Published online 29 June 2020

Nucleic Acids Research, 2020, Vol. 48, No. 14 ***7767–7785***
doi: 10.1093/nar/gkaa531

A unified view of the sequence and functional organization of the human RNA polymerase II promoter

Donal S. Luse[1,†], Mrutyunjaya Parida[2,†], Benjamin M. Spector[2], Kyle A. Nilson[2] and David H. Price[2,*]

[1]Department of Cardiovascular and Metabolic Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA and [2]Department of Biochemistry, The University of Iowa, Iowa City, IA 52242, USA

Received March 12, 2020; Revised May 31, 2020; Editorial Decision June 07, 2020; Accepted June 24, 2020

ABSTRACT

To better understand human RNA polymerase II (Pol II) promoters in the context of promoter-proximal pausing and local chromatin organization, 5′ and 3′ ends of nascent capped transcripts and the locations of nearby nucleosomes were accurately identified through sequencing at exceptional depth. High-quality visualization tools revealed a preferred sequence that defines over 177 000 core promoters with strengths varying by >10 000-fold. This sequence signature encompasses and better defines the binding site for TFIID and is surprisingly invariant over a wide range of promoter strength. We identified a sequence motif associated with promoter-proximal pausing and demonstrated that cap methylation only begins once transcripts are about 30 nt long. Mapping also revealed a ∼150 bp periodic downstream sequence element (PDE) following the typical pause location, strongly suggestive of a +1 nucleosome positioning element. A nuclear run-off assay utilizing the unique properties of the DNA fragmentation factor (DFF) coupled with sequencing of DFF protected fragments demonstrated that a +1 nucleosome is present downstream of paused Pol II. Our data more clearly define the human Pol II promoter: a TFIID binding site with built-in downstream information directing ubiquitous promoter-proximal pausing and downstream nucleosome location.

INTRODUCTION

In spite of decades of research, the metazoan Pol II promoter remains poorly understood. Biochemical and mechanistic studies of Pol II transcription have achieved major advances using promoters with canonical TATA elements that support single transcription start sites (TSSs) (1–3). This work has led to the identification of a set of general transcription factors required for initiation by Pol II, culminating in the determination of structures of the complete Pol II preinitiation and elongation complexes (4–9). However, it is now appreciated that the very large majority of Pol II promoters do not have TATA elements. While Pol II TSSs are often associated with a minimal initiator (Inr) and some TATA-less promoters depend on specific downstream elements (DSEs), many promoters lack clearly defined sequence motifs (10). More importantly, most Pol II promoters do not support single or tightly grouped TSSs. Instead, TSSs are often scattered within regions which can span a hundred bp or more. Such TSS groups are often referred to as dispersed or diffuse promoters (10–17). Critically, it is not known whether such regions consist of groups of promoters analogous to the well-characterized examples with tightly grouped TSSs or instead represent a mechanistically distinct promoter class. Indeed, in light of these observations the definition of a Pol II promoter is not entirely clear.

This fundamental uncertainty about Pol II promoter structure is linked to unanswered questions concerning the earliest stages of transcript elongation. Pausing by Pol II after synthesizing ∼20–60 nt of RNA occurs at essentially all metazoan promoters (18–20), but the relationship of pause sites to template sequence, promoter strength, and promoter class is not understood. It has been suggested that the immediately downstream +1 nucleosome seen in bulk chromatin drives pausing (21–24), but other roles for that nucleosome have been proposed including facilitating PIC assembly (25–27). Alternatively, the +1 nucleosome might be positioned by the paused polymerase (16,28,29). Proposals for the function of TSS-proximal metazoan nucleo-

*To whom correspondence should be addressed. Tel: +1 319 335 7910; Email: david-price@uiowa.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present address: Kyle A. Nilson, Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA.

somes often rely on a stereotypical spacing relative to the TSS, with a nucleosome-depleted region immediately upstream and the +1 nucleosome located with its proximal edge ∼50 nt downstream (30–32). However, such models are very difficult to reconcile with the reality of widely dispersed TSSs. They are also limited by the possibility that promoter-proximal nucleosomes are differentially sensitive to micrococcal nuclease typically used for nucleosome mapping, relative to bulk nucleosomes (24,33–36). In fact, it has not been rigorously demonstrated that there is a +1 nucleosome downstream on the small fraction of templates that are occupied by paused Pol II.

Further advances in unifying current results on transcription complex assembly, pausing and promoter-proximal chromatin structure depend on obtaining a much deeper and more precise understanding of the Pol II promoter. To address this problem, we have essentially relied on Pol II itself for information. We have developed a method for generating promoter-proximal nascent RNAs in nuclei in which both TSSs and pause sites are determined with accurate, base pair precision for hundreds of thousands of promoters, supporting levels of RNA synthesis varying by up to 10 000-fold. We began our study with the simple goal of examining cap methylation during transcription, but we quickly realized that the datasets generated had so much information on Pol II promoters that they demanded further analysis. Most of our efforts then turned to developing tools to visualize the complex data and this resulted in new hypotheses that required additional experimental approaches. Our collective findings lead to a strikingly simple and profound conclusion: Pol II promoters are TFIID binding sites that have downstream sequence information directing both the ubiquitous promoter proximal pause and the proximal nucleosome location.

MATERIALS AND METHODS

Generation of NasCap datasets

Initial steps. Adherent HeLa cells were grown in T150 flasks at 37°C and 5% CO_2 in 30 ml DMEM (Gibco 11965-092) supplemented with 5% fetal bovine serum (FBS, Gibco 26140-079). 1 h prior to harvest, when cells were at 80–90% confluence, 10 ml of media was removed from each flask and either 30 μl of DMSO or 1 mM flavopiridol in DMSO was added before the mixture was returned to the flask (final concentrations 0.1% DMSO or 1 μM flavopiridol and 0.1% DMSO). All rapid nuclei isolation steps were performed on wet ice using ice-cold buffers as previously described (37,38). The nuclear run-on (NRO) with biotinylated NTPs, streptavidin M-280 selection, and 3′ adapter (containing a 4 nt redundant unique molecular identifier) ligation and clean up steps were performed as previously described (39).

Removal of uncapped transcripts. The two RNA samples from control and flavopiridol treated cells were individually treated sequentially with RNA 5′ polyphosphatase, Terminator 5′-phosphate dependent exonuclease, and shrimp alkaline phosphatase as previously described for the PRO-Cap procedure (39).

Preparation of anti-m7G beads. To overcome problems of RNase contamination of commercial preparations of cap binding antibody beads, we started with Ascites fluid containing anti-2,2,7-trimethylguanosine antibodies (Calbiochem CS214155). The preparation was diluted 1:10 in 35 mM KCl HGKEDP (25 mM HEPES pH 7.6, 15% glycerol, 35 mM KCl, 0.1 mM EDTA, 1 mM DTT and 0.1% isopropanol-saturated PMSF) containing 0.5% Triton X-100 and fractionated with an 80 ml gradient (50 mM to 1 M KCl HGKEDP) over a Mono Q HR 10/10 column. Fractions eluting around 100 mM KCl containing the pure IgG were pooled. 37.5 μg antibody was bound to 37.5 μl of protein G Sepharose beads (Sigma P3296) after washing the beads sequentially with 200 μl each of RIPA buffer (25 mM Tris, pH 7.8,150 mM sodium chloride,1% Triton X-100, 0.1% SDS,1 mM EDTA, 0.01 U/μl SUPERase-In and 0.1% PMSF (saturated in isopropanol and added fresh), then binding buffer (25 mM Tris, pH 7.8, 50 mM sodium chloride,1 mM EDTA, 0.02% Tween20, 0.01 U/μl SUPERase-In and 0.1% PMSF), binding buffer with 0.2 mg/ml BSA, RIPA buffer and finally binding buffer. Incubation with the protein and beads was in 300 ul binding buffer for 1 h at 4°C. Antibody beads were washed with RIPA buffer and then binding buffer and stored at 4°C.

Separation of m7G capped RNAs from non-methylated capped RNAs. 1 μl of SUPERase-In was added to 20 μl of each sample before they were incubated with 15 μl of the anti-m7G beads. After an hour of rotation at 4°C beads were settled and the supernatant was removed (non-methylated RNAs). Beads (m7G capped RNA) were washed twice with 200 μl of RIPA buffer and once with 200 μl of binding buffer. RNA was isolated using Trizol from the supernatant and the beads. After precipitation with added glycogen and washing of the pellet with 70% ethanol, the RNA was dried and resuspended in water. RNA decapping and 5′ adaptor (containing a second 4 nt redundant unique molecular identifier) ligation was performed on the resulting RNA before a third streptavidin purification (39).

Library amplification and size selection. Test amplifications were performed to determine the number of cycles needed to obtain enough library material and finally a full-scale amplification using barcoded primers was performed. Libraries were analyzed by gel electrophoresis, Qubit and finally using the Agilent Bioanalyzer. Libraries were pooled and size selected 135–600 bp using a BluePippin and reanalyzed before being diluted for sequencing on an Illumina HiSeq 4000.

NasCap analysis

Raw sequences were first trimmed using trim_galore 0.4.4 (https://github.com/FelixKrueger/TrimGalore) and then aligned with the UCSC hg38 assembly using bowtie v1.2.2 (40). Additionally, bowtie was used to trim the 4 bp Unique Molecular Identifiers (UMI) from both ends of each read prior to alignment with a minimum insert size of 26 bp. Aligned samples were deduplicated using the dedup program to collapse identical mapped reads with redundant UMIs and remove the biotinylated NTP from the 3′ end

Table 1. Read statistics and average transcript length

Datasets	Total reads	Max to Max	Select TSSs	Average paused transcript length (nt)			
				Max to Max	Select TSSs	Top 10%	Top 1%
m7G capped control	39,093,914	177,098	522,186	38.6	41.2	41.7	42.1
m7G capped flavo	15,113,070	95,557	240,896	39.5	41.8	42.2	42.6
capped w/o m7 control	3,024,848	16,241	26,892	33.6	32.1	31.5	29.3
capped w/o m7 flavo	1,433,860	9,997	15,480	35.5	35.4	35.1	33.1

Total paired-end read counts for the four datasets and for the MaxTSS to MaxTPS (Max to Max) and Selected TSS subsets are indicated. In addition, the average lengths of transcripts associated with paused Pol II from the Max to Max and Selected TSS subsets as well as for the Top 10% and 1% of the Selected TSS subset are shown.

(https://github.com/P-TEFb/dedup). 5′ and 3′ site tracks for each strand and sample were first generated in bedGraph format using bedtools v2.26 (41), and then converted into bigwig format using the Kent UCSC utility program called bedGraphToBigWig (42), for display on the UCSC Genome Browser.

Nuclear run-off reactions

HeLa nuclei treatments, isolations, and nuclear walk-on reactions were done as previously described with slight modification to the walk-on procedure (37). In short, 1–3 × 10^5 nuclei were incubated in 20 mM HEPES pH 7.6, 5 mM $Mg(Ac)_2$, 5 mM DTT, 100 mM K(Ac), and 0.25 U/μl SUPERase-In with and without both 1.33 μg/ml α-amanitin (Sigma A2263) and approximately 2.5 μg of purified DFF40 at 37°C for 10 min in 24 μl. Immediately after digestion, the solution was raised to 30 μl of 0.5% Sarkosyl, 150 mM K(Ac), and 0.0833 μM α-^{32}P-CTP to allow for radio-label incorporation for 3 min. Reactions were chased with 500 μM of cold ATP, UTP, GTP, and CTP for 10 minutes. Due to the dramatic increase in viscosity due to the release of DNA from chromatin with Sarkosyl, chasing was performed by tripling the volume to 90 μl and down-up pipetting with a cut pipette tip. Elongation was stopped with stop solution containing 20 mM EDTA, 0.1M Tris, 1% Sarkosyl and 200 μg/ml Torula Yeast RNA to a final volume of 120 μl. Transcripts were isolated by Trizol LS (Ambion 10296028), precipitated by the addition of three volumes of 95% ethanol and 500 mM $NH_4(Ac)$, washed with 70% ethanol, and analyzed using 6% urea–PAGE. Total RNA was visualized by ethidium bromide staining and radiolabeled RNAs were visualized with a Fujifilm Typhoon FLA-7000 phosphorimager.

