Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs

Kata Fejes-Toth, Vihra Sotirova, Ravi Sachidanandam, Gordon Assaf, Gregory J. Hannon, Philipp Kapranov, Sylvain Foissac, Aaron T. Willingham, Radha Duttagupta, Erica Dumais, et al.

Vol 457|19 February 2009|doi:10.1038/nature07759

LETTERS

Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs

Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project*

The transcriptomes of eukaryotic cells are incredibly complex. Individual non-coding RNAs dwarf the number of protein-coding genes, and include classes that are well understood as well as classes for which the nature, extent and functional roles are obscure[1]. Deep sequencing of small RNAs (<200 nucleotides) from human HeLa and HepG2 cells revealed a remarkable breadth of species. These arose both from within annotated genes and from unannotated intergenic regions. Overall, small RNAs tended to align with CAGE (cap-analysis of gene expression) tags[2], which mark the 5′ ends of capped, long RNA transcripts. Many small RNAs, including the previously described promoter-associated small RNAs[3], appeared to possess cap structures. Members of an extensive class of both small RNAs and CAGE tags were distributed across internal exons of annotated protein coding and non-coding genes, sometimes crossing exon–exon junctions. Here we show that processing of mature mRNAs through an as yet unknown mechanism may generate complex populations of both long and short RNAs whose apparently capped 5′ ends coincide. Supplying synthetic promoter-associated small RNAs corresponding to the c-*MYC* transcriptional start site reduced *MYC* messenger RNA abundance. The studies presented here expand the catalogue of cellular small RNAs and demonstrate a biological impact for at least one class of non-canonical small RNAs.

The repertoire of RNAs found in eukaryotic cells is unexpectedly complex, with virtually the entire non-repeat portions of many genomes being transcribed[1]. Genic regions are often populated by interleaved transcription units, which give rise to both protein-coding RNAs and long and short non-coding RNAs[1]. Promoter-associated short RNAs (PASRs) and termini-associated short RNAs (TASRs) are recent additions to the pantheon of short RNAs[4]. Although their functions are unknown, several of their characteristics support biological significance. For example, PASRs and TASRs cluster at 5′ and 3′ termini of annotated genes[4]. Overall, the presence of PASRs correlates with the expression of a given locus, but not all expressed loci generate these species. Moreover, the production of PASRs and TASRs from particular loci is a conserved feature of the human and mouse genomes[4]. As part of our ongoing effort to understand the full repertoire of small RNAs, their mechanisms of biogenesis and their biological impacts, we analysed the small RNAs (<200 nucleotides (nt)) of HepG2 and HeLa cell lines using next-generation sequencing[5].

Nearly 80 million short sequence reads (30–35 bases) were generated, representing RNAs <200 nt in both cell lines (Supplementary Fig. 1). Our sequencing protocols favoured RNAs with 5′ mono-, di- and tri-phosphate groups and capped RNAs. Nearly 30 million of these could be matched perfectly to the hg18 release of the human genome, with 9.5 million reads mapping to unique sites (Supplementary Fig. 1).

Sequences derived from mitochondria, chromosome Y, repeats, annotated small RNAs, predicted RNA genes[6], and known and predicted small nucleolar RNA (sno)RNAs[7] were excluded from further analysis (Fig. 1a and Supplementary Fig. 1). This resulted in 232,805 sequences representing new small RNAs. Independent libraries from the same cell line only modestly overlap, indicating that our studies have not saturated small RNA content. Sequences were collapsed based on their mapping positions, assigned as the 5′ nucleotide of each read. This resulted in 102,159 distinct 5′ ends ('rest', Fig. 1a). A large fraction of sequences are derived from unannotated intergenic regions (Fig. 1b). Sequences also matched the sense and antisense strands of exonic and intronic regions of annotated genes. Notably, nearly half of all reads could be assigned to the sense strand of annotated exons, with a strong representation of first exons.

