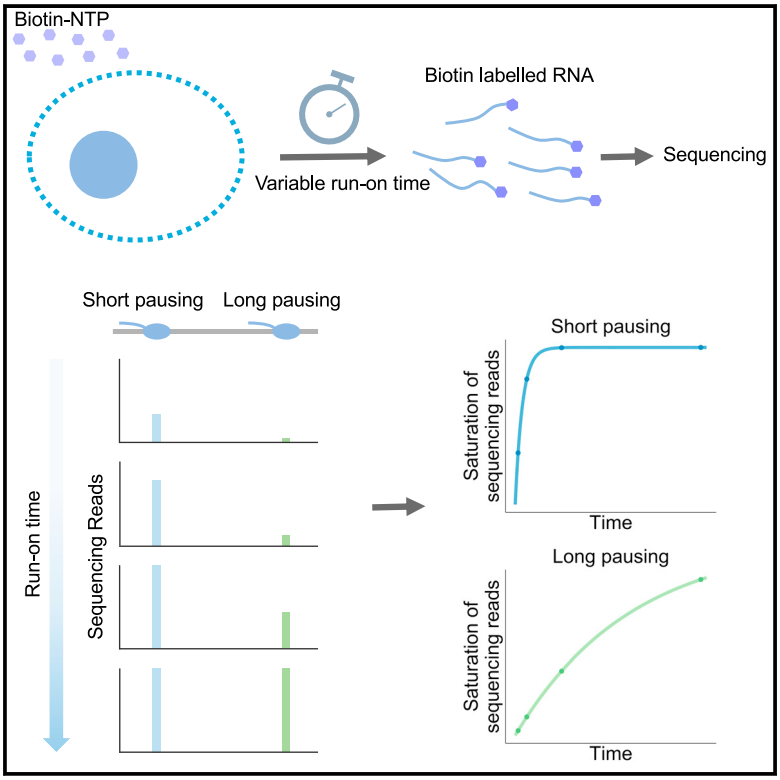
Cell Reports
Methods

Article

Timing RNA polymerase pausing with TV-PRO-seq

Graphical abstract



Highlights

- Measurement of polymerase pausing times genome-wide at single-base resolution
- Promoter-proximal regions have more frequent but shorter individual pausing events
- Genes with high transcriptional noise have more and longer pausing
- H3K36me3 correlates with pausing

Authors

Jie Zhang, Massimo Cavallaro, Daniel Hebenstreit

Correspondence

j.zhang.50@warwick.ac.uk (J.Z.),
d.hebenstreit@warwick.ac.uk (D.H.)

In brief

Zhang et al. develop a next-generation sequencing method based on PRO-seq that allows genome-wide measurement of pausing times of RNA polymerases at single-base resolution. TV-PRO-seq reveals frequent short pausing events in promoter-proximal regions and uncovers links between pausing times and gene expression as well as chromatin states.

Zhang et al., 2021, Cell Reports Methods *1*, 100083
October 25, 2021
https://doi.org/10.1016/j.crmeth.2021.100083

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Timing RNA polymerase pausing with TV-PRO-seq

Jie Zhang,[1,*] Massimo Cavallaro,[1,2] and Daniel Hebenstreit[1,3,*]
[1]School of Life Sciences, Gibbet Hill Campus, the University of Warwick, CV4 7AL Coventry, UK
[2]Mathematics Institute and Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research, the University of Warwick, CV4 7AL Coventry, UK
[3]Lead contact
*Correspondence: j.zhang.50@warwick.ac.uk (J.Z.), d.hebenstreit@warwick.ac.uk (D.H.)
https://doi.org/10.1016/j.crmeth.2021.100083

MOTIVATION Various next-generation sequencing-based methods, including Pol II ChIP-seq, GRO-seq, NET-seq, and PRO-seq, have been able to reveal the presence of polymerase pausing. However, these methods are mainly based on detection of polymerase occupancy, which is affected by factors other than the pausing. As an example, recent research revealed that polymerase has a high abortive transcription rate, especially at genes featuring strong pausing, which also contributes to high polymerase occupancy in the promoter-proximal regions. Further, the accuracy of the most widely used method to measure pausing—treating cells with triptolide (Trp)—is challenged by the latter's slow uptake. A method that outputs the actual pausing time is still required to dissect the pausing profile. We have developed TV-PRO-seq, which reveals polymerase pausing times at single-base resolution genome-wide and is not influenced by the polymerase turnover and other confounding factors.

SUMMARY

Transcription of many genes in metazoans is subject to polymerase pausing, which is the transient stop of transcriptionally engaged polymerases. This is known to mainly occur in promoter-proximal regions but it is not well understood. In particular, a genome-wide measurement of pausing times at high resolution has been lacking. We present here the time-variant precision nuclear run-on and sequencing (TV-PRO-seq) assay, an extension of the standard PRO-seq that allows us to estimate genome-wide pausing times at single-base resolution. Its application to human cells demonstrates that, proximal to promoters, polymerases pause more frequently but for shorter times than in other genomic regions. Comparison with single-cell gene expression data reveals that the polymerase pausing times are longer in highly expressed genes, while transcriptionally noisier genes have higher pausing frequencies and slightly longer pausing times. Analyses of histone modifications suggest that the marker H3K36me3 is related to the polymerase pausing.

INTRODUCTION

RNA polymerase II (Pol II) does not move uniformly after transcription initiation at metazoan genes; it frequently pauses during transcription (Levine, 2011; Adelman and Lis, 2012; Mayer et al., 2017; Jonkers and Lis, 2015; Porrua and Libri, 2015). The most conspicuous phenomenon that has been extensively studied in this context is pausing at promoter-proximal regions (PPRs) (Core et al., 2008; Nojima et al., 2015; Kwak et al., 2013). Several protein factors, such as negative elongation factor (NELF) (Yamaguchi et al., 1999) and DRB (5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole) sensitivity-inducing factor (DSIF) (Wada et al., 1998), have been found to influence pausing, along with more generic factors, such as DNA sequence (Szlachta et al., 2018) and/or nucleosomes (Gilchrist et al., 2010; Core et al., 2008). Another factor, positive transcription elongation factor-b (P-TEFb), mediates phosphorylation of NELF and DSIF, and facilitates Pol II release from promoter-proximal pausing into active elongation (Peterlin and Price, 2006; Patel et al., 2013).

While polymerase pausing was discovered decades ago (Rasmussen and Lis, 1993; Maizels, 1973; Gariglio et al., 1981), its purpose remains uncertain. Several examples suggest a role in expression regulation, in particular for genes that need to respond quickly, as upon heat shocks, for instance (Mahat et al., 2016b). On the other hand, pausing appears to be common; it was reported to occur at roughly a third of all genes (Adelman and Lis, 2012), as also demonstrated by small-molecule inhibition (with flavopiridol [FP]) of P-TEFb, which leads to a widespread suppression of all active genes (Peterlin and Price,

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

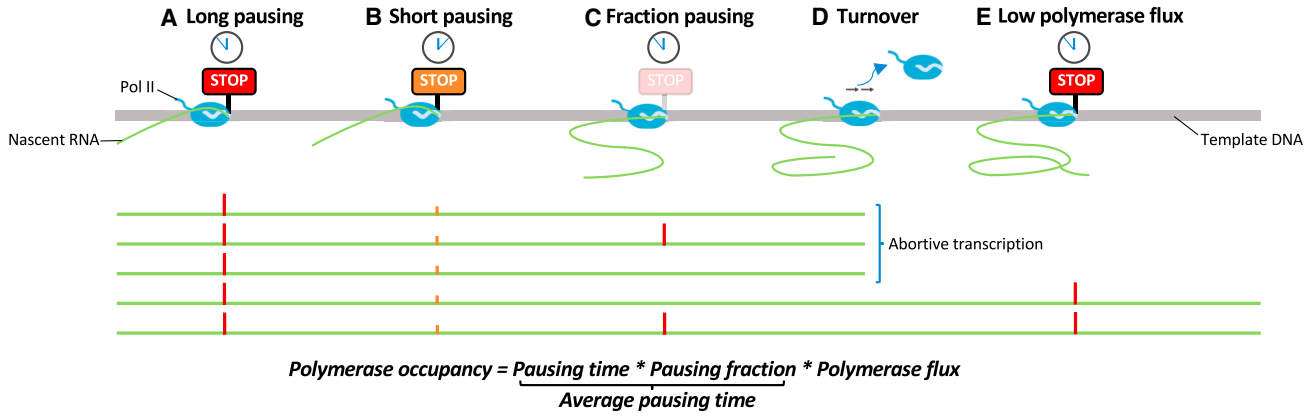


Figure 1. Schematic illustration of pausing-related phenomena

(A) The green lines refer to transcribed RNA of the example gene on top. Red bars correspond to sites where long pausing occurred during transcription.

(B) Polymerases pause shorter at this position (orange bars), resulting in a lower polymerase occupancy.

(C) Only part of the polymerases pause at this position, thus, even though (C) has similar pausing times for each paused polymerase to (A), the average pausing time is lower due to the low pausing fraction, and finally results in a lower polymerase occupancy.

(D) Polymerases can drop off the DNA template at this point and result in fewer polymerases reaching downstream positions (lower polymerase flux).

(E) The polymerase flux downstream of the turnover site (D) will be lower and so will the polymerase occupancy. A generally low gene expression level will have the same effect and lower both.

2006; Jonkers and Lis, 2015; Rahl et al., 2010). These results point toward a fundamental function of pausing in the transcriptional machinery. On the other hand, recent research suggests that high promoter-proximal Pol II densities, which are usually interpreted as signatures of pausing, rather reflect a high turnover rate of nascent transcripts, i.e., abortive transcription (Krebs et al., 2017; Steurer et al., 2018; Erickson et al., 2018). In fact, *in vivo* experiments show that only less than 10% of polymerases can escape from the PPR and enter productive elongation (Steurer et al., 2018).

Understanding of polymerase pausing has been greatly advanced by several types of assays based on next-generation sequencing; these include chromatin immunoprecipitation sequencing (ChIP-seq) and ChIP-nexus (chromatin immunoprecipitation experiments with nucleotide resolution through exonuclease, unique barcode, and single ligation) (He et al., 2015), global nuclear run-on sequencing (GRO-seq) (Core et al., 2008), mammalian native elongating transcript sequencing (Churchman and Weissman, 2011) and precision nuclear run-on sequencing (PRO-seq) (Kwak et al., 2013), along with more recent developments, such as coordinated PRO-seq, which can correlate pausing with transcriptional start sites (Angers-Loustau et al., 2018) and 5′ capping states (Tome et al., 2018). These assays are mostly based on the sequencing of polymerase-associated DNA fragments or nascent mRNA. After mapping the resulting sequencing reads to the genome, locations with higher read counts ("peaks") are thought to reflect greater polymerase occupancies, which are then used as proxies for pausing locations. Aside from revealing Pol II accumulations in the PPR, these assays led to many other important findings (Mayer et al., 2017; Tome et al., 2018), including pausing sites in gene bodies (Nojima et al., 2015) and at 3′ ends of genes (Core et al., 2008).

A fundamental problem with these methods based on measuring polymerase occupancy is that they cannot separate the influence of pausing and the turnover of aborted transcripts in the PPRs. This is due to the methods' inability to discriminate between few slow polymerases and many fast polymerases detected at a genomic position during a given amount of time; both cases will result in identical peaks of sequencing reads, which prevents measuring the actual pausing times. The latter has been accomplished only via blocking transcription initiation with the covalent TFIIH subunit XPB inhibitor Trp (Titov et al., 2011) and measuring Pol II release dynamics from the PPRs and at low resolution (Jonkers et al., 2014; Shao and Zeitlinger, 2017); to the best of our knowledge, genome-wide data for pausing times at single positions are lacking.

We present here an extension of the PRO-seq assay, which we termed time-variant PRO-seq (TV-PRO-seq), that achieves this goal. TV-PRO-seq removes the influence of many factors that affect Pol II occupancy (such as expression level, abortive transcription, and pausing fraction; see Figure 1), thus allowing us to directly study the pausing times *in vitro*. The TV-PRO-seq results are consistent with the more limited data obtained from *in vivo* Trp treatment followed by sequencing, and go beyond the latter in revealing, based on genome-wide analyses, the novel finding that Pol II pauses more often but for shorter times at each base in the PPRs than in other regions of a gene. The pausing related to NELF can be lifted by sarkosyl treatment. Our results also show that pausing within the genes with higher expression levels lasts longer. Genes with higher transcriptional noise tend to have higher NELF levels in their PPRs, along with higher densities of pausing sites that extend to their gene bodies but differ little in terms of pausing times. Our analysis of individual pausing sites has also yielded insights into the pausing profiles associated with productive elongation; we find that the active elongation marker H3K36me3 surprisingly relates to pausing, suggesting pausing could establish H3K36 methylation and/or a dynamic

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

equilibrium of H3K36me3 and histone acetylation that contributes to elongation-rate regulation.

RESULTS

Estimating pausing times for individual pausing sites in genome-wide fashion

Numerous advanced sequencing methods have been developed for studying pausing based on RNA polymerase occupancy (Core et al., 2008; Churchman and Weissman, 2011; Kwak et al., 2013; He et al., 2015). However, RNA polymerase occupancy correlates with various parameters, such as gene expression level, polymerase turnover rate, pausing fraction, and pausing time, thus preventing independent measurement of the latter (Figure 1). To overcome this problem, we developed TV-PRO-seq; it can extract pausing times of polymerases at individual pausing sites in genome-wide fashion. This enables us to dissect pausing profiles to gain a mechanistic understanding.

We developed TV-PRO-seq based on a detailed analysis of the principles underlying PRO-seq. The assay relies on the replacement of native NTPs (nucleoside triphosphates) in the nuclei with biotin-labeled ones (biotin-NTPs), which become incorporated into the 3′ end of nascent RNA (Mahat et al., 2016a) over a short period of time (run-on time). This blocks further transcription and makes polymerase drop off the template, thus marking the exact location of incorporation. The biotin tag is then used to isolate newly synthesized RNA, followed by library preparation and sequencing. The longer the run-on time, the more polymerases will be released. Eventually, all polymerases will have been released and no more reads can result. Although the kinetics of biotin-NTP incorporation are not necessarily identical to those of physiological NTP, we argue that they are correlated in rank order to those *in vivo*, thus permitting inference of biological dynamics. Preparing several PRO-seq reactions using different run-on times allows us to fit saturation curves, whose slopes permit estimation of pausing (release) times (Figure 2A).

In TV-PRO-seq, polymerases theoretically can only move a maximum of one nucleotide after release from pausing during run-on (Figure 2A). This allows us to record data for each nucleotide, thus enabling genome-wide estimation of pausing times at single-nucleotide resolution. The most widely used sequencing method for estimating pausing times, Trp treatment-based sequencing, is incapable of achieving this (Jonkers et al., 2014; Gressel et al., 2017; Shao and Zeitlinger, 2017; Henriques et al., 2013). Trp is a covalent inhibitor of the TFIIH subunit XPB (Titov et al., 2011) and therefore blocks transcription initiation. Fitting decay curves to the (declining) polymerase occupancy of the region downstream of TSS (transcription start site) upon a Trp treatment time series allows estimation of the average pausing times at the PPRs of all genes (Figure 2B). As it prevents initiation, this method can only measure the average pausing times (Figure 1) in a small region downstream of the TSS; its application in the gene body or to individual pausing sites in the PPR is not possible, which is a significant limitation.

