Regulation of cytoplasmic mRNA level by chromatin retention.

Callum Henfrey[1], Shona Murphy[1] and Michael Tellier[1,*]

[1] Sir William Dunn School of Pathology, University of Oxford, Oxford OX1 3RE, UK
* To whom correspondence should be addressed. Tel: +44 1865275583; Email: michael.tellier@path.ox.ac.uk

ABSTRACT

Transcription and co-transcriptional processes, including pre-mRNA splicing and mRNA cleavage and polyadenylation, regulate the production of a mature mRNA. The carboxyl terminal domain (CTD) of RNA polymerase (pol) II, which comprises 52 repeats of the Tyr1Ser2Pro3Thr4Ser5Pro6Ser7 peptide, is involved in the coordination of transcription with co-transcriptional processes. The pol II CTD is dynamically modified by protein phosphorylation, which regulates recruitment of transcription and co-transcriptional factors. We have investigated whether cytoplasmic levels of mature mRNA from intron-containing protein-coding genes can be inferred from RNA processing efficiency, CTD phosphorylation, and/or association of proteins complexes regulating RNA production. Surprisingly, we found that genes associated with efficient RNA processing have relatively low phosphorylation of the pol II CTD. In contrast, protein-coding genes that produce a low level of mature mRNA are associated with high pol II CTD phosphorylation, poor RNA processing, and increased chromatin retention. Unexpectedly, the transcripts from this subset of genes are not actively degraded by the RNA exosome. RNA exosome-regulated genes are instead characterised by Tyr1 hyperphosphorylation and Integrator recruitment around the poly(A) site. Our results indicate that in addition to the transcription level, CTD phosphorylation and RNA processing efficiency play important roles in the regulation of chromatin retention of transcripts.

INTRODUCTION

Transcription of human protein coding genes by RNA polymerase (pol) II is a highly complex process requiring the coordination of multiple proteins. In addition to the transcription cycle, composed of transcription initiation, pol II pausing and release, transcription elongation, and transcription termination, co-transcriptional processes, including capping, splicing, cleavage and polyadenylation, and co-transcriptional loading of mRNA export factors are required for the production of mature mRNA (1,2). A key element regulating the crosstalk between transcription and co-transcriptional processes is the carboxyl terminal domain (CTD) of the large subunit of pol II, which comprises 52 repeats of the heptapeptide Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7. The pol II CTD can be modified by several post-translational modifications (PTMs), including protein phosphorylation, methylation, acetylation, and proline isomerization (3,4). Phosphorylation of the pol II CTD is one of the major PTMs and can occur on five residues, Tyr1, Ser2, Thr4, Ser5, and Ser7 (3,4). Kinases, including several cyclin-dependent kinases (CDKs), and phosphatases regulate the CTD phosphorylation pattern and level across the transcription cycle (3,4).

Phospho-Ser5 (Ser5P) and phospho-Ser2 (Ser2P) are the most studied modifications and are found at the promoter region and in the gene body/downstream of the poly(A) site, respectively (4). Ser5P is associated with the recruitment of the mRNA capping complex and pre-mRNA splicing factors while Ser2P is linked to the recruitment of elongation factors and proteins of the mRNA cleavage and polyadenylation complex (CPA). The roles of the three other residues, Tyr1, Thr4, and Ser7, are less well understood (5). On protein-coding genes, phosphorylation of Tyr1 is present at promoter and transcription termination regions. Tyr1P has been found to be higher on antisense promoters (PROMPTs) and enhancers compared to protein-coding genes (6,7) and to be increased following DNA double-strand breaks and UV irradiation (8,9). In addition, mutations of three quarters of the tyrosine residues to alanine promotes transcriptional readthrough and a loss of the Mediator

and Integrator complexes from the pol II. Phosphorylation of Thr4 is found at the 3'end of protein-coding genes, indicating a potential role in transcription termination (10,11). In addition, Thr4P has been found to be higher on the gene bodies of long non-coding (lnc)RNAs, which are known to be prone to premature transcription termination (PTT) (10,11). Mutation of Thr4 residues to alanine results in a transcription elongation defect on protein-coding genes and in a 3'end-processing defect of histone gene transcripts (10,12). Phosphorylation of Ser7 is currently the least understood but follows a similar pattern to Ser2P. The combination of Ser2P and Ser7P has been shown to be involved in the recruitment of the Integrator complex to small nuclear (sn)RNA genes (13). Mutation of Ser7 residues to alanine results in a decreased transcription of snRNA genes and 3' processing of transcripts while protein coding genes do not seem to be affected (13,14).

Modification of the pol II CTD is therefore critical for coordinating co-transcriptional processes during transcription. In turn, co-transcriptional processes can also affect transcription (2). A major example is the coupling between pre-mRNA splicing, mRNA cleavage and polyadenylation, and transcription termination. It has been shown via long-read sequencing approaches that protein-coding genes transcripts that are poorly processed are associated with pol II transcriptional readthrough, likely due to a failure to recognize the poly(A) site (15,16). Additionally, transcriptional readthrough can also be promoted by knockdown of CPA factors, cellular stresses, or viral infections (17-19).

While important steps in the regulation of expression of protein-coding genes occur at the 5' end of genes, including transcription initiation and pol II pause release, it is becoming increasingly clear that premature transcription termination (PTT) is a major regulator of gene expression. PTT can happen across the whole gene unit, from pol II pausing site to poly(A) site, depending on the mechanism used. PTT can be mediated by mRNA decapping followed by Xrn2 degradation (20), the Integrator complex (21-24), co-transcriptional recruitment of the RNA exosome (25), U1 telescripting and intronic poly(A) site usage (26-30), and at the poly(A)-associated checkpoint (31-34).

To better understand how transcription and co-transcriptional processes regulate the production of mature mRNAs from intron-containing protein-coding genes (termed protein-coding genes in the rest of the manuscript), we took advantage of genome-wide data available for HeLa cells. We found that protein-coding genes producing a high level of mature mRNA have relatively low pol II CTD phosphorylation levels. In contrast, pol II is hyperphosphorylated on protein-coding genes that produce relatively low levels of mature mRNA. Interestingly, the reduced production of mature mRNAs from these protein-coding genes is not due to degradation by the nuclear RNA exosome but rather poor RNA processing efficiency coupled to chromatin retention of the transcripts.

Our results indicate that the regulation of RNA processing efficiency plays an important in controlling gene expression through chromatin retention of transcripts. This regulation of chromatin retention of transcripts through RNA processing efficiency is shared between lncRNAs and protein-coding genes, indicating a general mechanism regulating the level of mature mRNA.

MATERIAL AND METHODS

Genome-wide datasets

The genome-wide data used in this study are summarized in Table 1.

RNA-seq and POINT-seq analysis

Chromatin, nucleoplasm, and cytoplasmic RNA-seq were analysed as previously described (35). Briefly, adapters were trimmed with Cutadapt version 1.18 (36) in paired-end mode with the following options: --minimum-length 10 -q 15,10 -j 16 -A GATCGTCGGACTGTAGAACTCTGAAC -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. The remaining rRNA reads were removed by mapping the trimmed reads to the rRNA genes defined in the human ribosomal DNA complete repeating unit (GenBank: U13369.1) with STAR version 2.7.3a (37) and the parameters --runThreadN 16 --readFilesCommand gunzip -c -k --outReadsUnmapped Fastx --limitBAMsortRAM 20000000000 --outSAMtype BAM SortedByCoordinate. The unmapped reads were mapped to the human RNU2 gene or to the GRCh38.p13 reference sequence with STAR version 2.7.3a and the parameters: --runThreadN 16 --readFilesCommand gunzip -c -k --limitBAMsortRAM 20000000000 --outSAMtype BAM SortedByCoordinate. SAMtools version 1.9 (38) was used to retain the properly paired and mapped reads (-f 3) and to create strand-specific BAM files. FPKM-normalized bigwig files were created with deepTools2 version 3.4.2 (39) bamCoverage tool with the parameters -bs 10 -p max –normalizeUsing RPKM.

mNET-seq analysis

Adapters were trimmed with Cutadapt version 1.18 in paired-end mode with the following options: --minimum-length 10 -q 15,10 -j 16 –A GATCGTCGGACTGTAGAACTCTGAAC –a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. Trimmed reads were mapped to the human GRCh38.p13 reference sequence with STAR version 2.7.3a and the parameters: --runThreadN 16 --readFilesCommand gunzip -c -k --limitBAMsortRAM 20000000000 --outSAMtype BAM SortedByCoordinate. SAMtools version 1.9 was used to retain the properly paired and mapped reads (-f 3). A custom python script (40) was used to obtain the 3' nucleotide of the second read and the strandedness of the first read. Strand-specific bam files were generated with SAMtools. FPKM-normalized bigwig files were created with deepTools2 bamCoverage tool with the parameters -bs 1 -p max –normalizeUsing RPKM.

ChIP-seq and mNuc-seq analysis

Adapters were trimmed with Cutadapt version 1.18 in paired-end mode with the following options: --minimum-length 10 -q 15,10 -j 16 -A GATCGTCGGACTGTAGAACTCTGAAC -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. Trimmed reads were mapped to the human RNU2 gene or to the GRCh38.p13 reference sequence with STAR version 2.7.3a and the parameters: --runThreadN 16 --readFilesCommand gunzip -c -k --limitBAMsortRAM 20000000000 --outSAMtype BAM SortedByCoordinate. SAMtools version 1.9 was used to retain the properly paired and mapped

reads (-f 3) and to remove PCR duplicates. Reads mapping to the DAC Exclusion List Regions (accession: ENCSR636HFF) were removed with BEDtools version 2.29.2 (41). FPKM-normalized bigwig files were created with deepTools2 bamCoverage tool with the parameters -bs 10 -p max –e –normalizeUsing RPKM.

Proteomic analysis section

The proteome data of HeLa cells were obtained from (42). The list of genes found in the RNA-seq data was overlapped with the list of proteins found in the proteomic data to keep only the genes found in both lists.

Differential expression analysis

For differential expression analysis, the number of aligned reads per gene was obtained with STAR –quantMode GeneCounts option during the mapping of raw reads to the human genome or with HTSeq version 1.99.2 (43). The lists of differentially expressed genes were obtained with DESeq2 version 1.30.1 (44) and apeglm version 1.18.0 (45) keeping only the genes with a fold change < -2 or > 2 and an adjusted p-value below 0.05.

Human gene annotation and selection of subset of genes

Gencode V38 annotation, which is based on the hg38 version of the human genome, was used to obtain the list of all protein-coding genes. Intronless and histone genes were removed to obtain intron-containing protein-coding genes. For each gene, we kept the annotation (TSS and poly(A) site) of the highest expressed transcript isoform, which was obtained with Salmon version 1.2.1 on four HeLa chromatin RNA-seq experiments. Only transcripts that are expressed (TPM > 0) in at least three of the four biological replicates were retained. The list of similarly expressed of not chromatin retained RNA (NOT-CRR) genes, chromatin retained RNA (CRR) genes, and unchanged genes was generated through iterative random-subsampling to achieve subsets of 500 genes with the most similar expression level and distribution.

Splicing efficiency

The splicing efficiency on POINT-seq and RNA-seq was calculated by first parsing each bam file to obtain the list of spliced and unspliced reads with the awk command (awk '/^@/ || $6 ~ /N/' for spliced reads and awk '/^@/ || $6 !~ /N/' for unspliced reads). The splicing efficiency was then calculated as the number of spliced reads over total reads with BEDtools multicov –s –split. The splicing efficiency of each transcript was then normalised to the number of exons.

Transcription termination index

The transcription termination index is defined as from (40), termed readthrough index in this paper: $RTI = \log_2(GB\ /\ TES + c)$, $c = (Min(GB\ /TES) > 0\ /2)$.

Correlation heatmap

The mNET-seq heatmap was computed with deepTools2 multiBamSummary tool with the following parameters: bins –bs 10000 –distanceBetweenBins 0 –p max –e. The resulting matrix was plotted with deepTools2 plotCorrelation and the following parameters: --corMethod pearson --skipZeros –colorMap RdYlBu_r --plotNumbers.

Gene Ontology (GO) enrichment

Gene ontology (GO) enrichment analysis was performed with the Gene Ontology resource website (46,47).

Metaprofiles, boxplots, and violin plots

Metaprofiles, boxplots, and violin plots were generated with R version 4.0.5. Quantifications have been performed across the gene body, TSS to TES, except for H3K4me1 and H3K4me3 where the quantification has been done between TSS and TSS + 2.5 kb. ChIP-seq and mNuc-seq metaprofiles are shown as IP / Input signal.

Statistical tests

The statistical tests are indicated in the figures legends and were performed with R version 4.0.5.

RESULTS

A subset of expressed protein-coding genes produces a low amount of mature mRNA and protein

Previous studies have shown that a subset of long intergenic non-coding (linc)RNAs, named lincRNA-like protein-coding genes, are similar to mRNAs in that they undergo RNA processing and produce stable nuclear RNA. In contrast, a limited number of protein coding gene transcripts have been found to have features common to lincRNAs, including higher chromatin reads and RNA exosome sensitivity (11,48). However, the RNA exosome is only one of the protein complexes regulating RNA production. We have therefore more widely investigated the amount of mature mRNA produced from intron-containing protein-coding genes in relation to transcription and chromatin retention of transcripts.

