Sequence, biogenesis, and function of diverse small RNA classes bound to the Piwi family proteins of *Tetrahymena thermophila*

Mary T. Couvillion,[1] Suzanne R. Lee,[1] Brandon Hogstad,[1] Colin D. Malone,[2,3] Leath A. Tonkin,[4] Ravi Sachidanandam,[2,3] Gregory J. Hannon,[2,3] and Kathleen Collins[1,5]

[1]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA; [2]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; [3]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; [4]Vincent J. Coates Genomics Sequencing Laboratory, QB3/University of California at Berkeley, Berkeley, California 94720, USA

PAZ/PIWI domain (PPD) proteins carrying small RNAs (sRNAs) function in gene and genome regulation. The ciliate *Tetrahymena thermophila* encodes numerous PPD proteins exclusively of the Piwi clade. We show that the three *Tetrahymena* Piwi family proteins (Twis) preferentially expressed in growing cells differ in their genetic essentiality and subcellular localization. Affinity purification of all eight distinct Twi proteins revealed unique properties of their bound sRNAs. Deep sequencing of Twi-bound and total sRNAs in strains disrupted for various silencing machinery uncovered an unanticipated diversity of 23- to 24-nt sRNA classes in growing cells, each with distinct genetic requirements for accumulation. Altogether, Twis distinguish sRNAs derived from loci of pseudogene families, three types of DNA repeats, structured RNAs, and EST-supported loci with convergent or paralogous transcripts. Most surprisingly, Twi7 binds complementary strands of unequal length, while Twi10 binds a specific permutation of the guanosine-rich telomeric repeat. These studies greatly expand the structural and functional repertoire of endogenous sRNAs and RNPs.

[*Keywords*: Argonaute; Dicer; RNA silencing; RNA-dependent RNA polymerase; endogenous small RNA]

Supplemental material is available at http://www.genesdev.org.

Received May 15, 2009; revised version accepted July 17, 2009.

Eukaryotic ~20- to 30-nucleotide (nt) small RNAs (sRNAs) have an astounding diversity of sequence, developmental expression specificity, biogenesis mechanism, abundance, and turnover (Farazi et al. 2008; Ghildiyal and Zamore 2009; Kim et al. 2009). Despite several years of fast-paced discovery, appreciation for sRNA complexity remains far from complete. Classes of sRNA have been designated based on transcript origin, processing pathway, and/or protein partner. MicroRNAs (miRNAs) originate from an imperfectly base-paired hairpin precursor, which in the canonical biogenesis pathway is processed by sequential steps of dsRNA cleavage by enzymes of the RNase III family (Carthew and Sontheimer 2009; Kim et al. 2009). siRNAs originate from other forms of dsRNA, including products of endogenous RNA-dependent RNA polymerase (Rdr) multisubunit complexes (RDRCs) and regions of duplex proposed to arise by annealing within or between primary transcripts (Nilsen 2008; Okamura and Lai 2008; Ghildiyal and Zamore 2009). A typical siRNA biogenesis pathway involves cleavage by a Dicer enzyme, with different Dicer family members generating different sRNA size classes (Carthew and Sontheimer 2009; Jinek and Doudna 2009). Recent studies also reveal Dicer-independent pathways for production of *Caenorhabditis elegans* secondary siRNAs and siRNAs of *Entamoeba histolytica* (Aoki et al. 2007; Pak and Fire 2007; Sijen et al. 2007; Zhang et al. 2008). These classes possess a distinguishing 5′ polyphosphate, reflecting their origin as unprocessed primary transcripts of an Rdr subgroup specialized for short product synthesis.

A third category, Piwi-interacting RNAs (piRNAs), includes several high-complexity classes speculated to derive from processing of long, ssRNA precursors (Seto et al. 2007; Klattenhoff and Theurkauf 2008; Kim et al. 2009; Malone and Hannon 2009). The piRNAs of multicellular eukaryotes show developmentally restricted accumulation, detected predominantly in germline, germline-supportive, or self-renewing cell lineages. Where possible to discern, the piRNAs associated with a particular Piwi

[5]Corresponding author.
E-MAIL kcollins@berkeley.edu; FAX (510) 643-6791.
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.1821209.

family protein have an extreme bias of strand-specific accumulation. Subsets of partially complementary piRNAs are amplified by a cascade of reciprocal sense and antisense strand targeting, but mechanisms that give rise to the initial strand polarity of piRNA biogenesis are unknown.

Diverse sRNAs are unified by their association with PAZ/PIWI domain (PPD) proteins to form effector RNPs. Some PPD proteins retain the catalytic residues of their RNase H-like active site and possess a "slicer" cleavage activity that is important for sRNA maturation and/or effector RNP function. PPD proteins are classified into subfamilies, two of which are evolutionarily widespread (Cerutti and Casas-Mollano 2006; Höck and Meister 2008). The Argonaute (Ago) subfamily carries miRNAs and most siRNAs (Carthew and Sontheimer 2009; Ghildiyal and Zamore 2009; Kim et al. 2009). Ago RNPs function in transcriptional gene silencing (TGS) by histone and/or DNA modification and instigate mRNA decay and translational inhibition pathways of endogenous post-transcriptional gene silencing (PTGS) and exogenously induced RNAi. Roles for Ago RNPs in transcriptional and translational activation have also been described.

The second conserved PPD protein subfamily encompasses members of the Piwi clade (Seto et al. 2007; Klattenhoff and Theurkauf 2008; Malone and Hannon 2009). In animals, Piwi proteins function in particular stages of germline or stem cell lineage specification. Different Piwi protein family members expressed by the same organism can have distinct or overlapping expression with regard to developmental timing, cell type, and subcellular localization. Piwi RNPs have been shown to reduce transposon expression by mechanisms involving DNA modification and to contribute to epigenetic regulation of transposon mobility. Piwi family proteins are also expressed in some single-celled organisms, including the ciliated protozoans *Tetrahymena thermophila* and *Paramecium tetraurelia*. The *T. thermophila* genome harbors predicted ORFs for up to 12 PPD proteins exclusively of the Piwi clade (Cerutti and Casas-Mollano 2006; Seto et al. 2007). *T. thermophila* also encodes three Dicer family proteins (Dcl1, Dcr1, and Dcr2) and a single Rdr (Rdr1) that is assembled into multiple RDRCs (Malone et al. 2005; Mochizuki and Gorovsky 2005; Lee and Collins 2006, 2007; Lee et al. 2009).

T. thermophila reproduce asexually or sexually as two alternative phases of the ciliate life cycle (Chalker 2008). In rich media, cells in asexual or "vegetative" growth divide by fission, maintaining both the silent, diploid, germline micronucleus (MIC) and the expressed, polyploid, somatic macronucleus (MAC). If cells are starved in the company of another mating type, a sexual cycle of conjugation is initiated. Paired cells undergo MIC meiosis, gamete exchange and fusion, zygotic nuclear mitoses, and differentiation of new MICs and MACs. New MAC differentiation requires production of Twi1-bound 27- to 30-nt sRNAs termed scan RNAs (scnRNAs) by the conjugation-specific dicer Dcl1 (Mochizuki and Gorovsky 2004; Chalker 2008). Twi1 RNPs guide heterochromatin formation in the developing MAC, marking MIC-limited sequences for histone H3 Lys 9 (H3K9) methylation and subsequent elimination. As a result, the transcriptionally active mature MAC retains little repetitive DNA, no centromeres or transposons, and no H3K9-modified heterochromatin (Liu et al. 2004; Eisen et al. 2006).

Until recently, unicellular eukaryotes were thought to generate only sRNAs that guide heterochromatin formation. The first of many exceptions came with the discovery of a second class of *T. thermophila* sRNAs: In addition to the 27- to 30-nt scnRNAs that guide heterochromatin formation and DNA elimination in conjugating cells, a class of 23- to 24-nt sRNAs is constitutively expressed (Lee and Collins 2006). Limited sequencing of bacterially cloned *T. thermophila* 23- to 24-nt sRNAs revealed a strand-asymmetric accumulation of unphased sRNAs produced in clusters from unique loci, properties shortly thereafter reported for mammalian pachytene piRNAs (Seto et al. 2007). *T. thermophila* 23- to 24-nt sRNAs are also generated from long hairpin transgenes that post-transcriptionally silence cognate mRNA by RNAi (Howard-Till and Yao 2006). Two essential enzymes, Rdr1 and Dcr2, cooperate to produce 23- to 24-nt sRNAs in vitro and in vivo (Lee and Collins 2007; Lee et al. 2009). *T. thermophila* Rdr1 assembles several distinct RDRCs by mutually exclusive interaction of Rdr1 with one of two nucleotidyl transferases (Rdn1 or Rdn2) and by Rdr1–Rdn1 interaction with one of two additional factors (Rdf1 or Rdf2) (Supplemental Fig. S1 summarizes the developmental expression and associations of *T. thermophila* Dicer, RDRC, and Twi proteins). Each RDRC has a similar biochemical activity of processive dsRNA synthesis and a similar association with Dcr2 in vitro, but in vivo, knockout of the Rdf1, Rdf2, or Rdn2 subunit specific to a particular RDRC imposes distinct growth or conjugation phenotypes and has differential impact on accumulation of 23- to 24-nt sRNAs (Lee et al. 2009).

Considering the plethora of putative *T. thermophila* PPD proteins and the functional distinctions among RDRCs, we suspected that sRNA complexity was much greater than reflected by our previous sequencing. Here we use affinity purification, deep sequencing, and assays of genetic requirements for accumulation to discover and characterize an unanticipated diversity of 23- to 24-nt sRNA classes in growing cells. By analysis of individual Twi-bound sRNA populations and total sRNA populations in strains lacking RDRC components, we uncover multiple pathways of endogenous RNA silencing. We find that the expanded family of Piwi proteins recognizes and segregates sRNAs derived from silenced pseudogenes, silenced structured RNA transcripts, distinct categories of DNA repeats including telomeres, and expressed protein-coding genes. Some of the Twi-bound sRNA classes resemble endogenous sRNAs characterized in other organisms, while other sRNA classes are novel, likely to have escaped detection in other systems.

