Cell Reports
Resource



# Defining the Status of RNA Polymerase at Promoters

Leighton J. Core,[1] Joshua J. Waterfall,[1,4] Daniel A. Gilchrist,[2] David C. Fargo,[3] Hojoong Kwak,[1] Karen Adelman,[2] and John T. Lis[1,*]
[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
[2]Laboratory of Molecular Carcinogenesis
[3]Laboratory of Integrated Bioinformatics
National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, USA
[4]Present address: Genetics Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA
*Correspondence: jtl10@cornell.edu
http://dx.doi.org/10.1016/j.celrep.2012.08.034

## SUMMARY

**Recent genome-wide studies in metazoans have shown that RNA polymerase II (Pol II) accumulates to high densities on many promoters at a rate-limited step in transcription. However, the status of this Pol II remains an area of debate. Here, we compare quantitative outputs of a global run-on sequencing assay and chromatin immunoprecipitation sequencing assays and demonstrate that the majority of the Pol II on *Drosophila* promoters is transcriptionally engaged; very little exists in a preinitiation or arrested complex. These promoter-proximal polymerases are inhibited from further elongation by detergent-sensitive factors, and knockdown of negative elongation factor, NELF, reduces their levels. These results not only solidify the notion that pausing occurs at most promoters, but demonstrate that it is the major rate-limiting step in early transcription at these promoters. Finally, the divergent elongation complexes seen at mammalian promoters are far less prevalent in *Drosophila*, and this specificity in orientation correlates with directional core promoter elements, which are abundant in *Drosophila*.**

## INTRODUCTION

Transcription regulation is a major and primary mode by which developmental, nutritional, and environmental signals control gene expression. This regulation must ultimately target the activity of RNA polymerase II (Pol II), which encodes all mRNAs and many critical noncoding RNAs. Chromatin immunoprecipitation (ChIP) studies in *Drosophila* and mammals have shown that Pol II accumulates disproportionately at a large fraction of promoters relative to downstream gene regions (Baugh et al., 2009; Guenther et al., 2007; Muse et al., 2007; Zeitlinger et al., 2007), thereby identifying what appears to be a rate-limiting step in transcription. At least a portion of the accumulated Pol II at promoters has initiated transcription (Core et al., 2008; Nechaev et al., 2010), but whether this polymerase is predominantly bound and uninitiated in a preinitiation complex (PIC) with general transcription factors (Juven-Gershon et al., 2008) or exists as an elongation complex proximal to the promoter requires a quantitative analysis. Additionally, accumulated Pol II at promoters could be either paused, transcribing and undergoing rapid cycles of initiation and termination, or backtracked to an arrested state that is incapable of elongation. A quantitative determination of which of these forms of polymerase predominates at a given gene promoter would provide a basis for understanding how that gene is regulated; however, no single assay determines this in vivo.

Two assays that are commonly used to examine the density of polymerases along DNA are the ChIP assay and the nuclear run-on (NRO) assay. The ChIP assay can quantify Pol II levels across the genome, but it cannot distinguish whether Pol II is transcriptionally engaged, backtracked and arrested, or bound in a PIC, nor can ChIP assess the orientation of engaged polymerases. NRO assays measure polymerases that are transcriptionally engaged and competent to elongate and can determine the direction of transcription (Lis, 1998), but, on their own cannot determine what fraction of the total polymerase present at a given location is in this form. Also, engaged polymerases could be transiently passing through the promoter or could be stably held in a paused state as seen at the extensively characterized *Drosophila Hsp70* gene (Lis, 1998), and the human *c-myc* gene (Krumm et al., 1995; Strobl and Eick, 1992). At these promoters, the paused Pol II is thought to be physically held back since conditions that disrupt protein-protein and protein-DNA interactions, but do not affect transcriptionally engaged polymerases (i.e., high concentrations of salt or addition of the detergent Sarkosyl) are required for efficient run-on transcription of promoter-proximal Pol II (Hawley and Roeder, 1985; Rougvie and Lis, 1988). These inhibitory interactions led to the hypothesis that this step is likely to be regulated in vivo (Rougvie and Lis, 1988), and is now consistent with our current knowledge of the mechanism of promoter-proximal pausing: Pol II is held paused by the cooperative action of Spt5 and negative elongation factor (NELF) protein complexes. Regulated recruitment of positive elongation factor b (P-TEFb), alleviates this negative block, resulting in escape of Pol II from the pause site and entry into productive elongation (Nechaev and Adelman, 2011). However, not all promoters have been characterized to extent of the *Hsp70* gene, making it difficult to extrapolate these characteristics of the *Hsp70* promoter to other genes.

We developed a sensitive global run-on sequencing assay (GRO-seq) that maps the position, amount, and orientation of transcriptionally engaged polymerases genome wide (Core et al., 2008). Application of GRO-seq to a human primary cell line showed transcription occurring within 70% of genes, with 40% of these genes experiencing a significant accumulation of promoter-proximal polymerase that has properties of transcriptionally paused Pol II. We also observed that the majority of active promoters in human cells have a peak of transcriptionally engaged polymerase that is upstream and divergent relative to the annotated gene. This finding has initiated a debate over whether these upstream divergent transcripts are functional, or if they instead represent aberrant, ''sloppy'' transcription initiation events that result from open promoter chromatin (Buratowski, 2008; Seila et al., 2009).

Here, we used GRO-seq in *Drosophila* S2 cells to assess the genome-wide transcription pattern and characterize promoters. Our GRO-seq data show that transcription is tightly associated with annotated genes, with very little evidence of complete genomic transcription or initiation at 3′ ends of genes. We also report, as suggested elsewhere (Nechaev et al., 2010), that *Drosophila* promoters generally lack divergently engaged Pol II seen at the majority of human promoters. In this work, we show evidence that a well-known DNA element can specify increased directionality at human promoters, thereby providing a simple explanation for the strong directionality in *Drosophila* promoters, which are inherently rich in orientation specific elements (FitzGerald et al., 2006). To then quantify the status of polymerase at promoters, we use a normalized comparison of the polymerase densities at promoters as seen by ChIP-seq and GRO-seq, to conclude that the majority of polymerases at promoters are transcriptionally engaged and competent for elongation under steady state conditions. Moreover, we find that paused polymerases are physically tethered or blocked at promoters as they transcribe efficiently only in the presence of the anionic detergent sarkosyl. These observations establish not only that pausing occurs at most promoters, but that the predominant form of Pol II at promoters is paused in a manner that is similar to pausing at the *Drosophila Hsp70* gene. Altogether, these observations provide a framework with which to study transcription factor function during basal and activated states.

## RESULTS

### Transcription Is Predominantly Associated with Annotated Genes

We performed GRO-seq assays under several conditions in *Drosophila* S2 cells (Table S1). Under standard conditions that detect all transcriptionally competent polymerases, 67% of engaged polymerases occupy the sense strand of gene annotations, and 15% occupy the antisense strand of annotated genes (82% of total)(Figures S1 and S2). These numbers increase to 78% and 19% (98% of total), respectively, if gene boundaries are expanded by 0.5 kb. Thus, as we reported with a human primary lung fibroblast line (Core et al., 2008) and mouse embryonic stem cells (Min et al., 2011), the vast majority of transcription in *Drosophila* is associated with annotated gene regions.

Debate of whether or not genomes are ''pervasively'' transcribed depends on different assays of accumulated RNAs (Kapranov et al., 2007; van Bakel et al., 2010) and on semantics. The GRO-seq assay, which has high sensitivity and low background (libraries are estimated to be >99% pure), measures the distribution of transcriptionally competent polymerases. The snapshot of transcriptome activity provided by GRO-seq does not depend on RNA processing rates or transcript stability. The assay reveals that the vast majority, 98%, of transcriptionally competent RNA polymerases are focused within or near currently annotated genes and these genes cover~46% of the genome. Thus, while our GRO-seq data do not deal with the sum of transcripts produced in multiple cell types, they do argue that any ''pervasive'' transcription of the genome in *Drosophila* S2 cells must occur at levels that are indistinguishable from the low background of our assay.

### *Drosophila* Promoters Are More Directional Than Mammalian Counterparts

Alignment of all reads relative to observed transcription start sites (TSSs) (Nechaev et al., 2010), or plotting of the distribution of sense versus antisense reads at promoters, revealed a prominent lack of divergent transcription at *Drosophila* promoters compared to human promoters (Figures 1B, 1C, and S1F). In support of this, 95% of promoter-associated reads map in the direction of the annotated gene at *Drosophila* promoters compared to 58% for human promoters. ChIP-seq and ChIP-chip data sets in human and mouse cells show that Pol II and histone marks associated with initiation coincide with divergent initiation upstream of TSSs (Seila et al., 2008; Core et al., 2008). Consistent with this, Pol II ChIP-seq and the H3K4me3 initiation mark are strongly associated with the direction of transcription at *Drosophila* TSSs (Figures S1C–S1E). In addition, in *Drosophila*, only unidirectional profiles are evident in data sets comprised of small, 5′-OH or 5′-capped RNAs (Nechaev et al., 2010; Taft et al., 2009). The GRO-seq data confirm that the inability to detect divergent transcription in small RNA pools is not due to preferential capping or processing of the nascent RNA in one direction versus the other since GRO-seq will detect nascent RNAs regardless of how the RNA end is modified. Likewise, failure to detect divergent transcription in GRO-seq is not due to an alternative form of Pol II that is undetectable by nuclear run-on. Combined, these results reinforce the notion that marks of initiation, such as H3K4me3, coincide with promoter direction (Seila et al., 2008; Core et al., 2008).

The position and direction of transcription initiation are specified by a variety of core promoter sequence motifs. *Drosophila* promoters are enriched for several directional motifs, whereas human promoters appear to be enriched mainly for nondirectional motifs and CpG islands (FitzGerald et al., 2006). To test the hypothesis that directional motifs in *Drosophila* may be responsible for specifying unidirectional transcription, we generated an orientation index (OI) for all human promoters. The OI is defined as the fraction of GRO-seq density at promoters that is orientated in the sense direction. We then compared the OI of human promoters that contain directional and nondirectional motifs identified in a comparative analysis between *Drosophila* and human promoters (FitzGerald et al., 2006). Of these motifs,

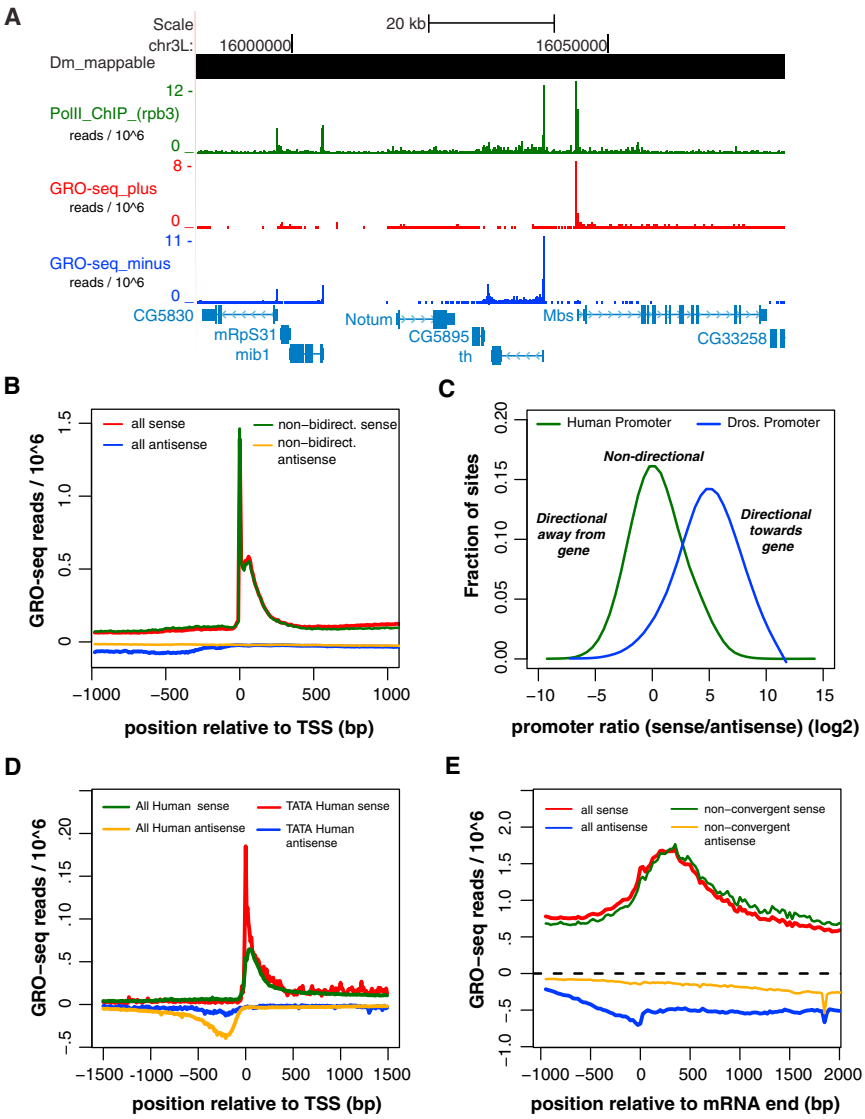


**Figure 1. RNA Polymerase Distribution on mRNA-Encoding Genes Using GRO-Seq**

(A) A representative view of GRO-seq data from S2 cells in the UCSC genome browser (Kent et al., 2002). GRO-seq reads (reads/base) aligning to the plus strand are shown in red; minus strand in blue. ChIP-seq for total Pol II (α-Rpb3) is shown in green (reads/25 bp bin), and gene annotations are shown at the bottom in blue.