We have used chromatin and nucleoplasm RNA-seq data available from HeLa cells to identify the protein-coding genes whose transcripts are enriched in the nucleoplasm compared to the chromatin (not chromatin retained RNA genes (NOT-CRR genes)) or that are enriched on the chromatin compared to the nucleoplasm (chromatin-retained RNA genes (CRR genes)) (Figure 1A). We found 4,803 NOT-CRR protein-coding genes and 6,131 protein-coding CRR genes (Figure 1B). For comparison purposes, we selected ~4,000 expressed genes from each of the NOT-CRR, CRR, and unchanged (no significant difference between nucleoplasm RNA-seq and chromatin RNA-seq) sets of genes. To confirm that the NOT-CRR and CRR subgroups are not explained by unclean cellular fractionation, we re-analysed HeLa cytoplasmic RNA-seq data, which confirms the chromatin and

nucleoplasm data (Figure 1C). To determine whether a lower nucleoplasm and cytoplasm RNA-seq signal also results in a lower protein production, we compared our RNA-seq results with a previously published re-analysis of HeLa proteome datasets (42). We could only match ~5,000 genes of the RNA-seq/proteome data but found that fewer of the CRR genes produce protein products compared to NOT-CRR genes, and CRR genes produce fewer peptides of each protein (Figure 1D).

We also reanalysed chromatin RNA-seq and total RNA-seq from Raji cells (49) (Figure 1E and F). Comparison of the genes found in HeLa and Raji shows that ~31% of the NOT-CRR genes and ~13% of CRR genes are common between both cell lines (Figure 1G). We analysed the gene ontology (GO) enrichment of the NOT-CRR and CRR genes that are common between HeLa and Raji cells (Figure 1H and Supplementary Table 1). We found GO terms related to protein localization for the NOT-CRR genes and terms associated with taste receptors and cilium for the CRR genes.

Higher levels of pol II Tyr1 and Thr4 phosphorylation are associated with poor expression and chromatin retention of transcripts

Chromatin retention of lncRNA transcripts is associated with a different pol II CTD phosphorylation pattern and with poor co-transcriptional RNA processing, including defective pre-mRNA splicing and mRNA CPA (11). We therefore investigated whether the pol II CTD phosphorylation patterns and/or levels also differ between NOT-CRR and CRR genes. We re-analysed HeLa mNET-seq data for total pol II and the different CTD phosphorylation marks, using Empigen-treated mNET-seq datasets when available as these identify bone fide mNET-seq signals without non-nascent RNA associated with the pol II (Figure 2A and Supplementary Figure 1A). The different pol II CTD phosphorylation profiles follow the expected mNET-seq pattern for the three groups of genes, with a higher signal for total pol II and CTD phosphorylation on NOT-CRR genes compared to CRR genes (Figure 2B) (4,11). Interestingly, when we ratioed each CTD phosphorylation signal to total pol II, we found that the NOT-CRR genes have generally less phosphorylated CTD than the CRR genes or unchanged genes (Figure 2C and D). In contrast, there are higher Tyr1 and Thr4 phosphorylation levels in the gene body of the CRR genes compared to the unchanged genes or the NOT-CRR genes. Interestingly, there is a negative correlation between each CTD phosphorylation mark ratioed to pol II and the nucleoplasmic RNA vs chromatin-associated RNA ratio, particularly for Tyr1P and Thr4P (Figure 2E). As the total pol II level is lower on the CRR genes compared to the NOT-CRR genes, we investigated the relationship between chromatin RNA-seq expression and gene body CTD phosphorylation ratioed to total pol II across all intron-containing protein-coding genes (Supplementary Figure 1B). Whereas there is a positive correlation between chromatin RNA-seq levels and phosphorylation of Ser2, Ser5, and Ser7, there is a negative correlation between the chromatin RNA-seq level and Tyr1 and Thr4 phosphorylation.

As Tyr1P and Thr4P level are associated with lower levels of nascent transcription, we wondered whether the higher Tyr1P and Thr4P we observed for CRR genes could be due to the generally lower expression of these genes rather than a chromatin retention-specific CTD phosphorylation pattern (Figure 2B and Supplementary Figure 1C and D). We selected 500 genes from each category that

have similar levels in chromatin RNA-seq (see Methods, Figure 3A-C, Supplementary Figure 1E and F). Re-analysis of the mNET-seq data on these three subsets of 500 genes indicates that the NOT-CRR genes have less Ser2, Thr4, Ser5, and Ser7 phosphorylation while the CRR genes still have higher Tyr1 and Thr4 phosphorylation (Figure 3D-F). Re-analysis of Raji pol II CTD datasets on the groups of Raji genes we described in Figure 1F (6,10) also indicates that there is higher Tyr1 and Thr4 phosphorylation on CRR genes while NOT-CRR genes have less Ser2 phosphorylation (Supplementary Figure 2A and B).

These findings indicate that lower pol II CTD phosphorylation is associated with higher production of mature mRNA in the nucleoplasm and cytoplasm while hyperphosphorylation of Tyr1 and Thr4 is associated with poor transcription and poor production of mature mRNAs.

Chromatin-retained RNA genes express poorly processed RNA

As pol II CTD phosphorylation is associated with co-transcriptional processes and we observed differences in CTD phosphorylation levels between NOT-CRR and CRR genes, we investigated whether RNA processing efficiency, including pre-mRNA splicing and mRNA CPA, also differs between the two groups of genes. For pre-mRNA processing, we re-analysed HeLa POINT-seq data, which captures nascent RNA transcription and co-transcriptional splicing (Figure 4A) (16). We calculated co-transcriptional splicing efficiency as the ratio of spliced reads over total reads across each intron-containing protein-coding transcript. As expected, we observed a correlation between number of exons and our measure of splicing efficiency (Supplementary Figure 3A). As the distribution of number of exons per gene differ between the CRR, NOT-CRR, and unchanged genes (Supplementary Figure 3B), we normalized the splicing efficiency of each transcript to its number of exons (Figure 4B and C). We found for the three datasets (POINT-seq, chromatin RNA-seq, and nucleoplasm RNA-seq) that the CRR gene transcripts have the lowest splicing efficiency while the NOT-CRR gene transcripts have the highest splicing efficiency. Importantly, analysis of the Raji chromatin RNA-seq and total RNA-seq gave similar results (Supplementary Figure 3C). We show as examples *PKM* and *NFIA*, a NOT-CRR and a CRR gene, respectively, with co-transcriptional splicing events indicated by a star (Figure 4D). We have also investigated whether there is a more general correlation between splicing efficiency in the POINT-seq or RNA-seq data and the production of mature mRNAs (ratio of nucleoplasm RNA-seq to chromatin RNA-seq) (Figure 4E). Interestingly, we found that co-transcriptional splicing efficiency of nascent RNA does not correlate as well as the splicing efficiency of chromatin and nucleoplasm RNA-seq with the Nucleoplasm/Chromatin ratio (R=0.25 for POINT-seq vs R=0.38-0.42 for RNA-seq) (Figure 4E).

As co-transcriptional splicing is associated with deposition of trimethylation on histone H3 lysine 36 (H3K36me3) by SETD2 (50), we re-analyzed HeLa mNuc-seq datasets for open chromatin marks, H3K4me1, H3K4me3, and H3K36me3, and for the heterochromatin mark, H3K27me3 (51) (Supplementary Figure 4). Interestingly, the CRR genes have lower H3K4me3 signals on the promoter and decreased H3K36me3 across the gene body, in line with lower co-transcriptional

splicing, while the heterochromatin mark H3K27me3 is higher in CRR genes than for unchanged and NOT-CRR genes.

As poor co-transcriptional splicing is also associated with transcriptional readthrough due to failure to recognise the poly(A) site (15,16), we analysed HeLa cell ChIP-seq of two CPA factors, CPSF73 and PCF11 (Figure 5A) (29). The NOT-CRR genes have clear peaks of CPA factors around the poly(A) site while the CRR genes do not, suggesting inefficient mRNA CPA, as shown on *PKM* and *NFIA* (Figure 5B). To investigate mRNA CPA efficiency, we calculated the read-through index (RTI), which measures pol II pausing downstream of the poly(A) site (40), from Ser2P mNET-seq and observed that CRR genes have higher transcriptional readthrough compared to the NOT-CRR and unchanged genes (Figure 5C and D). We confirmed this observation by re-analysing mNET-seq and chromatin RNA-seq data treated with siLuc or siCPSF73 (Figure 5E and F) (11,40). The CRR genes are less sensitive to the loss of CPSF73 compared to unchanged and NOT-CRR genes, which have more transcriptional readthrough following siCPSF73 treatment.

These findings indicate that CRR gene transcripts are poorly processed, which could explain their chromatin retention.

Chromatin-retained RNA gene transcripts are not generally sensitive to RNA exosome

The nuclear RNA exosome complex promotes RNA degradation, for example of pre-mRNAs with processing defects, such as those with retained introns or transcriptional readthrough (52). We wondered if the nuclear RNA exosome complex could degrade the CRR gene transcripts, which are poorly processed. We re-analysed previously published HeLa nucleoplasm RNA-seq treated with siLuc or siEXOSC3 (siEX3), a core component of the nuclear RNA exosome activity (Figure 6A) (11). We found 551, including 527 protein-coding genes were downregulated and 1,926, including 450 protein-coding genes, were up-regulated after depletion of the RNA exosome. Comparison of siEX3 upregulated genes (siEX3(+)) with CRR or NOT-CRR genes shows only a weak correlation, indicating that CRR gene transcripts are not usually degraded by the nuclear RNA exosome (Figure 6B). To determine whether siEX3(+) genes have a unique CTD phosphorylation pattern, which could mark them as targets of nuclear RNA exosome activity, we re-analysed the total pol II and CTD mNET-seq and found that the siEX3(+) genes have a higher level of Tyr1 phosphorylation and lower phosphorylation levels for the other residues, especially the serine residues (Figure 6C-E). As the siEX3(+) genes possess a different CTD phosphorylation pattern, we wondered if the knockdown of the RNA exosome results in a higher transcriptional level of the siEX3(+) genes or just in the stabilization of the nucleoplasmic mRNAs. We re-analysed the HeLa chromatin RNA-seq and total pol II mNET-seq treated with siLuc or siEX3 (11) and found an increase in chromatin RNA-seq signal for the siEX3(+) genes compared to the control genes (Figure 6F). Interestingly, the total pol II mNET-seq signal on the siEX3(+) genes is not as affected as the chromatin RNA-seq RNA level, apart from higher pol II pausing at the 5'end, which indicates a potential effect of the nuclear RNA exosome on chromatin-associated RNA rather than at the level of transcription (Figure 6G and H).

The Tyr1 residue of the pol II CTD has been recently associated with recruitment of the Integrator complex (24,49), a multiprotein complex with an RNA endonuclease activity on noncoding and coding RNA (53). As siEX3(+) genes are hyperphosphorylated on Tyr1, we wondered whether this subset of genes have higher Integrator recruitment, which could in turn cleave the RNA and promote its degradation by the RNA exosome. We therefore re-analysed previously published Integrator ChIP-seq in HeLa cells (Supplementary Figure 5A and B) (21,51). While a higher signal of Integrator complex is observed on siEX3(+) genes compared to the control genes, the pol II signal is also higher in the RNA exosome sensitive set of genes. When the Integrator signal is ratioed to the pol II signal, the Integrator signal is not higher but instead lower than on control genes across the gene body. However, we found a specific Integrator signal around the poly(A) site of the RNA exosome sensitive genes, which is conserved even after normalization to pol II level. In addition, we tested whether the Integrator complex recruitment is correlated to Tyr1 phosphorylation of the pol II CTD (Supplementary Figure 5C). We could not find any strong correlation, suggesting that Integrator recruitment is not strongly dependent on Tyr1 phosphorylation.

These findings show that the CRR gene transcripts, which are poorly processed and chromatin retained, are not degraded by the nuclear RNA exosome. However, protein-coding genes whose chromatin transcripts are targeted by the nuclear RNA exosome have a specific hyperphosphorylation of the pol II CTD on Tyr1 residues.

DISCUSSION

Production of mature mRNA requires both transcription and co/post-transcriptional RNA processing. Regulation of gene expression via the control of transcription initiation and pol II pause release are well established (54). We show here that the efficiency of RNA processing is also an important factor controlling chromatin retention of transcripts as we found a large subgroup of protein-coding genes that are transcribed but the transcripts are poorly processed and chromatin retained, which results in a low production of mature mRNAs and proteins. Interestingly, we found that this subset of CRR genes shares transcriptional and co-transcriptional similarities to lncRNA genes (11). These include higher CTD Thr4 phosphorylation, poor pre-mRNA splicing and CPA, higher transcriptional readthrough, decreased sensitivity to CPSF73 KD, and higher chromatin retention (Figure 7A). Together these observations explain the low production of mature mRNAs and indicate that the cellular mechanisms that regulate lncRNA gene expression are also used to regulate expression of some protein-coding genes. Of interest, Schlackow et al (11) also found a small subset of lncRNA genes whose transcripts are processed efficiently and not retained on the chromatin. These observations indicate that there is overlap between protein-coding genes and lncRNA genes in terms of the mechanisms operating at the transcriptional and co-transcriptional levels. The efficiency of transcription and co-transcriptional processes across transcription units, including protein-coding and non-coding genes, can be viewed as a continuum with poorly-expressed and poorly-processed lncRNAs at one end and highly-expressed and efficiently processed mRNAs at the other end with some overlap in the middle.

