Vol 466|29 July 2010|doi:10.1038/nature09190

LETTERS

New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism

Philipp Kapranov[1]*, Fatih Ozsolak[1]*, Sang Woo Kim[2]*, Sylvain Foissac[3]*, Doron Lipson[1], Chris Hart[1], Steve Roels[1], Christelle Borel[4], Stylianos E. Antonarakis[4], A. Paula Monaghan[5], Bino John[2] & Patrice M. Milos[1]

Small (<200 nucleotide) RNA (sRNA) profiling of human cells using various technologies demonstrates unexpected complexity of sRNAs with hundreds of thousands of sRNA species present[1–4]. Genetic and *in vitro* studies show that these RNAs are not merely degradation products of longer transcripts but could indeed have a function[1,2,5]. Furthermore, profiling of RNAs, including the sRNAs, can reveal not only novel transcripts, but also make clear predictions about the existence and properties of novel biochemical pathways operating in a cell. For example, sRNA profiling in human cells indicated the existence of an unknown capping mechanism operating on cleaved RNA[2], a biochemical component of which was later identified[6]. Here we show that human cells contain a novel type of sRNA that has non-genomically encoded 5′ poly(U) tails. The presence of these RNAs at the termini of genes, specifically at the very 3′ ends of known mRNAs, strongly argues for the presence of a yet uncharacterized endogenous biochemical pathway in cells that can copy RNA. We show that this pathway can operate on multiple genes, with specific enrichment towards transcript-encoding components of the translational machinery. Finally, we show that genes are also flanked by sense, 3′ polyadenylated sRNAs that are likely to be capped.

To date, all sequencing reports characterizing sRNAs rely on conversion of RNA into cDNA using ligation and/or amplification steps[7]. These steps could introduce bias in detecting any class of RNA, whether known or novel, as well as quantifying those RNAs. To minimize these biases, we developed an approach that skips ligation and amplification, and generates data via true single-molecule sequencing[8] of first-strand cDNA. This approach tails 3′ ends of RNAs with a homopolymeric stretch of cytosine using poly(A) polymerase, followed by conversion of tailed RNA into first-strand cDNA using a poly(G) primer. The cDNA is then 3′ poly(A)-tailed, blocked using ddTTP and directly sequenced without amplification[8]. Although this approach can be used for small RNA profiling experiments involving as low as 10 nanogram of RNA, there are also times when the amount of starting material is extremely limited or when it is desirable to sub-select specific species of RNAs, so an amplification-based approach has also been developed that uses ligation of a poly(T) oligo to the 5′ ends of RNAs, followed by the steps described above and PCR amplification of the resultant first strand cDNA (Supplementary Information).

Both unamplified and amplified protocols were applied to sequence the sRNAs from HeLaS3 cells. Following true single molecule sequencing (tSMS), the reads were mapped against sequences of 586 annotated sRNAs, including 178 microRNAs (miR-1–miR-299; see Supplementary Information). In total, ~16.6 million unamplified and ~4.7 million amplified reads could be mapped to known sRNAs with these parameters (Supplementary Table 1). In the no-amplification analysis, ribosomal RNAs were most abundant followed by small nuclear RNAs and small nucleolar RNAs (Supplementary Table 1). miRNAs represented less than 0.5% of the total mapped reads, the most abundant HeLaS3 miRNA being miR-21. Overall, 122/178 miRNAs could be detected in HeLaS3 cells with at least one read and 104 detected with at least five reads. Transfer RNAs were not heavily represented in the sequenced RNAs (Supplementary Table 1), consistent with previous results and probably owing to strong secondary structure that may prevent efficient reverse-transcription[9]. Most of the reads corresponding to miRNAs aligned to the mature miRNAs, indicating that the structured pre-miRNAs were not detected well either (not shown). The no-amplification protocol was highly strand-specific, as exemplified by the mapping of 985,516 out of 997,801 reads (98.8%) to annotated sno- and miRNAs (from the sno/miRNA track on the UCSC browser) on the same strand as the annotation.

Amplification preserved the general trend of abundances of different annotated sRNAs. However, normalized abundances of individual sRNAs varied up to 1-2 orders of magnitude between the amplification and no-amplification protocols (Supplementary Table 1). Correlation between two technical replicates of the no-amplification protocol performed separately on the same RNA preparation was >0.98, indicating that the difference between the amplified and non-amplified protocol originated with the amplification or ligation steps.