MaxTSS to MaxTPS, selected TSS and 200 bp TSR datasets

Three different methods were used to select TSSs from the m7G NasCap dataset generated from control cells. The first method involved identification of TSRs using tsrFinderM1 that identifies clusters of TSSs from 5′ read densities with user-defined parameters (https://github.com/P-TEFb/tsrFinderM1). Here, TSRs were identified using a 20 bp TSR with at least 20 reads that were 600 bp or less with an average read length of at least 30 bp. TSRs overlapping within 50 bp from the start and end of a genomic interval involving microRNA (miRNA), ribosomal RNA (rRNA), small cytoplasmic RNA (scRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), small Cajal body-specific RNA (scaRNA), and transfer RNA (collectively called the small RNA blacklist) were removed from further analysis (39). Additionally, chromosomes 1 through 22, X, and Y were retained for this analysis. In each TSR, the TSS containing the most mapped reads was defined as the MaxTSS and with rare ties being resolved by random selection of one of the two. For each MaxTSS a MaxTPS was assigned based on the number of 3′ reads (minimum of 2) with a maximum length of 100 bp. A total of 177,098 MaxTSS to MaxTPS intervals were found. The Selected TSS dataset was created from all TSSs (over 6 million total) present that have a read density of at least 10 reads and a maximum length of 100 bp. A total of 522 186 Selected TSSs were found. Refer to Table 1 for read statistics. The third method involved identification of TSRs using tsrFinderM2 that detects non-overlapping TSRs centered on MaxTSSs with user-defined parameters (https://github.com/P-TEFb/tsrFinderM2). Here TSRs were identified using a 200 bp TSR centered on the MaxTSS with at least 10 reads that were 600 bp or less. TSRs overlapping within 50 bp from the start and end of a genomic interval of the small RNA blacklist above were removed from further analysis (39). SINE and LINE element intervals were obtained from the UCSC hg38 repeat masker databases and used to remove overlapping TSRs. TSRs associated with chromosomes 1 through 22, X, and Y were retained for further analysis. For each MaxTSS a MaxTPS was assigned based on the number of 3′ reads (minimum of 2) with a maximum length of 100 bp. A total of 62 381 intervals were found. The standard deviation of the positions of TSSs within each TSR was calculated.

MaxTPS dataset

TPRs were identified in the four NasCap datasets with tprFinder that identifies clusters of TPSs from 3′ read densities with user-defined parameters. (https://github.com/P-TEFb/tprFinder). Here, TPRs were identified using a 40 bp TPR with at least 20 reads that were 600 bp or less with an average read length of at least 30 bp. TPRs overlapping within 50 bp from the start and end of a genomic interval of the small RNA blacklist were removed from further analysis (39). Additionally, chromosomes 1 through 22, X, and Y were retained for this analysis. TPSs with most mapped reads in each TPR were defined as MaxTPS. The total number of MaxTPSs are indicated on the logos.

Purification of DFF

The human DFF40 and DFF45 protein coding sequences were inserted into a pET-21a vector bi-cistronically for co-expression with a 6x His-tag at the C-terminus of DFF40 and TEV sites (ENLYFQS) inserted immediately downstream of the two Caspase-3 (117–118, 224–225) digestion sites within DFF45 as done previously (43). This vector was then transformed into *Escherichia coli* BL21 Star (DE3) cells and four individual colonies were picked and grown in one liter preps. Preps were then induced with 0.1 mM IPTG overnight at 18°C once the OD_{600} reached 0.6. Cells were pelleted, resuspended with PBS and combined, pelleted again, and then resuspended in buffer containing 1× PBS, 1% Triton X-100, 5 mM Imidazole, and 0.1% PMSF. Lysates were sonicated and NaCl was subsequently increased to 150 mM prior to a high speed spin for 45 minutes at 244 000 × g at 4°C. The resulting supernatant was applied to 2 ml Ni-NTA agarose beads for 1 h with rotation at 4°C and washed with high salt washes (10 mM Tris 7.8, 150 mM NaCl, 35 mM Imidazole, 1% PMSF) and low salt washes (10 mM HEPES 7.8, 50 mM KCl, 35 mM Imidazole, 1% PMSF). Bound proteins were eluted with 10 mM HEPES 7.8, 50 mm KCl, 300 mM Imidazole and 1% PMSF. The elution was then treated with 500 μg of tobacco etch virus protease (TEV) in a solution that contained 50 mM KCl, 5% glycerol, 1 mM DTT for 1 h at 30°C. TEV protease was obtained from the Protein and Crystallography Facility at the University of Iowa. The resulting product was then spun for 15 min at 22 500 × g and the supernatant was FPLC purified on a Mono S 5/50 GL column. DFF40 eluted at approximately 470 mM KCl, similar to a previous purification of caspase-3 activated DFF40 (44).

DFF-Seq

Approximately 600 000 nuclei from HeLa cells treated with 1 μM flavopiridol for 1 h were digested with 5 μg of DFF for 30 min to generate primarily (71%) mononucleosomes. Reactions were stopped with the addition of EDTA to 50 mM and subsequently treated with RNase A (0.1 mg/ml) for 1 h at 37°C in the presence of 20 mM HEPES (7.6) 100 mM potassium acetate, 1 mM DTT and then proteinase K (0.25 mg/ml) for 1 h at 50°C after addition of SDS to 0.68%. The digested DNA was then isolated by phenol extraction and precipitated by addition of three volumes of 95% ethanol containing 0.5 M ammonium acetate. The resulting pellet was washed with 70% ethanol and resuspended in 10 mM Tris pH 8. Libraries were prepared for sequencing from the purified DNA fragments using the KAPA Hyper Prep Kit (Roche 7962312001) without PCR amplification or size selection. 331 290 346 151 bp paired end reads were obtained from an Illumina HiSeq 2500 by the Iowa Institute of Human Genetics. Raw sequences were trimmed using trim_galore 0.4.4 (https://github.com/FelixKrueger/TrimGalore) and aligned to the human hg38 assembly using bowtie 1.2.2 (40) resulting in 259 878 094 mapped reads. The alignment file (.bam) was filtered for specific fragment lengths and used to generate heatmaps and metaplots as specified in the text. Bigwig tracks for the unfiltered and filtered data were generated using bedtools genomeCoverageBed, sortBed programs, and Kent UCSC utilities including bedGraphToBigWig (42). DFF-Seq tracks generated in this study can be viewed on the UCSC genome browser by following simple instructions provided in Supplementary Data.

Heatmaps

High precision heatmaps were created by controlling the aspect ratio, the number of pixels, and intensities assigned. The aspect ratio was maintained around 1:2 (width:height). The number of pixels chosen to display genomic intervals gave a discrete number of pixels for each base or bases for each pixel (exactly 3 pixels for each base in the sequence heatmaps or 2 bp for each pixel for DFF heatmaps). The number of pixels in the height were chosen based on the desired aspect ratio. Data were first vertically randomized before being sorted based on the strength of MaxTSSs (read counts) or the length of the transcripts using a python script. The large number of genomic intervals were grouped and each group averaged such that the number of groups matched the height in pixels of the heatmap. The choice of pixel number was made based on a trade-off between file size and resolution with file sizes varying between ∼0.2 and 1 megapixel. The values at each horizontal position of the average intervals were used to assign intensities using the gray.colors function in R. A linear relationship between average read value and intensity was utilized. Values of 0 or 1 were used to indicate the absence or presence of a specific nucleotide in the sequence heatmaps that were centered on MaxTSS or MaxTPS and read values were used in the TSS and DFF-Seq heatmaps that were centered on MaxTSS. Genomic intervals were grouped and then averaged. The number of groups matched the height of each heatmap. Black was set at a read value of 2 for 18–120 bp DFF-Seq heatmaps, 15 for TSS heatmaps, and 20 for 140–185 bp DFF-Seq heatmaps that were ±1000 bp. Additionally, a subset of 1896 MaxTSSs containing the first T of TATA between −34 to −29 bp upstream were used to draw heatmaps, with a fixed value of 20.0 reads as black. To enhance perception of dark and light patterns on heatmaps we applied a gamma correction of 0.6 on all heatmaps. Colors for each sequence heatmap was scaled between 0 (white) and 1 (black), prior to gamma correction by replacing two read values at the bottom right with 0.0 and 1.0.

truView plots

Plots of transcript length versus the genomic position of TSSs were generated to provide an informative visualization of the distribution of transcript lengths from each TSS. For each chosen region a table was generated that contained the number of transcripts of each length from each TSS. The transcript frequency information was converted into a linear heatmap with the maximum transcript frequency being black. To increase the ability to see less frequently used TSSs with lower transcript frequencies the intensity scale was saturated at 30% or 90% of the maximum value with the remaining values displayed linearly.

Metaplots

Metaplots displaying average TSS or DFF-Seq read densities at each genomic position across the specific intervals centered on TSSs from different groups of TSSs were generated and plotted in MS Excel. Intervals, groups, and subsets of TSSs used to generate the metaplots are indicated in the figure legends. Additionally, average base distributions were plotted displaying the fraction of each nucleotide (or sequence motif) within a chosen region centered on TSSs or TPSs derived from the indicated datasets.

Logos

Weblogo v3.6.0 24826 (45) was used to generate sequence logos using the following color code: A was gray (hex code #bbbbbb), T was red (hex code #ff0000), G was yellow (hex code #dddd00), and C was blue (hex code #2222ff). The program computed nucleotide composition probabilities at each genomic position, from input fasta files. Bedtools v1.2.2 getfasta program was used to generate fasta files while maintaining a positive strand orientation for ±100 bp genomic intervals centered on selected TSSs, selected TPSs, MaxTSSs, MaxTPSs, and unique 3′ ends of regions of productive elongation (ROPE). Additionally, the top 0.1%, 1% and 10% of each dataset, based on their TSS strength sorted from high to low were used to generate sequence logos. Moreover, sequence logos were made for ±100 bp sequences centering on 77 MaxTSSs associated with RPGs.

Transcript length frequency distribution plots

Transcript length frequencies were obtained by counting the number of mapped fragments of each length between 17 and 100 from the selected list of TSSs (MaxTSS to MaxTPS, Selected TSS or selected subsets) as specified in the figures.

RESULTS

Where do the transcription start and pause sites map within the genome?

Our study began with the generation of PRO-Seq datasets from control HeLa cells or from HeLa cells that were treated with flavopiridol for 1 h to block P-TEFb-dependent productive elongation. We incorporated a modified front end in which cells were lysed and nuclei bathed in EDTA within 20 s of removal from the 37°C incubator so sites of pausing in cells could be accurately determined (37). Libraries of nascent transcripts were prepared using an approach we call NasCap (Figure 1A). Biotinylated RNAs initially isolated were selected for those containing m7G caps using an antibody. Unbound RNAs were subjected to enzymatic selection for capped transcripts resulting in a second population that was capped, but not methylated. 150 bp paired-end sequencing identified both the 5′ and 3′ ends of nascent transcripts which correspond to the TSS and transcription pause site (TPS), respectively, for each RNA. Tracks generated in this study, including those showing accumulation of nascent transcripts as well as 5′ and 3′ end tracks for all four NasCap datasets, can be viewed on the UCSC genome browser by following simple instructions provided in Supplementary Data. Initial observations suggested that TSS distributions ranged from very highly focused to dispersed over broad regions, in some cases hundreds of base pairs. However, in the large majority of cases it was apparent from inspection that TSSs could be naturally grouped into clusters covering <20 bp. We generated a collection of transcription start regions (TSRs) from the control m7G capped dataset that were 20 bp wide, non-overlapping, and contained a total of 20 reads or more. There were 177 098 TSRs in this control NasCap m7G dataset.

To better visualize the patterns of initiation and pausing we developed a tool, truView, with which we examined 201 randomly selected genome regions spanning a wide range of RNA read levels. Four representative examples are shown in Figure 1B, with the positions of TSSs plotted versus the lengths of RNAs initiated at each position. Read levels are either displayed with a linear intensity (upper set) or with saturation increased to 90% (lower set) to bring out the weaker TSSs. The truView heatmaps provide a simple, intuitive way to examine TSS clustering (horizontal axis) and the range of TPSs associated with each TSS (vertical axis) at the same time. Each of the selected locations contained a TSR with read totals indicating high (CNNM4, ING1-AS) or lower (STARD7, ACBD3) transcription levels. While in some locations such as CNNM4 initiation occurs primarily at a single template location, a high level of RNA synthesis does not require a single TSS (for example, ING1-AS). One of the most striking observations made during examination of hundreds of TSRs was the appearance of a pattern with a −1 slope for the transcript lengths arising from closely spaced TSSs (Figure 1B). This indicates that there are highly preferred sites of pausing downstream of the TSRs, with more downstream TSSs supporting shorter transcripts (note in particular the no saturation panels). The distribution of these sites was different for each TSR, suggesting that pausing is not simply a function of transcript length, but rather depends, in part, on local sequence.

TSSs identified and the preferred pause positions downstream of each TSS are nearly identical between the m7G datasets obtaining from cells with or without flavopiridol treatment, demonstrating the reproducibility of the data (Figure 1B). However, we did notice for a substantial subset of genes that the relative distribution of TSSs utilized shifted upstream for the flavopiridol samples, suggesting a modest preference for P-TEFb-dependent, productive elongation to arise from slightly upstream promoters. Examples of this effect are shown in Supplementary Figure S1. To demonstrate wide applicability of the results, the HeLa NasCap data was compared to PRO-seq data obtained from contact inhibited, primary human foreskin fibroblasts (HFF) (39). TSS usage and pause sites found in HFFs are highly similar to those seen with HeLa cells (Figure 1B). The transcripts from HFF PRO-seq are shorter on average than those from HeLa NasCap and this will be addressed below. Significantly, for locations where there was an upstream shift of TSSs with flavopiridol treatment in HeLa cells, this effect was also seen with HFFs (Supplementary Figure S1).

