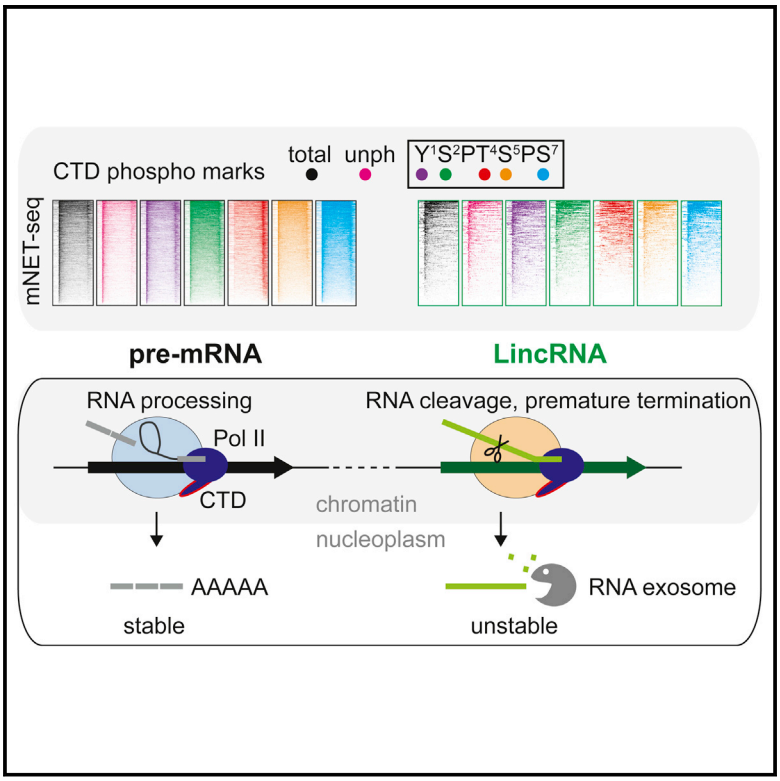
Article

Molecular Cell

Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs

Graphical Abstract



Highlights

- lincRNAs and pre-mRNAs are transcribed by different Pol II phospho-CTD isoforms
- lincRNAs are rarely spliced and mainly non-polyadenylated
- lincRNAs are stabilized in the nucleoplasm following exosome inactivation
- lincRNAs are co-transcriptionally cleaved

Authors

Margarita Schlackow, Takayuki Nojima, Tomas Gomes, Ashish Dhir, Maria Carmo-Fonseca, Nick J. Proudfoot

Correspondence

taka.nojima@path.ox.ac.uk (T.N.), nicholas.proudfoot@path.ox.ac.uk (N.J.P.)

In Brief

Schlackow and Nojima et al. show that mammalian pre-mRNAs and long intergenic noncoding (linc) RNAs employ radically different transcription and RNA-processing strategies. Pre-mRNAs are transcribed by defined RNA polymerase (Pol) II isoforms reflecting co-transcriptional splicing and polyadenylation. Instead, lincRNAs are mainly transcribed by deregulated Pol II and simultaneously degraded.

Accession Numbers

GSE81662

Schlackow et al., 2017, Molecular Cell 65, 25–38
January 5, 2017 © 2017 The Authors. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.molcel.2016.11.029

Molecular Cell
Article

Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs

Margarita Schlackow,[1,3] Takayuki Nojima,[1,3,*] Tomas Gomes,[2] Ashish Dhir,[1] Maria Carmo-Fonseca,[2] and Nick J. Proudfoot[1,4,*]
[1]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK
[2]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal
[3]Co-first author
[4]Lead Contact
*Correspondence: taka.nojima@path.ox.ac.uk (T.N.), nicholas.proudfoot@path.ox.ac.uk (N.J.P.)
http://dx.doi.org/10.1016/j.molcel.2016.11.029

SUMMARY

Numerous long intervening noncoding RNAs (lincRNAs) are generated from the mammalian genome by RNA polymerase II (Pol II) transcription. Although multiple functions have been ascribed to lincRNAs, their synthesis and turnover remain poorly characterized. Here, we define systematic differences in transcription and RNA processing between protein-coding and lincRNA genes in human HeLa cells. This is based on a range of nascent transcriptomic approaches applied to different nuclear fractions, including mammalian native elongating transcript sequencing (mNET-seq). Notably, mNET-seq patterns specific for different Pol II CTD phosphorylation states reveal weak co-transcriptional splicing and poly(A) signal-independent Pol II termination of lincRNAs as compared to pre-mRNAs. In addition, lincRNAs are mostly restricted to chromatin, since they are rapidly degraded by the RNA exosome. We also show that a lincRNA-specific co-transcriptional RNA cleavage mechanism acts to induce premature termination. In effect, functional lincRNAs must escape from this targeted nuclear surveillance process.

INTRODUCTION

Approximately 20,000 protein-coding genes are transcribed by RNA polymerase II (Pol II) from the human genome. These transcripts are modified by pre-mRNA processing events, such as 5′ capping, pre-mRNA splicing, 3′ end cleavage, and polyadenylation during Pol II transcription (Moore and Proudfoot, 2009). Pre-mRNA processing as well as generating translatable mature mRNA also acts to enhance mRNA stability and cytoplasmic export. Even though protein-coding genes occupy a limited proportion of the mammalian genome, transcription analyses reveal the widespread occurrence of long noncoding RNAs (lncRNAs), which lack significant protein-coding capacity (St Laurent et al., 2015). In general, lncRNA can be subdivided into different classes based on their positional relationship to protein-coding transcripts. Thus, Pol II promoters as well as generating pre-mRNAs also form promoter upstream transcripts in antisense orientation, called CUTs in *S. cerevisiae* or PROMPTs in mammals (Jensen et al., 2013). Additionally, in higher eukaryotes, multiple Pol II enhancers exist upstream or within protein-coding genes that act to guide Pol II to promoters by *trans* interactions. These numerous enhancers also generate bidirectional transcripts called eRNAs (Kim et al., 2010; Kowalczyk et al., 2012). Finally, some lncRNA initiate independently of protein-coding gene promoters and enhancers to generate separate transcription units (TUs) called long intervening noncoding RNA (Ulitsky and Bartel, 2013). It is the focus of this study to better understand how long intervening noncoding RNAs (lincRNAs) are synthesized and processed and how this may differ from protein-coding genes.

Whereas PROMPTs and enhancer RNAs (eRNAs) likely form as a consequence of Pol II accumulation at transcription initiation sites, it is more plausible that lincRNAs, with their independently defined transcription units, have specific biological significance. However, their low sequence conservation and often very low steady-state levels imply that many of these ephemeral transcripts reflect transcriptional noise (Struhl, 2007). One often proposed argument for lincRNA functionality is that they are at least partially capped, spliced, and polyadenylated, based on high-throughput cDNA analysis. This has led to the view that lincRNAs are mRNA like (Cabili et al., 2011; Derrien et al., 2012; Garber et al., 2011; Grabherr et al., 2011). Although the function of most lincRNAs remains unknown, some, such as XIST, HOTAIR, NORAD, and FENDRR, have established biological roles (Grote et al., 2013; Lee et al., 2016; Mattick, 2009; St Laurent et al., 2015; Wang and Chang, 2011).

Defining the TUs of lincRNAs is a challenging problem of sequence annotation. Often transcription start sites are inferred from 5′ end cap selection methods, such as CAGE (Kodzius et al., 2006) or Cap-seq (Gu et al., 2012). However, some degree of recapping has been shown to occur on cytoplasmic RNA (Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009), so that some capped lincRNAs may derive from RNA degradation intermediates. Also a recent description of chromatin-associated lncRNAs included many cases of low-level read-through transcription from upstream protein-coding gene TUs

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

(Werner and Ruthenburg, 2015). The realization that cellular stress can increase readthrough transcription for protein-coding genes (Vilborg et al., 2015) may exacerbate such problems of mis-annotation. For lincRNAs, 3′ end mapping by poly(A) selection methods are often employed, such as the 3P-seq method (Ulitsky and Bartel, 2013). Such approaches may not be appropriate for lincRNAs, as these transcripts are often unpolyadenylated, such as those harboring pre-microRNA (miRNA) sequences (Dhir et al., 2015). Also, lincRNA 3′ ends may be subject to rapid 3′ end degradation by the nuclear exosome (Pefanis et al., 2015; Lubas et al., 2015). Finally, previous annotations of lincRNAs have focused on spliced transcripts as a way to increase specificity. However, we show that lincRNAs are generally only weakly spliced and so may be excluded from such analysis (Cabili et al., 2011). Indeed, transcription regulation of lincRNA genes remains poorly characterized due to a lack of detailed information on how they are synthesized and processed.

Recently, we have developed mammalian native elongating transcript sequencing (mNET-seq) to precisely define nascent transcription across the human genome (Nojima et al., 2015). In particular, we have focused on the C-terminal domain (CTD) of the largest subunit of Pol II, which has a 52 times repeated heptad domain ($Y_1S_2P_3T_4S_5P_6S_7$) that is differentially phosphorylated during Pol II transcription (Heidemann et al., 2013). mNET-seq allows the determination of which CTD phosphorylation marks correlate with different stages of TU synthesis and processing. Here, we obtained mNET-seq profiles using a full range of Pol II CTD antibodies to compare the expression profiles between protein-coding and lincRNA TUs. We show that most lincRNAs, unlike protein-coding genes, are poorly co-transcriptionally spliced, and Pol II pauses inefficiently at their promoters. Furthermore, the CTD T4P mark that correlates with protein-coding gene termination is distributed more evenly across the gene body of lincRNAs. This implies that lincRNA termination occurs at multiple positions within the TU. Also, mRNA 3′ end processing endonuclease CPSF73 shows little effect on lincRNA 3′ end formation. These observations in general indicate that lincRNA and pre-mRNA processing differ both quantitatively and qualitatively.

RESULTS

Widespread lincRNAs have been defined in several comprehensive studies (Ulitsky and Bartel, 2013). Although combined transcription profiles from multiple cell types show that most human intergenic sequences (regions between annotated protein coding genes) are transcribed, within one specific cell type, lincRNA expression is more restricted. We have analyzed lincRNA expression in human HeLa cells where about 35% of the non-repetitive genome is transcriptionally active (Djebali et al., 2012). Of roughly 50,000 annotated Tus, about 20,000 are protein coding. To define the gene units of expressed lincRNAs for our analyses, we employed ENSEMBL and NONCODE databases as reference gene annotation (Flicek et al., 2014; Xie et al., 2014). We then cross-checked these annotations by visual identification of their transcription start and end sites (TSSs and TESs) using our own HeLa cell RNA sequencing (RNA-seq) data from chromatin and nucleoplasm fractions (Nojima et al., 2015). We excluded low-level expressed lincRNAs as well as lincRNAs that were close to other TUs either at their TSSs or TESs, including those annotated as an antisense biotype in the ENSEMBL annotation. This generated a list of 285 lincRNAs that are expressed at sufficiently high levels separate from other adjacent transcription units to allow their independent analysis (Tables S1 and S2). In the later stages of this study, we included the antisense biotype to effectively add 500 additional lincRNAs (antisense RNAs).

Pol II CTD Phosphorylation Profiles Differ between Pre-mRNAs and lincRNAs

Pol II CTD phosphorylation states are well established to match different transcriptional stages: Ser5P (S5P) with early elongation, 5′ end capping, and active splicing and Ser2P (S2P) with later elongation and 3′ end processing (Heidemann et al., 2013; Hsin and Manley, 2012). mNET-seq methodology sequences genome-wide nascent RNA at single nucleotide resolution (Nojima et al., 2015) by isolating RNA from immunoprecipitated (IP) Pol II. We previously employed Pol II antibodies against total, S2P, S5P, and unphosphorylated (unph) CTD to isolate specific nascent RNA fractions (Nojima et al., 2015). Here, we have added three additional phospho-CTD-specific antibodies, Y1P, T4P, and S7P, allowing a closer comparison between protein-coding and lincRNA genes (Figure 1A).

Meta-analysis of protein coding as compared to lincRNA genes reveals significant differences in mNET-seq profiles. Both heatmaps and metagene profiles (Figures 1B, 1C, and S1A) are shown. In particular, the unph followed by Y1P profiles show highest promoter peaks for protein-coding genes. In contrast, lincRNA genes show less pronounced unph and Y1P TSS peaks with a generally more even distribution of mNET-seq reads across their gene bodies. A wider set of lincRNA TUs that are partly overlapping with other TUs (ENSEMBL antisense biotype) looks closely similar to the separate lincRNA TU class (Figure S1A). We also included analysis of TSS-associated eRNAs (both strands), which derive from unph Pol II with some from Y1P Pol II, but very little with other phospho-CTD isoforms (Figure S1A, bottom panel). We next compared the promoter escape indexes between protein coding and lincRNA genes, taken as the ratio of reads in TSS regions versus gene body. Lower Pol II pausing was observed over the TSS regions of lincRNA than protein-coding genes, as shown in data replicates (Figure S1B, top; $p < 1e{-}5$ for unph [both replicates] and $p < 1e{-}6$ for Y1P [all three replicates]).

A notable feature of the TES region in protein-coding genes is the high T4P signal, which is indicative of Pol II termination (Figure 1B). This observation is consistent with previous chromatin immunoprecipitation sequencing (ChIP-seq) results (Hintermair et al., 2012). In contrast, T4P signal over lincRNA genes is more evenly distributed across the whole TU, with less TES-associated accumulation (Figures 1C and S1A), suggesting that Pol II termination occurs at multiple positions across lincRNA TUs. These replicated TES effects were quantitated by their termination indices, which are taken as the ratio of reads in termination regions versus gene body (Figure S1B, bottom). We observe a lower T4P termination index in lincRNA compared

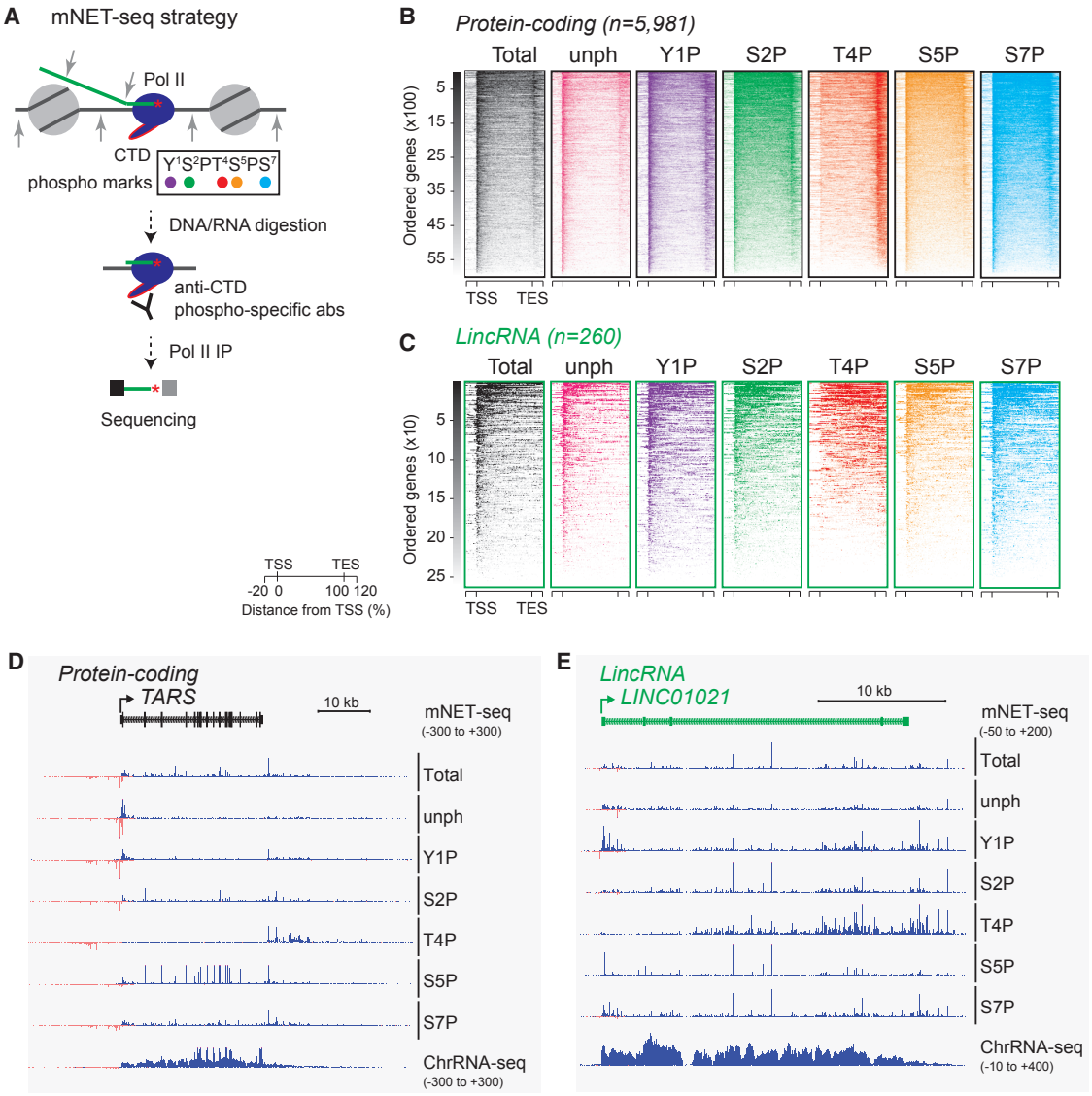


Figure 1. Differential mNET-Seq Profiles for Protein Coding and lincRNA genes

(A) mNET-seq strategy with each Pol II phospho-CTD modification color coded.

