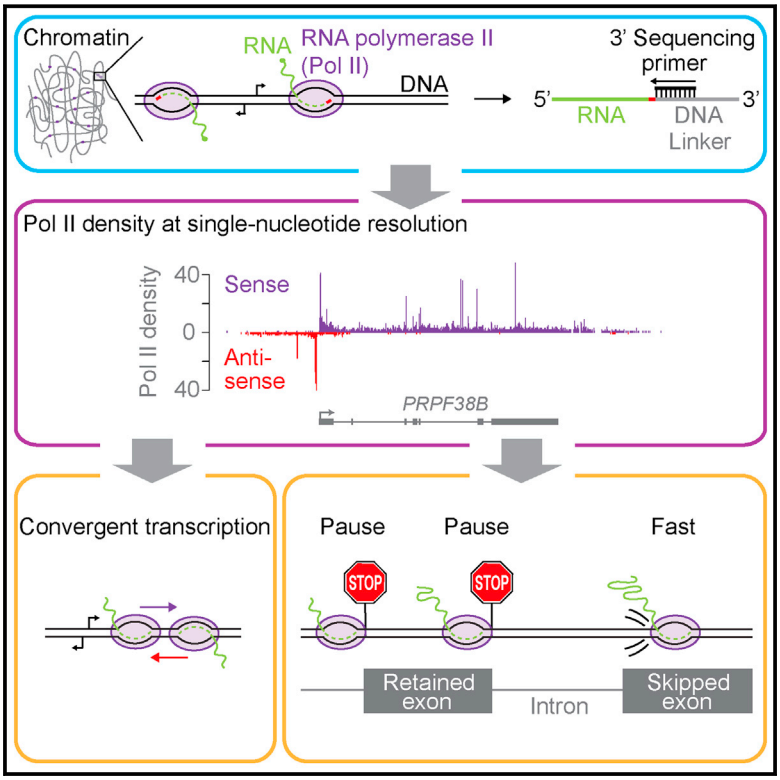
Article

Cell

Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution

Graphical Abstract



Authors

Andreas Mayer, Julia di Iulio, ..., John A. Stamatoyannopoulos, L. Stirling Churchman

Correspondence

churchman@genetics.med.harvard.edu

In Brief

Native elongating transcript sequencing in human uncovers convergent sense and antisense transcription within promoter-proximal regions of lower-expressed genes and unravels details about Pol II pausing.

Highlights

- Human NET-seq maps global RNA polymerase II (Pol II) density at high resolution
- Widespread convergent transcription occurs near promoters of lower-expressed genes
- Strong Pol II pausing at sites of occupied transcription factors, including YY1 and CTCF
- NET-seq reveals pronounced Pol II pausing at the boundaries of retained exons

Accession Numbers

GSE61332

Mayer et al., 2015, Cell *161*, 541–554
April 23, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.03.010

Article

Cell

Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution

Andreas Mayer,[1,4] Julia di Iulio,[1,4] Seth Maleri,[1] Umut Eser,[1] Jeff Vierstra,[2] Alex Reynolds,[2] Richard Sandstrom,[2] John A. Stamatoyannopoulos,[2,3] and L. Stirling Churchman[1,*]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[2]Department of Genome Sciences, University of Washington, Seattle, WA 98104, USA

[3]Department of Medicine, Division of Oncology, University of Washington, Seattle, WA 98104, USA

[4]Co-first author

*Correspondence: churchman@genetics.med.harvard.edu

http://dx.doi.org/10.1016/j.cell.2015.03.010

SUMMARY

Major features of transcription by human RNA polymerase II (Pol II) remain poorly defined due to a lack of quantitative approaches for visualizing Pol II progress at nucleotide resolution. We developed a simple and powerful approach for performing native elongating transcript sequencing (NET-seq) in human cells that globally maps strand-specific Pol II density at nucleotide resolution. NET-seq exposes a mode of antisense transcription that originates downstream and converges on transcription from the canonical promoter. Convergent transcription is associated with a distinctive chromatin configuration and is characteristic of lower-expressed genes. Integration of NET-seq with genomic footprinting data reveals stereotypic Pol II pausing coincident with transcription factor occupancy. Finally, exons retained in mature transcripts display Pol II pausing signatures that differ markedly from skipped exons, indicating an intrinsic capacity for Pol II to recognize exons with different processing fates. Together, human NET-seq exposes the topography and regulatory complexity of human gene expression.

INTRODUCTION

High-throughput sequencing analyses of transcription have discovered new classes of RNAs and new levels of regulatory complexity. Many of these results were obtained with two experimental strategies to measure RNA polymerase density genome wide. The first, RNA polymerase II (Pol II) ChIP-seq or ChIP-chip, identifies DNA bound to RNA polymerase. The second set of approaches, global run-on sequencing (GRO-seq) and precision nuclear run-on and sequencing (PRO-seq), restarts RNA polymerase in vitro with labeled nucleotides to purify and sequence nascent RNA (Core et al., 2008; Kwak et al., 2013). GRO-seq and Pol II ChIP detect strong transcriptional pauses ~50 bp downstream of many transcription start sites, demonstrating that promoter-proximal pausing is more prevalent than initially observed (Core et al., 2008; Krumm et al., 1992; Kwak et al., 2013; Muse et al., 2007; Rahl et al., 2010; Rougvie and Lis, 1988; Strobl and Eick, 1992; Zeitlinger et al., 2007). Abundant unstable transcripts upstream of and antisense to promoters revealed that divergent transcription is a common feature of eukaryotic promoters (Core et al., 2008; Neil et al., 2009; Preker et al., 2008; Seila et al., 2008; Xu et al., 2009). Despite progress in understanding how these transcripts are terminated and degraded (Almada et al., 2013; Ntini et al., 2013; Preker et al., 2008; Schulz et al., 2013), their roles remain unknown (Wu and Sharp, 2013). Finally, recent studies confirm that splicing is largely co-transcriptional and splicing outcome is kinetically tied to elongation rate (Bhatt et al., 2012; Davis-Turak et al., 2015; Dujardin et al., 2014; Fong et al., 2014; Ip et al., 2011; de la Mata et al., 2003; Roberts et al., 1998; Shukla et al., 2011; Tilgner et al., 2012). However, it has been impossible to determine whether such kinetic coupling in human cells is mediated by pausing events genome wide, due to the high resolution required to measure pausing on short human exons.

The strongly stereotyped locations of promoter-proximal pauses and divergent antisense transcription can be exposed by averaging Pol II density from many genes (metagene analysis) obtained at low resolution (Core et al., 2008; Neil et al., 2009; Preker et al., 2008; Rahl et al., 2010; Seila et al., 2008; Xu et al., 2009). Yet, the precise architecture of promoter-associated transcriptional activity and of pausing outside of promoter regions has been obscured by the resolution limitations of current methodologies, preventing deeper insight into the underlying regulatory mechanisms. Indeed, the interplay between chromatin structure, transcription factors, and the transcription machinery is largely undefined. Pol II ChIP-seq is typically limited in its resolution to >200 bp resolution and lacks strand specificity. GRO-seq is similarly limited to ~50 bp resolution, and although PRO-seq has higher resolution, both run-on methods require transcription elongation complexes to resume polymerization in vitro, a variable process sensitive to the experimental conditions and the Pol II pausing state (Core et al., 2008; Weber et al., 2014). Recently, we showed that

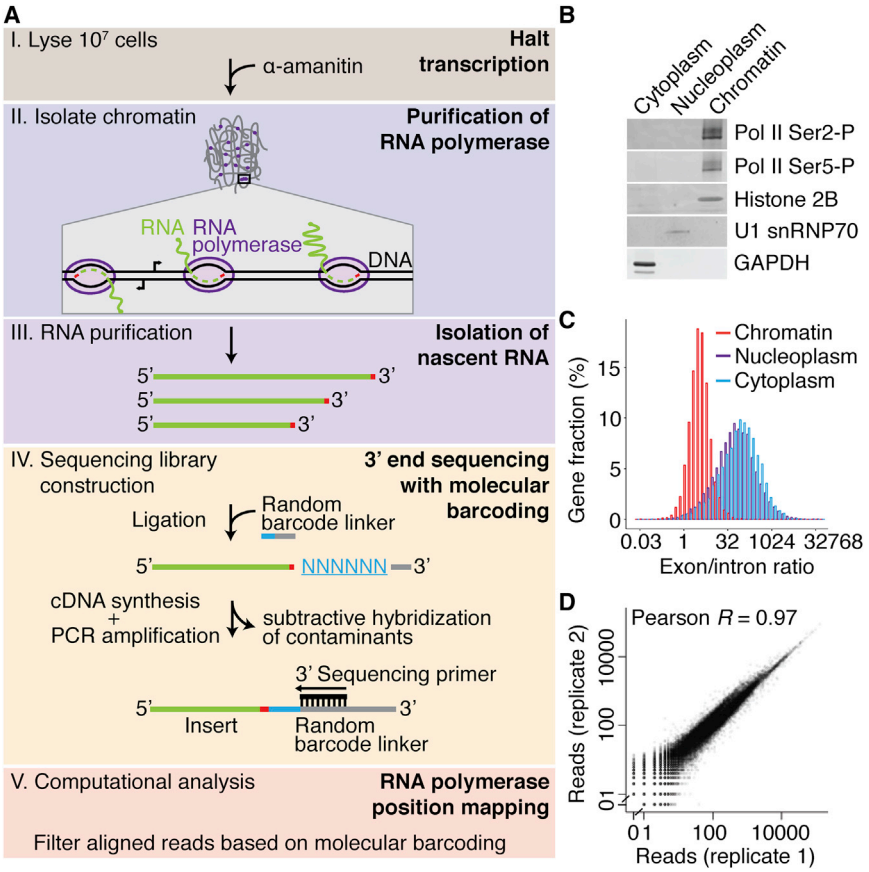


Figure 1. A Robust and Simplified NET-Seq Approach for Human Cells

(A) Schematic view of the key steps of the human NET-seq approach. The transcription inhibitor, α-amanitin, is introduced at cell lysis and is maintained through all purification steps. Engaged RNA polymerase is purified through the isolation of chromatin. The 3′ end of the co-purified nascent RNA (red) is ligated to a linker containing a mixed random hexameric sequence (blue) that serves as a molecular barcode. After cDNA synthesis, contaminant species are removed by hybridization. PCR amplification results in a DNA-sequencing library with the sequencing primer binding site proximal to the random hexamer barcode. Finally, the 3′ ends of the sequenced nascent RNA are aligned to the human genome, yielding RNA polymerase density at nucleotide resolution. Analysis of the molecular barcode allows reads arising from DNA library construction artifacts to be filtered out.

(B) Representative western blot analysis of cellular fractions in HeLa S3 cells. Subcellular localization markers were also probed (chromatin marker, histone 2B; nucleoplasm marker, U1 snRNP70; cytoplasm marker, GAPDH).

(C) Histograms of the size-normalized ratio of subcellular RNA-seq reads that map to exons versus introns for each gene.

(D) Number of uniquely aligned reads per Pol II gene for two biological replicates from HeLa S3 cells (Pearson's correlation, R = 0.97). 0.5 pseudocounts were added to genes with zero counts in one of the replicates. The data set with higher coverage was randomly downsampled to match the total number of reads of the other data set.

See also Figure S1 and Table S1.

the extraordinary stability of the RNA-DNA-RNA polymerase ternary complex can be exploited to capture nascent RNA (Churchman and Weissman, 2011). Native elongating transcript sequencing (NET-seq) quantitatively purifies Pol II complexes and sequences the 3′ end of nascent RNA to reveal the strand-specific position of Pol II with single-nucleotide resolution. NET-seq detects all transcriptionally engaged Pol II, including productively transcribing Pol II, paused Pol II, and Pol II recovering from pausing (Churchman and Weissman, 2011).

Here, we develop a NET-seq approach that quantitatively defines the full spectrum of transcriptional activity in a strand-specific manner and at nucleotide resolution in human cells. We find that many promoters display antisense transcription downstream of a promoter-proximal pause, resulting in convergent sense and antisense transcriptional activities that face one another in close proximity. Convergent transcription is associated with a distinct chromatin conformation and is a feature of lower-expressed genes, suggesting a possible regulatory role. NET-seq reveals that Pol II density profiles differ between retained exons, skipped exons, and introns in human cells, indicating generalized kinetic coupling of transcription and splicing. Human NET-seq is readily applicable to diverse cell types and provides a general strategy to study transcriptional complexity.

RESULTS

A Robust Human NET-Seq Methodology

The first step of NET-seq purifies nascent RNA through its tight interaction with RNA polymerase. In yeast, this is achieved through an epitope-tagged Pol II subunit that enables highly quantitative purification and specific elution (Churchman and Weissman, 2011). In adapting NET-seq to human cells, we biochemically purify >99% of all engaged RNA polymerase in a highly specific manner that can be applied to any mammalian cell line or tissue without genetic modification (Figure 1A). This method avoids using Pol II antisera, which could bias the population of isolated Pol II complexes due to posttranslational modifications and epitope masking by heterogeneous Pol II binding partners and structural conformations. Instead, human NET-seq exploits the high stability of the RNA-DNA-RNA polymerase ternary complex, even in the presence of high salt and urea (Cai and Luse, 1987; Wuarin and Schibler, 1994), to purify engaged RNA polymerase, along with its nascent RNA, through an association with chromatin after cellular fractionation into cytoplasm, nucleoplasm, and chromatin (Bhatt et al., 2012; Pandya-Jones and Black, 2009; Wuarin and Schibler, 1994). To prevent transcriptional run-on during fractionation, lysate is kept at ≤4°C, and α-amanitin, a potent transcriptional inhibitor (Lindell et al., 1970), is included in every step. Through optimization of current

fractionation approaches, we identified buffers and washing conditions that cleanly purify >99% of elongating RNA polymerase II (C-terminal domain [CTD] Ser2-P, Ser5-P, and the general CTD hyper-phosphorylated form of Pol II) in the chromatin fraction (see Experimental Procedures and Figures 1B and S1A). Western blot analyses of Pol II isoforms and factors with well-defined subcellular localizations verify the stringency of our fractionation conditions (Figure 1B).

To confirm that our purification strategy specifically isolates nascent RNA, we sequenced the RNA in each fraction. Unprocessed RNA species, such as intron-containing Pol II transcripts and spacer-containing Pol I transcripts (Figures 1C and S1B), are heavily enriched in the chromatin fraction. Importantly, the large majority of intron-containing RNAs observed in the nucleoplasm persist to the cytoplasm, indicating that these RNAs are products of intron-retaining alternative splicing and not nascent transcripts (Figure S1C). Together, these results demonstrate that RNA polymerase and nascent RNA are quantitatively purified through the isolation of chromatin.

