Title
Piwi-bound Small RNAs in Tetrahymena thermophila

Author
Couvillion, Mary Therese

Publication Date
2011

# Piwi-bound Small RNAs in *Tetrahymena thermophila*

by

Mary Therese Couvillion

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kathleen Collins, Chair
Professor Gary Karpen
Professor Lin He
Professor John Taylor

Spring 2011

# ABSTRACT

Piwi-bound Small RNAs in *Tetrahymena thermophila*
by
Mary Therese Couvillion

Doctor of Philosophy in Molecular and Cell Biology
University of California, Berkeley
Professor Kathleen Collins, Chair

Small RNA (sRNA)-mediated silencing is a strategy used by almost all eukaryotes to regulate gene expression. The only component common to all sRNA-mediated silencing pathways is a PAZ, PIWI domain-containing (PPD) protein, more commonly called Argonaute/Piwi (Ago/Piwi), after the founding members of the family, which have the same domain structure but fall into two subfamilies based on primary sequence similarity. The conserved PAZ and PIWI domains are required for binding the 3' and 5' ends of the sRNA, respectively. The ribonucleoprotein (RNP) complex then sequence-specifically targets nucleic acid substrates through base pairing between the sRNA guide and the target.

Argonaute subfamily proteins and bound sRNAs are ubiquitously expressed in many organisms and have been intensely studied over the past decade. Several mechanisms used by Argonuates to silence gene expression are known in detail. Piwi subfamily proteins and bound RNAs on the other hand are germline-restricted in many organisms and have been intensely studied only for the last five years. They are known to be important for silencing transposons, but detailed mechanisms of their action and other roles are still unknown.

*Tetrahymena thermophila* expresses eight distinct Piwi-subfamily proteins, no Argonaute subfamily proteins, and no transposons, allowing simplified discovery of other Piwi-bound sRNA classes. I have used protein tagging and affinity purification approaches along with deep sequencing to characterize sRNAs bound to each of these *Tetrahymena* Piwis (Twis), and have uncovered a great diversity of sRNA classes in this organism (Chapter One). In particular, the only essential Piwi in *Tetrahymena*, Twi12, specifically binds tRNA fragments (Chapter Two). My further studies of Twi12 have revealed its association with the 5' to 3' exoribonuclease Xrn2 in the nucleus (Chapter Three). In the final chapter I describe biochemical techniques and approaches useful for studies of RNPs in *Tetrahymena* (Chapter Four).

# ACKNOWLEDGMENTS

Thank you to everyone who helped to make this thesis possible. Above all, Kathy Collins who has been not just an advisor, but also a mentor and role model. I feel so lucky to have learned from someone with such a vast knowledge of biochemistry, a brilliant mind, and endless energy, combined with a talent for teaching and a mentoring style so carefully personalized to each student. Thank you Kathy, for supporting me in so many ways.

My thesis committee has also helped me along the way. Gary Karpen sparked my interest in small RNAs during my rotation in his lab, and has always offered a welcome and constructive critical perspective. Lin He taught me about the history of the small RNA field, as well as many other useful lessons and skills, and helped me focus on the most promising approaches. John Taylor has always been encouraging and reminded me starting early on to think about my future plans.

I could not ask for better colleagues than my labmates in the Collins lab over the years. Suzanne Lee took me under her wing when I started my rotation and taught me how to do everything. I have always strived to be as careful and thoughtful as Suzanne. Kristin Talsky has been my constant partner through the years on Team RNA Silencing, and Brandon Hogstad was a rich source of ideas and contributed so much during his time in the lab. Kyungah Hong has become a fast expert at many techniques and has been extremely helpful. The newest member of the RNA silencing team, Nicole Beier, has brought so much enthusiasm and will surely continue to progress our understanding of this system. I also want to thank Kasper Andersen and George Katibah ("tRNA club") for providing so many thought provoking, illuminating, interesting and fun conversations about science in general. Of the telomerase guys, I could always count on Aaron Robart for an interesting new recipe, and on Alec Sexton for a clever joke. I look forward to getting to know the newest member of the lab, Alex Wu. Finally, I want to thank Bosun Min, who was my baymate through most of my time in graduate school and taught me something new almost every day, about science or otherwise, and brought me so many delectable treats. Also Emily Egan and Barbara Eckert, who have been caring friends and confidants as well as incredibly smart women who I admire and look to for advice both in and out of lab.

My family and friends outside of lab have also been very supportive. I especially want to thank Abby Daane and Jesse Noffsinger for frequent fun, dinners, adventures, and sanity. Brad Malone has been my partner in life since before I even chose a major in college. It is because of his active support, involvement, and strict encouragement to go after what I want, that I am where I am today.

# TABLE OF CONTENTS

# CHAPTER ONE

## Sequence, Biogenesis, and Function of Diverse Small RNA Classes Bound to the Piwi-family Proteins of *Tetrahymena thermophila*

Based on Couvillion et al., Genes & Development 2009

### Abstract

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 23-24 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.

## Introduction

Eukaryotic ~20-30 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 double-stranded RNA (dsRNA) cleavage by enzymes of the RNase III family (Carthew and Sontheimer, 2009; Kim et al., 2009). Small interfering RNAs (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 (Ghildiyal and Zamore, 2009; Nilsen, 2008). 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 *C. 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, single-stranded RNA precursors (Kim et al., 2009; Klattenhoff and Theurkauf, 2008; Malone and Hannon, 2009; Seto et al., 2007). The piRNAs of muticellular eukaryotes show developmentally restricted accumulation, detected predominantly in germ-line, germ-line supportive, or self-renewing cell lineages. Where possible to discern, the piRNAs associated with a particular Piwi-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 RNaseH-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; Hock 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 RNA interference (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 (Klattenhoff and Theurkauf, 2008; Malone and Hannon, 2009; Seto et al., 2007). In animals, Piwi proteins function in particular stages of germ-line 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 open reading frames for up to twelve 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, Dcr2) and a single Rdr (Rdr1) that is assembled into multiple RDRCs (Lee and Collins, 2006; Lee and Collins, 2007; Lee et al., 2009a; Malone et al., 2005; Mochizuki and Gorovsky, 2005).