We performed TV-PRO-seq for HEK293 cells using 0.5-, 2-, 8-, and 32-min run-on times.

The standard PRO-seq preparation buffer contains sarkosyl, an anionic detergent that has been reported to facilitate pausing release especially in the PPR of coding genes and enhancers (Rougvie and Lis, 1988; Core et al., 2012). To remove its effects toward pausing, we excluded sarkosyl from the run-on buffer of all TV-PRO-seq samples, except for an initial comparison.

To analyze the resulting data, we first called peaks based on a heuristic thresholding (STAR Methods; Figure S1A), which resulted in 66,089 individual peaks for the HEK293 data. Plotting the distribution of peaks around TSSs reproduced the familiar pattern of promoter-proximal peaks and divergent transcription on the other strand (Figure 1C). A replicate, with 47,713 peaks detected, shows a similar pattern of peak distribution (Figure S1B), and the two replicates' pausing times were consistent among them (Figure S1C).

For comparison, we also generated two replicates of TV-PRO-seq with sarkosyl (Figure S1D). Due to the lack of elongation-rate measurements under sarkosyl treatment and the latter's effects on pausing in general, we do not expect our estimates for the sarkosyl run-on samples to reflect the real pausing times (Jonkers et al., 2014; Fuchs et al., 2015). The pausing times of these sarkosyl run-on samples have better consistency compared with sarkosyl-free samples (Figures S1C and S1D), as the sarkosyl increases the run-on efficiency (Core et al., 2012). In contrast, a complex pattern emerges when comparing sarkosyl and sarkosyl-free samples, indicating substantially stronger treatment effects than replicate variation in our datasets and confirming their informative value (Figure S1E).

We constructed a mathematical model that takes account of our theoretical considerations; the model predicts the saturation curve as a function of the pausing release rate and the TV-PRO-seq run-on time (STAR Methods). Fitting our saturation model to the time course data of a set of peaks allows the inference of their pausing release rates, whose reciprocals are the pausing times. The model is embedded into a Bayesian framework, detailed in STAR Methods. Examples of fitted curves corresponding to two close individual peaks are shown in Figure 2D (saturation curves from the replicate data for the same peaks are shown in Figure S2). Note that the saturations are subject to trade-offs in terms of sequencing reads with peaks with extremely high pausing times (whose polymerase occupation remains virtually unchanged during the time course experiment): normalizing the samples' total read numbers to stay constant throughout the run-on time, the latter's sizes will decrease (Figure 2D, bottom left; see STAR Methods for details). Our estimates of the individual pausing times are based on average elongation rates from published data (Jonkers et al., 2014; Fuchs et al., 2014), which were incorporated into the model as a Bayesian informative prior (STAR Methods).

Trp treatment followed by sequencing has been widely used for estimation of pausing times of PPR regions (Jonkers et al., 2014; Gressel et al., 2017; Shao and Zeitlinger, 2017; Henriques et al., 2013). The quicker the reduction of polymerase occupancy upon after Trp treatment, the shorter the pausing times are assumed to be (Figure 2B). As an *in vitro* method, TV-PRO-seq shows a good consistency with Trp treatment. We took the 2,000 genes with the highest nascent transcription signal in PPR region. Within these, we identified the 500 genes with the most reduction of polymerase occupancy after 10-min Trp

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

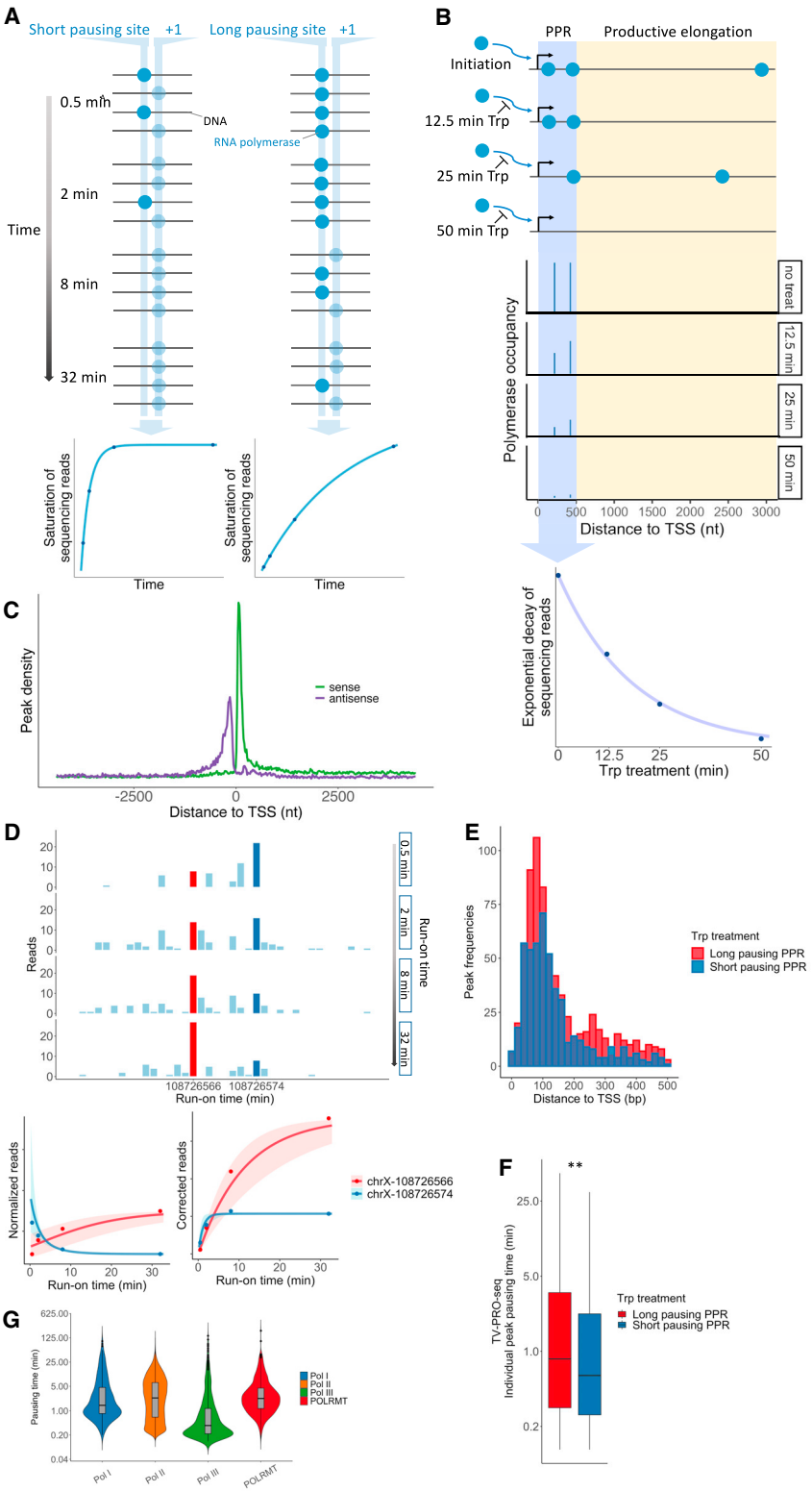


Figure 2. Principle of TV-PRO-seq

(A) The black horizontal lines symbolize a generic DNA region with a short (left graphics) and a long (right graphics) pausing site. The blue dots symbolize RNA polymerases that either are stationary or have just moved by one position (and incorporated a biotinylated NTP) as indicated by the lighter blue shades. A sequencing read results at a position if a polymerase steps forward by one base. Eventually, all polymerases will have moved, i.e., all positions will be saturated. The saturation takes longer at the position (+1) adjacent to the long pausing site, since the polymerases are released at a lower rate than from the short pausing site. Saturation curves (lower plots) can be inferred by reads from a run-on time course at each position, genome-wide.

(B) Trp blocks transcription initiation, thus decreasing the polymerase occupancies at the PPRs. The decay rate at different pausing sites is also influenced by their distance to TSS. Two pausing sites with the same pausing times are represented in the diagram; the decay rate of polymerase occupancy of the most downstream peak is underestimated by the presence of persisting polymerases upstream. The total reads of the PPR from Trp-treatment-based sequencing can be used to estimate the average pausing time in the PPR by fitting an exponential decay curve.

(C) Distributions of sense and antisense read around TSSs from pooled TV-PRO-seq samples confirm high library quality.

(D) Read numbers from two neighboring peaks (red and blue bars) in chromosome X obtained at the different run-on times (top). Normalizing these by the total-genome reads permits parameter estimation and produces the curves at bottom left. Correcting by the total-genome read trend reveals the saturation curves at bottom right (details in STAR Methods). Shaded regions are interquartile posterior ranges.

(E) More peaks are found in the long-pausing PPR. The 2,000 genes with the highest polymerase occupancy in the PPR (first 500 bp downstream of TSS) were used for analysis. Five-hundred genes each retaining the highest and lowest polymerase occupancies after 10 min of Trp treatment were grouped as long-pausing PPR and short-pausing PPR, respectively. Seven-hundred and two peaks were identified in the long pausing PPR and 493 peaks were found in the short-pausing PPR (exact binomial test, $p < 10^{-8}$).

(F) Peaks in the long-pausing PPR have longer pausing times as measured by TV-PRO-seq (Mann-Whitney U test, $p < 0.01$). Peaks were grouped same as in (E).

(G) Distributions of estimated pausing times for peaks in loci transcribed by Pol I, II, III, and POLRMT. For all pairwise comparisons except Pol II versus POLRMT and Pol II versus Pol I (non-significant), $p < 0.01$, Bonferroni-corrected Mann-Whitney U test.

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

CellPress
OPEN ACCESS

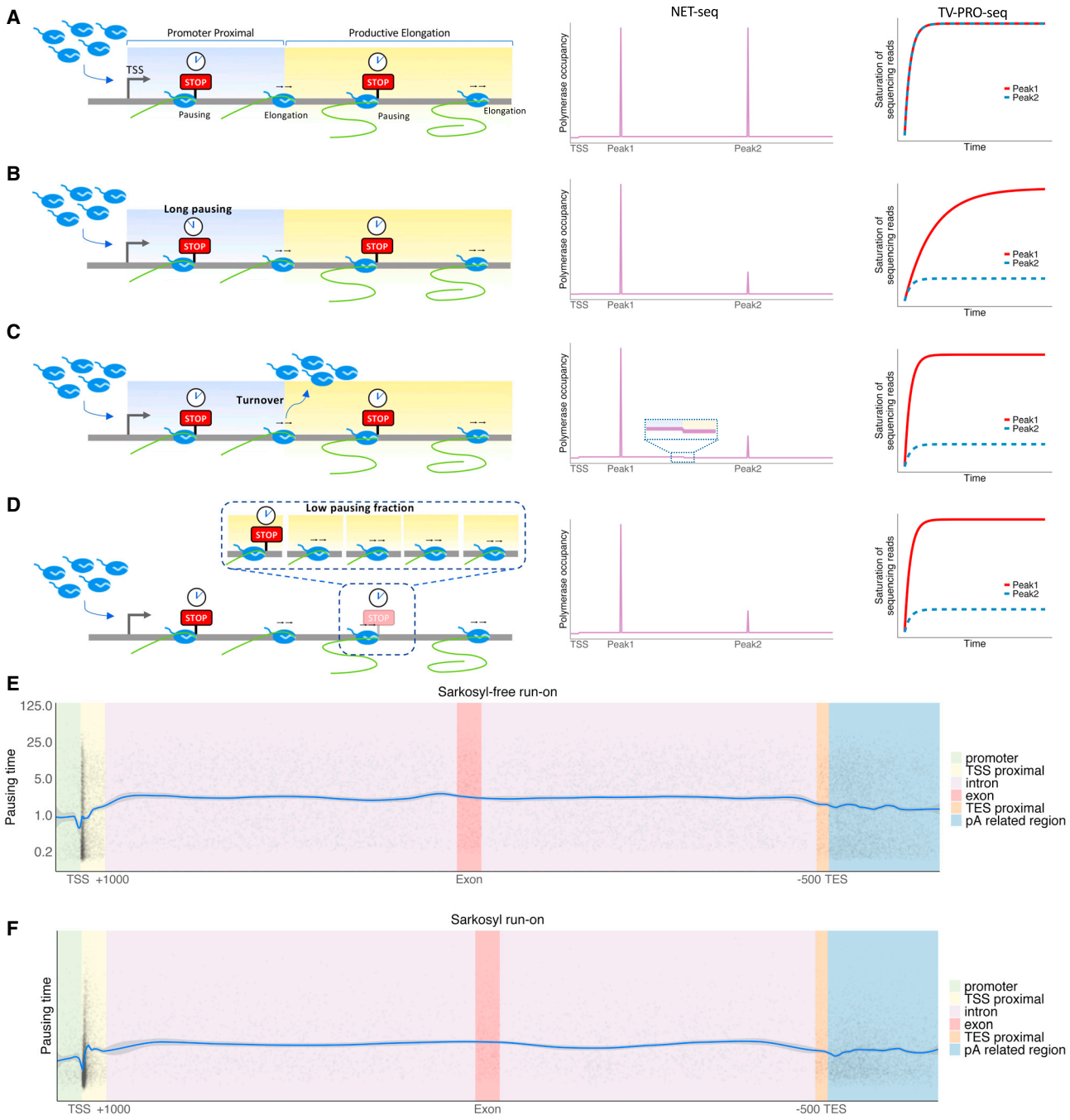


Figure 3. Several factors influence polymerase occupancy

(A) The left panel shows a schematic example case by *in silico* simulations; two regions were designated as promoter proximal (blue shading) and productive elongation (yellow shading), each with a single pausing site (peak 1 and peak 2) with identical properties. Polymerase occupancies measured by NET-seq (middle) and the saturation curves resulting from TV-PRO-seq (right) of the two peaks will be the same.

(B) As (A), but the pausing time of peak 1 was set five times longer (clock symbols) than peak 2's. Both polymerase occupancy and pausing time ($1/\beta_i$; see STAR Methods) of peak 1 would be measured to be five times higher than peak 2's.

(C) As (A) and (B), but 80% of polymerase is assumed to abort transcription at the boundary of the PPRs, thus reducing by 80% polymerase occupancy in the productive elongation region. Therefore, the measured polymerase occupancy of peak 1 would still be 5-fold higher than at peak 2, for both NET-seq and TV-PRO-seq. However, in contrast to NET-seq, TV-PRO-seq is still able to correctly measure the pausing times at the two peaks to be equal despite their differing sizes. In contrast to (B) and (D), high abortive transcription would also decrease the polymerase occupancy in the productive elongation region (magnified section).

