Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality

Evgenia Ntini[1], Aino I. Järvelin[2,6], Jette Lange[3,6], Yun Chen[3,6], Mette Boyd[3], Mette Jørgensen[3], Robin Andersson[3], Ilka Hoof[3], Aleks Schein[1,4], Peter R. Andersen[1], Pia K. Andersen[1], Pascal Preker[1], Eivind Valen[3,5], Xiaobei Zhao[3], Vicent Pelechano[2], Lars M. Steinmetz[2], Albin Sandelin[3] and Torben Heick Jensen[1,7]

[1]Centre for mRNP Biogenesis and Metabolism, Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark; [2]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany; [3]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [4]Present address: Omics Science Center, RIKEN Yokohama Institute, Yokohama City, Japan; [5]Present address: Department of Molecular and Cellular Biology, Harvard University, Cambridge MA 02138, USA; [6]Equal contribution; [7]Corresponding author: thj@mb.au.dk.

Active human promoters produce PROMoter uPstream Transcripts (PROMPTs). Why these RNAs are coupled to decay, while their neighboring promoter-downstream mRNAs are not, is unknown. Here, high-throughput sequencing demonstrates that PROMPTs generally initiate in the antisense direction closely upstream of the transcription start site (TSS) of the associated gene. PROMPT TSSs share features with mRNA-producing TSSs, including stalled RNA polymerase II (RNAPII) and the production of small TSS-associated (TSSa) RNAs. Importantly, motif analyses around PROMPT 3'ends reveal polyadenylation (pA)-like signals. Mutagenesis studies demonstrate that PROMPT pA signals are functional, but link to RNA degradation. Moreover, pA signals are under-represented in promoter-downstream vs. upstream regions, allowing for more efficient RNAPII progress in the sense direction from a gene promoter. We conclude that asymmetric sequence distribution around human gene promoters serves to provide a directional RNA output from an otherwise bi-directional transcription process.

The non-protein-coding part of the human genome is pervasively transcribed into a large diversity of non-coding (nc) RNA[1]. A substantial fraction of this material derives from, or near, active gene promoters, that are producing a range of small- ([1-6]) and long non-coding RNA (lncRNA)[7]. Indeed, it has been estimated that >60% of lncRNAs in human embryonic stem cells derive from promoters of active protein-coding genes[8]. Although some lncRNAs have reported functions, these species are generally kept at low abundance by cellular degradation activities[9,10]. For example, we previously coupled depletion of the major nuclear 3'-5' exonucleolytic activity, the RNA exosome, with tiling microarrays to reveal PROMPTs closely upstream of active human gene promoters[9]. PROMPTs are 5'capped, >100nt long and 3'end adenylated in the absence of exosome activity[11]. The mechanism underlying the efficient exosome-mediated suppression of these lncRNAs, while preserving the promoter-downstream mRNA, remains enigmatic.

Here, we couple exosome-depletion to high-throughput 5'end-, 3'end- and regular RNA-sequencing (RNAseq) to create a genome-wide map of PROMPTs. Our results demonstrate that PROMPT transcription initiates antisense with respect to the downstream gene. We suggest that such initiating RNAPII, if stalled at a PROMPT-TSS proximal position, can elicit the production of previously reported TSSa-RNA. Sequence motifs around PROMPT 3'ends adhere to a pA site consensus and are significantly more abundant upstream than downstream of gene promoters. This provides a directional RNA output from human promoters by rapidly terminating antisense transcription and enforcing degradation of its RNA product.

RESULTS

PROMPTs initiate from bi-directional promoter activity

To obtain strand-specific and positional information of PROMPTs, we first subjected total RNA from HeLa cells, that had been treated with either a control (ctrl) eGFP siRNA or RRP40 siRNA (Supplementary Fig. 1a), to regular RNA sequencing (RNAseq) as well as cap-selected RNA 5'end sequencing (Cap Analysis of Gene Expression (CAGE)). We focused our analysis on protein-coding genes and therefore considered reads mapping to the -3kb to +1kb regions of 2428 UCSC gene promoters, which were selected not to overlap any other annotated mRNAs. When aligned to the TSSs of these promoters, both RNAseq- and CAGE- data disclosed a strong presence of exosome-sensitive transcripts originating closely upstream of the gene TSS

(average peak CAGE position at -110bp) and commencing in the antisense direction relative to the neighboring mRNA TSS (Fig. 1a bottom panel, compare 'ctrl' and 'RRP40' plots). Only minor signal was detected in the sense direction of the same region. Whereas the abundance of antisense PROMPT (asPROMPT) CAGE tags increased by an average of 8-fold upon RRP40 depletion, the corresponding sense CAGE signals of the same region were largely unaffected (Fig. 1b, P<2e-16, two-sided t-test). This predominant occurrence of asPROMPTs was also visible from select examples (Supplementary Fig. 1b-i), two of which were verified by northern blotting analysis (Supplementary Fig. 1b and 1c). At, and downstream of, TSSs, CAGE and RNAseq data exposed the expected expression of sense-RNA (Fig. 1a, top panel). Collectively, the data therefore revealed a symmetric RNA production profile with TSSs firing in both directions. We conclude that the clear majority of PROMPT transcription initiates from bi-directional promoter activity and runs antisense with respect to the neighboring gene.

Small uncapped TSSa RNAs have been reported to emanate narrowly from both up- and down-stream of human promoters[1,4-6]. Promoter-downstream TSSa RNA 3'ends overlap with the position of stalled RNAPII present at the majority of human genes[5,12]. This, and their sizes (~18-22nt), suggested they may be pieces of the nascent RNA chain protected by RNAPII from 5'-3' exonucleolysis[5]. To analyze whether promoter-upstream and downstream TSSa RNAs could be made by a similar mechanism, we aligned HeLa 18-30nt long RNAs[5] to either mRNA TSSs or asPROMPT CAGE 5'ends as the upstream antisense equivalents, together with RNAPII ChIP reads[13]. Antisense TSSa RNA 3'ends positioned almost precisely as their sense counterparts with respect to the TSS and RNAPII (Fig. 1c); the most common distance between TSSa RNA 3'ends and their corresponding TSSs were 34nt and 29nt, respectively (Fig. 1d) and both kinds of TSSa RNA 3'ends located in the center of their RNAPII ChIP peaks (Fig. 1e). Thus, bi-directional promoter transcription is remarkably symmetric, suggesting that the mechanisms initiating mRNA- and asPROMPT-transcription are very similar. Moreover, two main fates of promoter antisense transcription exist: i) RNAPII stalling leading to antisense TSSa production and ii) escape from the RNAPII stalling site leading to production of unstable lncRNA in the form of exosome-sensitive asPROMPTs. Consistently, PROMPTs are exosome-sensitive while antisense TSSa RNAs are not (Fig. 1c).