The second step of the NET-seq approach requires sequencing the 3′ ends of the nascent RNA, which localizes Pol II genome wide at nucleotide resolution (Churchman and Weissman, 2011; Ferrari et al., 2013; Weber et al., 2014). In large part, our yeast library construction protocol is used (Churchman and Weissman, 2012), with two important changes to account for the increased complexity of the human genome. First, we addressed reverse transcription (RT) artifacts that arise from the significant size of human nascent RNA. We found that reverse transcription frequently initiates within the RNA if there are stretches as short as six nucleotides of complementarity to the RT primer (Figure S1D). When the 3′ ends of the nascent RNA are ligated to a linker pool, consisting of a random hexamer at the 5′ end followed by a common sequence, ligation efficiency increases and mispriming events are dramatically reduced (Figure S1D). Furthermore, the hexamer serves as a molecular barcode for each molecule and enables the computational removal of reads arising from residual mispriming events and PCR duplicates. Second, we deplete abundant mature snRNAs, snoRNAs, rRNA, and others through subtractive hybridization targeting their 3′ ends (Figure S1E, Table S1) to increase sequencing coverage for nascent transcripts. Finally, library construction steps are optimized to be highly efficient (>90%) and are continually monitored through quality controls to minimize bias. Together, our optimized library construction protocol faithfully converts the 3′ ends of nascent human RNA to a DNA sequencing library that allows the high-fidelity mapping of strand-specific Pol II density.

To observe genome-wide transcriptional activity, a NET-seq library was prepared from HeLa S3 cells and sequenced to high coverage (768 million total reads, 360 million uniquely aligned). Each sequencing read was aligned to the human genome, and the genomic location of the 3′ end of the nascent RNA was recorded to map RNA polymerase density with nucleotide resolution. As expected, we recovered nascent RNA from all three nuclear RNA polymerases (Pol I, Pol II, Pol III), as well as mature chromatin-associated RNAs, such as snRNAs, and splicing intermediates (Figures S1F and S1G). Here, we focused our analysis on Pol-II-synthesized RNAs, but our results suggest that the NET-seq approach is amenable to the study of other RNA polymerases. Importantly, comparison of a biological replicate library (175 million total reads, 83 million uniquely aligned) shows strong agreement, indicating the robustness of the approach (Pearson's coefficient, 0.97, Figure 1D). To demonstrate that NET-seq is easily adaptable to other cell lines, we applied our approach to HEK293T cells and obtained data from two replicates with similar reproducibility (replicate 1: 1.203 billion total reads, 555 million uniquely aligned; replicate 2: 358 million total reads, 135 million uniquely aligned; Figure S1H). From these analyses, we conclude that human NET-seq is capable of quantitatively monitoring transcriptional activity across the human genome and adaptation to new cell lines is straightforward.

NET-Seq Reveals Transcriptional Activity at Nucleotide Resolution Genome Wide

The resolution afforded by NET-seq and the sequencing coverage obtained provide an in-depth view of genome-wide transcriptional activity. The highest coverage is observed across promoter-proximal regions, which we conservatively defined as the region between the earliest annotated transcription start site and +1 kb. Within this region, >50% of genes have coverage of >1 read per kb per million uniquely aligned reads (RPKM) in both HeLa S3 (Figures 2A and 2B) and HEK293T cells (Figure S2C). When coverage is calculated across entire genes, the percentage decreases to <30% in both cell lines due to the prevalence of promoter-proximal pausing (Figures 2A and 2B). Indeed, most (89% in HeLa S3 cells and 94% in HEK293T cells) expressed genes display promoter-proximal pausing defined by a traveling ratio (coverage ratio between a narrow promoter-proximal region and the gene body) of ≥ 2, consistent with earlier observations in mouse embryonic stem cells (Figures 2C and S2D) (Rahl et al., 2010). Furthermore, we detect unstable RNA production, antisense transcription upstream of many promoters (89% in HeLa S3 cells and 82% in HEK293T cells), transcription downstream of many polyadenylation sites (95% in HeLa S3 cells and 88% in HEK293T cells), and enhancer RNAs (Figures S2A, S2B, S2E,and S2F).

NET-seq data describe transcriptional activity at many length scales. At the single-gene level, strong signal is observed at promoter regions and across introns (Figure 2D, top and middle). Signal variation across the gene body suggests that transcription elongation is discontinuous following release from promoter-proximal regions and that pausing is a general feature during productive Pol II transcription. Near transcription start sites (TSSs), NET-seq detects sense and antisense transcription of divergent promoters at single-nucleotide resolution, revealing that promoter-proximal pausing does not occur at only one position; instead, there are narrow regions of high Pol II density (Figure 2D, bottom). Together, NET-seq data uncover key features of human transcription activity, and the high resolution and the coverage of the data provide deeper insight into these complexities.

Widespread Convergent Transcription in Promoter-Proximal Regions

Several previous studies showed widespread divergent transcription at eukaryotic promoters (Churchman and Weissman,

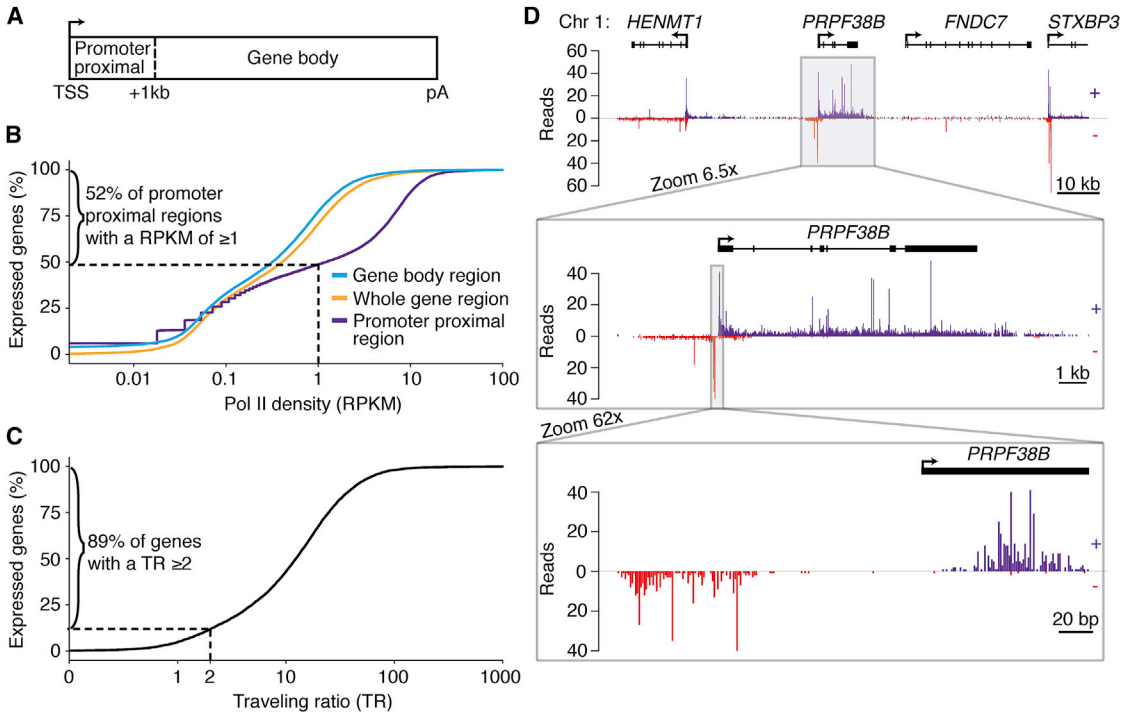


Figure 2. NET-Seq Reports on Transcription Globally and Locally

(A) Schematic defining gene regions used in analysis of NET-seq data.

(B) Distributions of the percent of expressed Pol-II-transcribed protein-coding genes (n = 19,108), with a given Pol II coverage for different gene regions as defined in Figure 2A.

(C) Distributions of the percent of well-expressed Pol II protein-coding genes (n = 8,912) with a given traveling ratio. Well-expressed Pol II genes are defined as those genes with an RPKM of 1 or greater in a tight promoter-proximal region (−30 bp to +300 bp of the TSS). Traveling ratio (TR) is defined as the RPKM of the tight promoter-proximal region divided by the RPKM of the gene body region.

(D) Number of NET-seq reads at three zoom levels around the *PRPF38B* locus for HeLa S3 cells. Reads that aligned to the positive strand (+) are in violet, and reads that aligned to the negative strand (−) are in red. The TSS and the direction of transcription are indicated by an arrow. Annotation of exonic and intronic regions is shown as boxes and lines, respectively.

RPKM, reads per kb per million uniquely aligned reads at Pol-II-transcribed genes; TSS, transcription start site; pA, polyadenylation site. See also Figure S2.

2011; Core et al., 2008; Neil et al., 2009; Seila et al., 2008; Xu et al., 2009). We analyzed this phenomenon for a stringently defined set of Pol-II-transcribed genes that do not overlap other genes within 2.5 kb of the TSS and polyadenylation site and are longer than 2 kb to avoid misinterpreting transcription from other genes as antisense transcription (n = 3,937 genes). Analysis of regions 2 kb upstream and downstream of transcription start sites with broad coverage and no sign of missing overlapping annotation (n = 1,488; see Experimental Procedure) reveals divergent transcription in 77% of promoter-proximal regions, consistent with other studies (Figure 3A, left) (Core et al., 2008; Seila et al., 2008). Surprisingly, close inspection of our data revealed an unappreciated form of antisense transcription near promoters. At 25% of promoter-proximal regions, we observe antisense transcription originating downstream of sense transcription (Figure 3A, right), which we term convergent transcription. Convergent transcription is clearly observed at single-promoter regions (Figures 3B and 3C), and in most cases, such as near the *KLHL9* promoter, convergent transcription is accompanied by divergent transcription (Figure 3B). However, it also occurs in the absence of divergent transcription (for example, *FAM133B*, Figure 3C). Furthermore, GRO-seq also detects these transcripts. A re-analysis of mouse embryonic stem cell data reveals convergent antisense transcription (Jonkers et al., 2014) (Figure S3A).

To characterize the structural attributes on these modes of transcriptional activity, distances between sense and antisense peaks were determined for each promoter-proximal region (Figure 3D). A stereotypical distance (250 ± 50 bp) separates the sense and antisense peaks in divergent transcription, while the sense and antisense peaks in convergent transcription are also separated by a stereotypical distance (150 bp ± 50 bp), indicating that convergent antisense transcription is not simply the result of spurious antisense transcription initiation events across the promoter-proximal region (Figure 3D).

A Distinct Chromatin Structure Associated with Convergent Transcription

Many chromatin modifiers control antisense transcription (Churchman and Weissman, 2011; DeGennaro et al., 2013; Kim et al., 2012; Marquardt et al., 2014; Whitehouse et al., 2007), and we asked how promoter-proximal transcriptional

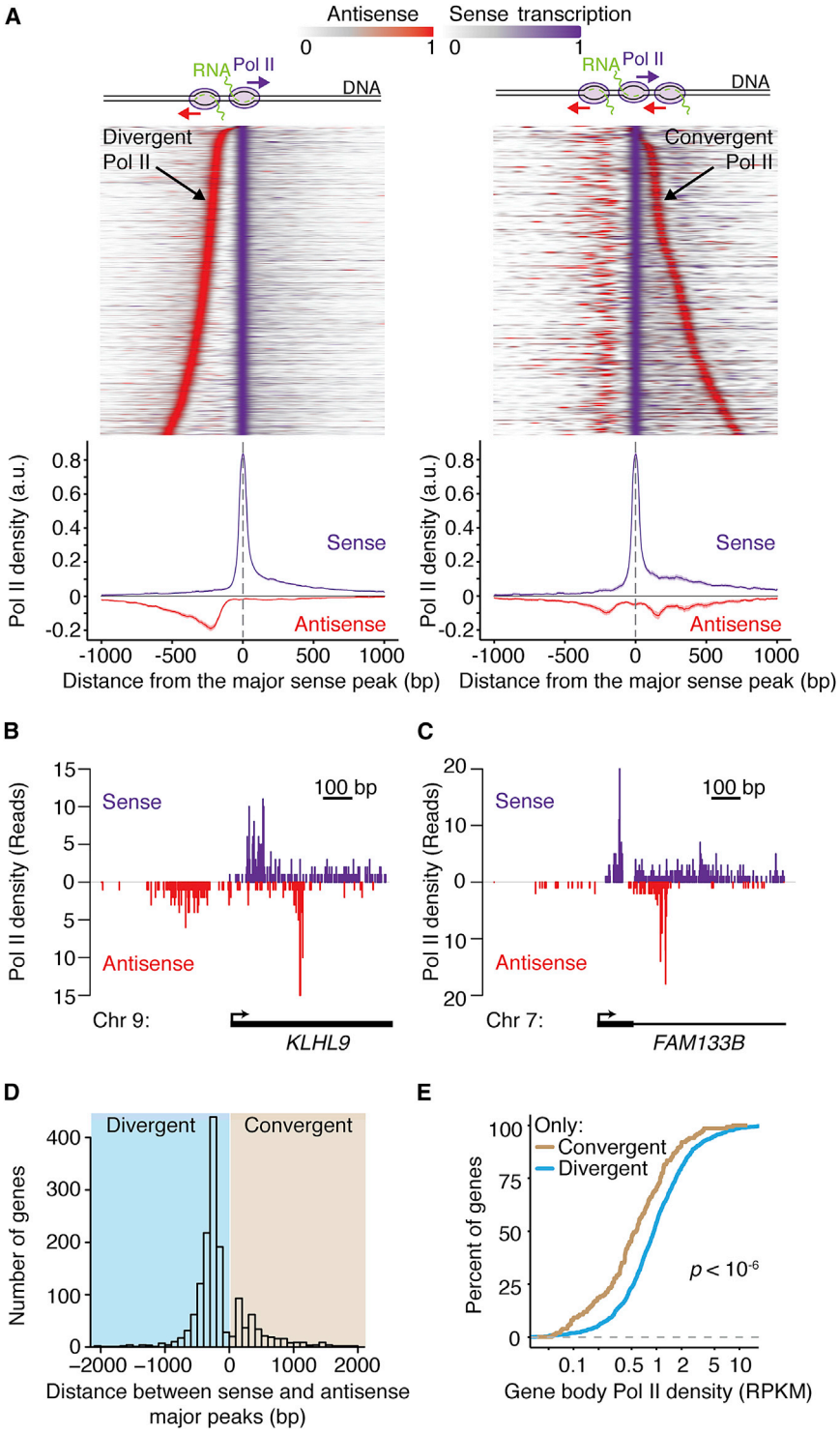


Figure 3. Convergent Transcription Observed at the Promoter-Proximal Regions of Lower-Expressed Genes

(A) Promoters are classified depending on whether they contain a peak of convergent antisense Pol II transcription, as illustrated by the cartoon above the heat maps. A stringent set of promoter-proximal regions was selected for analysis to ensure that transcription arising from other transcription units would not bias classification (see Experimental Procedures). Heatmaps of Pol II scaled density are displayed for each class (left, no convergent peak, n = 931 genes; right, convergent peak, n = 373). For each gene, the sense (violet) and antisense (red) raw signal is analyzed separately and normalized to vary from 0 to 1. Both signals are superposed, centered on the major sense transcription peak. Genes are sorted by the distance between the sense and antisense peaks. Mean Pol II density profile is displayed below the heatmaps, where raw sense and antisense data are normalized together to vary between 0 and 1 and smoothed with a 50 bp sliding window average. Solid lines indicate the mean values, and shading shows the 95% confidence interval. Sense transcription is shown in violet, and antisense transcription is shown in red.

(B and C) Examples of NET-seq reads in two promoter-proximal regions that display convergent Pol II transcription.

(D) Histogram of distances between the major peak of Pol II density in the sense direction and the peak(s) in the antisense direction for all promoters with convergent and/or divergent peaks (n = 1,304).