Results

Distinct developmental expression, localization, and function of Twi proteins

To investigate the expression specificity of *T. thermophila* Twi family members, we used Northern blot

hybridization and RT–PCR to detect putative mRNAs encoding each of the 12 *T. thermophila TWI* genes (Fig. 1A; additional data not shown). For intron-containing genes, we first cloned mRNAs to determine correct ORF sequences (see the Materials and Methods). Curiously, we found that *TWI8* undergoes alternative splicing, which is extremely rare in *T. thermophila* (Coyne et al. 2008). We then compared expression of each *TWI* gene in the alternative life cycles of vegetative growth and conjugation. *TWI1* and *TWI11* were undetectable in vegetative

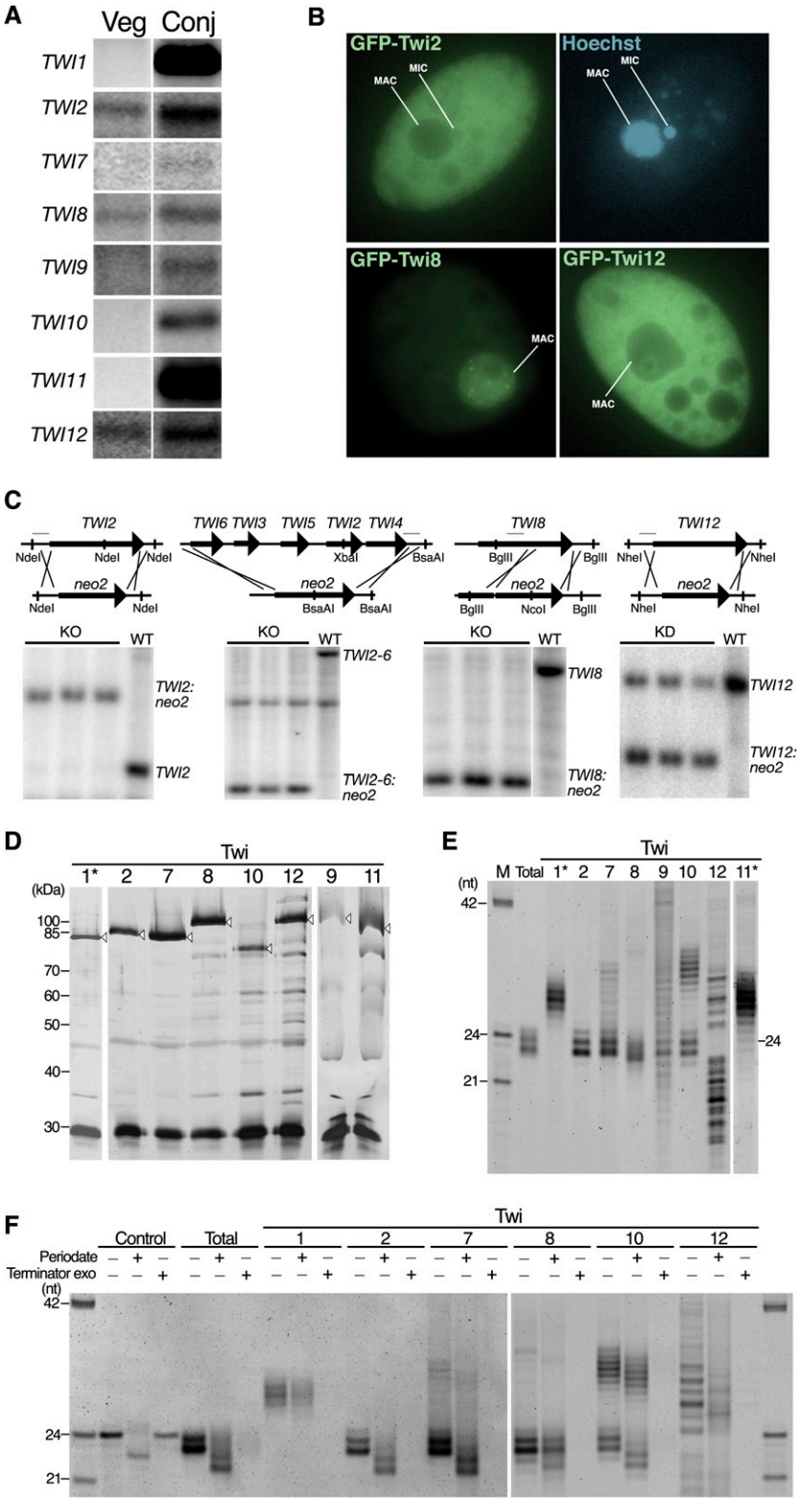

Figure 1. Twi protein expression, localization, function, and bound sRNAs. (*A*) Northern blots to detect each *TWI* mRNA using total RNA isolated from cells in vegetative growth (Veg) or 9 h after initiating conjugation (Conj). (*B*) Imaging of live cells that expressed the indicated GFP-Twi fusion protein and were stained with the membrane-permeable dye Hoechst to visualize nuclei. (*C*) Southern blots to assess locus disruption by the neo2 selectable marker cassette. Restriction enzymes used for genomic DNA digestion are indicated along with the region used for probe (the thin gray line *above* the wild-type [WT] locus). (*D*) Silver-stained SDS-PAGE gel showing proteins obtained by affinity purification. Open arrowheads indicate the Twi protein after elution by TEV protease. The ~30-kDa protein common to all lanes is recombinant TEV protease. Asterisk indicates purification from extract of conjugating cells harvested 6 h after conjugation initiation. (*E*) RNAs copurified with each Twi resolved by denaturing PAGE and stained with SYBR Gold. (M) Marker lane; (Total) gel-purified 23- to 24-nt sRNAs from strain CU522. Asterisk indicates purification from extract of conjugating cells harvested 4 or 10.5 h after conjugation initiation for Twi1 or Twi11, respectively. (*F*) End structure of sRNAs examined by β-elimination (Periodate) or 5′-monophosphate-dependent exonuclease treatment (Terminator exo). The control 24-nt RNA oligonucleotide has 2′- and 3′-hydroxyl groups but not a 5′ monophosphate.

growth, even by RT–PCR, but were dramatically induced in cells undergoing sexual reproduction. Previous studies have demonstrated *TWI1* expression throughout the time course of conjugation (Mochizuki et al. 2002; Miao et al. 2009), whereas *TWI11* expression was restricted to late conjugation stages after completion of MIC meiosis and zygotic nuclear mitoses. *TWI9* and *TWI10* were also conjugation-induced, although less dramatically than *TWI1* and *TWI11*. In contrast, *TWI2*, *TWI8*, and *TWI12* were robustly expressed in growing cells with mRNA levels readily detected by Northern blot hybridization. *TWI7* expression was also detected in growing cells using RT-PCR. Each of the Piwi family proteins expressed in vegetative growth was expressed at some level during conjugation as well. Notably, these expression profiling conclusions are fully consistent with analysis of primary microarray data from a recently published survey of gene expression over the *T. thermophila* life cycle (Miao et al. 2009).

The remaining putative genes encoding *TWI3* through *TWI6* are clustered in the genome as an uninterrupted tandem array surrounding *TWI2*. The high level of DNA sequence similarity between potential ORFs within the *TWI2-6* locus prevents discrimination of their putative mRNAs by Northern blot hybridization. From several lines of evidence, including sequencing of RT–PCR products and Northern blot assays using strains with disruption of *TWI2* versus the entire ~20-kb *TWI2-6* locus (see below), we infer that only *TWI2* is highly expressed. In the absence of *TWI2*, a transcript becomes detectable that is likely to derive from *TWI4* (Supplemental Fig. S2). Predicted proteins from *TWI2* and *TWI4* would share >90% identity, a greater level than shared by any other pairwise comparison of Twi proteins (Supplemental Fig. S3A). Overall, these findings establish an inventory of eight distinct *T. thermophila* Piwi family proteins in wild-type cells. Among these Piwi family proteins, only Twi1 harbors the active site residues necessary for "slicer" cleavage activity (Supplemental Fig. S3B).

To evaluate functional similarities or differences among the three Twi proteins robustly expressed in vegetative growth (Twi2, Twi8, and Twi12), we first examined their subcellular localization. Strains were created to express N-terminally GFP-tagged proteins from the cadmium-inducible *MTT1* promoter (Shang et al. 2002). Twi2 and Twi12 concentrated in the cytoplasm and were largely excluded from the MAC, while Twi8 adopted the inverse distribution (Fig. 1B). None of the three GFP-Twi proteins was detectably enriched in the transcriptionally silent MIC. These patterns did not change when cells were examined in growth versus nutrient starvation or at different times following cadmium induction (data not shown).

We also assessed whether the loci encoding Twi2, Twi8, and Twi12 are essential. Gene disruption was performed by targeting the endogenous ORF for replacement with the neo2 cassette encoding paromomycin resistance. Initial transformants harbor the drug resistance cassette in substitution of only a few of the 45 copies of each wild-type MAC chromosome, but increasing selective pressure over generations of amitotic MAC chromosome segregation yields complete replacement of nonessential genes. Strains targeted to disrupt a nonessential locus possess only recombinant chromosomes, while strains targeted to disrupt an essential locus retain some wild-type chromosome copies that can back-assort to increased copy number after release from selection. We found that all of the macronuclear copies of *TWI2*, the *TWI2-6* locus, or *TWI8* could be fully replaced by the drug resistance cassette, indicating that these genes are not essential for growth (Fig. 1C). In contrast, in multiple independent selections, *TWI12* was only partially replaced (Fig. 1C). *TWI12* is therefore essential for vegetative growth.

Comparing the three Twi proteins robustly expressed during vegetative growth, each can be functionally resolved from the others based on the combination of intracellular distribution and essentiality. Below, the gene knockout strains deleted for *TWI2*, *TWI2-6*, or *TWI8* (Twi2KO, Twi2-6KO, and Twi8KO) and the gene knockdown strain depleted for *TWI12* (Twi12KD) were used to investigate the cellular requirements for sRNA accumulation. We also used strains generated previously that are knockouts for the nonessential subunits of compositionally distinct RDRCs sharing essential Rdr1 and Dcr2 (Rdf1KO, Rdf2KO, and Rdn2KO) (see Supplemental Fig. S1).

Biochemical comparison of sRNAs bound to individual Twis

Ago and Piwi proteins are loaded with endogenous sRNAs by specificity principles that incorporate requirements for sRNA length, structure, 5′-nucleotide identity, and/or presence of 5′ polyphosphate (Jinek and Doudna 2009; Siomi and Siomi 2009). Pilot sequencing of bacterially cloned *T. thermophila* 23- to 24-nt sRNAs yielded insight about an abundant sRNA class with extreme strand asymmetry of accumulation and a strong bias for 5′ uridine (U). In that study, approximately half of the sRNAs were extended by a 3′ U that did not match the sequenced MAC genome (Lee and Collins 2006). Such untemplated addition of U or other ribonucleotides (A, C, G) has been linked in other organisms to increased or decreased turnover of specific sRNAs or entire sRNA classes (Li et al. 2005; Katoh et al. 2009). To investigate whether individual Twi proteins have inherent differences in their specificity of loading with endogenous sRNAs, and to characterize the potentially diverse classes of sRNA enriched by association with individual Twis, we created strains that would allow affinity purification of each of the eight distinct Twi proteins under parallel conditions for sRNA biochemical analysis and deep sequencing.

Transgenes were designed to express each Twi ORF fused to an N-terminal tag of tandem Protein A domains (ZZ) followed by a cleavage site for Tobacco Etch Virus (TEV) protease. Expression was placed under control of the cadmium-inducible *MTT1* promoter. Each transgene was integrated in complete replacement of the taxol-hypersensitive, nonessential β-tubulin 1 gene of strain CU522, an efficient strategy for positive selection (Gaertig

et al. 1994). Twi RNPs were purified by retention on IgG agarose and elution with TEV protease. We note that for Twi7 and especially for Twi9 and Twi10, the tagged Twi mRNA was expressed at a level substantially over endogenous mRNA in the growing cells used for RNP affinity purification. However, pilot experiments using different expression and purification conditions and in some cases using transgene strains with disruption of the endogenous *TWI* locus suggested that the level of Twi protein expression was not a primary determinant of sRNA loading specificity. Instead, the tagged Twis associated with largely distinct populations of sRNAs described in detail below.

Purified Twi RNP complexes were examined for protein recovery by SDS-PAGE and silver staining (Fig. 1D) and for RNA recovery by denaturing gel electrophoresis and SYBR Gold staining (Fig. 1E). Purification of Twi1 or Twi11 expressed during vegetative growth did not recover any associated sRNAs (data not shown). However, purification of either protein from conjugating cells enriched 27- to 30-nt sRNAs (Fig. 1E), as reported previously for Twi1 (Mochizuki et al. 2002). Thus, forced expression of Twi1 or Twi11 in vegetative growth is not sufficient to accomplish their loading with sRNAs, likely due to their specificity for binding sRNAs produced by the conjugation-specific Dcl1. Other than Twi1 and Twi11, each Twi expressed in vegetative growth did copurify sRNA, but each Twi enriched electrophoretically distinct sRNA populations (Fig. 1E). Twi2 and Twi8 bound predominantly 23- to 24-nt sRNAs that were offset from each other by an approximately half-nucleotide step of migration on gels with sufficient resolution. Twi7 and Twi10 copurified 23- to 24-nt sRNAs and also larger ~32- to 34-nt or ~33- to 36-nt RNAs, respectively. Twi9 copurified the most heterogeneously sized RNA, while Twi12 copurified a defined range of RNA lengths predominantly longer or shorter than 23–24 nt.