Most of the CRR genes produce some mature mRNA that can be translated into proteins. Surprisingly, there is only limited overlap between the CRR genes and the nuclear RNA exosome-sensitive genes. This observation indicates that nuclear degradation by the RNA exosome does not explain why most of the CRR genes are in this category. The transcripts of the CRR genes therefore seem to be retained on the chromatin due to their inefficient RNA processing. It remains to be determined why these chromatin-retained transcripts are not degraded by the nuclear RNA exosome as they share known features of sensitivity to the RNA exosome (52).

Some of the CRR genes are well transcribed but produce hardly any mature mRNAs and proteins, which begs the question: what is the cellular advantage of transcribing a protein-coding gene without producing a protein? It is possible that these genes are transcribed but poorly processed until their proteins are required, which would require only the activation of RNA processing rather than both transcription activation and RNA processing. From the GO enrichment analysis, we observed an overrepresentation of GO terms related to taste receptors and to a lower extent, cilium biology. Downregulation of the expression of taste receptors genes is a known feature of cancer cells (55). Our data indicate that the downregulation of these genes is occurring via poor co-transcriptional RNA processing and chromatin retention rather than low transcription. In addition, overlap between the HeLa and Raji datasets show a higher proportion of common genes between NOT-CRR genes (~31%) compared to CRR genes (~13%), which indicates a higher diversity in transcribed but poorly processed genes, at least between these two cancer cell lines.

A surprising observation is the lower pol II CTD phosphorylation level on the NOT-CRR genes. As pol II CTD phosphorylation is known to recruit splicing proteins and mRNA CPA factors (56), it is unexpected that lower Ser2P and Ser5P levels are associated with better RNA processing. Slow pol II elongation has been shown to result in a hyperphosphorylation of the CTD Ser2 residues at the 5' end of genes, promoting a higher dwell time at start sites and a reduced transcriptional polarity (57). In addition, hyperphosphorylation of the CTD during the M phase inhibits pol II, which contributes to mitotic gene silencing (58,59). While more work is required, these observations suggest that the level of pol II CTD phosphorylation could play an important role in controlling transcription activity and co-transcriptional processing efficiency.

Apart from CTD phospho-Ser7 available only in mNET-seq, CTD phosphorylation data is available by both ChIP-seq (Raji) (6,10) and mNET-seq (HeLa) (11,40). An interesting observation is that while Tyr1 and Thr4 phosphorylation marks look similar in both ChIP-seq and mNET-seq, Ser2 and Ser5 phosphorylation profiles markedly differ between the two techniques (Figure 2 and Supplementary Figure 2). In ChIP-seq, Ser5P is mostly found at the TSS while Ser2P is observed in the gene body and after the poly(A) site. In mNET-seq, the two CTD phosphorylation marks look similar with a peak at the TSS and after the poly(A) site. It remains unclear why the patterns of Ser2 and Ser5 phosphorylation differ between the two techniques. Some non-exclusive possibilities are that: mNET-seq has no crosslinking step, which could affect the presence of CTD interactors, particularly when the mNET-seq has been treated with Empigen which removes some pol II interactors; due to the size

selection step in mNET-seq, the first ~35 nucleotides are lost, with some pol II and CTD signals at the promoter lost; in mNET-seq, only pol II associated with nascent RNA is detected, which will represent the subset of active pol II compared to the total pol II obtained by ChIP-seq. The mNET-seq CTD profiles could therefore represent the actively transcribing pol II while the ChIP-seq CTD profiles would provide an average picture of inactive and active pol II. This could explain why Ser2P is observed at the 5'end of genes in mNET-seq but not in ChIP-seq.

Interestingly, we found that intron-containing protein-coding genes whose transcripts are degraded in the nucleoplasm (siEX3(+)) by the nuclear RNA exosome complex are specifically enriched for pol II CTD Tyr1 phosphorylation during nascent transcription (Figure 6C and D). Knockdown of EXOSC3 shows that the transcripts from these siEX3(+) genes are also degraded on the chromatin while nascent transcription is barely affected by the loss of the RNA exosome. These observations indicate that the RNA exosome likely degrades chromatin-associated transcripts rather than nascent RNA. This is in line with a previous finding showing that EXOSC10, the second exoribonuclease of the RNA exosome, is recruited in HEK293 cells to the chromatin to degrade poorly-processed transcripts that are associated with pol II (60). Another factor, MYCN, was recently shown to be able to recruit the nuclear RNA exosome to pol II in neuroblastoma (25). However, MYCN is not expressed in HeLa cells, which rules it out as a recruiting factor. A future question would be to determine whether Tyr1 hyperphosphorylation is involved in the recruitment of the nuclear RNA exosome to the chromatin. Interestingly, Integrator activity is required for transcription termination of some protein-coding genes (61). While Integrator recruitment is not enriched in the gene body of the siEX(3)+ genes, we found a specific Integrator signal peak at the 3'end of these RNA exosome sensitive genes. This could indicate that the Integrator complex participates in the 3'end processing of these genes, which could then be degraded on the chromatin by the nuclear RNA exosome.

The Integrator complex has been shown to act on both coding and non-coding RNAs (53). The Integrator complex was shown previously to be recruited via pol II CTD Ser2 and Ser7 phosphorylation to the snRNAs, a class of non-coding RNAs (62). Recently, mutation of the three quarters of the Tyr1 residues to alanine was found to inhibit Integrator recruitment to the pol II, which was not observed with a Ser2 to alanine mutation (49). In addition, structural work has shown that Integrator interacts with the Tyr1 residue of the pol II CTD (24). However, it was unclear whether the Tyr1 residue itself and/or Tyr1 phosphorylation is needed for Integrator recruitment. We have found only a weak correlation (R=0.14-0.17) between ChIP-seq of Integrator subunits and Tyr1 phosphorylation level by mNET-seq, indicating that it is likely that the Tyr1 residue, rather than its phosphorylation, mediates Integrator recruitment to the pol II.

A limitation of our study is the lack of pol II elongation rate, mRNA stability, and mRNA export data, which could affect the nucleoplasmic vs chromatin-associated RNA ratio. However, the CRR genes also have also a low number of cytoplasm RNA-seq reads, a lower number of proteins with peptides, and fewer peptides per proteins while we see the opposite for the NOT-CRR genes, which supports the observations made. In addition, some results were observed in both HeLa and Raji cells from experiments performed by independent research groups.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank Dr Andrew Angel (University of Aberdeen, UK) and Prof Nick. J Proudfoot (University of Oxford, UK) for discussion.

Author contributions: C.H carried out all the bioinformatics analysis with supervision from M.T. S.M. supervised C.H. and M.T. M.T. and S.M. designed the project. M.T. wrote the paper with contributions from all the authors.

FUNDING

Wellcome Trust Investigator Awards [WT106134AIA and WT210641/Z/18/Z to S.M.]. Funding for open access charge: Wellcome Trust.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Cramer, P. (2019) Organization and regulation of gene transcription. *Nature*, **573**, 45-54.
2. Tellier, M., Maudlin, I. and Murphy, S. (2020) Transcription and splicing: A two-way street. *Wiley Interdiscip Rev RNA*, **11**, e1593.
3. Eick, D. and Geyer, M. (2013) The RNA polymerase II carboxy-terminal domain (CTD) code. *Chem Rev*, **113**, 8456-8490.
4. Zaborowska, J., Egloff, S. and Murphy, S. (2016) The pol II CTD: new twists in the tail. *Nat Struct Mol Biol*, **23**, 771-777.
5. Yurko, N.M. and Manley, J.L. (2018) The RNA polymerase II CTD "orphan" residues: Emerging insights into the functions of Tyr-1, Thr-4, and Ser-7. *Transcription*, **9**, 30-40.
6. Descostes, N., Heidemann, M., Spinelli, L., Schuller, R., Maqbool, M.A., Fenouil, R., Koch, F., Innocenti, C., Gut, M., Gut, I. *et al.* (2014) Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. *Elife*, **3**, e02105.
7. Hsin, J.P., Li, W., Hoque, M., Tian, B. and Manley, J.L. (2014) RNAP II CTD tyrosine 1 performs diverse functions in vertebrate cells. *Elife*, **3**, e02112.
8. Burger, K., Schlackow, M. and Gullerova, M. (2019) Tyrosine kinase c-Abl couples RNA polymerase II transcription to DNA double-strand breaks. *Nucleic Acids Res*, **47**, 3467-3484.
9. Yamazaki, T., Liu, L. and Manley, J.L. (2021) Oxidative stress induces Ser 2 dephosphorylation of the RNA polymerase II CTD and premature transcription termination. *Transcription*, **12**, 277-293.

10. 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.
11. Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M. and Proudfoot, N.J. (2017) Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Mol Cell*, **65**, 25-38.
12. Hsin, J.P., Sheth, A. and Manley, J.L. (2011) RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing. *Science*, **334**, 683-686.
13. Egloff, S., Zaborowska, J., Laitem, C., Kiss, T. and Murphy, S. (2012) Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. *Mol Cell*, **45**, 111-122.
14. Chapman, R.D., Heidemann, M., Albert, T.K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E. and Eick, D. (2007) Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science*, **318**, 1780-1782.
15. Reimer, K.A., Mimoso, C.A., Adelman, K. and Neugebauer, K.M. (2021) Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. *Mol Cell*, **81**, 998-1012 e1017.
16. Sousa-Luis, R., Dujardin, G., Zukher, I., Kimura, H., Weldon, C., Carmo-Fonseca, M., Proudfoot, N.J. and Nojima, T. (2021) POINT technology illuminates the processing of polymerase-associated intact nascent transcripts. *Mol Cell*, **81**, 1935-1950 e1936.
17. 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.
18. Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C.C. *et al.* (2015) Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun*, **6**, 7126.
19. Bauer, D.L.V., Tellier, M., Martinez-Alonso, M., Nojima, T., Proudfoot, N.J., Murphy, S. and Fodor, E. (2018) Influenza Virus Mounts a Two-Pronged Attack on Host RNA Polymerase II Transcription. *Cell Rep*, **23**, 2119-2129 e2113.
20. Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., Kiemele, L., Hansen, K., Davis, R., Lykke-Andersen, J. *et al.* (2012) mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell*, **46**, 311-324.
21. Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S. *et al.* (2014) Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat Commun*, **5**, 5531.
22. Beckedorff, F., Blumenthal, E., daSilva, L.F., Aoi, Y., Cingaram, P.R., Yue, J., Zhang, A., Dokaneheifard, S., Valencia, M.G., Gaidosh, G. *et al.* (2020) The Human Integrator Complex Facilitates Transcriptional Elongation by Endonucleolytic Cleavage of Nascent Transcripts. *Cell Rep*, **32**, 107917.
23. Lykke-Andersen, S., Zumer, K., Molska, E.S., Rouviere, J.O., Wu, G., Demel, C., Schwalb, B., Schmid, M., Cramer, P. and Jensen, T.H. (2021) Integrator is a genome-wide attenuator of non-productive transcription. *Mol Cell*, **81**, 514-529 e516.
24. Fianu, I., Chen, Y., Dienemann, C., Dybkov, O., Linden, A., Urlaub, H. and Cramer, P. (2021) Structural basis of Integrator-mediated transcription regulation. *Science*, **374**, 883-887.
25. Papadopoulos, D., Solvie, D., Baluapuri, A., Endres, T., Ha, S.A., Herold, S., Kalb, J., Giansanti, C., Schulein-Volk, C., Ade, C.P. *et al.* (2022) MYCN recruits the nuclear exosome complex to RNA polymerase II to prevent transcription-replication conflicts. *Mol Cell*, **82**, 159-176 e112.
26. Oh, J.M., Di, C., Venters, C.C., Guo, J., Arai, C., So, B.R., Pinto, A.M., Zhang, Z., Wan, L., Younis, I. *et al.* (2017) U1 snRNP telescripting regulates a size-function-stratified human genome. *Nat Struct Mol Biol*, **24**, 993-999.

27. So, B.R., Di, C., Cai, Z., Venters, C.C., Guo, J., Oh, J.M., Arai, C. and Dreyfuss, G. (2019) A Complex of U1 snRNP with Cleavage and Polyadenylation Factors Controls Telescripting, Regulating mRNA Transcription in Human Cells. *Mol Cell*, **76**, 590-599 e594.
28. Dubbury, S.J., Boutz, P.L. and Sharp, P.A. (2018) CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, **564**, 141-145.
29. Kamieniarz-Gdula, K., Gdula, M.R., Panser, K., Nojima, T., Monks, J., Wisniewski, J.R., Riepsaame, J., Brockdorff, N., Pauli, A. and Proudfoot, N.J. (2019) Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Mol Cell*, **74**, 158-172 e159.
30. Wang, R., Zheng, D., Wei, L., Ding, Q. and Tian, B. (2019) Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. *Cell Rep*, **26**, 2766-2778 e2766.
31. Laitem, C., Zaborowska, J., Isa, N.F., Kufs, J., Dienstbier, M. and Murphy, S. (2015) CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nat Struct Mol Biol*, **22**, 396-403.
32. Tellier, M., Ferrer-Vicens, I. and Murphy, S. (2016) The point of no return: The poly(A)-associated elongation checkpoint. *RNA Biol*, **13**, 265-271.
33. Tellier, M., Zaborowska, J., Caizzi, L., Mohammad, E., Velychko, T., Schwalb, B., Ferrer-Vicens, I., Blears, D., Nojima, T., Cramer, P. *et al.* (2020) CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation. *Nucleic Acids Res*, **48**, 7712-7727.
34. Tellier, M., Zaborowska, J., Neve, J., Nojima, T., Hester, S., Fournier, M., Furger, A. and Murphy, S. (2022) CDK9 and PP2A regulate RNA polymerase II transcription termination and coupled RNA maturation. *EMBO Rep*, **23**, e54520.
35. Tellier, M. and Murphy, S. (2020) Incomplete removal of ribosomal RNA can affect chromatin RNA-seq data analysis. *Transcription*, **11**, 230-235.
36. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**, 3.
37. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.
38. 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.
39. Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, **44**, W160-165.
40. Nojima, T., Gomes, T., Grosso, A.R.F., 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.
41. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
42. Robin, T., Bairoch, A., Muller, M., Lisacek, F. and Lane, L. (2018) Large-Scale Reanalysis of Publicly Available HeLa Cell Proteomics Data in the Context of the Human Proteome Project. *J Proteome Res*, **17**, 4160-4170.
43. Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166-169.
44. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.
45. Zhu, A., Ibrahim, J.G. and Love, M.I. (2019) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, **35**, 2084-2092.

46. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.
47. Gene Ontology, C. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*, **49**, D325-D334.
48. Mukherjee, N., Calviello, L., Hirsekorn, A., de Pretis, S., Pelizzola, M. and Ohler, U. (2017) Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat Struct Mol Biol*, **24**, 86-96.
49. Shah, N., Maqbool, M.A., Yahia, Y., El Aabidine, A.Z., Esnault, C., Forne, I., Decker, T.M., Martin, D., Schuller, R., Krebs, S. *et al.* (2018) Tyrosine-1 of RNA Polymerase II CTD Controls Global Termination of Gene Transcription in Mammals. *Mol Cell*, **69**, 48-61 e46.
50. Kim, S., Kim, H., Fong, N., Erickson, B. and Bentley, D.L. (2011) Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proc Natl Acad Sci U S A*, **108**, 13564-13569.
51. Nojima, T., Tellier, M., Foxwell, J., Ribeiro de Almeida, C., Tan-Wong, S.M., Dhir, S., Dujardin, G., Dhir, A., Murphy, S. and Proudfoot, N.J. (2018) Deregulated Expression of Mammalian lncRNA through Loss of SPT6 Induces R-Loop Formation, Replication Stress, and Cellular Senescence. *Mol Cell*, **72**, 970-984 e977.
52. Kilchert, C., Wittmann, S. and Vasiljeva, L. (2016) The regulation and functions of the nuclear RNA exosome complex. *Nat Rev Mol Cell Biol*, **17**, 227-239.
53. Kirstein, N., Gomes Dos Santos, H., Blumenthal, E. and Shiekhattar, R. (2021) The Integrator complex at the crossroad of coding and noncoding RNA. *Curr Opin Cell Biol*, **70**, 37-43.
54. Core, L. and Adelman, K. (2019) Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev*, **33**, 960-982.
55. Zehentner, S., Reiner, A.T., Grimm, C. and Somoza, V. (2021) The Role of Bitter Taste Receptors in Cancer: A Systematic Review. *Cancers (Basel)*, **13**.
56. Pineda, G., Shen, Z., de Albuquerque, C.P., Reynoso, E., Chen, J., Tu, C.C., Tang, W., Briggs, S., Zhou, H. and Wang, J.Y. (2015) Proteomics studies of the interactome of RNA polymerase II C-terminal repeated domain. *BMC Res Notes*, **8**, 616.
57. Fong, N., Saldi, T., Sheridan, R.M., Cortazar, M.A. and Bentley, D.L. (2017) RNA Pol II Dynamics Modulate Co-transcriptional Chromatin Modification, CTD Phosphorylation, and Transcriptional Direction. *Mol Cell*, **66**, 546-557 e543.
58. Lu, K.P., Hanes, S.D. and Hunter, T. (1996) A human peptidyl-prolyl isomerase essential for regulation of mitosis. *Nature*, **380**, 544-547.
59. Xu, Y.X., Hirose, Y., Zhou, X.Z., Lu, K.P. and Manley, J.L. (2003) Pin1 modulates the structure and function of human RNA polymerase II. *Genes Dev*, **17**, 2765-2776.
60. de Almeida, S.F., Garcia-Sacristan, A., Custodio, N. and Carmo-Fonseca, M. (2010) A link between nuclear RNA surveillance, the human exosome and RNA polymerase II transcriptional termination. *Nucleic Acids Res*, **38**, 8015-8026.
61. Dasilva, L.F., Blumenthal, E., Beckedorff, F., Cingaram, P.R., Gomes Dos Santos, H., Edupuganti, R.R., Zhang, A., Dokaneheifard, S., Aoi, Y., Yue, J. *et al.* (2021) Integrator enforces the fidelity of transcriptional termination at protein-coding genes. *Sci Adv*, **7**, eabe3393.
62. Egloff, S., Szczepaniak, S.A., Dienstbier, M., Taylor, A., Knight, S. and Murphy, S. (2010) The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. *J Biol Chem*, **285**, 20564-20569.

TABLE AND FIGURES LEGENDS

Sample name (number of	GEO study	Reference

biological replicates)		
HeLa Chromatin RNA-seq (5)	GSE81662, GSE110028	(11,51)
HeLa Nucleoplasm RNA-seq (4)	GSE81662, GSE110028	(11,51)
HeLa Cytoplasmic RNA-seq (2)	GSE110028	(51)
HeLa Chromatin RNA-seq siEXOSC3 and associated siLuc (2)	GSE81662	(11)
HeLa Nucleoplasm RNA-seq siEXOSC3 and siLuc (2)	GSE81662	(11)
HeLa Chromatin RNA-seq siCPSF73 and associated siLuc (2)	GSE60358	(40)
HeLa Total pol II mNET-seq siEXOSC3 and associated siLuc (2)	GSE81662	(11)
HeLa Total pol II mNET-seq (2)	GSE60358, GSE81662	(11,40)
HeLa Pol II Y1P mNET-seq, Empigen treated (2)	GSE81662	(11)
HeLa Pol II S2P mNET-seq, Empigen treated (2)	GSE81662	(11)
HeLa Pol II T4P mNET-seq, Empigen treated (2)	GSE81662	(11)
HeLa Pol II S5P mNET-seq, Empigen treated (2)	GSE81662	(11)
HeLa Pol II S7P mNET-seq (2)	GSE81662	(11)
HeLa Total pol II POINT-seq (2)	GSE159326	(16)
HeLa H3K4me1, H3K4me3, H3K27me3, H3K36me3, and	GSE110028	(51)

associated Input, mNuc-seq (2)		
HeLa INTS3 and associated Input, ChIP-seq (2)	GSE110028	(51)
HeLa CPSF73 and associated Input, ChIP-seq (2)	GSE127256	(29)
HeLa PCF11 and associated Input, ChIP-seq (2)	GSE127256	(29)
HeLa Total pol II, ASUN, INTS11, INTS3 and associated Input, ChIP-seq (1)	GSE60586	(21)
Raji total pol II, Ser2P, Ser5P, and Thr4P (1)	GSE37519	(10)
Raji total pol II, Ser2P, and Tyr1P (1)	GSE52914	(6)
Raji chromatin and total RNA-seq (2)	GSE94330	(49)

Table 1. List of genome-wide data and their associated GEO accession numbers and references used in this study.

Figure 1. A subset of expressed protein-coding genes produces a low amount of mature mRNA and protein.

(A) MA plot in HeLa cells of the intron-containing protein-coding genes found to be differentially enriched in the nucleoplasm (NOT-CRR, purple) or in the chromatin (CRR, blue) fraction. The set of non-enriched genes (red) and the remaining genes (grey) are also indicated. **(B)** Metagene profiles in HeLa cells of chromatin RNA-seq and nucleoplasm RNA-seq of NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(C)** Metagene profiles in HeLa cells of cytoplasm RNA-seq of NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(D)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the number of peptides per protein found for NOT-CRR (purple), CRR (blue), and unchanged (red) genes. The number of proteins found to have at least one peptide are indicated at the top of each category. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(E)** MA plot in Raji cells of the intron-containing protein-coding genes found to be differentially enriched in the total (NOT-CRR, purple) or in the chromatin (CRR, blue) fraction. A set of non-enriched genes (red) and the remaining genes (grey) are also indicated. **(F)** Metagene profiles in Raji

cells of chromatin RNA-seq and total RNA-seq of NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(G)** Overlap between NOT-CRR genes or between CRR genes found in HeLa and Raji cells. **(H)** Gene Ontology (GO) enrichment analysis of the common HeLa and Raji NOT-CRR and CRR genes.

Figure 2. Higher levels of pol II Tyr1 and Thr4 phosphorylation are associated with poor expression and chromatin retention of transcripts.

(A) Schematic of the total pol II mNET-seq experiments. **(B)** Metagene profiles of mNET-seq in HeLa cells of total pol II and the different pol II CTD phosphorylation mark for NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(C)** Metagene profiles of mNET-seq in HeLa cells of each pol II CTD phosphorylation mark ratioed to total pol II for NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(D)** Boxplots, shown as min to max with first quartile, median, and third quartile, of each pol II CTD phosphorylation mark ratioed to total pol II across the gene body of NOT-CRR (purple), CRR (blue), and unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: ns: not significant, ** < 0.01, **** < 0.0001. **(E)** XY correlation plots of the nucleoplasm fold enrichment, defined as the fold change between nucleoplasm RNA-seq versus chromatin RNA-seq, and each pol II CTD phosphorylation mark ratioed to total pol II. The Pearson correlation with p-value is indicated on each plot. NOT-CRR (purple), CRR (blue), unchanged (red), and remaining (grey) genes are shown.

Figure 3. Not chromatin-retained RNA genes are associated with lower levels of pol II CTD phosphorylation.

(A) MA plot in HeLa cells of the 500 selected intron-containing protein-coding genes found to be differentially enriched in the nucleoplasm (NRCC, purple) or in the chromatin (RCC, blue) fraction. The set of 500 non-enriched genes (red) and the remaining genes (grey) are also indicated. **(B)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the expression (TPM) in chromatin RNA-seq of the selected 500 NOT-CRR (purple), CRR (blue), and unchanged (red) genes. **(C)** Metagene profiles in HeLa cells of chromatin RNA-seq and nucleoplasm RNA-seq of the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(D)** Metagene profiles of mNET-seq in HeLa cells of total pol II and the different pol II CTD phosphorylation mark for the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(E)** Metagene profiles of mNET-seq in HeLa cells of each pol II CTD phosphorylation mark ratioed to total pol II for the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(F)** Boxplots, shown as min to max with first quartile, median, and third quartile, of each pol II CTD phosphorylation mark ratioed to total pol II across the gene body of the selected 500 NOT-CRR (purple), CRR (blue), and unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: ns: not significant, ** < 0.01, **** < 0.0001.

Figure 4. Chromatin-retained RNA genes are less co-transcriptionally spliced.

(A) Schematic of the total pol II POINT-seq experiments. **(B)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the splicing index of each transcript normalized to the number

of exons from the POINT-seq data in HeLa cells of the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: * < 0.05, **** < 0.0001. **(C)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the splicing index of each transcript normalized to the number of exons from the chromatin RNA-seq and nucleoplasm RNA-seq data in HeLa of the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(D)** Screenshot of the genome browser POINT-seq, chromatin RNA-seq, and nucleoplasm RNA-seq tracks of the protein-coding genes *PKM* (NOT-CRR) and *NFIA* (CRR). The read density of one biological replicate for each ChIP-seq is shown in colour while the other biological replicates density is shown below. Stars indicate the location of co-transcriptional splicing events. The arrow indicates the sense of transcription. **(E)** XY correlation plots of the nucleoplasm fold enrichment, defined as the fold change between nucleoplasm RNA-seq versus chromatin RNA-seq, and of the splicing index of each transcript normalized to the number of exons from POINT-seq, chromatin RNA-seq, or nucleoplasm RNA-seq. The Pearson correlation with p-value is indicated on each plot. NOT-CRR (purple), CRR (blue), unchanged (red), and remaining (grey) genes are shown.

Figure 5. Chromatin-retained RNA genes have a weak mRNA cleavage and polyadenylation.

(A) Metagene profiles in HeLa cells of CPSF73 and PCF11 ChIP-seq of the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(B)** Screenshot of the genome browser ChIP-seq tracks of the protein-coding genes *PKM* (NOT-CRR) and *NFIA* (CRR). The read density of one biological replicate for each ChIP-seq is shown in colour while the other biological replicates density is shown below. The arrow indicates the sense of transcription. **(C)** Metagene profiles in HeLa cells of Ser2P mNET-seq treated with siLuc (grey) or siCPSF73 (coloured) of the 500 NOT-CRR, CRR, or unchanged genes. **(D)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the read-through index calculated on the Ser2P mNET-seq treated with siLuc of the 500 (purple), CRR (blue), or unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: ns: not significant, **** < 0.0001. **(E)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the read-through index calculated on the Ser2P mNET-seq treated with siLuc (grey) or siCPSF73 (red) of the 500 NOT-CRR, CRR, or unchanged genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(F)** Metagene profiles in HeLa cells of chromatin RNA-seq treated with siLuc (black) or siCPSF73 (blue) of the 500 NOT-CRR, CRR, or unchanged genes.

Figure 6. Chromatin-retained RNA gene transcripts are not generally sensitive to RNA exosome.