*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, germ-line 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 27-30 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 lysine 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; Taverna et al., 2002).

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 27-30 nt scnRNAs that guide heterochromatin formation and DNA elimination in conjugating cells, a class of 23-24 nt sRNAs is constitutively expressed (Lee and Collins, 2006). Limited sequencing of bacterially cloned *T. thermophila* 23-24 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* 23-24 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 23-24 nt sRNAs in vitro and in vivo (Lee and Collins, 2007; Lee et al., 2009a). *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; (Lee et al., 2009a). 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 23-24 nt sRNAs (Lee et al., 2009a).

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 23-24 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 twelve *T. thermophila TWI* genes (Fig. 1A; additional data not shown). For intron-containing genes, we first cloned mRNAs to determine correct ORF sequences (see 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 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 (Miao et al., 2009; Mochizuki et al., 2002), 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 kbp *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 more than 90% identity, a greater level than shared by any other pair-wise 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 non-essential genes. Strains targeted to disrupt a non-essential 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, 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 non-essential subunits of compositionally distinct RDRCs sharing essential Rdr1 and Dcr2 (Rdf1KO, Rdf2KO, 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* 23-24 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 (Katoh et al., 2009; Li et al., 2005). 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 a 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, non-essential beta-tubulin 1 gene of strain CU522, an efficient strategy for positive selection (Gaertig et al., 1994b). 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 27-30 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 23-24 nt sRNAs that were offset from each other by an approximately half-nt step of migration on gels with sufficient resolution. Twi7 and Twi10 copurified 23-24 nt sRNAs and also larger ~32-34 nt or ~33-36 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 beta-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 27-30 nt sRNAs are modified by ribose methylation at their 3' end, which prevents beta-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 23-24 nt sRNAs and Twi-bound 23-24 nt sRNA pools were generally increased in mobility following beta-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 27-30 nt sRNAs from conjugating cells were also resistant to beta-elimination (Fig. 1F), as recently shown for total 27-30 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; Zhang et al., 2008). *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 23-24 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 22-24 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 22-25 nt RNA from size-enriched total

RNA of the cognate background strain CU522. In addition, we gel-purified 22-25 nt RNA from size-enriched total RNA of the gene knockout strains individually lacking each non-essential 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 23-24 nt sRNA loci contribute most abundantly to the total sRNA population (Lee and Collins, 2006), but these strains derive from the same inbred parent 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' nt 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). Non-mappers 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 non-mappers 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 22-24 nt mappers), we analyzed the frequency and identity of untemplated 3' nt 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 22-24 nt sRNAs. To evaluate the sequence complexity of each library, we calculated the percentage of total genome-mapping 22-24 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 5 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 to 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 23-24 nt sRNAs (Lee et al., 2009a).

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 22-24 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 kbp. 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. Small RNA clusters identified and subsequently analyzed are listed in Supplemental Table S1. Ultimately we identified one or more sRNA classes associated with Twi2, Twi7, Twi8, or Twi10 that together are likely to constitute the overall pool of 23-24 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 ~30-85 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 through 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 position 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 non-consecutive 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, 3D; top panels are Northern blots and bottom panels are gels before transfer stained with SYBR Gold to show 23-24 nt sRNA loading). Some enrichment by Twi7 was also detectable, although this was substantially less than enrichment by Twi2 when normalized to loading of 23-24 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 23-24 nt sRNAs that mapped to degenerate repeat loci with interrupted tandem arrays of ~ 150 bp repeat units, spanning 2-20 kbp at any given genome location (Figs. 4A, 4B). 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 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).

We also found sRNAs cognate to low-copy repeats present at fewer than 10 genome locations, with repeat units of 100-300 bp spanning 2-4 kbp. 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 22-24 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, ~32-34 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 exclusively the G-rich strand and begin with fixed phasing in one of six possible permutations of the $T_2G_4$ repeat: 5'-UGGGGU (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 23-24 nt and ~33-36 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 (Allen et al., 2005; Molnar et al., 2007; Vazquez et al., 2004; Zhao et al., 2007) or from both sides of an RNA structure (Czech et al. 2008; Babiarz 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 23-24 nt sRNAs, but their accumulation does not strictly require the presence of Twi2 or Twi8 (Fig. 5A, 5B). 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 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 22-24 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 23-26 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 23-24 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* (Ghildiyal and Zamore, 2009; Nilsen, 2008). *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* (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 (compare 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 22-24 nt sRNAs in growing cells. Several classes of abundant sRNAs show unique specificities of Twi protein association and distinct genetic requirements for accumulation (Figs. 7A, 7B). 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.

## Discussion

Small RNAs 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 non-coding 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 easily detected (Lee and Collins, 2006). 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 non-essential, 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 precursor transcripts. Based on these observations and Twi8 localization to the MAC rather than the cytoplasm, it seems likely that Twi8 mediates co-transcriptional 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 Twi-8-mediated transcriptional regulation may be a less repressive and/or more dynamic form of transcriptional 'dampening' compared to 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 *A. thaliana* triggers the production of long ~30-40 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 20-30 nt sRNA library from *Giardia intestinalis* (Ullu et al., 2005), but to our knowledge G-strand telo-sRNAs have not been previously reported 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* 23-24 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 (Siomi and Siomi, 2009);(Jinek and Doudna, 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 utilization of different RDRC assemblies that share the same Rdr1 and Dcr2 enzymes. *T. thermophila* Rdf2 is the only non-essential RDRC subunit preferentially expressed in vegetative growth (Lee et al., 2009a). 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., 2009a).