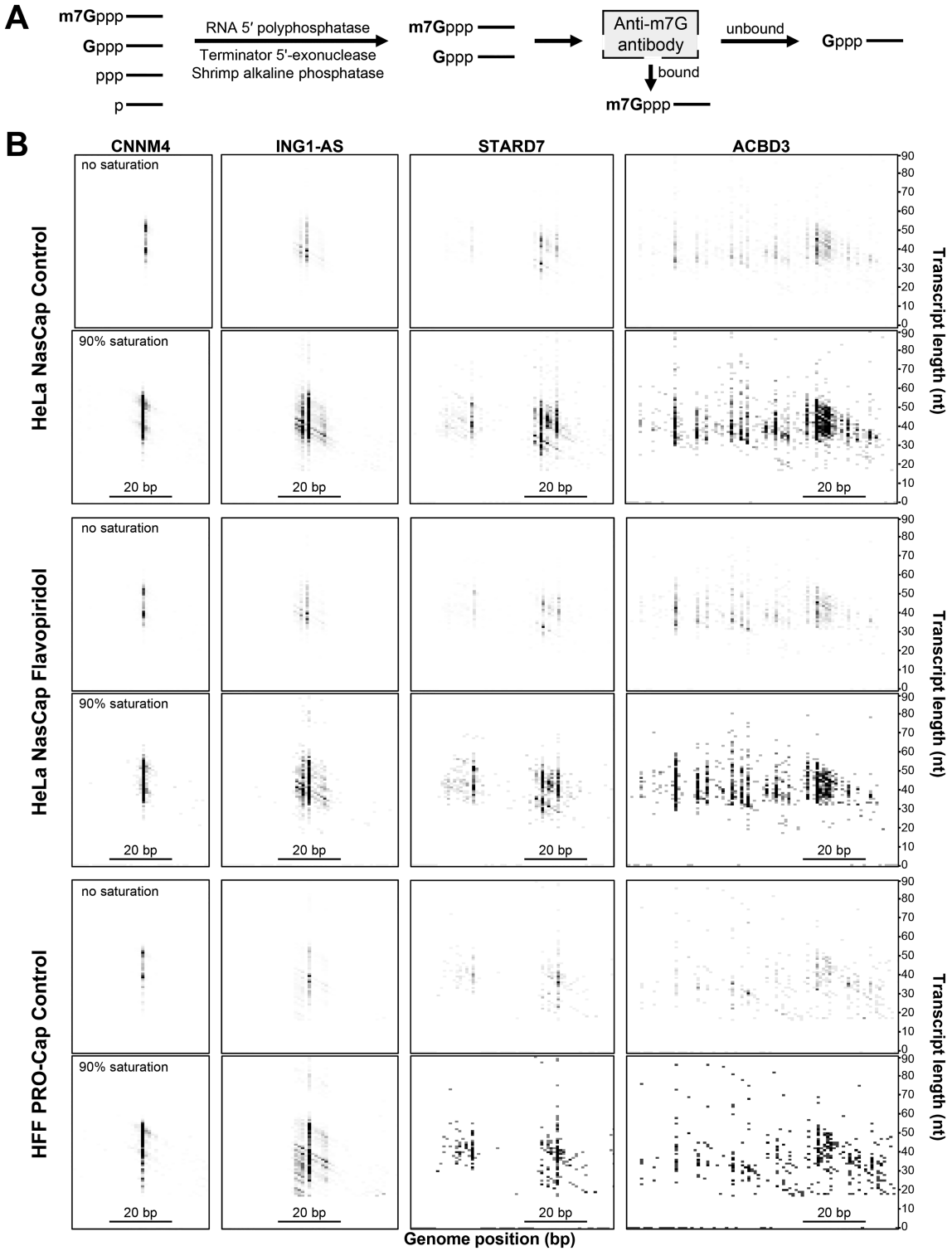

Figure 1. Analysis of nascent transcripts. (**A**) Diagram of the separation of biotinylated nascent transcripts that are capped and methylated (m7G) from those that have non-methylated caps. The steps shown occur after the first streptavidin isolation of nascent transcripts and after addition of the 3′ adaptor. Uncapped transcripts are removed by degradation and any remaining uncapped transcripts are blocked for 5′ adaptor ligation by removal of the 5′ phosphate. m7G capped transcripts bind to antibody beads and non-methylated caps flow through. After ligation of the 5′ adaptor, NasCap library preparation and sequencing results in determination of both 5′ and 3′ ends of nascent transcripts containing 5′ caps that are methylated or non-methylated. (**B**) The sequences of nascent transcripts obtained from NasCap libraries from control or flavopiridol treated HeLa cells or from control primary human foreskin fibroblasts were used to generate nucleotide resolution, truView plots of TSS (genome position) versus transcript length as described in the text and in the Methods section. Four examples of regions around the 5′ ends of genes are shown. ING1-AS is an antisense promoter within ING1 (chr13:110 714 873–110 715 815). The intensity of the spots represents the relative usage of the 3′ ends. In all cases, both an unmodified view and a 90% saturated view are shown.

Visualization of sequence features around transcription start and pause sites

Within any TSR there is a most frequently used TSS (MaxTSS) and downstream of that site is the most favored transcription pause site (MaxTPS). The RNA extending from the MaxTSS to the accompanying MaxTPS for each TSR is the most likely transcript arising from that genome region. We aligned the 177 098 MaxTSS to MaxTPS RNAs from the HeLa control NasCap m7G data centered on the TSS or TPS and displayed the genomic sequences as single-base heatmaps sorted by increasing RNA length or decreasing number of reads (TSS strength). The sequence patterns in the heatmaps revealed highly utilized elements around both the TSSs and the TPSs (Figure 2A and Supplementary Figure S2A). In addition, sequence logos (Figure 2B) and graphs of the average base distributions (Figure 2C and D) were generated from the MaxTSS to MaxTPS data as visual aids for the identification of common sequence elements.

A strikingly similar pattern of preferred sequences was evident extending from roughly −35 to +30 for TSSs when sequences were sorted by MaxTSS strength covering over a 10 000-fold range of RNA read values (Figure 2A-C). The most robust sequence motif was found around the TSS corresponding to the initiator (Inr). A preference for AT-richness between −25 and −30 and clear sequence preferences downstream of the Inr as well as between the AT rich region and the Inr are evident. The enrichment of specific sequences varied. The consensus across the 177 098 TSSs for the Inr was $CA_{+1}GT$ with enrichments being 2.0-, 4.0-, 1.4- and 1.8-fold for C, A, G and T respectively. T and A residues peaked between −30 and −25 with a 1.4-fold enrichment from the local average. A distinctive pattern of enrichment of G's was found downstream of the TSS at +7/+8, +12/+13, +18/+19, +23/+24 and +28/+29 ($GGN_3GGN_4GGN_3GGN_3GG$) and a depletion in C residues from +24 to +29 was evident. These downstream sequences are in positions that overlap with functional promoter elements described for a limited number of mammalian genes (10,13,16). Enrichment of C at three locations between −25 and −10 is also apparent, which is a region predicted to contact the TAF4 subunit of TFIID (7). Most importantly, the overall −35 to +30 region of sequence similarity exactly encompasses the footprint of TFIID on selected metazoan promoter-bearing templates *in vitro* (4,7,46,47). It is important to stress that the apparent TFIID interaction region emerges from averaging many individual TSSs together. All of the sequence preferences are not present around every TSS. Nevertheless, our results strongly suggest that the Pol II promoter centers on a TFIID binding site. For sake of simplicity, we will refer to each region that supports a TSS as a promoter and we equate the number of reads at that TSS with the strength of that promoter.

The MaxTSS to MaxTPS dataset is based on a 20 bp TSR discovery algorithm that does not allow overlapping of TSRs, followed by selection of the MaxTSS present in each TSR and its corresponding MaxTPS. To determine whether this initial selection of TSSs strongly influenced the sequence patterns we observe in Figure 2, we also selected from the control NasCap m7G data a second dataset which simply consists of the 5′ ends of all nascent RNAs with 10 or more reads. This 'Selected TSS' set contains 522 186 TSSs. As shown in Supplementary Figure S2B,C, the sequence similarities from −35 to +30 seen in Figure 2 are also readily apparent with the Selected set, indicating that these patterns are not dependent on the initial assignment of MaxTSSs within TSRs.

There is an important distinct category of previously-described Pol II promoters which would not be expected to center on an Inr, namely the promoters for the ribosomal protein genes (RPGs) (48,49). Supplementary Figure S2D shows the sequence logo of the 200 bp surrounding the MaxTSS from 77 RPG promoters from the control NasCap m7G dataset. The TCT element ($TTCC_{+1}TTTT$) is present around the TSS as expected. These promoters, which use TRF2 instead of TBP (49), nevertheless feature an AT rich element at about −30 which is as prominent as the AT-richness seen with the strongest Inr-based promoters (Figure 3A). The sequences downstream of the TSS in RPG promoters differ significantly from those seen with the Inr-centered promoters (Supplementary Figure S2D), which are recognized by TAF1 and TAF2 (4). It is not yet known which factors accompany TRF2 at RPG promoters, although a recent study reported TAF1 association with RPG promoters in Drosophila (50).

To what extent is promoter strength related to promoter sequence?

The distinctive sequence signature corresponding to the TFIID binding site in the single-base sequence heatmaps is apparent over the entire 10 000-fold range of promoter strength (Figure 2A), a striking and unexpected characteristic. To better assess the connection of the sequence features with strength, we generated logos for the strongest 10%, 1% or 0.1% of the control m7G MaxTSS to MaxTPS set. As seen in Figure 3A, the sequence similarities discernable in the heatmaps of the entire set become more pronounced when the strongest promoters are considered. The consensus Inr for the strongest promoters is more clearly defined as $CA_{+1}GT$ (14,51). At these positions, C, A, G and T are enriched 2.3-, 5.6-, 1.9- and 2.8- fold over the surrounding sequence for the top 1% promoters. The preference for G at +12/+13, +24/+25 and +28/+29 is 1.2- to 1.4-fold higher than the surrounding sequence for the strongest 1% of promoters. Interestingly, while a particularly high AT content from −30 to −25 is associated with very strong promoters, a preference for AT in this region is characteristic of all promoters in the MaxTSS to MaxTPS set and the larger Selected TSS dataset (Supplementary Figure S3A). To compare stronger and weaker promoters more directly, we separated out top and bottom quartiles of the MaxTSS to MaxTPS set. Average base distributions indicated that GC content declined as promoter strength decreased (Supplementary Figure S3B). The G-rich elements from +7 to +29 are enriched in both the top and bottom quartiles with only a small correlation with promoter strength when compared to the general decrease in GC content. Interestingly, the C-poor region from +24 to +29 was more prominent in the top quartile.

