www.sciencemag.org/cgi/content/full/339/6122/950/DC1

Supplementary Materials for

Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing

Hojoong Kwak, Nicholas J. Fuda, Leighton J. Core,* John T. Lis*

*To whom correspondence should be addressed. E-mail: ljc37@cornell.edu (L.J.C.); jtl10@cornell.edu (J.T.L.)

Published 22 February 2013, *Science* **339**, 950 (2013)
DOI: 10.1126/science.1229386

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Source codes for analysis scripts

Supporting Online Materials

Table of Contents: **Pages**

Materials and Methods

Cell culture and nuclei isolation

Drosophila S2 cells were maintained in Shields and Sang M3 insect medium supplemented with 10% Fetal Bovine Serum, Bacto-Peptone and Yeast Extract at 30℃. At 16~20 passages, nuclei were isolated as described previously with modifications*(1-3)*. All temperatures were at 4℃ or ice cold unless otherwise specified. Briefly, cells were washed in PBS and resuspended in Buffer S (10 mM Tris-HCl pH 7.5, 10% glycerol, 3 mM $CaCl_2$, 2 mM $MgCl_2$, 0.5 mM DTT, protease inhibitors cocktail (Roche), 4 u/ml RNase inhibitor (SUPERaseIN, Ambion) at the cell density of 2×10^7 cells/ml. After 5 min of incubation, 9× volume of Buffer L was added and immediately homogenized using a tight fitting pestle until over 90% of the nuclei were released. Nuclei were fractionated by centrifugation at 1000 g for 4 min and recovered from the pellet fraction. Recovered nuclei were washed twice in Buffer L and once in Buffer D (50 mM Tris-HCl pH 8.0, 25% glycerol, 5 mM $MgAcetate_2$, 0.1 mM EDTA, 5 mM DTT). Washed nuclei were finally resuspended in Buffer D at a density of 2×10^7 nuclei/100 µl) and immediately frozen in liquid nitrogen. Nuclei were stored in -80˚C until usage.

PRO-seq and PRO-cap library preparations

Four parallel run-on reactions of PRO-seq_{ATP}, PRO-seq_{CTP}, PRO-seq_{GTP} and PRO-seq_{UTP} were carried out as follows. 2×10^7 nuclei were added to the same volume of 2× Nuclear Run-On (NRO) reaction mixture (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% Sarkosyl, 5 mM $MgCl_2$, 1 mM DTT, 500 µM biotin-11-A/C/G/UTP (Perkin-Elmer), 0.8 u/µl RNase inhibitor) and incubated for 3 min at 30˚C. Alternatively, 375 µM of each of all 4 biotin-11-NTPs were supplemented in the reaction for an abbreviated protocol (PRO-seq_{4NTP}) or PRO-cap. Nascent RNA was extracted using Trizol and precipitated in 75% ethanol. Extracted nascent RNA was fragmented by base hydrolysis in 0.2 N NaOH on ice for 10~12 min, and neutralized by adding 1× volume of 1 M Tris-HCl pH 6.8. For PRO-cap, the fragmentation step was omitted. Excessive salt and residual NTPs were removed by using P-30 column (Bio-rad). Fragmented nascent RNA was bound to 30 µl of Streptavidin M-280 magnetic beads (Invitrogen) following the manufacturer's instructions. The beads were washed once in high salt (2 M NaCl, 50 mM Tris-HCl pH 7.4, 0.5% Triton X-100), once in medium salt (300 mM NaCl, 10 mM Tris-HCl pH 7.4, 0.1% Triton X-100), and once in low salt (5 mM Tris-HCl pH 7.4, 0.1% Triton X-100). Bound RNA was extracted from the bead using Trizol (Invitrogen) in two consecutive extractions, and the RNA fractions were pooled, followed by ethanol precipitation.

For the first ligation reaction, fragmented nascent RNA was redissolved in H_2O and incubated with 10 pmol of reverse 3' RNA adaptor (5'p-rGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArC rUrCrUrGrArArC-/3'InvdT/) and T4 RNA ligase I (NEB) under manufacturer's condition for 6 hr at 20˚C. For PRO-cap, the standard 3' RNA adaptor (Illumina) was used. Ligated RNA was

enriched with biotin-labeled products by another round of Streptavidin bead binding and extraction. To repair 5’ ends, the RNA products were treated with Tobacco Acid Pyrophosphatase (TAP, Epicentre) and Polynucleotide Kinase (PNK, NEB). Each reaction was followed by an ethanol precipitation step. For PRO-cap, PNK treatment step was omitted, and Antarctic phosphatase (AP, NEB) was used to treat the RNA preparation prior to TAP treatment to enrich for 5’ capped RNA. Since these procedures repair 5’ ends after the 3’ ligation, self-circularized products were not expected during the first ligation step.

5’ repaired RNA was ligated to reverse 5’ RNA adaptor (5'-rCrUrGrArArCrArArGrCrArGrArA rGrArCrGrGrCrArUrArCrGrA-3' or 5'-rCrCrUrUrGrGrCrArCrCrCrGrArGrArArUrUrCrCrA-3’ for using TruSeq barcodes (Illumina)). Standard 5’ RNA adaptors were used for PRO-cap. Ligated RNA products were further enriched for biotin-labels by the third round of streptavidin bead binding and extraction. Adaptor ligated nascent RNA was reverse transcribed using 25 pmol RT primer (5’-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3’(GX2 primer, Illumina) or 5’- AATGATACGGCGACCACCGAGATCTACACGTTCAGAG TTCTACAGTCCGA-3’ for TRU-seq barcodes (RP1 primer, Illumina). Standard Illumina RT primers were used for PRO-cap.

A portion of the RT product was removed and used for trial amplifications to determine the optimal number of PCR cycles. For the final amplification, 12.5 pmol of GX1 primer (Illumina) or RPI-index primers (for TRU-seq barcodes, Illumina) was added to the RT product with Phusion polymerase (NEB) under standard PCR condition. Excess RT primer served as one primer of the pair used for the PCR. The product was amplified 12~18 cycles and PAGE purified before being analyzed by Illumina’s GenomeAnalyzer 2 or HiSeq 2000 machines.

Hsp70 specific Run-On RNA length measurement

Lengths of biotin-NRO RNA originating from *Hsp70* TSS were analyzed using a modified ligation-mediated PCR described in a previous study*(4)*. Briefly, 5’-PRO-seq libraries were made as described above with all 4 biotin-NTPs separately and together, except that a short GX2 primer (GX2short; 5’-CAGAGTTCTACAGTCCGA-3’) was used for the amplification. To isolate *Hsp70* specific fragments, a 5’-biotin labeled *Hsp70* specific primer starting at the TSS (5’-bio-ATTCTATTCAAACAAGCAAAGT-3’) was used for an initial primer extension cycle followed by Streptavidin bead binding and extraction. Extracted biotin-labeled *Hsp70* specific primer contained 3’ extended fraction of the library that served as a gene specific template for PCR amplification with GX2short primer. This enrichment step prevents nonspecific priming and significantly reduces background. The PCR products were analyzed in sequencing gels. The resulting amplicon size would be 18 bp greater than the nascent RNA size. ImageJ*(5)* was used to analyze the intensities of the lanes and generate a composite profile.

Processing raw sequence data for polymerase active site mapping

Raw sequences were preprocessed using FASTX-Toolkit*(6)*. Adaptor sequences were removed from the raw sequences using 'fastx_clipper', and the first 26 bases were trimmed with 'fastx_trimmer'. Sequence reads shorter than 16 bases were removed. The first bases, which were the reverse complements of 3' end bases, were counted for each library to verify that the 3' ends represent the polymerase active sites (table S1). Indeed, the majority of the sequences had the same 3' end base as the biotin-NTP that was added in the run-on reaction, indicating that the identified sequences define the 3' ends that are exactly at or near the Pol II active sites.

Reverse complements of the sequence reads, which were the sense sequences of nascent RNA, were generated using 'fastx_reverse_complement'. Each of the 4 biotin-NTP libraries was aligned to the *Drosophila melanogaster* (Dm3) reference genome using Bowtie*(7)* allowing 2 mismatches and excluding any non-uniquely aligned reads. The histograms of the 3' end positions in base pair resolution were generated in the 'bedgraph' format.

For the normalization of the 4 biotin-NTP libraries to generate a composite profile, we first assumed that the probability of finding polymerase on difference bases in the bodies of the genes (GB) were uniform. Under this assumption, a normalization factor should be multiplied to a library such that the sum of the normalized reads mapped to the gene body divided by the corresponding base counts in the gene body regions becomes uniform throughout different base libraries. The normalization factor for each library is calculated as follows,

$$Normalization\ factor_{base} = \sum_{i=A,C,G,U} \frac{Reads\ mapped\ to\ GB_i}{Base\ counts\ in\ GB_i} \Big/ 4\frac{Reads\ mapped\ to\ GB_{base}}{Base\ counts\ in\ GB_{base}}$$

, where *GB* is the set of all gene body positions (table S4).

Using the normalization factors, composite PRO-seq histogram (in bedgraph format) was generated.

$$PROseq(pos) = \sum_{base=A,C,G,U} Normalization\ factor_{base} \cdot PROseq_{base}(pos)$$

This composite PRO-seq dataset was used for the downstream analysis unless specified otherwise.

Analysis of pausing level and gene activity

For the analysis of pausing level and gene activity, we first generated a list of genes for which the PRO-seq densities could be measured without having interference from other genes. From the scRNA-seq based re-annotated gene list*(8)*, we defined promoter upstream, promoter downstream, and 5' genic regions as -300 to -100 bp, +300 to +500 bp, and 0 to +500 bp from TSS respectively. For each region of the individual genes, we calculated 'active site coverage', which is the fraction of positions covered by 3' end of PRO-seq reads within each region. Because of the normalization, some positions have read counts less than 1 and we considered these positions partially covered. Active site coverage can be formulated as follows.

$$Active\,site\,coverage(region) = \sum_{pos \in region} \max\left(PROseq(pos), 1\right) \Big/ length\,of\,the\,region$$

We called genes 'upstream clear' if promoter upstream region had the active site coverage of less than 0.01, or less than one fourth of the downstream region active site coverage (n=11,584). This was intended to filter out genes that have polymerase transcribing through from the upstream genes that can interfere with the downstream levels. Among the 'upstream clear' genes, we called genes 'active' if the active site coverage in 5' genic regions was greater than 0.01 (n=5,471).

To calculate the pausing level, PRO-seq read counts per million normalized mapped reads (RPM) from -50 to +150 relative to TSS were obtained and the sum of the read count was divided by the length of the region (0.2 kb) to generate RPM per kb, or RPKM which equivalent to the commonly used definition of RPKM in RNA-seq. For the consistency of the unit usage, we also used RPKM to describe PRO-seq profiles for individual genes along the positions on the genome, regarding that a RPM read count on a single base position can be considered as a RPKM density for a 0.001 kb region.

For the gene body activity, we used active site coverage instead of read counts to minimize the effects of unexpected spikes or unannotated transcription initiation within the gene body region. This modified PRO-seq density was calculated by multiplying a conversion factor to the active site coverage from +300 from the TSS to the 3' end of the gene and RPKM normalized. For genes that contain another annotated TSS within the gene body, we truncated their gene body region to -300 bp from any downstream annotated gene starts. The conversion factor is given below.

$$Conversion\ factor = \frac{\sum_{pos \in GB} PROseq(pos)}{Active\,site\,coverage(GB) \cdot length\,of\,GB}$$

All the densities were adjusted by the mappability of 26 bp sequence uniquely to the genome.

Data visualization using scatterplots, average profiles and heatmaps

The scatterplots were generated using in-house scripts. Briefly, on a 1000×1000 pixel bitmap, each data-point was represented as filler circles with 11 pixel diameter on log axes. For each pixel, data-point counts were stackable. After plotting all the data-points, the counts in each pixel was converted to a color code, and the image was anti-aliased. The scale-bars for the color code were shown together with the colored scatterplot images. Pearson coefficients were presented on the scatterplots

Average profiles relative to position lists were generated using modified bootstrap methods and permutation tests. Briefly, position lists of N genes were randomly partitioned into [N/100] subsets (integer part of N/100) each containing ~100 members. The average profile of each subset was calculated removing 2 greatest and least outliers per relative positions. The average and the standard error of the subset profiles were calculated and usually plotted together respectively as a line and margins surrounding the line in shades. Gaussian smoothing was applied to the profiles if necessary using the formula below,

$$smoothed\ profile(pos) = \sum_{i=-3b/2}^{3b/2} \phi(2i/b) \cdot raw\ profile(pos+i) \Big/ \sum_{i=-3b/2}^{3b/2} \phi(2i/b)$$

where $\phi(x)=exp(-x^2/2)$ is the Gaussian density function, and b is the bandwidth of smoothing which is twice the standard deviation of regular Gaussian distribution. A data value is smoothed over 3 bandwidths around the data point. For most profiles, smoothing bandwidth of 2 bp was used unless specified otherwise.