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 fresh-water ponds. Growth in sterile rich media under laboratory conditions may allow some otherwise critical pathways of sRNA-mediated regulation to become non-essential, 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 germ-line 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 x $10^5$ 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% non-fat 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 (1-5 x $10^5$ cells/mL), 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 23-24 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. Beta-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 amino-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 amino-terminal tag with tandem Protein A domains and a TEV protease cleavage site. The tag was preceded by an ~1 kbp 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 (1-5 x $10^5$ cells/mL). 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 ~3 x $10^5$ cells/mL 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 40X objective of an Olympus model BX61 microscope equipped with a Hamamatsu Digital camera. Images were captured using Metamorph software.

*Small RNA cloning, sequencing, and analysis*

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 downloaded from 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.

*Library preparation*

Twi-bound sRNA libraries were cloned as described previously (Brennecke et al., 2007). SB210, Rdf1KO, Rdf2KO, Rdn2KO total sRNA libraries were cloned similarly except for use of SYBR Gold staining during gel purification steps. Gel-purified sRNAs (40-100 ng) were first ligated to 50 pmol pre-adenylated 3' adaptor (IDT-miRNA cloning linker 1) using truncated T4 RNA Ligase 2 (NEB) in 50 mM Tris-HCl at pH 7.5, 10 mM $MgCl_2$, 10 mM DTT, and 60 µg/mL BSA. Ligated products were resolved, SYBR Gold stained, and gel purified as described above, then ligated to 50 pmol RNA 5' adaptor using T4 RNA ligase (Ambion) in buffer provided and 1/10 volume DMSO. Ligated products were again gel-purified and reverse transcribed with SuperScript II (Invitrogen) and RT primer. First strand cDNA was amplified with Taq polymerase using RT primer and PCR primer, with an annealing temperature of 54°C for 5 cycles and 60°C for 14 cycles. After phenol:chloroform extraction and ethanol precipitation, products were separated from primers on a 2% low-melt agarose gel. Products were retrieved by melting agarose slices in 0.3 M NaCl and extracting with warmed phenol, phenol:chloroform, then chloroform, followed by ethanol precipitation.

3' adaptor (IDT-miRNA cloning linker 1):
5'-rAppCTGTAGGCACCATCAATdideoxyC-3'
5' adaptor:
5'-rArCrArCrUrCrUrUrUrCrCrCrUrArCrArCrGrArCrGrCrUrCrUrUrCrCrGrArUrC-3'
RT primer:
5'-CAAGCAGAAGACGGCATACGATTGATGGTGCCTACAG-3'
PCR primer:
5'-AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACG-3'

*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. Small RNA libraries are deposited at Gene Expression Omnibus (accession number GSE17006, data sets GSM425486–GSM425496).

