A Two-Layered Targeting Mechanism Underlies Nuclear RNA Sorting by the Human Exosome

Graphical Abstract

Highlights

- Most nuclear exosome targeted loci produce NEXT- and PAXT-sensitive RNA isoforms
- NEXT targets non-polyadenylated RNA 3’ ends distributed over kilobase-wide regions
- PAXT targets polyadenylated RNA 3’ ends generated by CPSF at canonical poly(A) sites
- NEXT targets are polyadenylated upon NEXT depletion, allowing exosomal decay via PAXT

Authors
Guifen Wu, Manfred Schmid, Leonor Rib, Patrik Polak, Nicola Meola, Albin Sandelin, Torben Heick Jensen

Correspondence
thj@mbg.au.dk

In Brief
Wu et al. demonstrate that the nuclear exosome adaptor NEXT targets poly(A)^− RNAs with poorly defined 3’ ends, whereas the PAXT connection targets poly(A)^+ RNAs derived from canonical poly(A) sites. NEXT substrates become polyadenylated in the absence of NEXT, causing fail-safe RNA decay via PAXT.
A Two-Layered Targeting Mechanism Underlies Nuclear RNA Sorting by the Human Exosome

Guifen Wu,1,4 Manfred Schmid,1,4 Leonor Rib,2,4 Patrik Polak,1 Nicola Meola,1,3 Albin Sandelin,2 and Torben Heick Jensen1,5,*
1Department of Molecular Biology and Genetics, Aarhus University, C.F. Møllers Allé 3, Building 1130, 8000 Aarhus C, Denmark
2The Bioinformatics Centre, Department of Biology and Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark
3Present address: Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark
4These authors contributed equally
5Lead Contact
*Correspondence: thj@mbg.au.dk
https://doi.org/10.1016/j.celrep.2020.01.068

SUMMARY

Degradation of transcripts in human nuclei is primarily facilitated by the RNA exosome. To obtain substrate specificity, the exosome is aided by adaptors; in the nucleoplasm, those adaptors are the nuclear exosome-targeting (NEXT) complex and the poly(A) (pA) exosome-targeting (PAXT) connection. How these adaptors guide exosome targeting remains enigmatic. Employing high-resolution 3′ end sequencing, we demonstrate that NEXT substrates arise from heterogenous and predominantly pA ends defined by canonical pA site use. Irrespective of this clear division, NEXT and PAXT act redundantly in two ways: (1) regional redundancy, where the majority of exosome-targeted transcription units produce NEXT- and PAXT-sensitive RNA isoforms, and (2) isoform redundancy, where the PAXT connection ensures fail-safe decay of post-transcriptionally polyadenylated NEXT targets. In conjunction, this provides a two-layered targeting mechanism for efficient nuclear sorting of the human transcriptome.

INTRODUCTION

Eukaryotic genomes are pervasively transcribed, leaving a large proportion of the output RNA prey to the ribonucleolytic nuclear exosome (Jensen et al., 2013; Kilchert et al., 2016; Schmid and Jensen, 2018; Zinder and Lima, 2017). Such intense decay of nuclear RNA prevents deleterious accumulation of non-functional material and includes removal of incorrectly/prematurely processed transcripts from protein-coding genes as well as post-transcriptional suppression of a large group of unstable long non-coding RNAs (lncRNAs). In addition, the exosome is engaged in turnover of nucleus-residing transcripts and 3′ end processing of numerous stable non-coding RNA (ncRNA) species, such as rRNAs, tRNAs, small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs) (Schmid and Jensen, 2018; Zinder and Lima, 2017). Hence, the diversity of RNAs targeted by the nuclear exosome is vast, and, unsurprisingly, the protein complex is essential for cell growth (Fakken et al., 2017; Mitchell et al., 1997; Snee et al., 2016).

In human cells, the multi-subunit nuclear exosome complex comprises two active exonucleases, EXOSC10/hRRP6 and DIS3/hRRP44, whose in vivo activity critically depends on association of the exosome with adaptor complexes that individually facilitate recognition of specific substrates (Schmid and Jensen, 2018; Zinder and Lima, 2017). Two such adaptors have been described to function in the nucleoplasm: the nuclear exosome-targeting (NEXT) complex and the poly(A) exosome-targeting (PAXT) connection (Lubas et al., 2011; Meola et al., 2016). NEXT and PAXT commonly contain the RNA helicase MTR4, the trimeric NEXT complex consists of the Zn-knuckle protein ZCCHC8 and the RNA binding protein RBM7 (Lubas et al., 2011). The more complex PAXT connection comprises at its core a stable dimer of MTR4 and the Zn-finger protein ZFC3H1 (Meola et al., 2016). Additionally, ZFC3H1-MTR4 associates more transiently with the nuclear poly(A) (pA) binding protein PABPN1, the Zn-finger protein ZC3H3, and presumably one of the two RNA binding protein paralogs RBM26 and RBM27 (Meola et al., 2016; Silla et al., 2020). Although the exact architecture of an active PAXT connection has not been established, individual depletion of any of the abovementioned PAXT components leads to stabilization of a specific subset of nuclear exosome targets (Silla et al., 2020).

NEXT and PAXT both connect, independently, to the nuclear cap binding complex (CBC) (Andersen et al., 2013; Meola et al., 2016), which, together with their strict nucleoplasmic localization (Lubas et al., 2011; Meola et al., 2016), predicts affinities toward capped RNAs. Consistently, the NEXT complex is required for exosomal targeting of promoter upstream transcripts (PROMPTs), enhancer RNAs (eRNAs), 3′ extended snRNA, 3′ extended histone RNA, and intronic RNA (Lubas et al., 2015; Lubas et al., 2011), whereas PAXT mediates the exosomal decay of...
diverse IncRNAs, including spliced transcripts deriving from snoRNA host genes and prematurely terminated RNAs produced by intronic pA sites (PASs) within protein-coding genes (Meola et al., 2016; Ogami et al., 2017). NEXT substrates have therefore been suggested to comprise mostly short and unprocessed RNAs, whereas PAXT substrates have been proposed to target primarily longer polyadenylated transcripts (Meola et al., 2016). However, at apparent odds with this trend, PROMPT and eRNA transcription appears to be terminated by polyadenylation (CPA) complex has not been established, and resulting in pA+/C0 et al., 2016; Ntini et al., 2013). However, whether these PASs always lead to canonical 3′ end processing by the cleavage and polyadenylation (CPA) complex has not been established, and at least some eRNA transcription appears to be terminated by the integrator-dependent pathway (Lai et al., 2015), presumably resulting in pA− RNA substrates. Taking these observations together, the RNA biochemical properties dictating NEXT versus PAXT targeting remain to be clarified.

Here we report a detailed genome-wide analysis of NEXT and PAXT substrate targeting in human HeLa cells. Our data corroborate the notion that NEXT and PAXT provide complementary substrate recognition by targeting pA− RNAs shortly after their release from the transcription machinery and pA+ RNAs processed by conventional CPA-mediated polyadenylation, respectively. However, despite this clear distinction, a surprisingly large degree of overlap exists. First, individual transcription units often produce a complex mix of both NEXT- and PAXT-targeted transcript isoforms that can only be distinguished by their precisely mapped 3′ ends. Second, even though NEXT substrates are normally not polyadenylated, NEXT depletion allows their post-transcriptional polyadenylation and ensuing PAXT-dependent turnover. Therefore, our results delineate targeting requirements of nucleoplasmic RNA decay adaptor complexes and reveal an unprecedented mechanism for fail-safe decay of unwanted transcripts in case a first line of substrate recognition is prohibited.

RESULTS

NEXT and PAXT Target Both Common and Distinct PROMPT Regions

Previous substrate analysis, based on a factor depletion/RNA sequencing (RNA-seq) approach, estimated that NEXT primarily targets short immature RNAs, such as PROMPTs, whereas PAXT targets longer and more processed transcripts (Meola et al., 2016). However, this distinction was not absolute because some PROMPTs were also affected by depletion of PAXT components (Meola et al., 2016; Silla et al., 2020). This blurry partition might reflect targeting of different subsets of PROMPTs by each of the two adaptor complexes or, alternatively, may be due to redundant functions of NEXT and PAXT. To distinguish these possibilities, we re-analyzed published RNA-seq datasets of small interfering RNA (siRNA)-mediated depletions of ZCCHC8, ZFC3H1, and RRP40 (siZCCHC8, siZFC3H1, and siRRP40; Meola et al., 2016) with an initial focus on PROMPT regions because they constitute an overall well-defined and uniform group of exosome-sensitive loci.

Manual inspection of these data, derived from steady-state (“total”) RNA samples, readily yielded evidence of PROMPTs upregulated in both PAXT and NEXT depletion samples. However, a fraction of PROMPTs was biased to be upregulated solely in siZFC3H1 or siZCCHC8 samples, which we termed “PAXT PROMPTs” and “NEXT PROMPTs,” respectively (see Figures 1A and S1A for individual examples). The pathway-specific sensitivities of these PROMPTs were further validated by qRT-PCR analysis (Figure 1B). In both the RNA-seq and qRT-PCR data-sets, upregulation of NEXT-sensitive PROMPTs under the siZCCHC8 condition was generally less pronounced than upon RRP40 depletion. In contrast, PAXT PROMPTs were often more upregulated in siZFC3H1 versus siRRP40 samples. Similar trends have been observed previously (Andersen et al., 2013; Lubas et al., 2011; Meola et al., 2016), indicating a biologically relevant distinction between NEXT and PAXT targets, which will be addressed in detail below.