Sequence features around PROMPT 3'ends adhere to a pA site consensus

Why does sense transcription efficiently elongate downstream from promoters to produce mRNA while upstream antisense transcription creates shorter and unstable PROMPTs? To answer this question, we subjected polyA purified RNA to 3'Tag sequencing to obtain a genome-wide view of PROMPT 3'ends. Two main criteria were employed to acquire sufficient read counts: i) RNA was purified from RRP40-depleted cells, and ii) cDNA was produced using a dT primer annealing to the polyA tail addition site and size-selected for molecules of 100-800nt in length. Although this excludes PROMPTs of a larger size, it deselects for mRNA 3'ends that would otherwise dominate the library. To facilitate an overview of where RNA 3'ends are positioned relative to their respective TSSs, we aligned these to mRNA- and asPROMPT-TSSs, respectively and overlaid the RRP40 RNAseq data. This revealed an expected distribution of asPROMPT 3'ends upstream of their corresponding CAGE defined 5'ends with an estimated average size of asPROMPTs, in the interrogated size-range, of 296nt (Fig. 2a, bottom panel and Supplementary Fig. 2a). We note that some of the RNAseq signal extends beyond this position due to the over-representation of longer molecules by the RNAseq protocol (see Methods) and the 100-800nt size selection of RNAs subjected to 3'end analysis. Still, 28% of interrogated promoters have at least two-fold more RNAseq tags in the first 500nt downstream of the asPROMPT TSS (see Supplementary Fig. 1b-i for individual examples). 3'ends could also be detected downstream of mRNA TSSs, although less commonly (see below).

To inquire for sequence features that might direct the formation of PROMPT 3'ends, motif analysis around these was conducted, which revealed a remarkable similarity to sequence information encompassing mRNA 3'ends (Fig. 2b, top panel). PROMPTs generally harbor the AWTAAA hexamer of the consensus mRNA 3'end processing site at the conventional position of 10-30nt upstream of the RNA 3'end[14]. Analyzing for the occurrence of AWTAAA, or its single substitution variants ('weak hexamers'[15,16]), demonstrated that while mRNAs harbor these motifs within the 10-30nt upstream region at 58% and 53% of their 3'ends, the respective fractions were 37% and 37% for 3'Tag sequenced PROMPTs (Fig. 2b, bottom panel). Moreover, downstream associated '2GT/T' and 'T-rich' motifs, known to compose binding sites for the mRNA 3'end processing factor CstF64[17,18] were also detectable at similar

positions downstream of PROMPT 3'ends (Fig. 2b). Consistently, these sites were able to recruit CstF64 as revealed by its overlapping cross-linking and immunoprecipitation (CLIP) interaction profile[19] (Supplementary Fig. 2b). Given the observed similarity between sequences around PROMPT and mRNA 3'ends, we next tested experimentally the effect of positioning two bona fide pA sites within a 300-400nt distance from the CMV promoter. These small 'loci' were introduced into HeLa cells either by plasmid transfection or by stable integration into the recipient site of HeLa Flp-In cells. Consistent with the result from 3'Tag sequencing, both bovine growth hormone (BGH)- and simian virus 40 (SV40) late-pA (L-pA) sites yielded RNA products, which were highly exosome-sensitive (Fig. 2c).

PROMPT pA sites are functional and trigger transcription termination

To directly demonstrate that PROMPT pA sites direct 3'end formation, we cloned two PROMPT loci, proIFNAR1 and proTMEM97, in between the CMV promoter and the SV40 late pA signal (Fig. 3a and 3b, see schematics on top), and mutated their predicted pA sites. When stably expressed in HeLa cells both constructs yielded exosome-sensitive PROMPT RNA, confirming the utility of this approach (Fig. 3a, bottom left image, and Fig. 3b). Mutation of the proIFNAR1 '2GT/T' element increased usage of the downstream, and more distal, SV40L-pA over the proIFNAR1-pA site by ~3 fold relative to the wt construct, while mutating the proIFNAR1 AATAAA hexamer completely abrogated 3'end formation at this site (Fig. 3a, compare lanes 8 and 10 to lane 6). Similarly, mutation of the proTMEM97 hexameric sequence resulted in the appearance of read-through RNAs (Fig. 3b). This was also the case when mutating the hexamer of the TSS-proximally positioned SV40 late pA site of the construct employed in Fig. 2c (Fig. 3c). Moreover, depleting key mRNA 3'end processing factors CPSF73 and PCF11 provoked read-through of the pA site of this construct (Supplementary Fig. 3a-c), demonstrating that at least some aspects of regular mRNA 3'end formation are taking place at TSS-proximal pA sites. Taken together, these data provide evidence that PROMPT pA sites are being used as predicted by the 3'Tag sequencing. However, in contrast to their promoter-distal counterparts, PROMPT pA sites yield unstable product. Indeed, extending the distance between the CMV promoter and the SV40 late pA site by 1kb, from ~ 0.4kb to ~1.4kb, resulted in a 87% reduction in exosome-sensitivity (Supplementary Fig. 4).

As pA sites are known triggers of transcription termination[14,20,21], we next analyzed transcription within the PROMPT region. To this end, we aligned sequence reads from a global run on (GRO)-seq experiment[22] to the peak position of tags at either asPROMPT- or their associated mRNA-TSSs. This analysis showed that while GROseq signal declines in a similar fashion from the respective TSSs for the first 100bp downstream of the GROseq peak, further progression is more readily achieved for mRNA-producing RNAPII (Fig. 4). The point of deviation of the two profiles is where asPROMPT pA sites start to accumulate (Fig. 2a). These data are therefore consistent with the notion that the high density of pA sites in the PROMPT region serves to elicit transcription termination.

Asymmetric sequence distribution around gene promoters

To inquire whether the different GROseq profiles could be explained by asymmetric sequence patterns around human promoters, we plotted the distance from either mRNA- or asPROMPT-TSSs that RNAPII has to travel before transcribing a AWTAAA hexamer. This demonstrated that a PROMPT-transcribing RNAPII has a substantially higher chance of encountering a pA hexamer in the first 500nt downstream of its TSS compared to an mRNA-transcribing RNAPII (Fig. 5a, left panel). Interestingly, this is mirrored by a higher likelihood of the mRNA-transcribing RNAPII to encounter a 5'splice site (5'SS) (Fig. 5a, right panel). A large majority (82%) of these predicted 5'SS are overlapping the first annotated exon-intron junction of the respective gene, and as 5'SS-bound U1 snRNP has been reported to suppress pA sites[23], it thus appears that strong measures are in place to avoid premature pA site utilization downstream of gene TSSs. To address whether this asymmetric sequence distribution around human gene promoters had any functional impact, we counted the fraction of promoters displaying an experimentally defined RNA 3'end as a function of its distance downstream of PROMPT- and mRNA-TSSs, respectively. While 22% of regions yielded a PROMPT 3'end within the first transcribed 500nt the result was 5,6% for mRNA 3'ends (Fig. 5b). Thus, the predicted higher density of promoter-upstream pA sites is also reflected in its utilization. We note that of the 3'ends detected downstream of mRNA TSSs, a good fraction presumably arise from sequences with a discernible pA site consensus (Supplementary Fig. 5). Thus, although clearly decreased in promoter downstream regions, pA site usage does occur. In fact, the incidence of AWTAAA hexamers up to 500nt downstream of the

mRNA TSS is higher for the genes where 3'ends can be detected in both of two independent 3'Tag libraries vs. genes without detectable 3'ends in this region, (Fig. 5c, left panel). Interestingly, regions displaying more pA site utilization also exhibited a decreased presence of 5'SS information, suggesting a less efficient suppression by U1 snRNP (Fig. 5c, right panel).