(B and C) Color-coded heatmaps showing phospho-CTD profiles across individual (B) protein coding TUs and (C) lincRNA TUs ordered based on their transcription levels. Profiles are aligned to TSS and TES as indicated. Genes >1,000 nt (excluding some smaller protein coding and lincRNA genes) were divided into 100 bins.

(D and E) (D) mNET-seq profiles across *TARS* (black for protein-coding gene) and (E) *LINC01021* (green for lincRNA gene) using seven different Pol II antibodies as indicated. Gene maps show exons filled in and introns hatched. A chromatin-seq profile is run below the mNET-seq profiles. Blue reads are sense and red reads antisense transcripts. Reads per 10^8 mapped reads are indicated in brackets.

See also Figure S1.

to coding genes ($p < 1e{-}10$; all three replicates). The metagene analysis is consistent with individual gene profiles of mNET-seq for the protein-coding gene *TARS* and a specific lincRNA gene (Figures 1D and 1E). Overall, mNET-seq reveals significant differences in Pol II CTD phosphorylation between protein-coding and lincRNA genes.

lincRNAs Are Inefficiently Spliced

We have previously identified a characteristic mNET-seq pattern associated with co-transcriptional splicing. In particular, a prominent splicing intermediate derived from RNA cleavage at 5′ splice sites (5′ss) is evident in mNET-seq/S5P profiles of protein-coding genes (Nojima et al., 2015), as seen for the multi-intronic protein-coding gene *TARS* (Figure 1D). These 5′ss peaks are indicative of co-transcriptional splicing, where upstream exons are tethered to Pol II S5P CTD prior to splicing with the downstream exon to complete the splicing reaction. mNET-seq/S5P also detects several peaks on the lincRNA gene *LINC01021*. However, these were not S5P CTD specific, showing similar patterns for S7P and S2P analysis, nor were they exon specific, appearing to

derive from intronic regions (Figure 1E). We next extended our analysis of specific lincRNAs using splicing specific mNET-seq/S5P profiles and tested their sensitivity to pretreatment of the HeLa cells with the chemical inhibitor Pla-B. This blocks splicing by direct binding to the SF3B complex (Kotake et al., 2007). As previously reported (Nojima et al., 2015), Pla-B erased most of the S5P CTD-specific 5′ss peaks on protein-coding genes as shown for *PTCD3* (Figure 2A). This confirms that these peaks derive from an active splicing process. Notably, a few *PTCD3* intronic peaks were either unaffected or enhanced by Pla-B treatment. These may reflect the maturation of small RNAs from intronic locations, such as *SNORD94*. In contrast, Pla-B treatment had a more limited effect on S5P peaks seen across various lincRNA genes (Figures 2B and S2A). Indeed, only two Pla-B-sensitive splicing events were detectible for these specific lincRNAs: 5′ss of *LINC00472* intron 3 and *LINC00263* intron 1.

To establish generality for lower co-transcriptional splicing on lincRNAs, we obtained mNET-seq/S5P meta-analysis profiles across the exon-intron boundaries of about 70,000 annotated introns for protein coding versus 1,000 for lincRNA genes with or without Pla-B treatment. Both average signals and heatmaps (Figures 2C and 2D) of the whole dataset show Pla-B-sensitive 5′ss signals occur less frequently for lincRNA than protein-coding genes. Quantitation of these data in all biological replicas indicates that 55%–70% of protein-coding introns give 5′ss peaks. Possibly those that lack detectible peaks reflect unspliced exons due to alternative splicing events or retained introns (Boutz et al., 2015). In contrast, only 20%–30% of lincRNA exons gave 5′ss peaks, reflecting lower levels of co-transcriptional splicing (Figure 2D, bottom).

The above data focus on the levels of co-transcriptional splicing based on 5′ss mNET-seq/S5P signals and clearly indicates reduced lincRNA co-transcriptional splicing. To directly measure splicing efficiency, we prepared duplicate HeLa cell transcript libraries from either pA+ or pA− nuclear RNA. pA+ reads across the specific protein-coding gene *WDR13* were exon restricted, indicative of efficient co-transcriptional spicing with little signal detected in the pA− NpRNA-seq profile. In contrast, for the lincRNA *TUG-1* pA+ profile, significant levels of intron reads were detected over its annotated intron regions, even though some splicing is evident. Furthermore, the pA− profile revealed a higher level of intron signal (Figure 2E). We performed quantitative analysis of splicing efficiency between protein coding and lincRNA transcripts. Comparison of splicing events between these two transcript classes for pA+, pA−, and total nucleoplasmic RNA showed a consistently lower splicing for lincRNAs in duplicate experiments (Figure S2B). We finally computed the splicing index of protein coding versus lincRNA by comparing the ratios of spliced exon-exon to unspliced intron-exon reads across active 3′ss in NpRNA-seq, either pA+, pA−, or total (Figures 2F and S2C). This quantitation reveals that lincRNA are inefficiently spliced as compared to protein-coding genes. Note that the duplicated pA+ and pA− NpRNA-seq analyses were closely consistent (Figure S2D).

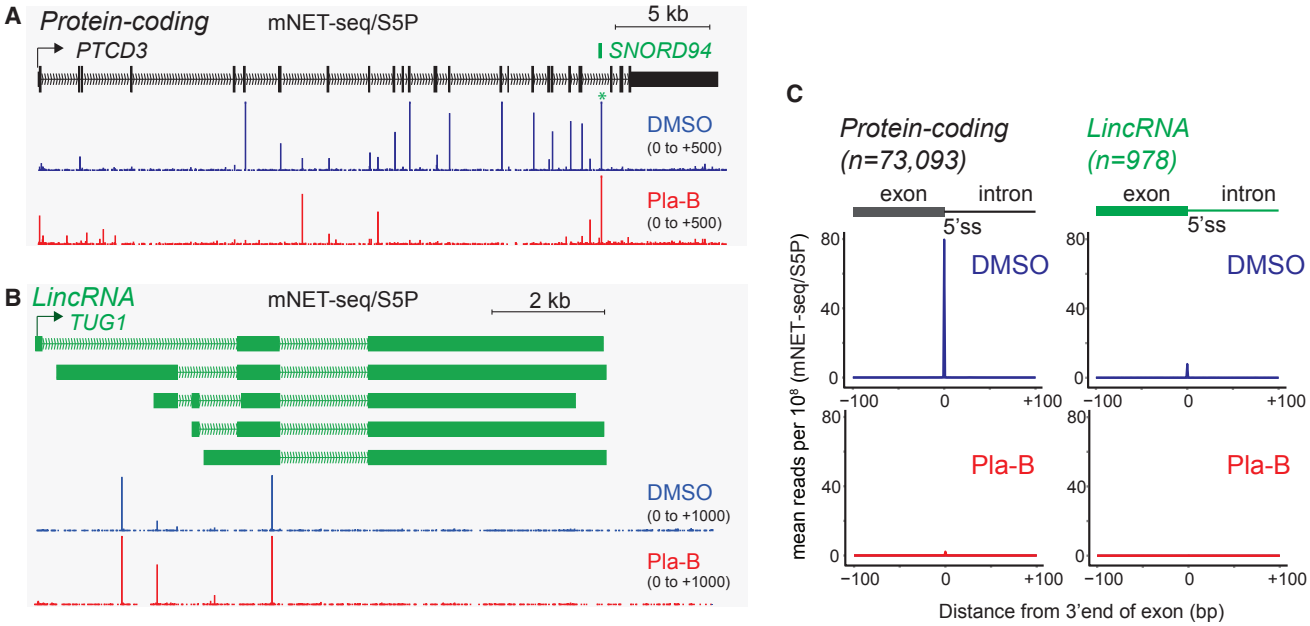
lincRNAs Are Inefficiently Polyadenylated

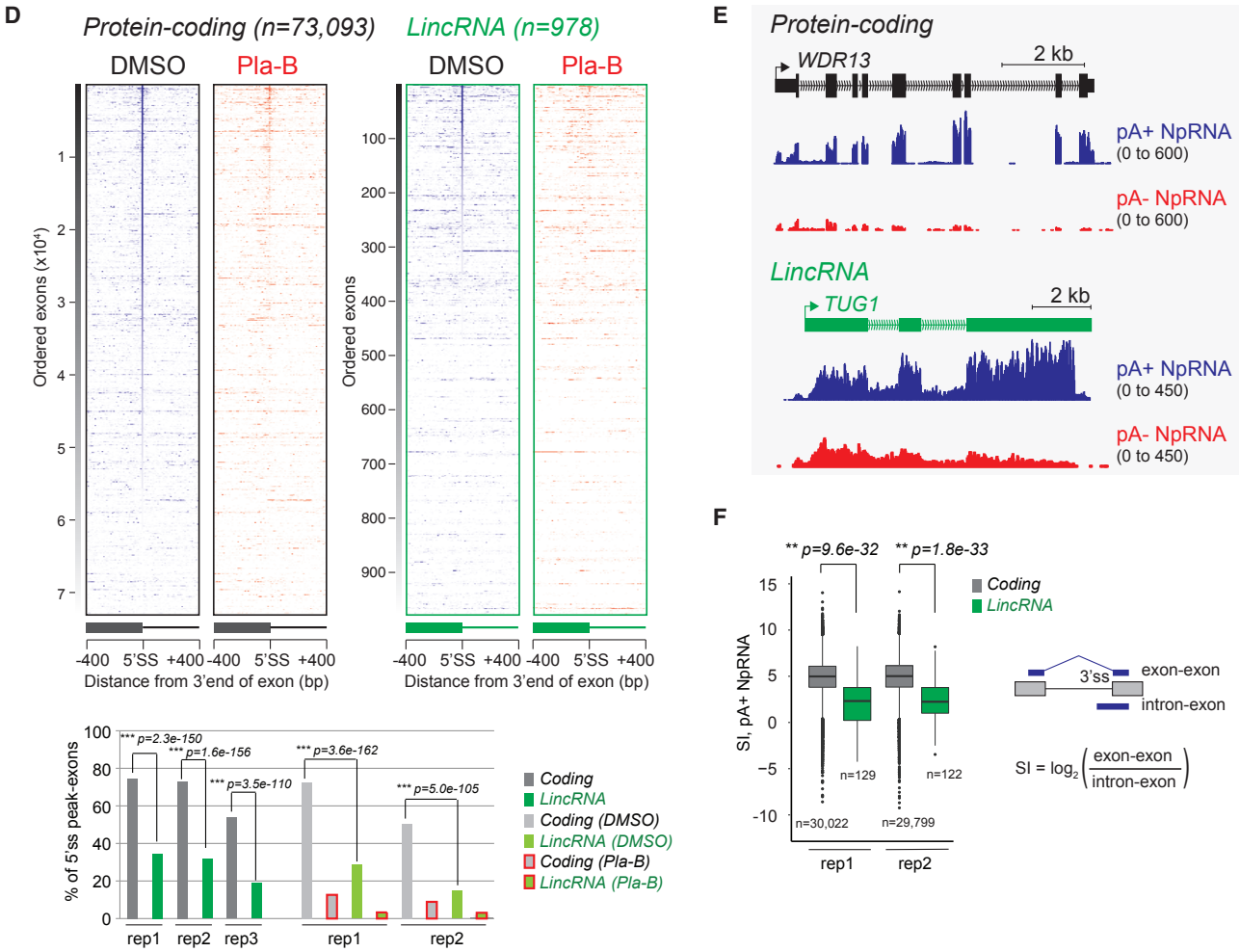
Our mNET-seq/T4P datasets show a close correlation between the CTD T4P mark and protein-coding gene termination (Figure 1). In contrast, lincRNAs show reduced T4P 3′ end association, with many showing a more widespread T4P profile across the whole TU. We previously demonstrated, based on mNET-seq/S2P analysis, that Pol II pauses over the 3′ end of protein-coding genes in a cleavage and polyadenylation factor (CPA)-dependent manner (Nojima et al., 2015). Thus, RNAi depletion of either CPSF73, the CPA endonuclease, or CstF-64/64tau, which recognize pA signal (PAS) downstream regions, markedly reduces this pausing effect.

We extended our previous data by testing the effect of CPSF73 depletion (Figure S3A) on mNET-seq/T4P profiles in duplicate. First, the specific patterns obtained for *GAPDH* versus *TUG-1* underlie the differences generally seen for protein coding versus lincRNA genes. Thus, *GAPDH* shows a clear accumulation of mNET-seq reads over the termination region that substantially shifts downstream following CPSF73 depletion (Figure 3A). Even though *GAPDH* shows a loss of PAS-dependent termination following CPSF73 depletion, a further downstream termination region is evident based on an abrupt loss of mNET-seq/T4P reads at a downstream position. We generally see this effect for protein-coding genes (Figure S3B), which may reflect a CPA-independent fail-safe termination process. Whereas the lincRNA *TUG1* profile for mNET-seq/T4P also detects some 3′ end peaks, depletion of CPSF73 does not affect this profile, suggesting *TUG1* termination is CPSF73 independent (Figure 3B). Four other lincRNAs gave similar results (Figure S3C), although *LINC00052* displayed some CPA-dependent termination especially visible in the ChrRNA-seq profiles. Again, we performed meta-analyses on the duplicate databases (Figures 3C and S3D), showing that protein coding, but not lincRNA gene termination, is strongly affected by CPSF73 depletion. We finally quantitated the effect of CPSF73 depletion on TES pausing and show that there is a significant effect on protein-coding genes compared to lincRNAs ($p = 6.2e{-}4$; Figure 3D).

To examine the degree of 3′ end polyadenylation in lincRNAs, we again employed our pA+ and pA− NpRNA-seq libraries. As expected, protein-coding transcripts were predominantly pA+, as exemplified by the *CDK9* gene (Figure 3E, top). In contrast, histone RNAs were exclusively in the pA− fraction (Figure 3E, middle), because histone mRNA is maturated by a PAS-independent mechanism (Dominski and Marzluff, 2007). Notably, lincRNAs, such as *LINC01021*, display higher pA− than pA+ reads (Figure 3E, bottom). In general, lincRNAs are inefficiently polyadenylated as compared to protein-coding transcripts as shown in our duplicated experiments (Figure 3F).

We also investigated the mNET-seq and ChrRNA-seq profiles of the lincRNA *MALAT1*. This lincRNA lacks a pA tail, being processed by RNase P to generate a 3′ terminal tRNA-like RNA, known as *MALAT1*-associated small cytoplasmic RNA (mascRNA) (Wilusz et al., 2008). The upstream *MALAT1* RNA is stabilized by the formation of a 3′ terminal triple helical structure (Brown et al., 2014). Notably, mNET-seq/T4P-detected reads peak at a TES position several kilobases downstream of mascRNA. Interestingly, this pause region is decreased by CPSF73 knockdown, suggesting *MALAT1* termination is CPA dependent (Figure 3G). Consistent with this possibility, a PAS is known to be present at the end of this downstream region (Wilusz et al., 2008). Whereas *MALAT1* is mainly present in the





(legend on next page)

pA− nucleoplasm RNA fraction due to RNase P cleavage, a small fraction of *MALAT1* RNA extending beyond the RNase P site to the PAS was detected in the pA+ fraction (Figure 3H). We also analyzed mNET-seq/S2P profiles for *MALAT1*, showing a clear termination defect following CstF64/64tau depletion (Nojima et al., 2015; Figure S3E). Furthermore, these CPA factors crosslink to the *MALAT1* PAS region based on PAR-CLIP analysis (Martin et al., 2012). Overall, these results imply a kinetic model for *MALAT1* 3′ end processing, where Pol II termination is mediated by the CPA complex at a downstream PAS, followed by co- or post-transcriptional RNase P cleavage in the nucleoplasm.

lincRNAs Are Degraded Post-transcriptionally by the Nuclear Exosome

Even though some lincRNAs have been reported to be functional (Quinn and Chang, 2016), we show above that this transcript class is both poorly spliced and polyadenylated (Figures 2 and 3). This led us to a study of lincRNA stability. We initially compared the levels of transcript reads over the TSS regions of protein coding versus lincRNA and also the antisense lncRNA class (Table S2). As shown (Figures 4A and 4B), whereas lincRNA and protein-coding gene transcripts are often similar in abundance in the chromatin fraction, lincRNA levels are substantially reduced in the nucleoplasm. In particular, we show transcription profiles for a tandem lincRNA and protein-coding gene *LBR* (Figure 4C). Whereas ChRNA-seq read levels are similar across these two adjacent TUs, little lincRNA is detectable in the nucleoplasm, suggesting that it is degraded post-transcriptionally. We also interrogated published RNA-seq data (Mayer et al., 2015) for lincRNA expression in the cytoplasm to exclude the possibility of rapid nuclear export. Again, much less cytoplasmic lincRNA is detected as compared to chromatin-associated lincRNA (Figure 4D).

