CapSeq and CIP-TAP Identify Pol II Start Sites and Reveal Capped Small RNAs as *C. elegans* piRNA Precursors

Weifeng Gu,[1] Heng-Chi Lee,[1] Daniel Chaves,[1] Elaine M. Youngman,[1] Gregory J. Pazour,[1] Darryl Conte, Jr.,[1] and Craig C. Mello[1,2,*]
[1]Program in Molecular Medicine
[2]Howard Hughes Medical Institute
University of Massachusetts Medical School, 373 Plantation Street, Worcester, MA 01605, USA
*Correspondence: craig.mello@umassmed.edu

http://dx.doi.org/10.1016/j.cell.2012.11.023

SUMMARY

Piwi-interacting (pi) RNAs are germline-expressed small RNAs linked to epigenetic programming. *C. elegans* piRNAs are thought to be transcribed as independent gene-like loci. To test this idea and to identify potential transcription start (TS) sites for piRNA precursors, we developed CapSeq, an efficient enzymatic method for 5′ anchored RNA profiling. Using CapSeq, we identify candidate TS sites, defined by 70–90 nt sequence tags, for >50% of annotated Pol II loci. Surprisingly, however, these CapSeq tags failed to identify the overwhelming majority of piRNA loci. Instead, we show that the likely piRNA precursors are ∼26 nt capped small (cs) RNAs that initiate precisely 2 nt upstream of mature piRNAs and that piRNA processing or stability requires a U at the csRNA +3 position. Finally, we identify a heretofore unrecognized class of piRNAs processed from csRNAs that are expressed at promoters genome wide, nearly doubling the number of piRNAs available for genome surveillance.

INTRODUCTION

Argonaute (AGO) proteins associate with small RNAs to form sequence-directed gene-regulatory complexes that are deeply conserved in eukaryotes (Hutvagner and Simard, 2008). Most organisms encode multiple functionally distinct AGO family members. These AGOs are loaded with a diversity of small RNA cofactors, produced through a similarly diverse repertoire of small-RNA biogenesis mechanisms (Siomi and Siomi, 2009). AGO-associated small RNAs include micro (mi) RNAs and short-interfering (si) RNAs that are processed from double-stranded (ds) RNA precursors by the RNase-III-related enzyme Dicer (Bernstein et al., 2001). In some organisms, AGO-associated small-RNA species are produced, independent of Dicer, by RNA-dependent RNA polymerase (RdRP) (Gu et al., 2009; Pak and Fire, 2007; Sijen et al., 2007).

piRNAs are Dicer-independent small RNAs that interact with AGOs related to *Drosophila* Piwi (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006; Lau et al., 2006; Ruby et al., 2006). Many piRNA species originate from large genomic clusters, and they direct Piwi-dependent transposon silencing, heterochromatin modification, and germ cell maintenance (Aravin et al., 2007; Batista et al., 2008; Brennecke et al., 2007; Das et al., 2008; Lin, 2007). In flies and mammals, transposon-directed piRNAs typically map to both strands and are produced by a "ping-pong" amplification cycle, whereby sense piRNAs direct Piwi-dependent cleavage of a primary transcript to generate the 5′ ends of antisense piRNAs, and vice versa (Aravin et al., 2007; Brennecke et al., 2007; Gunawardane et al., 2007; Houwing et al., 2007). Recent work suggests that Piwi-bound precursor piRNAs are trimmed by a 3′-to-5′ exonuclease and then methylated on the 2′-OH of the 3′ residue of the mature piRNA (Kawaoka et al., 2011). In mice an abundant class of piRNAs, pachytene piRNAs, originates from large genomic clusters (Aravin et al., 2006). These piRNAs are not generated by the ping-pong cycle (Aravin et al., 2006; Beyret et al., 2012) but instead appear to be processed directly from a single-strand precursor by an unknown mechanism.

The *C. elegans* piRNAs, known as 21U-RNAs, are an abundant class of germline-expressed small RNAs that interact with the Piwi ortholog PRG-1 (Batista et al., 2008; Das et al., 2008; Ruby et al., 2006). Similar to mammalian pachytene piRNAs, 21U-RNAs are diverse in sequence, and the overwhelming majority lack perfectly complementary RNA targets. Unlike mammalian piRNAs, however, 21U-RNAs do not appear to be processed from long RNA precursors. Instead, they derive from individual gene-like loci that are dispersed within two large clusters on chromosome IV (Cecere et al., 2012; Ruby et al., 2006). Within these clusters, more than 15,000 distinct 21U-RNAs are expressed from both strands and reside within introns and intergenic regions, but are rarely found in coding regions (Batista et al., 2008; Ruby et al., 2006). The presence of a conserved 8 nucleotide (nt) motif and A/T-rich region upstream of each 21U-RNA led Ruby et al. (Ruby et al., 2006) to suggest that 21U-RNAs are independently expressed loci. Consistent with this idea, a recent study identified Forkhead-family transcription factors that associate with the 8 nt motif and whose activity was correlated with 21U-RNA expression (Cecere et al., 2012). Using 5′ RACE, this study also amplified 70 nt of a longer transcript that initiated 2 nt upstream of one of two overlapping 21U-RNAs (21ur-3372 and 21ur-14,222), suggesting that these 21U-RNAs may be processed from a transcript greater than 70 nt in length.

In this paper we explore the biogenesis of 21U-RNAs genome wide. To do this, we use two complementary 5′ anchored RNA deep-sequencing approaches called CIP-TAP and CapSeq. Unlike published cap analysis of gene expression (CAGE)-related methods involving affinity purification (de Hoon and Hayashizaki, 2008), the CapSeq protocol, developed in this study, utilizes an efficient enzymatic approach to dramatically reduce the background of structural RNA reads and to enrich for 70–90 nt sequence tags corresponding to the capped 5′ ends of longer RNAs transcribed by RNA polymerase II (Pol II). We show that CapSeq identifies pre-mRNAs, *trans*-spliced mRNAs, primary (pri-) miRNAs, and noncoding RNAs, thus defining candidate transcription start (TS) sites for >50% of annotated genes. This information is absent for most current WormBase annotations. Surprisingly, however, we show that CapSeq reads derive from less than 0.5% of annotated 21U-RNA loci.

Instead, using CIP-TAP cloning (Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009), we show that piRNA loci express an endogenous capped small (cs) RNA species. csRNAs are native 18–40 nt Pol II products thought to arise through early termination or polymerase pausing, and previous studies have shown that csRNAs are expressed, often bidirectionally, at promoter regions in a variety of organisms (Haussecker et al., 2008; Nechaev et al., 2010; Seila et al., 2009). However, csRNAs have not been described to date in *C. elegans*. Our CIP-TAP data, which are far from saturation, detect csRNAs at Pol II promoters genome wide, including >50% of annotated 21U-RNA loci.

Interestingly, we show that csRNA transcription initiates precisely 2 nt upstream of the corresponding mature 21U-RNA, suggesting that csRNAs are processed into piRNAs by removing the cap plus two nucleotides and by trimming the 3′ end. Furthermore, we provide evidence that a U residue at the +3 position of a csRNA is critical for piRNA processing or stability. In summary, we present a fully enzymatic approach, CapSeq, for 5′ anchored RNA profiling, and use this approach to define the first TS site annotations for many Pol II loci in *C. elegans*. We show that csRNAs identified by CIP-TAP cloning, rather than longer RNAs identified by CapSeq, are the likely piRNA precursors. Our findings also identify csRNAs transcribed at promoters genome wide as 21U-RNA precursors, nearly doubling the repertoire of piRNAs available for genome surveillance.

RESULTS

CapSeq Is an Efficient Method for 5′ Anchored Profiling of Pol II Transcripts

21U-RNAs are thought to be expressed from thousands of independent loci (Ruby et al., 2006), and could thus comprise ~40% of the transcription units in the *C. elegans* genome. However, systematic RNA-seq methods to profile RNA expression (Allen et al., 2011; Lamm et al., 2011) have not identified likely 21U-RNA precursors. Therefore, to increase the chance of identifying 21U-RNA precursors, we sought to generate a comprehensive profile of 5′ anchored RNA sequence tags. To do this we developed CapSeq, which employs a straightforward series of enzymatic treatments to enrich the 5′ ends of Pol II transcripts (Figure 1), with as little as 0.5–2 μg of total RNA as starting material (see Extended Experimental Procedures).

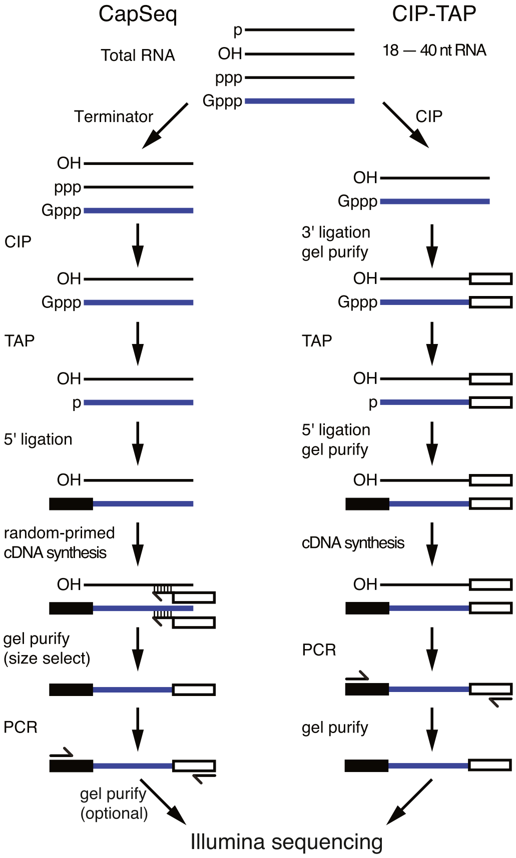


Figure 1. Flowcharts Illustrating the CapSeq and CIP-TAP Protocols

Using CapSeq, we generated five libraries from three different developmental stages (L1, L3, and adult) and obtained ~61 million reads that mapped to the *C. elegans* genome, including 46 million that mapped to nonstructural RNAs. Visual inspection using the genome browser software “GBrowse” (Stein et al., 2002) revealed that most CapSeq reads are indeed enriched at the 5′ ends of genes transcribed by RNA Pol II (Figure 2A, CapSeq). To estimate the fidelity with which CapSeq defines the actual 5′ ends of transcripts, as opposed to internally truncated RNAs, we took advantage of the fact that many worm transcripts contain a capped *trans*-spliced 5′ leader sequence. Using two different methods, we found that CapSeq enriches approximately 15,000-fold for reads starting at the actual 5′ end (indicated by the 5′-most nt of the spliced leader sequence) over reads starting at any other nucleotide downstream (see Extended Discussion available online).

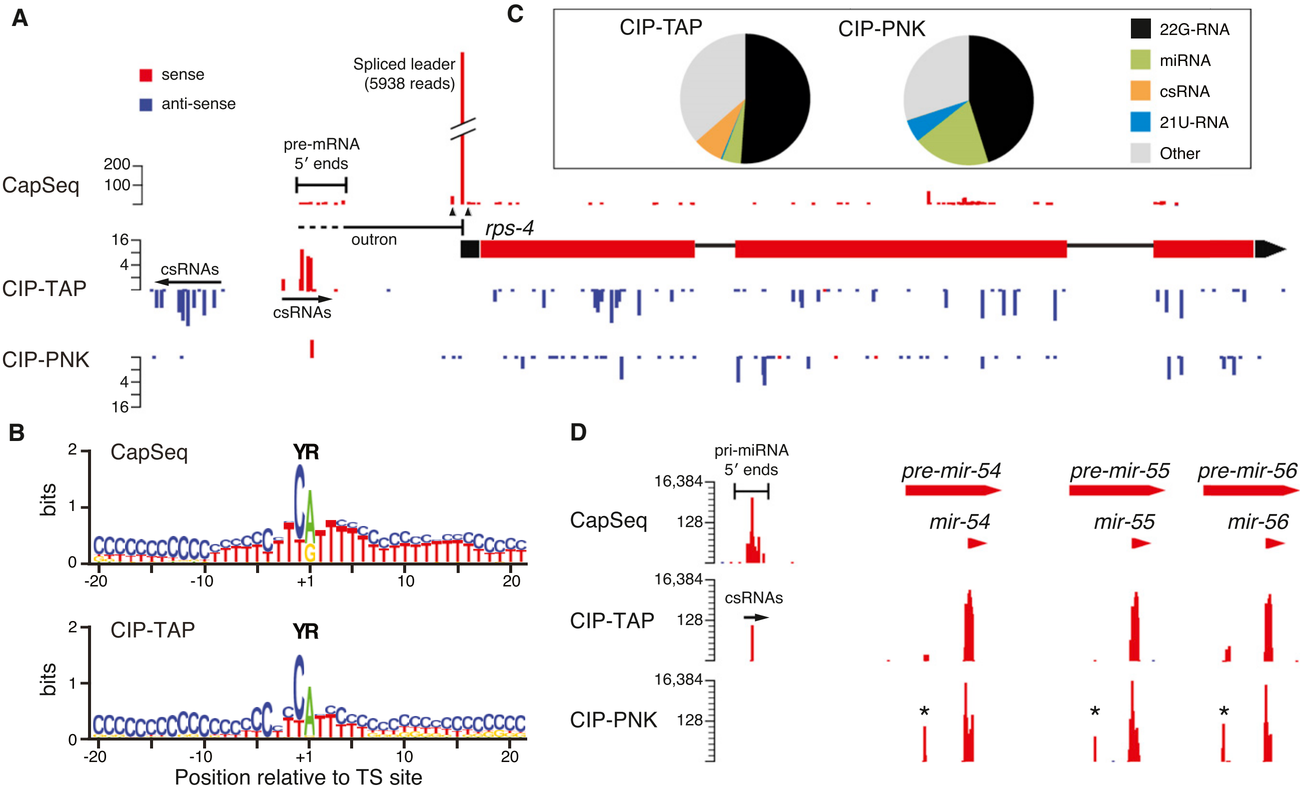


Figure 2. Comparative Analysis of RNA-Seq Protocols that Enrich Long- and Short-Capped RNAs and a Protocol that Clones Uncapped Short RNAs such as siRNA, piRNA, and miRNA Species

(A) Histograms representing the 5′ ends of mapped reads, as indicated, at a typical protein-coding locus, *rps-4*. The height of each histogram bar is proportional to the number of reads sharing the same 5′ nt and the scale (log2) is shown. Candidate pre-mRNA 5′ ends and csRNAs are indicated. *Trans*-splicing at some genes including *rps-4* results in removal of the 5′ UTR of the pre-mRNA, called an "outron," and the addition of a "Spliced leader." The major *trans*-splice site for *rps-4* is off the scale, as indicated by a break in the bar, and the total number of SL-containing reads is indicated. Two minor *trans*-splice sites flank the major *trans*-splice site as indicated by triangles below the CapSeq reads. The outron is indicated by a line below the CapSeq reads; dashes indicate the variable 5′ end of the outron. The blue bars beneath the *rps-4* coding sequences in the CIP-TAP and CIP-PNK samples correspond to antisense 22G-RNAs.