Discussion

Most mammalian promoters are inherently bi-directional[4,8,22,24] but transcription only elongates productively in one direction. It has been suggested that an unfavorable chromatin environment upstream of gene TSSs and/or insufficient post-translational modification of anti-sense transcribing RNAPII may account for such failure to conclude transcription[4]. Data presented here demonstrate that, at least part of, the answer instead lies in the asymmetric distribution of pA site sequences around human gene promoters, causing termination of upstream antisense transcription. As a further means of restricting bi-directional promoter activity to a uni-directional RNA output, the exosome effectively removes transcripts whose 3'ends are formed by TSS-proximal pA signals. Why these sites yield unstable product remains to be fully elucidated. Activity of the pre-mRNA cleavage factor CPSF73 (Fig. S3a-c) fits the notion that the distance between the pA site hexamer and the 3'end of PROMPTs confers to the consensus length observed for mRNAs (Fig. 2b) and implies that PROMPT 3'ends are defined by endocleavage. We speculate that aspects of the ensuing polyadenylation process occur sub-optimally, allowing the exosome to efficiently intervene. Since exosome-sensitivity of transcripts decreases with increased distance between the TSS and the pA site (Supplementary Fig. 4), it may be that RNAPII C-terminal domain (CTD) status is not optimal for supporting polyadenylation at TSS-proximal pA sites. Further analysis is required to evaluate whether a slightly relaxed consensus of PROMPT vs. mRNA pA sites can in some cases also contribute to exosome-sensitivity.

We note that while the non-canonical pA polymerase hTRF4-2 can adenylate PROMPT 3'ends in the absence of exosome activity, it does not generally contribute to PROMPT turnover[11]. Thus, the mechanism of PROMPTs turnover does not fully mirror that of *S. cerevisiae* cryptic unstable transcripts (CUTs)[25]. However, with regard to other mechanistic aspects, such sequence-specific removal of short RNA is indeed reminiscent of the *S. cerevisiae* Nrd1p-Nab3p-Sen1p pathway, which couples

transcription termination to exosomal decay of CUTs[26,27], the majority of which also initiates antisense from protein-coding gene loci[28,29]. Whereas *S. cerevisiae* utilizes short Nrd1p and/or Nab3p binding motifs in the nascent RNA to accomplish this task, higher eukaryotes employ pA site-induced removal of PROMPTs. While gene promoters take advantage of this mechanism to suppress promoter upstream transcripts, other mammalian RNAPII-transcribed loci producing small, but stable, RNAs employ terminator/3'end formation systems that are pA site-independent.

Accession codes. RNA sequencing data have been submitted to GEO at ncbi (http://www.ncbi.nlm.nih.gov/) under accession number XXX.

Acknowledgements. We thank M. Schmid and S. Lykke-Andersen for comments on the manuscript as well as P. Carninci and H. Takahashi for help with the CAGE protocol and S. Wilkening with the 3'Tag protocol. This work was supported by the Danish National Research Foundation (grant DNRF58), the Villum Kann Foundation and the Danish Cancer Society (to T.H.J.); the EU 7th Framework Programme (FP7/2007–2013)/ERC grant agreement 204135, the Novo Nordisk and Lundbeck Foundations (to A.Sa.) and University of Luxembourg—Institute for Systems Biology Program and the Deutsche Forschungsgemeinschaft (to L.M.S.). A.Sc. was the recipient of a European Molecular Biology Organization postdoctoral grant. The authors declare that they have no competing financial interests.

Author Contributions. E.N., J.L., A.Sa. and T.H.J. designed the experiments. E.N., J.L., M.B., A.Sc., P.R.A., P.K.A. and P.A.P. performed the experiments. A.I.J., Y.C., M.J., R.A., I.H., E.V., X.Z., J.B-L., V.P. and A.Sa. did the bioinformatics analyses. E.N., A.Sa. and T.H.J. wrote the manuscript. All additional authors have critically read and approved of the manuscript.

References:

1 Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028-1032, (2009).

2 Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature reviews. Genetics* **10**, 833-844, (2009).

3 Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature reviews. Genetics* **13**, 233-245, (2012).
4 Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-1851, (2008).
5 Valen, E. *et al.* Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nature structural & molecular biology* **18**, 1075-1082, (2011).
6 Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature genetics* **41**, 572-578, (2009).
7 Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488, (2007).
8 Sigova, A. A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2876-2881, (2013).
9 Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851-1854, (2008).
10 Flynn, R. A., Almada, A. E., Zamudio, J. R. & Sharp, P. A. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10460-10465, (2011).
11 Preker, P. *et al.* PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic acids research* **39**, 7179-7193, (2011).
12 Nechaev, S. & Adelman, K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochimica et biophysica acta* **1809**, 34-45, (2011).
13 Anamika, K., Gyenis, A., Poidevin, L., Poch, O. & Tora, L. RNA polymerase II pausing downstream of core histone genes is different from genes producing polyadenylated transcripts. *PloS one* **7**, e38769, (2012).
14 Kuehner, J. N., Pearson, E. L. & Moore, C. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature reviews. Molecular cell biology* **12**, 283-294, (2011).
15 Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome research* **22**, 1173-1183, (2012).
16 Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome research* **10**, 1001-1010, (2000).
17 Salisbury, J., Hutchison, K. W. & Graber, J. H. A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. *BMC genomics* **7**, 55, (2006).
18 Martin, G., Gruber, A. R., Keller, W. & Zavolan, M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell reports* **1**, 753-763, (2012).
19 Yao, C. *et al.* Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 18773-18778, (2012).
20 Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes & development* **25**, 1770-1782, (2011).

21 Richard, P. & Manley, J. L. Transcription termination by nuclear RNA polymerases. *Genes & development* **23**, 1247-1269, (2009).
22 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848, (2008).
23 Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664-668, (2010).
24 Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell reports* **2**, 1025-1035, (2012).
25 Wyers, F. *et al.* Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**, 725-737, (2005).
26 Gudipati, R. K., Villa, T., Boulay, J. & Libri, D. Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. *Nature structural & molecular biology* **15**, 786-794, (2008).
27 Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Molecular cell* **36**, 541-546, doi:10.1016/j.molcel.2009.10.019 (2009).
28 Neil, H. *et al.* Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**, 1038-1042, (2009).
29 Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033-1037, (2009).
30 Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L. & Hovorun, D. M. Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic acids research* **31**, 1375-1386 (2003).

Figure legends:

Figure 1 Exosome-depletion reveals similar transcription profiles up- and down-stream of promoters. a Full RNAseq reads and CAGE 5'nucleotides derived from RRP40-depleted and control cells were aligned on 2428 protein-coding gene promoters. Y-axis shows tags/million (TPM) reads (left axis for CAGE and right axis for RNAseq); for RNA-seq, intronic regions will have signal if reads are spanning supported exon-exons junctions. X-axis shows the relative distance to the mRNA TSS. Upper and lower panels display sense and antisense tags, respectively. **b** Distribution of CAGE RRP40/Ctrl library fold changes in the upstream -1 to -3000bp region of Fig. 1a split by strand relative to mRNA TSS. **c** Distribution of TSSa RNA 3'end nucleotide[5] and RNAPII ChIP[13] signal around mRNA TSSs (left panel) and 'CAGE-RRP40'-defined asPROMPT TSSs (right panel). Y-axis shows signal normalized to the maximal observed in respective data set and window; TSSa RNA 3'end tags are smoothed by a 10nt window. X-axis shows the nucleotide positions relative to the respective TSS. ChIP signals are not strand specific while TSSa RNA 3'end tags are. **d** Cross-correlation between CAGE- and TSSa RNA 3'end-tag counts

(both obtained from RRP40 depletion conditions) from the same strand and focused on mRNA (blue) or PROMPT (red) TSSs. Y-axis shows the average correlation between the two data sets when one is slid over the other (X-axis). Lags with the highest mean correlation scores are indicated. **e** Cross-correlation between RNAPII ChIP signal and TSSa RNA 3'end-tags, focused on mRNA (blue) and PROMPT (red) TSSs as in **d**.