It has been previously established that lncRNAs are substrates of the RNA exosome in mouse embryonic stem cells (ESCs) (Pefanis et al., 2015). However, in this study, total cellular RNA was analyzed so that it was not determined where in the cell such RNA degradation occurs. Exosome-mediated degradation of lncRNA may be triggered by the nuclear complex NEXT, which acts as an adaptor to recruit exosome to susceptible capped Pol II transcripts (Andersen et al., 2013; Lubas et al., 2015). We therefore depleted the RNA exosome component EXOSC3 (Figure S4A), which is essential for exosome activity (Chlebowski et al., 2013), and performed duplicate ChrRNA-seq and NpRNA-seq. Interestingly, lincRNAs were all significantly increased in the nucleoplasm by EXOSC3 knockdown, although RNA levels in chromatin (both ChRNA-seq and mNET-seq) were unaffected (Figures 4E and S4B). We also compared the ratio of chromatin to nucleoplasm RNA levels between protein-coding and definable classes of lncRNA genes following exosome depletion (Figures 4F and S4C). Notably, protein-coding RNA levels (first 500 nt) were slightly stabilized, suggesting some low-level turnover by the exosome of possibly mis-spliced mRNAs (Davidson et al., 2012). In contrast, tRNAs and structural ncRNAs (such as small nuclear RNAs [snRNAs]) were significantly destabilized by exosome inactivation, consistent with the known role of the exosome in tRNA and snRNA maturation (Schneider et al., 2012). Remarkably, all categories of lncRNAs (PROMPTs, eRNAs, antisense RNAs, and lincRNAs) show significant nucleoplasmic stabilization following exosome depletion. Because EXOSC3 depletion does not affect mNET-seq profiles (Figures 4E and S4B), we conclude that lincRNAs are downregulated by the nuclear RNA exosome in the nucleoplasm (Figure S4D).

Co-transcriptional RNA Cleavage of lincRNAs

We predict from the widespread profiles of mNET-seq/T4P reads across lincRNA TUs that Pol II terminates sporadically across this gene class (Figure 1). Additionally, the nuclear exosome degrades lincRNAs post-transcriptionally (Figure 4). These observations lead to the hypothesis that co-transcriptional RNA cleavage activity acting on lincRNAs might induce premature termination and that the cleaved RNA so formed can then act as a substrate for the nuclear exosome. To investigate this possibility, we searched for evidence of co-transcriptional RNA cleavage activity in our mNET-seq profiles.

The mNET-seq technique involves the ligation of a linker oligonucleotide onto any RNA 3′ end protected from micrococcal nuclease digestion. These principally derive from the Pol II active site, reflecting nascent transcription. However, co-precipitated RNA processing complexes, such as the spliceosome or microprocessor, can also generate RNA 3′ ends (detected by mNET-seq), such as splicing intermediates or microRNA precursors (Nojima et al., 2015). Because the positions of such RNA cleavage intermediates are well known (i.e., 5′ splice sites or pre-microRNA Drosha cleavage sites), their identification proved straightforward. However, RNA 3′ ends formed by unidentified RNA-processing complexes may also be co-precipitated with Pol II. To separate mNET-seq reads derived from Pol II active site RNA 3′ ends and those derived from co-precipitated RNA processing complexes, we employed the detergent Empigen to separate the Pol II core machinery from Pol-II-associated complexes, such as the spliceosome and microprocessor. Empigen is known to weaken many protein-protein interactions, but not high-affinity antigen-antibody interactions (Choi and Dreyfuss, 1984), suggesting that strong interactions should be

Figure 2. lincRNAs Are Inefficiently Spliced

(A and B) (A) mNET-seq/S5P analysis of protein-coding gene *PTCD3* and (B) lincRNA *TUG1*. HeLa cells were treated with Pla-B (red) or DMSO control (blue). Only sense transcripts are shown.

(C) Meta-analysis across exon-intron junctions (5′ss) of annotated introns for protein-coding TUs versus lincRNAs.

(D) Heatmaps for protein-coding versus lincRNA genes aligned to 5′ss −400 to +400 nt upstream and downstream. Percent of introns showing co-transcriptional 5′ss peaks is shown below, including all data repetitions, either with untreated, DMSO mock-treated, or Pla-B-treated HeLa cells.

(E) pA+ and pA− NpRNA-seq profiles are shown for *WDR13* versus lincRNA *TUG1*.

(F) Splicing index from pA+ NpRNA-seq for protein-coding and lincRNA TUs (duplicates shown).

See also Figure S2.

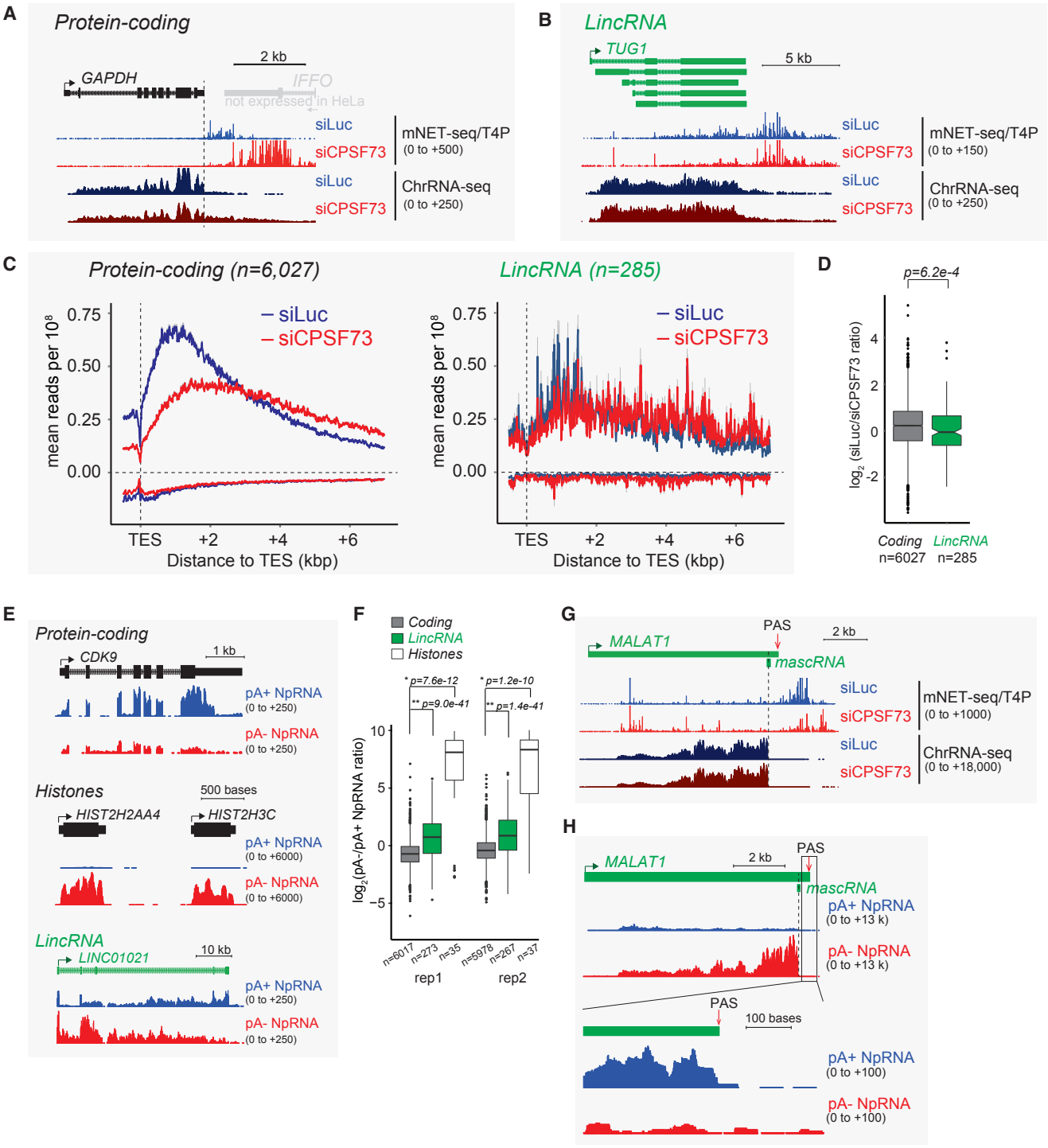


Figure 3. lincRNAs Are Largely Unpolyadenylated and CPA Independent

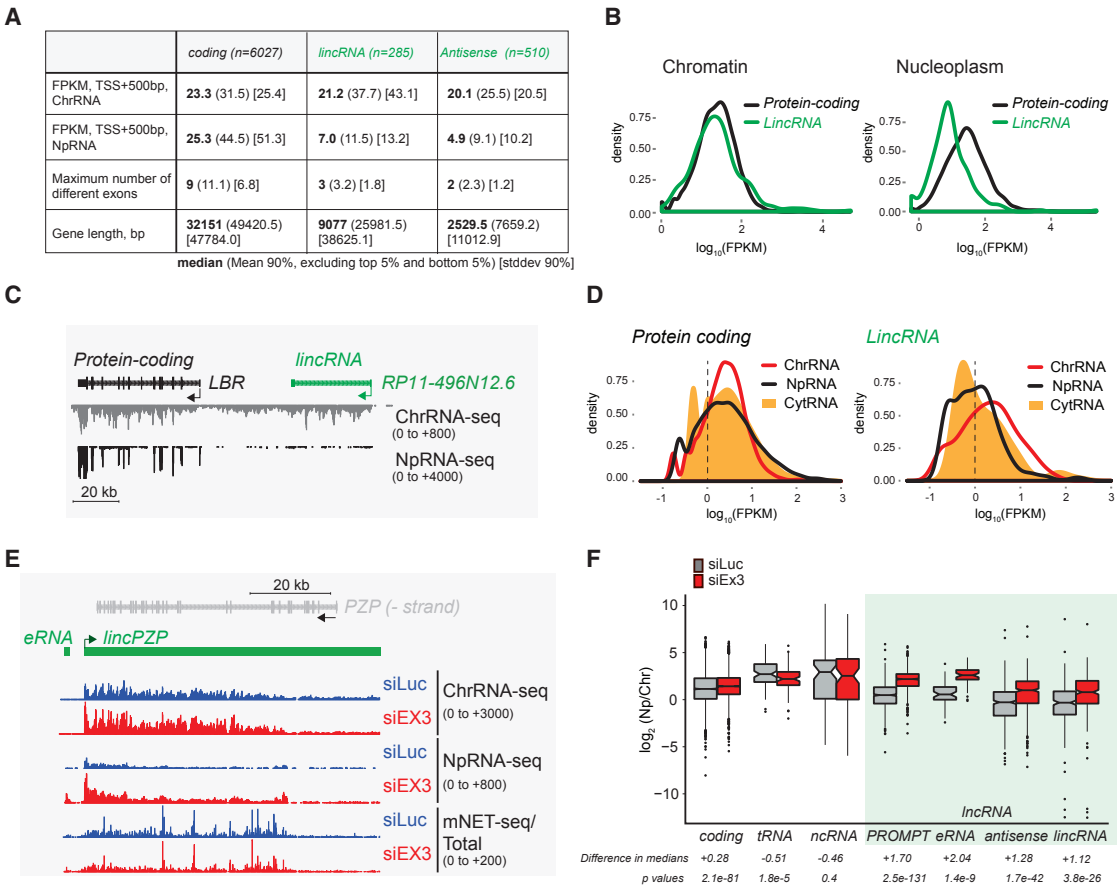
(A and B) (A) mNET-seq/T4P analysis of *GAPDH* and (B) lincRNA *TUG1*. Vertical dotted line over *GAPDH* denotes PAS.

(C) Meta-analysis of termination region (up to 7 kbp 3′ to TES) associated mNET-seq/T4P profiles, ±CPSF73 depletion by small interfering RNA (siRNA) treatment. siLuc indicates siRNA control treatment. Protein-coding TUs are shown on the left and lincRNA TUs on the right.

(D) Quantitation of readthrough transcript levels following CPSF73 depletion characterized by GB-signal-normalized siLuc to siCPSF73 signal ratio in 10 kbp downstream of TES.

(E) Gene-specific profiles (*CDK9*, histone *H2A*, histone *H3*, and *LINC01021*) for pA+ and pA−NpRNA-seq.

(legend continued on next page)



	coding (n=6027)	lincRNA (n=285)	Antisense (n=510)
FPKM, TSS+500bp, ChrRNA	**23.3** (31.5) [25.4]	**21.2** (37.7) [43.1]	**20.1** (25.5) [20.5]
FPKM, TSS+500bp, NpRNA	**25.3** (44.5) [51.3]	**7.0** (11.5) [13.2]	**4.9** (9.1) [10.2]
Maximum number of different exons	**9** (11.1) [6.8]	**3** (3.2) [1.8]	**2** (2.3) [1.2]
Gene length, bp	**32151** (49420.5) [47784.0]	**9077** (25981.5) [38625.1]	**2529.5** (7659.2) [11012.9]

median (Mean 90%, excluding top 5% and bottom 5%) [stddev 90%]

Figure 4. lincRNAs Are Chromatin Restricted and Degraded by the Nuclear Exosome

(A) Transcription levels for coding, lincRNA, and antisense RNA in chromatin or nucleoplasm as well as exon numbers and gene lengths.

(B) Density plots of chromatin and nucleoplasm fragments per kilobase of transcript per million of mapped reads (FPKM) levels (first 500 bp) for protein-coding and lincRNA TUs.

(C) ChrRNA-seq versus NpRNA-seq for tandem lincRNA and *LBR* locus.

(D) Density plots of FPKM levels in chromatin, nucleoplasm, and cytoplasm comparing protein-coding and lincRNA TUs.

(E) Comparison of ChrRNA-seq, Np-RNA seq, and mNET-seq/total Pol II for *lincPZP* ± exosome (EXOSC3). *lincPZP* is antisense to the protein-coding gene *PZP* (not expressed in HeLa cells).

(F) Quantitation of ratios of nucleoplasm to chromatin RNA levels for different classes of transcript as indicated. Non-coding RNA (ncRNA) denotes stable RNA, such as snRNA and snoRNA.

See also Figure S4.

resistant to Empigen treatment. We therefore added Empigen to the Pol II IP step in the mNET-seq procedure. As shown for mNET-seq analysis of the *MYC* gene, S5P-specific 5′ss peaks are specifically lost with Empigen treatment, presumably because the co-immunoprecipitated spliceosome containing this splicing intermediate is now released from the Pol II complex (Figure 5A). This was confirmed for a specific protein component of the spliceosome (U5 116k; Figure S5A). Similarly, the S5P-/S2P-specific microprocessor-mediated RNA cleavage intermediate is lost from the lincRNA *MIR17HG* following Empigen treatment (Figure 5B). Importantly, Y1P and T4P CTD mNET-seq signals were unaffected by Empigen treatment, implying that they all derive from the Pol II active site (Figures 5A and 5B). In addition, other signals, such as TSS-associated peaks, were unaffected (data not shown). All Empigen-treated mNET-seq libraries were duplicated and show highly consistent profiles.

Our mNET-seq analysis of individual lincRNAs, unlike protein-coding genes, reveals numerous Empigen-sensitive peaks, as shown for *MALAT1* and *LINC01021* in mNET-seq/S5P and S2P profiles (Figure 5C) and several other lincRNAs (Figure S5B). In

(F) Quantitation of levels of pA−/pA+ transcripts for protein coding versus lincRNA TUs based on number of fragments overlapping TUs. Duplicate data are shown.

(G) mNET-seq/T4P versus ChrRNA-seq profiles for *MALAT1*. mascRNA and PAS positions are indicated.

(H) pA+/pA− RNA-seq for MALAT1. 3′ end of TU is expanded.

See also Figure S3.