Scaled ‘metagene’ profile was generated as described previously*(2)*, with modification to the scaled region of the gene body being TSS +1 kb to 3’ end −1 kb of the gene. A smoothing bandwidth of 200 bp was used.

Heatmaps were generated using in-house scripts. Briefly, a data matrix of PRO-seq read counts, with genes on the rows and relative position to each TSS on the columns, was scaled to a 200×1000 matrix with an algorithm that uses incremental accumulators for each pixel. The data values were converted to color codes and the image was anti-aliased. Typically, this generated moderately averaged profiles for gene lists containing more than 10,000 genes, but represented individual genes relatively well for comparing gene subsets containing up to 1,000 genes.

Pause peak identification and paused gene clustering by pausing pattern

With the ‘active’ genes (n=5,471) listed above, we defined PRO-seq peaks using a clustering algorithm (fig. S6A). Briefly, for each gene, we scaled the number of reads to 1,000 pseudo-reads maintaining their relative positions in the promoter proximal region (-50 to +150 from TSS), and applied a k-means clustering algorithm*(9)* by their positions to identify the peaks. The number of clusters (k), *i.e.* the number of peaks, was determined by taking the minimum k for which the variance to the cluster centroid was less than 5 (bp×bp). The k value was modified within ±1 range to have the local maximum of the average silhouette*(10)*. For each peak, the total read count of the actual reads was calculated, and major peaks greater than one fourth of the

maximum peak of the region were selected. Each peak is assigned with two parameters, average position and total read count. We repeated the same peak calling algorithm in promoter downstream regions (+300 to +500), and called a gene 'paused' if the total read count of the maximum peak at promoter proximal region is greater than 4 times the read count of the maximum peak at the promoter downstream region (n=3,225). These cut-offs are chosen for the purpose of relative comparison between groups, but they do not necessary define pausing *per se (1)* For each paused gene, we calculated the median position of the peaks and the average dispersion of the peaks weighted by the read counts. We calculated the percent rank of the median position and the average dispersion within the paused genes subset and defined them as 'position percentile' and 'dispersion percentile' of the peaks respectively for each gene.

For the 2D heatmap representation of the pausing pattern on the position-dispersion axes, we used an in-house script (Fig. 2B). Briefly, each gene was added as a 2D Gaussian peak,

$$\phi(x,y) = \frac{1}{2\pi(b/2)^2} e^{-(x^2+y^2)/2(b/2)^2}$$

where x and y are the relative position to the position vs dispersion percentile coordinate of the gene on a 2D space [0,1]×[0,1], and b is the bandwidth of the peak (20%). The overall density on the 2D space was normalized by dividing by the total gene number. Therefore, the integral over an area reflects the probability of finding a gene in the corresponding position-dispersion range, and the integral over the whole 2D space, which is the probability of finding a gene over the whole region, equals 1.

To further identify the two apparent clusters of the genes- 'Clustered proximal' (*Prox*) on the lower left quadrant and 'Dispersed distal' (*Dist*) on the upper right quadrant- in Fig. 2B, we employed an Expectation-Maximization algorithm. This was done using the 'Mclust' package in R software*(11)*. Briefly, we performed Mclust on the position-dispersion dataset with the prior specification of 2 clusters with ellipsoidal model (variable volume, shape, and orientation of covariance matrix: 'VVV' model), and initialization of a Poisson noise model (p=0.25). This was done iteratively and a representative set was chosen. Two clusters were generated allowing outliers, and we determined the cut-offs for their z-scores (*Prox*: 0.15, *Dist*: 0.08) to have maximum number of non-overlapping elements and least difference in cluster sizes (fig. S6B). Genes were assigned to *Prox* or *Dist* clusters if their z-scores were smaller than the cut-offs (n=848, 846 respectively). The average profiles show similar pausing distributions as anticipated by individual cases (fig. S6C). Since the majority of the genes lie on the diagonal of the position-dispersion space, we defined the Pausing Proximity Index (PPI) as the average of the pausing position percentile and the pausing dispersion percentile.

Analysis of the initiation from PRO-cap

PRO-cap results were processed in the same way as PRO-seq, except that the 5' ends of non-reverse complemented sequence reads were used and the promoter proximal window was set to be -100 to +100 from the TSS. Average relative profile was generated by first dividing the read

counts by total number of reads in the promoter proximal window for each gene, and then calculating the average plots afterwards (fig. S7B). This allowed us to examine the average pattern of initiation at the TSS without having the pattern be overly affected by genes with the highest read counts.

To compare the dispersion pattern of initiation and pausing, we used the identical script to identify the initiation peaks. For a direct comparison, the dispersions were shown in the actual number of base pairs instead of the percentiles in boxplots (fig. S7C).

To assess the focusing of initiation, we defined the 'read fraction at TSS' parameter (frTSS) for each gene as described previously with modifications*(12)*.

$$frTSS = \frac{\sum_{pos \in TSS \pm 1bp} PROcap(pos)}{\sum_{pos \in TSS \pm 25bp} PROcap(pos)}$$

Scoring the positions and the strengths of the DNA elements

Promoter DNA elements were identified from promoter proximal regions using existing position weight matrices (PWM) or consensus sequences by a fast permutated string-match algorithm. First we extracted a sequence substring from the promoter proximal sequence of each gene on every position, and calculated the PWM score. The score was compared to a cumulative distribution function (CDF) of the scores of 100,000 permutated sequences that were randomly generated using the same background letter frequencies. From the CDF, p-values were obtained for every position on the promoter proximal region of a gene. When the PWM was not available, we built a PWM from the log-likelihood of the consensus match at matched letters and 0 at non-matched letters.

We identified the positions and the strengths of DNA elements by generating a log likelihood array (-log p-value) for each gene at every position (fig. S8B-D). The array was smoothed with a modified Gaussian smoothing,

$$Smoothed \log L(pos) = \max\left(\phi(2i/b) \cdot \log L(pos+i); -\tfrac{3}{2}b < i < \tfrac{3}{2}b\right)$$

where *Smoothed* log$L$(*pos*) is the smoothed profile for relative position *pos*, log$L$(*pos*) is the log likelihood profile at position *pos*, *b* is the bandwidth of the Gaussian curve for smoothing, and $\phi(x)=exp(-x^2/2)$ is the Gaussian density function. This type of smoothing reports the strongest element nearby that is modified by a distance factor dependent on the bandwidth, and reflects the probability of finding a factor binding at the position within the bandwidth. A bandwidth of 10 bp was used for core promoter elements and 40 bp for GAGA elements.

Generation of the fly lines with sequence modified *Hsp70* promoter transgenes

Adult flies carrying sequence modified *Hsp70* promoter transgenes (Fig. 4A) were made as described below. First, the *Hsp70* gene was amplified and cloned into pBluescriptII KS+. The 5 bp insert and 10 bp insert transgene were created using site-directed mutagenesis with the following primer sets; +5 bp forward: 5'-CGACGGAGAGTCAATTCAATTCAAACAAAACA AGCAAAGTGAACACATCG C-3'; +5 bp reverse: 5'-GCGATGTGTTCACTTTGCTTGTTTT GTTTGAATTGAATTGACTCTCCGTCG-3'; +10 bp forward: 5'-CGACGGAGAGTCAATTC AATTCAAACAATGAGTCACAAGCAAAGTGAACACATCGC-3'; +10 bp reverse: 5'-GCG ATGTGTTCACTTTGCTTGTGACTCATTGTTTGAATTGAATTGACTCTCCGTCG-3'.

The Hsp70 transgenes were created to minimize disruption of any core promoter elements. The placement of the insert 10 base downstream of the TSS is between the Initiator element (-2 to +4) and the downstream elements (DPE +28 to +32, MTE +18 to +27). In addition, the 5 and 10 base inserts were a 5 base duplication of the adjacent sequence and the 5 base duplication plus 5 bases randomized to separate the duplication, respectively. These inserts minimize the disruption of any unknown elements and do not alter the base composition of the region.

The gene was cut out of the wild type (*Hsp70wt*), 5 bp insert (*Hsp70+5*), and 10 bp insert (*Hsp70+10*) plasmids with XbaI to yield a fragment from -245 to +1863 (relative to the TSS), and cloned into a modified pCasper4 containing the *attB* site*(13)*. The ΦC31-mediated transformation was performed by Best Gene Inc. to insert each transgene into 22A3 (*PBac{yellow[+]-attP-3B}VK00037*). The lines were balanced and crossed to *Hsp70* null flies*(14)* (Bloomington 8841: *w1118*; *Df(3R)Hsp70A*, *Df(3R)Hsp70B*) to create homogenous stocks.

PRO-seq and the analysis of transgenic *Hsp70* promoter fly lines

The nuclear isolation from adult flies was adapted as described previously*(15)*. One gram of flies were homogenized in 15 ml cold Buffer A (10 mM Tris-HCl pH 8.0, 30 0mM sucrose, 3 mM $CaCl_2$, 2 mM $MgOAc_2$, 0.1% TritonX-100, 0.5 mM DTT) for 1 minute using the Omni-mixer, the homogenate was filtered through 100 um nylon mesh into a 40 ml Dounce homogenizer. After 40 strokes in the homogenizer, the homogenate was filtered through 35 um nylon mesh and mixed with an equal volume of Buffer B (10 mM Tris-HCl pH 8.0, 2 M sucrose, 5 mM $MgOAc_2$, 0.5 mM DTT). The homogenate was then layered over 10 ml Buffer B in 35 ml Ultracentrifuge tube, and centrifuged through the Buffer B cushion at 12 krpm for 25 minutes at 4˚C in a SW28 swinging bucket rotor. The supernatants were removed, and the nuclei were resuspended to the density of 1×10^8 nuclei/ml in 1 ml buffer C (50 mM Tris-HCl pH 8.0, 25% glycerol, 5 mM $MgOAc_2$, 0.1 mM EDTA, 5 mM DTT).

RT-qPCR analysis of heat shock induction of the transgenic *Hsp70* promoters was carried out as follows. Five to ten larvae from the indicated lines were transferred to a 1.5 ml tube and

incubated at 37˚C for the indicated times (5, 10 and 20 min). The larvae were immediately homogenized in the homogenization buffer and the RNA was isolated using Omega E.Z.N.A. Total RNA kit I (R6834), and quantified using NanoDrop 1000 spectrophotometer. Duplicate reverse transcription reactions were performed with 200 ng of total RNA using SuperscriptIII reverse transcriptase (Invitrogen 18080) with oligo(dT) primer. After the reactions were diluted 10-fold with 10 mM Tris-Cl (pH 8.0), 2 ul was used in 10 ul qPCR reactions to quantify the cDNAs using the following primer sets; *Hsp70Ab* +2155F: 5'-GGTCGACTAAGGCCAAAGA GTCTA-3'; *Hsp70Ab* +2266R: 5'-TCGATCGAAACATTCTTATCAGTCTCA-3'; *Hsp70* +1649F: 5'-GGGTGTGCCCCAGATAGAAG-3'; *Hsp70* +1754R: 5'-TGTCGTTCTTGATCGT GATGTTC-3'; *Hsp26* +580F: 5'-CAAGGTTCCCGATGGCTACA-3'; *Hsp26* +667R: 5'-CTGC GGCTTGGGAATACTGA-3'; *Rp49* +549F: 5'-CCCAAGGGTATCGACAACAGA-3'; *Rp49* +613R: 5'-CGATGTTGGGCATCAGATACTG-3'; *Actin5C* +1781F: 5'-GGAAATCCGCATT CTTTCCA-3'; *Actin5C* +1848R: 5'-CGACAACCAGAGCAGCAACTT-3'. The qPCR was run on the Roche LightCycler480, and the level of each relative Rp49 was calculated using $2^{-\Delta C(t)}$.

Supplementary Text

Validation of the PRO-seq method

In PRO-seq, only one of the four biotin-nucleotide triphosphates (biotin-A/C/G/UTP) was supplied without any unmodified NTP in each of the four parallel run-on reactions. Pol II would incorporate the biotinylated base only if the provided base were complementary to the active site DNA template. After incorporating the template base and encountering a different base, Pol II would stops, ensuring a near base-pair resolution match of the 3' end of the run-on RNA to the Pol II active site. We noticed that in control run-on assays where the biotin-NTP was supplemented with the other 3 unlabeled NTPs and a radioactive NTP tracer that very little radioactivity was incorporated relative to reactions containing no biotin-NTP. Nonetheless, the biotin-NTP was incorporated. This suggested that the biotin-NTP once incorporated into the nascent run-on RNA blocks further elongation. This then suggested an alternative, more convenient assay might work, where we used all 4 biotin-NTPs together in a single reaction (PRO-seq$_{4NTP}$).

