www.sciencemag.org/cgi/content/full/1162228/DC1

Supporting Online Material for

Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters

Leighton J. Core, Joshua J. Waterfall, John T. Lis*

*To whom correspondence should be addressed. E-mail: jtl10@cornell.edu

Published 4 December 2008 on *Science* Express
DOI: 10.1126/science.1162228

This PDF file includes:

Supplementary online material for:

Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters

AUTHORS/AFFILIATIONS

Leighton J. Core*, Joshua J. Waterfall* and John T. Lis‡

Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853 USA

‡Corresponding author: Prof. John T. Lis, Department of Molecular Biology and Genetics, 416 Biotechnology Building, Cornell University, Ithaca, NY 14853, USA. Tel (607) 255-2442, Fax (607) 255-2428, Email: jtl10@cornell.edu

* These authors contributed equally to this work

Table of contents: **Page**

The nuclear run-on assay

Background:

Nuclear Run-On (NRO) assays have been used to measure the density of transcribing polymerases over specific targeted regions of the genome, and variations of the assay are capable of mapping the position of polymerases with high precision*(1-4)*. Traditionally, nuclei are isolated, endogenous nucleotides are removed by washing, and radionucleotides are added back allowing transcriptionally engaged polymerases to resume elongation. The incorporated radiolabel is restricted to sequences immediately downstream of the original position of the transcriptionally-engaged polymerase by keeping run-on reaction times short. The anionic detergent sarkosyl, which does not interfere with elongating polymerases, is often added to the nuclear run-on reaction to ensure that new transcription initiation events do not occur, and to remove physical impediments that can block elongation*(3, 5)*. Thus all new transcription is produced by polymerases that are engaged at the time of nuclear isolation. The RNA is then isolated and hybridized to filters containing genes or gene regions of interest. These measurements have been shown to represent the level of transcriptionally-engaged polymerase on genes at the time of nuclei isolation, and have also been used to identify Pol II that is paused at the 5' ends of genes as well as the distance Pol II travels beyond the 3'-ends of genes prior to termination*(6-9)*.

Previous attempts at scale-up have hybridized radiolabeled NRO-RNAs to cDNA probes spotted on macroarrays to analyze how steady state transcription of genes relates to mRNA accumulation*(10, 11)*. These methods can give reasonable approximations for steady state transcription levels for some genes, however, they suffer from low sensitivity, lack of whole genome coverage, and no resolution within gene regions. Whole genome coverage is important for detection of novel transcription units as well as transcripts that are not present in cDNA libraries. The lack of resolution of cDNA arrays is of concern since genes that have a promoter-proximal paused Pol II, and do not produce full-length transcripts will produce detectable signal that does not reflect actual levels of full-length transcription of those genes*(12)*. In addition, the distribution of transcribing polymerases within genes provides information on how a particular gene is regulated, and when combined with our knowledge of promoter DNA sequences, transcription factor binding sites, and nucleosomes and their modifications, can further our knowledge of how these elements cooperate to specify distinct transcriptional outcomes.

Transcription in nuclei: reflection of in vivo transcription status:

Given that the nuclear run-on assay is performed in vitro, it is conceivable that some polymerases might bind and initiate transcription and/or elongate during isolation of the nuclei - prior to the addition of sarkosyl. However, several considerations suggest that very little if any transcription initiation or elongation occurs during nuclei isolation. Immediately before preparing nuclei, the cells are brought to ~4°C within seconds of removing the media. Under these conditions, even if a polymerase comes in contact with a promoter, it is unlikely to form a proper preinitiation complex (PIC) within the timeframe of the procedure (30min), and due to the high energy requirements of promoter DNA unwinding, even less likely to initiate transcription. Nucleotides are removed by washing within in the first 15 min of the procedure, thus initiation becomes impossible after this point. Also, experiments from Peter Cook's lab that utilized a combination of in vivo labeling of nascent transcripts with BrU followed by in vitro labeling with biotin-CTP have shown that no new initiation occurs in nuclei since all biotin-CTP sites also labeled with BrU *(13)*. These experiments by the Cook lab were carried out in the absence of sarkosyl, thus we think the event of observing initiation in isolated nuclei in the presence of sarkosyl, is very unlikely. Finally, high-resolution mapping experiments of pausing at the *Drosophila HSP70* gene have shown that Pol II does not elongate during the nuclei isolation site *(4, 14)*.

Further support that the nuclear-run-on reflects the in vivo state of transcription can be obtained by comparing the GRO-seq results with other assays that start with whole cell preparations. In the accompanying paper, Seila et al. *(15)*, show that small transcription start site RNAs (TSS-RNAs) are produced by promoters in both the forward and divergent direction. This is evidence that the transcription we detect at promoters with GRO-seq occurs in vivo. Also, ChIP data that shows that promoter regions are bound by Pol II is generated by cross-linking whole cells, thus Pol II-DNA interactions are occurring in vivo at the time of cross-linking *(16)*. These peaks of polymerase binding show nearly complete overlap with promoters called active by GRO-seq. GRO-seq identifies additional active promoters because of the increased sensitivity afforded by sequencing. In addition, recent Pol II ChIP-seq data from Sultan et al. *(17)* shows that Pol II is present in a peak that is resolvable from the peak at the transcription start site. Sultan et al. hypothesize that the upstream peak could be an upstream pre-initiation complex, or some sort of storage site for Pol II. We show that

this peak represents transcriptionally engaged Pol II complexes that are oriented in the opposite direction of gene transcription. This ChIP-seq data is further evidence that divergent polymerases can be detected from whole-cell preparations, and are not a consequence of polymerase binding during preparation of nuclei.

Development of GRO-seq

Incorporation of Br-UTP by nuclear RNA polymerases:

Given that the NRO-RNA represents a small fraction of the total RNA in nuclei (see below), analysis of NRO-RNA with conventional genomic platforms requires specific isolation of this RNA. To adapt nuclear run-ons for a global analysis, we reasoned that a nucleotide with an affinity purifiable tag could be added to the run-on reaction, and sought to test the incorporation and purification efficiencies as outlined below.

We first tested whether 5-Bromo-UTP (BrUTP) could be efficiently incorporated into RNA by nuclear RNA polymerases by also incorporating a radioactive nucleotide (α^{32}P-CTP) in a run-on time course experiment. Consistent with previous results*(18)*, addition of Br-UTP allowed incorporation at ~80% efficiency compared with UTP, and approximately 10 fold over the control that lacked both UTP and Br-UTP (Figure S9). These radiolabled RNAs made in the presence of Br-UTP bind very well to anti-Br-deoxy-U beads, which cross-reacts well with BrU (Figure S10) (see below). Although BrU is sometimes used as a mutagen, sequencing clones from GRO-seq libraries indicated the misincorporation rate by nuclear RNA polymerases is low. We also tested the propensity of BrU to cause misincorporation during reverse transcription by comparing sequencing results of cDNA clones that were generated from RT reactions that contain a BrU or U RNA template of known sequence. The results showed that there is no appreciable level of misincorporation by reverse transcriptase when BrU is incorporated into the RNA template. We thus chose BrU as our affinity tagged nucleotide for further development of the assay.

Control of resolution for GRO-seq:

The goal of the GRO-seq method is to isolate and obtain a high resolution and unbiased map of all RNAs as they are being transcribed. High resolution requires that run-on distances are kept short, whereas unbiased mapping requires efficient incorporation of the affinity-tagged nucleotide analog into all RNAs. We titrated nucleotide concentrations during

the run-on step and defined the minimum distance for library preparation as the lowest concentration that allows maximum binding of the run-on RNAs to beads. To determine the length of the run-on transcription, nuclei were first pre-treated with RNase in order to trim the nascent RNAs*(13)*. RNA polymerases can protect the nascent RNA from 15-25 bases upstream from the active site *(19, 20)*. The RNase activity was then removed through extensive washing and treatment with RNase inhibitor. The distance polymerases run-on was then controlled by titrating limiting concentrations of CTP. Since we primarily wanted to identify locations of RNA polymerase II (Pol II), we also examined the distance transcribed by polymerases in the presence of α-amanitin and actinomycin-D. α-amanatin is an efficient inhibitor of Pol II, but works much less effectively on Pol III, and is completely innocuous for Pol I transcription*(13)*. Actinomycin-D, when added to cells prior to nuclei isolation, primarily inhibits Pol I. By comparing the length of nascent transcripts produced from RNase treated nuclei and in the presence of inhibitors we were able to deduce the distance Pol II transcribes under various limiting nucleotide concentrations (Figure S11). Analysis of the efficiency of bead binding under similar conditions shows that with nuclei from IMR90 cells, 1uM CTP is sufficient to allow near maximum bead binding (Fig. S12). This corresponds to a run-on extension of ~80-100 nucleotides (Figure S11), which is the average length of the RNAs (~100 -120 nucleotides) subtracted by the length of RNAs protected by the polymerase (~20 nucleotides). We therefore consider 1uM CTP as the optimum concentration for these nuclei.

In non-RNase treated nuclei (which are used for creating the NRO-library) base hydrolysis of the nascent RNAs to an average size that is equal to the length of the run-on transcripts will then result in a final mapping resolution of approximately half this distance. Base hydrolysis of the RNA improves the resolution of this assay by severing the extended portions of the nascent RNA transcript that contain the nucleotide analog from distal regions that were transcribed prior to the run-on reaction. In this study, we allowed Pol II to run-on approximately 80-100 bases, thus we estimate our resolution to be 40-50bp from the location of the polymerase active site at the time of the assay.

Yield, enrichment and purity of nascent RNA after triple selection:

High sensitivity and specificity is desired in any genomic assay in order to decrease both false negative and false positive results. These parameters require that both the yield and enrichment of run-on RNAs be high relative to contaminant RNAs.

Enrichment by tracking radiolabled NRO-RNAs

To assess the specificity and efficiency of the purification, we first measured the enrichment of the nascent RNAs by incorporating a radiolabeled nucleotide (α-^{32}P-CTP) in run-on reactions performed in the presence of either UTP or Br-UTP. Quantification of the bound and unbound fractions from each reaction by scintillation counting showed that the enrichment by this method is ~450 fold for a single round of bead binding (Figure S13). Successive enrichment could not be examined because the amount of radioactivity in the UTP-RNA was below the limit of detection in the bound fraction after binding to a new set of beads. In order to assess whether contaminant RNA was able to cross-hybridize with BrU-RNA, we also performed a bead binding with α^{32}P-CTP radiolabeled, UTP-containing RNA in the presence of non-radioactive, BrU RNA. Under these conditions the level of radioactivity in the bound fraction was the same as CTP-labeled samples containing only UTP suggesting that cross-hybridization is negligible.

Measurement of enrichment and purity by RT-qPCR:

Since the amount of radiolabeled NRO-RNA measured in the above experiments is a minor fraction of the total RNA isolated from nuclei, it is possible that a significant amount of contaminant RNA still exists after triple selection. The total mass of RNA in the bound fraction after triple selection was near the limit of detection, and beyond the limit of detection for Br-U and U-RNA, respectively, thus we could not reliably measure the enrichment by UV spectrometry alone. We could determine that there was 50μg in the starting pool and 300ng in the elution from the third round of bead binding for the Br-UTP samples. We therefore added spiking controls consisting of multiple small (~100base) RNAs that were in vitro transcribed in the presence of either UTP or Br-UTP. The cDNAs used for in vitro transcription were reverse transcribed and amplified from Arabidopsis thaliana total RNA. U-RNAs were added in 10-fold dilutions from 1X10^{10}-1X10^{7} copies and a BrU-RNA was added at 1X10^{7} copies. After triple selection, reverse transcription followed by quantitative PCR

(RT-qPCR) was carried out on the final elution for each RNA. The Br-U RNA was present at 50% relative to input, and all U-RNAs were at or below background for the assay. The lowest amount of the input that we could detect was 1:10,000, therefore non-BrU RNAs could be present at 1:10,000th relative to the starting amount. This corresponds to 5ng since we start with 50μg of nuclear RNA. Since the final elution contains 300ng of RNA, U-RNA represents 1.6% of the final mass, corresponding to >98% purity for BrU-RNA.

In addition to the above results, several computational analyses suggest that our NRO-RNA libraries were highly enriched for NRO-RNA relative to accumulated RNAs. First, an estimation of background was determined by binning reads in 500kb windows genome wide. The distribution of windows with the lowest densities fits a Poisson distribution corresponding to spreading 2-3% of the aligned reads randomly over the mappable portion of the genome, agreeing well with the above experimental results and suggesting that background for the assay approaches 0.04 reads on a single strand per 1kb. Second, transcription is detected in regions of transcription units that are not present in fully processed mRNAs, including introns and regions beyond the site of nascent RNA cleavage and polyadenylation. The ratios of read density within introns vs. exons is 0.9 (Pearson correlation = 0.83), and is not significantly different from 1 (P = 0.71, Figure S14). Third, known gene deserts ranging from 0.6 Mb to 3 Mb, have an average density of reads on both strands together of 0.07 reads/1kb, which also agrees well with our experimental and computational analyses of background (Table S2).

Overview of GRO-seq method

This is intended to be a description of the overall method to accompany Figure S1. For a detailed description of the steps involved, please see the methods section below. Nuclei isolation and run-on reactions are performed using standard protocols with the exception that 5-Bromo-UTP is used in place UTP, and the concentration of CTP is adjusted to 1μM to keep the run-on distance to ~100 nucleotides (see above). α-32P-CTP is also used as a tracer in order to follow the purification steps, and analyze the products on denaturing PAGE. RNA is isolated and base hydrolyzed to the desired size. RNA fragments are then isolated by binding to anti-deoxy-BrU beads to select against accumulated nuclear RNAs, washed several times, and eluted from the beads. Because base hydrolysis of RNA leaves a molecule with a 5'-hydroxyl and a 3'-phosphate, neither of which are substrates for

ligation of adapter oligos, the RNA ends must be repaired. First, the RNAs are treated at low pH with tobacco acid pyrophosphatase to remove 5-methyl guanosine caps *(4)*, and then are treated at low pH with T4 polynucleotide kinase (PNK) to remove the 3'-phosphate*(21)*. The pH is then raised and the RNA is treated again with PNK, except now in the presence of ATP, to add a 5'-phosphate. An adapter is then added to the 5'-end with T4-RNA ligase and the RNA is bound to anti-deoxy-BrU beads to remove excess linkers and further enrich the RNA. This process is then repeated for the addition of a 3'-adapter. The affinity-enriched RNAs are then reverse transcribed, amplified, and PAGE purified. Analysis of a fraction of each step by denaturing polyacrylamide gel electrophoresis (Figure S15) shows that the RNA remains largely intact throughout the procedure. After amplification and PAGE purification (Figure S16), the library appears to be, on average, 100 bases in length (~190 base – 90 base adapters). A known amount of the library is re-amplified to determine the primer efficiency from which the original complexity of the cDNA library can be extrapolated. In the two libraries we constructed in this study, we obtained complexities of $1x10^9$ molecules prior to amplification. We also cloned and sequenced by conventional methods 50 molecules to verify the size and ensure the quality of the library before massively parallel sequencing on the Illumina 1G genome analyzer. Correlation of the read densities between the two replicates produced in this study show that replicates agree remarkably well (Spearmann correlation = 0.96, Figure S17).