Inspired by these findings, we refined our previous bioinformatics pipeline to define a list of 6140 PROMPT TSSs in HeLa cells and used these to compute downstream transcript sensitivities to ZCCHC8, ZFC3H1, and RRP40 depletion (Figure S1B; STAR Methods). As expected, the majority (5,426) of these PROMPTs were upregulated more than 2-fold (sensitivity score > 0.5; STAR Methods) upon siRRP40 treatment (Table S1), whereas ZCCHC8 or ZFC3H1 depletion by the same criteria led to upregulation of 3,946 and 2,982 PROMPTs, respectively. Plotting siZCCHC8 versus siZFC3H1 sensitivity confirmed our notion from selected examples that, although the majority of PROMPTs increased upon both ZCCHC8 and ZFC3H1 depletion, a sizeable number displayed preferential upregulation under only one of the conditions (Figure 1C). To uncover PAXT- or NEXT-specific PROMPT features, we therefore selected the 576 PAXT PROMPTs sensitive to siZFC3H1 but not siZCCHC8 (blue dots in Figure 1C) and the corresponding 1,540 ZCCHC8- but not ZFC3H1-sensitive NEXT PROMPTs (orange dots in Figure 1C). Metagene analysis of total RNA from siGFP control cells revealed that NEXT PROMPTs are generally shorter and less abundant than PAXT PROMPTs (Figure S1C, left panel), which supported our previous analysis (Meola et al., 2016). This was despite comparable TSS-proximal RNA levels for both PROMPT classes upon RRP40 depletion (Figure S1C, second panel) and roughly equal RNA polymerase II (Pol II) loading at their respective TSSs, as evaluated by mammalian native elongating transcript sequencing (mNETseq; Schlackow et al., 2017; Figure S1D). Altogether, this pointed to distinct post-transcriptional fates of NEXT and PAXT PROMPTs. In addition, although the metagene plots in Figure S1C confirmed specific siZCCHC8 and siZFC3H1 effects on NEXT and PAXT PROMPTs, respectively, they also reinforced the abovementioned notion that NEXT PROMPTs are only partially stabilized upon ZCCHC8 depletion (Figure S1C, compare the siZCCHC8 and siRRP40 plots).

Analyses so far were based on total RNA samples. We reported previously that the effect of NEXT depletion was enhanced in fractions of newly transcribed RNA purified from cells subjected to 1-h metabolic labeling with 5-bromouridine (BrU) (Meola et al., 2016). Surprisingly, analysis of these
BrU-RNA-seq data revealed that the majority of transcripts defined as PAXT PROMPTs, using the total RNA samples, lost their PAXT sensitivity in BrU-RNA-seq and instead became NEXT sensitive, whereas NEXT PROMPTs were upregulated largely equally between BrU and total RNA fractions (Figures 1D and S1E, left panel). These trends were also apparent from our selected PROMPT examples (compare Figure S1F with Figures 1A and S1A).

Hence, our initial analyses demonstrated that PROMPT regions are complex and express transcripts that can be targeted by both NEXT and PAXT pathways. Given the presence of NEXT- and PAXT-specific regions, these exosome adaptor systems do not appear to operate in a strictly redundant fashion. Still, our observation that siZFC3H1 treatment yielded a gradual increase in PAXT PROMPTs, becoming most visible in the total RNA fraction, and that the same PROMPTs were ZCCHC8

Figure 1. NEXT and PAXT Target Both Common and Distinct PROMPT Regions
(A) Genome browser views of RNA-seq data derived from total RNA samples of HeLa cells treated with siRNAs against GFP (siGFP control), RRP40 (siRRP40), ZCCHC8 (siZCCHC8), or ZFC3H1 (siZFC3H1), as indicated by color code. Only data from the strand of the depicted PAXT (top) or NEXT (bottom) PROMPTs are shown. RNA-seq signals represent the average of normalized values from triplicate libraries. Tracks of individual PROMPTs are shown on the same scale, and annotated genes are shown below the screen shots, with arrows indicating the major PROMPT TSSs as derived from siRRP40 cap analysis of gene expression (CAGE) data (Andersson et al., 2014b).
(B) qRT-PCR analysis of PAXT (top) and NEXT (bottom) PROMPTs from (A) and Figure S1A. Individual data points from three biological replicates are shown as circles. Bars and error bars represent mean values ± SD. Statistical analysis was performed using two-sided t tests. *p < 0.05; **p < 0.01; ***p < 0.001; ns, not significant.
(C) Scatterplot displaying PROMPT sensitivities to siZCCHC8 (x axis) and siZFC3H1 (y axis) conditions based on RNA-seq data from total RNA samples as in (A). Each dot represents a PROMPT. PAXT- and NEXT-sensitive PROMPTs, defined by y and x axis thresholds, are highlighted in blue and orange, respectively. Histograms showing distribution of sensitivities are added at the top and to the right of the scatterplot.
(D) Scatterplot and histograms as in (C) but for BrU RNA samples. Dot colors correspond to the classification from (C).
sensitive in the BrU-RNA fraction suggested that some redundancy exists.

3' End Patterns Differ Substantially between NEXT and PAXT PROMPTs

Given the non-trivial nature of PROMPT catabolism, we next aimed to measure the exact position as well as the polyadenylation status of the 3’ ends of both steady-state and newly synthesized RNA under relevant depletion conditions. To this end, HeLa cells were treated with siRNAs against GFP (control), RRP40, ZCCHC8, or ZFC3H1, and after sufficient factor depletion (Figure S2A), RNAs were labeled by a brief 10-min 4-thiouridine (4sU) pulse before harvesting. From these cell samples, both total (steady-state) and 4sU-labeled (nascent and newly synthesized) RNAs were isolated and subjected to pA⁺ RNA 3’ end sequencing (3’ end-seq) (pA⁺ libraries) directly or, after in vitro polyadenylation with E. coli pA polymerase (E-PAP) to capture both pA⁺ and pA⁻ RNA 3’ ends (pA⁻⁻ library) (Figures 2A, S2B, and S2C). Data were normalized to mRNA 3’ ends so that the 3’ ends of pA⁻⁺ RNA were equally abundant in pA⁺ and pA⁻⁻ libraries, whereas pA⁻ RNAs were only present in the pA⁺ libraries. Initial inspection of RNA 3’ end profiles within our previously selected PAXT PROMPT regions revealed well-defined peaks where signals concentrated in all analyzed libraries (Figures 2B and S2D, left panels). As expected, all of these 3’ ends were more strongly detected under siRRP40 and siZFC3H1 conditions within the same library type. Selected NEXT PROMPTs, on the other hand, showed striking spreading of their 3’ ends over 1- to 2-kb-wide regions, with substantial heterogeneity between individual libraries (Figures 2B and S2D, right panels). These 3’ ends were enriched in the siRRP40 and siZCCHC8 libraries and were generally more abundant in the pA⁻⁻ than in the pA⁺ libraries and in 4sU versus total RNA samples. Using S. cerevisiae, we demonstrated previously that pA⁻⁻ 4sU RNA samples contain nascent RNA 3’ ends (Schmid et al., 2018),...
which was likely to also be the case with the data analyzed here. Even so, signals within the pa\(^{-}\)–4sU RNA 3' end fraction were specifically increased upon both siRRP40 and siZCCHC8 treatment (Figures 2B and S2D; see also below), despite the fact that RRP40 depletion does not significantly alter Pol II recruitment to PROMPT regions (lasillo et al., 2017). Hence, we interpreted such increased signals to be due to post-transcriptional transcript stabilization. More generally, we conclude that 3' ends of the interrogated NEXT targets are generally poorly defined and only partially polyadenylated, whereas PAXT targets harbor well-defined and polyadenylated 3' ends.

To address the generality of these single-example observations, we characterized RNA 3' end signals within the regions of PAXT and NEXT PROMPTs classified above by RNA-seq data. For this analysis, we defined PROMPT transcript end sites (TESs) as individual positions that harbor the highest 3' end signal intensity within 5 kb downstream of the relevant PROMPT TSSs (STAR Methods). Metagene analysis anchored at these TES positions confirmed our RNA-seq data analysis, showing that PAXT PROMPT 3' ends were generally more abundant in control cells (Figure S2E) and further upregulated by siRRP40 and siZFC3H1 treatments in all library types, whereas NEXT PROMPT 3' ends were upregulated in siRRP40 and siZCCHC8 libraries, with the strongest relative increases in the pa\(^{-}\) and 4sU-labeled fractions (Figures 2C and S2E). In addition, the metagene plots revealed that the different patterning of NEXT and PAXT PROMPT 3' ends was general, with TES signals sharply focused around single positions for PAXT PROMPTs but much more spread for NEXT PROMPTs, which was especially pronounced in the pa\(^{-}\) libraries. Finally, and as initially noted using RNA-seq data (Figure 1D), PAXT PROMPTs displayed a pronounced sensitivity to siZCCHC8 treatment in the 4sU RNA samples, which was mostly visible in the pa\(^{-}\) libraries (Figure 2C, bottom left, and S3A). Interestingly, this NEXT sensitivity was not confined to the PAXT PROMPT TESs but extended to neighboring parts of the interrogated 2-kb region (Figure 2C, bottom left).