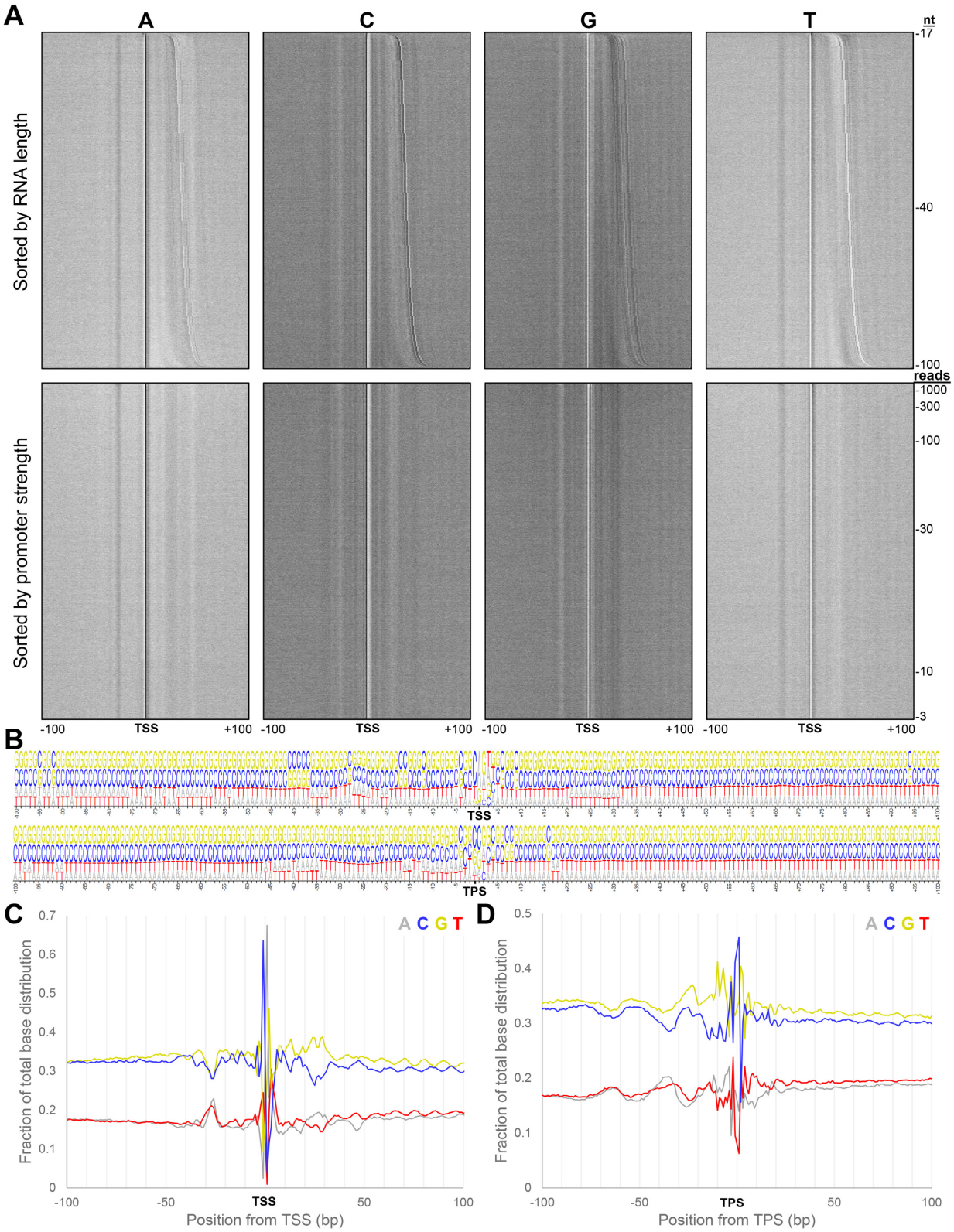

Figure 2. Sequences around TSSs from the MaxTSS to MaxTPS control m7G dataset. The MaxTSS to MaxTPS HeLa NasCap control m7G dataset (n = 177,098) was used to analyze the sequences −100 to +100 surrounding the TSSs or TPSs. (**A**) Sequence heatmaps (600 × 1200 pixels) were generated displaying the relative distribution of each individual nucleotide. The 200 bp intervals with the MaxTSS at center (+1) were vertically sorted by RNA length (17 to 100 nt) or promoter strength (number of reads for each TSS/TPS pair). The gray scale represents the fraction of the indicated nucleotide at each position (1, black; 0, white). (**B**) Fractional base distributions around the TSSs and TPSs are displayed by sequence logo. Fractional base distributions from (B) were graphed directly for the MaxTSSs (**C**) or MaxTPSs (**D**).

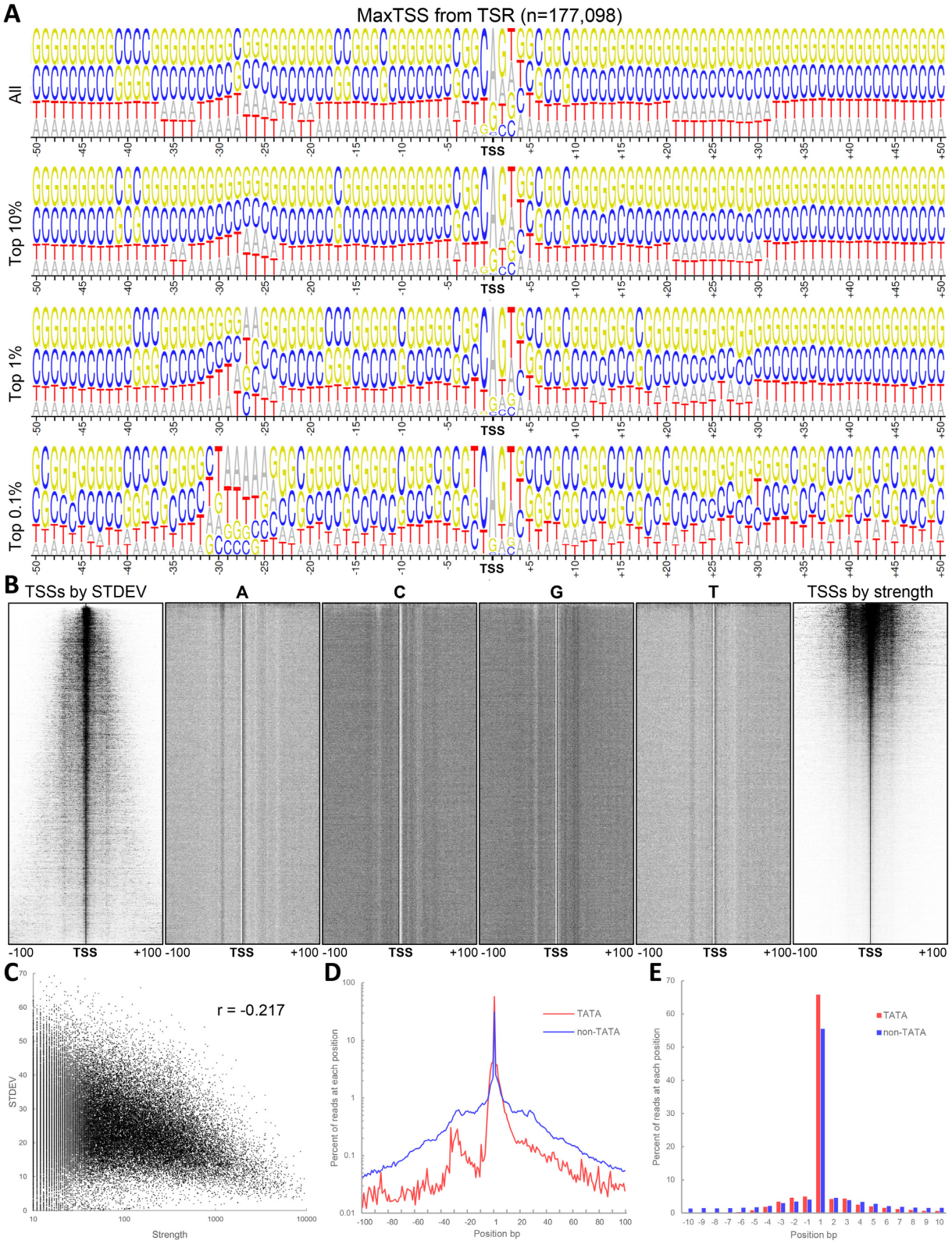

Figure 3. Correlations of promoter elements with MaxTSS strength and focus of surrounding TSSs. (**A**) The number of reads from each TSS was used to sort the MaxTSS to MaxTPS control m7G dataset (n = 177 098). Logos were created for the indicated portions of that dataset. (**B**) Heatmaps (600 × 1200 pixels) for TSS distribution and sequence (A, C, G and T) were generated for the 200 bp TSR dataset (n = 62 381) after sorting by standard deviation of TSS utilization across the 200 bp interval with highly focused TSRs at the top and least focused TSRs at the bottom. The right panel shows TSS distribution heatmap sorted by MaxTSS strength. (**C**) Correlation of focus (STDEV) of the 200 bp TSRs versus strength of the MaxTSS. Pearson coefficient is indicated. (**D**) Metaplot of the distribution of TSSs for TATA (n = 1408) or non-TATA (n = 60 973) TSRs normalized to total reads (%) in each dataset. (**E**) Distribution of TSSs from −10 to +10 around the MaxTSS for the TATA and non-TATA datasets after normalization of the reads (%) over that 20 bp region.

The TATA element is most often associated with promoter strength, and the preference for T and A between −30 and −25 is particularly prominent for the strongest promoters (Figure 3A and Supplementary Figure S3A). However, as well established in earlier work, not all strong promoters have TATA elements (10,14). For example, the very strong CNNM4 promoter in Figure 1 is TATA-less. We defined TATA promoters in the MaxTSS to MaxTPS set as those in which the upstream T of TATA is located from −34 to −29, giving a total of 1,896 promoters. Of these, only 446 (24%) are found in the top decile of promoter strength. The remainder are roughly equally distributed among the other nine strength deciles. A comparison of the average base distributions for the TATA promoters (Supplementary Figure S3B) with all MaxTSS to MaxTPS promoters indicates that some of the downstream sequence preferences are not as strong for the TATA group (note in particular G at +28/+29). This is consistent with earlier proposals that consensus TATA promoters are less dependent on downstream elements (10,13,16,17). Interestingly, the gradual rise in GC content with promoter strength does not affect the −30 to −25 region. Thus, the AT content in that area is more pronounced for the stronger promoters, which could be important for focusing the position of PIC formation. This localized preference for AT presumably reflects the importance of the sharp bend in the templates at about −30 within the PIC associated with the interaction of TBP with the DNA (4,7).

As noted earlier, Pol II promoters are often divided into those that support initiation over a relatively narrow range of TSSs (focused) and those that do not (diffuse/dispersed) (10–12,52). It is often thought that focus is linked to promoter strength. To determine to what extent the sequence patterns we observe are connected to the local spread of TSSs, we generated a new set of 200 bp wide, non-overlapping TSRs in which the MaxTSS has 10 or more reads and is centered in the TSR. This approach is optimal for assessing the relative use of locally strong TSSs in comparison to alternative starts over a range typically associated with diffuse/dispersed promoters. We computed the standard deviation of the distribution of all TSSs around the MaxTSS for each TSR in this 62,381 member set and sorted the promoters from most to least focused (Figure 3B, left panel). Single-base heatmaps for these TSSs were also sorted in the same way (Figure 3B). Alternatively, we sorted the TSS heatmap by promoter strength, from most to least MaxTSS reads (Figure 3B, right panel).

Several important points are apparent from these figures. The single-base sequence heatmaps show that the sequence similarity from −35 to +30 is uniformly evident regardless of the spread of additional TSSs around the local max TSS. In contrast to previous suggestions, the strongest promoters are more likely to be accompanied by flanking TSSs in comparison to the weaker promoters (Figure 3B, right panel). When strength is plotted against standard deviation, there is only a weak negative Pearson correlation ($r = -0.217$) for the very large majority of promoters; only the exceptionally strongest show less spread of TSSs (Figure 3C). Within this TSR set, MaxTSSs from TATA-containing promoters are surrounded by relatively fewer TSSs in comparison to the remaining MaxTSSs from non-TATA promoters (Figure 3D). However, when only the 20 bp regions surrounding the MaxTSSs are compared, the spread of flanking TSSs is similar for TATA and non-TATA promoters (Figure 3E). All of these results strongly reinforce the idea that there is only one type of Pol II promoter, centered on a TFIID binding site. The spread of TSSs that each promoter supports is confined to roughly ±5 bp around the optimal TSS. The juxtaposition of several such simple promoters within a larger genomic region (e.g. ±100 bp as in Figure 3B) accounts for TSS groups that in earlier studies were assigned to single diffuse promoters. This point is explored further in the Discussion.

What parameters affect the location of pausing and methylation of the cap?

We generated metaplots to determine the effect of cap methylation and flavopiridol treatment of cells on nascent transcript length. The large majority of paused HeLa RNAs with m7G caps are 30–50 nt in length in the MaxTSS to MaxTPS set or the larger Selected TSS set of RNAs (Figure 4A, Table 1). It has been suggested that paused RNAs might be shorter when driven by the strongest promoters (21). However, a significant effect of promoter strength on transcript length is not apparent with the m7G-capped RNAs from the MaxTSS to MaxTPS or Selected TSS sets (Figure 4A, Table 1). Flavopiridol treatment of the HeLa cells led to a very slight (∼1 nt) increase in pause RNA lengths. Comparison of the transcript lengths from the four NasCap datasets demonstrates that nascent transcripts containing non-methylated caps are substantially shorter on average compared to m7G RNAs (Figure 4A and Table 1). There are two subpopulations of cap-only RNAs, with one group 20–30 nt long and a second longer group which is slightly shorter than the peak of m7G RNAs (Figure 4A). The 20–30 nt RNAs represent a larger proportion of the non-methylated RNAs when only the more abundant non-methylated RNAs are considered (Figure 4A). Importantly, our results indicate that cap methylation normally begins as transcripts reach 30 nt in length.