GRO-seq reads and annotated transcript boundaries

Most reads align within or near the boundaries of known transcription units or expressed sequence tags (ESTs). 62.8% of reads align on the coding strand within Refseq genes. An additional 9.6% of reads align to the coding strand within the boundaries of Human mRNA, and a further 13.4% within EST coding regions (Figure S2). These values increase to 74.0%, 10.2%, and 12.8%, respectively, for a total of 97%, if the boundaries are expanded by 5kb from both the 5' and 3' ends of the annotated features.

Comparison of GRO-seq with Pol II ChIP-chip data from the Ren lab

To assess the relationship between promoters identified by transcription factor binding (i.e. ChIP) assays and the presence of engaged polymerase, we compared our GRO-seq densities with the list of over 10,000 active promoters identified in a previous study performed in the same cell line*(16)*. Active promoters in that study were identified genome-wide by

binding of TAF1, a component of the general transcription factor TFIID that is critical for specifying most sites of initiation by Pol II*(16)*. That study identified 9,324 TFIID binding sites within 2.5kb of annotated transcripts (referred to as transcript-matched) and 1,239 novel promoters that were greater than 2.5kb from known 5'-ends of genes. Of the promoters associated with annotated transcripts, 9,217 (98.9%) have coding-strand GRO-seq densities within the body of the associated gene significantly above background. Because the novel promoters have no associated orientation by ChIP, we assayed the neighboring +/- 1 kb region and found that 1,185 (95.6%) had GRO-seq densities significantly above background. Details of the statistical methods are described in the Methods section below. GRO-seq not only confirms these sites as active promoters, but also provides the direction and extent of transcription from these novel promoters (Figure S18). When we used GRO-seq densities alone to identify the number of active promoters within +/- 1 kb of RefSeq annotated 5'-ends, we find 16,882 active promoters. The increase in active promoters found here could be a consequence of different sensitivities, but may also reveal a class of promoters that are independent of TFIID*(22)*.

The Kim et al. study also reported that Pol II was bound to 97% of confirmed TFIID binding sites by performing ChIP-chip with an antibody that recognizes Pol II (antibody: 8WG16). This represented the most comprehensive Pol II ChIP data set at the time we began GRO-seq development, which is why we chose the IMR90 cell line. The 8WG16 antibody preferentially recognizes the hypo-phosphorylated form of the largest subunit of Pol II that is found at the 5'ends of genes. It has been demonstrated at many genes that as Pol II progresses further into a gene it becomes hyperphosphorylated, and thus a less suitable substrate for the antibody. Thus, in some cases the antibody will show a reduction in the downstream portions of a gene, that actually reflects a reduced affinity for Pol II in these regions. Therefore, we cannot directly compare GRO-seq density and ChIP density in the downstream region of most genes, since GRO-seq detects transcriptionally engaged Pol II regardless of phosphorylation state. In addition, the array used to analyze the Pol II ChIP data was essentially a promoter array, so there is no data in the downstream portion of longer genes. The above reasons explain why, in some of the figures presented in the main text and herein, Pol II ChIP signal appears concentrated only at the promoter regions, when in fact it is a result of the antibody used and the extent of the array design.

Comparison of GRO-seq to microarray expression data

We additionally determined how GRO-seq transcript densities in the sense orientation within gene regions compared to the microarray expression data available for this cell line*(16)*. First, microarray expression values plotted against GRO-seq densities reveal that accumulated, fully processed mRNA levels generally correlate with steady state transcription of genes obtained by GRO-seq (Figure S19). However, GRO-seq densities have a wider dynamic range that extends below the limit of detection by microarray (compare Figures S19A,B with S19C,D). To gauge the increase in sensitivity, we compared genes called absent or present by microarray to genes that could be called active or inactive by GRO-seq. For a gene to be called active by GRO-seq, we required the density within the downstream portions of genes to be significantly above background ($P < 0.01$). The first 1 kb was excluded from the analysis to avoid signals produced by promoter-proximal paused polymerases (see methods). When considering all RefSeq genes, 16,882 genes (68%) were classified as active by GRO-seq. When considering the genes covered by probes on the microarray, 16,858 genes were called active by GRO-seq, while only 8,438 were called active by microarray hybridization (Figure S19 Table S3). Active gene calls for GRO-seq span more than four orders of magnitude, whereas microarray experiments are restricted to approximately 2.5 orders of magnitude. The increased number of active genes in our GRO-seq analysis can be attributed to the increased sensitivity of sequencing technologies versus hybridization methodologies*(23, 24)*, and possibly due to the fact that nascent RNA libraries may be enriched for rare or unstable transcripts relative to highly accumulated RNAs. We confirmed the expression of several genes that were called active by GRO-seq but inactive by microarray by RT-qPCR (See Below).

<u>Validation of GRO-seq gene activity by RT-qPCR:</u>

Transcripts that are regulated by post-transcriptional mRNA turnover can be identified by comparing mRNA levels to GRO-seq densities. A highly stable transcript would be expected to have a high level of mRNA expression compared to the GRO-seq density within the corresponding gene, while unstable transcripts would be expected to have higher GRO-seq densities relative to mRNA expression level. By comparing GRO-seq with expression microarray data we identified candidates as stable or unstable transcripts by searching for genes that were microarray active : GRO-seq inactive or microarray inactive : GRO-seq active, respectively. We then compared several of these genes to genes that were found to

be active in both assays by performing RT-qPCR. We first ranked the genes from each class into deciles of gene activity as determined from the GRO-seq density within gene bodies. We then chose genes from a range of activity deciles to validate. The results show that all genes tested that are called active by GRO-seq can be detected by RT-qPCR after priming the reverse transcription with either random hexamers or oligo-dT to extents that generally mirror their level of GRO-seq transcription (Figure S20). In addition, genes that were not detected by the microarray had similar RT-qPCR levels as those that were not detected by the arrays. These results verify GRO-seq as a general and sensitive method for detecting active genes, and suggest that many genes are not detected by the microarray due to insufficient sensitivity or incorrect probe design. Two genes (COL1A1, IGFBP5) may be highly stabilized transcripts because they are called active by both microarray and GRO-seq, but were detected by microarray at much higher levels than other genes that are inactive by microarray but have similar GRO-seq densities.

Accumulated mRNA levels and GRO-seq density on the body of genes, generally showed a strong concordance in IMR90 cells (Figure S19, S20). The relatively limited dynamic range and sensitivity of the microarray data may have caused some less stable RNAs to be missed. Also, classes of genes that are regulated by mRNA stability might be more readily detectable in response to changing environments *(10, 11)*. Comparison of GRO-seq to RNA-seq data should also improve the efficiency of identifying mRNAs that are regulated by mRNA turnover rates *(23-26)*.

Supplemental information on promoter-proximal pausing

Pausing vs Stalling terminology:

We chose to describe genes that have transcriptionally engaged Pol II accumulated at the 5'-end as 'paused' since this pattern mirrors that of several human and drosophila genes that have been identified as paused (see below). Pausing refers to a polymerase that is engaged in transcription, is either not moving forward or moving slowly, but nonetheless retains its elongation potential. Since the nuclear run-on (NRO) assay that we have used here requires the polymerase to be transcriptionally competent, it is fitting to describe the polymerases that we see accumulated at promoters as paused. The term 'stalled' is sometimes used to describe a polymerase that is found at higher levels at the 5' ends of a gene*(29-31)*. Stalling refers to an engaged polymerase complex, but makes no assumption about whether that polymerase is competent to resume elongation *(32)*. That is, a stalled

polymerase could be paused, backtracked and arrested, or could exist is some form of dynamic equilibrium between the two states. The potassium permangate footprinting assay can be used to map the location of a paused or stalled polymerase *(33)*. This technique maps the unwound portion of the template DNA that is associated with an engaged polymerase. In the absence of further experimentation that examines transcriptional competence, genes that have excessive permanganate reactivity at the 5' end corresponding to the position of a paused polymerase are generally described as experiencing stalling *(29, 30, 34)*.

Pausing, termination, or both?:

Whereas we have clarified our use of terminology here, we are, however, uncertain whether the engaged complexes that we detect at the 5'-end of genes will actually proceed to transcribe to the end of the associated gene given the proper signal. It is possible that some of these polymerases will eventually terminate prematurely in a manner that has been observed for transcription of HIV genes in the absence of the transactivator Tat *(35, 36)*. For instance, the presence of promoter-proximal engaged polymerase peaks could also be observed if a promoter experienced high rates of initiation but also high rates of premature termination relative to the amount of polymerases that escape into productive elongation. Under these circumstances, one could expect to detect high levels of engaged polymerases immediately prior to the point of termination. Further experimentation and development of new methodologies are required to distinguish be these possibilities *in vivo*. However, our results do show that the transition from initiation to elongation can be rate limiting to gene transcription, whether or not it occurs through holding back a polymerase and causing it to pause or by causing premature termination, or though a combination of pausing and termination.

Pausing and gene activity:

As gene activity increases, it is expected that the occupancy of Pol II at promoters will also increase. This is borne out in ChIP data, as well as here in our GRO-seq data. Figure 3B shows that GRO-seq density within promoter-proximal regions generally increases as the density of reads in the body of genes increase. However, pausing indices have an inverse correlation with gene activity. This relationship could reflect that highly expressed genes

either do not experience pausing, or they transition through pausing faster, allowing more polymerase to enter into productive elongation. When we examine the fraction of paused genes according to gene activity deciles (Figure S21), we find that the fraction of paused genes increases with increasing gene activity and represent 63% of the highest decile of gene transcription. This result, in combination with the inverse correlation between gene body density and pausing indexes, indicates that highly active genes, relative to genes with lower activity, not only recruit more polymerase and stimulate faster pause site entry rates, but they must also increase pause site escape to a greater extent in order to account for these profiles.

Gene Ontology:

Significantly paused genes are enriched with biological processes such as cell cycle regulation, stress response, and protein biosynthesis (ribosomal proteins), and are de-enriched for developmentally regulated genes (Figure S22). Although previous studies identified developmentally regulated genes as enriched in the paused class *(29, 30, 37)*, these studies used either embryonic stem cells, an embryonic-derived cell line, or developmentally staged *Drosophila* embryos. The differences likely reflect the more differentiated state of the primary fibroblasts used in this study.

GRO-seq results for known paused genes:

Several human genes have been shown to have a high level of transcriptionally engaged Pol II at the 5'-end relative to the downstream portions either by traditional NRO-hybridization assays, or by potassium permanganate footprinting. The genes include MYC *(38, 39)*, FOS *(40)*, DHFR *(41)*, ACTG1 (γ-Actin) *(41)*, and HSPA1A (HSP70) *(42)*. The first four genes do exhibit a pattern consistent with pausing (Figure S23), and are called significantly paused by our analysis. The human genome has two nearly identical copies of the HSP70 gene, and could not be analyzed, because reads mapping to multiple locations were removed before any analysis performed.

Divergent transcription and histone modifications

The histone modifications (H3K4me3, H3K4me2, H3/H4ac) that mark active promoters generally occur at the +1 and -1 nucleosomes relative to the transcription start site (TSS).

The +1 nucleosome is downstream of the TSS, and is thus associated with initiation, but the modification of the upstream (-1) nucleosome is generally assumed to occur due to the simple proximity of the -1 nucleosome to the control sequences of the promoter. This would suggest that the mechanism by which these modifications are laid down has no strict directionality. Based on our GRO-seq results, and the ubiquity of divergent transcription, an alternative explanation could be that the -1 nucleosome has these modifications either as a consequence of, or perhaps, to allow formation of divergently-engaged polymerase. These modified histones in nucleosomes –1 and +1 are immediately downstream of the divergent polymerase peak and pause peak respectively. To examine this hypothesis, we replotted the available histone modification data produced by the Ren lab versus genes that do or do not have significant levels of divergent transcription (P>0.001)(the plot in Figure 4F was of all genes). As shown in Figure S24, the peak of H3K4me2 and H3ac at these genes is less defined in the upstream region compared to the region downstream of the TSS. The presence of these modifications in the upstream region can likely be accounted for by the small but identifiable peak of anti-sense GRO-seq reads at ~-250. Thus, the mechanism of placing these histone modifications might be tightly associated with the mechanism of forming these early elongation complexes.

Antisense transcription in gene regions

A number of studies have reported that gene regions are transcribed in the reverse orientation with unanticipated high frequency. Transcript pairs have been identified that overlap at the 5'-ends, 3'-ends, or with full overlap*(27, 28)*. Although antisense reads in gene regions account for only 6% of the total reads, ~14,545 genes (58.7%) have antisense transcription significantly above background (P < 0.01). Of these genes, 273 are accounted for by active annotated genes that overlap at the 5'-end, 4,407 by active convergent genes with a maximum separation of 10kb, and 242 by active annotated genes with full overlap (Figure S3).

Methods

Isolation of nuclei

Isolation of nuclei was carried out as described in*(39)*, with several modifications. 15cm^2 plates of IMR90 cells (~6X10^6 cells at 80% confluency) were washed directly on the plate 3X with ice cold PBS. 10ml of ice cold swelling buffer (10mMTris-cl pH7.5, 2mM MgCl2, 3mM CaCl2) was added and allowed to swell on ice for 5 min. Cells were removed from the plate with a plastic cell scraper, transferred to a 15 ml conical, and pelleted for 10 min at 4°C at setting 3 on an IEC clinical centrifuge. Cells were resuspended in 1ml of lysis buffer (swelling buffer + 0.5% Igepal, + 10% glycerol + 2units/ml SUPERase In (ambion)), and gently pipetted up and down 20 times using a p1000 tip with the end cut off to reduce shearing. The volume was brought to 10 ml and nuclei pelleted at setting 4 on an IEC clinical centrifuge. The nuclei were washed and pelleted once in Lysis buffer, resuspended in 1ml Freezing buffer (50mM Tris-CL pH 8.3, 40% glycerol, 5mM MgCl2, 0.1 mM EDTA), and transferred to a 1ml tube. Nuclei were pelleted at 1000Xg, and resuspended in 100ul of Storage Buffer / 5X10^6 nuclei.

NRO-RNA library construction

Construction of a NRO-library for sequencing involves the run-on reaction, base hydrolysis, immuno-purification, end repair, 5'- and 3'- adapter ligation, amplification, and PAGE purification.