To characterize the Twi-bound sRNAs at a biochemical level, we compared the structures of their 5′ and 3′ ends. An aliquot of each Twi-enriched sRNA pool was subject to 3′ truncation by β-elimination, which requires both 2′- and 3′-hydroxyl groups. The piRNAs bound to Piwi family proteins, some Ago-bound siRNAs and miRNAs, and *T. thermophila* total 27- to 30-nt sRNAs are modified by ribose methylation at their 3′ ends, which prevents β-elimination (Chen 2007; Kurth and Mochizuki 2009). For plant sRNAs, methylation correlates with increased sRNA accumulation and protection from untemplated 3′-nucleotide addition in vivo (Li et al. 2005). We found that *T. thermophila* total 23- to 24-nt sRNAs and Twi-bound 23- to 24-nt sRNA pools were generally increased in mobility following β-elimination, with the notable exception of Twi8-bound sRNAs (Fig. 1F). The implied 3′-end modification of Twi8-bound sRNAs could account for their offset electrophoretic mobility in denaturing gels (Fig. 1E). Twi1-bound 27- to 30-nt sRNAs from conjugating cells were also resistant to β-elimination (Fig. 1F), as shown recently for total 27- to 30-nt sRNAs (Kurth and Mochizuki 2009).

We next examined the 5′-end structure of sRNAs by treatment with the 5′-monophosphate-dependent exonuclease Terminator. Each Twi-bound sRNA population was labile to nuclease degradation (Fig. 1F). This suggests that none of them harbor the 5′ polyphosphate characteristic of unprocessed RDRC initiation products such as *C. elegans* secondary siRNAs (Pak and Fire 2007; Sijen et al. 2007). *T. thermophila* RDRCs synthesize long dsRNA products in vitro, unlike the RDRC formed by the *C. elegans* RRF-1 responsible for secondary siRNA production (Aoki et al. 2007; Lee and Collins 2007). Although *T. thermophila* Dcr2 preferentially cleaves an Rdr1 product 5′ end to produce triphosphate-capped 23- to 24-nt sRNA in vitro, sRNA produced by Dcr2 in vivo could lack the 5′ triphosphate due to activity of a phosphatase such as PIR1 (Deshpande et al. 1999; Duchaine et al. 2006) or preferential Twi protein loading with internal dsRNA fragments.

Sequence characterization of 22- to 24-nt sRNAs

For deep sequencing, we gel-purified sRNAs associated with Twi2, Twi7, Twi8, Twi9, Twi10, and Twi12 in growing cells. We also gel-purified 22- to 25-nt RNA from size-enriched total RNA of the cognate background strain CU522. In addition, we gel-purified 22- to 25-nt RNA from size-enriched total RNA of the gene knockout strains individually lacking each nonessential RDRC subunit (Rdf1, Rdf2, or Rdn2) and the cognate background strain SB210. The closely related strains CU522 and SB210 differ in which of the known 23- to 24-nt sRNA loci contribute most abundantly to the total sRNA population (Lee and Collins 2006), but these strains derive from the same inbred parent and are not known or thought to differ substantially in MAC genome sequence.

High-throughput sequencing was performed with an Illumina 1G Genome Analyzer. For each library, we obtained between ~600,000 and ~6,000,000 sequences (Fig. 2A). Here we focus on analysis of sRNAs with 22–24 nt of genome-matching sequence, with or without the additional untemplated 3′ nucleotide observed previously (Lee and Collins 2006). This represents the sRNA size class generated in vivo by cooperation of genetically essential Dcr2 and Rdr1. RNAs with genome-matching lengths of 22–24 nt composed ~80,000 to ~3,500,000 of the sequences in each library (Fig. 2A). Nonmappers were more abundant in total RNA than Twi-bound sRNA, consistent with the presence of some fungal and bacterial RNAs ingested from the growth media in the total RNA population (Lee and Collins 2006). Additional nonmappers may correspond to incorrectly trimmed sequence reads, spliced exon junctions, base-modified RNAs, unassembled genome loci, or RNAs with more than one untemplated 3′ nucleotide.

Among sRNAs with 22–24 nt of genome-matching sequence (the 22- to 24-nt mappers), we analyzed the frequency and identity of untemplated 3′-nucleotide addition (Fig. 2B). Most sRNA sequences mapped to the genome only after allowing for the presence of an untemplated 3′ U. Exceptions to the typically high frequency of untemplated 3′-U addition were the Twi8 library, likely due to 3′-end modification of most of these sRNAs (Fig.

1F), and the Twi9 and Twi12 libraries, which had proportionally few 22- to 24-nt sRNAs. To evaluate the sequence complexity of each library, we calculated the percentage of total genome mapping 22- to 24-nt sRNAs, with or without untemplated 3′ uridylation, that represented a unique sequence inventory (Fig. 2C). About 20% of sequences from the CU522 total sRNA library were unique (an average of five reads per sequence). The Twi2 and Twi7 libraries had similar complexity, with greater complexity in the Twi8 library and particularly high complexity in the Twi10 and Twi12 libraries. Compared with the library from the wild-type SB210 background, there was relatively low complexity in the sRNA library from the Rdf2KO strain. This matches expectation from previous work, which demonstrated that Rdf2KO cells have reduced accumulation of 23- to 24-nt sRNAs (Lee et al. 2009).

We examined the overall sequence specificity of sRNA association with each Twi by compiling nucleotide frequencies for each sRNA position from the 5′ end, using sRNA sequences that match the genome completely (Fig. 2D; note that some sequences may contain an untemplated U that matches the genome; see Supplemental Fig. S4 for analysis of sRNAs with a definitively untemplated 3′ U). In all libraries except Twi9 and Twi12, which had proportionally few 22- to 24-nt sRNAs, a strong bias to 5′ U was observed. This 5′ U bias was not an inherent bias of nucleotide composition for a sRNA-generating strand at any sRNA locus examined (data not shown). Curiously, the 5′ U bias was accompanied by a strong second position bias against U (Fig. 2D). The 5′ U bias is a common feature of *T. thermophila* sRNAs and many classes of siRNA, miRNA, and piRNA in other eukaryotes. An overall sequence composition bias for A/U is observed in all libraries except that of Twi9, reflective of the high A/T content of *T. thermophila* genomic DNA. The generally similar sequence composition of various Twi-bound sRNA pools suggests that Twi loading specificity does not depend on post-biogenesis sorting of sRNA by sequence features alone.

Below we describe the features of distinct Twi-bound sRNA classes that are produced from different types of genomic loci, each with different genetic requirements for accumulation. To comprehensively examine *T. thermophila* sRNA classes, we scanned output plots of sRNA density across the genome for peaks of abundant sRNAs using a cluster window of 10 kb. To detect classes of differential abundance, we performed this analysis independently for total sRNAs from wild-type strains and from each RDRC subunit knockout strain, as well as for the sRNA pools enriched by association with each Twi. sRNA clusters identified and subsequently analyzed are listed in Supplemental Table S1. Ultimately, we identified one or

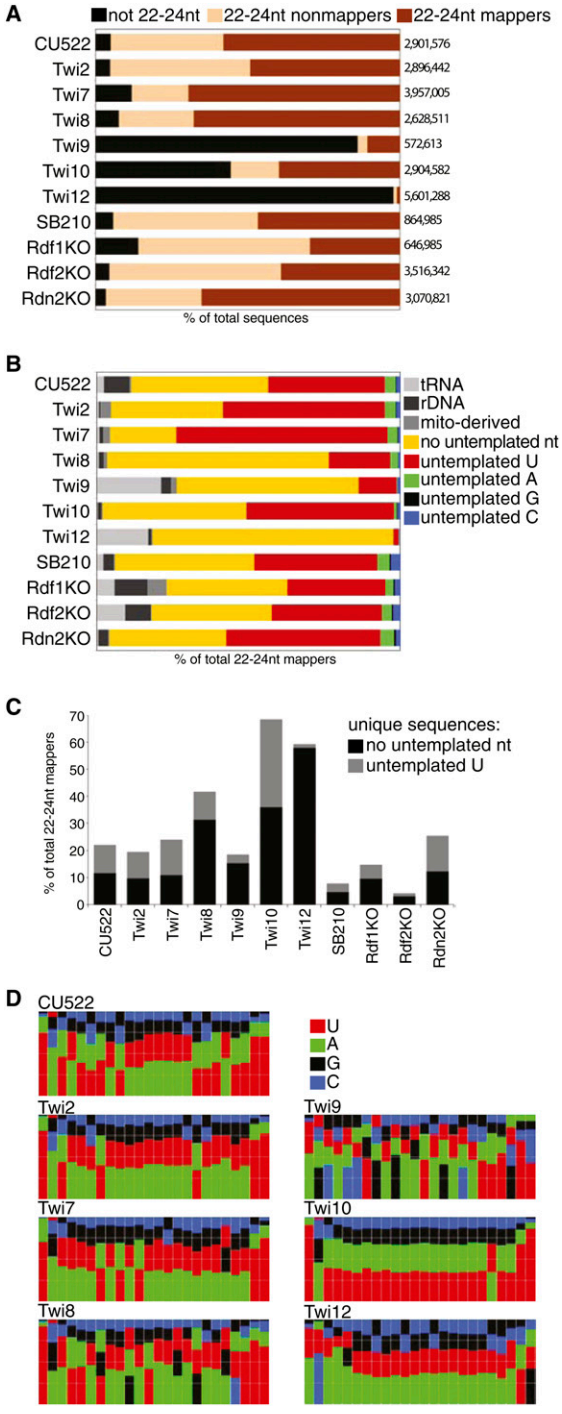

Figure 2. Deep sequencing library characteristics. (*A*) Proportions of sequence types in each unfiltered library. The number of total sequences from each library is indicated at *right*. (*B*) Proportions of 22- to 24-nt sequences in each library that mapped to the genome allowing a single 3′-nucleotide mismatch. Sequences indicated as rDNA were derived from the chromosome encoding the large ribosomal RNA precursor. Sequences indicated as mito-derived were from the mitochondrial genome. Further analysis was carried out on sequences represented by the yellow and red bars. (*C*) Fraction of distinct sequences (the inventory of unique sequences) calculated separately for genome-matching or definitively 3′-U-extended sRNAs. (*D*) Nucleotide frequency plots for genome-matching 22- to 24-nt sRNA sequences from each library. Nucleotide frequency plots for sRNAs with a definitively untemplated 3′ U are shown in Supplemental Figure S4.

more sRNA classes associated with Twi2, Twi7, Twi8, or Twi10 that together are likely to constitute the overall pool of 23- to 24-nt sRNA produced in vegetative growth. Among these sRNA classes are some eliminated by the loss of a specific RDRC, decreased in abundance by loss of any RDRC, or increased in abundance by the loss of a specific RDRC. The sRNA classes also show differential dependence on expression of the Twi proteins themselves. The diversity of genetic requirements for sRNA accumulation implies a remarkable complexity of handling pathways for sRNA precursors.