The sequences surrounding pause sites were examined using several different sets of TPSs. TPSs from the MaxTSS to MaxTPS set feature a distinctive base composition (Figure 2B). C is enriched 1.4-fold at the last base incorporated by Pol II in cells (called −1) and 1.5-fold at +1, the next base to be incorporated. C is also strongly depleted at +2 (0.5-fold relative to the surrounding sequence). G is enriched at −10 (1.3-fold), −7 (1.2-fold) and +2 (1.3-fold), while A is depleted at −3 (0.6-fold relative to surrounding A levels). The Selected TPS set (725,644 RNAs) containing all paused RNAs with at least 10 reads gave a much less feature-rich logo compared to the MaxTSS to MaxTPS set (compare Figure 4B with 2B). One explanation for the difference is that each TPS in the MaxTSS to MaxTPS set is a locally maximum TPS, while the Selected TPS set contains many pause sites downstream of strong promoters that are not the most frequent pause site in any local region. To eliminate the effects of promoter strength on pause site selection, we developed a Transcription Pause Region (TPR) Finder (tprFinder) and used it to identify the local MaxTPS within

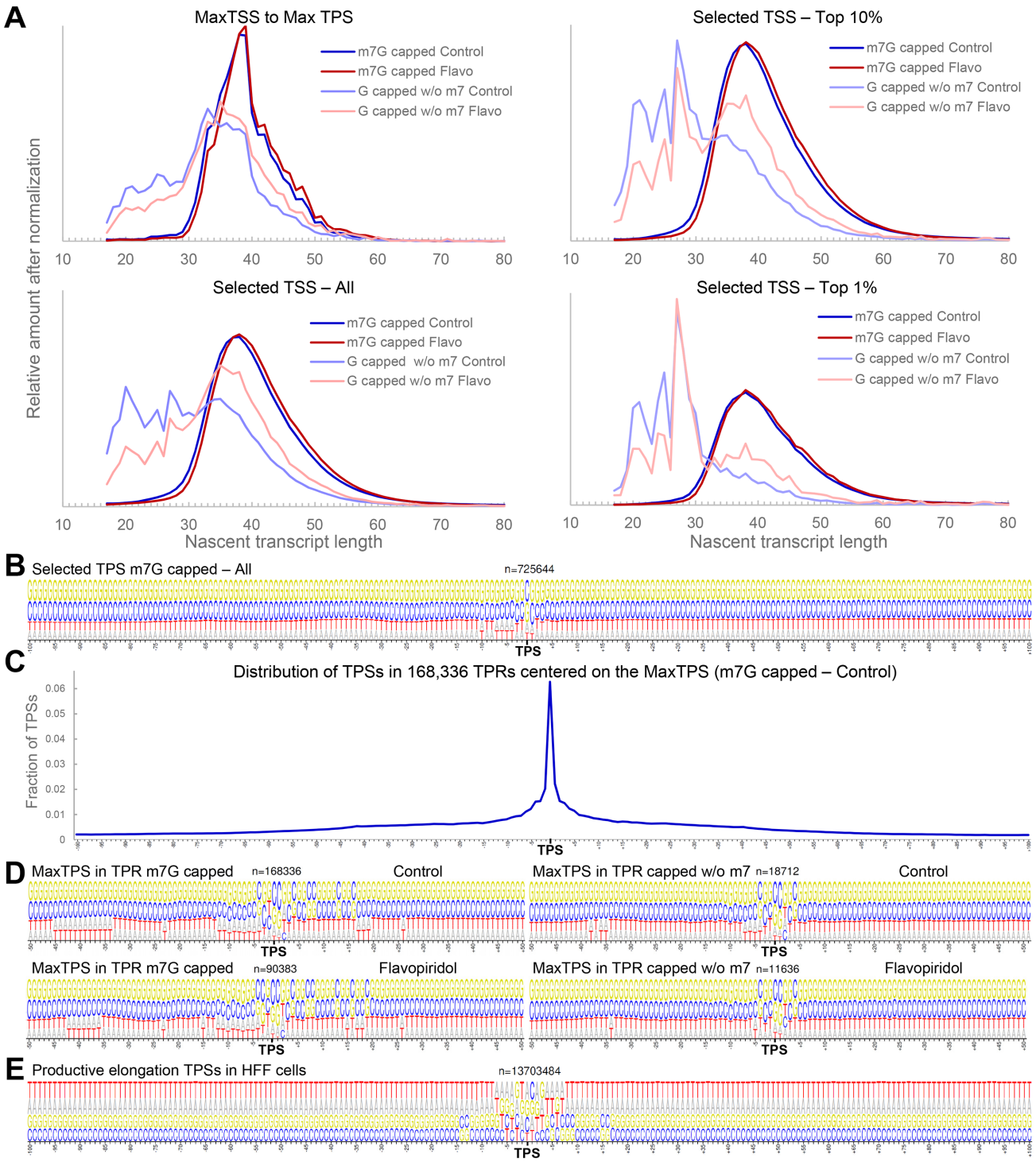

Figure 4. Distribution of nascent transcript sizes and analysis of TPSs. (**A**) The frequency distributions of transcripts between 17 and 100 nt from the MaxTSS to MaxTPS dataset and the indicated portions of the strength-sorted Selected TSS dataset were plotted to compare both m7G capped and capped but non-methylated nascent transcripts from control or flavopiridol treated cells. The number of reads in each dataset is found in Table 1. (**B**) Logos were created for –100 to +100 sequences surrounding the TPSs for the Selected TPS m7G control dataset, (**C**) Metaplot of TPSs in the TPR dataset ($n = 168\,336$). (**D**) Logos in the area surrounding the MaxTPSs from MaxTSS to MaxTPS datasets derived m7G capped transcripts or non-methylated transcripts from control or flavopiridol treated cells. (**E**) Logo of sequences surrounding the TPSs in regions of genes undergoing productive elongation from PRO-Seq data from HFF cells (39).

non-overlapping 40 bp regions. A metaplot of the 168,336 TPSs from TPRs demonstrated that MaxTPSs predominate among possible pause sites (Figure 4C). Logos centered on the MaxTPSs for m7G capped transcripts from the TPR set were very similar to those seen from the MaxTSS to MaxTPS set (compare Figures 4D to 2B). There was little difference between the logos for m7G capped RNA and the capped but non-methylated transcripts (Figure 4D). Treatment of cells with flavopiridol to block entry into productive elongation reduced the prevalence of C at the first position after the pause (Figure 4D) for both m7G and non-methylated capped RNAs. Analysis of PRO-Seq data in primary HFFs that employed the same rapid nuclear isolation method used here (39) demonstrated that sequences of pause sites during productive elongation within gene bodies do not resemble the sequences around the major promoter-proximal pause sites (Figure 4E). Overall, the analysis of TPSs indicates that the major pause sites downstream of TSSs are distinctive and different from the average sites of pausing around promoters and in gene bodies.

Tome *et al.* (19) recently examined promoter-proximal human RNAs in which both 5′ and 3′ ends were determined using a PRO-seq approach they termed CoPRO. That study reported a broader range of paused RNA lengths in comparison to what we observe for m7G RNAs. MaxTSS to MaxTPS and Selected TSS datasets were generated from the Tome *et al.* (19) datasets exactly as was done from our NasCap datasets. As seen in Supplementary Figure S4A, the TSS logos for the CoPRO RNAs are very similar to our TSS logos. However, the CoPRO RNAs show a much broader and somewhat biphasic length distribution compared to our cap-modified RNAs (Supplementary Figure S4B). Importantly, CoPRO does not distinguish between methylated and non-methylated caps. Our results strongly suggest that the shorter paused RNAs in the CoPRO set primarily correspond to non-methylated caps. The shorter average RNA lengths we observed for the HFF RNAs, in comparison to the HeLa transcripts (Figure 1B), likely resulted from this same effect, since the HFF RNAs were prepared by the PRO-cap procedure which does not distinguish methylated and non-methylated caps. A possible reason for the presence of a tail of much longer RNAs in the CoPRO set can be seen in the logos for the CoPRO pause sites, which differ from ours by a much greater tendency to stop before adding a C (Figure 5A). This is likely caused by differences in the procedures for isolating nuclei. The nuclear walk-on method we developed (37) and used here halts NTP incorporation within 20 seconds of removal of the cells from 37°C incubation and eliminates all NTPs during nuclear isolation, whereas the cell washing and nuclear isolation in the standard PRO-seq method allows for continued elongation in the presence of diluted cellular NTPs. The appearance of primarily C stops in the CoPRO data is likely due to the fact that CTP is the limiting nucleotide in mammalian cells (53).

Analysis of far downstream sequences

When the base distribution plots were extended to 500 bp downstream of the TSSs, we were surprised by a clear 10 bp periodicity of G, C, T and A from about +50 to +200 in both the MaxTSS to MaxTPS and Selected TSS datasets (Figure 5A). This periodic downstream element (PDE) stops abruptly before +200. The 10 bp periodicity within the PDE is strongly suggestive of a nucleosome positioning element. Recent work (54) has shown that in human nucleosomal DNA, maximum levels of SS and WW dinucleotides both vary with a 10 bp spacing, with the SS and WW peaks displaced from each other by 5 bp. More recent studies determined that among the decamers that maintain the appropriate spacing of SS and WW in human promoter-proximal (+1) nucleosomes, SSNYYWWNRR was the most frequently found (55). This decamer sequence defines the relative phases of the individual bases as shown in Figure 5B. Importantly, this is the base distribution found in our data downstream of TSSs (Figure 5B). A plot of the location of the SSNYYWWNRR decamer demonstrated that it is found in a 10 bp periodic distribution from +50 to at least +180 downstream of the TSSs (Figure 5C). When promoters were separated into the top and bottom quartiles by promoter strength, the SS and WW periodicity was still visible, although fainter, even for the weakest quartile (Figure 5D and Supplementary Figure S5A). Visual examination of single and double nucleotide distribution plots led to the discovery of peaks of CT and TC that overlapped by 1 nt. As seen in Figure 5E, the sequence CTC is highly enriched from +47 to +49. Both the PDE and the CTC peak were clearly evident in the control m7G Selected TSS dataset (Supplementary Figure S5B). Although we do not know the function of the CTC sequence, it would break the sequence periodicity in the PDE and could reinforce nucleosome positioning downstream of +50.

Localizing the +1 nucleosome relative to the TSS using DFF cleavage and nuclear run-off

It is often asserted (30–32) that the promoter-proximal edge of the +1 nucleosome is positioned at about +50 relative to the TSS in mammalian cells. The location of the PDE is suggestively consistent with that model. However, in earlier studies nucleosome locations were based on overall patterns of protection from micrococcal nuclease and imprecise TSS assignment. Significantly, the large majority of promoters are utilized infrequently (38,56,57); thus, it is not clear if average nucleosome positions reflect nucleosome locations on the crucial subset of templates that are actually being transcribed. We therefore examined potential occupancy of downstream nucleosomes on transcribed DNA directly in isolated nuclei using Pol II as the reporter. The nuclear run-off strategy is shown schematically in Figure 6A. Micrococcal nuclease, which degrades RNA and nicks DNA within nucleosomes, is not suitable for this method so instead we expressed and purified the DNA Fragmentation Factor (DFF) (Figure 6B) (43,58). DFF digestion of nuclei generates a ladder of nucleosome-length DNAs, reminiscent of ladders obtained with micrococcal nuclease but without the production of easily detectable sub-nucleosomal fragments and without any RNase activity (Figure 6C). We anticipated that DFF cleavage would occur downstream of the transcription complex or in cases where Pol II was immediately adjacent to the nucleosome, downstream of the putative +1 nucleosome (Figure 6A).

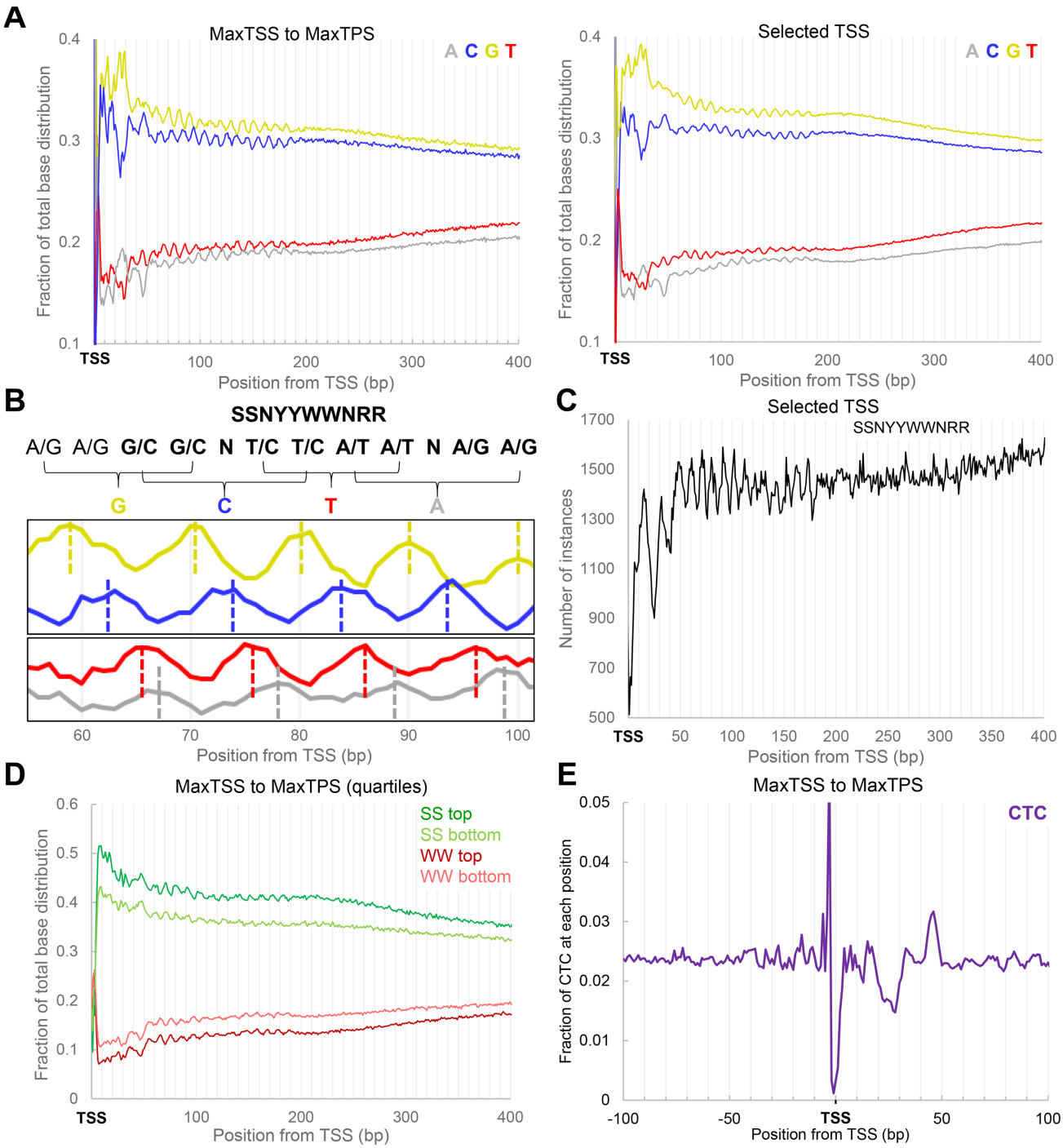

Figure 5. Far downstream promoter elements. (**A**) Extended fractional base distributions for MaxTSS to MaxTPS and Selected TSS control m7G datasets. (**B**) Identification of a 10 bp nucleosome positioning element. Nucleotide frequency distributions of individual bases from the MaxTSS to MaxTPS control m7G dataset correlate with the promoter-proximal nucleosome positioning element previously reported (55). (**C**) Distribution of SSNYYWWNRR downstream of the TSSs from the Selected TSS control m7G dataset.(**D**) Comparison of the distribution of SS and WW from the top and bottom quartiles (based on TSS utilization) from the MaxTSS to MaxTPS control m7G dataset. (**E**) Distribution of CTC from the MaxTSS to MaxTPS control m7G dataset.