(B) GRO-seq data aligned to transcription start sites (TSSs). For all genes, reads aligning to the sense strand of the gene are in red; antisense strand in blue. For nonbidirectional genes (head-to-head promoters within 1 kb removed), reads aligning to the sense strand of the gene are in green; antisense strand in orange.

(C) Comparison of directionality of *Drosophila* and human promoters. The distribution of the ratios of sense and antisense reads around promoters (log2) is plotted for active promoters (>25 reads) in IMR90 cells (green) and *Drosophila* S2 cells (blue). How different types of directionality of transcription from promoters are reflected in the ratio are indicated in italicized lettering.

(D) GRO-seq profiles from ±1.5 kb relative to TSS are shown for all human promoters (green, sense; orange, antisense) or human promoters that contain a TATA box (red, sense; blue, antisense).

(E) GRO-seq data aligned to gene end for all genes (red, sense; blue, antisense), and after convergent genes within 1.5 kb are removed (green, sense; orange, antisense).

See also Figures S1 and S2.

## RNA Pol II Accumulates at the Majority of Active Promoters and to a Lesser Extent at Gene Ends

Alignment of reads to the 3′ end of genes showed much smaller peaks in both the sense and antisense directions (Figure 1E). Neither peak at the 3′ end appears to be associated with genuine initiation at the 3′ end of genes, because there is no corresponding enrichment of small, capped RNAs (Figure S2). Thus, this 3′-sense peak likely represents Pol II that slows down after the polyadenylation signal is exposed. In support of this, the antisense peak is dramatically reduced when convergent genes are removed from the analysis (Figure 1E).

The striking accumulation of GRO-seq density in the promoter-proximal region indicates the existence of a rate-limiting step following transcription initiation. Accordingly, when we define active genes based on GRO-seq signal in gene bodies ($p$ value < 0.01, Fisher's exact test; Extended Experimental Procedures), we find that 6,044 of 9,544 (63%) of these genes have significantly enriched GRO-seq signal at the 5′ end ($p$ value < 0.01, Fisher's exact test). This fraction is likely an underestimate since overlapping transcription from neighboring genes can result in a false positive call for gene transcription when the actual promoter is not active. When we use 7,336

the TATA box (TATAWAAR) (Juven-Gershon et al., 2008), is the only one to show a clear bias toward unidirectional transcription at human promoters (OI = 0.86 compared to OI = 0.57 for all promoters) (Figure 1D; Table S2). Interestingly, the composite profile at human TATA-containing promoters more closely resemble *Drosophila* promoters (Figure S1F). We also found that promoters with a TATA box embedded within a CpG island also produce directional transcription (Figure S1G), suggesting that the TATA box can act dominantly in the context of human CpG islands to enhance initiation in the direction of the gene. However, because only 5%–20% of *Drosophila* and mammalian promoters contain an identifiable TATA box (Fitz-Gerald et al., 2006; Kutach and Kadonaga, 2000; Sandelin et al., 2007), it is likely that other DNA elements or protein factors that specify unidirectional transcription in *Drosophila* are either not present or not functional in the context of mammalian promoters.

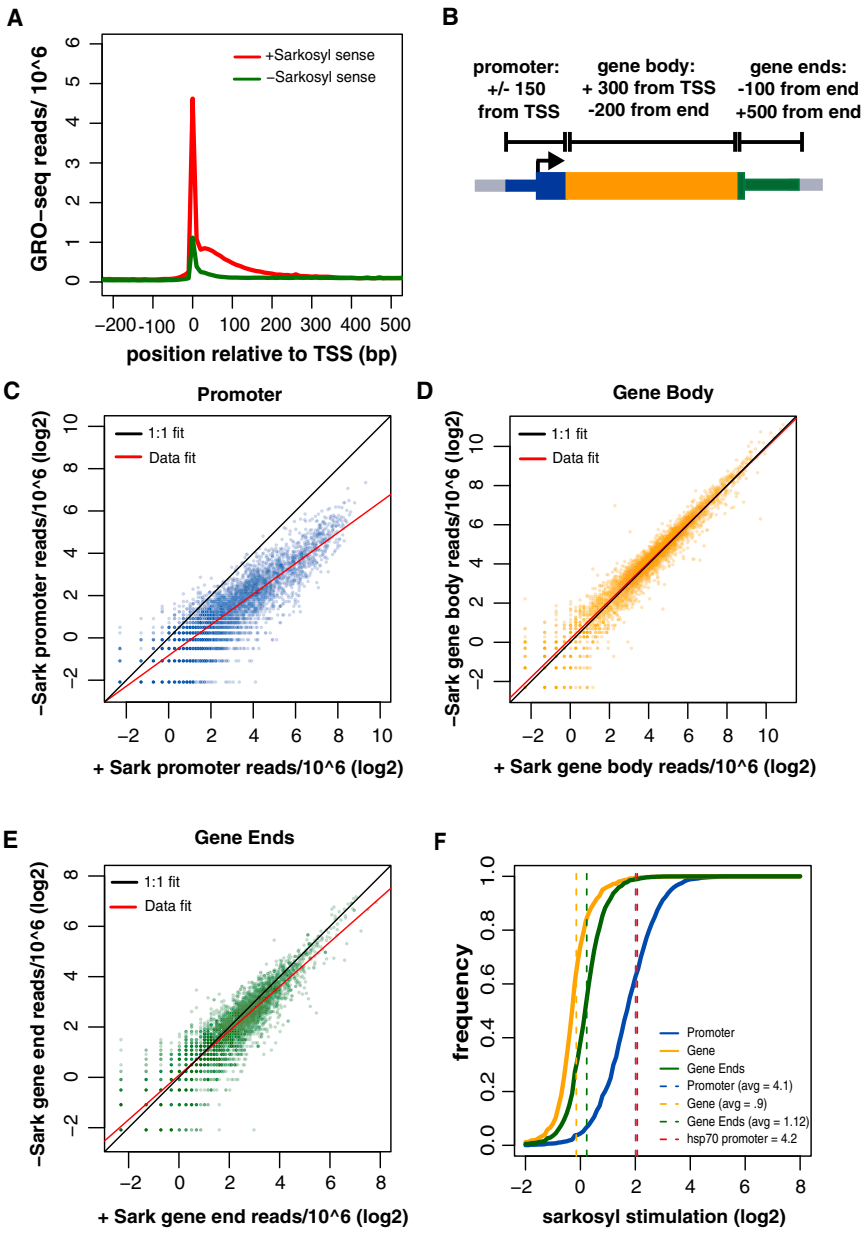


**Figure 2. Sarkosyl-Dependent Run-ons Identify Distinct Forms of Polymerase at Promoters Relative to Downstream Gene Regions**

(A) Composite profile of GRO-seq data showing the density reads in 10 bp windows from −200 bp to +500 bp relative to TSSs for run-ons performed with or without sarkosyl. y axis represents read/window/million reads sequenced. The number of genes shown = 11,800.

(B) Schematic showing how GRO-seq signal was quantified at promoters, the gene body, or at gene ends. After removing genes based on overlaps and filtering for genes that have active promoters, the number of genes for this analysis = 4,652.

(C–E) Scatter plots showing the effects of sarkosyl on run-on signal in promoters (C) or genes (D), or at gene ends (E).

(F) Cumulative distribution plots showing the differential effect of sarkosyl at promoters (blue) versus within genes (orange), or at gene ends (green). The average effect in the gene, promoter, and the *Hsp70* promoter are denoted by the hashed vertical lines. The effect at the Hsp70 promoter is shown as a hashed vertical line in red. The nonlogged value for the fold effect after sarkosyl stimulation is shown in the legend.

See also Figure S3.

promoters defined as active by sequencing small, capped RNAs from nuclei (>10 reads within ±50 bases from TSS) (Nechaev et al., 2010), or 3,168 promoters called bound by Pol II from a ChIP-seq experiment (Nechaev et al., 2010), we find that 5,166 (70%) and 2,784 (89%) of promoters, respectively, show significantly enriched Pol II in our GRO-seq analysis. Thus, post-initiation regulation occurs at the majority of promoters that show signs of Pol II binding or transcription activity. These polymerases that accumulate at promoters could be in the form of stably paused polymerases or polymerases that are actively transcribing within the promoter region, for example, undergoing cycles of initiation and rapid early termination. Thus, we sought to distinguish these two forms.

## Engaged Pol II Is Physically Tethered to Promoters

A genuinely paused Pol II that is held near the promoter by pause stabilizing factors requires the disassociation of these factors by the addition of high salt or the anionic detergent sarkosyl to resume transcription in a run-on assay (Rougvie and Lis, 1988). In contrast, Pol II that is undergoing active elongation transcribes efficiently with or without high salt or sarkosyl (Hawley and Roeder, 1985, Rougvie and Lis, 1988). We therefore produced matched GRO-seq data sets in the presence or absence of sarkosyl to test for pausing genome-wide. Our results show that run-on signal at nearly all promoters is dependent on sarkosyl, with the average promoter showing an ~4-fold increase in signal in the presence of sarkosyl (Figures 2A–2C, 2F, and S3). In contrast, read densities in gene bodies are unaffected by sarkosyl (Figures 2D and 2F). The stimulation by sarkosyl at gene ends (1.12-fold, Figures 2E and 2F) was much less pronounced than at promoters, indicating that the slowing down of polymerase near gene ends immediately prior to termination occurs through a different mechanism than pausing at promoters.

Interestingly, the effect of Sarkosyl at the *Hsp70* gene in the GRO-seq data set is equivalent to the genome-wide average (Figure 2F). Thus, the majority of promoters in the *Drosophila* genome behave in a manner similar to the *Hsp70* gene, which

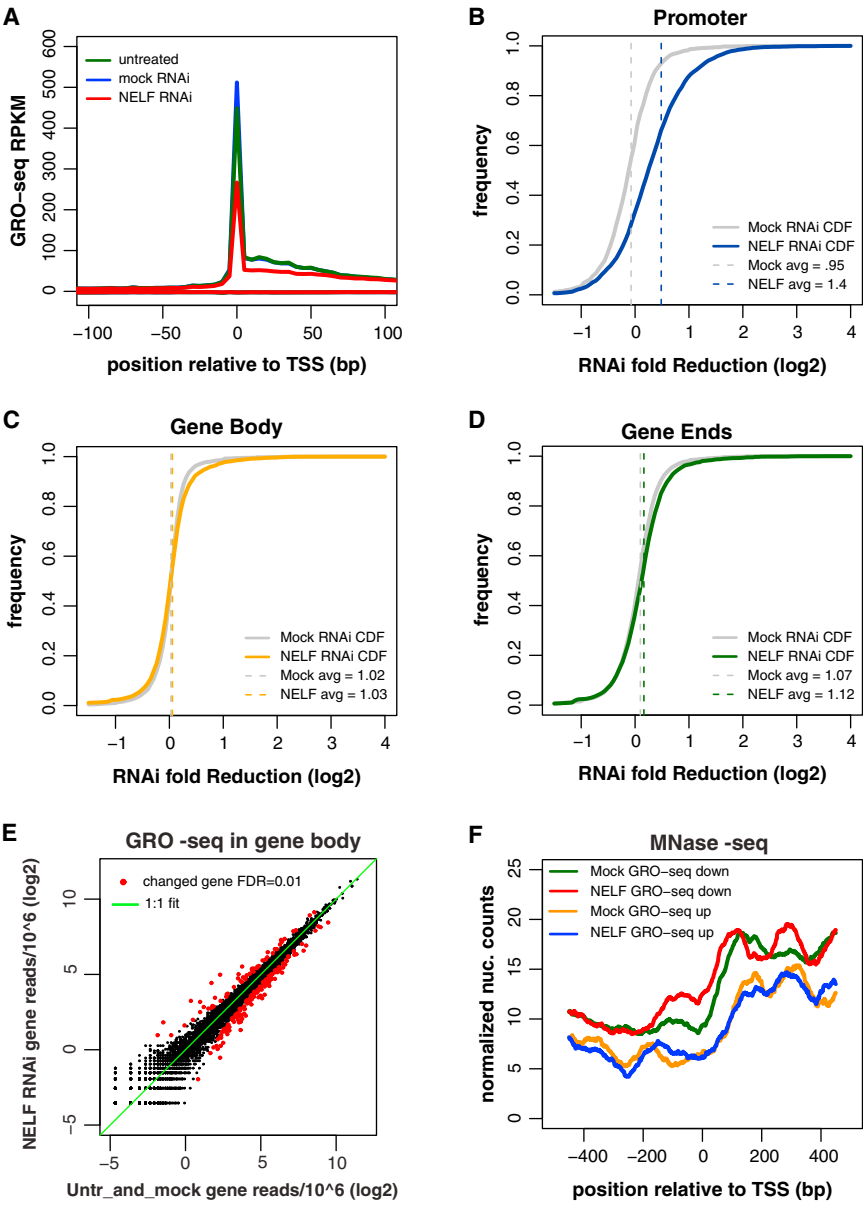


**Figure 3. Use of GRO-Seq to Examine Function of NELF at Promoters and Identification of Genes Affected by NELF Knockdown**

(A) Composite profile of GRO-seq data showing the density reads in 10 bp windows from ±100 bp relative to TSSs for untreated (green), mock- (blue), and NELF-depleted (red) cells. The number of genes shown = 11,800.