(legend continued on next page)

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

treatment as short-pausing PPR, and the 500 with the least reduction as long-pausing PPR. We then looked at the pausing times of peaks we identified by TV-PRO-seq within the two groups of PPRs. It shows that the genes with overall longer pausing in the PPR (Trp treatment following sequencing) have both more pausing sites (TV-PRO-seq; Figure 2E) and longer pausing time for each peak (TV-PRO-seq; Figure 2F).

The positions losing polymerase quicker after 10-min Trp treatment only have short average pausing times according to TV-PRO-seq, which amount to less than half of those with high polymerase occupancy after Trp treatment (Figure S3A). As the Trp acts from the TSS, the positions far from TSS would lose polymerase slower than the upstream ones (Figure 2B). This bias does not occur in TV-PRO-seq as the interruption of transcription happens at the position the polymerase is located at rather than the TSS. Thus we zoomed in to pausing sites within the first 500 bp or even 100 bp downstream of TSS to reduce this bias, and the difference remains significant (Figures S3B and S3C).

Pol I and Pol III transcription contribute about 80% of the total RNA production in rapidly growing cells (Willis and Moir, 2018). Many noncoding RNAs that are essential for cells are transcribed by Pol I and Pol III. Some of these noncoding RNAs, such as RNase P, RNase MRP, or transfer RNAs (tRNAs) are only about 100 nucleotides long (Oler et al., 2010; Willis and Moir, 2018; Dumay-Odelot et al., 2014). Previous methods for estimating pausing levels cannot work for these short genes, as they are mostly based on the so-called pausing index (PI; also referred to as stalling index or escaping index; STAR Methods), which normalizes polymerase occupancy within the first ~200 to 300 nt downstream of TSSs by that in the gene body. The purpose of this normalization is to correct for the gene's expression level by treating promoter-proximal pausing as the (increased) relative occupancy over the gene body's (Core et al., 2008; Nojima et al., 2015). In contrast, TV-PRO-seq estimates pausing times for individual pausing sites, thus providing a way to study pausing in both short and long genes, and at any position. We pooled the pausing positions in Pol III-transcribed genes and compared their times with those of pausing sites related to Pol I, Pol II, and mitochondrial polymerase (POLRMT). We found that pausing time varies for different type of polymerases. The long pausing times are mostly associated with Pol II, and, surprisingly, Pol III pauses for the shortest times, overall (Figure 2G).

TV-PRO-seq estimates pausing times independently of confounding factors

Pol II is found to be enriched in PPRs (Muse et al., 2007; Core et al., 2008), which is usually interpreted as longer Pol II residence time, and thus pausing, in these regions than elsewhere (Figures 3A and 3B). This pausing is often claimed to be of particularly long duration, from 2 min to even 30 min for some genes, based on studies blocking transcription initiation with Trp and the slow reduction of polymerase occupancy that ensues (Jonkers et al., 2014; Gressel et al., 2017; Shao and Zeitlinger, 2017; Henriques et al., 2013). However, measuring pausing time in this way relies on a rapid uptake of Trp, but 500 nM Trp treatment, which is the usual concentration used for initiation inhibition prior to pausing time measurement in PPR regions (Jonkers et al., 2014; Shao and Zeitlinger, 2017), has proved ineffective in this respect (Nilson et al., 2017).

Indeed, polymerase densities higher in the PPR than in the regions downstream can have different causes. If transcription aborts before entering productive elongation, the polymerase occupancy will appear higher in the PPR (Figures 3A and 3C); in fact, recent research shows that Pol II does have a high turnover rate in PPRs (Krebs et al., 2017; Steurer et al., 2018; Erickson et al., 2018). Modeling based on *in vivo* experiments suggests that only about one-thirteenth of Pol II can escape from the PPR and progress into productive elongation (Steurer et al., 2018); in contrast to reported average pausing times from 7 min to up to half an hour proximal to promoters (Jonkers et al., 2014; Gressel et al., 2017; Shao and Zeitlinger, 2017), this *in vivo* study claims polymerase residence times in the PPR of only about 42 s (Steurer et al., 2018). Similarly, a lower pausing fraction, i.e., lower utilization of pausing sites (Figure 1) by polymerase engaged in productive elongation, yields lower occupancy downstream of the PPR (Figures 3A and 3D). Also, a single pausing site with long pausing time versus many short-pausing sites will result in the same contribution to the PPR's overall polymerase occupancy.

Even though various sequencing methods have been developed/used for the research on transcriptional dynamics and similar topics (He et al., 2015; Core et al., 2008; Churchman and Weissman, 2011; Kwak et al., 2013; Tome et al., 2018), these are in fact restricted to revealing polymerase occupancy only. Extending the pausing time, increasing the turnover rate, or boosting utilization of a pausing site can affect polymerase occupancy in similar ways (Figures 3A–3D, NET-seq). To distinguish the source of polymerase enrichment among these three possibilities, we developed TV-PRO-seq. In contrast to previous approaches using a time series of Trp treatment (Figure 2B) followed by ChIP-seq or GRO-seq (Henriques et al., 2013; Jonkers et al., 2014; Gressel et al., 2017; Shao and Zeitlinger, 2017), TV-PRO-seq not only can remove the influence of early-terminated transcripts but can also measure pausing times without Trp treatment. Both of these advantages prevent over-estimating

(D) As (A), but only one-fifth of the polymerase is assumed to pause at peak 2 (i.e., the pausing fraction is a fifth), thus its polymerase occupancy would decrease to one-fifth of peak 1's. The pausing time of peak 2, however, would be about the same as peak 1's.

(E) Pausing times at mRNA-transcribing metagene. Each gray dot represents a pausing peak, with corresponding pausing time given by its y axis value. The x axis values correspond to the absolute position within ± 1,000 nt of the TSS (green and yellow tinged regions, respectively). The intron/exon regions (purple/red, respectively) start after +1,000 nt of the TSS and end before −500 of the TES (introns were split into an upstream and a downstream group at the gene's middle point) and 500 nt upstream and 4,500 nt downstream of the TES (polyA-related sites) were indicated (orange and blue, respectively). The blue line corresponds to the moving average (locally estimated scatterplot smoothing [LOESS] fit). The gray shading indicates the 0.95 confidence interval and is negligible on this scale, hence invisible over most of the graph. The widths of exons and introns have been scaled to their relative average lengths.

(F) Similar to (E), but including sarkosyl during the run-on reactions.

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

CellPress
OPEN ACCESS

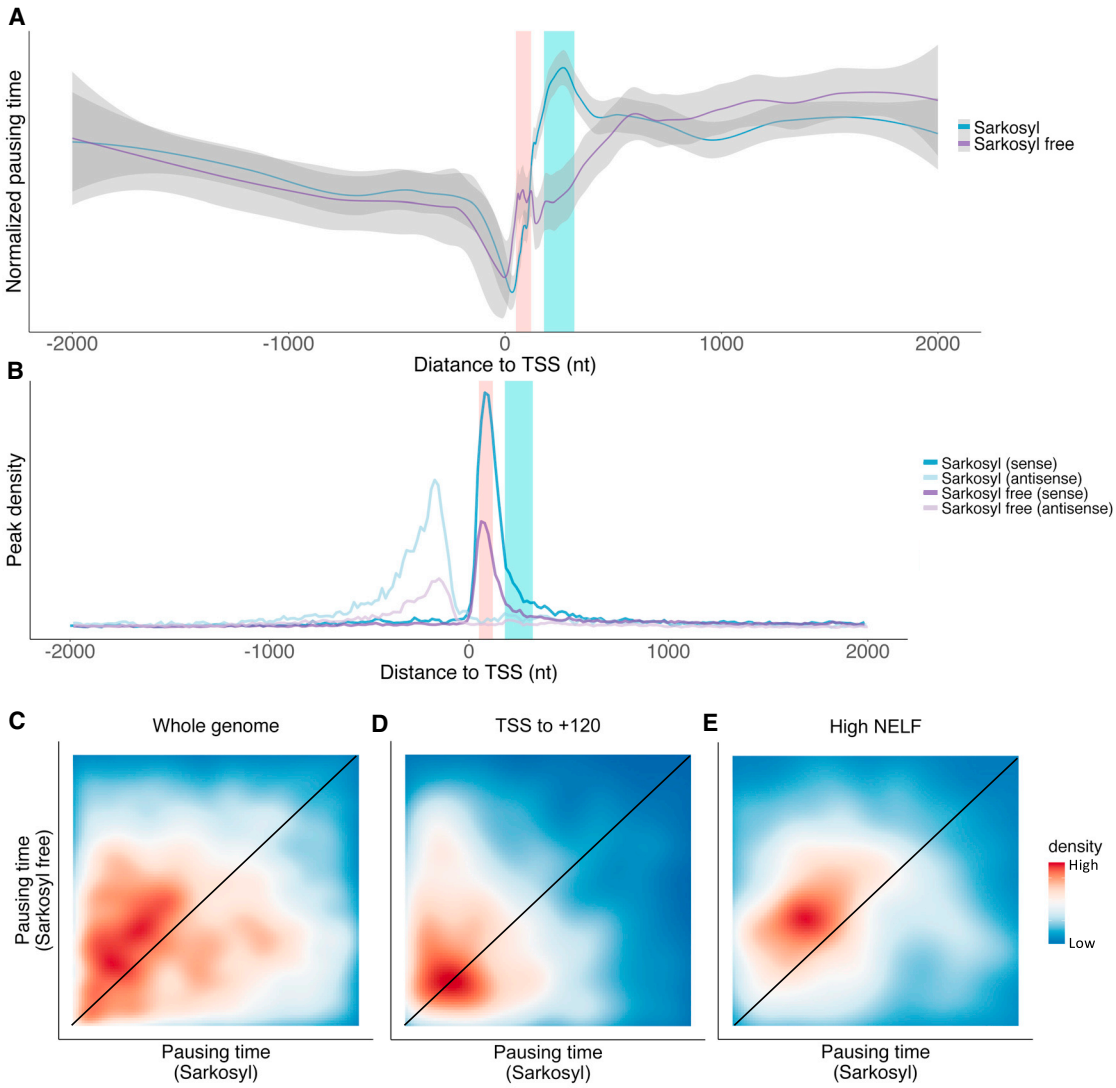


Figure 4. Influence of sarkosyl on pausing

(A) Pausing time of peaks around TSS. For removing the systematic bias of pausing time estimation, the average pausing time is normalized to the same value for samples with sarkosyl (blue line) or with a sarkosyl-free (purple line) run-on. Pausing in the +50 to +120 region (pink shading) is sensitive to sarkosyl, while pausing in the +180 to +320 region (cyan shading) shows resistance to sarkosyl.

(B) Sarkosyl increases the PPR peak density for both sense and divergent transcription.

(C) 2D density plots show the pausing time rank of the equivalent peaks in sarkosyl sample and sarkosyl-free sample. The black line reflects peaks with intermediate influence on pausing time by sarkosyl. Peaks above the black line correspond to pausing sites releasing paused polymerase after sarkosyl treatment.

(D) Similar to (C), only for the peaks within the first 120 bp of genes.

(E) Similar to (C), but only peaks with top 10% of NELF level.

pausing times in the PPR, since the artifactual extra occupancies due to aborted transcripts and slow Trp uptake do not influence the results.

As shown in Figure 3E, even though we find a much higher frequency of polymerase pausing in the PPR, TV-PRO-seq demonstrates that, in fact, individual pausing events close to TSSs last shorter times on average than those in other regions. Our interpretation of this is that pausing in the PPR is more akin to a collection of check points with high possibilities to pause polymerase for short times rather than a unitary long pausing apparatus that holds it back from moving into productive elongation. Even though individual pausing events in the PPR are shorter, considering the higher pausing frequency and potentially higher pausing fraction, the average elongation speed of polymerase in the PPR might still be lower than further downstream.

Sarkosyl facilitates release from NELF-mediated long pausing

Sarkosyl specifically facilitates pausing release in the PPR (Core et al., 2012). This effect is reflected in TV-PRO-seq by the

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

deepening of the dip in pausing times downstream of TSSs (Figures 3E, 3F, and 4A). Also, more peaks in the PPR are found in the sarkosyl run-on sample compared with the sarkosyl-free one (Figure 4B). This is due to PRO-seq's selective detection of active polymerases; sarkosyl appears to denature NELF, thus causing a pausing release in the PPR and boosting peak density (Core et al., 2012).

NELF and DSIF are the best-characterized factors involved in promoter-proximal pausing (Liu et al., 2015; Adelman and Lis, 2012; Vihervaara et al., 2018) and their depletion significantly reduces polymerase occupancy in the PPRs (Gilchrist et al., 2010; Yamaguchi et al., 2013). P-TEFb is a necessary factor for productive elongation, which facilitates dissociation of NELF and converts DSIF to a positive elongation factor by phosphorylating it (Adelman and Lis, 2012; Liu et al., 2015). Inhibition of P-TEFb can prevent polymerase from entering productive elongation at nearly all active genes (Peterlin and Price, 2006; Jonkers and Lis, 2015; Rahl et al., 2010).

Interestingly, we found that the effect of pausing release caused by sarkosyl is mostly restricted to the first 120 bp downstream of TSS (Figure 4A), which coincides with the region with the highest NELF levels (Figures S4A and S4B). The situation differs further downstream; polymerases show long pausing times between +180 and +320 (Figure 4A). But when we plot the pausing time of peaks with different NELF coverage around TSSs of the sarkosyl-free sample (Figure S4C), we find NELF to correlate with higher pausing times only within the first region, +120 downstream of TSS. In the region further downstream, pausing sites with higher NELF levels actually have shorter pausing times than the other peaks at the same distance toward the TSS (Figure S4C).

To further dissect the relation between NELF and pausing, we revisit the sarkosyl-treated samples. Sarkosyl disturbs pausing in the PPR (Figure S4D), but its effect on paused polymerase varies with distance to TSS. For the first 120 bp downstream of TSS, all peaks with different NELF levels show a reduction of pausing time. Pausing with low NELF levels does not show a pausing time decrease within the +180 to +320 region, though. We suspect that this indicates that pausing is established by different mechanisms whose prevalence varies with the type of region. Candidates include pausing related to G-quadruplex DNA secondary structures, which, similar to NELF/DSIF, are also enriched in the PPRs (Szlachta et al., 2018). Nucleosomes also can pause polymerase, specifically at the +1 nucleosome (Liu et al., 2015; Adelman and Lis, 2012; Vihervaara et al., 2018), but also further downstream in the gene body (Kwak et al., 2013; Kireeva et al., 2005; Hodges et al., 2009; Gilchrist et al., 2010). Pause elements (Vvedenskaya et al., 2014; Saba et al., 2019) and nascent RNA structures (Kang et al., 2018) can induce pausing in the whole gene. In fact, peaks in the first 120 bp were not affected by sarkosyl more than other peaks (Figures 4C and 4D). In contrast, a strong effect can be seen for the peaks with high NELF level (Figure 4E). Specifically, the peaks that are most sensitive to sarkosyl are those peaks in gene body rather than PPR (Figures S4E and S4F).