(B) Schematic representation of the nucleotide composition around candidate TS sites (the +1 position) identified by CapSeq and CIP-TAP reads (here only sense csRNAs). The nucleotide height (in bits) represents the log2 ratio of the frequency observed relative to the expected frequency based on genomic nucleotide composition. The enriched YR motif is indicated.

(C) Pie charts indicating the relative composition of small RNAs recovered in the CIP-TAP and CIP-PNK samples.

(D) Histograms representing the 5′ ends of mapped reads at the *mir-54–56* miRNA cluster. Candidate pri-miRNA 5′ ends and csRNAs are indicated. The asterisks indicate reads corresponding to miRNA star strands.

See also Figure S1 and Table S1.

Identification of RNA Pol II TS Sites and *trans*-Splice Sites

The TS sites for *C. elegans* genes are poorly mapped, in part because many mRNAs are *trans*-spliced to a capped ~22 nt spliced leader (SL), which results in removal of the pre-mRNA 5′ end (Blumenthal and Steward, 1997). Our CapSeq data identified 70% of the 15,759 SL *trans*-splice sites annotated in the WS215 genome release, as well as 5,711 new *trans*-splice sites (Figure S1A, top; Table S1A). Most gene annotations in WormBase simply indicate the 5′ end of the *trans*-spliced exon, or the position of the AUG codon. The 5′ ends of many non-SL CapSeq reads mapped near, and often upstream of, current 5′ end annotations (Figure 2A; data not shown). We hypothesized that the 5′ ends of these reads could represent TS sites. Consistent with this idea, we noticed that these reads exhibit a strong bias for a 2 nt motif of pyrimidine (Y) purine (R), or YR, in which R represents the first nt (+1) in the CapSeq read and Y represents the adjacent 5′ nt (Figure 2B). The YR motif is part of an extended consensus yY$\underline{R}$yyy (lower case indicates weaker preference), with a strong preference for A as the R at position +1 and a slight preference for T at flanking positions. The YR motif and extended consensus resemble the initiator element required for RNA Pol II transcription initiation in mammals, plants, and flies (de Hoon and Hayashizaki, 2008; Juven-Gershon et al., 2008; Smale and Baltimore, 1989). Using a cutoff of one CapSeq read per 10 million total reads, and a requirement for a YR motif, our CapSeq data predicted approximately 64,000 candidate TS sites genome wide (Table S1B).

In order to pair candidate TS sites with existing annotations, we considered CapSeq reads within a window from 1,000 nt

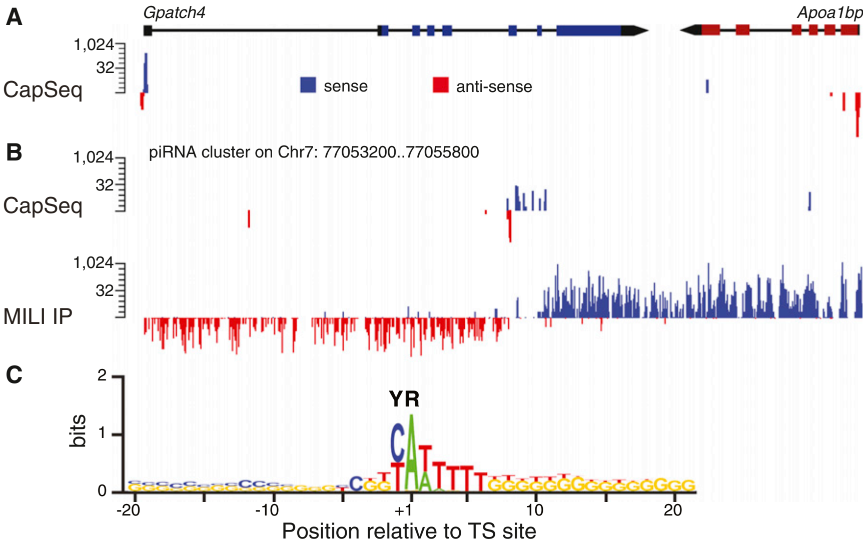


Figure 3. CapSeq Analysis of Mouse Testes RNA

(A and B) Browser views of representative protein-coding and piRNA cluster regions are shown. The histograms (log2 scale) represent the frequency of reads sharing the same 5′ ends from CapSeq or MILI IP (Robine et al., 2009), as indicated. Bidirectional reads were observed around the TS sites of *Gpatch4*.

(C) Schematic representation of the nucleotide composition around candidate TS sites (YR). The nucleotide height (in bits) represents the log2 ratio of the frequency observed relative to the expected frequency based on genomic nt composition.

See also Table S2.

upstream to 100–200 nt downstream of annotated 5′ ends. The 1,000 nt upstream distance was chosen as a conservative upper limit to allow for the possibility of nonannotated 5′ exon sequence or long distances between the TS site and the first splice-acceptor site required for *trans*-splicing. For most genes, the 3′ limit was arbitrarily set at 200 nt. However, in order to reduce the chance of scoring degradation products as TS sites, this 3′ limit was reduced to 100 nt for very abundantly transcribed genes whose total read counts exceeded 1,000 reads per 10 million. Using these criteria, we could assign candidate TS sites to more than 50% of annotations in WS215 (Table S1B), including 52% of annotated protein-coding genes (10,667 genes), 15% of annotated pseudogenes (226), 54% of annotated noncoding RNAs (137), 74% of snoRNAs (102), and 37% of snRNA genes (42). We found that TS sites often appeared to be clustered over regions of several to sometimes more than 50 nt (Figure 2A, *CapSeq*), suggesting that there is an inherent flexibility in transcription initiation mediated by RNA Pol II at these promoters.

We also identified ~20,000 candidate TS sites that did not pair with annotations based on our criteria (Table S1B). These included 12,457 clusters of TS sites that resembled TS-site clusters typical of annotated Pol II genes. The majority of these (84%) were separated from other annotations or from each other by greater than 1 kb. These findings suggest that there are many as yet nonannotated Pol II loci in the *C. elegans* genome and/or that many of the existing annotations are separated from their actual 5′ ends by greater than the arbitrarily set 1 kb limit used for our analysis.

Identification of Primary miRNA TS Sites

miRNAs are sequentially processed from primary transcripts (pri-miRNAs) synthesized by Pol II. Drosha processes pri-miRNAs into stem-loop precursors (pre-miRNAs) that are exported to the cytoplasm and processed by Dicer into mature miRNAs (Hutvagner and Simard, 2008). In most organisms, the TS sites of pri-miRNAs have not been identified, likely because the original 5′ end is rapidly removed during miRNA maturation.

To identify candidate TS sites for miRNA genes, we analyzed CapSeq reads mapping upstream of annotated miRNAs. Because many miRNAs are coexpressed in a single primary transcript (Lau et al., 2001), there are only about 100 unique miRNA loci annotated in the *C. elegans* genome, encoding ~140 miRNAs. We identified at least one candidate TS site for 56 of the 100 annotated miRNA loci corresponding to 74 individual mature miRNAs (Table S1E). As with other Pol II loci, we found that CapSeq reads that mapped upstream of the pre-miRNAs were often clustered within a 50 nt interval (Figure 2D; Table S1E). We found evidence for only a single group of TS sites upstream of each miR cluster, including the *mir-54–56, mir-35–41* and *mir-229/64–66* clusters (Figure 2D; Table S1E), indicating that each cluster is coexpressed (Lau et al., 2001). Pri-miRNAs were rarely *trans*-spliced; a total of five SL-containing reads were associated with pri-miRNAs, and all of these were spliced to pri-let-7. The five SL-containing reads mapped ~30 nt upstream of the Drosha-processed pre-let-7 RNA, while 20 non-SL reads mapped approximately 200 nt further upstream. Interestingly, we found that some pri-miRNAs were expressed at levels comparable to the pre-mRNAs of common protein-coding genes (Figure S1E), a finding that differs from previously published RNA-Seq data (Lamm et al., 2011; see Discussion).

Mouse CapSeq

The CapSeq protocol we describe should be useful for 5′ anchored RNA profiling from small quantities of tissue. To confirm that CapSeq can identify TS sites from other species, we performed a pilot study using mouse testis RNA. As shown in Figure 3A, we found that CapSeq reads were strongly biased for the 5′ end of annotated mouse genes. By searching for reads upstream of annotated miRNA loci, we identified candidate TS sites for hundreds of primary mouse miRNA genes (Table S2). We also analyzed reads mapping to mouse piRNA clusters. We found that multiple mouse piRNAs appeared to share a TS site (Figure 3B), consistent with previous studies suggesting that piRNAs may be processed from longer precursor RNAs (Aravin et al., 2006; Girard et al., 2006). Finally, we analyzed the motif surrounding candidate mouse TS sites and observed a clear YR motif within a broader motif of YRNyy, in which R (usually an A) corresponds to the predicted 5′ nt (Figure 3C). Thus our data show that CapSeq is generally useful for identifying Pol II TS sites, and that *C. elegans* TS sites are similar to mammalian TS sites.

CIP-TAP Cloning Identifies Promoter-Associated csRNAs Genome Wide

Despite the considerable depth of our CapSeq libraries, we were surprised to find that only 217 of >9,000 annotated unique 21U-RNA loci were identified by CapSeq reads. These included a single 88 nt CapSeq read mapping to the locus from which Cecere et al. (2012) identified a 70 nt 5′ RACE product. In contrast, our mouse CapSeq data identified CapSeq reads associated with most of the annotated piRNA loci. These findings could indicate that pre-21U-RNAs are exceptionally unstable or, alternatively, that the actual precursors are shorter than the 70 nt (minimum length) sequence tags amplified by CapSeq. To test the latter idea, we employed CIP-TAP cloning to identify native csRNAs (Figure 1; Extended Experimental Procedures). Total RNA was size-fractionated to recover 18–40 nt RNA species. To select against the recovery of abundant uncapped small RNA species, including 22G-RNAs, miRNAs and mature piRNAs, we treated the sample with CIP to remove 5′ mono- or triphosphates, reducing the accessibility of these species to 5′ ligation. The 3′ end of the small RNA was ligated to a linker and gel purified. Three fourths of the sample was then treated with TAP to decap csRNAs, thus exposing a 5′ monophosphate for 5′ ligation. The remaining one fourth was treated with polynucleotide kinase (PNK) to add a 5′-phosphate onto noncapped small RNA species. The CIP-TAP and CIP-PNK samples were then ligated to a 5′ linker, gel purified, reverse transcribed, and PCR amplified.

Deep sequencing of the CIP-PNK sample revealed abundant 22G-RNAs, miRNAs and piRNAs, but very few small RNAs that mapped to Pol II promoters (Figure 2A). In contrast, the CIP-TAP sample dramatically enriched small RNAs upstream of Pol II loci (Figure 2A). After normalizing to total nonstructural reads that match the genome, csRNA reads that map within 1,000 nt upstream of WS215 5′ end annotations (including 21U-RNA annotations, see below) were enriched 60-fold in the CIP-TAP sample relative to the CIP-PNK sample (Figure 2C). In contrast, mature miRNAs and 21U-RNA reads were depleted ∼4-fold and 17-fold, respectively, in the CIP-TAP sample. The relative rate at which 22G-RNAs were recovered did not change significantly between the CIP-TAP and CIP-PNK samples. We found that 42% of the sense-oriented reads identified in the CIP-TAP sample corresponded exactly to the 5′ ends of candidate TS sites identified in our CapSeq analysis (Figure 2A; Tables S1B and S1C), supporting the idea that CIP-TAP treatment enriches for *C. elegans* TS-site-associated csRNAs.

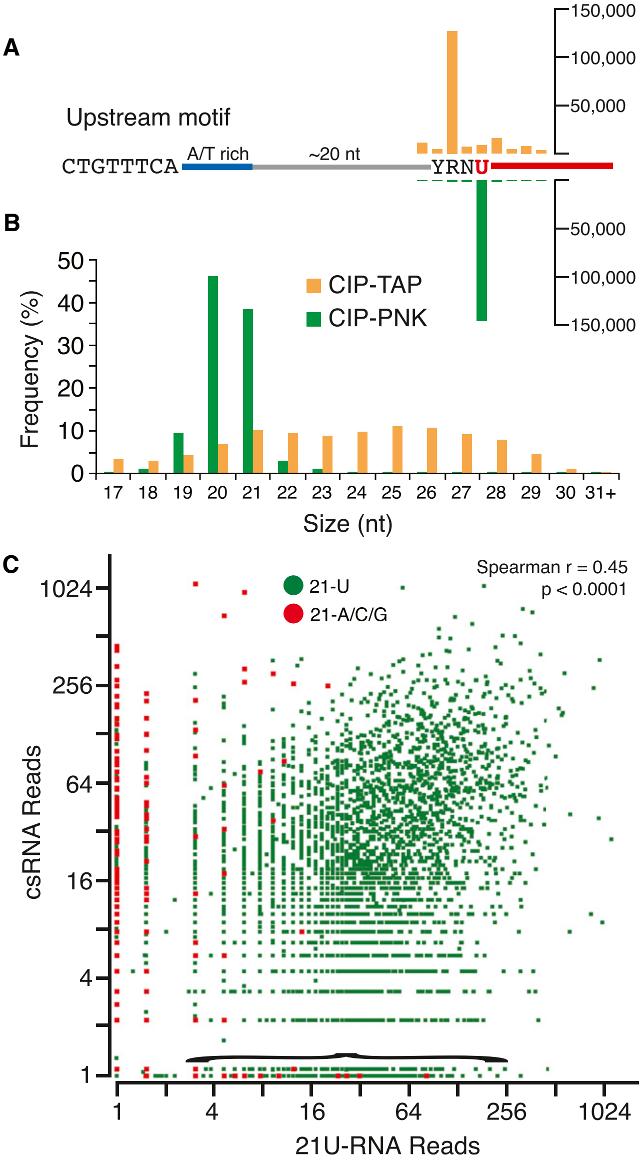


Figure 4. Analysis of Annotated 21U-RNA Loci

(A) Cumulative analysis of the 5′ ends of unique CIP-TAP (orange) and CIP-PNK (green) sequences with respect to the YRNU motif of a consensus 21U-RNA locus. The scale (linear) is shown. The red segment preceded by U indicates the mature 21U-RNA.

(B) Graph showing the length distribution of CIP-TAP/csRNA reads (orange) and CIP-PNK/ 21U-RNA reads (green) mapping to 21U-RNA loci.

(C) Graph of csRNA levels plotted against corresponding 21U-RNA levels for annotated 21U locus. The points in red indicate previously annotated 21A/G/C piRNAs. Points near the *x* axis (under the bracket) include 22G-RNAs previously misannotated as piRNAs.