We compared the run-on length of a PRO-seq$_{4NTP}$ run-on reaction with the four parallel conditions using one of the biotin-NTPs (fig. S2A). We measured the lengths of biotin-NRO RNA originating from *Hsp70* TSS using a modification of a ligation mediated PCR described in a previous study*(4)*. The result shows an agreement between the composite profile of the four parallel run-on reactions and the alternative 4NTP condition (fig. S2A). This indicates that Pol II does not elongate transcripts beyond the first active site with biotin-NTPs even in the presence of the correct substrate. In addition, considering that the transcription bubble extends from about -14 to +1 relative to the active site, this distribution matches the typical permangate footprint of *Hsp70* gene at +22, +30, and +34 positions (fig. S2A). We also compared the high resolution genome-wide profiles between PRO-seq and PRO-seq$_{4NTP}$, and confirmed this finding (fig. S2B~D). These results indicate the PRO-seq version with all four NTPs can be used as a convenient and economical protocol to identify the polymerase active sites at base pair resolution across the genome.

To ensure that the positions of polymerase mapped in isolated nuclei by PRO-seq represent the location of polymerase *in vivo*, we tested whether there was leakage elongation of Pol II during the nuclei preparation procedure (fig. S2E). Cells were homogenized to release nuclei and immediately divided into three aliquots, each reproducing typical conditions during the nuclei preparation procedures. In the first and the second conditions (L1 and L2), a trace amount of radioactive ^{32}P-UTP was added to the intracellular NTP pool in the homogenate. The estimated UTP concentration in the homogenate is at maximum ~10 μM, under the assumptions that the

intracellular UTP concentrations are ~1 mM, the average diameter of S2 cell is ~ 5 μm, and the cell density of the homogenate is 2×10^7 cells/ml. Nuclei were left on ice (L1) or centrifuged at 4˚C (L2) for 10 minutes to mimic the run-on leakage of the polymerase that might occur during nuclei preparation, and then the RNA was extracted. The amount of radioactivity detected on the nascent RNA per polymerase molecule will be the expected number of U bases incorporated by leakage run-on (= run-on length / 4) multiplied by the specific activity of ^{32}P-UTP.

For the third condition (R), nuclei were washed to remove all intracellular NTP pool and only single type of base (UTP+^{32}P-UTP) were added under the standard sarkosyl nuclear run-on conditions, to force all polymerases to incorporate UTP in a semi-terminating manner. Each polymerase molecule will have on average 1/4 chance of having UTP incorporated at the active site, and after incorporating the base, it will have another 1/4 chance of UTP incorporation, and so on. The average number of UTP incorporation per polymerase molecule will then be a geometric series of $1/4 + 1/4^2 + 1/4^3 + \ldots = 1/3$. Since UTP was supplied to have a similar specific activity as L1 or L2 conditions, the expected radioactivity of the 'R' condition per polymerase molecule will be 1/3 multiplied by the specific activity of ^{32}P-UTP. By measuring the relative radioactivity ratio between the leakage conditions and the standard run-on condition (L/R), average run-on lengths under the leakage conditions could be estimated to be $4/3 \times L/R$ (relative activity). In all the size ranges of nascent RNA, these estimated run-on lengths were much less than 1.0 (fig. S2E), indicating that the leakage run-on during the nuclei preparation step was negligible.

Also, we checked if shorter nascent RNA is selectively lost during the library preparation step or during the read alignment step to generate biases in the positions of pausing (fig. S2F). First, we examined distribution of the read lengths of the inserted sequences, and confirmed that RNA longer than 18 nt are not selectively lost in this distribution. Then we calculated the 'mappability' plot depending on the read length from the TSS, and reads longer than 21 nt are not selectively lost during the read alignment step. The actual distribution of paused Pol II active site starts to rise downstream of +25, showing that this distribution of pausing is not likely to be biased by selective loss of short nascent RNA originating from the TSS.

Finally, we also compared the PRO-cap, PRO-seq and permanganate footprints*(16)* in individual genes. 26 genes that have the permanganate reactivity near TSS annotation based on short capped RNA-seq (scRNA-seq*(8)*) were shown (fig. S2G). The average region of permanganate reactivity was compared to the average PRO-seq profile in these genes, showing the expected overlap of the average positions.

To validate that PRO-seq agrees with other types of genome-wide transcription assays, we generated scatterplots for the promoter proximal and gene body/exon densities between PRO-seq and GRO-seq *(3)*, nuclear short capped RNA-seq (scRNA-seq)*(8)* and RNA-seq (modENCODE_3138; GSM461182)*(17)* (fig. S3).

Analysis of the splicing junctions

The splicing sites list was generated from the Refseq gene list*(18)*. Any sites within 2 kb distance from annotated TSS were removed to produce the final lists of 3' splicing sites (exon start sites; n=41,005) and 5' splicing sites (exon end sites; n=41,710). Average profiles were generated relative to the splicing sites as describe above, except that the modified PRO-seq density derived from PRO-seq read coverage was used to minimize the effects of unexpected spikes or unannotated transcription starts (fig. S4A).

To identify low usage exons, we used an existing modENCODE RNA-seq dataset in S2 cells (modENCODE_3138; GSM461182)*(17)*. The alignment data was collapsed to a histogram in bedgraph format, and the average RNA-seq read counts per million mapped reads was generated for each exons. Exon starts were defined from the 3' splicing site list above (n=41,005) and their ends were chosen from the nearest 5' splicing sites if alternative exons existed. For every gene, the average RNA-seq density was calculated by averaging the exon densities weighted by exon lengths, and relative exon densities were calculated by dividing the exon densities by average gene density. Exons were called 'low usage' if the relative exon densities were less than 0.05, and the differences between adjacent exons were greater than 0.5 (n=242). Any exon that had zero PRO-seq density was removed to avoid the possibility of sequencing or alignment biases. Exons that were adjacent to the 'low usage' exons were selected as matched-pair controls. The RNA-seq densities of 'low usage' and their neighboring exons were examined in boxplots to ensure that the exons were skipped (fig. S4B). The PRO-seq densities were also compared to show that the exons were transcribed at similar levels (fig. S4C). The average PRO-seq profiles at the start sites of these exons are shown in Fig. 1E.

To test whether a specific sequence composition or the actual exon usage is associated with the Pol II accumulation at the junctions, we selected annotated 3' splicing sites with CAGRT consensus*(19)* (n=4,102). Non-spliced consensus sites within the same gene bodies are matched to the spliced consensus sites to generate a matched control group (n=10,552). The average PRO-seq profiles were generated as described above (fig. S4D). The sequence logo (fig. S4E) were generated by in-house scripts, using the rules described previously*(20)*.

We also checked whether base composition near the splicing site was contributing to the PRO-seq by examining each of PRO-seq$_{A/C/G/UTP}$ libraries separately. The average profiles of 4 separate libraries were obtained as described above, except that each profile was normalized by its average for comparison, which showed similar accumulation pattern immediately downstream of 3' splicing sites regardless of the base (fig. S4F).

It has been previously documented that introns have lower sequence complexity and are less 'mappable' than exons*(21)*. To assess the effect of mappability difference in our *Drosophila* PRO-seq analysis, we first extracted varying lengths of short sequences around 5' splicing sites and plotted the fraction of sequences that were uniquely aligned to the reference genome by the relative positions of their 3' ends (fig. S4G). The alignments were done using the same parameters for the alignment of the PRO-seq reads. Sequence lengths of 21, 26, and 36 bases were tested, and we choose 26 bases to be an acceptable minimum read length at which the mappability would not affect the interpretation of average profiles. We then generated average

PRO-seq profiles from all reads and reads equal to or greater than 26 bases in PRO-seq$_{4NTP}$ library (fig. S4H). From this we confirmed that the same accumulation pattern was observed with longer reads. We further reproduced Fig. 1E and fig. S4D with the longer reads, and confirmed that the findings were not biased by mappability (fig. S4I/J).

Analysis of the pausing at nucleosomes

For this analysis, we compared the PRO-seq density relative to the nucleosomes (Fig 1F). The positions of the dyad center of nucleosomes were obtained from a micrococcal nuclease sequencing (MNase-seq) data from the Adelman lab*(12)* in S2 cells (fig S5A). When we compared the average PRO-seq density relative to the nucleosome centers in the gene bodies, we saw accumulation of Pol II active site at around -40 from the nucleosome centers. This is consistent with the observation in yeast from a previous NET-seq study and prediction from the DNA-nucleosome interaction models*(22,23)*.

But the PRO-seq density relative to the first nucleosome is different. The average PRO-seq density relative to the first nucleosome shows maximum accumulation at around -80 from the nucleosome centers. This is inconsistent with a simple nucleosome barrier model. This becomes more clear when we compare the PRO-seq distribution of the *Prox* and the *Dist* genes with the first nucleosome (fig. S5C). The *Prox* genes are more sharply positioned at around -80 from the nucleosome centers, while *Dist* genes are more dispersed. The *Dist* genes show the same peak at -80 but also a possible subpeak at around -40. This indicates that *Dist* genes may have a component of pausing that is established by direct nucleosome barriers, but the majority of *Prox* genes are more related to the promoter structure. Moreover, 37% of the *Prox* genes didn't have detectible MNase-seq signals to assign the first nucleosome between +0 to +200, whereas only 22% of the *Dist* genes didn't have prominent first nucleosomes. Also, the average occupancy of the first nucleosome in *Prox* gene is about half that of the *Dist* gene, suggesting that for some *Prox* genes, the first nucleosome may even be absent.

Analysis of the association between the initiation and pausing

PRO-cap results were processed in the same way as PRO-seq, except that the 5' ends of non-reverse complemented sequence reads were used and the promoter proximal window was set to be -100 to +100 from the TSS. Average relative profile was generated by first dividing the read counts by total number of reads in the promoter proximal window for each gene, and then calculating the average plots afterwards (fig. S7B). This allowed us to examine the average pattern of initiation at TSS without having the pattern be overly affected by genes with the highest read counts. To compare the dispersion pattern of initiation and pausing, we used the identical script used for identifying pause peaks to identify the initiation peaks. For a direct comparison, the dispersions were shown in the actual number of base pairs instead of the percentiles in boxplots (fig. S7C), although this calculation was more affected by some genes that have multiple TSSs.

As defined in the main text, the fraction of PRO-cap reads at TSS (frTSS) between *Prox* and *Dist* genes were compared in boxplots (Fig. 2E). Conversely, genes that are either *Prox* or *Dist* groups were divided up into quartiles by frTSS value, and the top and the bottom quartiles were chosen as 'focused' or 'dispersed' initiation genes. The pausing proximity index (PPI) was compared between the focused and dispersed initiation groups (Fig. 2E). In addition, we also used the matched-pair controls described above for the same analyses to exclude the confounding effect of the absolute level of pausing on TSS focusing (fig. S7D/E).

Analysis of the DNA elements and the association with pausing

Promoter DNA elements were identified from promoter proximal regions using existing position weight matrices (PWM) or consensus sequences by a fast permutated string-match algorithm. We used the PWMs for core promoter elements (fig. S8A) from Ohler *et al.(24)*. When the PWM was not available, we built a PWM from the log-likelihood of the consensus match at matched letters and 0 at non-matched letters. This was the case for the consensus sequences from a collection of *Drosophila* functional motifs from Stark *et al. (25)*, and the consensus for 'Pause Button' from Hendrix *et al(26)*. For the *Drosophila* functional motifs, we used 145 'predicted transcription factor motifs' (ME1~145) and 87 'recovered known transcription factor motifs' (TE1~87). Many of the ME and TE motifs overlap.

To generate an average motif frequency plot in a group of genes (Fig. 3A/B, fig. S8B-D), we scored every position relative to the TSS for a motif match with the p-value of < 0.001 as a count. The count at each position was divided by the total number of genes and per 100 bp, and was smoothed using the regular Gaussian smoothing. This gave us the expected frequency of a motif per 100 bp at the position. The average difference of motif frequency between *Prox* and *Dist* genes for the entire set of DNA elements in Stark *et al.(25)* were presented as heatmaps (fig. S8E).

The "optimal" position of each DNA element was determined based on the following criteria. First is the peak consensus position of motif frequency relative to TSS in the Drosophila genome. Second is the position from the previous literature on the core promoter elements. For TATA, Inr, MTE, and DPE, the positions based on the motif frequencies match well with the positions and from in vitro and in vivo tests of these elements mainly by the Kadonaga Lab. For Pause Button (PB), its similarity to MTE and DPE was considered. The peak frequency positions of PB are at +18, +25, and +38, but only +38 peak position was taken, because the +18 and +25 positions corresponded to MTE and DPE positions. Finally, positions within ±5 from the peak frequency positions were considered optimal.