To determine what additional classes of sRNA were represented among the sequenced sRNAs, the tSMS reads were mapped to the human genome using more stringent criteria (see Supplementary Information). Although ensuring higher quality matches, these parameters eliminate RNAs shorter than 25 bases, such as mature miRNAs, owing to increased requirement for the length of alignment. Reads mapping to unique locations were dominated by ribosomal RNAs, snoRNAs and reads mapping to repetitive elements included RNA repeats representing tRNA and snRNA sequences (Supplementary Table 2). In the no-amplification protocol, these classes of sequences accounted for ~90% of all reads. The distribution of reads among different classes of annotations was more variable among replicates of the amplification protocol (data not shown); however, these classes of sequences still accounted for the majority of reads (~90%). Approximately 1.7% of all reads constituted unannotated sRNAs (Supplementary Table 2), enriched for a recently identified class of promoter-associated short RNAs (PASRs)[1,2].

Thus, tSMS technology can detect sRNAs belonging to various well-characterized, as well as recently discovered classes of sRNAs. Amplification caused substantial perturbations in the abundances of many known sRNAs.

[1]Helicos BioSciences Corporation, 1 Kendall Sq. Ste B7301 Cambridge, Massachusetts 02139-1671, USA. [2]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15260, USA. [3]Integromics, S.L., Grisolía 2, 28760 Tres Cantos, Madrid, Spain. [4]Department of Genetic Medicine and Development, University of Geneva Medical School, University of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland. [5]Department of Neurobiology, University of Pittsburgh, 3501 Fifth Avenue, Pittsburgh, Pennsylvania 15260, USA.
*These authors contributed equally to this work.

Interestingly, we observed one class of sRNAs that included 26,815 reads: 0.4% of all unique reads or 24% of all novel reads in the no-amplification sample. These sRNAs were found within 50 bp and antisense to the 3′ untranslated regions (UTRs) of annotated transcripts. As an example, the *FAU* locus is shown in Fig. 1a. The antisense sRNAs had a peculiar feature: their 5′ ends were very close to the 3′ ends of mRNAs to which they were antisense (Fig. 1a, b), similar to the termini-associated short RNAs (TASRs) identified previously using tiling arrays[1]. Furthermore, these sRNAs were not identified using a different sequencing platform (Fig. 1a)[2] To estimate the fraction of antisense reads of all reads that mapped near the 3′ end, sRNA mapping in the region up to 500 bp from the annotated 3′ end of all genes were summed up. Of the novel reads 62,000/158,454 mapped near the 3′ end and these sRNAs were predominantly on the antisense strand (42,045/62,000), indicating presence of bona fide antisense RNAs given the high strand-specificity of this protocol as described above.

Because this novel class of sRNAs is specifically on the antisense strand, unlike the previously characterized termini-associated short RNAs (TASRs), we refer to them as antisense TASRs or aTASRs. Given that they start almost exactly (see Supplementary Information) at the end of the transcripts they are antisense to, we thought that these RNAs could potentially be produced by copying of polyadenylated mRNAs using an endogenous enzymatic activity similar to an RNA-dependent RNA polymerase (RdRP) starting from the poly(A) tail (Fig. 1c). In such a scenario, the aTASRs would have a stretch of U residues at their 5′ ends. Thus, they would not be detected with a strategy based on short reads where sequencing starts directly from the 5′ ends of sRNAs, because such reads would be discarded as appearing to contain only U residues. However, with the strategy we used, a naturally-occurring 5′ poly(U) stretch would be converted into a 3′ poly(A) stretch on the cDNA and this would allow the resulting cDNA molecules to bind to the HeliScope Sequencer flow cell surface[8,10] (Fig. 1c). The 3′ poly(A) stretch on the cDNA would then be copied before initiation of sequencing and thus the detected sequence would correspond to the first or second non-U base of an RNA molecule (Fig. 1c; also see Supplementary Information)[8].

To test this, we used two strategies (Fig. 2). First, we performed the amplification protocol with and without ligation of a poly(T) adaptor to the 5′ ends of 3′ poly(C)-tailed RNA molecules and amplified the resultant mixes with oligo(G) and oligo(A). The amplification reaction with a ligated poly(T) adaptor serves as a control in which all sRNAs should be amplified whether they originally contained a 5′ poly(U) or not. In the reaction with sRNAs containing no ligated product, only aTASRs and any other sRNA with a 5′ poly(U) should be amplified. For the sample with a ligated poly(T) adaptor, less than 0.1% of all reads (4,077/4,807,191) were found to be within 50 bp of gene ends. In contrast, 416,069 of the 905,492 mapped sequences (see Supplementary Information) from the unligated sample were found antisense and within 50 bp of gene ends (snoRNAs were removed from the analysis). Most of the sRNA mapping to the 3′ ends of genes (92.7% or 416,069/448,968) were antisense to the 3′ UTRs, thus representing aTASR molecules. In total, aTASRs were found at 5,573 distinct 3′ends of the 37,138 annotated transcripts we examined.