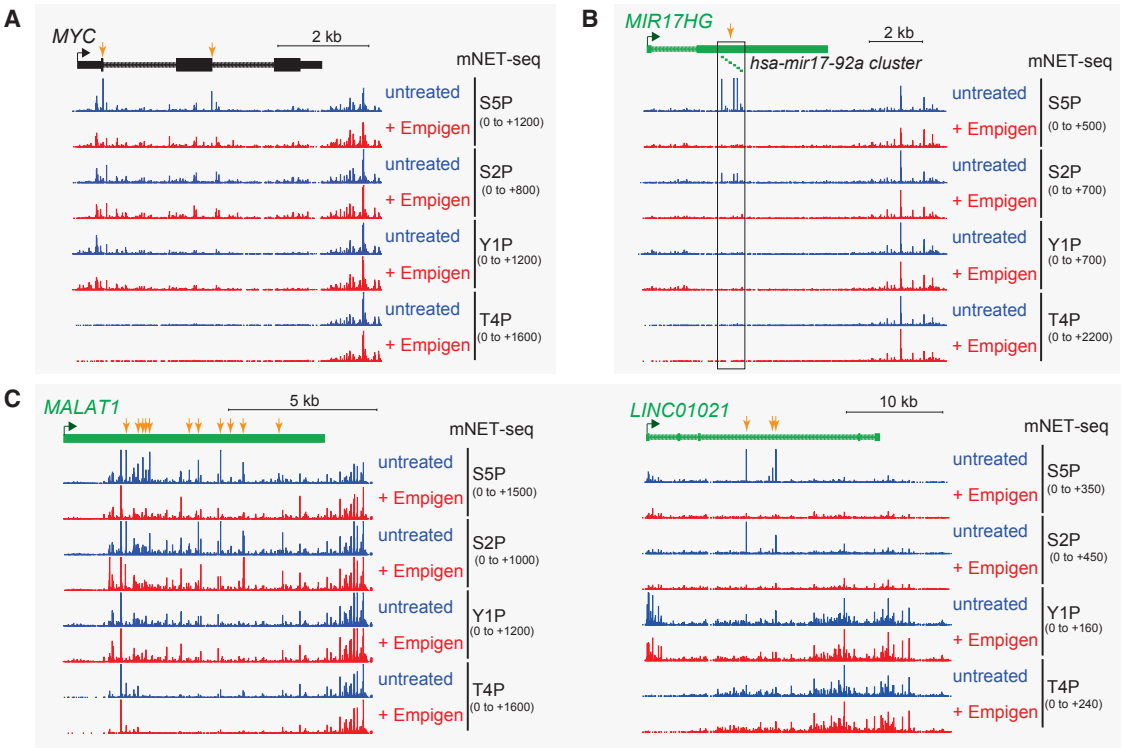


Figure 5. Identification of Co-associated RNA-Processing Complexes with Pol II

Comparison mNET-seq/S5P, S2P, Y1P, and T4P profiles with or without Empigen treatment for (A) *MYC*, (B) *MIR17HG*, (C) *MALAT1*, and (C) *LINC01021*, respectively. Orange arrows denote Empigen-sensitive peaks.

See also Figure S5.

many cases, peak levels reduced rather than completely disappeared. These Empigen-sensitive peaks indicate that lincRNAs are co-transcriptionally cleaved at multiple positions across their TUs. Notably, most Empigen-sensitive lincRNA peaks are insensitive to Pla-B treatment (Figure S2A), indicating that they are distinct from splicing intermediates (Nojima et al., 2015). Overall, we show that Empigen treatment can be employed to distinguish co-transcriptional RNA cleavage activity from ongoing transcription in the Pol II active site.

Role of RNAi Factors in lincRNA Degradation

We reasoned that possible endonucleases responsible for lincRNA degradation could be either nuclear Drosha as part of the microprocessor (with DGCR8) or the related RNase III endonuclease Dicer. Although Dicer activity is predominantly cytoplasmic, where it acts to process pre-microRNA into microRNA (Ha and Kim, 2014), nuclear Dicer has been reported in recent studies to play various roles in nuclear RNAi pathways (Burger and Gullerova, 2015). We therefore generated mNET/S5P datasets using chromatin from HeLa cells depleted for either DGCR8 or Dicer (Figure S6A). Note that DGCR8 depletion also inactivates Drosha as an integral part of the microprocessor (Dhir et al., 2015). Neither DGCR8 nor Dicer depletion affected mNET-seq/S5P profiles on the protein-coding gene *CCND1* (Figures 6A and S6B). In contrast, for *MIR17HG*, which encodes the miR17-92a cluster, mNET-seq peaks corresponding to release of these pre-miRNAs were abolished and a transcription termination defect was detected (Figures 6B and S6C) following DGCR8, but not Dicer, depletion. This confirms that microprocessor-mediated cleavage of linc-pre-miRNAs induces Pol II termination defects (Dhir et al., 2015). However, neither loss of the microprocessor (by DGCR8 knockdown) nor Dicer caused a general loss of lincRNA mNET-seq/S5P peaks (Figures 6C, 6D, S6D, and S6E), arguing against a role for these endonucleases in lincRNA cleavage.

Recent studies show that DGCR8, the RNA-binding component of the microprocessor, interacts with nuclear RNA exosome components, independently of the endonuclease Drosha (Macias et al., 2015). In this situation, it facilitates exosome recruitment to degrade abundant lncRNAs, such as small nucleolar RNAs (snoRNAs) and human telomerase RNA component (hTERC). Because we show that the nuclear RNA exosome degrades lincRNAs, we investigated whether DGCR8 is also involved in lincRNA turnover. Interestingly, DGCR8, but not Dicer, depletion acted to selectively increase Empigen-sensitive mNET-seq/S5P peaks on lincRNA genes, such as *MALAT1* and *LINC01021* (Figures 6C and 6D). This suggests that DGCR8 also acts to recruit the exosome to co-transcriptionally cleaved lincRNA, independently of miRNA. Consistent with our mNET-seq data, some lincRNA levels increase at a steady-state level based on whole-cell RNA-seq analysis (Figure S5C; Macias et al., 2015).

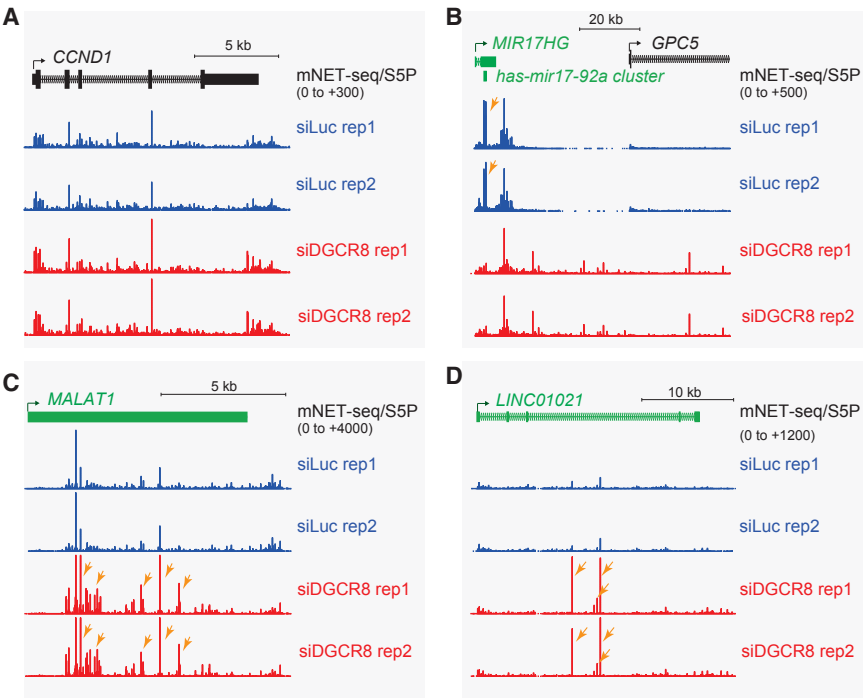


Figure 6. Effect of DGCR8 Depletion on Co-transcriptional Processing

mNET-seq/S5P profiles for (A) *CCND1*, (B) *MIR17HG-GPC5*, (C) *MALAT1*, and (D) *LINC01021* with DGCR8 siRNA-mediated depletion or control siLuc treatment. Orange arrow indicates loss of pre-miRNA cleavage for *MIR17HG* or elevated levels of cleavage products following DGCR8 depletion for *MALATI* and *LINC01021*. Duplicate mNET-seq/S5Ps are presented to underline data reproducibility.

See also Figure S6.

PCA Reveals lincRNAs Are Generally Distinct from Protein-Coding Genes

We employed principal-component analysis (PCA) to compare protein-coding versus lincRNA TUs based on multiple parameters. Because our restricted lincRNA set displays very similar profiles to the larger antisense lncRNA set (Figure 4F), these were combined for PCA. The effects of exosome knockdown on levels of nuclear RNA, nuclear-to-chromatin-associated RNA ratio, cytoplasmic-to-chromatin-associated RNA ratio, and the pA− to pA+ RNA ratio were collapsed into a two-dimensional representation in the principal components PC1 and PC2. The vectors depicted by arrows show the projection of the original four descriptors onto the PC1 and PC2 planes (Figure 7A). The main descriptor of lincRNA TUs is their upregulation upon exosome knockdown and their general lack of polyA. In contrast, the most distinguishing feature for protein-coding TUs is their stability within the nucleoplasm and cytoplasm. We note that a few lincRNAs behave in a similar manner to protein-coding TUs and are therefore potentially functional. Two clear examples are lincRNA *LINC00493* and *TINCR*, which are spliced, polyadenylated, and show an accumulation of nucleoplasm-spliced reads that lack exosome sensitivity (Figure S7A). Further examples of such potentially functional lincRNAs are listed (Table S2). We also analyzed protein-coding TUs, which have similar values in PC1 and PC2 to bulk lincRNA. Remarkably, the majority of these transcripts originate from an upstream promoter with respect to the main gene TSS (defined by higher chromatin-seq reads) and show significantly higher exosome sensitivity than transcripts from the main TSS (Figure S7B). In many cases, they derive from antisense transcripts (PROMPTs) emanating from an adjacent divergent protein-coding gene that will then read into the open reading frame (ORF) of the downstream gene (Table S2).

The full list of principal component (PC) values and the identified lincRNA-like protein-coding genes and protein-coding-like lincRNAs can be found in Table S2. Finally, it should be noted that PCA of lincRNAs derived from NONCODE without the elimination of overlapping TUs fails to show significant pattern differences with protein-coding genes (Figure S7C). Most of these lincRNAs behave similarly to protein-coding genes because they overlap with protein-coding genes or fall within their extended transcription termination regions. This emphasizes the importance of defining separate TUs to avoid lincRNA misidentification. Overall, we demonstrate that lincRNAs behave as a separate class of transcripts to protein-coding genes. They are co-transcriptionally cleaved by a Pol-II-associated endonuclease complex, which may in turn act to promote premature termination across lincRNA TUs (marked by T4P-specific mNET-seq profiles). Coupled to this, DGCR8 recognizes these 3′ ends and recruits the nuclear exosome to fully degrade these short-lived lincRNAs (Figure 7B).

DISCUSSION

We have analyzed HeLa cell nascent transcription using mNET-seq methodology (Nojima et al., 2015, 2016), employing a full set of CTD phosphorylation-specific antibodies (Figure 1). Armed with this wide repertoire of CTD-specific nascent transcript profiles, we have been able to scrutinize potential differences between protein-coding and lincRNA genes. In general, protein-coding genes show higher selectivity for specific CTD modifications. Thus, unphosphorylated CTD (together with Y1P) is a hallmark of TSS-paused protein-coding gene transcripts whereas T4P CTD precisely defines their termination regions. S5P and S2P CTD profiles then match key co-transcriptional pre-mRNA processing states (splicing and 3′ end cleavage and polyadenylation). In contrast, lincRNA CTD profiles appear less selective with all the above-mentioned CTD tendencies of protein-coding genes diminished. Whereas Pol II pausing at the TSS and TES of protein-coding genes appears to be a tightly regulated process, this is generally absent for lincRNA

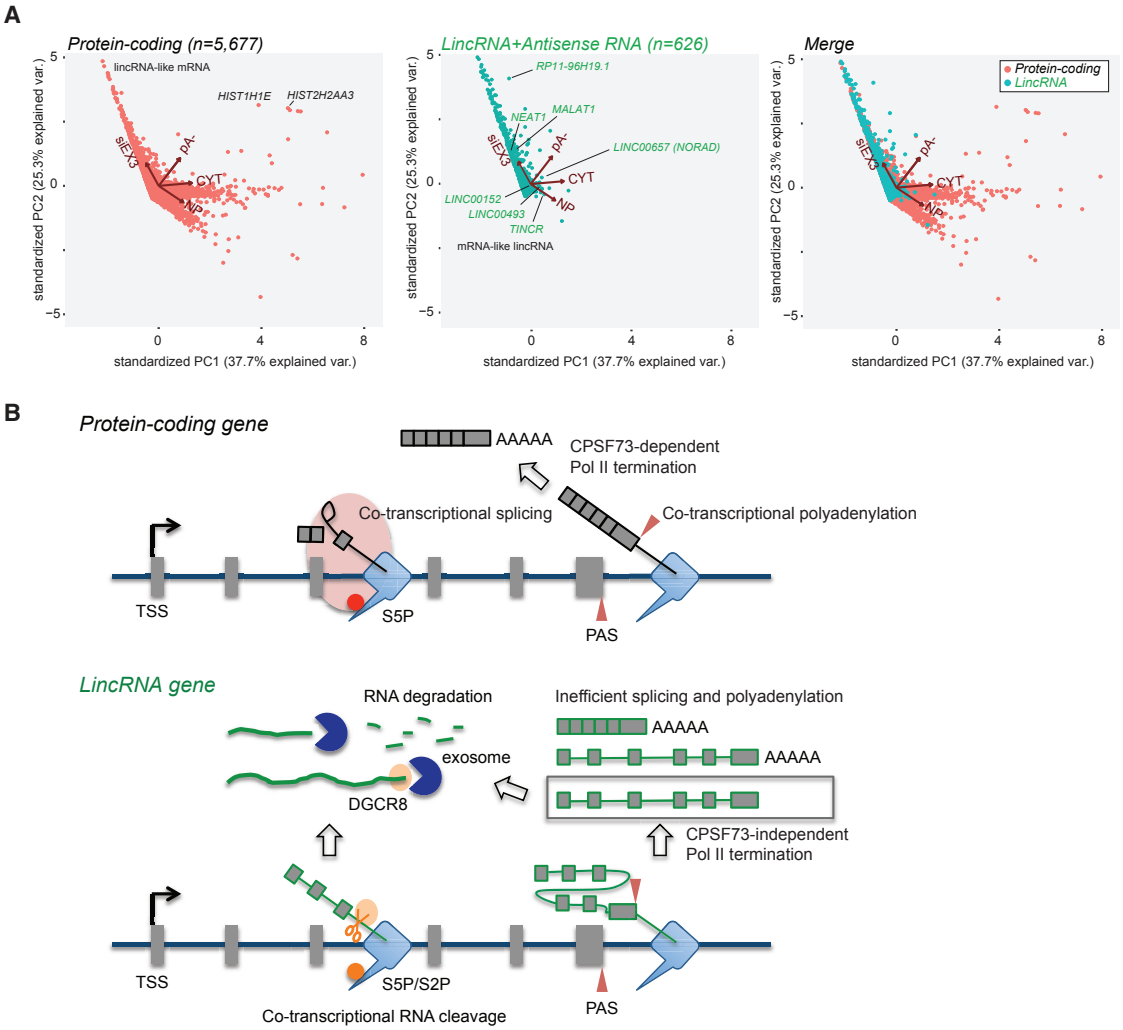


Figure 7. Protein Coding versus lincRNA Defining Features: PCA and Model

(A) Principal-component analysis applied to protein-coding and lincRNA TUs shown separately and merged. Vectors indicating key parameters compared are shown by arrows: these are exosome sensitivity, pA−/pA+ levels, cytoplasmic/chromatin, and nucleoplasmic/chromatin levels. Some key lincRNAs are identified as well as some protein-coding transcript outliers. The graph has been cropped for better visualization, but PC1 and PC2 values of all data points are available in Table S2.

(B) Model for protein-coding versus lincRNA co-transcriptional processing. Protein-coding genes are transcribed by Pol II with spliceosome (pink oblong) associated with CTD S5P (red dot). mRNA 3′ ends are generated co-transcriptionally by CPSF73 as part of CPA complex, which in turn promotes Pol II termination. lincRNA genes are weakly spliced and polyadenylated, resulting in CPSF73-independent termination and DGCR8-stimulated exosome degradation with co-transcriptional cleavage (scissors) associated with CTD S2P and S5P (orange dot) and exosome-mediated degradation on chromatin.

See also Figure S7.

genes. Similarly, the dominant RNA-processing reactions, co-transcriptional splicing, and 3′ cleavage and polyadenylation are associated with precise CTD marks S5P and S2P. Again, lincRNAs, which are largely unspliced (Figure 2) and generally not 3′ end processed (Figure 3), lack these dominant phospho-CTD features. Because this RNA processing is required to generate translatable mRNAs, it appears logical that noncoding lincRNAs lack the transcriptional CTD code that enhances these processes.

We observe less Pol II pausing over the TES region of lincRNA genes, compared to protein-coding genes (Figure S1B, bottom). Protein-coding gene TES pausing depends on CPA factors, such as CPSF73 and CstF64/64 tau using unph, S2P, and S5P Pol II CTD antibodies (Nojima et al., 2015). Here, we show that the mNET-seq/T4P profile gives the largest Pol II read accumulation in the TES region of protein-coding genes. Whereas this pausing effect at the TES is decreased by depletion of CPSF73 protein (Figure 3C, left), the profile switches to other T4P CTD peaks further downstream (Figures 3A and S3B). We hypothesize that the observed downstream CPA-independent termination is a failsafe mechanism. Possibly, additional terminators beyond CPA-dependent mechanisms are generally present to restrict

transcriptional interference caused by uncontrolled transcriptional readthrough (Greger and Proudfoot, 1998; Rutkowski et al., 2015). Interestingly, mNET-seq/T4P peaks at lincRNA TES are in general CPSF73 independent (Figures 3B, 3D, and S3C). Some lincRNAs retain CPA-independent termination, even though they lack CPA-dependent mechanisms. Consistent with this result, we also confirm lincRNAs are in general inefficiently 3′ end polyadenylated (Figure 3F). We note that mNET-seq/T4P signals in the lincRNA gene body are often decreased by CPSF73 knockdown (Figure S3C). This suggests that premature termination of lincRNAs may still be regulated by CPA factors.

