Supplementary Methods

Subcellular fractionation

Two procedures were used. 1. Cells were harvested, washed twice in PBS, and lysed in 250 μl/10^6 lysis buffer (10mM PIPES_KOH pH 7.0, 10mM KCl, 1.5mM $MgCl_2$, 0.34M sucrose, 1mM DTT, and 40 μg/ml digitonin) on ice for 10 min. Nuclei were pelleted at 600xg for 5 min at 4°C and the pellet was washed twice in lysis buffer to avoid cross-contamination. The supernatant (cytoplasmic fraction) was re-spun to clarify debris. Cytoplasmic and nuclear RNA was extracted with acidic phenol and TRIZOL respectively. 2. Cells were harvested, washed in PBS, and lysed in (50mM Tris-Cl pH 8.0, 140mM NaCl, 1.5mM $MgCl_2$, 0.5% (v/v) Ipegal CA- 630 (Sigma)) on ice for 5 min. Nuclei were pelleted at 500xg for 5 min at 4°C and the pellet was washed once in lysis buffer to avoid cross-contamination. The small RNA fraction was directly purified using the Qiagen RNA/DNA kit according to manufacturers' instructions.

RNA extraction

Total RNA from HeLa and HepG2 cells was extracted with TRIZOL with subsequent enrichment for small RNAs (< 200 b) using the *mir*Vana (Ambion) or the Qiagen RNA/DNA kit according to manufacturers' instructions.

Ribosomal RNA depletion

Small RNA from HepG2 cells was depleted of 5S and 5.8S rRNAs using the Human/Mouse RiboMinus™ Transcriptome Isolation Kit (Invitrogen) according to the manufacturer's protocol except that we used custom-designed biotinylated LNA probes against the 5S and 5.8S sequences with every 3rd base substituted for LNA.

5S-1 - 5'/5Biosg/TT+CCC+AGG+TCT+CCC+AT;
5S-2 - 5'-/5Biosg/TC+AGG+GTG+GTA+TGG+GCG+TA;
5.8S-1 - 5'-/5Biosg/CT+TCA+TCG+ACG+CAC+GAG+CC;
5.8S-2 - 5'-/5Biosg/CG+CTC+AGA+CAG+GCG+TAGC-3'.

Cloning of small RNAs

Libraries were generated from 5 μg small RNA starting material. RNA was denatured at 85°C for 5' and a polyC-tail was added using the polyA kit (Ambion) substituting ATP for CTP. Samples were phenol extracted, ethanol precipitated and adapters were ligated to the 5' end using T4 RNA ligase Ambion or NEB at 4°C overnight, followed by one hour at room temperature with addition of 1 μl fresh enzyme. Reactions were incubated further at room temperature for 1h followed by phenol extraction and precipitation. Reverse transcription was carried out with adapters bearing a 15nt polyG tail using Superscript III (Invitrogen) according to the manufacturers instructions or Thermoscript (Invitrogen) with a slow cool down ramp from the 70°C denaturing to 60°C followed by a 90min extension at 60°C and inactivation at 75°C for 15 min. 20-25 cycles of PCR were carried out with primers complementary to the ligated/RT-added adapters. PCR

also introduced the Solexa/Illumina p5 and p7 anchors to adapt the small RNA library for Illumina sequencing. Two sets of linkers and primers were used independently with similar results. PCR products were size excised from low melting point agarose gels (100-300 bp range) and either phenol extracted or purified using the Illustra PCR DNA and Gel Band Purification Kit.

REACTION		OLIGO
5' ligation	Protocol 1	5'- rArCrArCrUrCrUrUrUrCrCrCrUrArCrArCrGrArCrGrCrUrCrUrUrCrCrGrArUrC-3'
	Protocol 2	5'-rCrGrArCrUrGrGrArGrCrArCrGrArGrGrArCrArCrUrGrArCrArUrGrGrArCrUrGrArArGrGrArGrUrArGrArArA-3'
RT	Protocol 1	5'-ATTGATGGTGCCTACAGGGGGGGGG-3
	Protocol 2	5'-TCGCGAGCGGCCGCGGGGGGGGGGGGGGG-3'
PCR	Protocol 1	5'-AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACG-3' 5'-CAAGCAGAAGACGGCATACGATTGATGGTGCCTACAG-3'
	Protocol 2	5'-AATGATACGGCGACCACCGACACGAGGACACTGACATGGACTGAAGGAGTAGAAA-3' 5'-CAAGCAGAAGACGGCATACGATCGCGAGCGGCCGCGGGGGG-3'
sequencing	Protocol 1	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC-3'
	Protocol 2	5'-CACGAGGACACTGACATGGACTGAAGGAGTAGAAA-3'

Enzymatic treatments

For 5' end analysis RNA samples were treated with Calf intestinal alkaline phosphatase (NEB) for 1 hour according to the manufacturer's instructions followed by phenol extraction and ethanol precipitation or with 1U/µg Tobacco Acid Pyrophosphatase (Epicentre) for 2 h at 37°C prior to cloning.