Unphased sRNAs strictly antisense to families of potential pseudogenes

The few previously sequenced *T. thermophila* sRNAs map to the antisense strand of predicted ORFs with one unifying feature: a pseudogene-like, genome-encoded ~30- to 85-nt tract of polyadenosine 3′ of the ORF on the sense strand (Lee and Collins 2006). Predicted ORFs with these features are clustered in sequence-related groups that altogether can be classified into five families based on ORF sequence similarity (Families I–V). Deep sequencing of total or Twi2-bound sRNAs revealed major density peaks at these pseudogene loci. The sRNA distribution plot across a representative pseudogene cluster can be used to illustrate preferential enrichment of sRNAs from pseudogene loci by Twi2 relative to Twi7 or Twi8 (Fig. 3A shows sRNA density at pseudogene Family III cluster B or IIIB). Because the total number of sequences obtained for Twi2, Twi7, and Twi8 libraries is similar (Fig. 2A), density plots provide a readily visualized estimate of relative enrichment. Predictions based on density plot comparisons were then verified by Northern blot analysis. In the density plots shown, sRNAs are mapped by their 5′-end positions on either the top or bottom strand, indicated as above or below the zero axis, respectively. As shown for the representative cluster IIIB (Fig. 3A), sRNA accumulation at all pseudogene loci was highly strand-asymmetric, always antisense to the predicted ORFs (ORF annotation is illustrated below sRNA density). Close inspection revealed that most sRNAs overlap, such that nearly every nonconsecutive A in the coding strand mapped as the 5′ U of an antisense strand sRNA. The lack of sRNAs with a 5′ end complementary to the first of tandem adenosines is consistent with the observed bias against second position U (Fig. 2D).

We estimated relative sRNA enrichment at each pseudogene cluster based on the total number of sRNA sequences obtained from Twi2, Twi7, and Twi8 libraries. Twi2-associated sRNAs were the most abundantly represented at all pseudogene cluster loci with major sRNA density peaks (Fig. 3B, top panel). In agreement with this prediction, for every pseudogene cluster sRNA examined by Northern blot hybridization using a sRNA complementary oligonucleotide probe, sRNAs were most enriched in association with Twi2 (Fig. 3C,D, top panels are Northern blots and bottom panels are gels before transfer stained with SYBR Gold to show 23- to 24-nt sRNA loading). Some enrichment by Twi7 was also detectable, although this was substantially less than enrichment by Twi2 when normalized to loading of 23- to 24-nt sRNAs.

For sRNAs from most pseudogene clusters, sRNA accumulation specifically required Twi2: Knockout of *TWI2* or *TWI2-6* resulted in sRNA loss in vivo (Fig. 3C). Accumulation of IIIB sRNA also specifically required the RDRC subunit Rdn2. Comprehensive analysis uncovered an unanticipated disparity in the requirements for accumulation of sRNAs from Family I pseudogene loci. As shown for cluster IB (Fig. 3D), sRNA accumulation was eliminated by knockout of *TWI8*. Also, sRNA accumulation was reduced but not eliminated by knockout of *TWI2*, *TWI2-6*, or any of the individual RDRCs (Fig. 3D). This distinction between Family I and all other pseudogene families is evident in a comparison of sRNA sequence abundance across RDRC subunit knockout strains (Fig. 3B, bottom panel), which shows increased representation of Family I sRNAs in the Rdn2KO library and increased representation of sRNAs from other pseudogene families in the Rdf2KO library. The genetic requirements for accumulation of pseudogene Family I sRNAs also characterize high-copy repeat sRNAs described below, suggesting that loci representing an entire pseudogene family can switch from generating sRNAs through a mechanism that is most sensitive to disruption of Twi2 and Rdn2 to a different pathway, shared with high-copy repeat sRNAs, that is most sensitive to disruption of Twi8.

Because some gene knockouts greatly reduced pseudogene loci sRNA accumulation, we assayed for a corresponding increase in accumulation of primary transcripts. Transcripts from the pseudogene loci were only weakly detectable by RT–PCR and may not be quantitatively detected due to cDNA synthesis inhibition by the dsRNA complementary strand. Northern blot hybridization failed to detect transcripts from any pseudogene locus in any strain (data not shown). These results suggest that the pseudogene loci are either minimally transcribed or that in the absence of the pathway necessary for sRNA biogenesis, product(s) from the loci can undergo sRNA-independent degradation.

Three classes of sRNAs from three types of repeats

Although the MAC genome is relatively free of multicopy elements, repeat structure per se does not fate loci to elimination, and thus repeats can persist through the genome restructuring process (Eisen et al. 2006). We discovered sRNAs abundant enough to detect by Northern blot hybridization of total 23- to 24-nt sRNAs that mapped to degenerate repeat loci with interrupted tandem arrays of ~150-base-pair (bp) repeat units, spanning 2–20 kb at any given genome location (Figs. 4A,B). The same degenerate sequence motif was present at least once on 185 different genome scaffolds. Profiles of sRNA density suggest preferential binding of these high-copy repeat sRNAs by Twi2 (Fig. 4A), which matches the specificity determined by Northern blot hybridization (Fig. 4B). Like pseudogene sRNAs, high-copy repeat sRNAs show an extreme strand bias of accumulation (Fig. 4A). Because none of the ESTs that map to high-copy repeat loci map uniquely or

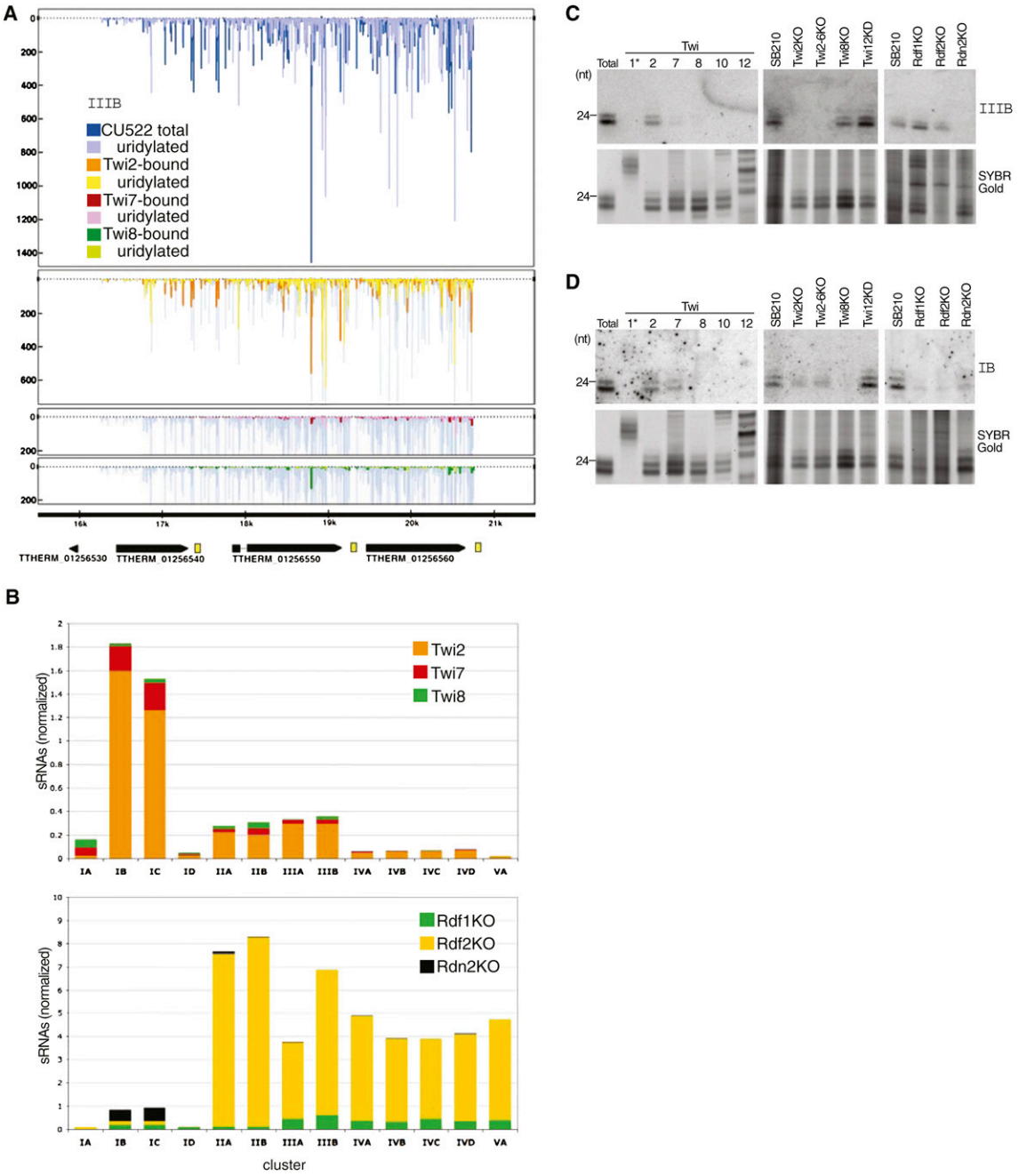

Figure 3. Pseudogene-derived sRNA clusters. (*A*) sRNA density plots for a representative pseudogene locus (IIIB). Density of sRNAs from each Twi-enriched library is shown superimposed on density from the total sRNA library of CU522. Peak height indicates the number of sRNA reads with the same 5′-end position at that genome location. Peaks *above* the zero axis represent sequences on the top strand and peaks *below* the zero axis represent sequences on the bottom strand. Annotations for predicted protein-coding genes (TTHERM numbers shown) are indicated at the *bottom* of the panel; yellow boxes indicate A-rich tracts. (*B*) Comparisons of sRNA numbers mapping to 13 genome locations of pseudogene loci. Number of sRNAs enriched by each Twi protein (*top* panel) or present in total sRNA from RDRC subunit knockout strains (*bottom* panel) was normalized to the corresponding wild-type library by the number of filtered sequences. Loci are indicated by the name of the ORF family (I, II, III, IV, or V) and cluster within that family (A, B, C, D). Note the difference in scale between the *top* and *bottom* panels of the figure. (*C*,*D*) Northern blot hybridization with an oligonucleotide probe complementary to a sRNA from pseudogene cluster IIIB or IB, as indicated. *Below* the blots, RNA loading is shown by SYBR Gold staining.

extensively across the repeats, the orientation of sRNAs relative to primary transcripts from the region is uncertain. In vivo accumulation of high-copy repeat sRNAs was eliminated by knockout of *TWI8* and reduced by knockout of *TWI2*, *TWI2-6*, or any individual RDRC (Fig. 4B), as described for pseudogene Family I sRNAs above. Curiously, high-copy repeat sRNAs were not subject to 3′ uridylation in association with any Twi protein (Fig. 4A).