We had previously noted a class of small RNAs, namely PASRs, which associated with transcriptional start sites (TSSs) and mapped both to promoter regions and to annotated first exons. We therefore plotted the distribution of the unannotated RNA category (rest, Fig. 1a, b) with respect to known TSSs. A clear pattern emerged with an enrichment of small RNAs on both strands directly adjacent to the TSS (Fig. 1c). The sense and antisense strands surrounding TSSs showed mirror-image profiles, with small RNAs on the sense strand accumulating most strongly downstream of the annotated TSS and small RNAs on the antisense strand accumulating mainly upstream of the annotated TSS. This is similar to what had been previously observed on high-resolution genomic tiling arrays[4]. A gap of ~50 nt on the antisense strand separated the precise TSS and PASRs, an observation we have yet to understand (Fig. 1c).

This PASR class (previously defined as small RNAs mapping within 500 nt of an annotated TSS) is found in both genic and intergenic regions of the genome and comprises 16.2% or 17.7% of filtered sequence tags (un-collapsed or collapsed, respectively). PASRs contribute to the small RNAs placed within both strands of exons and introns and small RNAs annotated as intergenic (Fig. 1b, shaded inner circles in each pie chart). On the basis of these definitions, PASRs form the most abundant individual class of defined small RNAs within the non-annotated fraction of our sequences.

Because PASRs strongly associate with TSSs, we posited that transcription initiation *per se* might generate PASR 5′ ends. Thus, PASRs might contain cap structures. To probe this possibility, we prepared small RNA libraries using methods that require the presence of a 5′ phosphate and 3′ hydroxyl group[8]. Capped RNAs should be refractory to capture by this cloning protocol, but could be made susceptible to cloning by removing cap structures. We therefore prepared three different small RNA libraries from HepG2 cells. One was from untreated RNA. The second was from RNA treated with tobacco acid pyrophosphatase (TAP) to leave clonable monophosphorylated ends on RNAs with caps, or with di- or tri-phosphate termini. The third library was from RNA treated with calf intestinal alkaline phosphatase (CIP) before TAP treatment. Pre-treating with CIP removes phosphates to leave unclonable 5′ OH termini on all uncapped

*Lists of participants and their affiliations appear at the end of the paper.

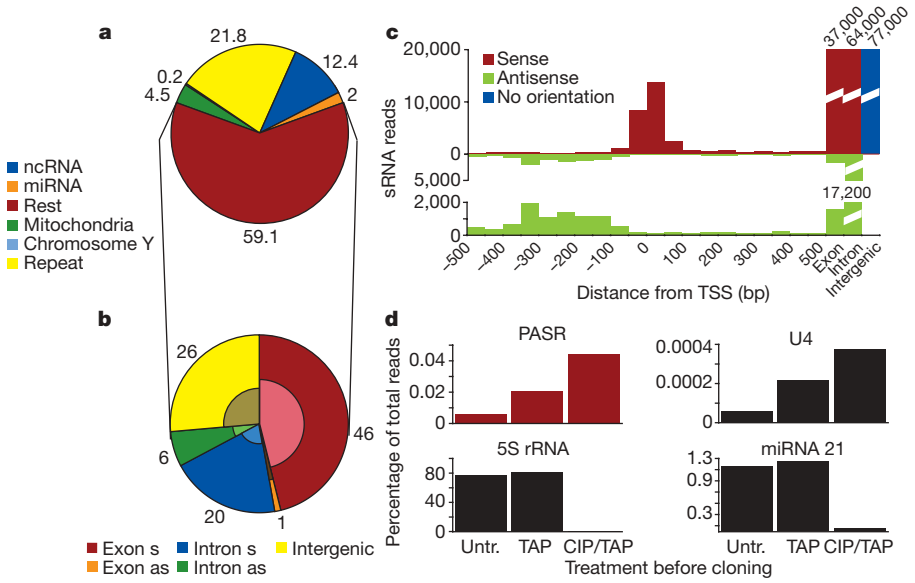

Figure 1 | Genomic distribution of small RNAs. **a**, Annotation of sRNAs from sequencing. 'Rest' represents unannotated sRNAs filtered for mitochondria, chromosome Y, repeats and known sRNAs. miRNA, microRNA; ncRNA, non-coding RNA. (Collapsed data are shown.) **b**, Mapping of unannotated sequences to annotated genomic landmarks. as, antisense to corresponding transcript; s, sense. (Collapsed data.) Shaded inner portions represent the fraction that are PASRs. **c**, Distribution of sRNAs over TSSs. Orientation is with respect to the long transcript. Antisense sRNAs are plotted with a different *y*-axis beneath. (Uncollapsed data.) **d**, Characterization of PASR 5′ ends. Untr., untreated. Sequences corresponding to the 5′ end of U4, 5S rRNA and mir-21 were extracted as controls. (Uncollapsed data.)