To test the dependency of pausing position upon the position of core promoter elements, we first chose the gene set which the pausing sites and DNA element positions were unambiguously determined. For each DNA element, we picked genes that have only one instance of the element within ±40 bp from the consensus position from either *Prox* or *Dist* subsets. The genes were then classified by the position of the element, upstream (Up), consensus (Cs) or downstream (Dn) (Fig. 3D: TATA and PB, fig. S9A: Inr). Alternatively, the first digit of the element position relative to TSS was used to identify the element position where the consensus positions were different from canonical positions (fig. S9B: MTE, fig. S9C: DPE). The number of genes for

each subset are described in table S2. The pausing position percentiles in genes with different DNA element positions were compared in boxplots. In addition, pausing indices of these subsets were also compared between these genes to show that the extent of pausing also correlates with the optimum positioning of the DNA elements (fig. S9D).

To determine the strength of a DNA element at the consensus position for each gene, we generated smoothed likelihood plots of sequence match to the DNA elements 'Active' genes (n=5,471) were divided into subsets according to the smoothed likelihood level for each DNA element and the promoter proximal (-50 to +150) PRO-seq levels were compared in boxplots (Fig 3E, table S3).

Sequence modified *Hsp70* promoter transgenes

We made PRO-seq$_{4NTP}$ libraries from the adult flies carrying sequence modified *Hsp70* promoter transgenes (Fig. 4A). The insert sequences do not significantly change the energy landscape of DNA-RNA hybrid base pairing generated by the transcription bubble (fig. S10A). For these libraries, RNA fragmentation was done for a limited time to conserve full-length nascent transcripts at *Hsp70* promoters. For the validation of each library, the sequencing data were processed in the same way as describe in the previous sections. The correlations of the promoter proximal levels of non-*Hsp70* genes between the different fly line libraries were shown in scatter-plots (fig. S10B). Since *Hsp70wt*, *Hsp70+5* and *Hsp70+10* libraries have high correlation with each other in non-*Hsp70* genes, we calculated the standard deviation of promoter proximal PRO-seq densities for each gene. The standard deviations had high correlation with the average of pausing level, and we used them to estimate the amplitude of the errors in the level of sequence modified transgenic *Hsp70* read counts.

For the mapping of the sequence reads to *Hsp70* promoters, we inserted +5 bp or +10 bp elements at the +15 position from TSS (Fig. 4A) of the endogenous *Hsp70* sequence (-100bp to +1.8kb). After the clipping the adaptors and making the reverse-complements of the sequence reads, we aligned reads that are greater than 20 bases from each library to the corresponding reference using Bowtie*(7)* as described above. The sequence reads in this case were not trimmed so that both 5' and 3' ends of the reads could be analyzed to give the initiation site and pausing site information simultaneously. This was possible because *Hsp70* genes have relatively upstream pausing positions and shorter paused nascent RNA lengths. Finally, the reads were normalized by dividing by the total number of sequence reads from TSS to +80 to compare the pattern of initiation and pausing positions (Fig. 4B,D).

The PRO-seq results from transgenic fly lines with directed insertion mutations of the *Hsp70* genes showed the level of pausing was reduced. These results indicate that disrupting the spacing of these pausing sites can affect the level of pausing. To assess whether the promoters of these transgenes could induce transcription, we used RT-qPCR to measure *Hsp70* mRNA levels and control genes after heat-shock (fig. S10C~F). Comparable amounts of RNA from each time points were assayed based on the levels of *Actin5C* transcripts (fig. S10C). In addition, the levels of *Hsp26* mRNA show that all lines had the similar levels of heat shock induction (fig. S10D).

Since the transgenes do not contain the 3’ end of *Hsp70*, they can be distinguished from the endogenous genes by reverse transcribing the RNA with primers to the 3’ end. As expected, HS induced wild type flies (*w1118*) had high levels of 3’ *Hsp70* RNA, while *Hsp70 null* and all of the transgene lines did not have any RNA within this region (fig. S10E). In contrast, when a primer complementary to a region contained in the transgenes was used to reverse transcribe the RNA, all of the transgene lines produced RNA with this region. The *Hsp70+10* transgene was induced to 70-75% of the *Hsp70wt* and *Hsp70+5* transgene levels (fig. S10F). Although this may suggest that pausing does not affect induction level, the dramatic reduction in pausing on *Hsp70+10* does not eliminate pausing entirely. In addition, disruption of pausing through RNAi depletion of NELF identified many genes that were dependent on pausing to maintain an open promoter, but *Hsp70* was not one of these genes*(3, 27)*. There is evidence from previous studies that the multiple core promoter elements present on *Hsp70* allow robust induction even when the downstream core promoter element is disrupted*(28)*, and reduction of pausing reduces, but does not eliminate, induction*(15)*. Thus, the promoter architecture of *Hsp70* can maintain robust heat shock transcription even when pausing is reduced.

Figure S1

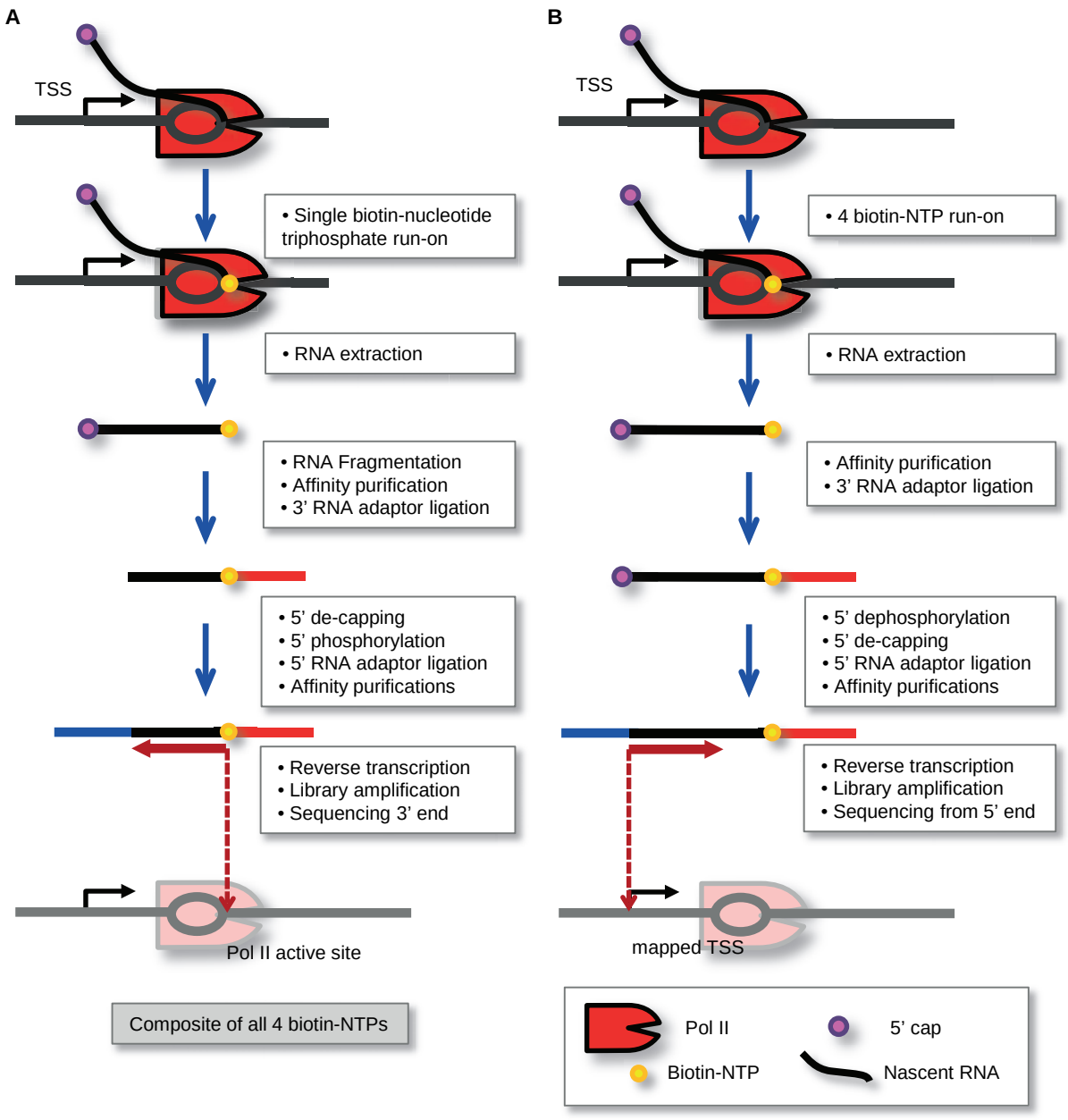


Figure S1. Detailed schematics of PRO-seq procedures.
(**A**) PRO-seq is used to identify polymerase pausing sites. (**B**) PRO-cap is used to identify transcription start sites at nascent RNA production stage.