Immunoprecipitation

Cells were trypsinized, washed with PBS and lysed in 3 volumes of lysis buffer (20mM Tris, pH8, 8.5mM KCl, 0.5% NP40 1mM DTT, complete protease inhibitor), dounced and incubated on ice for 15 min. Supernatant was clarified by centrifuging at 14000g for 10 min. Samples were diluted 10x in IPP buffer (10mM Tris, pH7.4, 150mM KCl, 0.1% NP40). Samples were incubated with m3m7G antibody (Synaptic Systems) overnight prior to addition of protein G agarose (Roche) for an additional 3-4 hours. Beads were washed 3x 10 min at 4°C with IPP buffer. RNA was extracted with phenol/chloroform and ethanol precipitated.

Sequencing

Sequencing was performed on the GA1 and GA2 Illumina platforms according to the manufacturer's recommendations, except that a custom sequencing primer (CACGAGGACACTGACATGGACTGAAGGAGTAGAAA) was used for some samples. Base calling was in part performed with Alta-Cyclic [21].

PASR transfection and RT-QPCR

PASR sequences were identified using high-resolution tiling array-based maps of small RNAs from HeLa and HepG2 cell lines[4]. Sequence coordinates are based on the hg18 v35 annotation of the human genome. RNA oligonucleotides were synthesized and HPLC purified (IDT) (Table S2). HeLa and HepG2 cells were transfected with single stranded RNA oligos using methods similar to those employed for siRNAs[20]. Briefly, 15nM of each oligo was reverse-transfected onto cells in a 6-well dish using Lipofectamine 2000 (Invitrogen) and total RNA was harvested 72hrs later using RNeasy minipreps (Qiagen) and DNase treatment (Qiagen). Quantitative RT-PCR was performed as described[4] using an ABI 7300 realtime PCR instrument and SYBR Green 1-step PCR reagents (Applied Biosystems)(myc forward primer: ATCCCGGAGTTGGAAAACAAT, myc reverse primer: TGAGCTTTTGCTCCTCTGCTT, CTGR forward primer: ACTCCCAAAATCTCCAAGCCTAT, CTGF reverse primer: ATCGGCCGTCGGTACATACT).

The MYC responsive reporter (Clontech) was transfected into cells using Fugene (Roche) and then transfected with RNAs 24 hours later. Luciferase readings were gathered 48 hours later using the Bright Glo reagent (Promega) and a multimode plate reader (Tecan).

Data Origin and Formatting

All analysis was performed on the hg18 assembly version of the human genome. Annotations were downloaded from the UCSC Genome Browser (Tables section, track Old Known UCSC Genes version 2). Some of the annotations used for filtrating small RNAs (sRNAs) such repeated regions, annotated sno- and miRNAs and predicted RNA genes were also obtained from the following UCSC browser tracks: RepeatMasker, sno/miRNAs, RNA genes (see Fig. S1 and the main text for additional details). Finally, the current version of the predicted snoRNAs used for filtering was obtained from the snoSeeker[7] website: http://genelab.zsu.edu.cn/HSsnoRNA.html . Promoter regions were defined as the set of genomic positions no more distant than 500 bp away from a Transcription Start Site (TSS) of the annotated RNAs.

CAGE tag data

CAGE tag sequences were downloaded from the RIKEN institute web site (http://fantom3.gsc.riken.jp/)[9]. Sequences were aligned onto the human genome (assembly version hg18), requiring the sequences to align with 100% similarity over their full length at a unique position in the genome. Other matches were discarded from this version of the analysis. Alternative mapping methods have been tested -allowing mismatches and/or partial lengths- to ensure that resulting conclusions were not affected by using a high mapping stringency (not shown).

A total of 1,159,964 reads mapped uniquely on the genome, representing 429,641 different sequences. 390,419 distinct genomic positions are defined by considering the 5' extremity of each mapped read. We refer to this set of distinct 5' ends as "collapsed" data. Small RNA sequencing reads were mapped in a

similar fashion, requiring 100% match over the entire sequence after removing the polyC track and the adaptor oligo sequences (Fig. S1). The same collapsing method was applied to small RNA sequence reads (Fig. S1).

