tag{5}$$

where $p_s = \mathbb{P}[\Delta > P_{max}]$ is the probability that the waiting time is longer than the length of the short movie.

In other words, for short movies, the survival function is obtained from the conditional survival function by an affine transformation.

3.2 Waiting times from long movies: Δ_l

Active and inactive periods, threshold parameter

The frame rate for long movies is $1/3min^{-1}$ and the typical length is $9h$. Let us notice that the deconvolution procedure is not possible for long movies, because the number of polymerases is too large. Therefore, long waiting times are obtained directly from the signal. For long movies, there is no need to calibrate the fluorescence intensity, nor to deconvolve the signal. An intensity threshold is defined and a given transcription site is considered active in a given frame if its intensity is larger than the threshold, inactive if not, see Figure 8.

Outliers handling

We define the fraction of inactivity (FI) as the ratio of cumulative total inactivity time to the total cumulative time in the long movie and for all the transcription sites.

Some transcription sites in long movies data set also show unusual behaviours being active (FI=0) or inactive (FI=1) during the entire movie. We exclude these outliers as we did it for the short movies, but based on the fraction of inactivity for each transcription site.

We will only consider the transcription sites from the long movies where

$$Q1 - 2.5(Q3 - Q1) < FI < Q3 + 2.5(Q3 - Q1)$$

Corrected waiting times, parameter Δ_0

In the long movies the waiting times Δ_l between successive transcription initiations correspond roughly to the inactive periods Δ_I. As a matter of fact, these waiting times can be longer than the inactive periods by a time varying between 0 and 6min (because the signal needs about 3min to vanish and starts about 3min before

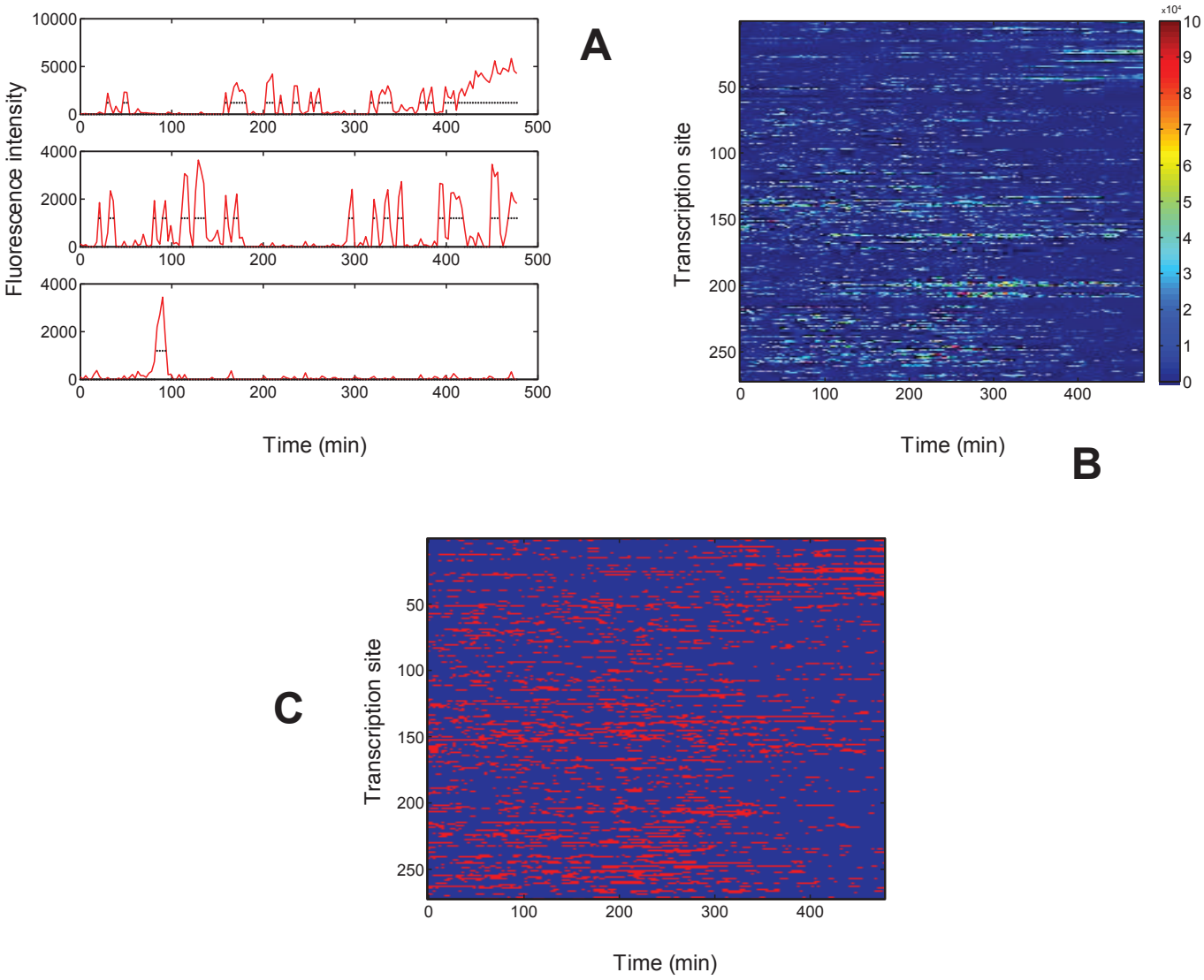


Figure 8: Long movie data for a HIV-1 promoter (no tat condition). A) intensity of fluorescence vs. time for several transcription sites; the real valued intensity are transformed into a binary signal (dots) by thresholding (here the threshold value is 1200). B) intensities for all transcription sites. C) Binary valued intensities for all transcription sites.

it is detected). This unknown time is a parameter of the method and its values are discretized to $\Delta_0 = 0, 3, 6$. All waiting times are computed as $\Delta_l = \Delta_I + \Delta_0$.

Affine transformation, parameter p_l

The signal from a single polymerase lasts roughly 3min (see Figure 4). In this case, waiting times shorter

than P_{min} where P_{min} is roughly 3min can not be observed. The observed conditional survival function is now $S_{>P_{min}}(t) = \mathbb{P}[\Delta > t | \Delta > P_{min}]$.

We can note that for $t > P_{min}$, one has $\mathbb{P}[\Delta < t | \Delta < P_{min}] = 1$.

Once again, from the total probability theorem (4), it follows:

$$S_l(t) = p_l S_{>P_{min}}(t), \text{for } t > P_{min}, \tag{6}$$

where $p_l = \mathbb{P}[\Delta > P_{min}]$ is the probability that the waiting time is longer than P_{min}, S_l is the survival function for long movies.

Estimate of parameter p_l

The probability p_l is estimated by combining information extracted from the long and the short movies. p_l is precisely the probability that waiting times are observed as inactive periods in the long movie. Let $N_{inactive}$ and N_{active} be the number of waiting times observed as inactive periods, and hidden within active periods of the long movie, respectively. $N_{inactive}$ can be determined directly from the long movie, it represents the number of inactive periods. N_{active} is obtained as the ratio $P_{active}/\mathbb{E}[\Delta | \Delta < P_{min}]$ where P_{active} is the cumulative time of all active periods in the long movie and $\mathbb{E}[\Delta | \Delta < P_{min}]$ is the conditional expectancy of the waiting time provided that this is smaller than P_{min} therefore undetectable by the long movie. By definition one has

$$\mathbb{E}[\Delta | \Delta < P_{min}] = \frac{\int_0^{P_{min}} u f(u)\, du}{\mathbb{P}[\Delta < P_{min}]},$$

where f is the probability density function of Δ. Taking the derivative of $S(t) = \mathbb{P}[\Delta > t] = \int_t^{\infty} f(u)\, du$ we get $f(t) = -S'(t)$. Using the integral by parts formula we find

$$\mathbb{E}[\Delta | \Delta < P_{min}] = \frac{-P_{min} S(P_{min}) + \int_0^{P_{min}} S(u)\, du}{1 - S(P_{min})}.$$

Summarizing, we find

$$p_l = \frac{N_{inactive}}{N_{inactive} + \frac{P_{active}(1 - S(P_{min}))}{-P_{min} S(P_{min}) + \int_0^{P_{min}} S(u)\, du}}. \tag{7}$$

Both $S(t)$ and the integral above are computed using the survival function of the short movie. $N_{inactive}$ and P_{active} are determined from the long movie data.

Estimate of parameter p_s

p_s is estimated by optimization. We look for the value of p_s that minimizes the square distance between the solutions (5) and (6) on the overlap interval $[P_{min}, P_{max}]$.

The survival functions after calculation of p_l, p_s and affine transformations are shown in Figure 9.

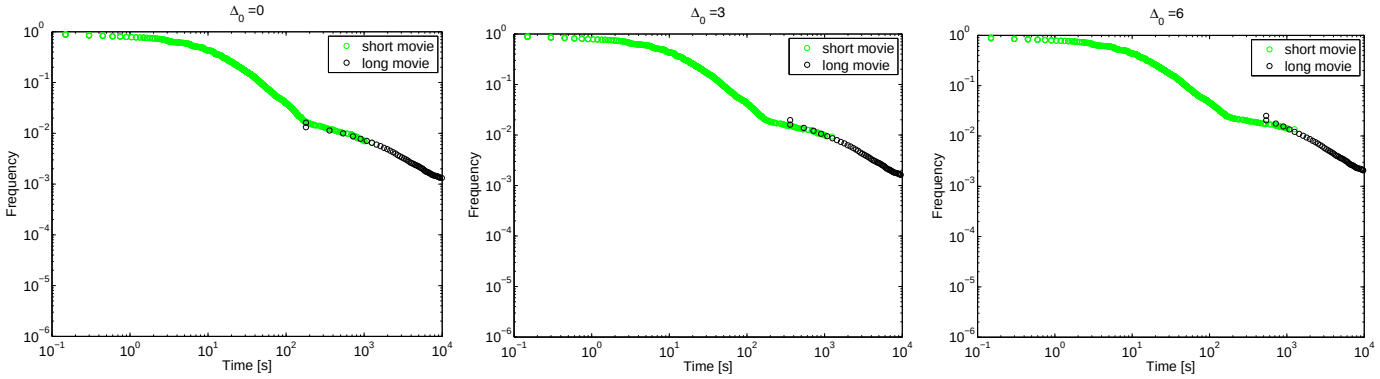


Figure 9: Survival functions of the waiting time after affine transformations for several values of the shift Δ_0 (HIV promoter, no tat condition, see text).

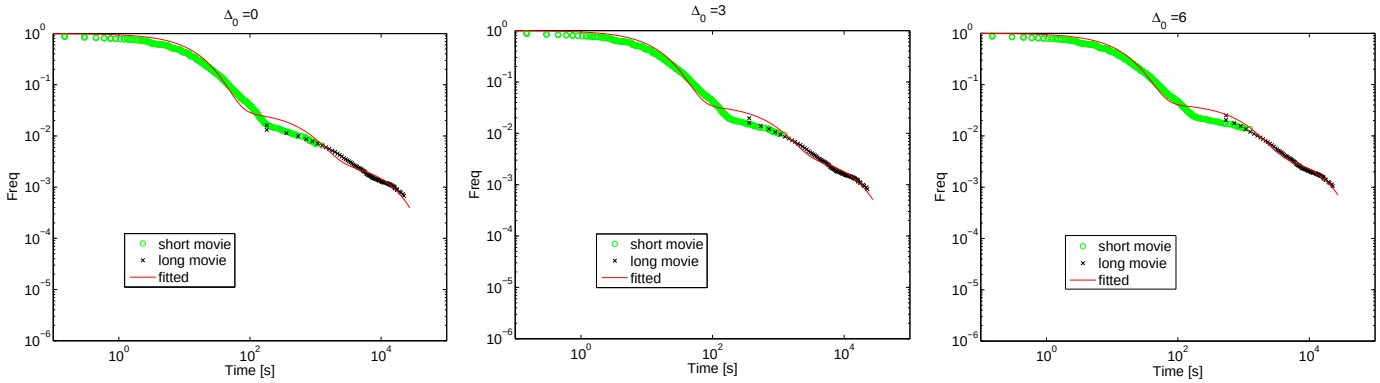


Figure 10: Multi-exponential regression ($N = 3$) for several values of the shift Δ_0 (HIV promoter, no tat condition, see text).

3.3 Multiexponential regression

The previously determined survival function function (5), (6) is modelled by a multiexponential function:

$$S(t) = A_1 \exp(\lambda_1 t) + A_2 \exp(\lambda_2 t) + \ldots + (1 - A_1 - A_2 - \ldots - A_{n-1}) \exp(\lambda_n t), \tag{8}$$

where $A_1, \ldots, A_{n-1}$, $\lambda_1, \ldots, \lambda_n$ are $2n - 1$ parameters.

Because $\lim_{t\to\infty} S(t) = 0$, these parameters must satisfy the constraints $\lambda_i < 0, 1 \leq i \leq n$. For practical reasons we can consider that all λ_i are distinct. Degenerate cases, when two or more λ_i are equal can be uniformly approximated by formula (8) with distinct λ_i (see the section 4.2). Up to relabelling we can consider that $|\lambda_1| > |\lambda_2| > \ldots |\lambda_n|$. Furthermore, because the complementary distribution function is always decreasing

we have

$$S'(t) < 0, \forall t \geq 0. \tag{9}$$

The condition (76) implies that $\sum_{i=1}^{n} A_i \lambda_i < 0$, where $A_n = 1 - \sum_{i=1}^{n} A_i$ (follows from $S'(0) < 0$) and that $A_n > 0$ (this follows from $\lim_{t\to\infty} S'(t) exp(-\lambda_n t) = A_n \lambda_n < 0$ and $\lambda_n < 0$). The hyperplanes $\sum_{i=1}^{n} A_i \lambda_i = 0$, $A_n = 0$ together with other manifolds delineate the domain of valid parameters A_i. This domain depends on the exponents λ_i as illustrated for $n = 3$ in Figure 11.

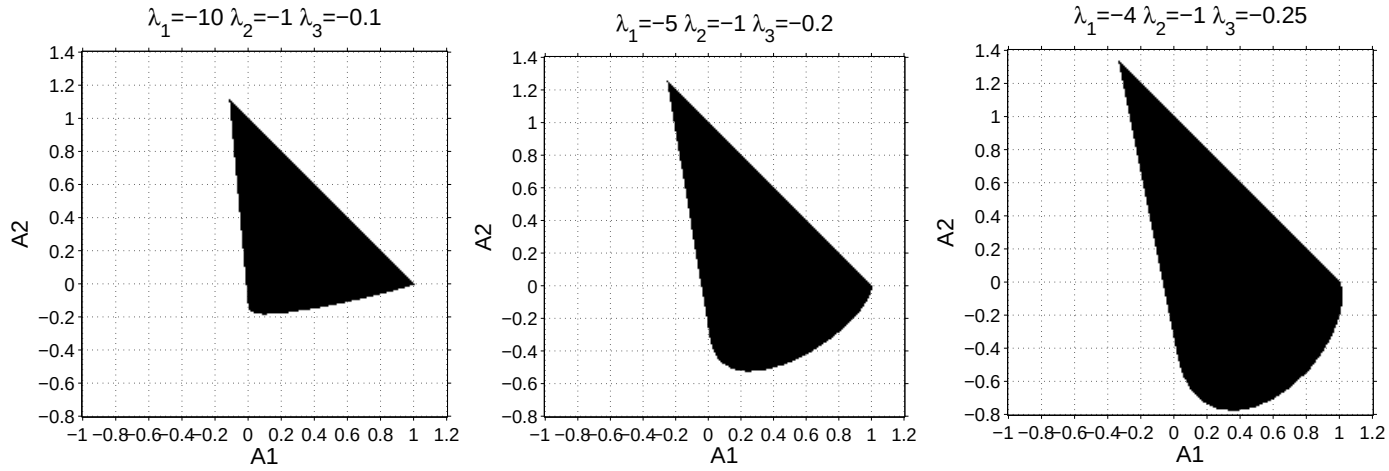


Figure 11: Permitted values of $A_i, 1 \leq i \leq n$ for $n = 3$ are represented in black for various λ_i. These parameters values are defined by the condition $S'(t) = \sum_{i=1}^{3} \lambda_i A_i \exp(\lambda_i t) < 0, \forall t \geq 0$, where $A_3 = 1 - A_1 - A_2$. The permitted values are limited at the top right by the line $A_1 + A_2 = 1$ and at the left by the line $\lambda_1 A_1 + \lambda_2 A_2 + \lambda_3(1 - A_1 - A_2) = 0$.