Figure S2

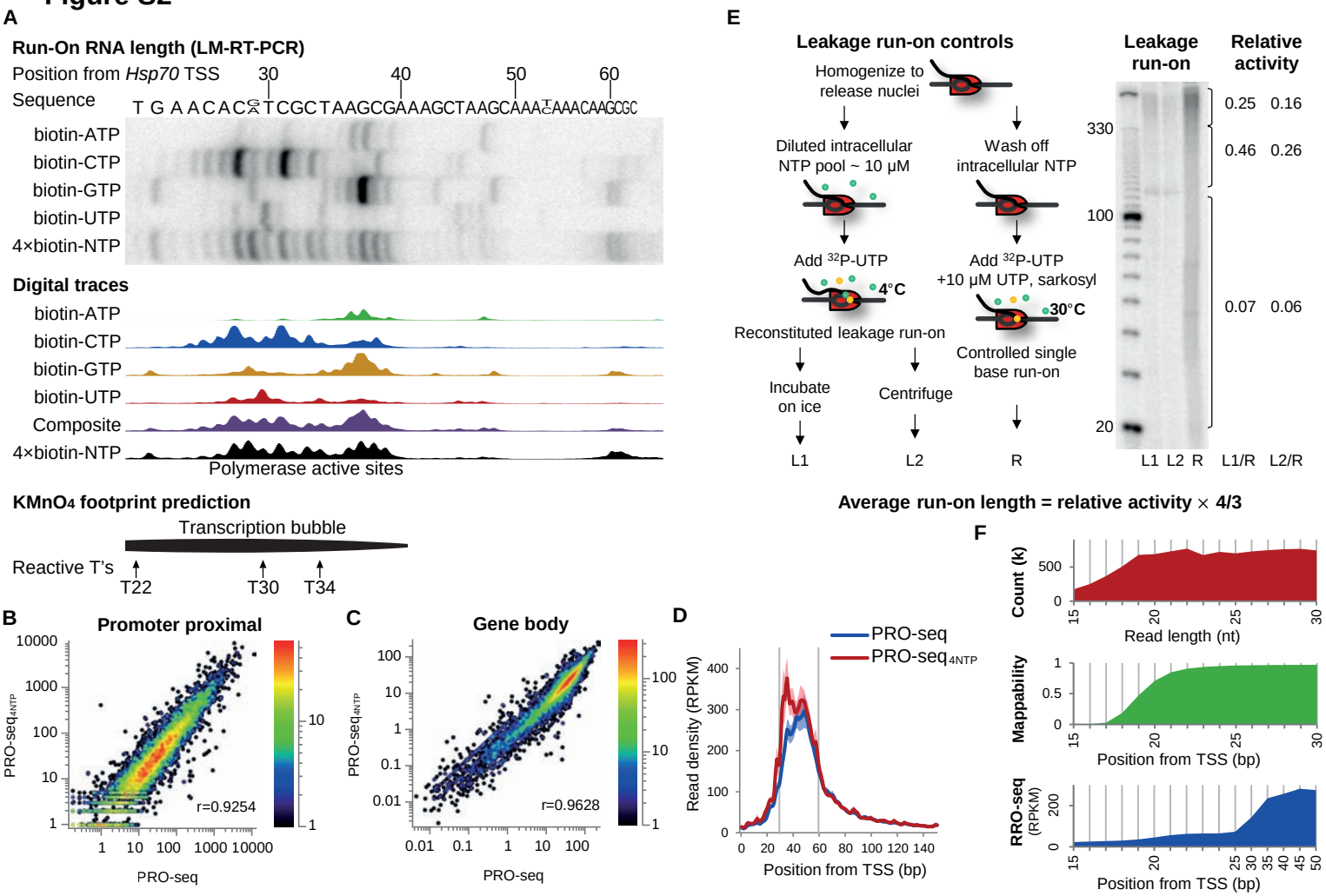
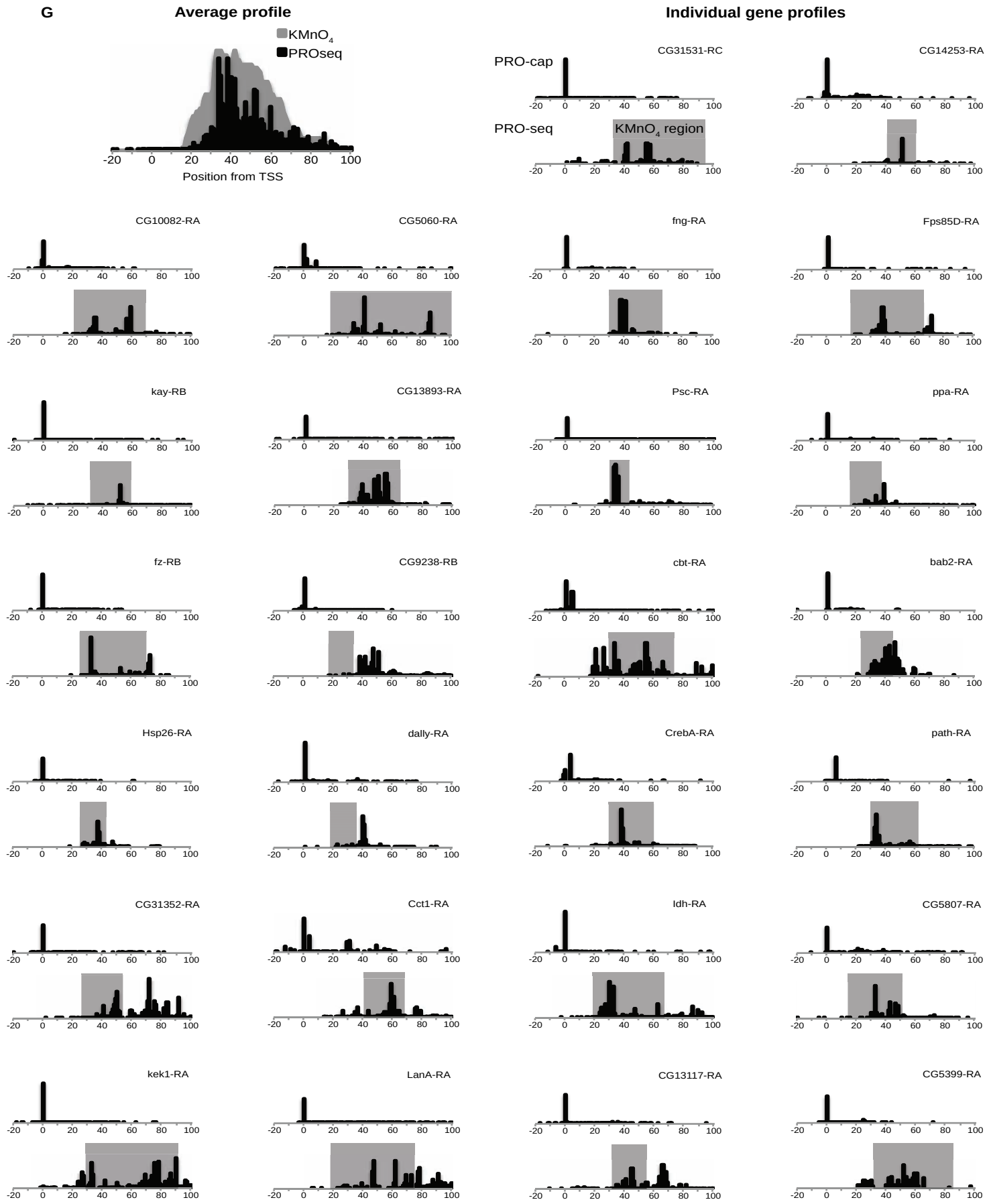


Figure S2. Validation of the PRO-seq method. (S2G on the next page)

See 'Validation of the PRO-seq method' pp. 11-12 in the supplementary text. (**A**) Analysis of 3' ends of *Hsp70* nuclear run-on RNA. Nuclear run-on RNA from each of the 4 biotin-NTP and all 4 biotin-NTPs in a single reaction were analyzed by ligation mediated RT-PCR. *Hsp70* gene-specific 5' PCR primer and universal 3' adaptor primer were used for amplification. The lengths of the PCR products represent the 3' end position of the nascent RNA after the length of the 3'-adapter (18bases) is subtracted. The sequences from the 6 copies of endogenous *Hsp70* genes are shown at corresponding positions. Image traces of the biotin-NTP run-ons and the computed composite trace are shown above the 4NTP run-on gel intensity plot. Note that the composite of the 3' end positions in 4 separate libraries matches the 4NTP library pattern for *Hsp70*, and the expected position of the transcription bubbles (bottom) fits well with previously known permanganate reactivity sites at *Hsp70*. (**B**) Promoter-proximal density correlation between PRO-seq and PRO-seq$_{4NTP}$ at each gene. (**C**) Gene body density correlation between PRO-seq and PRO-seq$_{4NTP}$. (**D**) Average profiles of promoter-proximal PRO-seq density in PRO-seq and PRO-seq$_{4NTP}$ as described in Fig. 1C. (**E**) Leakage run-on experiment controls for the nuclei isolation procedure. ^{32}P-UTP substrate was added to trace the length of leakage run-on under typical conditions during nuclei isolation. A standard single base nuclear run-on reaction was carried out with similar specific activity of ^{32}P-UTP as a reference to estimate the leakage run-on lengths. The average run-on length equals to 4/3 × relative activity (see the supplementary text), which is less than 1.0 nt on average. The ~130 nt bands are speculated to be a single species of nascent transcripts that is produced by non-Pol II polymerase. **(F)** Nascent RNA library read length (shown by insert size) and read alignment steps (shown by mappability), do not explain the actual pattern of pausing. Average PRO-seq profile downstream of +25 has been scaled by one fifth to compare the relative pausing level on the shorter length region. (**G**: **Continued on the next page**) Comparison of the PRO-seq , PRO-cap, and permanganate footprint*(16)* average profiles in selected genes and individual profiles.

Figure S2 (Continued)



Permanganate footprints were obtained from Lee *et. al.* at http://mcb.asm.org/content/suppl/2008/04/18/28.10.3290.DC1/S1_ChIP_KMnO4_59genes_111607.zip

Figure S3

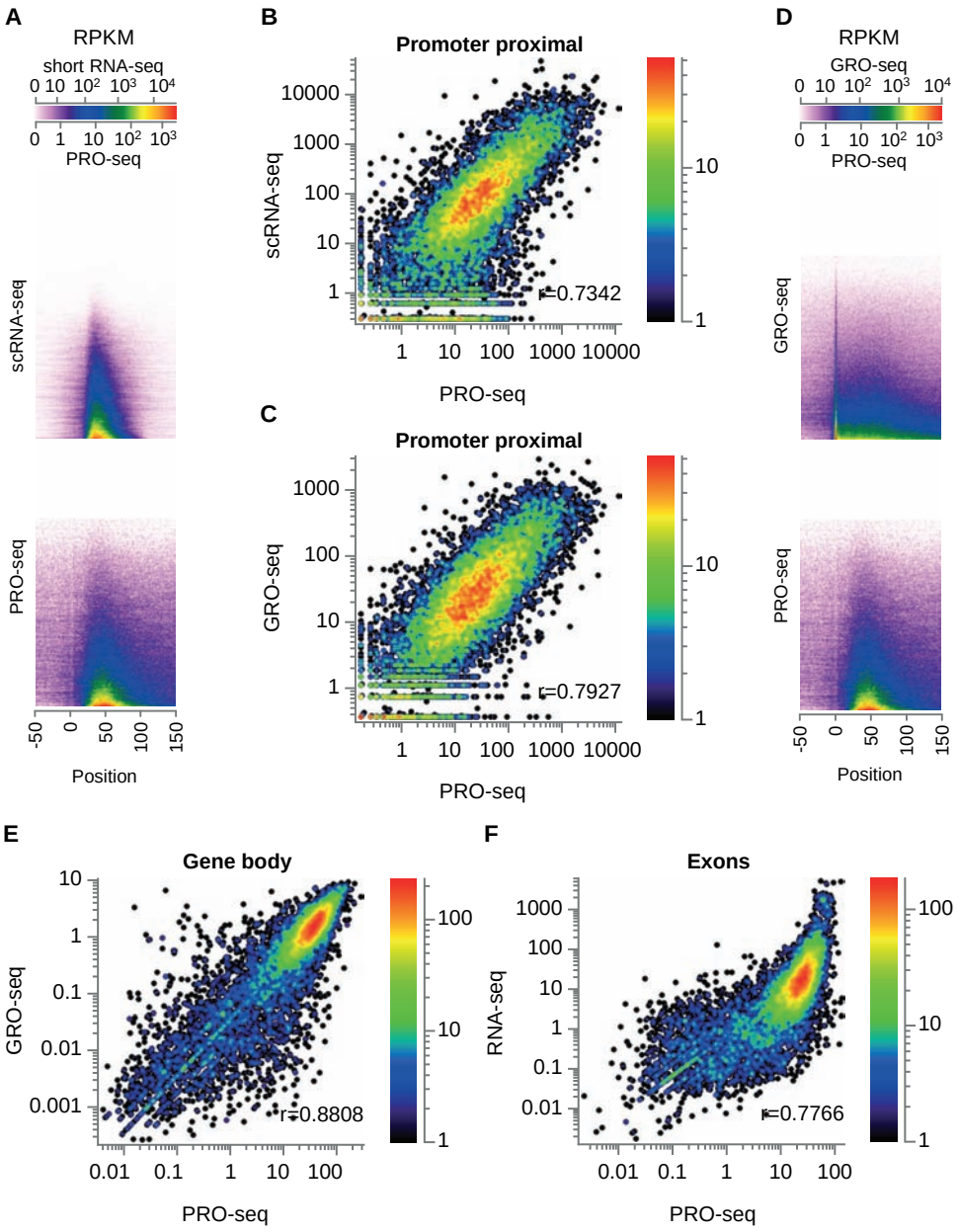


Figure S3. PRO-seq compared with other genome-wide transcription assays.
(**A**) High resolution heatmap of short capped nuclear RNA-seq (scRNA-seq)*(8)* and PRO-seq profiles near the TSS. Each profile is sorted by the sum of the reads in the region for each gene. (**B,C**) Scatterplots of promoter-proximal densities between PRO-seq and scRNA-seq (**B**) or GRO-seq (**C**)*(3)*. Promoter-proximal densities are the read densities from -50 to + 150 around TSS, RPKM normalized. (**D**) High resolution heatmap of GRO-seq and PRO-seq profiles near TSS. (**E**) PRO-seq and GRO-seq density correlations in gene body regions. (**F**) PRO-seq and poly-A RNA-seq (modENCODE_3138; GSM461182)*(17)* densities in exons. The upward concave curvature of the distribution suggests the possibility of elongation rate, nascent RNA processing efficiency, or mRNA stability being positively correlated with density of transcribing polymerases.