Polymerase pausing and expression level

Counterintuitively, the presence of polymerase pausing is not associated with low gene expression. As a matter of fact, most pausing is found in active genes (Adelman and Lis, 2012). In line with this, paused polymerases in the PPR have been suggested to keep the chromatin in an open state, thus keeping the gene active (Gilchrist et al., 2010). On the other hand, these paused polymerases have been described to block initiation of successive polymerases (Shao and Zeitlinger, 2017).

Studies of polymerase pausing and expression levels have been mostly focused on the PPR. With TV-PRO-seq, we can produce a more detailed picture across the whole genome. Highly expressed genes should have more detectable pausing sites due to higher polymerase volume. As expected, we find about 10 times more pausing sites in highly expressed genes (top 20% quantile; Figure 5A). Interestingly, highly expressed genes have not only more pausing sites but also higher pausing times at each site (Figure 5B). This difference is maintained across the whole gene, but is especially prominent in the PPR and downstream of TESs (transcription end sites) (Figures 5C and 5D).

Despite this, highly expressed genes do not appear to elongate more slowly (Veloso et al., 2014). This may be due to biased sampling, since elongation-rate measurements can only be taken for long genes. Conversely, TV-PRO-seq is not restricted in this way, which allowed us to revisit the pausing time analysis for peaks in extremely long genes and relatively short genes (>100 kb and 3–10 kb, respectively). Interestingly, we found that pausing in the extremely long genes tends to be shorter than in the short genes (Figure S5A), the difference being significant only for highly expressed genes (Figures S5B and S5C). This dilutes the difference in pausing times between highly expressed and less expressed genes among long genes (it results in a non-significant difference between these groups), albeit highly expressed long genes do have slightly higher pausing times than their less expressed counterparts (Figure S5D).

We suggest that these different pausing profiles relate to the regulation of gene expression. The rate-limiting step of the less expressed genes is considered to be the activation/deactivation of the promoter. In highly expressed genes, instead, the promoter is thought to be always active, thus, the regulation of expression must, to a certain degree, rely on post-initiation mechanisms, including the pausing. Short genes might provide insufficient space in their gene bodies for the complex regulatory machinery to adjust expression, and therefore denser and longer pausing might be utilized in these genes instead.

Polymerase pausing and transcriptional noise

A gene's expression level is determined by its initiation rate, the fraction of nascent RNA that is turned into mature RNA, and the latter's degradation rate. Polymerase pausing also influences the gene expression by adjusting the elongation process, but this chiefly results in the dispersed distribution of mRNAs among individual cells rather than contributing to the mean expression (Rajala et al., 2010). This dispersion, or "noise," is quantified by the square of coefficient of variation (CV^2) and can be obtained in genome-wide fashion from single-cell RNA-seq (e.g., droplet sequencing [Drop-seq]) data. To study the relation between noise and pausing, we used Drop-seq data for HEK293 cells (Macosko et al., 2015) and classified genes based on their CV^2 for a moving average of mean expression levels. This

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

CellPress
OPEN ACCESS

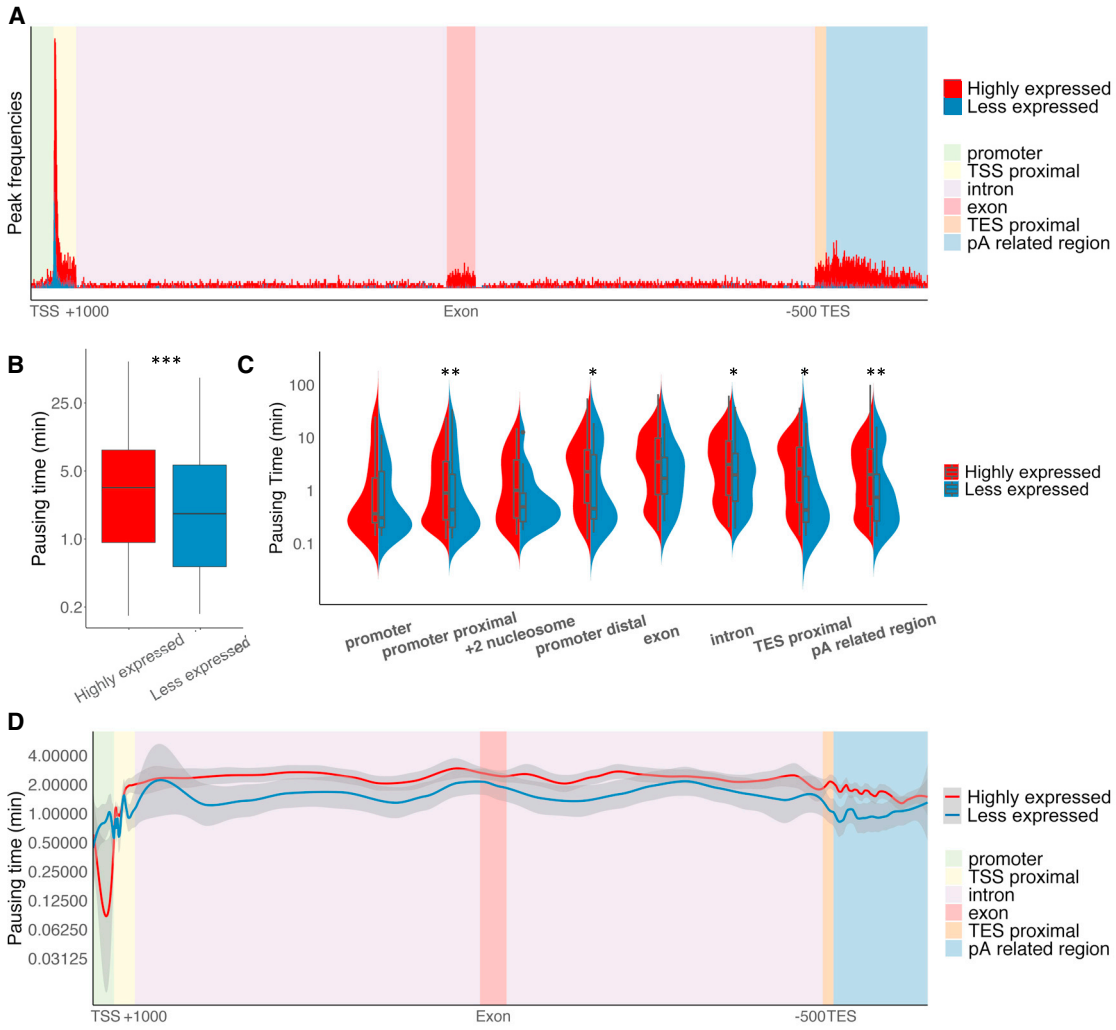


Figure 5. Pausing profiles and expression level

(A) Absolute peak density at mRNA-transcribing metagene as in Figure 3E, for genes classified into different expression levels (highly expressed, less expressed; red, blue, respectively).

(B) Pausing times of pausing sites in highly expressed genes are longer than less expressed ones. $p < 10^{-23}$, Mann-Whitney U test.

(C) Pausing times of different regions of highly expressed and less expressed genes. Definitions of the region are the same as in Figure 3E; TSS proximal has been split into promoter proximal (TSS to +120), +2 nucleosome (+180 to +320) and promoter distal (+500 to +1,000) according to the different effects of sarkosyl on these regions. For promoter-proximal and pA-related region, $p < 0.01$; promoter distal, intron, and TES proximal, $p < 0.05$; Mann-Whitney U test.

(D) Pausing times of pausing peaks among genic regions for low- and high-expression genes at the metagene as in (A) shown as LOESS fits as in Figure 3E.

reduces the influence of the latter, which the noise depends on (Klein et al., 2015; Dar et al., 2016) (Figure S6A).

We assigned genes to low- and high-noise classes (Figures S6B and S6C). We find that, overall, noisier genes have significantly higher pausing frequency (the number of pausing peaks in a given region) throughout gene bodies (Figure 6A). Polymerase also tends to pause longer in noisier genes, albeit a statistically significant difference emerges only for introns (Figures 6B–6D). Our TV-PRO-seq-based analyses agree with previous theoretical considerations that predict more and longer pausing for high-noise genes (Rajala et al., 2010; Ribeiro et al., 2010; Dobrzynski and Bruggeman, 2009). Overall, our result suggests highly expressed genes with high transcriptional noise tend to have more and longer pausing than other genes.

We also find NELF levels to correlate with the noise of genes (Figure 6E). For divergent transcription, NELF levels differ little between high- and low-noise genes (Figure 6E). This can explain why the divergent transcription of genes with different noise levels has similar pausing frequencies in both (Figure 6F).

Histone modification and polymerase pausing

After these results, we turned our attention toward pausing and chromatin states. Different types of histone modifications

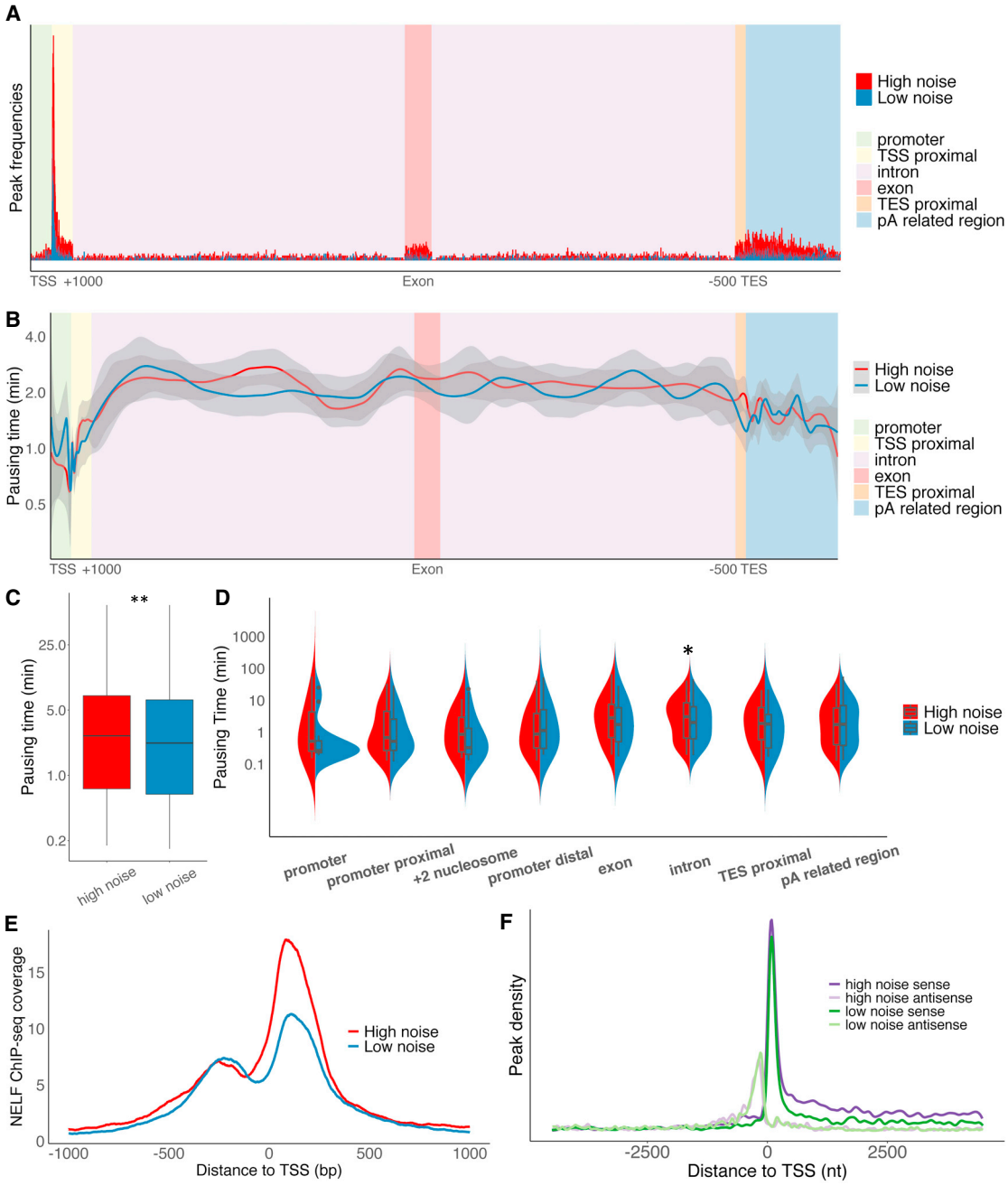


Figure 6. Pausing profiles and transcriptional noise

(A) Absolute peak density at mRNA-transcribing metagene as in Figure 3E, for genes classified into different levels of transcriptional noise (high, low; red, blue, respectively).

(B) Pausing times of pausing peaks among genic regions for low- and high-noise genes at the metagene as in (A) shown as LOESS fits as in Figure 3E.

(C) Pausing times of pausing sites in high-noise genes are longer than those of low-noise genes. $p < 0.01$, Mann-Whitney U test.

(D) Pausing times of different regions of high- and low-noise genes. Definitions of the regions are the same as in Figure 5C. For introns, $p < 0.01$, Mann-Whitney U test.

(E) NELF coverage at TSSs of genes with high or low noise.

(F) Absolute peak densities of both sense and antisense transcription of high- or low-noise genes.

influence transcription in various ways and vice versa (Li et al., 2007). For instance, new histone acetylation is found at many genes after a heat shock (Vihervaara et al., 2017, 2018). Histone acetylation can also accelerate the release of paused polymerase (Hodges et al., 2009; Bintu et al., 2012; Stasevich et al., 2014; Galvani and Thiriet, 2015).

We thus investigated the effects of chromatin states on pausing times. To this end we classified peaks as long and short

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

according to their pausing times and quantified their presence around different chromatin features. We found relations between pausing times and DNA accessibility and/or regulatory character; open chromatin regions as determined by DNase-seq display strong enrichment of short pausing (Figure 7A). This is consistent with our other results, as open chromatin is found at the PPR, which in turn is enriched for NELF (Figure S4A). The dramatic drop of DNA accessibility we see after the short pausing sites suggests that polymerases tend to pause in front of closed chromatin. A similar drop, albeit of reduced magnitude, is also seen for long pausing (Figure 7A).

Activating histone modifications (Li et al., 2007) such as H3K4 methylations and H3K27 acetylation exhibit similar profiles around long and short pausing sites within the PPR (Figures 7B and S7). H3K36me3 is an elongation marker that is usually found enriched at exons of active genes (Guenther et al., 2007; Kolasinska-Zwierz et al., 2009). In contrast to other active markers, H3K36me3 shows a clear pattern of enrichment downstream of long pausing sites in the PPR (Figure 7C) and is also enriched at the pausing sites in the gene body (Figure 7D). This suggests that H3K36me3 is involved in pausing, specifically long pausing. This contrasts with other activating histone modifications, whose profiles appear flat in gene bodies despite having higher coverages around long pausing sites (Figures 7E and S7). Its association with active genes could mean that H3K36me3 is involved in the more intensive pausing activity we find in highly expressed genes (Figures 5A and 5B).