Like usually in machine learning, the choice of n is guided by a parcimony principle. One can start with $n = 2$ and progressively increase n until the goodness of fit stops improving (at equal goodness of fit, one favors the model with lowest complexity, lowest n and/or with lowest parameter uncertainty).

The objective function is defined as follows:

$$\begin{aligned}\mathcal{O}_2 = &\frac{\alpha}{n_s}\sum_{i=1}^{n_s}(S(t_i^s) - S_s(t_i^s))^2 + \frac{\alpha}{n_l}\sum_{i=1}^{n_l}(S(t_i^l) - S_l(t_i^l))^2 + \frac{1-\alpha}{n_s}\sum_{i=1}^{n_s}(\log(S(t_i^s)) - \log(S_s(t_i^s)))^2 + \\ &+ \frac{1-\alpha}{n_l}\sum_{i=1}^{n_l}(\log(S(t_i^l)) - \log(S_l(t_i^l)))^2,\end{aligned} \tag{10}$$

where $S(t)$ is defined by (8); S_s and S_l are computed by (5),(6), respectively; t_i^s, t_i^l are sampling times for short and long movies, respectively; α is a positive weight representing the relative importance of the linear scale

compared to the logarithmic scale in the representation of the survival function.

We minimize (10) by local optimization (Levenberg-Marquardt algorithm implemented in the Matlab function *lsqnonlin*) starting with N_p (in our program $N_p = 100$) random values of the regression parameters $A_1, \ldots, A_{n-1}$, $\lambda_1, \ldots, \lambda_n$. The initial parameters $A_1, \ldots, A_{n-1}$ are chosen uniformly distributed in the cube $[-M, M]^{n-1}$ (we used $M = 2$), whereas the initial parameters $\lambda_1, \ldots, \lambda_n$ are all negative and log-uniformly distributed in absolute value. More precisely, $\log(|\lambda_i|)$ are uniform in a cube $(l_1, l_2, \ldots, l_n) + [-K, K]^n$, where $l_1 < l_2 < \ldots < l_n$.

The optimization is repeated for all values of Δ_0 and each time repeated N_p times with different initial parameters (Figure8). We keep the lowest value $\mathcal{O}_2^{min}$ of (10) as well as sub-optimal solutions with $\mathcal{O}_2 < 1.5\mathcal{O}_2^{min}$. The suboptimal parameters are utilized to estimate the parameter uncertainty. For each parameter we compute an uncertainty interval defined by the minimum and the maximum values over the set of all optimal and suboptimal parameters. Uncertain parameters have large uncertainty intervals.

An example of multi-exponential fit is given in Figure 12.

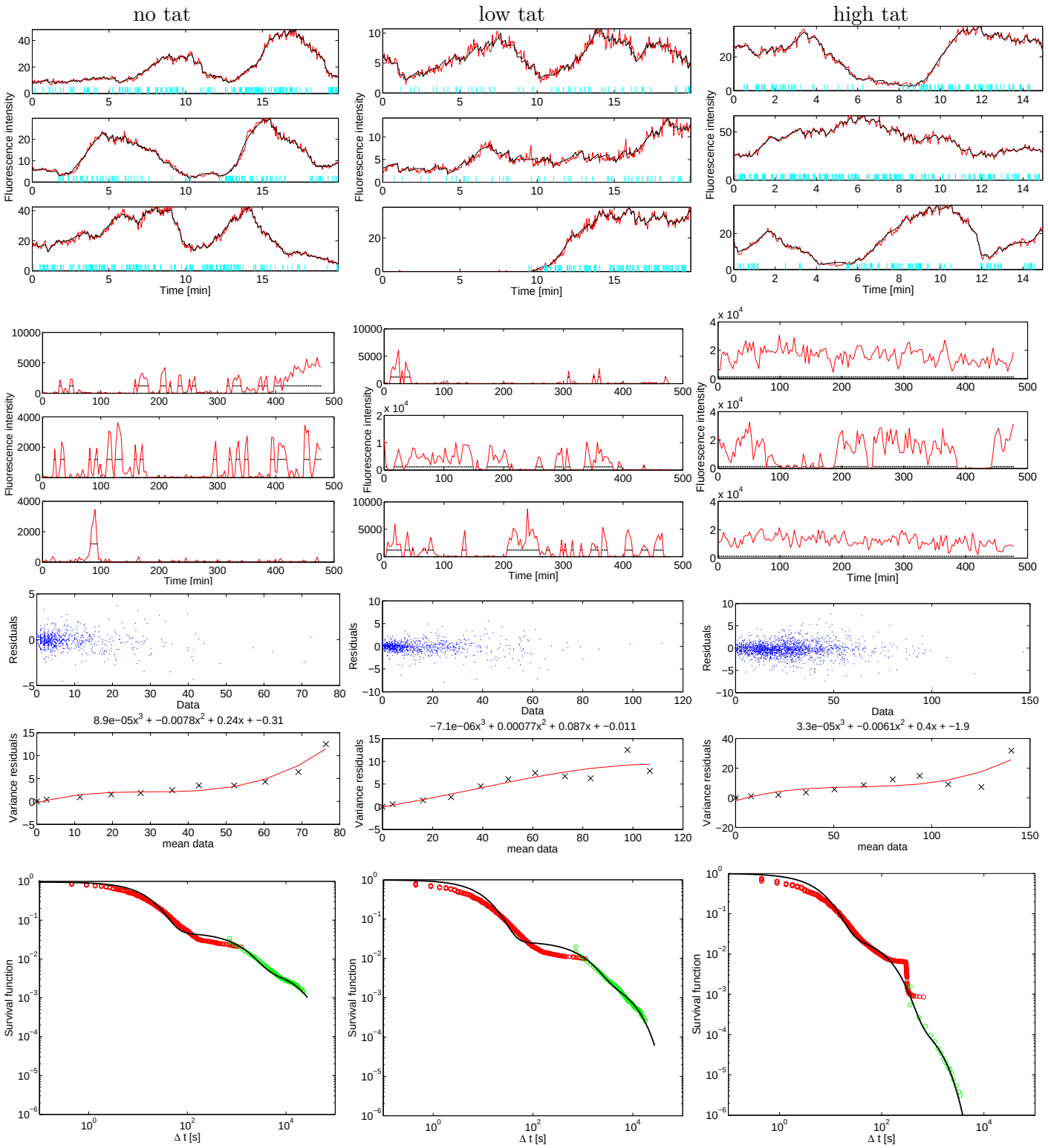


Figure 12: Results of the unconstrained three-exponential fit. First row: short movie data with reconstructed polymerase positions. Second row: long movie data. Third row: noise interpolation. Fourth row: most optimal fit for $\alpha = 0.30$.

4 Symbolic solution to the inverse problem

4.1 General model and waiting time distribution

We consider a continuous time Markov chain promoter model with N states $P_i, \quad i \in [1, N]$. One of these N states, P_o, is the "ON" state from which polymerase can start transcription, and all the other states are "OFF" states (non-processive). A supplementary state P_{N+1}, designates the start of processive elongation. From P_{N+1}, there is systematic return to P_o. The models have parameters $k_{i,j}, 1 \leq i, j \leq N+1$ indicating the transition rates from the promoter state i to the promoter state j. We consider that processive elongation immediately frees the operator and the promoter returns to the "ON" state. In mathematical terms

$$k_{N+1,o} = \infty. \tag{11}$$

We also consider that only one state, denoted X_N, can lead to processive elongation X_{N+1}: $k_{N,N+1} \neq 0$, $k_{i,N+1} = 0$, for $1 \leq i \leq N-1$. X_N is not necessarily X_o, for instance it can be a paused transcription state.

Because the movies are always started when transcription sites are in the active state and supposing that after each transcription initiation there is return to the active state, we model the experimental waiting time as the first time when the promoter reaches the state P_{N+1} starting from P_o. This is a first hitting time (or first passage time) problem. Because the lifetime of the state P_{N+1} is zero and P_{N+1} is always followed by P_o (see (11)), the same waiting time is also the first return time to P_o.

In order to compute the distribution of the first hitting time we use the following standard method.

Let $M(t)$ be the state of the Markov chain at the time t. For the purposes of this calculation, we can consider that $M(t)$ stops when it reaches P_{N+1}. Let $X_i = \mathbb{P}[M(t) = P_i | M(0) = P_o]$. Because $M(t)$ is stopped in P_{N+1}, one has $X_{N+1} = \mathbb{P}[M(t) = P_{N+1} | M(0) = P_o] = \mathbb{P}[\Delta \leq t]$. Thus, X_{N+1} is the cumulative distribution function of the waiting time Δ to reach P_{N+1} from P_o. The survival function of Δ is $S(t) = 1 - X_{N+1}(t)$.

The variables $X_i(t)$, $1 \leq i \leq N+1$, satisfy the following system of linear differential equations (the master equation):

$$\frac{d\boldsymbol{X}}{dt} = Q\boldsymbol{X}, \tag{12}$$

with the initial conditions $X_i(0) = \delta_{i,o}$, where δ is the Kronecker symbol; Q is the transpose transition rate matrix whose elements are defined by $Q_{j,i} = k_{i,j}$, $Q_{i,i} = -\sum_{j \neq i} k_{i,j}$.

Because $M(t)$ is stopped in P_{N+1}, the last column of the matrix Q is zero, namely $Q_{i,N+1} = 0$.

Let $\tilde{\boldsymbol{Q}}$ the $N \times N$ matrix obtained by eliminating the last line and the last column of the $(N+1) \times (N+1)$ matrix $\boldsymbol{Q}$.

Then $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_N)$ is the solution of the reduced equation

$$\frac{d\tilde{\boldsymbol{X}}}{dt} = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{X}}, \tag{13}$$

with initial conditions $X_i = \delta_{i,o}$ and reads

$$\tilde{\boldsymbol{X}}(t) = \sum_{i=1}^{N} C_i \boldsymbol{u}_i e^{\lambda_i t}, \tag{14}$$

where λ_i and $\boldsymbol{u}_i$, $i \in [1, N]$ are eigenvalues and eigenvectors of $\tilde{\boldsymbol{Q}}$, respectively.

Although (14) is written with the non-degenerate case in mind, when $\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_N$ (in general (14) is valid when $\tilde{Q}$ is diagonalizable), our final results, relating parameters of the survival function and kinetic parameters, can be extended to the degenerate case by continuous extension (see Section 4.2).

Furthermore, X_{N+1} can be obtained from the equation

$$\frac{dX_{N+1}}{dt} = k_{N,N+1} X_N, \tag{15}$$

with the initial condition $X_{N+1}(0) = 0$.

Without restricting generality, all eigenvectors $\boldsymbol{u}_i$ can be chosen such that their o-th coordinate is $u_o^i = 1$. Therefore, from (14), it follows

$$X_o = \sum_{i=1}^{N} C_i e^{\lambda_i t},$$

and from (15) it follows

$$X_{N+1}(t) = k_{N,N+1} \sum_{i=1}^{N} C_i u_N^i \frac{e^{\lambda_i t} - 1}{\lambda_i}. \tag{16}$$

From $\lim_{t\to\infty} X_{N+1}(t) = 1$ and (16) we get $k_{N,N+1} \sum_{i=1}^{N} \frac{C_i u_N^i}{\lambda_i} = -1$. Using again (16) we find

$$S(t) = 1 - X_{N+1}(t) = -k_{N,N+1} \sum_{i=1}^{N} \frac{C_i u_N^i e^{\lambda_i t}}{\lambda_i} = \sum_{i=1}^{N} A_i e^{\lambda_i t}. \tag{17}$$

Hence

$$A_i = -\frac{k_{N,N+1} u_N^i C_i}{\lambda_i},\ 1 \leq i \leq N. \tag{18}$$

In particular, when $N = o$

$$A_i = -\frac{k_{o,N+1} C_i}{\lambda_i},\ 1 \leq i \leq N. \tag{19}$$

Eq(17) implies that in the non-degenerate case the survival function is a combination of exponential functions, implying that the waiting time has a mixed exponential distribution. However, although the mixture coefficients satisfy $\sum_{i=1}^{N} A_i = 1$, they are not guaranteed positive in general (see Figure 11).

4.2 Degenerate case

The matrix $\tilde{\boldsymbol{Q}}$ is a linear function in the model parameters k_{ij}. According to classical results (see Kato), the eigenvalues of this matrix are branches of analytic functions in the parameters with only algebraic singularities. Moreover, the number of distinct eigenvalues is constant with the exception of a zero measure set of parameter values where this number is different. Excluding the permanently degenerate case when the matrix $\tilde{\boldsymbol{Q}}$ has a number of distinct eigenvalues smaller than N almost everywhere, we may consider that $\tilde{\boldsymbol{Q}}$ has N distinct eigenvalues except in a finite number of parameter values where it is degenerate, i.e. where $\lambda_i = \lambda_j$ for at least two distinct indices $i \neq j$.

Each degenerate case is arbitrarily close in the parameter space to a non-degenerate case. The solutions of the linear differential system (12) are continuous in the transition rates parameters $k_{i,j}$, therefore the survival function computed in a degenerate case can be approximated by survival functions computed for non-degenerate cases. Because all the survival functions are monotone, by Dini's theorem, this approximation can be made uniform for $t \in [0, T]$, for any T. Using the inequality $|A_i \exp(\lambda_i t) - A'_i \exp(\lambda'_i t)| < C \exp(\lambda'_i T), \forall t > T$, where $\lambda_i \leq \lambda'_i < 0$, we can show that the uniform approximation is valid for all times.

Let us now compute the survival function in the degenerate case.

When, in spite of having degenerate eigenvalues (there are N independent eignevectors), the matrix $\tilde{\boldsymbol{Q}}$ is diagonalizable, then Eqs.(14),(16),(17) hold. Therefore, in the diagonalizable case with degenerate eignevalues the survival function is a sum of less than N exponentials.

When the matrix $\tilde{\boldsymbol{Q}}$ is not diagonalizable (there are less than N independent eigenvectors), (14) no longer holds.

Let $g_i \leq n_i$ be the geometric multiplicity (number of independent eigenvectors) of the eigenvalue λ_i. Here n_i is the algebraic multiplicity of the eigenvalue λ_i, representing the number of times this eigenvalue occurs as a root of the characteristic polynomial ($\det(\tilde{\boldsymbol{Q}} - \lambda_i \boldsymbol{I}) \sim (\lambda - \lambda_i)^{n_i}$) and one has $\sum'_i n_i = N$, where the sum is over all distinct eigenvalues. Let us consider that $g_i < n_i$ for at least one i. In this situation, $\tilde{\boldsymbol{Q}}$ is not diagonalizable but can be reduced to a Jordan normal form. For each eigenvalue, there are g_i Jordan blocks. After reindexing the eigenvalues and Jordan blocks we have $N = \sum_{i=1}^{p} m_i$, where p is the total number of Jordan blocks $p = \sum'_i g_i$

(the sum is over distinct λ_i) and m_i is the dimension of a block i.