Distributions of CAGE tags/sRNAs around TSS regions (Fig 1c and 2a).
Each mapped sequence read (CAGE tags and sRNAs for Fig. 1c and 2a respectively) was assigned to the closest annotated TSS. Distances are computed from the position of the TSS to the 5' end of the sequence. All of these distances were considered for these histograms (no collapsing). Reads further than 500 bp away from the closest TSS were assigned to different functional categories according to the UCSC genomic annotation. Annotations on the same strand or on the opposite strand with respect to the mapped sequence were considered separately for plotting the sense (red) and the antisense (green) distribution respectively. The height of each bar represents the total number of reads with a 5' end in the corresponding 50 bp bin. Small RNAs within 500 bp of a known TSS represent promoter-associated short RNAs as described previously[4].

Distribution of sRNAs over CAGE tags (Fig 2b and 3d).
For each of the 390,419 mapped and collapsed CAGE tags, we reported the genomic distance to the closest mapped sRNA of a given category (5' end to 5' end). These categories are for the Fig 2b: genic PASR (from a TSS to 500 bp downstream), intergenic PASR (within the 500 bp upstream of a TSS and not included in any transcript), non-PASR genic and non-PASR intergenic. For Fig 3b, sequences from sRNAs immunoprecipitated with anti-cap antibody and mapped within internal exons were analyzed. Since we consider only the shortest distance to a sRNA from each CAGE tag, the distribution is identical whether collapsed or non collapsed sRNAs are used. If two sRNAs are equidistant (one upstream and one downstream), one of them is randomly picked. Note that this configuration is exceptional and alternative strategies -representing both sRNAs- did not lead to significant changes in the distributions (not shown). Only distances between reads that are mapped on the same strand were considered.

Statistical significance of the proximity between CAGE tags and cap-IP sRNAs in internal exons (Fig 3)
To assess whether the observed coincidence between the mapping positions of CAGE tags and immunoprecipitated sRNAs in internal exons was higher than expected by chance, we performed the following sampling experiment: from the set of all 16,673 sRNAs located in internal exons (and further than 500 bp away from any TSS), 1,000 distinct subsets were randomly sampled. Each control subset was generated independently, and contained the same number of sRNAs found in the set of cap-IP sRNAs mapped in internal exons (1,874). The proportion of sRNAs located at 10 bp or closer from a CAGE tag was computed in each random subset (10% on average, compared to 28% in the set of cap-IP sRNAs). Despite the stringency of this control –as the total population of sRNAs

should also contain the capped sRNAs- none of the random subset produced such a prevalence of proximal sRNAs, which clearly indicates that the immunoprecipitation protocol results in a significant enrichment of sRNAs that correlate with CAGE tags in internal exons.

Distribution of CAGE tags over internal exons (Fig 3a).
From the UCSC genomic annotation, we retrieved all internal exons that do not overlap any known initial or terminal exon. To remove redundancy across identical exons from different annotated transcripts, only one representative was kept for a given instance of chromosome/start/end/strand values. Note that exons can still overlap in their genomic space though, such as the ones that are generated from alternative splicing. For each of these internal exons, distances from the acceptor site (5' end of the exon) to the mapped CAGE tags (5' end of the read, collapsed data, same strand) were converted in proportion of the length of the corresponding exon. Fig 3a (left panel) represents the distribution of these relative positions within and outside the exonic boundaries (intronic flanking regions are also displayed).

To confirm that the decay of CAGE tags mapping at the end of the exons was due to tags coming from spliced RNAs, we looked among the tags that could not be mapped on the genome for those that could be mapped to known Exon-Exon Junctions (EEJ). From each pair of consecutive exons longer than 50 bp in the set of annotated transcripts, EEJs were retrieved by merging the last 50 bp of the upstream exon and the first 50 bp of the downstream exon. After filtering for redundancy, the final set contained 171,875 distinct EEJ sequences.

CAGE tag sequences with no perfect match on the genomic sequence (obtained from the RIKEN web site, see "Data origin") were aligned on the EEJ set by using the MUMmer software[22] with the longest common substring method and no mismatches allowed. 4,673 CAGE tags could be mapped on these EEJ, which represents about 6% of the non-collapsed tags from the "internal exonic" category (78,917 non collapsed CAGE tags in these genomic regions). The distribution of these positions is displayed in the right panel of the Fig 3a.