NRO reaction

5X106 IMR90 nuclei (100ul) were mixed with an equal volume of reaction buffer (10mM Tris-Cl pH 8.0, 5mM MgCl2, 1mM DTT, 300mM KCL, 20 units of SUPERase In, 1% sarkosyl, 500uM ATP, GTP, and Br-UTP, 2μM CTP and 0.33μM α-32P-CTP (3000Ci/mmole)). The reaction was allowed to proceed for 5 min at 30°C, followed by the addition of 23ul of 10X DNAseI buffer, and 10ul RNase free DNase I (Promega). Proteins were digested by addition of an equal volume of Buffer S (20mM Tris-Cl pH 7.4, 2% SDS, 10mM EDTA, 200ug/ml Proteinase K (invitrogen)), followed by incubation at 55°C for 1 hour. RNA was extracted twice with acid Phenol:chloroform, and once with chloroform, and precipitated at a final concentration of 300mM NaCl, with 3 volumes of -20°C ethanol. The pellet was washed in 75% ethanol before resuspending in 20ul of DEPC-treated water.

Base hydrolysis of RNA

Base hydrolysis was performed on ice by addition of 5ul 1M NaOH and incubated on ice for 30min. The reaction was neutralized by addition of 25 ul 1M Tris-Cl pH 6.8. The reaction was then run twice through a p-30 RNAse-free spin column (BioRad), according to the manufacturer's instructions. Before moving on to the immuno-purification, DNA was further removed by another digestion with RNase-free DNaseI for 10 min at 37°C, and the reaction stopped by addition of 10mM EDTA.

Immuno-purification of Br-U RNA.

Anti-deoxyBrU beads (Santa Cruz Biotech) were blocked in 0.5X SSPE, 1mM EDTA, 0.05% tween, 0.1% PVP, and 1mg/ml ultrapure BSA (Ambion). NRO-RNAs were heated to 65°C, added to 100ul beads in 500ul of binding buffer (0.5XSSPE, 1mM EDTA, 0.05% tween), and allowed to bind 1hour while rotating. The beads were washed once in low salt buffer (0.2X SSPE, 1mM EDTA, 0.05% Tween), twice in high salt buffer, 0.5% SSPE, 1mM EDTA, 0.05% Tween, 150mM NaCl), and twice in TET buffer (TE + 0.05% Tween). The Br-U RNA is then eluted 4X 125ul of Buffer E (20mM DTT, 300mM NaCl, 5mM Tris-cl pH 7.5, 1mM EDTA, and 0.1% SDS). The RNAs are then extracted and precipitated as above.

End Repair

Enriched RNAs were resuspended in 20ul DEPC-treated water, and incubated with 2.5ul Tobacco acid pyrophosphatase (TAP, Epicentre Biotechnologies), 1X TAP buffer, and 1ul SUPERase Inhibitor in a final volume of 30ul at 37°C for 1hour. 1ul of Polynucleotide Kinase (PNK, NEB), and 0.5ul of 5mM MgCl2 is then added and the reaction continued for 30min. 20 ul PNK buffer, 2ul 100mM ATP, and 145ul water, and 1ul PNK is then added and the reaction continued for another 30 min. 90ul water and 10ul 500mM EDTA, is then added, followed by extraction and precipitation of the RNA.

Adapter ligations

For adapter ligations the RNA was resuspended in 8.5ul, and incubated with 2.5ul of either the 5'- or 3'- adapter oligo (Small RNA Isolation Kit, Illumina), 1ul SUPERase In, 2ul RNA ligase-1 buffer, 5ul 50% PEG 8000, and 1.5ul of T4 RNA ligase-1 (NEB). The reactions were incubated on the lab bench for 4 hours. After both the first and second adapter ligations the RNAs were enriched over anti-deoxy-BrU beads as described above.

Reverse transcription and amplification and PAGE purification of NRO-RNA libraries

The RNAs were reverse transcribed (otherwise according to the manufacturer's specifications) in two separate 10ul reactions, with 0.5ul 100uM RT-Primer (Illumina Small RNA Isolation Kit), and 1ul SIII reverse transcriptase (invitrogen), at 44°C for 15min, followed by 52°C for 45 min. The RNAs were degraded by addition of RNAse cocktail (Ambion), and RNAse H (Ambion), and amplified 15 cycles, with Phusion high fidelity DNA polymerase (Finnzymes) using the PCR primers specified by Illumina. The NRO-cDNA libraries were then run on a non-denaturing 1XTBE, 8% acrylamide gel, and cDNAs greater than 90 nucleotides were excised from the gel and eluted by incubating in TE + 300mM NaCl overnight while rotating. The library was then extracted, precipitated, and then sent to Illumina for sequencing on the 1G Genome Analyzer.

Data Analysis

Alignment of GRO-seq reads to the human genome:

Two independent biological replicates were submitted for sequencing at Illumina. Library 1 was sequenced on three channels and yielded 13,818,931 total reads while library 2 was sequenced on two channels and yielded 9,389,058 reads. All reads were 33 bases long. Alignments to the hg18 assembly of the human genome were performed with the Eland alignment tool from Illumina. 5,316,960 full length reads from library 1 aligned uniquely to the human genome and 4,459,581 full length reads from library 2 aligned uniquely to the human genome. Alignments allowed up to two mismatches per sequence to account for sequencing errors and SNPs between the IMR90 cell line and the sequenced genome. To increase the coverage of our libraries, we then iteratively trimmed one base from the 3' end of reads that did not align uniquely and checked if they now aligned uniquely at the reduced length. Trimming was done from the 3' end, because the quality score for reads was highest at the 5' end and lowest in the 3' end, and because it is possible that some of our amplified library was shorter than the 33 bases sequenced. Analysis of the correlation between the two libraries as a function of trimming extent showed that 29 bases was the optimal minimum length to be included (Figure S25). Alignments were done to the full (non-repeat masked) human

genome. While unique alignments can be achieved in repeat masked sequences, we analyzed the number of reads mapping to such repeat masked sequences to be sure they were trust worthy. With the exception of rRNA repeats, the density of alignments to repeat regions mirrored the average overall density of surrounding regions, suggesting that they were indeed accurate. The rRNA repeats however had an average density roughly five orders of magnitude above the average genome wide level. Since rRNA is the most abundant mature RNA in the cell, we reasoned that this would be the major non-NRO RNA contaminant in our purifications, and thus we removed all alignments to rRNA repeats in the genome. These steps increased the total number of reads aligned to the genome to 5,800,577 for library 1 and 4,950,956 for library 2, for a total of 10,751,533 unique alignments. Since sequencing was performed from the 5' end of the BrU purified NRO RNA, the 5' coordinate of each read was used as the position of engaged polymerase for all future analyses.

Identifying mappable bases in the genome:

To assess the fraction of the genome where reads could be expected to align, all unique 32 base sequences from both strands of the hg18 assembly were identified. This is a total of 2,414,845,175 32-mers per strand from a total possible 3,080436,051 per strand. A mappable or unmappable base refers to the 5' base of a given mappable or unmappable 32-mer. All calculations of read densities in future analyses were relative to these mappable bases.

Background calculation from low-density windows:

To assess the background GRO-seq density, the genome was divided into 500 kbp windows and the density of reads in each window was calculated. The distribution of low-density windows is described very well by placing 3% of the total GRO-seq reads randomly on the mappable portion of the genome (Figure S26). The blue theoretical curve is described by

$$p(x) = \frac{\lambda^{x*l} e^{-\lambda}}{(x*l)!}$$

where x is the density of reads on both strands per base pair, l is the window size (500 kb), and λ is the background density of reads (in units of reads/bp).

$$\lambda = \frac{f * N_{reads}}{L_{mappable}}$$

f is the fraction of all reads that are from background (0.03 in Figure S26), N_{reads} is the total number of reads aligning to the genome (10,751,533) and $L_{mappable}$ is the total number of mappable 32-mers in the genome summed over both strands (4,829,690,350).

Background calculation from gene deserts:

Sixteen separate 'gene deserts' were identified where most GRO-seq alignments should represent background. These regions ranged in size from roughly 500 kb to nearly 7 Mb. The details of the coordinates of these gene deserts and the number of GRO-seq reads are in Table S2.

Calculation of gene activity:

Gene activity was defined as *N*/*L* where *N* is the number of coding strand GRO-seq reads from +1kb (relative to the TSS) to the end of each gene, and *L* is the number of mappable bases in this region. The significance of a given gene's activity level was determined by the probability of observing at least *N* reads in an interval of length *L* from a Poisson distribution of mean λ = 0.04 reads/kb (the background density of our libraries).

$$P = \sum_{n=N}^{\infty} \frac{(\lambda * L)^n e^{-\lambda * L}}{n!}$$

If the probability was less than 0.01, the gene was called active. The first kilobase of each gene was omitted to better gauge the density of polymerase that actively elongates through the gene and to avoid over-counting from the increased density of paused polymerase in the 5' end of the gene. All analyses were done with the complete RefSeq gene list for the hg18 assembly of the human genome reduced to include only genes at least 3kb in length so that the measurement of GRO-seq density in the body of the gene would be robust.

Correlation of GRO-seq densities with microarray expression data:

The previous expression microarray work *(16)* was performed on the Affymetrix U133Plus2 array. To correlate the GRO-seq data with this expression array data, the original array data was downloaded from the supplementary material of that paper, and the knownToRefSeq and knownToU133Plus2 tracks from the UCSC genome browser were used to map RefSeq

genes to probe IDs. The analysis of the array data was performed as in the original paper*(16)*. That is, a probe had to be present or absent in both replicates to be called present or absent. If all probes mapping to a particular gene are absent then the gene is absent and if any probes mapping to a particular gene are present then the gene is present. All other genes are considered ambiguous and removed from future analyses.

Identification of promoter proximal peaks:

The exact position of many TSSs are not precisely annotated and many promoters in fact do not have a single well defined TSS*(43)*. Therefore, in order to identify the peak of promoter proximal coding strand GRO-seq reads, we tiled around each annotated TSS 1kb upstream and downstream in 50 bp windows, shifting by 5 bp. In each window we counted the number of coding strand reads and the number of mappable bases. We could then calculate the significance of the density in each window by comparing to the background density of 0.04 reads/kb in a manner similar to how gene activity significance was calculated (see above). The most significant window was chosen as the promoter proximal peak, and if multiple windows had the same significance, then the most 5' of these windows was chosen. If the promoter proximal peak had a p value less than 0.001, the gene was identified as having a significant promoter proximal activity. To identify the divergent peak, a similar approach was used but tiling was done +/- 1 kb from the identified promoter proximal peak and only reads on the noncoding strand were counted. The same p value cutoff of 0.001 was used to classify genes as having a significant peak of divergent transcription.

Identification of paused genes:

Significantly paused genes were identified by using the Fisher exact test to compare the density of reads in the sense strand promoter proximal peak to the density of reads in the body of the gene as compared to a uniform distribution of all these reads based on the number of mappable bases. A p value cutoff of 0.01 was used to call significantly paused genes.

Extending peaks to transcribed regions

To measure how far the significant promoter proximal peaks could be extended into transcribed regions we began by identifying the 3' most read within the peak (in a strand specific manner), and calculated d(n), the distance from the current read to the n^{th}

downstream read on the same strand. If this distance was less than the cutoff distance, the 3' boundary of the peak was extended to this n^{th} read and the process was repeated by shifting one read downstream. This process continued until the peak could no longer be extended. The value of *n* used in this analysis was 5 and the length cutoff was 2.5 kb.

Correlation of GRO-seq and ChIP-chip data:

The previous ChIP-chip data was reported for positions relative to the hg16 assembly of the human genome*(16)*. The UCSC liftOver tool was used to convert these coordinates to the hg18 assembly. To assess GRO-seq levels around the TAF1 peaks identified in the previous work we looked either at the GRO-seq density of the associated gene for the transcript-matched promoters, or 1kb upstream and downstream for the novel promoters. For the transcript-matched promoters gene activity values and significance were calculated as described above. For the novel promoters, we counted the total number of reads on both strands and the number of mappable bases. To identify significant transcription, we used a p value cutoff of 0.01 when comparing to the probability of obtaining that number of reads or more from a Poisson distribution with a rate of ~0.08 reads/kb because both strands are being counted.

Summary of GRO-seq method:

We have presented here a new methodology for documenting transcribed regions in the human genome by isolation and large-scale sequencing of nascent RNAs. GRO-seq is efficient, requiring only ~$5x10^6$ cells/library, and the resulting NRO-cDNA library is highly enriched relative to total RNA. We have shown that this technology can map polymerase locations with precision, and that this allows the identification of active promoters and their directionality. The distribution of transcriptionally engaged polymerases around gene regions can identify interesting characteristics of promoters and gene regions such as promoter-proximal pausing, internal pausing, co-transcriptional cleavage of the nascent RNA, the distance Pol II travels beyond annotated 3' ends before termination, and the level antisense transcription within genes.

References

1. P. Gariglio, J. Buss, M. H. Green, *FEBS Lett.* **44**, 330 (1974).

2. P. Gariglio, M. Bellard, P. Chambon, *Nucleic Acids Res.* **9**, 2589 (1981).

3. A. E. Rougvie, J. T. Lis, *Cell* **54**, 795 (1988).

4. E. B. Rasmussen, J. T. Lis, *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7923 (1993).

5. D. K. Hawley, R. G. Roeder, *J. Biol. Chem.* **260**, 8163 (1985).

6. J. Lis, *Cold Spring Harb. Symp. Quant. Biol.* **63**, 347 (1998).

7. I. Faro-Trindade, P. R. Cook, *Biochem. Soc. Trans.* **34**, 1133 (2006).

8. N. J. Proudfoot, *Trends Biochem. Sci.* **14**, 105 (1989).

9. N. Gromak, S. West, N. J. Proudfoot, *Mol. Cell. Biol.* **26**, 3986 (2006).

10. M. Schuhmacher *et al.*, *Nucleic Acids Res.* **29**, 397 (2001).

11. J. Garcia-Martinez, A. Aranda, J. E. Perez-Ortin, *Mol. Cell* **15**, 303 (2004).

12. L. J. Schilling, P. J. Farnham, *Nucleic Acids Res.* **22**, 3061 (1994).

13. D. A. Jackson, F. J. Iborra, E. M. Manders, P. R. Cook, *Mol. Biol. Cell* **9**, 1523 (1998).

14. A. E. Rougvie, J. T. Lis, *Cell* **54**, 795 (1988).

15. A. Seila, et al., *Science*(in press).

16. T. H. Kim *et al.*, *Nature* **436**, 876 (2005).

17. M. Sultan *et al.*, *Science* **321**, 956 (2008).

18. F. J. Iborra, A. Pombo, D. A. Jackson, P. R. Cook, *J. Cell. Sci.* **109 (Pt 6)**, 1427 (1996).

19. W. Gu, M. Wind, D. Reines, *Proc Natl Acad Sci U S A* **93**, 6935 (1996).

20. M. L. Kireeva *et al.*, *Mol Cell* **18**, 97 (2005).

21. V. Cameron, O. C. Uhlenbeck, *Biochemistry* **16**, 5120 (1977).

22. K. L. Huisinga, B. F. Pugh, *Mol. Cell* **13**, 573 (2004).

23. B. Wold, R. M. Myers, *Nat. Methods* **5**, 19 (2008).

24. B. T. Wilhelm *et al.*, *Nature*(2008).

25. U. Nagalakshmi *et al.*, *Science* **320**, 1344 (2008).

26. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, *Nat. Methods*(2008).