Our analysis of HeLa cell lincRNAs by subcellular RNA-seq analysis reveals a clear pathway to their rapid degradation (Figure 7). First, we show that lincRNAs are mainly restricted to the nuclear chromatin fraction, as observed for eRNAs, PROMPTs, and antisense RNAs. We also demonstrate that chromatin-restricted lincRNAs are degraded by the nuclear exosome as soon as they are made (Figures 4E, 4F, S4B, and S4C). However, to be substrates for exosome-associated 3′ exonuclease, lincRNAs must first be cleaved by endonucleases to generate accessible 3′ ends. Our mNET-seq analysis of lincRNAs using Empigen treatment indicates the presence of a separable endonuclease complex associated with Pol II. Thus, Empigen treatment removes multiple cleavage sites across lincRNAs, which are detectable as peaks in the mNET-seq analysis. These RNA 3′ ends do not derive from splicing because their appearance is insensitive to the splicing inhibitor Pla-B.

We examined the possibility that lincRNA endonucleolytic cleavage could be generally mediated by the microprocessor. Components of microprocessor, Drosha, and DGCR8 proteins cleave pre-miRNA structures co-transcriptionally (Morlando et al., 2008; Nojima et al., 2015). We therefore suspected that lincRNAs might possess multiple pre-miRNA-like secondary structures and so be cleaved by the microprocessor. Depletion of DGCR8 (which causes inactivation of the microprocessor) followed by mNET-seq analysis removed mNET-seq peaks corresponding to authentic pre-miRNAs (Figure 6B). However, unexpectedly, Empigen-sensitive cleavage sites on lincRNAs were generally increased by DGCR8 knockdown (Figures 6 and S5B). Because DGCR8 is both associated with elongating Pol II and with RNA exosome components, it is likely to enhance exosome activity. It is, however, also possible that DGCR8 plays a regulatory role in the recruitment or activity of the presumptive lincRNA endonuclease. Overall, we propose a model for lincRNA degradation in which these weakly spliced and polyadenylated transcripts are largely degraded post-transcriptionally by DGCR8-mediated recruitment of the nuclear exosome (Figure 7B). Another feature of lincRNA transcription is that many transcripts prematurely terminate well before reaching the distal TES.

We ended our bioinformatics comparison of lincRNA TUs versus protein-coding TUs by subjecting them to PCA (Figure 7A). Remarkably, lincRNAs gave a characteristic profile showing high exosome sensitivity. However, a few lincRNAs display more protein-coding-like properties (Figure S7A; Table S2) and so may represent transcripts with specific functions. Notably, protein-coding TUs gave a mainly non-overlapping PC profile with lincRNA TUs. Those that did significantly match the lincRNA PC profile correspond to transcripts derived from upstream start sites and often come from divergent gene PROMPTs. These can therefore be viewed as lincRNA TUs. Overall, our bioinformatics comparison of lincRNA versus protein-coding TUs underlies substantial differences between these two transcript classes. In general, lincRNAs appear unlikely to possess sequence-specific functions. Possibly, the act of transcription rather than the nature of the transcript underlies their biological purpose. However, it remains an attractive possibility that tissue-specific RNA-binding proteins (possibly absent in HeLa cells) may selectively restrict lincRNA turnover and so allow their sufficient accumulation to promote functional roles at least for some of these RNAs.

EXPERIMENTAL PROCEDURES

mNET-Seq and Fractionated RNA-Seq

Detailed protocols for mNET-seq, ChrRNA-seq, and NpRNA-seq were previously described (Nojima et al., 2015, 2016). For mNET-seq/total, unph, S2P, and S5P, published data were used (Nojima et al., 2015).

Transcription Unit Annotation

Hg19/GRCh37 was used as a reference genome. TUs were extracted based on ENSEMBL (GRCh37.75; Flicek et al., 2014), NONCODE v4 (Xie et al., 2014), and UCSC tRNA (Lowe and Eddy, 1997). PROMPTs were extracted based on published data (Ntini et al., 2013), and ubiquitously expressed eRNAs were taken from PrESSTo (FANTOM 5 project; Andersson et al., 2014). Overlapping, expressed TUs and exons were reduced to the most upstream and downstream boundaries. Some overlapping TUs with different biotypes were excluded from further analysis. Defined TUs were categorized by biotype (Tables S1 and S2).

Data Processing

RNA-sequencing reads were trimmed by Cutadapt 1.8.3 and then mapped to the human hg19 reference sequence with Tophat 2.0.13. All sequencing data were processed to only include properly paired, properly mapped reads with SAMtools 1.2. mNET-seq profiles were created by only using the most 3′ nucleotide of the second sequencing read. Data were visualized with Bedtools 2.23.0 and scaled to each library size (genomeCoverageBed).

Bioinformatic Analysis

Heatmaps were created using the MATLAB R2015b image function. All other graphs were created using ggplot2 in R. p values are computed via a Wilcoxon test in R or a Fisher exact test in MATLAB (Figure 2D). PCA is based on the R prcomp function and visualized with ggbiplot.

ACCESSION NUMBERS

The accession number for the sequencing data reported in this paper is GEO: GSE81662.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2016.11.029.

AUTHOR CONTRIBUTIONS

M.S., with advice from T.G., performed all bioinformatics analyses. T.N. performed all molecular biology and transcriptomic experiments with help from A.D. on the microprocessor. N.J.P. and T.N. designed the project and wrote the paper with help from M.S. and M.C.-F.

ACKNOWLEDGMENTS

We thank Lars Steinmetz at EMBL for hosting M.S. as an EMBO STF and in particular Vicent Pelechano and Aaron Brooks for useful discussions. We are also grateful to the N.J.P. lab for helpful advice. This work was supported by grants to N.J.P. (European Research Council advanced grant [339270] and Wellcome Trust Investigator Award [107928/Z/15/Z]) and to M.C.-F. (Fundação para a Ciência e Tecnologia, Portugal grant [PTDC/BEX-BCM/5899/2014]). We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z) for the generation of the Sequencing data.

Received: May 24, 2016
Revised: July 19, 2016
Accepted: November 17, 2016
Published: December 22, 2016

REFERENCES

Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009). Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature *457*, 1028–1032.

Andersen, P.R., Domanski, M., Kristiansen, M.S., Storvall, H., Ntini, E., Verheggen, C., Schein, A., Bunkenborg, J., Poser, I., Hallais, M., et al. (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nat. Struct. Mol. Biol. *20*, 1367–1376.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al.; FANTOM Consortium (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Boutz, P.L., Bhutkar, A., and Sharp, P.A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. *29*, 63–80.

Brown, J.A., Bulkley, D., Wang, J., Valenstein, M.L., Yario, T.A., Steitz, T.A., and Steitz, J.A. (2014). Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. Nat. Struct. Mol. Biol. *21*, 633–640.

Burger, K., and Gullerova, M. (2015). Swiss army knives: non-canonical functions of nuclear Drosha and Dicer. Nat. Rev. Mol. Cell Biol. *16*, 417–430.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Chlebowski, A., Lubas, M., Jensen, T.H., and Dziembowski, A. (2013). RNA decay machines: the exosome. Biochim. Biophys. Acta *1829*, 552–560.

Choi, Y.D., and Dreyfuss, G. (1984). Monoclonal antibody characterization of the C proteins of heterogeneous nuclear ribonucleoprotein complexes in vertebrate cells. J. Cell Biol. *99*, 1997–2004.

Davidson, L., Kerr, A., and West, S. (2012). Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. EMBO J. *31*, 2566–2578.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Dhir, A., Dhir, S., Proudfoot, N.J., and Jopling, C.L. (2015). Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. Nat. Struct. Mol. Biol. *22*, 319–327.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dominski, Z., and Marzluff, W.F. (2007). Formation of the 3′ end of histone mRNA: getting closer to the end. Gene *396*, 373–390.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. Nucleic Acids Res. *42*, D749–D755.

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods *8*, 469–477.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Greger, I.H., and Proudfoot, N.J. (1998). Poly(A) signals control both transcriptional termination and initiation between the tandem GAL10 and GAL7 genes of Saccharomyces cerevisiae. EMBO J. *17*, 4771–4779.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., and Herrmann, B.G. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev. Cell *24*, 206–214.

Gu, W., Lee, H.C., Chaves, D., Youngman, E.M., Pazour, G.J., Conte, D., Jr., and Mello, C.C. (2012). CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. Cell *151*, 1488–1500.

Ha, M., and Kim, V.N. (2014). Regulation of microRNA biogenesis. Nat. Rev. Mol. Cell Biol. *15*, 509–524.

Heidemann, M., Hintermair, C., Voß, K., and Eick, D. (2013). Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. Biochim. Biophys. Acta *1829*, 55–62.

Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier, P., Flatley, A., Kremmer, E., et al. (2012). Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. EMBO J. *31*, 2784–2797.

Hsin, J.P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev. *26*, 2119–2137.

Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol. Cell *52*, 473–484.

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.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. (2006). CAGE: cap analysis of gene expression. Nat. Methods *3*, 211–222.

Kotake, Y., Sagane, K., Owa, T., Mimori-Kiyosue, Y., Shimizu, H., Uesugi, M., Ishihama, Y., Iwata, M., and Mizui, Y. (2007). Splicing factor SF3b as a target of the antitumor natural product pladienolide. Nat. Chem. Biol. *3*, 570–575.

Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D., et al. (2012). Intragenic enhancers act as alternative promoters. Mol. Cell *45*, 447–458.

Lee, S., Kopp, F., Chang, T.C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y., and Mendell, J.T. (2016). Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. Cell *164*, 69–80.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Lubas, M., Andersen, P.R., Schein, A., Dziembowski, A., Kudla, G., and Jensen, T.H. (2015). The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. Cell Rep. *10*, 178–192.

Macias, S., Cordiner, R.A., Gautier, P., Plass, M., and Cáceres, J.F. (2015). DGCR8 acts as an adaptor for the exosome complex to degrade double-stranded structured RNAs. Mol. Cell *60*, 873–885.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. Cell Rep. *1*, 753–763.

Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. PLoS Genet. *5*, e1000459.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541–554.

Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. Cell *136*, 688–700.

Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I., and Proudfoot, N.J. (2008). Primary microRNA transcripts are processed co-transcriptionally. Nat. Struct. Mol. Biol. *15*, 902–909.

Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. Cell *161*, 526–540.

Nojima, T., Gomes, T., Carmo-Fonseca, M., and Proudfoot, N.J. (2016). Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. Nat. Protoc. *11*, 413–428.

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

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.P., et al. (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. Cell *161*, 774–789.

Quinn, J.J., and Chang, H.Y. (2016). Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. *17*, 47–62.

Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C.C., and Dölken, L. (2015). Widespread disruption of host transcription termination in HSV-1 infection. Nat. Commun. *6*, 7126.

Schneider, C., Kudla, G., Wlotzka, W., Tuck, A., and Tollervey, D. (2012). Transcriptome-wide analysis of exosome targets. Mol. Cell *48*, 422–433.

St Laurent, G., Wahlestedt, C., and Kapranov, P. (2015). The landscape of long noncoding RNA classification. Trends Genet. *31*, 239–251.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. *14*, 103–105.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26–46.

Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2015). Widespread inducible transcription downstream of human genes. Mol. Cell *59*, 449–461.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol. Cell *43*, 904–914.

Werner, M.S., and Ruthenburg, A.J. (2015). Nuclear fractionation reveals thousands of chromatin-tethered noncoding RNAs adjacent to active genes. Cell Rep. *12*, 1089–1098.

Wilusz, J.E., Freier, S.M., and Spector, D.L. (2008). 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell *135*, 919–932.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. *42*, D98–D103.

Molecular Cell, Volume *65*

Supplemental Information

Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs

Margarita Schlackow, Takayuki Nojima, Tomas Gomes, Ashish Dhir, Maria Carmo-Fonseca, and Nick J. Proudfoot

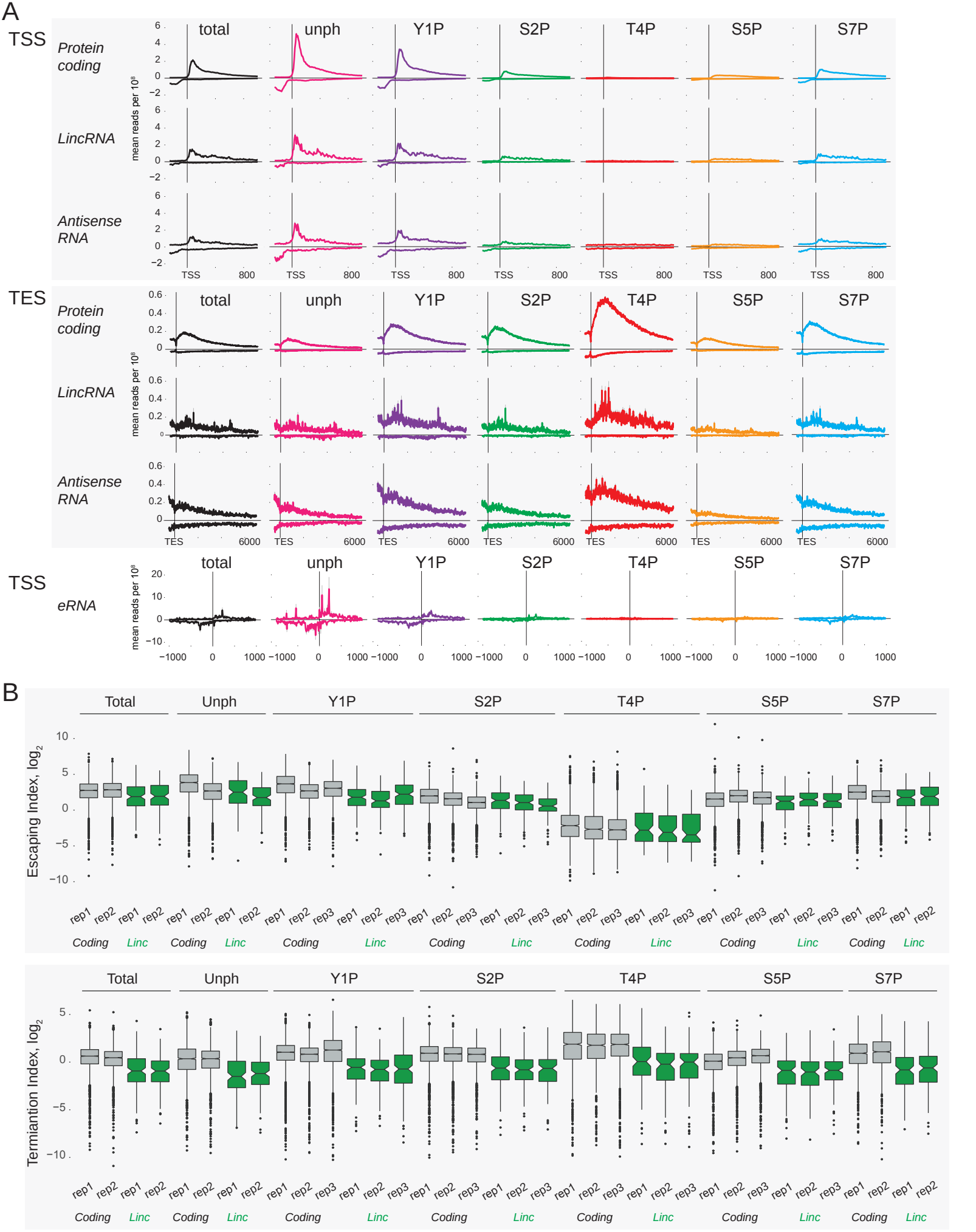

A
TSS
Protein coding
LincRNA
Antisense RNA
total
unph
Y1P
S2P
T4P
S5P
S7P
mean reads per 10⁸
TSS
800
TES
6000
eRNA
B
Escaping Index, log2
Termiantion Index, log2
Total
Unph
rep1
rep2
rep3
Coding
Linc