Genes producing CAGE tags in internal exons (Fig 3b)
We sought to estimate the prevalence of genes that are involved in the production of CAGE tags from internal exons. For this purpose, we computed for every gene the proportion of internal exons that contained a mapped CAGE tag by proceeding as follows: first, all internal exons that have no overlap with any initial or terminal exon were projected on the genomic space in order to merge identical or overlapping exons into non overlapping contigs. Thus, each exonic cluster is defined as the union of the overlapping exons and is referred as a single contig, simply called "internal exon" in this part of the study. This projection is performed in order to assign a simple number of internal exons to each gene in a genomic space, regardless of alternatively spliced transcripts. Obviously, genes with fewer than 3 exons were not considered in this analysis. Next, we

reported for each exon how many distinct mapped 5' ends of CAGE tags it contains (collapsed data). For any gene name, it was then possible to define two values (Fig 3b): the total number of internal exons (X-axis) and the number of internal exons that contain one or more CAGE tag (Y-axis). The number of genes that correspond to a given (X;Y) configuration is represented on the heat map by a color scale (Fig 3b).

Given the fact that 15.5% of the internal exons overall contain a CAGE tag, it is not surprising to find several exons from the same locus. To estimate if a substantial fraction of signal in the heat map was unlikely to occur by chance, we proceeded to perform a random permutation test. For each possible number of internal exons in a gene, an equal number of internal exons was randomly picked independently (without ensuring that they come from the same locus), and the number that contain one or more CAGE tags was reported. For each position of the X-axis, 1,000 of such sub-samples were tested. The continuous line indicates the maximum values observed after the randomizations, defining the upper part of the heat map as highly significant (P-value<.001).

As illustrated in the APOB example (Fig 3c), several positions in the same exon can be marked by CAGE tags. Here, we observed 606 genes (about 2%) in which all internal exons contain at least one CAGE tag; we can also identify 96 with at least three different CAGE tags in every internal exon.

The following table indicates the proportion of genes that have at least one internal exon in which a given number of CAGE tags are found.

Min. number of CAGE tags in the same internal exon	Number of genes with such an internal exon	Corresponding proportion of genes (%)
1	15360	48.6
2	6793	22.5
3	3822	12.1
4	2510	8.0
5	1763	5.6
6	1266	4.0
7	970	3.0

Note that the number of CAGE tags refers here to collapsed data, which means that distinct 5' ends of sequence reads are mapped within the same internal exon.

Correlation assessment between CAGE density and gene expression

To test whether the observed densities of CAGE tags could be simply a function of the expression level of the gene, we tried to correlate CAGE tag data with tiling array expression data for HepG2 long cytosolic polyA+ RNA[4].

The data is available from the UCSC genome browser under URL http://hgdownload.cse.ucsc.edu/goldenPath/hg18/affyTxnPhase3/lRNA.affyTxnPhase3HepG2Cyto.wigVar.gz . For consistency with the data from the tiling array experiment, we focused in this experiment on the CAGE tags derived from the HepG2 libraries only.

Several measurements were performed in different regions of the annotated transcripts, defining 4 categories of data:

Group A: CAGE tag density in PASR regions. The number of CAGE tags (non collapsed data) was computed in the -500/+500 bp window around the TSS of each transcript.

Group B: CAGE tag density in internal exon. For each transcript, the number of CAGE tags (non collapsed data) included in the internal exons was measured and normalized by the total length of these internal exons (expressed in # of tag/Kb of internal exon).

Group C: Expression of the first exon obtained from tiling array. For a given transcript, we reported the median intensity value among all of the probes that are included within the first exon.

Group D: Expression value in internal exons. Considering the probes spanned by the internal exons of a transcript, we defined the expression as their median intensity value.

The following table summarized the results of the correlation assessments in terms of R-square values (regression on a linear model computed with the statistical package R).

	A	B	C	D
A	1	0.09	0.25	NA
B	-	1	NA	0.21
C	-	-	1	0.4
D	-	-	-	1

Results show that some correlation can be observed within the same transcript between the expression value of the initial exon and that of the internal exons (R-square=0.4). However, only a weak correlation could be detected between CAGE tag data and gene expression data, whether initial or internal exons are probed (R-square=0.25 and 0.21 respectively). This absence of correlation argues against the hypothesis that internal CAGE tags might result from systematic degradation as a function of transcript abundance. Interestingly, the fact that no obvious correlation could be measured between the densities of CAGE tags in

the first exon vs. internal exons suggests two or more partially independent regulation mechanisms.