Two hypotheses could be proposed for explaining the reason that H3K36me3 is an active marker of expression but also associates with long pausing. The first one posits that pausing could help the recruitment and function of the SET2 complex. Since H3K36me3 is deposited co-transcriptionally, increasing Pol II pausing time would also give SET2 more time to act. The longer the pausing lasts, the higher the methylation level of H3K36 would thus be expected to be (Figure 7F). The second explanation is that the mark is deposited in the wake of elongating Pol II rather than functioning as a pre-set, static marker. Methylation of H3K36 is carried out co-transcriptionally by the SET2 complex, which is recruited by the carboxy-terminal domain (CTD) of Pol II (Venkatesh and Workman, 2015). By facilitating histone deacetylation via activation of EAF3 (Carrozza et al., 2005) and remodeling of repressive chromatin (Venkatesh et al., 2012; Wan et al., 2013), H3K36me3 might thus act as a "speed bump" to prevent collisions between succeeding polymerases (Figure 7G). This would also explain why a loss of SET2 only slightly influences expression levels of H3K36me3 positive genes (Zentner and Henikoff, 2013). Interestingly, a longer continuous H3K36me3 region in the gene body will form into a stronger speed bump that blocks polymerase for a longer time (Figure 7E). A tug of war between H3K36me3 and histone acetylation may function as speed control for elongation: paused polymerase is released by demethylation of H3K36me3 and histone acetylation in response to stimuli such as heat shocks (Vihervaara et al., 2018), thus raising the elongation rate of polymerase (Figure 7H). These two mechanisms could also function together.

In order to further gauge the quality and informative value of these TV-PRO-seq based results, we carried out a side-by-side comparison with NET-seq data for the same cells and chromatin states in different regions. NET-seq shows good agreement overall but completely different patterns for H3K36me3 in the PPR compared with TV-PRO-seq (Figure S7). The reason is probably that its signal correlates not only with pausing time but also with pausing fraction, abortive transcription, and expression level (Figures 3A–3D and 1). As H3K36me3 has been identified as an elongation marker, high polymerase occupancies are expected for genes with high H3K36me3 loads. The high polymerase occupancy contributes to a low PI and leads to opposite results for TV-PRO-seq and NET-seq, further demonstrating our assay's benefit.

DISCUSSION

While the phenomenon of polymerase pausing has been known for decades (Mayer et al., 2017; Maizels, 1973), the methods to investigate it are still mainly based on polymerase occupancy (Mayer et al., 2017; Kwak et al., 2013; Churchman and Weissman, 2011; Shao and Zeitlinger, 2017), which can be actually confounded by various other factors.

Highly expressed genes tend to have more polymerases on their bodies regardless of the pausing; therefore, a non-pausing position of a highly expressed gene can have an even higher polymerase occupancy than a pausing site of a low-expression gene. As all the positions in a gene are associated with the same expression level, using the polymerase occupancy of non-pausing sites to normalize the occupancy of pausing sites can reduce the influence of the expression level toward the occupancy.

It has been suggested that the high density of polymerases in PPRs is due to long pausing in these regions (Core et al., 2008). However, this suggestion is based on the assumption that all nascent transcripts from a gene are eventually expressed as full-length RNA and/or share TSS and TES. In fact, genes have wide transcription initiation domains (Tome et al., 2018), and most transcription terminates before entering productive elongation (Krebs et al., 2017; Steurer et al., 2018; Erickson et al., 2018). This means that only a fraction of polymerases will reach a gene's 3′ end, leading to the high polymerase occupancy observed in the PPR. Based on our findings, we propose that the activity of pausing sites can be regulated. When the pausing site has been turned off, polymerase can pass it unimpededly. A pausing site with high pausing fraction and low pausing time can thus have the same average pausing time as a pausing site with low pausing fraction and high pausing time (Figures 1 and 3A–3D). Since TV-PRO-seq compares reads from the same positions, it can remove the influence of the polymerase flux. Furthermore, since non-pausing polymerase makes only tiny contributions to the polymerase occupancy (Figures 3D and 1), TV-PRO-seq can also reduce the influence of the pausing fraction and measure pausing time of paused polymerase only. Our results suggest that NELF can stabilize pausing (Figures S4 and 4E) and that polymerase indeed frequently pauses in the PPR. However, the median pausing time of individual pausing sites in PPR is less than the other regions' peaks (Figure 3E).

Unlike measurements from Trp-related methods, which are largely limited to studying the PPR, TV-PRO-seq yields results

CellPress
OPEN ACCESS

Cell Reports Methods
Article

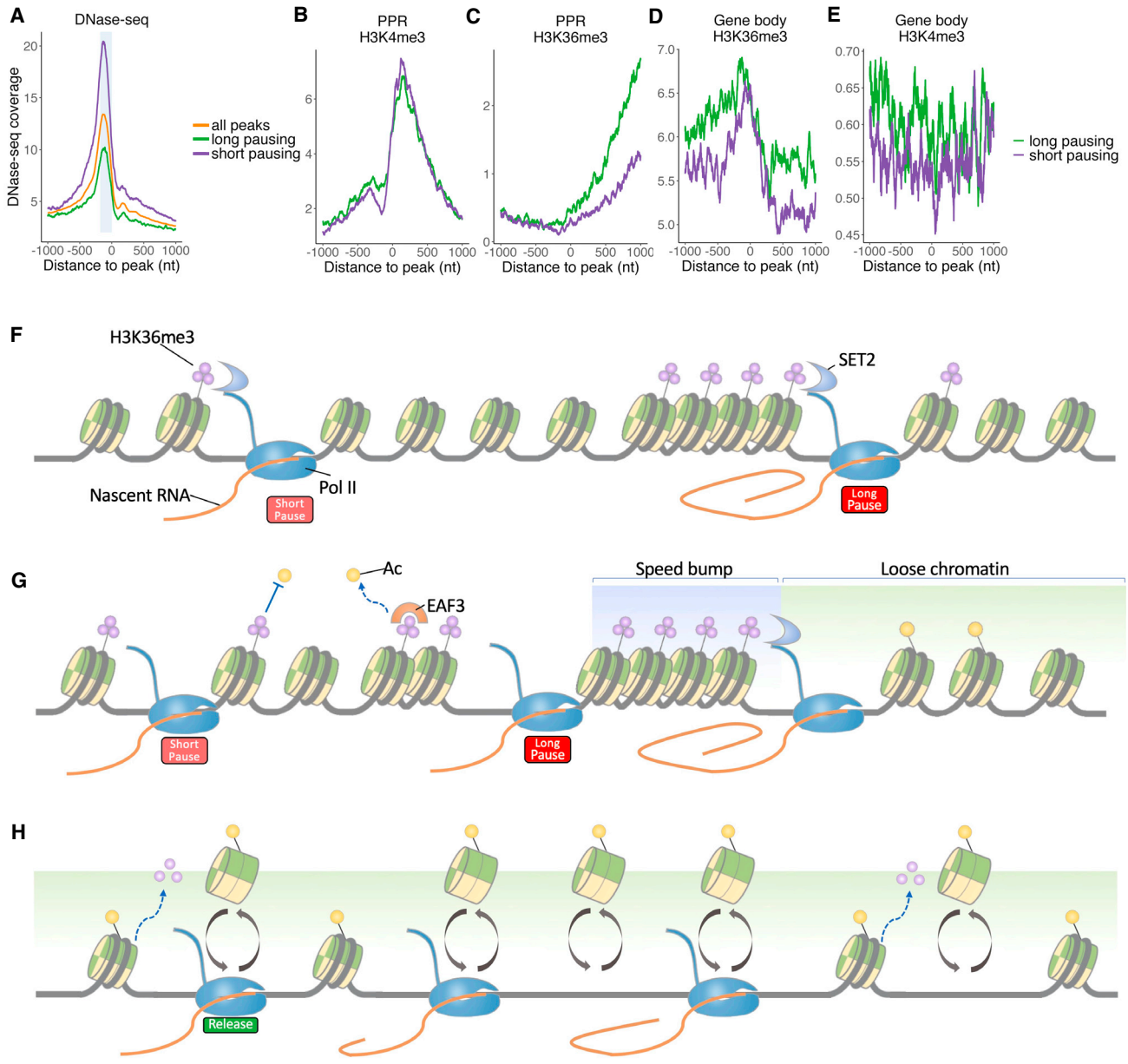


Figure 7. Chromatin state and pausing times

(A) Peaks were classified as long and short according to their pausing times. The average signal of DNase-seq data is displayed in the vicinity of the two classes of peaks (and all peaks). The region from −180 to the peak is shaded in light blue.

(B) Peaks were classified as in (A); signal profiles of H3K4me3 ChIP-seq data of peaks within first 500 bp of gene are shown.

(C) Similar to (B), for H3K36me3.

(D) Similar to (B), for the peaks within the gene body (except the first 2,000 bp and last 1,500 bp of gene).

(E) Similar to (D), for H3K4me3.

(F) CTD of paused Pol II recruits SET2 to trimethylated H3K36. H3K36me3 level increases if the pausing lasts longer.

(G) Model of the dynamic equilibrium between H3K36me3 and histone acetylation under homeostasis. Two H3K36me3-related pausing sites have been set. Packaged H3K36me3 can form into a "speed bump", which establishes long pausing, while shorter pausing might correspond to isolated marks. The Pol II CTD can recruit SET2 to methylate H3K36. The H3K36me3 then can facilitate deacetylation of histones (by active EAF3) and/or inhibit histone acetylation.

(H) Histone acetylation releases paused polymerase after removal of H3K36me3, resulting in a transcriptional burst.

at single-nucleotide resolution, genome-wide (Figures 2A and 2B). This advantage allowed us to analyze pausing times in much greater detail and bigger scale than what was previously achieved. We found that, unlike peaks in the first 120 bp of genes, pausing in the +180 to +320 region appears resistant to sarkosyl treatment (Figure 4A). Also, we were able to investigate

pausing times at locations far from TSSs; TV-PRO-seq revealed pausing times of peaks further downstream within genes and even beyond the genes' 3′ ends (Figure 3F). Trp-based studies would not be able to easily show this due to their focus on the PPR (Figure 2B).

Beyond these Pol II-related analyses, TV-PRO-seq is suited to examine transcription by other RNA polymerases. Exploiting this potential revealed that Pol III has significantly shorter pausing times than Pol II and Pol I (Figure 2G). This highlights the relevance of pausing for transcription by the two former polymerase types, too.

TV-PRO-seq also allowed us to integrate the pausing time profiles with other genome-wide data. By grouping genes with different expression levels and different transcriptional noise levels according to droplet single-cell RNA-seq data (Macosko et al., 2015), we were able to uncover an intriguing relationship between expression and pausing; that is, highly expressed genes have not only more pausing sites (Figure 5A) but also longer pausing times at each of these (Figure 5C). For noisier genes, we find both more (Figure 6A) and longer pausing (Figure 6C); this is consistent with predictions of modeling works (Rajala et al., 2010; Ribeiro et al., 2010; Dobrzynski and Bruggeman, 2009). We further integrated pausing times with ChIP-seq data of histone markers and found the active transcription marker H3K36me3 to correlate with pausing in the gene body (Figure 7E). We propose that paused polymerase could recruit SET2 for methylation of H3K36 and/or H3K36me3 pauses polymerase by repressing histone acetylation (Figures 7F–7H).

In summary, TV-PRO-seq provides a powerful tool to time polymerase pausing. It permits genome-wide estimation of pausing release times at single-base resolution. Our analyses illustrate the rich new insights that can be obtained with our approach with regard to the different polymerase types, the dynamics associated with the different pausing sites, the chromatin state, and, more generically, the process of stochastic transcription. These findings would be hard to obtain with competing techniques, such as NET-seq, which reflect only the polymerase occupancy. Our data provide promising starting points for further investigations, including the study of pausing at short genes such as tRNA and lncRNA loci, the mechanisms involved in pausing, and several other related subjects.

Limitations of the study

TV-PRO-seq is based on PRO-seq; it can only reflect pausing profiles *in vitro* as the nucleotide run-on of biotin-NTP is performed in permeabilized cells. Plus, polymerase that is paused in the region very close to TSSs cannot be fully detected as short reads are removed before alignment.

TV-PRO-seq requires preparation of $\geq$4 different parallel PRO-seq samples with independent cell permeabilization and library building efforts; this is a source of noise in terms of sample variation, along with technical noise that commonly affects PRO-seq data quality. As TV-PRO-seq provides single-base resolution on genome-wide scale, even high sequencing depths will result in relatively low read counts on individual pausing sites, limiting sensitivity and precision. This effect can be alleviated, though, by considering and analyzing ensembles of pausing sites.

While we exploited the effects of sarkosyl to perturb pausing in some TV-PRO-seq samples, it should typically not be included in TV-PRO-seq, lest polymerase is artificially released via the run-on buffer as an unspecific side effect. This makes library building more demanding than with conventional PRO-seq. Also note that FP might generally be a better choice than sarkosyl for investigating links between NELF and pausing owing to its higher specificity.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Processing of sequencing data
  - Peak calling
  - Inference of single-nucleotide transcription rates
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Peak annotation to 3′ and 5′ ends of exons
  - Consistency of TV-PRO-seq and Trp treatment following PRO-seq
  - Peak annotation within genic regions
  - Metagene analysis about pausing peaks
  - Gene expression level and transcriptional noise estimation and selection
  - Histone modification and chromatin accessibility for TV-PRO-seq data
  - Histone modification and chromatin accessibility for mNET-seq data

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100083.

ACKNOWLEDGMENTS

We thank Andrew Nelson and Keith Leppard for reading the manuscript and making valuable suggestions. This work was supported by BBSRC grants BB/L006340/1 and BB/M017982/1 and EPSRC grant EP/T002794/1. Parts of the work were carried out by M.C. during an earlier affiliation with the Department of Statistics, University of Warwick.

AUTHOR CONTRIBUTIONS

J.Z. designed the study and carried out experimental work. J.Z., M.C., and D.H. analyzed the data, carried out theoretical work, and wrote the manuscript. D.H. supervised the work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 3, 2019
Revised: August 3, 2021
Accepted: August 18, 2021
Published: September 27, 2021

REFERENCES

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. *13*, 720–731.

Al-Rfou, R., Guillaume, A., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., et al. (2016). Theano: A Python Framework for Fast Computation of Mathematical Expressions (arXiv), 1605.02688.

Angers-Loustau, A., Petrillo, M., Bengtsson-Palme, J., Berendonk, T., Blais, B., Chan, K.G., Coque, T.M., Hammer, P., Hess, S., Kagkli, D.M., et al. (2018). The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. F1000Res *7*, ISCB Comm J-459.

Bintu, L., Ishibashi, T., Dangkulwanich, M., Wu, Y.-Y., Lubkowska, L., Kashlev, M., and Bustamante, C. (2012). Nucleosomal elements that control the topography of the barrier to transcription. Cell *151*, 738–749.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P., and Workman, J.L. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581–592.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep. *2*, 1025–1035.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Dar, R.D., Shaffer, S.M., Singh, A., Razooky, B.S., Simpson, M.L., Raj, A., and Weinberger, L.S. (2016). Transcriptional bursting explains the noise-versus-mean relationship in mRNA and protein levels. PLoS One *11*, e0158298.

