Timing Polymerase Pausing with TV-PRO-seq

Jie Zhang[1], Massimo Cavallaro[1,2], Daniel Hebenstreit[1]*

[1]School of Life Sciences, Gibbet Hill Campus, The University of Warwick, Coventry, CV4 7AL, UK

[2]Department of Statistics, The University of Warwick, Coventry, CV4 7AL, UK

*Correspondence to: D.Hebenstreit@warwick.ac.uk

Transcription of many genes in metazoans is subject to polymerase pausing, which corresponds to the transient arrest of transcriptionally engaged polymerase. It occurs mainly at promoter proximal regions and is not well understood. In particular, a genome-wide measurement of pausing times at high resolution has been lacking.

We present here an extension of PRO-seq, time variant PRO-seq (TV-PRO-seq), that allowed us to estimate genome-wide pausing times at single base resolution. Its application to human cells reveals that promoter proximal pausing is surprisingly short compared to other regions and displays an intricate pattern. We also find precisely conserved pausing profiles at tRNA and rRNA genes and identified DNA motifs associated with pausing time. Finally, we show how chromatin states reflect differences in pausing times.

Transcription of metazoan genes often involves the phenomenon of polymerase pausing, the transient arrest of RNA polymerase II (Pol II) at promoter proximal regions after transcription initiation [1]. While polymerase pausing was discovered early [2-4], 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 [5]. On the other hand, the commonness of pausing, which is observed for roughly a third of genes [1], points towards a more fundamental function in the transcriptional machinery. Several protein factors such as negative elongation factor (NELF) [6] and DRB sensitivity-inducing factor (DSIF) [7] have been found to influence pausing, along with more generic factors, such as DNA sequence and/or nucleosomes (e.g. [8, 9]).

Understanding of polymerase pausing has been greatly advanced by several types of assays based on next generation sequencing, including ChIP-seq, GRO-seq (Global nuclear Run-On sequencing)[9], (m)NET-seq (mammalian Native Elongating Transcript sequencing)[10] and PRO-seq (Precision nuclear Run-On sequencing)[11]. These revealed accumulations of Pol II at positions other than promoter proximal regions, such as exons [12] and 3' ends of genes [9], and led to many other important findings [13]. The 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 proxy for pausing.

A major limitation of all these methods is their inability to discriminate between *few slow* polymerases and *many quick* polymerases detected at a genomic position, since the observations are aggregated over many cells; both cases will result in identical peaks of sequencing reads, which prevents measuring the actual pausing times. The latter is

accomplished only indirectly, at low resolution [14-16]; 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. We developed TV-PRO-seq based on a detailed analysis of the logic underlying PRO-seq. The principle of the latter is to replace native NTPs in the nuclei with biotin-labelled ones (biotin-NTPs), which become incorporated into the 3' end of nascent RNA [17] 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.

Each polymerase in a PRO-seq sample thus moves theoretically a maximum of one position, and movement is a necessary condition for producing a sequencing read. An individual move will occur upon release of polymerase from its original position, one base upstream. The rate of this release will relate inversely to the time the polymerase resides at the upstream position (Fig. 1A). 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. What this means is that individual positions will gradually saturate with reads if a run-on time course is performed. If polymerase pauses, a flat saturation curve will be observed for the downstream position, whereas swift elongation will produce a steeper curve (Fig. 1A).

This is the principle of TV-PRO-seq: preparation of several PRO-seq reactions using different run-on times allows preparation of saturation curves, the slopes of which permit estimation of pausing (-release) times. Depending on sequencing depth and time resolution, TV-PRO-seq potentially has genome-wide single base resolution.

To test TV-PRO-seq, we prepared PRO-seq reactions with 0.5, 2, 8, and 32 min run-on times from human HEK293 cells as independent duplicates, and sequenced all samples to depths of ~50m reads. To confirm successful PRO-seq reactions, we pooled the data after read alignment and selected peaks based on heuristic thresholding (Supp Methods). Plotting the distribution of peaks around transcriptional start sites (TSSs) reproduced the familiar pattern of promoter-proximal peaks and divergent transcription on the other strand (Fig. 1B).

PRO-seq in principal does not discriminate between different types of RNA polymerases, which we can exploit to explore these issues and carry out internal comparisons; transcription of the mitochondrial genome is subject to polymerase pausing as well, but is carried out by a highly processive single-subunit polymerase [18, 19]. We presumed that positions on mtDNA would thus saturate early and so increase only by a small degree or stay approximately constant, allowing normalization of the total data with mtDNA data. We verified this approach and our peak thresholding (Supp Methods, Fig. S1) and constructed a mathematical model that takes account of our theoretical considerations; the model predicts the saturation curve as function of the pausing release rate and the TV-PRO-seq timepoint (Supp. Methods). Fitting this model to the time course of an individual peak allows inference of the pausing release rate as a free parameter, the reciprocal of which yields the pausing time at that position. We embedded this procedure in a Bayesian framework and applied it to the peaks in our data to study the resulting genome-wide pausing times from different angles. Examples for fitted curves to two close individual peaks are shown in Fig. 1C (see Fig. S2 for an alternative normalization).

We note that run-on methods are influenced by technical noise and that GRO- and PRO-seq are based on permeabilized cells, thus not an optimal reflection of the situation *in vivo*; we therefore regard our pausing time figures as estimates, which are however powerful for relative comparisons and aggregate analyses of multiple peaks. TV-PRO-seq also has significant advantages over its main alternative (m)NET-seq for investigating pausing times; pausing results obtained with the latter assay can be difficult to interpret (please see Supplementary Discussion, Fig. S3).

We revisit analysis of different polymerase types as a first application of our approach. Plotting distributions of pausing times for peaks in Pol I, II, and III transcribed loci reveals significant differences among the polymerases. Pol III pausing is shortest, with relatively little variation, while Pol I and II display broader distributions, with higher median pausing times that are greatest for Pol II (Fig. 2A; for all pairwise comparisons except Pol I vs Pol II (n.s.), $P < 10^{-10}$, Bonferroni-corrected Mann-Whitney U test). Mitochondrial polymerase pausing times are more constrained and significantly shorter than on nuclear DNA (Fig. 2A), while individual nuclear chromosomes have similar distributions (Fig. S4).

Promoter proximal pausing has been considered a rate limiting step for transcription due to the higher polymerase occupancies in this region [20]. However, TV-PRO-seq strikingly reveals that promoter proximal pausing is significantly shorter than pausing in other parts of a gene as a metagene profile demonstrates (Fig. 2B). In fact, pausing time follows a serpentine curve, with short pausing within ~100 bp downstream of TSS, followed by increased times over a slightly longer stretch (a magnified section is shown in Fig. S5). While peak densities in exons are higher than in introns as noticed previously (Fig. S6)[21, 22], pausing times in these regions appear to be similar (Fig. 2B). This suggests that the pausing frequency, rather than time, is lower in introns. We also observe an interesting pausing profile at the metagene's 3' end; pausing time appears to be slightly elevated close and upstream to TES, followed by a dip downstream of the poly-adenylation site (Fig. 2B).

We then took a closer look at the 5' region. Taking the reverse approach of the metagene plot and first classifying peaks into short and long pausing times, followed by analysis of their positional distribution, confirms the two regions (Fig. 2C). As an independent verification of this observation, we prepared TV-PRO-seq samples for a different cell line, the human chronic myelogenous leukemia line KBM-7. After applying the same processing and analysis pipeline to these samples as before, we obtained virtually identical results (Fig. S7). As an additional test, we prepared PRO-seq samples at 6 min run-on time after cells had been pre-treated for 10 min with Triptolide (Trp) [14]. Since Trp blocks transcription initiation, the TSS proximal region should become vacated, while the more distal region should remain occupied by polymerase due to the longer pausing time. To test this, we calculated foldchanges for peaks between treated and untreated samples and classified these into 'high' and 'low'. Plotting their positional distributions reveals that peaks with biggest Trp-induced changes are close to TSS, confirming our previous results (Fig. 2D). Our finding of surprisingly short promoter proximal pausing appears to agree with recent studies hinting at rapid Pol II turnover in this region [23, 24]. This suggests that, instead of pausing, this region rather features a high rate of abortive transcription, which might be a standard feature of metazoan Pol II transcription.

We further explored the characteristics of nuclear non-Pol II transcription. Pol III transcribes mainly short structured RNAs, most prominently including tRNAs and 5S rRNA [25]. Pausing profiles of tRNA genes display a remarkably clear picture; short intragenic pausing concentrates in three peaks that appear conserved across genes, while the TES is followed by a region with much longer pausing times (Fig. 3A , Fig. S8). An interesting pausing time pattern also emerges if we focus on Pol I. Pol I transcribes exclusively ribosomal RNAs, which are expressed from operons that are repeated many times on the genome with various variations. An rDNA repeat unit encodes a copy of 18S, 5.8S and 28S rRNA, and several spacer and repeat sequences (Fig. 3B)[26]. Similar to the tRNA genes, the average rDNA operon features several relatively focused pausing locations with varying pausing times, followed by a long-pausing region at the 3' end (Fig. 3B).

We now sought to investigate the relation between pausing and transcriptional bursting by integrating our TV-PRO-seq data with single-cell transcriptomics data. The latter permits analysis of transcriptional dynamics, since bursting will result in more dispersed distributions of mRNAs among cells [27]. This dispersion, or 'noise', is quantified by the CV^2 and is a function

of the mean expression level [28, 29], as is the size of pausing peaks; higher transcription in general means higher polymerase occupancy and thus more PRO-seq reads along a gene.

To this end, we used Drop-seq data for HEK293 cells [30] and classified genes based on their CV^2 for a moving average of mean expression levels to reduce influence of the latter (Fig. S9). We assigned genes to 'low', and 'high noise' classes. We find that, overall, noisy genes have significantly higher peak densities throughout gene bodies (Fig. S10), while pausing times in most genic regions are similar (Fig. 3C). An exception is the region following the promoter proximal dip in pausing times, where Pol II pauses significantly longer at noisy genes (Fig. S11). We term this region the variable pausing region. If we consider the relative distributions of pausing peaks within genes, we observe a mild shift of pausing positions away from the promoter proximal region to other parts, including the variable pausing region and exons (Fig. 3d). These results shift the focus of potential links between polymerase pausing and transcriptional noise away from promoters [16] and towards internal genic regions. This would agree with previous theoretical considerations that proposed this [31-33].

TV-PRO-seq also offers an opportunity to study candidates for DNA motifs that might be involved in pausing. Such motifs have been identified for Drosophila [34], but less is known for human systems in this regard. To investigate this, we extracted 100bp of sequence surrounding each pausing peak and applied a *de novo* motif detection algorithm to the total sequence set. This revealed a list of enriched motifs, which we narrowed down to motifs with a well conserved distance to the pausing peak. We termed these the 'Accurate Pausing Motifs' (APM; Table S1). With the exception of APM5, pausing at all motifs is significantly different (either greater or less) from the overall pausing time distribution at all peaks (Fig. 4A; Mann-Whitney U tests). APM3 is notable for appearing a second time as its reverse complement, but with the pausing peak at a different position. To understand how pausing at these motifs relates to the local pausing environments, we also compared pausing times for the motif-associated peaks and other peaks located between 20 and 40bp away from these. We saw differences now for some APMs, but not all (Fig. S12, Mann-Whitney U tests). This suggests that the various APMs differ when considering the size of regions subject to their possible effects on pausing times.

We chose to focus on the motif with the highest enrichment score, APM1, with the consensus sequence ACAGTCCT (Fig. 4B). This motif appears upstream of pausing positions, with the peak at the last 'C', implying pausing on the adjacent upstream 'C'. The position of this pausing site is highly precise, with large differences between peak frequency at the precise pausing site and the surrounding background; variants of the last 'C' completely abolish pausing (Fig. 4B; all variants are shown in Fig. S13).

Interestingly, if we consider dinucleotide variants of the first two positions, we observe systematic effects of individual bases on the pausing times of the downstream peaks (Fig. 4C). This pattern would be unlikely to appear by chance (Kendall tau test, all $P < 10^{-6}$; background pausing times do not show such a pattern, Fig. S14) and agrees with elementary biochemical considerations relating affinity to lifetime of an interaction; it suggests functional relevance of the motif.