Implementation

Apart from cases specifically mentioned in the text, the methods and algorithms designed for the computational analyzes have mainly been implemented in a collection of in-house Perl, Awk and Shell scripts.

Correlation of ChIP-chip data and PASR density

ChIP chip data from HeLa and HepG2 cells for different chromatin marks and proteins as well as DNase sensitivity measurements were extracted from the UCSC ENCODE database. Promoters were grouped depending on the ChIP-chip signal intensity as indicated in figure S3. The number of collapsed and uncollapsed PASR reads in genes comprising each of the groups was analyzed.

Supplementary Methods References

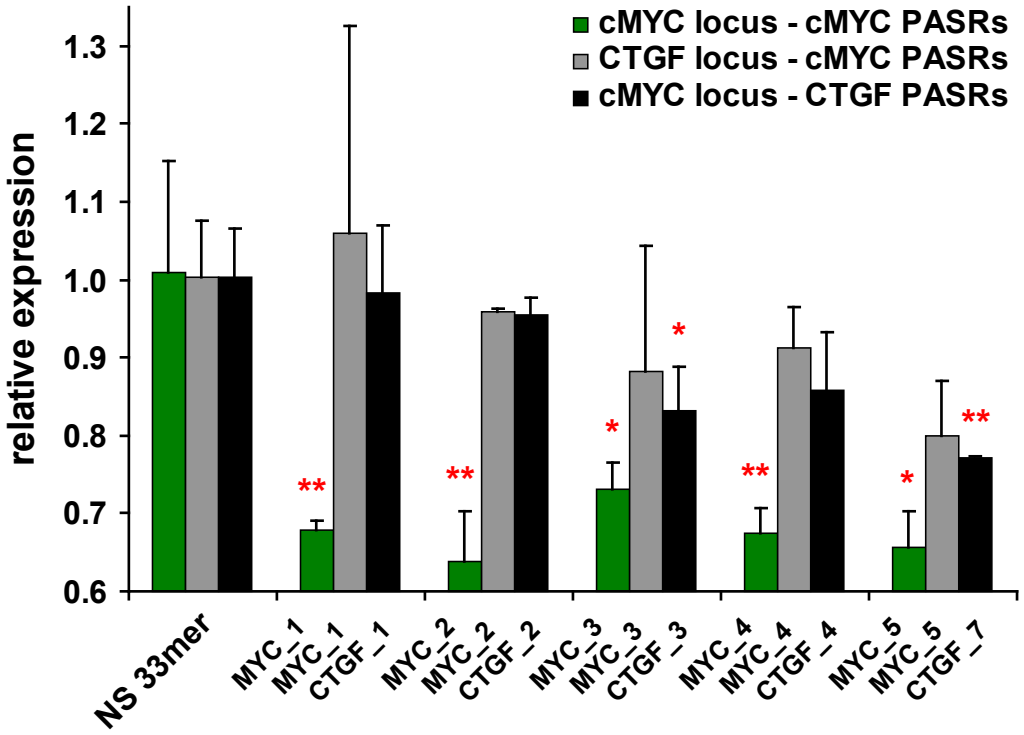
21. Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. & Hannon, G. J. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods 5, 679-82 (2008).
22. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics Chapter 10, Unit 10 3 (2003).