RNAs. Sequence tags corresponding to the 5′ end of the capped U4 small nuclear (sn)RNA are enriched in libraries by TAP treatment and further enriched by CIP addition before TAP treatment (Fig. 1d). MicroRNA 21, which has a 5′ monophosphate terminus, is lost from the library on CIP treatment, as is 5S ribosomal RNA, which has a 5′ triphosphate terminus (Fig. 1d). Small RNAs defined as PASRs follow the pattern established by U4, consistent with them bearing some type of cap structure. The observation that PASRs are revealed to the cloning protocol by TAP alone indicates that they also contain 3′ OH termini. Considered together, these data indicate that PASRs are likely to arise either as independent capped transcripts emanating from annotated TSSs on both genomic strands or as processing products from longer capped RNAs. A candidate for the latter are PALRs (promoter-associated long RNAs), which often extend through the first exon and into the first intron[4].

CAGE tagging protocols take advantage of the 5′ cap structure to capture sequence reads from the 5′ ends of long RNAs[2]. Substantial databases of such tags have been produced from long polyadenylated RNAs from more than 20 human tissues[9,10]. Plotting CAGE tags with respect to annotated TSSs revealed patterns similar to those observed for PASR class small RNAs (Fig. 2a). In both genic and intergenic regions, we also observed a strong tendency for a precise identity between CAGE and PASR 5′ ends (Fig. 2b).

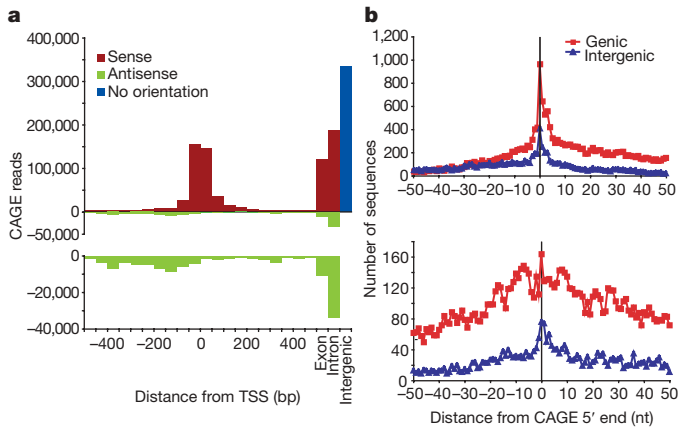

Figure 2 | Correlation of sRNAs and CAGE tags. **a**, Distribution of CAGE tags over annotated TSSs. Orientation is with respect to the long transcript. Antisense sRNAs are plotted with a different *y*-axis beneath. (Uncollapsed data.) **b**, Distribution of PASR (top) and non-PASR sRNAs (bottom) around CAGE tag 5′ ends. The distance to closest short RNA5′ end was plotted for each CAGE tag. (Collapsed data.)

We also noted a substantial population of small RNAs mapping more than 500 nt away from annotated TSSs (Fig. 1c), which contributed to intronic, exonic and intergenic classes (Fig. 1b, outer portions of each pie). Similarly we noted a large population of CAGE tags located >500 base pairs (bp) from the TSS (59.0%) that could be assigned to exons, introns and intergenic regions (11.36, 18.8 and 28.9% of total uncollapsed CAGE tags, respectively; Supplementary Table 1).