Let us remind that a generalized eigenvector $\boldsymbol{v}$ is any vector from the kernel $Ker((\tilde{\boldsymbol{Q}}-\lambda_i)^{m_i})$. The subspace $Ker((\tilde{\boldsymbol{Q}}-\lambda_i)^{m_i})$ corresponds to a Jordan block and is generated by a chain of generalized eigenvectors u_i, $(\tilde{\boldsymbol{Q}}-\lambda_i)\boldsymbol{u}_i, \ldots, (\tilde{\boldsymbol{Q}}-\lambda_i)^{m_i-1}\boldsymbol{u}_i$, where $\boldsymbol{u}_i$ is a generalized vector that satisfies $(\tilde{\boldsymbol{Q}}-\lambda_i)\boldsymbol{u}_i \neq 0, \ldots, (\tilde{\boldsymbol{Q}}-\lambda_i)^{m_i-1}\boldsymbol{u}_i \neq 0$. Furthermore, the solution of (13) starting from any generalized vector $\boldsymbol{v}$ reads:

$$\tilde{\boldsymbol{X}}(t) = \exp(\lambda_i t)\sum_{j=1}^{m_i}\frac{t^{j-1}}{(j-1)!}(\tilde{\boldsymbol{Q}}-\lambda_i)^{j-1}\boldsymbol{v}. \tag{20}$$

Let us consider that

$$\tilde{\boldsymbol{X}}(0) = \sum_{i=1}^{p}\sum_{j=1}^{m_i} C_{i,j}(\tilde{\boldsymbol{Q}}-\lambda_i)^{j-1}\boldsymbol{u}_i.$$

By the definition of the generalized eignvectors, $(\tilde{\boldsymbol{Q}}-\lambda_i)^{m_i}\boldsymbol{u}_i = 0$.

Therefore, in the non-diagonalizable case, (14) must be replaced by:

$$\tilde{\boldsymbol{X}}(t) = \sum_{i=1}^{p}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i+2-j} C_{i,j}\exp(\lambda_i t)\frac{t^{k-1}}{(k-1)!}(\tilde{\boldsymbol{Q}}-\lambda_i)^{j+k-2}\boldsymbol{u}_i. \tag{21}$$

Then, (16) should be replaced by

$$X_{N+1} = k_{N,N+1}\int_0^t X_N(s)\,ds = k_{N,N+1}\sum_{i=1}^{p}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i+2-j} C_{i,j}u_N^{i,j+k-2}\frac{\gamma(k,\lambda_i t)}{\lambda_i^k(k-1)!}, \tag{22}$$

where $u_N^{i,j}$ is the N^{th} coordinate of $(\tilde{\boldsymbol{Q}}-\lambda_i)^j\boldsymbol{u}_i$, and γ is the incomplete gamma functions. It follows that (17) should be replaced by

$$S(t) = 1 - X_{N+1}(t) = k_{N,N+1}\sum_{i=1}^{p}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i+2-j}\frac{C_{i,j}u_N^{i,j+k-2}}{\lambda_i^k}\left[1-\frac{\gamma(k,\lambda_i t)}{(k-1)!}\right]. \tag{23}$$

Eq. (23) implies that in the non-diagonalizable case, the survival function is a combination of gamma functions, implying that the waiting time has a mixed gamma distribution.

As an example illustrating this case let us consider the irreversible chain $P_1 \overset{k}{\to} P_2 \overset{k}{\to} P_3 \overset{k_{ini}}{\to} P_4$ where P_3 is the ON state and P_4 is the EL state. in this case we have

$$\tilde{\boldsymbol{Q}} = \begin{bmatrix} -k & 0 & 0 \\ k & -k & 0 \\ 0 & k & -k_{ini} \end{bmatrix}.$$

There are two distinct eigenvalues $\lambda_1 = -k$ and $\lambda_2 = -k_{ini}$. Each eigenvalue contributes with one Jordan block of dimensions 2 and 1, respectively. The chains of generalized eigenvectors are

$$\boldsymbol{u}_1 = \begin{bmatrix} -\frac{k-k_{ini}}{k} \\ 1 \\ 0 \end{bmatrix}, (\tilde{\boldsymbol{Q}} - \lambda_1)\boldsymbol{u}_1 = \begin{bmatrix} 0 \\ k_{ini} - k \\ k \end{bmatrix}, \boldsymbol{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Suppose we want to compute the distribution of the waiting time to reach P_4 starting from P_1. Then

$$\tilde{\boldsymbol{X}}(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = -\frac{k}{k-k_{ini}} \begin{bmatrix} -\frac{k-k_{ini}}{k} \\ 1 \\ 0 \end{bmatrix} - \frac{k}{(k-k_{ini})^2} \begin{bmatrix} 0 \\ k_{ini} - k \\ k \end{bmatrix} + \frac{k^2}{(k-k_{ini})^2} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

and

$$\tilde{\boldsymbol{X}}(t) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = -\frac{k}{k-k_{ini}} e^{-kt} [\begin{bmatrix} -\frac{k-k_{ini}}{k} \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ k_{ini} - k \\ k \end{bmatrix}] - \frac{k}{(k-k_{ini})^2} e^{-kt} \begin{bmatrix} 0 \\ k_{ini} - k \\ k \end{bmatrix} + \frac{k^2}{(k-k_{ini})^2} e^{-k_{ini}t} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The survival function reads

$$S(t) = A_1 \left[1 - \gamma(1, -kt)\right] + A_2 \exp(-kt) + A_3 \exp(-k_{ini}t),$$

where $A_1 = -k_{ini}/(k - k_{ini})$, $A_2 = -k_{ini}k/(k - k_{ini})^2$, $A_3 = k^2/(k - k_{ini})^2$ satisfy $A_1 + A_2 + A_3 = 1$.

The waiting time is distributed according to a mixture of gamma and exponential distributions. If $k_{ini} >> k$, then $A_1 \approx 1$, $A_2, A_3 \approx 0$, meaning that the waiting time is distributed according to a gamma distribution of shape parameter 2 and scale parameter $1/k$.

If in the previous model we make $k_{ini} = k$,

$$\tilde{\boldsymbol{Q}} = \begin{bmatrix} -k & 0 & 0 \\ k & -k & 0 \\ 0 & k & -k \end{bmatrix}.$$

Then, $\tilde{\boldsymbol{Q}}$ has only one eigenvalue $\lambda = -k$ and one Jordan block of dimension 3. The chain of generalized eigenvectors is

$$\boldsymbol{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, (\tilde{\boldsymbol{Q}} + k\boldsymbol{I})\boldsymbol{u} = \begin{bmatrix} 0 \\ k \\ 0 \end{bmatrix}, (\tilde{\boldsymbol{Q}} + k\boldsymbol{I})^2\boldsymbol{u} = \begin{bmatrix} 0 \\ 0 \\ k^2 \end{bmatrix}$$

The survival function reads

$$S(t) = 1 - \frac{\gamma(3, -kt)}{2!},$$

meaning that the waiting time is distributed according to a gamma distribution with scale parameter $1/k$ and shape parameter 3. This result is obvious from the structure of the model. If $k = k_{ini}$, in order to reach P_4 from P_1 one needs three exponentially distributed steps of equal mean time $1/k$; a sum of three independent equally distributed exponential variables is a gamma distribution of shape parameter 3.

In general, if the chain contains n limiting steps of constant k, the waiting time to reach the end of the chain starting from the beginning is distributed approximately according to a gamma distribution with shape parameter n and scale parameter $1/k$.

4.3 Inverse problem

In order to formulate a well posed inverse problem, we have to choose a structure of the model. The structure is defined by the directed graph G whose vertices are the promoter states and such that there is an edge from i to j if and only if $k_{i,j} \neq 0$. Thus the model structure specifies which transitions are allowed between the promoter states. We also need to specify which one of the promoter states is ON.

Given a model structure, the inverse problem consists in computing the kinetic constants $k_{i,j}, 1 \leq i, j \leq N$ and $k_{N,N+1}$ from the $2N-1$ parameters of the survival function. This is possible only if there are at most $2N-1$, kinetic constants. Uniqueness and thus well-posedness of the solution is possible only if there are exactly $2N-1$ parameters. However, not all models with $2N-1$ parameters have unique solutions of the inverse problem (an example is the model M3, see Figure 1 and Section 4.6).

In order to solve the inverse problem, we must write down the equations relating the parameters $k_{i,j}$, A_i and λ_i.

Let us consider that all the nonzero kinetic parameters are the $2N-1$ elements of a vector $\boldsymbol{k} \in \mathbb{R}^{2N-1}$.

Vieta's formulas

Let us introduce the elementary symmetric polynomials of eigenvalues

$$L_1 = \sum_{i=1}^{N} \lambda_i \tag{24}$$

$$L_2 = \sum_{i<j} \lambda_i \lambda_j \tag{25}$$

$$\vdots \tag{26}$$

$$L_N = \lambda_1 \lambda_2 \dots \lambda_N \tag{27}$$

The characteristic polynomial of $\tilde{\boldsymbol{Q}}$ is

$$P(\lambda) = det(\tilde{\boldsymbol{Q}} - \lambda \boldsymbol{I}) = (-1)^N \lambda^N + a_{N-1}(\boldsymbol{k})\lambda^{N-1} + \ldots + a_1(\boldsymbol{k})\lambda + a_0(\boldsymbol{k}) \tag{28}$$

where the coefficients a_i are multivariate polynomial functions of the kinetic constants.

The coefficients of the characteristic polynomial are related to the symmetric polynomials of eigenvalues by the so-called Vieta's formulas. We have the following N equations for the kinetic constants:

$$L_j = (-1)^{N-j} a_{N-j}(\boldsymbol{k}), \quad j \in [1, N] \tag{29}$$

Eigenvectors

The eigenvectors of $\tilde{\boldsymbol{Q}}$ are solutions of the system of linear equations $(\tilde{\boldsymbol{Q}} - \lambda \boldsymbol{I})\boldsymbol{u} = 0$ and are chosen of the form $\boldsymbol{u} = (u_1(\lambda, \boldsymbol{k}), \ldots, u_{o-1}(\lambda, \boldsymbol{k}), 1, u_{o+1}(\lambda, \boldsymbol{k}), \ldots, u_N(\lambda, \boldsymbol{k}))$, where $u_n(\lambda, \boldsymbol{k}), n \in [1, N-1]$ are rational functions (ratios of polynomials) of λ and $\boldsymbol{k}$.

The initial conditions satisfied by the variables X_i provide a linear system of equations for the constants C_i:

$$\sum_{j=1}^{N} u_i(\lambda_j, \boldsymbol{k}) C_j = \delta_{i,o}, \quad i \in [1, N] \tag{30}$$

Let $C_i(\boldsymbol{\lambda}, \boldsymbol{k}), \quad i \in [1, N]$ be the unique solution of (30).

From (18),(19) we get $N-1$ equations for the kinetic constants $\boldsymbol{k}$:

$$C_i(\boldsymbol{\lambda}, \boldsymbol{k}) = -A_i \lambda_i / (u_o^i k_{N,N+1}), \quad i \in [1, N-1] \tag{31}$$

Inverse problem

The solution of the inverse problem is the solution of the system of $2N-1$ equations (29) and (31).

In the next sections we solve this system symbolically. When a solution of the inverse problem exists, the kinetic parameters $k_{i,j}$ can be expressed as functions in λ_i and A_i. These functions are symmetric in the pairs

(λ_i, A_i) and homogeneous of degree -1 in λ_i. These functions are not always rational. For instance, they can have branching singularities, allowing, eventually, to pass from one solution to another, equivalent one. In general, multiple solutions are equivalent with respect to symmetries of the model. For instance, the model M2 in the Figure 1 is symmetric with respect to the permutation of the two lateral chains. In this case there are two solutions of the inverse problem, one solution being obtained from the other by permuting the parameters $k_1^{\pm}$ with $k_2^{\pm}$. The general solutions will be presented elsewhere. In the sequel we provide full solutions for some models with $N \leq 4$.

Recursion relations for eigenvectors

The eigenvector components $u_i(\lambda_j, \boldsymbol{k})$ can be obtained by recursion along the structure digraph.

We consider models such that any state of the promoter is connected to the ON state, in both directions, by directed paths on the structure digraph.

In the sequel, we discuss two representative cases.

The type I (*single chain*) model is a reversible chain ending with the P_N state:

$$P_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} P_2 \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} P_3 \ldots P_{N-1} \underset{k_{N-1}^-}{\overset{k_{N-1}^+}{\rightleftharpoons}} P_N \overset{k_N}{\to}$$

For this model, an eigenvector $(b_1, b_2, \ldots, b_N)$ satisfies the equations

$$k_1^- b_2 - (k_1^+ + \lambda) b_1 = 0, \tag{32}$$

$$k_{n-1}^+ b_{n-1} + k_n^- b_{n+1} - (k_n^+ + k_{n-1}^- + \lambda) b_n = 0, \text{for } 2 \leq n \leq N-1. \tag{33}$$

We can choose $b_1 = 1$ and then from (32) $b_2 = \frac{k_1^+ + \lambda}{k_1^-}$. Threfore, b_n satisfy the recursion

$$\begin{aligned} b_1 &= 1, b_2 = \frac{k_1^+ + \lambda}{k_1^-} \\ b_{n+1} &= \frac{k_n^+ + k_{n-1}^- + \lambda}{k_n^-} b_n - \frac{k_{n-1}^+}{k_n^-} b_{n-1}, 2 \leq n \leq N-1. \end{aligned} \tag{34}$$

In order to have $u_o^i = 1$ for all $1 \leq i \leq N$, we define

$$u_n(\lambda, \boldsymbol{k}) = \frac{b_n(\lambda, \boldsymbol{k})}{b_o(\lambda, \boldsymbol{k})}, \quad n \in [1, N], \tag{35}$$

where b_n are rational functions of λ and $\boldsymbol{k}$ computed with the recursion (34).

The type II model is a reversible chain with the P_N state inside the chain:

$$P_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} P_2 \ldots P_{N-r} \underset{k_{N-r}^-}{\overset{k_{N-r}^+}{\rightleftharpoons}} P_N \underset{k_{N-r+1}^-}{\overset{k_{N-r+1}^+}{\rightleftharpoons}} P_{N-r+1} \ldots P_{N-2} \underset{k_{N-1}^-}{\overset{k_{N-1}^+}{\rightleftharpoons}} P_{N-1}$$

The type II model can be also described as two reversible chains branching from the P_N state. We can call this type *two chain model*.

For this model, an eigenvector $(b_1, b_2, \ldots, b_N)$ satisfies the recursion

$$k_1^- b_2 - (k_1^+ + \lambda) b_1 = 0, \tag{36}$$

$$k_{n-1}^+ b_{n-1} + k_n^- b_{n+1} - (k_n^+ + k_{n-1}^- + \lambda) b_n = 0, \text{for } 2 \le n \le N - r, \tag{37}$$

$$k_n^+ b_{n-1} + k_{n+1}^- b_{n+1} - (k_{n-1}^+ + k_n^- + \lambda) b_n = 0, \text{for } N - r + 1 \le n \le N - 2, \tag{38}$$

$$k_{N-1}^+ b_{N-2} - (k_{N-1}^- + \lambda) b_{N-1} = 0, \tag{39}$$

The recursion (36),(37),(38),(39) can be solved in the following way:

i) Choose $b_1 = 1$ and compute b_2 from (36).

ii) Use (37) to compute b_n, $3 \le n \le N - r$ and b_N.

iii) Use (39) and (38) to compute b_n, $N - r + 1 \le n \le N - 1$ and b_N as multiples of b_{N-1}.

iv) Determine b_{N-1} from b_N, already computed at step ii).

Below we study several examples of type I and type II models.