Figure 1, part 1

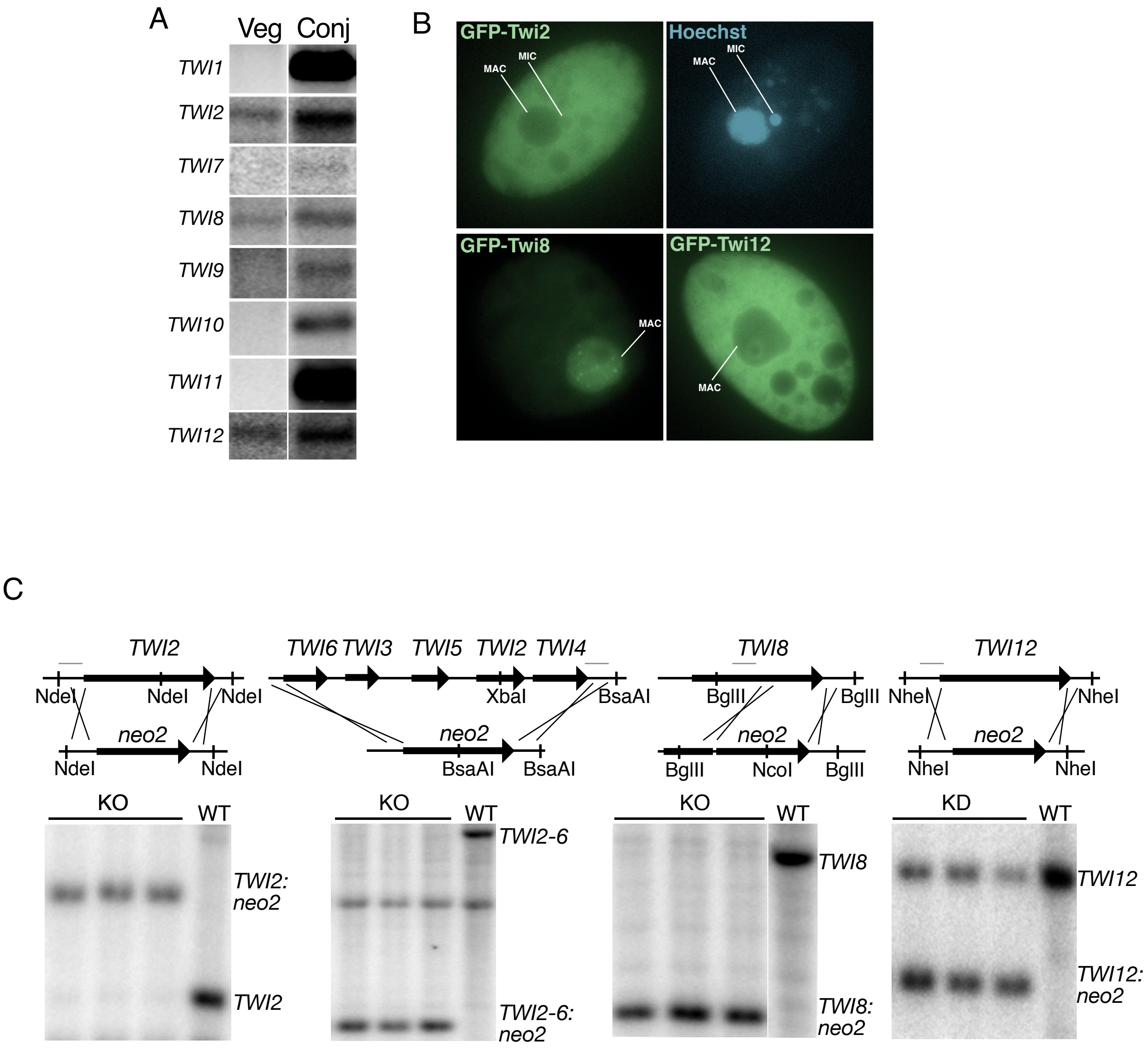


Figure 1, part 2

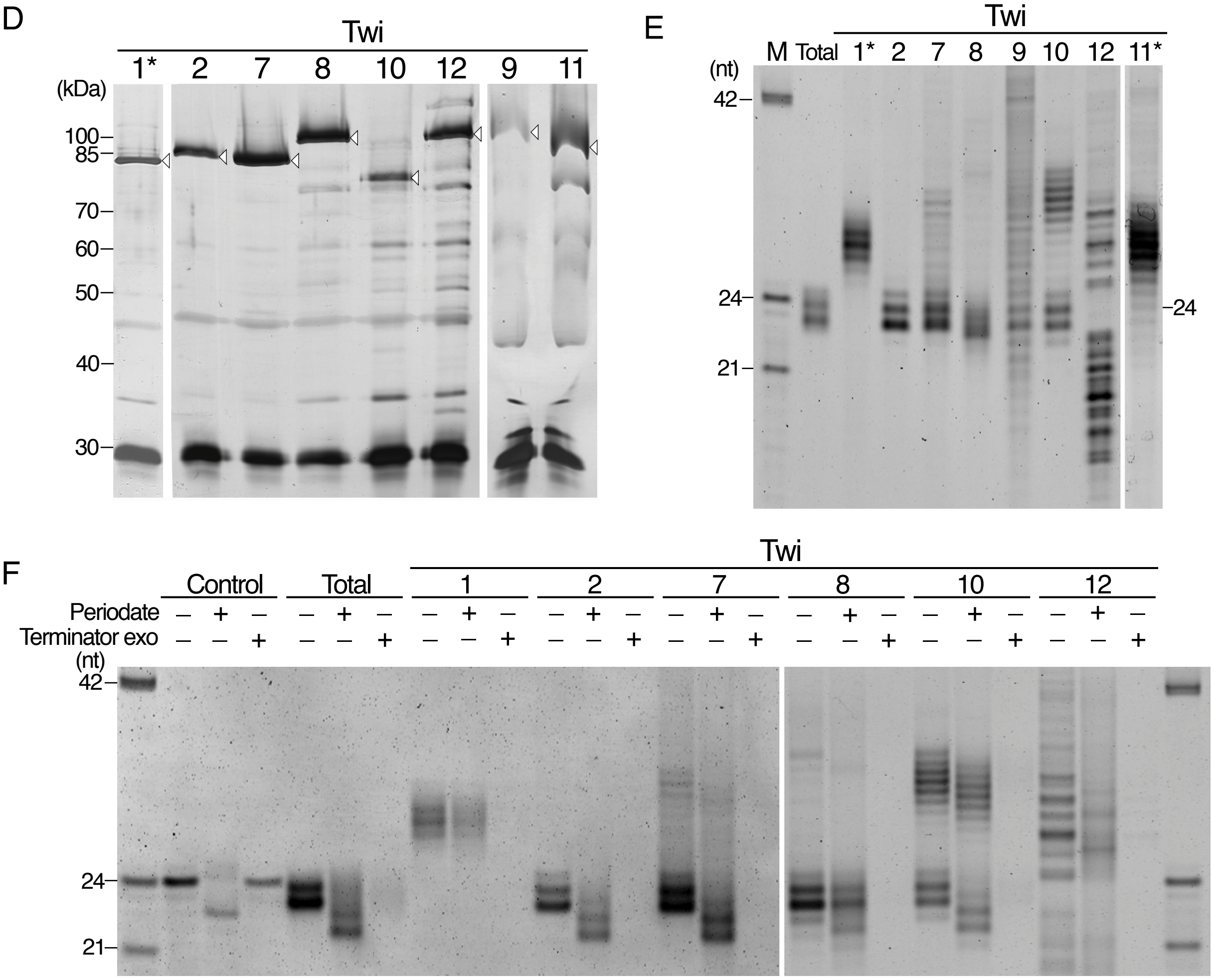

D
Twi
1* 2 7 8 10 12 9 11
(kDa)
100–
85–
70–
60–
50–
40–
30–
E
Twi
M Total 1* 2 7 8 9 10 12 11*
(nt)
42–
24–
21–
–24
F
Twi
Control
Total
1
2
7
8
10
12
Periodate – + – – + – – + – – + – – + – – + – – + – – + –
Terminator exo – – + – – + – – + – – + – – + – – + – – + – – +
(nt)
42–
24–
21–

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 grey line above the 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 indicates marker lane. Total represents gel-purified 23-24 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 beta-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.