In the second verification approach, sRNAs were not amplified and also not poly(A)-tailed. sRNAs were poly(C)-tailed, reverse-transcribed into cDNA using oligo(G), and then sequenced directly without poly(A) tailing normally required to provide a sequence for binding to the poly(T) sequencing flow cell surface. This sample showed even higher levels of enrichment with 97,707 out of 151,930 or 64.3% of all uniquely-mapping reads corresponded to aTASRs. These data confirm that aTASRs must have poly(U)-like 5′ tails that are not genomically encoded and thus represent a class of RNAs not described previously.

To confirm that aTASRs are not artefacts of reverse transcription, we investigated the presence of aTASRs in the locus encoding eukaryotic translation elongation factor 2 EEF2, an essential factor for protein synthesis, using northern blots and RNase protection assays (RPA; see Supplementary Information). For northern blots based on a recently published protocol[11], we designed three 26-nucleotide long locked nucleic acid (LNA) probes (Fig. 3a) complementary to the EEF2 aTASR (Supplementary Information). Probe 1 (Fig. 3a) was designed to test the presence of 5′ poly(U) tail and the presence of EEF2 aTASR. Probe 2 contained an identical LNA spiking pattern as probe 1, but without significant complementarity to the 5′ poly(U) tail. Probe 2 therefore served as a control to demonstrate that the probe1 signal is due to hybridization to both the genomic sequences and the 5′ poly(U) tail. Probe 3 was designed as an additional validation test for EEF2 aTASR. Signals from probes 1 and 3 clearly showed the presence of EEF2 aTASR (Fig. 3b) in both HeLaS3 and MCF7 cell lines, whereas signals from probes 1 and 2 substantiated the presence of a 5′ poly(U) tail. We also detected a sense EEF2 TASR using another probe (probe 4), also validating our sequencing results (see later). Interestingly, the sizes of the sense and antisense TASRs in the EEF2 locus were similar

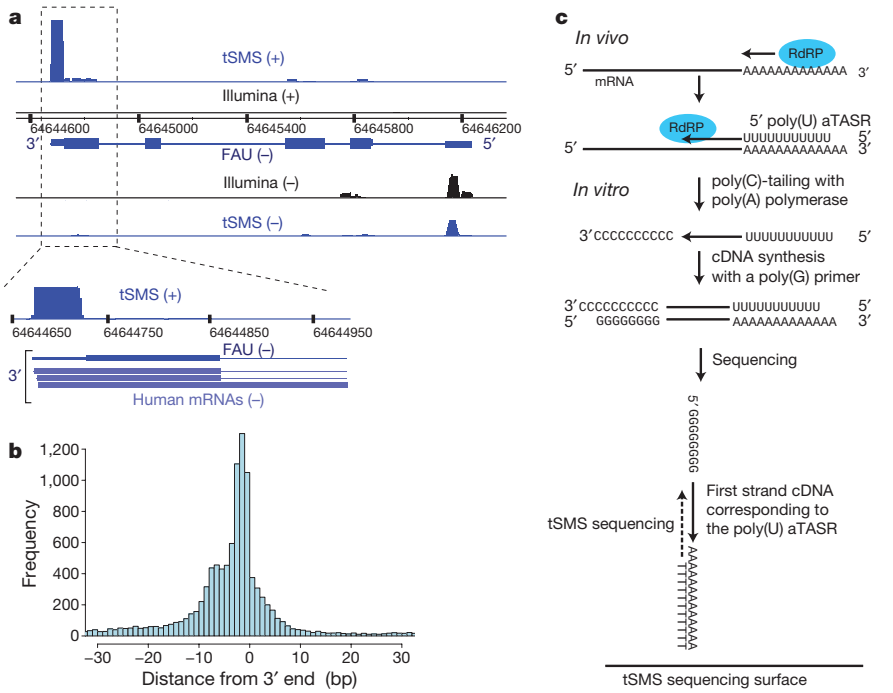

Figure 1 | Novel class of sRNAs antisense to 3′ ends annotations. a, Distribution of sRNA detected using tSMS using the no-amplification protocol and Illumina platforms[2] on both strands of the genome within the *FAU* locus. **b**, Distribution of distances between the 5′ ends of tSMS sequences obtained using the non-amplified protocol and 3′ ends of annotations of the opposite strand. Base to the right of '0' represents the most 3′ base of an RNA, see Supplementary Information for additional explanation and details. **c**, Schematic representation of a possible mode of generation of sRNAs antisense to 3′ UTRs and their detection using tSMS.