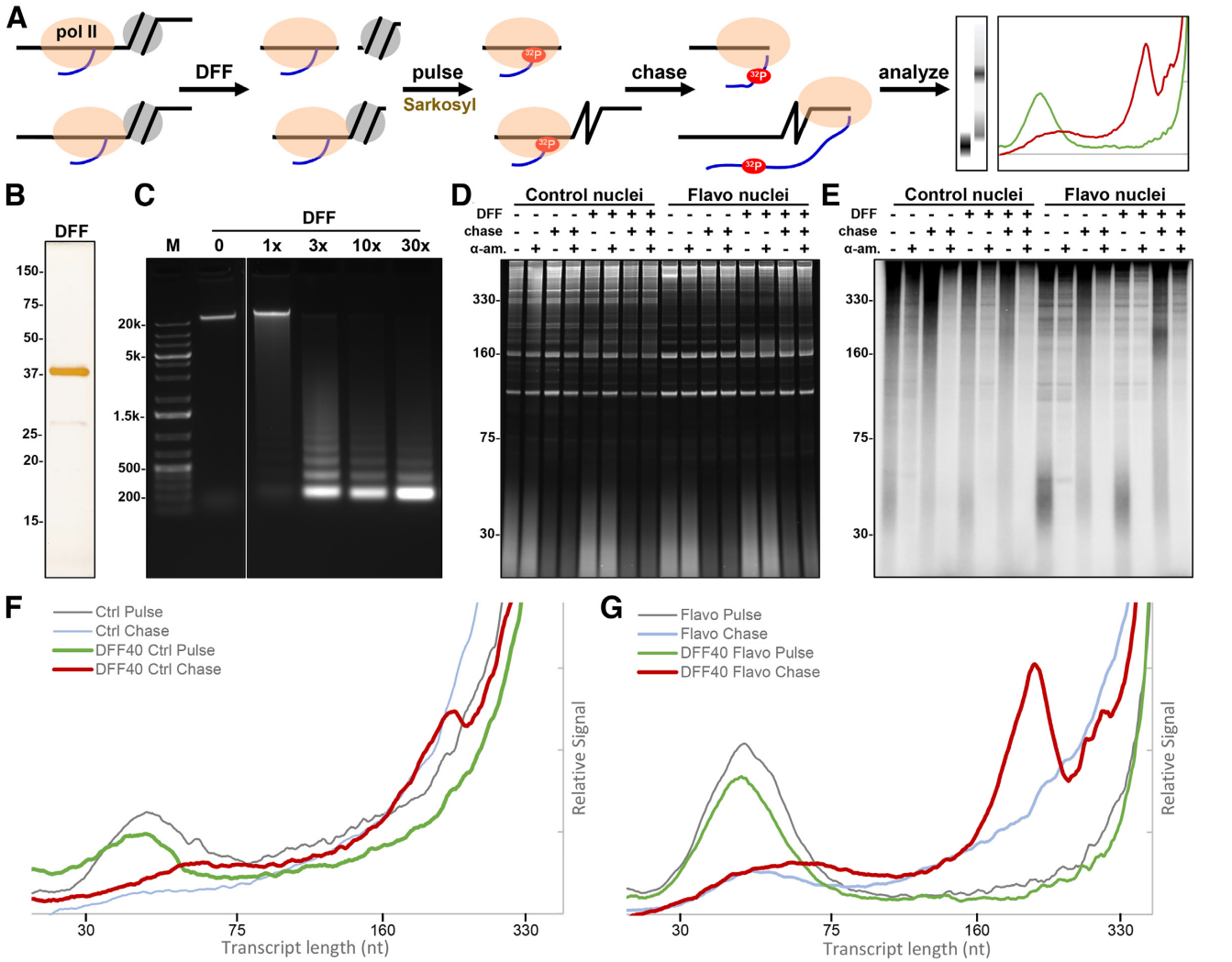

Figure 6. Nuclear run-off assay. (**A**) Diagram of the nuclear run-off assay. (**B**) Purified DFF40 (silver stained gel); positions of molecular weight markers are indicated. (**C**) Digestion of nuclei with increasing amounts of DFF and analysis of the resulting DNA fragments in a native agarose gel. (**D**) Ethidium bromide stained denaturing RNA gel showing endogenous RNAs from nuclei. (**E**) Phosphorimage of gel in (D) showing radiolabeled RNAs synthesized in the isolated nuclei. (**F**) Pol II profile analysis from amanitin-sensitive radiolabeled RNAs synthesized in control and DFF digested nuclei from untreated cells after pulse or pulse and chase. (**G**) Pol II profile analysis from amanitin-sensitive radiolabeled RNAs synthesized in control and DFF digested nuclei from flavopiridol treated cells after pulse or pulse and chase.

After DFF treatment, nascent RNAs were pulse-labeled and then chased with excess unlabeled NTPs to the closest downstream DFF cleavage site. The pulse and chase were performed in the presence of Sarkosyl to negate the influence of factors and nucleosomes on elongation. The labeled RNAs were resolved on gels and the Pol II transcript profiles (difference between minus and plus α-amanitin) were compared after slight corrections for loading based on the EtBr staining of endogenous RNAs (Figure 6D–G). For both control and flavopiridol-treated nuclei, nearly all the paused RNAs chased in the absence of DFF cleavage. DFF treatment did not have a major effect on the ability of engaged Pol II to incorporate label during the pulse. Two populations of chase products were observed in DFF-cleaved nuclei. One population of RNAs was only extended 15–30 nt indicating that for those complexes, DFF could digest the DNA immediately downstream of the polymerase. The second set chased to about +200, which is consistent with the paused Pol II being very close to a well-positioned +1 nucleosome. This second population dominated in nuclei from cells treated with flavopiridol to block the transition into productive elongation. The change in transcript patterns generated by DFF treatment has been highly reproducible (Supplementary Figure S6).

To directly observe the products of DFF digestion of nuclei from flavopiridol treated HeLa cells, a DFF-Seq library was generated and sequenced resulting in about 260 million mapped, paired-end reads. Heatmaps and metagene analyses for the 2000 bp around the TSSs from the 177,098 member MaxTSS to MaxTPS dataset were generated directly from 185 million fragments (71% of the total) between 140 and 185 bp. A DFF-protected region is located with remarkable consistency downstream of the TSSs when either promoter strength or the length of the most preva-

lent paused transcript length was used to sort the data (Figure 7A). Similar results were obtained when heatmaps were generated from the 62 381 member set of 200 bp TSRs (from Figure 3B) that were sorted by promoter strength or by standard deviation (focus) (Figure 7B). As expected, the average base distribution plots indicate that the PDE and the CTC elements are also present in the 200 bp TSR dataset (Figure 7C). The results from both datasets provide strong evidence for a well positioned +1 nucleosome consistent with the presence of the PDE element and the length of the RNAs in the nuclear run-off experiment after DFF cleavage. The apparent +1 nucleosome fragments from both datasets (Figure 7D) center on +114 ± 2, suggesting an average upstream edge at +42 for that nucleosome. This nucleosome location is very similar for promoters with TATA elements (Figure 7E) and the strongest (top 1%) promoters (Figure 7F). A minimum of DFF protection was observed centered about −53 ± 4 for the two promoter sets, which shifts slightly upstream relative to the +1 nucleosome for the TATA promoters. The RPG promoters also have a +1 nucleosome at nearly the same location as the Inr-based promoters but their DFF protection pattern upstream is distinct, with the minimum near the TSS (Figure 7G).

The distribution of fragments in the DFF-Seq library provides evidence for the unique properties of DFF. The vast majority of fragments were slightly larger than expected for protection by nucleosomes (Supplementary Figure S7A). Fragments less than 120 bp (0.3% of the total) were only visible when a log scale was used to display the number of fragments (Supplementary Figure S7A, inset). However, heatmaps and metaplots generated from DFF-Seq fragments less than 120 bp show a striking pattern, with these fragments concentrated upstream of the TSS and highly depleted in the +50 to +200 region (Supplementary Figure S7B). This upstream signal correlated with promoter strength but not TSR focus (Supplementary Figure S7B) and likely represents protection by transcription regulatory factors. In contrast, the density of nucleosome sized fragments upstream of the promoter decreased with increasing promoter strength (Supplementary Figure S7C). The average transcript lengths varied by 13 bp for the four quartiles of length but the position of the +1 nucleosome varied <3 bp, indicating that transcript length was not affected directly by the position of the +1 nucleosome (Supplementary Figure S7D).

DISCUSSION

We have determined with base pair precision the 5′ and 3′ ends of exceptionally high numbers of promoter-proximal nascent RNAs from HeLa cells. When we aligned the TSSs from these RNAs, striking sequence patterns were evident. The distinctive sequence signature from −35 to +30 includes all documented Pol II promoter elements from functional studies (10,13,16) and exactly encompasses the reported footprint of TFIID on DNA (4,7,46,47,59). The presence of this sequence signature is remarkably robust. It is readily apparent for 177 098 MaxTSSs contained within 20 bp wide TSRs, or for 62 381 MaxTSSs centrally positioned within 200 bp TSRs, or simply when we consider all 522 186 TSSs with more than 10 reads (Figures 2A, Supplementary Figure S2A, 3B). Importantly, it is also apparent for promoters whose strength varies by more than three orders of magnitude. These observations point unavoidably to the conclusion that human Pol II promoters are all organized around a single core, which is a TFIID binding site (Figure 8A). The only exceptions are the RPG promoters (Supplementary Figure S2D) and the Pol II promoters for U series RNAs, which we excluded from our analysis. While this paper was in review, Fant *et al.* (60) reported that TFIID is essential for pausing downstream of both Drosophila and human promoters. Given that pausing occurs at nearly all Pol II promoters (18–20), this finding independently supports the idea that Pol II promoters are centered on TFIID binding sites.

Stringent nuclear isolation protocols to freeze Pol II in its paused location in the cell prior to run-on analysis (37) allow greater insight into sequence elements that direct pausing. The TPSs from the MaxTSS to MaxTPS control m7G dataset and the MaxTPSs from the TPR sets are the most frequently used pause sites from nearby TSSs. We observe clear sequence preferences at and upstream of these pause sites (Figures 1B, 4D, 8A). The range of these preferred sequences coincides with the expected position of the transcription bubble or in the case of G at −10, the point at which the DNA strands initially reanneal (61). This G residue is particularly noteworthy because this base is favored at the analogous location for paused bacterial RNA polymerase (62).

Once the nascent RNA emerges from within Pol II the cap is added immediately, at about 18 nt (63). We now show that the m7G modification does not take place until the nascent RNA is at least 30 nt (Figure 4A). For those RNAs with m7G caps, pausing occurs primarily over a relatively narrow range, from +30 to +50, and that range is essentially independent of promoter strength. However, while the lengths of RNAs without cap modification overlap the range of m7G RNAs, a substantial fraction of cap-only RNAs are <30 nt and this fraction does increase with promoter strength (Figure 4A). This effect could, in part, be the basis for the earlier observation in Drosophila cells that the strongest promoters preferentially support shorter paused RNAs (21).

A sequence element called the pause button (KCRWCG) was correlated in earlier work to high levels of promoter proximal pausing in Drosophila cells (64). We have not explored the relative extent of escape into productive elongation for the HeLa cell promoters so we cannot comment directly on the relationship of the elements we report here with pause release. The pause button did not map to pause sites but instead frequently coincided with downstream promoter elements (DPEs). The Drosophila DPE would correspond to a downstream segment of the core promoter signature we identify. Very recently, Shao *et al.* (65) reported a correlation of efficient escape into productive elongation in Drosophila with features of the Inr, particularly a preference for G at +2. We also see G as the preferred +2 base, especially for the strongest promoters (Figure 3A), but we have not ranked the HeLa promoters based on relative efficiency of escape from pausing.