Figure S1

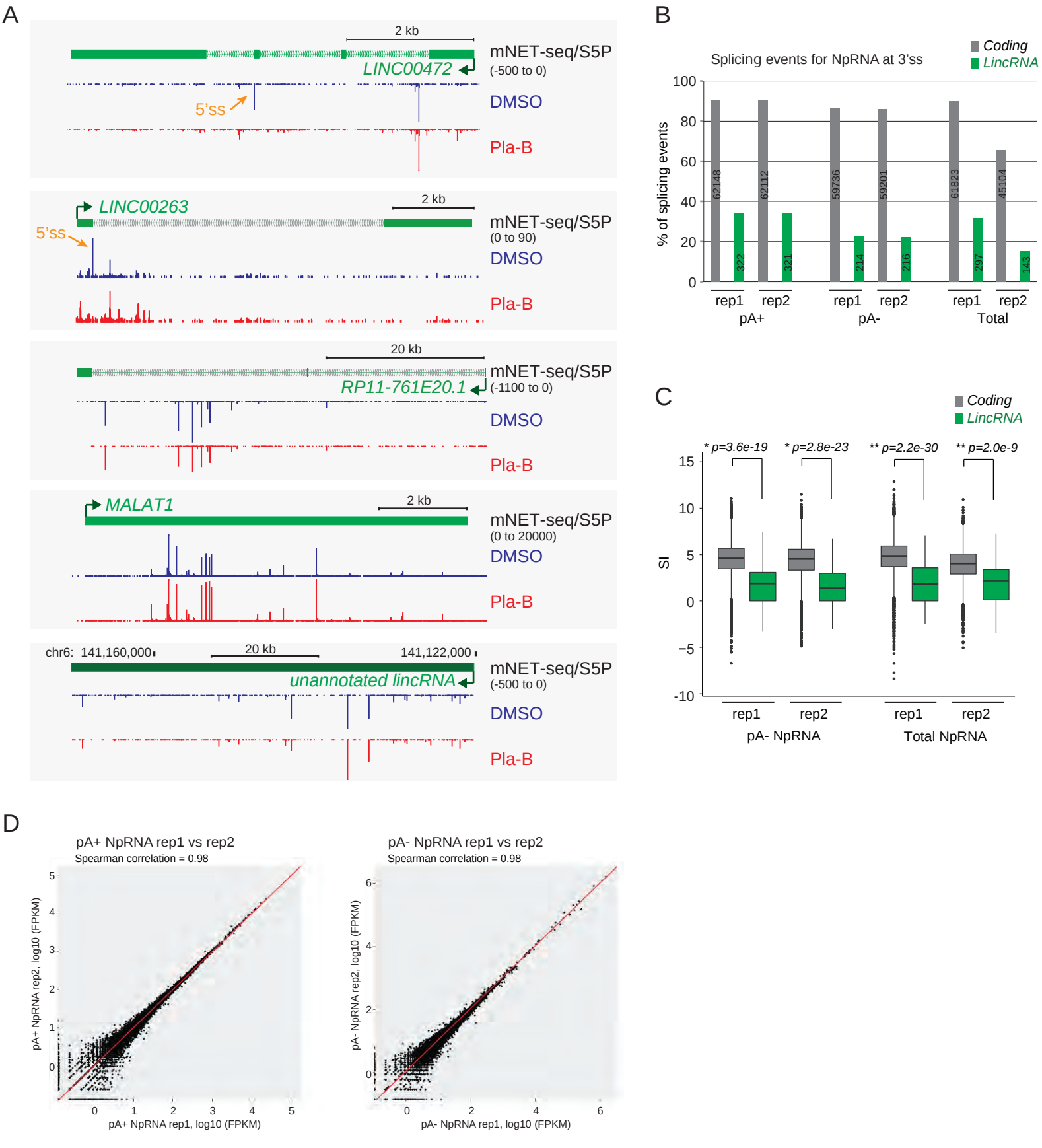

A
2 kb
LINC00472
5'ss
mNET-seq/S5P
(-500 to 0)
DMSO
Pla-B
LINC00263
5'ss
2 kb
mNET-seq/S5P
(0 to 90)
DMSO
Pla-B
20 kb
RP11-761E20.1
mNET-seq/S5P
(-1100 to 0)
DMSO
Pla-B
MALAT1
2 kb
mNET-seq/S5P
(0 to 20000)
DMSO
Pla-B
chr6: 141,160,000
20 kb
141,122,000
unannotated lincRNA
mNET-seq/S5P
(-500 to 0)
DMSO
Pla-B
B
Splicing events for NpRNA at 3'ss
Coding
LincRNA
% of splicing events
100
80
60
40
20
0
62148
322
62112
321
59736
214
59201
216
61823
297
45104
143
rep1
rep2
pA+
pA-
Total
C
Coding
LincRNA
* p=3.6e-19
* p=2.8e-23
** p=2.2e-30
** p=2.0e-9
SI
15
10
5
0
−5
-10
rep1
rep2
pA- NpRNA
Total NpRNA
D
pA+ NpRNA rep1 vs rep2
Spearman correlation = 0.98
pA+ NpRNA rep2, log10 (FPKM)
pA+ NpRNA rep1, log10 (FPKM)
pA- NpRNA rep1 vs rep2
Spearman correlation = 0.98
pA- NpRNA rep2, log10 (FPKM)
pA- NpRNA rep1, log10 (FPKM)

Figure S2

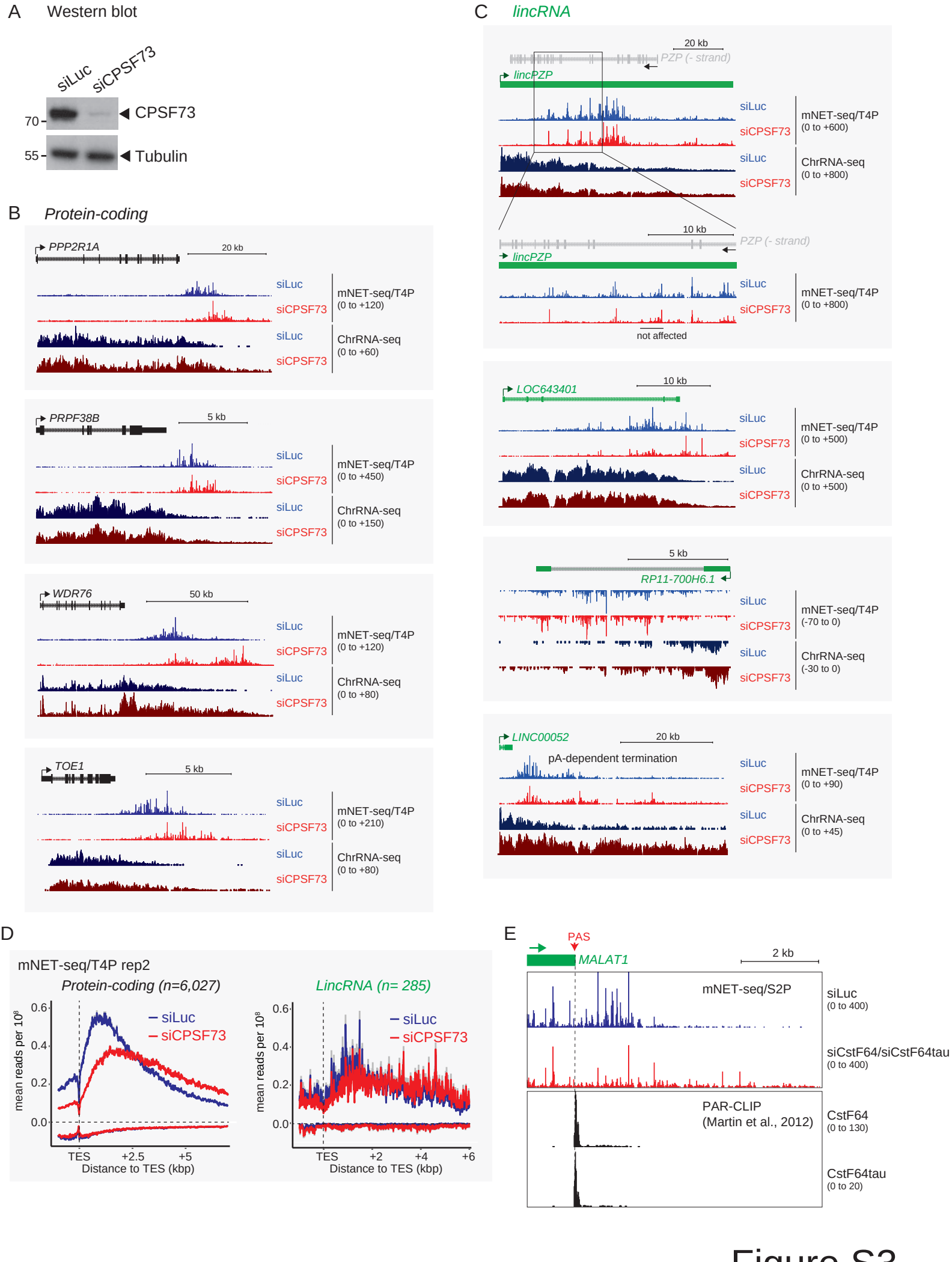

A
Western blot
siLuc
siCPSF73
70
CPSF73
55
Tubulin
B
Protein-coding
PPP2R1A
20 kb
siLuc
siCPSF73
mNET-seq/T4P
(0 to +120)
ChrRNA-seq
(0 to +60)
PRPF38B
5 kb
mNET-seq/T4P
(0 to +450)
ChrRNA-seq
(0 to +150)
WDR76
50 kb
mNET-seq/T4P
(0 to +120)
ChrRNA-seq
(0 to +80)
TOE1
5 kb
mNET-seq/T4P
(0 to +210)
ChrRNA-seq
(0 to +80)
C
lincRNA
20 kb
PZP (- strand)
lincPZP
mNET-seq/T4P
(0 to +600)
ChrRNA-seq
(0 to +800)
10 kb
mNET-seq/T4P
(0 to +800)
not affected
LOC643401
10 kb
mNET-seq/T4P
(0 to +500)
ChrRNA-seq
(0 to +500)
5 kb
RP11-700H6.1
mNET-seq/T4P
(-70 to 0)
ChrRNA-seq
(-30 to 0)
LINC00052
20 kb
pA-dependent termination
mNET-seq/T4P
(0 to +90)
ChrRNA-seq
(0 to +45)
D
mNET-seq/T4P rep2
Protein-coding (n=6,027)
LincRNA (n= 285)
mean reads per 10⁸
0.6
0.4
0.2
0.0
siLuc
siCPSF73
TES
+2.5
+5
+2
+4
+6
Distance to TES (kbp)
E
PAS
MALAT1
2 kb
mNET-seq/S2P
siLuc
(0 to 400)
siCstF64/siCstF64tau
(0 to 400)
PAR-CLIP
(Martin et al., 2012)
CstF64
(0 to 130)
CstF64tau
(0 to 20)

Figure S3

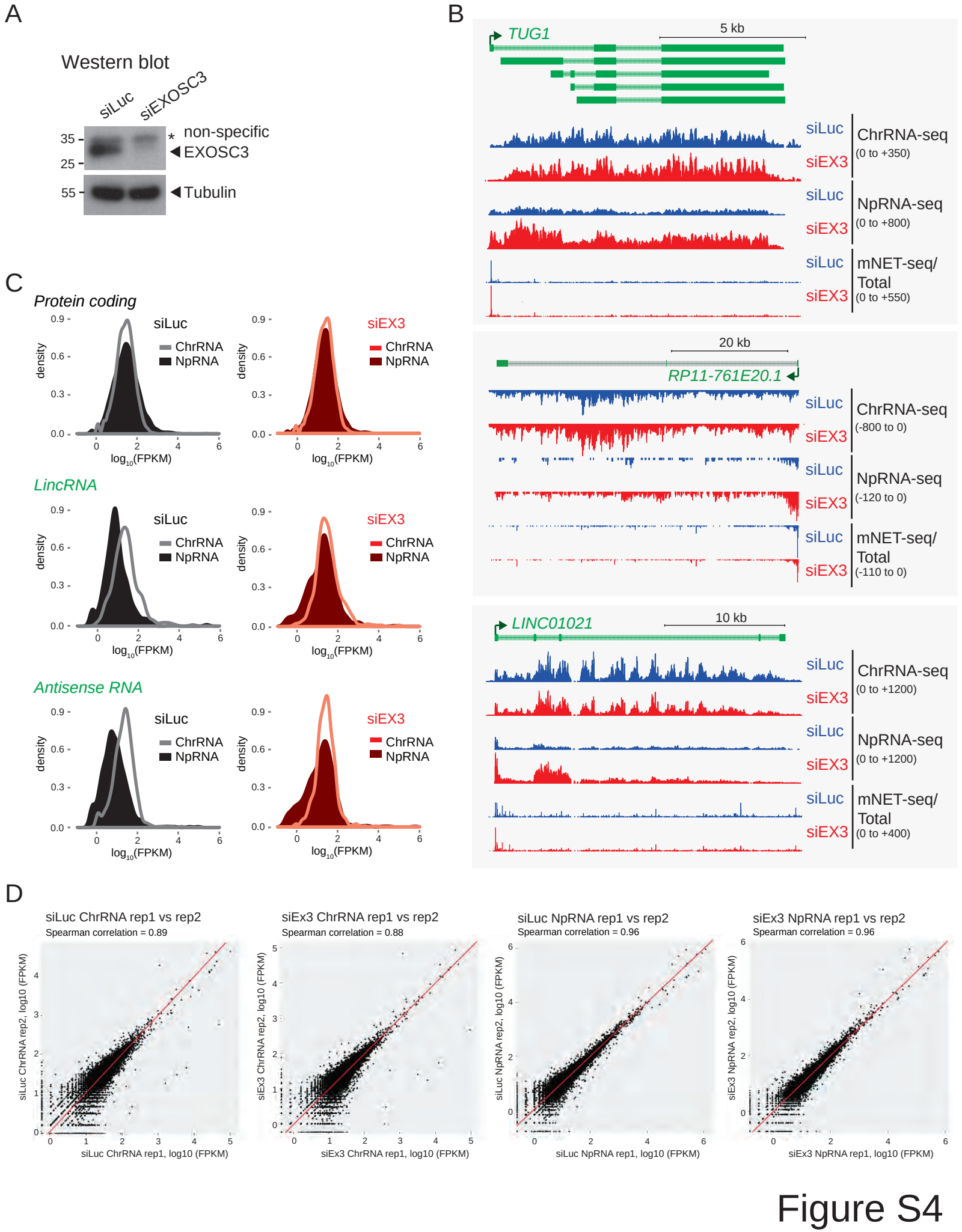

A
Western blot
siLuc
siEXOSC3
35
25
55
non-specific
EXOSC3
Tubulin
B
TUG1
5 kb
siLuc
siEX3
ChrRNA-seq
(0 to +350)
NpRNA-seq
(0 to +800)
mNET-seq/
Total
(0 to +550)
20 kb
RP11-761E20.1
(-800 to 0)
(-120 to 0)
(-110 to 0)
LINC01021
10 kb
(0 to +1200)
(0 to +1200)
(0 to +400)
C
Protein coding
LincRNA
Antisense RNA
density
ChrRNA
NpRNA
log10(FPKM)
D
siLuc ChrRNA rep1 vs rep2
Spearman correlation = 0.89
siEx3 ChrRNA rep1 vs rep2
Spearman correlation = 0.88
siLuc NpRNA rep1 vs rep2
Spearman correlation = 0.96
siEx3 NpRNA rep1 vs rep2
Spearman correlation = 0.96
siLuc ChrRNA rep2, log10 (FPKM)
siLuc ChrRNA rep1, log10 (FPKM)
siEx3 ChrRNA rep2, log10 (FPKM)
siEx3 ChrRNA rep1, log10 (FPKM)
siLuc NpRNA rep2, log10 (FPKM)
siLuc NpRNA rep1, log10 (FPKM)
siEx3 NpRNA rep2, log10 (FPKM)
siEx3 NpRNA rep1, log10 (FPKM)