Certainly, a fraction of these could arise as products from unannotated TSSs, giving rise to both CAGE tags and PASRs. However, the correlation between the 5′ ends of non-PASR small RNAs and CAGE tags was less precise (±10 nt) than was noted at annotated TSSs (Fig. 2b). This indicated that non-PASR classes of small RNAs might arise by mechanisms that differ from canonical PASRs.

Both small RNAs and CAGE tags accumulate more strongly in internal exons than in introns or in intergenic space, if these regions are normalized by their cumulative length (Fig. 3a and not shown). By examining the distribution of both CAGE tags and small RNAs across annotated internal exons, we noted a strongly decreasing number of CAGE sequences beginning about 20 nt from the 3′ end of the average exon/intron boundary (splice donor site; Fig. 3a). If CAGE tags crossed splice junctions, they would not be co-linear with the genome at these sites and would, therefore, not have been mapped in the initial analyses[10], possibly giving rise to the observed pattern. We therefore extracted previously published CAGE tags, which had failed to map to the genome, and probed these against sequences of known exon–exon junctions. We uncovered a substantial population of CAGE tags that crossed splice junctions, and that therefore must have arisen from at least partially processed mRNAs (Fig. 3a).

CAGE tags are well established as markers of capped 5′ ends. Certainly, internal exons might contain unknown sites of transcriptional initiation that could give rise to both CAGE tags and small RNAs, which would then be defined as PASRs. However, we observe numerous tags that both initiate less than 20 bases from exon boundaries and cross exon–exon junctions. Very short exons splice inefficiently[11], and naturally occurring 5′ exons less than 20 bases in length are rare (not shown). Thus, the CAGE tags that we observe probably represent cleaved products of mature mRNAs that somehow acquire a 5′ modification analogous to a cap structure that renders them sensitive to the CAGE tagging method. Such a reaction would represent a previously unrecognized RNA processing pathway and a previously unknown fate for spliced mRNAs. Although the CAGE tags used for comparison in this study were derived from polyadenylated RNAs, we cannot determine whether small RNAs originated from poly(A)$^+$ or poly(A)$^-$ transcripts.