(B–D) Cumulative distribution plots showing the overall effect of NELF RNAi on polymerase density in promoters (B) or genes (C), or at gene ends (D). Figure panels and legends are displayed as in Figure 2.

(E) Scatter plot showing the comparison of GRO-seq signal in the gene body region in mock – or RNAi – treated cells. Genes were identified as significantly affected (red) using edgeR (Robinson et al., 2010), with an FDR of 0.01. The green line represents a 1:1 fit.

(F) MNase-seq patterns relative to TSSs in mock, or RNAi, -treated cells for genes that are either up or downregulated after NELF-RNAi (identified in E). Genes that are upregulated after NELF RNAi (orange and blue) have overall lower nucleosome density around their promoters than genes that are downregulated (red and green). As seen previously (Gilchrist et al., 2010) genes that are downregulated by NELF RNAi have increased encroachment of nucleosomes over the TSS after NELF RNAi.

See also Figure S4.

has served as a classic gene model for regulation through Pol II pausing. These results indicate not only that a high degree of stable pausing likely occurs at most promoters, but also that transcription elongation is inherently different at promoters versus downstream regions (Pal et al., 2001; Saunders et al., 2006). This implies that regulatory mechanisms are in place to control the level of pausing, presumably by modulating interactions that retain stably paused Pol II or release it into productive elongation.

## NELF Increases Promoter Occupancy of Paused Pol II

If promoter-proximal pausing is a rate-limiting step in transcription governed by the interactions of pausing factors with the transcribing complex, then we expect that disruption of a factor involved in stabilizing the paused complex would reduce the accumulation of Pol II in the promoter-proximal region (Muse et al., 2007; Wu et al., 2003; Yamaguchi et al., 1999). Indeed, RNA interference (RNAi) knockdown of NELF leads to a general decrease in the GRO-seq signal on promoters (Figure 3A and 3B) relative to gene bodies and 3′ ends (Figures 3B–3D and S4). This moderate decrease in Pol II at promoters following NELF knockdown is not surprising, because residual NELF, or its partner DRB sensitivity inducing factor (DSIF), could still be sufficient to induce pausing (Figure S4A).

The reduction of Pol II at promoters after NELF RNAi could be accounted for by either increased escape of polymerase into the gene without immediate entry of a new polymerase into the pause site, or by decreased initiation due to increased nucleosome occupancy at promoters (Gilchrist et al., 2010). Previous studies relying on ChIP-chip have been unable to determine conclusively at which genes the reduced amount of Pol II at promoters is due to increased escape of Pol II into the gene, or decreased initiation (Gilchrist et al., 2010). The highly sensitive GRO-seq assay can detect both significant increases and decreases in the polymerase density in the downstream portion of genes (Figure 3E; Table S3). Since GRO-seq measures nascent RNA transcription, the significantly changed genes are

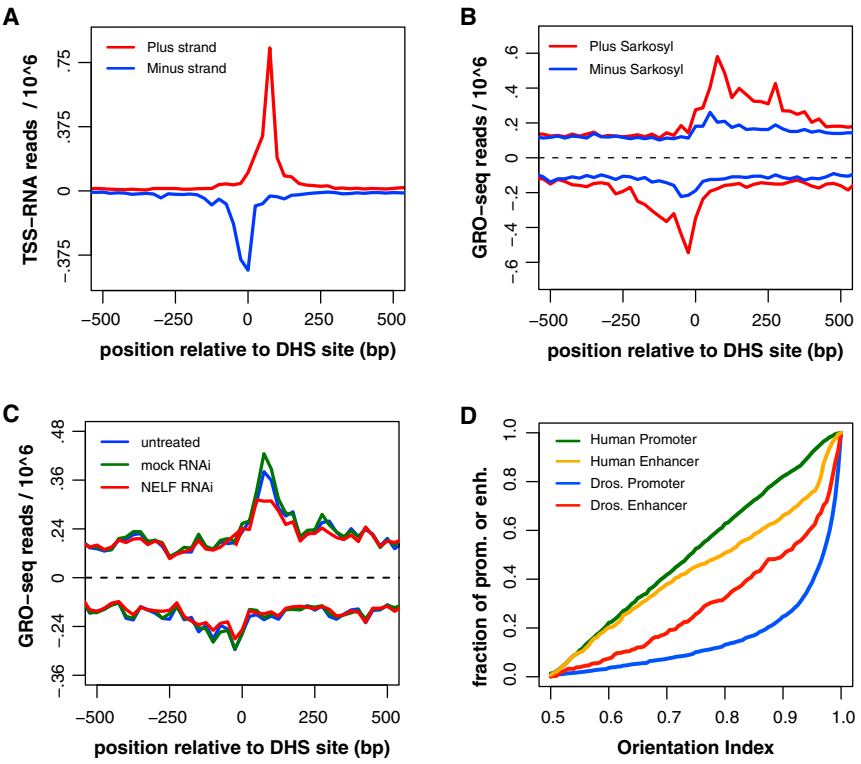


**Figure 4. Pausing and Directionality of RNA Polymerase at Enhancers**

(A) A composite profile of TSS-RNA reads (Nechaev et al., 2010) surrounding putative *Drosophila* enhancers as identified by (Kharchenko et al., 2011), n = 533. Data are plotted relative to the DNase hypersensitive site (DHS) site in 25 bp bins, and the y axis is in reads/bin/million reads sequenced.

(B) GRO-seq data around the same sites in the presence and absence of sarkosyl. The positive signal (above dotted line) is from the plus stand, the negative signal; the minus strand.

(C) GRO-seq data set after RNAi of the pausing factor NELF.

(D) Plots of the distribution of orientation indexes for human promoters (green) and enhancers (orange); and *Drosophila* promoters (blue) and enhancers (red). For ease of comparison between promoters and enhancers, the "direction" of the promoter or enhancer is defined by the strand (plus or minus) with the greatest intensity. Thus, the orientation index here will be equal to or greater than 0.5. See also Figure S5.

more likely to be directly affected by NELF RNAi than those identified by microarray, providing a high confidence gene list with which to investigate the molecular phenotypes of NELF knockdown, and the effects on promoter chromatin. Therefore, we examined the effect of NELF RNAi on MNase-seq pattern around promoters of genes that were identified as up- or downregulated by GRO-seq. As seen previously, downregulated genes have increased nucleosome density at the promoter (Figure 3F), consistent with the model that a paused polymerase competes with nucleosomes for occupancy of some promoters (Gilchrist et al., 2010). In contrast, the MNase pattern at upregulated genes does not change after NELF knockdown, and these promoters have an overall lower level of nucleosome occupancy before or after NELF RNAi (Figure 3F). These data indicate that each promoter has an inherent propensity to displace or position nucleosomes around the promoter and this influences the net effect on transcription caused by removing a pausing factor.

## Pol II Pauses at Enhancers

Transcripts originating from enhancers, or eRNAs, are a newly identified class of RNAs with unknown regulatory functions (Kim et al., 2010). Transcription at enhancers is associated with active enhancers, and the resulting eRNAs can emanate bidirectionally from enhancers. The eRNAs can be spliced and polyadenylated, but they have little coding potential (Kim et al., 2010; Wang et al., 2011). Enhancers themselves can be found within or outside of genes and are enriched in monomethylation of histone 3 at lysine 4 (H3K4me1) but have lower levels of trimethylation at the same site (H3K4me3) (Heintzman et al., 2007). In contrast, active promoters are highly enriched with H3K4me3 (Kim et al., 2005). To characterize the status and directionality of polymerase at *Drosophila* enhancers, we examined putative intergenic enhancers as identified by the ModENCODE group (Kharchenko et al., 2011). 5′-RNA sequencing (Nechaev et al., 2010) provides evidence of initiation and pausing at these sites (Figure 4A). In addition, the polymerase at enhancers appears similar to that at promoters in that it is stimulated by sarkosyl during the run-on (Figure 4B), colocalizes with NELF (Figure S5), and has reduced occupancy after NELF RNAi (Figure 4C).

Given that human promoters and enhancers both produce divergent transcripts, we compared the orientations of polymerase for *Drosophila* and human enhancers (Figure 4D). Since enhancers do not have inherent directionality, we specified the "direction" of the enhancer or gene to be the strand with the highest signal, making all OIs >0.5 for this analysis. Interestingly, the distribution of OIs at *Drosophila* enhancers resembles a mixed distribution, with many showing strong directionality and a similar number appearing to be bidirectional (Figure 4D). Since the putative directional enhancers could be a result of non-annotated promoters, it is difficult to say whether this represents the true distribution of enhancer orientations. Nonetheless, it appears that *Drosophila* enhancers could more closely resemble human enhancers in their directionality (or lack thereof), and emphasizes that there is some difference between *Drosophila* enhancers and promoters.

## Majority of Pol II at Promoters Is Engaged and Competent for Transcription

Limitations of currently available assays have prevented a quantitative characterization of the form of Pol II at promoters. The ChIP assay cannot distinguish between Pol II that is in a PIC, paused, or backtracked and arrested. Sequencing of small (<100 nt) RNAs from nuclei can identify RNAs generated by

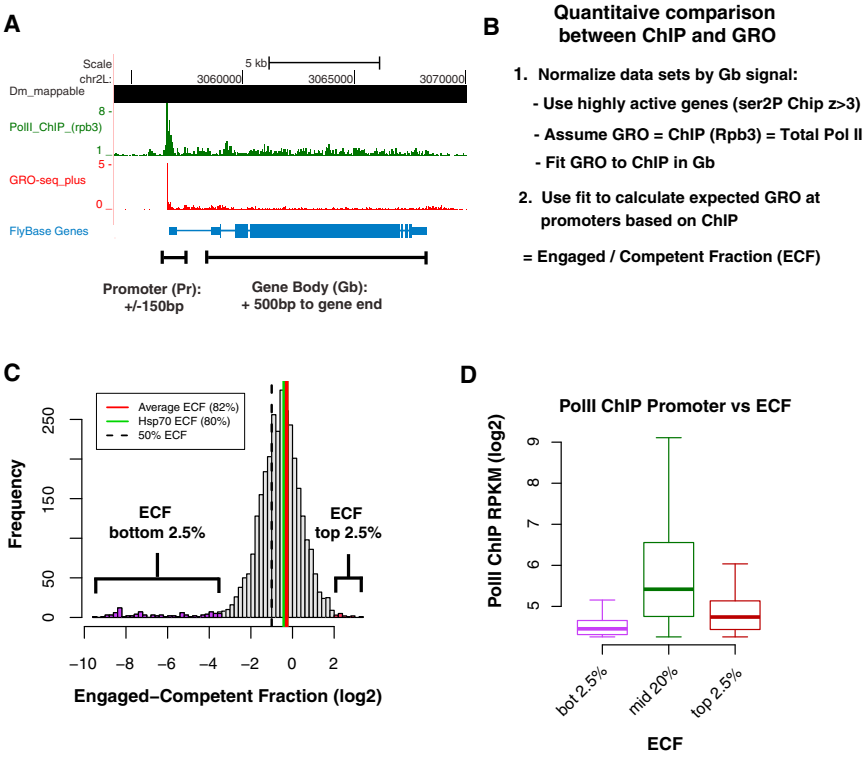


**Figure 5. Pol II at Promoters Is Predominantly Engaged and Competent for Elongation**

(A) Representative browser shot showing Pol II Chip-seq (green) and GRO-seq (red) with y axis in reads/bp/10exp6. The regions used for calculating the engaged and competent fraction (ECF) at promoters are indicated below.