(A) MA plot in HeLa cells of the intron-containing protein-coding genes found to be upregulated (siEX3(+), purple) or downregulated (siEX3(-), blue) after siEXOSC3. A set of non-enriched genes (red) and the remaining genes (grey) are also indicated. **(B)** XY correlation plots of the nucleoplasm fold enrichment, defined as the fold change between nucleoplasm RNA-seq versus chromatin RNA-seq, and of the expression level defined from the chromatin RNA-seq. The siEX3(+) genes are indicated in purple. The Pearson correlation with p-value is indicated on each plot. **(C)** Metagene

profiles of mNET-seq in HeLa cells of total pol II and the different pol II CTD phosphorylation mark for siEX3(+) (purple) or unchanged (red) genes. **(D)** Metagene profiles of mNET-seq in HeLa cells of each pol II CTD phosphorylation mark ratioed to total pol II for siEX3(+) (purple) or unchanged (red) genes. **(E)** Boxplots, shown as min to max with first quartile, median, and third quartile, of each pol II CTD phosphorylation mark ratioed to total pol II across the gene body of siEX3(+) and unchanged genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(F)** Metagene profiles of chromatin and nucleoplasm RNA-seq in HeLa cells following siEXOSC3 in siEX3(+) (purple) or unchanged (red) genes. The siLuc condition is shown in grey. **(G)** Metagene profiles of mNET-seq in HeLa cells of total pol II following siEXOSC3 in siEX3(+) (purple) or unchanged (red) genes. The siLuc condition is shown in grey. **(H)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the total pol II mNET-seq fold change between siEXOSC3 versus siLuc across the gene body of siEX3(+) (purple) or unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001.

Figure 7. Regulation of cytoplasmic mRNA level by chromatin retention.

(A) The high production of mature mRNAs of NOT-CRR genes is associated with lower pol II CTD phosphorylation level and a more efficient pre-mRNA splicing and mRNA CPA. In contrast, CRR genes are associated with higher pol II CTD phosphorylation level, especially on Tyr1 and Thr4 residues, less efficient pre-mRNA splicing and mRNA CPA, and chromatin retention of the poorly processed transcripts. **(B)** Protein-coding genes found to be upregulated after the loss of the nuclear RNA exosome are specifically associated with a hyperphosphorylation of Tyr1 residues of the pol II CTD. As the nuclear RNA exosome has been shown to be recruited co-transcriptionally and to act post-transcriptionally, it remains to be determined whether Tyr1 hyperphosphorylation participates in the recruitment of the nuclear RNA exosome and/or serves as a signal, via interactions with other proteins, for the transcripts to be degraded.

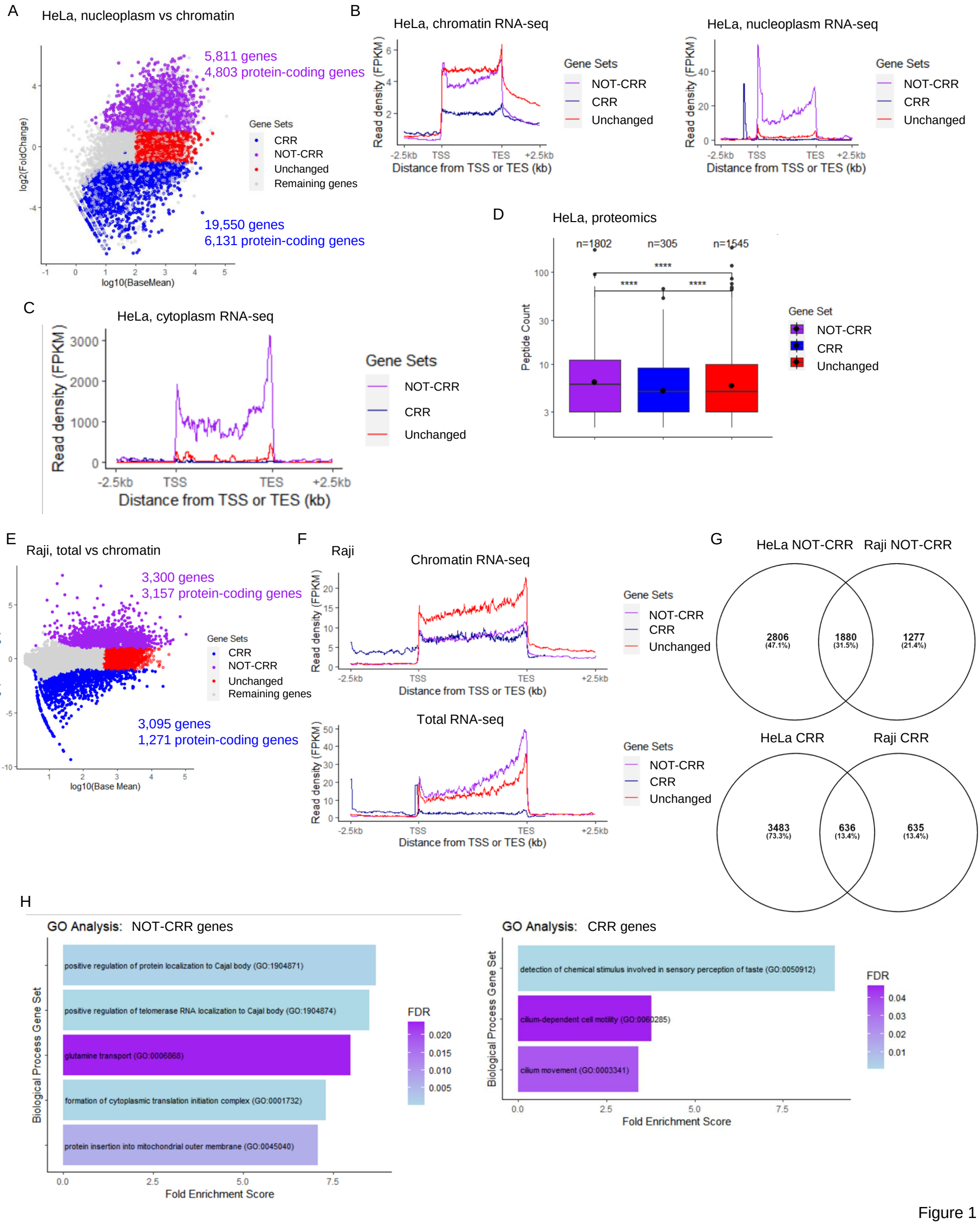

A
HeLa, nucleoplasm vs chromatin
5,811 genes
4,803 protein-coding genes
19,550 genes
6,131 protein-coding genes
Gene Sets
CRR
NOT-CRR
Unchanged
Remaining genes
log2(FoldChange)
log10(BaseMean)
B
HeLa, chromatin RNA-seq
HeLa, nucleoplasm RNA-seq
Read density (FPKM)
Distance from TSS or TES (kb)
-2.5kb
TSS
TES
+2.5kb
Gene Sets
NOT-CRR
CRR
Unchanged
C
HeLa, cytoplasm RNA-seq
D
HeLa, proteomics
n=1802
n=305
n=1545
Peptide Count
Gene Set
E
Raji, total vs chromatin
3,300 genes
3,157 protein-coding genes
3,095 genes
1,271 protein-coding genes
log2(Fold Change)
log10(Base Mean)
F
Raji
Chromatin RNA-seq
Total RNA-seq
G
HeLa NOT-CRR
Raji NOT-CRR
2806 (47.1%)
1880 (31.5%)
1277 (21.4%)
HeLa CRR
Raji CRR
3483 (73.3%)
636 (13.4%)
635 (13.4%)
H
GO Analysis: NOT-CRR genes
positive regulation of protein localization to Cajal body (GO:1904871)
positive regulation of telomerase RNA localization to Cajal body (GO:1904874)
glutamine transport (GO:0006868)
formation of cytoplasmic translation initiation complex (GO:0001732)
protein insertion into mitochondrial outer membrane (GO:0045040)
Biological Process Gene Set
Fold Enrichment Score
FDR
GO Analysis: CRR genes
detection of chemical stimulus involved in sensory perception of taste (GO:0050912)
cilium-dependent cell motility (GO:0060285)
cilium movement (GO:0003341)

Figure 1

Figure 1. A subset of expressed protein-coding genes produces a low amount of mature mRNA and protein.

(A) MA plot in HeLa cells of the intron-containing protein-coding genes found to be differentially enriched in the nucleoplasm (NOT-CRR, purple) or in the chromatin (CRR, blue) fraction. The set of non-enriched genes (red) and the remaining genes (grey) are also indicated. **(B)** Metagene profiles in HeLa cells of chromatin RNA-seq and nucleoplasm RNA-seq of NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(C)** Metagene profiles in HeLa cells of cytoplasm RNA-seq of NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(D)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the number of peptides per protein found for NOT-CRR (purple), CRR (blue), or unchanged (red) genes. The number of proteins found to have at least one peptide are indicated at the top of each category. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(E)** MA plot in Raji cells of the intron-containing protein-coding genes found to be differentially enriched in the total (NOT-CRR, purple) or in the chromatin (CRR, blue) fraction. A set of non-enriched genes (red) and the remaining genes (grey) are also indicated. **(F)** Metagene profiles in Raji cells of chromatin RNA-seq and total RNA-seq of NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(G)** Overlap between NOT-CRR genes or between CRR genes found in HeLa and Raji cells. **(H)** Gene Ontology (GO) enrichment analysis of the common HeLa and Raji NOT-CRR and CRR genes.

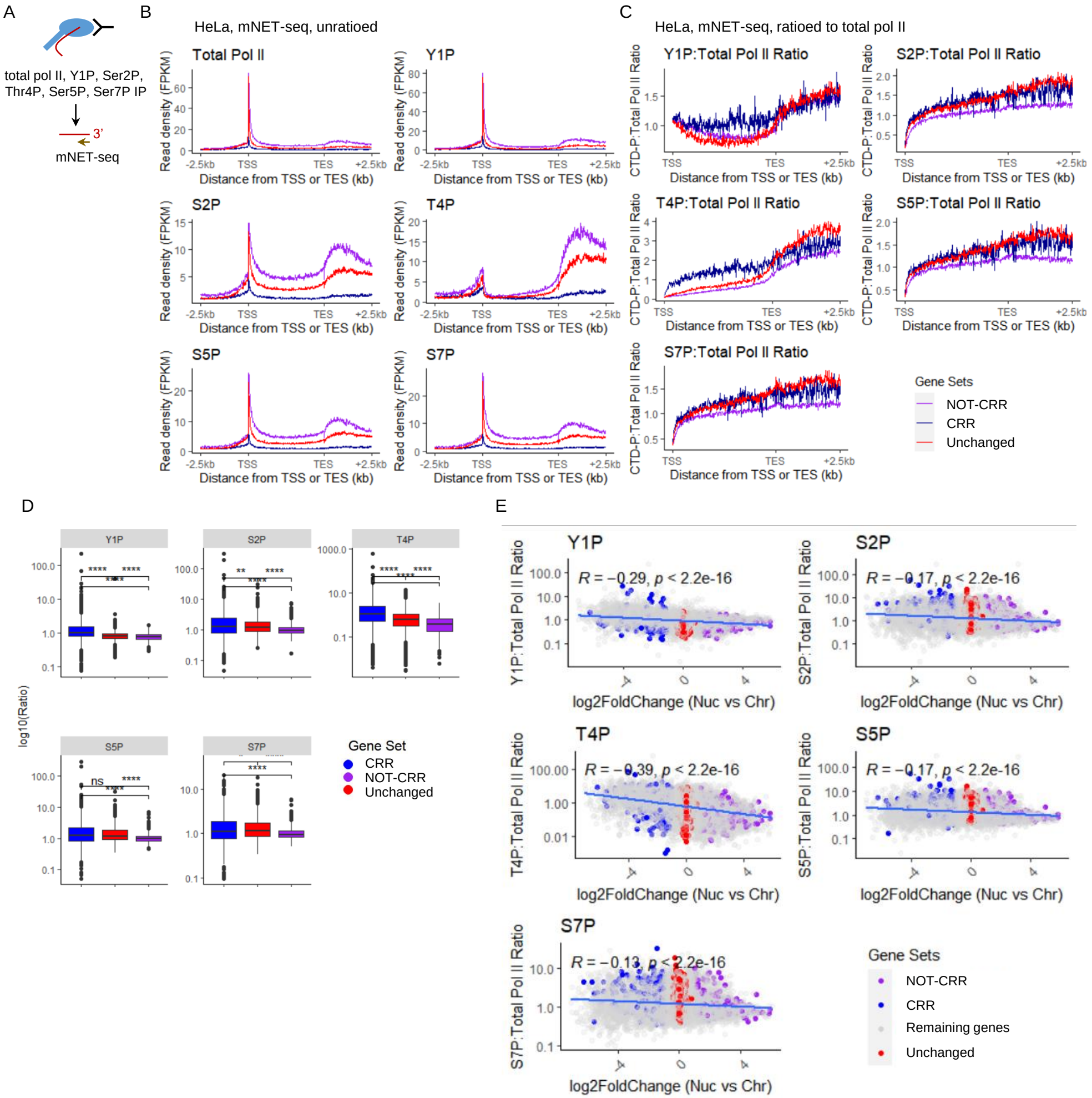

A
total pol II, Y1P, Ser2P, Thr4P, Ser5P, Ser7P IP
3'
mNET-seq
B
HeLa, mNET-seq, unratioed
Total Pol II
Y1P
S2P
T4P
S5P
S7P
Read density (FPKM)
-2.5kb
TSS
TES
+2.5kb
Distance from TSS or TES (kb)
C
HeLa, mNET-seq, ratioed to total pol II
Y1P:Total Pol II Ratio
S2P:Total Pol II Ratio
T4P:Total Pol II Ratio
S5P:Total Pol II Ratio
S7P:Total Pol II Ratio
CTD-P:Total Pol II Ratio
Gene Sets
NOT-CRR
CRR
Unchanged
D
Y1P
S2P
T4P
S5P
S7P
log10(Ratio)
Gene Set
CRR
NOT-CRR
Unchanged
E
Y1P
S2P
T4P
S5P
S7P
R = −0.29, p < 2.2e-16
R = −0.17, p < 2.2e-16
R = −0.39, p < 2.2e-16
R = −0.17, p < 2.2e-16
R = −0.13, p < 2.2e-16
Y1P:Total Pol II Ratio
S2P:Total Pol II Ratio
T4P:Total Pol II Ratio
S5P:Total Pol II Ratio
S7P:Total Pol II Ratio
log2FoldChange (Nuc vs Chr)
Gene Sets
NOT-CRR
CRR
Remaining genes
Unchanged

Figure 2

Figure 2. Higher levels of pol II Tyr1 and Thr4 phosphorylation are associated with poor expression and chromatin retention of transcripts.