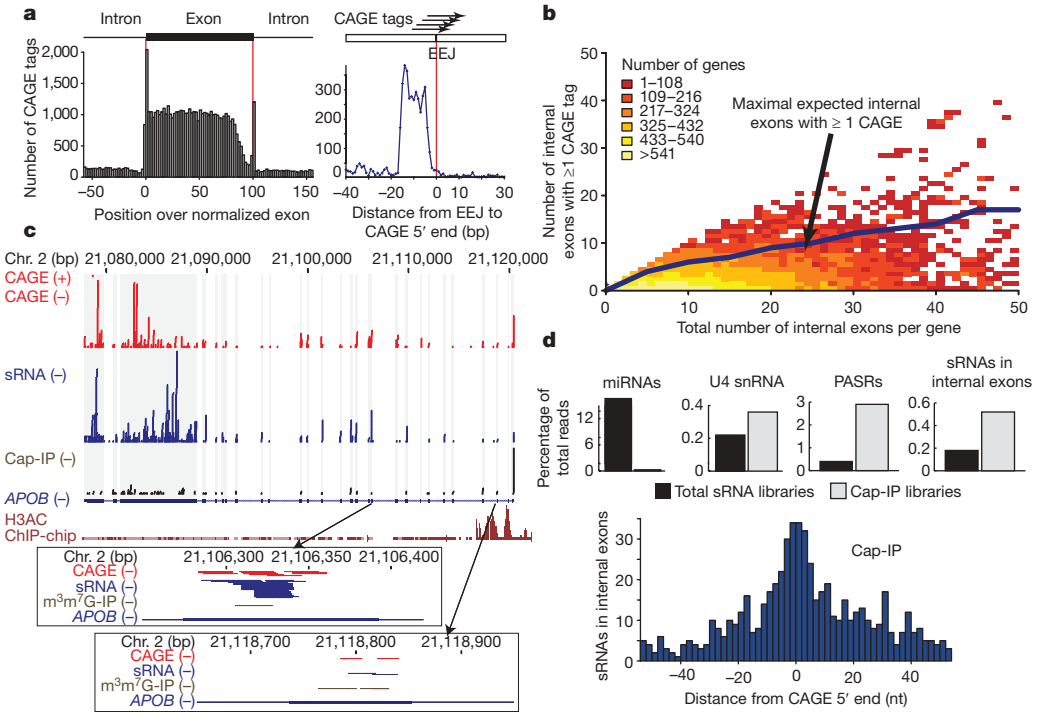

Figure 3 | Correlation between CAGE tags, sRNAs and internal exons of annotated transcripts. **a**, Left: distribution of mapped CAGE tag 5′ ends across internal exons. Exon length was normalized to 100 segments. Right: distribution of CAGE tags not mapping to the genome but mapping to exon–exon junctions (EEJ) of internal exons. **b**, Prevalence of internal CAGE tags. Black line represents the maximum expected exons in random samplings (see Methods). Colour corresponds to number of transcripts represented by each data point. **c**, CAGE tag and sRNA coverage of the *APOB* gene. sRNAs from cap-immunoprecipitation (IP) are shown separately. Histone H3 acetylation (H3AC) pattern[13] is shown below. Two internal exons are magnified. **d**, Characterization of libraries from anti-cap-immunoprecipitated RNA. Top panel: representation of sRNAs in total and IP libraries (uncollapsed data). For all but the U4 fraction, uniquely mapping sequences were considered. Bottom panel: distance to closest sRNA 5′ end from CAGE 5′end in internal exons (collapsed data).

Generation of CAGE tags from internal exons is not confined to a small number of genes. In fact, 49% of all human genes generate a CAGE tag mapping to an internal exon. For 2% of them, one or more CAGE tags are found in all of the internal exons (Fig. 3b). This exceeds the number expected by chance (P value < 0.001, Fig. 3b). Prevalent and systematic generation of both small RNAs and CAGE tags from internal exons is illustrated by *APOB* (Fig. 3c), a gene encoding the apolipoprotein B protein that regulates cholesterol metabolism[12]. In this instance, mapping of acetylated histone H3 (ref. 13) is consistent with a single prevalent TSS, which correlates with the presence of both CAGE tags and PASRs (Fig. 3c). However, both CAGE tags and small RNAs are even more abundant in internal *APOB* exons, and these often coincide at specific sites (Fig. 3c, insets). This provokes a model in which mature transcripts from the *APOB* gene are processed post-transcriptionally and in which processing products become modified by some type of cap structure. This possibility gains support from sequencing of small RNAs recovered by immunoprecipitation with a methylguanosine cap antibody[14]. This not only recovers PASRs but also enriches small RNAs that map less than 10 bp away from a CAGE tag in internal exons (P value < 0.001, see also Fig. 1d). In libraries prepared from unfractionated small RNAs, 10.0% of filtered sequences mapping within internal exons lie within 10 bp of a CAGE tag. This number increases to 27.0% in libraries prepared from RNAs immunoprecipitated with the anti-m^7G antibody.

As illustrated by *APOB*, genic CAGE tags and small RNAs are approximately ten times more likely to map within exonic than intronic regions (Fig. 3a, b). As with CAGE tags crossing exon–exon junctions, this result is consistent with a model in which CAGE tags can be derived from products of processed mRNAs.

Considering the prevalence of the PASR class, we sought to probe its potential biological function. As with many genes, PASRs are found at the annotated TSSs of the *MYC* oncogene[4]. We synthesized a collection of 30–35-nt, single-stranded RNAs that share their 5′ ends with three PASRs from the sense genomic strand and two from antisense strand upstream of the annotated TSS (Fig. 4a). These were transfected individually into HeLa cells, and their effect on the abundance of *MYC* mRNA was measured (Fig. 4b). In each case, transfection of the PASR mimetic reduced the expression of c-*MYC* mRNA. The consequences of these effects were measured by co-transfection of a *MYC*-responsive luciferase reporter construct, which showed reduced activity in the presence of each PASR (Fig. 4c). Similar results were obtained for five PASR mimetics corresponding to the connective tissue growth factor (*CTGF*) gene (Supplementary Fig. 2). The presence of PASRs is associated with marks of active transcription, including association with RNA polymerase II, histone H3 and H4 acetylation, and H3K4 tri-methylation, as well an increased susceptibility to DNase treatment (Supplementary Fig. 3). Our data indicate a causal connection between PASRs and active MYC expression, although we have not yet investigated the impact of delivering ectopic PASRs on the active marks with which the presence of endogenous species is correlated.