27. S. Katayama *et al.*, *Science* **309**, 1564 (2005).

28. P. Kapranov, A. T. Willingham, T. R. Gingeras, *Nat. Rev. Genet.* **8**, 413 (2007).

29. G. W. Muse *et al.*, *Nat. Genet.* **39**, 1507 (2007).

30. J. Zeitlinger *et al.*, *Nat. Genet.* **39**, 1512 (2007).

31. S. Nechaev, K. Adelman, *Cell. Cycle* **7**(2008).

32. R. N. Fish, C. M. Kane, *Biochim Biophys Acta* **1577**, 287 (2002).

33. C. Giardina, J. T. Lis, *J. Biol. Chem.* **268**, 23806 (1993).

34. C. Lee *et al.*, *Mol. Cell. Biol.* **28**, 3290 (2008).

35. S. Y. Kao, A. F. Calman, P. A. Luciw, B. M. Peterlin, *Nature* **330**, 489 (1987).

36. R. A. Marciniak, P. A. Sharp, *EMBO J.* **10**, 4189 (1991).

37. M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, R. A. Young, *Cell* **130**, 77 (2007).

38. A. Krumm, T. Meulia, M. Brunvand, M. Groudine, *Genes Dev* **6**, 2201 (1992).

39. L. J. Strobl, D. Eick, *Embo J* **11**, 3307 (1992).

40. J. Fivaz, M. C. Bassi, S. Pinaud, J. Mirkovitch, *Gene* **255**, 185 (2000).

41. C. Cheng, P. A. Sharp, *Mol Cell Biol* **23**, 1961 (2003).

42. S. A. Brown, A. N. Imbalzano, R. E. Kingston, *Genes Dev* **10**, 1479 (1996).

43. P. Carninci *et al.*, *Nat. Genet.* **38**, 626 (2006).

Supplementary figures S1-S8, Table S1

Figure S1

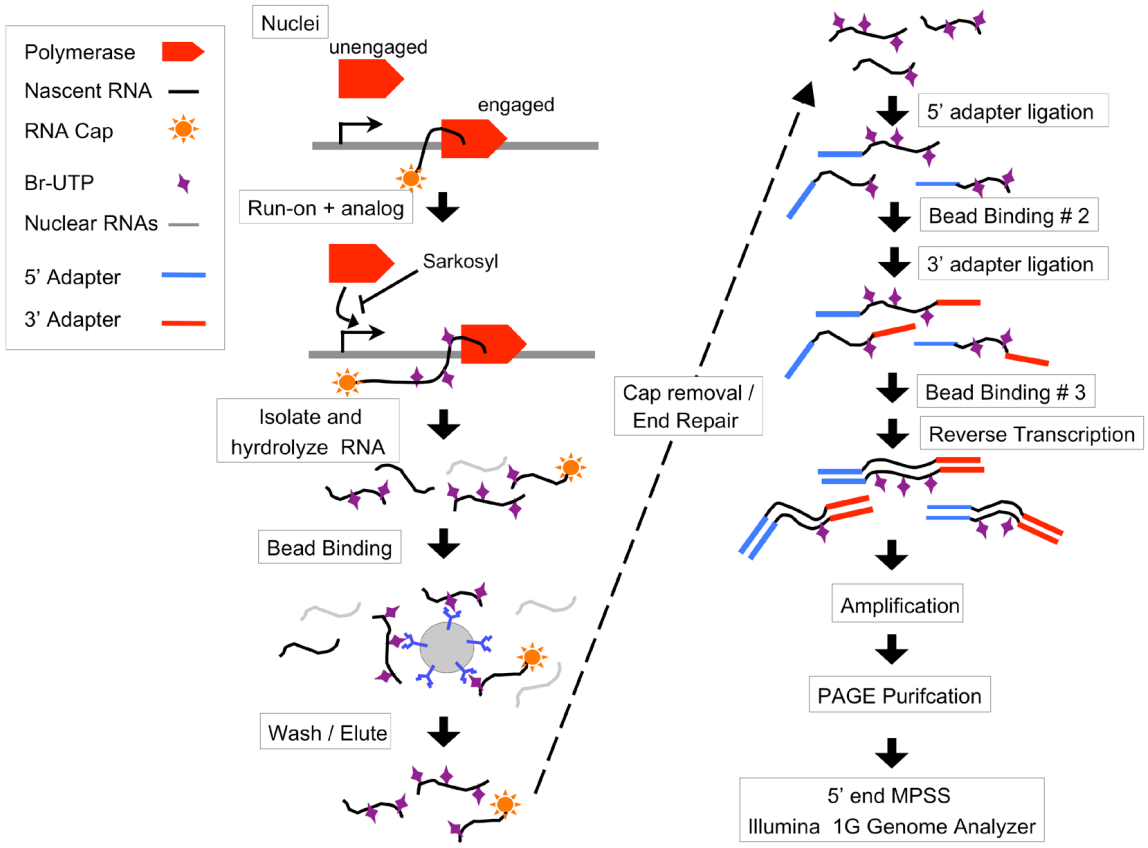


Figure S1: Overview of the GRO-seq method. Polymerases are allowed to run-on ~100 bases in isolated nuclei in the presence of sarkosyl and Br-UTP. The RNA is then base hydrolyzed to ~100 bases and bound to agarose beads that are coated with an α-BrdUTP antibody. 5'-7meG caps are then removed, and the ends of the RNA are prepared for adapter ligations. Illumina small RNA adapters are added to the 5' end follwed by the 3'end, with an additional round of immuno-enrichment after each adapter ligation. The RNAs are then reverse transcribed, amplified, and PAGE purified prior to sequencing from the 5'end on the Illumina 1G genome analyzer. See SOM for detailed description and protocol.

Figure S2

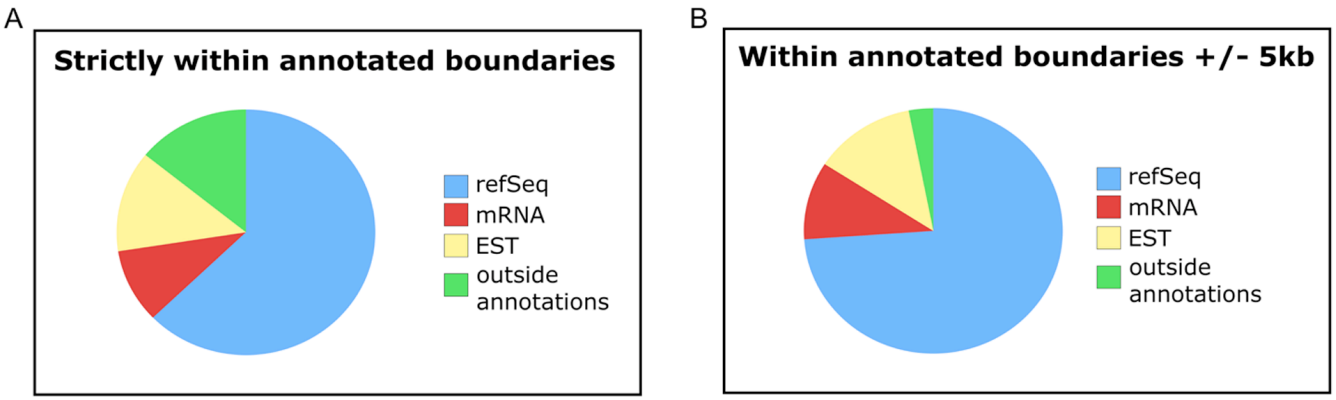


Fig. S2. GRO-seq reads relative to gene annotations. The fraction of reads aligning to the coding strand and strictly within the annotated boundaries (A) or within the annotated boundaries expanded by 5 kb (B). Reads were first mapped to RefSeq genes (blue), then unmapped reads were mapped to Human mRNA (red), then reads that were still unmapped were mapped either to Human ESTs (yellow) or outside annotations (green).

Figure S3

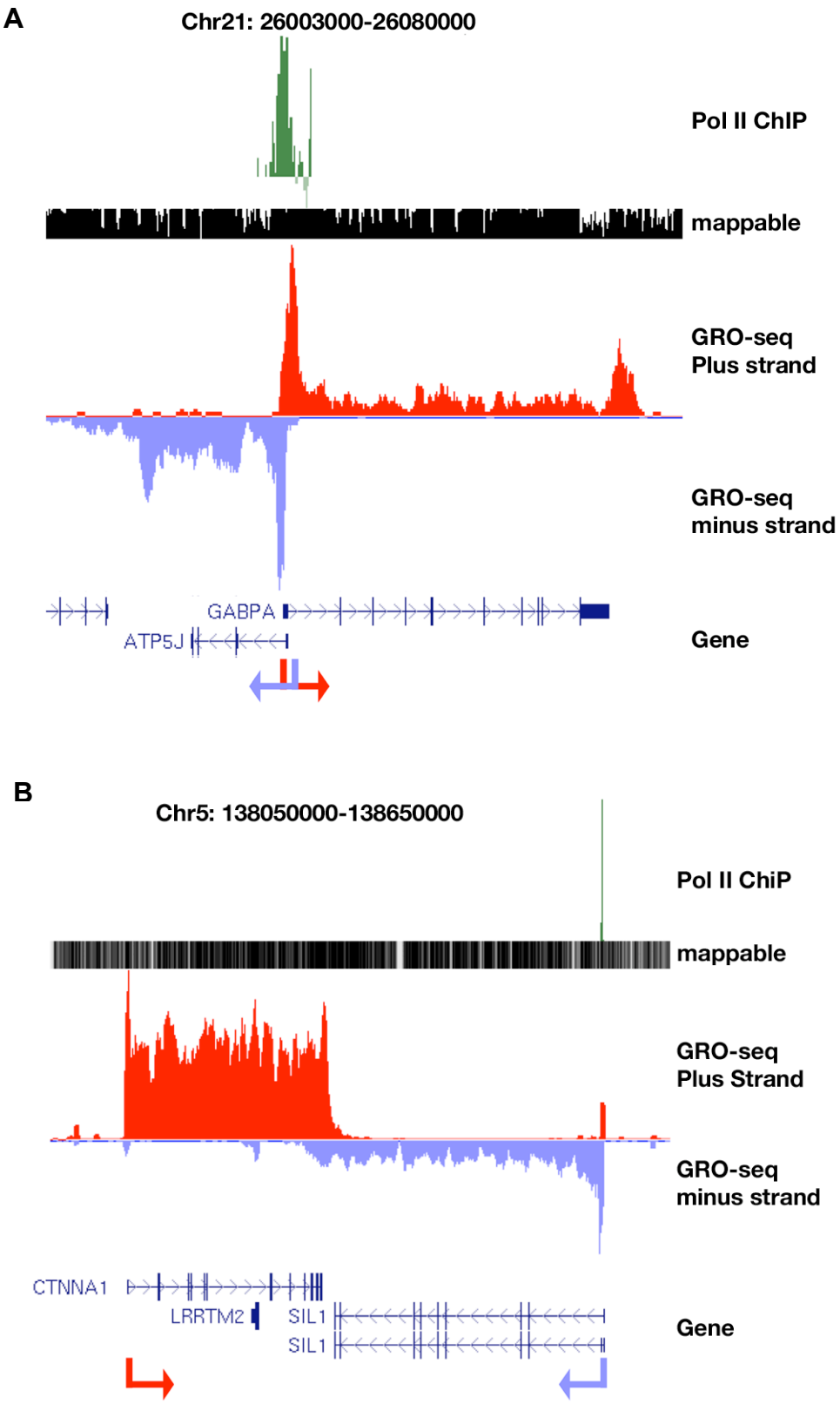

A
Chr21: 26003000-26080000
Pol II ChIP
mappable
GRO-seq
Plus strand
GRO-seq
minus strand
GABPA
ATP5J
Gene
B
Chr5: 138050000-138650000
Pol II ChiP
mappable
GRO-seq
Plus Strand
GRO-seq
minus strand
CTNNA1
LRRTM2
SIL1
SIL1
Gene

Figure 3 (continued)

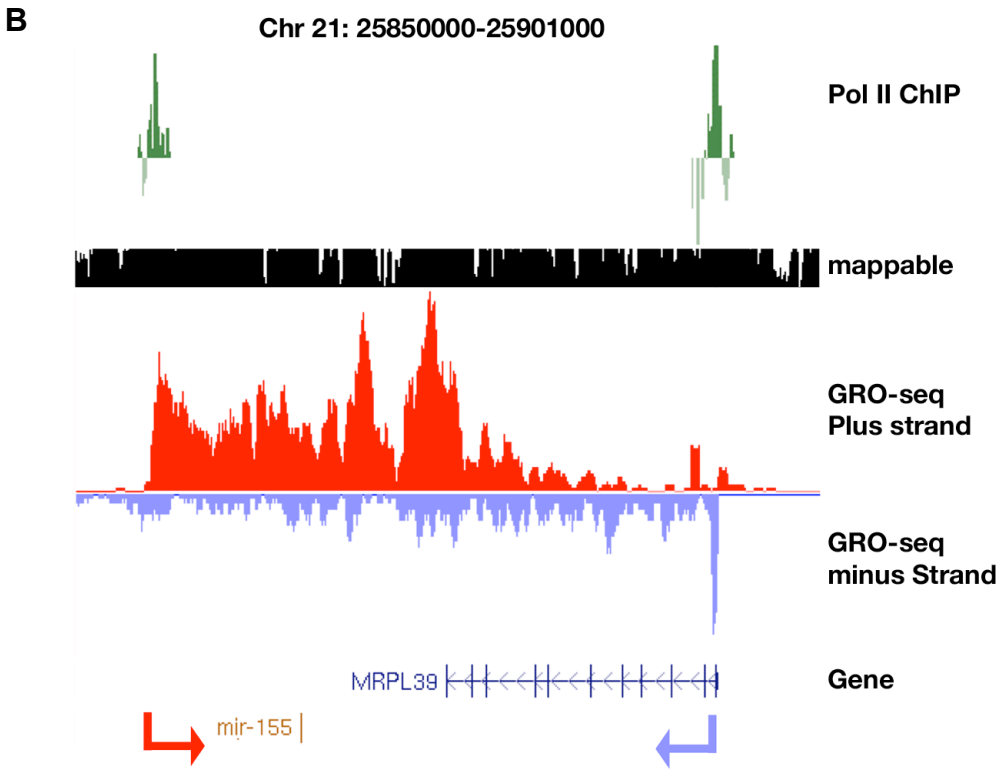


Figure S3. Identification of antisense transcription. Thee representative loci that show three types of antisense transcription identified previously by others, and presently in this study. The number of occurrences of (A) 5'-overlapping, (B) 3'-overlapping (convergent), and (C) fully overlapping antisense transcription is 273, 4,407, and 242, respectively.