Figure 2

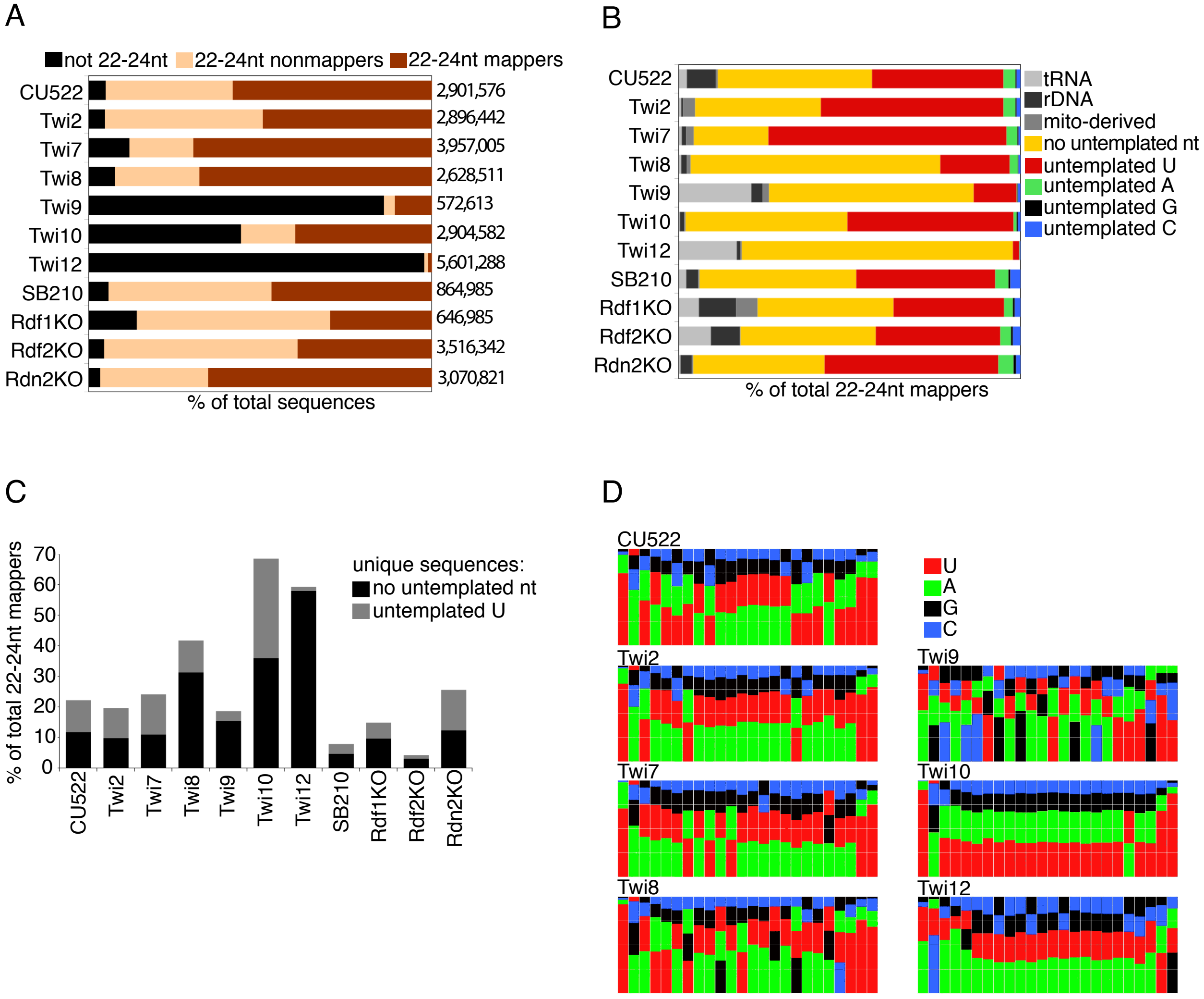


Figure 2. Deep sequencing library characteristics.
(A) Proportions of sequence types in each unfiltered library. The number of total sequences from each library is indicated at right.
(B) Proportions of 22-24 nt sequences in each library that mapped to the genome allowing a single 3' nt mismatch. Sequences indicated as rDNA 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 22-24 nt sRNA sequences from each library. Nucleotide frequency plots for sRNAs with a definitively untemplated 3' U are shown in Supplemental Figure S4.