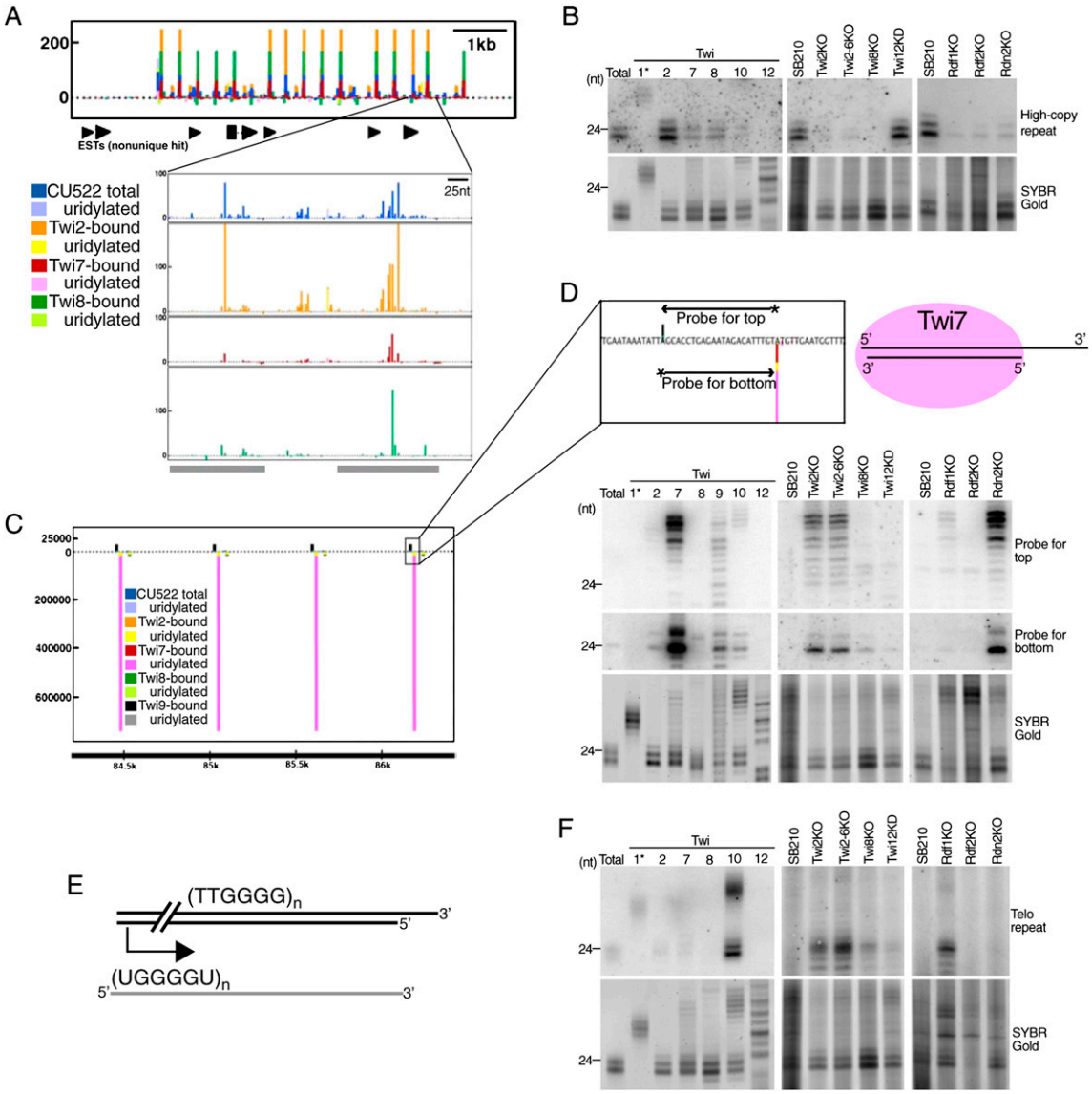

Figure 4. Repeat-derived sRNA clusters. (*A*) sRNA density plots at a high-copy repeat locus. Gray-shaded bars represent the degenerate repeat unit. (*B*) Northern blot hybridization with an oligonucleotide probe. *Below* the blots, RNA loading is shown by SYBR Gold staining. (*C*) sRNA density plot at a low-copy repeat locus. Colored bars that are not visible are not hidden; they are not large enough to see on the given scale. (*D*) Northern blot hybridization with oligonucleotide probes indicated in the illustration at *top*. *Below* the blots, RNA loading is shown by SYBR Gold staining. The possible structure of sRNAs bound to Twi7 is also shown. (*E*) Illustration of a putative telo-sRNA precursor derived from telomere transcription. (*F*) Northern blot hybridization with an oligonucleotide probe complementary to the G-strand of telomeric repeats. *Below* the blots, RNA loading is shown by SYBR Gold staining.

We also found sRNAs cognate to low-copy repeats present at <10 genome locations, with repeat units of 100–300 bp spanning 2–4 kb. The highest-density peaks of the Twi7 library mapped to these low-copy repeat loci, with particular enrichment of sequences extended by an untemplated 3′ U (Fig. 4C). The Twi7-bound 22- to 24-nt sRNAs showed highly strand-asymmetric accumulation, but a minor peak was detected on the complementary strand in the Twi9 library (Fig. 4C). Using oligonucleotide probes for strand-specific detection of low-copy repeat sRNAs by Northern blot (Fig. 4D, top), we discovered that sRNAs from the strand represented in the Twi9 library were most strongly associated with Twi7 in the form of longer, ~32- to 34-nt sRNAs visible by SYBR Gold staining (Fig. 4D, bottom). Because Twi9 is not normally expressed in vegetative growth, both strands are likely to be bound exclusively to Twi7, potentially as a duplex (Fig. 4D, top). Neither size class of Twi7-bound sRNA is abundant enough to detect by Northern blot in total sRNA, but both sRNA strands are coordinately and dramatically increased in accumulation by knockout of *TWI2*, *TWI2-6*, or the RDRC subunit Rdn2 (Fig. 4D).

A third, unexpected class of repeat-derived sRNAs was discovered as peaks of sRNA density in the Rdf1KO strain

library. These telomeric repeat sRNAs (telo-sRNAs) represent the G-rich strand exclusively and begin with fixed phasing in one of six possible permutations of the T_2G_4 repeat: 5′-UGGGGU-3′ (Fig. 4E). If precursor transcripts have the G-rich telomeric repeat sequence, telo-sRNAs would arise from transcription initiated within the chromosome (Fig. 4E). Northern blot hybridization with an oligonucleotide complementary to the G-rich repeat revealed strong enrichment of telo-sRNAs in association with Twi10, with hybridization to both the 23- to 24-nt and ~33- to 36-nt Twi10-enriched sRNA size classes visible by SYBR Gold staining (Fig. 4F). Telo-sRNAs are not abundant enough to detect by Northern blot in size-enriched total sRNA from wild-type strains, but they are dramatically increased in accumulation by knockout of *TWI2*, *TWI2-6*, or the RDRC subunit Rdf1 (Fig. 4F). The absolute strand bias and precise phasing of telo-sRNAs could arise from processing requirements and/or from selective stabilization of sRNAs with 5′ U but not second position U (Fig. 2D).

Phased strand-asymmetric sRNAs initiated from a predicted hairpin

Apart from the pseudogene loci that generate abundant sRNAs described above, seven additional loci designated Ph1–Ph7 generated a high proportion of the clustered sRNAs in wild-type total sRNA and Twi2-bound sRNA libraries. These loci were generally unannotated in the genome, although parts of two clusters overlapped portions of predicted protein-coding genes in the antisense orientation. Unlike the case for any other class of *T. thermophila* sRNA, sRNA density peaks at these loci were spaced at intervals predominantly 24 nt apart (shown for cluster Ph3 in Fig. 5A). In common with other *T. thermophila* sRNA classes, sRNA accumulation at these phased cluster loci was strand-asymmetric. Strikingly, at the edge of all clusters, the strand complementary to the sRNAs could form a predicted stem–loop structure of 50–100 bp in a configuration that would prime Rdr-mediated synthesis of the sRNA-generating strand (Fig. 5A). Curiously, sRNA density was highest adjacent to but not within the region of transcript self-complementarity. These features of *T. thermophila* phased sRNAs differ from phased sRNAs identified in other organisms, which accumulate from both strands (Vazquez et al. 2004; Allen et al. 2005; Molnár et al. 2007; Zhao et al. 2007) or from both sides of an RNA structure (Babiarz et al. 2008; Czech et al. 2008).

If the characteristic stem–loop feature is a novel mechanism for soliciting RDRC and thus stimulating sRNA production, genetic requirements for production of phased cluster sRNAs may differ from those of other sRNA classes. Phased cluster loci sRNAs associate with Twi2 as well as other Twis that carry 23- to 24-nt sRNAs, but their accumulation does not strictly require the presence of Twi2 or Twi8 (Fig. 5A,B). Unique among *T. thermophila* sRNA classes, phased cluster loci sRNAs absolutely require the RDRC subunit Rdf2 for their accumulation (Fig. 5B). By Northern blot hybridization, we detected an ~700-nt transcript from Ph3 that accumulates only in the absence of Rdf2 (Fig. 5C). RT–PCR assays suggest that this transcript derives from the strand complementary to the sRNAs (data not shown), representing the hypothetical transcript depicted in Figure 5A. To confirm the loss of transcript silencing at phased cluster loci, we probed for the putative primary transcript of another phased cluster locus, Ph2. Again, accumulation of an ~500-nt transcript was detected only in the

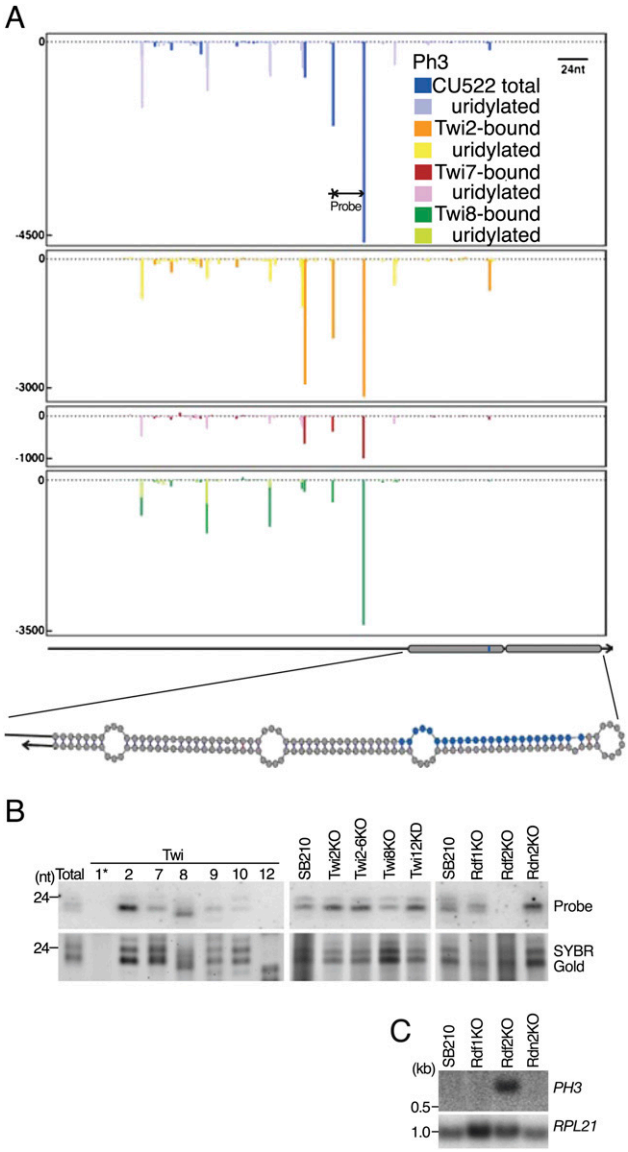

Figure 5. Phased sRNA clusters. (*A*) sRNA density plots at the phased cluster locus Ph3. A hypothetical sRNA complementary transcript is shown *below* the plot as well as a predicted secondary structure for the region shaded in gray. Blue shading in the secondary structure depiction indicates the complement of the sRNA in that region. (*B*) Northern blot hybridization with a sRNA complementary oligonucleotide probe. *Below* the blots, RNA loading is shown by SYBR Gold staining. (*C*) Northern blot hybridization with a hexamer-labeled probe covering much of the region shown in *A*. *RPL21* expression is shown as a loading control.

absence of Rdf2 (Supplemental Fig. S5). The hairpin structure may be a conserved feature that marks these transcripts for silencing as virus-like elements. It is also possible that the structured transcripts serve to generate sRNAs that regulate mRNAs from unlinked loci, if only imperfect complementarity between the sRNA and target mRNA is required.