4.4 Symbolic solution to the inverse problem for the M1 model ($N = 3$)

This model is described by the transitions $P_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} P_2 \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} P_3 \overset{k_3}{\to}$. It is a type I model. In this case P_3 is the ON state. The matrix of kinetic rates reads

$$\tilde{Q} = \begin{bmatrix} -k_1^+ & k_1^- & 0 \\ k_1^+ & -(k_2^+ + k_1^-) & k_2^- \\ 0 & k_2^+ & -(k_3 + k_2^-) \end{bmatrix}.$$

The characteristic polynomial of $\tilde{\boldsymbol{Q}}$ is $P(\lambda) = det(\tilde{\boldsymbol{Q}} - \lambda \boldsymbol{I}) = -\lambda^3 - (k_3 + k_1^- + k_2^- + k_1^+ + k_2^+)\lambda^2 - (k_3 k_1^- + k_3 k_1^+ + k_3 k_2^+ + k_1^- k_2^- + k_2^- k_1^+ + k_1^+ k_2^+)\lambda - k_3 k_1^+ k_2^+$.

The Vieta formulas read

$$k_3 k_1^+ k_2^+ = -L_3 \tag{40}$$

$$k_3 k_1^- + k_3 k_1^+ + k_3 k_2^+ + k_1^- k_2^- + k_2^- k_1^+ + k_1^+ k_2^+ = L_2 \tag{41}$$

$$k_3 + k_1^- + k_2^- + k_1^+ + k_2^+ = -L_1 \tag{42}$$

The solution of the recursion (34) is

$$b_1 = 1 \tag{43}$$

$$b_2 = \frac{k_1^+ + \lambda}{k_1^-} \tag{44}$$

$$b_3 = \frac{k_1^+ k_2^+ + (k_1^- + k_1^+ + k_2^+)\lambda + \lambda^2}{k_1^- k_2^-} \tag{45}$$

The system (30) has the solution

$$\begin{aligned} C_1 &= \frac{k_1^+ k_2^+ + (k_1^- + k_1^+ + k_2^+)\lambda_1 + \lambda_1^2}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)} \\ C_2 &= \frac{k_1^+ k_2^+ + (k_1^- + k_1^+ + k_2^+)\lambda_2 + \lambda_2^2}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} \\ C_3 &= \frac{k_1^+ k_2^+ + (k_1^- + k_1^+ + k_2^+)\lambda_3 + \lambda_3^2}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)} \end{aligned}$$

The unique solution of (29) and (31) is

$$k_3 = -S_1, \tag{46}$$

$$k_2^+ = \frac{S_2^2 - S_1 S_3}{S_1(-S_1^2 + S_2)}, \tag{47}$$

$$k_2^- = S_1 - \frac{S_2}{S_1}, \tag{48}$$

$$k_1^+ = \frac{L_3(-S_1^2 + S_2)}{S_2^2 - S_1 S_3}, \tag{49}$$

$$k_1^- = \frac{A_1 A_2 A_3 S_1 (\lambda_1 - \lambda_2)^2 (\lambda_1 - \lambda_3)^2 (\lambda_2 - \lambda_3)^2}{(-S_1^2 + S_2)(S_2^2 - S_1 S_3)}, \tag{50}$$

where

$$L_1 = \lambda_1 + \lambda_2 + \lambda_3, \tag{51}$$

$$L_2 = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3, \tag{52}$$

$$L_3 = \lambda_1\lambda_2\lambda_3, \tag{53}$$

$$S_1 = A_1\lambda_1 + A_2\lambda_2 + A_3\lambda_3, \tag{54}$$

$$S_2 = A_1\lambda_1^2 + A_2\lambda_2^2 + A_3\lambda_3^2, \tag{55}$$

$$S_3 = A_1\lambda_1^3 + A_2\lambda_3^3 + A_3\lambda_3^3. \tag{56}$$

4.5 Symbolic solution to the inverse problem for the M2 model ($N = 3$)

This model is described by the transitions $P_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} P_3 \overset{k_3}{\uparrow}$ $P_3 \underset{k_2^+}{\overset{k_2^-}{\rightleftharpoons}} P_2$. In this case P_3 is the ON state. Model M2 is a type II model. It has a matrix of kinetic rates

$$\tilde{Q} = \begin{bmatrix} -k_1^+ & 0 & k_1^- \\ 0 & -k_2^+ & k_2^- \\ k_1^+ & k_2^+ & -(k_3 + k_1^- + k_2^-) \end{bmatrix}.$$

The characteristic polynomial of $\tilde{\boldsymbol{Q}}$ is $P(\lambda) = det(\tilde{\boldsymbol{Q}} - \lambda \boldsymbol{I}) = -\lambda^3 - (k_3 + k_1^- + k_2^- + k_1^+ + k_2^+)\lambda^2 - (k_3 k_1^+ + k_3 k_2^+ + k_1^- k_2^+ + k_2^- k_1^+ + k_1^+ k_2^+)\lambda - k_3 k_1^+ k_2^+$.

The Vieta formulas read

$$k_3 k_1^+ k_2^+ = -L_3 \tag{57}$$

$$k_3 k_1^+ + k_3 k_2^+ + k_1^- k_2^+ + k_2^- k_1^+ + k_1^+ k_2^+ = L_2 \tag{58}$$

$$k_3 + k_1^- + k_2^- + k_1^+ + k_2^+ = -L_1 \tag{59}$$

The solution of the recursion (36),(37),(38),(39) reads

$$b_1 = 1 \tag{60}$$

$$b_2 = \frac{k_2^-(k_1^+ + \lambda)}{k_1^-(k_2^+ + \lambda)} \tag{61}$$

$$b_3 = \frac{k_1^+ + \lambda}{k_1^-} \tag{62}$$

The system (30) has the solution

$$\begin{aligned} C_1 &= \frac{k_1^+ k_2^+ + (k_1^+ + k_2^+)\lambda_1 + \lambda_1^2}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)} \\ C_2 &= \frac{k_1^+ k_2^+ + (k_1^+ + k_2^+)\lambda_2 + \lambda_2^2}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} \\ C_3 &= \frac{k_1^+ k_2^+ + (k_1^+ + k_2^+)\lambda_3 + \lambda_3^2}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)} \end{aligned}$$

Up to the permution symmetry $P_1 \leftrightarrow P_2$ the solution of (29) and (31) is unique and described by

$$k_3 = -S_1, \tag{63}$$

$$k_2^+ = \frac{1}{2}\left[-L_1 + \frac{S_2}{S_1} - \frac{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}{S_1}\right], \tag{64}$$

$$k_2^- = \frac{1}{2}\left[S_1 - \frac{S_2}{S_1} + \frac{-S_1^2L_1 + S_1S_2 + S_1L_2 - L_3 + \frac{S_2^2}{S_1} - S_3}{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}\right], \tag{65}$$

$$k_1^+ = \frac{1}{2}\left[-L_1 + \frac{S_2}{S_1} + \frac{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}{S_1}\right], \tag{66}$$

$$k_1^- = \frac{1}{2}\left[S_1 - \frac{S_2}{S_1} - \frac{-S_1^2L_1 + S_1S_2 + S_1L_2 - L_3 + \frac{S_2^2}{S_1} - S_3}{\sqrt{(S_1L_1 - S_2)^2 - 4L_3S_1}}\right], \tag{67}$$

4.6 Symbolic solution to the inverse problem for the M3 model ($N = 3$)

The chain without P_4 is the same as the model M1. Like M1, M3 is a type I model. The difference is the position of the ON state which is in the middle of the chain (P_2 is the ON state).

The matrix $\tilde{Q}$ and its characteristic polynomial are the same as in the section 4.4. In particular, the Vieta relations remain the same:

$$\begin{gathered} k_3k_1^+k_2^+ = -L_3, \\ k_3k_1^- + k_3k_1^+ + k_3k_2^+ + k_1^-k_2^- + k_2^-k_1^+ + k_1^+k_2^+ = L_2, \\ k_3 + k_1^- + k_2^- + k_1^+ + k_2^+ = -L_1. \end{gathered} \tag{68}$$

However, instead of computing the waiting time for reaching P_4 starting from P_3, we compute the waiting time for reaching P_4 starting from P_2. In this model, the significance of the states P_2 and P_3 is ON and PAUSE, respectively. The observed waiting time is from ON to EL, therefore from P_2 to P_4.

We look for solutions of the master equation (13) with initial conditions $\boldsymbol{X}(0) = (0, 1, 0, 0)$.

Like in section in order to compute solutions of (13) we need the eigenvectors of $\tilde{Q}$. For the new initial conditions it is convenient to impose the normalization condition $u_2 = 1$, where u_i, $1 \leq i \leq 3$ are the components of the eigenvector $\boldsymbol{u}$. We get

$$u_1 = k_1^-/(k_1^+ + \lambda), \tag{69}$$

$$u_2 = 1, \tag{70}$$

$$u_3 = k_2^+/(k_3 + k_2^- + \lambda). \tag{71}$$

A solution of the (13) reads $\boldsymbol{X}(t) = C_1\boldsymbol{u}_1\exp(\lambda_1 t) + C_2\boldsymbol{u}_2\exp(\lambda_2 t) + C_3\boldsymbol{u}_3\exp(\lambda_3 t)$. From the initial conditions, it follows

$$\begin{gathered} C_1\frac{k_1^-}{k_1^+ + \lambda_1} + C_2\frac{k_1^-}{k_1^+ + \lambda_2} + C_3\frac{k_1^-}{k_1^+ + \lambda_3} = 0, \\ C_1 + C_2 + C_3 = 1, \\ C_1\frac{k_2^+}{k_3 + k_2^- + \lambda_1} + C_2\frac{k_2^+}{k_3 + k_2^- + \lambda_2} + C_3\frac{k_2^+}{k_3 + k_2^- + \lambda_3} = 0. \end{gathered} \tag{72}$$

The system (72) has the solution

$$\begin{aligned} C_1 &= \frac{(k_1^+ + \lambda_1)(k_3 + k_2^- + \lambda_1)}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)}, \\ C_2 &= -\frac{(k_1^+ + \lambda_2)(k_3 + k_2^- + \lambda_2)}{(\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)}, \\ C_3 &= \frac{(k_1^+ + \lambda_3)(k_3 + k_2^- + \lambda_3)}{(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)}. \end{aligned} \tag{73}$$

X_4 obeys $\frac{dX_4}{dt} = k_3 X_3$ and the survival function is

$$s(t) = \sum_{i=1}^{3} A_i exp(\lambda_i t) = 1 - X_3 = -\sum_{i=1}^{3} \frac{C_i k_2^+}{k_3\lambda_i(k_3 + k_2^- + \lambda_i)} exp(\lambda_i t).$$

The relation between C_i and A_i reads:

$$-\lambda_i A_i = k_3 C_i u_3(\lambda_i) = k_3 C_i k_2^+/(k_3 + k_2^- + \lambda_i). \tag{74}$$

By definition $s(0) = 1$, therefore

$$A_1 + A_2 + A_3 = 1. \tag{75}$$

Using (74) and (29) we can show that

$$A_1\lambda_1 + A_2\lambda_2 + A_3\lambda_3 = 0. \tag{76}$$

Eq.76 is very important. It implies that in this case, instead of 5 independent parameters, the survival function has only 4 independent parameters A_1, λ_1, λ_2, λ_3. Using (76) and (75) we can compute the remaining parameters as

$$\begin{aligned} A_2 &= -\frac{\lambda_3 + A_1(\lambda_1 - \lambda_3)}{\lambda_2 - \lambda_3}, \\ A_3 &= \frac{\lambda_2 + A_1(\lambda_1 - \lambda_2)}{\lambda_2 - \lambda_3}. \end{aligned} \tag{77}$$

If the condition (76) is not satisfied, then the system formed from eqs. (68) and (74) (with $i = 1, 2$) is incompatible.

If the condition (76) is satisfied, then the system (68), (74) is indeterminate and has an infinity of solutions. In this case, all the solutions can be expressed as functions of a free parameter. In the sequel we will choose k_3 as free parameter. This choice leads to the following symmetric expressions:

$$\begin{aligned} k_1^+ &= L_3/S_2, \\ k_2^+ &= -S_2/k_3, \\ k_2^- &= \frac{S_2 - 2k_3^2 - k_3 L_1 \pm \sqrt{k_3(k_3(L_1^2 - 4L_2 - 4S_2) - 2S_3 - 2L_3) + S_2^2}}{2k3}, \\ k_1^- &= -k_3 - S_3/S_2 + S_2/k_3 - k_2^-. \end{aligned} \tag{78}$$

4.7 Symbolic solution to the inverse problem for the two state ON-OFF model ($N = 2$)

The two states ON-OFF model (telegraph model) reads $P_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} P_2 \overset{k_2}{\to}$.

In order to identify this model we use a two exponential fit of the survival function $S(t) = A_1 \exp(\lambda_1 t) + A_2 \exp(\lambda_2 t)$. Without restricting the generality, we can consider that $\lambda_1 < \lambda_2 < 0$. Then, from $S'(t) \leq 0$ it follows $\frac{\lambda_2}{\lambda_2 - \lambda_1} \leq A_1 \leq 1$, $A_2 = 1 - A_1$.

From the parameters of the survival function we can compute the model parameters as follows

$$\begin{aligned} S_1 &= A_1\lambda_1 + A_2\lambda_2, \\ S_2 &= A_1\lambda_1^2 + A_2\lambda_2^2, \\ S_3 &= A_1\lambda_1^3 + A_2\lambda_2^3, \\ k_2 &= -S_1, \\ k_1^- &= S_1 - S_2/S_1, \\ k_1^+ &= (S_3S_1 - S_2^2)/S_1/(S_1^2 - S_2). \end{aligned} \tag{79}$$

4.8 Symbolic solution to the inverse problem for the four state chain model ($N = 4$)

This model is described by the transitions

$$P_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} P_2 \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} P_3 \underset{k_3^-}{\overset{k_3^+}{\rightleftharpoons}} P_4 \overset{k_4}{\to}.$$

In this case P_4 is the ON state. The model is of type I.