Dobrzynski, M., and Bruggeman, F.J. (2009). Elongation dynamics shape bursty transcription and translation. Proc. Natl. Acad. Sci. U S A *106*, 2583–2588.

Dumay-Odelot, H., Durrieu-Gaillard, S., El Ayoubi, L., Parrot, C., and Teichmann, M. (2014). Contributions of in vitro transcription to the understanding of human RNA polymerase III transcription. Transcription *5*, e27526.

Erickson, B., Sheridan, R.M., Cortazar, M., and Bentley, D.L. (2018). Dynamic turnover of paused Pol II complexes at human promoters. Genes Dev. *32*, 1215–1225.

Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol. *15*, R69.

Fuchs, G., Voichek, Y., Rabani, M., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2015). Simultaneous measurement of genome-wide transcription elongation speeds and rates of RNA polymerase II transition into active elongation with 4sUDRB-seq. Nat. Protoc. *10*, 605–618.

Galvani, A., and Thiriet, C. (2015). Nucleosome dancing at the tempo of histone tail acetylation. Genes *6*, 607–621.

Gariglio, P., Bellard, M., and Chambon, P. (1981). Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. Nucleic Acids Res. *9*, 2589–2598.

Gilchrist, D.A., Dos Santos, G., Fargo, D.C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell *143*, 540–551.

Gressel, S., Schwalb, B., Decker, T.M., Qin, W., Leonhardt, H., Eick, D., and Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. eLife *6*, e29736.

Grothendieck, G. (2013). Non-linear Regression with Brute Force (R package) version 0.2. .

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88.

He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat. Biotechnol. *33*, 395.

Henriques, T., Gilchrist, D.A., Nechaev, S., Bern, M., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2013). Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. Mol. Cell *52*, 517–528.

Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M., and Bustamante, C. (2009). Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science *325*, 626–628.

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

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 167–177.

Kang, J.Y., Mishanina, T.V., Bellecourt, M.J., Mooney, R.A., Darst, S.A., and Landick, R. (2018). RNA polymerase accommodates a pause RNA hairpin by global conformational rearrangements that prolong pausing. Mol. Cell *69*, 802–815 e5.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kireeva, M.L., Hancock, B., Cremona, G.H., Walter, W., Studitsky, V.M., and Kashlev, M. (2005). Nature of the nucleosomal barrier to RNA polymerase II. Mol. Cell *18*, 97–108.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. *41*, 376–381.

Krebs, A.R., Imanci, D., Hoerner, L., Gaidatzis, D., Burger, L., and Schubeler, D. (2017). Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. Mol. Cell *67*, 411–422 e4.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.

Levine, M. (2011). Paused RNA polymerase II as a developmental checkpoint. Cell *145*, 502–511.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, X., Kraus, W.L., and Bai, X. (2015). Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. Trends Biochem. Sci. *40*, 516–525.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016a). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat. Protoc. *11*, 1455–1476.

Mahat, D.B., Salamanca, H.H., Duarte, F.M., Danko, C.G., and Lis, J.T. (2016b). Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. Mol. Cell *62*, 63–78.

Maizels, N.M. (1973). The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of *Escherichia coli*. Proc. Natl. Acad. Sci. U S A *70*, 3585–3589.

Mapendano, C.K., Lykke-Andersen, S., Kjems, J., Bertrand, E., and Jensen, T.H. (2010). Crosstalk between mRNA 3' end processing and transcription initiation. Molecular Cell *40*, 410–422.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. *17*. https://doi.org/10.14806/ej.17.1.200.

Mayer, A., Di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541–554.

Mayer, A., Landry, H.M., and Churchman, L.S. (2017). Pause & go: from the discovery of RNA polymerase pausing to its functional implications. Curr. Opin. Cell Biol. *46*, 72–80.

Morgan, M.A.J., Rickels, R.A., Collings, C.K., He, X., Cao, K., Herz, H.M., Cozzolino, K.A., Abshiru, N.A., Marshall, S.A., Rendleman, E.J., et al. (2017). A cryptic Tudor domain links BRWD2/PHIP to COMPASS-mediated histone H3K4 methylation. Genes Dev. *31*, 2003–2014.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat. Genet. *39*, 1507.

Nilson, K.A., Lawson, C.K., Mullen, N.J., Ball, C.B., Spector, B.M., Meier, J.L., and Price, D.H. (2017). Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. Nucleic Acids Res. *45*, 11088–11105.

Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. Cell *161*, 526–540.

Oler, A.J., Alla, R.K., Roberts, D.N., Wong, A., Hollenhorst, P.C., Chandler, K.J., Cassiday, P.A., Nelson, C.A., Hagedorn, C.H., Graves, B.J., and Cairns, B.R. (2010). Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. Nat. Struct. Mol. Biol. *17*, 620–628.

Patel, M.C., Debrosse, M., Smith, M., Dey, A., Huynh, W., Sarai, N., Heightman, T.D., Tamura, T., and Ozato, K. (2013). BRD4 coordinates recruitment of pause release factor P-TEFb and the pausing complex NELF/DSIF to regulate transcription elongation of interferon-stimulated genes. Mol. Cell. Biol. *33*, 2497–2507.

Peterlin, B.M., and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Mol. Cell *23*, 297–305.

Polanski, K., Gao, B., Mason, S.A., Brown, P., Ott, S., Denby, K.J., and Wild, D.L. (2018). Bringing numerous methods for expression and promoter analysis to a public cloud computing service. Bioinformatics *34*, 884–886.

Porrua, O., and Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. Nat. Rev. Mol. Cell Biol. *16*, 190.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445.

Rajala, T., Hakkinen, A., Healy, S., Yli-Harja, O., and Ribeiro, A.S. (2010). Effects of transcriptional pausing on gene expression dynamics. PLoS Comput. Biol. *6*, e1000704.

Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three *Drosophila* heat shock genes. Proc. Natl. Acad. Sci. U S A *90*, 7923–7927.

Ribeiro, A.S., Hakkinen, A., Healy, S., and Yli-Harja, O. (2010). Dynamical effects of transcriptional pause-prone sites. Comput. Biol. Chem. *34*, 143–148.

Rougvie, A.E., and Lis, J.T. (1988). The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of *D. melanogaster* is transcriptionally engaged. Cell *54*, 795–804.

Saba, J., Chua, X.Y., Mishanina, T.V., Nayak, D., Windgassen, T.A., Mooney, R.A., and Landick, R. (2019). The elemental mechanism of transcriptional pausing. eLife *8*, e40981.

Saldi, T., Fong, N., and Bentley, D.L. (2018). Transcription elongation rate affects nascent histone pre-mRNA folding and 3′ end processing. Genes Dev. *32*, 297–308.

Salvatier, J., Wiecki, T.V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. PeerJ Comput. Sci. *2*. https://doi.org/10.7717/peerj-cs.55.

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

Stasevich, T.J., Hayashi-Takanaka, Y., Sato, Y., Maehara, K., Ohkawa, Y., Sakata-Sogawa, K., Tokunaga, M., Nagase, T., Nozaki, N., McNally, J.G., and Kimura, H. (2014). Regulation of RNA polymerase II activation by histone acetylation in single living cells. Nature *516*, 272–275.

Steurer, B., Janssens, R.C., Geverts, B., Geijer, M.E., Wienholz, F., Theil, A.F., Chang, J., Dealy, S., Pothof, J., Van Cappellen, W.A., et al. (2018). Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA polymerase II. Proc. Natl. Acad. Sci. U S A *115*, E4368–E4376.

Szlachta, K., Thys, R.G., Atkin, N.D., Pierce, L.C.T., Bekiranov, S., and Wang, Y.H. (2018). Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. Genome Biol. *19*, 89.

Titov, D.V., Gilman, B., He, Q.L., Bhat, S., Low, W.K., Dang, Y., Smeaton, M., Demain, A.L., Miller, P.S., Kugel, J.F., et al. (2011). XPB, a subunit of TFIIH, is a target of the natural product triptolide. Nat. Chem. Biol. *7*, 182–188.

Tome, J.M., Tippens, N.D., and Lis, J.T. (2018). Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. Nat. Genet. *50*, 1533–1541.

Veloso, A., Kirkconnell, K.S., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. Genome Res. *24*, 896–905.

Venkatesh, S., Smolle, M., Li, H., Gogol, M.M., Saint, M., Kumar, S., Natarajan, K., and Workman, J.L. (2012). Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. Nature *489*, 452–455.

Venkatesh, S., and Workman, J.L. (2015). Histone exchange, chromatin structure and the regulation of transcription. Nat. Rev. Mol. Cell Biol. *16*, 178.

Vihervaara, A., Duarte, F.M., and Lis, J.T. (2018). Molecular mechanisms driving transcriptional stress responses. Nat. Rev. Genet. *19*, 385–397.

Vihervaara, A., Mahat, D.B., Guertin, M.J., Chu, T., Danko, C.G., Lis, J.T., and Sistonen, L.J.N.C. (2017). Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat. Commun. *8*, 255.

Vvedenskaya, I.O., Vahedian-Movahed, H., Bird, J.G., Knoblauch, J.G., Goldman, S.R., Zhang, Y., Ebright, R.H., and Nickels, B.E. (2014). Interactions between RNA polymerase and the "core recognition element" counteract pausing. Science *344*, 1285–1289.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G.A., Winston, F., et al. (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes Dev. *12*, 343–356.

Wan, Y., Arens, C.E., Wang, S., Zuo, X., Zhuo, Y., Xing, J., and Liu, H. (2013). Role of the repressor Oaf3p in the recruitment of transcription factors and chromatin dynamics during the oleate response. Biochem. J. *449*, 507–517.

Wickham, H. (2011). ggplot2[J]. Wiley Interdisciplinary Reviews: Computational Statistics *3*, 180–185.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).

Willis, I.M., and Moir, R.D. (2018). Signaling to and from the RNA polymerase III transcription and processing machinery. Annu. Rev. Biochem. *87*, 75–100.

Yamaguchi, Y., Shibata, H., and Handa, H. (2013). Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. Biochim. Biophys. Acta *1829*, 98–104.

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell *97*, 41–51.

Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat. Struct. Mol. Biol. *20*, 259–266.

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
DEPC water	Invitrogen	PN: 10514065
NaCl	Sigma-Aldrich	PN: S9888
KCl	Sigma-Aldrich	PN: P9333
MgCl2 . 6H2O	Sigma-Aldrich	PN: M2670
EDTA	Sigma-Aldrich	PN: E9884
NH4Ac	Sigma-Aldrich	PN: A1542
MgAc2	Sigma-Aldrich	PN: M5661
EGTA	Sigma-Aldrich	PN: E3889
Sarkosyl	Sigma-Aldrich	PN: L5125
Sucrose	Sigma-Aldrich	PN: S0389
NaOH	Fisher Chemical	PN: 10396240
DTT	Sigma-Aldrich	PN: D0632
Glycerol	Sigma-Aldrich	PN: G5516
DMSO	Sigma-Aldrich	PN: 41640
TWEEN-20	Sigma-Aldrich	PN: P9416
Triton X-100	Sigma-Aldrich	PN: X100
Tris-HCl pH 6.8 1M	AESAR	PN: J63831.K2
Tris-HCl pH 7.4 1M	AESAR	PN: J60202.K2
Tris-HCl pH 8.0 1M	AESAR	PN: J62726.K2
1 X PBS pH 7.4	Gibco	PN: 10728775
Absolute Ethanol	Fisher Chemical	PN: BP2818
Isopropanol	Fisher Chemical	PN: BP2816
Chloroform	Fisher Chemical	PN: 10488400
Biotin-11-CTP	PerkinElmer	PN: NEL542001EA
Biotin-11-UTP	PerkinElmer	PN: NEL543001EA
Biotin-11-ATP	PerkinElmer	PN: NEL544001EA
Biotin-11-GTP	PerkinElmer	PN: NEL545001EA
ATP	New England Biolabs	PN: P0756S
P-30 column	Bio-Rad	PN: 732-6250
Streptavidin Dynabeads M-280	Invitrogen	PN: 10465723
Trizol	Invitrogen	PN: 15608948
Trizol LS	Invitrogen	PN: 15867521
GlycoBlue	Invitrogen	PN: 10301575
Phenol:chloroform	Sigma-Aldrich	PN: 77617
SUPERase RNase inhibitor	Invitrogen	PN: 10773267
T4 RNA ligase I	New England Biolabs	PN: M0204
RppH	New England Biolabs	PN: M0356
T4 Polynucleotide Kinase	New England Biolabs	PN: M0201L
Superscript III	Invitrogen	PN: 12087539
dNTP mix	New England Biolabs	PN: N0447
Q5 master mix	New England Biolabs	PN: M0544
TEMED	Fisher Chemical	PN: 10549960

(*Continued on next page*)

CellPress
OPEN ACCESS

Cell Reports Methods
Article

RESOURCE AVAILABILITY

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
APS	Sigma-Aldrich	PN: A3678
30% Acrylamide	Sigma-Aldrich	PN: A3449
Orange loading dye 6X	New England Biolabs	PN: B7022
SYBR Gold	Invitrogen	PN: 10358492
Deposited data		
Raw and analyzed data	This paper	GEO: GSE118957
HEK293 NELF ChIP-seq data	Published data	GEO: GSE109652
HEK293 H3K4me1, H3K4me2, H3K4me3 and H3K27ac ChIP-seq data	Published data	GEO: GSE101646
HEK293 H3K36me3 ChIP-seq data	Published data	ENCODE: ENCSR372WXC
HEK293 DNase-seq data	Published data	ENCODE: ENCSR000EJR
Experimental models: Cell lines		
Human: HEK293	Mapendano et al. 2010	β pA+
Oligonucleotides		
VRA3: GAUCGUCGGACUGUAGAACUCUGAAC- / inverted dT/	IDT	N/A
VRA5: CCUUGGCACCCGAGAAUUCCA	IDT	N/A
RP1: AATGATACGGCGACCACCGAGATCTACAC GT TCAGAGTTCTACAGTCCGA	IDT	N/A
RPI-N: CAAGCAGAAGACGGCATACGAGAT NNNNNN GTGACTGGAGTT CCTTGGCACCCGAGAATTCCA	IDT	N/A
Software and algorithms		
Cutadapt	Martin, 2011	https://cutadapt.readthedocs.io/en/stable/
hisat2	Kim et al., 2015	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	Li et al. 2009	http://samtools.sourceforge.net/
Bedtools	Quinlan and Hall 2010	https://bedtools.readthedocs.io/en/latest/index.html
ggplot2	Wickham 2011	https://cran.r-project.org/web/packages/ggplot2/index.html
Custom code	This paper	Github: https://github.com/Jiezhangwarwick/TV-PRO-seq https://doi.org/10.5281/zenodo.5201598

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Daniel Hebenstreit (D.Hebenstreit@warwick.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- TV-PRO-seq data and Trp treatment following PRO-seq have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at Github and Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HEK293 cells were grown to 60% confluency at 37°C and 5% CO_2 in a 175 cm^2 flask in DMEM supplemented with 10% FBS. One day before permeabilization of cells, the culture medium was replaced with fresh medium. For Trp treatment, Trp was added at a concentration of 500 nM, then cells were incubated at 37°C for 10 min, followed by permeabilization. Cell permeabilization was carried out following the PRO-seq protocol (Mahat et al., 2016a). Permeabilized cells were stored at −80°C.