Short RNA processing pipeline

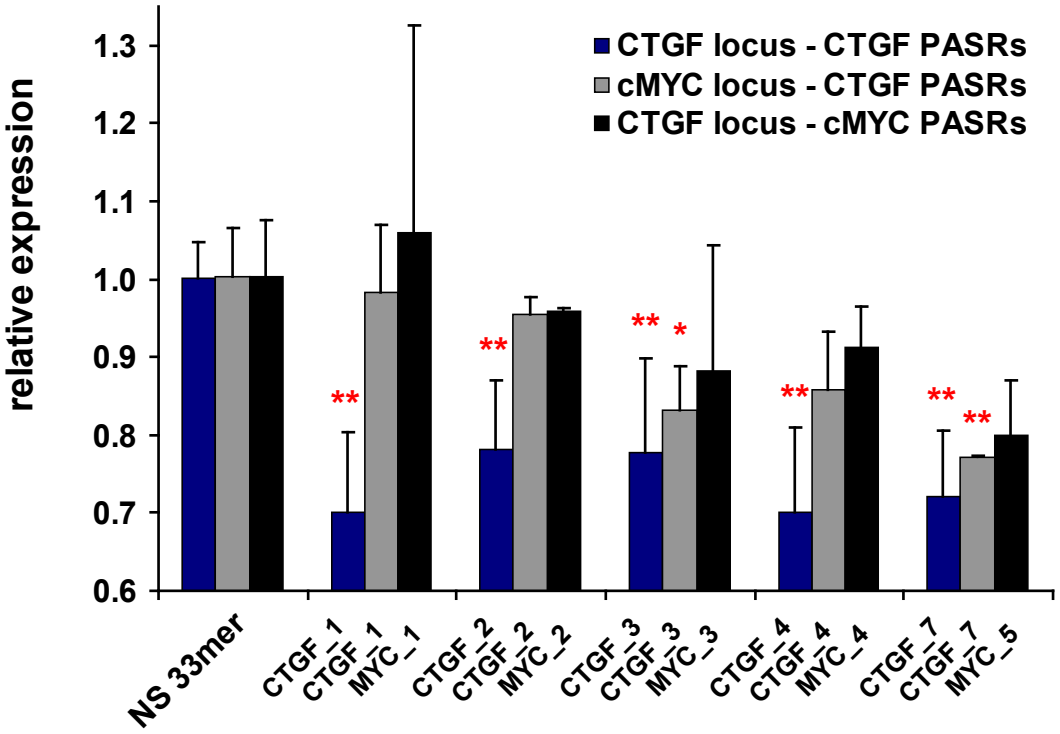
Step	Description	# of remaining sequences
I. Sequencing	5' end reads using Illumina platform	78,586,728
II. Mapping reads to the genome	1. After clipping of C-tail and adapter sequence, map to version hg18 with complete alignement of entire sequence NNNNNNNNNNNNNNNNNNNNCCCCCCCCC Used for mapping | C-tail and adapter	29,020,429
	2. Keep only sequences that map uniquely to the genome	9,506,432
III. Filtration	1. Removal of sequences derived from mitochondria and chrY	8,773,534
	2. Removal of sequences overlapping repeats (RepeatMasker)	6,801,591
	3. Removal of sequences overlapping sno- or mi-RNAs (miRBase)	642,427
	4. Removal of sequences overlapping a hand-curated set of RNAs (U12, RNAse P RNA component etc)	600,376
	5. Removal of sequences overlapping predicted RNA genes ("RNA Genes" track on the UCSC browser)	536,120
	6. Removal of sequences overlapping predicted snoRNAs (http://genelab.zsu.edu.cn/HSsnoNRA.html)	232,805
IV. Treatment of sequences post filtration	1. Collapse sequences that have the same 5' ends	102,159
	2. Segregate sequences into promoter-associated (PASRs) and non promoter-associated (nonPASRs). Promoters are defined as +/- 500 bp from the start sites of UCSC Old Known Gene annotations	
	3. Segregate sequences from PASR and nonPASR class into further subclasses: exonic (internal exons for nonPASRs), intronic and intergenic etc.	

Figure S1

Figure S1. Short RNA processing pipeline. A simple representation of the steps applied to short RNA sequences is shown for reference.



P-values (t-test) are noted by asterisks: * <0.01, ** <0.005



P-values (t-test) are noted by asterisks: * <0.01, ** <0.005

Figure S2. PASRs act specifically on their gene of origin. (**A**) The activity of MYC PASRs on mRNA expression was measured at the MYC locus (green bars) and compared to the activity of these PASRs at the CTGF locus (gray bars). Both experiments were performed in HeLa cells. Also, the activity of CTGF PASRs on the transcription of the MYC locus was measured in HepG2 cells (black bars). For all samples 2-3 biological replicates were evaluated. (**B**) The activity of CTGF PASRs on mRNA expression at the CTGF locus was measured (blue bars) with 5-7 biological replicates per sample. This data was compared to the activity of CTGF PASRs on the MYC locus (gray bars). Both experiments were performed in HepG2 cells. The affect of MYC PASRs on the CTGF locus was also measured in HeLa cells (black bars). P-values (t-test) are noted by red asterisks: * < 0.01 and ** < 0.005. Note that the MYC PASRs on the CTGF locus (gray bars in panel A and black bars in panel B) represent the same data points. Similarly, the CTGF PASRs on the MYC locus (black bars in panel A and gray bars in panel B) are drawn from the same data points. All experiments were normalized to a 33bp RNA control as described in the methods.