(B) Schematic explaining the workflow used to calculate the ECF for Pol II at promoters.

(C) Histogram showing the distribution of ECF values for significantly bound promoters (n = 3,168). The vertical lines represent a 50% (black), the average (red), and the Hsp70 (green) ECFs. Promoters with the lowest ECFs are highlighted in purple.

(D) Boxplots showing Pol II ChIP-seq levels at promoters with different ranges of ECF. Promoters with the lowest (purple) and the highest (dark red) ECF values have less Pol II bound at promoters in ChIP-seq experiments than promoters with less extreme ECF values (middle 20% shown), suggesting that the ChIP and GRO discrepancies here could be due to experimental noise.

The box spans the first quartile (Q1, bottom) to third quartile (Q3, top), the horizontal line in the box represents the median, and the whiskers extend as follows: (Q1 or Q3 + 1.5 )*(Q3-Q1). See also Figures S6 and S7.

Pol II, but can't discern between Pol II that have paused, arrested, or terminated. GRO-seq can only detect Pol II that is engaged in transcription with the 3′ end of the nascent RNA in register with the active site and competent to transcribe during a nuclear run-on assay. Notably, promoter signals from each of the three assays correlate very well (Figures S6A–S6C), but these correlations alone do not explicitly identify the major form of Pol II at promoters. For instance, in a population of cells, a promoter could contain a PIC in some cells and a paused Pol II in others. Thus, in an ensemble-type assay like ChIP- and GRO-seq, it is possible that one could see a peak of Pol II at promoters in GRO-seq, even though in most cells the polymerase was still in a PIC. Determining which is the predominant form is a critical distinction for understanding how gene regulation works.

We reasoned that a more quantitative comparison of ChIP-seq and GRO-seq signals at promoters would reveal what fraction of the ChIP signal at promoters is represented by engaged and elongation-competent Pol II. As an internal standard, we used the ChIP-seq and GRO-seq signal in the body of the gene to normalize the gene-specific signal for each assay (Figures 5A and 5B). Because of the presumably high background in the ChIP-seq data (Figures S6 and S7), we focused on genes with highest levels of ser2-P ChIP signal ($Z$ score > 3), assuming these will contain the highest densities of transcribing Pol II over background. Good quantitative agreement between GRO-seq and total Pol II ChIP-seq levels in these 1,874 genes suggests that the ChIP-seq signal here represents engaged polymerases complexes that are competent for transcription (Figure S6D). With this gene set, we generated a conversion factor that was then used to calculate the fraction of the total Pol II at promoters that can be accounted for by the GRO-seq signal. We call this fraction the engaged/competent fraction (ECF). Approximately 80% of the polymerase found by ChIP-seq can be accounted for by the signal from the GRO-seq data set (average ECF = 0.82, Figures 5B–5D). We identified candidate promoters that were likely to contain PICs in the leftward tail in the ECF distribution (Figure 5C). However, these promoters are likely false positives, because outliers on both ends of the distribution (top and bottom 2.5%, ECF < 0.06, ECF > 2.5) have low levels of Pol II binding as seen in ChIP-seq (Figure 5D). In cases where the relative ChIP-seq signal is greater than GRO-seq at promoters, the "noncompetent" polymerase could be in the process of forming a functional PIC or could be backtracked and arrested. However, since the data fit a normal distribution around the mean and there are theoretically impossible instances where relative GRO-seq signal at promoters is greater than the ChIP-seq, we believe that the major discrepancies between the two assays are due to inherent experimental noise or counting biases associated with next-generation sequencing. We therefore conclude that the major form of Pol II found at promoters by ChIP is engaged and competent for elongation.

We also compared the promoter ECF with several other data sets, including level of association of TFIIA, NELF, or SPT5 with promoters as measured by ChIP or levels of TSS RNAs, NELF RNAi sensitivity, sarkosyl sensitivity, or the presence of promoter elements and were unable to identify candidate PICs (Figure S7; data not shown). In all data sets, the genes that are the most likely candidates for PICs (i.e., those with the lowest ECF), displayed signals approaching background, further suggesting that these genes are false positives and result from noise

inherent to the low signal range. However, if these candidate promoters truly maintain a PIC, they do so at a very low occupancy compared to the occupancy of a paused polymerase. Taken together, these data argue against the notion of a stable preinitiation complex and indicate that once Pol II is recruited to a promoter, it rapidly initiates RNA synthesis and undergoes pausing.

## DISCUSSION

### Unlike Mammals, *Drosophila* Promoters Lack an Upstream Divergent Peak of Pol II

Here, we have mapped the nascent transcriptome of *Drosophila* S2 cells using GRO-seq. A striking difference between the *Drosophila* and human transcriptomes is the lack of divergent transcription at *Drosophila* promoters. *Drosophila* has a collection of directional core promoter elements that serve to direct the transcription complex to the promoter (Juven-Gershon et al., 2008). We searched for several of these directional elements in human promoters and found that the most were either not prevalent or were nonfunctional because the corresponding protein that binds the element does not exist. Interestingly, the one core element that is present in a subset of human promoters, the TATAWAAR box, does correlate with a subclass of human promoters that show unidirectional transcription. This supports a model where core promoter elements are powerful directors of Pol II direction at a promoter. Human promoters are predominantly characterized by unmethylated CpG islands that by themselves do not specify orientation.

Our analysis of *Drosophila* enhancers reveals that the polymerase initiates and pauses at these locations. In *Drosophila*, an interesting difference from promoters is that a higher proportion of enhancers can produce bidirectional transcription. Thus, transcription from human and *Drosophila* enhancers appears to be more similar than their promoter counterparts. Although the enhancer transcripts themselves may be functional, it seems equally plausible that the act of transcription itself could provide an important function for maintaining enhancer activity. Alternatively, transcription at enhancers could result from nonspecific initiation of transcription in a region of chromatin that is both generally accessible and attracting a high localized concentration of polymerases.

### Promoter-Proximal Pol II Is Predominantly in an Elongationally Paused State

Previous ChIP assays have shown that Pol II accumulates at high concentrations on promoters of a large fraction of *Drosophila* genes in what is apparently a rate-limiting step in transcription (Muse et al., 2007; Zeitlinger et al., 2007). We show here by a quantitative comparison of Pol II in ChIP and GRO-seq assays that the majority of this promoter-associated Pol II seen across the genome is in a paused configuration and thus competent for transcription elongation. The properties of paused Pol II originally uncovered for *Drosophila Hsp70* and other heat shock genes: transcription of a short transcript (Rasmussen and Lis, 1993), its CTD phosphorylation state (Boehm et al., 2003; O'Brien et al., 1994), the association of pausing factors (Saunders et al., 2006), and the stimulation of their transcription in nuclear run-on assays by treatments that strip chromatin of repressive factors (Rougvie and Lis, 1988, 1990), are shared by a majority of *Drosophila* genes (Nechaev et al., 2010; and this work). Consistent with this last point and extrapolating from previous data (Gilchrist et al., 2010), we show that knockdown of a pausing factor reduces the occupancy of Pol II at promoters and that the overall effect on gene transcription after of disrupting pausing is dependent on whether the promoter itself allows for a competing nucleosome or perhaps another protein complex to occlude the initiation site in the absence of pausing.

### The Fate of Promoter Proximal Paused Pol II

Our quantitative analyses argue that the bulk of promoter-associated Pol II exists largely in a relatively stable paused configuration, and that this polymerase is a target of regulation. We expect that a paused polymerase turns over both by termination (Brannan et al., 2012), and by escape into productive elongation. The rates of either of these processes must be relatively slow to account for the high levels of accumulation of Pol II at pause sites 30–60 bases downstream of the TSS. Although our data do not definitively establish that the paused Pol II is the same Pol II that transcribes through the gene to produce a full mRNA transcript, evidence from our labs supports this view. First, the majority of polymerases are engaged and competent for transcription in a nuclear run-on assay; thus, the paused polymerase has the proper alignment to the 3′ end of the RNA and the Pol II active site to transcribe the gene following activation. Second, many genes are firing productive Pol II's into the body of the gene, some quite rapidly, e.g., the induced *Hsp70* fires every 4 s, yet most active genes still have a peak of promoter paused Pol II. Thus, Occam's razor directs us to propose that the Pol II molecules that undergo pausing subsequently elongate through the gene.

### Promoter-Proximal Pausing as a Step in Transcription Regulation

The biological significance of pausing has both experimental support and compelling speculation. First, some classes of activators directly stimulate pause escape rather than initiation and vice versa (Blau et al., 1996; Rahl et al., 2010; Yankulov et al., 1994), suggesting that different transcription factors could integrate different cellular signals to specify initiation and escape from pausing. Second, pausing of Pol II is accompanied by the capping of its associated short mRNA (Rasmussen and Lis, 1993) and by phosphorylation of the CTD of Pol II to a form that provides a scaffold for RNA processing factors that are coupled to transcription elongation (Phatnani and Greenleaf, 2006). This suggests that pausing may be a critical checkpoint in metazoans ensuring that RNA capping and the proper maturation of Pol II has an opportunity to occur for efficient transcription elongation and coupled splicing (Mandal et al., 2004, Rasmussen and Lis, 1993). Third, the residence time of a paused Pol II allows it to directly compete with nucleosomes for high affinity nucleosome positioning sequences at promoters, thus maintaining promoters in an active state (Gilchrist et al., 2008, 2010), and allowing for regulatory factor binding (Shopland et al., 1995). Fourth, maintenance of promoters in an open configuration provides a means for promoters to be primed for rapid,

synchronous regulation in response to a variety of signals (Adelman et al., 2009; Boettiger and Levine, 2009). Fifth, the knockdown of factors important for establishing pausing causes defects in both transcription activation and repression, which can be mediated through pausing mechanisms (Adelman et al., 2005; Aida et al., 2006; Missra and Gilmour, 2010). Finally, pause site escape is modulated by the recruitment of P-TEFb kinase (Peterlin and Price, 2006) that acts to phosphorylate and thereby inactivate pause stabilizing complexes, DSIF and NELF, and phosphorylate Pol II at Ser2 of its CTD to generate the elongationally modified form of Pol II. Evidence that this is a rate-limiting step is supported by the observation that the direct recruitment of P-TEFb to promoters is sufficient to produce high level of activation of *Drosophila Hsp70* (Lis et al., 2000) and other genes (Bieniasz et al., 1999; Majello et al., 1999). Together, these observations suggest that pausing serves to potentiate transcription, and at the same time allow a repertoire of transcription factors to fine tune transcript levels both up and down by changing the rate of escape of Pol II from pausing.

## EXPERIMENTAL PROCEDURES

### RNAi Treatment and Generation of ChIP-Seq Data

RNAi in *Drosophila* S2 cells were performed as described (Gilchrist et al., 2010). Further details regarding the published ChIP-seq data can be found in the Extended Experimental Procedures.

### Isolation of Nuclei for GRO-Seq

Nuclei were isolated as described previously (Core et al., 2008), with several modifications. Details regarding the specific protocols used for isolating nuclei from RNAi-treated cells and nuclei for the plus—and minus—sarkosyl data sets can be found in the Extended Experimental Procedures.