We next investigated the effects of chromatin states on pausing times. To this end we classified peaks into long and short according to their pausing times and quantified their presence around different chromatin features. We found striking 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, while long pausing is shifted away from these regions (Fig. 4D). 'Activating' histone modifications such as H3K4 methylations and H3K27 acetylation exhibit similar profiles (Fig. 4D and Fig. S15). The reverse is seen for repressive chromatin; long pausing is enriched at the heterochromatin marker H3K9me3 [35], while short pausing is strongly reduced (Fig. 4D). We observe a similar

pattern for the H3K36me3 modification, where the pausing time differences interestingly extend over broader regions (Fig. 4D), possibly relating to the repressive role among its diverse functions [36]. 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 (using the Pausing Index to estimate the extent of pausing; see Supplementary Discussion). This demonstrates overall agreement between the two assays, confirming data quality of both, but also reveals intriguing differences; TV-PRO-seq appears to produce clearer profiles in gene bodies and for some histone marks (H3K9me3; this is potentially due to ineffective ChIP-seq from compacted heterochromatin) and shows opposite results for others (H3K36me3; Fig. S16). This illustrates the value of TV-PRO-seq to produce novel insights.

In summary, TV-PRO-seq provides a powerful new 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 in regards to different polymerases, the dynamics associated with different pausing sites, stochastic transcription, and chromatin state; we find that promoter proximal pausing of Pol II is unexpectedly short for the average gene and precedes a longer pausing region further downstream. We observe characteristic patterns also for different polymerases and show how epigenetic marks relate to pausing times in intriguing ways, potentially hinting at unknown mechanisms. These findings would be hard to obtain with competing techniques, such as NET-seq, which do not target actively transcribing polymerase. Our data provide promising starting points for further investigations.

Acknowledgments:
We would like to thank Andrew Nelson and Keith Leppard for reading the manuscript and making valuable suggestions, and Thijn R. Brummelkamp for providing KBM7 cells.

Funding: This work was supported by BBSRC grants BB/L006340/1 and BB/M017982/1;

Author contributions: JZ designed the study and carried out experimental work. JZ, MC and DH analysed the data, carried out theoretical work, and wrote the manuscript. DH supervised the work;

Competing interests:

Authors declare no competing interests;

Data and materials availability:

Data accompanying the study have been deposited at Gene Expression Omnibus, accession number GSE118957. Scripts are available in the supplementary materials.

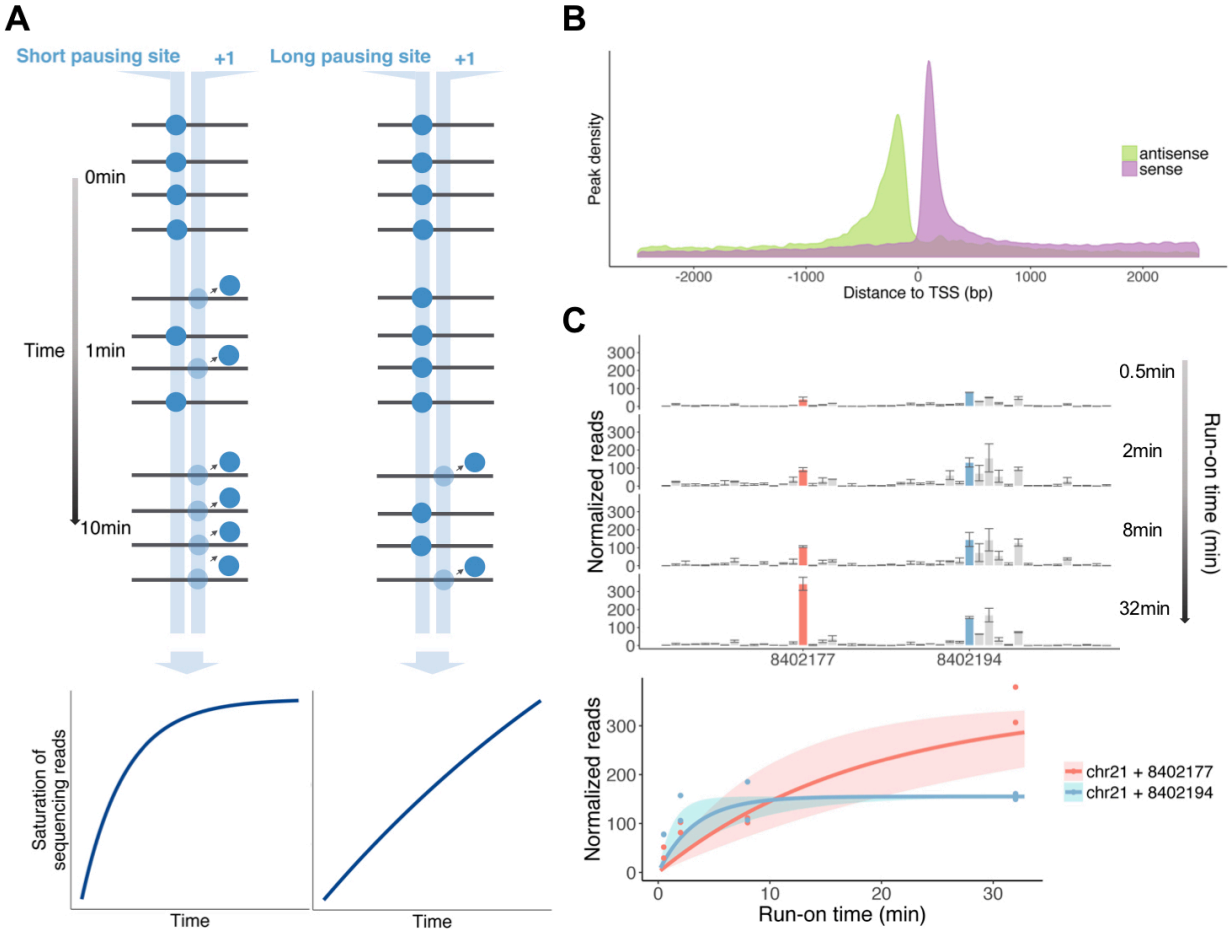

Fig. 1. Principle of TV-PRO-seq (**A**) The black horizontal lines symbolize a particular DNA region. The blue dots symbolize RNA polymerases that either are stationary or that have just moved by one position (and dropped off after incorporating biotinylated NTP), as indicated by the lighter blue shades and the small arrows. A sequencing read can result at a position if a polymerase moves by one base. Eventually, all polymerases will have moved, i.e. all positions have become saturated. If particular positions ('+1') along the genome are considered, the saturation will take longer adjacent to pausing sites, since the polymerase will be released at lower rates. Genome-wide saturation curves for each position can be resolved by run-on time courses. (**B**) Distributions at TSS of sense and antisense reads of pooled TV-PRO-seq samples confirm library qualities. (**C**) Normalized (by chrM reads) and rescaled (by 10^6) read numbers at two example peaks (red, blue) in close vicinity on chr21 at different run-on times (left panel) are shown. Error bars correspond to the values of two biological replicates. The resulting curve fits are shown (right panel). The shaded regions correspond to lower and upper quartiles of the posteriors.

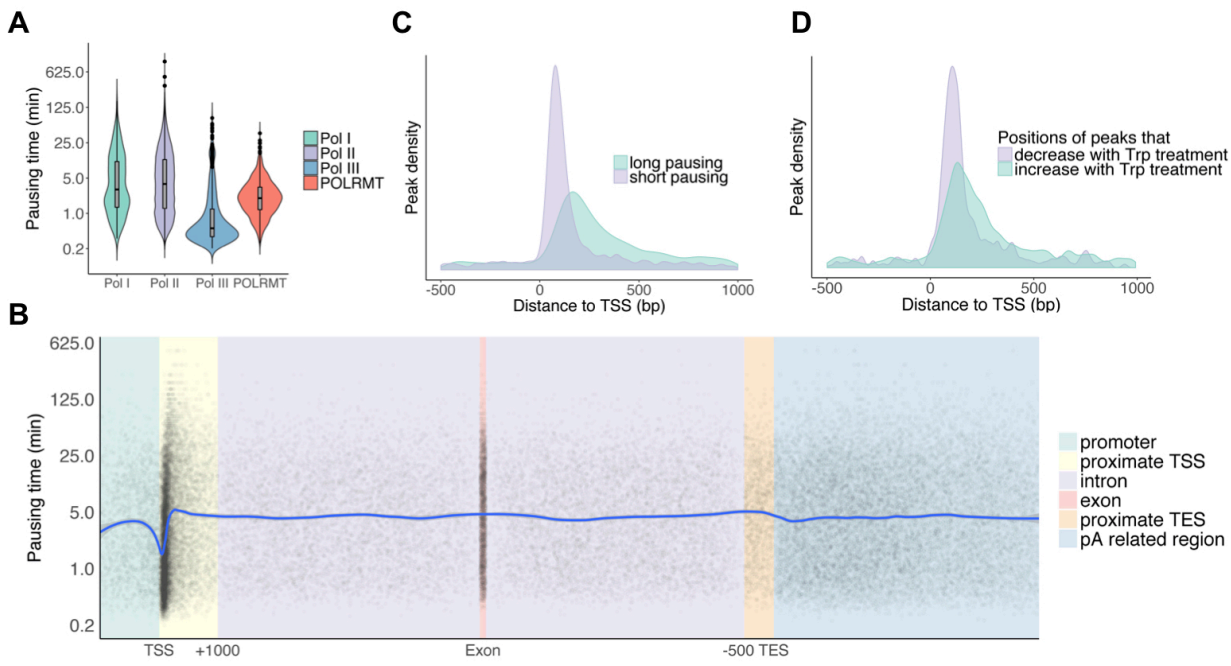

Fig. 2. Pausing times at different loci (**A**) Distributions of estimated pausing times for peaks in loci transcribed by Pol I, II, III and POLMT. For all pairwise comparisons except Pol I vs Pol II (n.s.), $P < 10^{-10}$, Bonferroni-corrected Mann-Whitney U test. (**B**) Pausing times at mRNA-transcribing metagene. Each gray dot represents a pausing peak, with corresponding pausing time given by its y-axis value. The x-axis values corresponds to absolute position within -/+ 1kb of TSS (green and yellow tinged regions, respectively), 500bp upstream and 4.5kb downstream of TES (orange and blue, respectively), or relative position within the other genic sections (color code as indicated). The blue line corresponds to the moving average (LOESS fit). The gray shading indicates the 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. (**C**) Peaks within -500 to +1000 of TSS were classified into 'long' and 'short' according to their pausing times and were displayed as distributions regarding their distances to TSS, $P < 10^{-100}$, Mann-Whitney U test. (**D**) Pre-treatment of cells with Triptolide (Trp) to block transcription initiation leads to differential vacation of pausing sites near TSS; peaks with increased relative sizes after Trp treatment are further from TSS than decreasing peaks, $P < 10^{-10}$, Mann-Whitney U test.