See also Figure S2 and Table S3.

csRNAs Originate 2 nt Upstream of Mature 21U-RNAs

Analysis of our CIP-TAP data revealed that csRNAs strongly correlate with 21U-RNAs. There are 9,079 21U-RNA loci that each express a single (nonoverlapping) 21U-RNA that matches uniquely in the genome. Among a total of 4.5 million CIP-TAP reads, we identified candidate csRNAs that map to approximately 6,000 of the 9,079 unique 21U-RNA loci. Interestingly, we observed a very strong bias for the 5′ ends of csRNA reads to map 2 nt upstream of the mature 21U-RNA species (∼4,600 of 6,000 csRNA/21U-RNA pairs, one-tail $p = 0$, binomial distribution; Figure 4A).

csRNA reads with 5′ ends that align 2 nt upstream of mature 21U-RNAs (−2 csRNAs) peaked broadly at a length of 25–26 nt

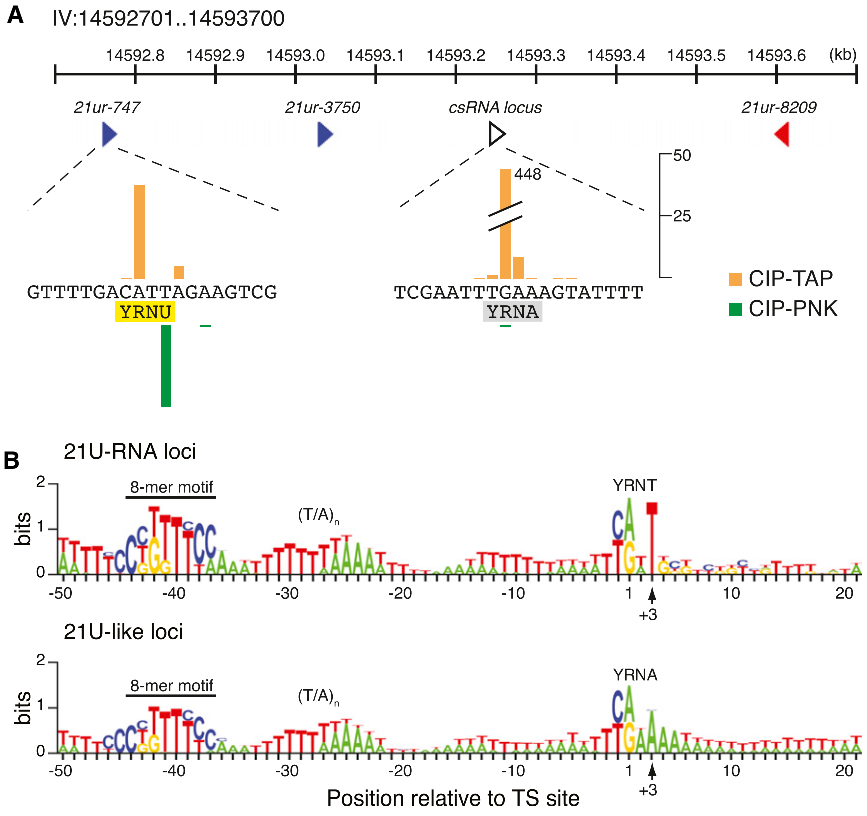


Figure 5. Analysis of 21U-like Loci

(A) An example of a 21U-like locus with a piRNA cluster on chromosome IV. Both *21ur-747* (blue triangle at left) and a nearby 21U-like or csRNA locus (open triangle) are enlarged to single nucleotide resolution. The bars indicate the number of reads (linear scale provided) sharing the corresponding 5′ nt from CIP-TAP (orange) and CIP-PNK (green), relative to the YRNU motif (yellow) of *21ur-747* and the YRNA motif (gray) of the 21U-like locus (open triangle).

(B) Schematic representation of the nucleotide composition at canonical 21U-RNA loci (top) and 21U-like loci (bottom). The nucleotide height (in bits) represents the log2 ratio of the frequency observed relative to the expected frequency based on genomic nucleotide composition. The upstream and TS-site (YR) motifs are indicated. The observed 5′ end of mature 21nt-RNAs is indicated by the arrow at the +3 position.

See also Figure S2 and Table S3.

(Figure 4B). These −2 csRNAs were enriched approximately 60-fold by CIP-TAP cloning relative to the CIP-PNK cloning method (Gu et al., 2009), which enriches for mature 21U-RNAs. Consistent with a precursor-product relationship, we found that the level of −2 csRNAs recovered by CIP-TAP significantly correlates with the level of mature 21U-RNAs recovered by CIP-PNK (Figure 4C). In contrast, the longer −2 CapSeq reads were not only rarely detected (44 of the 217 loci mentioned above; Figure S2A; Table S3C), but also poorly correlated with corresponding mature 21U-RNA levels, as discussed below. Some loci previously annotated as highly expressed 21U-RNA loci failed to produce detectable csRNAs (Figure 4C, points just above the *x* axis). Visual inspection using GBrowse revealed that several of these loci are likely to be derived from degraded 22G-RNAs that were misannotated as 21U-RNAs (data not shown). Loci with abundant csRNAs that lack corresponding mature 21U-RNAs (Figure 4C, points along the *y* axis) define a set of “21U-like” loci discussed below.

The YRNT Motif Associated with 21U Loci Represents a Transcription Start Site

As observed for CapSeq reads, CIP-TAP reads genome wide exhibited a strong bias for initiating at a YR motif (Figures 2B and S1A), consistent with the idea that csRNAs are independently initiated Pol II products. As noted above, csRNAs tend to initiate 2 nt upstream of the corresponding mature 21U-RNAs at 21U-RNA loci, which may explain why the majority of 21U-loci exhibit a YRNT motif (Ruby et al., 2006), where R is the first nt of the csRNA and T corresponds to the 5′U of the mature 21U-RNA.

To look for a correlation between the presence of a YR motif and the levels of csRNA and 21U-RNA expression, we considered ∼1000 pairs of 21U-RNAs, for which the 5′ ends in each pair are separated by 1 nt. Due to the 1 nt separation, a −2 YR motif can only exist for one member of each paired 21U-RNA. These paired loci, which account for 40% of 21U-RNAs that lack a −2 YR motif, provided an opportunity to examine the expression levels of YR- and non-YR-associated transcripts driven from the same promoter. Consistent with the idea that the YR motif is important for transcription, we found that, regardless of their arrangement (5′ or 3′) at tandem loci, both the csRNAs and corresponding mature 21U-RNAs were 10-fold more abundant for the YR-containing sister than for the non-YR-containing sister (paired t test, $p < 0.0001$). Taken together, these findings suggest that the previously defined YRNT motif is a transcription initiation site where R (usually an A) encodes the +1 nt of a pre-21U (csRNA) and T encodes the +3 U, which corresponds to the 5′ end of a mature 21U-RNA (Figure 4A).

A +3 U Is Required for piRNA Production or Stability

While analyzing the CIP-TAP data, we noticed that there were many loci within the 21U-RNA clusters on chromosome IV for which abundant csRNAs were detected but mature 21U-RNAs were not. Altogether, we identified 2,309 csRNA-producing loci that fail to produce mature 21U-RNAs (Table S3A). The csRNA reads obtained from these loci were similar in both size and abundance to those derived from canonical 21U-RNA loci (Figures 5A and S2B). Furthermore, most of these loci (65%; Table S3A) exhibited an adjusted motif score greater than 7, typical of canonical 21U-RNA loci with the upstream 8 nt motif (Ruby et al., 2006). Interestingly, the csRNAs produced at these

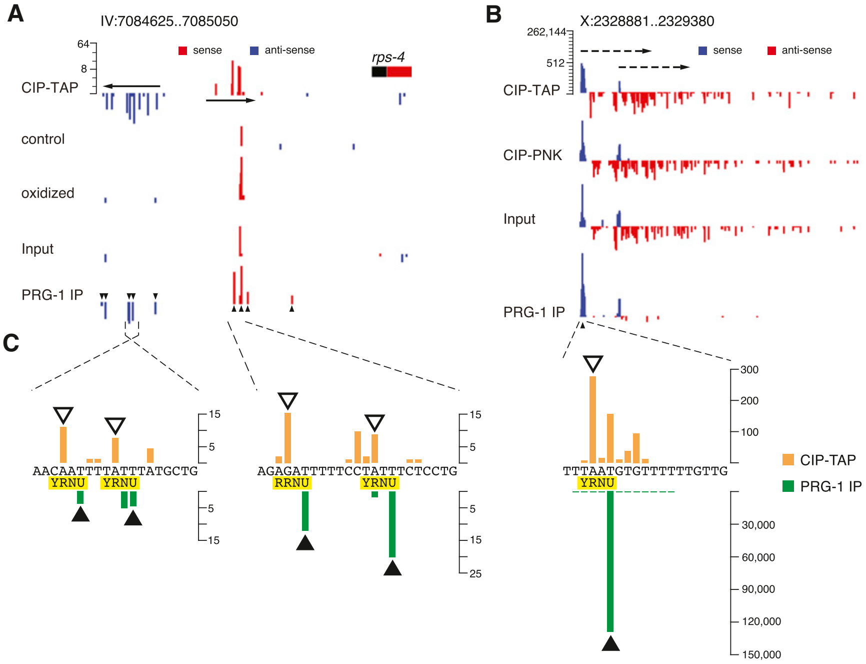


Figure 6. Analysis of Type 2 21U-RNA Loci

(A and B) Comparative analysis of reads from RNA-seq protocols (as indicated). The histograms represent the frequency of reads sharing the same 5′ end at the *rps-4* locus (A) and at a nonannotated locus on chromosome X (B). The genome coordinates are indicated. A log2 scale is provided for each set of histograms. Mature 21U-RNA reads are enriched in the oxidized versus control (A) and PRG-1 IP versus Input (A and B) samples, as indicated. Closed triangles indicate the positions of mature 21U-RNAs enriched in the PRG-1 IP. In (B), the red bars correspond to WAGO-dependent 22G-RNAs likely targeting nonannotated transcripts (dashed arrows) of unknown length.

(C) Enlarged regions indicating the precise positions of corresponding CIP-TAP (orange) and PRG-1 IP (green) reads relative to the YRNU motif indicated below the sequence. Note that one of the TS sites for *rps-4* is RR rather than YR. The open triangles above the CIP-TAP bars point to the likely precursor of the mature 21U-RNAs, which are indicated by the closed triangles below the PRG-1 IP bars.

See also Figure S3 and Table S3.

loci lack a +3U—the majority (~60%) contained a YRNA rather than a canonical YRNT motif (Figure 5B). Approximately 400 previously annotated 21U-RNAs actually start with a 5′ nt other than U (~3% of annotated 21U-RNAs; Batista et al., 2008; Ruby et al., 2006), and we noted that ~60% of these previously detected 21nt-RNAs exhibit corresponding csRNAs. Further examination of these 21U-like loci revealed that the mature 21nt-RNAs (Figure 4C, red) were an average of 10-fold less abundant relative to their corresponding csRNAs than were 21U-RNAs from canonical 21U-RNA loci (Figure 4C, green). These findings suggest that 21U-like loci express csRNAs at normal levels, but that the mature piRNAs are either inefficiently processed or unstable.

Given the large number of 21U-like loci, we reasoned that polymorphisms in *C. elegans* wild isolates might convert the +3 residue of a csRNA to a U at one or more of these loci. Consistent with this possibility, we identified two 21U-RNAs (IV: 17159702-22 and IV: 15903563-83) cloned from JU1580 (Félix et al., 2011) and CB4856 (E.M.Y. and C.C.M., unpublished data), respectively, that mapped to 21U-like loci in the N2 background. In both cases, independent deep-sequencing data confirmed that the wild isolates contain SNPs in the corresponding 21U-like loci that change the +3 residue of the csRNA to a U. Taken together, these findings suggest that the YR portion of the YRNT motif is likely sufficient for transcription initiation, while a U at position +3 of the csRNA is important for 21U-RNA processing and/or stability.

Capped Small RNAs Produced throughout the Genome Are Processed into 21U-RNAs

The majority of csRNAs produced throughout the genome lack the conserved 8 nt consensus (CTGTTTCA) that was weighted heavily in the annotation of canonical 21U-RNA loci (compare Figures 2B and S1A to Figure 5B). However, as expected by chance, many csRNAs lacking the 8 nt motif nevertheless exhibit a YRNT motif (Figures 2B and S1A) and thus contain a U at position +3. Therefore, we asked whether this subset of csRNAs, which are associated with TS sites of protein-coding and other Pol II transcripts, might also be processed into mature 21U-RNAs and loaded onto the Piwi Argonaute PRG-1. Indeed, a number of previously annotated 21U-RNAs coincide with csRNAs proximal to protein-coding genes on chromosomes other than chromosome IV, where the piRNA clusters reside (data not shown). To investigate this further, we deep-sequenced piRNAs enriched by PRG-1 immunoprecipitation (IP). This new IP deep-sequencing data was consistent with previously published and unpublished PRG-1 IP deep-sequencing data (Batista et al., 2008). However, the cloning methods used to generate the previous PRG-1 IP data sets also generated much more noise from degraded mRNAs than did the TAP cloning procedure used here. Altogether, we identified 12,183 new 21U-RNA species (Table S3B), of which ~10,000 exhibit a poor motif score and are associated with TS sites throughout the genome (Figure 6). We refer to this class as "type 2" 21U-RNAs.

We next asked whether type 2 21U-RNAs exhibit the same 2′-O-methyl modification found on the 3′ ends of canonical 21U-RNAs. Consistent with this idea, examination of a previously published data set (Vasale et al., 2010) revealed that the majority of type 2 21U-RNAs were resistant to 3′ end oxidization (Figure 6A). Like canonical (or type 1) 21U-RNAs, type 2 21U-RNAs were only expressed in the germline, consistent with the germline-specific expression of PRG-1. Soma-specific loci, such as the gut-specific gene *vit-1* that produces abundant +3U containing csRNAs, did not give rise to 21U-RNAs (data not shown).

Altogether, our data define more than 12,000 new 21U-RNA species, nearly doubling the total number of piRNAs in

C. elegans. Moreover, type 2 21U-RNAs include several extremely abundant 21U-RNA species. In fact, the single most abundant 21U-RNA is a type 2 21U-RNA expressed from an X chromosome locus (Figures 6B and 6C). This X-locus is intriguing in that it is one of six homologs (all on X) with extensive sequence identity flanking distinct 21U-RNAs (Figures S3A and S3B). The SL1-spliced leader locus and several snRNA loci were also found to produce very abundant 21U-RNA species (Figures S3C and S3D). However, the majority of type 2 21U-RNA loci produce relatively low levels of mature 21U-RNAs (Table S3B), accounting for approximately 5% of total 21U-RNA levels (see Extended Experimental Procedures).

csRNAs Are Unlikely to Be Processed from Longer Pol II Transcripts

The above findings suggest that −2 csRNAs are very likely the precursors for mature 21U-RNAs, but they do not address whether csRNAs might be processed from longer, exceptionally rare transcripts. To address this possibility, we compared 21U-RNA expression levels to the levels of −2 csRNAs and longer −2 CapSeq RNA reads at type 1 and type 2 21U-RNA loci. There are only 44 type 1 loci with −2 CapSeq reads (Table S3C; Figure 7A). Despite the small sample size, we found a very significant correlation between csRNA and 21U-RNA levels at these loci (Spearman $r = 0.69$, $p < 0.0001$; Figure 7A). The correlation coefficient r is similar to that observed previously for all type 1 21U-RNA loci (Figure 4C). However, we found no correlation between CapSeq reads and 21U-RNA levels (Spearman $r = –0.08$, $p < 0.64$; Figure 7A). We also examined the 982 type 2 loci with −2 csRNA and −2 CapSeq reads, and again found a significant correlation between csRNA and 21U-RNA levels but no significant correlation between 21U-RNA levels and −2 CapSeq read levels (Figure 7B).