To provide robust 3' end classification, we extended the NEXT and PAXT RNA targeting analysis to a genome-wide inquiry. First, we aggregated neighboring signals into “3' end clusters” without considering gene annotations to permit unbiased analysis of 3' ends from both intergenic and genic regions (Figure S3B; STAR Methods). For each cluster, signals in siZFC3H1 and siZCCHC8 libraries were compared relative to signals from the siGFP controls, separately for each of the four library types (total RNA, pa\(^{-}/\)pa\(^{+}\); 4sU RNA, pa\(^{-}/\)pa\(^{+}\)). Clusters displaying significant upregulation (false discovery rate [FDR] < 0.1; STAR Methods) in at least one of these four siZFC3H1/siGFP comparisons, but no significant upregulation in any of the four siZCCHC8/siGFP comparisons, were classified as “PAXT 3' end clusters” and, vice versa, as “NEXT 3' end clusters.” This classification distinguished PAXT- and NEXT-dependent 3' ends more stringently than the above classification of PROMPTs (see below). Of the total 208,069 3' end clusters considered, 21,547 and 6,082 were classified as NEXT or PAXT specific, respectively. Taken together, NEXT and/or PAXT-sensitive clusters encompassed 38% of all total pa\(^{-}\) RNA 3' ends detectable in 3' end clusters, underscoring the large effect of the nuclear exosome on the transcriptome.

As observed for PAXT PROMPT TESs, the genomic set of all PAXT 3' end clusters generally constituted narrow regions with a clear center peak (Figure S3D, blue profiles), which included 3' ends from within intergenic regions, introns, and lncRNAs (Figure S3E, top panel). In contrast, NEXT 3' end clusters were heterogeneous (Figures S3D, orange profiles, and S3E, bottom panel), akin to the 3' end positions of NEXT PROMPTs. NEXT-sensitive 3' end clusters were also detected downstream of snRNA gene bodies and at the 3' splice sites of introns hosting snoRNAs (Figure S3C), which is consistent with the reported role of NEXT in targeting the exosome for decay and/or processing of snRNAs and snoRNAs (Andersen et al., 2013; Lubas et al., 2015). Finally, as also seen for PROMPT TESs, PAXT and NEXT 3' end clusters were most upregulated in total RNA and 4sU RNA samples, respectively (Figure S3D).

We further inquired about 3' end cluster behavior within eRNA transcription units, which have been reported previously to primarily produce NEXT targets but harbor high PAS densities (Andersson et al., 2014a; Chen et al., 2016; Lubas et al., 2011), and within the first introns of multi-exonic protein-coding genes, which reportedly harbor premature termination sites targeted by the PAXT pathway (lasillo et al., 2017; Kamieniarz-Gdula et al., 2019; Ogami et al., 2017). Unexpectedly, both NEXT and PAXT 3' end clusters were abundant in each of the annotation types, with appearances reminiscent of the genome-wide pattern (compare Figures 2D and 2E with Figure S3D), which was also apparent at the level of individual examples (Figure S3E, left panels). Whether intronic 3' ends are produced by premature termination of transcripts starting at the TSS of the analyzed genes or by exosomal decay of spliced-out introns could not be distinguished.

Taking all these analyses together, our observations from PROMPT regions could be extended to other RNA types, all containing 3' ends biased toward NEXT- or PAXT-dependent decay but without any particular biotype-dependent behavior. Within this general notion, PAXT-targeted 3' ends are generally extremely localized, whereas NEXT-targeted 3' ends span larger genomic regions. This presumably highlights marked distinctions in NEXT- and PAXT-directed RNA 3' end biology; in particular, that a potential redundancy between the NEXT and PAXT pathways might primarily be based on recognition of overlapping transcript isoforms, with the NEXT complex targeting heterogeneous 3' ends surrounding distinct PAXT-sensitive sites.

**PAXT- but Not NEXT-Sensitive 3' Ends Are Defined by Canonical PASs**

Because our analysis so far suggested that PAXT- but not NEXT-mediated decay initiates at well-defined 3' end positions, we wondered whether PAS-directed 3' end formation might be specific for PAXT-dependent exosome targets. To examine this possibility, we first compared the frequency of the consensus PAS hexamer motif AWTTAAA around the summit position of PAXT and NEXT 3' end clusters, which revealed that PAXT clusters contain strong enrichment of PAS hexamer sequences ~20 nt upstream of cluster summits, whereas NEXT clusters display less clear enrichment (Figure 3A). A smaller peak of AWTTAAA at both, NEXT and PAXT, cluster summits was due to the general enrichment of As and Ts around the PAS...
endonucleolytic cleavage position. That is, other A/T-rich motifs were equally enriched at this position, whereas AATAAA was the most enriched 6-mer sequence at positions −30 to −10 nt upstream of the cleavage site, as expected for canonical PASs (data not shown). Consistently, estimating PAS strength in the same genomic regions using a recently published neuronal network-based approach that interprets not only hexamer presence but also the local sequence context, including additional PAS-proximal motifs, demonstrated that NEXT 3’ end clusters generally display poorer PAS strengths than PAXT 3’ end clusters (Figure 3B; Bogard et al., 2019).

We then examined whether the observed PAS enrichment upstream of PAXT sensitive 3’ ends would also be reflected by the presence of relevant RNA binding proteins (RBPs). To this end, we compared cross-linking immunoprecipitation (CLIP) profiles from a CLIP database (CLIPdb; Yang et al., 2015) between NEXT and PAXT 3’ end clusters. Ranking of available bait proteins by their relative affinities to PAXT versus NEXT 3’ end clusters placed a subset of CPA components as the proteins most prone to bind PAXT 3’ ends (Figure 3C).

Although our analyses demonstrated a robust link between conventional PAS positioning and PAXT activity, lesser but still distinct PAS enrichment could also be observed upstream of NEXT 3’ end clusters (Figure 3A), which was even more apparent when exploring PAS motif density (Figure S4A) and strength (Figure S4B) upstream of the TESs of our defined PROMPT classes. It was also reflected by our set of individual PROMPTs, where all four PAXT PROMPTs contain an AWTAAA consensus sequence but after dividing each class into groups lacking (TES without hexamer, solid lines) or containing (TES with hexamer, dashed lines) AWTTAAA in the region −10 to −30 nt from the TES. The significance of differences in log2FC values at the TES (position 0) was verified by a one-sided log2FC(ties) test between the TESs of PAXT and NEXT groups. The inset shows the top eight PAXT-enriched CLIPdb datasets derived from factors related to PAS processing: CPSF (FIP1L and WDR35), CSTF (CSTF2 and CSTF2T), or CF1m (NUDT21, CPSF6, and CPSF7) complexes.

Metagene plots of 3’ end signals around the TESs of PAXT and NEXT PROMPTs as in Figure 2C but after dividing each class into groups lacking (TES without hexamer, solid lines) or containing (TES with hexamer, dashed lines) AWTTAAA in the region −10 to −30 nt from the TES. The significance of differences in log2FC values at the TES (position 0) was verified by a one-sided log2FC(ties) test for each of the panels. Higher log2FCs in PROMPT TESs with versus without hexamers were highly significant (p < 0.00001) in all siZCCHC8 library types for both PAXT and NEXT PROMPT TESs. Log2FCs were not significantly increased (p > 0.1) in most siZCCHC8 library types except for the total PA* RNA libraries, where the increased log2FCs for the subset with hexamer was significant for PAXT PROMPTs (p = 0.011) and NEXT PROMPTs (p = 0.000036).

Figure 3. PAXT- but Not NEXT-Targeted 3’ Ends Are Defined through PASs
(A) Frequency per nucleotide of the PAS hexamer motif AWTTAAA ± 100 nt around the summit of NEXT and PAXT 3’ end clusters.
(B) Combined violin density and boxplots depicting PAS scores derived as logit(prediction score) from the neuronal network model “APARENT” (APA [alternative polyadenylation] regression network) (Bogard et al., 2019) for PAXT versus NEXT PROMPT 3’ end cluster summits. ****p < 0.00001 from a two-sided Mann-Whitney U test.
(C) CLIPdb factor affinities ranked by log2FC within regions ± 100 nt around PAXT/NEXT 3’ end clusters. Individual CLIP bait proteins are listed on the x axis, and y axis values indicate log2FC between CLIP signals at PAXT versus NEXT 3’ end clusters. Clusters overlapping annotated splice junctions were omitted from analysis. The color code indicates FDR values from a two-sided Mann-Whitney U test between the PAXT and NEXT groups. The inset shows the top eight PAXT-enriched CLIPdb datasets derived from factors related to PAS processing: CPSF (FIP1L and WDR35), CSTF (CSTF2 and CSTF2T), or CF1m (NUDT21, CPSF6, and CPSF7) complexes.

(E) Frequency plot of AWTTAAA motifs in 500-nt intervals of the first 5,000 nt from the TSSs of NEXT (orange) and PAXT (blue) PROMPTs. (F) Mean PAS scores based on Bogard et al. (2019) as in (B) but for the regions in (E). Shaded areas show 95% confidence intervals. (G) Nucleotide frequencies for regions in (E) using 100-nt intervals.
Figure 4. Many PROMPTs Embody Both NEXT- and PAXT-Sensitive Transcript Isoforms

(A) Genome browser views of 3’ end-seq data as in Figure 2B but from four manually selected NEXT+PAXT PROMPT regions. Blue and orange bars below the browser views indicate regions of PAXT and NEXT sensitivity, respectively.

(B) Left: selection of NEXT+PAXT PROMPTs (green dots) from total RNA-seq data as in Figure 1C, based on sensitivity thresholds. Right: metagene profiles of 3’ end-seq data around the TESs of NEXT+PAXT PROMPTs. Display as in Figure 2C except that siZCCHC8 (dark green) and siZFC3H1 (light green) are shown in the same panel.