### Preparation of GRO-Seq Libraries

Untreated, mock and NELF-depleted GRO-seq libraries were prepared as in Core et al. (2008), with the following modifications. Trizol (Invitrogen) was used to stop the reaction instead of DNase I and proteinase K treatment. The RNA was further extracted once with acid phenol:chloroform, and once with chloroform before precipitating with 2.5 volumes of −20°C ethanol. Bead binding buffers all contained 4 units/ml of SUPERaseIN (Ambion) and the following buffers were slightly modified. Bead blocking buffer: 0.25 × SSPE, 1 mM EDTA, 0.05% Tween 20, 0.1% PVP, and 1 mg/ml ultrapure BSA (Ambion); Binding buffer: 0.25 × SSPE, 37.5 mM NaCl, 1 mM EDTA, 0.05% Tween 20; low-salt wash buffer: 0.2 × SSPE, 1 mM EDTA, 0.05% Tween 20. High-salt wash buffer: 0.25% SSPE, 137.5 mM NaCl, 1 mM EDTA, 0.05% Tween 20. The end repair steps were modified as follows. Pelleted RNA from the first bead binding was resuspended in 20 μl, and heated to 70°C for 5 min, followed by incubation on ice for 2 min. 1.5 μl tobacco acid pyrophosphatase (TAP) buffer, 4.5 μl water, 1 μl SUPERaseIn, and 1.5 μl TAP (Epicenter) were then added and the reaction incubated at 37°C for 1.5 hr. One microliter of 300 mM $MgCl_2$ and 1 μl T4 polynucleotide kinase (PNK) were added to the reaction for an additional 30 min. for phosphorylating the 5′ ends, 20 μl T4 PNK buffer, 2 μl 100 mM ATP, 145 μl water, 1 μl SUPERaseIn, and an additional 2 μl of PNK were added for 30 min at 37°C. The reaction was then stopped by addition of 20 mM EDTA followed by acid phenol extraction and precipitation.

Plus- and minus-Sarkosyl matched GRO-seq libraries (cells grown in Lis lab) and the Circ-Ligase libraries (grown in Adelman lab for ECF analysis) were made with three sequential bead enrichment steps as above, but a RNA cloning strategy developed by Ingolia (2010), was used to prepare the samples for sequencing with the following modifications. PNK treatment to remove 3′ phosphates was performed after the first bead enrichment. NRO-RNA (24.5 μl) was mixed with 3 μl 10X PNK buffer (NEB), 1.5 μl T4-PNK, and 1 μl SUPERase Inhibitor (Ambion) for 30 min at 37°C. Poly-A tailing of RNAs was performed prior to the third bead enrichment, and performed as described in Ingolia (2010). Triple-enriched and poly-A tailed nascent RNAs were then reverse transcribed and circularized as in Ingolia (2010). cDNAs were not linearized or PAGE purified after circularization because the range of sizes (~150–350 bp) of the cDNA prevented efficient separation of the circularized and linearized cDNAs. Samples were amplified and PAGE purified as described (Core et al., 2008) and quantified before submission for sequencing.

### Data Acquisition and Analysis

GRO-seq libraries were sequenced on the Illumina Genome Analyzer II, using standard protocol at the Cornell bioresources center (http://www.BRC.cornell.edu). Bowtie (Langmead et al., 2009) was used to map 26-mer, with up to two mismatches to the DM3 version on the *Drosophila* genome. Reads were also mapped to a representative of repetitive genes transcribed specifically by Pol I (rRNA gene; GenBank accession number M21017.1), and Pol III (transfer RNAs [tRNAs]; parsed from flybase gene set described below). The rRNA included the extragenic spacers, and tRNAs, were extended ±100 bases to account for nascent transcripts that are processed and not part of the annotated tRNA. A summary of sequencing yields and the number of reads mapping uniquely to the genome or other annotations is contained in Table S1.

Details on gene and enhancer lists, and the analyses contained throughout the manuscript can be found in the Supplemental information.

## ACCESSION NUMBERS

The Gene Expression Ominbus accession number for the Pol II ChIP-seq and all GRO-seq data sets is GSE23544.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2012.08.034.

## LICENSING INFORMATION



## ACKNOWLEDGMENTS

We would like to thank Charles Danko and Andre Martins for help with R-programming, Peter Kharchenko for providing a *Drosophila* enhancer list, and Bing Ren for providing a human enhancer list from IMR90 cells. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101987) to K.A. and by NIH grants GM25232 and HG004845 to J.T.L. L.J.C., K.A., and J.T.L. conceived the study and designed the experiments. L.J.C. produced the GRO-seq datasets. K.A. and D.A.G. performed RNAi treatments and produced the ChIP data sets. L.J.C., D.A.G., J.J.W., K.A., D.C.F., and H.K. performed the data analysis. L.J.C., J.T.L., and K.A. wrote the paper. The authors declare that they filed a provisional patent (US 2010/0062946 A1) that describes the GRO-seq technology.

Received: June 21, 2012
Revised: August 24, 2012
Accepted: August 30, 2012
Published online: October 11, 2012

## REFERENCES

Adelman, K., Marr, M.T., Werner, J., Saunders, A., Ni, Z., Andrulis, E.D., and Lis, J.T. (2005). Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. Mol. Cell *17*, 103–112.

Adelman, K., Kennedy, M.A., Nechaev, S., Gilchrist, D.A., Muse, G.W., Chinenov, Y., and Rogatsky, I. (2009). Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. Proc. Natl. Acad. Sci. USA *106*, 18207–18212.

Aida, M., Chen, Y., Nakajima, K., Yamaguchi, Y., Wada, T., and Handa, H. (2006). Transcriptional pausing caused by NELF plays a dual role in regulating immediate-early expression of the junB gene. Mol. Cell. Biol. *26*, 6094–6104.

Baugh, L.R., Demodena, J., and Sternberg, P.W. (2009). RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science *324*, 92–94.

Bieniasz, P.D., Grdina, T.A., Bogerd, H.P., and Cullen, B.R. (1999). Recruitment of cyclin T1/P-TEFb to an HIV type 1 long terminal repeat promoter proximal RNA target is both necessary and sufficient for full activation of transcription. Proc. Natl. Acad. Sci. USA *96*, 7791–7796.

Blau, J., Xiao, H., McCracken, S., O'Hare, P., Greenblatt, J., and Bentley, D. (1996). Three functional classes of transcriptional activation domain. Mol. Cell. Biol. *16*, 2044–2055.

Boehm, A.K., Saunders, A., Werner, J., and Lis, J.T. (2003). Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock. Mol. Cell. Biol. *23*, 7628–7637.

Boettiger, A.N., and Levine, M. (2009). Synchronous and stochastic patterns of gene activation in the Drosophila embryo. Science *325*, 471–473.

Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., Kiemele, L., Hansen, K., Davis, R., Lykke-Andersen, J., and Bentley, D.L. (2012). mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. Mol. Cell. *46*, 311–324.

Buratowski, S. (2008). Transcription. Gene expression—where to start? Science *322*, 1804–1805.

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.

FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B., and Vinson, C. (2006). Comparative genomics of Drosophila and human core promoters. Genome Biol. *7*, R53.

Gilchrist, D.A., Nechaev, S., Lee, C., Ghosh, S.K., Collins, J.B., Li, L., Gilmour, D.S., and Adelman, K. (2008). NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. Genes Dev. *22*, 1921–1933.

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

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.

Hawley, D.K., and Roeder, R.G. (1985). Separation and partial characterization of three functional steps in transcription initiation by human RNA polymerase II. J. Biol. Chem. *260*, 8163–8172.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Ingolia, N.T. (2010). Genome-wide translational profiling by ribosome footprinting. Methods Enzymol. *470*, 119–142.

Juven-Gershon, T., Hsu, J.Y., Theisen, J.W., and Kadonaga, J.T. (2008). The RNA polymerase II core promoter - the gateway to transcription. Curr. Opin. Cell Biol. *20*, 253–259.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488.

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

Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., et al. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature *471*, 480–485.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876–880.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Krumm, A., Hickey, L.B., and Groudine, M. (1995). Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. Genes Dev. *9*, 559–572.

Kutach, A.K., and Kadonaga, J.T. (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol. Cell. Biol. *20*, 4754–4764.

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

Lis, J. (1998). Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. Cold Spring Harb. Symp. Quant. Biol. *63*, 347–356.

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.

Majello, B., Napolitano, G., Giordano, A., and Lania, L. (1999). Transcriptional regulation by targeted recruitment of cyclin-dependent CDK9 kinase in vivo. Oncogene *18*, 4598–4605.

Mandal, S.S., Chu, C., Wada, T., Handa, H., Shatkin, A.J., and Reinberg, D. (2004). Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. Proc. Natl. Acad. Sci. USA *101*, 7572–7577.

Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev. *25*, 742–754.

Missra, A., and Gilmour, D.S. (2010). Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. Proc. Natl. Acad. Sci. USA *107*, 11301–11306.

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.

Nechaev, S., and Adelman, K. (2011). Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. Biochim. Biophys. Acta *1809*, 34–45.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335–338.

O'Brien, T., Hardin, S., Greenleaf, A., and Lis, J.T. (1994). Phosphorylation of RNA polymerase II C-terminal domain and transcriptional elongation. Nature *370*, 75–77.

Pal, M., McKean, D., and Luse, D.S. (2001). Promoter clearance by RNA polymerase II is an extended, multistep process strongly affected by sequence. Mol. Cell. Biol. *21*, 5815–5825.

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

Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev. *20*, 2922–2936.

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.

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

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

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.

Rougvie, A.E., and Lis, J.T. (1990). Postinitiation transcriptional control in Drosophila melanogaster. Mol. Cell. Biol. *10*, 6041–6045.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. *8*, 424–436.

Saunders, A., Core, L.J., and Lis, J.T. (2006). Breaking barriers to transcription elongation. Nat. Rev. Mol. Cell Biol. *7*, 557–567.

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.

Shopland, L.S., Hirayoshi, K., Fernandes, M., and Lis, J.T. (1995). HSF access to heat shock elements in vivo depends critically on promoter architecture defined by GAGA factor, TFIID, and RNA polymerase II binding sites. Genes Dev. *9*, 2756–2769.

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., Grimmond, S.M., Schroder, K., et al. (2009). Tiny RNAs associated with transcription start sites in animals. Nat. Genet. *41*, 572–578.

van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol. *8*, e1000371.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A., et al. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature *474*, 390–394.

Wu, C.H., Yamaguchi, Y., Benjamin, L.R., Horvat-Gordon, M., Washinsky, J., Enerly, E., Larsson, J., Lambertsson, A., Handa, H., and Gilmour, D. (2003). NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. Genes Dev. *17*, 1402–1414.

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

Yankulov, K., Blau, J., Purton, T., Roberts, S., and Bentley, D.L. (1994). Transcriptional elongation by RNA polymerase II is stimulated by transactivators. Cell *77*, 749–759.

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.

Cell PRESS
Open ACCESS

## EXTENDED EXPERIMENTAL PROCEDURES

### RNAi Treatment and Generation of ChIP-Seq Data

Cells received RNAi targeting the NELF-B and NELF-E subunits or "mock" dsRNA against β-galactosidase for 96 hr prior to harvesting cells. Pol II ChIP-seq (Rpb3 antibody) data from Nechaev et al. (Nechaev et al., 2010) was combined with an additional biological replicate from (Gilchrist et al., 2010) for a total of ~13 million mappable reads. Data analysis was performed as described previously (Nechaev et al., 2010). MNase-seq data was from (Gilchrist et al., 2010). Ser2-P ChIP data was generated by muse et al. (Muse et al., 2007). H3K4me data was generated by (Gan et al., 2010), and downloaded from Gene Expression Omnibus (GEO accession no. GSM480156).

### Isolation of Nuclei from RNAi-Treated Cells

0.5-1x$10^8$ S2 cells grown in M3 + BPYE+10%FBS, were pelleted at 1000Xg for 4min in a swinging bucket centrifuge at 4°C. Cells were washed once in 10 ml ice cold PBS, and then pelleted again at 1000 Xg for 4min. Cells were swollen by resuspension in 1ml swelling buffer (10mM Tris-Cl pH 7.5, 10% glycerol, 3mM $CaCl_2$, 3mM MgCl2, 1X protease inhibitior cocktail (roche tablets, EDTA free), and 4units/ml of SUPERaseIn (ambion)), followed by incubation at 4°C for 5min. 4ml of ice cold lysis buffer (10mM Tris-Cl pH7.5, 300mM sucrose, 10mM NaCl, 3mM CaCl2, 2mM MgCl2, 0.5% igepal, 0.5mM DTT, protease inhibitors and SUPERaseIn) was then added directly to the cells, and the cells were transferred to a 5ml capacity dounce homogenizer and lysed by 5 strokes with a Teflon dounce. Nuclei were pelleted at 1000Xg for 4min at 4°C. Nuclei were washed once with 10ml of lysis buffer, pelleted, and resuspended in 1ml storage buffer (50mM Tris-Cl pH 8.0, 25% glycerol, 5mM MgCl2, 0.1mM EDTA, 5mM DTT), and transferred Nuclei were pelleted by centrifugation in a fixed angle rotor at 1000Xg for 5min at 4°C. Nuclei were resuspended at 2x$10^7$ nuclei/100ul, snap frozen in liquid nitrogen and stored a −80°C until use.