We have

$$\tilde{Q} = \begin{bmatrix} -k_1^+ & k_1^- & 0 & 0 \\ k_1^+ & -(k_2^+ + k_1^-) & k_2^- & 0 \\ 0 & k_2^+ & -(k_3^+ + k_2^-) & k_3^- \\ 0 & 0 & k_3^+ & -(k_4 + k_3^-) \end{bmatrix}.$$

The Vieta formulas read

$$k_4 k_1^+ k_2^+ k_3^+ = L_4 \tag{80}$$

$$k_4 k_1^- k_2^- + k_4 k_2^- k_1^+ + k_4 k_1^- k_3^+ + k_4 k_1^+ k_2^+ + k_4 k_1^+ k_3^+ + k_4 k_2^+ k_3^+ + k_1^- k_2^- k_3^- + k_2^- k_3^- k_1^+ + k_3^- k_1^+ k_2^+ +$$
$$+ k_1^+ k_2^+ k_3^+ = -L_3 \tag{81}$$

$$k_4 k_1^- + k_4 k_2^- + k_4 k_1^+ + k_4 k_2^+ + k_4 k_3^+ + k_1^- k_2^- + k_1^- k_3^- + k_2^- k_3^- + k_2^- k_1^+ + k_1^- k_3^+ + k_3^- k_1^+ + k_3^- k_2^+ +$$
$$+ k_1^+ k_2^+ + k_1^+ k_3^+ + k_2^+ k_3^+ = L_2 \tag{82}$$

$$k_4 + k_1^- + k_2^- + k_3^- + k_1^+ + k_2^+ + k_3^+ = -L_1 \tag{83}$$

The solution of the recursion (34) is

$$b_1 = 1 \tag{84}$$

$$b_2 = \frac{k_1^+ + \lambda}{k_1^-} \tag{85}$$

$$b_3 = \frac{k_1^+ k_2^+ + (k_1^- + k_1^+ + k_2^+)\lambda + \lambda^2}{k_1^- k_2^-} \tag{86}$$

$$b_4 = \frac{\lambda^3 + (k_1^- + k_2^- + k_1^+ + k_2^+ + k_3^+)\lambda^2 + (k_1^- k_2^- + k_2^- k_1^+ + k_1^- k_3^+ + k_1^+ k_2^+ + k_1^+ k_3^+ + k_2^+ k_3^+)\lambda + k_1^+ k_2^+ k_3^+}{k_1^- k_2^- k_3^-} \tag{87}$$

The system (30) has the solution

$$C_1 = \frac{\lambda_1^3 + (k_1^- + k_2^- + k_1^+ + k_2^+ + k_3^+)\lambda_1^2 + (k_1^- k_2^- + k_2^- k_1^+ + k_1^- k_3^+ + k_1^+ k_2^+ + k_1^+ k_3^+ + k_2^+ k_3^+)\lambda_1 + k_1^+ k_2^+ k_3^+}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)(\lambda_1 - \lambda_4)}$$

$$C_2 = \frac{\lambda_2^3 + (k_1^- + k_2^- + k_1^+ + k_2^+ + k_3^+)\lambda_2^2 + (k_1^- k_2^- + k_2^- k_1^+ + k_1^- k_3^+ + k_1^+ k_2^+ + k_1^+ k_3^+ + k_2^+ k_3^+)\lambda_2 + k_1^+ k_2^+ k_3^+}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)(\lambda_2 - \lambda_4)}$$

$$C_3 = \frac{\lambda_3^3 + (k_1^- + k_2^- + k_1^+ + k_2^+ + k_3^+)\lambda_3^2 + (k_1^- k_2^- + k_2^- k_1^+ + k_1^- k_3^+ + k_1^+ k_2^+ + k_1^+ k_3^+ + k_2^+ k_3^+)\lambda_3 + k_1^+ k_2^+ k_3^+}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)(\lambda_3 - \lambda_4)}$$

The eqs. (29) and (31) provide

$$k_4 = -S_1 \tag{88}$$

$$k_3^+ = (S_2^2 - S_1S_3)/(-S_1^3 + S_2S_1) \tag{89}$$

$$k_3^- = S_1 - S_2/S_1 \tag{90}$$

$$k_2^+ = -\frac{(-S_1^2 + S_2)P_9}{(-S_2^2 + S_1S_3)P_6} \tag{91}$$

$$k_2^- = \frac{S_1P_6}{(-S_2^2 + S_1S_3)(-S_1^2 + S_2)} \tag{92}$$

$$k_1^+ = -\frac{L_4P_6}{P_9} \tag{93}$$

$$k_1^- = \frac{(-S_2^2 + S_1S_3)P_{12}}{P_6P_9} \tag{94}$$

where $S_n = \sum_{i=1}^4 A_i\lambda_i^n$,

$L_n = \sum_{1\le i_1<i_2<\ldots<i_n\le 4} \lambda_{i_1}\lambda_{i_2}\ldots\lambda_{i_n}$,$P_6 = L_3S_1^3 - L_2S_1^2S_2 + L_1S_1^2S_3 - L_4S_1^2 - 2S_1S_2S_3 - L_3S_1S_2 + S_2^3 + L_2S_2^2 - L_1S_2S_3 + L_4S_2 + S_3^2$,

$P_9 = L_1^2S_2^2S_3 - L_1L_2S_1S_2S_3 - L_1L_2S_2^3 + L_1L_3S_1^2S_3 + L_1L_3S_1S_2^2 - L_1L_4S_1S_3 - L_1L_4S_2^2 - 2L_1S_2S_3^2 + L_2^2S_1S_2^2 - 2L_2L_3S_1^2S_2 + 2L_2L_4S_1S_2 + L_2S_1S_3^2 + L_2S_2^2S_3 + L_3^2S_1^3 - 2L_3L_4S_1^2 - 3L_3S_1S_2S_3 + L_3S_2^3 + L_4^2S_1 + L_4S_1^2S_3 - L_4S_1S_2^2 + 2L_4S_2S_3 + S_3^3$,

$P_{12} = L_1^3S_2^3S_3 - 2L_1^2L_2S_1S_2^2S_3 - L_1^2L_2S_2^4 + L_1^2L_3S_1^2S_2S_3 + L_1^2L_3S_1S_2^3 + L_1^2L_3S_2^2S_3 + L_1^2L_4S_1^3S_3 - 3L_1^2L_4S_1S_2S_3 - L_1^2L_4S_2^3 - 3L_1^2S_2^2S_3^2 + L_1L_2^2S_1^2S_2S_3 + 2L_1L_2^2S_1S_2^3 - L_1L_2L_3S_1^3S_3 - 3L_1L_2L_3S_1^2S_2^2 - L_1L_2L_3S_1S_2S_3 - L_1L_2L_3S_2^3 - L_1L_2L_4S_1^3S_2 + L_1L_2L_4S_1^2S_3 + 5L_1L_2L_4S_1S_2^2 + L_1L_2L_4S_2S_3 + 4L_1L_2S_1S_2S_3^2 + 2L_1L_2S_2^3S_3 + L_1L_3^2S_1^3S_2 + L_1L_3^2S_1^2S_3 + L_1L_3^2S_1S_2^2 + L_1L_3L_4S_1^4 - 4L_1L_3L_4S_1^2S_2 - 2L_1L_3L_4S_1S_3 - L_1L_3L_4S_2^2 - L_1L_3S_1^2S_3^2 - 4L_1L_3S_1S_2^2S_3 + L_1L_3S_2^4 - 2L_1L_3S_2S_3^2 - L_1L_4^2S_1^3 + 3L_1L_4^2S_1S_2 + L_1L_4^2S_3 - L_1L_4S_1^2S_2S_3 - L_1L_4S_1S_2^3 + 3L_1L_4S_1S_3^2 + 5L_1L_4S_2^2S_3 + 3L_1S_2S_3^3 - L_2^3S_1^2S_2^2 + 2L_2^2L_3S_1^3S_2 + L_2^2L_3S_1S_2^2 - 2L_2^2L_4S_1^2S_2 - L_2^2L_4S_2^2 - L_2^2S_1^2S_3^2 - 2L_2^2S_1S_2^2S_3 - L_2L_3^2S_1^4 - 2L_2L_3^2S_1^2S_2 + 2L_2L_3L_4S_1^3 + 4L_2L_3L_4S_1S_2 + 5L_2L_3S_1^2S_2S_3 - L_2L_3S_1S_2^3 + L_2L_3S_1S_3^2 + L_2L_3S_2^2S_3 - L_2L_4^2S_1^2 - 2L_2L_4^2S_2 - 2L_2L_4S_1^3S_3 + 3L_2L_4S_1^2S_2^2 - 4L_2L_4S_1S_2S_3 - 2L_2L_4S_2^3 - L_2L_4S_3^2 - 2L_2S_1S_3^3 - L_2S_2^2S_3^2 + L_3^3S_1^3 - 3L_3^2L_4S_1^2 - L_3^2S_1^3S_3 - 3L_3^2S_1S_2S_3 + L_3^2S_2^3 + 3L_3L_4^2S_1 - L_3L_4S_1^3S_2 + 5L_3L_4S_1^2S_3 - L_3L_4S_1S_2^2 + 3L_3L_4S_2S_3 + 3L_3S_1S_2S_3^2 - L_3S_2^3S_3 + L_3S_3^3 - L_4^3 - L_4^2S_1^4 + 4L_4^2S_1^2S_2 - 4L_4^2S_1S_3 - 2L_4^2S_2^2 - 2L_4S_1^2S_3^2 + 4L_4S_1S_2^2S_3 - L_4S_2^4 - 4L_4S_2S_3^2 - S_3^4$.

Using the relation $A_1 + A_2 + A_3 + A_4 = 1$, the above expressions can be simplified to

$P_6 = -(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)(\lambda_1 - \lambda_4)(\lambda_2 - \lambda_4)(\lambda_3 - \lambda_4)^2A_3A_4(A_1 + A_2) - (\lambda_1 - \lambda_2)(\lambda_3 - \lambda_2)(\lambda_1 - \lambda_4)(\lambda_3 - \lambda_4)(\lambda_2 - \lambda_4)^2A_2A_4(A_1 + A_3) - (\lambda_1 - \lambda_2)(\lambda_4 - \lambda_2)(\lambda_1 - \lambda_3)(\lambda_4 - \lambda_3)(\lambda_2 - \lambda_3)^2A_2A_3(A_1 + A_4) - (\lambda_2 - \lambda_1)(\lambda_3 -$

$\lambda_1)(\lambda_2-\lambda_4)(\lambda_3-\lambda_4)(\lambda_1-\lambda_4)^2A_1A_4(A_2+A_3)-(\lambda_2-\lambda_1)(\lambda_4-\lambda_1)(\lambda_2-\lambda_3)(\lambda_4-\lambda_3)(\lambda_1-\lambda_3)^2A_1A_3(A_2+A_4)-$
$(\lambda_3-\lambda_1)(\lambda_4-\lambda_1)(\lambda_3-\lambda_2)(\lambda_4-\lambda_2)(\lambda_1-\lambda_2)^2A_1A_2(A_3+A_4)$,

$P_{12}=-A_1A_2A_3A_4(\lambda_1-\lambda_2)^2(\lambda_1-\lambda_3)^2(\lambda_1-\lambda_4)^2(\lambda_2-\lambda_3)^2(\lambda_2-\lambda_4)^2(\lambda_3-\lambda_4)^2$.

4.9 Symbolic solution to the inverse problem for the four state model with branching ($N=4$)

This model is described by the transitions $P_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} P_2 \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} P_4 \underset{k_3^+}{\overset{k_3^-}{\rightleftharpoons}} P_3$. In this case P_4 is the ON state. The model is of type II.

We have

$$\tilde{Q}=\begin{bmatrix} -k_1^+ & k_1^- & 0 & 0 \\ k_1^+ & -(k_2^+ + k_1^-) & 0 & k_2^- \\ 0 & 0 & k_3^+k_3^- & \\ 0 & k_2^+ & k_3^+ & -(k_4+k_3^-+k_2^-) \end{bmatrix}.$$

The Vieta formulas read

$$k_4k_1^+k_2^+k_3^+ = L_4 \tag{95}$$

$$k_4k_1^-k_3^+ + k_4k_1^+k_2^+ + k_4k_1^+k_3^+ + k_4k_2^+k_3^+ + k_1^-k_2^-k_3^+ + k_2^-k_1^+k_3^+ + k_3^-k_1^+k_2^+ +$$
$$+k_1^+k_2^+k_3^+ = -L_3 \tag{96}$$

$$k_4k_1^- + k_4k_1^+ + k_4k_2^+ + k_4k_3^+ + k_1^-k_2^- + k_1^-k_3^- + k_2^-k_1^+ + k_1^-k_3^+ + k_3^-k_1^+ + k_2^-k_3^+ + k_3^-k_2^+ +$$
$$+k_1^+k_2^+ + k_1^+k_3^+ + k_2^+k_3^+ = L_2 \tag{97}$$

$$k_4 + k_1^- + k_2^- + k_3^- + k_1^+ + k_2^+ + k_3^+ = -L_1 \tag{98}$$

The eigenvectors of $\tilde{Q}$ are

$$u_1 = \frac{k_1^-k_2^-}{k_1^+k_2^+ + k_1^-\lambda + k_1^+\lambda + k_2^+\lambda + \lambda^2} \tag{99}$$

$$u_2 = \frac{k_2^-(k_1^+ + \lambda)}{k_1^+k_2^+ + k_1^-\lambda + k_1^+\lambda + k_2^+\lambda + \lambda^2} \tag{100}$$

$$u_3 = \frac{k_3^-}{k_3^+ + \lambda} \tag{101}$$

$$u_4 = 1 \tag{102}$$

The system (30) has the solution

$$\begin{aligned}
C_1 &= -\frac{k_1^-\lambda_1^2+k_1^+\lambda_1^2+k_2^+\lambda_1^2+k_3^+\lambda_1^2+\lambda_1^3+k_1^+k_2^+k_3^++k_1^-k_3^+\lambda_1+k_1^+k_2^+\lambda_1+k_1^+k_3^+\lambda_1+k_2^+k_3^+\lambda_1}{(\lambda_1-\lambda_2)(\lambda_1\lambda_3+\lambda_1\lambda_4-\lambda_3\lambda_4-\lambda_1^2)}\\
C_2 &= -\frac{(k_3^++\lambda_2)(k_1^+k_2^++k_1^-\lambda_2+k_1^+\lambda_2+k_2^+\lambda_2+\lambda_2^2)}{(\lambda_1-\lambda_2)(\lambda_2-\lambda_3)(\lambda_2-\lambda_4)}\\
C_3 &= \frac{(k_3^++\lambda_3)(k_1^+k_2^++k_1^-\lambda_3+k_1^+\lambda_3+k_2^+\lambda_3+\lambda_3^2)}{(\lambda_3-\lambda_4)(\lambda_1\lambda_2-\lambda_1\lambda_3-\lambda_2\lambda_3+\lambda_3^2)}\\
C_4 &= -\frac{(k_3^++\lambda_4)(k_1^+k_2^++k_1^-\lambda_4+k_1^+\lambda_4+k_2^+\lambda_4+\lambda_4^2)}{\lambda_1\lambda_4^2+\lambda_2\lambda_4^2+\lambda_3\lambda_4^2-\lambda_4^3+\lambda_1\lambda_2\lambda_3-\lambda_1\lambda_2\lambda_4-\lambda_1\lambda_3\lambda_4-\lambda_2\lambda_3\lambda_4}
\end{aligned} \tag{103}$$

The eqs. (29) and (31) provide

$$k_4 = -S_1$$

$$k_3^- = \frac{-S_1(k_3^+)^3-S_1(L_1-S_1)(k_3^+)^2-S_1(L_2+S_2-L_1S_1)k_3^+-S_1(L_3-S_3+L_1S_2-L_2S_1)}{3S_1(k_3^+)^2+(2L_1S_1-2S_2)k_3^++S_3-L_1S_2+L_2S_1},$$

$k_2^+ = (S_1^3(k_3^+)^4+(-S_1^4+L_1S_1^3)(k_3^+)^3+(-L_1S_1^4+S_1^3S_2+L_2S_1^3+3S_3S_1^2-3S_1S_2^2)(k_3^+)^2+(L_3S_1^3-L_2S_1^4-S_1^3S_3+2S_2^3-2S_1S_2S_3-2L_1S_1S_2^2+2L_1S_1^2S_3+L_1S_1^3S_2)k_3^++L_2S_1^2S_3-L_2S_1S_2^2-L_1S_1S_2S_3+S_1S_3^2+L_1S_2^3-S_2^2S_3)/(S_1^3(k_3^+)^3+(S_1(2S_1^3+L_1S_1^2-3S_2S_1))(k_3^+)^2+(S_1(L_1S_1^3-S_1^2S_2+L_2S_1^2-2L_1S_1S_2+2S_2^2))k_3^+-S_1(-L_3S_1^2+L_2S_1S_2-L_1S_2^2+S_3S_2)),$

$k_2^- = (S_1^2(k_3^+)^3+(2S_1^3+L_1S_1^2-3S_2S_1)(k_3^+)^2+(L_1S_1^3-S_1^2S_2+L_2S_1^2-2L_1S_1S_2+2S_2^2)k_3^++L_3S_1^2-L_2S_1S_2+L_1S_2^2-S_3S_2)/((3S_1^2)(k_3^+)^2+(-S_1(2S_2-2L_1S_1))k_3^++S_1(S_3-L_1S_2+L_2S_1)),$