(legend continued on next page)
depletion (Figure S4D, compare the siRRP40 and siZCCHC8 tracks). Collectively, this raised the question of whether the PAS presence within NEXT substrate territory is functionally relevant for NEXT-dependent decay. In relation to this question, we noted that a low level of siZFC3H1 dependence was evident at NEXT PROMPT TESs (Figure 2C) and, to a lesser extent, at NEXT 3’ end cluster summits (Figure S3D), which suggested that such PAS occurrence might trigger PAXT-mediated RNA decay. To address this possibility more directly, we stratified the PROMPT classes into subgroups with or without an AWTAAA sequence within region −10 to −30 nt from the TES. Subsequent metagene analysis highlighted that siZFC3H1 treatment sensitivity at PROMPT TESS was strongest in the AWTAAA-containing subgroups of both PAXT and NEXT PROMPTs (Figure 3D, left panels). Sensitivity to siZCCHC8 treatment, on the other hand, appeared to be independent of the presence of an AWTAAA sequence (Figure 3D, right panels). Similar results were obtained when sub-classing 3’ end clusters (Figure S4E). Taken together, this strongly implied that NEXT-sensitive 3’ ends are not directed by PASs and that “PAS occurrence” within regions of NEXT sensitivity reflects overlapping PAXT action at CPA-produced 3’ ends.

We then wondered whether PAS positioning at PAXT- but not NEXT-sensitive 3’ ends could be explained by genomic context; i.e., whether NEXT PROMPT regions are generally depleted for PAS motifs. To analyze this, we measured the abundance of the PAS hexamer motif AWTAAA starting from the TSSs of PAXT or NEXT PROMPTs. In addition, we also measured PAS strength in the same genomic regions (Bogard et al., 2019). Surprisingly, these analyses demonstrated that NEXT PROMPTs, on average, contain as many, if not more, AWTAAA motifs (Figure 3E) and ensuing higher PAS probabilities (Figure 3F). This mirrored a general nucleotide bias because TSS-proximal regions of NEXT PROMPTs were found to be generally more AT-rich compared with their PAXT PROMPT counterparts (Figure 3G).

Taking these results together, we conclude that PAXT-targeted 3’ ends are generally created by the CPA machinery acting on canonical PASs. The exact mechanism producing NEXT-dependent 3’ ends is more enigmatic, but evidence so far indicates that these are produced independent of the CPA complex. Even though Pol II encounters PAS sequences during transcription, this implies that these are either not recognized adequately or do not produce stable polyadenylated 3’ ends.

**PROMPT Regions Commonly Express Overlapping and Pathway-Specific Transcript Isoforms**

The PROMPT connection and the NEXT complex form mutually exclusively (Meola et al., 2016), and our analyses so far indicated that PAXT and NEXT targets are separable. However, the majority of analyzed regions express transcripts sensitive to both siZFC3H1 and siZCCHC8 treatment, which was particularly evident by the siZCCHC8 responsiveness of PAXT PROMPTs in the BrU and 4sU RNA datasets (Figures 1D and 2C). To understand how individual transcription units are targeted by two seemingly independent pathways, we first interrogated four PROMPT regions displaying dual NEXT and PAXT sensitivities (“NEXT+PAXT PROMPTs”) in both BrU and total RNA-seq samples (Figure S5A). Although these PROMPT regions displayed extensive 3’ end spreading, like those of NEXT PROMPTs, many 3’ ends accumulated around well-defined PAXT-sensitive summits, with the surrounding 3’ ends responding only to NEXT depletion (Figure 4A). This phenomenon was also evident from metagene analysis of NEXT+PAXT PROMPTs selected from RNA-seq data (Figure 4B) and for genome-wide 3’ end clusters responsive to both ZCCHC8 and ZFC3H1 depletion (Figure S5B). Moreover, the four individual PAXT-sensitive 3’ end summits all arose 17–28 nt downstream of a PAS hexamer (Figure S5C) and were enriched in the pA+ libraries, whereas the surrounding 3’ end signal was most enriched in 4sU/pA+− samples (Figure 4A). It therefore appeared that NEXT+PAXT PROMPT regions produce distinct transcript isoforms that are individually targeted by NEXT or PAXT pathways in a manner resembling the targeting of transcripts from regions biased toward expressing only one of the pathway isoforms. To further validate the notion that 3’ ends within NEXT+PAXT PROMPT regions surrounding the PAXT-sensitive summit are targeted by NEXT, we conducted qRT-PCR analysis using primer sets targeting regions upstream (5’ amplicon) or downstream (3’ amplicon) of the summits (Figure 4C, top panels). In line with the 3’ end seq data, ZFC3H1 depletion resulted in increased signals only when monitored with the 5’ amplicons (Figure 4C, left panel), whereas both RRP40 and ZCCHC8 depletion caused increased RNA output, as measured with both 5’ and 3’ amplicons (Figure 4C).

Taken together, we conclude that NEXT+PAXT sensitive loci are common and produce transcript isoforms with pathway-specific sensitivities similar to those of NEXT- and PAXT-biased regions.

**PAXT Targeting of NEXT Substrates in NEXT-Depleted Cells**

Given that NEXT and PAXT target transcripts from common loci, we next investigated directly whether these exosome adaptors may also redundantly target the same transcript isoforms. To this end, we treated HeLa cells with siRNAs co-depleting ZCCHC8 and ZFC3H1 (siZCCH8-siZFC3H1) and compared this with cell samples treated with either siGFP control or individual ZCCHC8, ZFC3H1, or RRP40 siRNAs as before (Figure S6A). qRT-PCR analysis of total RNA samples using the 5’ and 3’ amplicons from Figure 4C revealed an additive increase of NEXT+PAXT PROMPTs upon siZCCH8-siZFC3H1 double depletion (Figure 5A). Interestingly, however, this was not only observed when using the 5’ amplicon (Figure 5A, left panel), detecting both NEXT- and PAXT-specific isoforms, but also when using the NEXT-specific 3’ amplicon (Figure 5A, right panel). Similarly, NEXT PROMPT levels were also elevated in the

(C) Experimental validation of NEXT versus PAXT 3’ end sensitivities. Top: scheme depicting the prediction that PAXT sensitive transcript isoforms (blue arrow) are PAS directed and will be detected only with the 5’ amplicon, whereas extended NEXT-sensitive isoforms (orange arrow) will be detected with both the 5’ and 3’ amplicons. Bottom: qRT-PCR analysis of the NEXT+PAXT PROMPTs from (A) using the 5’ and 3’ amplicons. Presentation as in Figure 1B.
Figure 5. PAXT Redundantly Targets NEXT Substrates

(A) qRT-PCR analysis of NEXT+PAXT PROMPT levels in cells subjected to the indicated single or double depletions. Presentation as in Figure 4C.
(B) qRT-PCR analysis of NEXT PROMPTs as in (A).
(C) qRT-PCR analysis of PAXT PROMPTs as in (A).
(D) Metagene profiles of 3' end-seq data at NEXT+PAXT, NEXT, and PAXT PROMPT TESs. Log$_2$FC values from siZCCHC8 (green), siZFC3H1 (light blue), and siZCCHC8-siZFC3H1 (dark blue) samples relative to their siGFP controls are shown.
(E) qRT-PCR analysis as in (B) but of NEXT PROMPT levels in cells subjected to the indicated single and double depletions. Note the similar effect of co-depleting any individual PAXT component with ZCCHC8.
siZCCHC8-siZFC3H1 compared with the siZCCHC8 depletion sample and approaching levels of the siRRP40 sample (Figure 5B), whereas PAXT PROMPT levels did not markedly change between siZCCHC8-siZFC3H1 and siZFC3H1 samples (Figure 5C). Altogether, this provided an explanation for our previous observation that NEXT PROMPTs were only partially stabilized in siZCCHC8 compared with siRRP40 samples (Figure 1) because efficient decay of NEXT PROMPTs requires inactivation of not only NEXT but also of PAXT.

Intrigued by these observations, we prepared total and 4sU RNA from siZCCHC8-siZFC3H1 samples and converted them into 3' end-seq libraries that were directly comparable with those from single-factor depletion samples (Figure S6B). As expected from the qRT-PCR experiments, metagene analysis of NEXT+PAXT PROMPT TESs showed their enhanced accumulation in the siZCCHC8-siZFC3H1 compared with the siZCCHC8 and siZFC3H1 libraries (Figure 5D, NEXT+PAXT). This effect was most evident for the pA+/4sU sample with particularly elevated levels of RNA 3' ends in NEXT-sensitive territory surrounding PROMPT TESs. Similar additive NEXT-PAXT sensitivity could be observed for NEXT PROMPTs (Figure 5D, NEXT) and for NEXT-sensitive regions surrounding the TESs of PAXT PROMPTs (Figure 5D, PAXT). Again, the enhanced stabilization of such NEXT-sensitive 3' ends from the siZCCHC8-siZFC3H1 condition was most pronounced in the pA+/4sU sample but also clearly significant for total RNA pA+ ends (Figures 5D and S6C). In general, such additive effects were also visible at single NEXT-sensitive PROMPT loci (Figure S6D), and similar results were obtained when comparing signals from 3' end cluster classes (Figure S6E).

We recently identified the additional PAXT components PABPN1 and ZC3H3 and the paralogous RBM26 and RBM27 factors as being required for PAXT-mediated RNA decay (Meola et al., 2016; Silla et al., 2020) and therefore inquired whether additive upregulation of NEXT targets could also be achieved upon inactivation of these PAXT components. Although their individual depletion (Figure S6F) led to only minor upregulation of interrogated NEXT PROMPTs, co-depletion of any of these factors with ZCCHC8 (Figure S6F) resulted in an additive increase of NEXT PROMPT levels (Figure 5E). Hence, PAXT components generally cause enhanced stabilization of NEXT targets when co-depleted with siZCCHC8.