Figure 3, part 1

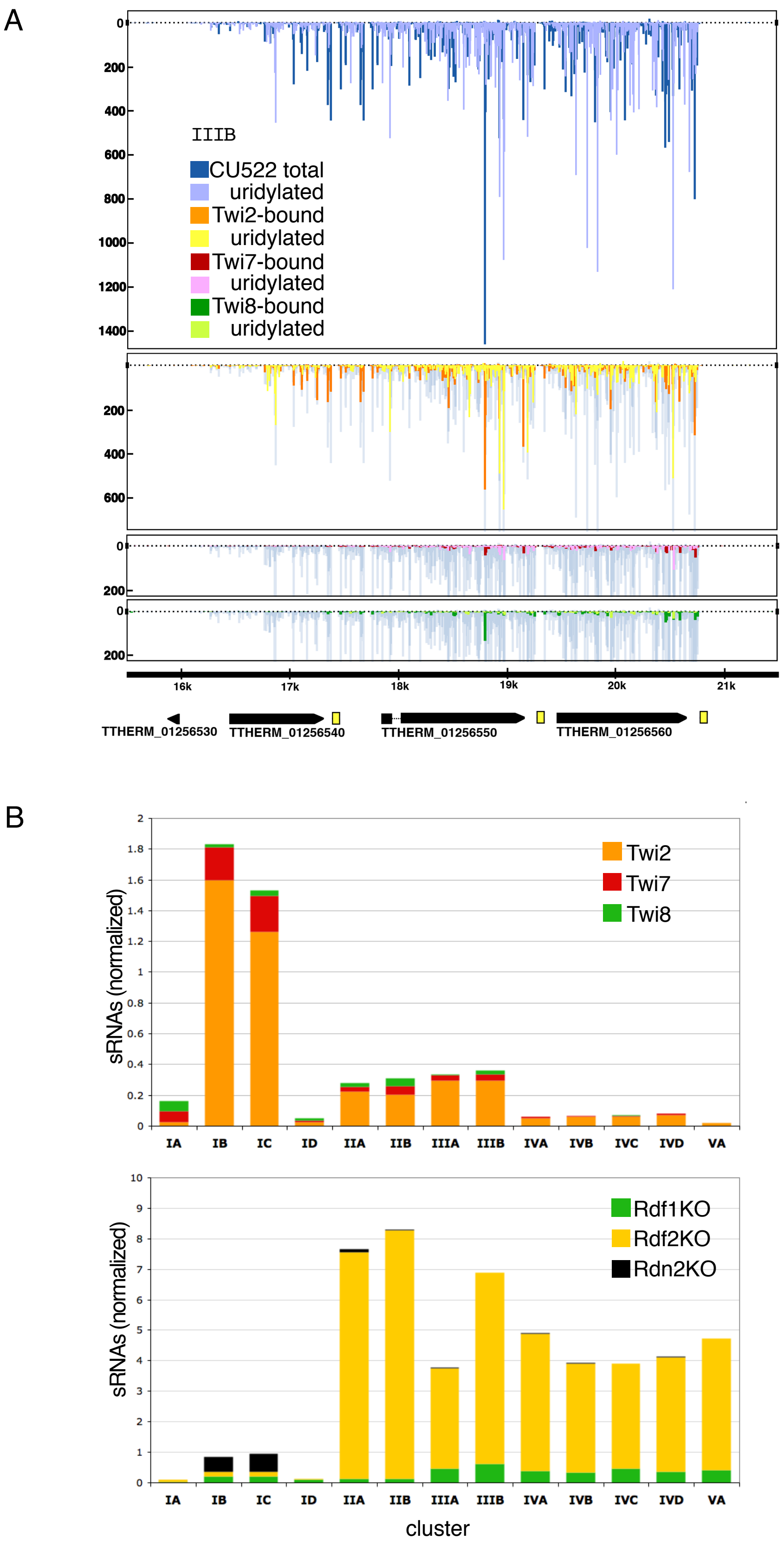


Figure 3, part 2

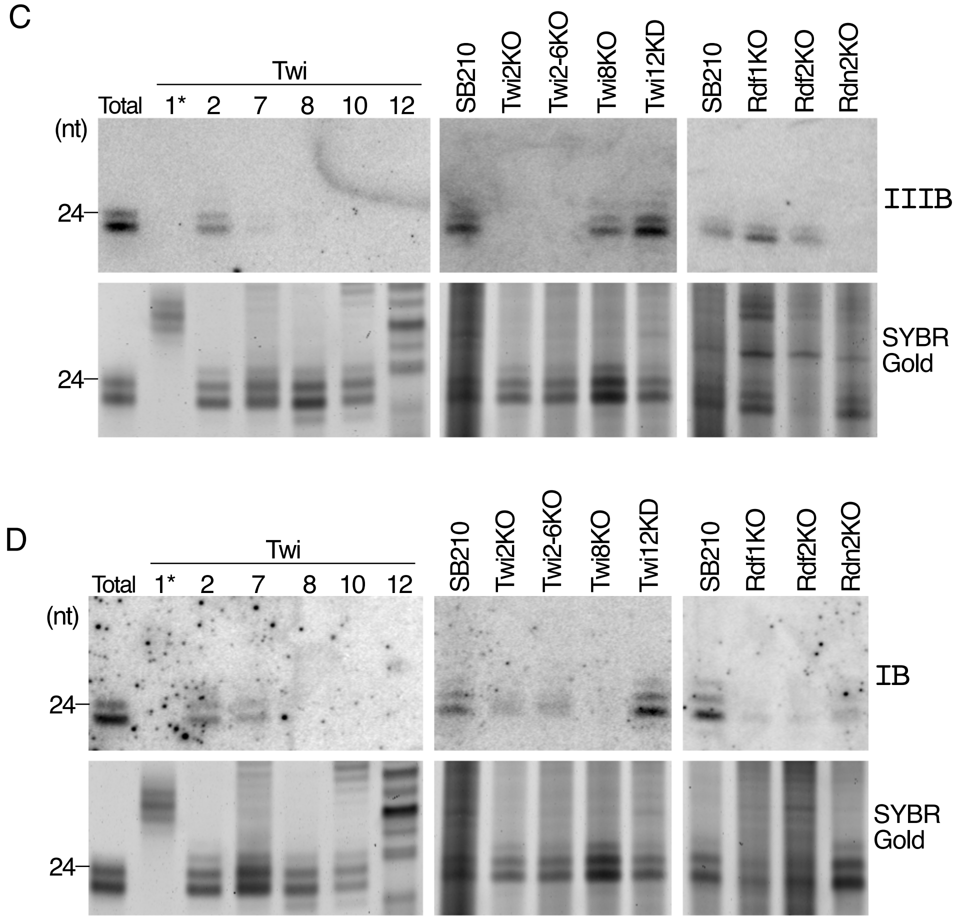


Figure 3. Pseudogene-derived sRNA clusters.
(A) Small RNA density plots for a representative pseudogene locus (IIIB). Density of sRNAs from each Twi-enriched library is shown superimposed on density from the total sRNA library of CU522. Peak height indicates the number of sRNA reads with the same 5' 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 number of filtered sequences. Loci are indicated by name of 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.