Figure 2 Exosome-sensitivity is triggered by TSS-proximal pA sites.

a Distribution of sequence reads from RNAseq-, CAGE- and 3'Tag libraries prepared from RRP40-depleted cells. Full RNAseq-, CAGE 5'nucleotide- and 3'Tag 3'nucleotide-tags are displayed as in Fig. 1c. Upper and lower panels display sense and antisense tags as indicated. **b** Motif analysis of asPROMPT 3'ends. Top panels: Positional distribution of pA site-related motifs (as indicated by schematics on top) within ±100nt windows around asPROMPT- and mRNA-3'ends. Bottom panels: Plots comparing cumulative fractions of asPROMPT- and mRNA-3'ends having a particular predicted site within a given window range. Light shadowing indicates the 10-30nt region upstream of the RNA 3'end where pA site hexamers usually occurs. Motifs were from ([16-18,30]). **c** Exosome-sensitivity of RNA produced using TSS-proximal BGH- or SV40L-pA sites. Constructs harboring either of these pA sites within 400nt downstream of a CMV promoter (shown schematically on top) were transiently transfected or stably integrated into HeLa cells. After cellular administration with ctrl (-) or RRP40 (+) siRNA as indicated total RNA was subjected to northern blotting analysis using probes against 7SL- or 18S-RNA as loading controls.

Figure 3 Mutagenesis of PROMPT pA signals underscore their functionality.

a Top: The proIFNAR1 PROMPT locus ('wt'), and its mutated variants ('hex_mut' and 'GT_mut' (mutated nucleotides marked in grey)), were cloned in between a CMV promoter and a SV40L-pA site (schematics on top) and stably integrated into HeLa cells. Bottom: Northern blotting analysis of RNA harvested from cells expressing the indicated constructs and treated with control (egfp) or RRP40-directed siRNAs (left image). Migrations of products terminated at the PROMPT- and SV40L-pA sites, respectively, are indicated. Total RNA was treated with RNaseH and a dT20-oligo

(dT) or no oligo (-). 7SL RNA was probed as a loading control. **b** The proTMEM97 PROMPT locus ('wt'), or its mutant variant ('hex_mut') (schematics on top), were stably expressed from HeLa cells, which were pre-treated with egfp- or RRP40-siRNAs as indicated. Total RNA was subjected to quantitative RT-PCR analysis, employing a random hexameric DNA primer and one of two PCR amplicons '5'AMPL' or '3'AMPL'. RT-qPCR results are plotted relative to levels from 'egfp' samples. Error bars indicate standard deviations from three technical replicates. **c** Northern blotting analysis of total RNA harvested from RRP40-depleted cells stably expressing the CMV-SV40L-pA construct ('wt', Fig. 2C) or its AATAAA→AAGAAA variant ('hex_mut') using same probe as in **a**. Total RNA was treated with RNaseH and no (-), dT-, ol1-, or rt-DNA oligonucleotides as indicated (relative positions of ol1 and rt oligonucleotides indicated on top schematics). Note that the hexameric mutation causes collapsing of 3'extended RNAs when employing the 'rt' oligonucleotide. Migration of a species arising from dT cleavage at an internal A-stretch is indicated.

Figure 4 Transcription declines more rapidly downstream of asPROMPT TSSs than mRNA TSSs. a Distribution of GROseq[22] RNA 5'end reads downstream of mRNA TSSs (black) and asPROMPT TSSs (grey) as defined in Fig. 1c. The Y-axis shows reads smoothed by a 10nt window, where the distributions are aligned on the maximal peak downstream of the TSS. X-axis shows nt positions downstream of respective peaks (which tend to lie 50nt downstream of TSSs)[22]. X-axis shows nucleotide positions downstream of the respective TSSs. Dashed line indicates where the two profiles start to deviate (~100nt).

Figure 5 Asymmetric sequence distribution around promoters ensures transcription directionality. a Analysis for AWTAAA- and 5'SS-motifs (see Methods) in the 500nt regions downstream of mRNA (black)- or asPROMPT (grey)-TSSs. The Y-axis shows the cumulative fraction of regions having at least one predicted site after traversal of a given number of nt as indicated on the X-axis. The 2428 gene set was subjected to analysis. **b** Analysis of experimentally defined 3'ends downstream of mRNA (black)- or asPROMPT (grey)-TSSs. The Y-axis shows the cumulative fraction of regions having at least one 3'end after traversal of a given

number of nt as indicated on the X-axis. Note that analyzed regions were extended to 1000nt downstream of the respective TSSs. **c** AWTAAA and 5'SS motif analysis interrogating the 1000nt region downstream of mRNA TSSs and scrutinizing experimentally verified transcripts harboring RNA 3'ends in both (black) or neither (grey) of two 3'Tag libraries.

Online methods:

Cell culture, protein depletion and plasmid transfection: HeLa cells were grown in DMEM medium supplemented with 10% fetal bovine serum at 37°C and 5% CO_2. siRNA transfections were done twice with 22 nM of siRNA with the second hit performed after 48 h using Lipofectamin2000 as transfecting agent (Invitrogen). For plasmid transfections 0.5 μg of plasmid per ml of culture were co-transfected with siRNA during the second hit. Cells were harvested 48 h after the second hit, and protein depletion was verified by western blotting analysis as described[9]. RNA was extracted using Trizol (Invitrogen). Stable cell lines using 'pCMV-BGH' (pCDNA5_FRT/TO (Invitrogen)), 'pCMV_SV40L-pA' ('YLR0LpA+'[31]), 'pCMV_SV40L-pA_mut' ('YLR0LpA-'[31]), wt and point-mutant versions of 'pCMV_proIFNAR1_SV40L-pA' and 'pCMV_proTMEM97_SV40L-pA' were established in HeLa Flp-In, by co-transfecting 1 μg of plasmid with 10 μg pOG44 (Invitrogen) in 6 ml cultures, followed by hygromycin selection (2 μl/ml medium).

Plasmid construction: The proIFNAR1 sequence was amplified with primers AflII_proIFNAR1_F and XhoI_proIFNAR1_R (sequences in Supplementary Table 4) and cloned into AflII/XhoI of pCMV_SV40L-pA. proIFNAR1_hex_mut- and GTmut-variants were generated by overlapping PCR using respective point mutagenesis primers (Supplementary Table 4). pCMV_proTMEM97_SV40L-pA -wt and -hex_mut were generated as described above using the respective primers (Supplementary Table 4). Vectors pCMV_SV40L-pA wt and -mut were used in a previous study under the names 'YLR0LpA+' and 'YLR0LpA-' respectively[31]. pCMV_YLR 1.4kb_SV40L-pA was constructed by NotI/XmaI ligation of 1kb YLR454 sequence amplified using primers 'YLR454_FW_NotI' and 'YLR454_1/8_RE_XmaI'. Plasmid '272_LpA' used in transient transfection was generated from 'pCMV_SV40L-pA' by site directed mutagenesis using primers '272_LpA_F' and '272_LpA_R' (Supplementary Table 4).

Datasets: All sequencing data were mapped to hg19 genome (see below for data-set-specific details). For published datasets, we used already existing mappings but converting them to hg19 using the LiftOver tool if necessary. Additional post-processing is noted below. For tags per million (TPM) calculations, we divide by the number for mapped reads. For all plotted data, to avoid undue influence of outliers, we assign a value corresponding to the 99th TPM percentile of respective dataset to values that were over this threshold.