The sequences surrounding Pol II TSSs not only revealed a characteristic promoter signature, but also the presence of

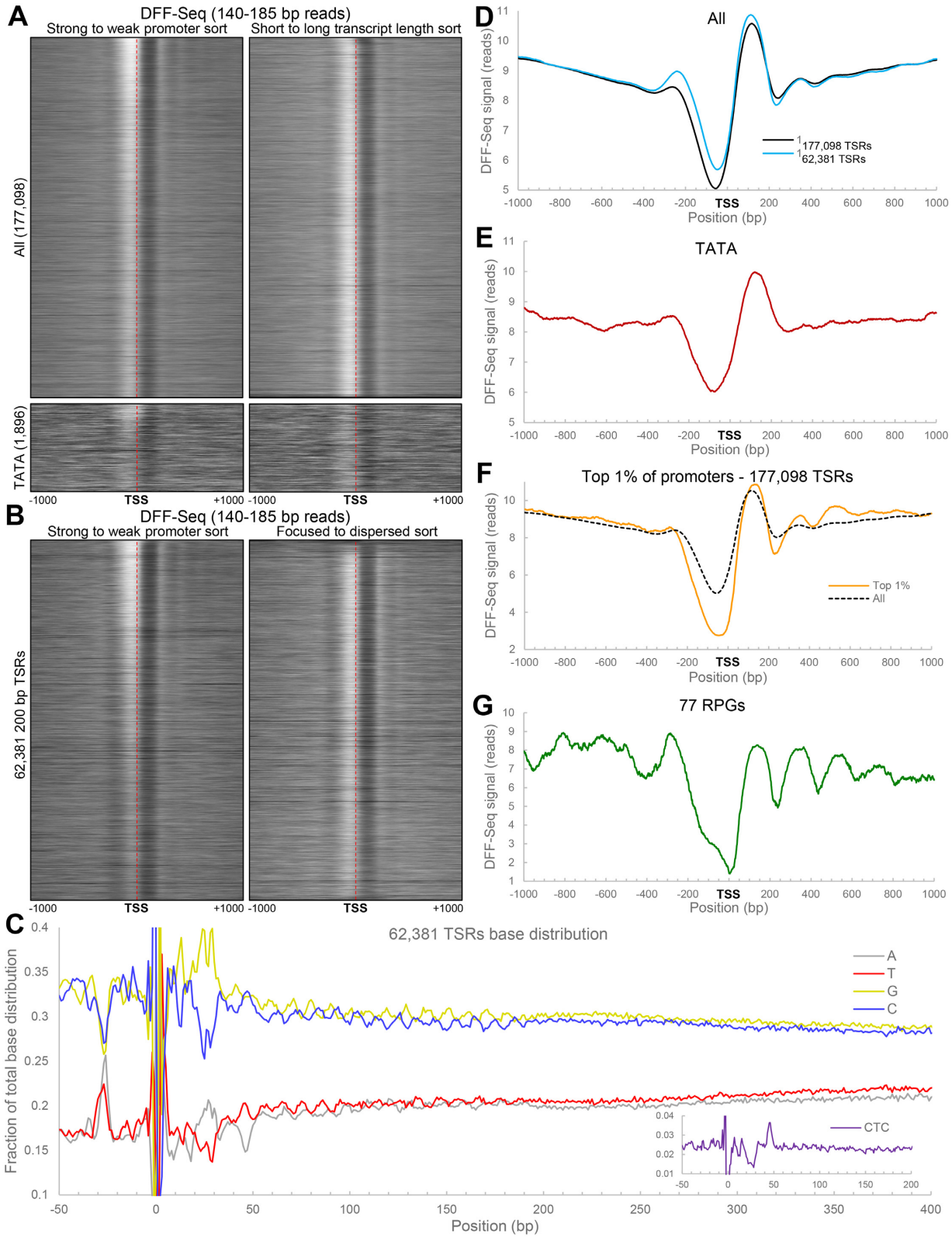

Figure 7. DFF-Seq. Nuclei from flavopiridol treated cells were digested with DFF as described in Methods and libraries from the isolated DNA were prepared and sequenced. (**A**) Fragment lengths from 140 to 185 bp were used to generate a heatmap (1000 × ∼1700 pixels) for the interval ±1 kb centered on the MaxTSS of the MaxTSS to MaxTPS control m7G dataset, sorted as indicated (All). In addition, intervals for only those promoters with TATA elements (first T of TATA from −34 to −29) were used for a separate heatmap (1000 × 632 pixels). (**B**) Fragment lengths from 140 to 185 bp were used to generate a heatmap (1000 × ∼1700 pixels) for the interval ±1 kb centered on the MaxTSS in the 62 381 member control m7G dataset selected using 200 bp TSRs, sorted as indicated. (**C**) Fractional base distributions for A, C, G, T and CTC (inset) for the 200 bp TSR dataset for the regions indicated. (**D**) Metaplots of the 177 098 and 62 381 member datasets in (A) and (B). (**E**) Metaplot of the TATA promoters from (A). (**F**) Comparison of metaplots of all 177,098 TSRs with the top 1% based of TSS strength. (**G**) Metaplot of the 77 ribosomal protein gene promoters.

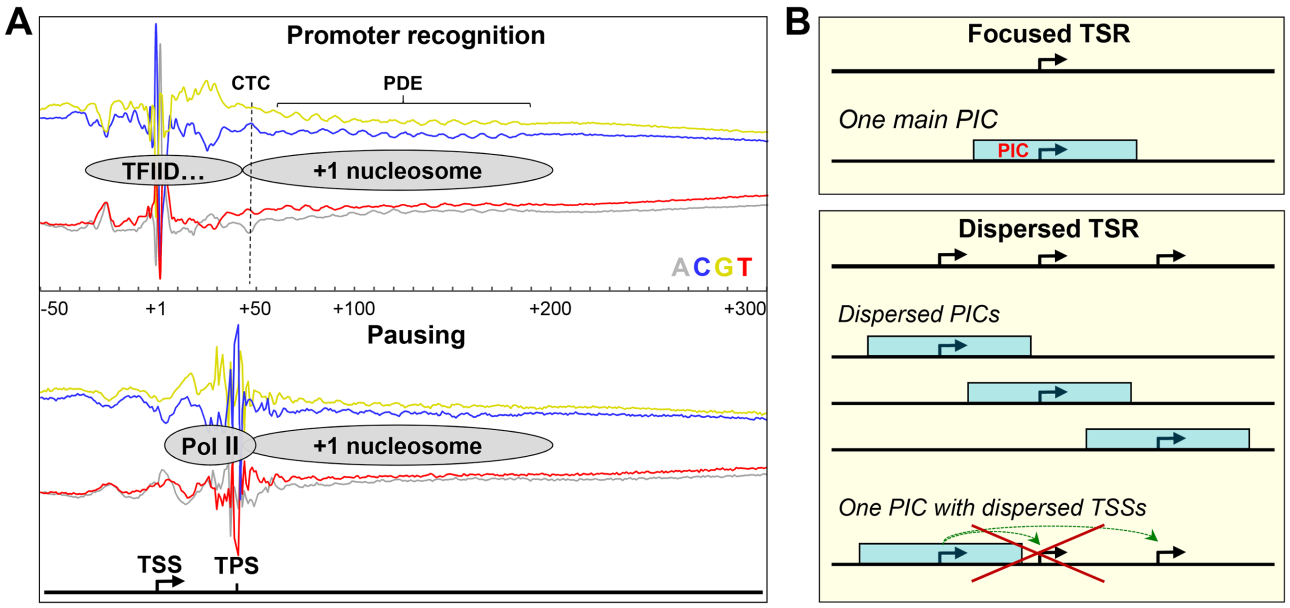

Figure 8. Models. (**A**) Model summarizing our findings. Promoter recognition occurs through interaction of TFIID with sequences surrounding the TSS and with a positioned +1 nucleosome. Pausing is influenced by both downstream sequence and through blockage by the downstream +1 nucleosome. The location of CTC and PDE sequence elements are indicated. (**B**) Unification of focused and dispersed promoters through focused and dispersed TSRs. Each major TSS is generated by a promoter directing a single preinitiation complex which can support only a narrow range (~±5 bp) of alternative TSSs. Focused TSRs have one major promoter and dispersed TSRs have several or many dispersed promoters. Human transcription does not follow the yeast model with a single PIC generating many TSSs over a very wide (up to 100 bp) range.

another sequence, the PDE, which could impact the location of the +1 nucleosome. From +50 to +200 downstream of the TSSs the individual bases in the MaxTSS to MaxTPS, Selected and 200 bp TSR datasets (Figures 5A and 7C) all vary with a 10 bp periodicity characteristic of a nucleosome positioning element. The presence of +1 nucleosomes in the locations predicted by the PDE was demonstrated by sequencing nucleosome-length fragments protected from the DFF nuclease (Figure 7A, B). Since DFF lacks RNase activity, we could independently verify that +1 nucleosomes are present at the expected locations by nuclear run-off after DFF cleavage (Figure 6F, G). Crucially, use of Pol II as the reporter of nucleosome position demonstrates directly that paused polymerases are in many cases actually in contact with the +1 nucleosome, thereby significantly extending our understanding of the mechanisms governing the earliest stages of transcript elongation. All of our results place a +1 nucleosome at a relatively fixed distance from the TSS (Figure 8A), which would be consistent with a positive interaction between the downstream edge of the PIC and the nucleosome (see also (15)). Earlier work reported such an interaction between TAFs and nucleosomal histones that depended on specific histone modifications (25,26), but those studies did not indicate a preferred rotational specificity and spacing between the nucleosome and TSS that our results demonstrate.

Does the +1 nucleosome play a role in pausing? The average pause location is +41 and the upstream edge of the +1 nucleosome is roughly +42, so the +1 nucleosome is clearly positioned to contribute to pausing in most cases. As just noted, run-off RNA synthesis after DFF cleavage shows that at least half of Pol II pauses in close contact with the +1 nucleosome (Figure 6). Blocking the transition into productive elongation by flavopiridol treatment of cells leads to increases in the fraction of Pol II interacting with the first nucleosome. However, the +1 nucleosome is not substantially displaced downstream for promoters with longer paused RNAs (Supplementary Figure S7D). Pausing preferentially occurs within particular sequences which are distinct from the sequences associated with pausing well downstream of the promoter (Figures 2D and 4D, E). Thus, promoter-proximal pausing is most likely driven by a combination of sequences that favor pausing, the immediate proximity of the +1 nucleosome, and the known factors that antagonize early elongation (Figure 8A).

The core promoter sequence signature is nearly invariant over most of the range of TSS read values. Therefore, the strength of Pol II promoters is not primarily determined by core promoter sequence. The region immediately upstream of the core promoter is typically occupied by complexes which protect <120 bp from DFF cleavage. Importantly, the level of this occupancy is correlated with promoter strength (Supplementary Figure S7B). Regulatory factors associating with this region will orchestrate the loss of nucleosomes from the promoter and recruit the transcriptional machinery. The effectiveness of that recruitment will play a major role in determining promoter strength. Within the nucleosome-free regions, Pol II and the GTFs will search for the best local match(es) to the core promoter signature sequence. For most transcriptionally active regions, more than one usable match may be identified (for example, ACBD3, Figure 1B). Presumably, these matches will be less than optimal for the majority of promoters. Such suboptimal promoter matches should occur far more often

in the genome than we actually detect in functional assays in nuclei. This reinforces the essential role of the upstream factors in clearing away promoter-bound nucleosomes to reveal any local promoter possibilities. Interestingly, late in human cytomegalovirus infection hundreds of HCMV genomes accumulate that are thought to lack nucleosomes. We recently showed that such HCMV templates support much higher levels of transcript initiation per bp than the host genome (39).

Our findings provide a more complete, unified model of the human RNA polymerase II promoter, in which a core promoter signature encompassing the TFIID DNA footprint directs PIC assembly. The possible TSSs for each PIC have a narrow (∼±5 bp) spread (Figure 3E and 8B, top) and thus, all Pol II promoters are essentially focused. It is unnecessary to postulate distinct focused and dispersed promoter classes. Instead, individual focused promoters may be dispersed to varying extents within transcriptionally active regions (Figure 8B, middle). We note that our results are not consistent with the Pol II promoter architecture in yeast (Figure 8B, bottom), in which a single PIC supports TSSs distributed over a broad (up to 100 bp) downstream range (66,67).

DATA AVAILABILITY

All raw sequencing data and BigWig tracks are available (GSE139237).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Justin Ling for initial efforts in cloning DFF subunits for bacterial expression and Craig Kaplan for helpful comments on the manuscript. K.A.N. developed NasCap and generated the NasCap datasets. B.M.S. purified DFF and generated the DFF-Seq data. K.A.N. and D.H.P. conceived of NasCap. D.S.L and D.H.P. spent countless hours over two years extensively discussing and refining the analyses and their significance. M.P. performed all the bioinformatics. D.S.L. wrote the manuscript.