We conclude that PAXT and NEXT cooperate to ensure efficient decay of cryptic transcripts. More specifically, PAXT appears to provide a “second line of defense” for exosome targets that, for one reason or another, evade early NEXT-mediated decay. PAXT-specific 3’ ends, on the other hand, are not affected by NEXT.

**Polyadenylation Is Required for Handover of NEXT Targets to PAXT**

How is the PAXT connection capable of targeting NEXT substrates when these are only poorly adenylylated, if at all? We previously noted that NEXT-sensitive 3' ends could be detected as pA+ species, especially under siRRP40 and siZCCHC8 depletion conditions (Figures 2B, 2C, 4A, and 4B), which prompted us to investigate whether handover of NEXT targets to PAXT-mediated decay might require their 3’ end adenylation. To this end, we first inquired whether our respective factor depletions would affect the polyadenylation status of individual transcripts. This was done by directly comparing the abundance of pA+ relative to pA− RNAs within the established PAXT and NEXT 3’ end clusters. Although these analyses underscored the efficient polyadenylation of PAXT targets in all libraries (Figure 6A), they also uncovered a striking effect of siZCCHC8 and siRRP40 depletion on NEXT target polyadenylation. That is, even though the pA+ RNA fraction varied between conditions, NEXT targets were consistently more polyadenylated in both RRP40- and ZCCHC8-depleted samples (Figure 6A, compare siRRP40, siZCCHC8, and siZCCHC8/siZFC3H1 with siGFP).

This implied that polyadenylation of NEXT targets is a post-transcriptional event prevented by early NEXT/exosome-dependent decay and that such polyadenylation of transcripts failing to undergo efficient decay (as, for example, under the siZCCHC8 condition) might allow PAXT recognition. In this scenario, NEXT targets would theoretically accumulate as polyadenylated species upon ZCCHC8/ZFC3H1 co-depletion, which was evident for NEXT 3’ end clusters in both total and 4sU RNA samples (Figure 6A). To more directly analyze whether polyadenylation would be required for PAXT-mediated decay of NEXT substrates, we treated ZCCHC8-depleted cells with the chain-terminating derivative of adenosine, cordycepin, which blocks RNA polyadenylation and has been shown previously to inhibit PAXT-dependent RNA decay in vivo (Bresson et al., 2015; Meola et al., 2016). In accordance, cordycepin treatment led to additive accumulation of NEXT PROMPTs, much like the siZCCHC8/siZFC3H1 double depletion condition (Figure 6B). Moreover, cordycepin treatment did not cause any additional PROMPT stabilization in siZCCHC8/siZFC3H1-treated cells. This strongly indicated that polyadenylation of NEXT targets allows their fail-safe decay by the PAXT-dependent pathway (Figure 6C).

**DISCUSSION**

Target recognition and decay by the nucleoplasmic RNA exosome requires particular adaptors: the NEXT complex and the PAXT connection (Lubas et al., 2011; Meola et al., 2016; Ogami et al., 2017; Silla et al., 2020). Previous characterization of NEXT and PAXT activities was predominantly based on an RNA biotype-centric division of substrates based on their genomic location and length, whereas exact substrate properties that provide pathway specificity were left uncharacterized. However, this remained an important problem that needed to be resolved toward offering an understanding of which abilities eukaryotic cells possess to handle the pervasive RNA production from their genomes. We demonstrate here that NEXT substrates arise from poorly defined and heterogeneous 3’ ends of predominately pA− RNAs, whereas PAXT targets pA+ RNAs deriving from canonical PAS use. Despite these distinct substrate features, the two adaptor complexes act complementarily by amply targeting of overlapping genomic regions and even, in some cases, by the ability to target the same transcript isoforms (Figure 6C). This two-layered targeting provides efficient removal of unwanted and spurious transcription products in the human cell nucleus.

Our analyses reiterate previous findings that NEXT targets are overall less abundant and mature than PAXT targets (Meola et al., 2016).
et al., 2016). That said, direct transcription measurements demonstrated that Pol II initiation at NEXT- versus PAXT-sensitive loci is comparable (Figure S1D). Instead, mNETseq and RNA-seq indicated that NEXT-sensitive loci are only transcribed for short distances, often less than 1–2 kb (Figures S1C and S1D). Taken together with the low steady-state abundance of, e.g., NEXT-sensitive PROMPT in control cells, this all points to rapid decay of NEXT-targeted transcripts, possibly shortly after their transcriptional termination. How such efficiency is achieved remains a matter of speculation (see below).

PAXT targets, on the other hand, generally accumulate to higher levels, compatible with longer nuclear half-lives. This is in line with PAXT primarily targeting pA⁺ RNA produced by the CPA machinery, which is further consistent with the importance of PABPN1 for PAXT-dependent RNA decay. Because pA tail addition and PABPN1 association are also critical for the stability and nuclear export of RNA from, e.g., protein-coding loci (Tudek et al., 2018; Wigington et al., 2014), a nuclear timer model has been suggested (Libri, 2010; Meola and Jensen, 2017), which posits that assembly of the full PAXT connection onto target RNA is in competition with nuclear export of the same transcript. Hence, only poorly exported RNAs attain a nuclear residence time long enough to allow formation of a capable exosome-recruiting PAXT complex. Although this might explain the longer nuclear half-lives of PAXT versus NEXT substrates, it concomitantly raises the question of how nuclear export of PAXT
substrates is prevented. Here it is interesting to note that, even though PAXT targets contain 5’ end caps and 3’ end pA tails, both of which stimulate export (Stewart, 2019; Wickramasinghe and Laskey, 2015); these transcripts are still generally shorter and less intron-containing than typical protein-coding RNAs, which likely causes their less efficient exit from the nucleus. Taken together with their apparent canonical employment of the CPA complex at consensus-positioned PASs as well as their robust 3’ end polyadenylation, we favor the notion that PAXT substrates are subject to exosomal decay because of their nuclear export incompetence rather than their inefficient 3’ end processing.

In contrast to PAXT-targeting, the mechanism(s) producing NEXT-sensitive 3’ ends are less clear. We and others have suggested previously that TSS-proximal PASs direct, at least in part, PROMPT and eRNA 3’ end formation in preparation for exosome targeting (Almada et al., 2013; Andersson et al., 2014a; Ntini et al., 2013). In possible contradiction to this, it was suggested that eRNA 3’ ends are formed by cleavage of the integrator complex (Lai et al., 2015), best known for its production of pA 3’ ends of snRNAs (Bailat and Wagner, 2015; Guiro and Murphy, 2017). The results presented here now reconcile these previous findings by showing that many PROMPT and eRNA loci express both NEXT- and PAXT-targeted 3’ ends (Figures 2C and 2D). Moreover, PASs often reside within these NEXT- and PAXT-sensitive regions; however, their correspondingly polyadenylated 3’ ends are only susceptible to PAXT targeting. Still, this is in line with previous observations because NEXT-sensitive pA 3’ ends are regularly scattered around the stronger PAXT-dependent signals. This explains why PAS-independent PROMPT 3’ ends were previously overlooked and establishes that PROMPTs and eRNAs can be targeted by NEXT and/or PAXT, depending on the exact genomic sequence context.

More noteworthy perhaps, is that our present study provides a radically different view of possible NEXT targeting mechanisms. Substrates of this adaptor complex have poorly defined and largely non-adenylated 3’ ends, often appearing in a spread-out fashion from the RNA-producing TSS. Although it cannot be excluded that a fraction of these 3’ ends arise by ribonuclease nibbling from fewer major sites, we note that they also appear downstream of PAXT-targeted 3’ ends, which, consequently, cannot trigger their appearance. We also note that levels of these heterogenous 3’ ends are elevated upon both NEXT and exosome inactivation. As a consequence, they cannot simply be termini of nascent RNAs tied in actively transcribing RNA polymerase, and we suspect that instead they are produced by promiscuous transcription termination. These considerations therefore suggest that the early transcription elongation process is very termination prone. Our analysis of protein-coding genes argues that this phenomenon is not just restricted to loci producing cryptic transcripts. Moreover, it suggests that TSS-proximal termination events may extend well beyond previously reported examples (Eirod et al., 2019; Iasillo et al., 2017; Kamieniarz-Gdula et al., 2019; Ogami et al., 2017; Tatomer et al., 2020).

How is NEXT-mediated exosome recruitment coupled to such promiscuous transcription termination? Although the situation conceptually resembles the transcription termination-coupled degradation of short S. cerevisiae cryptic unstable transcripts (CUTs) by the Nrd1-Nab3-Sen1 (NNS) and TRF4/5-Air1/2-Mtr4 polyadenylation (TRAMP) complexes (LaCava et al., 2005; Porrua and Libri, 2015), comparable links in human cells are unlikely because the NNS complex does not appear to be conserved in higher eukaryotic cells (Gregersen et al., 2019). Moreover, in contrast to the direct interactions of the NNS and TRAMP complexes with S. cerevisiae Pol II (Porrua and Libri, 2015), NEXT components have no reported physical links to factors of the human transcription elongation and Pol II machineries. Instead, it has been suggested that NEXT might be recruited early in the RNA synthesis process via its presence in the so-called CBC-NEXT (CBCN) complex assembling on capped nascent RNA (Andersen et al., 2013; Giacometti et al., 2017; Hallais et al., 2013). Still, whether the CBC-NEXT linkage provides a sufficiently strong platform for early exosome recruitment to deal with the promiscuous by-products of early transcription elongation failure remains to be further elucidated.