METHOD DETAILS

The permeabilized cells were placed on 37°C for 3 min for thawing. Thawed cells were further processed by adding biotin-labeled NTPs. Two replicates of 4-biotin run-on samples were prepared for HEK293 cells following the PRO-seq protocol. Furthermore, duplicates of 4-biotin run-on samples were prepared in HEK293 cells in a run-on buffer without sarkosyl. The main TV-PRO-seq experiment consisted of 4 independent PRO-seq samples of the 4 run-on times 30 sec, 2 min, 8 min and 32 min. For the Trp treatment sample, 8 min run-on with sarkosyl was performed.

After run-on, the experiment followed the PRO-seq protocol (Mahat et al., 2016a). In brief, total RNA was extracted with Trizol LS and further fragmented with 0.2N NaOH on ice for 10 min. Then biotin labelled RNA fragments were enriched by streptavidin beads M-280. The 3′ adaptor was add to RNAs with T4 RNA ligase I. The 5′ adaptor was also ligated to RNA by T4 RNA ligase I after RppH and T4 PNK treatment at 37°C for 1 hour each. The RNA fragments with adapters on both sides were reverse transcribed by Superscript III RT enzyme into cDNA and amplified by Q5 master mix. The DNA products above 130bp were purified via 8% native PAGE gel selection. The final products were quantified by Qubit and then sent for sequencing.

Processing of sequencing data

Sequencing was performed on an Illumina NextSeq 500 for 51bp single end. Raw data was converted into FASTQ format by bcl2fastq with 0 index mismatches allowed.

Reads were trimmed with Cutadapt version 1.14 (Martin, 2011), to remove sequences starting with the adaptor sequence 'TGGAATTCTCGGGTGCCAAGG' from the 3′ end of reads, and reads shorter than 20bp after trimming were discarded:

cutadapt -a TGGAATTCTCGGGTGCCAAGG -m 20 -e 0.05

Trimmed reads were aligned to the best matched position of hg38 genome with Hisat2 version 2.1.0 (Kim et al., 2015) , resulting in alignment rates above 80%:

hisat2 -p 4 -k 1 –no-unal -x ∼/hg38/genome -U data_2.fastq.gz -S data.sam

Because the ends of sequencing reads have lower sequencing quality, Hisat2 uses soft clipping for the reads, which moves the detected pausing site upstream of the actual pausing site. A custom script Sam_enlong.pl was used on the SAM files to extend the soft clipped reads to their original lengths.

Because sequencing depth also has an influence during the process of peak calling of TV-PRO-seq, another script Sam_cutter.pl was used to reduce the 4 TV-PRO-seq SAM files for each PRO-seq sample to the same sizes by randomly selecting a subset of reads for each.

The processed SAM files were further converted to BAM files and were sorted with samtools version 0.1.19 using samtools view -S -b and samtools sort (Li et al., 2009).

The sorted BAM files were then converted to BEDGRAPH files (Quinlan and Hall, 2010). The 5′ end of a read corresponds to the position of the paused polymerase release site on the opposite strand:

Pausing on plus strand: genomeCoverageBed -strand - -5 -bga -ibam

Pausing on minus strand: genomeCoverageBed -strand + -5 -bga -ibam

We then combined the BEDGRAPH files for the various replicates and time points into two files, one for each strand, with the custom script TV_bedGraph_merger.pl. These files corresponded to tables with rows for each position and columns containing the read numbers across the samples, and were used for the further analysis.

Peak calling

We developed a custom procedure for peak calling from single-base resolution strand-specific sequencing experiments such as TV-PRO-seq. Rather generically, we require that the transcription level μ at a peak exceeds a threshold value Q_{bio} which depends on local fluctuations:

$$\mu \geq Q_{bio} \quad \text{(Equation 1)}$$

The actual procedure is based on the aggregated reads from all the experiments at different run-on times and for a specific position (hereafter, such total reads per nt will be simply referred to as the "total reads") and is detailed below.

1. A threshold t for the minimum number of reads on each single genomic position was set. More precisely, genomic positions with total reads higher than t were selected as 'candidate peaks' for further analysis. The basic threshold t has been heuristically set to 10 and will vary with sequencing depth. In addition to this, we discard the candidate peaks if the number of reads is

zero for all the replicates corresponding to a single one run-on time, at least.

2. Secondly, we address the fact that some polymerase pausing regions are wider than one nt (Kwak et al., 2013). An example of such a dispersed pausing region is illustrated in Figure S1A, within a 50-nt fragment of plus strand of chromosome 1. In Figure S1A, we consider the position with most reads in the dispersed pausing region. To deal with this, we exclude a 'candidate peak' if another 'candidate peak' has more reads in its +/− three-nt neighborhood. This ensures that only a single position is selected from a dispersed peak.

For highly expressed genomic regions, it is likely that some positions have a large number of reads (viz., higher than the threshold *t*) and pass selection step 1, even if they correspond to regions with constant elongation rate and do not have significant pausing. Similarly, along the same non-pausing regions, step 2 returns the genomic positions that have the highest amount of reads, even if this is just due to random fluctuations. As an example, the genomic positions marked as purple in the fragment illustrated in Figure S1A correspond to such a case. Therefore, a third step is necessary to filter the candidate peaks that are likely to be located in a region of constant elongation rate but cannot be discarded during steps 1 and 2. We perform a two-(sub)step procedure as explained below.

3.1. The first sub-step consists of assessing the local biological fluctuations in the polymerase occupation and deriving the threshold *Q* of condition (1). We assume that the polymerase occupancy in a constant elongation-rate region follows the Poisson distribution with parameter *b.* As the average elongation rate across the mammalian genome is about 33.3 nt/sec (Jonkers et al., 2014), we expect that, in such non-pausing regions, all the polymerases are released by the time of the first run-on experiment (i.e., 30 seconds); therefore, for these regions, the differences observed between experiments at different run-on times are presumably due to statistical fluctuations, suggesting that we can actually ignore the dependence on run-on time and aggregate the reads across all experiments. We then focus on the reads across the +/−100-nt neighborhood around each candidate peak. Their mean reads, averaged over both the replicates and the 201 nts, yields the expected number of reads *b* per nt (*b* is ideally estimated from the sample mean of read numbers at each of the 201 positions; however, many peaks are close to the TSS, which has many more reads downstream than upstream. To take account of this asymmetry, we assume that all the reads are downstream and average over the half-interval. This overestimates the background noise, and is thus a conservative estimate) (in the neighborhood). Based on a null local Poissonian assumption, as if reads were Poisson distributed with rate *b*, we associate an upper *q*th quantile Q_{bio} to each neighborhood, where the value of *q* is heuristically chosen to control the number of (false positives) bases whose read number exceeds Q_{bio} purely due to statistical fluctuations. Our (rather conservative) choice would be to allow only one false positive in the whole 'active genome'. We define the latter as all positions with at least one read. Since from our experiment there are 111868728 such bases, we heuristically set q=1/111868728.

3.2. Secondly, we need to assess the sequencing noise as a function of the transcription level. To this end, we sequenced one of the replicates (specifically, the second 32-minute run-on replicate) twice, and trimmed the technical replicate with the highest total aligned reads to the same level as the other one. This trick gave us two replicates of identical total aligned reads, from which we computed the average reads for each nt. Further, we gathered the positions whose average read equals a certain number μ and computed their CV^2, which appears to closely follow the fitted standard noise model $CV^2 = A/\mu + B$, and which can be expressed as

$$\varepsilon_\mu \sim N\left(0, \sigma^2(\mu)\right)$$

where

$$\sigma^2(\mu) = A\ \mu + B\mu^2 \qquad \text{(Equation 2)}$$

Based on this model, the (observed) peak read is randomly drawn from

$$X = \mu + \varepsilon_\mu \qquad \text{(Equation 3)}$$

from which it follows that selecting the candidate peaks with more reads than the 0.99th quantile Q_{seq} of the normal distribution centred at Q_{bio} with variance $\sigma^2(\mu)$ satisfies condition (1) with probability 0.99,

$$Q_{seq} = \left\{x : \text{Prob}\left(x > Q_{bio} + \varepsilon_\mu\right) = 0.99\right\}$$

Since we don't know the value of μ to insert into Equation 2, we replace it with either Q_{bio} or the peak reads itself; the first choice underestimates $Q_{seq\ as}$ $Q_{bio} < \mu$ (for all the non-trivial cases) and hence $\sigma^2(Q_{bio}) < \sigma^2(\mu)$, while the second choice has not such a bias as *X* is centred at μ. It is worth noting that there is an alternative but equivalent choice: one can compute the lower quantile of the distribution centred at the peak read *x*, $Q'_{seq}=\{q: \text{Prob}(q < x+\varepsilon)\}$, and require that $Q'_{seq} > Q_{bio}$.

In conclusion, we incorporate the polymerase noise model of point 3.1 and the sequencing noise model of point 3.2 into condition (1) by choosing the candidate peaks such that $x \geq Q_{seq}$, where Q_{seq} depends on Q_{bio}.

Inference of single-nucleotide transcription rates

In this section, we derive a simple Bayesian model for TV-PRO-seq data and detail the procedure to infer the single-nucleotide transcription rates β_i. We are interested in the stochastic dynamics of biotin-NTP incorporation into a nascent mRNA which can be represented as the following simple reaction:

$$\text{nascent mRNA} + \text{biotin} - \text{NTP} \rightarrow \text{biotin} - \text{labeled mRNA}.$$

Such a reaction corresponds to one transcription step and is specific to the genomic position i complementary to the 3′-end nucleotide of the nascent mRNA. Assuming that the biotin-NTP population is large and remains constant during the reaction progress, we obtain

$$\text{nascent mRNA} \xrightarrow{\beta_i} \text{biotin} - \text{labeled mRNA},$$

which occurs at constant single-nucleotide transcription rate β_i. The average time that the PolII spends on the base i is the reciprocal $1/\beta_i$, which we refer to as the *pausing time*.

Let $y_i(t)$ and $x_i(t)$ denote the average populations of nascent-mRNA and biotin-labelled mRNA (specific to the genomic position i), respectively. The following rate equation is satisfied:

$$\frac{d}{dt}x_i(t) = \beta_i y_i(t).$$

As the presence of the biotin prevents further elongation and no new transcription is initiated, $y_i(t)$ naturally decays according to

$$\frac{d}{dt}y_i(t) = -\beta_i y_i(t).$$

Solving this simple system of ODEs with initial conditions

$$x_i(0) = 0,$$

$$y_i(0) = A_i,$$

yields

$$x_i(t) = A_i\left(1 - e^{-t\beta_i}\right),$$

$$y_i(t) = A_i e^{-t\beta_i},$$

which predicts that the average population of the biotin-labelled mRNA increases up to the saturation point A_i while the unlabelled nascent mRNA is depleted according to exponential law.

Our analysis focuses on a subset of genomic positions $i \in S$, which we refer to as *peak* positions, where transcription level saturates to A_i at rate β_i. We speculate that a large number of genomic positions displays negligible pausing with Pol IIs stepping forwards shortly after biotin-NTP treatment and with transcription level concentrating around A_{bck}. We refer to such positions as *background*. Therefore, the expression level of the whole genome $x_{tot}(t) = \sum_{i \in S} x_i(t) + x_{bck}(t)$ grows according to

$$x_{tot}(t) = \sum_{i \in S} A_i\left(1 - e^{-\beta_i t}\right) + A_{bck}\left(1 - e^{-\beta_{bck} t}\right).$$

While we have a model for the average transcription level $x_i(t)$ at genomic position $i \in S$ and run-on time t, the average number of reads $N_i(t)$ depends on the sequencing depth $\kappa(t)$ which is different for each sequencing experiment and therefore depends on the run-on time t, i.e.,

$$N_i(t) = \kappa(t) A_i\left(1 - e^{-\beta_i t}\right).$$

It is convenient to study the ratio $x_i = N_i(t)/N_{tot}(t)$, where $N_{tot}(t) = \kappa(t) x_{tot}(t)$, as the dependence on $\kappa(t)$ cancels out. This represents the expected number of reads from the region of interest (e.g., from a peak position) normalised by the average total-genome reads at the same run-on time t.

We obtain the normalised model

$$x_i(t) = \frac{x_i(t)}{x_{tot}(t)} = \frac{\left(1 - e^{-\beta_i t}\right)}{\sum_{j \in S} \rho_{ij}\left(1 - e^{-\beta_j t}\right) + \rho_{i,bck}\left(1 - e^{-\beta_{bck} t}\right)}, \qquad i \in S,$$

where $\rho_{ij} = A_j/A_i$ and $\rho_{i,bck} = A_{bck}/A_i$. We will later consider an approximated model where the growth curve $x_{tot}(t)$ is described by a single effective rate β_{tot}.

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

The quantities $x_i(t)$, $i \in S$, can be organised into an $|S| \times T$ matrix X where T is the number of predictor observation run-on times. This allows us to use the compact notation

$$X = \left(1 - e^{-\beta^T t}\right) \circ \left[\left(\varrho 1 - e^{-\beta^T t}\right) + \varrho_{bck}^T \left(1 - e^{-\beta_{bck} t}\right)\right]^{\circ -1}, \quad \text{(Equation 4)}$$

where $t = (t_1, t_2, \ldots, t_T)$ is the vector of predictor observation run-on times, $\beta = (\beta_1, \beta_2, \ldots, \beta_{|S|})$ is the vector of rates, and $\varrho = \{\rho_{ij}\}$, $i,j \in S$, and $\varrho_{bck} = (\rho_{1,bck}, \rho_{2,bck}, \cdots \rho_{|S|,bck})$ incorporates the relative saturation points. The notation $A \circ B$ is the Hadamard (element-wise) product of A and B while $A^{\circ -1}$ is the Hadamard inverse of A.

To simplify this model, we use the naïve form

$$N_{tot}(t) = \kappa(t) x_{tot}(t) = \kappa(t) A_{tot}\left(1 - e^{-\beta_{tot} t}\right),$$

which approximates the growth of the average of total reads. The total number of mitochondrial reads x_{chrM} appears to saturate much quicker than the pausing site reads and, to a first approximation, we assume that $x_{chrM} = \kappa(t) A_{chrM}$. We divide the total reads by the chromosome-M reads, and fit the model

$$\frac{x_{tot}(t)}{x_{chrM}} = \rho_{chrM,tot}\left(1 - e^{-\beta_{tot} t}\right), \quad \text{(Equation 5)}$$

where $\rho_{chrM,tot} = A_{tot}/A_{chrM}$, to such data using the random-search algorithm of the nls2 R package (Grothendieck, 2013), which returned a fit with estimated parameters reported in the table below.