### Isolation of Sarkosyl-Dependent Nuclei

Isolation of sarkosyl dependent nuclei was performed as in (Rougvie and Lis, 1988). Nuclei were tested for sarkosyl dependence prior to creating GRO-seq libraries by conventional run-on hybridization as in (Rougvie and Lis, 1988). Briefly 2 x$10^7$ nuclei were incubated with 2x Run-on reaction buffer (Core et al., 2008; Rougvie and Lis, 1988). RNA was isolated and hybridized under standard conditions to positively charged nylon membranes that were spotted with single stranded M13 phagemid probes containing various regions of the Hsp70 gene as well as several internal controls. The nuclei used for creating GRO-seq libraries showed a 4.5 fold decrease in promoter signal at the *Hsp70* gene in the absence of sarkosyl (not shown).

### Reannotation of TSSs

Gene annotations were processed to customize them for certain analyses. Starting with *D. melanogaster* build 5.23 .gff file from flybase (//ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.23_FB2009_10/gff/), which contained 22,213 elements, we first reannotated each gene transcription start site (TSS), by searching for evidence of a TSS within ±150 bp from the annotated TSS using the 5′-capped short RNA data generated by Nechaev et al. The base with the highest number of reads within this window was designated the observed TSS. If two bases within this window both displayed the same maximum number of reads, the most 5′-base was used as the TSS. Finally, if there was no evidence of a TSS-RNA for this cell type, the annotated TSS was used. We then generated a list of genes with unique transcription start sites, keeping one representative gene name at random for genes with identical TSSs. This list of 16,746 unique TSSs was then used as a starting point for creation of other gene lists used in analyses described below.

### Defining Active Genes and Promoters

To identify genes as actively transcribed in the GRO-seq data set we first chose a region corresponding to the gene body of each gene (+300 to gene end), in order to avoid interference from the signal at promoters. Genes were considered active if there was significant (pval < 0.01) enrichment of GRO-seq density (reads/mappable bases) relative to a background distribution of 1% of the total GRO-seq reads in a poisson distribution in the mappable genome. Enrichment p-values were calculated using Fisher's exact test. We chose a 1% as the background as this is thought to be a conservative estimate of background (Core et al., 2008; Min et al., 2011).

We also used small-capped RNA and Pol II ChIP-seq data sets to define active genes in *Drosophila* as in Nechaev et al., 2010.

IMR90 active promoters were defined as promoters with a significant enrichment of GRO-seq density near the TSS (highest 250 bp window ±500 bp) from the annotated TSS. Significantly enriched windows were identified using Fisher's exact test as above. By this criteria, 9005 promoters were considered active.

### Classification of Paused Genes with GRO-seq

Genes were identified as having significant enrichment of GRO-seq density in the promoter regions as previously (Core et al., 2008), with the following changes due to the differences in the structure of the *Drosophila* and human genomes and availability of accompanying data. Promoter regions were defined as ±150 bp from the TSS, and the promoter density was designated as the 100bp window within this promoter region that had the highest density. The gene body density was defined by the density of GRO-seq reads from +300 to the annotated end of the gene. Only genes greater than 800bp were used because short gene regions have low statistical power when calling pausing (Core et al., 2008), bringing our unique gene list to 14,427 genes. Genes were considered active by

GRO-seq if the enrichment of GRO-seq density of the gene was significant (pval < 0.01, Fisher's exact test) relative to our assumption of background (1% of reads spread in a poisson distribution over the nonrepetitive portion of the genome). The assumption of background comes from our estimates of the NRO libraries being > 99% pure. This resulted in 9,544 active genes. To call pausing at active promoters we used the a list as defined by Nechaev et al. (Nechaev et al., 2010). This list of 7,336 active promoters (min length 800) was considered in order avoid counting signal at promoters that is due to transcription read-through from other promoters. Bound promoters were defined as previously (Nechaev et al., 2010), which resulted in 3,138 genes. Promoters from these three lists were then classified as paused as done previously (Core et al., 2008).

### Designation of Gene Ends for Analysis of Transcription Changes within Genes for the NELF RNAi Data Sets

In order to avoid unannotated, promoters that are nested within genes from influencing the apparent gene body signal under various conditions, the annotations in the unique TSS gene list were truncated as follows. Genes were scanned for TSS-RNAs (Nechaev et al., 2010) from +300 to the end in 100bp windows at 25bp steps. Windows with greater than 25 reads were designated as an annotated TSS, and the end of the gene was then designated as 100 bases upstream from this window. If multiple windows were found, the most 5′-window was used as the anchor point for designating the new gene end. Newly annotated genes less than 800bp were then removed, yielding 11,719 'truncated' genes that were used for quantifying the level of transcription at promoters and in genes for the RNAi treated and control data sets.

### Construction of a Single Gene List Used for Analyzing the Signal at Promoters, Gene Bodies and Gene Ends

In order to clearly assess the effect of Sarkosyl treatment and NELF RNAi on different gene regions (promoters, gene bodies, and gene ends), it was necessary to remove genes where these gene regions overlap. For simplicity in the analysis, a single gene list was constructed with which to analyze all three gene regions. Starting with 16,746 genes for which we reannotated the TSS as described above, we then removed genes with an overlapping promoter from −500bp to +1000bp relative to the gene end. Duplicate TSSs were then removed, and a representative isoform was selected randomly. To clean up the overlaps of promoters in genes, this list was then merged with the 'truncated' genes described above to yield 8,266. Finally, to avoid quantifying low levels of signal coming from read-through transcription from nearby genes, we selected the active promoters (described above) from this list to obtain 4,652 genes with active promoters but no overlapping promoters in the 3′-end or gene body.

### Construction of *Drosophila* Putative Enhancers List

A list of 6,553 putative enhancers from *Drosophila* S2 cells was obtained from the Peter Kharchenko and Peter Park (Kharchenko et al., 2011). These locations were filtered against our reannotated genes as well as refseq annotations for being 2kb from a TSS and 500 bp from a gene end, yielding 533 putative distal enhancers that were used in calculations of orientation indexes and generation of composite profiles.

### Construction of Human Putative Enhancer List

A list of 54,251 putative human enhancer locations from IMR90 cells was obtained from Bing Ren, based on published work (Heintzman et al., 2007). These locations were then overlayed with DNase hypersensitivity (DHS) mapping data from the same cell line (GEO accession: GSM530665). The center of the DHS peak was considered the location of the enhancer. The putative enhancers were then filtered for those that were distal to genes by removing locations that were within refseq genes, within 2kb from an annotated TSS, or 5kb from an annotated gene end, yielding 18,835 gene distal locations with a DHS peak that were used for generating the composite profile in Figure S5. To avoid problems associated with regions of low or zero signal, these putative enhancers were filtered further by requiring 25 GRO-seq reads from ±1 kb from the DHS center on either strand (p-val = < 0.01, fisher's exact test). This yielded 952 'active' enhancers that were used for calculations of orientation index.

### Generation of Composite Profiles and Orientation Indices around Enhancers

When generating composite profiles of genomic data around regions with low signal, a small number of regions with intense signal can disproportionately influence the profile. To avoid this problem, we repeated the generation of the average composite profile 1,000 times on 10% of the locations. The locations were chosen at random with replacement. We then plot the median signal for these averages as the composite signal for the list of enhancers.

Since enhancers have no known directionality, we chose the strand with the highest signal within 1kb downstream (plus strand) or 1 kb upstream (minus strand) and the 'direction' for the enhancer. For ease of comparison with promoters in Figure 4D, the promoter 'direction' was chosen in the same way, but after tabulating the signal from ±500 bp from the promoter.

### Comparison of GRO-seq, ChIP-seq, and TSS-RNA at Promoters

Comparisons between GRO-seq and ChIP-seq were performed with the untreated GRO-seq data set, since these nuclei were made from the same cultures of S2 cells. Promoters with any overlapping annotations were cleared from the gene lists, and then the genes were truncated, if there was a downstream promoter on either strand relative to each gene. Gene less than 800bp were then removed from the truncated list, resulting in a final list 12,541 genes that were compared by their densities at promoter regions

(Figures S5A–S5C, black). This list was merged with significantly bound genes (FDR = 0.01) from the Pol II ChIP-seq data set to yield 3168 genes (Figures S5A–S5C, red).

### Comparison of GRO-seq and ChIP in Gene Bodies for Normalization of ChIP and GRO

Because of the high and variable background associated with ChIP-seq, we sought a list of genes with gene where the gene body signal could be easily separated from background with which we could use to quantitatively compare ChIP-seq and GRO-seq. We reasoned that genes with the highest Ser2-P signal (mark of active transcription elongation) would represent the genes with the highest density of active polymerases and thus, control for background. Thus, we further trimmed the 12,541 genes by selecting those genes from a Ser2-P ChIP data set with Z score > 3. This resulted in 1874 genes for which to compare ChIP-seq and GRO-seq (Figure S5D, red).

### Calculation of ECF

Promoters that are significantly bound by Pol II (n = 3168) transformed by the equation resulting from GRO-ChIP comparison in genes ($GRO_{exp}$ = 1.2(ChIP)-.25) to obtain an expected amount of GRO-seq signal at promoters. The actual GRO-seq ($GRO_{act}$) signal at promoters is then divided by the expected GRO-seq to obtain the Engaged, Competent Fraction (ECF) of polymerase at promoters.

### Motif Search within Human Promoters and Calculation of Orientation Indexes

Consensus sequences for core promoter motifs were obtained from the literature (FitzGerald et al., 2006; Juven-Gershon et al., 2008; Ohler et al., 2002) (Table S2), and these sequences were used to search near human promoters (from −500 to +500 relative to the annotated RefSeq TSSs), using custom scripts. Genes with each core promoter motif were thus identified and used to determine the effect of each motif on transcription directionality.

Total promoter reads were determined by adding the highest density 100bp window from ±500bp relative to the TSS from the sense and antisense strands. An orientation index was then expressed as the fraction of total promoter reads found in the sense orientation for Table S2 and as a ratio of sense to antisense for Figure 1C. For each motif in Table S2, the orientation index is the average of all genes found with that motif.

## SUPPLEMENTAL REFERENCES

Gan, Q., Schones, D.E., Ho Eun, S., Wei, G., Cui, K., Zhao, K., and Chen, X. (2010). Monovalent and unpoised status of most genes in undifferentiated cell-enriched Drosophila testis. Genome Biol. *11*, R42.

Ingolia, N.T. (2010). Genome-wide translational profiling by ribosome footprinting. Methods Enzymol. *470*, 119–142.

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

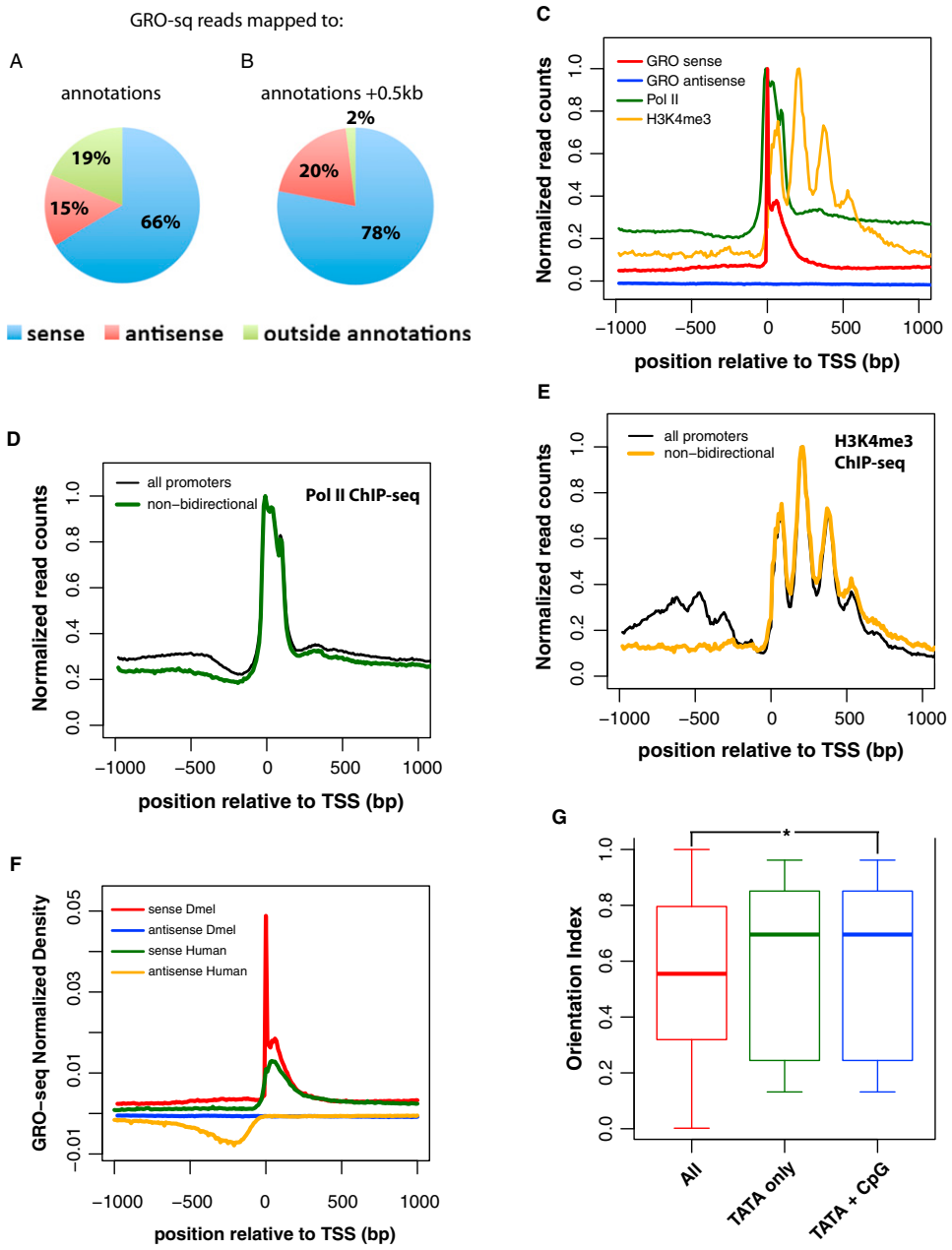


**Figure S1. Genomic Distribution of GRO-Seq Data and Alignments to Gene Features, Related to Figure 1**

(A and B) The majority of transcription occurs within or very near gene annotations. GRO-seq reads were counted whether they were coding/sense strand (blue) or noncoding/antisense strand (red), or outside (green) of flybase annotations (A) or flybase annotations extended by 500 bp on each end (B).