It also remains to be investigated how NEXT-dependent 3’ ends arise. Their heterogeneous nature indicates that 3’ end formation does not involve a sequence-specific endonuclease. Instead, these ends may arise from stochastic termination of Pol II occurring during early and termination-prone transcription. It is tempting to speculate that the high AT density of NEXT-sensitive regions plays a role, as suggested by previous studies for such sequence bias (Nojima et al., 2013; White et al., 2013). In the same vein, the presence of PASs in such regions might be instrumental to facilitate Pol II pausing in preparation for transcription termination (Proudfoot, 2016). Upcoming research will address these possibilities. We previously reported that depletion of CBP80 of the CBC and its interactor, the ARS2 protein, lead to transcriptional read-through at selected PROMPT and eRNA loci (Andersen et al., 2013; Iasillo et al., 2017). Given that CBC and ARS2, forming the CBCA complex (Andersen et al., 2013; Hallais et al., 2013), also contact NEXT and PAXT components, the CBCA complex could, in principle, serve a dual role in transcription termination and RNA decay, reminiscent of the S. cerevisiae NNS complex.

Regardless of their exact origin, the majority of NEXT-targeted pA RNA ends is efficiently repressed by nuclear RNA quality control. Although this task is primarily facilitated by NEXT-dependent recruitment of the RNA exosome, we provide evidence that the PAXT connection can serve as a backup system when substrates evade the NEXT complex. This conclusion is based on several observations: (1) the effect on NEXT targets by ZCCHC8 depletion was consistently weaker than upon RRP40 depletion and only reached siRRP40 levels when ZCCHC8 was co-depleted with ZCF3H1, (2) NEXT targets became adenylated in siZCCHC8 and siRRP40 conditions, and (3) treating ZCCHC8-depleted cells with cordycepin mimicked the siZCCHC8/ siZFC3H1 double depletion phenotype. Based on these findings, we propose that substrate handover to PAXT is a consequence of the 3’ end adenylation of NEXT targets. Which pA polymerases (PAPs) facilitate this activity remains to be shown. Our individual and combined depletion of PAPs and terminal uridylyl
transferases (TUTs) did not yield clear evidence for specific enzymes involved in this process (data not shown). It may therefore be that nuclear PAPs and TUTs act redundantly to adenylate unprotected 3’ ends. However, depletion of canonical PAPs, and to some extent TUTs, is likely to have a broad effect on RNA metabolism, rendering interpretation of such experiments ambiguous. Whatever the exact mechanism, the PAXT connection provides a fail-safe ensuring tight post-transcriptional control of cryptic transcription.

In conclusion, we provide evidence that the nuclear exosome adaptors NEXT and PAXT co-operate to ensure efficient removal of a wide range of cryptic transcripts that, taken together, comprise a substantial fraction of the HeLa cell transcriptome. Distinct targeting principles and extensive overlap of NEXT and PAXT activities can be observed at essentially all Pol II-dependent gene annotation types, which likely reflects the critical importance of efficient removal of cryptic transcripts. Given the vast and complex transcriptome, additional players may participate in this process, perhaps comprising additional exosome adaptors and/or endonuclease or 5’-3’ exonucleolytic pathways. Finally, many IncRNAs are expressed in a highly tissue-specific manner, warranting cell-type-specific strategies to evade nuclear decay, which may involve controlling NEXT and PAXT activities in ways still to be uncovered.

ACKNOWLEDGMENTS

We thank Søren Lykke-Andersen, Peter Refsing Andersen, and Domenico Libri for critical comments on the manuscript. Work in the T.H.J. laboratory was supported by ERC grant 339953, the Novo Nordisk Foundation, and the Lundbeck Foundation. The A.S. laboratory was supported by the Lundbeck Foundation.

AUTHOR CONTRIBUTIONS

G.W., P.P., and N.M. performed wet laboratory experiments. M.S. and L.R. performed computational analyses. G.W., M.S., and T.H.J. conceived the project. T.H.J. and A.S. supervised the project. G.W., M.S., and T.H.J. wrote the paper. All authors read and approved the manuscript.

DECLARATIONS OF INTERESTS

The authors declare no competing interests.

REFERENCES

Insight into the RNA Exosome Complex Through Modeling Pontocerebellar Hypoplasia Type 1b Disease Mutations in Yeast. Genetics 205, 221–237.


## STAR★METHODS

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibodies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse monoclonal anti-RRP40</td>
<td>Santa Cruz</td>
<td>Cat# SC-101038; RRID: AB_1122700</td>
</tr>
<tr>
<td>Rabbit polyclonal anti-ZCCHC8</td>
<td>Novus Biologicals</td>
<td>Cat# NB100-94995; RRID: AB_1262274</td>
</tr>
<tr>
<td>Rabbit polyclonal anti-ZFC3H1</td>
<td>Sigma-Aldrich</td>
<td>Cat# HPA-007151; RRID: AB_1846133</td>
</tr>
<tr>
<td>Rabbit monoclonal anti-PABPN1</td>
<td>Abcam</td>
<td>Cat# ab75855; RRID: AB_1310538</td>
</tr>
<tr>
<td>Rabbit polyclonal anti-ZC3H3</td>
<td>Sigma-Aldrich</td>
<td>Cat# HPA023658; RRID: AB_1858973</td>
</tr>
<tr>
<td>Rabbit polyclonal anti-RBM26</td>
<td>Atlas Antibody</td>
<td>Cat# HPA040252; RRID: AB_10672895</td>
</tr>
<tr>
<td>Rabbit polyclonal anti-RBM27</td>
<td>Bethyl Laboratories</td>
<td>Cat# A301-233A; RRID: AB_890638</td>
</tr>
<tr>
<td>Mouse monoclonal anti-Tubulin</td>
<td>Rockland</td>
<td>Cat# 200-301-880; RRID: AB_10705001</td>
</tr>
<tr>
<td>Chemicals, Peptides, and Recombinant Proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cordycepin</td>
<td>Sigma-Aldrich</td>
<td>Cat# C3394</td>
</tr>
<tr>
<td>4-thiouridine</td>
<td>Carbosynth</td>
<td>Cat# 13957-31-8</td>
</tr>
<tr>
<td>EZ-Link™ HPDP-Biotin</td>
<td>Thermo Fisher Scientific</td>
<td>Cat# 21341</td>
</tr>
<tr>
<td>Critical Commercial Assays</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TURBO DNA-free kit</td>
<td>Invitrogen</td>
<td>Cat# AM1907</td>
</tr>
<tr>
<td>SuperScript™ III reverse transcriptase</td>
<td>Invitrogen</td>
<td>Cat# 18080044</td>
</tr>
<tr>
<td>MEGAscript RNAi Kit</td>
<td>ThermoFisher Scientific</td>
<td>Cat# AM1626</td>
</tr>
<tr>
<td>Platinum SYBR Green qPCR SuperMix-UDG</td>
<td>Invitrogen</td>
<td>Cat# 11733046</td>
</tr>
<tr>
<td>μMacs streptavidin kit</td>
<td>Miltenyi Biotec</td>
<td>Cat# 130-074-101</td>
</tr>
<tr>
<td>miRNeasy Micro kit</td>
<td>QIAGEN</td>
<td>Cat# 217084</td>
</tr>
<tr>
<td>E. coli poly(A) polymerase</td>
<td>Invitrogen</td>
<td>Cat# AM2030</td>
</tr>
<tr>
<td>RiboLock RNase Inhibitor</td>
<td>ThermoFisher Scientific</td>
<td>Cat# EO0381</td>
</tr>
<tr>
<td>PureLink micro RNA purification kit</td>
<td>ThermoFisher Scientific</td>
<td>Cat# 12183016</td>
</tr>
<tr>
<td>Quant Seq 3’mRNA-Seq REV library prep kit</td>
<td>Lexogen GmbH</td>
<td>Cat# 016</td>
</tr>
<tr>
<td>ERCC RNA spike-in mix</td>
<td>Invitrogen</td>
<td>Cat# 4456740</td>
</tr>
<tr>
<td>RibocCop rRNA Depletion kit</td>
<td>Lexogen GmbH</td>
<td>Cat# 037.96</td>
</tr>
<tr>
<td>Lipofectamine 2000 Reagent</td>
<td>Invitrogen</td>
<td>Cat# 11668500</td>
</tr>
<tr>
<td>Deposited Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw and analyzed data</td>
<td>This paper</td>
<td>GEO: GSE137612</td>
</tr>
<tr>
<td>RNaseq datasets</td>
<td>Meola et al., 2016</td>
<td>GEO: GSE84172</td>
</tr>
<tr>
<td>CAGE data</td>
<td>Andersson et al., 2014b</td>
<td>GEO: GSE62047</td>
</tr>
<tr>
<td>NETseq data</td>
<td>Schlackow et al., 2017</td>
<td>GEO: GSM2357382</td>
</tr>
<tr>
<td>CLIP binding data</td>
<td>Yang et al., 2015</td>
<td><a href="http://lulab.life.tsinghua.edu.cn/postar/">http://lulab.life.tsinghua.edu.cn/postar/</a></td>
</tr>
<tr>
<td>Experimental Models: Cell Lines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human: HeLa cells</td>
<td>Torben Heick Jensen lab</td>
<td>N/A</td>
</tr>
<tr>
<td>Oligonucleotides</td>
<td></td>
<td></td>
</tr>
<tr>
<td>siRNA sequences</td>
<td>see Table S2</td>
<td>N/A</td>
</tr>
<tr>
<td>q-PCR primer sequences</td>
<td>see Table S3</td>
<td>N/A</td>
</tr>
<tr>
<td>Recombinant DNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pUC19- ERCC-00170</td>
<td>Schwab et al., 2016</td>
<td>N/A</td>
</tr>
<tr>
<td>pUC19- ERCC-00002</td>
<td>Schwab et al., 2016</td>
<td>N/A</td>
</tr>
<tr>
<td>pUC19- ERCC-00145</td>
<td>Schwab et al., 2016</td>
<td>N/A</td>
</tr>
<tr>
<td>pUC19- ERCC-00092</td>
<td>Schwab et al., 2016</td>
<td>N/A</td>
</tr>
<tr>
<td>pUC19- ERCC-00136</td>
<td>Schwab et al., 2016</td>
<td>N/A</td>
</tr>
<tr>
<td>pUC19- ERCC-00043</td>
<td>Schwab et al., 2016</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(Continued on next page)
LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Torben Heick Jensen (thj@mbg.au.dk). All unique/stable reagents generated in this study are available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HeLa Kyoto cells were cultured in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. Cells were maintained at 37°C and 5% CO₂.