The failure of CapSeq reads to correlate with 21U-RNA expression levels was not due to a lack of depth in the CapSeq data set. There were, in fact, >40-fold more CapSeq reads than csRNA reads analyzed in our young adult data sets. Furthermore, as noted above, we frequently observed promoter regions with bidirectional, divergent csRNAs (Figure 2A; Tables S1C and S1D). At these bidirectional loci, we found that oppositely oriented −2 csRNAs appeared equally likely to give rise to 21U-RNAs (Figure 6A), whereas longer CapSeq reads were almost exclusively sense oriented (Figure 2A). Moreover, as noted above, even when sense CapSeq reads were present at −2 relative to a type 2 21U-RNA species, CapSeq reads and mature 21U-RNA levels failed to exhibit a significant correlation. Together these findings suggest that RNA Pol II transcribes csRNAs directly, and that −2 csRNAs (not longer RNA species) are the 21U-RNA precursors.

DISCUSSION

C. elegans piRNA Precursors Are Expressed Individually by Pol II as Capped Small RNAs

Here we have explored the biogenesis of *C. elegans* piRNAs (21U-RNAs). To do so, we employed two approaches, CapSeq and CIP-TAP, both of which enrich for the 5′ ends of Pol II transcripts. The CapSeq protocol, designed to select for long-capped RNAs, identified reads mapping to the 5′ ends of thousands of other Pol II genes, but detected reads mapping to only 0.5% of 21U-RNA loci (44 out of 9,079 unique 21U-RNAs). The CIP-TAP protocol, on the other hand, designed to detect capped small RNAs, identified thousands of candidate 21U-RNA precursor transcripts (more than 50% of 21U loci) that average 26 nt in length and initiate 2 nt upstream of the mature piRNA species. In addition, CIP-TAP identified csRNAs that were associated with many other Pol II promoters, where they were frequently oriented divergently, with the sense csRNA often corresponding to a major TS site for the corresponding longer transcript detected by CapSeq.

Strikingly, germline-expressed csRNAs that contain a U at the +3 position were found to correspond to a previously overlooked class of 21U-RNAs associated with Pol II promoters genome wide. These findings indicate that the U in the YRNU motif is important for 21U-RNA stability, processing, or Piwi Argonaute loading, and that the YR is important for efficient transcription initiation (see Model, Figure 7C). Consistent with this idea, the distance between the conserved upstream 8 nt motif and the putative initiator element (YRNT) is similar to the distance between the TFIIB/TATA and the initiator elements of core TS sites described for other organisms (Juven-Gershon et al., 2008). Based on these findings, we now propose that *C. elegans* piRNAs be divided into two categories (Figure 7C): type 1 21U-RNAs, which correspond to the previously defined 21U-RNAs that share an 8 nt upstream motif and are clustered on chromosome IV (Batista et al., 2008; Ruby et al., 2006), and type 2 21U-RNAs, which need not have an 8 nt motif and are processed from csRNAs derived from the promoters of Pol II genes throughout the genome.

An Enzymatic Approach for 5′ End Anchored Transcription Profiling

Transcription profiling by deep sequencing has become an increasingly important tool for following gene expression. The CapSeq protocol described here facilitates transcription profiling by using a series of three enzymatic treatments that dramatically enrich for the 5′ ends of Pol II transcripts. Because CapSeq does not require affinity purification to remove structural RNA contaminants, it can be performed on relatively small quantities of RNA. Aside from a single size-selection step, the entire procedure is carried out in a PCR tube. Importantly, the CapSeq procedure anchors clones at the 5′ cap of Pol II transcripts and thus can clone RNAs with or without poly(A) tails. CapSeq provides a quantitative way to profile a diversity of Pol II transcripts, while providing insight into alternative transcription-initiation sites, which may be of potential developmental significance.

Genome-wide Identification of Pol II TS Sites

The data described here provide a systematic and comprehensive look at the TS sites of Pol II transcripts in *C. elegans*. The *trans*-splicing of SL sequences to the 5′ ends of many mature transcripts confounds the identification of TS sites in *C. elegans.* Consequently, only a handful of TS sites for *C. elegans* Pol II transcripts had been identified prior to the present study (Allen et al., 2011; Morton and Blumenthal,

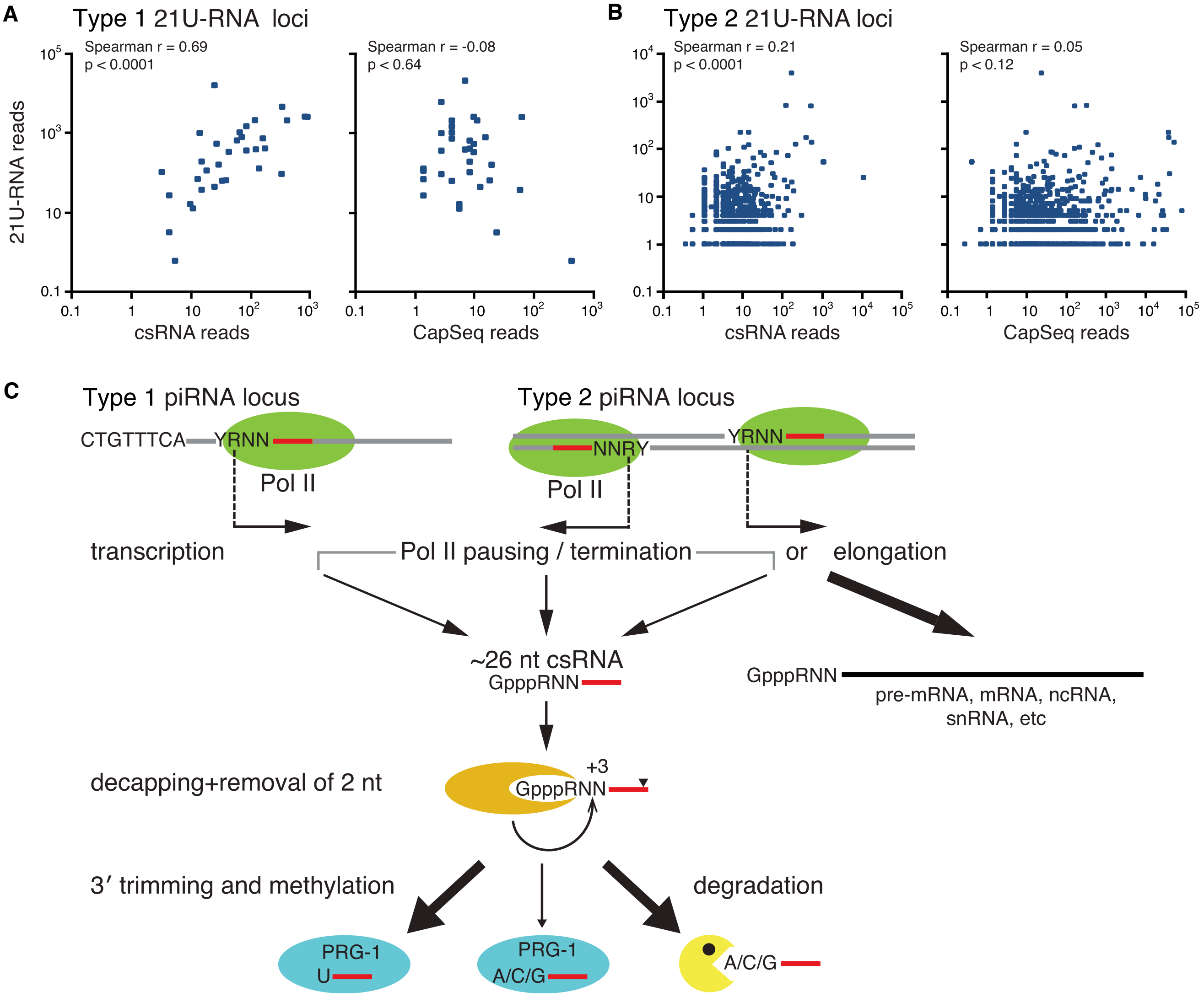


Figure 7. *C. elegans* piRNAs Are Processed from Capped Small RNAs

(A and B) Correlation analyses between 21U-RNAs and csRNAs or long-capped RNAs (as indicated) at the 44 type 1 loci where long-capped RNA reads were obtained by CapSeq (A), and at 982 type 2 loci where csRNA and CapSeq reads were both found at +1 (B), relative to the downstream (+3) U of a 21U-RNA. For a perfect correlation, the Spearman rank correlation coefficient (r) = 1 or −1, and for no correlation, r = 0. p values (p) were calculated using nonparametric correlation model.

(C) Model for the biogenesis of 21U-RNA. Arrows indicate TS sites of type 1 and type 2 piRNA loci.

2011). By using CapSeq to clone capped transcripts from several different developmental stages, we have identified candidate TS sites for approximately 50% of the annotated protein-coding genes in *C. elegans*. In addition, we have identified 5′ ends for Pol II transcripts that are typically under-represented in poly(A)-selected RNA-seq studies, including snRNAs, snoRNAs, SL RNA precursors, and histone mRNAs. In keeping with predictions from previous studies (Allen et al., 2011), we found that an overall 70% of annotated protein-coding genes have *trans*-spliced forms. Because of the abundance of SL-containing reads, our findings provide a comprehensive measure of alternative spliced-leader usage for most genes and also provide useful data for refining the prediction of SL splice-acceptor sites.

Sequencing of a mouse testes CapSeq library also revealed a strong enrichment for RNA 5′ ends and for a YR motif at TS sites of mouse Pol II genes. Our CapSeq analysis identified candidate TS sites for many primary miRNA transcripts in both *C. elegans* and mouse. Altogether, by combining CIP-TAP and CapSeq data, we were able to predict TS sites for 60% of the annotated *C. elegans* miRNA genes (Table S1E). Surprisingly, we found that expression levels, as inferred from read counts for pri-miRNAs, were comparable to that of many abundant protein-coding genes. In contrast, pri-miRNA transcripts were very rarely detected in data from a previous study that used poly(A) selection RNA-Seq protocols (Lamm et al., 2011), suggesting that pri-miRNAs either lose their poly(A) tail more rapidly than their 5′ cap or, perhaps, lack a poly(A) tail entirely. We conclude that CapSeq and CIP-TAP can be used to quantify the activity of a wide variety of Pol II promoters. The approach described here could readily be extended to produce a comprehensive profile of *C. elegans* or mouse TS sites.

Capped Small RNAs Are Associated with Promoters in *C. elegans*

Like the longer reads recovered by CapSeq, csRNAs exhibit a consensus Pol II initiator element yYRyyy. Indeed the 5′ ends of csRNAs often coincide with the 5′ ends of CapSeq reads. However, unlike CapSeq reads, csRNAs were frequently bidirectional at promoters, with divergent csRNAs separated by an average of approximately 150 nt. This finding is consistent with the idea that many eukaryotic promoters are intrinsically bidirectional (Seila et al., 2009). In general, for csRNA and CapSeq reads that share a common 5′ end, the abundance of csRNA reads was proportional to the abundance of CapSeq reads, suggesting that csRNAs might be associated with Pol II initiation at active promoters. Despite their correlation with active gene expression, our analysis suggests that csRNAs are relatively low-abundance transcripts compared to other small RNAs. Based on our CIP-TAP cloning experiments, we estimate that csRNAs represent less than 1% of the total small RNAs in adult *C. elegans*.

Capped small RNAs that flank the TS sites of active promoters have been identified in mammals and *Drosophila* (Core et al., 2008; Haussecker et al., 2008; Seila et al., 2008; Yamamoto et al., 2007). Our data suggest that csRNAs are most similar to promoter-associated short RNAs (PASRs), which were enriched using a CIP-TAP cloning method and had 5′ ends that frequently coincided with those of capped RNAs identified by CAGE (Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009; Kapranov et al., 2007). Although the biogenesis and function of PASRs remains unknown, it has been speculated that PASRs might reflect Pol II pausing or premature termination, or that they are processed from promoter-associated long-capped RNAs (Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009; Nechaev et al., 2010). Our data are most consistent with the idea that csRNAs are independent transcripts rather than processing intermediates derived from longer RNAs. If csRNAs were derived from long-capped RNAs, we would have expected a broader size range continuing up to 32 nt (the largest size we could sequence in this experiment). The size of *C. elegans* csRNAs is similar to the estimated size, ~28 nt, of nascent RNA that can be accommodated in the Pol II exit canal (Andrecka et al., 2008; Chen et al., 2009; Proudfoot et al., 2002). This size is also similar to that of csRNAs found associated with promoter-proximal pausing of Pol II, thought to occur at many genes throughout metazoan genomes (Nechaev et al., 2010; Rasmussen and Lis, 1995).

C. elegans piRNAs Are Processed from Capped Small RNAs

Here we have shown that csRNA loci genome wide give rise to 21U-RNAs, and that the relative levels of the −2 csRNA and corresponding mature 21U-RNA are well correlated (Figures 7A and 7B). The only requirement for 21U-RNA production was the presence of a U residue at the +3 position of the csRNA. These findings are consistent with a model in which csRNAs are precursors for 21U-RNA production (Figure 7C). The canonical 21U-RNA loci (type 1 loci) appear to be specialized to produce csRNAs primarily in one direction. The pattern of RNA expression at these loci was quite distinct from the pattern observed upstream of other Pol II transcripts. Type 1 21U-RNA loci typically produced abundant csRNAs and rarely, if ever, produced longer CapSeq reads (Figure 7C). When multiple csRNAs were produced at 21U-RNA loci, they typically shared the same orientation and their 5′ ends were often separated by less than 5 bp (data not shown). In contrast, other Pol II loci, such as protein-coding genes, produced abundant sense-oriented CapSeq reads, and multiple, relatively low-abundance csRNAs that were often oriented in both directions (Figure 7C). These observations suggest that type 1 21U-RNA loci somehow focus Pol II initiation and restrict elongation to promote csRNA biogenesis at the expense of longer transcripts. In the future, it will be interesting to learn whether the upstream motif or other features of type 1 21U-RNA loci govern their tendency to produce csRNAs but not longer Pol II transcripts.

Our findings suggest that a key step in 21U-RNA production is the removal of the cap and two nucleotides (Figure 7C). Although the exact cap structure of csRNAs is not known, we note that removal of a cap-2 structure (m^7GpppAmNm) would directly expose the +3 U of a pre-21U for PRG-1 loading. Thus, the piRNA processing machinery could be linked to pathways that decap and turnover csRNAs. However, our findings clearly indicate that a U at positions +2 or +4 cannot substitute for a U at +3, suggesting that PRG-1 does not randomly sample 5′ degradation products, as was proposed for Ago1 and priRNAs in *S. pombe* (Conte and Mello, 2010; Halic and Moazed, 2010). Additional biochemical and genetic studies are needed to determine the structure and processing of csRNA caps.

The importance of the YR motif for 21U-RNA expression was highlighted by a subset of 21U-RNA loci that produce two mature 21U-RNA species whose 5′ ends are adjacent nucleotides. It is, of course, only possible for one of these 21U-RNAs to be associated with a YR motif, and the YR motif was approximately equally likely to be associated with the 5′ or 3′ 21U-RNA. Interestingly, regardless of relative order, we found that the YR-associated 21U-RNA was an average of ~10 times more abundant than the non-YR-associated 21U-RNA. In cases where csRNAs were detected for both transcripts (which occurred at 21% of loci), a similar difference in expression level (~10-fold) was also observed between the YR-associated and non-YR-associated csRNAs. These observations support the idea that a YR motif is preferred for strong transcription initiation.