Figure S4

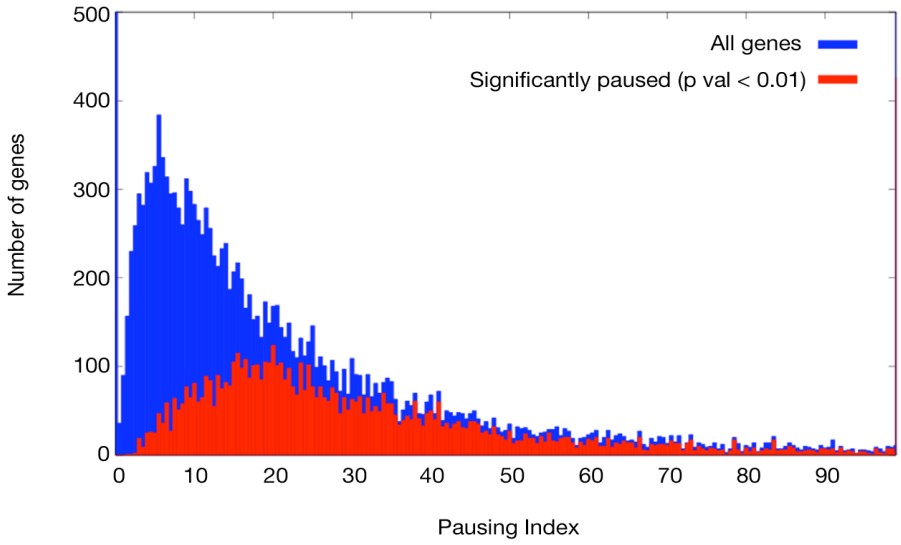


Figure S4. Histogram of pausing indices. Pausing indices for all genes (blue) or significantly paused genes (red) were binned in windows of width 0.5 from 0 to 100. There are 3,907 genes with a pausing index less than 0.5. The smallest pausing index amongst the significantly paused genes is 1.65 and the largest pausing index is 8661.2 in both distributions.

Figure S5

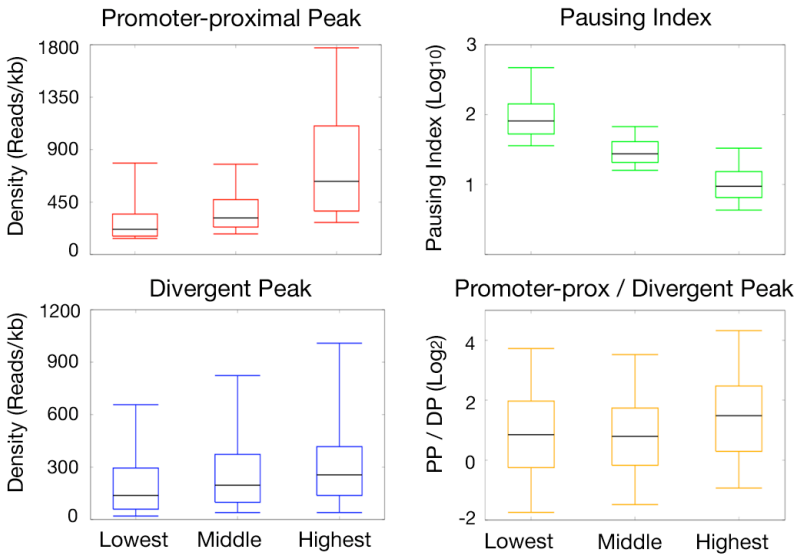


Figure S5. Correlation of promoter features with gene activity for paused genes only (A-D) Box plots that show the relationship of Promoter-proximal (PP) sense peaks (red), divergent peaks (DP) (blue), Pausing indices (green) and PP/DP ratios (orange) to the top, middle and bottom deciles of gene activity. Only genes with significant levels of pausing were used in the analysis. Box plot boundaries are as in Figure 3.

Figure S6

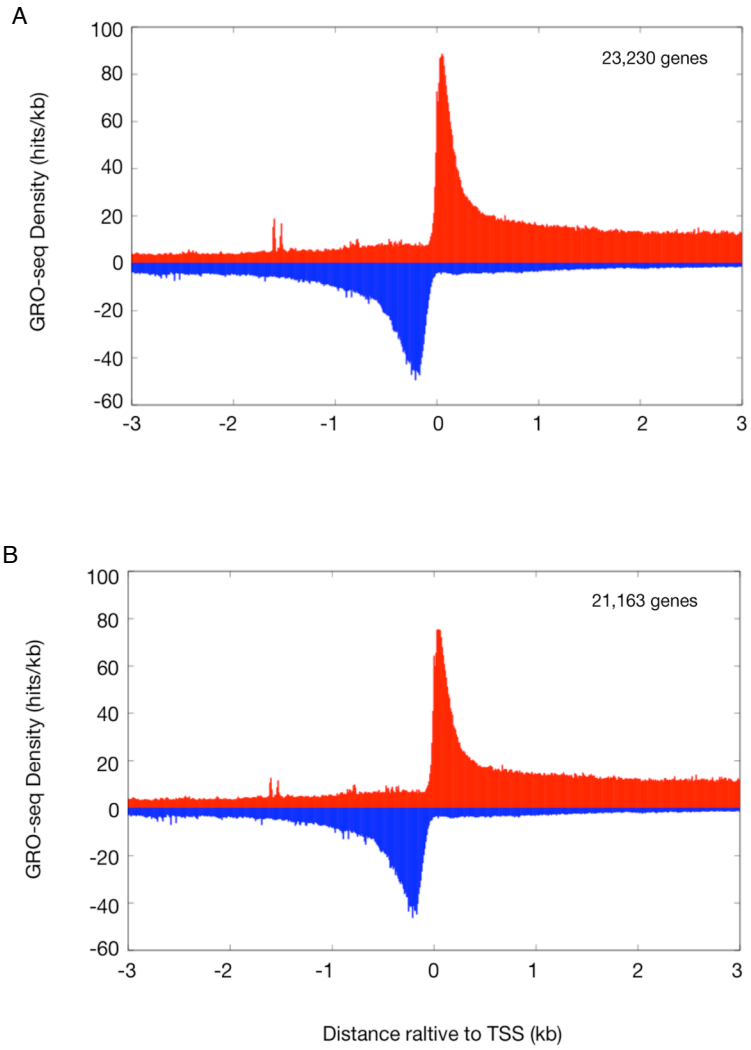


Figure S6. GRO-seq aligned to TSS without bidirectional promoters. (A) Comparison of composite GRO-seq profiles aligned to TSSs of all RefSeq genes, or **(B)** RefSeq genes minus the annotated bidirectional promoters (genes arranged head-to-head and within 1kb).

Figure S7

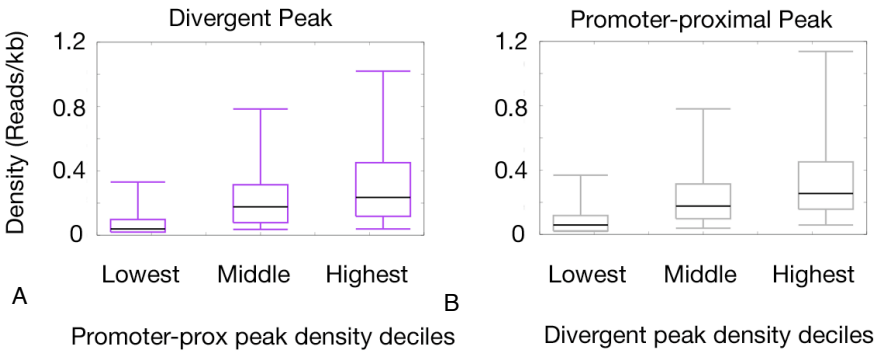


Figure S7. Reciprocal box plots that show the relationship of (**A**) divergent peak height to the lowest, middle, and highest deciles of Promoter-proximal (PP) sense peak heights (violet), and (**B**) Promoter-proximal (PP) sense peak heights with the lowest, middle and highest divergent peaks (DP) heights (gray). All deciles are significantly different from each other: $p < 10^{-9}$ for all comparisons. Box plots are defined as in Figure 3.

Figure S8

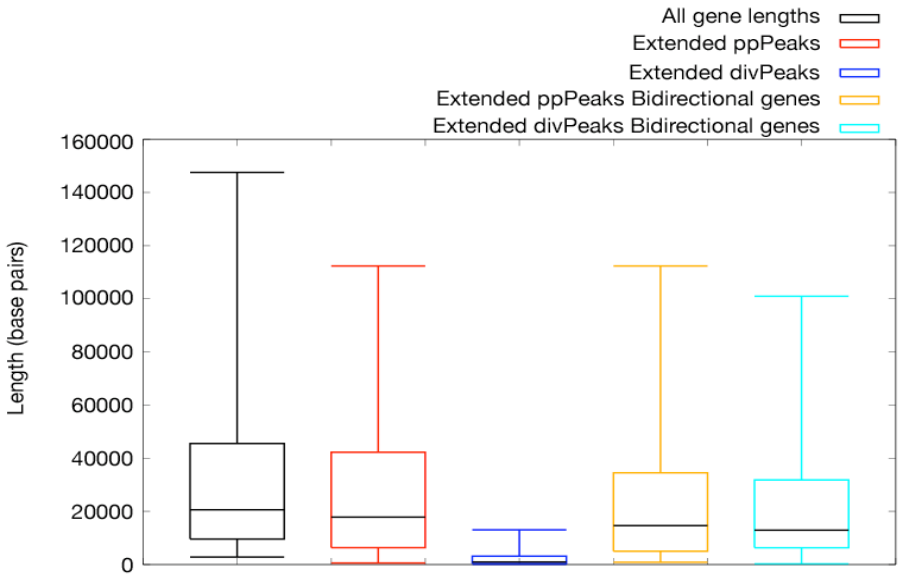


Figure S8. Extending peaks to transcribed regions. In black is the distribution of all RefSeq gene lengths to provide a scale for the other distributions. In red and orange are the transcribed regions extended from promoter proximal peaks on the sense strand of genes while dark blue and cyan are transcribed regions extended from the divergent peaks. The genes used for the red and dark blue data sets do not included pairs of annotated genes oriented head-to-head with less than 1kb between TSSs. The orange and cyan distributions are from just those annotated gene pairs alone.

Table S1

	Active	Paused	Divergent	CpG island
Active	16,882 (100%)	7,032 (99.6%)	13,087 (96.0%)	13,773 (85.2%)
Paused	7,032 (41.7%)	7,057 (100%)	6,614 (48.5%)	6,304 (39.1%)
Divergent	13,087 (77.5%)	6,614 (93.7%)	13,633 (100%)	12,053 (74.8%)
CpG island	13,773 (81.6%)	6,304 (89.3%)	12,053 (88.4%)	16,118 (100%)

Table S1. <u>Pairwise correlations between Gene Acitivity, Pausing, Divergent transcription, and CpG island promoters.</u> Four qualities of individual genes were found to significantly co-occur by pairwise tests. The four qualities were significant levels of gene activity, significant levels of pausing, a significant peak of divergent transcription, and having a CpG island-type promoter. The criteria for gene acitivity, pausing, and divergent transcription are described in the methods. To define whether a given promoter had a CpG island the CpG Islands track was downloaded from the UCSC Genome Browser. If there was an annotated CpG island within 1kb of a given TSS, the gene was classified as having a CpG island-type promoter. The percentages listed in the Table are the fraction of genes from the category on the left that are also in the category on the top.

Supporting figures for SOM text

Figure S9

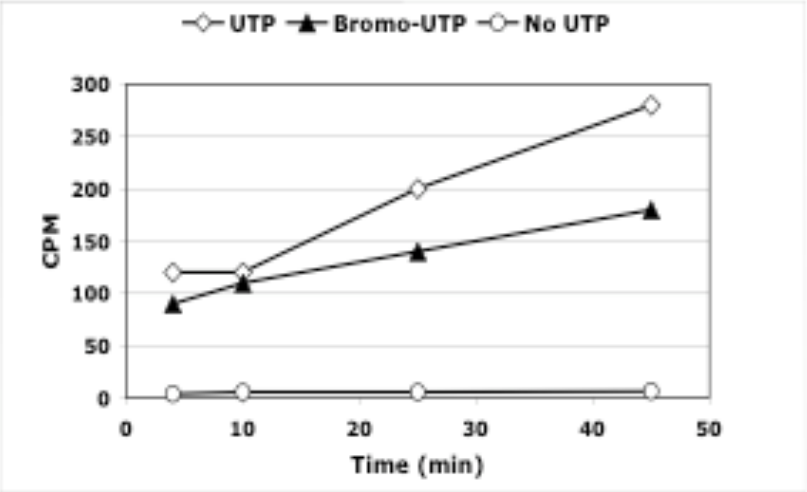


Figure S9. Incorporation of Br-UTP in a nuclear run-on.

Polymerases were run-on in nuclei supplemented with Sarkosyl, ATP, GTP , α-^{32}P-CTP and UTP (open diamonds), Br-UTP (closed triangles), or no UTP (open circles),Separate reactions were setup for each timepoint and the reactions were stopped at 5, 10, 25 or 45 min. The RNAs were isolated, and the radioactivity incorporated was assayed by scintillation counting.

Figure S10

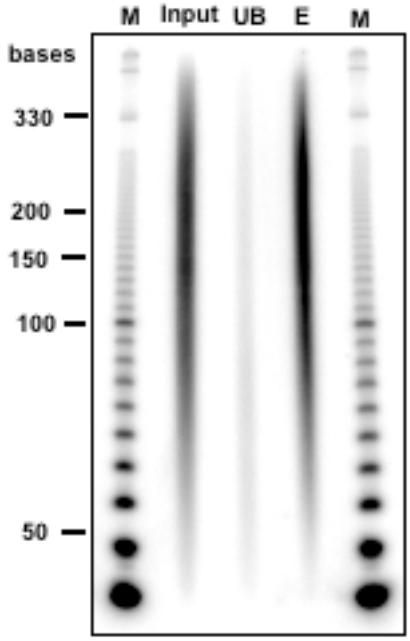


Figure S10. Binding and elution of base-hydrolyzed BrU-RNA to α-BrdU beads. Isolated RNA from a nuclear run-on containing Br-UTP and α-^{32}P-CTP was base hydrolyzed to an average size of 100 bases, and then bound to agarose beads that are conjugated with an antibody specific for α-BrdU. The beads were washed several times and then eluted. Equivalent amount of each fraction were run on an 8% denaturing PAGE gel to assess the efficiency of bead binding.