Figure S4

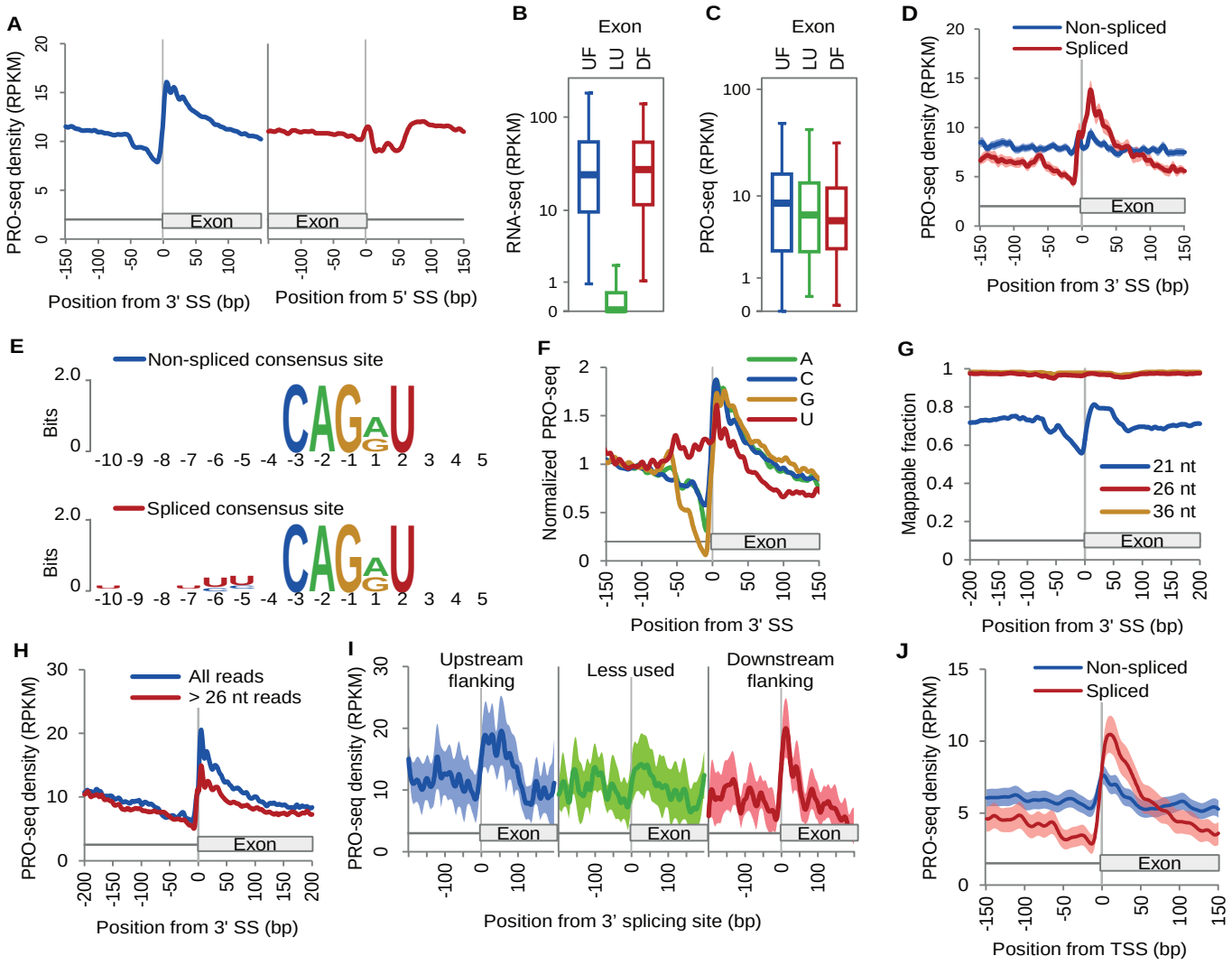


Figure S4. Pol II accumulation at 3' splicing junctions.
See 'Analysis of the splicing junctions' pp. 13-14 in the supplementary text. (**A**) Average PRO-seq profiles at 3' and 5' splicing junctions (n=41,005, 41,710 respectively). Splicing site list was generated from Refseq*(18)* annotations and sites positioned within 2 kb from any annotated TSS were removed to exclude the influence from promoter-proximal region. (**B**, **C**) Identification of low usage exons. Exons with significantly lower RNA-seq*(17)* density were selected (n=242). Their upstream and downstream flanking exons were used as controls. RNA-seq (**B**) and PRO-seq (**C**) densities of each group are shown as boxplots. Average PRO-seq profiles at these exons are compared in Fig. 1D. (**D**, **E**) Pol II accumulation at annotated 3' splicing sites or randomly selected non-splicing sites with 3' splicing consensus sequence. Annotated 3' splicing sites containing "CAGRT"*(19)* consensus sequence were selected (n=4,102), and non-spliced consensus sites within the same gene bodies were matched to the spliced consensus sites to generate a matched control group (n=10,552). The base composition of each group is shown in sequence logo representation. (**F**) Pol II accumulation at 3' splicing sites in separate PRO-$seq_{A/C/G/UTP}$ libraries prior to composite profiling. Average profile in each library is normalized to the average PRO-seq density from -150 to +150 bp relative to 3' splicing sites. (**G**) Sequence fragment size effect on read 'mappability' near 3' splicing sites. Sequence reads with varying lengths (21, 26, 36 nt) were artificially generated from the reference sequences around 3' splicing sites and mapped to the genome. The fraction of uniquely mappable reads is plotted for each read length group. (**H**) PRO-seq profile at 3' splicing junctions using only the highly mappable sequence reads that are longer than 26nt from PRO-seq_{4NTP} library. (**I**, **J**) PRO-seq profiles re-generated using sequence reads longer than 26nt for 'mappability' control.

Figure S5

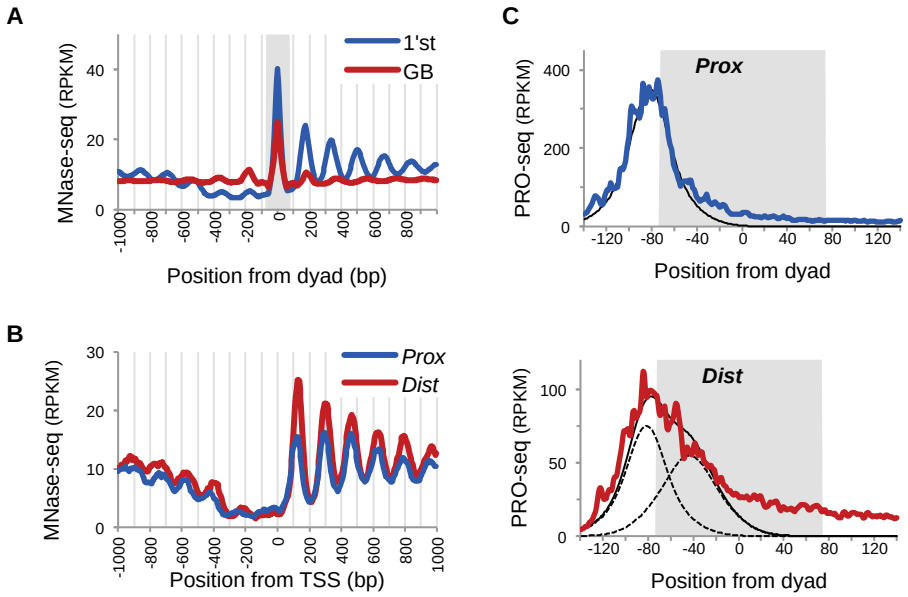


Figure S5. PRO-seq relative to nucleosomes.
(**A**) Micrococcal nuclease (MNase)-seq profile relative to nucleosome centers in gene body and the first nucleosome. Nucleosome centers were defined from a local Gaussian fit of the MNase-seq data*(12)* within 175 bp windows. Nucleosome center positions that were more than 2 kb downstream of the TSS were considered gene body, and those that are between 0 to +200 bp from the TSS were considered the first nucleosomes. Regions of the nucleosome occupancy are shaded in grey. (**B**) MNase-seq profile relative to TSS in *Prox* and *Dist* genes. (**C**) Average PRO-seq profile relative to the first nucleosome center in *Prox* and *Dist* genes (See fig. S6 for a description on the *Prox* and *Dist* genes). Fit curves with broken strokes for the *Dist* genes show possible combination of the *Prox* fit and the gene body nucleosome fit (Fig. 1F) that are centered at -83 and -44 from nucleosome centers, respectively.

Figure S6

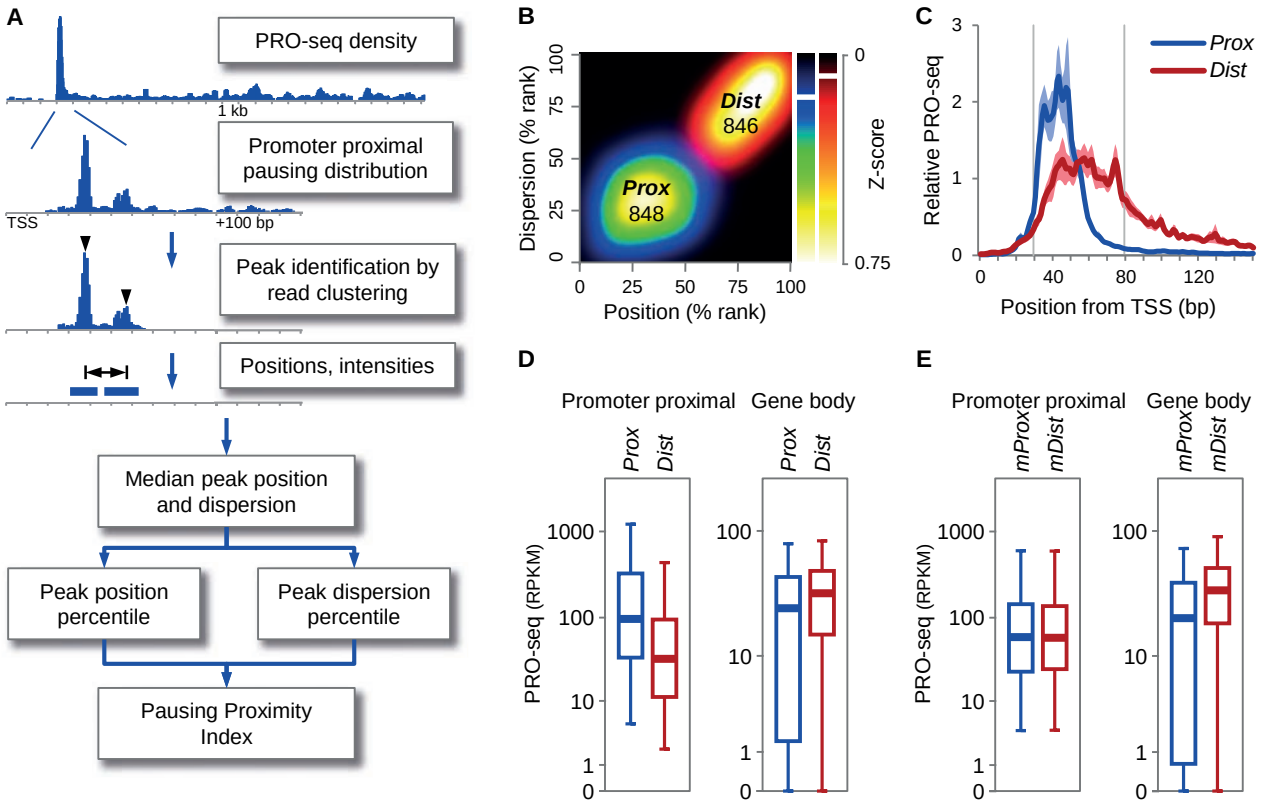


Figure S6. Parameters of the pausing patterns.
(**A**) Flowchart of pausing pattern analysis. PRO-seq profiles were obtained for individual genes (n=16,746) and analyzed as follows. To select the genes that were clearly paused, genes that have significant upstream interfering PRO-seq reads were excluded, and only the genes that have significant enrichment of promoter-proximal densities were included (n=3,225, Supplementary text). From the promoter-proximal profiles, peaks were identified from read clustering. Peak positions and dispersions were calculated from the identified peaks, and were each indexed by taking the percentile rank among the 3,225 clearly paused genes. Pausing proximity index (PPI) was further defined as the average of these two percentiles. (**B**) Expectation-Maximization clustering*(11)* of the paused genes according to the position and the dispersion percentiles. Paused genes (n=3,225) were clustered into *Prox* and *Dist* groups allowing outliers, and heatmaps of the Z-scores for either group is shown. The cut-offs were chosen to maximize the number of genes in each cluster without any overlapping genes. (**C**) Average PRO-seq profiles of *Prox* and *Dist* groups. To compare the pattern, the levels were normalized to the average of the read counts within the pausing region (+30 to +80) in each group. (**D**) Promoter-proximal and gene body PRO-seq levels of clustered proximal (*Prox*, n=848) and dispersed distal (*Dist*, n=846) gene classes of pausing. PRO-seq levels were normalized to RPKM. (**E**) Promoter-proximal and gene body PRO-seq levels of matched-pair controls for clustered proximal (*mProx*) and dispersed distal (*mDist*) gene classes of pausing (n=600 each). For every gene in a class, its counterpart gene with similar level of promoter-proximal PRO-seq density was paired in the other class.