(E) Distributions of the percentage of genes with a given Pol II density in the gene body region, as defined in Figure 2A. Genes with only convergent transcription (yellow, n = 151) or only divergent transcription (blue, n = 931) in their promoter-proximal regions are compared. The p value is calculated by the Kolmogorov-Smirnov test.

RPKM, reads per kb per million uniquely aligned reads at Pol-II-transcribed genes. See also Figure S3.

activity relates to local chromatin structure. We used DNase-seq to map regions of open chromatin and highly positioned nucleosomes in the same HeLa S3 cells used for NET-seq (Thurman et al., 2012). We examined the distribution of DNase I accessibility relative to promoter-proximal peaks in NET-seq data (Figure 4). At genes that have a sense Pol II peak (representing promoter-proximally paused Pol II), we observe strong DNase I hypersensitivity upstream of the peak, determining the canonical promoter (Figure 4A), and reduced DNase sensitivity downstream of the peak corresponding to the +1 nucleosome. Thus, promoter-proximal pausing occurs prior to the +1 nucleosome in mammalian cells. Comparison of DNase I data relative to the divergent antisense peak shows that this transcriptional activity originates from the 5′ side of the promoter hypersensitivity region, consistent with the model that divergent antisense transcription is a consequence of an open chromatin region (Seila et al., 2009) (Figure 4C). In contrast, genes with

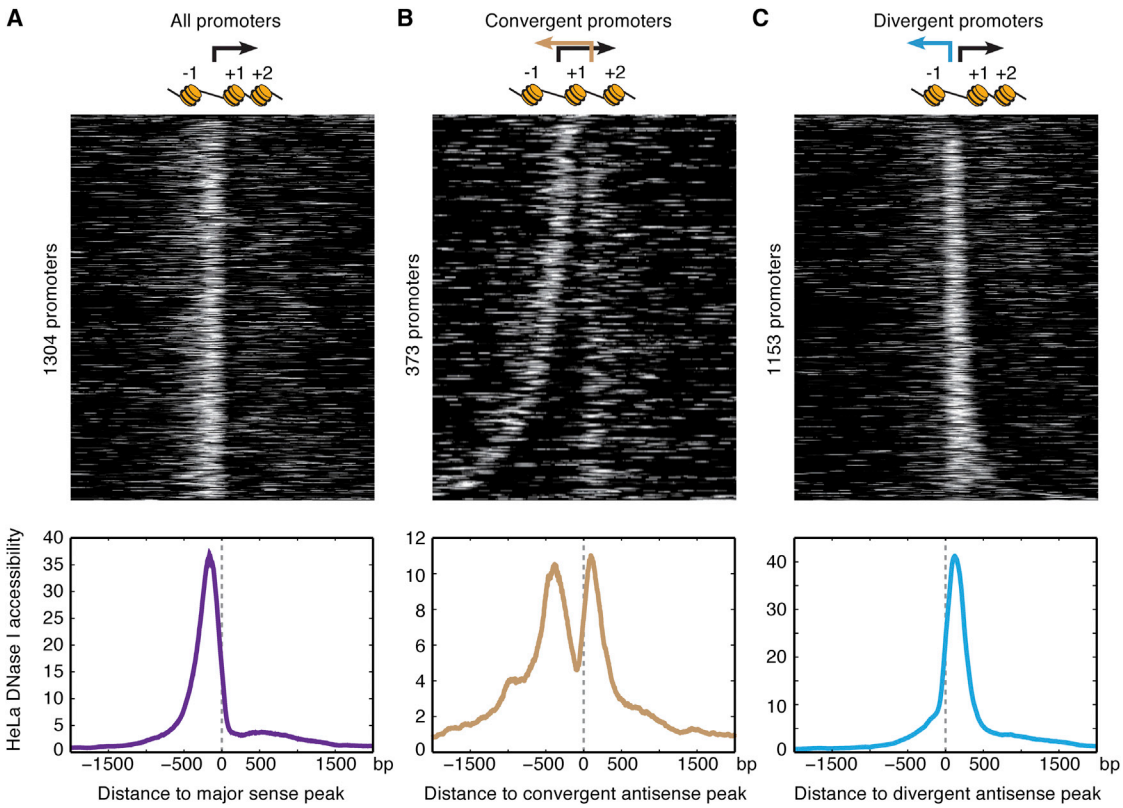


Figure 4. Convergent Transcription Is Associated with a Distinct Chromatin Structure

(A–C) Heatmaps showing DNase I accessibility in HeLa S3 cells surrounding all (A) promoters (n = 1,304) aligned to the sense NET-seq peak, (B) promoters that have convergent transcription (n = 373) aligned to the antisense convergent NET-seq peak, and (C) promoters that have divergent transcription (n = 1,153) aligned to the antisense divergent peak. Below each heatmap is the mean DNase I accessibility profile of the region shown in the heatmap. The raw data are smoothed with a 150 bp sliding window in 20 bp steps. Solid lines indicate the trimmed mean (removing 5% of extreme data points). Above each heatmap are arrows showing the transcriptional activity observed in each promoter-proximal region. A cartoon displays the chromatin structure determined by analysis of the DNase-seq data.

convergent transcription show two distinct peaks in DNase I hypersensitivity: a canonical promoter peak and a downstream peak located proximal to the convergent antisense peak (Figure 4B). Thus, convergent antisense transcription likely originates locally. Furthermore, the dip between the two peaks of DNase I hypersensitivity likely represents the +1 nucleosome, consistent with the ~150 bp spacing between the sense and convergent antisense Pol II peaks (Figures 3D and 4B). These results indicate that convergent transcription reflects sense and antisense transcription that initiates locally and undergoes promoter-proximal pausing flanking the +1 nucleosome.

Convergent Transcription Is a Feature of Lower-Expressed Genes

Convergent transcription can regulate gene expression through transcriptional interference mechanisms (Callen et al., 2004; Eldledge and Davis, 1989; Gullerova and Proudfoot, 2012; Hobson et al., 2012; Martens et al., 2004; Prescott and Proudfoot, 2002; Shearwin et al., 2005). Thus, we considered whether promoter-proximal convergent transcription may be involved in release of Pol II from promoter-proximal pausing into productive elongation. We compared Pol II density within the gene body (+1 kb after the transcription start site to the polyadenylation site, illustrated in Figure 2A) at genes that display only convergent transcription to genes that display only divergent transcription. Notably, genes with only convergent transcription near their promoters show consistently less transcription downstream of their promoter regions (Figure 3E) (1.8-fold less on average, Kolmogorov–Smirnov test, $p < 10^{-6}$). Comparison of less stringently defined sets of genes, such as all genes with convergent transcription to all genes without convergent transcription, showed a similar effect (Figure S3B). In agreement with this observation, analysis of ENCODE HeLa S3 ChIP-seq data reveals that H3K79me2 histone marks, which correlate with transcription elongation, occur at significantly lower levels in the gene bodies of genes with convergent antisense transcription (Figure S3C) (Consortium, 2012; Guenther et al., 2007; Wozniak and Strahl, 2014). Thus, convergent antisense transcription could interfere with productive transcription elongation or could be a consequence of less-productive elongation. Either of these possibilities could be directly mediated by Pol II or by another factor, such as chromatin.

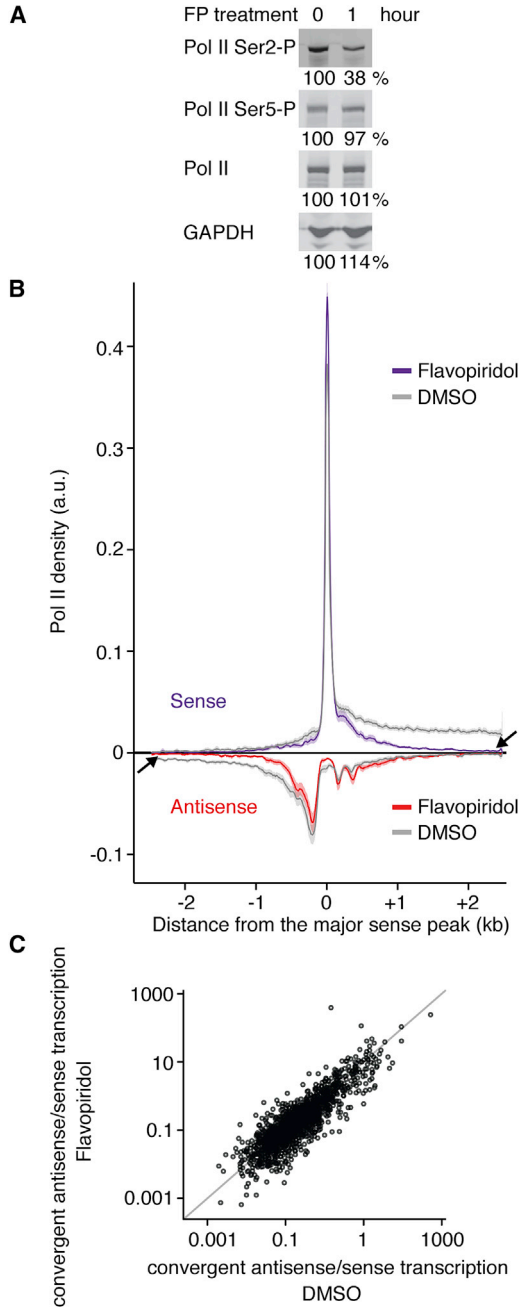


Figure 5. P-TEFb Inhibition Proportionally Affects Levels of Sense Transcription and Convergent Antisense Transcription

(A) Western blot analysis of whole-cell extract of HeLa S3 cells with flavopiridol (FP) treatment (1 hr). The percentage at the bottom of each lane is the amount of the respective protein (as determined by image quantification) before and after FP treatment. GAPDH serves as a loading control.

(B) Meta-gene analysis of NET-seq data from HeLa S3 cells treated with 1 μM FP (purple and red) or DMSO control (gray) for 1 hr. Arrows indicate regions where transcription is affected by FP treatment. Genes that had convergent and/or divergent peaks (described in Experimental Procedures) in both data sets are included in the analysis (n = 615). NET-seq signal from each promoter region (± 2.5 kb centered at the sense transcription peak) are binned into 10 bp windows. For each sample, sense and antisense signal are normalized together to vary between 0 and 1 and then smoothed with a 50 bp sliding window. Solid lines indicate the mean normalized Pol II density, and shading shows the 95% confidence interval.

(C) A scatterplot comparing the convergent-to-sense ratio after treatment with DMSO (control) and after FP treatment for a stringent subset of non-overlapping genes with at least 10 reads across the 500 bp region after TSS in both samples (n = 1,667). The ratio is the sum of NET-seq signal on the antisense strand versus the sense strand across the 500 bp region after the TSS. The handful of genes with a ratio of 0 are not plotted.

TSS, transcription start site. See also Figure S4.

To test whether convergent antisense transcription is a consequence of reduced sense transcription elongation, we globally suppressed productive elongation by inhibiting positive transcription elongation factor b (P-TEFb). Most promoter-proximally paused Pol II are released through recruitment of P-TEFb that phosphorylates multiple proteins, including Ser2 residues of the Pol II CTD (Kim and Sharp, 2001; Peterlin and Price, 2006). Therefore, active P-TEFb greatly facilitates the transition to productive elongation but does not affect transcription initiation (Lis et al., 2000; Peterlin and Price, 2006; Rahl et al., 2010). We performed NET-seq analysis on HeLa S3 cells exposed to the P-TEFb inhibitor flavopiridol (FP) (Chao and Price, 2001) or DMSO alone. As expected, after 60 min, FP reduced Pol II CTD Ser2 phosphorylation, but phosphorylation of Ser5 residues and overall Pol II levels remained unchanged (Figures 5A and S4). We generated NET-seq libraries from HeLa S3 cells after a 1 hr FP treatment or DMSO control (FP treatment NET-seq data set, 486 million total reads, 262 uniquely mapped reads; DMSO control NET-seq data set, 491 million total reads, 263 million uniquely mapped). In agreement with previous studies, we observe a global decrease in Pol II density outside of promoter-proximal regions compared to the DMSO control (Figure 5B, arrows) (Flynn et al., 2011; Jonkers et al., 2014; Rahl et al., 2010). Thus, FP treatment reduces productive elongation of most genes. To quantify the effect of FP treatment on convergent transcription, we calculated the ratio of convergent antisense to sense transcription at all promoter-proximal regions. If convergent transcription were a simple consequence of lower expression, it should not only be increased proportionally to promoter-proximal sense transcription following FP treatment, but importantly, it should appear in genes where it was not detected before. We observe that the convergent antisense-to-sense transcription ratio remains constant following FP treatment, indicating that sense and convergent antisense transcription levels covary, and we do not detect a new subpopulation of genes with convergent transcription in their promoter-proximal regions (Figure 5C). This result suggests that the lack of sense-productive transcription elongation is not sufficient to induce convergent transcription. Thus, if convergent antisense transcription is not a simple consequence of low sense expression, then it may contribute to the cause.

Impact of Transcription Factor Occupancy on Pol II Elongation

DNA-bound transcription factors (TFs) have the potential to obstruct elongating Pol II. To investigate the relationship between TF occupancy and Pol II progress, we expanded our

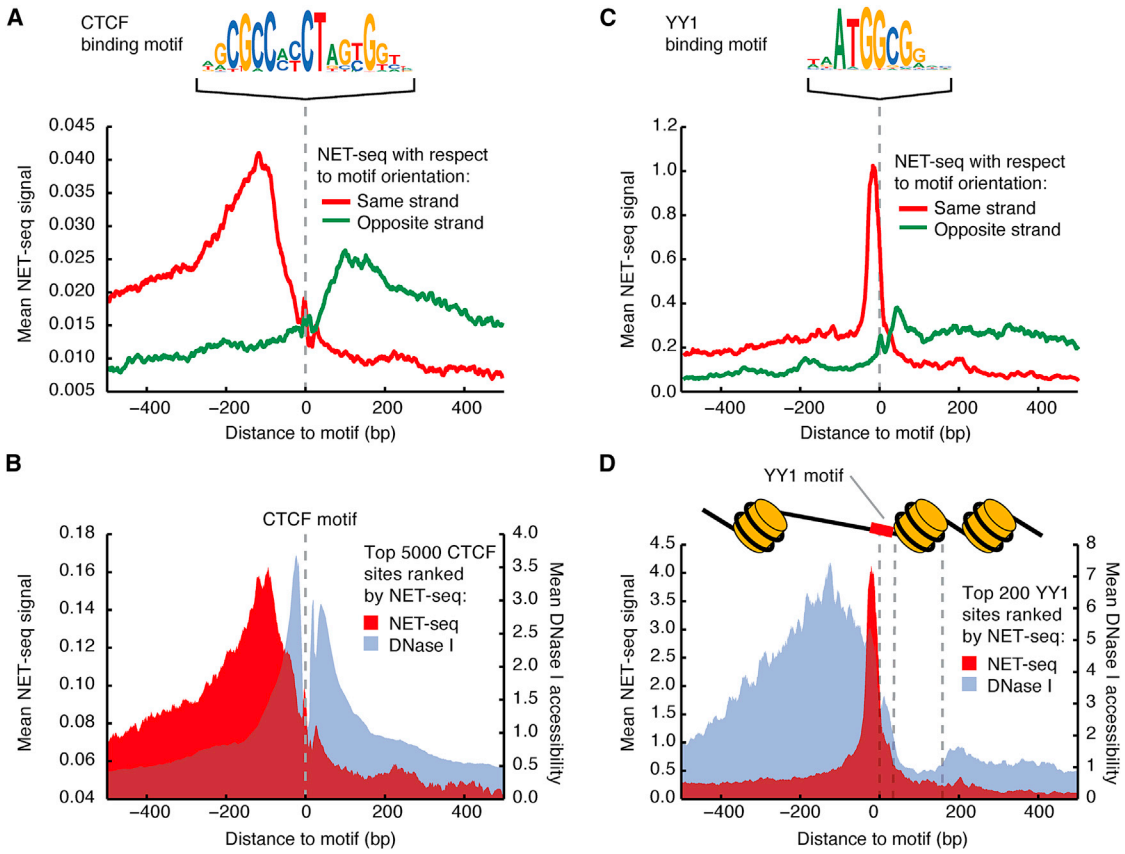


Figure 6. Pol II Pausing Associated with Transcription Factor Occupancy

(A) Average NET-seq signal around 16,339 CTCF motifs with accessible chromatin. In red is NET-seq signal oriented to the strand of the motif (Pol II transcription from left to right), and in green is NET-seq on the other strand (Pol II transcription from right to left). The CTCF binding motif is pictured above. The NET-seq data were smoothed by a 10 bp sliding window average.