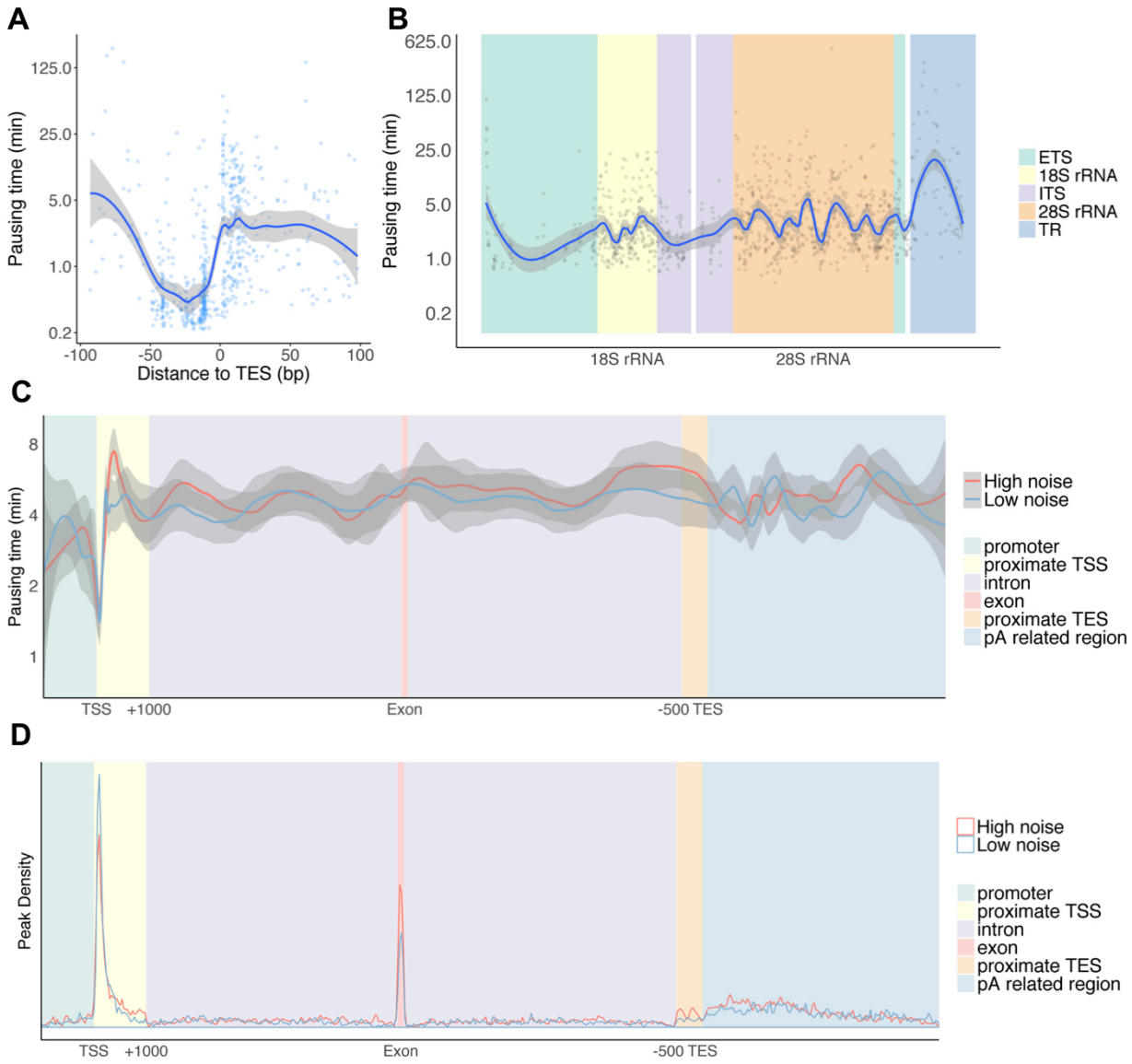

Fig. 3. Pausing characteristics associated with different transcript types. (**A**) Pausing times and positions at tRNA genes. Each dot corresponds to a pausing peak. The blue line corresponds to the moving average with the gray shading indicating the confidence interval (LOESS fit). The metagene is aligned at the TES due to the interesting pausing time increase at TES. Due to their shortness, alignment at TSS looks similar for these genes (Fig. S8). (**B**) As (A), for ribosomal RNA genes. ETS & ITS, external & internal transcribed spacers, respectively (5.8S not shown). TR, tandem repeat (**C**) Pausing times at mRNA-transcribing metagene as in Fig. 2B, for genes classified into different levels of transcriptional noise ('high', 'low'; red, blue, respectively). (**D**) Densities (so that the areas under the peaks are equal for the metagene) of pausing peaks among genic regions for 'low' and 'high' noise genes at the metagene as in (C).

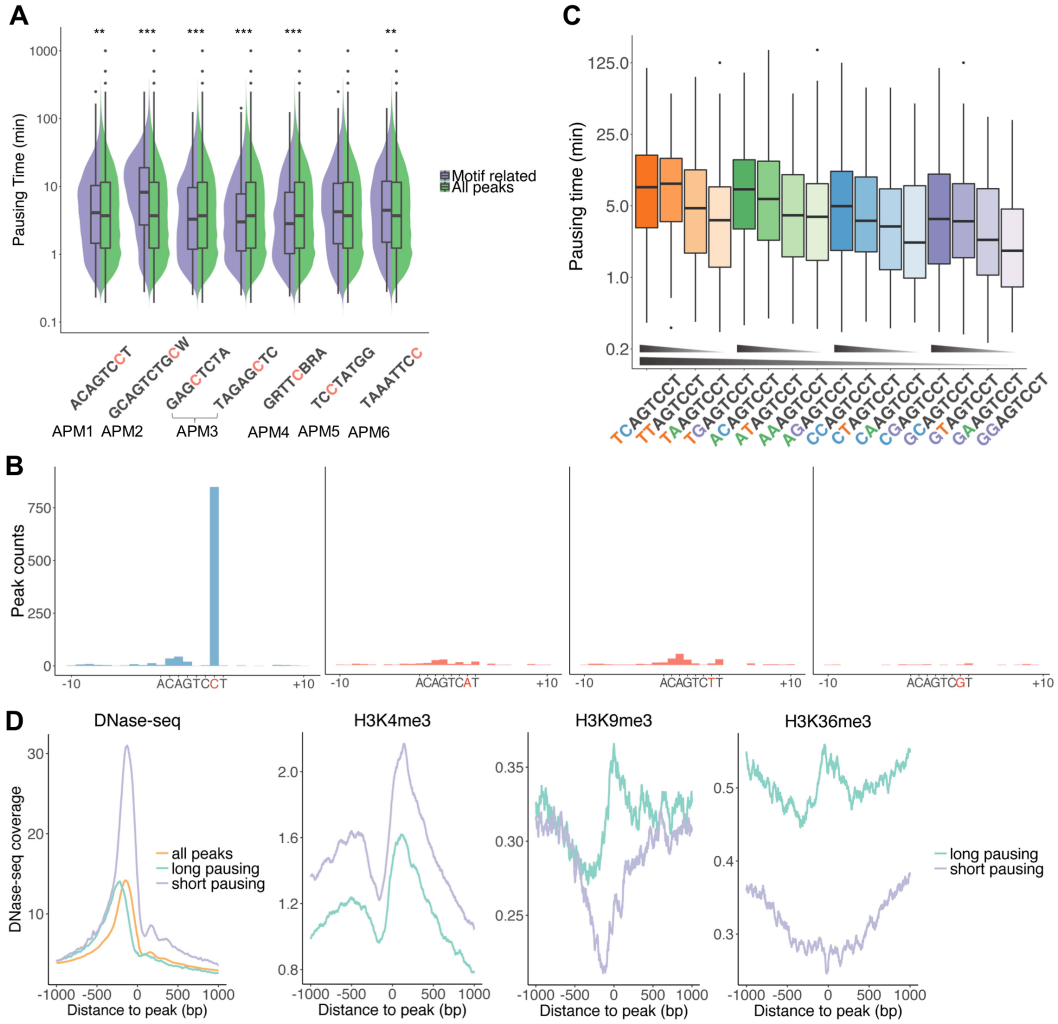

Fig. 4. Genomic features and pausing times. (**A**) Pausing time comparison for enriched motifs at peaks and nearby background sequences. **P<0.01, ***P<0.001, Mann-Whitney U test. (**B**) Histograms of peak frequencies at positions relative to the motif ACAGTCC and single base variants of it at the main position (red). (**C**) Dinucleotide variants of the motif of (A) show systematic effects on the pausing times. Black triangles were added to better illustrate the trends. Trends among all groups of four were assessed with Kendall's tau test and were found to have $P < 10^{-6}$ in all cases (H_1: tau $\neq 0$). (**D**) Peaks were classified into 'long' and 'short' according to their pausing times. The average signal of DNase-seq data and H3K4me3, H3K9me3, and H3K36me3 ChIP-seq data is displayed in the vicinity of the two classes of peaks.

1. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-731 (2012).
2. Rasmussen, E.B. & Lis, J.T. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 7923-7927 (1993).
3. Maizels, N.M. The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3585-3589 (1973).
4. Gariglio, P., Bellard, M. & Chambon, P. Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. *Nucleic acids research* **9**, 2589-2598 (1981).
5. Mahat, D.B., Salamanca, H.H., Duarte, F.M., Danko, C.G. & Lis, J.T. Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Mol Cell* **62**, 63-78 (2016).
6. Yamaguchi, Y. et al. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41-51 (1999).
7. Wada, T. et al. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes & development* **12**, 343-356 (1998).
8. Gilchrist, D.A. et al. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**, 540-551 (2010).
9. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).
10. Churchman, L.S. & Weissman, J.S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368-373 (2011).
11. Kwak, H., Fuda, N.J., Core, L.J. & Lis, J.T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950-953 (2013).
12. Carrillo Oesterreich, F., Preibisch, S. & Neugebauer, K.M. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Molecular cell* **40**, 571-581 (2010).
13. Mayer, A., Landry, H.M. & Churchman, L.S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Current opinion in cell biology* **46**, 72-80 (2017).
14. Jonkers, I., Kwak, H. & Lis, J.T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407 (2014).
15. Gressel, S. et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* **6** (2017).
16. Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nature genetics* **49**, 1045-1051 (2017).
17. Mahat, D.B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc* **11**, 1455-1476 (2016).
18. Posse, V., Shahzad, S., Falkenberg, M., Hallberg, B.M. & Gustafsson, C.M. TEFM is a potent stimulator of mitochondrial transcription elongation in vitro. *Nucleic acids research* **43**, 2615-2624 (2015).
19. Barshad, G., Marom, S., Cohen, T. & Mishmar, D. Mitochondrial DNA Transcription and Its Regulation: An Evolutionary Perspective. *Trends in genetics : TIG* (2018).
20. Liu, X., Kraus, W.L. & Bai, X. Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends Biochem Sci* **40**, 516-525 (2015).
21. Nojima, T. et al. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526-540 (2015).
22. Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**, 541-554 (2015).
23. Krebs, A.R. et al. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Molecular cell* **67**, 411-422 e414 (2017).
24. Steurer, B. et al. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E4368-E4376 (2018).
25. Khatter, H., Vorlander, M.K. & Muller, C.W. RNA polymerase I and III: similar yet unique. *Curr Opin Struct Biol* **47**, 88-94 (2017).
26. Kim, J.H. et al. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res* **46**, 6712-6725 (2018).

27. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4**, e309 (2006).
28. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
29. Dar, R.D. et al. Transcriptional Bursting Explains the Noise-Versus-Mean Relationship in mRNA and Protein Levels. *PLoS One* **11**, e0158298 (2016).
30. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
31. Rajala, T., Hakkinen, A., Healy, S., Yli-Harja, O. & Ribeiro, A.S. Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput Biol* **6**, e1000704 (2010).
32. Ribeiro, A.S., Hakkinen, A., Healy, S. & Yli-Harja, O. Dynamical effects of transcriptional pause-prone sites. *Comput Biol Chem* **34**, 143-148 (2010).
33. Dobrzynski, M. & Bruggeman, F.J. Elongation dynamics shape bursty transcription and translation. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 2583-2588 (2009).
34. Hendrix, D.A., Hong, J.W., Zeitlinger, J., Rokhsar, D.S. & Levine, M.S. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc Natl Acad Sci U S A* **105**, 7762-7767 (2008).
35. Nakayama, J., Rice, J.C., Strahl, B.D., Allis, C.D. & Grewal, S.I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110-113 (2001).
36. Wagner, E.J. & Carpenter, P.B. Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol* **13**, 115-126 (2012).

Methods:

Time variant PRO-seq library building

HEK293 cells were grown to 60% confluency at 37°C and 5% CO_2 in a 175cm^2 flask in DMEM supplemented with 10% FBS. One day before permeabilization of cells, the culture medium was replaced with fresh medium. KBM-7 cells were cultured in the same way, using IMDM instead of DMEM.

Cell permeabilization was carried out following the PRO-seq protocol [17]. Permeabilized cells were stored at -80°C. Prior to treatment or run-on, the cells were placed on 37°C for 3 min for thawing. For triptolide (Trp) treatments, 1µL of 100µM Trp was added to 100µl permeabilized KBM-7 cells and placed in a 37°C inhibitor for 10min. As control, 1 µL DMSO instead of Trp was used for 10min. Thawed cells were further processed by adding biotin-labeled NTPs; a '2-biotin' run-on with biotin-UTP and biotin-CTP was conducted for KBM-7 cells and a 4-biotin run-on for HEK293 cells, following the PRO-seq protocol. For the Trp-treatment experiment, the run-on duration was set to 6min. The main TV-PRO-seq experiment consisted of 8 independent PRO-seq samples: biological duplicates of the 4 run-on times 30sec, 2min, 8min and 32min. After run-on, the experiment followed the PRO-seq protocol [17].