Strand-unbiased sRNAs from EST-supported loci

Because Twi8-associated sRNAs carry a 3′ modification with potential to reduce ligation efficiency (Fig. 1F), sRNAs associated specifically with Twi8 may have been poorly represented in libraries of total 22- to 24-nt sRNAs. Correspondingly, genome mapping of sRNAs from the Twi8 library revealed peaks of density distinct from those described above. The highest density of Twi8-bound sRNAs was located in an unannotated region of the genome with a uniquely mapping EST. The sRNA distribution is not spread across the EST, but instead occurs as a major peak on the EST antisense strand and additional minor peaks on both strands (Fig. 6A, peaks numbered 1–4). The sRNA density peaks define a region within the EST that forms a snap-back hairpin structure by Mfold analysis (Fig. 6A, sequences of numbered sRNAs are highlighted in green). Northern blot hybridization using an oligonucleotide complementary to the most abundant sRNA sequence confirmed enrichment by Twi8 and also by Twi7 (Fig. 6B). Notably, the size profile of Twi7-associated sRNAs includes a wider range (23–26 nt) than the predominantly ~24-nt sRNAs associated with Twi8. This difference could reflect tighter size-selectivity of Twi8 versus Twi7 for loading with the distribution of sRNAs generated by nuclease processing in stem bulges of the precursor (Fig. 6A). As observed for Twi7-enriched low-copy repeat sRNAs (Fig. 4D), the heterogeneous 23- to 26-nt sRNA population increased in accumulation in strains with knockout of *TWI2*, *TWI2-6*, or the RDRC subunit Rdn2 (Fig. 6B).

Other high-density peaks of Twi8 library sRNAs showed the common feature of mapping to loci with protein-coding gene predictions supported by ESTs (Fig. 6C shows a representative locus). Individual sequences from the Twi8 library other than the most represented sRNA (Fig. 6A) were not detectable by Northern blot in total 23- to 24-nt sRNA (data not shown). This difficulty in detecting individual Twi8-enriched sRNAs by Northern blot is consistent with the relatively high complexity of the Twi8 sRNA library (Fig. 2C). Excepting the highest-density cluster of Twi8-bound sRNAs (Fig. 6A), sRNAs did not map to transcripts predicted to form extensive secondary structure. However, at every cluster, either ESTs did not map uniquely (indicating the presence of paralogous genes) or transcription units were closely spaced and often predicted to converge (suggesting transcriptional interference or overlap). Unlike all other *T. thermophila* sRNA loci, loci that generate predominantly Twi8-bound sRNAs produced sRNAs that mapped to both strands of the genome (Fig. 6C). At these loci, the few Twi2-bound sRNAs derived predominantly from the sense strand rather than the antisense strand. These properties suggest that loci producing Twi8-associated sRNAs form a different type of dsRNA precursor.

The high complexity and genome mapping features of Twi8-bound sRNAs resemble the features of endogenous siRNAs (endo-siRNAs) in mouse and *Drosophila* (Nilsen 2008; Okamura and Lai 2008; Ghildiyal and Zamore 2009). *Drosophila* endo-siRNAs have been mapped to originate from the *Ago2* locus (Okamura and Lai 2008). Curiously, one of the Twi8 sRNA clusters mapped to the tandem array of predicted ORFs at *TWI2-6* (Fig. 6D). Only *TWI2* yields abundant mRNA (discussed above; see Supplemental Fig. S2), and *TWI2* also gives rise to the vast majority of uniquely mapping sRNAs (Fig. 6D shows only uniquely mapping sRNAs; see Supplemental Fig. S6 for mapping of all sRNAs to all possible *TWI2-6* locations). Consistent with Twi8 RNPs exerting a silencing effect on *TWI2*, *TWI2* mRNA level increased in a strain lacking Twi8 and decreased in a strain overexpressing Twi8 (Supplemental Fig. S7). Curiously, *TWI2* mRNA also increased in a strain lacking Rdn2 (Supplemental Fig. S7). In addition to any direct effect of Twi8 or Rdn2 on *TWI2* mRNA level, there could be indirect influence from a change in intracellular production of dsRNA; in previous studies, dsRNA formation has been proposed to regulate *TWI2* expression (Howard-Till and Yao 2006).

In the course of this work, we generated strains lacking *TWI2* or *TWI2-6* (Fig. 1C). We recreated these gene disruptions in the transgene strain used for Twi2 affinity purification, again obtaining complete knockout of *TWI2* or *TWI2-6* (data not shown). Intriguingly, Northern blot analysis of *TWI2* expression in these strains revealed regulation of the transgene mRNA by the endogenous gene locus (Fig. 6E, top panels). Expression of the transgene mRNA was detectable at the basal level of expression from the transgene *MTT1* promoter in strains lacking *TWI2* or *TWI2-6* (Fig. 6E, lanes 3,4) but was undetectable in the presence of endogenous *TWI2* (Fig. 6E, lane 2). To confirm this unexpected difference in transgene expression, we used tag-specific antibody to detected tagged Twi2 protein accumulation, which was indeed repressed in the presence of *TWI2* (Fig. 6E, bottom panel). When cadmium was added to induce high-level *MTT1* transcription, transgene mRNA increased relative to the mRNA from *RPL21* used as a loading control (Fig. 6E, lanes 5–8, note from *RPL21* hybridization that lane 8 is underloaded). Surprisingly, along with the increase in transgene mRNA level, cadmium also increased mRNA from endogenous *TWI2* in the transgene strain (Fig. 6E, lane 6; see also Supplemental Fig. S7) but not the wild-type strain (Fig. 6E, cf. lanes 1 and 5). The cadmium-induced increase in mRNA from endogenous *TWI2* appeared to count against the full increase in transgene mRNA and transgene-encoded protein. These observations reveal *trans*-active modulation of total Twi2 mRNA and protein by the endogenous *TWI2* locus.

From an overall perspective, the isolation and characterization of *T. thermophila* sRNAs from Twi RNPs and from different strain backgrounds revealed an unexpectedly large complexity of 22- to 24-nt sRNAs in growing

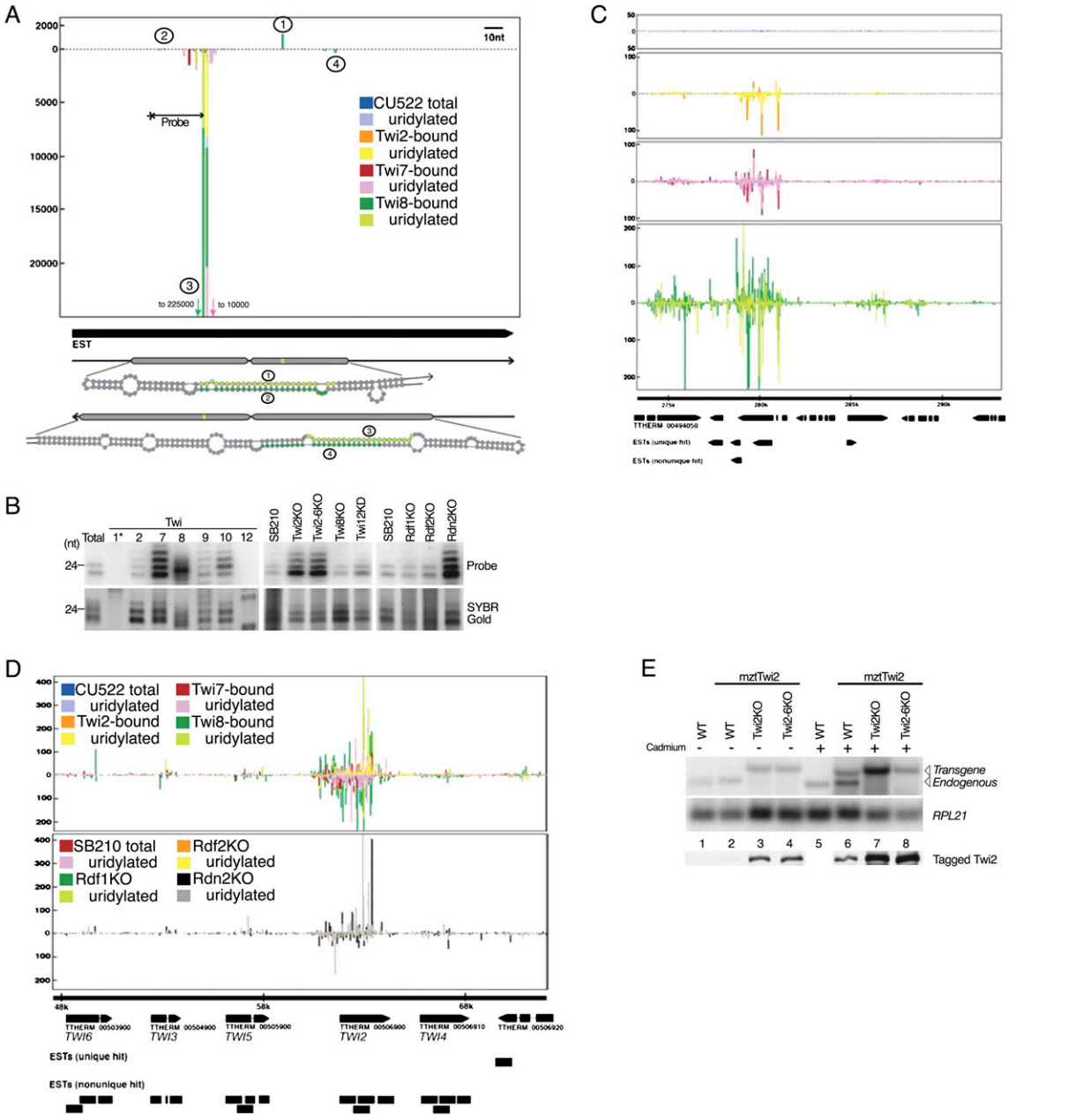

Figure 6. Twi8-associated sRNA clusters. (*A*) sRNA density plot at a structured RNA locus that is the most abundant cluster of Twi8 library sRNA. An EST that maps to the locus is indicated as well as predicted stem–loop structures for both sense strand and antisense strand transcripts. The green shading in stem–loop structures corresponds to the sRNAs numbered in the density plot. (*B*) Northern blot hybridization with an oligonucleotide probe complementary to the most abundant sRNA in *A*. *Below* the blots, RNA loading is shown by SYBR Gold staining. (*C*) sRNA density plots for a representative locus of Twi8-enriched sRNAs with convergently transcribed ORFs and/or ESTs that do not map uniquely. (*D*) sRNA density plots at *TWI2-6*. Only uniquely mapping sRNA sequences are included; see Supplemental Figure S6 for mapping of all sRNAs. Colored bars that are not visible are not large enough to see on the given scale. (*E*) Regulation of transgene-encoded Twi2 by the endogenous *TWI2* locus. The presence of the transgene is indicated as mztTwi2. The *top* panels are mRNA Northern blots for expression of *TWI2* and the *RPL21* loading control; note that endogenous and transgene mRNAs are equally detected by the probe but differ slightly in size, as indicated. The *bottom* panel is an immunoblot detecting tagged Twi2; equal loading was confirmed by total protein staining (not shown).

cells. Several classes of abundant sRNAs show unique specificities of Twi protein association and distinct genetic requirements for accumulation (Figs. 7A,B). Several additional classes of less abundant sRNA were uncovered only after enrichment for a particular Twi protein or in a genetic background that depleted abundant sRNAs, reinforcing the utility of sRNA profiling under different conditions of growth, development, and strain background.