$k_1^+ = (S_1^3(k_3^+)^5+(S_1(2S_1^3+L_1S_1^2-3S_2S_1)-S_1^2(S_2-L_1S_1))(k_3^+)^4+(S_1(L_1S_1^3-S_1^2S_2+L_2S_1^2-2L_1S_1S_2+2S_2^2)+S_1^2(S_3-L_1S_2+L_2S_1)-(S_2-L_1S_1)(2S_1^3+L_1S_1^2-3S_2S_1))(k_3^+)^3+((S_3-L_1S_2+L_2S_1)(2S_1^3+L_1S_1^2-3S_2S_1)-S_1(-L_3S_1^2+L_2S_1S_2-L_1S_2^2+S_3S_2)-(S_2-L_1S_1)(L_1S_1^3-S_1^2S_2+L_2S_1^2-2L_1S_1S_2+2S_2^2))(k_3^+)^2+((S_2-L_1S_1)(-L_3S_1^2+L_2S_1S_2-L_1S_2^2+S_3S_2)+(S_3-L_1S_2+L_2S_1)(L_1S_1^3-S_1^2S_2+L_2S_1^2-2L_1S_1S_2+2S_2^2))k_3^+-(S_3-L_1S_2+L_2S_1)(-L_3S_1^2+L_2S_1S_2-L_1S_2^2+S_3S_2)/(S_1^3(k_3^+)^4+(-S_1^4+L_1S_1^3)(k_3^+)^3+(-L_1S_1^4+S_1^3S_2+L_2S_1^3+3S_3S_1^2-3S_1S_2^2)(k_3^+)^2+(L_3S_1^3-L_2S_1^4-S_1^3S_3+2S_2^3-2S_1S_2S_3-2L_1S_1S_2^2+2L_1S_1^2S_3+L_1S_1^3S_2)k_3^++L_2S_1^2S_3-L_2S_1S_2^2-L_1S_1S_2S_3+S_1S_3^2+L_1S_2^3-S_2^2S_3),$

$k_1^- = -(S_1(S_3-2S_2k_3^++3S_1(k_3^+)^2-L_1S_2+L_2S_1+2L_1S_1k_3^+)(L_1^2S_1^4(k_3^+)^3-L_1^2S_1^3S_2(k_3^+)^2+L_1^2S_1^3S_3k_3^++L_1^2S_1^3(k_3^+)^4-3L_1^2S_1^2S_2(k_3^+)^3+L_1^2S_1^2S_3(k_3^+)^2+3L_1^2S_1S_2^2(k_3^+)^2-2L_1^2S_1S_2S_3k_3^+-L_1^2S_2^3k_3^++L_1^2S_2^2S_3+L_1L_2S_1^4(k_3^+)^2-L_1L_2S_1^3S_2k_3^++2L_1L_2S_1^3(k_3^+)^3-5L_1L_2S_1^2S_2(k_3^+)^2+L_1L_2S_1^2S_3k_3^++4L_1L_2S_1S_2^2k_3^+-L_1L_2S_1S_2S_3-L_1L_2S_2^3+L_1L_3S_1^4k_3^++2L_1L_3S_1^3(k_3^+)^2-3L_1L_3S_1^2S_2k_3^++L_1L_3S_1^2S_3+L_1L_3S_1S_2^2+L_1S_1^5(k_3^+)^3-L_1S_1^4S_2(k_3^+)^2+2L_1S_1^4(k_3^+)^4-$

$5L_1S_1^3S_2(k_3^+)^3 + L_1S_1^3S_3(k_3^+)^2 + 2L_1S_1^3(k_3^+)^5 + 4L_1S_1^2S_2^2(k_3^+)^2 - L_1S_1^2S_2S_3k_3^+ - 6L_1S_1^2S_2(k_3^+)^4 + 4L_1S_1^2S_3(k_3^+)^3 - L_1S_1S_2^3k_3^+ + 6L_1S_1S_2^2(k_3^+)^3 - 8L_1S_1S_2S_3(k_3^+)^2 + 2L_1S_1S_3^2k_3^+ - 2L_1S_2^3(k_3^+)^2 + 4L_1S_2^2S_3k_3^+ - 2L_1S_2S_3^2 + L_2^2S_1^3(k_3^+)^2 - 2L_2^2S_1^2S_2k_3^+ + L_2^2S_1S_2^2 + 2L_2L_3S_1^3k_3^+ - 2L_2L_3S_1^2S_2 + L_2S_1^5(k_3^+)^2 + L_2S_1^4(k_3^+)^3 - 4L_2S_1^3S_2(k_3^+)^2 - 2L_2S_1^3S_3k_3^+ + 2L_2S_1^3(k_3^+)^4 + 3L_2S_1^2S_2^2k_3^+ - 5L_2S_1^2S_2(k_3^+)^3 + 3L_2S_1^2S_3(k_3^+)^2 + 4L_2S_1S_2^2(k_3^+)^2 - 4L_2S_1S_2S_3k_3^+ + L_2S_1S_3^2 - L_2S_2^3k_3^+ + L_2S_2^2S_3 + L_3^2S_1^3 + L_3S_1^4(k_3^+)^2 - L_3S_1^3S_2k_3^+ + 2L_3S_1^3(k_3^+)^3 - 3L_3S_1^2S_2(k_3^+)^2 + 3L_3S_1^2S_3k_3^+ - 3L_3S_1S_2S_3 + L_3S_2^3 + S_1^5(k_3^+)^4 - S_1^4S_2(k_3^+)^3 + S_1^4S_3(k_3^+)^2 + S_1^4(k_3^+)^5 - 4S_1^3S_2(k_3^+)^4 + S_1^3(k_3^+)^6 + 4S_1^2S_2^2(k_3^+)^3 - 4S_1^2S_2S_3(k_3^+)^2 - 3S_1^2S_2(k_3^+)^5 - 2S_1^2S_3^2k_3^+ + 3S_1^2S_3(k_3^+)^4 + 4S_1S_2^2S_3k_3^+ + 3S_1S_2^2(k_3^+)^4 - 6S_1S_2S_3(k_3^+)^3 + 3S_1S_3^2(k_3^+)^2 - S_2^4k_3^+ - S_2^3(k_3^+)^3 + 3S_2^2S_3(k_3^+)^2 - 3S_2S_3^2k_3^+ + S_3^3))/((2S_1^3(k_3^+)^2 + L_1S_1^3k_3^+ - S_1^2S_2k_3^+ + S_1^2(k_3^+)^3 + L_1S_1^2(k_3^+)^2 + L_2S_1^2k_3^+ + L_3S_1^2 - 3S_1S_2(k_3^+)^2 - 2L_1S_1S_2k_3^+ - L_2S_1S_2 + 2S_2^2k_3^+ + L_1S_2^2 - S_3S_2)(S_1^3S_2(k_3^+)^2 - L_1S_1^4(k_3^+)^2 - L_2S_1^4k_3^+ - S_1^4(k_3^+)^3 + L_1S_1^3S_2k_3^+ - S_1^3S_3k_3^+ + S_1^3(k_3^+)^4 + L_1S_1^3(k_3^+)^3 + L_2S_1^3(k_3^+)^2 + L_3S_1^3k_3^+ + 3S_1^2S_3(k_3^+)^2 + 2L_1S_1^2S_3k_3^+ + L_2S_1^2S_3 - 3S_1S_2^2(k_3^+)^2 - 2L_1S_1S_2^2k_3^+ - L_2S_1S_2^2 - 2S_1S_2S_3k_3^+ - L_1S_1S_2S_3 + S_1S_3^2 + 2S_2^3k_3^+ + L_1S_2^3 - S_2^2S_3))$,

where k_3^+ is the solution of the cubic equation

$$S_1(k_3^+)^3 + (L_1S_1 - S_2)(k_3^+)^2 + (L_2S_1 - L_1S_2 + S_3)k_3^+L_4 = 0. \tag{104}$$

The equation (104) has the discriminant

$$\begin{aligned}\Delta = {} & (S_2 - L_1S_1)^2(S_3 - L_1S_2 + L_2S_1)^2 + 4L_4(S_2 - L_1S_1)^3 \\ & - 27L_4^2S_1^2 - 4S_1(S_3 - L_1S_2 + L_2S_1)^3 - 18L_4S_1(S_2 - L_1S_1)(S_3 - L_1S_2 + L_2S_1).\end{aligned} \tag{105}$$

When $\Delta < 0$, there is an unique real solution

$$k_3^+ = \frac{S_2 - L_1S_1}{3S_1} + (\Delta_3)^{1/3} + \frac{\frac{(S_2-L_1S_1)^2}{9S_1^2} - \frac{S_3-L_1S_2+L_2S_1}{3S_1}}{\Delta_3^{1/3}}, \tag{106}$$

where $\Delta_3 = \frac{(S_2-L_1S_1)^3}{27S1^3} - \frac{L_4}{2S_1} - \frac{(S_2-L_1S_1)(S_3-L_1S_2+L_2S_1)}{6S1^2} + \frac{\sqrt{-\Delta/108}}{S_1^2}$.

4.10 Symbolic solution to the inverse problem for four state chain with return in state P_3 ($N = 4$)

This model has exactly the same transitions as the 4 state chain model described in the Section 4.8 with the difference that the ON state is P_3.

Using the same methods as in Section 4.6 we show that in this case

$$A_1\lambda_1 + A_2\lambda_2 + A_3\lambda_3 + A_4\lambda_4 = 0. \tag{107}$$

Using (107) and $A_1 + A_2 + A_3 + A_4 = 1$ we can compute the remaining parameters as

$$\begin{aligned} A_3 &= -\frac{\lambda_4 + A_1(\lambda_1 - \lambda_4) + A_2(\lambda_2 - \lambda_4)}{\lambda_3 - \lambda_4}, \\ A_4 &= \frac{\lambda_3 + A_1(\lambda_1 - \lambda_3) + A_2(\lambda_2 - \lambda_3)}{\lambda_3 - \lambda_4}. \end{aligned} \tag{108}$$

If the condition (107) is not satisfied, then there is no solution to the inverse problem.

If the condition (107) is satisfied, then the inverse problem is not well posed and has an infinity of solutions. In this case, all the solutions can be expressed as functions of a free parameter. In the sequel we will choose k_4 as free parameter. Although we were able to obtain analytic solutions, these are too long to be displayed.

The following, simple relations are useful for the anaysis of this model:

$$\begin{aligned} k_3^+ &= -S_2/k_4, \\ k_3^- + k_2^- &= -S_3/S_2 + S_2/k_4 - k_4 \end{aligned} \tag{109}$$

5 Uncertainty estimation for the model parameters

In Section 3.3 we have used optimization with multiple initial parameters to estimate confidence intervals for each parameter of the multi-exponential survival function as lower and upper bounds of optimal and sub-optimal parameters. These intervals are presented as $A_i \in [A_i^{min}, A_i^{max}]$, $1 \leq i \leq N$ and $\lambda_i \in [\lambda_i^{min}, \lambda_i^{max}]$, $1 \leq i \leq N$.

In the sections above we have shown how to compute symbolically the kinetic parameters of various models from the parameters A_i, λ_i, $1 \leq i \leq N$ of the multi-exponential survival function. By applying the symbolic mapping to the confidence intervals $[A_i^{min}, A_i^{max}]$,$[\lambda_i^{min}, \lambda_i^{max}]$ one can get the confidence intervals of the kinetic parameters. However, finding intervals that bound the kinetic parameters from the confidence intervals of the survival function parameters is a non-convex optimization problem with constraints which may prove difficult. Therefore, in the current implementation of our software we decided to apply the symbolic mapping directly to the entire set of optimal and sub-optimal survival function parameters obtained in Section 3.3 and compute the lower and upper bounds of the resulting kinetic parameters.

6 Computing the mean mRNA at the steady state

The statistics of the waiting time between two successive transcription initiations can be used to compute the statistics of the number of mRNA molecules. Each elongating polymerase will generate one molecule of mRNA

that will survive in the average a time $T \approx 45\,\text{min}$. Therefore the mean mRNA number at the steady state is simply:

$$mRNA[min] = 45/w, \tag{110}$$

where w is the average waiting time.

The straightforward calculation

$$w = -\int_0^\infty tS'(t)\,dt = \int_0^\infty S(t)\,dt = -\sum_{i=1}^{N} \frac{A_i}{\lambda_i},$$

leads to two equivalent ways to compute the mean mRNA number, from the area under curve, or from the parameters of the survival function

$$mRNA[min] = 45/AUC = -45/\sum_{i=1}^{N} \frac{A_i}{\lambda_i}, \tag{111}$$

where AUC is the area under curve of the survival function.

7 Computing the probability of each state at stationarity

Although the computation of the distribution of waiting times does not require stationarity conditions (successive waiting times form a renewal process even without stationarity, as soon and as long as the model parameters are constant in time) it is usefull to have estimates for the stationary probabilities of being in each of the model's state. The sojourn time in the state P_{N+1} being nil the probability of being in this state is also nil. The remaining N probabilities $p_i = \mathbb{P}[M(t) = P_i]\,, 1 \leq i \leq N$ satisfy $p_1 + p_2 + \ldots + p_N = 1$ and the following homogeneous system of linear equations:

$$\tilde{\tilde{\boldsymbol{Q}}} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix} = 0, \tag{112}$$

where $\tilde{\tilde{\boldsymbol{Q}}}$ is obtained from $\tilde{\boldsymbol{Q}}$ by setting $k_{N,N+1} = 0$.

A few examples follow.

For the model M1,

$$\tilde{\tilde{\boldsymbol{Q}}} = \begin{bmatrix} -k_1^+ & k_1^- & 0 \\ k_1^+ & -(k_2^+ + k_1^-) & k_2^- \\ 0 & k_2^+ & -k_2^- \end{bmatrix},$$

$$p_1 = \frac{k_1^- k_2^-}{k_1^+ k_2^- + k_1^- k_2^- + k_1^+ k_2^+}, \tag{113}$$

$$p_2 = \frac{k_1^+ k_2^-}{k_1^+ k_2^- + k_1^- k_2^- + k_1^+ k_2^+}, \tag{114}$$

$$p_3 = \frac{k_1^+ k_2^+}{k_1^+ k_2^- + k_1^- k_2^- + k_1^+ k_2^+}. \tag{115}$$

For the model M2,

$$\tilde{\tilde{\boldsymbol{Q}}} = \begin{bmatrix} -k_1^+ & 0 & k_1^- \\ 0 & -k_2^+ & k_2^- \\ k_1^+ & k_2^+ & -(k_1^- + k_2^-) \end{bmatrix},$$

$$p_1 = \frac{k_1^- k_2^+}{k_1^+ k_2^+ + k_1^- k_2^+ + k_1^+ k_2^-}, \tag{116}$$

$$p_2 = \frac{k_1^+ k_2^-}{k_1^+ k_2^+ + k_1^- k_2^+ + k_1^+ k_2^-}, \tag{117}$$

$$p_3 = \frac{k_1^+ k_2^+}{k_1^+ k_2^+ + k_1^- k_2^+ + k_1^+ k_2^-}. \tag{118}$$

For the four states model,

$$\tilde{\tilde{\boldsymbol{Q}}} = \begin{bmatrix} -k_1^+ & k_1^- & 0 & 0 \\ k_1^+ & -(k_2^+ + k_1^-) & k_2^- & 0 \\ 0 & k_2^+ & -(k_3^+ + k_2^-) & k_3^- \\ 0 & 0 & k_3^+ & -k_3^- \end{bmatrix},$$

$$p_1 = \frac{k_1^- k_2^- k_3^-}{k_1^+ k_2^+ k_3^+ + k_1^+ k_2^+ k_3^- + k_1^+ k_2^- k_3^- + k_1^- k_2^- k_3^-}, \tag{119}$$

$$p_2 = \frac{k_1^+ k_2^- k_3^-}{k_1^+ k_2^+ k_3^+ + k_1^+ k_2^+ k_3^- + k_1^+ k_2^- k_3^- + k_1^- k_2^- k_3^-}, \tag{120}$$

$$p_3 = \frac{k_1^+ k_2^+ k_3^-}{k_1^+ k_2^+ k_3^+ + k_1^+ k_2^+ k_3^- + k_1^+ k_2^- k_3^- + k_1^- k_2^- k_3^-}, \tag{121}$$

$$p_1 = \frac{k_1^+ k_2^+ k_3^+}{k_1^+ k_2^+ k_3^+ + k_1^+ k_2^+ k_3^- + k_1^+ k_2^- k_3^- + k_1^- k_2^- k_3^-}. \tag{122}$$

For the two states (ON-OFF) model

$$\tilde{\tilde{\boldsymbol{Q}}} = \begin{bmatrix} -k_1^+ & k_1^- \\ k_1^+ & -k_1^- \end{bmatrix},$$

$$p_1 = \frac{k_1^-}{k_1^+ + k_1^-}, \tag{123}$$

$$p_2 = \frac{k_1^+}{k_1^+ + k_1^-}. \tag{124}$$

8 Testing the robustness of the method using artificial data

The numerical method is based on the assumption that the instrumental noise and other sources of noise are averaged out by the algorithm and therefore can be neglected. In this subsection we use artificial data to test the consequences of releasing this assumption. Furthermore, the optimization algorithm is stochastic and include approximate steps such as the estimation of the parameters p_s and p_l, and errors resulting from the analog to digital conversion of the long movie signals. Artificially generated data with well know parameters will also allow us to test the fidelity of the parameter identification in our method.