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 [37], 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 [38] , 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 TVPRO-seq, another script `Sam_cutter.pl` was used to reduce the 8 TV-PRO-seq SAM files for HEK293 cells 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` [39],

The sorted bam files were then converted to BEDGRAPH files [40]. 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:

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

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 bp 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 read higher than t were selected as 'candidate peaks' for further analysis. The basic threshold t has been heuristically set to 13 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 bp [11]. An example of such a dispersed pausing region is illustrated in Figure S17A, within a 50bp fragment of plus strand of chromosome 1. In Figure S17A, 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-bp 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, the 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 position 632561 in the fragment illustrated in Figure S17A corresponds 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 the steps 1 and 2. We perform a two-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 33.3bp/sec [14], 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 +/-100bp neighborhood around each candidate peak. Their mean read, averaged over both the replicates and the 201bps,

yields the expected number of reads b per bp[1] (in the neighborhood). Based on a null local Poissonian assumption, as if reads were Poisson distributed with rate b, we associate an upper qth 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 alignment 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 bp. Further, by gathering the positions whose average read equals a certain number μ and computing their CV^2 we obtain the scatter plot of Figure S17B, which appears to closely follow the fitted standard noise model $CV^2 = A/\mu + B$, and which can be expressed as

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

where

$$\sigma^2(\mu) = A\,\mu + B\mu^2 \qquad (2)$$

(As an example, see Figure S17B for the empirical distribution of the reads centered at μ=20 alongside its Poisson and normal fit). Based on this model, the (observed) peak read is randomly drawn from

$$X = \mu + \varepsilon_\mu \qquad (3)$$

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,

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

Since we don't know the value of μ to insert into equation (2), we replace it with either Q_{bio} or the peak read 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 centered at the peak read x, $Q'_{seq}=\{q: \mathrm{Prob}(q < x+\varepsilon)\}$, and require that $Q'_{seq} > Q_{bio}$.

[1] 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 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}.

Normalization of reads from nuclear chromosomes by reads from mitochondria

As the polymerases are individually released during a (small) time interval, we predicted an increase in the number of nascent transcripts with increased run-on times. However, absolute reads are influenced by the sequencing depth, which cannot be easily controlled. These aspects must be taken into account to observe and investigate polymerase pausing with TV-PRO-seq.

We trimmed the aligned reads in SAM files from each replicate of all run-on times to the same total genomic read numbers. As a consequence, given that the total number of labeled nascent RNA increases during the run-on, the number of reads corresponding to peaks that would otherwise stay constant in size (if the experiments where performed at *identical* sequencing reads) decreases. We used the reads from the mitochondrial chromosome as an internal control. In fact, the mitochondrial chromosome is believed to lack the pausing elements typical of metazoans, therefore the average transcription levels at different run-on times can be thought of as being constant to a first approximation [41].

We subset the mitochondrial DNA positions into three groups based on thresholding their reads x: (i) positions such that $x > Q_{seq}$, which we referred to as 'peaks' in the previous section; (ii) positions such that $x < Q_{seq}$, which we label as 'background'; (iii) positions with read counts such that $x < Q_{seq}/2$, which we label as 'background/2'.

For each group, we summed up the total chrM reads at a run-on time and used these numbers to normalize the total reads from nuclear chromosomes corresponding to the same run-on time. We then further normalized the resulting curves to have equal values at the last run-on time (assuming that the plateaus were reached at the 32min run-on) and plotted the normalized reads vs run-on time.

For all three groups, the normalized reads result in saturation curves, in line with our considerations (Figure S1). Furthermore, the steepness of the curves scales with the height of the chosen threshold, confirming that the polymerase is released at a slower rate from peaks compared to background positions (Figure S1).

Model to calculate beta score

In this section, we derive a simple Bayesian model for TV-PRO-seq data and a procedure for their analysis on server CyVerse[42]. We are interested in the stochastic dynamics of biotin-NTP incorporation into a nascent mRNA which can be represented as the following simple reaction:

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

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

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

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 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:

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

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

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

Solving this simple system of ODEs with initial conditions

$$x_i(0) = 0,$$

$$y_i(0) = A_i,$$

yields

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

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

predicting 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 $x_{\mathrm{tot}}(t) = \sum_{i\in S} x_i\,(t) + x_{\mathrm{bck}}(t)$ grows according to

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

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.,

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

It is convenient to study the ratio $\mathrm{x}_i = N_i(t)/N_{\mathrm{tot}}(t)$, where $N_{\mathrm{tot}}(t) = \kappa(t) x_{\mathrm{tot}}(t)$, 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 to the average total-genome reads at the same run-on time t.

We obtain the normalised model

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

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

The quantities $\mathsf{x}_i(t)$, $i \in S$, 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

$$\mathsf{X} = (1 - \mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{t}}) \circ \left[\varrho\,(1 - \mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{t}}) + \varrho_{\mathrm{bck}}^{\mathrm{T}}\,(1 - \mathrm{e}^{-\beta_{\mathrm{bck}}\mathbf{t}})\right]^{\circ -1}, \qquad (4)$$

where $\mathbf{t} = (t_1, t_2, \dots, t_T)$ is the vector of predictor observation run-on times, $\beta = (\beta_1, \beta_2, \dots, \beta_{|S|})$ is the vector of rates, $\varrho = \{\rho_{ij}\}$, $i, j \in S$, and $\varrho_{\mathrm{bck}} = (\rho_{1,\mathrm{bck}}, \rho_{2,\mathrm{bck}}, \cdots \rho_{|S|,\mathrm{bck}})$ 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

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

to approximate the growth of the average of total reads. As in the previous section, the mitochondrial chromosome can be thought of as being constant to $x_{\mathrm{chrM}} = \kappa(t) A_{\mathrm{chrM}}$ to a first approximation. We use them as a reference level. We divide the total reads by the chromosome-M reads, and fit the model

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

where $\rho_{\mathrm{chrM,tot}} = A_{\mathrm{tot}}/A_{\mathrm{chrM}}$, to such data using the random-search algorithm of the nls2 R package [43], which returned a significant fit with estimated parameters reported in the table below, see also Figure S17C.

	Estimate	Std.err.	t value	Pr(>|t|)
$\rho_{\mathrm{chrM,tot}}$	46.621	2.769	16.839	0.000
β_{tot}	0.760	0.176	4.322	0.005

Based on this consideration, our choice is to use the exponential model to approximate the growth of the average total-genome reads $N_{\mathrm{tot}}(t)$, and study

$$\mathsf{x}_i(t) = \frac{1}{\rho_{i,\mathrm{tot}}} \frac{(1 - \mathrm{e}^{-\beta_i t})}{(1 - \mathrm{e}^{-\beta_{\mathrm{tot}} t})}, \qquad (5)$$

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

$$\mathsf{X} = (1 - \mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{t}}) \circ \left[\varrho_{\mathrm{tot}}^{\mathrm{T}}\,(1 - \mathrm{e}^{-\beta_{\mathrm{tot}}\mathbf{t}})\right]^{\circ -1}, \qquad (6)$$

where

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

We then chose the informative prior

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

where $\mathrm{Gamma}(\alpha, \beta)$ 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 rates corresponding to pausing elements can be significantly smaller. Based on such a heuristic consideration we choose the informative priors

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

which have mean and variance equal to 1 and 10, respectively, and place 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.,

$$\mathsf{Y} = \mathsf{X} \cdot \epsilon,$$

where

$$\log\epsilon \sim \mathcal{N}(0, \sigma^2).$$

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

$$\log\mathsf{Y} \sim \mathcal{N}(\log\mathsf{X}, \sigma^2).$$

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

$$\mathrm{CV}^2\mathsf{Y} \approx \mathrm{CV}^2\epsilon,$$

where $\mathrm{CV}^2\mathsf{Y}$ is estimated from aggregated data. As ϵ is lognormal, we have

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

and

$$\sigma^2 \approx \log[\mathrm{CV}^2\mathsf{Y} + 1],$$

which suggests the prior

$$\sigma \sim \mathrm{Gamma}(1.6,0.4).$$

An MCMC sampler to fit the model was implemented using the PyMC3 Library for Bayesian Statistical Modeling and Probabilistic Machine Learning [44]. PyMC3 relies on the **Theano** framework [45], which allows fast evaluation of matrix expressions, such as those in equations (4) and (6), 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 ~ 170000 peaks. To ease the computational burden, we divide the peak list into chunks of ~ 3000 randomly chosen peaks. 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 0.1 min^{-1} obtained from the half-life measured in Jonkers, Kwak, and Lis [14]. Aggregating the individual-chunk estimates using the laws of total mean and variance yields:

$$\beta_{\mathrm{tot}} = 0.147 \pm 0.007\ \mathrm{min}^{-1}.$$

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

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

which results in a wider range of inferred β_i, whilst maintaining the same rank order.

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 field from selected table [46]. The 5' and 3' ends of all exons from the mRNA list was transformed into another table with the custom script `Unique_annotation_maker.pl`. Column 1 to 3 was the chromosome, position and strand of annotation site; column 4 was the gene name; if the column 5 'type' equal to start it means it is the 5' end of exon, otherwise it is 3' end; the column 6 'number_min' and 7 'number_max' is the min and max number of exon in different variant of same gene, TES have been marked as -1; The column 8 'hit' showed how many variant of transcript have this splicing site; and column 9 'variant' reference to the number of variant the gene have.

For rRNA and tRNA genes, two tables were downloaded from RNAcentral (*https://rnacentral.org/*), the RNA gene classification information was in rfam_annotations.tsv and the genomic locations of these genes was in Homo_sapiens.GRCh38.bed, respectively. We used a custom script `rFAM_annotation_merger.pl` to merge these two tables for the further analysis [47].

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 ±4500bp 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 TSS were used to generate the density plot using the ggplot2 package [48] for *R* (Figure 1B). We also extracted peaks within 2500bp of unique TSSs of mRNA genes, and plotted with ggplot2 the distribution of peaks corresponding to the top and bottom 10% β values, respectively (Figure 2C).

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 [49]; we used the 'Potential Pol3 targets' table and converted it to human genome assembly GRCh38 with the UCSC liftOver tool [46]. 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. These were displayed as violin plots with inserted boxplots using the ggplot2 package for *R* (Figure 2A).

Trp Treatment data analysis

Triptolide (Trp) treatment inhibits transcription initiation [14]. Trp treatment will thus perturb the dynamic balance of polymerase occupancy; the percentage of reads from peaks corresponding to short pausing times around TSS will reduce after Trp treatment, as the polymerase will vacate these early while Trp prevents the influx of new polymerases from the TSS. In contrast, the longer pausing region will remain occupied longer and will thus receive an increased percentage of reads. We ranked the peaks around TSS by the ratio of reads in peaks in the

untreated sample and the same peaks in the Trp treated sample. Peaks whose reads decreased extremely after Trp treatment were considered likely to pause for a short time and vice versa for peaks with increased reads; we thus selected the peaks corresponding to the top and bottom 10% quantiles of these ratios and displayed their positional distributions (Figure 2D).

Meta gene analysis about pausing peaks

6562 genes which have unique TSSs and TESs and are longer than 3000bp were used for meta gene 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: 1000bp region upstream of TSS
TSS related region: 1000bp region downstream of TSS
region before TES: 500bp region upstream of TES
pA related region: 4500bp 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 1000bp or last 500bp 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 by the ggplot2 *R* package with parameter span=0.1 (Figure 2B). We also separately plotted the smoothed conditional mean plot for promoter and TSS related region only (Figure S5). Peaks around TSS and TES of tRNA genes were plotted in the same way (Figure 3A, Figure S8).

Gene expression noise estimation and selection

Gene expression noise is estimated from single-cell sequencing data as [28]:

$$\eta = CV^2 - 1/\mu,$$

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 [30] 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. 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 S9).

We generated the smoothed conditional mean plot of the 'high noise' and 'low noise' genes with the same strategy as for the meta gene analysis (Figure 3C) and plotted histograms to show the absolute frequencies of peaks from 'high noise' and 'low noise' genes (Figure S10).

Density plots (Figure 3D) and split violin plots (Figure S11) were generated with ggplot2 as before.

Pausing analysis of rDNA repeats

As rDNA loci are highly repeated in the genome, we used a special strategy for the analysis of pausing in rDNA. We built a Hisat2 index by combining the masked hg38 genome from UCSC and a standard reference rDNA sequence [38]. Thus reads corresponding to rDNA will align to the reference rDNA only. We then extracted these reads for peak calling and beta calculation. The betas of these peaks were used to plot pausing times for peaks in different regions of rDNA (Figure 3B). No peaks were annotated to the 5.8S rRNA region is because it was not masked in the hg38 genome from UCSC.