CAGE library preparations and data processing: CAGE libraries were prepared from 5 μg of total RNA purified from $2x10^6$ cells using the Purelink mini kit (Ambion cat.no. 12183018A) with 1% 2-Mercaptoethanol and on-column DNAse I treatment (Ambion cat no 12185010) as recommended by manufacturer. Libraries were prepared according to published protocol[32]. Prior to sequencing, four CAGE libraries with different barcodes were pooled and applied to the same sequencing lane. The libraries were sequenced using a HiSeq2000 instrument from Illumina. The libraries were sequenced using the HiSeq2000 platform from Illumina at the National High-throughput DNA Sequencing Centre, University of Copenhagen. To compensate for the low complexity in 5'end of the CAGE liberaries 30% Phi-X spike-in were added to each sequencing lane as recommended by Illumina. CAGE reads were assigned to their respective originating sample according to identically matching barcodes. Assigned reads were trimmed to remove linker sequences and subsequently filtered for a minimum sequencing quality of 30 in 50 % of the bases using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Mapping to the human genome (hg19) was performed using Bowtie (version 0.12.7), using standard settings but allowing for multiple alignments and subsequently filtering for uniquely mapping reads. Reads that mapped to unconventional chromosomes or chrM were discarded.

RNAseq library preparations and data processing: RNAseq libraries were prepared from total RNA depleted of ribosomal RNA using the RiboMinus kit (Invitrogen) using the ScriptSeq™ v2 RNA-Seq Library Preparation Kit (Epicentre) with multiplexing, according to manufacturer's directions. 100-nt-paired-end sequencing was performed on the Illumina HiSeq platform. Reads (8669268 and

11560746 from the ctrl and RRP40 libraries, respectively) were aligned to the human genome (hg19/GRCh37) using the STAR algorithm[33].

3'Tag-Seq library preparations and data processing: The 3'Tag-Seq protocol[34] was followed with minor modifications: 5 μg of DNaseI treated, polyA$^+$ enriched (MicroPolyA Purist kit, Ambion) RNA were ribosomal RNA-depleted using RiboZero kit (Epicentre) and 1 μg of the resulting RNA was sequentially treated with T4 PNK, 5'polyphospatase and Terminator enzyme (Epicentre) to enrich for capped molecules. Fragmentation of total RNA was ommited. Reverse transcription and second strand cDNA synthesis were performed according to the protocol and ~100-800 bp size selected double stranded cDNA was sonicated using Covaris (15 min, Duty cycle 10%, Intensity 10, 200 CpB). End repair, (A)-tailing and ligation of Illumina sequence adaptors were performed according to the protocol and final stringent size selection of 200bp fragments was done by E-Gel 2% SizeSelect electrophoresis (Invitrogen). Two Rrp40 libraries (biological replicates) were constructed and sequenced. Reads were processed as described[34,35]. Briefly, reads running into an A stretch of at least 10bp were trimmed and aligned to the human reference genome (hg19, ENSEMBL 37.66) using the GSNAP aligner[36]. Reads with unique and single good quality alignments were kept. Filters were applied to remove potential instances of internal priming: poor quality alignments and reads shorter than 20bp, read A content more than 65%, 8 out of 10 preceding bases in the genome As, or an otherwise high A content (65%) of +/-50bp region around p(A) site. Reads were collapsed to their 3'most base to capture the position of pA sites.

Definition of asPROMPT TSSs by CAGE: For all computational analyses unless otherwise mentioned, we used the -3kb to 1kb region around hg19 UCSC known genes which had i) both RRP40 CAGE and RRP40 3' end data in the upstream region, and ii) there were no other annotated mRNAs overlapping the region (including TSSs and TTSs), resulting in 2428 regions. The highest antisense RRP40 CAGE peak in the upstream region was considered to be the main PROMPT TSS for subsequent analyses.

Motif search: Scripts for motif searches were written in Python programming language for the following motifs: AWTAAA; A[AT]TAAA, T-rich; ([ATCG]TTTT)

or (T[ATCG]TTT) or (TT[ATCG]TT) or (TTT[ATCG]T) or (TTTT[ATCG]), 2GT/T; (GTGTT) or (TGTGT) or (GTTGT) or TGT[GC]T, weak hexamer (single substitutions of AWTAAA); ATAAA or ([ACT]AATAA) or (AAAATA) or (A[CG]TAAA) or (AATA[ACGT]A) or (AAGAAA) or (AATGAA). Motif search was done in ±100bp windows around asPROMPT 3'ends which were extracted from 5kb regions upstream and antisense to the first mRNA TSS, excluding regions with overlapping annotated features. mRNA ends were extracted from 3'UTR regions extended by 200bp downstream. Promoter proximal regions were defined as 1kb regions downstream of TSS of coding genes longer than 4 kb, excluding any containing overlapping features. Motif search profiling of Fig. S5 was performed using asPROMT pA sites within 500bp from the highest asPROMPT CAGE tag and promoter proximal pA sites within 500bp downstream of TSS. For Figure 4, we used the following two count matrices, using 0.9 relative score as a cutoff and the ASAP tool[37] (with standard settings). The 5'SS matrix is collected from the JASPAR database[38] (ID SD0001.1) while the AWTAAA matrix correspond to the consensus defined above.

Cross Correlation: Cross correlation plots are made by sliding one dataset across another in 1 nt increments and calculating the mean Pearson correlation over all the windows analyzed, as a function of the shift between the datasets.

Processing of external datasets: RNAPII ChIP data[13] were aligned to the reference genome. Uniquely aligned reads were shifted 40bp towards the centre of the read fragment to estimate RNAPII binding position. GRO-seq data[22] were aligned to the reference genome and the 5'base of uniquely aligned reads was used to estimate position of actively engaged RNAPII. The GRO-seq and sRNA data were smoothed by a 10 nt window.

Northern blotting and RNaseH assays: 5-50μg of DNase I treated RNA was subjected to 4-6% PAGE followed by wet transfer over-night at 15V after which membranes were UV-cross-linked and pre-hybridized for 1 hour in ULTRAhyb buffer (Ambion). For RNaseH assays, 5 μM of DNA oligonucleotide was annealed to 25μg of total RNA by 2 min incubation at 80°C followed by slow cooling to room

temperature, in annealing buffer (0.5M KCL, 0.5M Tris pH 8.3). 0.06 µl of RNaseH stock was added and the reaction (125 mM KCL, 250 mM Tris pH 8.3, 15 mM $MgCl_2$) was carried out for 30min at 37°C. Products were precipitated (0.3 M NaAC pH 5.3, 2.5 vol. EtOH) and respunded in 10µl of formamide loading buffer, followed by 2 min denaturation at 94°C prior immediate gel loading. α^{32P}-UTP labeled RNA northern probes (for BGH construct and detection of endogenous PROMPTs) were generated by T7 *in vitro* transcription reactions from gel purified PCR-generated DNA template. 18S, 7SL and LpA probes were γ^{32P}-ATP end-labeled using T4 PNK (Fermentas) and hybridized to the blots in ULTRAhyb-Oligo buffer (Ambion) at 42°C. Original images of blots from Figures 2 and 3 can be found in Supplementary Figures 6 and 7.

References for Online Methods:

31 Andersen, P. K., Lykke-Andersen, S. & Jensen, T. H. Promoter-proximal polyadenylation sites reduce transcription activity. *Genes & development* **26**, 2169-2179, (2012).