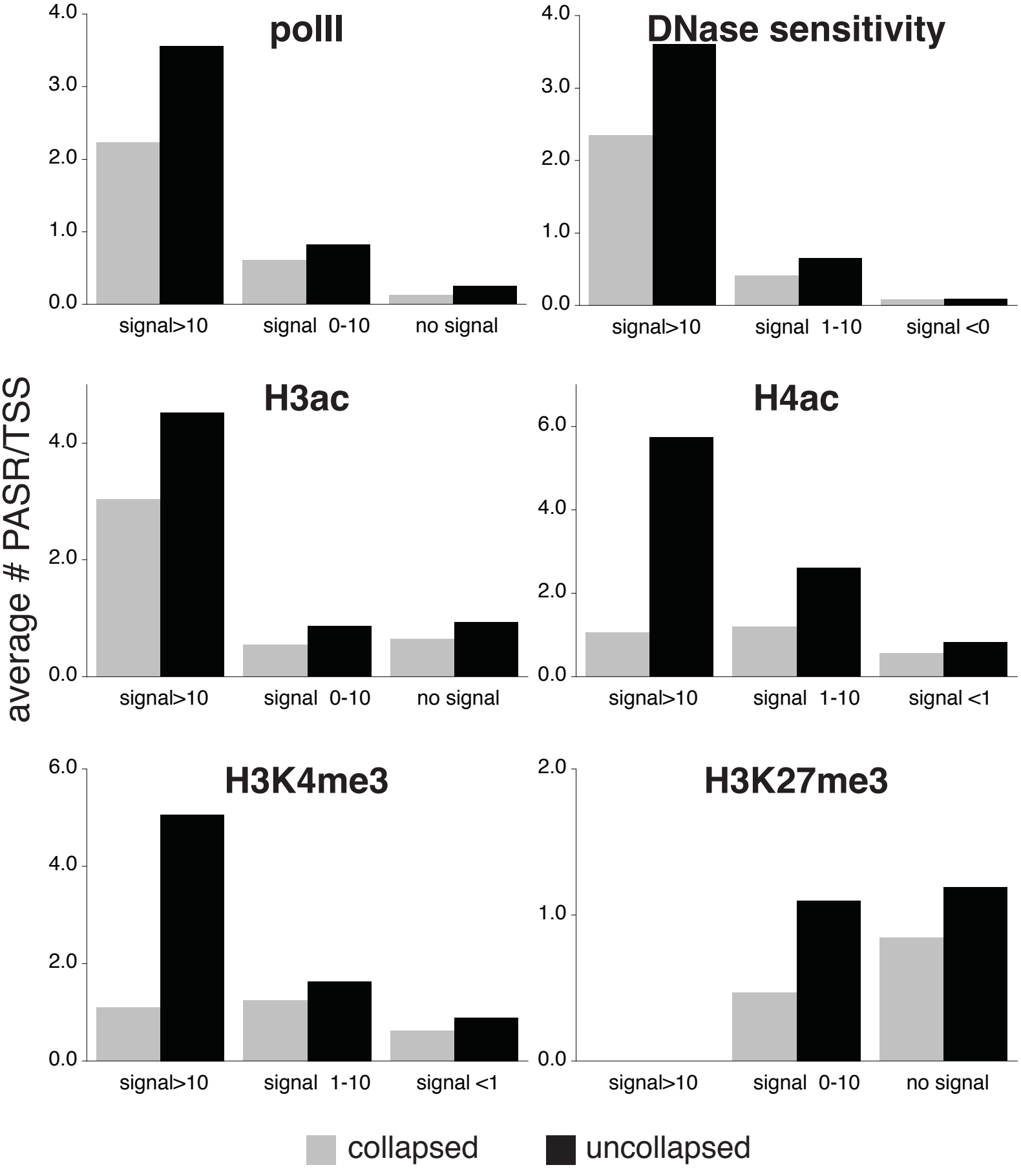

polII
DNase sensitivity
H3ac
H4ac
H3K4me3
H3K27me3
average # PASR/TSS
signal>10
signal 0-10
signal 1-10
no signal
signal <0
signal <1
collapsed
uncollapsed

Figure S3

Figure S3. Prevalence of PASR sequences correlates with chromatin state. PASR density (number per 1kb tss region) correlates with the ChIP signal intensity of active chromatin marks (H3 and H4 acetylation, H3K4-me3), the presence of RNA polymerase II at the promoter and DNase sensitivity (suggesting an open chromatin state). A negative correlation was observed with H3K27-me3, a repressive chromatin mark. Collapsed (gray) and uncollapsed (black) sequence numbers are shown.