Figure 4, part 1

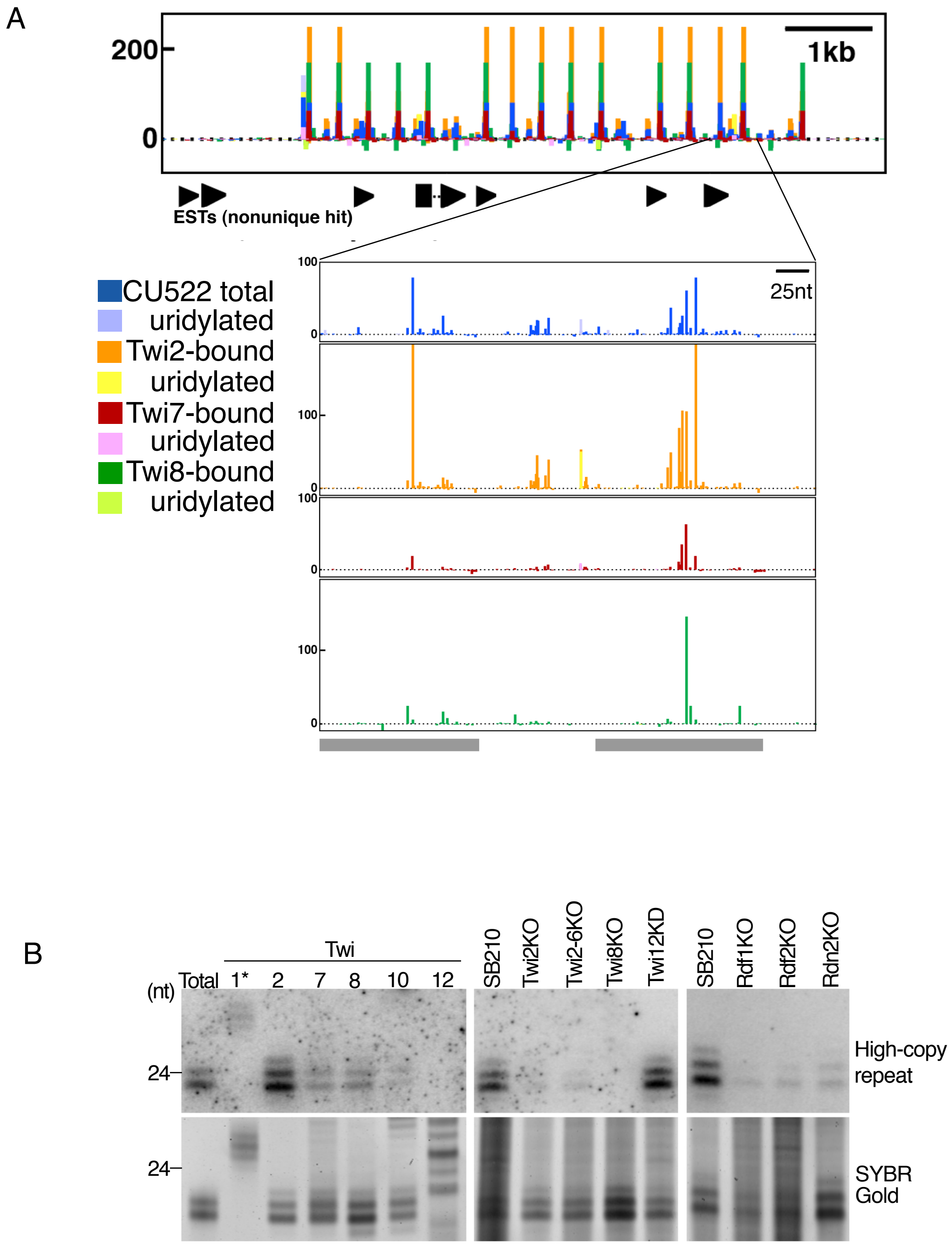


Figure 4, part 2

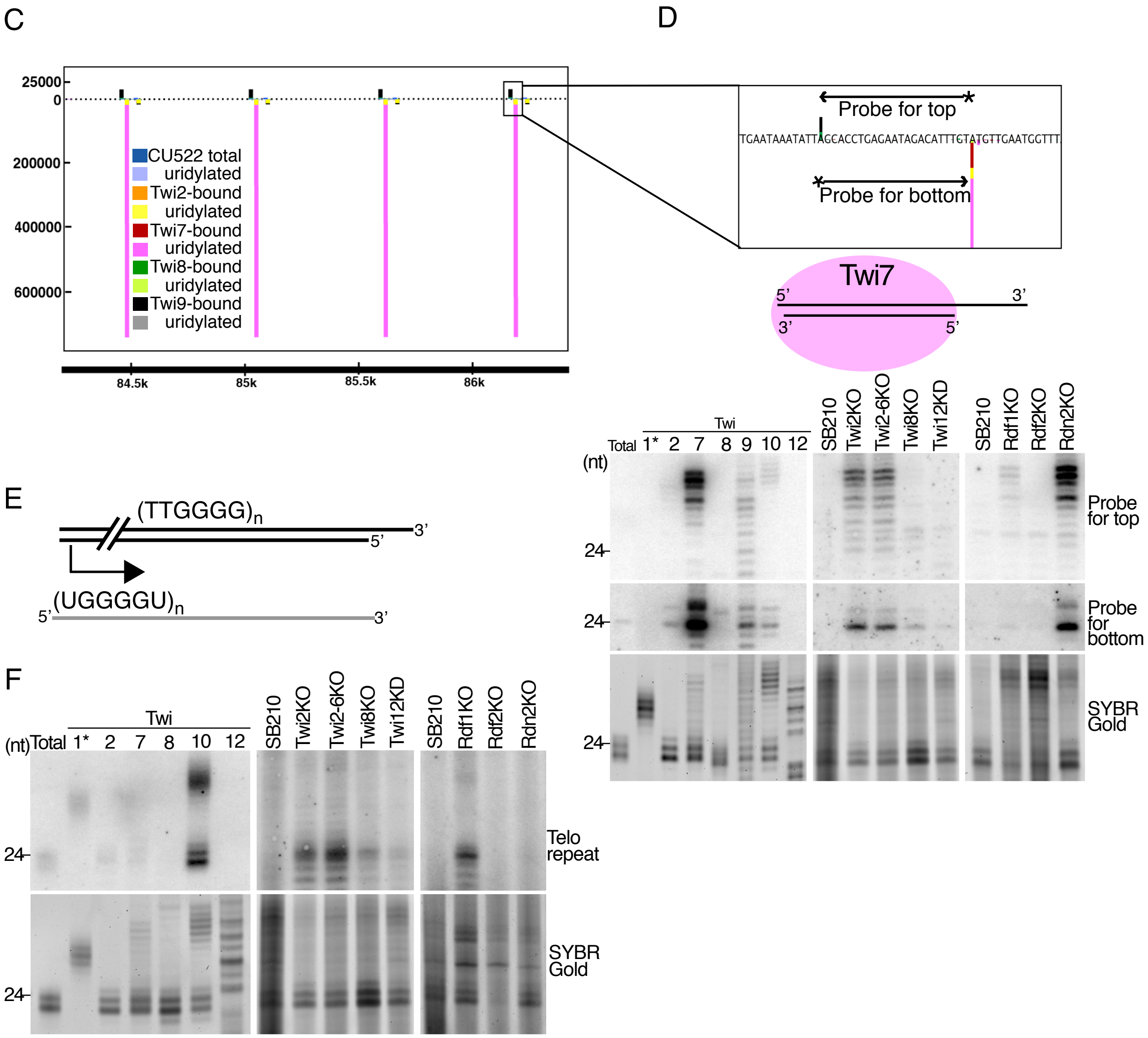


Figure 4. Repeat-derived sRNA clusters.
(A) Small RNA density plots at a high-copy repeat locus. Gray shaded bars represent the degenerate repeat unit.
(B) Northern blot hybridization with an oligonucleotide probe. Below the blots, RNA loading is shown by SYBR Gold staining.
(C) Small RNA density plot 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. 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.