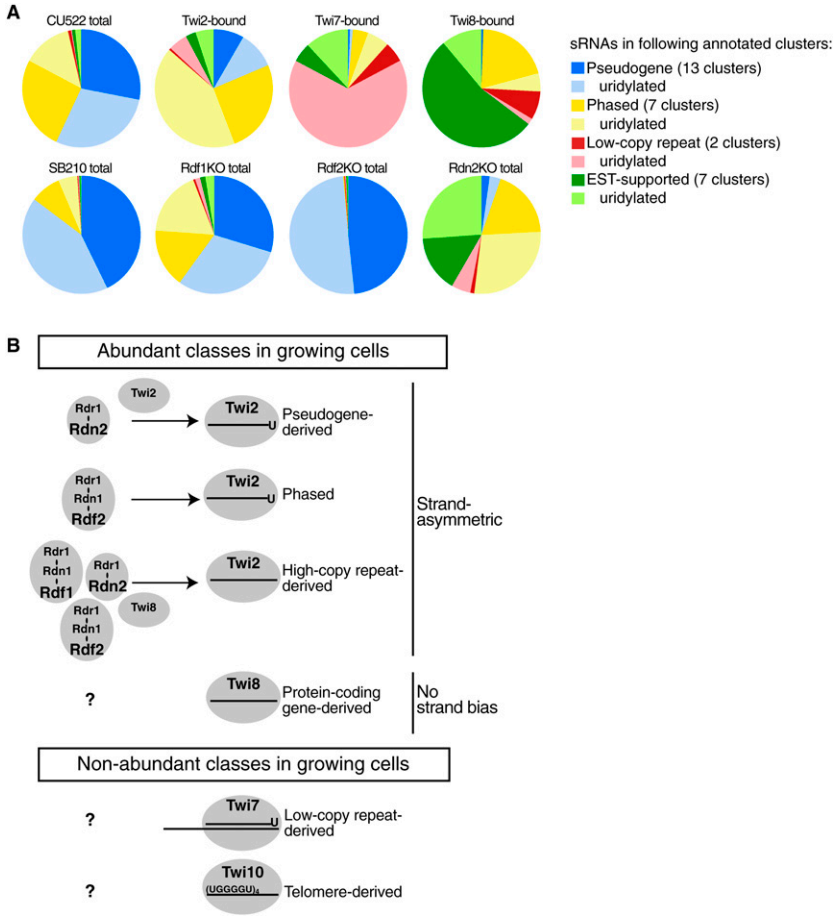

Figure 7. Summary of library representation and sRNA classes. (*A*) Representation of abundant and preferentially enriched sRNA classes. Only annotated classes with large enough representation to generate a visible pie slice in at least one library are included; these classes are listed in the legend at *right*. (*B*) Summary of sRNA classes, characteristics, and accumulation requirements.

Discussion

sRNAs provide sequence specificity for regulated gene expression and, with their associated PPD proteins, modulate the influence of viruses, transposons, and other challenges to genome stability. Conserved properties of sRNAs can be obscured by low sRNA abundance. New principles of sRNA function continue to emerge from studies that exploit a wide range of eukaryotic systems. Here, by profiling of endogenous sRNAs expressed from the *T. thermophila* MAC during asexual growth, we favored the discovery of sRNAs with roles other than repressive heterochromatin formation or transposon silencing. This approach yielded numerous unanticipated sRNA classes that are variously strand-asymmetric or strand-symmetric, phased or unphased, with or without untemplated uridylation, and originating from protein-coding or noncoding loci. At the sequence level, these sRNA classes derive from pseudogene families, distinct types of DNA repeats, RNAs with internal regions of secondary structure, and EST-supported mRNAs with convergently transcribed or paralogous genes.

Twi2 associates with the two most abundant sRNA classes from pseudogene and phased cluster loci, which share highly asymmetric antisense strand accumulation. These sRNAs are extremely abundant in both CU522 and SB210 wild-type backgrounds (Fig. 7A), even though the primary transcripts are not detected easily (Lee and Collins 2006; additional data not shown). The abundance of sRNAs and virtual absence of cognate mRNAs, combined with Twi2 cytoplasmic localization, support the PTGS function of Twi2 in mRNA decay proposed from studies of exogenously induced RNAi (Howard-Till and Yao 2006). The selective stabilization of antisense strand sRNAs would increase the sequence selectivity of PTGS, which may be particularly important in gene-rich ciliate genomes with >20,000 unique ORFs. Twi2 is nonessential, consistent with aberrant transcript destruction as a primary biological function rather than gene regulation in *trans*. In Twi2KO strains, Twi2-targeted transcripts may be shunted to degradation by another sRNA-mediated silencing pathway or to the *T. thermophila* nonsense-mediated and no-go mRNA decay machineries (Atkinson et al. 2008). Knockout of Twi2 (or Rdn2) increased the accumulation of sRNAs enriched by Twi7 and Twi8, which could reflect increased availability of shared sRNA biogenesis factors and/or sparing of precursors for Twi7 and Twi8 sRNAs from Twi2-dependent degradation.

Twi8 carries sRNAs that are distinct from Twi2-enriched sRNAs by several criteria: their origin from mRNA-producing loci, the lack of the otherwise pervasive strand asymmetry of accumulation, and 3′-end modification. Twi8-bound sRNAs are less abundant than either

major class of Twi2-bound sRNAs, suggesting a lower abundance of the precursors. Based on these observations and Twi8 localization to the MAC rather than the cytoplasm, it seems likely that Twi8 mediates cotranscriptional rather than post-transcriptional regulation. We suggest that Twi8 RNPs reduce mRNA production from convergent genes and also genes subject to antisense regulation by transcription of paralogous sequences, such as *TWI2*. This nuclear sRNA regulation pathway may also buffer the output of any bidirectionally transcribed MIC loci that escape elimination in the developing MAC. The relatively modest change in *TWI2* transcript level upon Twi8 overexpression or gene knockout (~10-fold or less over several experiments) (see Supplemental Fig. S7) suggests that Twi8-mediated transcriptional regulation may be a less repressive and/or more dynamic form of transcriptional "dampening" compared with the H3K9 methylation that occurs in conjugating cells and in formation of heterochromatin in other organisms. We suggest that endo-siRNAs bound to PPD proteins in other organisms could function similarly to Twi8 to establish a state of transcriptional dampening, which in other organisms could be subsequently acted on by additional chromatin-modifying activities to produce a heterochromatin state at least partially independent of the initiating sRNA pathway.

The low-copy repeat sRNAs and telo-sRNAs associated with Twi7 and Twi10, respectively, represent novel classes of sRNAs. The predominant class of Twi7-enriched sRNA is comprised of complementary strands with unequal lengths, apparently inconsistent with biogenesis by a Dicer. However, we note that bacterial infection of *Arabidopsis thaliana* triggers the production of long ~30- to 40-nt sRNAs from natural antisense transcripts in a process that requires DCL4 and DCL1, which on other substrates generate shorter sRNA products (Katiyar-Agarwal et al. 2007). The telomeric-repeat sRNAs bound by Twi10, like Twi10 itself, may play a larger role in conjugation-induced genome restructuring than in vegetative growth. Recent studies have reported long telomeric repeat RNA transcripts as structural components of mammalian telomeres (Azzalin et al. 2007; Schoeftner and Blasco 2008), and C-strand telomeric repeat sRNAs were identified as 1% of sequences from a total 20- to 30-nt sRNA library from *Giardia intestinalis* (Ullu et al. 2005), but to our knowledge, G-strand telo-sRNAs have not been reported previously to associate with a PPD protein in any organism. The *T. thermophila* MAC in growing cells contains a minimum of ~40,000 telomeres, each with ~300 bp of telomeric repeats, and even in this organism, telo-sRNAs were only detectable by Northern blot of total sRNA in strains lacking Twi2 or Rdf1. Thus, telo-sRNAs may be conserved but not yet detected in other organisms due to low abundance.

In addition to the numerous distinctions between the *T. thermophila* 23- to 24-nt sRNA classes, some similarities were also evident. All classes of sRNA share the features of a 5′-monophosphate and a 5′-U bias. Also, most Twis bound sRNAs extended by a 3′-untemplated nucleotide, predominantly U. These common properties of end structure suggest that the observed specificity of sRNA sorting to Twi partners does not occur by RNA sequence selectivity inherent in the PPD proteins themselves (Jinek and Doudna 2009; Siomi and Siomi 2009). Instead, this specificity is likely to derive from the pathway of sRNA biogenesis. Curiously, rather than exploiting multiple isoforms of Dicer or Rdr, *T. thermophila* sRNA biogenesis pathways in growing cells are diversified by use of different RDRC assemblies that share the same Rdr1 and Dcr2 enzymes. *T. thermophila* Rdf2 is the only nonessential RDRC subunit preferentially expressed in vegetative growth (Lee et al. 2009). Consistent with this developmental expression profile, Rdf2 is required for accumulation of Twi2-bound sRNAs from phased cluster loci, one of the most abundant sRNA classes in growing cells. The inability of the sequence-related Rdf1 to compensate for loss of Rdf2 may reflect the lower expression level of Rdf1 than Rdf2 in vegetative growth, an alternative subcellular compartmentalization, or a specialized role for Rdf2 in recognition of the hairpin structure in phased cluster loci primary transcripts. On the other hand, loss of Rdf1 but not Rdf2 increased the accumulation of telo-sRNAs, which could be linked to the defects in MAC DNA segregation observed in Rdf1KO cells (Lee et al. 2009).

By exploiting the ciliate life cycle stage of vegetative growth, this work uncovered a remarkable diversity of sRNAs including some previously unknown classes likely to be conserved. The *T. thermophila* expansion of Piwi family proteins and sRNA cargos may be important for optimal gene expression and growth in the ecological niche of temperate freshwater ponds. Growth in sterile rich media under laboratory conditions may allow some otherwise critical pathways of sRNA-mediated regulation to become nonessential; for example, pathways that respond to changing environmental conditions or foreign genome invasion. A major evolutionary motivation for the ciliate expansion of sRNA and Twi RNP diversity may be in the opportunity for an epigenetic influence of asexual growth history on subsequent sexual reproduction. Results here suggest future studies of the influence of somatic sRNA populations on germline differentiation through genome restructuring in the next sexual cycle.

Materials and methods

Gene cloning, nucleic acid and protein methods, and affinity purification

ORFs for *TWI1*, *TWI2*, *TWI8*, *TWI9*, *TWI10*, and *TWI11* were cloned by RT–PCR. ORFs for *TWI7* and *TWI12* were cloned by PCR from genomic DNA. Southern blots and Northern blots for mRNA detection were hybridized with hexamer-primed probes, while sRNA Northern blots used 5′-end-labeled oligonucleotide probes as described (Lee and Collins 2006). To perform immunoblots of total cellular protein, 2×10^5 cells from cultures in log phase were washed and resuspended in 50 μL of 10 mM Tris (pH 7.5). SDS-PAGE loading buffer was added, samples were immediately boiled, and 5% of the sample was resolved by SDS-PAGE. After protein transfer, membrane was blocked in 5% nonfat milk, incubated with rabbit IgG primary antibody at 1/10,000 dilution for 1 h, washed in PBS, incubated with goat anti-rabbit AF800 at 1/20,000 for 1 h, washed in PBS, and imaged using the LI-COR Odyssey system.

Immunopurification of each ZZ-tagged Twi complex was performed as described (Lee and Collins 2007). Transgene expression was induced by addition of 1 μg/mL $CdCl_2$ overnight to vegetative growth cultures, which were harvested during log phase (1×10^5 to 5×10^5 cells per milliliter), or by addition of 0.1 μg/mL $CdCl_2$ to conjugating cultures upon cell mixing. RNAs were isolated from the purified complexes by phenol:chloroform extraction. Total 23- to 24-nt sRNA was isolated from size-enriched TRIzol-extracted RNA of vegetative growth cultures as described (Lee and Collins 2006). For library construction, sRNAs were excised from denaturing PAGE gels stained with SYBR Gold and eluted in 0.3 M sodium acetate by soaking overnight with shaking at 37°C. β-Elimination was performed largely as described (Akbergenov et al. 2006). Terminator Exonuclease (Epicentre) was used based on the manufacturer's protocol.