Profiling of small RNAs, defined as those less than 200 nt in length, has revealed a substantial complexity in the output of both genic and intergenic regions of the genome. These studies have raised two possibilities for the origin of PASRs. First, they may be produced as capped, independent transcription products from promoters that also generate long RNAs. Second, they may be generated as post-transcriptional processing products of longer RNAs that initiate at annotated TSSs.

A notable outcome of these studies is the finding that both CAGE tags and small RNAs decorate not only intergenic spaces but also internal exons of protein coding and non-coding transcripts. The existence of a large class of CAGE tags that are both adjacent to and cross splice junctions provides a *prima facie* case for the conclusion that long RNAs are metabolized into short RNAs that bear cap-like structures at their 5′ ends (Fig. 5). The long RNAs, which

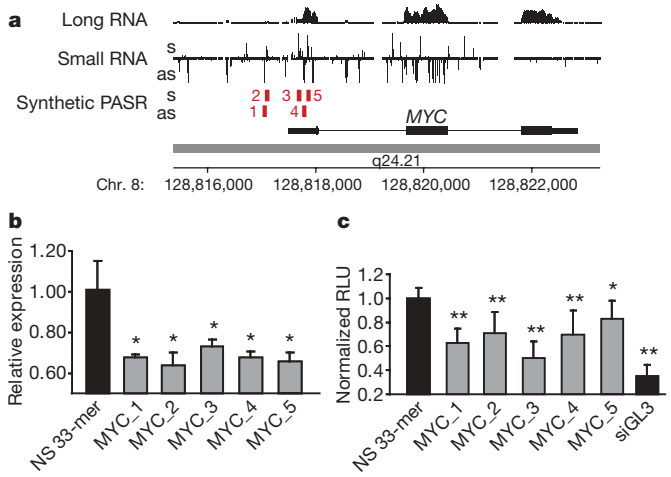

Figure 4 | Regulation of gene expression by PASRs. a, Expression profile of the *MYC* locus. The long and short RNA profile of HeLa cells on Affymetrix tiling arrays. Red rectangles indicate the designed synthetic PASRs (MYC_1–5 are denoted by numbers and sequence information is provided in Supplementary Table 2) corresponding to peaks in the sRNA array profile. **b**, *MYC* mRNA expression levels in HeLa cells as measured by quantitative PCR with reverse transcription ($n = 3$, P values < 0.01). **c**, Effects of PASR transfections on a *MYC*-responsive luciferase transcriptional reporter in HeLa cells was measured as relative light units (RLU) ($n = 2$, $^*P < 0.01$, $^{**}P < 0.001$). For reference, a control 33-mer and an siRNA directed against luciferase (siGL3) are shown.