CONCLUSIONS

Recent studies have shown that PRG-1 and its piRNA cofactors provide an important first line of defense in a surveillance pathway that distinguishes self from non-self (Ashe et al., 2012; Lee et al., 2012; Shirayama et al., 2012). Importantly, PRG-1/piRNA complexes function in a context that does not require perfect base-pairing, greatly increasing the repertoire of potential target RNAs in *C. elegans*. The findings described here add to the amazing variety of piRNA biogenesis mechanisms and identify a second type of piRNA that nearly doubles the number of piRNA species available for genome defense in *C. elegans*. The finding that small RNAs associated with TS sites are processed and loaded onto an Argonaute also raises an intriguing possibility that Argonaute-small RNA pathways might regulate promoter activity directly.

EXPERIMENTAL PROCEDURES

Worm and Mouse Strains

The Bristol N2 strain of *C. elegans* was used in this study and cultured essentially as described (Brenner, 1974). Mouse testes were dissected from the C57Bl/6 background. All mouse work was carried out at University of Massachusetts Medical School (UMMS) and was approved by the UMMS IACUC.

RNA Cloning and Sequencing

RNA was extracted using TRI Reagent (MRC, Inc.) or phenol. For CapSeq, 0.5 - 2 μg of total RNA was treated with Terminator 5′-phosphate-dependent exonuclease (Epicenter) to degrade rRNAs, calf intestinal phosphatase (CIP, NEB) to remove 5′ phosphates, and tobacco acid pyrophosphatase (TAP, Epicenter) to remove 5′ caps. The resulting long-capped RNAs were ligated to a 5′ adaptor. The first-strand cDNA was primed using a pool of random octamers containing a common 5′ sequence corresponding to a 3′ adaptor oligo. The first-strand cDNA was size selected and then amplified using Illumina adaptor oligos.

Small RNA libraries were prepared essentially as described (Gu et al., 2009, 2011). Briefly, for CIP-TAP cloning, 18–40 nt RNA was gel purified from 40 μg of total RNA using a 15% PAGE/8M urea gel. The RNA was dephosphorylated with 50 U of CIP in a 100 μl reaction containing 1X NEB Buffer 3 and 0.5 U/μl SUPERase·In (Ambion) at 37°C for 1 hr; it was then extracted with phenol/chloroform, precipitated with isopropanol, ligated to a 3′ linker, and gel purified. Three-fourths of the 3′ ligation product was decapped with 2–4 U TAP in a 10 μl reaction containing 1X TAP buffer and 1 U/μl SUPERase·In at 37°C for 1 hr (CIP-TAP sample). The remaining one-fourth was phosphorylated with 20 U of PNK (NEB) in a 20 μl reaction containing 1X PNK buffer, 0.5 U/μl SUPERase·In, and 2 mM ATP at 37°C for 1 hr (CIP-PNK control). Each sample was then ligated to a barcoded 5′ linker, gel purified, reverse-transcribed, and PCR amplified with Solexa linkers.

Additional details are provided in the Extended Experimental Procedures. Libraries were sequenced using an Illumina Genome Analyzer II or HiSeq instrument at the UMass Medical School Deep Sequencing Core (Worcester, MA).

Bioinformatics

Sequences were processed and mapped using custom PERL (5.10.1) scripts, Bowtie 0.12.7 (Langmead et al., 2009) and BLASTN 2.2.25 (Altschul et al., 1990). For *C. elegans* analysis, reads were mapped to the genome (WormBase release WS215), Repbase 15.10 (Jurka et al., 2005), and miRBase 16 (Kozomara and Griffiths-Jones, 2011). For mouse analysis, reads were aligned to genome assembly NCBIM37 (Ensembl 67), miRBase 18 and the non-coding RNA database fRNAdb 3.4 (Mituyama et al., 2009). The Generic Genome Browser (GBrowse; Stein et al., 2002) was used to visualize the alignments.

Immunoprecipitation

The PRG-1 IP was performed as described previously (Batista et al., 2008). Small RNAs were extracted from IP and input and cloned using a TAP cloning protocol, as described (Gu et al., 2009).

ACCESSION NUMBERS

Illumina data are available from GEO under the series number GSE40053.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Discussion, Extended Experimental Procedures, three figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.11.023.

ACKNOWLEDGMENTS

We thank members of the Mello lab for useful discussions, Tom Blumenthal for providing spliced leader sequences, Stephen Jones for providing mouse sample, Zhiping Weng for providing information about the mouse database, and Ellen Kittler for deep sequencing. H.-C.L. is supported by a Ruth L. Kirschstein National Research Service Award (GM099372). C.C.M. is a Howard Hughes Medical Institute Investigator and is supported by NIH grant GM058800.

Received: August 19, 2012
Revised: October 21, 2012
Accepted: November 14, 2012
Published: December 20, 2012

REFERENCES

Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project. (2009). Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature *457*, 1028–1032.

Allen, M.A., Hillier, L.W., Waterston, R.H., and Blumenthal, T. (2011). A global analysis of *C. elegans trans*-splicing. Genome Res. *21*, 255–264.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Andrecka, J., Lewis, R., Brückner, F., Lehmann, E., Cramer, P., and Michaelis, J. (2008). Single-molecule tracking of mRNA exiting from RNA polymerase II. Proc. Natl. Acad. Sci. USA *105*, 135–140.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203–207.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744–747.

Ashe, A., Sapetschnig, A., Weick, E.M., Mitchell, J., Bagijn, M.P., Cording, A.C., Doebley, A.L., Goldstein, L.D., Lehrbach, N.J., Le Pen, J., et al. (2012). piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans*. Cell *150*, 88–99.

Batista, P.J., Ruby, J.G., Claycomb, J.M., Chiang, R., Fahlgren, N., Kasschau, K.D., Chaves, D.A., Gu, W., Vasale, J.J., Duan, S., et al. (2008). PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. Mol. Cell *31*, 67–78.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363–366.

Beyret, E., Liu, N., and Lin, H. (2012). piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. Cell Res. *22*, 1429–1439.

Blumenthal, T., and Steward, K. (1997). RNA Processing and Gene Structure. In C. elegans II, D.L. Riddle, T. Blumenthal, B.J. Meyer, and J.R. Priess, eds. (Cold Spring Harbor: Cold Spring Harbor Laboratory Press).

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell *128*, 1089–1103.

Brenner, S. (1974). The genetics of *Caenorhabditis elegans*. Genetics *77*, 71–94.

Cecere, G., Zheng, G.X., Mansisidor, A.R., Klymko, K.E., and Grishok, A. (2012). Promoters recognized by forkhead proteins exist for individual 21U-RNAs. Mol. Cell *47*, 734–745.

Chen, C.Y., Chang, C.C., Yen, C.F., Chiu, M.T., and Chang, W.H. (2009). Mapping RNA exit channel on transcribing RNA polymerase II by FRET analysis. Proc. Natl. Acad. Sci. USA *106*, 127–132.

Conte, D., Jr., and Mello, C.C. (2010). Primal RNAs: the end of the beginning? Cell *140*, 452–454.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Das, P.P., Bagijn, M.P., Goldstein, L.D., Woolford, J.R., Lehrbach, N.J., Sapetschnig, A., Buhecha, H.R., Gilchrist, M.J., Howe, K.L., Stark, R., et al. (2008). Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. Mol. Cell *31*, 79–90.

de Hoon, M., and Hayashizaki, Y. (2008). Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques *44*, 627–628, 630, 632.

Félix, M.A., Ashe, A., Piffaretti, J., Wu, G., Nuez, I., Bélicard, T., Jiang, Y., Zhao, G., Franz, C.J., Goldstein, L.D., et al. (2011). Natural and experimental infection of *Caenorhabditis* nematodes by novel viruses related to nodaviruses. PLoS Biol. *9*, e1000586.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202.

Grivna, S.T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev. *20*, 1709–1714.

Gu, W., Shirayama, M., Conte, D., Jr., Vasale, J., Batista, P.J., Claycomb, J.M., Moresco, J.J., Youngman, E.M., Keys, J., Stoltz, M.J., et al. (2009). Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. Mol. Cell *36*, 231–244.

Gu, W., Claycomb, J.M., Batista, P.J., Mello, C.C., and Conte, D. (2011). Cloning Argonaute-associated small RNAs from *Caenorhabditis elegans*. Methods Mol. Biol. *725*, 251–280.

Gunawardane, L.S., Saito, K., Nishida, K.M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M.C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. Science *315*, 1587–1590.

Halic, M., and Moazed, D. (2010). Dicer-independent primal RNAs trigger RNAi and heterochromatin formation. Cell *140*, 504–516.

Haussecker, D., Cao, D., Huang, Y., Parameswaran, P., Fire, A.Z., and Kay, M.A. (2008). Capped small RNAs and MOV10 in human hepatitis delta virus replication. Nat. Struct. Mol. Biol. *15*, 714–721.

Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., et al. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell *129*, 69–82.

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nat. Rev. Mol. Cell Biol. *9*, 22–32.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Juven-Gershon, T., Hsu, J.Y., Theisen, J.W., and Kadonaga, J.T. (2008). The RNA polymerase II core promoter - the gateway to transcription. Curr. Opin. Cell Biol. *20*, 253–259.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488.

Kawaoka, S., Izumi, N., Katsuma, S., and Tomari, Y. (2011). 3′ end formation of PIWI-interacting RNAs in vitro. Mol. Cell *43*, 1015–1022.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. *39*(Database issue), D152–D157.

Lamm, A.T., Stadler, M.R., Zhang, H., Gent, J.I., and Fire, A.Z. (2011). Multimodal RNA-seq using single-strand, double-strand, and CircLigase-based capture yields a refined and extended description of the *C. elegans* transcriptome. Genome Res. *21*, 265–275.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science *294*, 858–862.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367.

Lee, H.C., Gu, W., Shirayama, M., Youngman, E., Conte, D., Jr., and Mello, C.C. (2012). *C. elegans* piRNAs mediate the genome-wide surveillance of germline transcripts. Cell *150*, 78–87.

Lin, H. (2007). piRNAs in the germ line. Science *316*, 397.

Mituyama, T., Yamada, K., Hattori, E., Okida, H., Ono, Y., Terai, G., Yoshizawa, A., Komori, T., and Asai, K. (2009). The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. Nucleic Acids Res. *37*(Database issue), D89–D92.

Morton, J.J., and Blumenthal, T. (2011). Identification of transcription start sites of *trans*-spliced genes: uncovering unusual operon arrangements. RNA *17*, 327–337.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. Science *327*, 335–338.

Pak, J., and Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. Science *315*, 241–244.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501–512.

Rasmussen, E.B., and Lis, J.T. (1995). Short transcripts of the ternary complex provide insight into RNA polymerase II elongational pausing. J. Mol. Biol. *252*, 522–535.

Robine, N., Lau, N.C., Balla, S., Jin, Z., Okamura, K., Kuramochi-Miyagawa, S., Blower, M.D., and Lai, E.C. (2009). A broadly conserved pathway generates 3′UTR-directed primary piRNAs. Curr. Biol. *19*, 2066–2076.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell *127*, 1193–1207.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849–1851.

Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle *8*, 2557–2564.

Shirayama, M., Seth, M., Lee, H.C., Gu, W., Ishidate, T., Conte, D., Jr., and Mello, C.C. (2012). piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. Cell *150*, 65–77.

Sijen, T., Steiner, F.A., Thijssen, K.L., and Plasterk, R.H. (2007). Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science *315*, 244–247.

Siomi, H., and Siomi, M.C. (2009). On the road to reading the RNA-interference code. Nature *457*, 396–404.

Smale, S.T., and Baltimore, D. (1989). The ‘‘initiator’’ as a transcription control element. Cell *57*, 103–113.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. (2002). The generic genome browser: a building block for a model organism system database. Genome Res. *12*, 1599–1610.

Vasale, J.J., Gu, W., Thivierge, C., Batista, P.J., Claycomb, J.M., Youngman, E.M., Duchaine, T.F., Mello, C.C., and Conte, D., Jr. (2010). Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. Proc. Natl. Acad. Sci. USA *107*, 3582–3587.

Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K., and Abe, T. (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics *8*, 67.

Supplemental Information

EXTENDED DISCUSSION

Estimating the Fidelity of CapSeq

To estimate the fidelity with which CapSeq defines the actual 5′ ends of transcripts, as opposed to internally truncated RNAs, we analyzed the degradation rate of the spliced leader SL1, a 22 nt 5′ leader sequence *trans*-spliced to many *C. elegans* genes (Allen et al., 2011). For this analysis, we chose five highly-expressed SL1 *trans*-spliced genes: *alh-8*, *rps-24*, *rps-3*, *rps-29*, and *rps-15*. We analyzed complete or partial SL1 reads appended with sequences that perfectly match to the first 30 nt of each gene. For technical and computational reasons, we considered all sequences with at least the last 3 nt of SL1 and excluded sequences missing only the first nt of SL1 (see Extended Experimental Procedures). We found that the average frequency of cloning a 5′-truncated product at each position (nt 3–20, positions counting from the last nucleotide) of SL1 was approximately 1 in 15,000 per nucleotide relative to intact SL1 reads.

As an additional and more general approach to place an upper limit on the frequency of cloning degradation products by CapSeq, we considered reads that map to all 6,967 *trans*-spliced transcripts both annotated by WormBase and confirmed by our data set. We then compared the number of non-SL reads that start at each position of the transcript to the total number of SL-containing reads. This analysis revealed that for every SL read detected, a non-SL read (potential degradation product) occurred at an average rate of 1 in 17,000 at each position along a transcript. Together, these data indicate that CapSeq strongly enriches for the 5′-most nt of capped transcripts.

EXTENDED EXPERIMENTAL PROCEDURES

CapSeq Protocol

Capped RNA Enrichment and Cap Removal

To destroy 5.8S, 18S, and 26S rRNA, total RNA (0.5–2 μg) was treated with 0.1 U/μl Terminator 5′-phosphate-dependent exonuclease (Epicenter, TER51020) in 20 μl volume containing 1U/μl of SUPERase·In (Ambion, AM2696) and 1X Buffer A at 30°C for 1-2 hr. To monitor the efficiency of the Terminator reaction (for troubleshooting), a control sample can be treated and resolved on a 5% denaturing polyacrylamide gel and visualized using Ethidium Bromide. To de-phosphorylate tRNA and 5S rRNA, the volume of the Terminator reaction was adjusted to 100 μl with 10 μl of 10X DNase I buffer, 20 U more of SUPERase·In, 30 U of CIP (NEB, M0290S), 5 U of DNase I (Ambion, AM2222), and H_2O, and the sample was incubated at 37°C for 30 min. DNase I treatment here removes any residual DNA contamination, if any. The RNA was extracted in a phase-lock column (5PRIME, 2302830) using 1 volume of phenol/chloroform (1:1 ratio, pH 7.5), and then precipitated with 0.3M NaAc (pH 5.2) and 1 to 1.3 volume of isopropanol plus 30 μg of glycoblue (Ambion, AM9515). To remove the cap structure, the RNA pellet was resuspended with a 10 μl reaction mixture containing 1U/μl SUPERase·In, 1X TAP buffer and 0.25 U/μl tobacco acid pyrophosphatase (TAP; Epicenter, T19050) and incubated at 37°C for 1 hr. The decapped RNA was extracted with phenol/chloroform and precipitated, as described previously, without adding additional glycoblue (still present from previous step).