Figure S4

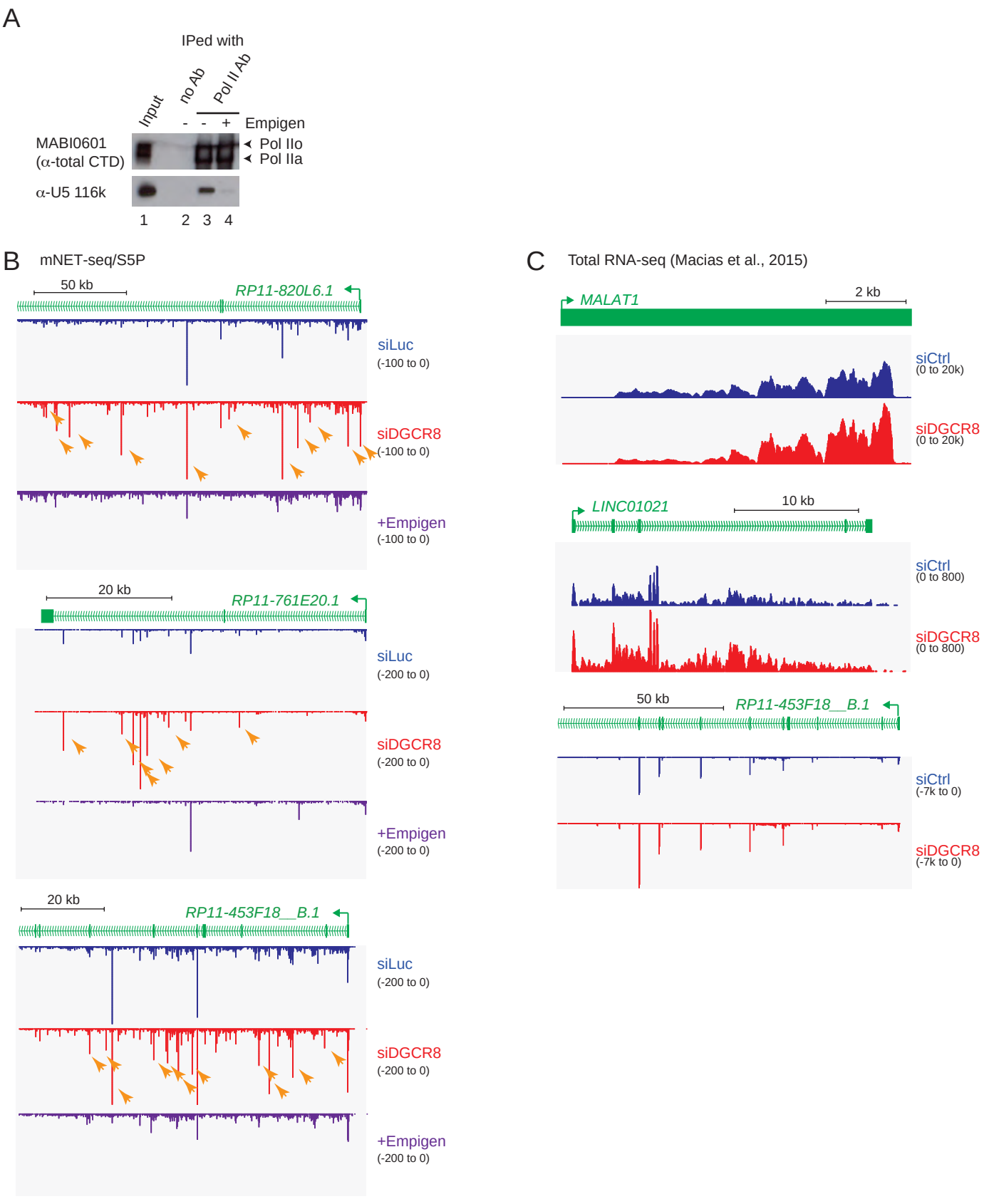

A
IPed with
Input
no Ab
Pol II Ab
- - + Empigen
MABI0601
(α-total CTD)
Pol IIo
Pol IIa
α-U5 116k
1 2 3 4
B
mNET-seq/S5P
50 kb
RP11-820L6.1
siLuc
(-100 to 0)
siDGCR8
(-100 to 0)
+Empigen
(-100 to 0)
20 kb
RP11-761E20.1
siLuc
(-200 to 0)
siDGCR8
(-200 to 0)
+Empigen
(-200 to 0)
20 kb
RP11-453F18__B.1
siLuc
(-200 to 0)
siDGCR8
(-200 to 0)
+Empigen
(-200 to 0)
C
Total RNA-seq (Macias et al., 2015)
MALAT1
2 kb
siCtrl
(0 to 20k)
siDGCR8
(0 to 20k)
LINC01021
10 kb
siCtrl
(0 to 800)
siDGCR8
(0 to 800)
50 kb
RP11-453F18__B.1
siCtrl
(-7k to 0)
siDGCR8
(-7k to 0)

Figure S5

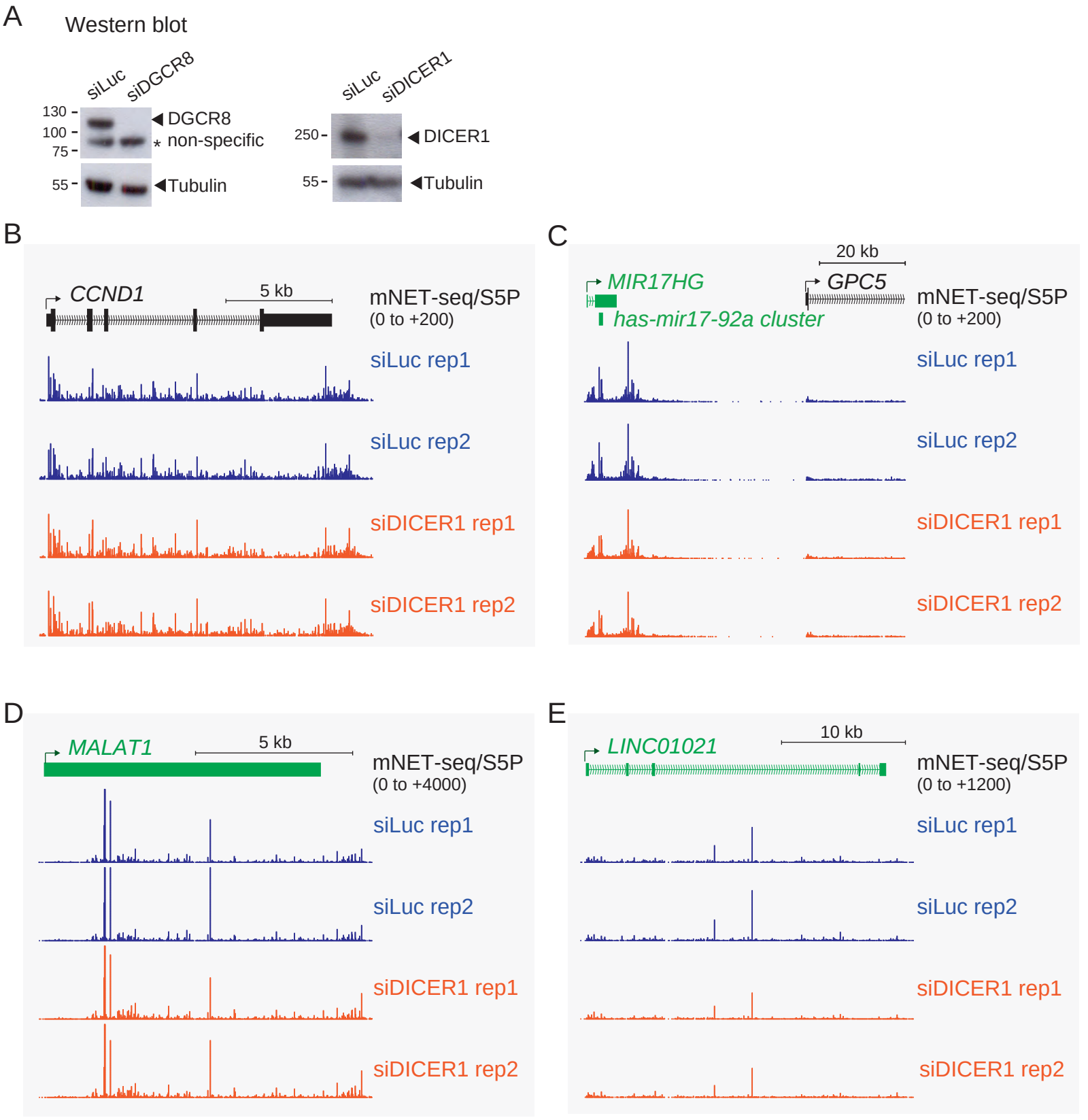

A
Western blot
siLuc
siDGCR8
130
100
75
DGCR8
* non-specific
55
Tubulin
siLuc
siDICER1
250
DICER1
55
Tubulin
B
CCND1
5 kb
mNET-seq/S5P
(0 to +200)
siLuc rep1
siLuc rep2
siDICER1 rep1
siDICER1 rep2
C
20 kb
MIR17HG
GPC5
has-mir17-92a cluster
mNET-seq/S5P
(0 to +200)
siLuc rep1
siLuc rep2
siDICER1 rep1
siDICER1 rep2
D
MALAT1
5 kb
mNET-seq/S5P
(0 to +4000)
siLuc rep1
siLuc rep2
siDICER1 rep1
siDICER1 rep2
E
LINC01021
10 kb
mNET-seq/S5P
(0 to +1200)
siLuc rep1
siLuc rep2
siDICER1 rep1
siDICER1 rep2

Figure S6

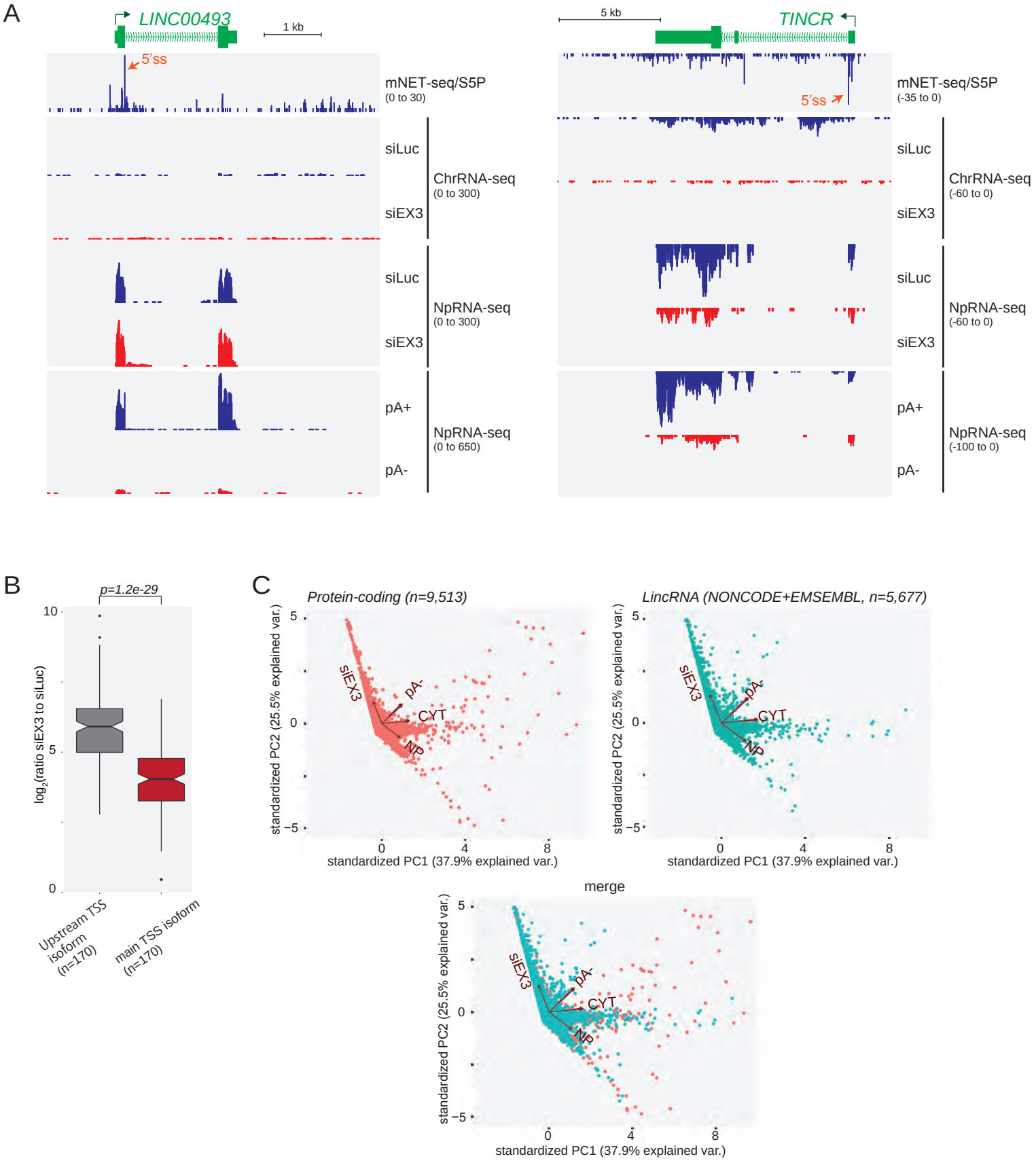

A
LINC00493
1 kb
5'ss
mNET-seq/S5P
(0 to 30)
siLuc
siEX3
ChrRNA-seq
(0 to 300)
NpRNA-seq
(0 to 300)
pA+
pA-
NpRNA-seq
(0 to 650)
5 kb
TINCR
mNET-seq/S5P
(-35 to 0)
ChrRNA-seq
(-60 to 0)
NpRNA-seq
(-60 to 0)
NpRNA-seq
(-100 to 0)
B
p=1.2e-29
log2(ratio siEX3 to siLuc)
Upstream TSS isoform (n=170)
main TSS isoform (n=170)
C
Protein-coding (n=9,513)
LincRNA (NONCODE+EMSEMBL, n=5,677)
merge
standardized PC2 (25.5% explained var.)
standardized PC1 (37.9% explained var.)
siEX3
pA-
CYT
NP

Figure S7

ENSEMBL Gene Biotype	Simplified category
IG_C_gene	coding
IG_D_gene	coding
IG_J_gene	coding
IG_LV_gene	coding
IG_V_gene	coding
TR_C_gene	coding
TR_J_gene	coding
TR_V_gene	coding
TR_D_gene	coding
IG_C_pseudogene	pseudogene
IG_J_pseudogene	pseudogene
IG_V_pseudogene	pseudogene
TR_V_pseudogene	pseudogene
TR_J_pseudogene	pseudogene
Mt_rRNA	ncRNA
Mt_tRNA	ncRNA
miRNA	ncRNA
misc_RNA	ncRNA
rRNA	ncRNA
snRNA	ncRNA
snoRNA	ncRNA
ribozyme	ncRNA
sRNA	ncRNA
scaRNA	ncRNA
Mt_tRNA_pseudogene	pseudogene
tRNA_pseudogene	pseudogene
snoRNA_pseudogene	pseudogene
snRNA_pseudogene	pseudogene
scRNA_pseudogene	pseudogene
rRNA_pseudogene	pseudogene
misc_RNA_pseudogene	pseudogene
miRNA_pseudogene	pseudogene
TEC	predicted
nonsense_mediated_decay	discarded
non_stop_decay	pseudogene
retained_intron	discarded
protein_coding	coding
processed_transcript	lincRNA
non_coding	ncRNA
ambiguous_orf	predicted
sense_intronic	discarded
sense_overlapping	ncRNA
antisense	antisense
known_ncrna	ncRNA

pseudogene	pseudogene
processed_pseudogene	pseudogene
polymorphic_pseudogene	pseudogene
retrotransposed	pseudogene
transcribed_processed_pseudogene	pseudogene
transcribed_unprocessed_pseudogene	pseudogene
transcribed_unitary_pseudogene	pseudogene
translated_unprocessed_pseudogene	pseudogene
unitary_pseudogene	pseudogene
unprocessed_pseudogene	pseudogene
artifact	discarded
lincRNA	lincRNA
macro_lncRNA	lncrna
LRG_gene	discarded
3prime_overlapping_ncrna	ncRNA
disrupted_domain	discarded
vaultRNA	ncRNA

Table S1

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY FIGURE AND TABLE LEGENDS

Figure S1 (related to Figure 1)

(A) Meta-analysis of different transcript categories (as indicated) using mNET-seq analysis with indicated Pol II antibodies centered over either TSS or TES except for eRNA category which is centered over sense and antisense TSS.

(B) Quantitation (box plots) of TSS Escaping index and TES Termination index for protein coding versus lincRNA meta-analysis. Replicated data is presented in all cases

Figure S2 (related to Figure 2)

(A) mNET-seq/S5P profiles for indicated lincRNAs with transcript read profiles aligned to above gene maps. HeLa cells were treated with Pla-B (in DMSO) or mock treated with DMSO. Arrow denotes promoter direction. Yellow arrows denote high 5'ss PLa-B sensitive mNET-seq signals.

(B) Tabulation of splicing event % (total numbers indicated in bars) for coding or lincRNA in total or pA+/- nuclear RNA. Duplicate data is presented.

(C) Splicing index derived from pA- or total NpRNA-seq. Duplicate data is presented.

(D) Reproducibility is tested via scatter plots for FPKM values (first 500 nt) in pA+ and pA- NpRNA-seq duplicates. Spearman coefficient confirms a high positive correlation.

Figure S3 (related to Figure 3)

(A) Western blot showing degree of CPSF73 depletion following siCPSF73 but not siLuc treatments of Hela cells. Blots with anti CPSF73 and control anti tubulin are shown.

(B) mNET-seq/T4P versus chromatin RNA-seq profiles are shown for four protein-coding genes as indicated from chromatin extracted siLuc or siCPSF73 treated HeLa cells.

(C) As for (B) but for 4 lincRNA TUs. For *lincPZP*, antisense the protein coding gene PZP (not expressed in HeLa cells) a blow up of boxed region is also shown.

(D) Meta-analysis of termination region as shown in Figure 3C but using duplicate data.

(E) mNET-seq/S2P and PAR-CLIP data for *MALAT1* 3' end region.

Figure S4 (related to Figure 4)

(A) Western blot showing degree of exosome component EXOSC3 depletion following siEXOSC3 but not siLuc treatments of Hela cells. Blots with anti EXOSC3 and control anti tubulin are shown. * denotes nonspecific band.

(B) Profile comparison between ChrRNA-seq, NpRNA-seq and mNET-seq profiles with and without EXOSC3 depletion for three lincRNA TUs as indicated.

(C) Density plots of FPKM levels for indicated RNA types with or without EXOSC3 depletion comparing protein-coding, lincRNA and antisense RNA TUs.

(D) Scatter plots of FPKM values (first 500 nt) showing high reproducibility (Spearman correlation coefficient indicated) of ChrRNA-seq and NpRNA-seq with mock or siEXOSC3 depleted HeLa cells.

Figure S5 (related to Figure 5 and 6)

(A) Co-immunoprecipitation of spliceosomal U5 116k protein with Pol II blocked by empigen treatment.