Strain construction, culture growth, and imaging

Strains expressing Twi2, Twi8, and Twi12 with an N-terminal-enhanced GFP tag were made in CU427 background using integration cassettes targeted to *MTT1*, with a neo2 cassette upstream of the *MTT1* promoter (Shang et al. 2002). Knockout and knockdown strains were made in SB210 background by targeting integration of the neo2 cassette in replacement of the endogenous ORF. Cells were selected for maximal locus assortment to the recombinant chromosome using paromomycin (Witkin et al. 2007). Strains expressing ZZ-tagged Twi1, Twi2, Twi7, Twi8, Twi10, and Twi12 for affinity purification were made using integration cassettes targeted to *BTU1* in strain CU522. Each endogenous ORF was fused to an N-terminal tag with tandem Protein A domains and a TEV protease cleavage site. The tag was preceded by an ~1-kb promoter region from *MTT1* (Shang et al. 2002), and the ORF was followed by the polyadenylation signal of *BTU1*. Cells were selected using paclitaxel (Witkin and Collins 2004). For ZZ-tagged Twi9 and Twi11 expression constructs, a neo2 cassette was included after the polyadenylation signal, and selection was performed using paromomycin.

Cell cultures were grown shaking at 30°C in 2% proteose peptone, 0.2% yeast extract, and 10 μM $FeCl_3$ to log phase (1×10^5 to 5×10^5 cells per milliliter). Conjugation was initiated by mixing equal numbers of starved cells and incubation at 30°C without shaking. To image GFP, growing cells at a density of $\sim 3 \times 10^5$ cells per milliliter were starved for 5 h in 10 mM Tris (pH 7.5) to reduce autofluorescence from food vacuoles, with 0.05–0.1 μg/mL cadmium chloride added 1–3 h after transfer to starvation media to induce tagged protein expression. Hoechst 33342 was added to a concentration of 5 μg/mL for 2 min at room temperature. Cells were then immobilized in 8–10 μL of 2% low-melt agarose under coverslips with corners dabbed with nail polish. Fluorescence was visualized using the 40× objective of an Olympus model BX61 microscope equipped with a Hamamatsu Digital camera. Images were captured using Metamorph software.

sRNA cloning, sequencing, and analysis

sRNA cloning was performed largely as described previously (see the Supplemental Material). Sequenced sRNAs were trimmed for 3′ linker and then matched to the November 2003 MAC genome assembly 2 (Eisen et al. 2006), excluding 763 MIC-limited scaffolds (Coyne et al. 2008). Only sRNAs with 100% sequence identity after allowing a single 3′ mismatch were retained. Genome annotations were from the August 2004 release and EST data were downloaded from http://www.ciliate.org. Due to the relative lack of MAC repeat sequences, most sRNAs mapped uniquely in the genome. Comparisons with and without normalization of sRNA density by number of genome hits did not affect interpretations; with exceptions noted in the text, no normalization was performed for the density plots shown.

Accession numbers

GenBank accession numbers and *Tetrahymena* Genome Database annotations for *TWI* sequences are as follows: *TWI2*: DQ855010, TTHERM_00506900; *TWI7*: EF507506, TTHERM_00600450; *TWI8*: DQ855965, TTHERM_00449120; *TWI9*: EU183124, TTHERM_00203030; *TWI10*: EF507505, TTHERM_01132860; *TWI11*; EU183125, TTHERM_00144830; *TWI12*: EF507507, TTHERM_00653810. sRNA libraries are deposited at Gene Expression Omnibus (accession no. GSE17006, data sets GSM425486–GSM425496).

Acknowledgments

We thank members of the *Tetrahymena* research community for sharing unpublished data and discussion. C.D.M. is a Beckman fellow of the Watson School of Biological Sciences and is supported by a National Science Foundation Graduate Research Fellowship. This work was supported in part by grants from the NIH (to G.J.H. and K.C.) and a kind gift from Kathryn W. Davis (to G.J.H.).

References

Akbergenov R, Si-Ammour A, Blevins T, Amin I, Kutter C, Vanderschuren H, Zhang P, Akbergenov R, Si-Ammour A, Blevins T, et al. 2006. Molecular characterization of geminivirus-derived small RNAs in different plant species. *Nucleic Acids Res* **34:** 462–471.

Allen E, Xie Z, Gustafson AM, Carrington JC. 2005. microRNA-directed phasing during *trans*-acting siRNA biogenesis in plants. *Cell* **121:** 207–221.

Aoki K, Moriguchi H, Yoshioka T, Okawa K, Tabara H. 2007. In vitro analyses of the production and activity of secondary small interfering RNAs in *C. elegans*. *EMBO J* **26:** 5007–5019.

Atkinson GC, Baldauf SL, Hauryliuk V. 2008. Evolution of nonstop, no-go and nonsense-mediated mRNA decay and their termination factor-derived components. *BMC Evol Biol* **8:** 290. doi: 10.1186/1471-2148-8-290.

Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J. 2007. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* **318:** 798–801.

Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R. 2008. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & Dev* **22:** 2773–2785.

Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136:** 642–655.

Cerutti H, Casas-Mollano JA. 2006. On the origin and functions of RNA-mediated silencing: From protists to man. *Curr Genet* **50:** 81–99.

Chalker DL. 2008. Dynamic nuclear reorganization during genome remodeling of *Tetrahymena*. *Biochim Biophys Acta* **1783:** 2130–2136.

Chen X. 2007. A marked end. *Nat Struct Mol Biol* **14:** 259–260.

Coyne RS, Thiagarajan M, Jones KM, Wortman J, Tallon LJ, Haas BJ, Cassidy-Hanley DM, Wiley EA, Smith JJ, Collins K, et al. 2008. Refined annotation and assembly of the *Tetrahymena thermophila* genome sequence through EST analysis, comparative genomic hybridization, and targeted gap

closure. *BMC Genomics* **9:** 562. doi: 10.1186/1471-2164-9-562.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, et al. 2008. An endogenous small interfering RNA pathway in *Drosophila*. *Nature* **453:** 798–802.

Deshpande T, Takagi T, Hao L, Buratowski S, Charbonneau H. 1999. Human PIR1 of the protein-tyrosine phosphatase superfamily has RNA 5′-triphosphatase and diphosphatase activities. *J Biol Chem* **274:** 16590–16594.

Duchaine TF, Wohlschlegel JA, Kennedy S, Bei Y, Conte DJ, Pang K, Brownell DR, Harding S, Mitani S, Ruvkun G, et al. 2006. Functional proteomics reveals the biochemical niche of *C. elegans* DCR-1 in multiple small-RNA-mediated pathways. *Cell* **124:** 343–354.

Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol* **4:** e286. doi: 10.1371/journal.pbio.0040286.

Farazi TA, Juranek SA, Tuschl T. 2008. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **135:** 1201–1214.

Gaertig J, Thatcher TH, Gu L, Gorovsky MA. 1994. Electroporation-mediated replacement of a positively and negatively selectable b-tubulin gene in *Tetrahymena thermophila*. *Proc Natl Acad Sci* **91:** 4549–4553.

Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: An expanding universe. *Nat Rev Genet* **10:** 94–108.

Höck J, Meister G. 2008. The Argonaute protein family. *Genome Biol* **9:** 210. doi: 10.1186/gb-2008-9-2-210.

Howard-Till RA, Yao MC. 2006. Induction of gene silencing by hairpin RNA expression in *Tetrahymena thermophila* reveals a second small RNA pathway. *Mol Cell Biol* **26:** 8731–8742.

Jinek M, Doudna JA. 2009. A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457:** 405–412.

Katiyar-Agarwal S, Gao S, Vivian-Smith A, Jin H. 2007. A novel class of bacteria-induced small RNAs in *Arabidopsis*. *Genes & Dev* **21:** 3123–3134.

Katoh T, Sakaguchi Y, Miyauchi K, Suzuki T, Kashiwabara S, Baba T, Suzuki T. 2009. Selective stabilization of mammalian microRNAs by 3′ adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes & Dev* **23:** 433–438.

Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10:** 126–139.

Klattenhoff C, Theurkauf W. 2008. Biogenesis and germline functions of piRNAs. *Development* **135:** 3–9.

Kurth HM, Mochizuki K. 2009. 2′-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in *Tetrahymena*. *RNA* **15:** 675–685.

Lee SR, Collins K. 2006. Two classes of endogenous small RNAs in *Tetrahymena thermophila*. *Genes & Dev* **20:** 28–33.

Lee SR, Collins K. 2007. Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. *Nat Struct Mol Biol* **14:** 604–610.

Lee SR, Talsky KB, Collins K. 2009. A single RNA-dependent RNA polymerase assembles with mutually exclusive nucleotidyl transferase subunits to direct different pathways of small RNA biogenesis. *RNA* **15:** 1363–1374.

Li J, Yang Z, Yu B, Liu J, Chen X. 2005. Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in *Arabidopsis*. *Curr Biol* **15:** 1501–1507.

Liu Y, Mochizuki K, Gorovsky MA. 2004. Histone H3 lysine 9 methylation is required for DNA elimination in developing macronuclei in *Tetrahymena*. *Proc Natl Acad Sci* **101:** 1679–1684.

Malone CD, Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136:** 656–668.

Malone CD, Anderson AM, Motl JA, Rexer CH, Chalker DL. 2005. Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of *Tetrahymena thermophila*. *Mol Cell Biol* **25:** 9151–9164.

Miao W, Xiong J, Bowen J, Wang W, Liu Y, Braguinets O, Grigull J, Pearlman RE, Orias E, Gorovsky MA. 2009. Microarray analyses of gene expression during the *Tetrahymena thermophila* life cycle. *PLoS One* **4:** e4429. doi: 10.1371/journal.pone.0004429.

Mochizuki K, Gorovsky MA. 2004. Small RNAs in genome rearrangement in *Tetrahymena*. *Curr Opin Genet Dev* **14:** 181–187.

Mochizuki K, Gorovsky MA. 2005. A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. *Genes & Dev* **19:** 77–89.

Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. 2002. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* **110:** 689–699.

Molnár A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC. 2007. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature* **447:** 1126–1129.

Nilsen TW. 2008. Endo-siRNAs: Yet another layer of complexity in RNA silencing. *Nat Struct Mol Biol* **15:** 546–548.

Okamura K, Lai EC. 2008. Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* **9:** 673–678.

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

Schoeftner S, Blasco MA. 2008. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat Cell Biol* **10:** 228–236.

Seto AG, Kingston RE, Lau NC. 2007. The coming of age for Piwi proteins. *Mol Cell* **26:** 603–609.

Shang Y, Song X, Bowen J, Corstanje R, Gao Y, Gaertig J, Gorovsky MA. 2002. A robust inducible-repressible promoter greatly facilitates gene knockouts, conditional expression, and overexpression of homologous and heterologous genes in *Tetrahymena thermophila*. *Proc Natl Acad Sci* **99:** 3734–3739.

Sijen T, Steiner FA, Thijssen KL, Plasterk RH. 2007. Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science* **315:** 244–247.

Siomi H, Siomi MC. 2009. On the road to reading the RNA-interference code. *Nature* **457:** 396–404.

Ullu E, Lujan HD, Tschudi C. 2005. Small sense and antisense RNAs derived from a telomeric retroposon family in *Giardia intestinalis*. *Euk Cell* **4:** 1155–1157.

Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crété P. 2004. Endogenous *trans*-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. *Mol Cell* **16:** 69–79.

Witkin KL, Collins K. 2004. Holoenzyme proteins required for the physiological assembly and activity of telomerase. *Genes & Dev* **18:** 1107–1118.

Witkin KL, Prathapam R, Collins K. 2007. Positive and negative regulation of *Tetrahymena* telomerase holoenzyme. *Mol Cell Biol* **27:** 2074–2083.

Zhang H, Ehrenkaufer GM, Pompey JM, Hackney JA, Singh U. 2008. Small RNAs with 5′-polyphosphate termini associate with a Piwi-related protein and regulate gene expression in the single-celled eukaryote *Entamoeba histolytica*. *PLoS Pathog* **4:** e1000219. doi: 10.1371/journal.ppat.1000219.

Zhao T, Li G, Mi S, Li S, Hannon GJ, Wang XJ, Qi Y. 2007. A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes & Dev* **21:** 1190–1203.