(B) Mean NET-seq (red) and DNase I cleavage (gray) signal (10 bp windowed averages) surrounding the top 5,000 CTCF motifs sorted by NET-seq signal.

(C) Average NET-seq signal (smoothed by a 10 bp sliding window average) around 731 YY1 motifs with accessible chromatin. In red is NET-seq signal oriented to the strand of the motif (Pol II transcription from left to right), and in green is NET-seq on the other strand (Pol II transcription from right to left). The YY1 binding motif is pictured above.

(D) Mean per nucleotide NET-seq (red) and DNase I cleavage (gray) signal surrounding the top 200 YY1 motifs sorted by NET-seq signal. Both signals are presented as 10 bp windowed averages. Schematic of nucleosome positioning relative to YY1 inferred from DNase I accessibility is above plot.

DNase-seq data from HeLa S3 cells to genomic footprinting depth (269 million uniquely mapped genomic reads), enabling detailed mapping of the occupancy of TF recognition sites within DNase I hypersensitivity sites (DHSs). As CTCF is implicated in Pol II pausing in vitro and within the cell (Shukla et al., 2011), we quantified NET-seq signal and DNase-seq signal around CTCF recognition sites within DHSs on both strands. We observed higher Pol II density just upstream of the CTCF sites, suggesting that CTCF might represent a barrier to Pol II elongation genome wide (Figures 6A and 6B). Interestingly, the NET-seq signal around these sites differs in magnitude for each strand, indicating that CTCF may pose strand-specific obstacles (Figure 6A).

As transcriptional pausing has been seen upstream of nucleosomes in yeast and *Drosophila* cells (Churchman and Weissman, 2011; Mavrich et al., 2008; Weber et al., 2014), we investigated Pol II density around YY1, a canonical promoter-centric transcription factor (Xi et al., 2007) thought to position +1 nucleosomes (Vierstra et al., 2014). Thus, we speculated that YY1 occupancy might impact Pol II elongation. Given that poly-zinc finger TFs engage DNA asymmetrically, we also speculated that any impact on Pol II might also be strand specific. We observed a peak in NET-seq signal precisely at YY1 sites in DHSs, consistent with YY1-directed pausing (Figures 6C and 6D). Strikingly, this effect was highly directional and is predominant when Pol II engages YY1 from the upstream direction (Figure 6D). These results indicate that TFs might directly regulate Pol II elongation in direction- or strand-specific ways.

Fine Structure of Pol II Pausing along Constitutive and Alternative Exons

Alteration to transcription elongation rates affects splicing outcomes, which has led to the proposal of the kinetic model of transcription and splicing coupling (Dujardin et al., 2014; Fong et al.,

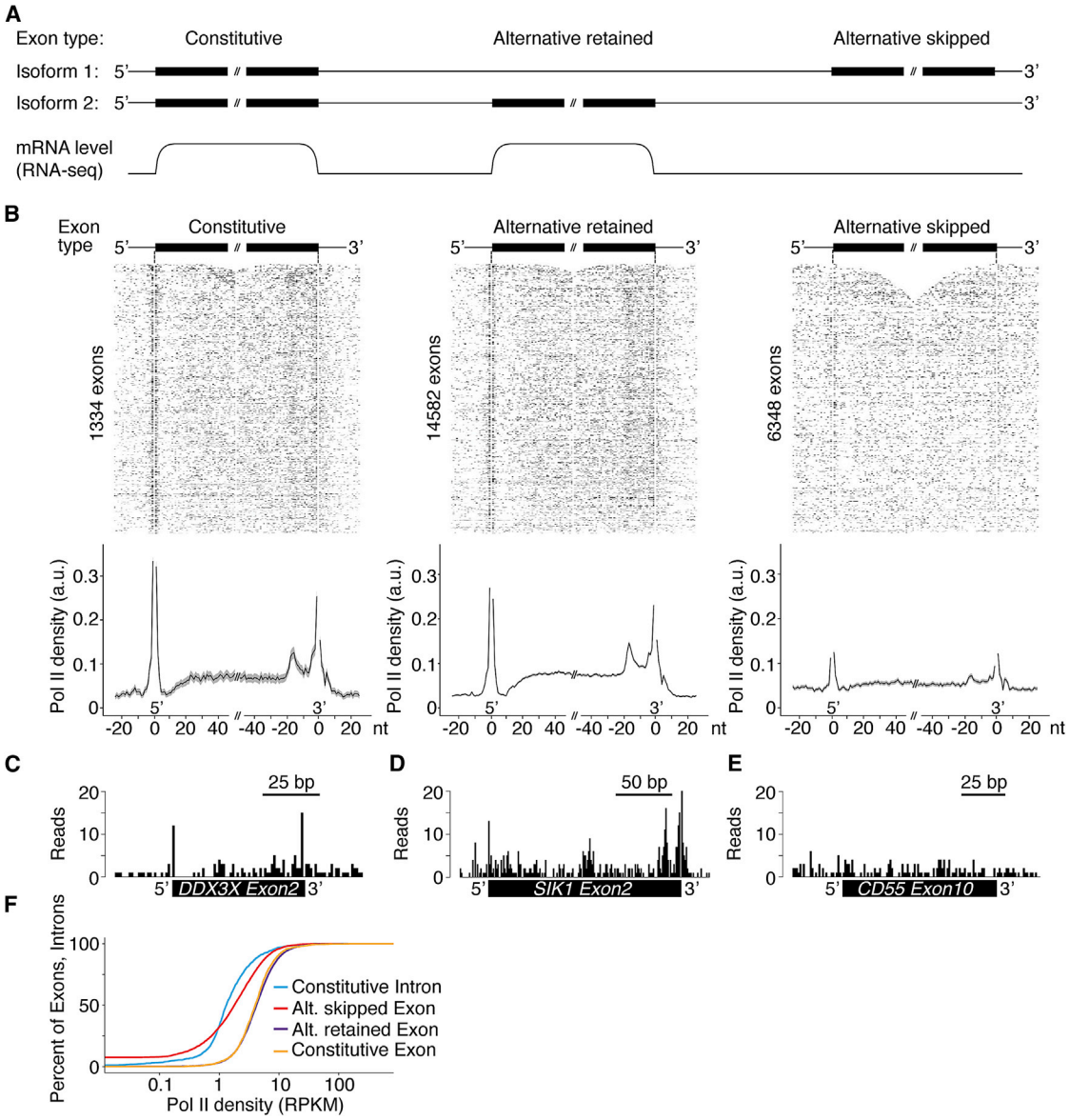


Figure 7. Pol II Density across Exons Reveals a Stereotypical Pausing Pattern that Depends on Splicing Outcome

(A) Schematic of the classification of constitutive, alternative retained, and alternative skipped exons based on annotated isoforms and detected levels in cytosolic RNA-seq data.

(B) A stringent set of exons was selected for analysis from genes containing NET-seq signal of ≥1 RPKM (B, see Experimental Procedures). Heatmaps and meta-exon analysis of HeLa S3 Pol II density across each type of exon, as defined in (A) (left to right: constitutive exons, n = 1,334; alternative retained, n = 14,582; and alternative skipped, n = 6,348). NET-seq signal from each exon (± 25 bp) is normalized to vary from 0 to 1 (white to black scale in the heatmaps). Solid lines on the meta-exon plots indicate the mean values, and the gray shading represents the 95% confidence interval. The single-nucleotide positions where splicing intermediates align (3′ ends of introns and exons) were entirely removed from analysis (see Experimental Procedures) and appear as a blank position in the figures.

(C-E) Raw NET-seq reads across the constitutive exon 2 within the *DDX3X* gene (C), alternative retained exon 2 within the *SIK1* gene (D), and the alternative skipped exon 10 within the *CD55* gene (E).

(F) Distribution of the percent of exons or introns with a given Pol II density.

RPKM, reads per kb per million Pol II uniquely aligned reads. See also Figure S5.

2014; Ip et al., 2011; de la Mata et al., 2003; Roberts et al., 1998). However, the degree to which transcription rate is modulated locally around exons is unclear. Higher Pol II density at human exons versus introns was reported using Pol II ChIP-seq and ChIP-chip (Brodsky et al., 2005; Schwartz et al., 2009), but in another study, no significant difference was observed (Spies et al., 2009). Furthermore, the precise pattern across individual exons could not be resolved. In *Drosophila* cells, PRO-seq observed high Pol II density across exons and detected a high enrichment of Pol II density at the 5′ ends (Kwak et al., 2013).

We analyzed NET-seq data at constitutive exons and revealed significantly higher coverage than at introns in both HeLa S3 (2.4 × higher) and HEK293T cells (2.2× higher) ($p < 10^{-15}$, Kolmogorov–Smirnov test), suggesting that transcription elongation is slower at exons in human cells (Figures 7A, 7B, 7F and S5A). Any contamination from processed mRNA would inflate these differences; however, our quality controls (Figures 1B, 1C, and S1A) suggest that this is a small effect, if at all. Strikingly, NET-seq signal across exons is not uniform: sharp increases in Pol II density occur in the few base pairs surrounding the 5′ and 3′ ends of constitutive exons, indicating strong Pol II pausing at exon boundaries (Figure 7B). As splicing intermediates are known NET-seq contaminants (Figure S1G), we removed the single bp positions where they align from analysis. Furthermore, a broader peak of RNA polymerase density is present ~17 nt before the 3′ end of exons. The general features of this pattern are observed at single exons, for example, exons 2 of the *DDX3X* and *SIK1* genes (Figures 7C and 7D). Finally, we observe similar trends in the NET-seq data from HEK293T cells (Figure S5B). This analysis suggests that exon borders impose a structured barrier to Pol II elongation.

Most human exons can be alternatively spliced, with retained exons varying between cell types (Pan et al., 2008; Wang et al., 2008). We expanded our analysis to alternatively spliced exons and investigated whether transcriptional pausing varies at exons with different splicing outcomes. We focused our analysis on genes with a NET-seq RPKM of >1 in the gene body (Figure S5A) and defined skipped exons as those undetected in the cytoplasmic RNA-seq data (Figure 7A). As for constitutive exons, retained alternative exons have higher Pol II density compared to the density across introns (Figure 7F). These exons also have a similar pausing pattern as constitutive exons, which is visible by meta-exon analysis and at the single exon (Figures 7B and 7D). Interestingly, Pol II density is lower at skipped exons than at alternative retained exons (Figure 7F). Strikingly, the Pol II density pattern is similarly shaped across skipped and retained exons, albeit significantly different in amplitude (Figures 7B and 7E). The residual pausing pattern at skipped exons could be due to the small number of retained exons that are misannotated as skipped. Finally, the same differences in the Pol II density patterns across retained and skipped exons are observed in HEK293T cells (Figure S5B). Together, these data show that Pol II recognizes exon structures with different processing fates, suggesting that alternative splicing is kinetically coupled to transcription elongation genome wide.

DISCUSSION

Here, we demonstrate that human NET-seq provides complete, strand-specific maps of transcription at single-nucleotide resolution. NET-seq thereby defines transcriptional pausing sites and directly measures unstable transcripts. Finally, NET-seq instantaneously reports the transcription status of genes, in contrast to RNA-seq, which reports the balance between RNA synthesis and degradation.

Our work describes an unappreciated aspect of promoter-proximal transcription: the presence of convergent transcription at many human genes. Importantly, we show that convergent transcription is characteristic of lower-expressed genes, suggesting a potential role in the regulation of promoter-proximal pausing. Prominent DNase I hypersensitivity sites flanking the convergent antisense peak indicate that promoter-proximal convergent transcription reflects initiation at a defined promoter located a characteristic distance from the canonical sense promoter.

Other than expression level, only one commonality is found between the genes with convergent transcription: the dinucleotide CC occurs slightly more frequently in regions displaying convergent transcription (12.4% ± 0.4% for convergent, 11.1% ± 0.2% for not convergent). Thus, it appears that convergent transcription is a prevalent phenomenon that is not restricted to a specific class of genes. An intriguing possibility is that paused antisense Pol II directly blocks or clashes with sense transcription, as can occur in yeast (Prescott and Proudfoot, 2002). The sense and convergent antisense peaks are too far apart (~150 bp) to reflect direct contact of paused polymerases, but the DNase-seq data reveal that this distance likely represents the +1 nucleosome that is positioned between them. Interference could arise through positioning of the +1 nucleosome or indirect mechanisms such as transcription-induced changes in DNA topology, chromatin modifications, or transcription factor occupancy. In any event, NET-seq data do not resolve whether sense and antisense transcription occur simultaneously, as the approach requires averaging over a population of cells. Therefore, potential roles of convergent transcription during initiation, elongation, and termination will have to be investigated within cell populations and at the single-cell level.

Our study yields a global picture of how transcription elongation is altered at alternatively spliced exons in human cells. Changes in transcription elongation influence alternative splicing, which is thought to be mediated either by the differential recruitment of splicing factors (recruitment model) or by biasing kinetic competition between multiple splicing outcomes (kinetic model) (Bentley, 2014; Dujardin et al., 2013; Kornblihtt et al., 2004). Here, we show that alternative splicing outcomes in human cells are associated with Pol II exon density and strong pauses at the 5′ and 3′ ends, consistent with the kinetic model. What causes pauses at exons is an important question. Nucleosomes can influence transcriptional pausing (Churchman and Weissman, 2011; Hodges et al., 2009; Izban and Luse, 1991; Skene et al., 2014), and, importantly, nucleosome occupancy and histone modifications transition at exon boundaries according to splice site strength (Andersson et al., 2009; Chodavarapu et al., 2010; Huff et al., 2010; Schwartz et al., 2009; Spies et al., 2009; Tilgner et al., 2009). DNA sequence and DNA methylation at exon boundaries could contribute to pausing because sequence elements have been shown to cause transcriptional pausing (Gelfman et al., 2013; Herbert et al., 2006; Kassavetis and Chamberlin, 1981; Larson et al., 2014; Maizels, 1973; Vvedenskaya et al., 2014). Additionally, transcription factors could underlie pausing at retained exons, as is the case with CTCF binding at exon 5 of the *CD45* gene (Shukla et al., 2011). The broad peak of Pol II density 17 bp from the 3′ end of the exon may reflect Pol II backtracking during the recovery from the strong pause at the 3′ end of the exon. Backtracking would

produce small cleavage products, consistent with the population of tiny RNAs that were previously identified in this region (Taft et al., 2009).