(C–E) *Drosophila* promoters are predominantly unidirectional. (C) Composite profiles for GRO-seq coding strand (sense, red), noncoding (antisense, blue), Pol II ChIP-seq (green), H3K4me3 ChIP-seq (orange), showing a strong correlation between the level and direction of transcription and H3K4me3. Bidirectional promoters within 1kb were removed for this analysis. Each window for each profile were are expressed as a fraction of the peak signal for that profile, which was set to 1. (D and E) ChIP-seq data for Pol II and H3K4me3, respectively, for all promoters (n = 14,454) versus nonbidirectional promoters (n = 10,031), showing that signal in both data sets is also highly associated with gene direction.

(F) Direct comparison between GRO-seq profiles at *Drosophila* and human promoters. GRO-seq profiles from −/+ 1.0 kb relative to TSS are shown for all unique flybase promoters (n = 14,454) (red, sense; blue, antisense) or all unique human promoters (n = 18,031)(green, sense; orange, antisense). When compared to Figure 1E, it is readily apparent that the directional TATA element human promoters resemble directional *Drosophila* promoters.

(G) TATA-containing promoters are directional when within CpG islands. Boxplots showing the distribution the orientation indexes (OI; sense versus antisense density) at promoters for all active human promoters (red; n = 6,017), active promoters with a TATA box (green; n = 8), and active promoters with a TATA box within a CpG island (blue; n = 63). Active promoters were designated as those with > 25 reads ±500 bases from the TSS on wither strand. The distributions of the OI for all genes and those with both a TATA box and CpG island are significantly different (p-val = 1.1 $e^{-04}$, two sample t test).

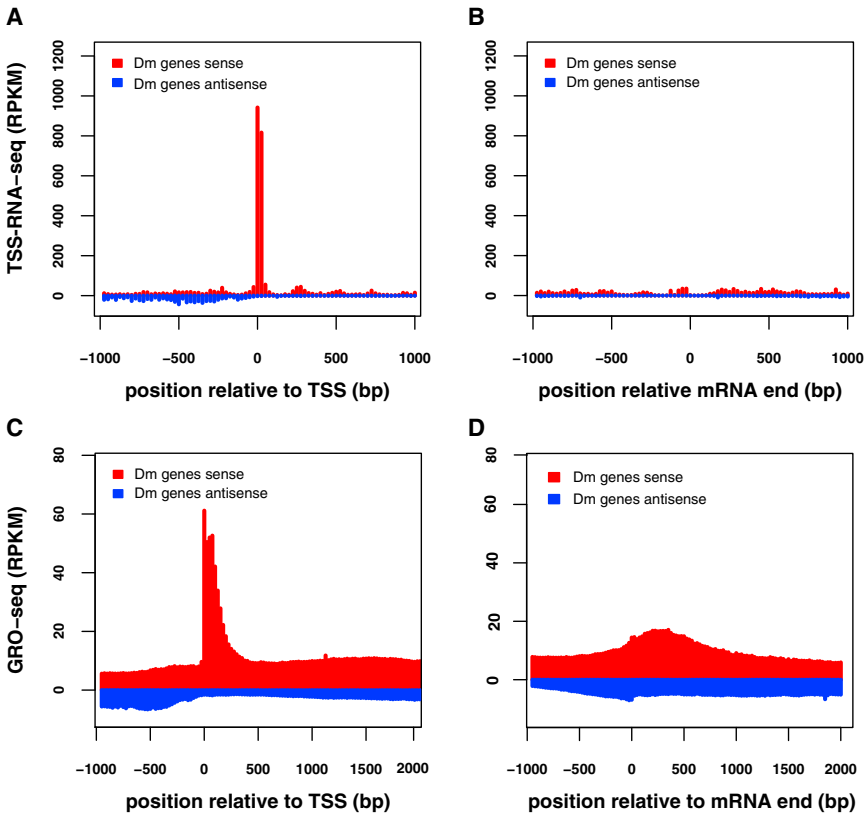


**Figure S2. Distribution of Small-Capped RNAs and GRO-Seq at Promoters and Gene Ends, Related to Figure 1**

(A–D) GRO-seq profile at gene ends indicates Pol II is slowing down prior to termination. TSS-RNA and GRO-seq profiles at promoters (A and C, respectively) have a similar structure; whereas profiles at the 3′-end (B and D respectively) do not, indicating the GRO-seq peak at 3′-ends is not due to initiation. Note that y axis scales are the same for A and B, and C and D.

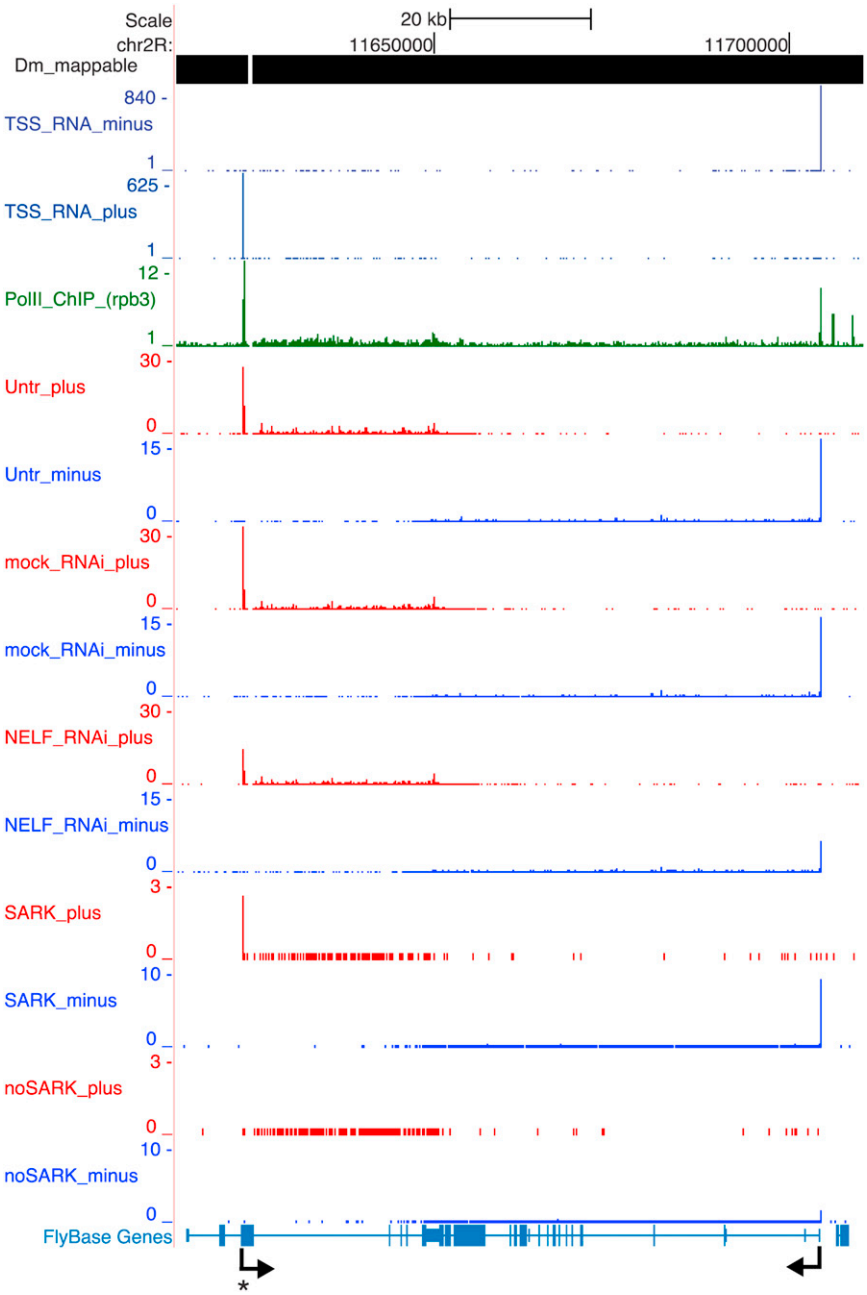


**Figure S3. UCSC Browser Screenshot of Sarkosyl Dependent and NELF RNAi GRO-Seq Experiments, Related to Figure 2**

TSS-RNA reads (Nechaev et al., 2010) marking TSSs are in dark blue for the plus and minus strand (read/base/10ˆ6 reads). GRO-seq reads (reads/base/10ˆ6 reads) aligning to the plus strand are shown in red; minus strand in blue. ChIP-seq for total Pol II (α-Rpb3) is shown in green (reads/25bp bin), and gene annotations are shown at the bottom in blue. The arrowheads depict TSSs and the * denotes a TSS that is reannoated with our data sets.

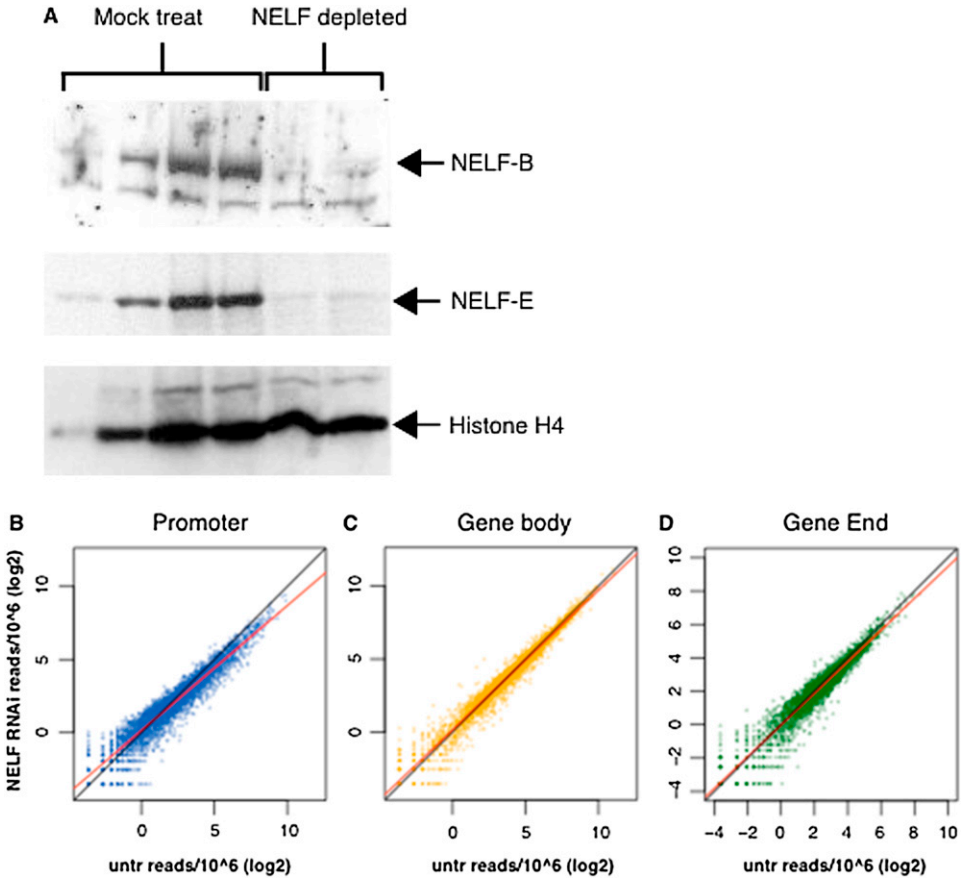


**Figure S4. NELF RNAi Effect on Gene Regions, Related to Figure 3**

(A) Western blot showing depletion of NELF-B and NELF-E after RNAi treatment. Serial dilutions NELF-B and NELF-E from mock-treated cells, compared with the RNAi treated cells show that both NELF subunits are depleted by > 90%. Histone H4 is used as a loading control.