(A) Schematic of the total pol II mNET-seq experiments. **(B)** Metagene profiles of mNET-seq in HeLa cells of total pol II and the different pol II CTD phosphorylation mark for NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(C)** Metagene profiles of mNET-seq in HeLa cells of each pol II CTD phosphorylation mark ratioed to total pol II for NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(D)** Boxplots, shown as min to max with first quartile, median, and third quartile, of each pol II CTD phosphorylation mark ratioed to total pol II across the gene body of NOT-CRR (purple), CRR (blue), or unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: ns: not significant, ** < 0.01, **** < 0.0001. **(E)** XY correlation plots of the nucleoplasm fold enrichment, defined as the fold change between nucleoplasm RNA-seq versus chromatin RNA-seq, and each pol II CTD phosphorylation mark ratioed to total pol II. The Pearson correlation with p-value is indicated on each plot. NOT-CRR (purple), CRR (blue), unchanged (red), and remaining (grey) genes are shown.

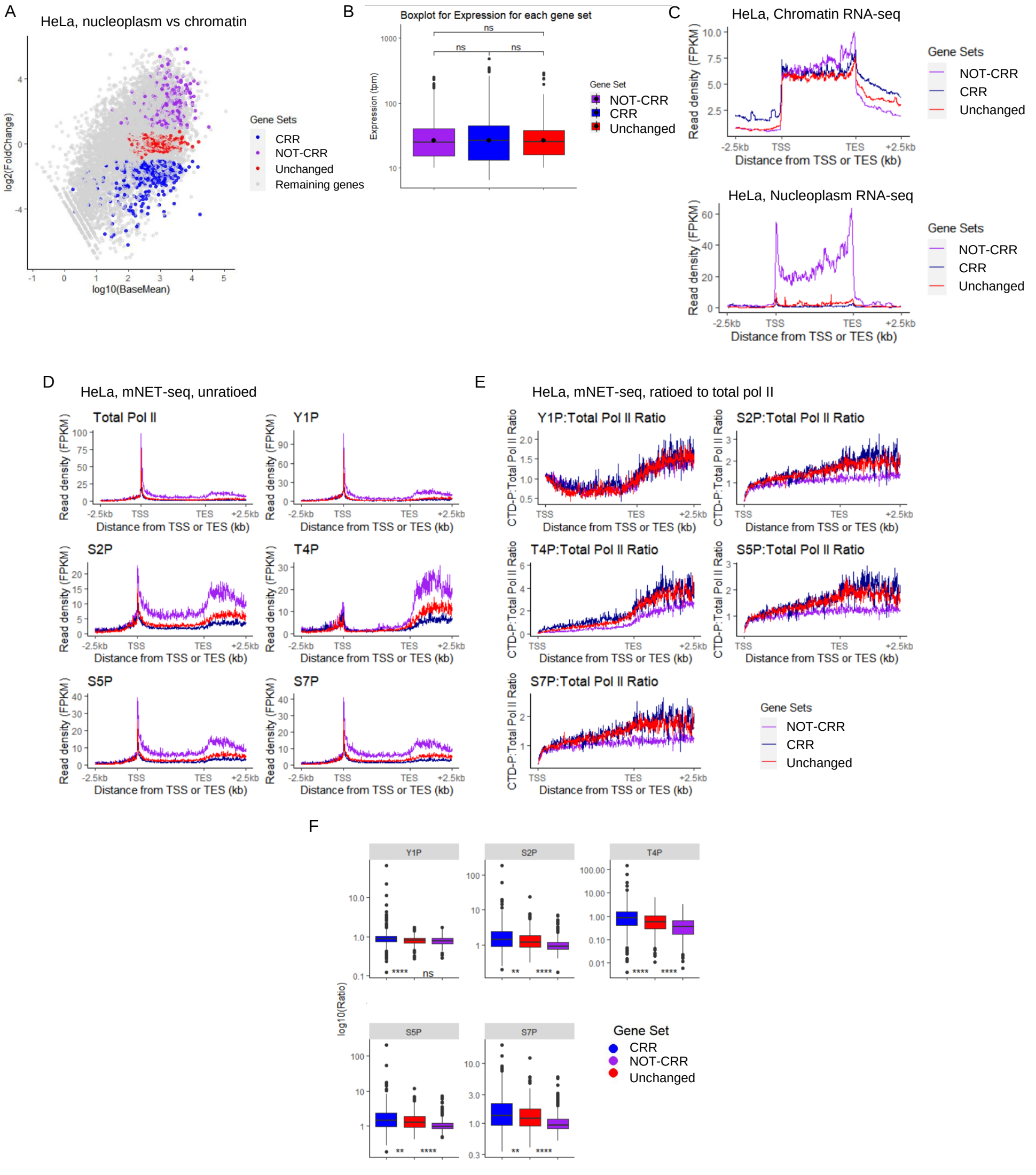

A
HeLa, nucleoplasm vs chromatin
log2(FoldChange)
log10(BaseMean)
Gene Sets
CRR
NOT-CRR
Unchanged
Remaining genes
B
Boxplot for Expression for each gene set
Expression (tpm)
ns
Gene Set
NOT-CRR
CRR
Unchanged
C
HeLa, Chromatin RNA-seq
HeLa, Nucleoplasm RNA-seq
Read density (FPKM)
Distance from TSS or TES (kb)
-2.5kb
TSS
TES
+2.5kb
Gene Sets
NOT-CRR
CRR
Unchanged
D
HeLa, mNET-seq, unratioed
Total Pol II
Y1P
S2P
T4P
S5P
S7P
Read density (FPKM)
E
HeLa, mNET-seq, ratioed to total pol II
Y1P:Total Pol II Ratio
S2P:Total Pol II Ratio
T4P:Total Pol II Ratio
S5P:Total Pol II Ratio
S7P:Total Pol II Ratio
CTD-P:Total Pol II Ratio
F
log10(Ratio)
Y1P
S2P
T4P
S5P
S7P
Gene Set
CRR
NOT-CRR
Unchanged

Figure 3

Figure 3. Not chromatin-retained RNA genes are associated with lower levels of pol II CTD phosphorylation.

(A) MA plot in HeLa cells of the 500 selected intron-containing protein-coding genes found to be differentially enriched in the nucleoplasm (NOT-CRR, purple) or in the chromatin (CRR, blue) fraction. The set of 500 non-enriched genes (red) and the remaining genes (grey) are also indicated. **(B)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the expression (TPM) in chromatin RNA-seq of the selected 500 NOT-CRR (purple), CRR (blue), and unchanged (red) genes. **(C)** Metagene profiles in HeLa cells of chromatin RNA-seq and nucleoplasm RNA-seq of the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(D)** Metagene profiles of mNET-seq in HeLa cells of total pol II and the different pol II CTD phosphorylation mark for the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(E)** Metagene profiles of mNET-seq in HeLa cells of each pol II CTD phosphorylation mark ratioed to total pol II for the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(F)** Boxplots, shown as min to max with first quartile, median, and third quartile, of each pol II CTD phosphorylation mark ratioed to total pol II across the gene body of the selected 500 NOT-CRR (purple), CRR (blue), and unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: ns: not significant, ** < 0.01, **** < 0.0001.

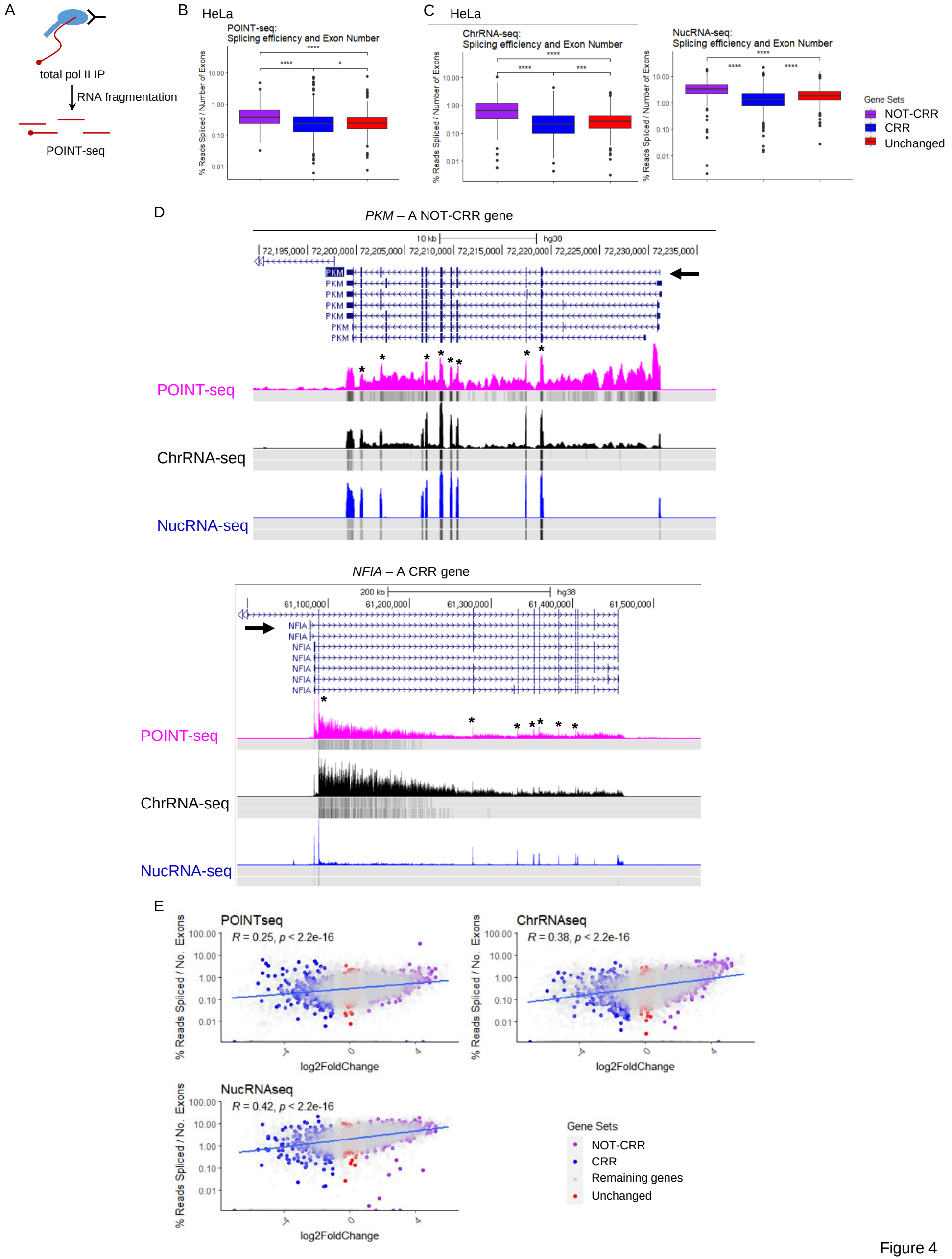
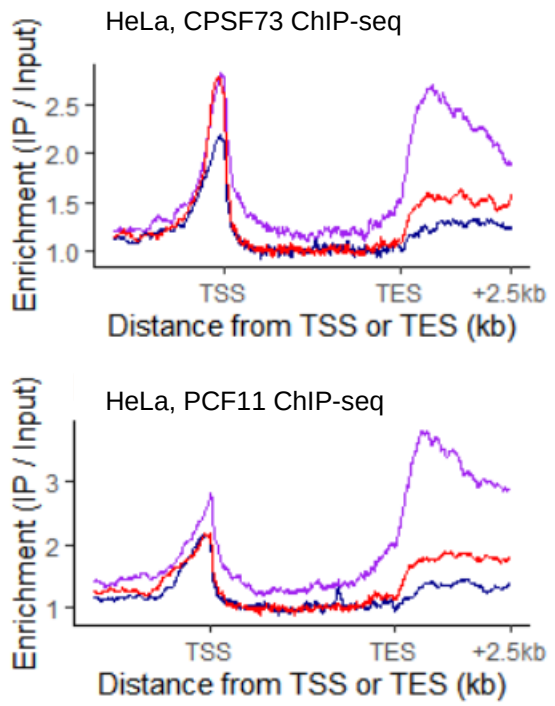