(B) mNET-seq/S5P profiles for indicated lincRNA TUs from HeLa cells treated with siLuc (control), siDGCR8 (see Figure S6A) or micrococcal nuclease digested chromatin pretreated with empigen (see Supplementary Methods). Positions of dominant multiple DGCR8

sensitive peaks are indicated by yellow arrows.

(C) Total RNA-seq profiles for indicated lincRNA with or without DGCR8 depletion by siRNA treatment.

Figure S6 (related to Figure 6)

(A) Western blots showing degree of DGCR8 and Dicer depletion following siDGCR8 or siDICER1 but not siLuc treatments of Hela cells. Blots with anti DGCR8 and anti DICER1 versus control anti tubulin are shown. * denotes nonspecific band.

(B-E) mNET-seq/S5P profiles for indicated lincRNA TUs with or without DICER1 depletion. For *MIR17HG* (C) downstream tandem protein-coding gene *GPC5* is also presented. Note that all duplicated profiles are shown.

Figure S7 (related to Figure 7)

(A) Specific examples of protein coding-like lincRNA showing profiles of mNET-seq/S5P, ChrRNA-seq and NpRNA-seq with or without EXOSC3 depletion and NpRNA-seq pA- and pA+ selected. Position of 5'SS mNET-seq peak is indicated by arrow.

(B) Protein coding genes, which appear EXOC3 sensitive show weakened EXOC3 sensitivity if a more dominant, further downstream TSS is considered.

(C) Principal component analysis applied to protein coding and lincRNA TUs shown separately and merged as in Figure 7. However in this case the NONCODE data base (Xie et al., 2014) is used as the main source of lincRNA and protein-coding TUs.

Table S1 (related to Experimental Procedures)

Simplified categories used instead of the original ENSEMBL gene biotype to guide transcription unit annotation.

Table S2 (related to Experimental Procedures)

Dataset 1 is a list of all genomic regions used for the analyses in Figures 1-4. Many overlapping genes and genes with low expression in HeLa cells were excluded (Supplemental Experimental Procedures).

Dataset 2 is a list of all lincRNA genes used for PCA from Dataset 1. Descriptors were computed based on RNA-Seq data from Mayer et al. 2015 (GEO:GSE61332). LincRNA genes were excluded if the chromatin signal from Mayer et al. 2015 was 0. Data is sorted in decreasing order of PC1.

Dataset 3 is as Dataset 2 but of antisense genes.

Dataset 4 is a list of all lincRNA genes used for PCA ENSEMBL and NONCODE. Less stringent criteria for allowing overlapping transcription units were used. Descriptors, inclusion of lincRNA genes and sorting is equivalent to Dataset 2.

Dataset 5 is equivalent of Dataset 2 but of protein coding genes.

Dataset 6 is equivalent of Dataset 4 but of protein coding genes.

Dataset 7 is a list of coding genes from the PCA with multiple TSSs, which behave similar to lincRNA (PC1<0, PC2>1). TSS1 refers to the upstream TSS used in the PCA. TSS2 refers to a downstream TSS, which was not used in the analysis. Green highlights the genes where the downstream TSS is mainly used according to higher Seq signal on the chromatin. Purple highlights the genes where the upstream TSS may be mainly used according to chromatin Seq signal. This data is based on our RNA Seq data from chromatin (Chr), nucleoplasm (NP) siLuc and NP siEX3.

Dataset 8 is equivalent to Dataset 7, but of coding genes with one annotated TSS. Comments indicate which genes may be overlapping a PROMPT giving rise to the observed effect. Comments also indicate where there may be a misannotation in the database or where UCSC and ENSEMBL entries don't match.

Dataset 9 is equivalent to Dataset 7 but with inclusion of lincRNA from ENSEMBL and NONCODE (corresponding to the lincRNA from Dataset 4).

Dataset 10 is equivalent to Dataset 8 but with inclusion of lincRNA from ENSEMBL and NONCODE (corresponding to the lincRNA from Dataset 4)

Methods

siRNA transfection

SMARTpool siRNA against human CPSF73 (CPSF3), EXOSC3 and Dicer were purchased from Thermo scientific. DGCR8 siRNA is previously described (Dhir et al., 2015). These siRNA (final concentration 30 nM) were transfected into HeLa cells using Lipofectamine RNAiMAX reagent (Life technologies) according to the manual and incubated for between 60 and 72 hr.

Antibodies

Pol II antibodies CMA601, CMA602 and CMA603 were purchased from MBL international (Nojima et al., 2016). Pol II antibodies 4E12 (phospho Ser7) and 6D7 (phospho Thr4) were purchased from active motif. Pol II antibody 3D12 (phospho Tyr1) was purchased from Millipore. Pol II antibody 8WG16, Dicer and Drosha antibodies were purchased from Abcam. DGCR8 and Tubulin antibodies were purchased from Novus bio and Sigma, respectively. CPSF73, SNRP116 (U5 116k) and EXOC3 antibodies were purchased from Bethyl.

mNET-seq method

The detailed protocol was as previously described (Nojima et al., 2016).

Fractionated RNA-seq methods

ChrRNA-seq and NpRNA-seq method including the library preparation method were as previously described (Nojima et al., 2015).

pA+/- RNA selection

pA + and pA- RNA were separated from 10 μg HeLa nucleoplasm RNA using Dynabeads mRNA purification kit (Thermo Fisher) according to the manual. Ribosomal RNA were

depleted only from pA- RNA with Ribo-Zero rRNA depletion kit (Illumina). 500 ng of the isolated RNA were used to prepare the library.

Cell culture and in vivo splicing inhibition

Cell culture, siRNA transfection, in vivo splicing inhibition with Pla-B were as previously described (Nojima et al., 2015).

Empigen treatment in mNET-seq

Chromatin was isolated and digested with microccocal nuclease as previously described (Nojima et al., 2016; Nojima et al., 2015). EGTA (25 mM) was added to inactivate microccocal nuclease and digested chromatin was centrifuged at 13,000 rpm for 10 min to collect supernatant as soluble fraction. The soluble fraction was ten times diluted with NET-2 buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl and 0.05% NP-40) containing 1% empigen BB (Sigma) and added to Pol II antibody-conjugated beads for 1 hour immunoprecipitation. The beads were washed six times with NET-2 buffer containing 1% empigen BB. The rest of mNET-seq protocols were as previously described (Nojima et al., 2016).

Transcription unit annotation

hg19/GRCh37 was used as a reference genome. The matching ENSEMBL gene annotation (GRCh37.75) was used to extract transcription units (Flicek et al., 2014). This annotation was further complemented by transcription units annotated in NONCODE v4 (Xie et al., 2014), UCSC tRNA (Lowe and Eddy, 1997), PROMPT (Ntini et al., 2013) and eRNA (Andersson et al., 2014).

PROMPTs were extracted for the PROMPT genes in HeLa cells as previously annotated (Ntini et al., 2013). PROMPT 5’ ends were defined as the position within [TSS-3000, TSS+1000] with the maximal CAGE signal within EXOSC3 (hRrp40) KD environment. The PROMPT 3’ end was defined as the maximal 3’ TAG signal in EXOSC3 KD within the same region, provided it was downstream of the PROMPT 5’ end.

eRNA were taken from the PrESSTo database, part of the FANTOM5 project. Only ubiquitously expressed eRNA across cells and organs were used as they were most likely to also be present in HeLa cells (Andersson et al., 2014).

The ENSEMBL gene biotype annotation as defined by the second column in the gtf file was simplified into reduced categories (Table S1).

All Genes were taken from the most 5' transcription start site (TSS) and most 3' transcription end site (TES). Overlapping exons were merged to only include the most 5' and most 3' exon borders. Each gene was considered as not expressed (silent) if the interval [TSS, min (TSS+500,TES)] lacked sufficient signal in the chromatin fraction as well as the nucleoplasm fraction (defined by total number of reads mapping, fpkm and maximal signal within the interval).

Once silent regions were excluded, a further number of overlapping features on the same strand were also excluded under the following conditions: ENSEMBL lincRNA were excluded from further analysis if they overlapped coding genes (1 kbp extended at 5' end and 3 kbp extended at 3' end) or an antisense biotype labeled gene. Coding genes were excluded if they overlapped a lincRNA. lincRNA obtained from the NONCODE database were excluded if they overlapped an ENSEMBL lincRNA, eRNA, antisense RNA, ncRNA, another NONCODE annotation, coding gene, tRNA, pseudogene or PROMPT. eRNA were excluded if they overlapped an ENSEMBL lincRNA, pseudogene, ncRNA, tRNA, antisense RNA, PROMPT or coding gene. ncRNA were excluded if they overlapped a lincRNA, eRNA, another ncRNA, antisense RNA, tRNA, pseudogene, coding gene or PROMPT. tRNA were excluded if they overlapped an ENSEMBL lincRNA, eRNA, ncRNA, NONCODE lincRNA, tRNA, PROMPT, pseudogene, coding gene or antisense RNA. Antisense RNA were excluded if they overlapped an ENSEMBL lincRNA, eRNA, pseudogene, ncRNA, tRNA or PROMPT. PROMPTs were excluded if they overlapped an ENSEMBL lincRNA or a tRNA.

For the final list of used annotated regions, ENSEMBL lincRNA and NONCODE lincRNA were manually cross-checked with the siLuc chromomatin fraction RNA-seq to ensure that they are adequately expressed and do not fall into regions, which are likely to be read-through

from neighboring genes. This resulted in the selection of 285 lincRNA from both databases. TUs on chrY were not taken into consideration.

The above procedure generated a final feature annotation file, which was used for the majority of the analysis (Table S2, Dataset 1).

Data processing and presentation

mNET-Seq data and chromatin RNA-seq was processed as follows: mNET-Seq adapters were trimmed with Cutadapt v. 1.8.3 ((Martin, 2011), https://cutadapt. readthedocs.io/en/stable/) in paired end mode with the following parameters: -A GATCGTCGGACTGTAGAACTCTGAAC -a TGGAATTCTCGGGTGCCAAGG --minimum-length 10. Obtained sequences were mapped to the human hg19 reference sequence with Tophat v. 2.0.13 ((Kim et al., 2013), https://ccb.jhu.edu/software/tophat/) and the parameters -g 1 -r 3000 --no-coverage-search. Only properly paired and mapped reads were used for subsequent analysis (samflags 0x63, 0x93, 0x53, 0xA3), which were extracted with SAMtools v. 1.2 ((Li et al., 2009), http://www.htslib.org/). For mNET-Seq profiles only the most 3' nucleotide of the second read was used with the strandedness of the first read. Data was visualized with Bedtools v. 2.23.0 (genomeCoverageBed) ((Quinlan and Hall, 2010), http://www.htslib.org/). Trackhubs in the UCSC browser were created by employing the UCSC bedGraphToBigWig tool (Kent et al., 2002).

Metagene profiles

All profiles have averaged sense and antisense coverage around the indicated 5' or 3' sites of genomic features, except eRNA genes, where center of the annotated eRNA gene coordinates is taken as a reference point. As eRNA genes are bi-directional and therefore not associated with sense/antisense direction, the coverage-values for enhancers are shown on the plus and minus strand. For each metagene plot ≤1% of most extreme coverage-values in each bp-location were trimmed before averaging. For smoothing purposes data was binned into 10 bp

bins, error bars indicate the SEM across each bin. On rare occasions where an overlapping ncRNA caused large noise in the metagene profile, the corresponding TU was excluded from the analysis resulting in a small variation in considered number of lincRNA. Graphs were created using ggplot2 (http://www.ggplot2.org/) in R (http://www.R-project.org/).

Heatmaps

CTD profile heatmaps at the TSS and TES (Figure 1B) were generated by binning coding genes and lincRNA genes of length greater than 1000 nt into 100 bins. Profiles are shown with an additional 20 bins 5' of the TSS and 3' of the TES. Genes were ordered in descending order according to signal throughout the binned vector in each CTD phosphorylation isoform. Splicing associated S5P signal heatmaps at the 5'ss were generated by extracting the annotated exons for coding genes and lincRNA. Overlapping exons were grouped into one with the most extreme 5' and 3' boundaries. Only non-terminal exons were considered for the signal description of the 5'ss. Heatmaps were generated for bins of 10 bps within the [5'ss – 400 bp, 5'ss +400 bp] region. Genes were sorted according to the maximal signal of the bin with coordinates [5'ss -9bp, 5'ss] in the control sample and the same sorting was applied to the Pla-B treated sample. The S5P signal at the 5'ss was considered a peak if it was the maximal signal in the [5'ss -100bp, 5'ss +100bp] interval, allowing the computation of the proportion of peaks in coding vs. lincRNA exons (Figure 2D bottom). Heatmaps were created using the MATLAB R2015b (The MathWorks, Inc., Natick, MA, US) image function.

Escaping and termination indices

Escaping and Termination indices were computed on a single nucleotide basis from the mNET-seq profiles. In detail, the last nucleotide of the second read with the strandedness of the first read was formatted into bed format with the associated read names for all CTD phospho isoforms. These bedfiles were overlapped with [TSS, TSS+500], [TES, TES+2000] and GB (referring to gene body, defined as the middle 50% of the interval [TSS+500, TES]) using bedtools intersect with –s –c parameters. All counts were normalised to the length of

the corresponding region. The Escaping index EI was then defined as

$$EI = \log_2 \left(\frac{\frac{[TSS, TSS + 500nt]_{counts}}{500}}{\frac{GB_{counts}}{length_{GB}}} \right)$$

and the termination index TI was then defined as

$$TI = \log_2 \left(\frac{\frac{[TES, TES + 2000nt]_{counts}}{2000}}{\frac{GB_{counts}}{length_{GB}}} \right)$$

Splicing index

The Splicing index was computed from the nucleoplasm, nucleoplasm pA+ and nucleoplasm pA- fractions. Only exons, which are not annotated as the first exon were used and only the most extreme boundaries of overlapping exons were considered to compute this index. Spliced reads were extracted from sam files of all properly paired, properly mapped reads based on the CIGAR string containing the 'N'-label and mapped to the corresponding 3' and 5' splice sites. The number of spliced reads mapping to the 3'ss of the used exons were computed and defined as splicing events. Reads spanning the intron-exon junction were computed. In detail, the 2 nucleotides around the 3'SS were extracted (last nucleotide in the intron and first nucleotide in the exon). Reads overlapping these two nucleotides were computed with the bedtools intersect tool using the –f 1 option, i.e. enforcing the read to overlap both nucleotides. Only reads where the first mate is on the same strand as the 3'SS were considered. The splicing index was computed only for 3'SS which have non-zero levels of spliced reads at the site and with non-zero levels of reads spanning the intron-exon junction. The splicing index was defined as SI = (reads spliced at 3'SS) / (reads spanning 3'SS Intron-Exon junction) – hence the larger the SI the more efficient the splicing.

Ratio of nucleoplasm RNA and chromatin RNA and FPKM values

The ratio of nucleoplasm and chromatin RNA for siLuc and siEX3 data was computed with the R DESeq package (Anders and Huber, 2010). For each TU, the region [TSS, TSS+500] was overlapped with the RNA-seq reads using the inbuilt function summarizeOverlaps. FPKM values were computed with the same package using the estimateSizeFactors and fpkm functions.

pA ratio

The strand specific overlap of pA+ and pA- fragments with whole length coding and lincRNA genes was counted using bedtools intersect –c. The log2 ratio was taken of $pA-_{counts}/pA+_{counts}$ for each TU.

Principal Component Analysis

Principal Component Analysis was performed on the following ratios: chromatin siEX3/siLuc, nucleoplasmic pA-/pA+, nucleoplasm RNA/chromatin RNA, cytoplasm RNA/chromatin RNA (Mayer et al., 2015). Principal components were computed with the R prcomp function and the data was centered and scaled to zero mean and unit variance. Prcomp was applied to data of protein coding genes and the principal component rotation was subsequently applied to the data of lincRNA and antisense RNA. To identify lincRNA most similar to coding genes the cutoffs PC1>0 and PC2<1 were employed. Similarly for identification of coding genes most similar to lincRNA PC1<0 and PC2>1 were used. PCA was visualised with ggbiplot (http://github.com/vqv/ggbiplot).

P-values , significance tests and boxplots

P-values for Figure 2D (bottom) were computed with a Fisher Exact Test. All other p-values were computed by a Wilcoxon rank sum test in R. For Figure 4 we employed the paired Wilcoxon signed rank test. All boxplots were created with ggplot2 in R.

Scatterplots

Scatterplots were generated in R based on the FPKM values for the first 500nt of all analysed TUs. The Spearman correlation coefficient was computed in R. The identity line is indicated.

SUPPLEMENTARY REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Dhir, A., Dhir, S., Proudfoot, N.J., and Jopling, C.L. (2015). Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. Nat Struct Mol Biol *22*, 319-327.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.*, et al.* (2014). Ensembl 2014. Nucleic acids research *42*, D749-755.

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

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research *25*, 955-964.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541-554.

Nojima, T., Gomes, T., Carmo-Fonseca, M., and Proudfoot, N.J. (2016). Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. Nature protocols *11*, 413-428.

Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526-540.

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

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic acids research *42*, D98-103.