Figure S11

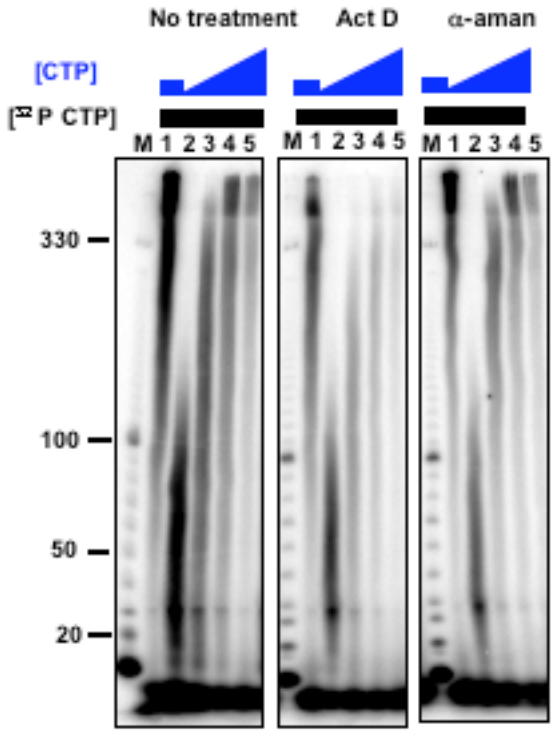


Figure S11. Control of nuclear run-on distance by limiting nucleotide concentration. Nuclei were pre-treated with RNase to reduce the nascent RNA to ~20 nucleotides, washed, and then allowed to run-on in separate reactions containing a α-32P-CTP and cold CTP for a total of 0.65μM (Lane 2), 1μM (lane 3), 5μM (lane 4) or 25μM (lane 5). Non-RNase treated nuclei supplemented with 1 μM total CTP were used as a control (Lane 1). Cells were treated with Act-D and nuclei were treated with α-amanatin.

Figure S12

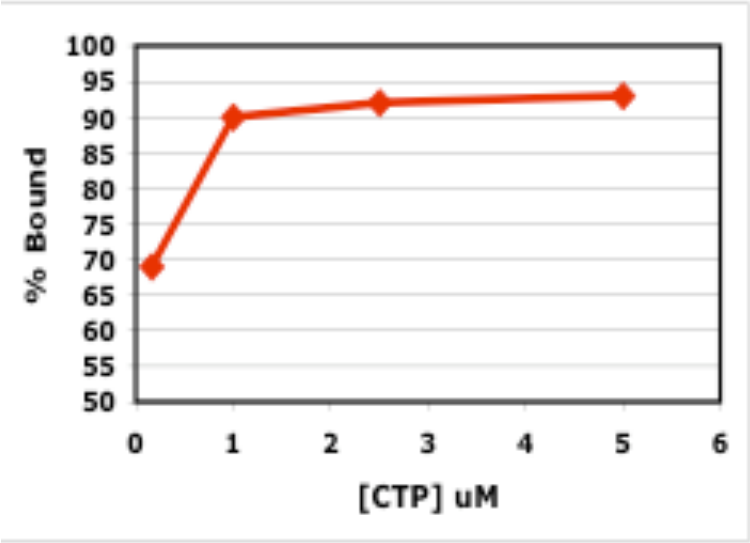


Figure S12. Bead binding efficiency in response to [CTP] titration.

Nuclear run-on were performed as described in figure S11, but without pre-treatment with RNase. Run-on RNAs from each sample were base hydrolyzed and bound to equivalent amounts of beads. The bound and unbound fractions were monitored for radioactivity by scintillation counting. The percent bound (y-axis) was calculated relative to input fractions and is displayed relative to the concentration of CTP in the reaction (x-axis).

Figure S13

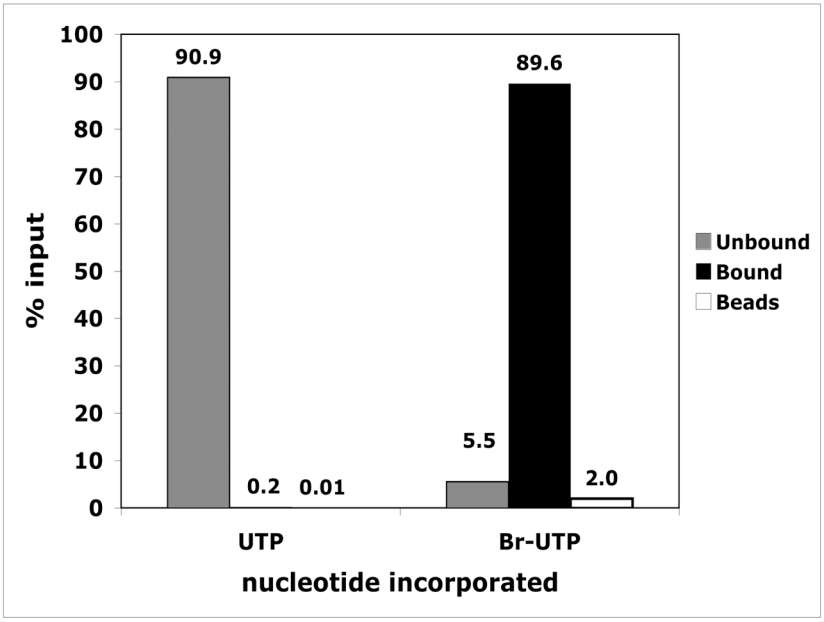


Figure S13. Specificity of α-BrdU beads.

Run-ons were performed in the presence of either UTP or Br-UTP, and handled as described previously. RNAs from each fraction were quantified by scintillation counting.

Figure S14

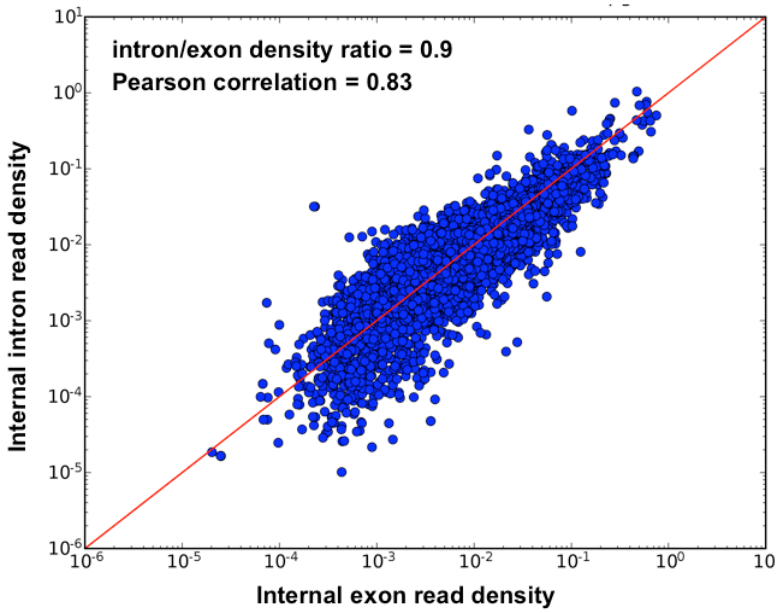


Figure S14. **Comparison of GRO-seq read density in Exon vs, intron.** Scatter plot showing the density of GRO-seq reads within introns (yaxis) vs exons (x-axis) for each RefSeq gene. Axes are in $\log_{10}$ scale. Only internal exons and introns were udes in the analysis to avoid inflation of signal due to promoter-proximal pausing or build up of polymerases that can occur near the 3'-end of genes.

Table S2

Chromosome	Start	Stop	Read count	Mappable Length	Read density (reads/bp)
Chr4	27900001	3500000	1667	6900326	$2.42*10^{-4}$
Chr2	144700001	148400000	927	3425237	$2.71*10^{-4}$
Chr1	79311112	81964443	170	2407985	$7.06*10^{-5}$
Chr1	185879619	188333419	149	2234374	$6.67*10^{-5}$
Chr2	139254282	140705466	56	1328410	$4.22*10^{-5}$
Chr2	56466815	57988288	67	1344339	$4.98*10^{-5}$
Chr2	33700268	36420000	125	2384667	$5.24*10^{-5}$
Chr2	139254283	140705464	56	1328410	$4.21*10^{-5}$
Chr2	155421262	156585290	42	1057143	$3.97*10^{-5}$
Chr2	192775891	196025184	147	2902767	$5.06*10^{-5}$
Chr2	222155254	222762851	21	550326	$3.82*10^{-5}$
Chr4	44473369	45702544	43	1099820	$3.91*10^{-5}$
Chr4	104870422	105599015	21	640337	$3.28*10^{-5}$
Chr4	116264481	118214158	71	1772986	$4.00*10^{-5}$
Chr4	135352353	137135534	61	1605865	$3.80*10^{-5}$
Chr5	104744250	106704250	65	1785211	$3.64*10^{-5}$

Table S2: Background calculation in gene deserts. The indicated large intergenic spaces were analyzed for the number of GRO-seq reads on either strand and the number of mappable bases.

Figure S15

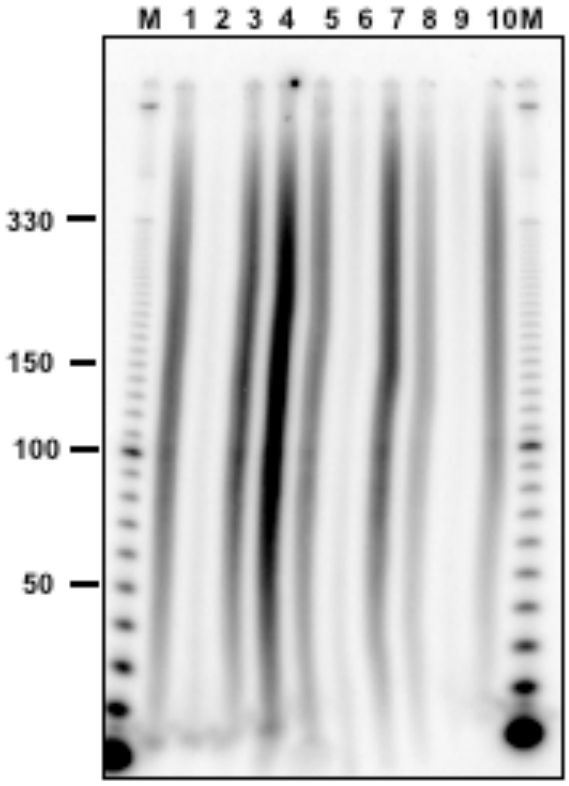


Figure S15. Denaturing PAGE analysis of fractions from GRO-seq library preparation. Lanes: 1) Input, 2) Unbound-1, 3) Elution-1, 4) After TAP-PNK treatment, 5) 5' adapter ligation, 6) Ubound 2, 7) Elution 2, 8) 3' adapter ligation, 9) Unbound 3, 10) Elution 3.

\

Figure S16

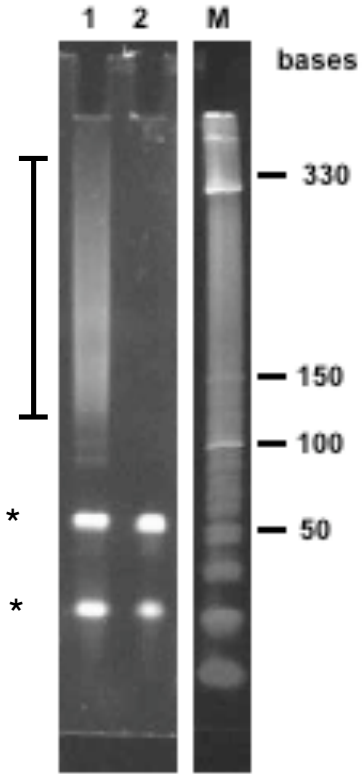


Figure S16. Example of amplified NRO-library cDNA. After the third elution the library was reverse transcribed amplified by 15 cycles of PCR, and then run on an 8% PAGE gel for purification away from the primers (*) Lane 1 cDNA library, Lane 2) No template control. Bracket indicates region cut from gel.

Figure S17

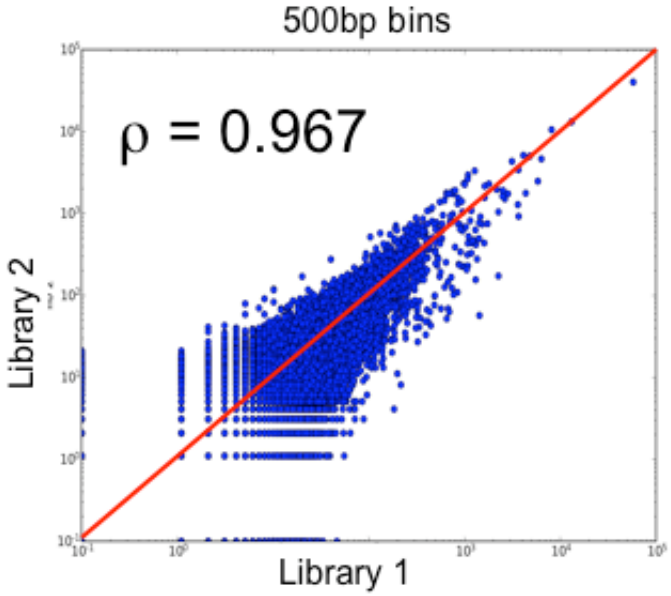


Figure S17: Correlation of GRO-seq biological replicates. GRO-seq transcript reads were mapped to the genome and unique reads were binned in 500bp windows. Of the 6,160,849 windows, 3,458,076 windows had no reads in each replicate. The replicates show a correlation coefficient of 0.967 (Spearmann correlation).