A
total pol II IP
RNA fragmentation
POINT-seq
B
HeLa
POINT-seq:
Splicing efficiency and Exon Number
% Reads Spliced / Number of Exons
C
HeLa
ChrRNA-seq:
Splicing efficiency and Exon Number
NucRNA-seq:
Splicing efficiency and Exon Number
Gene Sets
NOT-CRR
CRR
Unchanged
D
PKM – A NOT-CRR gene
10 kb
hg38
POINT-seq
ChrRNA-seq
NucRNA-seq
NFIA – A CRR gene
200 kb
E
POINTseq
R = 0.25, p < 2.2e-16
ChrRNAseq
R = 0.38, p < 2.2e-16
NucRNAseq
R = 0.42, p < 2.2e-16
% Reads Spliced / No. Exons
log2FoldChange
Gene Sets
NOT-CRR
CRR
Remaining genes
Unchanged

Figure 4

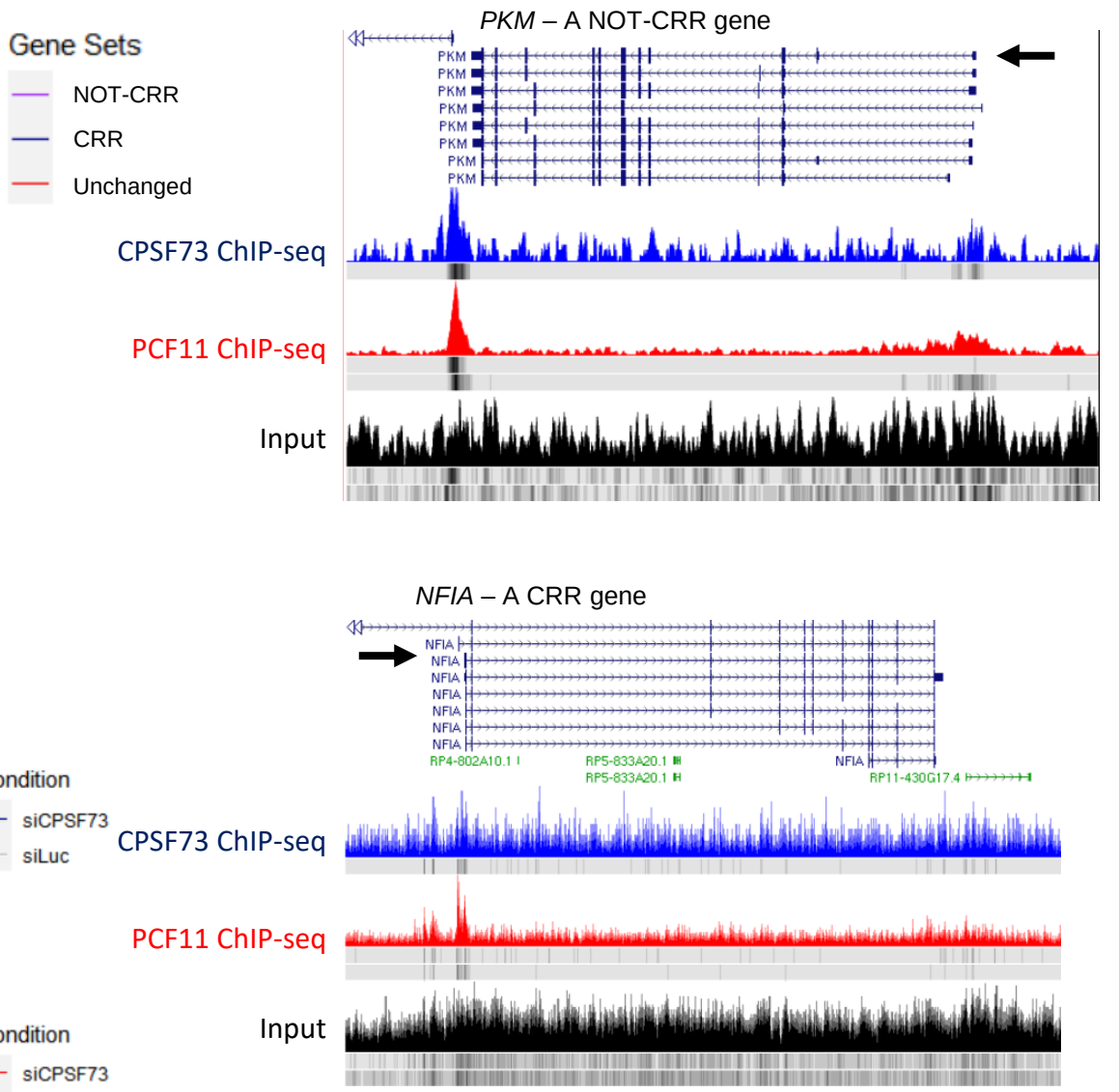
Figure 4. Chromatin-retained RNA genes are less co-transcriptionally spliced.

(A) Schematic of the total pol II POINT-seq experiments. **(B)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the splicing index of each transcript normalized to the number of exons from the POINT-seq data in HeLa cells of the 500 NOT-CRR (purple), CRR (blue), and unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: * < 0.05, **** < 0.0001. **(C)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the splicing index of each transcript normalized to the number of exons from the chromatin RNA-seq and nucleoplasm RNA-seq data in HeLa of the 500 NOT-CRR (purple), CRR (blue), and unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(D)** Screenshot of the genome browser POINT-seq, chromatin RNA-seq, and nucleoplasm RNA-seq tracks of the protein-coding genes *PKM* (NOT-CRR) and *NFIA* (CRR). The read density of one biological replicate for each ChIP-seq is shown in colour while the other biological replicates density is shown below. Stars indicate the location of co-transcriptional splicing events. The arrow indicates the sense of transcription. **(E)** XY correlation plots of the nucleoplasm fold enrichment, defined as the fold change between nucleoplasm RNA-seq versus chromatin RNA-seq, and of the splicing index of each transcript normalized to the number of exons from POINT-seq, chromatin RNA-seq, or nucleoplasm RNA-seq. The Pearson correlation with p-value is indicated on each plot. NOT-CRR (purple), CRR (blue), unchanged (red), and remaining (grey) genes are shown.

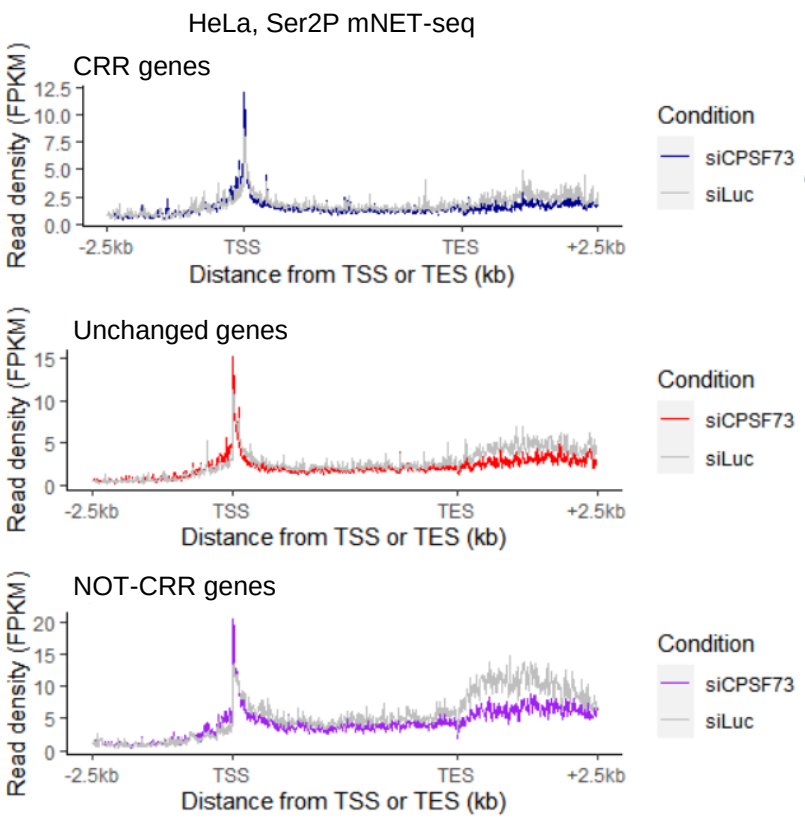
A



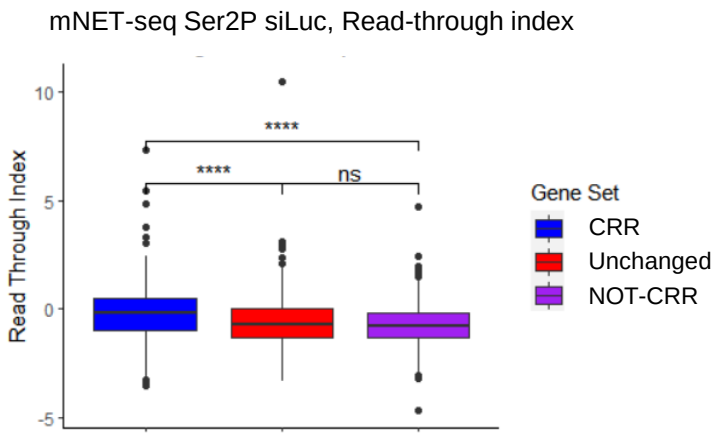
B



C



D



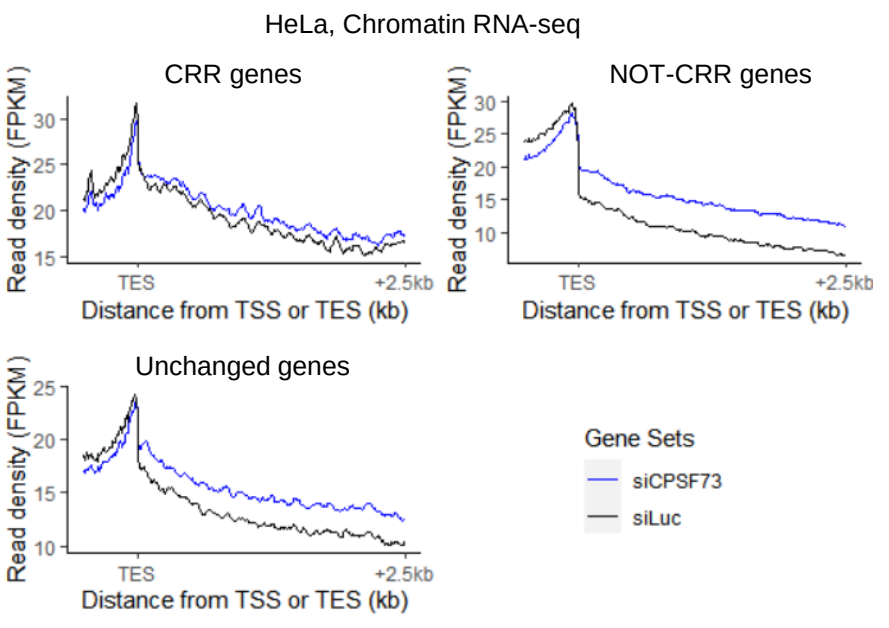


Figure 5

Figure 5. Chromatin-retained RNA genes have a weak mRNA cleavage and polyadenylation.

(A) Metagene profiles in HeLa cells of CPSF73 and PCF11 ChIP-seq of the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. **(B)** Screenshot of the genome browser ChIP-seq tracks of the protein-coding genes *PKM* (NOT-CRR) and *NFIA* (CRR). The read density of one biological replicate for each ChIP-seq is shown in colour while the other biological replicates density is shown below. The arrow indicates the sense of transcription. **(C)** Metagene profiles in HeLa cells of Ser2P mNET-seq treated with siLuc (grey) or siCPSF73 (coloured) of the 500 NOT-CRR, CRR, or unchanged genes. **(D)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the read-through index calculated on the Ser2P mNET-seq treated with siLuc of the 500 NOT-CRR (purple), CRR (blue), or unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: ns: not significant, **** < 0.0001. **(E)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the read-through index calculated on the Ser2P mNET-seq treated with siLuc (grey) or siCPSF73 (red) of the 500 NOT-CRR, CRR, or unchanged genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(F)** Metagene profiles in HeLa cells of chromatin RNA-seq treated with siLuc (black) or siCPSF73 (blue) of the 500 NOT-CRR, CRR, or unchanged genes.