Addition of 5′ Linker

To add a 5′ linker bearing a 4 nt barcode at the 3′ end, the decapped RNA pellet was resuspended in a 10 μl reaction containing 5 μM 5′ linker, 1X T4 RNA ligation buffer, 1 U/μl of SUPERase·In, 2 U/μl of T4 RNA ligase (Takara, 2050A), 10% DMSO, and 0.1 μg/μl BSA, and the sample was incubated at 15°C for 8 hr, and then extracted and precipitated, as described previously, without adding additional glycoblue.

Random-Primed cDNA

First-strand cDNA was generated by random priming using a library of random octamers appended with the 3′ linker sequence, which is required for PCR amplification later. First, the precipitated 5′ linker::RNA ligation products were annealed to 50 pmole of random octamer::3′ linker oligo in a 13 μl volume containing 10 nmole of each dNTP at 65°C for 5 min, and then chilled on ice for at least 2 min. The sample was then incubated in a final volume of 20 μl reaction with addition of 5 U/μl Superscript III (Invitrogen, 18080), 1X Superscript buffer, 10 mM DTT, and 1 U/μl SUPERase·In. The RT reaction was incubated at 15°, 25°, and 37°C each for 15 min, and then at 50°C for 30 min. Finally, the Superscript III enzyme was heat-inactivated at 85°C for 5 min. RNA was destroyed by adding 1 μl each of RNase A (Ambion, AM2270) and RNase H (Ambion, AM2292) and incubating at 37°C for 20 min.

Linear Amplification

To increase the cDNA quantity, a linear PCR was performed, as follows: the 20 μl RT reaction was diluted into a 100 μl PCR reaction containing 1X PCR buffer, 0.01 U/μl ExTaq (Takara, RR001B), 0.05 μM of oligo CMo13729, 0.2 mM dNTP, and H_2O. Linear PCR was performed for 12 cycles with the parameters: 94°C for 20 s, 55°C for 20 s, and 72°C for 30 s.

Gel Purification

The cDNA was fractionated on a 15% acrylamide/8M Urea denaturing gel and visualized using SYBRGold (Invitrogen, S-11494). Total RNA containing 5S RNA (~120 nt) and 5.8S RNA (~160 nt) was used as size markers for gel purification of cDNA (~130–170 nt or desired size including the 5′ and 3′ linkers) for all samples, except ya0217, which was ~50–130 nt). The cDNA was eluted in TE buffer (10 mM Tris pH 7.5, 1 mM EDTA) containing 0.3 M NaCl for 6 hr to overnight with constant shaking. The eluate was filtered through a 0.45 μm Spin-X column (Costar, 19442-758) and precipitated with 1–1.3 volume of isopropanol and glycoblue, as above.

Library Amplification

To determine the optimal cycle number for cDNA amplification, a 50 μl reaction containing 20% of the eluted linear PCR product, 1X ExTaq Buffer, 0.2 μM each of the oligos CMo13279 and Solexa3sh, 0.25 mM dNTP and 0.025 U/μl ExTaq was amplified for 15 cycles with the parameters: 94°C for 20 s, 52°C for 20 s, and 72°C for 30 s. Then 5 μl of each 10 μM oligo, Solexa3 and CMo13278, were added and the reaction continued for an additional 12 cycles. During these 12 cycles, a 3 μl of PCR product was sampled at each additional 3 cycles. These samples were then resolved on an 8% polyacrylamide native gel along with a 10 bp DNA marker (Invitrogen 10821-015) and visualized using Ethidium Bromide. The optimal cycle number was defined as the one at which PCR products of the desired size (∼160–200 nt; or ∼80–160 nt for ya0217) were obtained without obvious heteroduplexes, which result in a diffuse band that migrates more slowly (Gu et al., 2011). A final PCR reaction was performed using the optimized cycle conditions and gel purified or affinity purified (PCR purification), depending on the quality of the product.

Quality Control and Illumina Sequencing

A portion of each library was cloned using TOPO-TA (Invitrogen) and at least 10 individual clones from each library were sequenced by a standard method to ensure that the majority of amplicons had a cDNA insert flanked by the Illumina linkers. Each library was then analyzed by single-end long-read (75 or 100 nt) sequencing on an Illumina Genome Analyzer II or a Hi-Seq instrument at the UMass Medical School Deep Sequencing Core.

Oligos Used in CapSeq

5′ linker: DNA/RNA hybrid (ribonucleotides preceded by ‘r’), TCTACrArGrUrCrCrGrArCrGrArUrC + barcode; barcodes, (A) rTrGrArC, (B) rCrArGrT, (C) rGrCrTrG, (D) rArTrCrA.

RT oligo (N = random nucleotide): CAGAAGACGGCATACGANNNNNNNN.

PCR oligos: solexa3, CAAGCAGAAGACGGCATACGA; solexa3sh, GCAGAAGACGGCATACGA; CMo13279, GTTCTACAGTCCGACGATC; CMo13278, AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGACGATC.

Small RNA Cloning

CIP-TAP Cloning

To clone capped small (cs) RNA, 40 μg of total RNA was fractionated on a 15% polyacrylamide/8M Urea denaturing gel and stained with Ethidium Bromide. The region of the gel containing RNA of 18–40 nt was excised, and the small RNAs were eluted overnight in TE buffer (10 mM Tris pH 7.5, 1mM EDTA) containing 0.3 M NaCl with constant shaking. The eluted RNA was dephosphorylated with 50 U of CIP in a 100 μl reaction containing 1X NEB Buffer 3 and 0.5 U/μl SUPERase·In (Ambion) at 37°C for 1 hr, and then extracted with phenol/chloroform, precipitated with isopropanol, ligated to a 3′ linker (IDT, miRNA cloning linker 1) and gel purified, essentially as described (Gu et al., 2009, 2011). Three-fourths of the 3′ ligation product was decapped with 2–4 U TAP in a 10 μl reaction containing 1X TAP buffer and 1 U/μl SUPERase·In at 37°C for 1 hr (CIP-TAP sample). The remaining one-fourth was phosphorylated with 20 U of PNK (NEB) in a 20 μl reaction containing 1X PNK buffer, 0.5 U/μl SUPERase·In, and 2 mM ATP at 37°C for 1 hr (CIP-PNK control). Each sample was then ligated to a barcoded 5′ linker (as for CapSeq above), gel purified, reverse transcribed, and PCR amplified with Solexa linkers.

TAP Cloning

The small RNAs extracted from the PRG-1 IP and corresponding control (Input) samples, were cloned using the TAP cloning method, as described (Batista et al., 2008; Gu et al., 2009). The TAP cloning method clones far fewer degradation products than the CIP-PNK cloning or ligation-independent cloning methods (Batista et al., 2008; Gu et al., 2009). The oxidized sample enriches for 3′ modified small RNAs and was previously described (Vasale et al., 2010). In most cases samples were cloned using 5′-linker with a 4 nt barcode. Libraries were analyzed using an Illumina Genome Analyzer II at the UMass Medical School Deep Sequencing Core.

Bioinformatics Analysis

Sequences were processed and mapped to the genome using custom PERL (5.10.1) scripts, Bowtie 0.12.7 (Langmead et al., 2009) and BLASTN 2.2.25 (Altschul et al., 1990). For C. elegans experiments, reads were aligned to the C. elegans genome (WormBase release WS215), Repbase 15.10 (Jurka et al., 2005), and miRBase 16 (Kozomara and Griffiths-Jones, 2011). For mouse cloning experiments, reads were aligned to the mouse genome assembly NCBIM37 (Ensembl 67), miRBase 18 and the noncoding RNA database fRNAdb 3.4 (Mituyama et al., 2009). The Generic Genome Browser GBrowse 1.70, (Stein et al., 2002) was used to visualize the alignments. Detailed PERL scripts and related database files and analyses used in this study are available upon request.

(1) Mapping CapSeq reads to the C. elegans genome. Single-end reads of 75 nt (L10831, L30831 and YA0831) or 100 nt (avr0217 and ya0217) were obtained using the Genome Analyzer II and Hi-Seq platforms, respectively. A custom PERL script was used to remove the 5′ barcode and 3′ linker sequences. To remove the 3′ linker, each read was queried for the presence of the linker sequence TCGTATGCC. If the query sequence was found anywhere in the read, the entire 3′ part of the read beginning at the query sequence was removed. For reads where the query sequence was not found, the 3′ end of each read was queried sequentially to remove one of the following incomplete 3′ linkers: TCGTATGC, TCGTATG, TCGTAT, TCGTA, TCGT, TCG or TC. Reads without complete or partial 3 linker sequence were automatically included in the following analysis. To eliminate potential mutations introduced by random priming during cDNA synthesis, the last 8 nt of the read was trimmed after removal of the 3′ linker.

A custom PERL script was used to identify and remove spliced leader sequences (SL1–SL12) allowing up to 2 mutations. The processed reads were then mapped to the *C. elegans* genome WS215 and annotations with Bowtie 0.12.7 using the parameter “-n 2 -e

180 -a –best –strata -m 200''. The ''-e 180'' parameter indicates that up to 6 mutations were allowed; the ''-a –best –strata'' parameter only returns the best matches; and ''-m 200'' only reports RNA read with less than 200 best matches. The mutation rate allowed for the alignment was 0 for reads 17–19 nt, 1 for 19–21 nt, 2 for 22–24 nt, 3 for 25–49 nt, 4 for 50–74 nt, and 5–6 for 75 nt or bigger. Unmatched reads of at least 40 nt were mapped to the WS215 genome using BLASTN with the parameter 'e 0.01'. For reads that partially match the genome, such as exon junctions, a custom PERL script was used to search within 1000 nt of the mapped part to identify sequences that match the unmapped part of the reads according to splice junction rule. A custom PERL script was used to normalize the read counts to 10 million total reads mapped to the sense strand of protein coding genes, to combine the mapped reads according to their 5′ end genomic positions, and to generate a gff2 file for GBrowse display.

(2) Mapping Small RNA and RNA-Seq to *C. elegans*. A custom PERL script was used to remove the barcode and 3′ linker from small RNA reads. To remove the 3′ linker, each read was queried for the presence of the sequence CTGTAG. If the query sequence was found anywhere in the read, the entire 3′ part of the read beginning at the query sequence was removed. For reads where the query sequence was not found, the 3′ end of each read was queried sequentially to remove one of the following incomplete 3′ linkers: CTGTA, CTGT, CTG, and CT. Only reads of at least 17 nt were mapped to the genome and annotations.

The processed reads were mapped to the C. elegans genome WS215 and annotations using Bowtie 0.12.7 with a parameter ''-v 3 -a –best –strata -m 400''. A custom PERL pipeline was created to perform the postmatch analyses. The mutation rates allowed were: 0 mismatch for size 17–18 nt, 1 for 19–21 nt, 2 for 22–24 nt, and 3 for longer than 24 nt. If a RNA read mapped to multiple genomic loci, then the read number was split evenly among these loci. To account for differences in sequence volume between samples, the small RNA reads were normalized to 5 million nonstructural reads mapped. A custom PERL script was used to draw the scatter plot in Figure 4C. A custom PERL script was used to combine the mapped reads according to their 5′ ends and to generate a gff2 file for GBrowse display.

C. elegans RNA-seq data were obtained from GEO GSE22410 (Lamm et al., 2011), processed and mapped as described above. The RNA-seq (mRNA) data, however, was normalized to 10 million sense protein coding reads mapped.

(3) Analysis of C. elegans CapSeq quality − enrichment for capped RNA. To estimate the enrichment of CapSeq reads mapped to the very 5′ end of transcripts relative to reads that mapped internally (potentially noncapped but cloned anyway), we performed two analyses.

In analysis 1, normalized reads that perfectly match the first 30 nt of *alh-8*, *rps-24*, *rps-3*, *rps-29*, and *rps-15* were identified, including those with a complete or partial SL1 spliced leader. Because a mutation/insertion/deletion is allowed during barcode removal, the first nucleotide of SL1 could be removed. Therefore, reads that lack the first nucleotide of SL1 were not included in the analysis. We also required that at least the last 3 nt of SL1 be present because a partial SL1 sequence of 1-2nt size could be introduced by sequencing error. As such, reads containing SL1 sequence starting at positions 3–20 of SL1 (counting from the last nt of SL1) were used to calculate the degradation rate for each position of SL1, which was ~1/15297.

In analysis 2, we collected all of the sense reads that mapped to 6967 of the annotated *trans*-spliced protein-coding genes—our data confirm that these genes are *trans*-spliced. The number of SL-containing reads (T-SL) was compared to the number of non-SL-containing reads (T-non-SL), which could be noncapped RNA or non-*trans*-spliced capped RNA reads. Assuming that all of the non-SL-containing reads are derived from degraded RNAs, the relative enrichment of capped reads is equal to (T-SL/T-non-SL)*1500, where ''1500'' is the average gene size in bp. Consistent with analysis 1, full-length SL-containing reads were enriched ~17,000-fold over any other position. Importantly, this analysis underestimates the enrichment of capped 5′ ends over 5′ truncations because some of the non-SL-containing reads are actually capped RNAs but not *trans*-spliced.

(4) Mapping SL-containing reads to protein coding genes. All five CapSeq samples were used in this analysis. The SL sequence was removed from SL-containing reads, and the resulting reads of at least 20 nt that uniquely and perfectly matched the genome were used in this analysis. Sites (mapped 5′ ends) were removed if they were at least 10-fold less than the upstream or downstream neighbor because these sites could be generated by sequencing errors of the nearby abundant reads. RNA reads were first normalized to 10 million sense protein coding reads, and a histogram for the start sites of mapped reads was generated for each sample. The five histograms were combined and sites with less than 1 read were removed.

To associate an SL-containing read with a gene, a custom PERL script searched within the annotated genes. If failing, the script then searched for the nearest transcripts within 500nt downstream the mapped read (relative to the 5′ end). If still failing, the script searched within 500 nt upstream of the mapped read. If still failing, the script labeled this read as ''NA'', meaning it can not assign the read to any gene. If the read can be assigned to a gene, the script searched where the read was mapped relative to the gene, as shown in Table S1A, column 7.

(5) Identification of Pol II start sites for protein coding genes using CapSeq. Only uniquely mapped, 5′ perfectly matched, non-SL-containing CapSeq reads of at least 30 nt were included in this analysis. RNA reads were first normalized to 10 million sense protein coding reads, and a histogram for the 5′ ends of mapped reads was generated for each sample. The histograms were combined for all five CapSeq samples, and sites with less than 5 reads were removed. Sites (mapped 5′ ends) were also removed if they were at least 10-fold less than the upstream or downstream neighbor because these sites could be generated by sequencing errors of the nearby abundant reads.

In the sample YA0831, 944 genes each had 1000 or more reads out of ~7 million of sense protein coding reads, and therefore were defined as highly expressed genes. Reads mapped over 100 nt downstream the annotated 5′ ends of these genes were removed, because the 5-read cutoff could include degradation products from highly expressed genes. By this way, we can minimize such noise while at the same time identified the TS sites for these genes.