Motif analysis

The ±50bp surrounding sequence around each peak was extracted with the custom script `Peak_seq_getter.pl`, saved into a fasta file, and subjected to *de novo* motif detection. In addition, the regions from -550 to -450 and +450 to +550 at each peak were extracted to serve as control sequences. Motif detection was done with the program `findMotifs.pl` of the Homer software suite with default options and by using the control sequences as background [50], which resulted in a number of position probability matrices (PPM) for enriched motifs, which we term the PPM_e's. For each PPM_e, we used the `homer2 find` to obtain the distances between all motif occurrences and peaks in the input sequence set. We used parameter -strand to ensure strand-specific motif detection.

For each distance distribution resulting from a PPM_e, we compared the most frequently occurring distance d_1, to the second most frequently occurring distance d_2; we ranked PPM_es by the relative standard error ρ in estimating the proportion $\hat{p} = n_1/(n_1 + n_2)$, based on the heuristic assumption that n_1 is binomially distributed, where n_1 and n_2 are the numbers of occurrences of d_1 and d_2, respectively,

$$\rho = \frac{1}{\hat{p}}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n_1+n_2}}.$$

After ranking by ρ, the top 6 motifs were taken for further analysis. We considered these motifs to have a unique, precise pausing site at single base resolution. We then extracted the PPM for the motifs appearing at d_1 and termed this second PPM the precision PPM, PPM_p. We generated sequence logos for PPM_e and PPM_p with the ggseqlogo R package.

We then plotted pausing times of peaks at the precise pausing sites and considered these to be related to the motifs. Peaks at distances between 20bp to 40bp with regards to the precise pausing sites were used as controls of the surrounding neighbourhood, since different genomics regions have different overall pausing characteristics/times. Box plots were used to show the pausing time distributions between motif related peaks and these adjacent controls. A Mann–Whitney U test was used to test for significant differences (Figure 4A). We repeated this comparison for all peaks to test if the motif peaks' pausing times deviated from the genome-wide average.

The top motif output by Homer, 'Accurate pausing motif1', which corresponds to the sequence ACAGTCCT, was taken for further analysis. We identified 'variant motifs' from the consensus by changing individual positions of Accurate pausing motif1 and then determined their occurrences as described above (Figure 4B, Figure S13).

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. H3K4me1, H3K4me2, H3K4me3 and H3K27ac data were obtained from Gene Expression Omnibus, GSE101646 [51], and H3K9me3, H3K36me3 and DNase-seq data 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` [52], then aligned to hg38 under --no-spliced-alignment condition by Hisat2 [38]. The SAM files were converted to BAM files, then to BED files using Samtools [39] and Bedtools [40], 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` [40]. The BEDGRAPH files were annotated to TSS or pausing peaks with the custom script `Liner_bedgraph.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 +/-1000bp 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. Finally, we averaged the coverages of each class, respectively, and displayed the results using ggplot2 in *R* (Figure 4D, Figure S15).

Calculation of PI, histone modification and chromatin accessibility for mNET-seq data

HEK293 mNET-seq data was downloaded from Gene Expression Omnibus, GSE61332 [22]. We used the UCSC liftOver tool to convert the BEDGRAPH file to hg38 [46]. We then defined target genes for further analysis by selecting genes longer than 3000bp, with unique TSSs and TESs. Peak selection for the mNET-seq data followed the same strategy as for TV-PRO-seq; the peaks were annotated to target genes with the script `PI_annotater.pl`. We defined the genic regions from TSS +500bp to TES as gene body (GB) [21], and calculated a pausing index (PI) for each peak position by dividing reads in peaks by the average reads in GB of the same gene. We considered either peaks along the whole gene or peaks within TSS +500bp only. We implemented this by processing the UCSC mRNA gene annotation as above with the script `PI_reference_maker.pl`. We then used the script `PI_counter.pl` to count the GB reads of target genes.

The peak selection output file was processed with the script `Liner_bedgraph.pl` to extract histone modification states within +/-1000bp of peaks in the same way as for TV-PRO-seq; we removed the top 5% peaks with 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, or peaks within the TSS to +500 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 S16).

Reference

37. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17** (2011).
38. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).
39. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
40. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
41. Barshad, G., Marom, S., Cohen, T. & Mishmar, D. Mitochondrial DNA Transcription and Its Regulation: An Evolutionary Perspective. *Trends Genet* **34**, 682-692 (2018).
42. Polanski, K. et al. Bringing numerous methods for expression and promoter analysis to a public cloud computing service. *Bioinformatics* **34**, 884-886 (2018).
43. Grothendieck, G. Non-linear regression with brute force. *R package version 0.2* (2013).
44. Salvatier, J., Wiecki, T.V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2** (2016).
45. Al-Rfou R, A.G., Almahairi A, et al. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint* (2016).
46. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. The human genome browser at UCSC. Genome research. *Genome Res* **12**, 996-1006 (2002).
47. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**, D335-D342 (2018).
48. Wickham, H. ggplot2: elegant graphics for data analysis. (Springer, 2016).
49. Oler, A.J. et al. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17**, 620-628 (2010).
50. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
51. Morgan, M.A.J. et al. A cryptic Tudor domain links BRWD2/PHIP to COMPASS-mediated histone H3K4 methylation. *Genes Dev* **31**, 2003-2014 (2017).
52. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

Supplementary Text

Supplementary discussion

Deriving β from above process will provide a better estimate of pausing times at single base resolution than using (m)NET-seq; estimation of polymerase pausing using (m)NET-seq data is usually based on calculation of the 'pausing index' (PI) [1, 21]. The PI is based on the assumption that elevated read densities at certain positions in a gene body must reflect longer polymerase occupancy and thus pausing at these positions. At the same time, the overall read density along the gene will depend on its expression level, which needs to be taken account of. The PI is therefore calculated as the ratio of read counts at peaks (or suspected pausing positions) over the average reads from 'background' region that is expected to reflect normal elongation in the same gene. As a consequence, the PI needs a long background region in the gene of interest, while TV-PRO-seq can be used to estimate the pausing times of peaks in genes of arbitrary length (thus permitting, e.g., the study of PolIII transcription, genes transcribed by PolII are typically short), even if they don't display a long enough background region.

Further, (m)NET-seq/PI approach is not well suited to investigate pausing times for the following reasons:

(i) a pausing site is not expected to be utilized every single time a polymerase moves into its position. Instead, the polymerase will pause at an average fraction of transcription events and pass unimpeded at other times. Since the normal elongation background will contain reads from both cases, calculating the PI will underestimate the pausing time (Figure S3A). TV-PRO-seq does not have this problem, since the 'pass' fraction will contribute very little signal, if any, due to its short residence time at the position (saturation before the first TV-PRO-seq timepoint). It thus estimates the actual pausing time of the paused fraction (Figure S3A).

(ii) if significant fractions of transcripts of a gene terminate early, the flux of polymerase will not be constant throughout the gene. If the PI calculation uses a background region at the end of the gene to gauge promoter proximal pausing, pausing times will be overestimated due to lower read densities at the 3' end. TV-PRO-seq again is not affected by this; early termination will reduce the height of the plateau of the saturation curve, but will not affect the slope, which is used to estimate the pausing time (Figure S3B).

(iii) Different genes might have systematic differences in their pausing times, therefore using the reads in the gene body to calculate the PI can lead to systematic bias. This is of interest, for instance, for genes transcribed by different types of polymerases (Figure 2A).

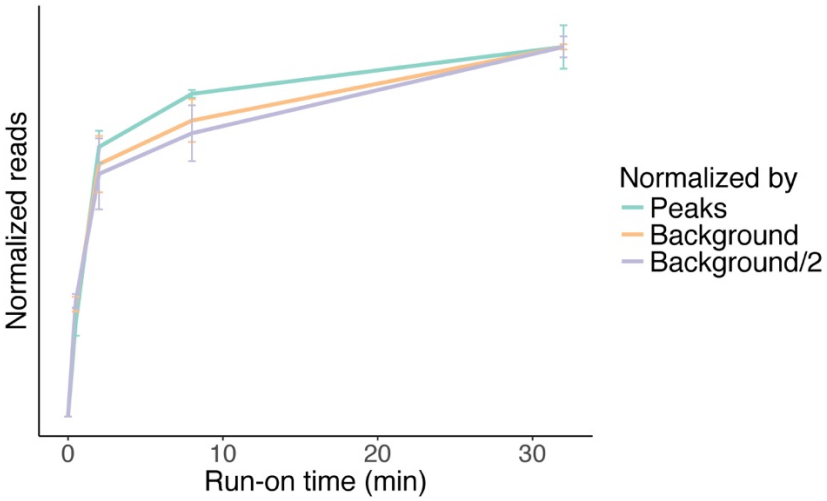

Fig. S1. *Total reads of TV-PRO-seq samples corresponding to different run-on times form saturation curves upon normalization by mtDNA reads selected at different heuristic thresholds (Background/2, Background, Peaks).*

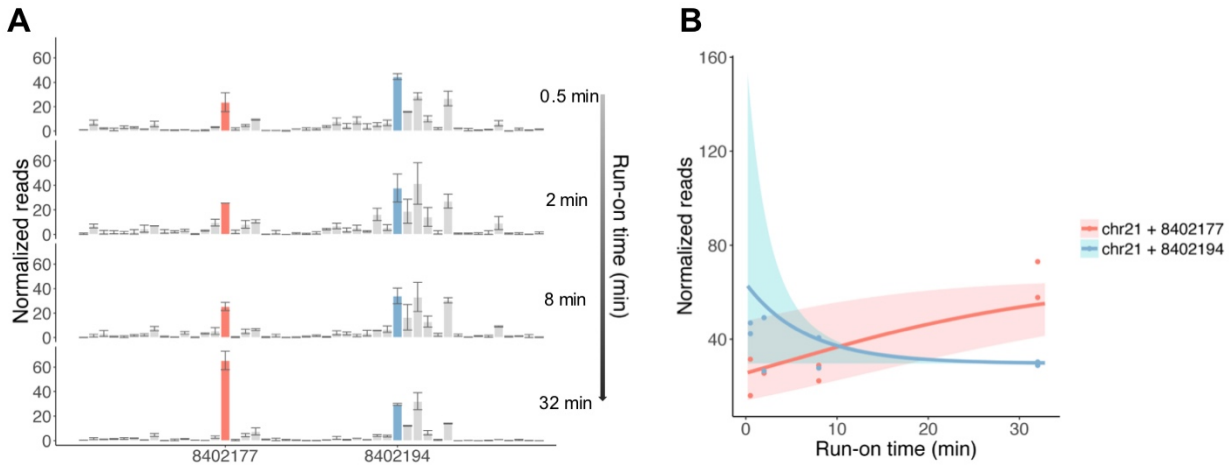

Fig. S2. *Read numbers normalised by total-genome reads, rescaled (by 10^7), and averaged over the biological replicates, as used for model fitting.*

*(**A**) Normalised reads of the example peaks of Figure 1D. Error bars correspond to the data range from two biological replicates.*

*(**B**) Normalised reads (points) and mean prediction curves (5) (solid lines) as functions of the run-on time for the selected peaks of (A). The shaded region boundaries correspond to curves obtained from the lower and upper quartiles of the posteriors.*

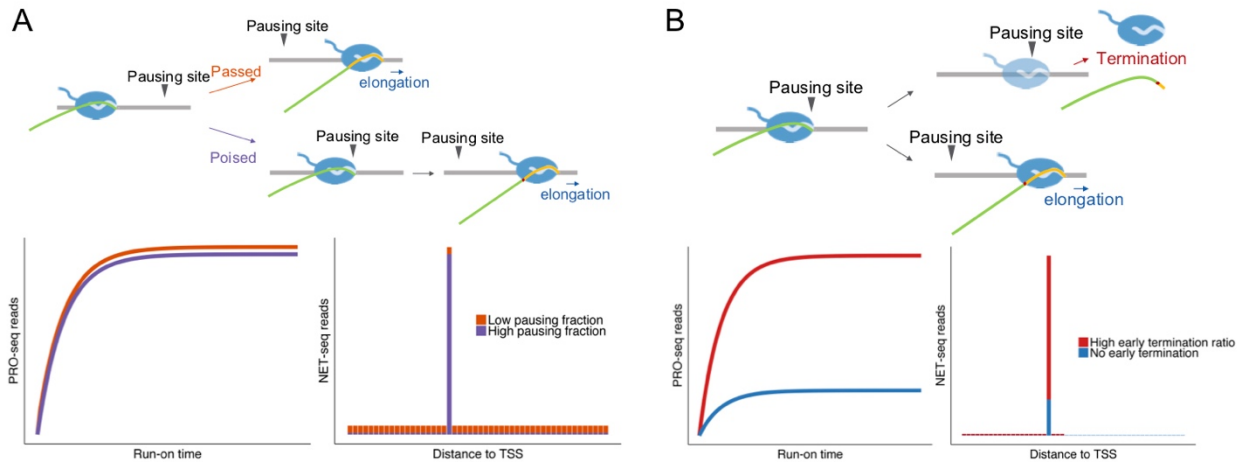

Fig. S3. Comparison of β and PI to gauge polymerase pausing.