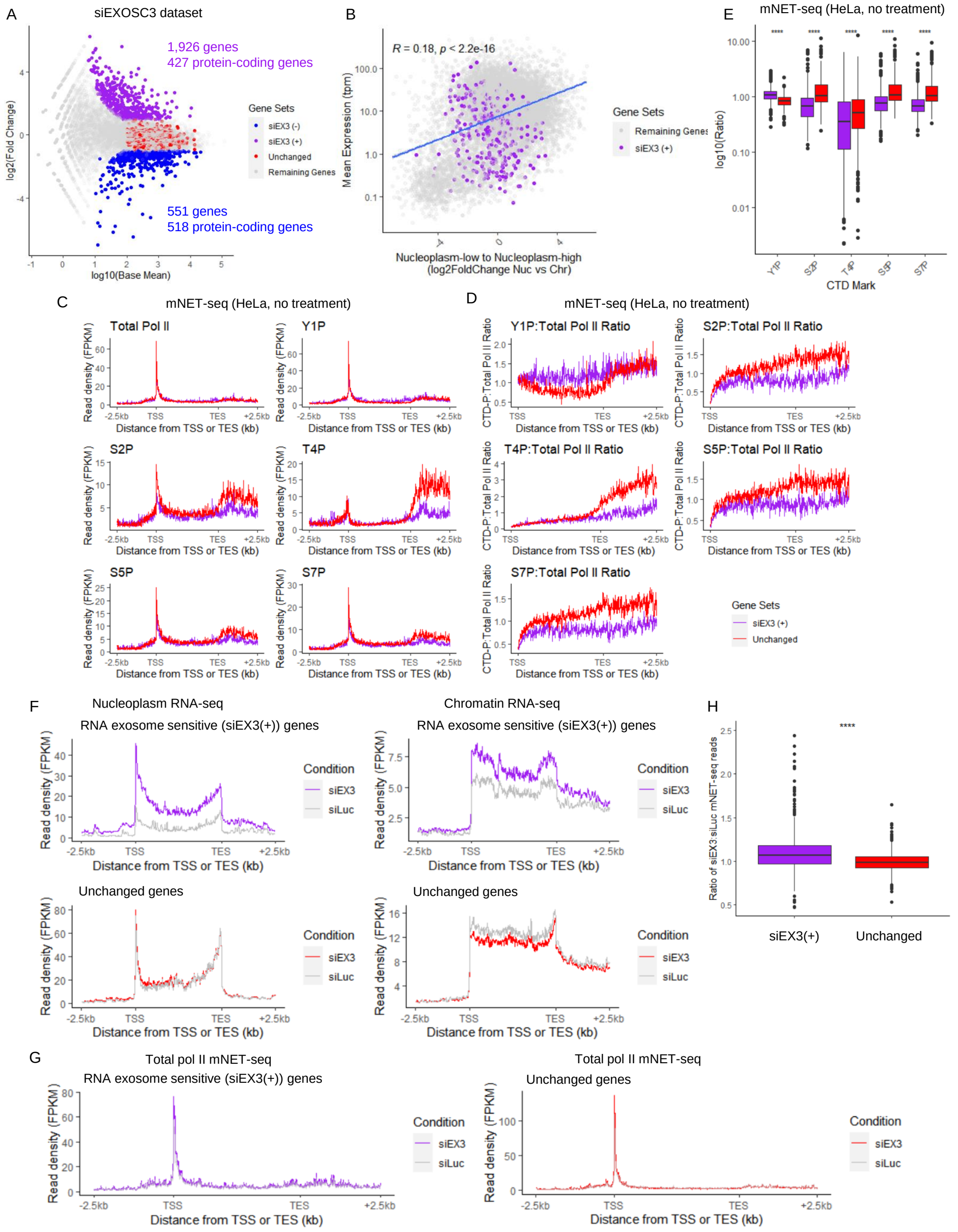

A
siEXOSC3 dataset
1,926 genes
427 protein-coding genes
551 genes
518 protein-coding genes
Gene Sets
siEX3 (-)
siEX3 (+)
Unchanged
Remaining Genes
log2(Fold Change)
log10(Base Mean)
B
R = 0.18, p < 2.2e-16
Mean Expression (tpm)
Nucleoplasm-low to Nucleoplasm-high
(log2FoldChange Nuc vs Chr)
Gene Sets
Remaining Genes
siEX3 (+)
E
mNET-seq (HeLa, no treatment)
log10(Ratio)
CTD Mark
Y1P
S2P
T4P
S5P
S7P
C
mNET-seq (HeLa, no treatment)
Total Pol II
Y1P
S2P
T4P
S5P
S7P
Read density (FPKM)
Distance from TSS or TES (kb)
-2.5kb
TSS
TES
+2.5kb
D
mNET-seq (HeLa, no treatment)
Y1P:Total Pol II Ratio
S2P:Total Pol II Ratio
T4P:Total Pol II Ratio
S5P:Total Pol II Ratio
S7P:Total Pol II Ratio
CTD-P:Total Pol II Ratio
Gene Sets
siEX3 (+)
Unchanged
F
Nucleoplasm RNA-seq
RNA exosome sensitive (siEX3(+)) genes
Chromatin RNA-seq
RNA exosome sensitive (siEX3(+)) genes
Unchanged genes
Unchanged genes
Condition
siEX3
siLuc
H
Ratio of siEX3:siLuc mNET-seq reads
siEX3(+)
Unchanged
G
Total pol II mNET-seq
RNA exosome sensitive (siEX3(+)) genes
Total pol II mNET-seq
Unchanged genes
Condition
siEX3
siLuc

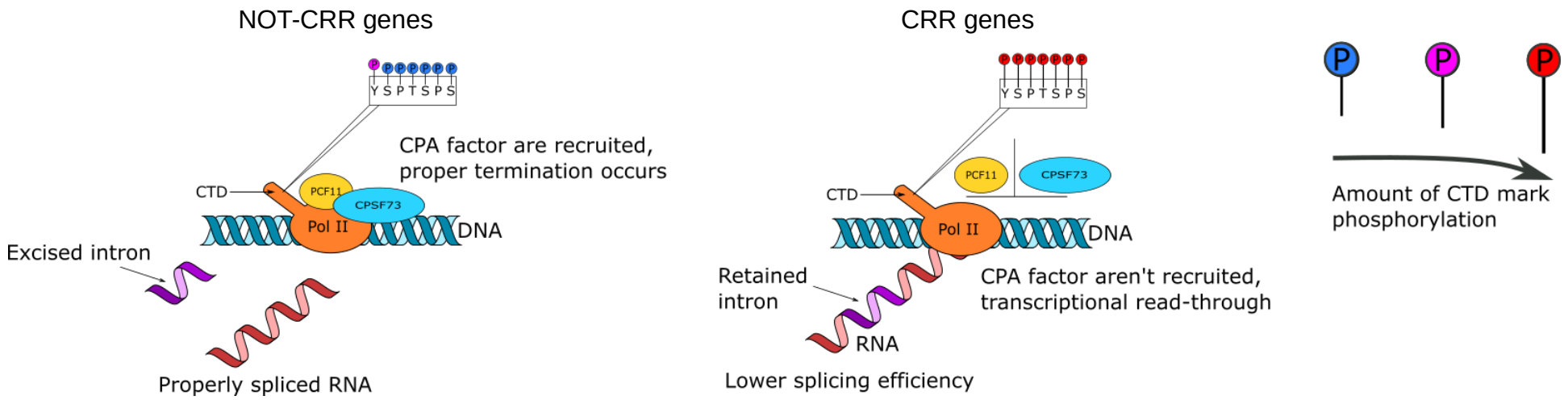
Figure 6

Figure 6. Chromatin-retained RNA gene transcripts are not generally sensitive to RNA exosome.

(A) MA plot in HeLa cells of the intron-containing protein-coding genes found to be upregulated (siEX3(+), purple) or downregulated (siEX3(-), blue) after siEXOSC3. A set of non-enriched genes (red) and the remaining genes (grey) are also indicated. **(B)** XY correlation plots of the nucleoplasm fold enrichment, defined as the fold change between nucleoplasm RNA-seq versus chromatin RNA-seq, and of the expression level defined from the chromatin RNA-seq. The siEX3(+) genes are indicated in purple. The Pearson correlation with p-value is indicated on each plot. **(C)** Metagene profiles of mNET-seq in HeLa cells of total pol II and the different pol II CTD phosphorylation mark for siEX3(+) (purple) or unchanged (red) genes. **(D)** Metagene profiles of mNET-seq in HeLa cells of each pol II CTD phosphorylation mark ratioed to total pol II for siEX3(+) (purple) or unchanged (red) genes. **(E)** Boxplots, shown as min to max with first quartile, median, and third quartile, of each pol II CTD phosphorylation mark ratioed to total pol II across the gene body of siEX3(+) and unchanged genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001. **(F)** Metagene profiles of chromatin and nucleoplasm RNA-seq in HeLa cells following siEXOSC3 in siEX3(+) (purple) or unchanged (red) genes. The siLuc condition is shown in grey. **(G)** Metagene profiles of mNET-seq in HeLa cells of total pol II following siEXOSC3 in siEX3(+) (purple) or unchanged (red) genes. The siLuc condition is shown in grey. **(H)** Boxplots, shown as min to max with first quartile, median, and third quartile, of the total pol II mNET-seq fold change between siEXOSC3 versus siLuc across the gene body of siEX3(+) (purple) or unchanged (red) genes. Statistical test: Wilcoxon rank sum test. P-value: **** < 0.0001.

A

Model of cytoplasmic mRNA level by chromatin retention



B

Model for RNA Exosome degradation

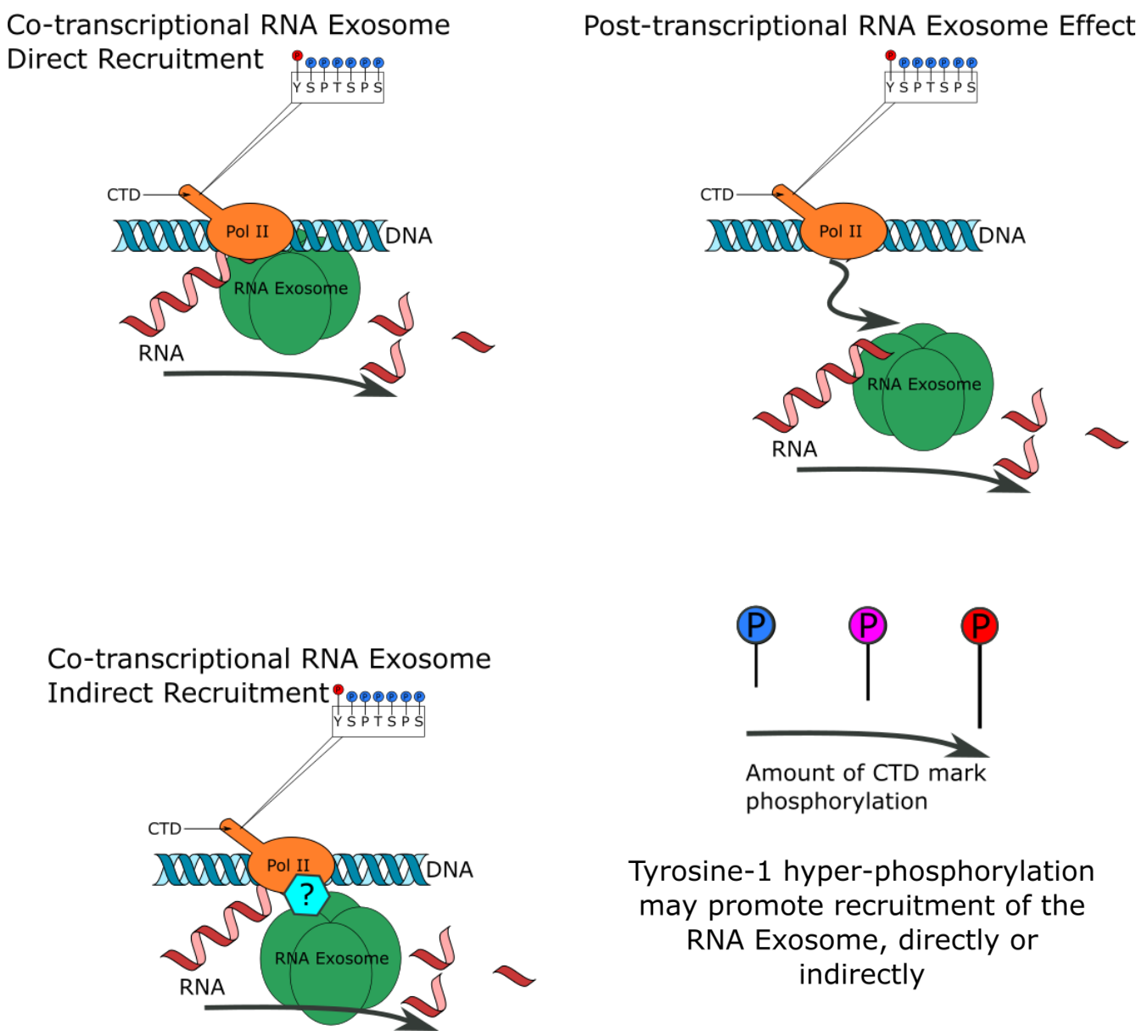


Figure 7

Figure 7. Regulation of cytoplasmic mRNA level by chromatin retention.

(A) The high production of mature mRNAs of NOT-CRR genes is associated with lower pol II CTD phosphorylation level and a more efficient pre-mRNA splicing and mRNA CPA. In contrast, CRR genes are associated with higher pol II CTD phosphorylation level, especially on Tyr1 and Thr4 residues, less efficient pre-mRNA splicing and mRNA CPA, and chromatin retention of the poorly processed transcripts. **(B)** Protein-coding genes found to be upregulated after the loss of the nuclear RNA exosome are specifically associated with a hyperphosphorylation of Tyr1 residues of the pol II CTD. As the nuclear RNA exosome has been shown to be recruited co-transcriptionally and to act post-transcriptionally, it remains to be determined whether Tyr1 hyperphosphorylation participates in the recruitment of the nuclear RNA exosome and/or serves as a signal, via interactions with other proteins, for the transcripts to be degraded.