32 Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods Mol Biol* **786**, 181-200, (2012).

33 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, (2013).

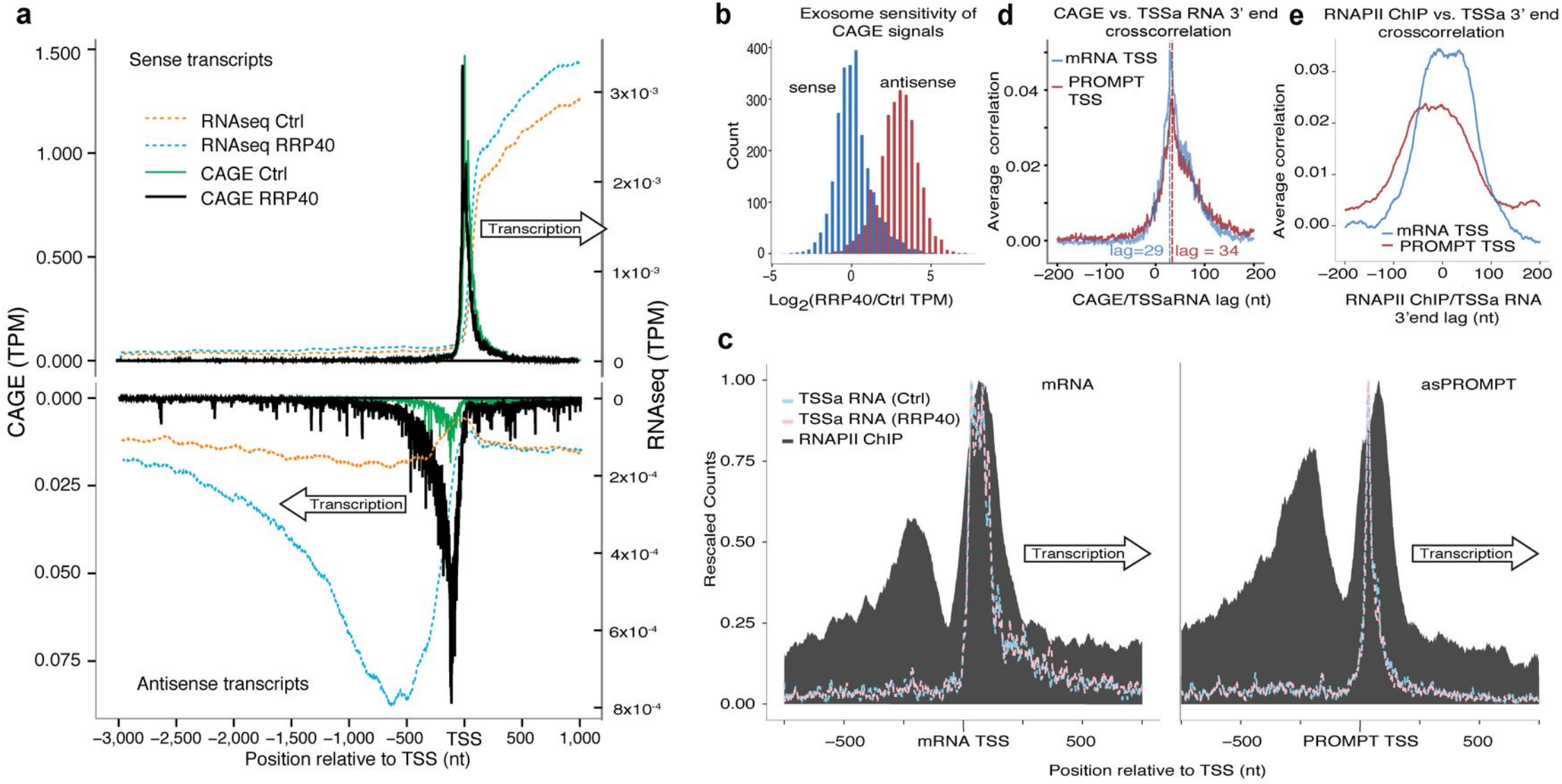
34 Pelechano, V., Wilkening, S., Jarvelin, A. I., Tekkedil, M. M. & Steinmetz, L. M. Genome-wide polyadenylation site mapping. *Methods in enzymology* **513**, 271-296, (2012).

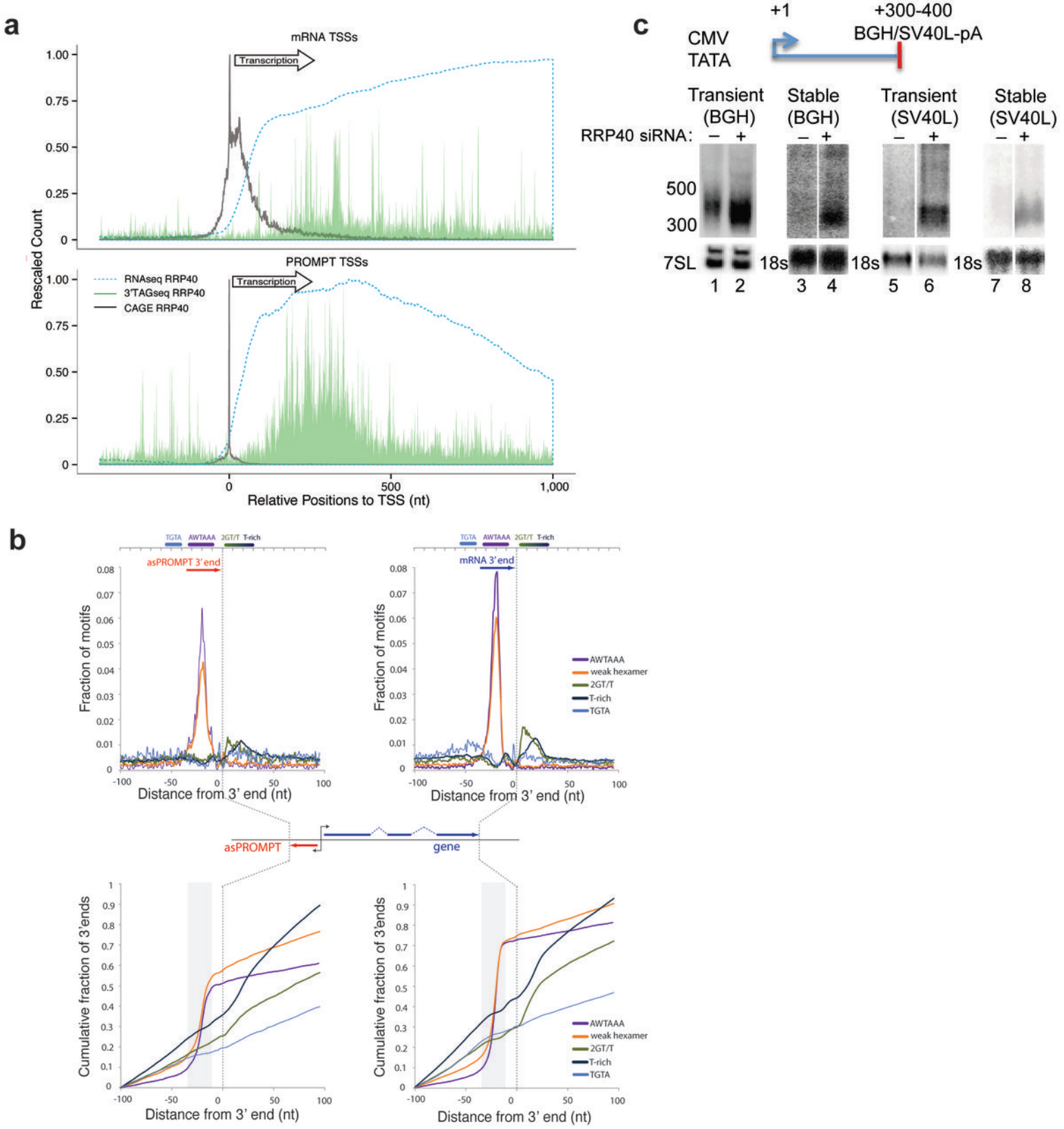
35 Wilkening, S. *et al.* An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic acids research*, (2013).