We expect adaptation of human NET-seq to any human cell type to be straightforward, resulting in a tool to illuminate a variety of biological processes. Future applications include high-resolution analyses of transcription regulation across cell types, responses to signaling pathways, and cellular differentiation.

EXPERIMENTAL PROCEDURES

Cell Fractionation and RNA Purification

Cell fractionation is performed as described by (Bhatt et al., 2012; Pandya-Jones and Black, 2009) and based on (Wuarin and Schibler, 1994) with modifications. All steps are conducted on ice or at 4°C and in the presence of 25 μM α-amanitin, 50 units/ml SUPERaseIN and Protease inhibitors cOmplete. HeLa S3 cells and HEK293T cells are grown in DMEM containing 10% FBS, 100 U/ml penicillin, and 100 μg/ml streptomycin to a confluency of 90%. Following lysis of 1×10^7 cells, the nuclei are washed with the nuclei wash buffer (0.1% Triton X-100, 1 mM EDTA, in 1× PBS) to remove cytoplasmic remnants. Nuclei lysis is performed without $MgCl_2$ (1% NP-40, 20 mM HEPES [pH 7.5], 300 mM NaCl, 1 M Urea, 0.2 mM EDTA, 1 mM DTT). The success of cell fractionation is monitored by western blot analysis and subcellular RNA-seq.

Sequencing Library Constructions

For NET-seq, the library preparation is performed as described by Churchman and Weissman (2011, 2012) with modifications. For 3′ RNA ligation, a pre-adenylated DNA linker with a mixed random hexameric barcode sequence at its 5′ end is used. cDNA containing the 3′ end sequences of a subset of mature and heavily sequenced snRNAs, snoRNAs, rRNAs, and mitochondrial tRNAs are specifically depleted using biotinylated DNA oligos (Table S1), as described by Ingolia et al. (2012). For subcellular RNA-seq, the sequencing libraries are prepared as described in Churchman and Weissman (2012), with the ribosomal RNA removed using the Ribo-Zero Magnetic Kit (Epicentre). DNA libraries are sequenced by the NextSeq 500 and HiSeq 2000 Illumina platforms.

Processing and Alignment of Sequencing Reads

Reads are trimmed and aligned using STAR (v2.4.0) (Dobin et al., 2013). For NET-seq data, only the position matching the 5′ end of the sequencing read (after removal of the barcode), corresponding to the 3′ end of the nascent RNA fragment, is recorded with a Python script using HTSeq package (Anders et al., 2015). Reverse transcription mispriming events are identified and removed when molecular barcode sequences match exactly to the genomic sequence adjacent to the aligned read. Reads that align to the same genomic position and contain identical barcodes are considered PCR duplication events and are removed. Splicing intermediates have 3′ hydroxyls and will enter NET-seq libraries and contribute to the reads aligning to the exact single-nucleotide 3′ ends of introns and 3′ ends of exons (Figure S1G). Therefore, reads that map precisely at the exact single-nucleotide ends of introns and exons are discarded, and the single 1 bp genomic positions are not considered in subsequent analysis.

Annotation of Exons and Introns

Clear exonic regions are identified by determining the minimum overlapping exonic region of all isoforms that have an exon at that position. If the region is present in all isoforms, it is considered a constitutive exon; otherwise, it is labeled alternative. Alternative skipped exons are classified by those alternative exons that are entirely undetected in the cytoplasm RNA-seq data, and the rest of the alternative exons are classified as retained. Constitutive intronic regions are identified as the minimum intronic overlapping regions present in all isoforms.

NET-Seq Exon Metagene and Heatmap Analysis

The set of exons included in the analysis are required to be within genes of an RPKM >1 in gene bodies (defined in Figure 2A) and not overlapping any other annotated exon. They are required to begin and end at the same position in all isoforms that contain the exon. First and last exons of genes are removed from analysis. NET-seq signal across each exon ±25 bp is normalized to range between 0 and 1 so that each exon contributes to the analysis with the same weight. Precise single-nucleotide genomic loci where splicing intermediates map (exact 3′ ends of introns and exons) are not included in the analysis, and those locations are left blank in any plots.

Analysis of Promoter-Proximal Regions

Promoter-proximal regions were carefully selected for analysis to ensure that there is minimal contamination from transcription arising from other transcription units. Starting with genes that are Pol II protein coding, non-overlapping within a region of 2.5 kb upstream of the TSS and 2.5 kb downstream of the polyA site, and longer than 2 kb, NET-seq data at promoter-proximal regions are required to have a coefficient of variation >0.5 and have at least 40 positions covered in the sense strand. Within a 4 kb window surrounding the TSS, peaks were identified in the sense from these genes. If >40 bases on the antisense strand have NET-seq signal, peaks were also identified on the antisense strand. Promoter regions with an antisense peak located downstream of the sense major peak are classified as displaying convergent transcription. Promoter regions with an antisense peak located upstream of the sense major peak are classified as displaying divergent transcription.

ACCESSION NUMBERS

All NET-seq and RNA-seq data sets are available at GEO under the accession number GSE61332. DNase-seq data sets are available at ENCODE under the ENCODE DCC accession number ENCBS229UDI.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.03.010.

AUTHOR CONTRIBUTIONS

A.M., J.d.I. and L.S.C. designed the NET-seq experiments; A.M. established NET-seq and subcellular RNA-seq experimental protocols, with input from J.d.I.; A.M. and S.M. carried out experiments; J.d.I. developed a bioinformatics analysis pipeline for human NET-seq and subcellular RNA-seq, with input from A.M.; A.R., J.V., R.S., and J.A.S. generated and analyzed the DNase-seq data; J.d.I., U.E., and L.S.C. analyzed NET-seq data; A.M., J.d.I., J.A.S., and L.S.C. wrote the manuscript.

ACKNOWLEDGMENTS

We thank F. Winston, J. Gray, M. Couvillion, S. Doris, and E. Feinberg for critical comments on the manuscript. We thank J.Gray and A. Snavely for help with eRNA analysis. We thank F. Winston, K. Struhl, S. Buratowski, A. Ciuffi, and A. Regev for advice and discussions. We thank M. Gebremeskel for tissue culture support; K. Waraska at the HMS Biopolymers Facility and Z. Herbert at the DFCI Molecular Biology Core Facilities for sequencing; and N. Pho and B.D. Kim at HMS Research Computing for computing support. This work was supported by US National Institutes of Health NHGRI grants R01HG007173 to L.S.C. and U54HG007010 to J.A.S.; a Damon Runyon Dale F. Frey Award for Breakthrough Scientists (to L.S.C.); and a Burroughs Wellcome Fund Career Award at the Scientific Interface (to L.S.C). A.M. was supported by Long-Term Postdoctoral Fellowships of the Human Frontier Science Program (LT000314/2013-L) and EMBO (ALTF858-2012). J.d.I. was supported by the Swiss National Science Foundation Postdoc Mobility

Fellowship. J.V. was supported by US National Science Foundation Graduate Research Fellowship under grant DGE-071824.

Received: October 8, 2014
Revised: November 26, 2014
Accepted: February 18, 2015
Published: April 23, 2015

REFERENCES

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360–363.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. *19*, 1732–1741.

Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nat. Rev. Genet. *15*, 163–175.

Bhatt, D.M., Pandya-Jones, A., Tong, A.-J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279–290.

Brodsky, A.S., Meyer, C.A., Swinburne, I.A., Hall, G., Keenan, B.J., Liu, X.S., Fox, E.A., and Silver, P.A. (2005). Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. Genome Biol. *6*, R64.

Cai, H., and Luse, D.S. (1987). Transcription initiation by RNA polymerase II in vitro. Properties of preinitiation, initiation, and elongation complexes. J. Biol. Chem. *262*, 298–304.

Callen, B.P., Shearwin, K.E., and Egan, J.B. (2004). Transcriptional interference between convergent promoters caused by elongation over the promoter. Mol. Cell *14*, 647–656.

Chao, S.H., and Price, D.H. (2001). Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. J. Biol. Chem. *276*, 31793–31799.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., et al. (2010). Relationship between nucleosome positioning and DNA methylation. Nature *466*, 388–392.

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

Churchman, L.S., and Weissman, J.S. (2012). Native elongating transcript sequencing (NET-seq). Curr. Protoc. Mol. Biol. *Chapter 4*, 1–17.

Consortium, T.E.P.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

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

Davis-Turak, J.C., Allison, K., Shokhirev, M.N., Ponomarenko, P., Tsimring, L.S., Glass, C.K., Johnson, T.L., and Hoffmann, A. (2015). Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing. Nucleic Acids Res. *43*, 699–707.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. Mol. Cell *12*, 525–532.

DeGennaro, C.M., Alver, B.H., Marguerat, S., Stepanova, E., Davis, C.P., Bähler, J., Park, P.J., and Winston, F. (2013). Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. Mol. Cell. Biol. *33*, 4779–4792.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L.I., Fiszbein, A., Godoy Herz, M.A., Nieto Moreno, N., Munoz, M.J., Alló, M., et al. (2013). Transcriptional elongation and alternative splicing. Biochim. Biophys. Acta *1829*, 134–140.

Dujardin, G., Lafaille, C., de la Mata, M., Marasco, L.E., Muñoz, M.J., Le Jossic-Corcos, C., Corcos, L., and Kornblihtt, A.R. (2014). How slow RNA polymerase II elongation favors alternative exon skipping. Mol. Cell *54*, 683–690.

Elledge, S., and Davis, R. (1989). Position and density effects on repression by stationary and mobile DNA-binding proteins. Genes Dev. *3*, 185–197.

Ferrari, F., Plachetka, A., Alekseyenko, A.A., Jung, Y.L., Ozsolak, F., Kharchenko, P.V., Park, P.J., and Kuroda, M.I. (2013). "Jump start and gain" model for dosage compensation in Drosophila based on direct sequencing of nascent transcripts. Cell Rep. *5*, 629–636.

Flynn, R.A., Almada, A.E., Zamudio, J.R., and Sharp, P.A. (2011). Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. Proc. Natl. Acad. Sci. USA *108*, 10460–10465.

Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., Diener, K., Jones, K., Fu, X.-D., and Bentley, D.L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. *28*, 2663–2676.

Gelfman, S., Cohen, N., Yearim, A., and Ast, G. (2013). DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. Genome Res. *23*, 789–799.

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

Gullerova, M., and Proudfoot, N.J. (2012). Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. Nat. Struct. Mol. Biol. *19*, 1193–1201.

Herbert, K.M., La Porta, A., Wong, B.J., Mooney, R.A., Neuman, K.C., Landick, R., and Block, S.M. (2006). Sequence-resolved detection of pausing by single RNA polymerase molecules. Cell *125*, 1083–1094.

Hobson, D.J., Wei, W., Steinmetz, L.M., and Svejstrup, J.Q. (2012). RNA polymerase II collision interrupts convergent transcription. Mol. Cell *48*, 365–374.

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

Huff, J.T., Plocik, A.M., Guthrie, C., and Yamamoto, K.R. (2010). Reciprocal intronic and exonic histone modification regions in humans. Nat. Struct. Mol. Biol. *17*, 1495–1499.

Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. *7*, 1534–1550.

Ip, J.Y., Schmidt, D., Pan, Q., Ramani, A.K., Fraser, A.G., Odom, D.T., and Blencowe, B.J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res. *21*, 390–401.

Izban, M.G., and Luse, D.S. (1991). Transcription on nucleosomal templates by RNA polymerase II in vitro: inhibition of elongation with enhancement of sequence-specific pausing. Genes Dev. *5*, 683–696.

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

Kassavetis, G.A., and Chamberlin, M.J. (1981). Pausing and termination of transcription within the early region of bacteriophage T7 DNA in vitro. J. Biol. Chem. *256*, 2777–2786.

Kim, J.B., and Sharp, P.A. (2001). Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain independently of cyclin-dependent kinase-activating kinase. J. Biol. Chem. *276*, 12317–12323.

Kim, T., Xu, Z., Clauder-Münster, S., Steinmetz, L.M., and Buratowski, S. (2012). Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics. Cell *150*, 1158–1169.

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., and Nogues, G. (2004). Multiple links between transcription and splicing. RNA *10*, 1489–1498.

Krumm, A., Meulia, T., Brunvand, M., and Groudine, M. (1992). The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. Genes Dev. *6*, 2201–2213.

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.

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.

Lindell, T.J., Weinberg, F., Morris, P.W., Roeder, R.G., and Rutter, W.J. (1970). Specific inhibition of nuclear RNA polymerase II by alpha-amanitin. Science *170*, 447–449.

Lis, J.T., Mason, P., Peng, J., Price, D.H., and Werner, J. (2000). P-TEFb kinase recruitment and function at heat shock loci. Genes Dev. *14*, 792–803.

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

Marquardt, S., Escalante-Chong, R., Pho, N., Wang, J., Churchman, L.S., Springer, M., and Buratowski, S. (2014). A chromatin-based mechanism for limiting divergent noncoding transcription. Cell *157*, 1712–1723.

Martens, J.A., Laprade, L., and Winston, F. (2004). Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature *429*, 571–574.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., et al. (2008). Nucleosome organization in the Drosophila genome. Nature *453*, 358–362.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat. Genet. *39*, 1507–1511.

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038–1042.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat. Struct. Mol. Biol. *20*, 923–928.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA *15*, 1896–1908.

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

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851–1854.

Prescott, E.M., and Proudfoot, N.J. (2002). Transcriptional collision between convergent genes in budding yeast. Proc. Natl. Acad. Sci. USA *99*, 8796–8801.

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

Roberts, G.C., Gooding, C., Mak, H.Y., Proudfoot, N.J., and Smith, C.W. (1998). Co-transcriptional commitment to alternative splice site selection. Nucleic Acids Res. *26*, 5568–5572.

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

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075–1087.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16*, 990–995.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849–1851.

Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle *8*, 2557–2564.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference—a crash course. Trends Genet. *21*, 339–345.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74–79.

Skene, P.J., Hernandez, A.E., Groudine, M., Henikoff, S., and Espinosa, J.M. (2014). The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. eLife *3*, e02042.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol. Cell *36*, 245–254.

Strobl, L.J., and Eick, D. (1992). Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. EMBO J. *11*, 3307–3314.

Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G.J., Lassmann, T., Forrest, A.R.R., Grimmond, S.M., Schroder, K., et al. (2009). Tiny RNAs associated with transcription start sites in animals. Nat. Genet. *41*, 572–578.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16*, 996–1001.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. *22*, 1616–1625.

Vierstra, J., Wang, H., John, S., Sandstrom, R., and Stamatoyannopoulos, J.A. (2014). Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. Nat. Methods *11*, 66–72.

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

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

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.

Whitehouse, I., Rando, O.J., Delrow, J., and Tsukiyama, T. (2007). Chromatin remodelling at promoters suppresses antisense transcription. Nature *450*, 1031–1035.