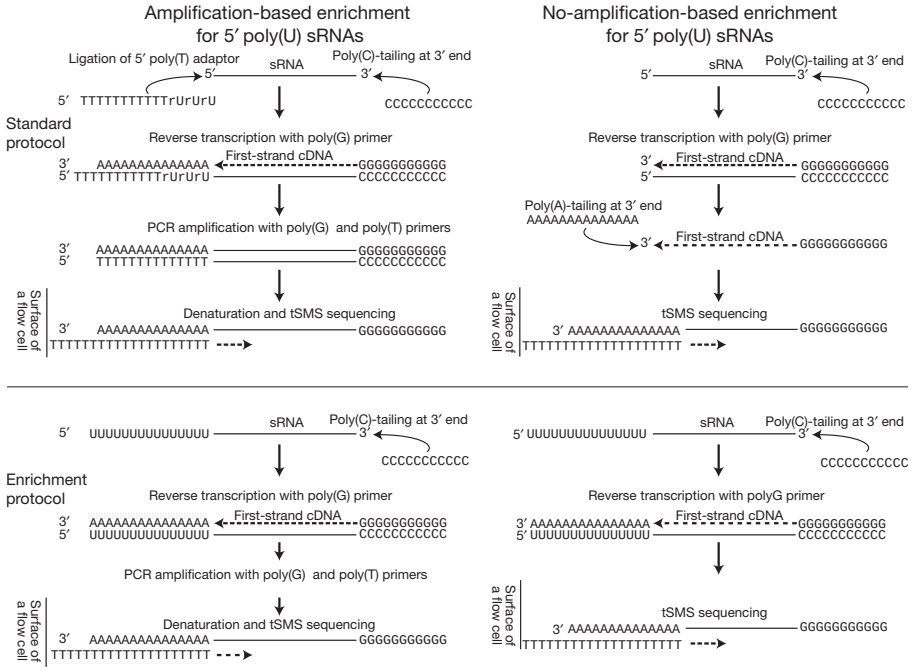

Figure 2 | Schematic representation of the two approaches (with and without amplification) used to enrich for 5′ poly(U) sRNAs (bottom) and comparison to regular tSMS protocols (top).

(Fig. 3b), again suggesting a possibility that one of these RNAs could be produced by copying from another. Having demonstrated the presence of an EEF2 aTASR, we used RPAs with a probe complementary to both the genomic region and the 5′ poly(U) tail to test further the presence of 5′ poly(U)-containing RNAs (Fig. 3c). Consistent detection of the protected RNA fragment of the expected size of 49 bases in three experiments (one of which is shown on Fig. 3c) confirmed the presence of 5′ poly(U) tail on the aTASR. We do see additional protected bands of smaller sizes that could represent alternative isoforms of the EEF2 aTASRs. Additional studies confirming these findings are described in the Supplementary Information ("aTASRs are not obvious artifacts of reverse transcription" and Supplementary Fig. 2).

To exclude the possibility that the presence of aTASRs is limited to cancerous cell lines like HeLaS3, we performed the no-amplification-based enrichment on a sRNA fraction from a normal liver tissue and obtained similar results as in HeLaS3 (Supplementary Figs 5–7). We have also tested whether aTASRs represent true RNA molecules or whether they are DNA copies of RNA. The sRNA fraction was treated with RNase before cDNA synthesis and this treatment virtually abolished the yield of sequences from the no-amplification-based based enrichment of 5′ poly(U)-containing RNAs from liver (data not shown). This strongly indicates that these species are in fact RNAs. Additional evidence for the presence of an endogenous RNA-copying mechanism comes from transcripts antisense to the spliced forms of the sense transcripts seen in the analysis of the cap-analysis of gene expression (CAGE) data[12] and human expressed sequence tags (ESTs; see Supplementary Information for details). Taken together with our sequencing results, these experiments confirm the presence of endogenous human aTASRs and sense TASRs and the presence of a 5′ poly(U) tail on the latter at least for the EEF2 locus.