FUNDING

National Institutes of Health [GM113935 to D.H.P., D.S.L., GM35500 and GM126908 to D.H.P., GM121428 to D.S.L.]. Funding for open access charge: National Institutes of Health [GM126908].
Conflict of interest statement. None declared.

REFERENCES

1. Hahn,S. (2004) Structure and mechanism of the RNA polymerase II transcription machinery. *Nat. Struct. Mol. Biol.*, **11**, 394–403.
2. He,Y., Fang,J., Taatjes,D.J. and Nogales,E. (2013) Structural visualization of key steps in human transcription initiation. *Nature*, **495**, 481–486.
3. Gupta,K., Sari-Ak,D., Haffke,M., Trowitzsch,S. and Berger,I. (2016) Zooming in on transcription preinitiation. *J. Mol. Biol.*, **428**, 2581–2591.
4. Louder,R.K., He,Y., Lopez-Blanco,J.R., Fang,J., Chacon,P. and Nogales,E. (2016) Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature*, **531**, 604–609.
5. Schilbach,S., Hantsche,M., Tegunov,D., Dienemann,C., Wigge,C., Urlaub,H. and Cramer,P. (2017) Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature*, **551**, 204–209.
6. Bernecky,C., Plitzko,J.M. and Cramer,P. (2017) Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp. *Nat. Struct. Mol. Biol.*, **24**, 809–815.
7. Patel,A.B., Louder,R.K., Greber,B.J., Grunberg,S., Luo,J., Fang,J., Liu,Y., Ranish,J., Hahn,S. and Nogales,E. (2018) Structure of human TFIID and mechanism of TBP loading onto promoter DNA. *Science*, **362**, eaau8872.
8. Dienemann,C., Schwalb,B., Schilbach,S. and Cramer,P. (2019) Promoter distortion and opening in the RNA polymerase II cleft. *Mol. Cell.*, **73**, 97–106.
9. Vos,S.M., Farnung,L., Urlaub,H. and Cramer,P. (2018) Structure of paused transcription complex Pol II-DSIF-NELF. *Nature*, **560**, 601–606.
10. Vo Ngoc,L., Wang,Y.L., Kassavetis,G.A. and Kadonaga,J.T. (2017) The punctilious RNA polymerase II core promoter. *Genes. Dev.*, **31**, 1289–1301.
11. Juven-Gershon,T. and Kadonaga,J.T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.*, **339**, 225–229.
12. Rach,E.A., Winter,D.R., Benjamin,A.M., Corcoran,D.L., Ni,T., Zhu,J. and Ohler,U. (2011) Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS genetics*, **7**, e1001274.
13. Roy,A.L. and Singer,D.S. (2015) Core promoters in transcription: old problem, new insights. *Trends. Biochem. Sci.*, **40**, 165–171.
14. Vo Ngoc,L., Cassidy,C.J., Huang,C.Y., Duttke,S.H. and Kadonaga,J.T. (2017) The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes. Dev.*, **31**, 6–11.
15. Haberle,V. and Stark,A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell. Biol.*, **19**, 621–637.
16. Muller,F. and Tora,L. (2014) Chromatin and DNA sequences in defining promoters for transcription initiation. *Biochim. Biophys. Acta.*, **1839**, 118–128.
17. Vo Ngoc,L., Kassavetis,G.A. and Kadonaga,J.T. (2019) The RNA polymerase II core promoter in *Drosophila*. *Genetics.*, **212**, 13–24.
18. Kwak,H., Fuda,N.J., Core,L.J. and Lis,J.T. (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, **339**, 950–953.
19. Tome,J.M., Tippens,N.D. and Lis,J.T. (2018) Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.*, **50**, 1533–1541.
20. Core,L. and Adelman,K. (2019) Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes. Dev.*, **33**, 960–982.
21. Weber,C.M., Ramachandran,S. and Henikoff,S. (2014) Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell.*, **53**, 819–830.
22. Jimeno-Gonzalez,S., Ceballos-Chavez,M. and Reyes,J.C. (2015) A positioned +1 nucleosome enhances promoter-proximal pausing. *Nucleic Acids. Res.*, **43**, 3068–3078.
23. Day,D.S., Zhang,B., Stevens,S.M., Ferrari,F., Larschan,E.N., Park,P.J. and Pu,W.T. (2016) Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome. Biol.*, **17**, 120.
24. Ramachandran,S., Ahmad,K. and Henikoff,S. (2017) Transcription and remodeling produce asymmetrically unwrapped nucleosomal intermediates. *Mol. Cell.*, **68**, 1038–1053.
25. Vermeulen,M., Mulder,K.W., Denissov,S., Pijnappel,W.W., van Schaik,F.M., Varier,R.A., Baltissen,M.P., Stunnenberg,H.G., Mann,M. and Timmers,H.T. (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, **131**, 58–69.
26. Lauberth,S.M., Nakayama,T., Wu,X., Ferris,A.L., Tang,Z., Hughes,S.H. and Roeder,R.G. (2013) H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, **152**, 1021–1036.

27. Rhee,H.S. and Pugh,B.F. (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, **483**, 295–301.
28. Mavrich,T.N., Ioshikhes,I.P., Venters,B.J., Jiang,C., Tomsho,L.P., Qi,J., Schuster,S.C., Albert,I. and Pugh,B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome. Res.*, **18**, 1073–1083.
29. Gilchrist,D.A. and Adelman,K. (2012) Coupling polymerase pausing and chromatin landscapes for precise regulation of transcription. *Biochim. Biophys. Acta.*, **1819**, 700–706.
30. Core,L.J., Martins,A.L., Danko,C.G., Waters,C.T., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.
31. Scruggs,B.S., Gilchrist,D.A., Nechaev,S., Muse,G.W., Burkholder,A., Fargo,D.C. and Adelman,K. (2015) Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell.*, **58**, 1101–1112.
32. Chen,Y., Pai,A.A., Herudek,J., Lubas,M., Meola,N., Jarvelin,A.I., Andersson,R., Pelechano,V., Steinmetz,L.M., Jensen,T.H *et al.* (2016) Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat. Genet.*, **48**, 984.
33. Ishii,H., Kadonaga,J.T. and Ren,B. (2015) MPE-seq, a new method for the genome-wide analysis of chromatin structure. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E3457–E3465.
34. Mieczkowski,J., Cook,A., Bowman,S.K., Mueller,B., Alver,B.H., Kundu,S., Deaton,A.M., Urban,J.A., Larschan,E., Park,P.J *et al.* (2016) MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.*, **7**, 11485.
35. Mueller,B., Mieczkowski,J., Kundu,S., Wang,P., Sadreyev,R., Tolstorukov,M.Y. and Kingston,R.E. (2017) Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes. Dev.*, **31**, 451–462.
36. Voong,L.N., Xi,L., Wang,J.P. and Wang,X. (2017) Genome-wide mapping of the nucleosome landscape by micrococcal nuclease and chemical mapping. *Trends Genet.*, **33**, 495–507.
37. Ball,C.B., Nilson,K.A. and Price,D.H. (2019) Use of the nuclear walk-on methodology to determine sites of RNA polymerase II initiation and pausing and quantify nascent RNAs in cells. *Methods*, **159–160**, 165–176.
38. Nilson,K.A., Lawson,C.K., Mullen,N.J., Ball,C.B., Spector,B.M., Meier,J.L. and Price,D.H. (2017) Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. *Nucleic Acids Res.*, **45**, 11088–11105.
39. Parida,M., Nilson,K.A., Li,M., Ball,C.B., Fuchs,H.A., Lawson,C.K., Luse,D.S., Meier,J.L. and Price,D.H. (2019) Nucleotide resolution comparison of transcription of human cytomegalovirus and host genomes reveals universal use of RNA polymerase II elongation control driven by dissimilar core promoter elements. *mBio.*, **10**, e02047–e02018.
40. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome. Biol.*, **10**, R25.
41. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, **26**, 841–842.
42. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome. Res.*, **12**, 996–1006.
43. Xiao,F., Widlak,P. and Garrard,W.T. (2007) Engineered apoptotic nucleases for chromatin research. *Nucleic Acids Res.*, **35**, e93.
44. Liu,X., Li,P., Widlak,P., Zou,H., Luo,X., Garrard,W.T. and Wang,X. (1998) The 40-kDa subunit of DNA fragmentation factor induces DNA fragmentation and chromatin condensation during apoptosis. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 8461–8466.
45. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome. Res.*, **14**, 1188–1190.
46. Nakatani,Y., Horikoshi,M., Brenner,M., Yamamoto,T., Besnard,F., Roeder,R.G. and Freese,E. (1990) A downstream initiation element required for efficient TATA box binding and *in vitro* function of TFIID. *Nature*, **348**, 86–88.
47. Purnell,B.A., Emanuel,P.A. and Gilmour,D.S. (1994) TFIID sequence recognition of the initiator and sequences farther downstream in *Drosophila* class II genes. *Genes. Dev.*, **8**, 830–842.
48. Parry,T.J., Theisen,J.W., Hsu,J.Y., Wang,Y.L., Corcoran,D.L., Eustice,M., Ohler,U. and Kadonaga,J.T. (2010) The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes. Dev.*, **24**, 2013–2018.
49. Wang,Y.L., Duttke,S.H., Chen,K., Johnston,J., Kassavetis,G.A., Zeitlinger,J. and Kadonaga,J.T. (2014) TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes. Dev.*, **28**, 1550–1555.
50. Baumann,D.G. and Gilmour,D.S. (2017) A sequence-specific core promoter-binding transcription factor recruits TRF2 to coordinately transcribe ribosomal protein genes. *Nucleic Acids Res.*, **45**, 10481–10491.
51. Kugel,J.F. and Goodrich,J.A. (2017) Finding the start site: redefining the human initiator element. *Genes. Dev.*, **31**, 1–2.
52. Tora,L. and Timmers,H.T. (2010) The TATA box regulates TATA-binding protein (TBP) dynamics in vivo. *Trends. Biochem. Sci.*, **35**, 309–314.
53. Traut,T.W. (1994) Physiological concentrations of purines and pyrimidines. *Mol. Cell. Biochem.*, **140**, 1–22.
54. Gaffney,D.J., McVicker,G., Pai,A.A., Fondufe-Mittendorf,Y.N., Lewellen,N., Michelini,K., Widom,J., Gilad,Y. and Pritchard,J.K. (2012) Controls of nucleosome positioning in the human genome. *PLoS Genet.*, **8**, e1003036.
55. Dreos,R., Ambrosini,G. and Bucher,P. (2016) Influence of rotational nucleosome positioning on transcription start site selection in animal promoters. *PLoS Comput. Biol.*, **12**, e1005144.
56. Darzacq,X., Shav-Tal,Y., de Turris,V., Brody,Y., Shenoy,S.M., Phair,R.D. and Singer,R.H. (2007) In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.*, **14**, 796–806.
57. Steurer,B., Janssens,R.C., Geverts,B., Geijer,M.E., Wienholz,F., Theil,A.F., Chang,J., Dealy,S., Pothof,J., van Cappellen,W.A *et al.* (2018) Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E4368–E4376.
58. Widlak,P., Li,P., Wang,X. and Garrard,W.T. (2000) Cleavage preferences of the apoptotic endonuclease DFF40 (caspase-activated DNase or nuclease) on naked DNA and chromatin substrates. *J. Biol. Chem.*, **275**, 8226–8232.
59. Bhuiyan,T. and Timmers,H.T.M. (2019) Promoter recognition: putting TFIID on the spot. *Trends Cell. Biol.*, **29**, 752–763.
60. Fant,C.B., Levandowski,C.B., Gupta,K., Maas,Z.L., Moir,J., Rubin,J.D., Sawyer,A., Esbin,M.N., Rimel,J.K., Luyties,O *et al.*2020) TFIID enables RNA polymerase II promoter-proximal pausing. *Mol. Cell.*, **78**, 785–793.
61. Holstege,F.C., van der Vliet,P.C. and Timmers,H.T. (1996) Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. *EMBO. J.*, **15**, 1666–1677.
62. Larson,M.H., Mooney,R.A., Peters,J.M., Windgassen,T., Nayak,D., Gross,C.A., Block,S.M., Greenleaf,W.J., Landick,R. and Weissman,J.S. (2014) A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, **344**, 1042–1047.
63. Mullen,N.J. and Price,D.H. (2017) Hydrogen peroxide yields mechanistic insights into human mRNA capping enzyme function. *PLoS One*, **12**, e0186423.
64. Hendrix,D.A., Hong,J.W., Zeitlinger,J., Rokhsar,D.S. and Levine,M.S. (2008) Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 7762–7767.
65. Shao,W., Alcantara,S.G. and Zeitlinger,J. (2019) Reporter-ChIP-nexus reveals strong contribution of the Drosophila initiator sequence to RNA polymerase pausing. *Elife*, **8**, e41461.
66. Murakami,K., Mattei,P.J., Davis,R.E., Jin,H., Kaplan,C.D. and Kornberg,R.D. (2015) Uncoupling promoter opening from start-site scanning. *Mol. Cell.*, **59**, 133–138.
67. Fazal,F.M., Meng,C.A., Murakami,K., Kornberg,R.D. and Block,S.M. (2015) Real-time observation of the initiation of RNA polymerase II transcription. *Nature*, **525**, 274–277.