Wozniak, G.G., and Strahl, B.D. (2014). Hitting the 'mark': Interpreting lysine methylation in the context of active transcription. Biochim. Biophys. Acta *1839*, 1353–1361.

Wu, X., and Sharp, P.A. (2013). Divergent transcription: a driving force for new gene origination? Cell *155*, 990–996.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol. Cell. Biol. *14*, 7219–7225.

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., and Weng, Z. (2007). Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res. *17*, 798–806.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033–1037.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat. Genet. *39*, 1512–1516.

EXTENDED EXPERIMENTAL PROCEDURES

Flavopiridol Treatment

Flavopiridol treatment was performed similar to Rahl et al., Cell, 2010. HeLa S3 cells (ATCC, CCL-2.2) were grown in DMEM containing 10% FBS, 100 U/ml penicillin and 100 μg/ml streptomycin until 90% confluency. Media was then replaced by DMEM containing 1 μM flavopiridol (F3055, Sigma), 10% FBS, 100 U/ml penicillin and 100 μg/ml streptomycin, and incubated for 1 hr. Flavopiridol was solubilized in DMSO. To assess potential effects that may arise from DMSO alone, DMSO only treated cells were processed in parallel. Following treatment, $1x10^7$ cells were applied to fractionation as described below.

For the time course experiment, cells were harvested by trypsinization before flavopiridol or DMSO containing media was added as well as 1, 2, 3, 4 and 6 hr after treatment. Cells were centrifuged at 60 g for 1 min at room temperature. Cell pellets were resuspended in 200 μl 1x SDS sample buffer and boiled for 5 min at 95°C. Western blot analysis were performed using antibodies directed against the CTD Ser2-phosphorylated (3E10, Active Motif), the CTD Ser5-phosphorylated forms of Pol II (3E8, Active Motif), Pol II (8WG16, NeoClone) as well as against GAPDH (6C5, Applied Biosystems).

Cell Fractionation

Cell fractionation was performed as described in (Bhatt et al., 2012; Pandya-Jones and Black, 2009) and based on (Wuarin and Schibler, 1994) with modifications described below. All subsequent steps have been conducted on ice or at 4°C and in the presence of 25 μM α-amanitin (Sigma, A2263), 50 Units SUPERaseIN (Life Technologies, AM2696) and Protease inhibitors cOmplete (Roche, 11873580001) according to manufacturer's instructions using RNase free equipment. All buffers have been pre-chilled on ice before use. The cell pellet corresponding to $1x10^7$ cells was gently resuspended in 200 μl cytoplasmic lysis buffer (0.15% NP-40, 10 mM Tris-HCl pH 7.0, 150 mM NaCl). The cell lysate was incubated for 5 min on ice, layered onto 500 μl sucrose buffer (10 mM Tris-HCl pH 7.0, 150 mM NaCl, 25% sucrose) and centrifuged at 16,000 g for 10 min. The supernatant corresponding to the cytoplasmic fraction was carefully removed. For subcellular RNA-seq (described below) the cytoplasmic fraction was mixed with 3.5x volumes of RLT buffer (74104, QIAGEN). RNA purification from the cytoplasmic and following fractions is described below.

The nuclei pellet was gently resuspended in 800 μl nuclei wash buffer (0.1% Triton X-100, 1 mM EDTA, in 1x PBS) and centrifuged at 1500 g for 1 min. The supernatant was removed and the pellet was gently resuspended in 200 μl glycerol buffer (20 mM Tris-HCl pH 8.0, 75 mM NaCl, 0.5 mM EDTA, 50% glycerol, 0.85 mM DTT). Next, 200 μl nuclei lysis buffer (1% NP-40, 20 mM HEPES pH 7.5, 300 mM NaCl, 1M Urea, 0.2 mM EDTA, 1 mM DTT) was added, vortexed and incubated on ice for 2 min. The lysed nuclei were centrifuged at 18,500 g for 2 min. The supernatant corresponding to the nucleoplasmic fraction was removed. For subcellular RNA-seq (described below) the nucleoplasmic fraction was mixed with 3.5x volumes of RLT buffer (QIAGEN). The chromatin pellet was resuspended in 50 μl chromatin resuspension solution (25 μM α-amanitin, 50 Units SUPERaseIN, Protease inhibitors cOmplete, in 1x PBS).

The success of cell fractionation was monitored by western blot analyses and subcellular RNA sequencing. For western blot analyses membranes were probed with the following primary antibodies: Pol II (F-12, Cruz Biotechnology), Pol II Ser2-P (3E10, Active Motif), Pol II Ser5-P (3E8, Active Motif), Histone 2B (FL-126, Santa Cruz Biotechnology), U1 snRNP70 (C-18, Santa Cruz Biotechnology) and GAPDH (6C5, Applied Biosystems). Next, membranes were probed with Cy5-conjugated secondary antibodies (Cy5 goat anti-mouse, A10524; Cy5 goat anti-rabbit, A10523; Cy5 goat anti-rat, A10525; Life Technologies) and scanned using a Typhoon 9400 scanner (GE Healthcare). Fluorescent signals were quantified with ImageJ 1.47v software.

RNA Preparation

For NET-seq, 700 μl Quiazol (QIAGEN) was added to the chromatin solution and thoroughly mixed. RNA was purified using the miRNeasy Mini Kit (QIAGEN, 217004) and an on-column DNase I digestion was performed using the RNase-free DNase set (QIAGEN, 79254) according to the manufacturer's instructions. Following the column washes, the RNA was eluted in 30 μl RNase-free H_2O and the concentration as well as the quality was determined using a NanoDrop 2000 spectrophotometer (Thermo Scientific). The RNA concentration was typically 800-1,000 ng/μL. The concentrations of RNA obtained from flavopiridol treated cells were in the range of 300-500 ng/μL. The corresponding A_{260}/A_{280} ratios were usually around 2.1.

For subcellular RNA-seq RNA was prepared as described by Bhatt et al., Cell, 2012. Briefly, 500 μl TRI-reagent (TR 118, Molecular Research Center) was added to the resuspended chromatin. Next, 100 μl chloroform (Sigma, 288306) was added, incubated for 5 min at room temperature and centrifuged at 13,000 rpm for 15 min at 4°C. The upper aqueous phase was mixed with 3.5x volumes of RLT buffer (QIAGEN). RNA was purified from the RLT dissolved samples, including the cytoplasmic and nucleoplasmic samples, by using the RNeasy Mini Kit (QIAGEN, 74104). Briefly, 2.5x volumes ethanol were added, mixed and loaded onto RNeasy columns. After column washes, RNA was eluted in 50 μl RNase-free H_2O. The RNA concentration as determined by NanoDrop 2000 spectrophotometer (Thermo Scientific) was typically 400-500, 600-700 and 3,000-3,800 ng/μL for chromatin, nucleoplasmic and cytoplasmic fractions, respectively. The corresponding A_{260}/A_{280} ratios were in the range of 2.05 to 2.10.

Sequencing Library Preparation for NET-Seq

Sequencing library preparation was performed as described in (Churchman and Weissman, 2012) with critical modifications described below. 3 μg of purified chromatin associated RNA was used to prepare one DNA sequencing library. 3′ ligation of RNA with a DNA linker was performed essentially as originally described, except that a DNA linker with a mixed random hexameric

sequence at its 5′ end was used (new DNA linker: 5′-AppNNNNNNCTGTAGGCACCATCAAT/3 ddC-3′). The ligation efficiency was monitored by ligating the new DNA linker to a 28 nt long RNA control oligo (oGAB11: 5′-AGUCACUUAGCGAUGUACACUGACUGUG-3′OH). The ligation efficiency as determined by polyacrylamide gel electrophoresis was ≥ 95%. After ligation the RNA was fragmented by partial alkaline hydrolysis essentially as originally described. The fragmentation time was adjusted so that most RNA molecules were fragmented into the range of 35 to 100 nt, size-selected and converted into cDNA. Reverse transcription was performed with the following modifications to the original protocol. First, for reverse transcription a new RT primer was used (oLSC007: 5′-Phos/ATCTCGTATGCCGTCTTCTGCTTG/iSp18/CACTCA/iSP18/ TCCGACGATCATTGATGGTGCCTACAG-3′). Second, the RT primer was applied at a lower final concentration (1.9 μM instead of 10.9 μM). Next, cDNA was circularized and vectors that contained cDNA corresponding to original mature snRNAs, snoRNAs, rRNAs and mitochondrial tRNAs that were heavily sequenced in first NET-seq experiments were specifically depleted. For specific depletion, biotinylated oligos were designed that were complementary to the 3′ ends of the 19 most heavily sequenced mature RNAs of HeLa S3 cells (see Table S1). Depletion was performed similar to the rRNA depletion approach described by (Ingolia et al., 2012). Per depletion 5.0 μl circularization reaction was used. 5.0 μl of circularization reaction was combined with 1.0 μl depletion oligo pool (10 μM of each oligo, prepared in 10 mM Tris-HCl, pH 8.0; see Table S1), 1.0 μl 20x SSC (Life Technologies, AM9763) and 3.0 μl DNase-free H_2O. Tubes were placed in a thermal cycler, denatured for 90 s at 99°C and then annealed at 0.1°C s^{-1} to 37°C. Tubes were incubated for 15 min at 37°C. 37.5 μl of MyOne Streptavidin C1 DynaBeads (10 mg/ml, Life Technologies, 65001) were used per depletion reaction. Magnetic beads were washed 3x with 1x bind/wash buffer (1M NaCl, 0.5 mM EDTA, 2.5 mM Tris-HCl pH 7.0, 0.1% (vol/vol) Triton X-100). After the last wash, magnetic beads were resuspended in 15 μl 2x bind/wash buffer (2 M NaCl, 1 mM EDTA, 5 mM Tris-HCl pH 7.0, 0.2% (vol/vol) Triton X-100) and incubated at 37°C. 10 μl of the depletion reaction were transferred directly to the bead aliquot, immediately mixed by pipetting and incubated for 15 min at 37°C with mixing at 1000 rpm. Finally, tubes were transferred to a magnetic rack and 20 μl of supernatant were recovered and isopropanol precipitated. After depletion, the circularized DNA served as template for PCR amplification. PCR amplification was performed as originally described (Churchman and Weissman, 2012), using minimal amplification cycles. The concentration, the quality and the size distribution of final DNA libraries were assessed by Qubit 2.0 Fluorometer (Invitrogen) and 2100 Bioanalyzer (Agilent) measurements. 3′ end sequencing was performed on NextSeq 500 (SE 75 nt) and HiSeq 2000 (SE 50 nt) Illumina sequencing platforms using the following custom sequencing primer: oLSC006: 5′-TCCGACGATCATTGATGGTGCCTACAG-3′.

Sequencing Library Preparation for Subcellular RNA-Seq

Sequencing libraries for the chromatin, nucleoplasmic and cytoplasmic fractions were prepared as described for NET-seq with the following modifications.

DNase I treatment was performed after RNA purification. 25 μg of RNA were treated with 500 Units RNase-free DNase I (13550-50-9, MO Bio Laboratories). DNase I digestion was performed for 30 min at 37°C. DNase I treated RNA was purified by phenol-chloroform extraction, followed by an isopropanol precipitation. The RNA pellet was resuspended in 30 μl RNase-free H_2O. Next, rRNA was removed from the DNase I treated RNA sample by using the Ribo-Zero Magnetic Kit (MRZH116, Epicenter). rRNA removal and subsequent purification of rRNA depleted RNA by ethanol precipitation was performed according to the manufacturer's instructions. 26 μl of the DNase I treated RNA sample were used for rRNA depletion.

3x 1 μg of rRNA depleted RNA was fragmented before ligation. RNA fragmentation was performed essentially as described for NET-seq library preparation.

Fragmentation resulted in RNA fragments with 3′ phosphate-groups that had to be removed before DNA linker ligation. 3′ phosphate-groups were removed enzymatically as follows, using the T4 Polynucleotide Kinase (PNK: M0201S, NEB). Fragmented RNA were denatured for 2 min at 80°C and cooled on ice for 3 min. Next, 22.5 μl PNK reaction mix (5 μl T4 Polynucleotide Kinase Buffer (B0201S, NEB), 1.25 μl SUPERaseIN (20 U/μl), 16.25 μl RNase-free H_2O) was added, mixed and combined with 3 μl (10 U) PNK. Dephosphorylation was performed for 1 hr 20 min at 37°C, followed by 10 min at 75°C to inactivate PNK. After heat inactivation RNA was isopropanol precipitated and resuspended in 6 μl RNase-free H_2O. 5 μl were used for DNA linker ligation. In this case the original pre-adenylated DNA linker (5′-AppCTGTAGGCACCATCAAT/3 ddC-3′) was ligated to the 3′ end of the RNA fragments, essentially as described in (Churchman and Weissman, 2012). The success of dephosphorylation and ligation was monitored by using a 28 nt long RNA control oligo with a 3′ phosphate-group (oGAB11-P: 5′-AGUCACUUAGCGAUGUACACUGACUGUG-PO_4^{3-}-3′). A complete dephosphorylation by PNK would result in a high ligation efficiency. The ligation efficiency after PNK dephosphorylation was determined by polyacrylamide gel electrophoresis and was ≥ 95%, also indicating a high dephosphorylation efficiency by PNK.

Processing and Alignment of Sequencing Reads

For NET-seq data and HEK293T cytoplasmic RNA-seq library, the six 5′ end nucleotides corresponding to the molecular barcode, are trimmed from the reads, but remain associated with the read using a custom python script. For all samples, reads are aligned using the STAR aligner (v2.4.0) (Dobin et al., 2013) with the following parameters: –clip3pAdapterSeq ATCTCGTATGCCGT CTTCTGCTTG –clip3pAdapterMMp 0.21 –clip3pAfterAdapterNbases 1 –outFilterMultimapNmax 101 –outSJfilterOverhangMin 3 1 1 1 –outSJfilterDistToOtherSJmin 0 0 0 0 –alignIntronMin 11 –alignEndsType EndToEnd. The reference genome used consisted in the human reference genome (hg19), with additional sequences containing the rDNA sequences (GenBank U13369.1, GenBank U67616.1), and the sequence of processed tRNA (with the additional CCA at their end, and intron removed, when present). For NET-seq data, to avoid any bias toward favoring annotated regions the alignment was performed without providing transcriptome

information, while for RNA-seq data, the GENCODE annotation v16 was used as the transcriptome. For NET-seq data, only the position corresponding to the 5′ end of the sequencing read (after removal of the barcode), which corresponds to the 3′ end of the nascent RNA fragment, is recorded with a custom python script using HTSeq package (Anders et al., 2015), while the whole read coverage is recorded for RNA-seq data. Reverse transcription mispriming events are identified where barcode sequences correspond exactly to the genomic sequence adjacent to the aligned read. These reads are removed from further analysis. The few reads that align to the same genomic position and contain identical barcodes are considered PCR duplication events and are filtered out. Evidence of splicing intermediates in the NET-seq data (Figure S1G) revealed that the exact single nucleotide 3′ ends of introns and exact single nucleotide 3′ ends of exons, where splicing intermediates map, contain alignments that are due to both splicing intermediates and the 3′ ends of nascent RNA. Thus, reads that map precisely at the exact single nucleotide ends of introns and exons are discarded and the single 1 bp genomic positions are not considered for subsequent analysis.