A significant fraction of reads (186,891/905,492 or 21% of the mapped reads, see Supplementary Information) in the HeLaS3 library enriched by amplification for 5′ poly(U) sRNAs were ±500 bp from a transcriptional start site (TSS), thus representing a previously identified PASRs[1]. The aTASR reads were excluded from this analysis to avoid double-counting of sRNAs antisense to 3′ UTRs in the vicinity of the 5′ ends of neighbouring or overlapping genes. Unlike the aTASRs, most of the PASR sequences from this amplified library were sense to the genes: 148,456/186,891 or 79.4% of all such reads. Such enrichment of PASRs was not detected in the no-amplification based enrichment approach (data not shown). A possible explanation for this discrepancy is that the 5′ poly(U) tails on the PASRs may be shorter than those on the aTASRs: a minimal length of ~50 residues is typically required for efficient binding to the HeliScope Sequencer surface. A stretch of less than 50 5′ U residues could be converted into the desirable length during the amplification-enrichment step with a fixed length oligo of 50 T residues, but not during the no-amplification-based enrichment approach. The presence of PASRs with 5′ poly(U) stretches was also fairly common with 2,612 different promoter regions having them. We next asked whether genes had a propensity to have both aTASRs and sense PASRs with the 5′ poly(U) stretches. To do so, we first collapsed the UCSC Known Gene annotations that had the same 5′ and 3′ ends. Of these 51,512 loci, 4,982 loci had 5′ poly(U) aTASRs, 1,856 had 5′ poly(U) sense PASRs and 564 had both. The latter is significantly higher than expected by chance (*P*-value of $<2.2\ 10^{-16}$, Pearson chi-square test). Interestingly, we have found that translation-related functions seem to be enriched among the genes that are flanked by 5′ poly(U) sRNAs, and this relationship may not be just due to their level of expression (see Supplementary Information).

These results strongly indicated that the 5′ poly(U) RNAs (aTASRs and PASRs) could be copied from 3′ poly(A) RNAs and prompted us to investigate the profile of the 3′-polyadenylated sRNA transcriptome. The sRNA fraction from the HeLaS3 cells was reverse-transcribed in the presence of a poly(dU)$_{25}$ oligonucleotide, the resulting cDNA was poly(A)-tailed and sequenced (see Supplementary Information). For this analysis, the definition of TASR regions was extended to be within 500 bp of a 3′ end. The resulting reads were clearly enriched in the sRNAs around the 5′ and 3′ ends of genes. Reads corresponding exclusively to PASRs or TASRs were represented respectively by 500,315 (20.5%) and 578,385 (23.7%) of the 2,436,066 non-ribosomal, non-mitochondrial reads; or 7% and 8.1% respectively from the 7,131,693 total mapped reads in this library. For comparison, the PASR and TASR regions used in the analysis each represented ~1% of the total genome and thus small RNAs represented ~20-fold enrichment of what would be expected randomly considering the reads that do not map to ribosomes or mitochondria. Unlike the 5′ poly(U) sequences at the 3′ ends of genes, the 3′ poly(A) sequences were clearly enriched to be sense to the genes, with 97% (484,831/500,315) and 92.6% (535,830/578,385) of poly(A)+ PASRs and TASRs being in the sense orientation with respect to their genes.

Comparison of the distribution of the 3′ poly(A) and 5′ poly(U) sRNA data relative to each other and to the locations of the 5′ and 3′ ends of genes revealed some interesting features of these sRNAs (Supplementary Fig. 3). First, the distance between the 5′ ends of the 3′ poly(A) sRNA and the 3′ ends of genes peaked at ~50–70 bp upstream of the poly(A) site (Supplementary Fig. 3). This indicates a