METHOD DETAILS

siRNA-mediated depletions and cordycepin treatment
siRNA transfections were carried out using lipofectamine 2000 (Invitrogen) according to the manufacturer’s protocol. All siRNA sequences are listed in Table S2. Cells were treated with 20 nM siRNA for 4 days, including a re-transfection 2 days after the initial transfection. For cordycepin treatment, cells were cultured in DMEM containing 20 μg/ml cordycepin (C 3394; Sigma-Aldrich) for 3 hr before harvesting.

Western blotting analysis
Cells were lysed with lysis buffer (10 mM Tris-Cl pH 7.4, 100 mM NaCl, 2.5 mM MgCl₂, 0.5% NP-40, 0.5% Triton X-100) on ice for 10 min, then centrifuged at 12000 rpm for 20 min. The protein concentration in the supernatant was measured using Bradford solution (Bio-Rad). Equal amounts of proteins were loaded onto PAGE gels. After running, proteins were transferred to PVDF membranes, which were blocked with 5% skimmed milk/PBS-T for 1 hr at room temperature (RT), and then incubated with primary antibodies (see Key Resources Table) diluted in PBS-T at 4°C overnight, followed by washing 3 x 10 min with PBS-T. Membranes were then incubated with HRP-conjugated secondary antibodies diluted in PBS-T for 1 hr at RT, followed by washing 3 x 10 min with PBS-T. SuperSignal West Femto HRP substrate (ThermoFisher Scientific) was applied to the membranes and the signal was detected with X-Ray film (Konica Minolta).

RNA isolation and RT-qPCR analysis
RNA was extracted using TRIzol (Invitrogen) and treated with TURBO DNase (Invitrogen) following the manufacturer’s instructions. RT-qPCR was performed using SuperScript™ III reverse transcriptase (Invitrogen) with 1 μg RNA and a mix of 80 pmol random primers and 20 pmol dT20 primers. qPCR was performed using Platinum SYBR Green qPCR SuperMix-UDG (Invitrogen) in a AriaMx Real-Time PCR machine (Agilent Technologies) using the primers listed in Table S3. qPCR data were normalized to GAPDH and siGFP control.

4sU RNA labeling and purification
The 4sU RNA labeling and purification method was modified from Schwalb et al. (2016). Briefly, HeLa cells were labeled in DMEM medium with 500 μM 4-thiouridine (4sU, Carbosynth) for 10 min, and harvested with TRIzol. Six spike-in RNAs were cloned from ERCC spike-in and in vitro transcribed using the MEGAscript RNAi Kit (ThermoFisher Scientific) with 0.6 mM 4sUTP (ERCC-00092, ERCC-00136 and ERCC-00043) or without 4sUTP (ERCC-00170, ERCC-00002 and ERCC-00145). In vitro transcribed...
RNAs were purified by phenol extraction and ethanol precipitation. RNA concentration was measured by a NanoDrop Spectrophotometer (Thermo Fisher Scientific) and equal amounts of spike-in transcripts were mixed. 24 ng of this mix was added to 400 μg total HeLa cell RNA before 4sU RNA purification for measurements of 4sU RNA purification efficiencies. For 4sU RNA purification, 400 μg total RNA was biotinylated with 0.2 mg/ml HPDP-biotin (Thermo Fisher Scientific) in biotinylation buffer (10 mM Tris, pH7.5; 1 mM EDTA) for 1.5 hr in a thermodenicer while shaking at 750 rpm at 24 °C in the dark. DNA was then cleaned up with one chloroform extraction followed by isopropanol precipitation and resuspension in 100 μl H2O. Biotinylated RNA was purified by incubating with 100 μL 100 μM Macs streptavidin beads (Miltenyi Biotec) for 16 min with gently shaking at 24 °C in the dark. Beads were then applied to a 1 γMACS column (Miltenyi Biotec) and washed with washing buffer (100 mM Tris, pH7.5; 10 mM EDTA; 1 M NaCl and 0.1% Tween 20) 3 times at 65 °C and 3 times at 25 °C. Finally, the purified RNA was eluted from beads with 200 μl of 100 mM DTT and cleaned up using the miRNeasy Micro kit (QIAGEN) with on-column DNase treatment to remove genomic DNA.

**In vitro polyadenylation**

1 μg TURBO DNase treated total RNA or 250 ng 4sU labeled RNA were incubated with or without (control) *E. coli* poly(A) polymerase (E-PAP, Invitrogen) at 30 °C for 30 min in 20 μl reactions, containing 1 × E-PAP buffer, 2.5 mM MnCl2, 0.4 U polymerase, 0.8 μM RiboLock RNAse inhibitor and 1 mM ATP. RNA was purified using PureLink micro RNA purification kit (Ambion) following the manufacturer’s instructions.

**Library preparation and RNA 3' end-seq**

The quality of total RNA was checked on an Agilent 2100 bioanalyzer. 1 μg total RNA or 250 ng 4sU RNA were used as starting material for each library. For total RNA samples an additional 2 μl of 1:100 diluted commercially available ERCC RNA spike-in mix was added. From each RNA sample 1 3' end-seq library was prepared without further end modification (pA +), while the other sample was treated with E-PAP (pA -) as described above. All RNAs were then RNA-depleted using RiboCop rRNA Depletion kit (Lexogen GmbH), according to the manufacturer’s instruction. Libraries were made using Quant Seq 3' mRNA-Seq REV library prep kit (Lexogen GmbH), according to the manufacturer’s instruction and multiplexed based on the concentrations measured by a Qubit Fluorometer (ThermoFisher Scientific) and a Bioanalyzer (Agilent 2100). 8 or 10 libraries were multiplexed and sequenced at the Vienna Biocenter Core Facilities with a NextSeq SR75 High Output run. Libraries from total and 4sU RNA samples were produced in triplicate and duplicate, respectively for all depletions, except for the siGFP control, which was included in all batches and therefore present in 5 (total RNA samples) and 4 (4sU RNA samples) replicates, respectively (see Table S4). Sequencing data was deposited to GEO: GSE137612.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**GEO datasets**

RNAseq datasets were described in Meola et al. (2016) and are available from GEO: GSE84172. Raw reads were downloaded from SRA and trimmed using Trimomatic v0.32 (Bolger et al., 2014) with settings 'PE ILLUMINA CLIP:/com/extra/Trimmomatic/0.32/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:12 LEADING:22 SLIDINGWINDOW:4:22 MINLEN:25' as described (Meola et al., 2016). Paired trimmed reads were then mapped to GRCh38 using HISAT2 v2.1.0 (Kim et al., 2019) using the genome index available from the HISAT2 website for UCSC hg38 and Refseq gene annotations with otherwise default settings. Mapped reads in proper pairs were selected using samtools view v.1.6.0 (Li et al., 2009) and genomic coverage obtained using bedtools genomecov v.2.25.0 (Quinlan and Hall, 2010) with setting '-split' and TPM-normalized by scaling raw read coverage by 1000000/(number of mapped reads). Genome browser views show average TPM values from the 3 biological replicates available for each depletion sample and library type.

CAGE data were described in Andersson et al. (2014b) and are available from GEO: GSE62047. Raw reads were downloaded from SRA, filtered using fastx-toolkit v0.13 with fastx_trimmer at settings '-Q 33 -f 12 -l 38' and fastx_quality_filter with settings '-Q 33 -q 30 -p 50' and then mapped to GRCh38 using bowtie2 v.2.2.8 with bowtie2 genome index available via NCBI at ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_genbank/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh38/seqs_for_alignment_pipelines/ GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index.tar.gz and otherwise default settings. Mapped reads with a mapping quality > = 10 were selected using samtools view v.1.6.0 ('-q 10') and coverage of read 5'ends for each strand was obtained using bedtools genomcov v.2.25.0 (Quinlan and Hall, 2010).

NETseq data were from total NET-seq sample of control cells (GEO: GSM2357382) from Schlackow et al. (2017). Track coverage for hg19 was downloaded from GEO and lifted to GRCh38 for downstream analysis.