Artificial data was generated by simulating the model M1 using the Gillespie algorithm. We use three parameter sets, similar to those identified from data in the three experimental conditions (previous subsection). The simulations generate artificial polymerase positions from which we first compute a noiseless signal using Eq. (1).

In a second step we add to the signal a centered Gaussian noise, whose variance is similar to the one in data, as follows

$$S_\eta(t) = S(t) + \eta(t), \tag{125}$$

where $S(t)$ is the noiseless signal and $\eta(t)$ is the noise.

The noise estimate is obtained from the short movies data. It is defined as the difference between the raw signal and the signal reconstructed by deconvolution (computed using Eq. (1)). We found that the noise variance is an increasing function of the signal amplitude. By using cubic polynomial interpolation we have derived analytic formulas for the variance in the three experimental conditions:

$$Var(\eta) = b_3 S^3 + b_2 S^2 + b_1 S + b_0, \tag{126}$$

where b_i, $0 \leq i \leq 3$ are parameters whose values can be found in the Table 1.

We applied our algorithm to a raw signal described by (125) and obtained estimates of the kinetic parameters. η is defined by (126) and Table 1. Together with η we have also tested the double 2η and four times 4η noise amplitude. These estimates were compared to the know values of the parameters that were used for simulating the artificial data. The result of the comparison is shown in Fig.13.

Data set	b_0	b_1	b_2	b_3
Low tat	0.27	0.026	0.0022	-1.5e-5
No tat	-0.97	0.23	-0.0021	9.8e-6
High tat	0.27	0.026	0.0022	-1.5e-5

Table 1: Noise parameters for various experimental conditions in the study of the HIV-1 promoter.

The method faithfully retrieves the parameter values, at least for noise amplitudes comparable to the ones determined from the data used in this study. For larger noise amplitudes some parameters may not be faithfully retrieved. As expected, some large kinetic parameters, corresponding to small time scales are not faithfully retrieved. However, the small parameters, corresponding to large time scales are faithfully retrieved even for large noise amplitudes. This proves the robustness of the method with respect to noise.

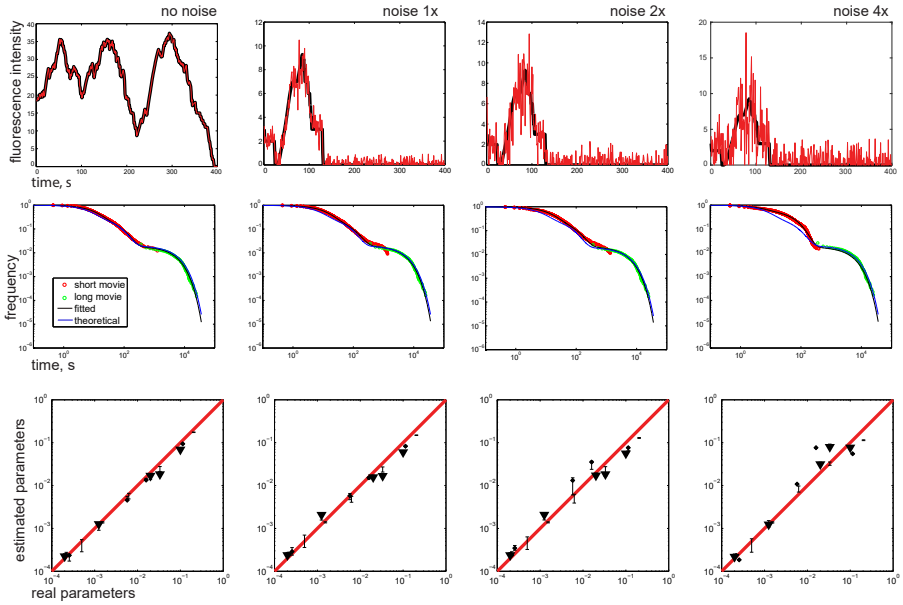


Figure 13: Testing the algorithm with artificial data for various noise amplitude. $\times 1$ represents artificial data with the same amplitude of noise as the real data. $\times 0$ is the noiseless artificial data. For each comparison we consider 3 sets of 5 parameters corresponding to the three experimental conditions in the HIV-1 data: no tat, low tat and high tat. The survival functions (middle row) and the artificial signal (upper row) are shown only for the no-tat conditions.

9 Results.

9.1 Identifying the parameters of the model M2.

Model M2 corresponds to the stochastic, facultative pausing (Figure 1). The model parameters can be identified from a unconstrained three-exponential fit (Figure 12).

The results of the fit are presented in the Table 2 and in the Figure 14.

Type	OBJ	mRNA	k_1^+	k_1^-	k_2^+	k_2^-	k_3
no tat optimal	0.028	16.7	6.0e-05	0.00035	0.00089	0.003	0.063
min		16.7	6.0e-05	0.00021	0.00089	0.00199	0.06
max		29.5	7.1e-05	0.00035	0.00130	0.003	0.063
low tat optimal	0.061	49.6	0.00015	0.00031	0.0012	0.0028	0.1
min		49.6	0.00015	0.00021	0.0012	0.0021	0.099
max		114	0.00022	0.00100	0.028	0.0180	0.15
high tat optimal	0.115	315	0.0015	4.9e-05	0.0100	0.0043	0.17
min		265	0.0014	4.9e-05	0.0052	0.003	0.16
max		315	0.0015	6.3e-05	0.0100	0.0043	0.17

Table 2: Results of the unconstrained three-exponential fit of the model M2. $\alpha = 0.30$

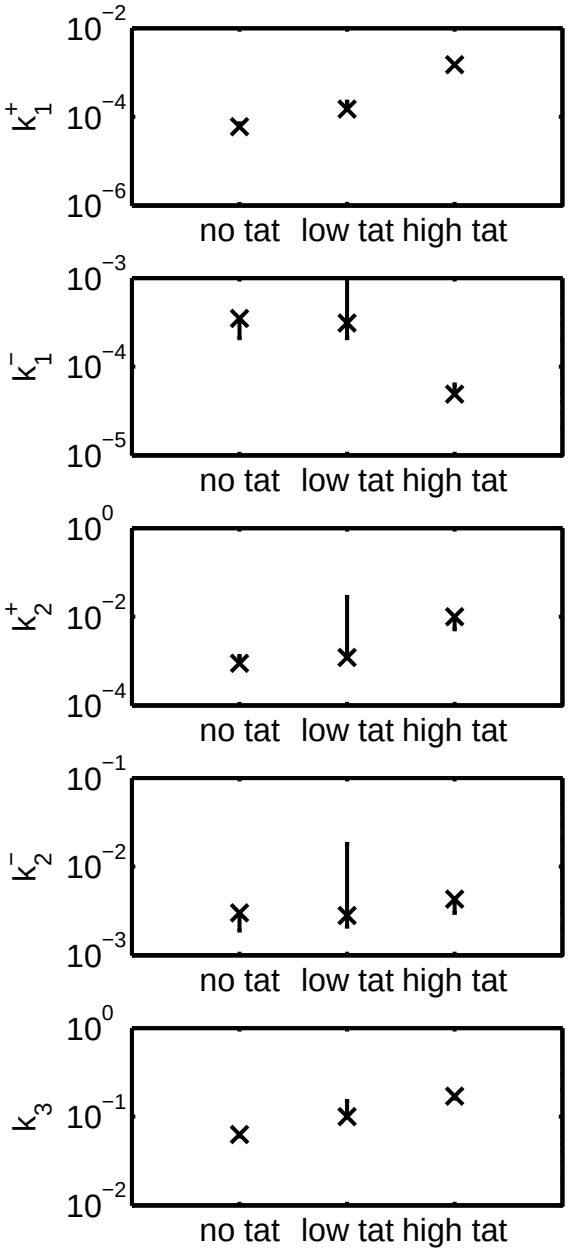


Figure 14: Results of the unconstrained three-exponential fit of the model M2. Parameter dependence on the experimental conditions for $\alpha = 0.30$. The vertical bars are uncertainty intervals.

9.2 Identifying the parameters of the two states ON-OFF model.

In order to identify this model we use a two exponential fit of the survival function $S(t) = A_1 \exp(\lambda_1 t) + A_2 \exp(\lambda_2 t)$. The model parameters are computed from the survival function parameters according to the Section 4.7.

The result of the fit is given in the Table 3 and in the Figure 15. The large values of the objective function suggest that this model is not suitable for our data.

Type	OBJ	λ_1	λ_2	A_1	A_2	$k2$	k_1^-	k_1^+	mRNA
no tat optimal	0.22046	-0.0478	-0.000141	0.984	0.0159	0.047	0.000755	0.000144	20.3
min		-0.0478	-0.000141	0.984	0.00646	0.0444	0.000287	0.000136	
max		-0.0446	-0.000135	0.994	0.0159	0.047	0.000755	0.000144	
low tat optimal	0.273	-0.0782	-0.00025	0.991	0.00943	0.0774	0.000733	0.000252	53.5
min		-0.0782	-0.00025	0.991	0.00368	0.0719	0.000264	0.000244	
max		-0.0721	-0.000243	0.996	0.00943	0.0774	0.000733	0.000252	
high tat optimal	0.64799	-0.12	-0.00222	0.996	0.00422	0.12	0.000488	0.00223	264.8
min		-0.12	-0.00222	0.00278	0.00113	0.097	0.000105	0.00218	
max		-0.0971	-0.00218	0.999	0.997	0.12	0.000488	0.00223	

Table 3: Results of the two-exponential fit, $\alpha = 0.30$. The objective function has large values compared to the three state model M_2, for the same value of α.

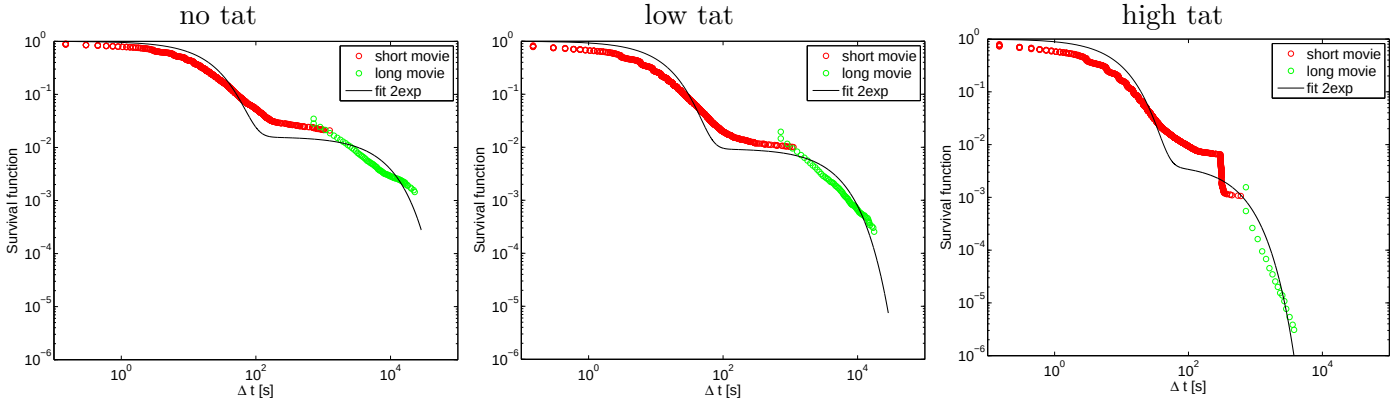


Figure 15: Results of the two-exponential fit: most optimal fit for $\alpha = 0.30$.

9.3 Identifying the parameters of the model M3

Model M3 corresponds to obligatory pausing (see Figure 1) and is identified using the constrained three-exponential fit described in the Section 4.6. However, even if (up to errors) the three-exponential fit provides a single best fit, the set of corresponding parameters of model M3 is a curve in the 5D space of parameters. The inverse problem for model M3 is not well posed as the relation between the parameters of the model M3 and the parameters of the three-exponential fit is many to one. The result of the constrained three-exponential fit is given in the Table 4.

The dependence of the parameters of the model M3 on the undetermined parameter k_3 is shown in the Figure 17. The parameters $k_2^{\pm}$ have very large values compared to all other parameters. The model $M3$ is in this case equivalent to the two states ON-OFF model and inherits the difficulty of this model to fit the data.

Type	OBJ	λ_1	λ_2	λ_3	A_1	A_2	A_3	mRNA
no tat optimal	0.22	-5660	-0.0478	-0.000141	-8.31e-06	0.984	0.0159	20.3
min		-7380	-0.0708	-0.000169	-0.00232	0.984	0.0060	
max		-2	-0.029	-0.000121	-6e-6	1.02	0.0161	
low tat optimal	0.27	-8480	-0.0782	-0.00025	-9.13e-06	0.991	0.00943	53.5
min		-12500	-0.148	-0.000278	-68.3	0.99	0.00306	
max		-0.144	-0.0556	-0.000224	-5.27e-06	69.3	0.0125	
high tat optimal	0.65	-26000	-0.12	-0.00222	-4.59e-06	0.996	0.00422	264.8
min		-54200	-0.166	-0.00239	-1.49	0.994	0.000797	
max		-0.254	-0.0692	-0.00202	-1.74e-06	2.48	0.00647	

Table 4: Results of the constrained three-exponential fit of the model M3, $\alpha = 0.30$. The objective function has large values (compared to different models and for the same α) and the fitted parameters are very uncertain.

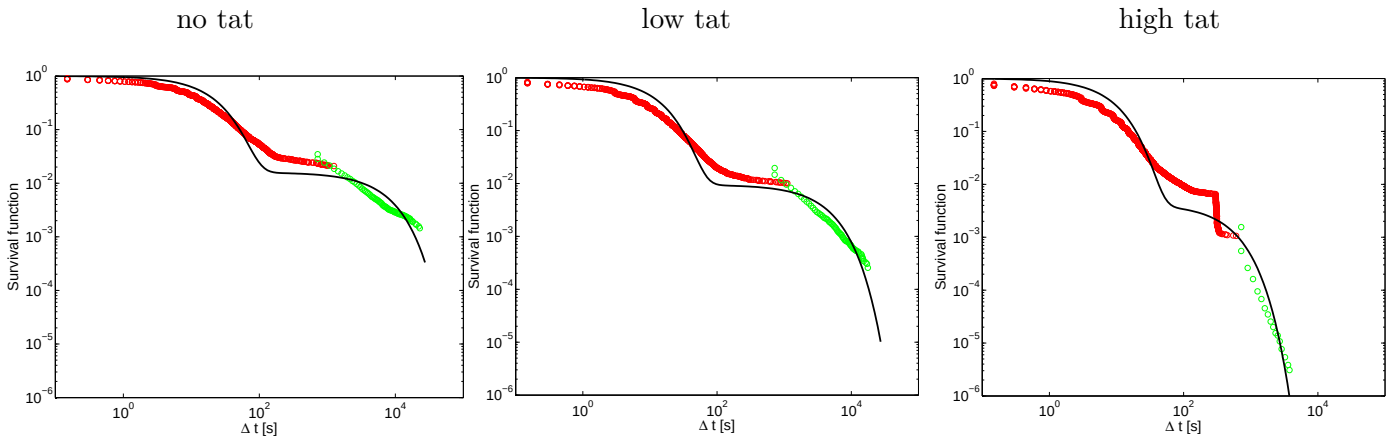


Figure 16: Results of the constrained three-exponential fit: most optimal fit for $\alpha = 0.30$.