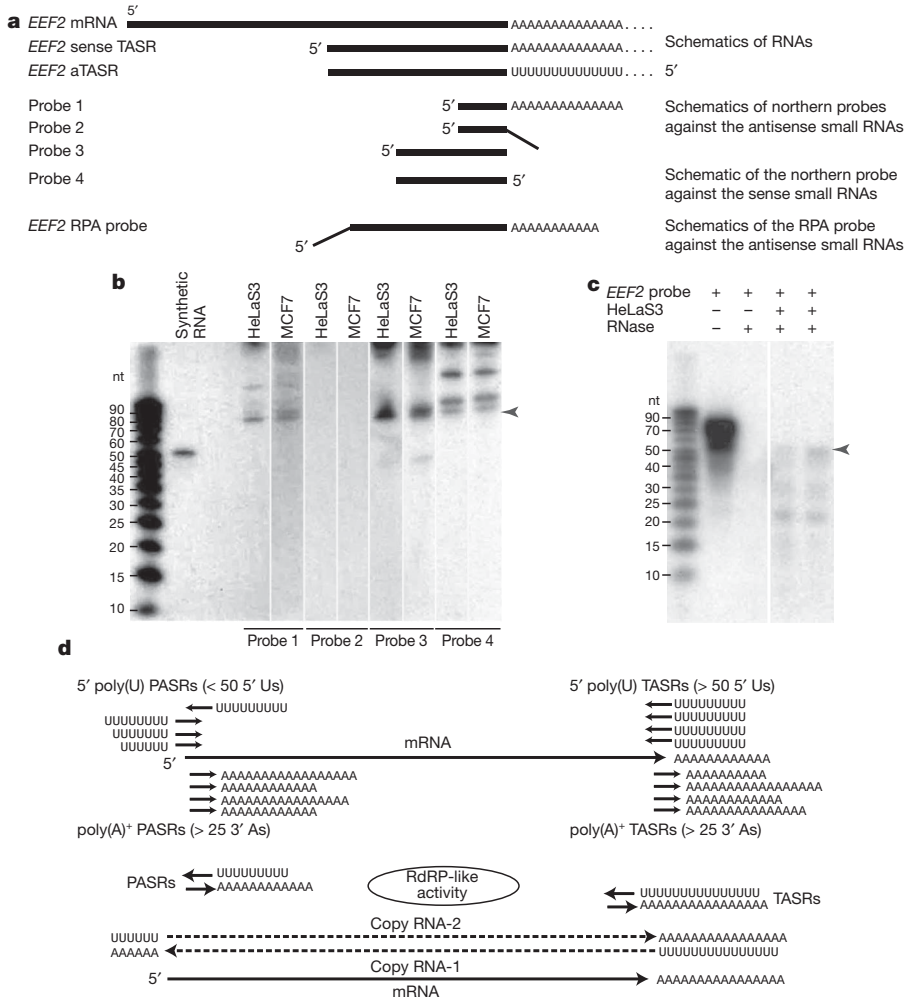

Figure 3 | Verification of EEF2 TASRs using independent biochemical methods and possible mechanism for their production using endogenous RdRP-like activity. **a**, Probes used in northern blot and RPA assays. Regions of no complementarity in the probes are shown as thin oblique lines. **b**, Northern blot results. The arrowhead marks the strongest consistent band. **c**, RPA assays results. The arrowhead marks the expected protected fragment. **d**, Schematic representation of 5′ poly(U) and 3′ poly(A) sRNAs flanking a gene locus. The existence of RNAs with both a 5′ poly(U) and a 3′ poly(A) sequence has not been demonstrated by these experiments and is thus hypothetical.

non-random mechanism that generates these sRNAs, presumably via cleavage of longer, poly(A)+ RNAs similar to what has been reported[2]. Interestingly, sRNAs detected by the Illumina technology as well as tags representing 5′ ends of capped RNAs (CAGE tags[12]) also peaked at this position (data not shown and Supplementary Fig. 3). This latter observation shows that the 5′ ends of the 3′ poly(A) small RNAs could be generated by cleavage of long RNAs 50–70 bases upstream of the poly(A) addition site followed by re-capping[2,6]. Second, the distribution of the 5′ ends of the 3′ poly(A) and 5′ poly(U) PASRs with respect to annotated TSSs was different, with the former peaking at the TSS, and the latter peaking at +10/+20 relative to the TSS (Supplementary Fig. 3). In this respect, the 3′ poly(A) sRNAs near the TSSs behave similarly to PASRs[1], whereas the poly(U) sRNAs at the location behave more similarly to the tiny promoter-associated RNAs (tiRNAs) reported elsewhere[3]. Together with the presence of the 5′ poly(U) tail, this data further indicates that, like the tiRNAs, the 5′ poly(U) PASRs are unlikely to represent mere truncated products of the 5′ ends of long RNAs that would be expected to peak at the TSS.

As reported previously for tiRNAs[3], PASRs and TASRs[1], the quantity of 3′ poly(A) PASRs and TASRs as well as 5′ poly(U) aTASRs correlated with the expression levels of long RNAs of the corresponding transcripts. Pearson correlation coefficients between the levels of long RNAs and the levels of sense, anti-sense 3′ poly(A) and 5′ poly(U) sRNAs were found to range from 0.51 with 5′ poly(U) aTASRs to 0.58 with 3′ poly(A) PASRs and TASRs, indicating that although correlated with gene expression, other factors influence the abundance of these sRNAs as well. Moreover, association with 5′ poly(U) aTASRs as well as with 3′ poly(A) PASRs and TASRs seems to represent a common property of human transcripts. Interestingly, the 5′ poly(U) PASRs had a much weaker correlation with gene expression (0.15).