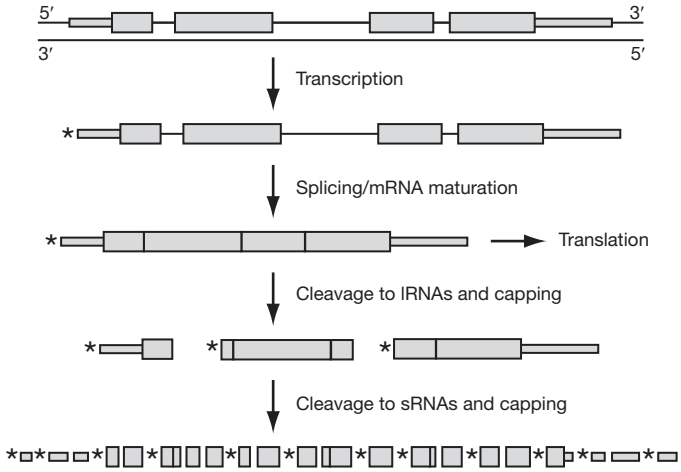

Figure 5 | A proposed model for the metabolism of genic transcripts into a diversity of long and short RNAs. Transcription of a genic region results in a precursor long RNA containing a 5′ cap structure, as shown by asterisks. After processing into spliced RNAs, protein-coding RNAs are destined either to be translated or to be further processed. This further processing entails cleavage followed in some cases by addition of a 5′ modification, possibly a cap structure. Additional cleavage of these intermediate products can generate a class of short RNAs, some also bearing a cap structure. lRNAs, long RNAs.

ultimately give rise to short RNAs, could be primary transcription products or processing products themselves. Moreover, our studies indicate that CAGE tags are capturing not only TSSs but also the 5′ ends of post-transcriptionally processed RNAs.

A key question remains as to whether the group of small RNAs that arise from internal exons represents transition products from mature mRNAs into recyclable ribonucleotides. Several lines of evidence argue against these representing simple degradation intermediates. First, there is a strong correlation between the precise 5′ ends of CAGE tags, derived from long RNAs, and small RNAs identified in our study. These maps were produced from various RNA and tissue sources, and by different groups. The results from independent samples are consistent and indicative of discrete processing sites. Second, based on chemical modification in the CAGE procedure and affinity purification for small RNA libraries, both types of tags significantly enrich under conditions that favour capped RNAs. Third, CAGE tags and small RNA species arise only from a discrete, although substantial, subset of genes, and the abundance of the non-PASR class does not correlate simply with the expression level of their generative loci (see Supplementary Methods).

Several studies have indicated that RNA interference directed to promoter regions and apparently non-transcribed portions of genes can have a regulatory impact. In some cases that impact is silencing[15,16], whereas in others activation was surprisingly observed[17]. Our analysis of PASRs indicates that providing their synthetic mimetics *in trans* can have a consistent, although modest, impact on gene expression. Although in the two cases tested, *MYC* and *CTGF*, increasing PASR levels decreased expression, it remains possible that the outcome of manipulating PASRs will be gene-specific, consistent with accumulating evidence that destroying promoter-associated RNA (PASR) species can have both positive and negative impacts[17].

The functions of small RNAs corresponding to intergenic and exonic regions remain obscure. Such species could have regulatory roles, *per se*, or they could participate more globally in a bookkeeping or quality control mechanism by which the cell records its transcriptional output and splicing patterns. This has been previously hypothesized as a role for non-protein-coding RNAs[18,19]. What is clear is that the transcriptional product of cells is captured in small, stable RNA populations to a degree that was unanticipated, and that at least a subset of these can serve as components of regulatory circuits.

METHODS SUMMARY

A detailed description of the methods is provided in Supplementary Information. Short RNA was extracted from HeLa and HepG2 cells and subjected to CIP and TAP treatments as indicated. Adaptor sequences were added by ligation at the 5′ end and by C-tailing and reverse transcription at the 3′ end. Libraries were PCR-amplified and sequenced on an Illumina GA2. RNAs bearing 5′ caps were enriched by immunoprecipitation using an m^3/m^7G antibody.

Analysis was performed on the hg18 assembly of the human genome using UCSC annotations. CAGE tag sequences were from the RIKEN institute website[9]. For short (s)RNA density, the distance to the closest annotated TSS or CAGE tag was calculated. Internal exons that do not overlap initial or terminal exons from the UCSC annotation were retrieved and distances from the acceptor sites to the mapped CAGE tags were normalized to the length of the corresponding exon. CAGE tags not mapping to the genome were mapped to exon–exon junctions containing the last and the first 50 bp of consecutive exons. To determine the prevalence of genes with internal CAGE tags, the number of internal exons containing at least one CAGE was plotted against the total number of exons in each gene. Transfection experiments were performed similar to those used for siRNAs[20] using single-stranded RNA oligonucleotides. Messenger RNA levels were determined 48 h later using quantitative PCR with reverse transcription. The *MYC*-responsive luciferase reporter was transfected 24 h before the sRNAs. ChIP-chip data and DNase sensitivity profiles from HeLa and HepG2 cells were extracted from the UCSC ENCODE (Encyclopedia of DNA Elements) database. Promoters were grouped depending on the ChIP-chip signal intensity, and the number of PASRs in each group was determined.