Annotation of Exons and Introns

Constitutive exons, alternative exons and constitutive introns were defined as follows. Starting with GENCODE v16 human gene annotation, clear exonic regions are identified by determining the minimum overlapping exonic region of all isoforms that have an exon at that position. If the region is present in all isoforms, it is considered a constitutive exon, otherwise it is labeled alternative. Alternative skipped exons are classified by those alternative exons that are entirely undetected in the respective cell line cytoplasmic RNA-seq data and the rest of the alternative exons are classified as retained. Constitutive intronic regions are identified as the minimum intronic overlapping regions present in all isoforms. For the calculations determining the ratio of Pol II density on introns compared to exons, intronic regions were also included for analysis in cases where they are not present in all isoforms, but do not overlap an exon (for example, 5′ extended isoforms). For this analysis, all Pol II protein coding non-overlapping any other genes were used. For exon metagene and heatmap analysis, Pol II protein coding exons that start and end at the same position in all isoforms that contain the exon, and that did not overlap any other annotated exons from another gene were considered for analysis.

NET-Seq Exon Metagene and Heatmap Analysis

Exons defined as constitutive, alternative retained and alternative skipped are further filtered by expression. The set of exons included in analysis were required to be within genes of an RPKM greater than 1 in gene bodies (defined in Figure 2A). RPKM is reads per kb per million uniquely aligned reads at Pol II-transcribed genes. First and last exons from the newly defined gene units are removed from analysis. NET-seq signal across each exon ± 25 bp is normalized to range between 0 and 1 (illustrated as white to black scale on the heatmaps) so that each exon contributes to the analysis with the same weight. The exons are aligned at their 5′ and 3′ ends, and the mean normalized signal per bp is calculated. Precise single nucleotide genomic loci where splicing intermediates map (exact 3′ ends of introns and exons) are not included in the analysis and those locations are left blank. Plots were performed in R (R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/) using ggplot2 (H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009) and fields (Reinhard Furrer, Douglas Nychka and Stephen Sain (2013). fields: Tools for spatial data. R package version 6.8. http://CRAN.R-project.org/package=fields) packages.

Exon to Intron Pol II Density Ratio Analysis

Intronic and exonic feature counts are extracted using HTSeq-count (Anders et al., 2015) using the annotations described above. The total number of exonic and intronic reads per gene are normalized by their respective length. The ratio of the size-normalized counts is calculated for each Pol II protein coding non-overlapping gene having at least 1 read in the defined exonic and intronic regions.

Analysis of Promoter-Proximal Regions

Promoter-proximal regions were carefully selected for analysis to ensure that there is minimal contamination from transcription arising from outside areas. Starting with genes that are Pol II protein coding, non-overlapping within a region of 2.5 kb upstream of the TSS and 2.5kb downstream of the polyA site, and longer than 2 kb, promoter-proximal regions are required to have a coefficient of variation greater than 0.5 and at least 40 positions covered in the sense transcription out of the 4000 bp surrounding the TSS. In the case of conflicting isoforms annotations, the most upstream annotated TSS and the most downstream annotated polyA sites are used. NET-seq signal in a 4 kb window surrounding the TSS is converted to a ranking score and convoluted with a Gaussian kernel (bandwidth = 500 bp) for sense and antisense signal. Promoter-proximal regions with a signal lower than 0.1 in the sense transcription throughout the studied region were removed from analysis, and only promoter-proximal regions with at least 40 positions covered in the antisense transcription in the same region were retained for peak calling. The promoters with less than 40 positions with signal were considered as having no divergent or convergent signal. For both strands, peaks are identified through analysis of the first and second derivatives of the convoluted data. For both strands, the major peak is identified and additional significant peaks are identified if it has a height > = 0.5 and 0.8 times the height of the major peak respectively for antisense and sense strands. Promoters with an antisense peak located downstream of the sense major peak are classified as displaying convergent transcription. Promoters with an antisense peak located upstream of the sense major peak are classified as displaying divergent transcription. Promoters with more than 3 antisense detected peaks throughout the region were discarded from analysis, as they could arise from missing annotation.

NET-Seq Promoter-Proximal Region Heatmap and Metagene Analysis

NET-seq data from each promoter-proximal region classified as either divergent and/or convergent are used for analysis and centered at the sense transcription peak (+/− 1kb). For the heatmaps, the sense (violet) and antisense (red) raw signal is separately ranked, convoluted with a Gaussian kernel (bandwidth = 500), normalized to vary from 0 to 1 and smoothed with a 50 bp sliding window average. Both signal are then superposed. Genes are sorted by the distance between the sense and antisense peaks. For metagenes, raw sense and antisense signal are normalized together to vary between 0 and 1 and smoothed with a 50 bp sliding window average. The mean of the normalized signal is then plotted in the metagene.

Traveling Ratio Calculations

The traveling ratio is calculated by dividing the RPKM in the region −30 bp to + 300 bp around transcription start sites by the RPKM in the region +1 kb to the polyadenylation site (Rahl et al., 2010). Analysis was performed for all genes with an RPKM of at least 1 in the promoter region, defined above. RPKM is reads per kb per million uniquely aligned reads at Pol II-transcribed genes.

NET-Seq Flavopiridol and DMSO Metagene Analysis

NET-seq data from each promoter-proximal region classified as either divergent and/or convergent (obtained with the same steps described above) in both DMSO and flavopiridol samples are used for analysis and centered at the sense transcription peak (+/− 2.5kb). Raw sense and antisense signal are binned in 10 bp window, normalized together to vary between 0 and 1 and smoothed with a 50 bp sliding window average. The mean of the normalized signal per bin is then plotted in the metagene.

DNase-Seq Analysis

Using the same cells as used for NET-seq, HeLa S3 DNase I data was generated and aligned to the human reference genome (hg19) as described (Thurman et al., 2012). To determine the DNase hypersensitivity at promoter-proximal regions, the raw DNase I per-nucleotide cleavage data was smoothed over 150 base pairs centered on 20 bp interval steps. The aggregate accessibility profiles were generated by calculating the trimmed mean (mean removing 5% of extreme data points) for each of 20 bp intervals surrounding a NET-seq peak. The CTCF and YY1 position weight matrices (PWMs) were obtained for a reference set of consensus sequences generated via SELEX-seq (Jolma et al., 2013). Transcription factor binding sites were identified by scanning the entire genome for consensus sequences using the FIMO (45) tool from the MEME Suite (version 4.6) (Bailey et al., 2009). A 5th order Markov model was generated from 36 bp mappable genome sequence and used as the background model. Putative binding sites with a FIMO $p < 10^{-5}$ were retained. For the analysis in Figure 6, per nucleotide NET-seq and DNase I cleavage data were smoothed in 10 bp windows.

Analysis of Enhancer Regions

Enhancer locations were identified as DNase hypersensitive sites in HeLa S3 cells (mapped in this study and by ENCODE) which overlap with H3K4me1 but not H3K4me3 chromatin marks (as mapped by ENCODE) and were more than 5 kb away from annotated genes.

NET-Seq Enhancer Region Heatmap and Metagene Analysis

Summed NET-seq signal from both strand at each enhancer region (+/− 2 kb from center of the DNase I hypersensitive region) passing the filtering step described above is normalized to vary between 0 and 1 and smoothed into 50 bp window average. The mean of the normalized signal is then plotted in the metagene. Only the top 50% enhancer regions with highest NET-seq signal are kept for analysis (n = 882).

GRO-Seq Promoter Proximal Region Metagene Analysis

Genes (mm9 NCBI37 ensembl annotation) that are Pol II protein coding, non-overlapping within a region of 2.5 kb upstream of the TSS and 2.5 kb downstream of the pA site, and longer than 2 kb were selected for analysis (n = 4,200). GRO-seq antisense signal (Jonkers et al., 2014) (GEO accession GSE48895, mouse embryonic stem cells treated with flavopiridol for 50 min) from each gene with non-zero signal is normalized to vary between 0 and 1 and smoothed by 50 bp sliding window average. The mean of the normalized signal is then plotted in the metagene.

SUPPLEMENTAL REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

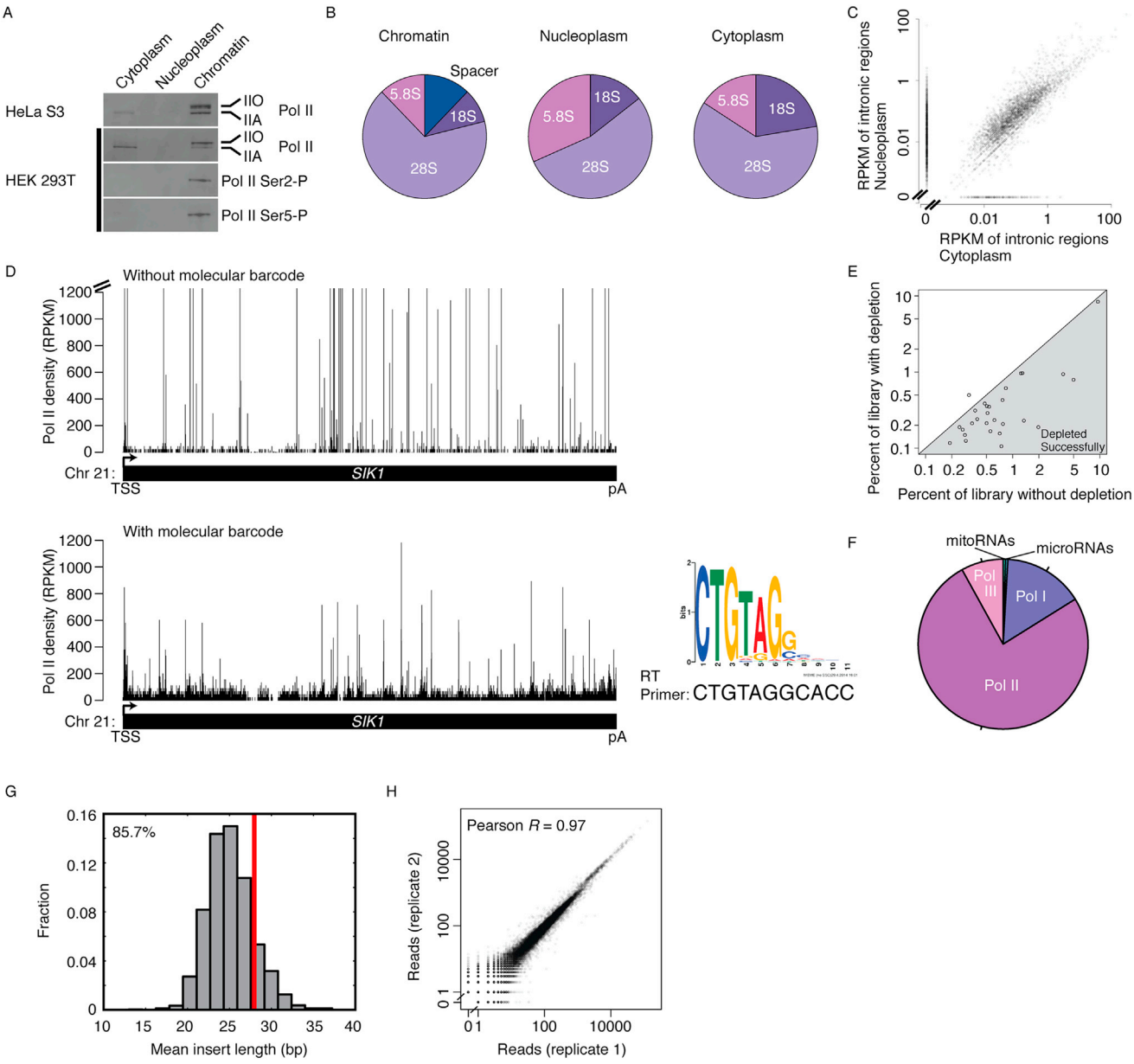


Figure S1. Human NET-Seq Quality Controls, Related to Figure 1

(A) Elongating Pol II is captured in the chromatin fraction of HeLa S3 and HEK293T cells. Cell lysates prepared from the cytoplasmic, nucleoplasmic and chromatin fractions of HeLa S3 (first panel) and HEK293T cells (second to fourth panels) were analyzed by western blotting using different antibodies against Pol II. Probing the different fractions with the F-12 antibody (Santa Cruz) that is directed against the N-terminal part of Rpb1, the largest subunit of Pol II, revealed the CTD hyper-phosphorylated form (IIO) as well as the hypo-phosphorylated form (IIA) of Pol II for HeLa S3 (first panel) and HEK293T cells (second panel). The three subcellular fractions were also probed with antibodies directed against the CTD serine 2 (Ser2-P, 3E10 Ab; third panel) and the CTD serine 5 phosphorylated (Ser5-P, 3E8 Ab; fourth panel) form of Pol II. (B) Distribution of Pol I transcribed rRNAs reads in the three cellular fractions, chromatin, nucleoplasm and cytoplasm, as determined by RNA-seq. Spacer, part of the precursor 45S rRNA, is found majorily in the chromatin fraction. Multialigned reads were kept for the purpose of this analysis and weighted accordingly. (C) RPKM of intronic regions in the nucleoplasm RNA-seq library compared to the RPKM of intronic regions in the cytoplasm RNA-seq library. Introns with an RPKM equal to 0 in one of the samples were given pseudocounts in order to be included on the plot. Intronic regions with an RPKM > 1 are probably intron retention events and are present in both nucleoplasm and cytoplasm. (D) RPKM in representative gene (*SIK1*) with (lower panel) and without (upper panel) use of a random hexamer barcode in library construction. Positions of extremely high reads were due to reverse transcription mispriming artifacts. Analysis of the sequences around those positions revealed a strong sequence motif that precisely matches the sequence of the RT primer. The sequence motif was identified using MEME software (http://meme.nbcr.net/meme/tools/meme). The sequence motif as well as parts of the reverse complement sequence of the RT primer are given (lower panel, right). The direction of transcription is indicated by an arrow. For the upper panel the y axis was cut at an RPKM of 1,200. The 10 highest RPKM values were in the range of 8,000 to 1,050,000. (E) Mature RNA species in high abundance are depleted by a custom hybridization subtraction protocol. Percentage of reads mapping to highly abundant RNA species before and after depletion are compared. As each primer can

(legend continued on next page)

target several species with identical 3′ end sequences, there are more depleted species than number of primers used in the subtraction protocol. (F) Distribution of reads mapping to microRNAs, Pol I, Pol II, Pol III and the mitochondrial RNA polymerase transcribed RNAs (mitoRNAs) in HeLa S3 NET-seq data. Multialigned reads were kept for the purpose of this analysis and weighted accordingly. (G) Evidence of co-purification of splicing intermediates. The branch point of lariat splicing intermediates will block the reverse transcription reaction and create a shorter cDNA product, leading to shorter insert sizes of sequencing reads. The insert size is the length of the sequencing reads after trimming the adaptor sequence when the RNA fragment is smaller than the read length. For each intron where ten or more reads align to the exact single nucleotide 3′ end, where the 3′ end of the lariat aligns, the insert size of the reads are recorded and the mean insert size is calculated. The result of the genome-wide analysis is plotted as a histogram. The red line indicates the mean insert size for all reads regardless of where they align. 85.7% of introns have reads mapping to the single nucleotide 3′ end that are shorter than the average NET-seq library insert size, indicating that NET-seq captures and sequences splicing intermediates. (H) Number of uniquely aligned reads per Pol II gene for two complete biological replicates generated from HEK293T cells (Pearson's correlation, R = 0.97). Genes with zero counts in only one of the replicates were added a 0.5 pseudocount to appear on the logarithmic scale and be taken into account for the correlation calculation. The dataset with higher coverage was randomly downsampled to match the total number of reads of the other dataset. RPKM, reads per kb per million Pol II uniquely aligned reads. RT, reverse transcription.