Thus, to summarize these observations, many gene loci are flanked by sRNAs that have either 5′ poly(U) stretches and/or 3′ poly(A) stretches (Fig. 3d). The 5′ poly(U) sRNAs at the 3′ ends of the genes (aTASRs) are almost exclusively antisense to the gene locus, whereas their counterparts at the 5′ ends (PASRs) tend to be on the same strand as the gene locus itself and seem to have shorter 5′ poly(U) tails. The 3′ poly(A) PASRs and TASRs tend to be on the same strand as the gene itself. The sRNAs with the 5′ poly(U)s, particularly the aTASRs, are proposed to be the products of RdRP-mediated copying from polyadenylated RNA molecules. Although aTASR could potentially be copied from the polyadenylated sRNAs at the 3′ ends, the 5′ poly(U) PASRs tend to have the same strand as the 3′ poly(A) PASRs (see above), indicating that most are unlikely to be made by copying from 3′ poly(A) sRNAs. This observation also goes against a general 'loop-back' artefact of reverse transcription (as shown in Supplementary Fig. 2) as a mechanism of generation of the 5′ poly(U) molecules. The artefact would be expected to generate palindromic cDNAs from all poly(A)+ molecules and thus the majority of the 5′ poly(U) PASRs would be antisense to the 3′ poly(A) PASRs, which is not the case. Thus, 5′ poly(U) PASRs are likely to be produced from polyadenylated antisense long RNAs (Fig. 3d). This suggests a model where an existing poly(A) RNA molecule is amplified via RdRP-mediated copying which starts at the 3′ poly(A) tail resulting in the generation of a complementary RNA molecule which is then polyadenylated at its 3′ end. This could potentially involve an RdRP-mediated mechanism similar to that by which the influenza virus adds poly(A) tails to its gene products[13]. This would then be followed by another round of copying to generate a cRNA molecule that is sense to the gene (Fig. 3d). In this model the 5′ poly(U) TASRs and PASRs either function as primers for this synthesis or represent short, potentially-incomplete products of this copying (Fig. 3d). Similarities

between the distributions of the 5′ poly(U) PASRs and the tiRNAs[3] may indicate that both classes of RNAs have similar function.

The existence of the endogenous RNA copying mechanism similar to the one described in Fig. 3d could potentially result in amplification of RNA molecules in a cell, similar to a model proposed previously[14] based on a discovery of 5′ poly(U) spliced mRNAs collinear and antisense to the globin mRNA. Here we show that this phenomenon may be widespread, potentially affecting thousands of loci. The prevalence of these RNAs and mechanisms will require further investigation.

Our data and observations do not completely rule out mechanisms which are distinct from that of an RdRP activity that may give rise to these novel 5′ poly(U)-like stretch-containing sRNAs. For instance, a novel mechanism may add nucleotides to the 5′ ends of RNAs in a step-wise manner. However, it would need to somehow favour only two classes of RNA located at the termini of annotated genes. This combined with the fact that 5′ poly(U) aTASRs tend to be located at the 3′ ends of genes just before the poly(A) tails, strongly argues for an RNA copying mechanism as a mode of their generation.

Recently, an RdRP activity was isolated from human cells as the ribonucleoprotein complex of telomerase reverse transcriptase catalytic subunit (TERT) and the RNA component of mitochondrial RNA processing endoribonuclease (RMRP)[15]. Unlike the previously identified RdRP, the TERT-RMRP enzyme has a specific requirement for an RNA template to be able to fold back on itself and thus provide a primer for RNA synthesis[15]. The presence of RdRP activity in mammalian cells was also suggested on the basis of the requirement of a host RNA-copying activity for replication of the hepatitis delta virus[16]. However, the protein(s) responsible for this activity has not been isolated yet[16], and it is thus possible that other RdRPs besides TERT-RMRP exist in human cells. For example, RNA polymerase II was shown to have a limited ability to copy RNA *in vitro*[17]. The TERT-RMRP RdRP was shown to have a restricted set of targets in human cells, presumably owing to the requirement for a template to form a secondary structure that could generate a 3′ loop-back[15]. Since thousands of loci were found to have aTASRs in this work, it indicates that additional RdRPs with broader target spectrums do exist in human cells. It also predicts the properties of the RdRP-like activity postulated to generate aTASRs: since the RNA copies start within the 3′ poly(A) tails of sense polyadenylated RNAs, this activity is likely to initiate copying of RNA opposite to its 3′ end similarly to some other RdRPs[18]. It remains to be seen whether the 5′ poly(U) sRNAs are themselves just copies of 3′ poly(A) sRNAs or whether they also serve as primers for copying of long poly(A)+ RNAs. Although the exact mechanism of generation of 5′ poly(U) RNAs and their function remain open, we have provided a starting point and the technology needed to address whether this novel class of sRNAs has a critical role in important biological processes. The ability of SMS to detect novel RNAs without amplification-induced biases will provide biologists with a powerful tool for better understanding gene expression.

METHODS SUMMARY

Small RNA (<200 nucleotides, sRNA) fractions from HeLaS3 total RNA was isolated using a mirVana miRNA isolation kit, 3′-tailed with poly(C) using poly(A) polymerase and converted into first-strand cDNA using a poly(G)-containing oligonucleotide. The cDNA was then 3′-tailed with poly(A) using terminal transferase, blocked at the 3′ end with biotin-ddATP and sequenced using Helicos single-molecules sequencing technology.