	Estimate	Std.err.	t value	Pr(>|t|)
$\rho_{chrM,tot}$	14.993	1.230	12.192	0.000
β_{tot}	4.837	1.985	2.437	0.050

Our choice is to use the exponential model to approximate the growth of the average total-genome reads $N_{tot}(t)$, and study

$$x_i(t) = \frac{1}{\rho_{i,tot}} \frac{\left(1 - e^{-\beta_i t}\right)}{\left(1 - e^{-\beta_{tot} t}\right)}, \quad \text{(Equation 6)}$$

where $i \in S$ and $\rho_{i,tot}$ are parameters fixed by data. In matrix form, we get

$$X = \left(1 - e^{-\beta^T t}\right) \circ \left[\varrho_{tot}^T \left(1 - e^{-\beta_{tot} t}\right)\right]^{\circ -1}, \quad \text{(Equation 7)}$$

where

$$\varrho_{tot} = \left(\rho_{1,tot}, \rho_{2,tot}, \cdots \rho_{|S|,tot}\right).$$

Dividing Equation 6 by $\left(1 - e^{-\beta_{tot} t}\right)$ yields the more intuitive saturation curves of Figure 2A. We then chose the informative prior

$$\beta_{tot} \sim \text{Gamma}(1.1, 1.1),$$

where $\text{Gamma}(\alpha, \beta)$ represents the Gamma distribution with mean α/β and variance α/β^2, which places substantial mass around 1 and little mass around 0^+. The peaks must have an average rate of the same order as the total growth rate, although the growth rates corresponding to pausing elements can be significantly smaller. Based on these considerations we chose the informative priors

$$\beta_1, \beta_2, \ldots, \beta_{|S|} \overset{i.i.d.}{\sim} \text{Gamma}(0.1, 0.1),$$

which have mean and variance equal to 1 and 10, respectively, and place a lot of mass at 0^+.

The next steps consist of incorporating noise and thus defining a Bayesian model to be fitted. We incorporate the noise in the model as follows. The sequencing reads are obtained after several amplification steps and are restricted to be positive. Hence we assume that the observables Y are subjected to multiplicative errors with lognormal distribution, i.e.,

$$Y = X \bullet \varepsilon,$$

where

$$\log \varepsilon \sim N(0, \sigma^2).$$

As $\varepsilon = e^{\sigma Z}$ with $Z \sim N(0,1)$, we get

$$\log Y \sim N(\log X, \sigma^2).$$

To empirically guess a prior distribution for σ given the coefficient of variation of Y, we use the error-propagation formula

$$CV^2Y \approx CV^2\varepsilon,$$

where CV^2Y is estimated from aggregated data. As ε is lognormal, we have

$$CV^2\varepsilon = e^{\sigma^2} - 1,$$

and

$$\sigma^2 \approx \log\left[CV^2Y + 1\right],$$

which suggests the prior

$$\sigma \sim \text{Gamma}(2, 1).$$

An MCMC sampler to fit the model was implemented using the PyMC3 Library for Bayesian Statistical Modeling and Probabilistic Machine Learning (Salvatier et al., 2016). PyMC3 relies on the Theano framework (Al-Rfou and Almahairi, 2016), which allows fast evaluation of matrix expressions, such as those in Equations 4 and 7, and offers the powerful NUTS sampling algorithm to fit models with thousands of parameters. Nevertheless, we aim to infer the growth rate of up to $\sim$ 60000 peaks. To ease the computational burden, we divide the peak list into chunks of $\sim$ 3000 randomly chosen peaks. The simulations were performed on CyVerse computational facilities (Polanski et al., 2018). Further, we averaged the reads over the replicates, and the averages at 32 minutes of run-on time are used as saturation levels.

In addition to the estimates of the peak rates, the method returns estimates of β_{tot} from each chunk. These are very close to the rate 0.1min^{-1} obtained from the half-life measured in Jonkers, Kwak, and Lis (Jonkers et al., 2014). Aggregating the individual-chunk estimates using the laws of total mean and variance yields:

$$\beta_{tot} = 0.139 \pm 0.007\text{min}^{-1}$$

In order to assess the sensitivity with respect to the prior distribution, we also ran the inference procedure using the vague prior distributions:

$$\beta_1, \beta_2, \ldots, \beta_{|S|}, \beta_{tot} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(0.001, 0.001),$$

which results in a wider range of inferred β_i, whilst conserving the overall rank order.

QUANTIFICATION AND STATISTICAL ANALYSIS

Peak annotation to 3′ and 5′ ends of exons

Two reference lists were used for annotating the ends of the target regions. For mRNA genes, the list was downloaded from UCSC table browser with parameters: assembly - hg38, group - mRNA and EST, table - UCSC RefSeq, output format – All fields from selected table (Kent et al., 2002). The 5′ and 3′ ends of all exons from the mRNA list were transformed into another table with the custom script Unique_annotation_maker.pl. Column 1 to 3 were the chromosome, position and strand of annotation site, respectively; column 4 was the gene name; if the column 5 'type' equals 'start', it means it is the 5′ end of exon, otherwise it is 3′ end; the column 6 'number_min' and 7 'number_max' are the min and max number of exons in different variants of the same gene, respectively, and the TES are marked as -1; The column 8 'hit' shows how many variants of a transcript have this splicing site and column 9 'variant' refers to the number of transcript variants the gene has.

The two annotation files were used for annotating peaks by another custom script Peak_annotater.pl, which identifies peaks located within a specified distance of the annotation site. For example, we can detect the peaks located in a ±4500 nt region of all the 5′ and 3′ ends of UCSC refgene mRNA genes with the following command:

perl Peak_annotater.pl All_mRNA Beta_summary 4500

The peaks that were annotated to have 'type' equal to 'start', 'number_max' equal to 1 and 'hit' equal to 'variant' were those near the TSS of genes with unique TSSs. The sense and antisense reads around these unique TSSs were used to generate the density plot using the ggplot2 package (Wickham, 2016) for *R* (Figure 2C).

Consistency of TV-PRO-seq and Trp treatment following PRO-seq

The top 2000 genes with the highest reads in the first 500bp 5′ were considered as expressed genes. 500 genes each were then classified as 'Long pausing PPR' or 'Short pausing PPR' according to the fold change of reads after 10 minutes Trp treatment. 702 peaks were identified in the long pausing PPR and 493 peaks were found in the short pausing PPR, Exact Binomial Test, $P < 10^{-8}$ (Figure 2E). Furthermore, the peaks within 'Long pausing PPR' had longer pausing times, Mann-Whitney U test, $P < 0.01$ (Figure 2F).

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

Peak annotation within genic regions

For mRNA transcripts by Pol II, UCSC2bed.pl was used on the same UCSC list as above, and for rRNA transcripts by Pol I, the script rFAM_region.pl was used for transforming the merged list from RNAcentral. Pol III target regions were taken from published data (Oler et al., 2010); we used the 'Potential Pol3 targets' table and converted it to human genome assembly GRCh38 with the UCSC liftOver tool (Kent et al., 2002). The output BED file contained 6 columns: chromosome, start of region, end of region, gene name, gene type/transcript ID and DNA strand.

The custom script Annotation_region.pl was used to extract peaks in the target regions according to the annotation lists generated above. The peaks annotated by Pol I, Pol II and Pol III were compared to the peaks detected on chrM in terms of their pausing time distributions. All pairwise comparisons except Pol II vs POLRMT and Pol II vs Pol I (n.s.), $P < 0.01$, Bonferroni-corrected Mann-Whitney U test. These were displayed as violin plots with inserted boxplots using the ggplot2 package for *R* (Figure 2G).

Metagene analysis about pausing peaks

15993 genes which have unique TSSs and TESs and are longer than 3000 nt were used for metagene analysis. We classified the peaks into 7 regions: 1. Promoter, 2. TSS related region, 3. earlier intron, 4. exon, 5. later intron, 6. region before TES and 7. pA related region.

We obtained regions 1, 2, 6 and 7 from the annotations of 3′ and 5′ ends of exons from the list generated with Peak_annotater.pl.

Promoter: 1000-nt region upstream of TSS

TSS related region: 1000-nt region downstream of TSS

region before TES: 500-nt region upstream of TES

pA related region: 4500-nt region downstream of TES

The peaks in the introns and exons were annotated with whole_gene_annotater.pl, using the annotation list generated with whole_gene_annotation_list_maker.pl. Only exons and introns not overlapping with the first 1000-nt or last 500-nt of transcripts were selected. If the intron's centre position was in the first half of the gene, we considered an intron to be an early intron. Otherwise we regarded it as a later intron.

Because most exons or introns have different lengths, we normalized the peak densities before plotting. First, the peaks in introns and exons were annotated with the relative location, that is the distance between the peak and the 5′ end of the annotated region, divided by the length of the annotated region. Then we calculated the average length for each region and multiplied it with the relative location.

To show the pausing times of the 7 regions defined above, a smoothed conditional mean plot with LOESS fitting was generated using the ggplot2 *R* package with parameter span=0.1 for the ggplot function (Figures 3E and 3F). We also separately plotted the smoothed conditional mean plot for peaks around TSSs (Figure 4A).

Gene expression level and transcriptional noise estimation and selection

Genes' expression levels were calculated as average UMI counts from single-cell sequencing data (Klein et al., 2015). 4146 genes with unique TSS and TES and at least one pausing peak in the gene body were taken into consideration. Genes within the top 20% of expression levels were identified as highly expressed and the bottom 20% as less expressed. Overall, longer pausing times of pausing peaks were found in highly expressed genes, $P < 10^{-23}$, Mann-Whitney U test. We generated the smoothed conditional mean plots of the 'highly expressed' and the 'less expressed' genes of Figure 5D using the same strategy as 3E. The boxplot data of 5C was extracted from 5D, but the first 500bp were split into promoter proximal (TSS to +120), +2 nucleosome (+180 to +320) and promoter distal (+500 to +1000). Highly expressed genes have longer pausing times at the promoter proximal region and the pA related region, $P < 0.01$, Bonferroni-corrected Mann-Whitney U test. Likewise, promoter distal, intron and TES proximal region, $P < 0.05$, Bonferroni-corrected Mann-Whitney U test.

To estimate transcriptional noise, we computed the 'above Poisson score' η of genes from single-cell sequencing data as in (Klein et al., 2015), where μ is the mean mRNA number for a gene, and CV is its coefficient of variation. We selected genes with the highest and the lowest noise heuristically, taking into account the dependence of η on μ as follows. We processed the single-cell sequencing dataset of (Macosko et al., 2015) with the custom script Rank_eta.pl. This first sorts the genes into a list by their mean expression. It then moves a sliding window of size $WS = 100$ along this list and, at each position of the window, ranks the genes with regards to the value of η and records these ranks. For each gene in the list, a number WS of ranks results, of which the top and bottom ranks are averaged to give the 'noise score'. We refer to genes within the top and bottom 5% noise scores as 'high noise' and 'low noise' genes, respectively. For genes with equal noise scores, this procedure was repeated for $WS = 20$ and $WS = 500$, and rescaling the resulting noise scores to the range 0 to 100, followed by averaging across the three noise scores (Figure S6).

We also generated Figures 6B and 6C similar to Figures 5D and 5C with 'high noise' and 'low noise' genes. High noise genes have longer pausing times ($P < 0.01$, Mann-Whitney U test) in the whole gene body (Figure 5C).

Histone modification and chromatin accessibility for TV-PRO-seq data

We used existing HEK293 cell ChIP-seq data for different histone modifications from published studies and/or public depositories for the analysis. NELF data were obtained from Gene Expression Omnibus, GSE109652 (Saldi et al., 2018); H3K4me1, H3K4me2, H3K4me3 and H3K27ac data were obtained from GSE101646 (Morgan et al., 2017); H3K36me3 and DNase-seq data

Please cite this article in press as: Zhang et al., Timing RNA polymerase pausing with TV-PRO-seq, Cell Reports Methods (2021), https://doi.org/10.1016/j.crmeth.2021.100083

were downloaded from ENCODE series ENCSR372WXC and ENCSR000EJR. The data were first trimmed with Trimmomatic-0.36 with options LEADING:24 TRAILING:24 SLIDINGWINDOW:4:20 MINLEN:20 (Bolger et al., 2014), then aligned to hg38 under –no-spliced-alignment condition by Hisat2 (Kim et al., 2015). The SAM files were converted to BAM files, then to BED files using Samtools (Li et al., 2009) and Bedtools (Quinlan and Hall, 2010), respectively. The read intervals in the BED files were adjusted to the same lengths with the custom script bed_normal_length.pl to make sure the coverages of reads bore equal weights for each read. We then converted the data to BEDGRAPH files with the genomeCoverageBed command from Bedtools, using the flags -bga (Quinlan and Hall, 2010). The BEDGRAPH files were annotated to TSS or pausing peaks with the custom script Liner_bedgraph_v4.pl.

We then classified peaks on nuclear chromosomes into those with the longest 5% and shortest 5% pausing times, and extracted the coverage from the BEDGRAPH files within +/-1000 nt of each peak in both classes. We then removed the top 5% of these coverage intervals since these had disproportionately strong influence on the results. The peaks were further classified by the position within genes. PPR refers to the first 500bp of a gene and gene body to the region after +1500 from TSS and before -1500 from TES. Finally, we averaged the coverages of each class, respectively, and displayed the results using ggplot2 in *R* (Figures 7A–7E).

The NELF levels of peaks were defined as the average ChIP-seq coverage in the region +/-80bp of peaks. The peaks within the top 10% of NELF levels were identified as high NELF level and the bottom 10% as low NELF level.

Histone modification and chromatin accessibility for mNET-seq data

HEK293 mNET-seq data were downloaded from Gene Expression Omnibus, GSE61332 (Mayer et al., 2015). We used the UCSC liftOver tool to convert the BEDGRAPH file to hg38 (Kent et al., 2002). We then defined target genes for further analysis by selecting genes longer than 3000 nt, with unique TSSs and TESs. Peak selection for the mNET-seq data followed the same strategy as for TV-PRO-seq; the peak selection output file was processed with the script Liner_bedgraph.pl to extract histone modification states within +/−1000 nt of peaks in the same way as for TV-PRO-seq. We estimate pausing of mNET-seq by the PI (pausing index), which divides the read numbers of peaks by the average read numbers in the gene body (+500 to TES). We removed the top 5% peaks with the highest average coverage of each group and plotted the average coverage of histone modification at peaks corresponding to the top and bottom 5% PI, respectively (for all peaks in target genes, peaks within the TSS to +500 region only, or peaks within the +1500 to TES region only).

In order to compare TV-PRO-seq and mNET-seq with regards to the chromatin state results, we needed to subset the TV-PRO-seq data to the same target genes as we used for the mNET-seq data. The script PI_TV_annotater.pl was used to extract the coverage information of individual TV-PRO-seq peaks located in the target genes. We then selected long pausing and short pausing peaks as above. The average ChIP-seq/DNase-seq coverages of long pausing and short pausing peaks were then used for comparison with the high PI and low PI peaks (Figure S7).