*(**A**) Influence of infrequently utilized pausing positions on assay data. The purple line represents the signal resulting from a pausing site at which all polymerase molecules need to pause before they can pass ('high pausing fraction'). The orange line reflects a position which has the same pausing time as the purple example for polymerases that do pause, but only 20% of polymerase molecules on average do actually pause at this site ('low pausing fraction').*

*(**B**) Influence of early transcriptional termination on assay data. The blue line represents the signal resulting from a pausing site located in a gene which does not have early termination. The red line corresponds to a site that has the same pausing time and production level of mature RNA as the blue example, but 80% of nascent RNAs will terminate after released from pausing site.*

*While TV-PRO-seq will allow robust pausing estimates in both cases, NET-seq will under- and overestimate pausing in cases orange and red, respectively, when the PI is used.

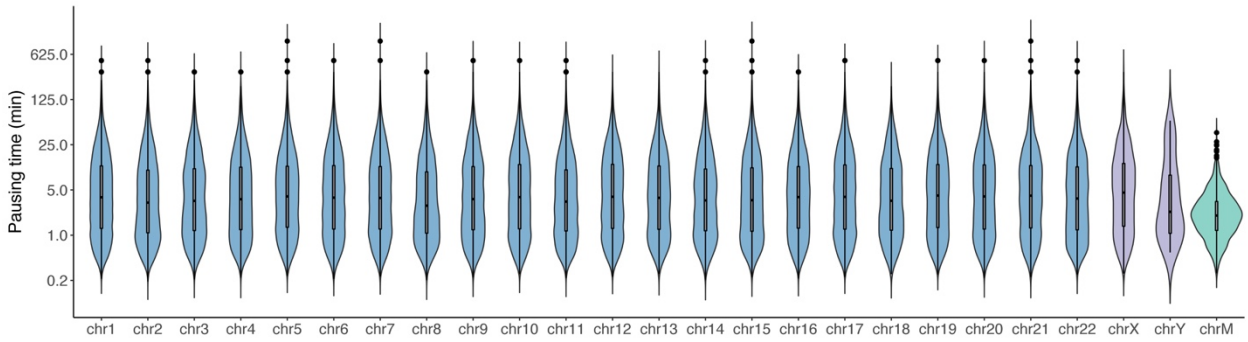

Fig. S4. Distributions of estimated pausing times for peaks on different chromosomes

Blue indicate autosomes, purple indicates *allosomes, green indicates the mitochondrial chromosome.*

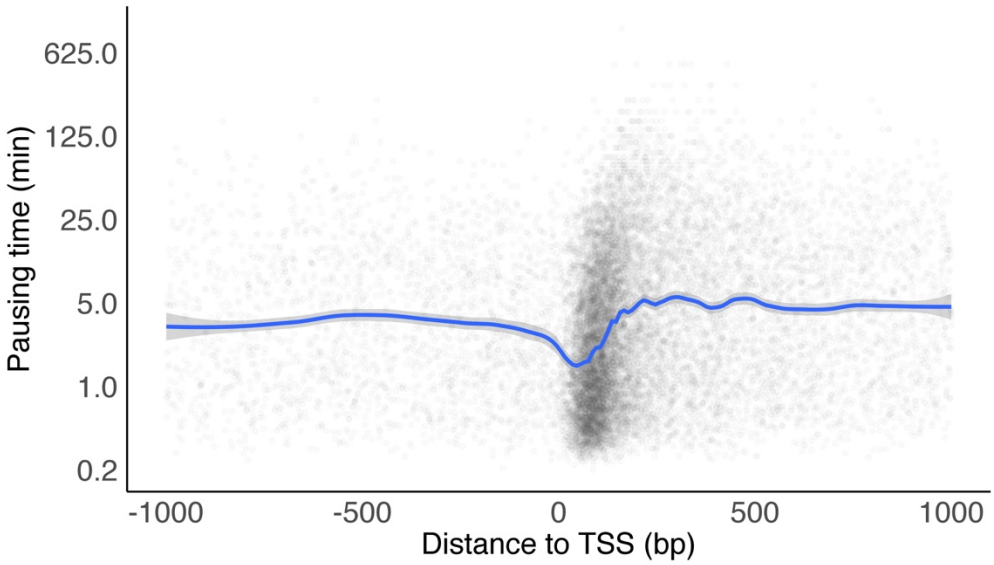

Fig. S5. Pausing times of peaks within 1000bp towards TSS of mRNA-transcribing metagene

Each grey dot represents a pausing peak. The blue line corresponds to the moving average (LOESS fit). The grey shading indicates the confidence interval and is negligible on this scale, hence invisible over most of the graph.

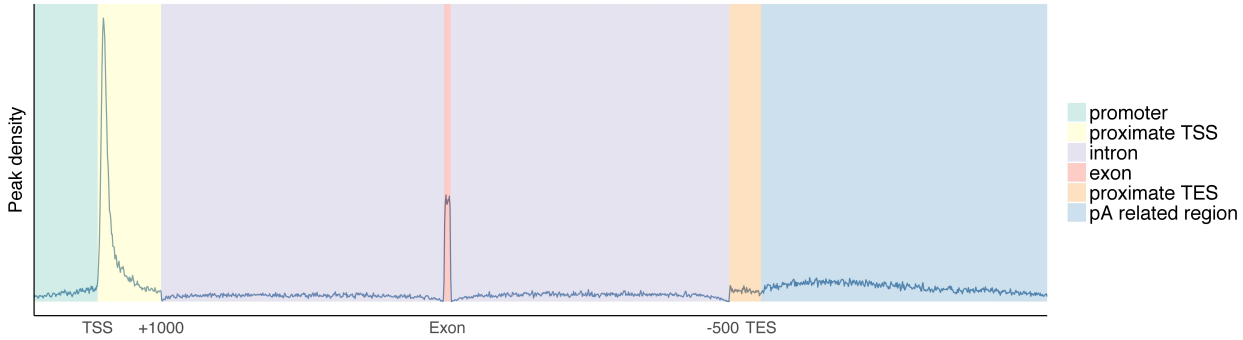

Fig. S6. Peak density of mRNA-transcribing metagene

Definition of region is similar as in Figure 2B, the blue line represents the density of peaks.

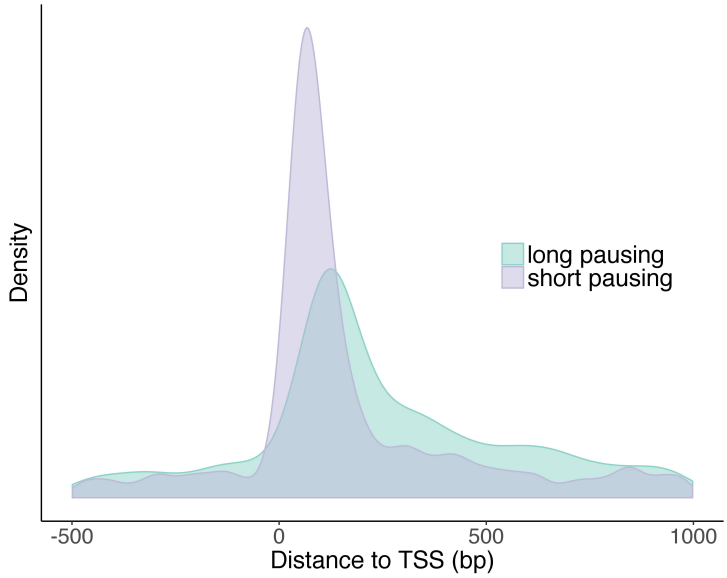

Fig. S7. Peak density of long and short pausing around TSS in KBM7 cells

Peaks within -/+ 1kb of TSS were classified into 'long' and 'short' according to their pausing times and were displayed as distributions regarding their distances to TSS.

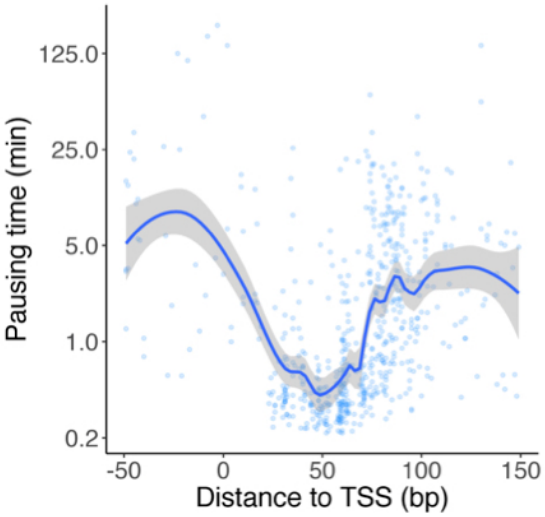

Fig. S8. Pausing times and positions around TSS of tRNA genes

Each dot corresponds to a pausing peak. The blue line corresponds to the moving average with the grey shading indicating the confidence interval (LOESS fit).

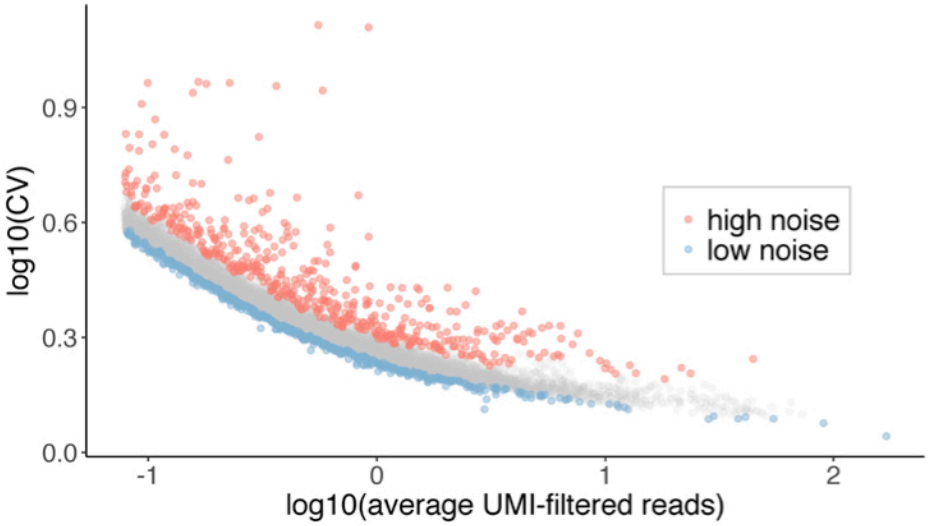

Fig. S9. Selection of high/low-noise genes.