**PROMPT annotations**

PROMPT candidate regions were selected from Gencode v.28 annotations and human genome release GRCh38. PROMPT candidate regions 1kb upstream and antisense of protein-coding-, lincRNA- and pseudo-gene TSSs were selected and intersected with combined CAGE signals across all libraries from GEO:GSE62047 mapped to GRCh38 as described above. The most prominent CAGE signal was selected for each candidate PROMPT region and only the highest CAGE peak was kept when neighboring signals where within 5kb on the same strand. Finally, CAGE signals overlapping Gencode v.28 transcript models, including 5kb upstream on
the same strand, were removed, except for transcripts belonging to annotation classes bidirectional promoter IncRNA, lincRNA, pseudogene and antisense genes, which were considered PROMPTs. This yielded 6140 PROMPT TSSs. Of these, 2134 overlapped with an annotated antisense RNA, 118 with a bidirectional promoter IncRNA, 833 with a lincRNA and 791 annotated as pseudogenes. PROMPT TESs were identified using 3'end-seq data presented herein. In brief, normalized signals (see below) from all single depletion libraries were combined and the positions with maximum 3'end signals within 5kb from the TSS were assigned as PROMPT TESs. Annotations are listed in Table S1.

**PROMPT sensitivities**

Normalized RNaseq signals were collected for PROMPT regions from the TSS to +1kb. The average of 3 replicates was computed for each region and sensitivity computed for each knock-down (KD) relative to the GFP control (CTRL), using the formula: sensitivity = (KD – CTRL)/max(KD,CTRL,1). Sensitivities were computed independently for total RNA and 1h BrU RNA samples and are available in Table S1.

**RNA 3' end seq quality control, filtering and mapping**

Quality control and trimming of raw reads was done as described in Schmid et al. (2018). For mapping, a custom index, concatenating human genome GRCh38 and sequences for ERCC spike ins, was prepared. This merged genome was indexed using STAR (v2.5.2b, Dobin et al., 2013) with settings ‘-runMode genomeGenerate–genomeSAindexNbases 11’ and otherwise default settings (providing no splice-junction information). Reads were then mapped to this index using the STAR aligner (v2.5.2b, Dobin et al., 2013) together with samtools (v1.6, Li et al., 2009) with settings ‘-outFilterType BySJout–outFilterMultimapNmax 20’. Bam files were then indexed using samtools and 5'end positions (marking the RNA 3'ends) of uniquely aligned reads were obtained using a custom python script using HTSeq (v0.6.0, Anders et al., 2015), that computes the coverage of uniquely mapping 5' ends in bedgraph format (see Schmid et al., 2018). Trimming and mapping statistics were obtained from the output of the bbduk script and the log files from the STAR mapper, respectively (Table S4). Mapped 3' ends aligning to genomic A-rich positions, likely deriving from priming of the reverse transcription primer to internal positions, were removed from analysis as described in Schmid et al. (2018).

**Normalization of 3' end seq datasets**

For normalization of RNAs between different factor depletions, library types and sample batches, we reasoned that mRNAs from highly expressed protein-coding genes produce primarily pA+ 3' ends, that are not globally sensitive to exosome depletion and therefore serve as suitable controls. 3'end-seq coverage values for last exons of protein-coding genes were therefore counted in all libraries and last exons with more than 100 reads in at least 1 library were selected for normalization (n = 9696). A pseudocount of 1 was added to all counts and the geometric mean of last exon counts computed within each library, and scaled relative to the mean of these values across all libraries (i.e., normfactor = GMlibrary/(1/m) ΣGMlibrary, where GMlibrary = e^(cntexon+1) / n and m is the number of libraries, n is the number of last exons considered and cntexon is the 3' seq read count within a last exon in a given library). A similar normalization strategy was applied using ERCC spike-ins, but yielded less consistent results, especially for the 4sU RNA fractions, and was therefore not used (data not shown). For all analysis presented, raw counts were divided by the last exon normalization factor as defined above. Genome browser tracks shown contain the mean normalized value from replicate experiments (see Table S4). Metagene analysis depict the mean of normalized signals across replicates.

**Aggregation and classification of 3' end clusters**

Closely positioned 3'end signals were aggregated separately for each strand into clusters by first summing up the coverage from all single factor depletion datasets from 4 library types into a master coverage dataset. Nucleotide positions covered with less than 5 reads in this merged coverage set were removed and remaining normalized signals within 25 bp were merged into clusters using the merge function from bedtools v2.25.0 (Quinlan and Hall, 2010). Finally, only clusters comprising a total signal of at least 20 reads were kept for further analysis and positions with maximum signal within each cluster (‘summit positions’) were obtained using the merged coverage. This yielded a total number of 208069 clusters covering roughly 4Mb (4048286 bp). Raw read coverage was then obtained from all individual non-normalized libraries. Counts from each of the 4 individual library types (ie pA+ total RNA; pA- total RNA; pA+ 4sU RNA; pA- 4sU RNA) were separately subjected to DESeq2 v1.20.0 (Love et al., 2014) differential expression analysis, applying the normalization factors from above as sizeFactors and including siRNA and batch information (Table S4) in the design matrix. Differential expression relative to siGFP samples was estimated and clusters were classified as 'NEXT 3'end clusters' when at least 1 of the 4 comparisons yielded a significant upregulation (DESeq2 FDR < 0.1) in siZFC3H1 samples but no upregulation in any of the 4 library types produced from siZFC3H1 cells. Conversely, 'PAXT 3'end clusters' were classified when at least 1 library type was significantly upregulated in siZFC3H1, but none of the siZFCCH8 samples and ‘NEXT+PAXT 3’end clusters’ when at least 1 significance call was obtained for both siZFC8H8 and siZFC3H1 samples. This yielded 6082 PAXT 3'end clusters, 21547 NEXT 3'end clusters 6986 NEXT+PAXT 3'end clusters and 173454 non-significant clusters. Clusters were overlapped to genomic annotations from R/Bioconductor package AnnotationHub v2.12.1 for hub genome = “GRCh38” and title = “Homo_sapiens.GRCh38.92.gtf” using R/Bioconductor package GenomicRanges v1.32.3 (Lawrence et al., 2013).

’eRNA cluster’ were constituted of the subset of clusters overlapping on the same strand the region from the TSS to +5kb of eRNAs as defined in Chen et al. (2016) and lifted to GRCh38. To define clusters overlapping intron1 of annotated genes, RNaseq data from
siEGFP control samples (Meola et al., 2016) were used for annotating HeLa cell transcripts. Only transcripts matching Gencode v28 annotations were considered and their expressions were estimated using stringtie v1.3.3b (Pertea et al., 2015) and R package ballgown v2.12.0 (Frazee et al., 2015). Only genes were a single transcript isoform contributed more than 50% of the total FPKM were considered and clusters overlapping first introns of those transcripts were selected.

PCA
PCA plots were performed using the prcomp function in R with default settings on the R/Bioconductor package DESeq2 objects, containing raw 3’end cluster counts with normalization factors as mentioned above as size factors. Before PCA, raw counts were first transformed using function ‘vst’ (variance stabilizing transformation) from DESeq2 and then used for PCA, and the resulting matrices filtered: we only analyzed the top 20000 (when analyzing all libraries together) or 500 (when analyzing individual library types) clusters ranked by variance across analyzed samples.

Metagene profiles
Metagene profiles were produced using computeMatrix tools from the deeptools software suite v3.0.2 (Ramírez et al., 2014) combined with custom python and R scripts for grouping and display (see Data and Code Availability). Metagene plots were computed as the mean normalized coverage across replicates in individual bins (typically 50bp wide, see code) displayed as log2 coverage after adding a pseudocount of 1 to all bins. For metagene plots of log2FC values, a pseudocount of 1 was added to all binned values before computing the log2 ratio between relevant depletions and siGFP. FDRs were computed by two-sided Mann-Whitney-U tests performed on each position bin and obtained p values from metagene profile multiple-test were corrected using the Benjamini-Hochberg method.

PAS motifs and strengths
Sequence motif frequencies were obtained using R/Bioconductor package ‘seqPattern’ (v1.12.0) based on sequences from ‘BSgenome.Hsapiens.UCSC.hg38’ (v1.4.1). PAS strengths were estimated using the python package ‘APARENT’ v0.1 (https://github.com/johli/aparent) with the neuronal network model ‘aparent_theano_legacy_30_31_34_padded.h5’ included in APARENT (Bogard et al., 2019). APARENT was designed to estimate alternative polyadenylation by comparing proximal to distal PASs. The model used here was trained to predict the isoform abundance and cut profile (not analyzed herein) of ‘proximal’ PASs given a fixed background distal PAS using as input a 205bp DNA sequence centered around the cleavage position. PAS strength was depicted as logit ‘isoform prediction’ for the 205bp surrounding given genomic positions. For scanning of 5kb PROMPT regions, 205bp windows with 25bp step size were used as input.

CLIPdb
Peak-called CLIP binding data for GRCh38 was obtained from CLIPdb v10-10-2018 (Yang et al., 2015 and http://lulab.life.tsinghua.edu.cn/postar/). Raw files were converted to bigwig format and signals ± 100bp from 3’end-seq cluster summits were obtained using R/Bioconductor package ‘rtracklayer’ v1.40.3. Total signals per cluster were obtained, a pseudocount = 1 added, and then used to compute the difference in mean log2(signals) between PAXT and NEXT sensitive clusters. For statistical testing, a 2-sided Mann-Whitney U test was used on the same log2(signals) and P-values multiple test corrected using Benjamini Hochberg (FDR) correction.

Statistical Tests
Statistical tests used are described in the respective figure legends.

DATA AND CODE AVAILABILITY
All high-throughput RNA-seq data generated during this study are available at Gene Expression Omnibus (GEO) under accession code GSE137612. Code for all bioinformatics analysis is available at GitHub (https://github.com/manschmi/Wu_etal).