Figure S18

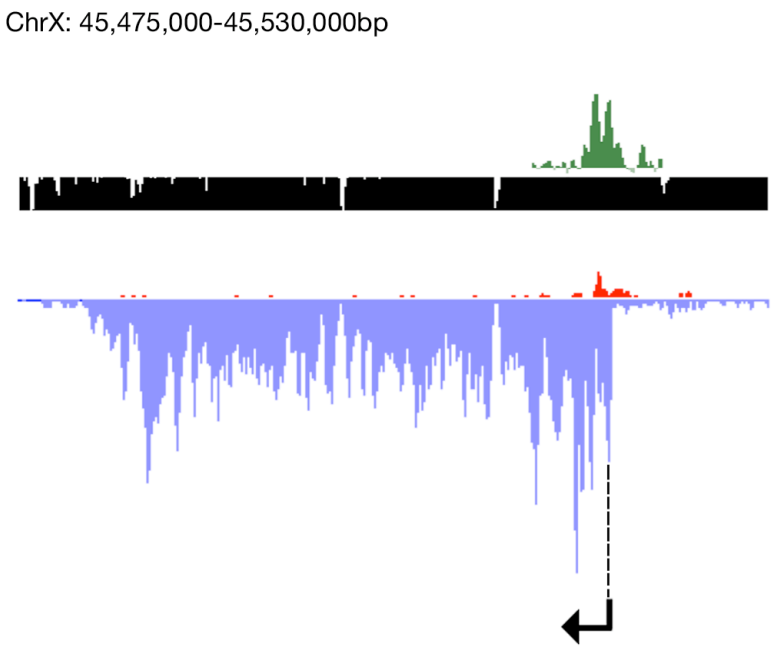


Figure S18. Novel promoter by ChIP and GRO-seq. A novel transcription unit on chrX: 45,475,000- 45,530,000bp is shown that is not annotated by any of the major databases or gene prediction tools. The promoter was identified as putative by Pol II ChIP shown in green.

Figure S19

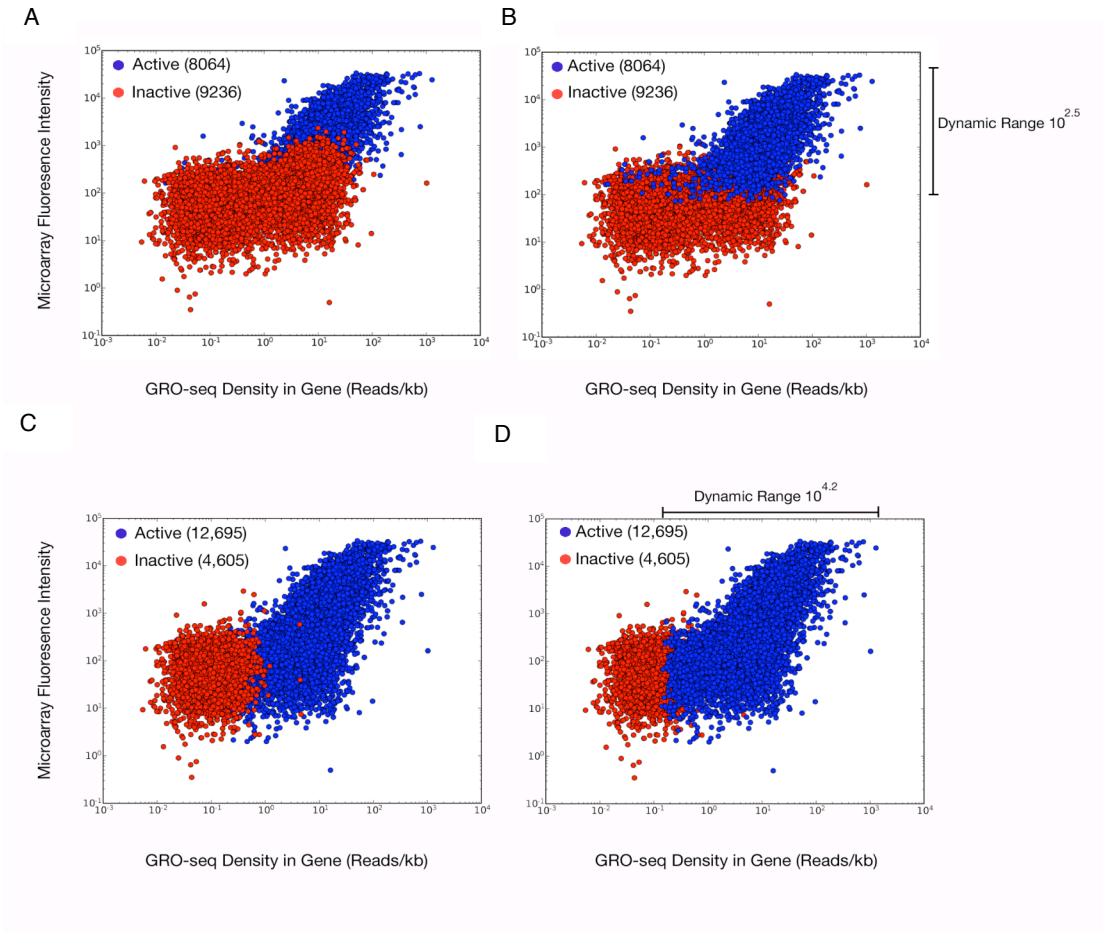


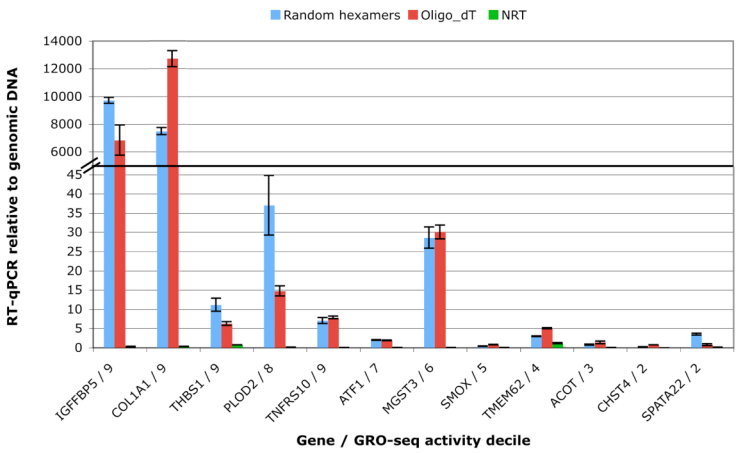
Figure S19. GRO-seq activity versus expression microarray. Scatter plots of gene expression levels by microarray versus GRO-seq. Inactive genes are colored in red and active genes are colored in blue. Only the 17,300 genes that were unambiguous by both methods are shown in the plots. The range for which genes can be called significantly active is shown to the right (**B**) or top (**D**) by microarray hybridizations or GRO-seq, respectively. Gene activity significance is determined by microarray in **A** and **B**, and by GRO-seq in **C** and **D**. Inactive genes are plotted on top of active genes in **A** and **C** while the order is reversed in **B** and **D**.

Table S3. GRO-seq vs. microarray gene activity calls

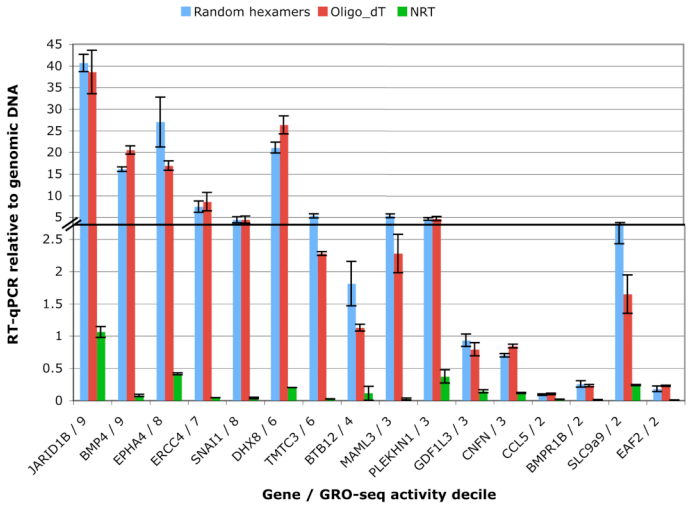
GRO-seq	Expression Microarray				
	Present	Absent	Ambiguous	Not on array	Total
Active	7983	4712	4163	24	16882
Inactive	81	4524	1101	6	5712
Less than 3kb	374	1159	683	3	2219
Total	8438	10395	5947	33	24813

Figure S20

A



B



C

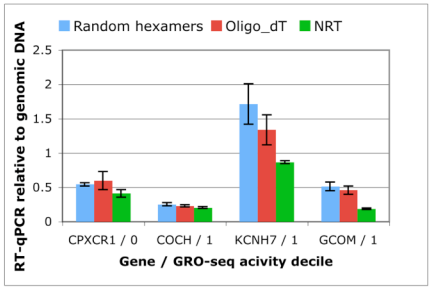


Figure S20. RT-qPCR validation of GRO-seq levels. Genes that were active by microarray and GROseq (A), inactive by microarray – active by GRO-seq (B), and active by microarray but not by GRO-seq (C) were analyzed by RT-qPCR. Reverse transcription was performed with random primers (blue), or oligo-dT (red), and compared to a known amount of genomic DNA. No reverse transcription reactions (green) . Error bars represent standard error of the mean, n=3.

Figure S21

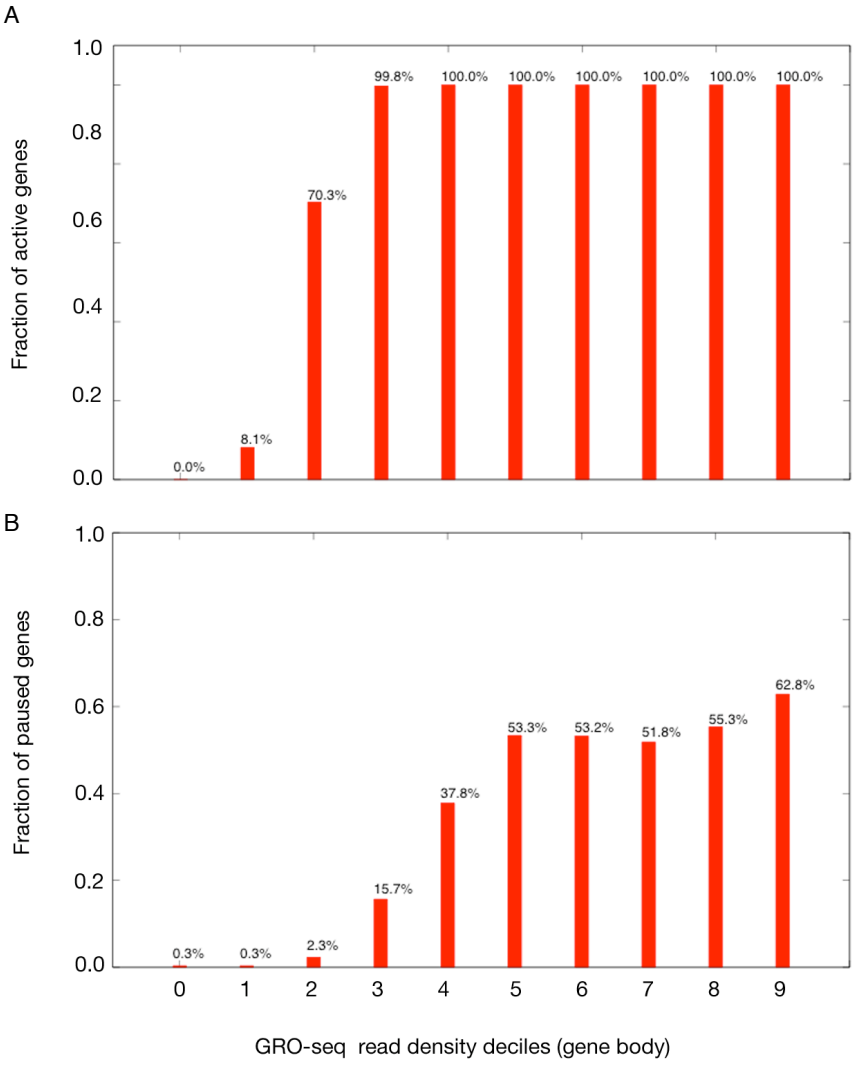


Figure S21. Fraction of paused genes and active genes by gene activity decile The percentage of significantly active (**A**) and significantly paused (**B**) genes in each decile of gene activity. See methods for calculation of gene activity levels and the criteria for significant pausing and significant gene acitivity.

Figure S22

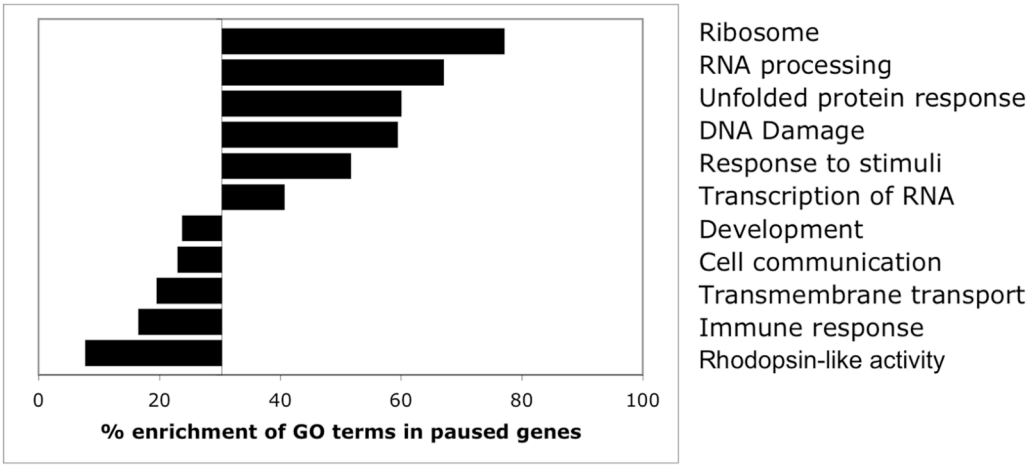


Figure S22. Gene onotology of paused genes. Bar plot show the summary of enriched and de-enriched gene ontology (GO) terms of significantly paused genes. The Y-axis is set to 28.3%. GO terms that are enriched in paused genes are to the right of the axis, and GO that are de-enriched are to the left. All terms are significant ($P < 10^{-10}$).