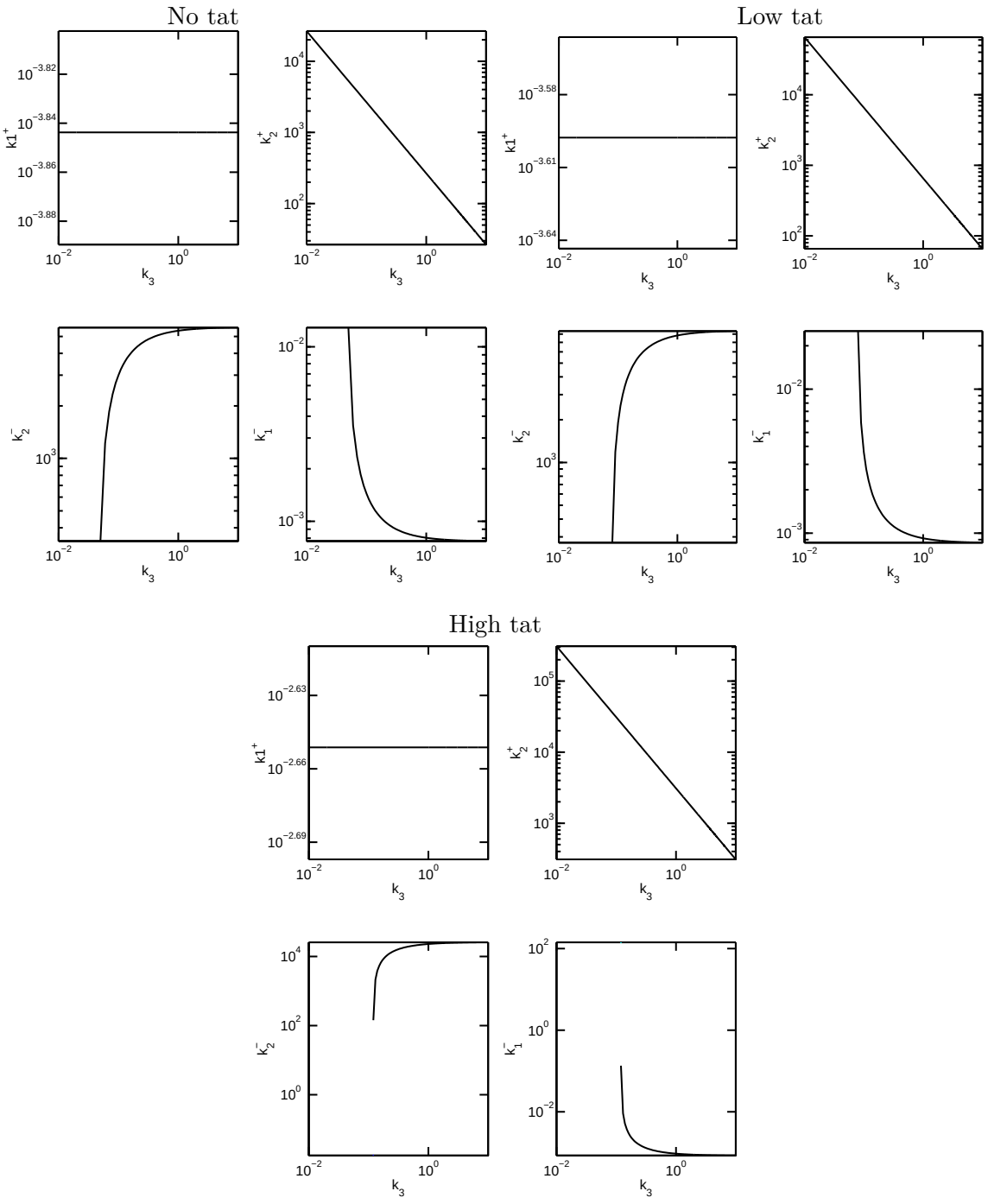


Figure 17: Results of the constrained three-exponential fit of the model M3. Parameter dependence on the undetermined parameter k_3 (pause exit rate) for $\alpha = 0.30$. The parameters $k_2^{\pm}$ have very large values compared to all other parameters and correspond to very fast processes (timescales smaller than $0.01s$). For such parameters the model $M3$ is equivalent to a two states ON-OFF model (the states ON and PAUSE can be pooled with no information loss in the model M3). In order to ensure positivity of kinetic parameters, one needs $k_3 > 0.1s^{-1}$.

9.4 Identifying the parameters of a four states model with pausing.

In order to identify four states models we use a four exponential fit of the survival function $S(t) = A_1 \exp(\lambda_1 t) + A_2 \exp(\lambda_2 t) + A_3 \exp(\lambda_3 t) + A_4 \exp(\lambda_4 t)$, where $A_1 + A_2 + A_3 + A_4 = 1$. Let us consider that $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < 0$. From $S'(t) \leq 0$ it follows $\lambda_1 A_1 + \lambda_2 A_2 + \lambda_3 A_3 + \lambda_4 A_4 \leq 0$, $A_4 \geq 0$.

The model M_4 is obtained by adding one more OFF state to the model M_3 (see Figure 18). It corresponds to the theoretical model described in the Section 4.10. The parameters of this model can be obtained from a constrained four exponential fit with six free parameters λ_1,λ_2,λ_3,λ_4,A_1,A_2 (see Eq.(108)). Although this model has more free parameters than the model M_2, the fit quality is lower.

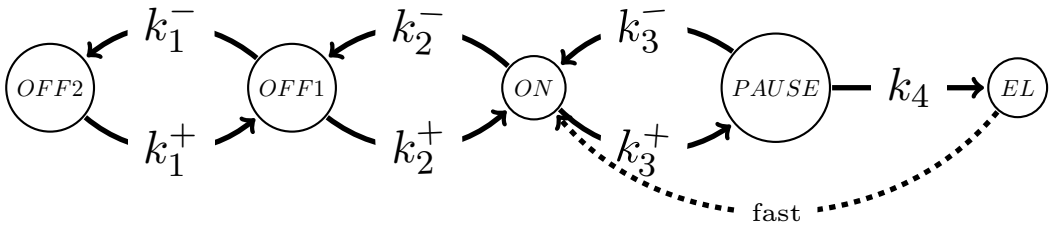


Figure 18: Model M4 with two OFF states and obligatory pausing. k_4 is the pause exit rate, k_3^- is the transcription abortion rate.

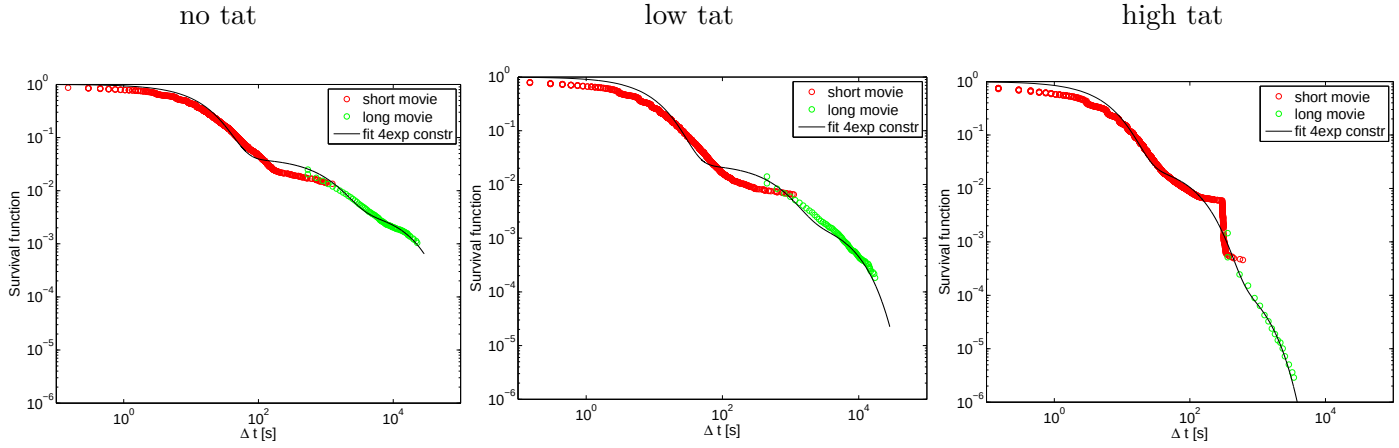


Figure 19: Results of the constrained four-exponential fit of the model M4: most optimal fit for $\alpha = 0.30$.

Type	OBJ	λ_1	λ_2	λ_3	λ_4	A_1	A_2	A_3	A_4
no tat optimal	0.036875	-8430	-0.0636	-0.00103	-6.6e-05	-7.23e-06	0.958	0.0373	0.0043
min		-19700	-0.087	-0.0121	-0.00011	-0.0698	0.869	0.0261	0.00306
max		-1.02	-0.0597	-0.00103	-6.6e-05	-3.06e-06	1.02	0.126	0.00477
low tat optimal	0.078907	-5550	-0.102	-0.0018	-0.000161	-1.79e-05	0.976	0.0221	0.00226
min		-36600	-1.74	-0.0487	-0.000228	-335	0.716	0.0198	0.00212
max		-0.202	-0.0975	-0.00157	-0.000156	-2.97e-06	336	0.453	0.00525
high tat optimal	0.11601	-9330	-0.174	-0.0101	-0.00148	-1.82e-05	0.972	0.0281	0.000288
min		-85200	-0.178	-0.0126	-0.00149	-0.0442	0.968	0.0218	0.000258
max		-4.1	-0.166	-0.00681	-0.00144	-1.91e-06	1.01	0.0323	0.000349

Table 5: Results of the constrained four-exponential fit of the model M4, $\alpha = 0.3$. The objective function shows that the fit is not better than the one of the model M2, for the same α and the fitted parameters are very uncertain.

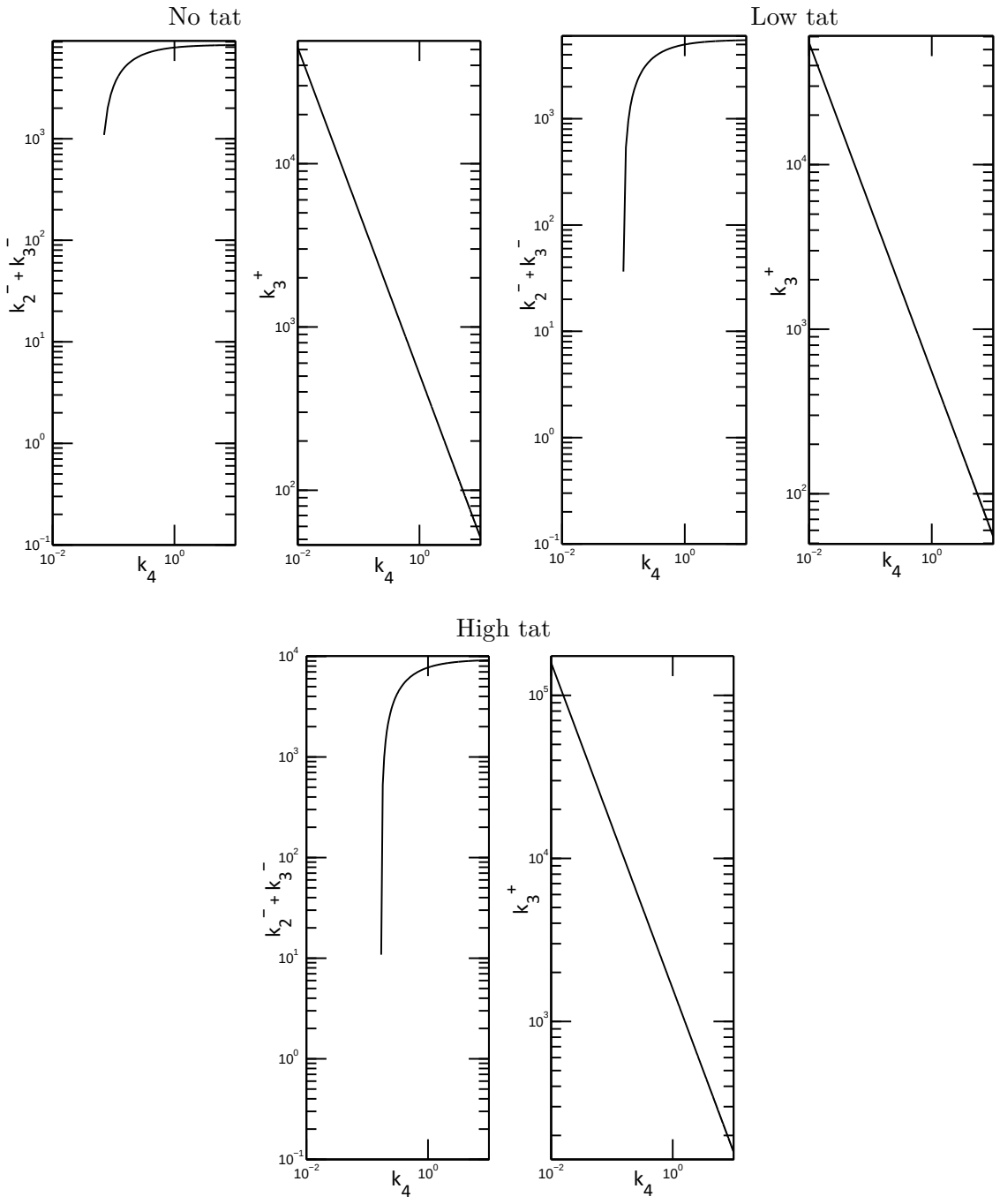


Figure 20: Results of the constrained four-exponential fit of the model M4. Parameter dependence on the undetermined parameter k_4 (pause exit rate) for $\alpha = 0.30$. The parameters k_2^m and $k_3^{\pm}$ have very large values compared to other parameters and correspond to very rapid processes (timescales smaller than $0.1s$). With such parameters the model M4 is equivalent to a three states model with the states ON and PAUSE pooled. For positivity of kinetic parameters one needs $k_4 > 0.1s^{-1}$.

References

[1] A. M. Corrigan, E. Tunnacliffe, D. Cannon, and J. R. Chubb. A continuum model of transcriptional bursting. *Elife*, 5:e13051, 2016.

[2] A. Coulon and D. R. Larson. Fluctuation analysis: dissecting transcriptional kinetics with signal theory. In *Methods in enzymology*, volume 572, pages 159–191. Elsevier, 2016.

[3] M. Dejean, V. L. Pimmet, C. Fernandez, A. Trullo, E. Bertrand, O. Radulescu, and M. L. Lagha. Quantitative imaging of transcription in living *Drosophila* embryos reveals the impact of core promoter motifs on promoter state dynamics. *preprint*, 2020.

[4] J. Desponds, H. Tran, T. Ferraro, T. Lucas, C. P. Romero, A. Guillou, C. Fradin, M. Coppey, N. Dostatni, and A. M. Walczak. Precision of readout at the hunchback gene: analyzing short transcription time traces in living fly embryos. *PLoS computational biology*, 12(12), 2016.

[5] M. L. Ferguson and D. R. Larson. Measuring transcription dynamics in living cells using fluctuation analysis. In *Imaging gene expression*, pages 47–60. Springer, 2013.

[6] N. C. Lammers, V. Galstyan, A. Reimer, S. A. Medin, C. H. Wiggins, and H. G. Garcia. Multimodal transcriptional control of pattern formation in embryonic development. *Proceedings of the National Academy of Sciences*, 117(2):836–847, 2020.

[7] J. Rodriguez, G. Ren, C. R. Day, K. Zhao, C. C. Chow, and D. R. Larson. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*, 176(1-2):213–226, 2019.