(B–D) Scatterplots showing a comparison of promoter (B), gene body (C), and gene ends (D) signals between the untreated samples, and NELF-RNAi treated samples. The black line represents a 1:1 fit, and the red line represent the fit of the data.

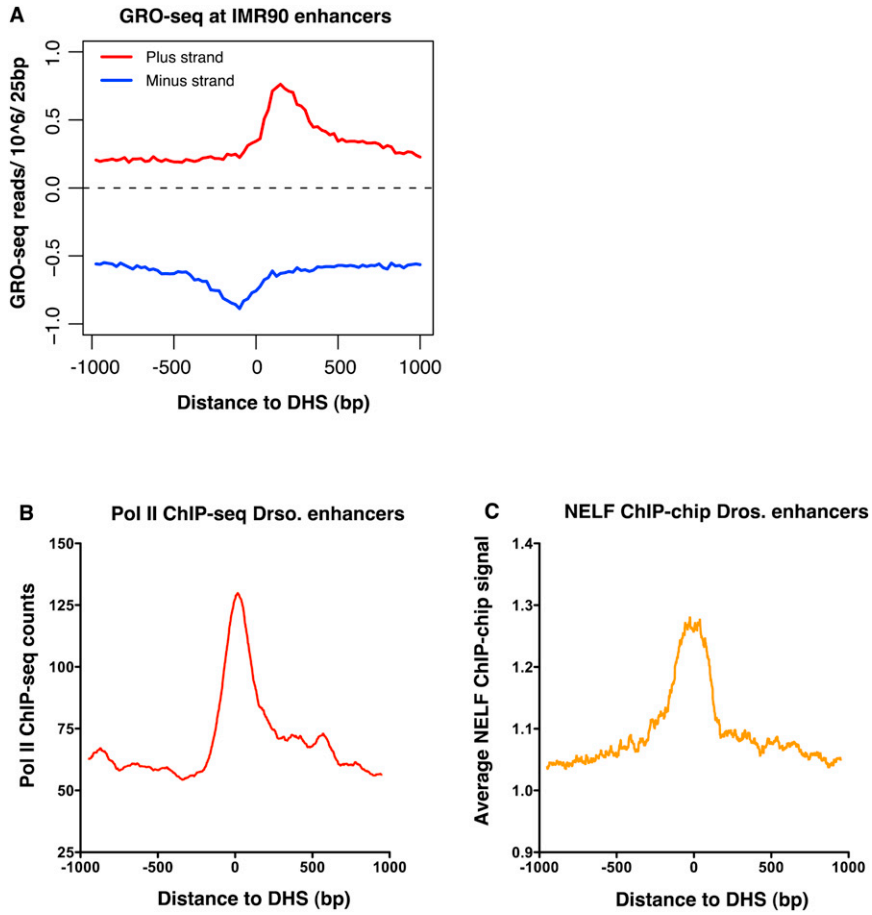


**Figure S5. Supporting Genomic Data at Enhancers in This Study, Related to Figure 4**

(A) GRO-seq data at putative human enhancers (n = 34,915). GRO-seq data is from IMR90 cells (Core et al., 2008). Data was compiled relative to the center of DHS sites.

(B and C) Pol II ChIP-seq and NELF ChIP-chip data (Gilchrist et al., 2010), respectively, around *Drosophila* putative enhancers.

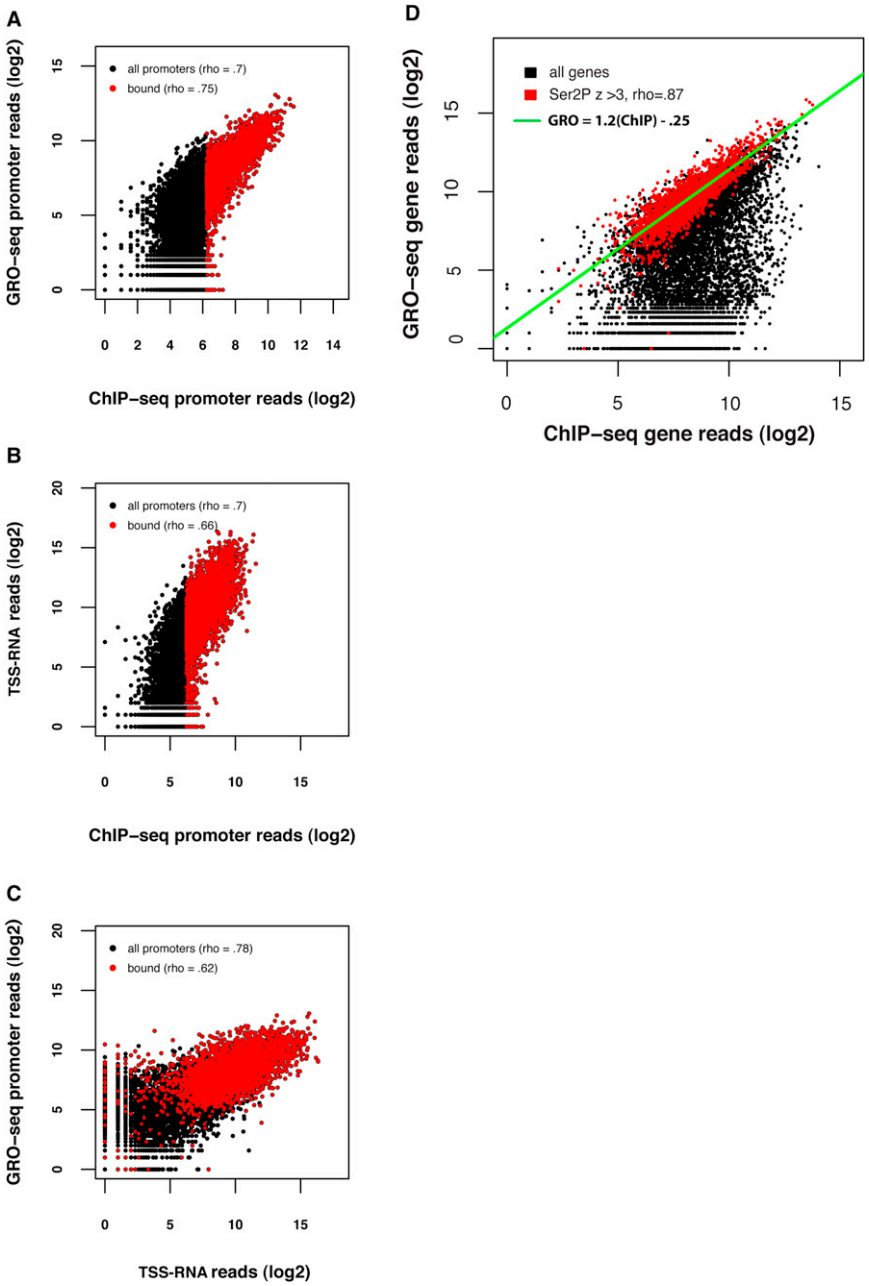


**Figure S6. Comparison between Assays that Detect Polymerase at Promoters and in Genes, Related to Figure 5**

(A–C) Shown are scatter-plots comparing amount of sequencing reads between (A) GRO-seq and Pol II ChIP-seq, (B) small-RNA-seq to ChIP-seq, and (C) GRO-seq to small-RNA-seq at promoters. All unique promoters are shown in black (n = 12,541); promoters called Pol II-bound by ChIP-seq (n = 3,168) in red. rho is Spearman's correlation coefficient between the two data sets.

(D) Normalization between ChIP-seq and GRO-seq data sets through fitting of signal within gene bodies. Plotting of the signal for each assay within all genes (black) shows a poor correlation (rho = 0.54), whereas plotting the signal for each after selecting genes that are highly active gives (red) an excellent correlation between data sets (rho = 0.87). Highly active genes are classified as those with the top 10% of Ser2P ChIP (n = 1874) signal within the gene (mark of active polymerases). The fit of the gene data for highly active genes is shown in green. This equation is used for calculating the engaged, competent fraction of polymerase at promoters.

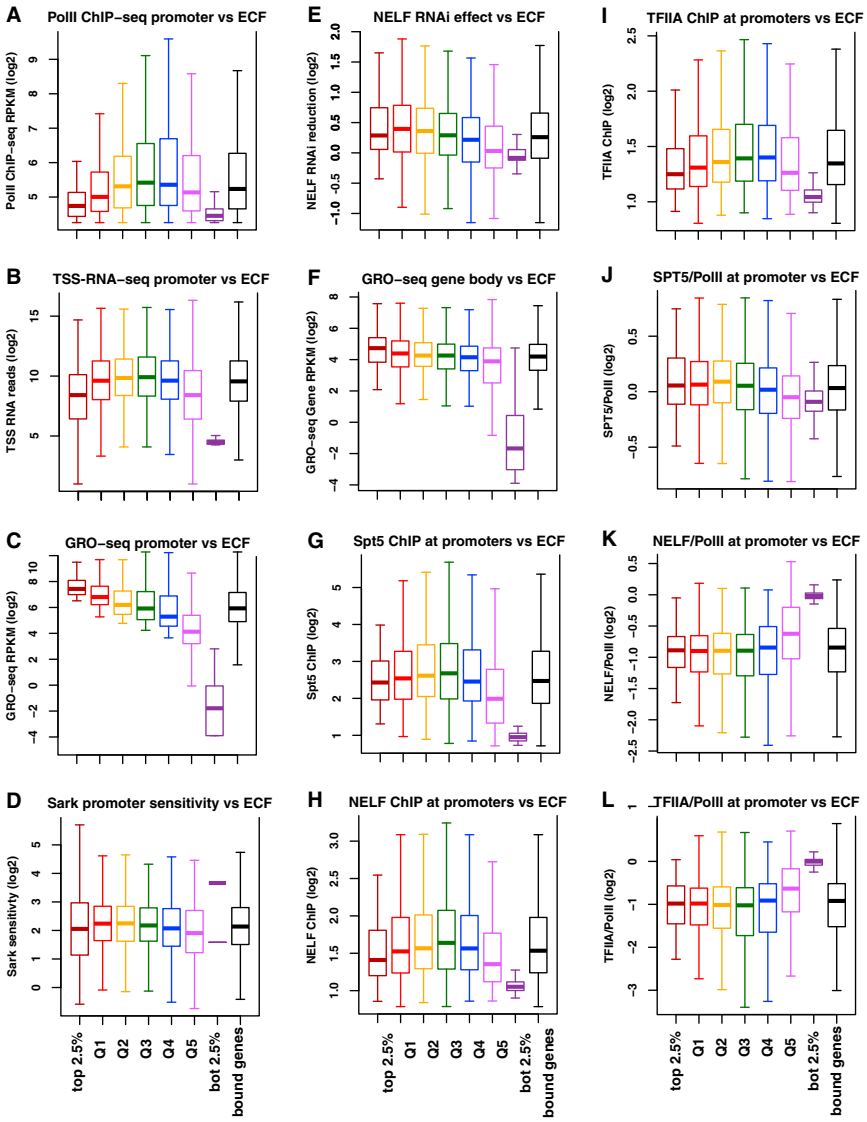


**Figure S7. Association of Various Genomic Data Sets with Engaged, Competent Fraction of Pol II, Related to Figure 5**

(A–L) Boxplots showing how various data sets or properties at promoters relate to Engaged, Competent Fraction at promoters. The list of bound promoters for which an ECF was generated (n = 3168), was broken into quintiles based on ECF (Q1-Q5: High ECF- low ECF) and the distribution of other data sets was plotted in boxplot format. The distribution for the top and bottom 2.5%, and all genes is also shown. Solid line = median, boxes encompass middle quartiles, and the whiskers encompass 95% of the data. Colors are consistent with Figure 5D.