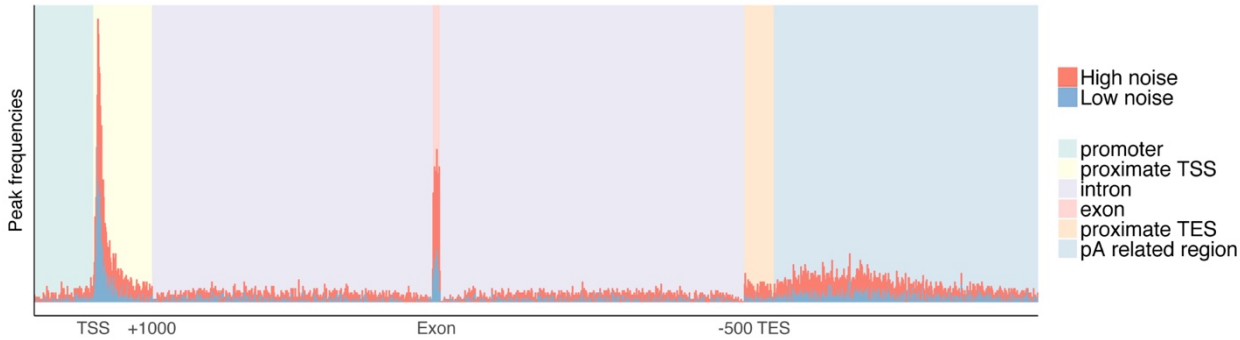

Fig. S10. Histogram of peak frequencies of high/low-noise genes

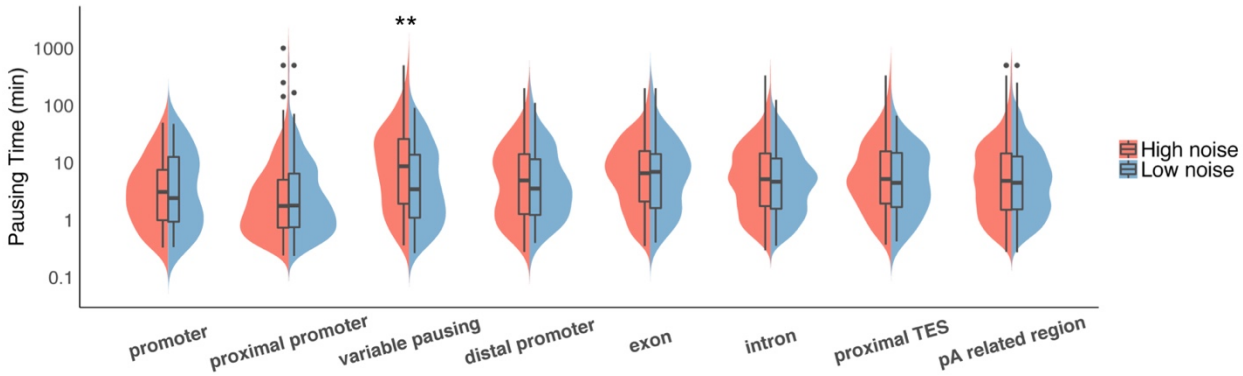

Fig. S11. Comparison of pausing times of high/low noise genes in different regions

Definitions of regions and p-values of Mann–Whitney U tests for significantly different pausing times between high noise and low noise genes in the regions are:
Promoter: -1000 to TSS, p-value=0.80;
Proximal promoter: TSS to +200, p-value=0.50;
Variable pausing: +200 to +500, p-value=0.00076;
Distal promoter: +500 to +1000, p-value=0.78;
Exon: p-value=0.77;
Intron: p-value=0.17;
Proximate TES: -500 to TES, p-value=0.32;
pA related region: TES to +4500, p-value=0.51;
*The significance thresholds were Bonferroni corrected to take account of multiple testing (** P<0.00125).*

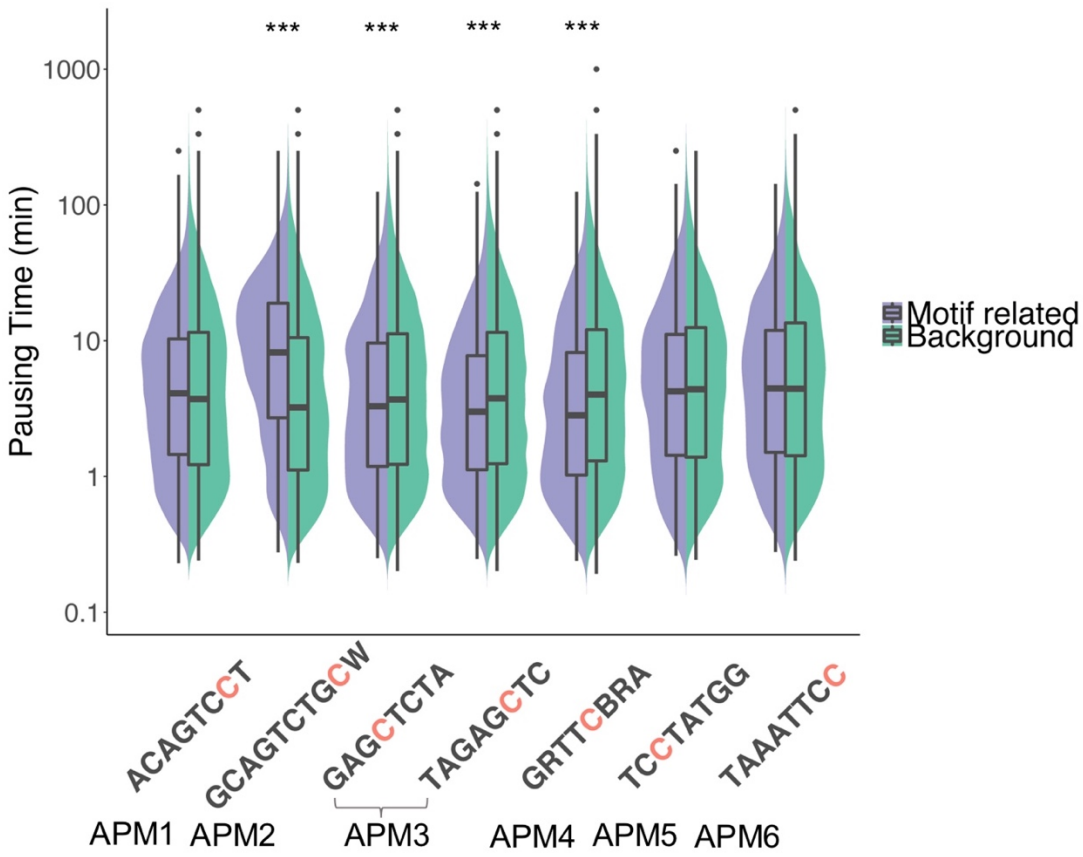

Fig. S12. Pausing time comparison for enriched motifs at peaks and nearby background sequences.

**P<0.05, ***P<0.001, Mann-Whitney U test.*

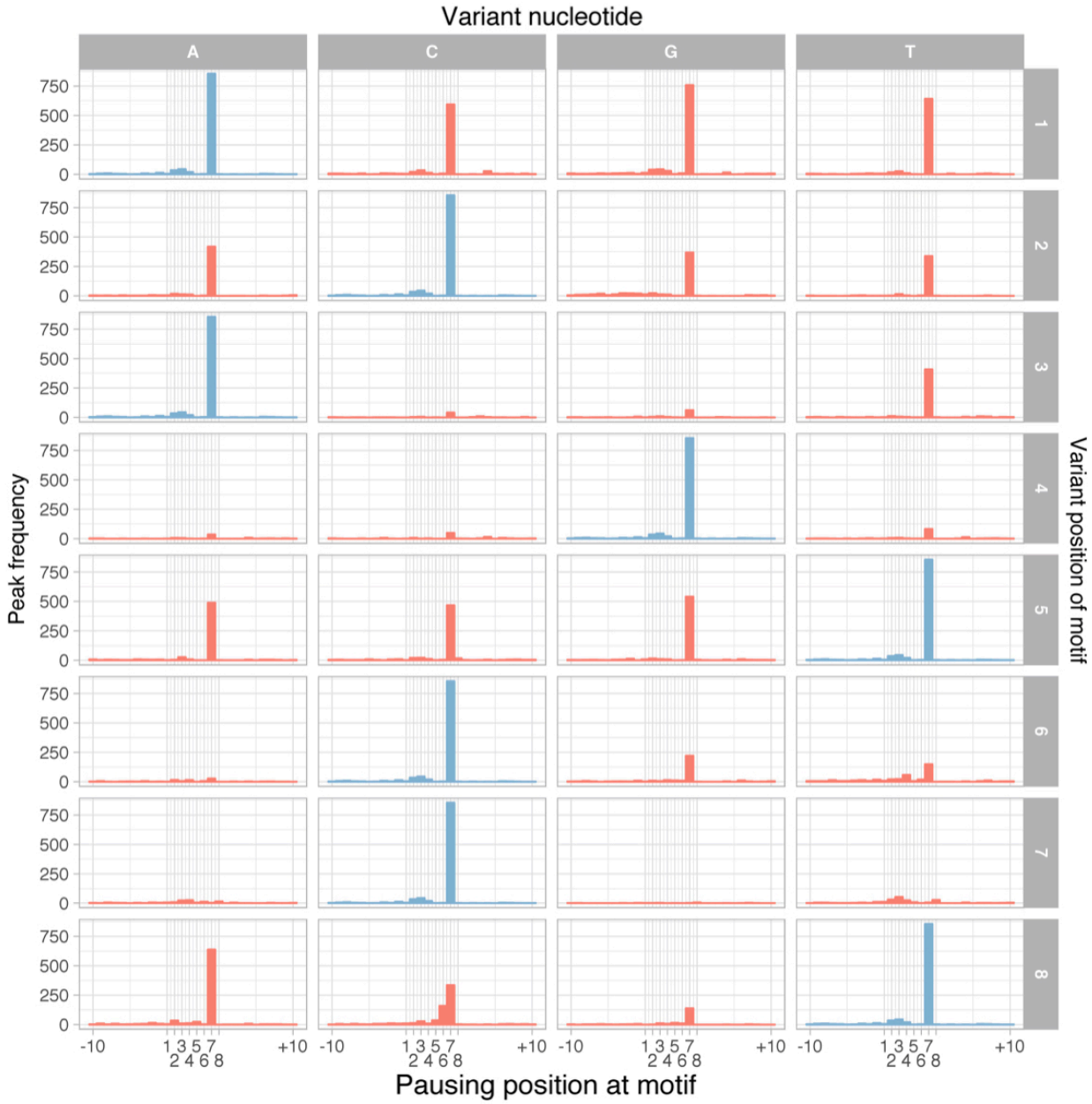

Fig. S13. Histograms of peak frequencies at positions relative to the motif ACAGTCC and single base variants (position variants are shown in red, whereas consensus positions are shown in blue.)

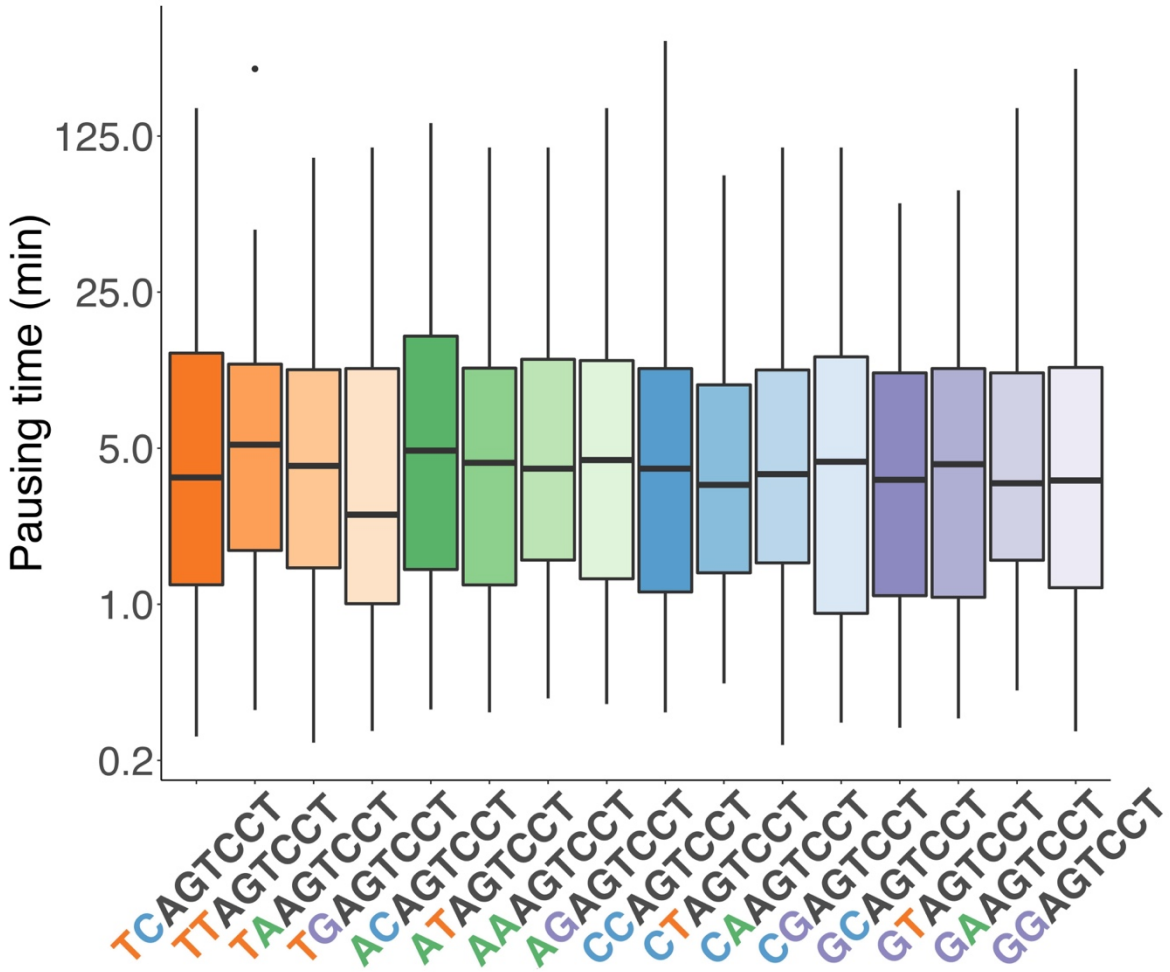

Fig. S14. Dinucleotide variants of APM1 (Fig. 4C) do not show systematic effects on background peak pausing times

'Background peak' refers to the pausing peaks within a distance of 20 to 40bp of the accurate pausing site. Trends among all groups of four were assessed with Kendall's tau test and were found to be not significant (H_1: tau ≠ 0).