Table S1

A. Distribution of CAGE tags.

<table>
<tr><td>Within 500 bp of a TSS</td><td colspan="5">>500bp from TSSs</td></tr>
<tr><td rowspan="6">475,158 (41.0%)</td><td colspan="4">683,331 (59.0%)</td><td></td></tr>
<tr><td colspan="3">GENIC</td><td colspan="2">INTERGENIC</td></tr>
<tr><td colspan="3">348,803 (30.11%)</td><td colspan="2" rowspan="4">334,528 (28.88%)</td></tr>
<tr><td colspan="2">Exonic</td><td>Intronic</td></tr>
<tr><td colspan="2">131,564 (11.36%)
121,683 over sense 10.5%
10,830 over antisense 0.93%</td><td>217,239 (18.75%)
187,539 in sense intron 16.2%
33,479 in antisense intron 2.9%</td></tr>
<tr><td>Initial</td><td>Internal</td><td>Terminal</td></tr>
<tr><td>6,417
0.55%</td><td>78,917
6.81%</td><td>41,860
3.61%</td><td></td></tr>
</table>

Percentages are fractions of the total number of un-collapsed CAGE tags: 1,158,489 sequences.

B. Distribution of small RNAs.

<table>
<tr><td>Within 500 bp of a TSS</td><td colspan="5">>500bp from TSSs</td></tr>
<tr><td rowspan="6">37,647 (16.2%)</td><td colspan="4">195,158 (83.8%)</td><td></td></tr>
<tr><td colspan="3">GENIC</td><td colspan="2">INTERGENIC</td></tr>
<tr><td colspan="3">118,351 (50.8%)</td><td colspan="2" rowspan="4">76,807 (33.0%)</td></tr>
<tr><td colspan="2">Exonic</td><td>Intronic</td></tr>
<tr><td colspan="2">38,281 (16.4%)
37055 over sense 15.9%
1,552 over antisense 0.67%</td><td>80,070 (34.39%)
63,938 in sense intron 27.5%
17,205 in antisense intron 7.4%</td></tr>
<tr><td>Initial</td><td>Internal</td><td>Terminal</td></tr>
<tr><td>493
0.21%</td><td>16,673
7.16%</td><td>20,837
8.95%</td><td></td></tr>
</table>

Percentages are fractions of the total number of un-collapsed sRNAs: 232,805 sequences.

Table S1. Distribution of mapped CAGE tags (A) and sRNAs (B) in different annotated categories of the genome. Sequence reads are non-collapsed.

TABLE S2: Synthetic PASRs from the *c-myc* and *CTGF* TSSs

Name	Chr	Start	Stop	Locus relation	Size (nucl.)	Synthesized RNA oligo
NS 33mer					33	GCGGAACCGAAGUGUAAUAGACCAGUGAUUACG
MYC_1	chr8	128817047	128817077	AS-promoter	30	CGCUAAGGCUGGGGAAAGGGCCGCGCUUUG
MYC_2	chr8	128817092	128817122	S-promoter	30	UGGUACGCGCGUGGCGUGGCGGUGGGCGCG
MYC_3	chr8	128817680	128817715	S-exon	35	GAGAGGCAGAGGGAGCGAGCGGGCGGCCGGCUAGG
MYC_4	chr8	128817781	128817811	AS-exon	30	GCGGGAGGGCUGGGCCAGAGGCGAAGCCCC
MYC_5	chr8	128817861	128817891	S-exon	30	CCCAUAGCAGCGGGCGGGCACUUUGCACUG
CTGF_1	chr6	132314158	132314207	S-promoter	49	UCACACAACAACUCUUCCCCGCUGAGAGGAGACAGCCAGUGCGACUCCA
CTGF_2	chr6	132313822	132313857	AS-intron	35	CGGCUGCAGGGAGGACAGGGCGGUCAGCGGCGCUC
CTGF_3	chr6	132313942	132313972	AS-exon	30	GCUGCAGAGGGCGAGGAGGACCACGAAGGC
CTGF_4	chr6	132313983	132314013	AS-exon	30	GGCCCAUACUGGCGGCGGUCAUGGUUGGCA
CTGF_7	chr6	132313767	132313797	S-exon	30	CGUGCCGGUGCCCGGACGAGCCGGCGCCGC
33-mer contr.					34	GCGGAACCGAAGUGUAAUAGACCAGUGAUUACG

Table S2. Artificial PASR sequences designed to the myc promoter region.