Figure S7

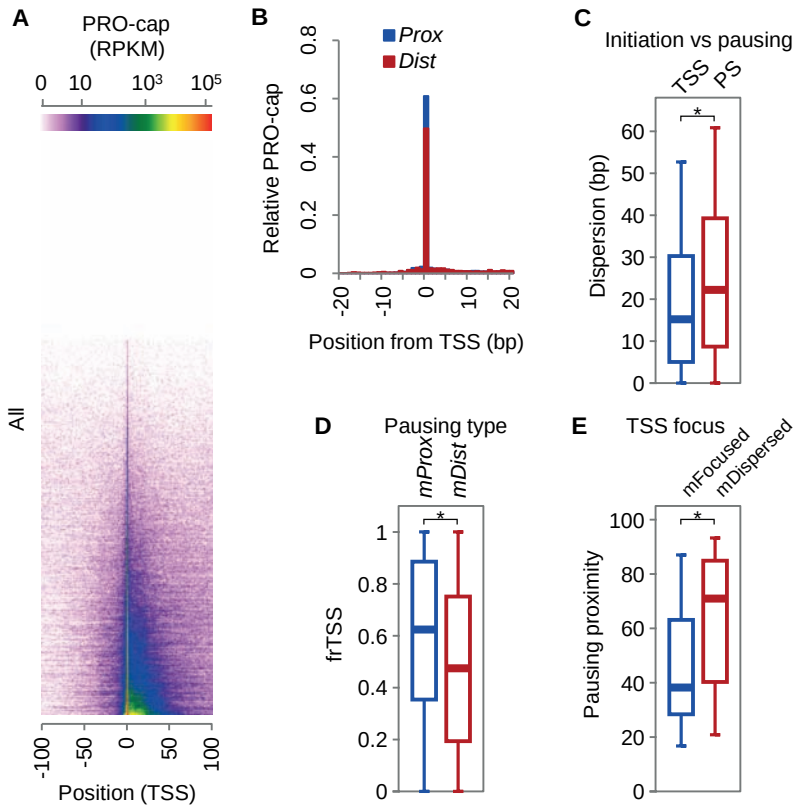


Figure S7. Pausing sites are associated with, but not fixed to, transcription initiation sites.
(**A**) Heatmap of 5'-PRO-seq profile for all annotated genes (n=16,746). The heatmap was generated as described in Fig. 1D, except for the position of the window (-100 to +100 bp from TSS). (**B**) Average profile of relative 5'-PRO-seq density around TSS in clustered proximal (*Prox*) and dispersed distal (*Dist*) pausing classes. To obtain the relative level in individual genes, PRO-cap read count for each gene was divided by the sum of PRO-cap reads from the window of -25 to +25 bp from TSS. (**C**) Dispersion of the pause peaks and initiation peaks in box plot presentation. The pause peaks from PRO-seq and the initiation peaks from PRO-cap were analyzed as in fig. S6A. To directly compare the extent of dispersion between initiation and pausing the peak dispersion is displayed in base pairs instead of the percentiles. (**D**) Analysis of TSS focusing in pair-matched proximal and distal pausing classes that have similar pausing level as in fig. S6E (n=600 each) was performed in order to isolate pausing level as a confounding variable. (**E**) Pausing proximity in pair-matched high and low TSS focusing classes that have similar pausing level (n=400 each). Asterisks indicate p-value < 0.001 in KS-test.

Figure S8

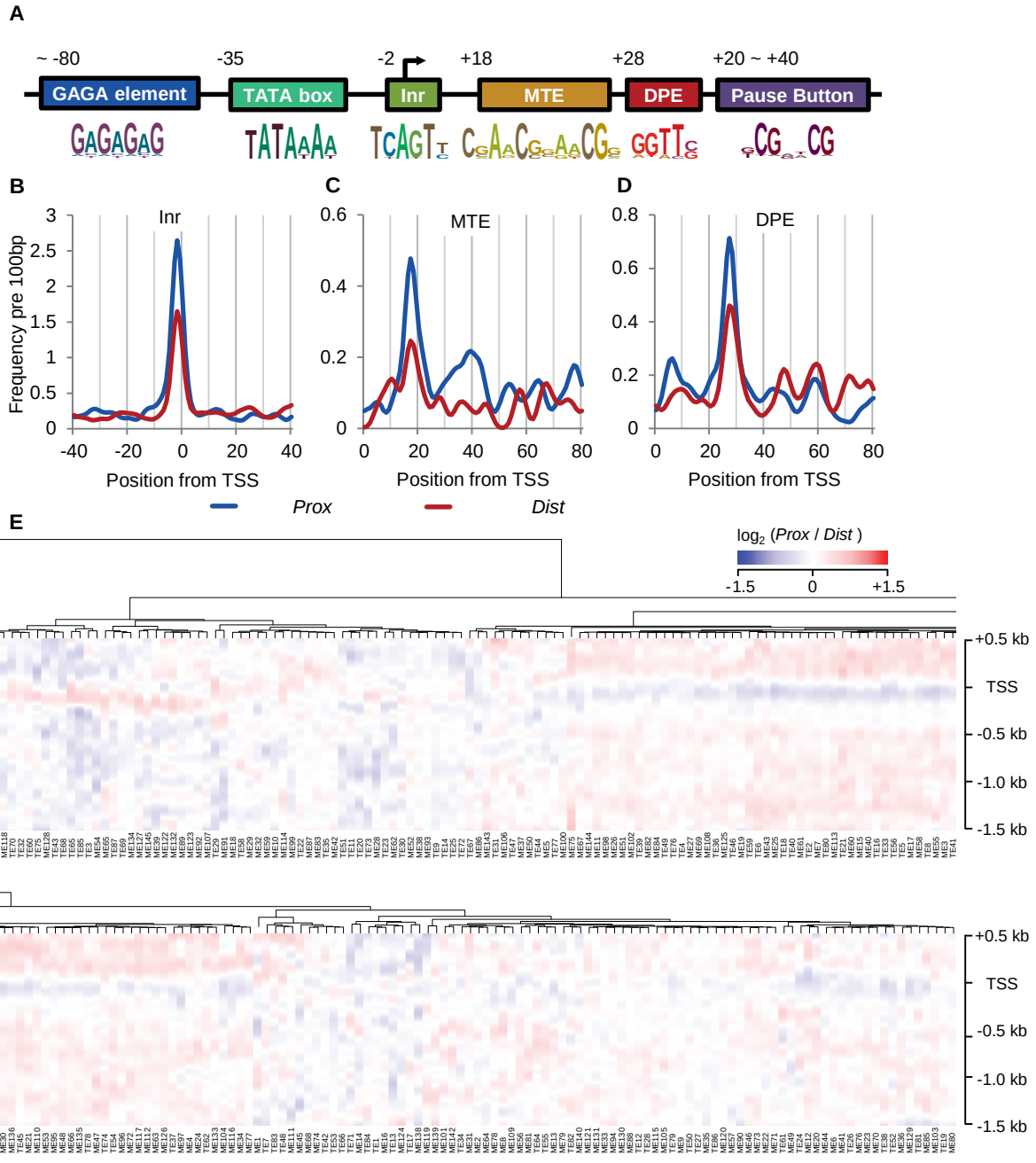


Figure S8. DNA elements in different pausing subsets

(**A**) Structure of promoter DNA elements*(24)*. GAGA element, TATA box, Initiator element (Inr), Motif Ten Element (MTE), Downstream Promoter Element (DPE), and Pause Button (PB)*(26)* are shown at the positions of their peak occurrences. Sequence logo*(20)* representation of each element is also shown. (**B**, **C**, **D**) Frequency of the core promoter elements in gene subsets of *Prox* or *Dist* pausing. Inr, MTE and DPE are shown respectively as described in Fig. 3B,C,D. (**E**) Heatmap of the difference in the occurrence of sequence elements between *Prox* and *Dist* groups. 232 regulatory DNA elements in two sub-panels from Stark *et al.(25)* were examined. The elements are ordered according to a hierarchical clustering. TE63 and ME137, both of which match the consensus for GAGA element, are the only elements showing noticeable enrichments.

Figure S9

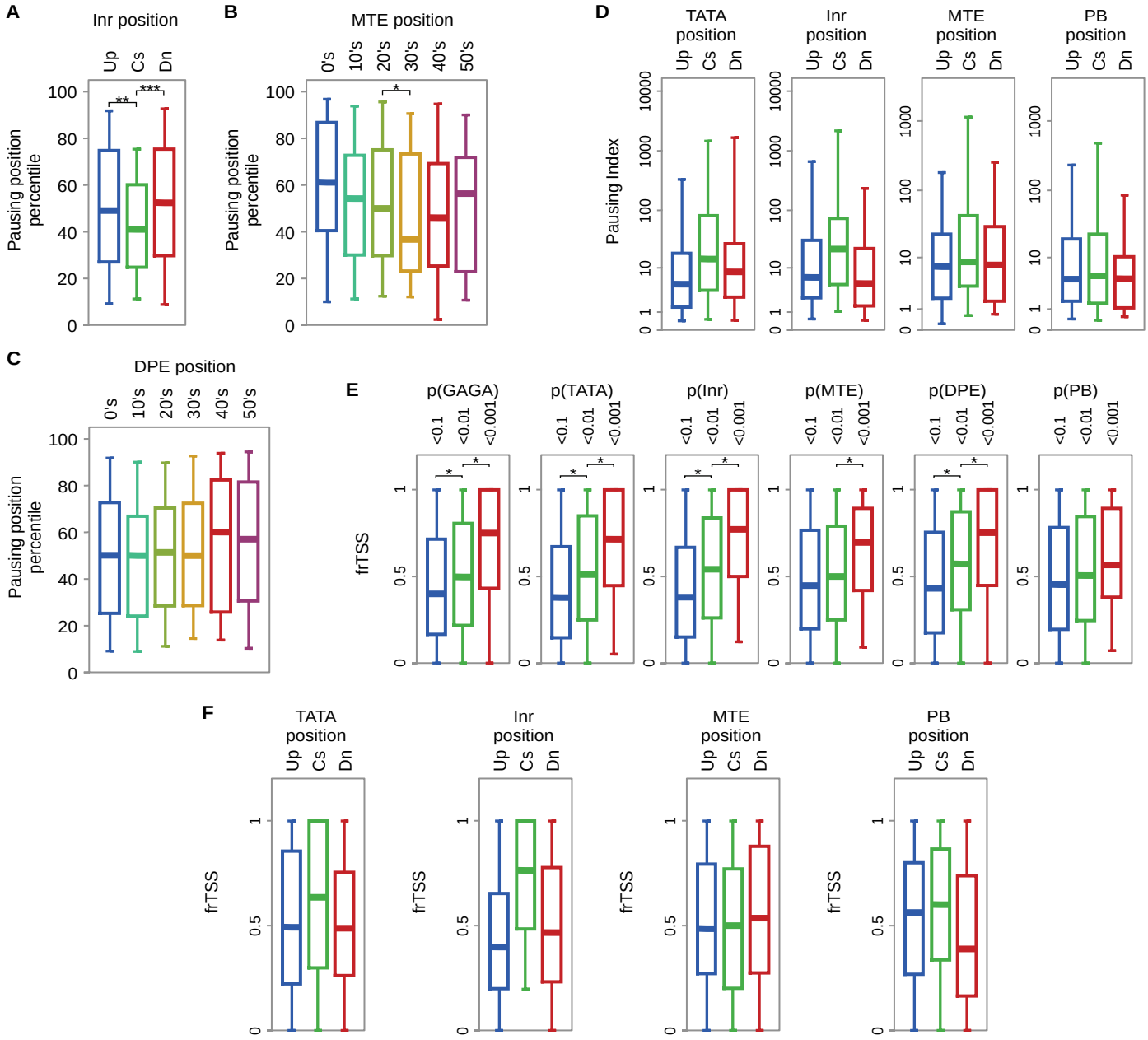


Figure S9. Association between the strength or the position of promoter DNA elements and Pol II pausing or initiation

(**A**) Pattern of positional association between Inr and pausing. Inr positions are -24 to -4 from TSS for upstream genes (Up, n=111), -4 to +1 from TSS for consensus positioned genes (Cs, n=131), +1 to +21 for downstream genes (Dn, n=110). (**B**) Pattern of positional association between MTE and pausing. Genes were divided based on the position of MTE in bins of 10 bp from TSS to +60 (n=56, 90, 70, 64, 52, 51). (**C**) Pattern of positional association between DPE and pausing. Genes were divided based on the position of DPE in bins of 10 bp from TSS to +60 (n=148, 102, 142, 105, 101, 97). Asterisks indicate p-values for the Kolmogorov-smirnov (KS) test (*** - p<0.0022, ** - p<0.021, * - p<0.066). (**D**) Pausing indices in genes with DNA elements at Up, Cs, and Dn positions. TATA and PB gene subsets are the same sets used in Fig. 3E; Inr subsets are the same sets used above (fig. S9B); for MTE subsets, subsets containing MTE at 20's, 30's, and 40's were used for Up, Cs, and Dn subsets respectively (fig. S9C). (**E**) Association of promoter DNA element strength at consensus positions with initiation focusing (frTSS). Gene sets are the same as in Fig. 3D (table S3). Asterisks indicate p-values for the KS test (p<0.001) (**F**) Initiation focusing (frTSS) in genes with DNA elements at Up, Cs, and Dn positions. Gene sets are the same as in panel D.