Figure S23

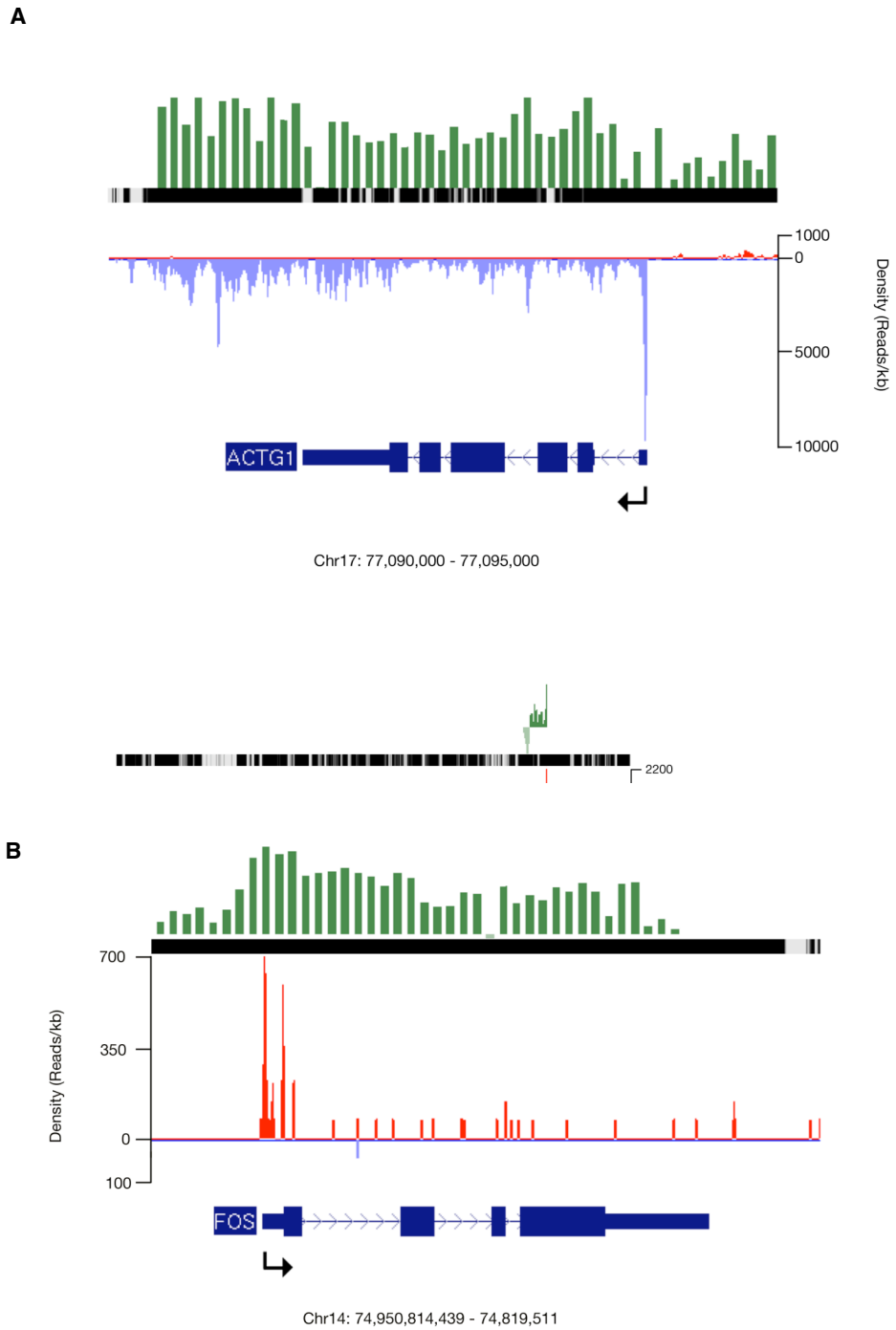

A
1000
0
5000
10000
Density (Reads/kb)
ACTG1
Chr17: 77,090,000 - 77,095,000
2200
B
700
350
0
100
Density (Reads/kb)
FOS
Chr14: 74,950,814,439 - 74,819,511

Figure S23 (continued)

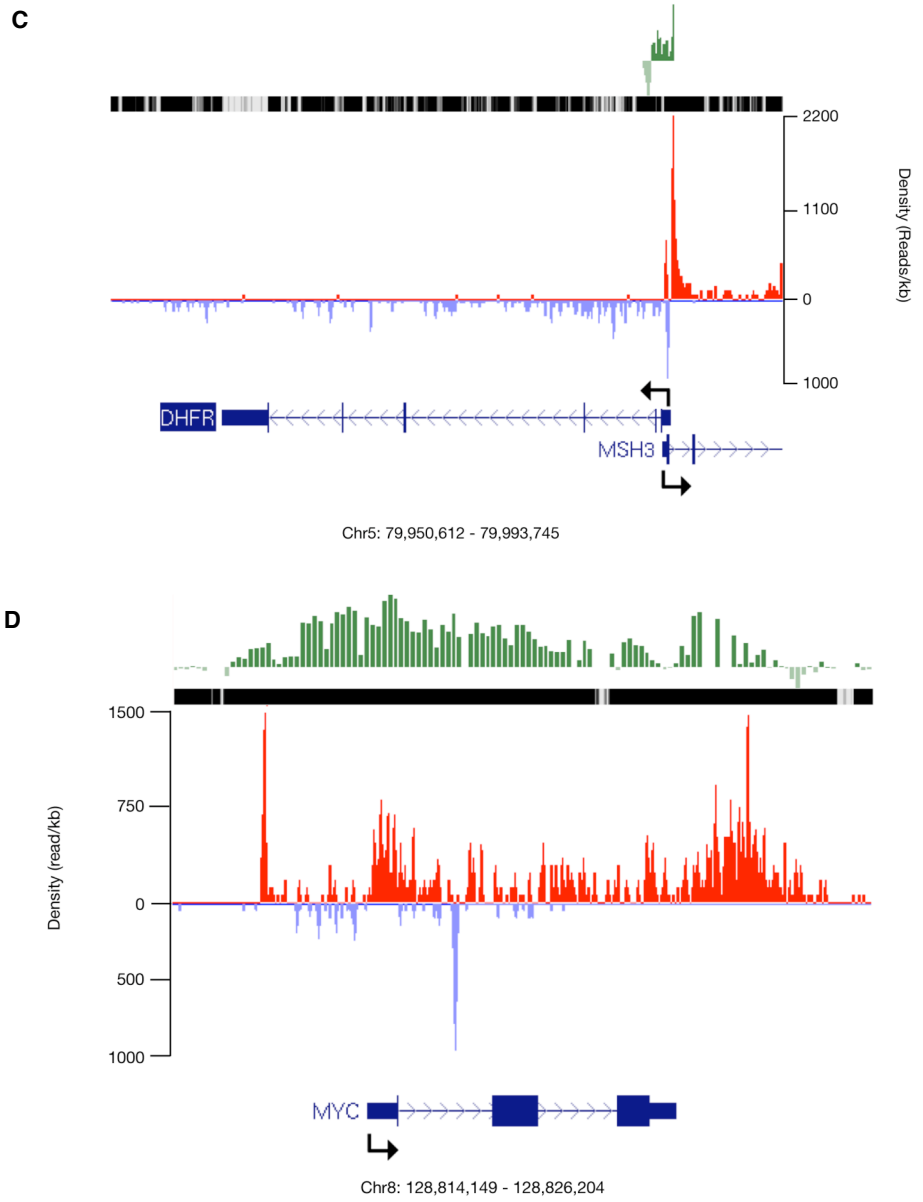


Figure S23. GRO-seq profiles for known paused genes. Snaphots from the UCSC genome browser showing the regions around genes previous characterized as paused. Gene names, pausing indexes, and associated P values are as follows: (A) ACTG1, 6.3, $8X10^{-30}$; (B) FOS, 43, 1.7 $X10^{-4}$; (C) DHFR, 25, 7.8 $X10^{-4}$; and (D)MYC, 5.7, 3.2 $X10^{-3}$. Pol II ChIP results are shown in green and the start site and direction of transcription of the gene is shown by the arrow (black). Y-axis (Reads/kb) is shown to delineate the scale between the images.

Figure S24

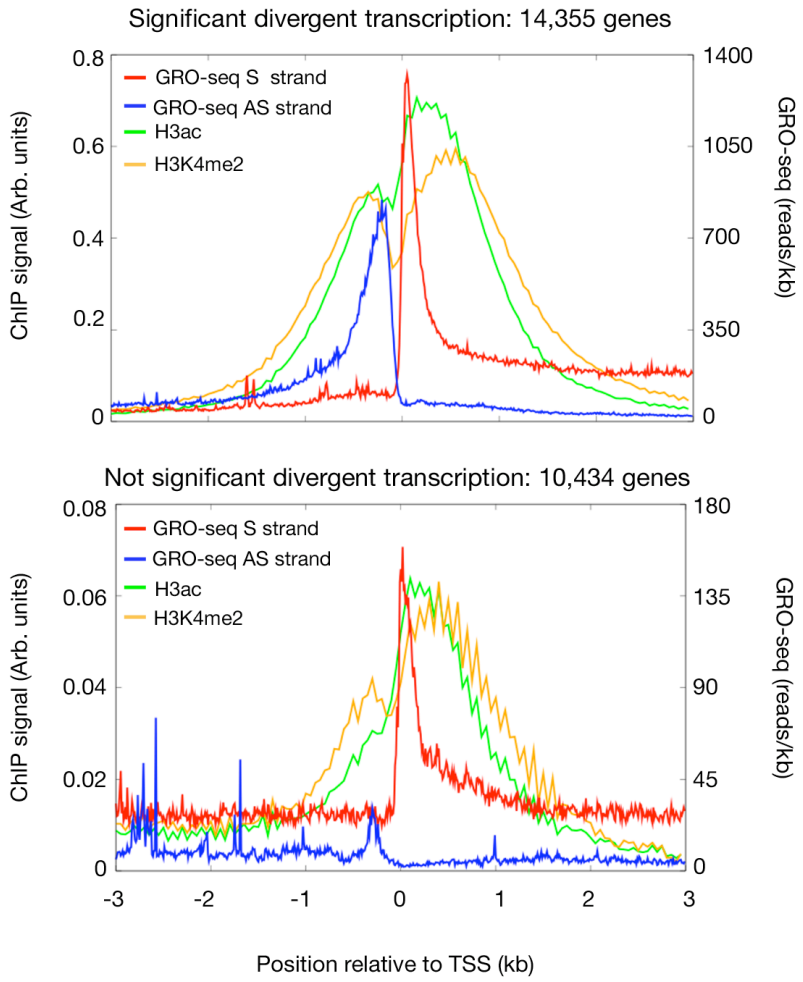


Figure S24. Histone modifications at promoters with or without significant divergent transcription. Genes were separated based on whether they had significant divergent transcription ($P < 0.001$) (A) or not ($P > 0.001$)(B). The profiles for histone modifications H3ac (green) and H3K4me2 (orange) were then plotted in arbitrary units against GRO-seq read density (reads/kb) for the plus strand (red) and minus strand (blue) reads. X-axis represents the distance (kb) relative to the TSS, which is set to zero.

Figure S25

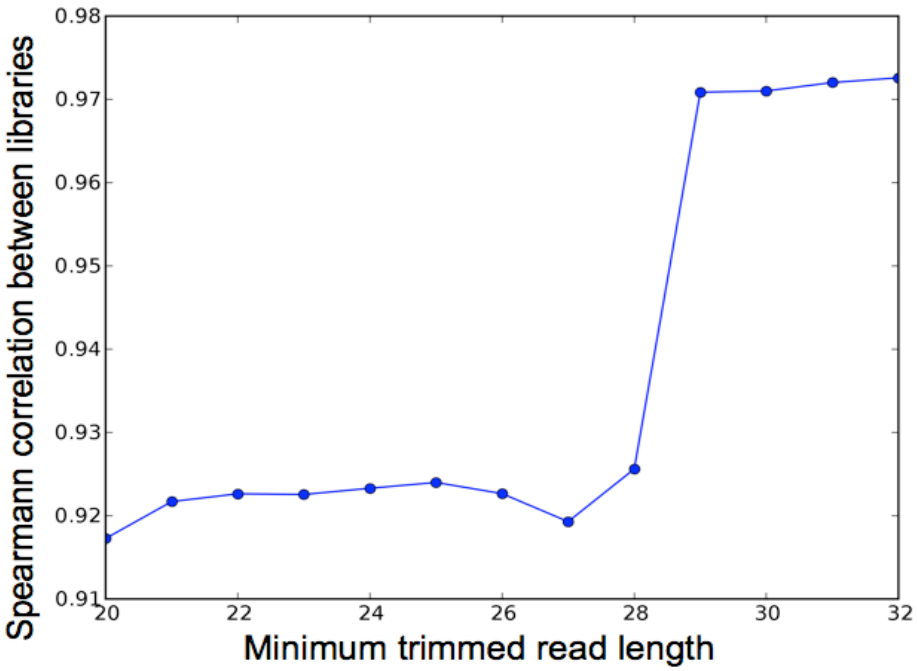


Figure S25. Interlibrary correlation versus read trimming. Reads that did not align uniquely were trimmed by one base at the 3' end and realigned to the genome in an iterative process. The Spearmann correlation between the two libraries is shown as a function of the minimum length of the reads included in the libraries. Because the correlation drops when 28mers are included, all analyses were performed with only 29mers and longer.

Figure S26

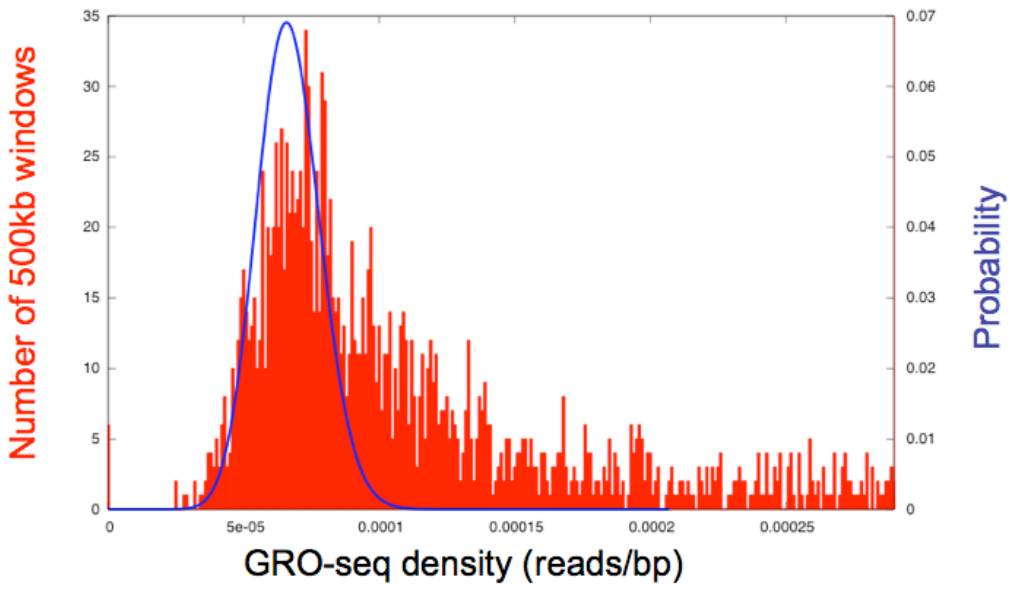


Figure S26. Background calculation by low-density windows. After aligning reads to genome, the density of GRO-seq reads was assessed in 500kb windows. Shown in red is a histogram of the lowest density windows and in blue is a Poisson distribution with a mean given by placing 3% of all GRO-seq reads at random throughout the mappable portion of the genome.