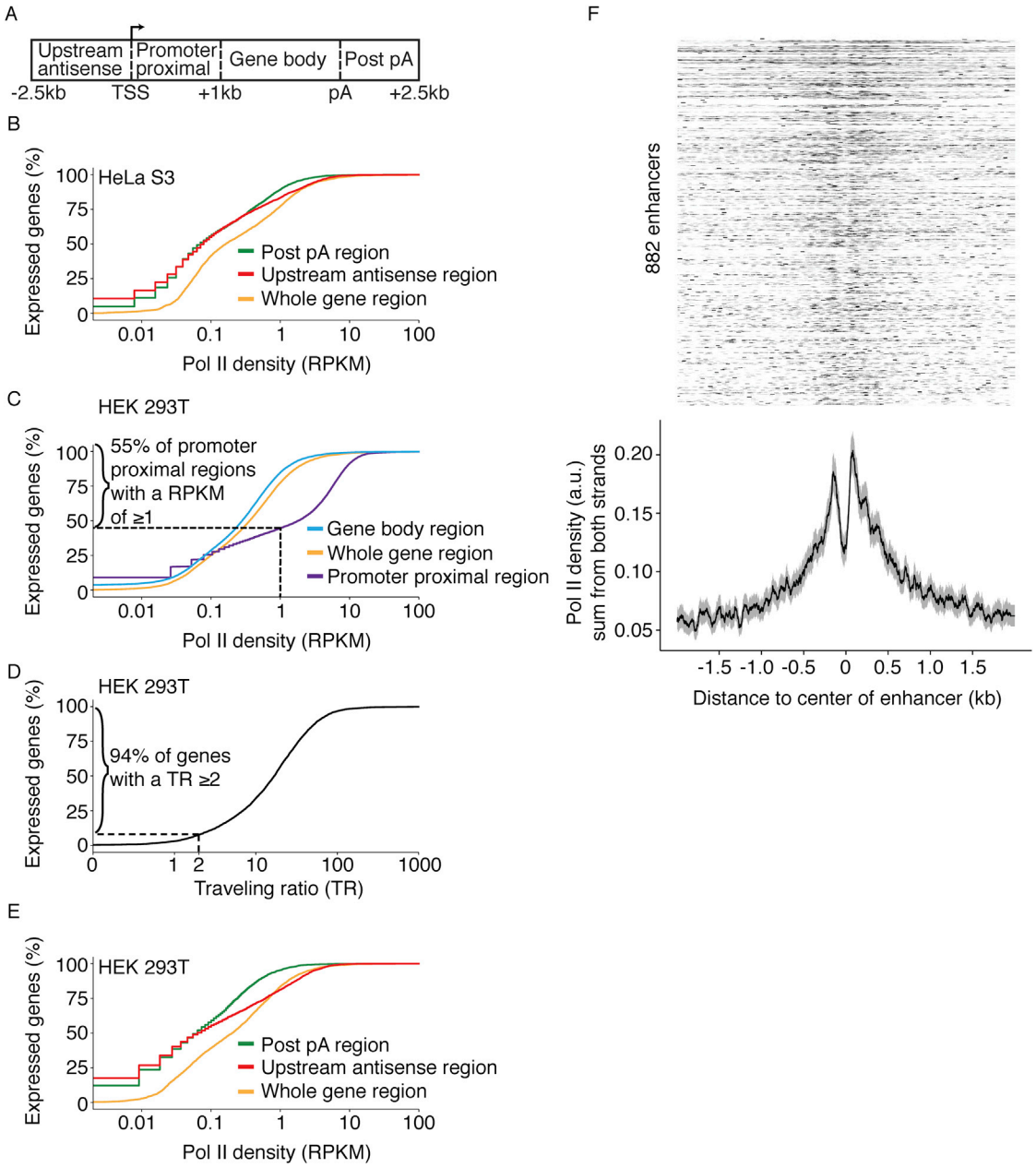


Figure S2. Distribution of NET-Seq reads, Related to Figure 2

(A) Schematic defining gene regions used in analysis of NET-seq data. −2.5kb and +1kb refers to the TSS; +2.5kb refers to the pA site; (B) Distributions of the percent of expressed Pol II transcribed protein coding non-overlapping (within 2.5kb upstream and downstream) genes (n = 4,309) with a given Pol II coverage for post pA regions and upstream antisense regions, as defined in Figure S2A for HeLa S3 cells. (C) Distributions of the percent of expressed Pol II transcribed protein coding genes (n = 18,779) with a given Pol II coverage for different gene regions as defined in Figure 2A for HEK293T cells. 55% of promoter-proximal regions have an RPKM of 1 or greater. (D) Distributions of the percent of well expressed Pol II protein coding genes (n = 8,749) with a given traveling ratio in HEK293T cells. Well expressed genes are defined as those genes with an RPKM of 1 or greater in the narrow promoter-proximal region (−30 bp to +300 bp of the transcription start site (TSS)). Traveling ratio is defined as the RPKM of the narrow promoter proximal region divided by the RPKM of the gene body region. 94% of genes have a traveling ratio of 2 or higher indicating that a vast majority of genes display promoter-proximal pausing. (E) Distributions of the percent of expressed Pol II transcribed protein coding non-overlapping (within 2.5kb upstream and downstream) genes (n = 4,162) with a given Pol II coverage for post pA regions and upstream antisense regions, as defined in Figure S2A for HEK293T cells. (F) Heat map shows the sum of Pol II density for both strands in selected enhancers, n = 882. NET-seq signal from each enhancer region (+/− 2 kb from center) is normalized from 0 to 1 (white to black scale in the heatmap) and smoothed by a 50 bp sliding window average. The average Pol II density profile is displayed below the heat maps. Solid line indicates the mean values and shading shows the 95% confidence interval. TSS, transcription start site; pA, polyadenylation site; RPKM, reads per kb per million Pol II uniquely aligned reads.

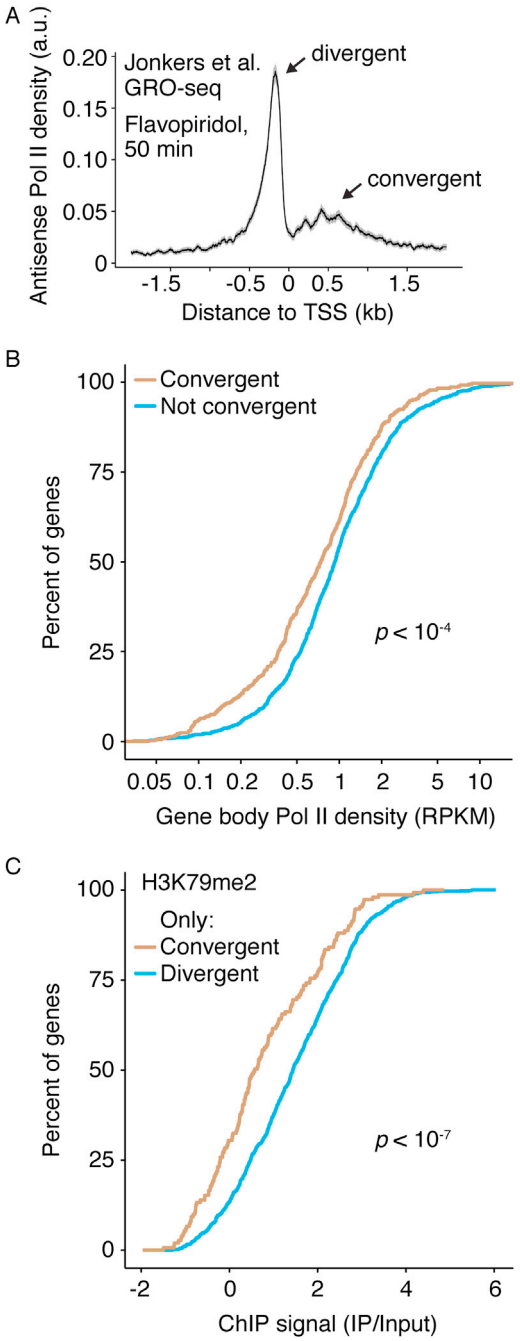


Figure S3. Histone Marks at Promoters, Related to Figure 3

(A) Antisense GRO-seq signal, in mouse embryonic stem cells treated with flavopiridol for 50 min, from each Pol II protein-coding non-overalapping (+/− 2.5 kb upstream and downstream) gene's promoter regions (+/− 2 kb from TSS), n = 4,200, is normalized from 0 to 1 and smoothed by a 50 bp sliding window average. Pol II density metagene profile is displayed. Solid line indicates the mean values and shading shows the 95% confidence interval. (B) Distributions of the percentage of genes with a given Pol II density in the gene body region, as defined in Figure 2A. Two sets of genes are compared, those that display convergent transcription in their promoter-proximal regions (yellow, n = 373), and those without convergent transcription in their promoter-proximal regions (blue, n = 931). (C) Distributions of the percentage of genes with a given H3K79me2 ChIP-seq signal fold change over input in the gene body region. The fold change is calculated as the difference of the ChIP-seq and input RPKM (log2). Data is from the ENCODE project (GEO accession number GSM733669) (Consortium, 2012). Two sets of genes are compared, those that display only convergent transcription in their promoter-proximal regions (yellow, n = 151) and those that display only divergent transcription in their promoter-proximal regions (blue, n = 931). The *p* value is calculated by the Kolmogorov–Smirnov test. The *p* value is calculated by the Kolmogorov–Smirnov test. TSS, transcription start site; RPKM, reads per kb per million Pol II uniquely aligned reads.

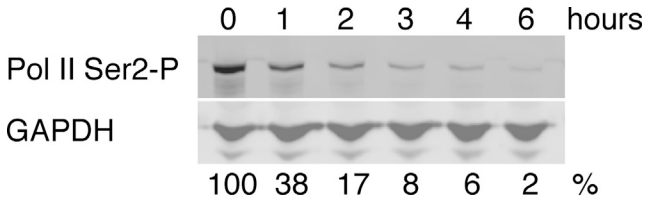


Figure S4. Flavopiridol Reduces Pol II CTD Ser2-P Levels, Related to Figure 5

Cell lysates were prepared from HeLa S3 cells before and 1, 2, 3, 4 and 6 hr after treatment with 1 μM flavopiridol and probed with antibodies directed against Pol II Ser2-P (3E10, Active Motif; first panel) and GAPDH (second panel). The percentage at the bottom is the amount of Pol II Ser2-P (as determined by image quantification) before and at the different time points after flavopiridol treatment. GAPDH serves as a loading control.

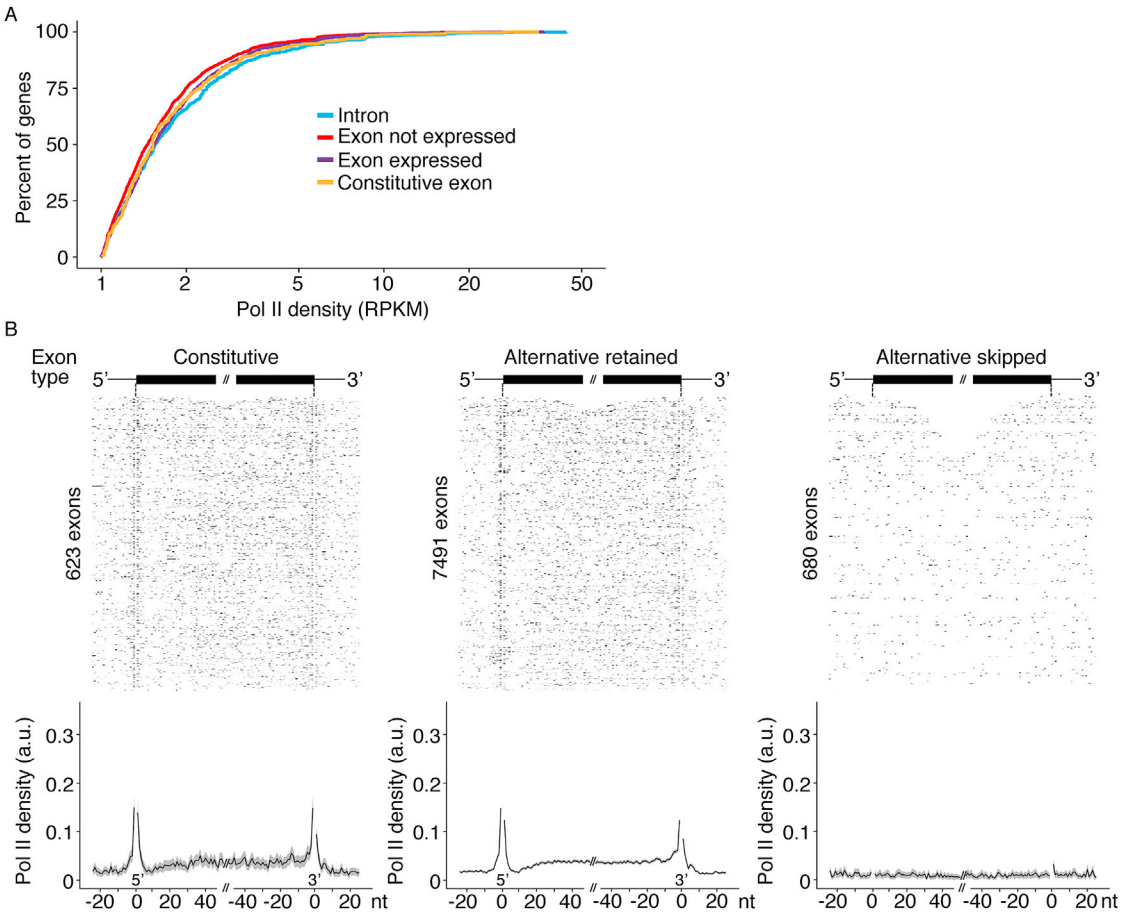


Figure S5. Meta-Exon Analysis of HEK293T Cells, Related to Figure 7

(A) Percentage of genes in HeLa S3 cells with a given Pol II density in gene body region as defined in Figure 2A. The genes are the one that contain the different types of exons and introns selected for Figure 7. (B) Heat maps show HEK293T Pol II density across constitutive exons, n = 623, alternative retained, n = 7,491 and alternative skipped, n = 680. The same criteria than used in HeLa S3 were used for exon selection (see Experimental Procedures). NETseq signal is normalized from 0 to 1 (white to black scale in the heatmaps) for each exon (+/− 25 bp). The average Pol II density profile is displayed below the heat maps. Solid lines indicate the mean values and shading shows the 95% confidence interval. The single nucleotide positions where splicing intermediates align (3′ ends of introns and exons) were entirely removed from analysis (see Experimental Procedures) and appear as a blank position in the figures. RPKM, reads per kb per million Pol II uniquely aligned reads.