Figure S10

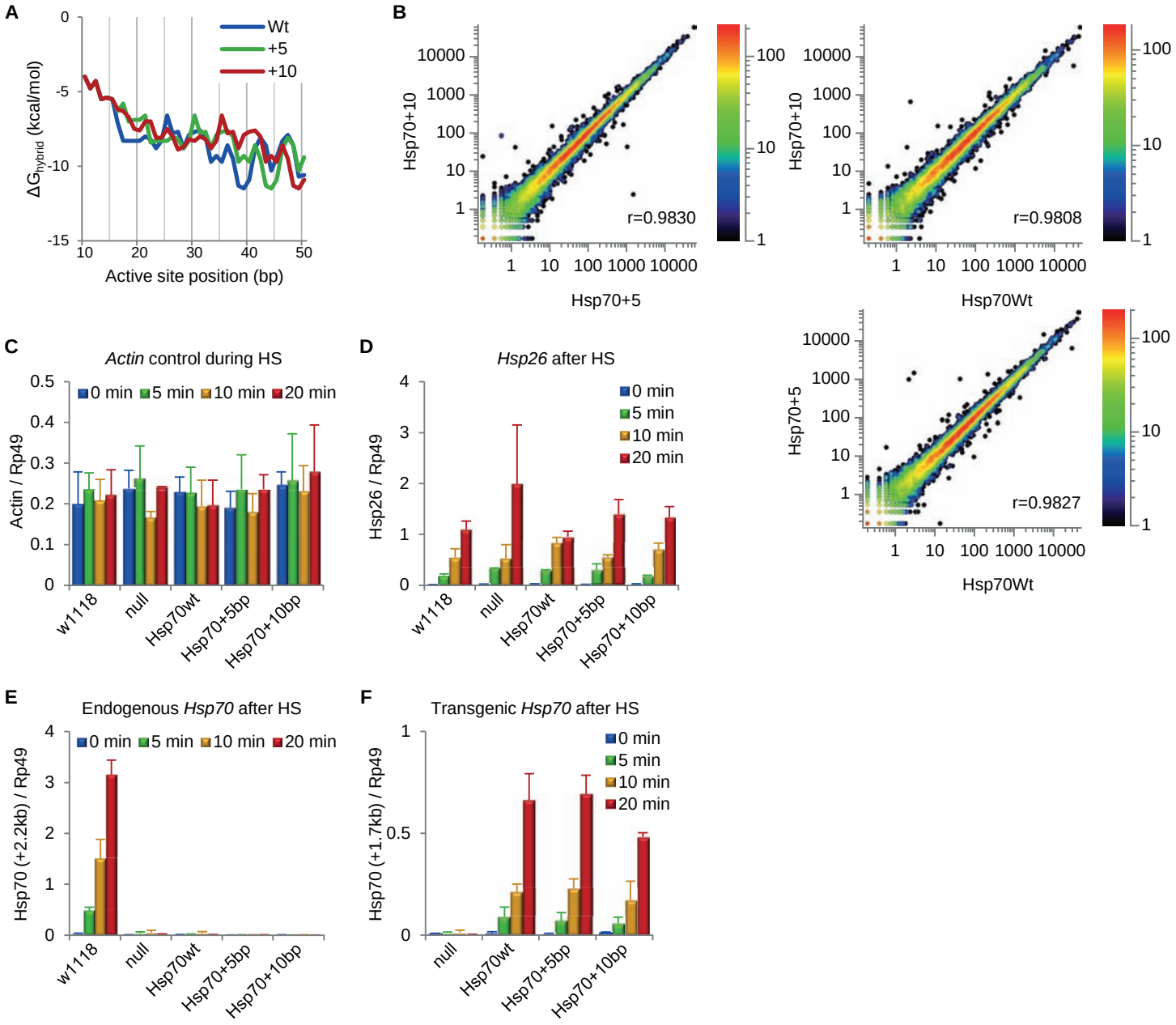


Figure S10. PRO-seq in animals with sequence modified *Hsp70* transgene promoters.
(A) Energy landscapes of the RNA-DNA hybrid*(8,29,30)* of the transcription bubble in *Hsp70* insertion mutants is shown as a function of position at the promoter. (**B**) Scatter-plot of PRO-seq densities in wild type (wt), 5 bp insert (+5), and 10 bp insert (+10) *Hsp70* transgenes. The transgenes contain -0.2 kb to + 1.8 kb of the *Hsp70* gene followed by polyadenylation signals, and were inserted into an *Hsp70* null background *Drosophila* line. Promoter-proximal (-50 to +150 from annotated TSS) PRO-seq densities were used, and the scatter-plots are otherwise the same as in fig. S2. (**C**) *Actin* mRNA control during heat-shock induction. All RT-qPCR values are normalized to *Rp49* mRNA level. (**D**) Induction of the endogenous *Hsp26* mRNA. (**E**) Induction of the endogenous *Hsp70* mRNA. RT-qPCR primers targeting +2.2 kb region of the endogenous *Hsp70* gene were used, which do not detect transgenic *Hsp70* copy. (**F**) Induction of the transgenic *Hsp70* mRNA. RT-qPCR primers targeting +1.7 kb region of the transgenic *Hsp70* gene were used.

Libraries	3’ end base				Substrate base percentage
	A	C	G	T	
PRO-seq$_{ATP}$	11,198,117	270,067	831,090	314,045	88.8%
PRO-seq$_{CTP}$	765,809	11,760,234	402,994	423,382	88.1%
PRO-seq$_{GTP}$	1,447,593	439,723	9,900,845	275,139	82.1%
PRO-seq$_{UTP}$	1,965,563	1,903,897	465,259	8,915,790	67.3%

Table S1. 3’ end uniformity in PRO-seq sequence reads.

For each of the 4 libraries, we counted the number of sequences reads ending in each of the 4 bases. Then we calculated the percentage of the reads ending in the same base as the nuclear run-on substrate. High substrate base percentage indicates that the majority of 3’ ends of the sequences were at the active sited of engaged polymerases.

DNA element	Subset	Position start	Position end	Gene count (n)
TATA box	Up	-44	-35	69
	Cs	-34	-20	67
	Dn	-29	-20	54
Inr	Up	-24	-5	111
	Cs	-4	+0	131
	Dn	+1	+21	110
PB	Up	+23	+32	94
	Cs	+33	+42	55
	Dn	+43	+52	45
MTE	0's	0	+9	56
	10's	+10	+19	90
	20's	+20	+29	70
	30's	+30	+39	64
	40's	+40	+49	52
	50's	+50	+59	51
DPE	0's	0	+9	148
	10's	+10	+19	102
	20's	+20	+29	142
	30's	+30	+39	105
	40's	+40	+49	101
	50's	+50	+59	97

Table S2. Gene subsets according to the position of a DNA element in Fig. 3D and fig. S9.

DNA element	Gene count (n)		
	0.1 > p > 0.01	0.01 > p > 0.001	0.001 > p
GAGA element	3,366	742	197
TATA box	1,712	352	207
Inr	2,484	886	648
MTE	2,829	658	166
DPE	2,550	472	92
PB	2,290	357	85

Table S3. Gene subsets according to the DNA element strength in Fig 3E.

Base	A	C	G	U	4N
Mapped reads total	3,606,748	7,011,991	4,932,695	5,348,348	4,889,866
Reads mapped to GB	2,821,812	5,601,559	3,624,817	3,822,842	3,505,727
Base counts in GB	29,421,288	22,348,555	21,783,485	28,572,783	102,126,111
Normalization factor	1.686	0.645	0.972	1.208	4.710

Table S4. Normalization of different PRO-seq libraries

Read counts dependent on the base composition are used to normalize each library as described in the supplementary information. GB; Gene Body.

	All genes	Constitutive genes (expected count)	
Active	5,471	3,557	
Paused	3,225	2,022 (2,097)	
Prox	848	492 (532)	]*
Dist	846	583 (530)	

Table S5. Enrichment of *Dist* pausing group among constitutive genes

A list of constitutive genes from a developmental transcriptome study*(31)* was examined for the number of genes in each *Prox* and *Dist* pausing groups. As defined in the original study, genes with short poly(A)+ RNA-seq levels greater than 1.0 FPKM in all 30 developmental conditions were considered constitutively active*(31)*. In addition, genes that have significant PRO-seq reads from upstream genes were removed and only the 'upstream clear' genes were selected as the 'active' genes. Expected counts of the constitutive genes are derived from the proportions of each group in all genes. Asterisk (*) indicate p-value < 0.0043 by χ^2 test.

Supplementary References:

S1. L. J. Core, J. J. Waterfall, J. T. Lis, "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters," *Science* **322**, 1845 (2008).

S2. E. Larschan *et al.*, "X chromosome dosage compensation via enhanced transcriptional elongation in *Drosophila*," *Nature* **471**, 115 (2011).

S3. L. J. Core *et al.*, "Defining the Status of RNA Polymerase at Promoters," *Cell Rep.* **2**, 1025 (2012).

S4. E. B. Rasmussen, J. T. Lis, "Short transcripts of the ternary complex provide insight into RNA polymerase II elongational pausing," *J. Mol. Biol.* **252**, 522 (1995).

S5. M. D. Abramoff, P. J. Magalhaes, S. J. Ram, "Image Processing with ImageJ," *Biophotonics International* **11**, 36 (2004).

S6. http://hannonlab.cshl.edu/fastx_toolkit/index.html

S7. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.* **10**, R25 (2009).

S8. S. Nechaev *et al.*, "Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*," *Science* **327**, 335 (2010).

S9. D. Author, S. Vassilvitskii. “k-means++: the advantage of careful seeding”, paper presented at the eighteenth annual ACM-SIAM symposium on Discrete algorithms (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007).

S10. P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math* **20**, 53 (1987).

S11. C. Fraley, A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Am. Stat. Assoc.* **97**, 611 (2002).

S12. D. A. Gilchrist *et al.*, "Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation," *Cell* **143**, 540 (2010).

S13. S. Maheshwari, D. A. Barbash, "Cis-by-Trans regulatory divergence causes the asymmetric lethal effects of an ancestral hybrid incompatibility gene," *PLoS Genet.* **8**, e1002597 (2012).

S14. W. J. Gong, K. G. Golic, "Genomic deletions of the *Drosophila melanogaster Hsp70* genes," *Genetics* **168**, 1467 (2004).

S15. H. Lee, K. W. Kraus, M. F. Wolfner, J. T. Lis, "DNA sequence requirements for generating paused polymerase at the start of *hsp70*," *Genes Dev.* **6**, 284 (1992).

S16. C. Lee *et al.*, "NELF and GAGA factor are linked to promoter-proximal pausing at many genes in *Drosophila*," *Mol. Cell. Biol.* **28**, 3290 (2008).

S17. A. N. Brooks *et al.*, "Conservation of an RNA regulatory map between *Drosophila* and mammals," *Genome Res.* **21**, 193 (2011).

S18. K. D. Pruitt et al., “NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy,” *Nucleic Acids Res.* **40**, D130 (2012)

S19. S. M. Mount *et al.*, "Splicing signals in *Drosophila*: intron size, information content, and consensus sequences," *Nucleic Acids Res.* **20**, 4255 (1992).

S20. T. D. Schneider, R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Res.* **18**, 6097 (1990).

S21. S. Schwartz, R. Oren, G. Ast, "Detection and removal of biases in the analysis of next-generation sequencing reads," *PLoS One* **6**, e16685 (2011).

S22. L. S. Churchman, J. S. Weissman, "Nascent transcript sequencing visualizes transcription at nucleotide resolution," *Nature* **469**, 368 (2011).

S23. M.A.Hall *et al.*, "High-resolution dynamic mapping of histone-DNA interactions in a nucleosome," *Nat. Struct. Mol. Biol.* **16**, 124 (2009).

S24. U. Ohler, G. C. Liao, H. Niemann, G. M. Rubin, "Computational analysis of core promoters in the *Drosophila* genome," *Genome Biol.* **3**, RESEARCH0087 (2002).

S25. A. Stark *et al.*, "Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures," *Nature* **450**, 219 (2007).

S26. D. A. Hendrix, J. W. Hong, J. Zeitlinger, D. S. Rokhsar, M. S. Levine, "Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo," *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7762 (2008).

S27. S. K. Ghosh, A. Missra, D. S. Gilmour, "Negative elongation factor accelerates the rate at which heat shock genes are shut off by facilitating dissociation of heat shock factor," *Mol. Cell. Biol.* **31**, 4232 (2011).

S28. C. H. Wu *et al.*, "Analysis of core promoter sequences located downstream from the TATA element in the hsp70 promoter from *Drosophila melanogaster*," *Mol. Cell. Biol.* **21**, 1593 (2001).

S29. N. Sugimoto *et al.*, "Thernodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes," *Biochemistry* **34**, 11211 (1995).

S30. J. SantaLucia *et al.*, "Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability," *Biochemistry* **35**, 3555 (1996).

S31. B. R. Graveley *et al.*, "The Developmental Transcriptome of *Drosophila melanogaster*," *Nature* **471**, 473 (2011)