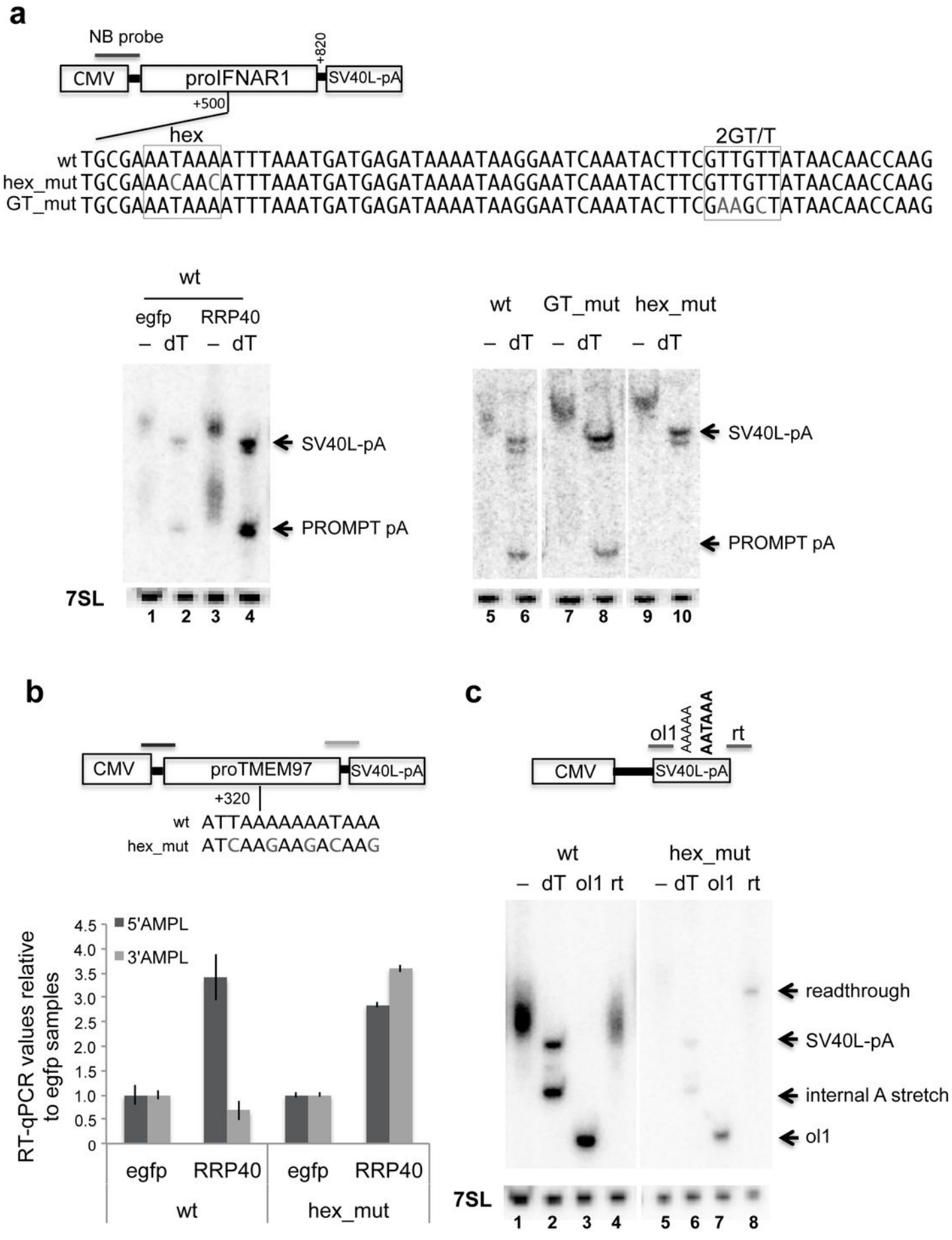
36 Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881, (2010).

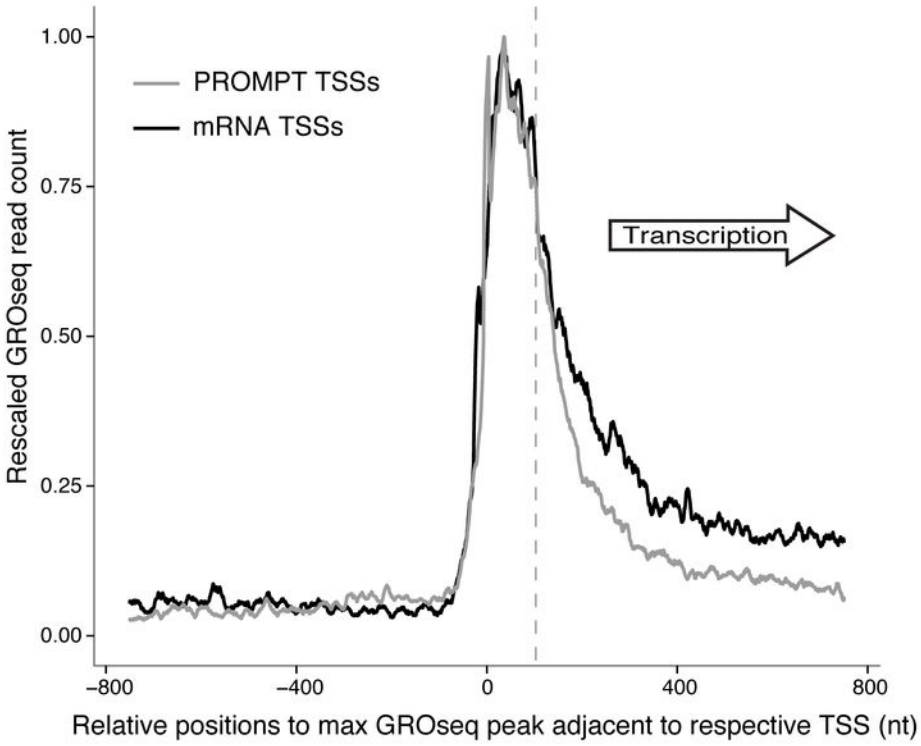
37 Marstrand, T. T. *et al.* Asap: a framework for over-representation statistics for transcription factor binding sites. *PloS one* **3**, e1623, (2008).