To enrich for 5′ poly(U)-containing sRNAs two methods were used, with and without amplification. In the former, the sRNAs from HeLaS3 and normal human liver were C-tailed and converted into cDNA using the protocol described above. The cDNA was blocked and hybridized directly to sequencing flow cells coated with oligo(dT) primers without A-tailing. In the latter, the sRNA fraction from HeLaS3 and liver was C-tailed, converted into cDNA and amplified using primers containing poly(T) and poly(G) without A-tailing of cDNA, blocked with ddTTP and sequenced.

To identify sRNAs with 3′ poly(A) tails, HelaS3 sRNA was converted into first-strand cDNA with oligo(dT/U) primer (5′-TTTTUTTUTUTTTUTTTTUTTTUTTV-3′), treated with RNases H and If and the USER enzyme to digest away the RNAs and 5′ T/dU stretches of the cDNAs, 3′ A-tailed, blocked with biotin-ddATP and sequenced.

Mapping of the reads was done either to known sRNAs or to the entire genome with the in-house indexDPgenomic program[19] freely available on the Helicos website (http://open.helicosbio.com/mwiki/index.php/Main_Page).

Northern blots were based on the LED protocol[11]. To reduce background noise, digoxigenin-labelling of LED protocol was substituted by ^{32}P-labelling of total RNA (10 μg) from HeLaS3 and MCF7 cells. Ribonuclease protection assays (RPAs) were performed using the mirVana miRNA detection kit by hybridizing 5×10^4 c.p.m. of the RNA probe with 5 μg of HeLaS3 total RNA at 37 °C. Unprotected RNA was digested using an RNase mixture containing diluted (1:100) RNase A+T1 and RNase ONE. The radioactive signals were detected using PhosphorImager.

Received 2 July 2009; accepted 20 May 2010.

1. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
2. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457,** 1028–1032 (2009).
3. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41,** 572–578 (2009).
4. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322,** 1849–1851 (2008).
5. Rassoulzadegan, M. *et al.* RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* **441,** 469–474 (2006).
6. Otsuka, Y., Kedersha, N. L. & Schoenberg, D. R. Identification of a cytoplasmic complex that adds a cap onto 5′-monophosphate RNA. *Mol. Cell. Biol.* **29,** 2155–2167 (2009).
7. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).
8. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–109 (2008).
9. Aspegren, A., Hinas, A., Larsson, P., Larsson, A. & Soderbom, F. Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development. *Nucleic Acids Res.* **32,** 4646–4656 (2004).
10. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnol.* **27,** 652–658 (2009).
11. Kim, S. W. *et al.* A sensitive non-radioactive northern blot method to detect small RNAs. *Nucleic Acids Res.* **38,** e98 (2010).
12. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
13. Fodor, E., Mikulasova, A., Mingay, L. J., Poon, L. L. & Brownlee, G. G. Messenger RNAs that are not synthesized by RNA polymerase II can be 3′ end cleaved and polyadenylated. *EMBO Rep.* **1,** 513–518 (2000).
14. Volloch, V., Schweitzer, B. & Rits, S. Antisense globin RNA in mouse erythroid tissues: structure, origin, and possible function. *Proc. Natl Acad. Sci. USA* **93,** 2476–2481 (1996).
15. Maida, Y. *et al.* An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature* **461,** 230–235 (2009).
16. Lai, M. M. RNA replication without RNA-dependent RNA polymerase: surprises from hepatitis delta virus. *J. Virol.* **79,** 7951–7958 (2005).
17. Lehmann, E., Brueckner, F. & Cramer, P. Molecular basis of RNA-dependent RNA polymerase II activity. *Nature* **450,** 445–449 (2007).
18. Ahlquist, P. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science* **296,** 1270–1273 (2002).
19. Giladi, E. *et al.* Error tolerant indexing and alignment of short reads with covering template families. *J. Comput. Biol.* (in the press).

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

Acknowledgements We wish to thank A. Willingham, J. Thompson and Z. Li for discussions and help in the preparation of the manuscript. S.E.A. is supported by the Swiss National Science Foundation, B.J. by the NIH (GM079756) and American Cancer Society (RSG0905401), and A.P.M. by the NIH (MH60774). S.F. acknowledges support by a grant from the Comunidad de Madrid and European Union (Exp. 11/2009).

Author Contributions P.K., F.O. and P.M.M. designed the study, performed the experiments and the data analysis. S.F. and D.L. performed additional bioinformatics analyses. C.H. and S.R. assisted with the development of the Helicos analysis pipeline. S.W.K., A.P.M. and B.J. performed the northern blot and RPA validation experiments. C.B. and S.E.A. contributed to the validation experiments.

Author Information Sequencing datasets described in this study have been deposited at the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA), accession no SRA012676. Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of this article at www.nature.com/nature. 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 P.K. (philippk08@gmail.com) and P.M.M. (pmilos@helicosbio.com).