A PERL script searched an interval from 200 nt upstream to 1,000 nt downstream of a given read for the nearest gene. The final output is the relative position between the CapSeq 5′ end and the start site of the gene assigned as above, with negative number indicating that the CapSeq 5′ end is upstream of the gene.

To analyze the motif around the start site, nt occurrence within a window of −50 (upstream) to +50 around all start sites was summarized. The weight for each start site was set up as 1, regardless of the number of the mapped reads. Otherwise, if using read number, the motif analysis could be biased for the highly expressed sites.

(6) Identification of csRNAs upstream of protein coding genes. Sense csRNAs and anti-sense csRNAs were defined separately using the CIP-TAP sample. To remove background noise, reads mapped sense to 21U-RNAs, sense/anti-sense to tRNAs/rRNAs/miRNAs, anti-sense to other structural RNAs such as snRNAs/snoRNAs, and anti-sense to protein coding genes/pseudogenes were removed. Only uniquely matched reads outside the two chromosome IV regions containing canonical 21U-RNAs were considered. In addition, the considered reads were enriched at least 10-fold in the CIP-TAP sample over the CIP-PNK sample, and were present at 2 reads or more per 5 million non-structural reads in the CIP-TAP sample. A script searched for sense csRNAs within a 500 nt interval upstream of any SL-containing transcripts, or from 50nt upstream to 50 nt downstream of any non-SL-containing transcripts. In either case above, the script stops searching if it reaches an upstream gene in the same direction.

To identify the anti-sense csRNAs, a script scans a 200 nt interval from 100–300 nt upstream of every sense csRNA. The script stops scanning if it reaches an upstream gene.

(7) Analysis of the motif and size of 22G-RNA. In this analysis the frequency of each nucleotide at positions flanking the 5′ ends (+1) of anti-sense 22G-RNA reads, from −50 nt (upstream) to +50 nt (downstream), was summarized. Only reads with the 1st nt perfectly matched and at least 1 read per million nonstructural reads in CIP-PNK sample were analyzed.

(8) Analysis of the Expression Levels of csRNAs and lcRNAs at TS Sites. Sense csRNA read levels (CIP-TAP) were compared to long-capped RNA read levels (CapSeq) at the same TS sites for protein coding genes. The 4,344 sites with both csRNAs and long-capped RNAs were used to draw a scatter plot as in Figure S1B.

(9) The Ratio of trans-Spliced Genes in the *C. elegans* Genome. The number of SL-containing genes and the number of non-SL-containing genes were obtained from analyses 4 and 5. CapSeq reads mapped to the 5′ ends of 14,337 genes including 10,276 genes with SL-containing 5′ end reads, 10,679 genes with non-SL-containing 5′ end reads, and 6,618 genes with both.

(10) Prediction of the pri-miRNA start sites. All miRNA loci were visually inspected using GBrowse, and 66 regions (typically within 500 nt upstream of individual miRNAs or miRNA clusters) that contained CapSeq and/or csRNA reads, were identified. Only uniquely mapped reads with no mutations at the first nucleotide were used. A combined histogram of CapSeq start sites was generated and sites were removed if they were present at less than 1 read per 10 million sense protein coding reads. A histogram of CIP-TAP 5′ ends was also generated, and sites with less than 1 read per 5 million non-structural reads were removed. For each miRNA, the sites with the most reads and a YR motif (R, the first nt of the mapped reads) was assigned as the major TS site. If several major sites with the same read number were identified for a gene from the same histogram, the nearest site was selected. After obtaining both major start sites for each pri-miRNA, one from CapSeq and the other from CIP-TAP sample, the distance between the two sites was analyzed.

(11) Comparison of CapSeq and RNA-Seq for miRNA Loci. In this comparison, we randomly chose 5 non-SL-containing protein coding genes with size of 2,156, 1,573, 2,560, 1,382, and 860 nt for *kin-31*, *glc-1*, *rpl-12*, *glh-1*, and *cel-1*, and 5 miRNAs, each of which represent the 5′ most miRNA, if in the miRNA clusters, with genomic size of 897, 941, 580, 548, and 402 nt for *mir-35*, *mir-61*, *mir-54*, *mir-58*, and *mir-229* (see Table S4 for the gene coordinates). The above strategy eliminates the complexity caused by trans-splicing and dramatically reduces the size difference between individual mRNAs and miRNAs. Using CapSeq and CIP-TAP data visualized by GBrowse, we expanded the coding region of each miRNA or mRNA, as annotated in WS215, to include the upstream capped RNA region. Only uniquely mapped reads of size ≥30 nt that fell within the desired region were included in the analysis. 11 RNA-seq samples (Lamm et al., 2011) were combined and so were all 5 CapSeq samples. In this way, both RNA-seq and CapSeq were considered as mixed stage samples because they both contain L1, L3 and YA samples. CapSeq reads basically represent capped RNAs, and RNA-seq reads mostly represent noncapped RNAs. Assuming there is no difference between RNA enrichment methods used in CapSeq (enzymatic enrichment of non-rRNAs) and RNA-seq (polyA selection), the capped RNAs (CapSeq) and noncapped RNAs (RNA-seq) should correlate positively, regardless of the targets, i.e., miRNAs or mRNAs. However, miRNAs as a whole were almost completely depleted from RNA-seq, as compared to mRNAs, while in CapSeq, we can detected similar levels of miRNA and mRNA. This observation strongly suggested that the enrichment method resulted in a representation bias.

(12) Definition of Unique 21U-RNAs. A custom PERL script was used to identify all annotated 21U-RNAs that mapped to unique genomic loci, i.e., unique sequence and no overlapping 21U-RNAs. A genomic locus is defined by chromosome, strand and start position. A total of 9,079 such 21U-RNA loci out of 15,073 annotated loci were identified in WS215.

(13) Identification of CapSeq Reads Overlapping with 21U-RNAs. Uniquely matched CapSeq reads (YA0831, young adult) without a mutation at the 1st nt were overlapped with the unique 21U-RNAs. Only CapSeq reads that overlap with a single 21U-RNA on the same strand were considered. The position of the 5′ end of a CapSeq read was determined with respect to the 5′ end (+1) of the 21U-RNA; a negative number indicates the 5′ end is upstream of the mature 21U-RNA. For the CapSeq reads starting 2 nt upstream of annotated 21U-RNAs (44 loci), we asked whether another gene annotation existed within an interval from 20 nt upstream to 200 nt downstream of the CapSeq reads.

(14) Analysis of csRNAs that Overlap with 21U-RNAs. To correlate 21U-RNA level with csRNA level, the CIP-PNK reads mapped to the 5′ ends of unique 21U-RNAs and CIP-TAP reads mapped 2 nt upstream of the same 5′ ends were compared. In order to draw a figure using log2 scale, only 21U-RNAs with at least 1 reads in either the CIP-PNK or CIP-TAP sample was considered. If one sample had at least one read and the other had less than 1 read, we assigned the latter sample to 1, thus avoiding negative number when using log2 scale.

(15) Identification of 21U-like RNA Loci. To identify loci similar to 21U-RNAs but failing to generate 21U-RNAs, we used the CIP-TAP sample with restraints either from inside or outside. This analysis included the reads that mapped uniquely within one of the two intervals on chromosome IV between 4,500,000–7,000,000 and 13,500,000–17,200,000. The inside filter removed sense RNA reads mapped to annotated 21U-RNA loci, antisense and sense reads mapped to miRNA/tRNA/rRNA loci, and antisense reads mapped to protein coding genes and pseudogenes. The outside restraints came from PRG-1 IP and CIP-PNK. Reads enriched at least 5-fold in the PRG-1 IP over the input sample were considered as a 21U-RNA candidate and the corresponding csRNA start sites at −2 was removed from the CIP-TAP sample. This way we removed the 21U-RNAs not annotated in WormBase because these 21U-RNAs did not meet the minimum read requirement for 21U-RNA identification and/or motif score previously (Ruby et al., 2006 and Batista et al., 2008). The CIP-PNK sample served as control for CIP-TAP, and we only considered in this analysis the CIP-TAP RNA reads (start sites) present at least 1 read / million of nonstructural reads and enriched > 10-fold as compared to CIP-PNK sample. This way we can get csRNAs with high confidence. Moreover, the CIP-PNK sample was used to remove CIP-TAP sites near which there were at least two other sites mapped in CIPPNK, because most 21U-RNAs are either well separated from each other or closely overlapped (spacing within 5 nt). We speculated 21U-like loci could have the same characteristics. The nearby regions searched above in CIP-PNK sample were the intervals of position −25 to −5 for upstream and +6 to 26 for downstream, relative to the start sites of the mapped reads in CIP-TAP sample. This way we can minimize the noise from 22G-RNA contamination in CIP-TAP sample. The motif and size analysis was performed using the ~2,300 loci obtained this way, and the position in the motif analysis was referred to the start sites of csRNAs.

(16) Identification of 21U-RNAs in Wild Isolates. A custom PERL script searched 21U-RNA candidates cloned in the wild isolates using criteria: 1) the 21U-RNA locus had a U at +1 position in the wild isolates but other nucleotides in wild-type N2, as annotated in WS215; 2) the 21U-RNA locus has a YR motif; 3) 21U-RNA had at least 5 reads after normalization to 5 million nonstructural reads, and mapped to a unique locus; 4) the 21U-RNA reads mapped to chromosome IV within the two intervals of 4,500,000–7,000,000 and 13,500,000–17,200,000.

(17) Identification of 21U-RNAs. 21U-RNAs were defined using PRG-1 IP with the inside and outside restraints. The inside restraints removed reads mapped to tRNAs or rRNAs, RNAs of non-20/21-U, and RNAs with mutation at the 1st position or with less than 1 read per 5 million nonstructural reads in PRG-1 IP. The outside restraints removed the reads without a YR motif in which Y is 3 nt upstream the start site of a read, and/or the reads enriched less than 5-fold, as compared to the input sample. csRNAs, loci enriched at least 10-fold in the CIP-TAP sample over CIP-PNK sample, were used to overlap with the 21U-RNA loci.

(18) Analysis of the Percentage of type 2 21U-RNAs. 21U-RNAs in this analysis were not the same as those defined above because those new 21U-RNAs could contain many canonical 21U-RNAs missed in the published list due to insufficient sequencing depth or other reasons. In this analysis, each small RNA sample was normalized to 5 million nonstructural reads. U-21nt RNA loci, starting with U and 21 nt long, was obtained as the RNA reads uniquely matched and enriched at least 5-fold in PRG-1 IP sample, as compared to the input sample. rRNAs and tRNAs were removed beforehand. These RNA loci are divided into 4 groups: 1) mapped within the two regions of 4,500,000–7,000,000 and 13,500,000–17,200,000 on chromosome IV and annotated as 21U-RNAs already; 2) mapped within the two regions above, but not annotated as 21U-RNAs; 3) mapped to other regions on chromosome IV or other chromosomes, but annotated as 21U-RNAs; 4) mapped to other regions on chromosome IV or other chromosomes, but not annotated as 21U-RNAs. Group 2 and group 4 U-21nt RNAs must have YR motif. Group 1 represented type 1 21U-RNAs, while group 4 represented type 2 21U-RNAs.

(19) Mouse CapSeq Analysis. Genome and annotations NCBIM37.67 were obtained from ftp.ensembl.org, CAGE data were obtained from http://fantom31p.gsc.riken.jp/cage/download/mm5/, noncoding RNA database fRNAdb v3.4 was obtained from http://www.ncrna.org/frnadb/, and miRNA database miRBase release 18 was obtained from miRBase. These annotations were mapped to mouse genome using a custom PERL script plus Bowtie. Bowtie was also used to map CapSeq reads of at least 19 nt long to the genome and annotations with parameters: "-n 2 -e 180 -a –best –strata -m 200". A more stringent mutation filter was used to exclude matches with mutations: 0 maximum mismatch for reads 19–24 nt long, 1 for 25–29 nt, 2 for 30–39 nt, 3 for 40–49 nt, 4 for 50–59 nt, 5 for 60–69, and 6 for ≥70 nt. A custom PERL script was used to obtain the start site histogram of the mapped reads, after normalization to 10 million nonstructural RNA reads. And all the alignments were visualized using GBrowse 1.70.

(20) Mapping Small RNAs to Mouse Genome. The mouse MILI IP data were obtained from GEO GSM475280 (Robine et al., 2009), and then reads of size ≥18 nt was mapped using Bowtie 0.12.7, with parameters "-n 2 -e 180 -a –best –strata –m 200". The mutation rate allowed was as described for mouse CapSeq analysis.

(21) Motif Analysis of Mouse CapSeq. Included in the analysis were RNA reads uniquely matched without any mismatch at the first position, after removal of reads mapped to tRNAs, rRNAs, snRNAs and snoRNAs. A start site histogram of mapped reads was made. The start sites included in the following analysis had at least 1 reads out of 1 million nonstructural reads, and there were no neighbor start sites with 10-fold more reads. The nucleotides around each start site (−50 to 50) was summarized to calculate the overall frequency table. To avoid bias for the start sites with abundant reads, the weight for each start site was set to 1. The $\log_2$ ratio of

foreground rate / background rate represented the relative enrichment for each nucleotide at each position. Here the background rate for A/G/C/T was based on the nucleotide frequency of all genomic sites 200 nt flanking each start site, and the foreground frequency for each nucleotide at each position was calculated using the −50 to 50 frequency table above.