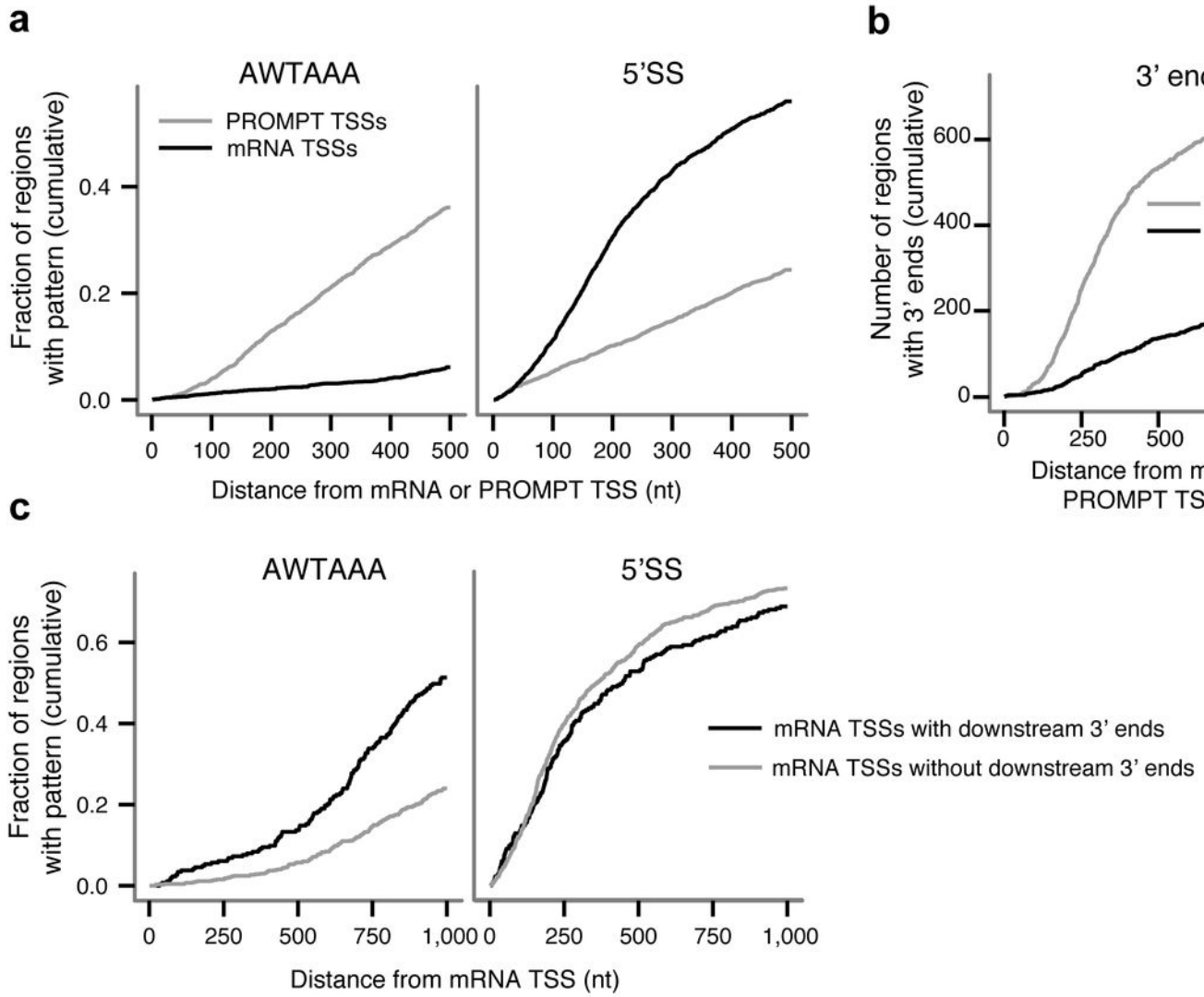
38 Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research* **38**, D105-110, (2010).


a
Sense transcripts
RNAseq Ctrl
RNAseq RRP40
CAGE Ctrl
CAGE RRP40
Transcription
CAGE (TPM)
RNAseq (TPM)
Antisense transcripts
Transcription
Position relative to TSS (nt)
b
Exosome sensitivity of CAGE signals
sense
antisense
Count
Log2(RRP40/Ctrl TPM)
d
CAGE vs. TSSa RNA 3' end crosscorrelation
mRNA TSS
PROMPT TSS
lag=29
lag = 34
Average correlation
CAGE/TSSaRNA lag (nt)
e
RNAPII ChIP vs. TSSa 3' end crosscorrelation
mRNA TSS
PROMPT TSS
Average correlation
RNAPII ChIP/TSSa RNA 3'end lag (nt)
c
TSSa RNA (Ctrl)
TSSa RNA (RRP40)
RNAPII ChIP
mRNA
asPROMPT
Transcription
Transcription
Rescaled Counts
mRNA TSS
PROMPT TSS
Position relative to TSS (nt)


a
mRNA TSSs
Transcription
PROMPT TSSs
RNAseq RRP40
3'TAGseq RRP40
CAGE RRP40
Rescaled Count
Relative Positions to TSS (nt)
b
TGTA AWTAAA 2GT/T T-rich
asPROMPT 3'end
mRNA 3'end
Fraction of motifs
Distance from 3' end (nt)
AWTAAA
weak hexamer
2GT/T
T-rich
TGTA
asPROMPT
gene
Cumulative fraction of 3'ends
c
CMV TATA
+1
+300-400 BGH/SV40L-pA
Transient (BGH)
Stable (BGH)
Transient (SV40L)
Stable (SV40L)
RRP40 siRNA: – + – + – + – +
500
300
7SL
18s
1 2 3 4 5 6 7 8


a
NB probe
CMV
proIFNAR1
SV40L-pA
+820
+500
hex
2GT/T
wt TGCGAAATAAAATTTAAATGATGAGATAAAATAAGGAATCAAATACTTCGTTGTTATAACAACCAAG
hex_mut TGCGAAACAACATTTAAATGATGAGATAAAATAAGGAATCAAATACTTCGTTGTTATAACAACCAAG
GT_mut TGCGAAATAAAATTTAAATGATGAGATAAAATAAGGAATCAAATACTTCGAAGCTATAACAACCAAG
wt
egfp RRP40
– dT – dT
SV40L-pA
PROMPT pA
7SL
1 2 3 4
wt GT_mut hex_mut
– dT – dT – dT
SV40L-pA
PROMPT pA
5 6 7 8 9 10
b
CMV
proTMEM97
SV40L-pA
+320
wt ATTAAAAAAATAAA
hex_mut ATCAAGAAGACAAG
RT-qPCR values relative to egfp samples
5'AMPL
3'AMPL
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
egfp RRP40 egfp RRP40
wt hex_mut
c
ol1
AAAAA
AATAAA
rt
CMV
SV40L-pA
wt hex_mut
– dT ol1 rt – dT ol1 rt
readthrough
SV40L-pA
internal A stretch
ol1
7SL
1 2 3 4 5 6 7 8


1.00
0.75
0.50
0.25
0.00
Rescaled GROseq read count
PROMPT TSSs
mRNA TSSs
Transcription
−800
−400
0
400
800
Relative positions to max GROseq peak adjacent to respective TSS (nt)


a
AWTAAA
5'SS
PROMPT TSSs
mRNA TSSs
Fraction of regions with pattern (cumulative)
0.0
0.2
0.4
0
100
200
300
400
500
Distance from mRNA or PROMPT TSS (nt)
b
3' ends
Number of regions with 3' ends (cumulative)
0
200
400
600
0
250
500
750
1,000
Distance from mRNA or PROMPT TSS (nt)
c
AWTAAA
5'SS
Fraction of regions with pattern (cumulative)
0.0
0.2
0.4
0.6
0
250
500
750
1,000
Distance from mRNA TSS (nt)
mRNA TSSs with downstream 3' ends
mRNA TSSs without downstream 3' ends