Received 16 September 2008; accepted 2 January 2009.
Published online 25 January 2009.

1. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
2. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
3. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8**, 413–423 (2007).
4. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
5. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
6. Eddy, S. The UCSC Genome Browser ⟨http://www.genome.ucsc.edu/cgi-bin/hgTables?db=hg18&hgta_group=genes&hgta_track=rnaGene&hgta_table=rnaGene&hgta_doSchema=describe+table+schema⟩ (2006).

7. Yang, J. H. *et al.* snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res.* **34,** 5112–5123 (2006).
8. Huttenhofer, A. *et al.* RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.* **20,** 2943–2953 (2001).
9. Kawaji, H. *et al.* CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. *Nucleic Acids Res.* **34,** D632–D636 (2006).
10. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
11. Berget, S. M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270,** 2411–2414 (1995).
12. Olofsson, S. O. & Boren, J. Apolipoprotein B: a clinically important apolipoprotein which assembles atherogenic lipoproteins and promotes the development of atherosclerosis. *J. Intern. Med.* **258,** 395–410 (2005).
13. Rada-Iglesias, A. *et al.* Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. *Genome Res.* **18,** 380–392 (2008).
14. Bochnig, P., Reuter, R., Bringmann, P. & Luhrmann, R. A monoclonal antibody against 2,2,7-trimethylguanosine that reacts with intact, class U, small nuclear ribonucleoproteins as well as with 7-methylguanosine-capped RNAs. *Eur. J. Biochem.* **168,** 461–467 (1987).
15. Morris, K. V., Chan, S. W., Jacobsen, S. E. & Looney, D. J. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305,** 1289–1292 (2004).
16. Ting, A. H., Schuebel, K. E., Herman, J. G. & Baylin, S. B. Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nature Genet.* **37,** 906–910 (2005).
17. Janowski, B. A. *et al.* Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. *Nature Chem. Biol.* **3,** 166–173 (2007).
18. Mattick, J. S. RNA regulation: a new genetics? *Nature Rev. Genet.* **5,** 316–323 (2004).
19. Mattick, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25,** 930–939 (2003).
20. Willingham, A. T. *et al.* A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309,** 1570–1573 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Cardone, D. Rebolini, M. Kramer, and W. R. McCombie for Illumina sequencing. We wish to thank J. Brosius, J. Schmitz and T. Rozhdestvensky for their help with the small RNA cloning protocol and J. Dumais for technical assistance. K.F.-T. was in part supported by the Schering Foundation. This work was supported in part by grants from the NIH and was performed as part of the ENCODE consortium (G.J.H. and T.R.G.). G.J.H is an investigator of the Howard Hughes Medical Institute.

Author Contributions K.F.-T. and P.K. performed experiments in collaboration with E.D., V.S., R.D. and A.T.W. P.K., S.F., R.S. and G.A. performed data analysis. G.J.H. and T.R.G. planned experiments and wrote the paper.

Author Information Sequences generated during this study have been deposited in GEO under accession number GSE14362. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.J.H. (hannon@cshl.edu) or T.R.G. (gingeras@cshl.edu).

Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project

Cold Spring Harbor Laboratory Katalin Fejes-Toth[1,2]*, Vihra Sotirova[1,2], Ravi Sachidanandam[1]†, Gordon Assaf[1,2], Gregory J. Hannon[1,2]; **Affymetrix** Philipp Kapranov[3]*, Sylvain Foissac[3], Aarron T. Willingham[3], Radha Duttagupta[3], Erica Dumais[3] & Thomas R. Gingeras[1,3]

[1]Watson School of Biological Sciences, [2]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [3]Affymetrix, Inc. Santa Clara, California 95051, USA. †Present address: Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, 5 East 98th Street, New York, New York 10029, USA.

*These authors contributed equally to this work.