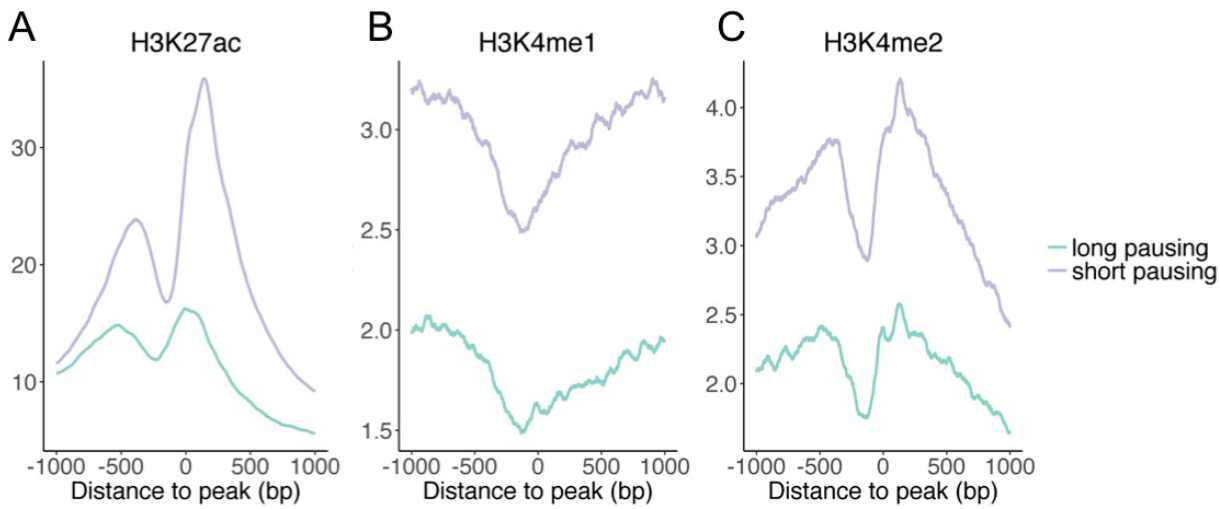

Fig. S15. The average signal of H3K27ac, H3K4me1, and H3K4me2 ChIP-seq data around long/short pausing sites as determined by TV-PRO-seq.

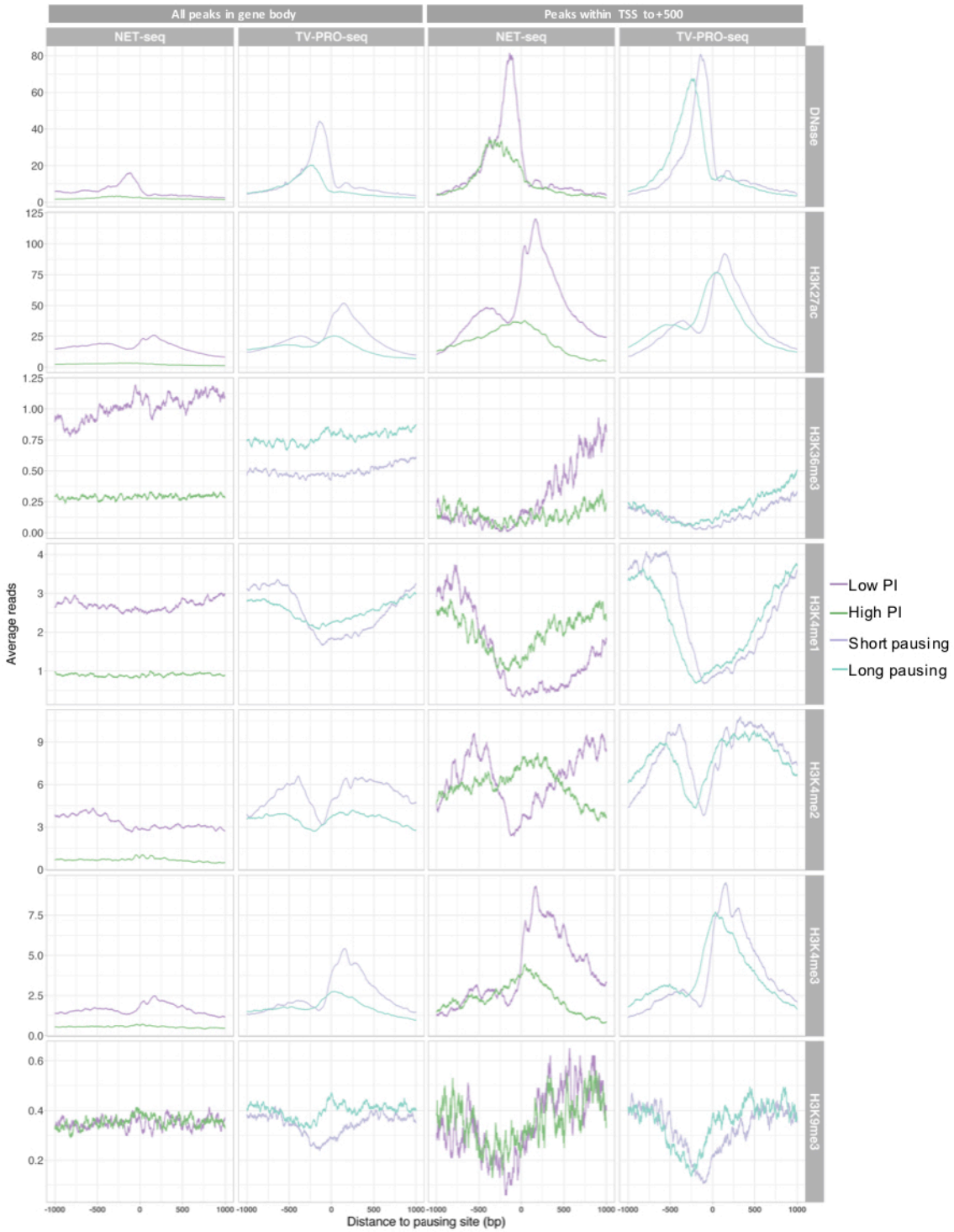

Fig. S16. Comparison of chromatin state profiles for TV-PRO-seq and NET-seq/PI.

Dark purple and dark green lines represent the low PI and high PI pausing positions from NET-seq data, respectively. Light purple and light green represent the short pausing and long pausing positions from TV-PRO-seq data. The type of chromatin feature (as determined by DNase-seq or ChIP-seq) is shown on the right hand side; all peaks in the gene body or peaks in the region from TSS to +500bp are shown in the first two and last two columns, respectively. The profiles are clearer for TV-PRO-seq in many cases, and often deviate from the NET-seq profiles, suggesting that TV-PRO-seq often produces better and sometimes different information

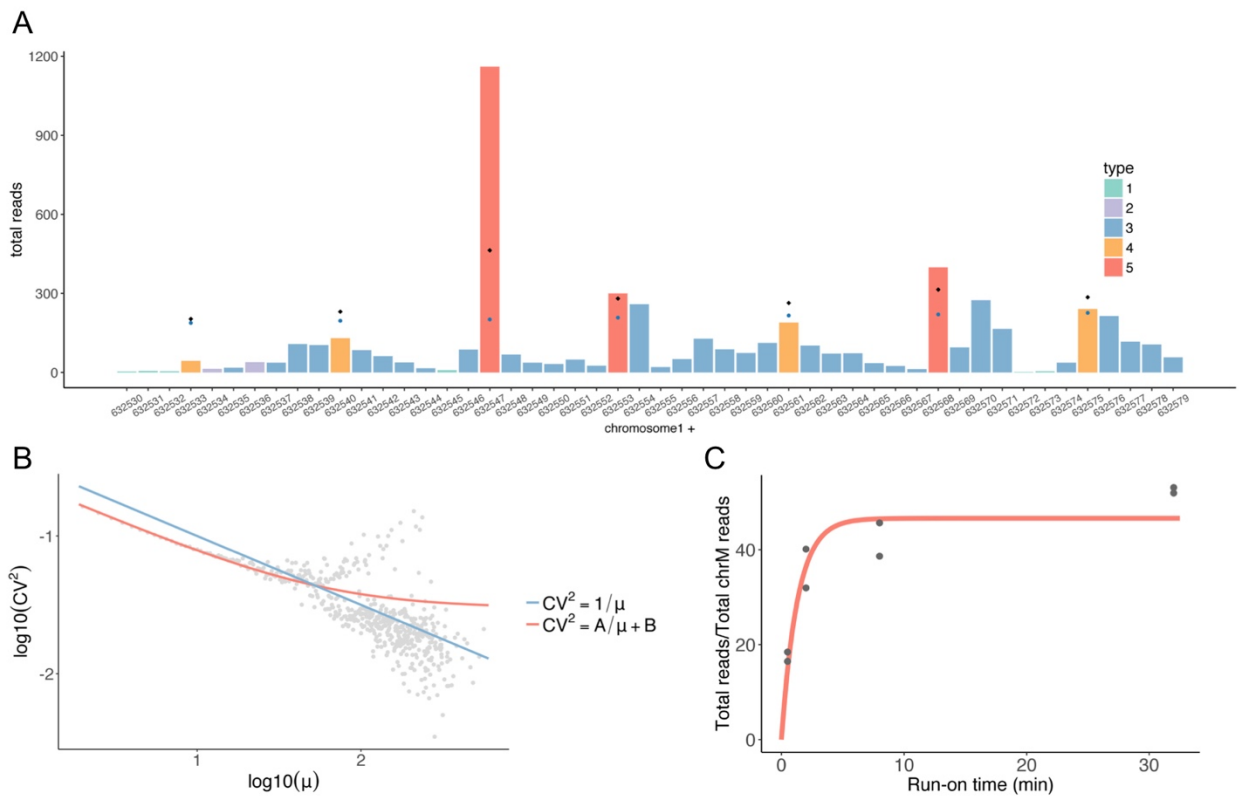

Fig. S17. Peak selection and β calculation

*(**A**) The total reads from a sample 50-bp fragment of the + strand of chromosome 1 are shown as an example for the peak selection procedure. Positions of color 1 are discarded at the first step as they have less than 13 total reads, and positions of color 2 are rejected for having zero reads at at least a single run-on time sample. Positions 632553-632554 correspond to a 2-bp dispersed peak, where only the first base is selected for further analysis while position 632547 corresponds to a peak concentrated on a single base. Peaks of color 3 are discarded at step 2 in favor of the highest peak in the 3-bp neighborhood. Peaks of color 4 are discarded at step 3. Red peaks are selected/called for further analysis. Blue points correspond to the quantile threshold Q_{bio}. Black points correspond to the threshold Q_{seq}.*

*(**B**) Sequencing noise scatterplot. Line is a weighted nonlinear least-square fit* $CV^2=A/\mu+B$*, with parameters (A, B) = (0.53, 0.009), estimated by means of the random-search algorithm of the* `nls2` *R package* [43].

*(**C**) Saturation plot of the total-genomic reads normalized to the total-chrM reads.*

Table S1. Accurate pausing motifs.

Motif name	Consensus sequence	Logo for PPM_e	Logo for PPM_p
Accurate Pausing Motif 1	ACAGTCCT		
Accurate Pausing Motif 2	GCAGTCTGCW		
Accurate Pausing Motif 3	GAGCTCTA		
	TAGAGCTC		
Accurate Pausing Motif 4	GRTTCBRA		
Accurate Pausing Motif 5	TCCTATGG		
Accurate Pausing Motif 6	TAAATTCC		

** Reverse complemented Accurate pausing motif3 has a different precision pausing site, i.e. the peak at a different position compared to the forward one.*