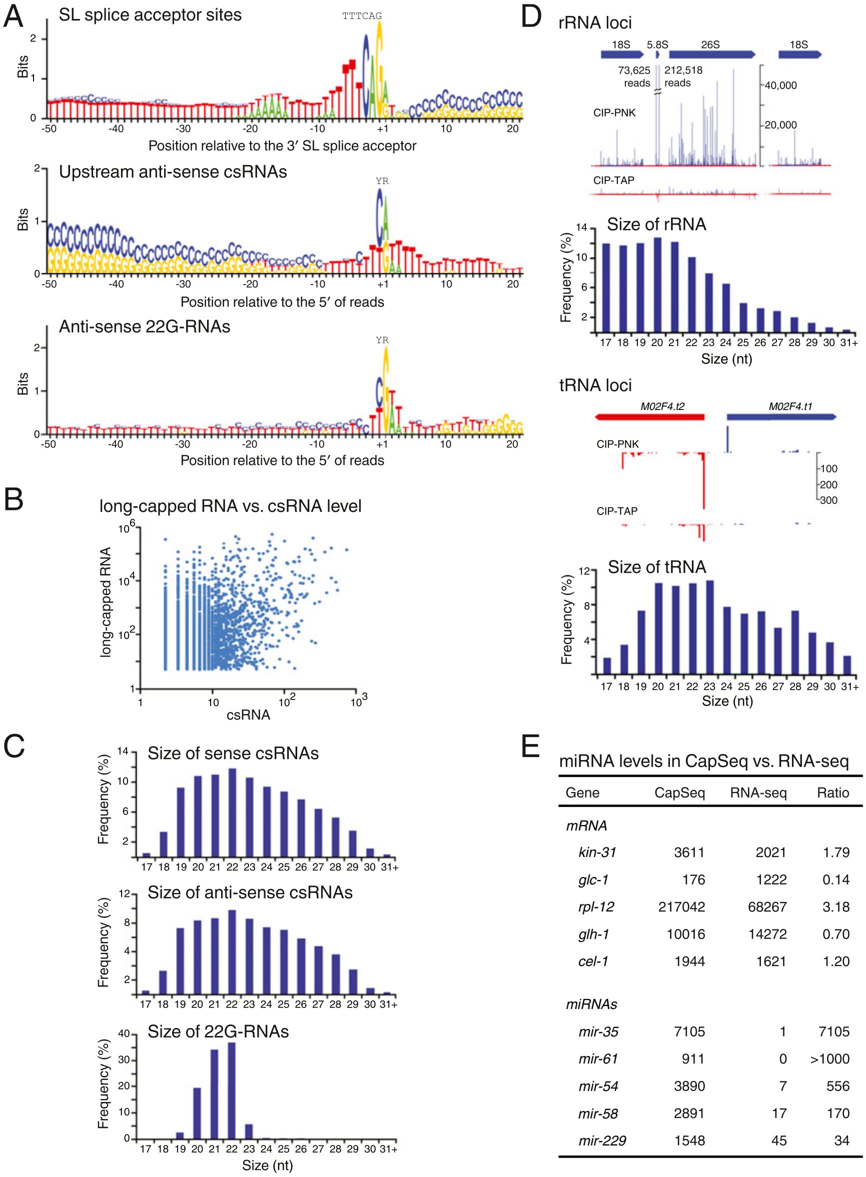
(22) Prediction of TS Sites for Mouse pri-miRNAs. Mouse testis samples at 4 weeks and 6 months were used in this analysis. Reads mapped to more than 3 genomic loci, or mapped with mutation at position +1, or mapped to snRNAs, snoRNAs, rRNAs, tRNAs, mRNAs and piRNAs were excluded. All reads were combined according to the start sites mapped, and then the start sites with less than 1 read per 5 million nonstructural RNA reads were excluded. Also excluded were the start sites with less than 1/10th reads compared to the neighbor start sites. The remaining start sites were assigned to the annotated pre-miRNAs as long as the start sites are within a 1 kb region upstream of the pre-miRNAs. Totally, 134 pre-miRNAs were assigned with TS sites.

(23) Analysis of the Level of csRNA or CapSeq RNA vs. that of 21U-RNA. The samples used were CapSeq YA0831 and CIP-TAP, both of which were made from young adult worms. 44 type 1 21U-RNA loci and 982 type 2 21U-RNA loci with −2 csRNA and −2 CapSeq reads were considered. A nonparametric correlation was used to assess the relationship between csRNA/CapSeq RNA and 21U-RNA.

SUPPLEMENTAL REFERENCES

Kato, M., de Lencastre, A., Pincus, Z., and Slack, F.J. (2009). Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. Genome Biol. *10*, R54.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.



miRNA levels in CapSeq vs. RNA-seq

Gene	CapSeq	RNA-seq	Ratio
mRNA			
kin-31	3611	2021	1.79
glc-1	176	1222	0.14
rpl-12	217042	68267	3.18
glh-1	10016	14272	0.70
cel-1	1944	1621	1.20
miRNAs			
mir-35	7105	1	7105
mir-61	911	0	>1000
mir-54	3890	7	556
mir-58	2891	17	170
mir-229	1548	45	34

Figure S1. Additional Capped RNA Analyses for mRNAs and miRNAs, Related to Figure 2

(A) Motif analyses of *trans*-splice sites (upper), upstream antisense csRNAs (middle) and antisense 22G-RNAs (lower). '+1' corresponds to either the first nt of the *trans*-splice acceptor site or the 5′ nt of small RNA read. The 22G-RNA analysis is to test if YR motif is also required for other RNA polymerase, like RNA-dependent RNA Polymerase (RdRP).

(B) Positive correlation of csRNA level and long-capped RNA level. The expression levels of csRNA and long-capped RNA derived from the same TS sites were compared.

(C) Size distribution of sense (upper) or antisense (middle) csRNAs and 22G-RNAs (lower). Again 22G-RNA analysis was used as an internal control.

(D) Small RNA reads mapped to rRNA (top) and tRNA (bottom). Genome browser views of CIP-PNK and CIP-TAP reads mapping to a rRNA repeat and a tRNA locus are shown. The size distribution of all rRNA or tRNA CIP-TAP reads is also shown. Sites (histogram bars) colored the way as the genes are sense-oriented. Otherwise, sites are antisense-oriented.

(E) Comparison of CapSeq and published RNA-seq reads that correspond to miRNA loci (Lamm et al., 2011). Each annotated gene locus was extended upstream to include the capped RNA loci, and then reads from CapSeq or RNA-seq were mapped to each gene locus, as shown in column 2 and 3 respectively. The relative ratio CapSeq/RNA-seq for each gene is shown in column 4.

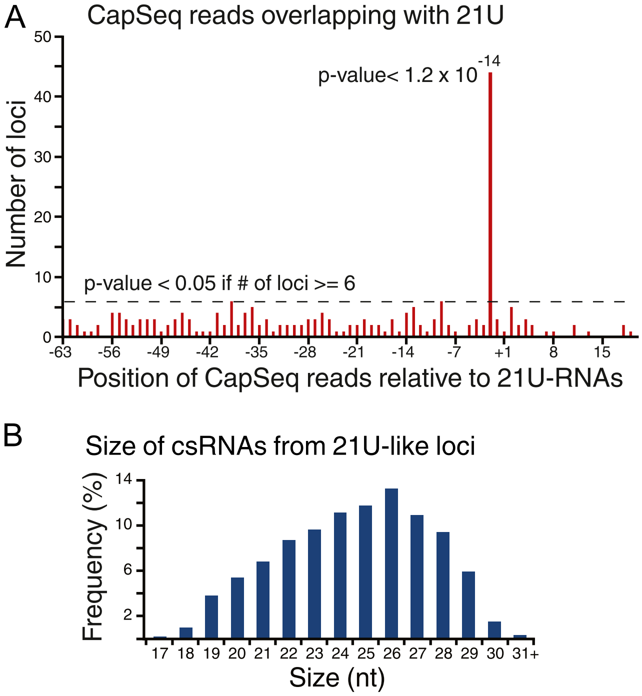
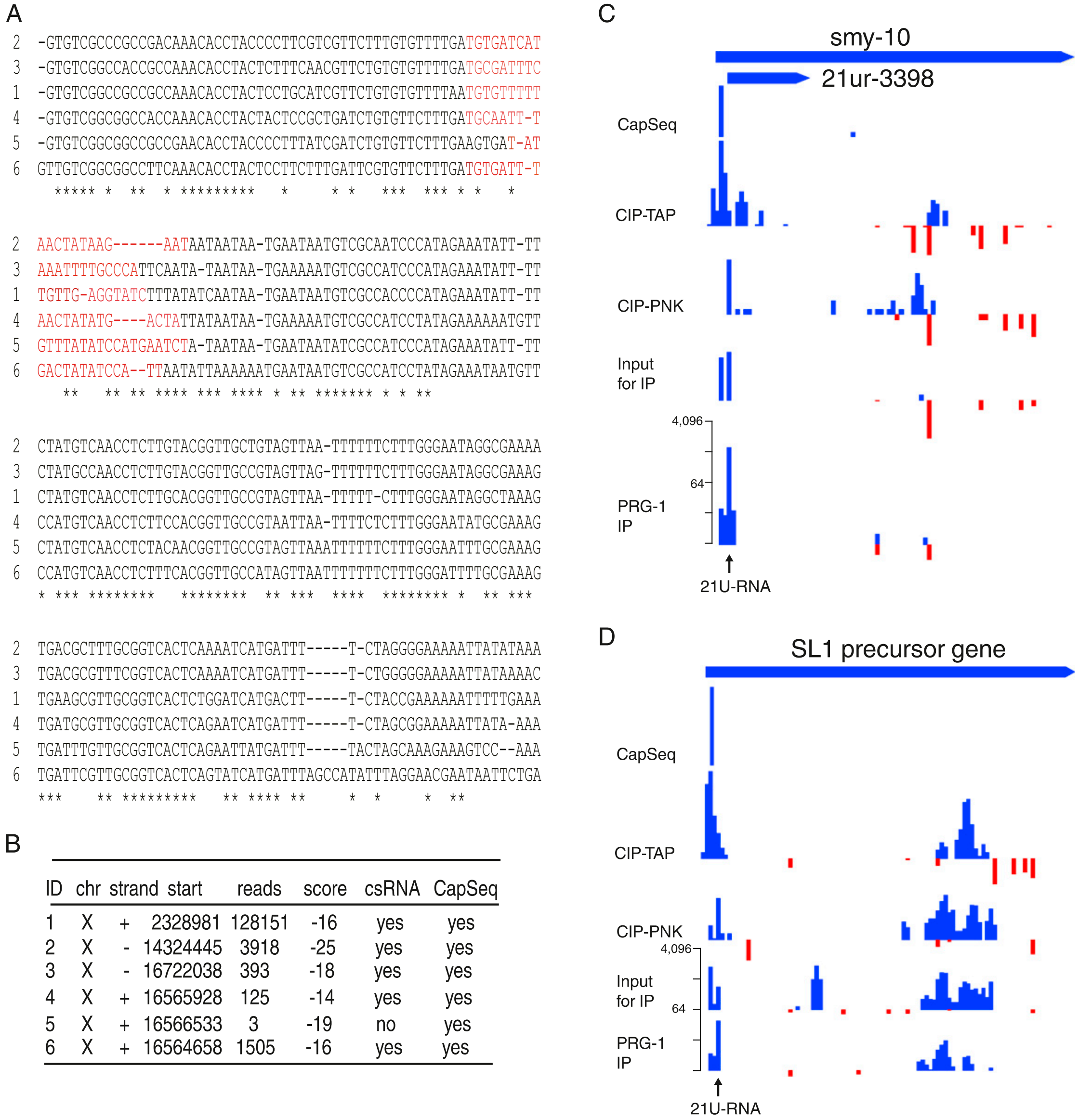


Figure S2. Additional Analysis of 21U-RNA Loci, Related to Figures 4 and 5

(A) Long-capped RNA loci (CapSeq) preferentially map 2nt upstream mature 21U-RNAs. Long-capped RNA reads that overlap with mature annotated 21U-RNAs were identified, and the relative distance between the 5′ end of the long-capped RNA and mature 21U-RNA (*x* axis) was plotted against the number of such cases. The negative number indicates that the long-capped RNA starts upstream of the mature 21U-RNAs. The p value was obtained using binomial distribution.

(B) Size distribution of csRNAs derived from 21U-like loci.



```
2   -GTGTCGCCCGCCGACAAACACCTACCCCTTCGTCGTTCTTTGTGTTTTGATGTGATCAT
3   -GTGTCGGCCACCGCCAAACACCTACTCTTTCAACGTTCTGTGTGTTTTGATGCGATTTC
1   -GTGTCGGCCGCCGCCAAACACCTACTCCTGCATCGTTCTGTGTGTTTTAATGTGTTTTT
4   -GTGTCGGCGGCCACCAAACACCTACTACTCCGCTGATCTGTGTTCTTTGATGCAATT-T
5   -GTGTCGGCGGCCGCCGAACACCTACCCCTTTATCGATCTGTGTTCTTTGAAGTGAT-AT
6   GTTGTCGGCGGCCTTCAAACACCTACTCCTTCTTTGATTCGTGTTCTTTGATGTGATT-T
      ***** *  **  * *********   *     * *   ***  *** * *   *

2   AACTATAAG------AATAATAATAA-TGAATAATGTCGCAATCCCATAGAAATATT-TT
3   AAATTTTGCCCATTCAATA-TAATAA-TGAAAAATGTCGCCATCCCATAGAAATATT-TT
1   TGTTG-AGGTATCTTTATATCAATAA-TGAATAATGTCGCCACCCCATAGAAATATT-TT
4   AACTATATG----ACTATTATAATAA-TGAAAAATGTCGCCATCCTATAGAAAAAATGTT
5   GTTTATATCCATGAATCTA-TAATAA-TGAATAATATCGCCATCCCATAGAAATATT-TT
6   GACTATATCCA--TTAATATTAAAAAATGAATAATGTCGCCATCCTATAGAAATAATGTT
       **   ** ** **** *** **** * ** ******* * * **

2   CTATGTCAACCTCTTGTACGGTTGCTGTAGTTAA-TTTTTTCTTTGGGAATAGGCGAAAA
3   CTATGCCAACCTCTTGTACGGTTGCCGTAGTTAG-TTTTTTCTTTGGGAATAGGCGAAAG
1   CTATGTCAACCTCTTGCACGGTTGCCGTAGTTAA-TTTTT-CTTTGGGAATAGGCTAAAG
4   CCATGTCAACCTCTTCCACGGTTGCCGTAATTAA-TTTTCTCTTTGGGAATATGCGAAAG
5   CTATGTCAACCTCTACAACGGTTGCCGTAGTTAAATTTTTTCTTTGGGAATTTGCGAAAG
6   CCATGTCAACCTCTTTCACGGTTGCCATAGTTAATTTTTTTCTTTGGGATTTTGCGAAAG
    * *** ********   ********  ** ***  ****  ******** *  ** ***

2   TGACGCTTTGCGGTCACTCAAAATCATGATTT-----T-CTAGGGGAAAAATTATATAAA
3   TGACGCGTTTCGGTCACTCAAAATCATGATTT-----T-CTGGGGGAAAAATTATAAAAC
1   TGAAGCGTTGCGGTCACTCTGGATCATGACTT-----T-CTACCGAAAAAAATTTTTGAAA
4   TGATGCGTTGCGGTCACTCAGAATCATGATTT-----T-CTAGCGGAAAAATTATA-AAA
5   TGATTTGTTGCGGTCACTCAGAATTATGATTT-----TACTAGCAAAGAAAGTCC--AAA
6   TGATTCGTTGCGGTCACTCAGTATCATGATTTAGCCATATTTAGGAACGAATAATTCTGA
    ***    ** *********   ** **** **     *  *     *  **
```

ID	chr	strand	start	reads	score	csRNA	CapSeq
1	X	+	2328981	128151	-16	yes	yes
2	X	-	14324445	3918	-25	yes	yes
3	X	-	16722038	393	-18	yes	yes
4	X	+	16565928	125	-14	yes	yes
5	X	+	16566533	3	-19	no	yes
6	X	+	16564658	1505	-16	yes	yes

Figure S3. Additional Analysis of Type 2 21U-RNA Loci, Related to Figure 6

(A) Alignment of six homologous 21U-RNA producing loci using CLUSTALW (Thompson et al., 1994) program. Colored are the 21U-RNA sequences obtained in PRG-1 IP. '*' indicates the identical positions. (B) 21U-RNA genomic loci with 'start' as the start site, 'reads' as the PRG-1 IP read number, and 'score' as 21U-RNA motif score. Column 7 and 8 are for existence of the −2 csRNA reads (CIP-TAP) and −2 long-capped RNA reads (CapSeq) respectively. (C) and (D) Type-2 21U-RNAs mapped at the 5′ end smy-10 gene (snRNA) and SL1 precursor gene, as indicated (black arrows) in the PRG-1 IP sample.