Figure 5

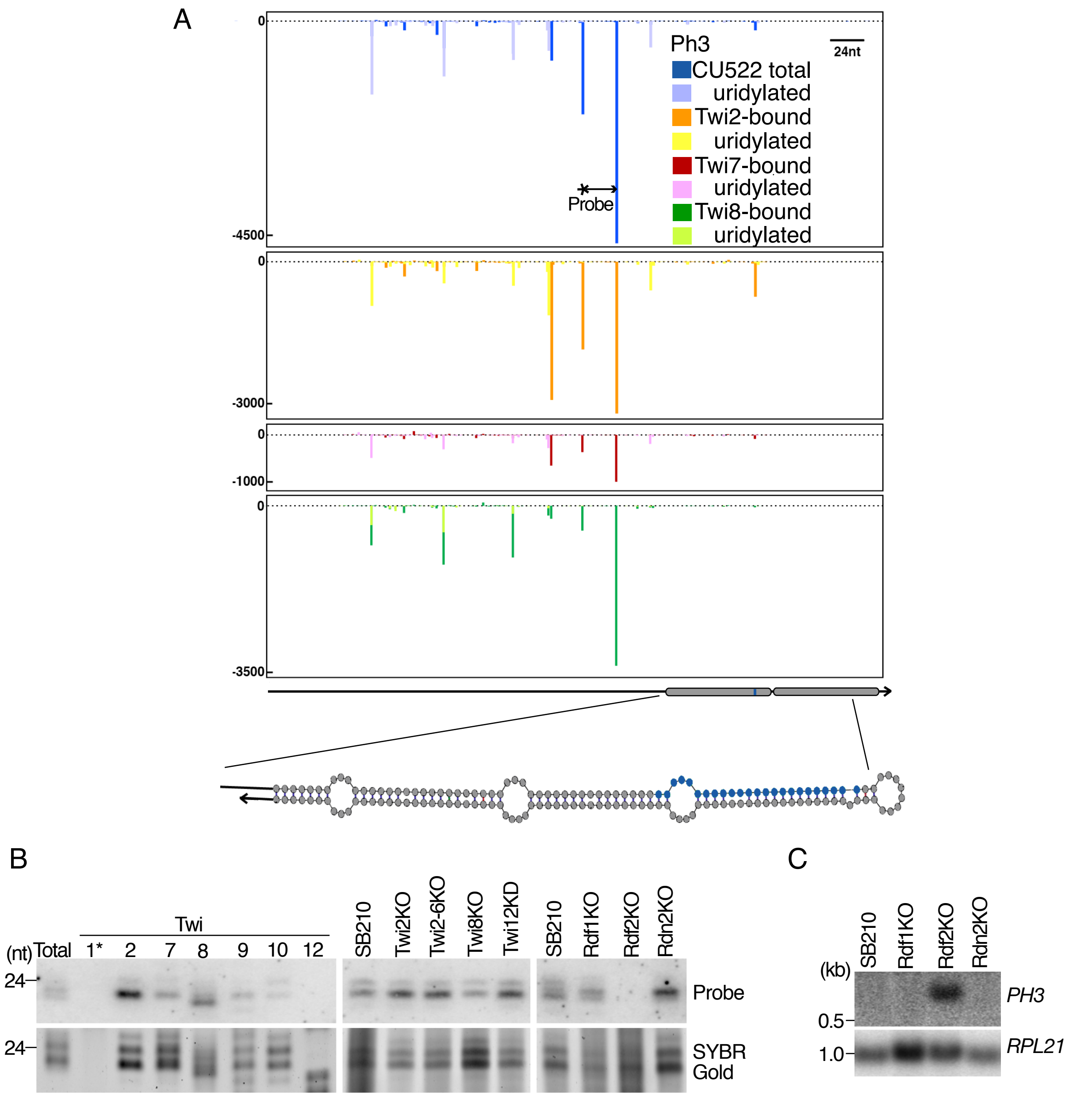


Figure 5. Phased sRNA clusters.
(A) Small RNA density plots at the phased cluster locus Ph3. A hypothetical sRNA-complementary transcript is shown below the plot as well as a predicted secondary structure for the region shaded in gray. Blue shading in the secondary structure depiction indicates the complement 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.

Figure 6, part 1

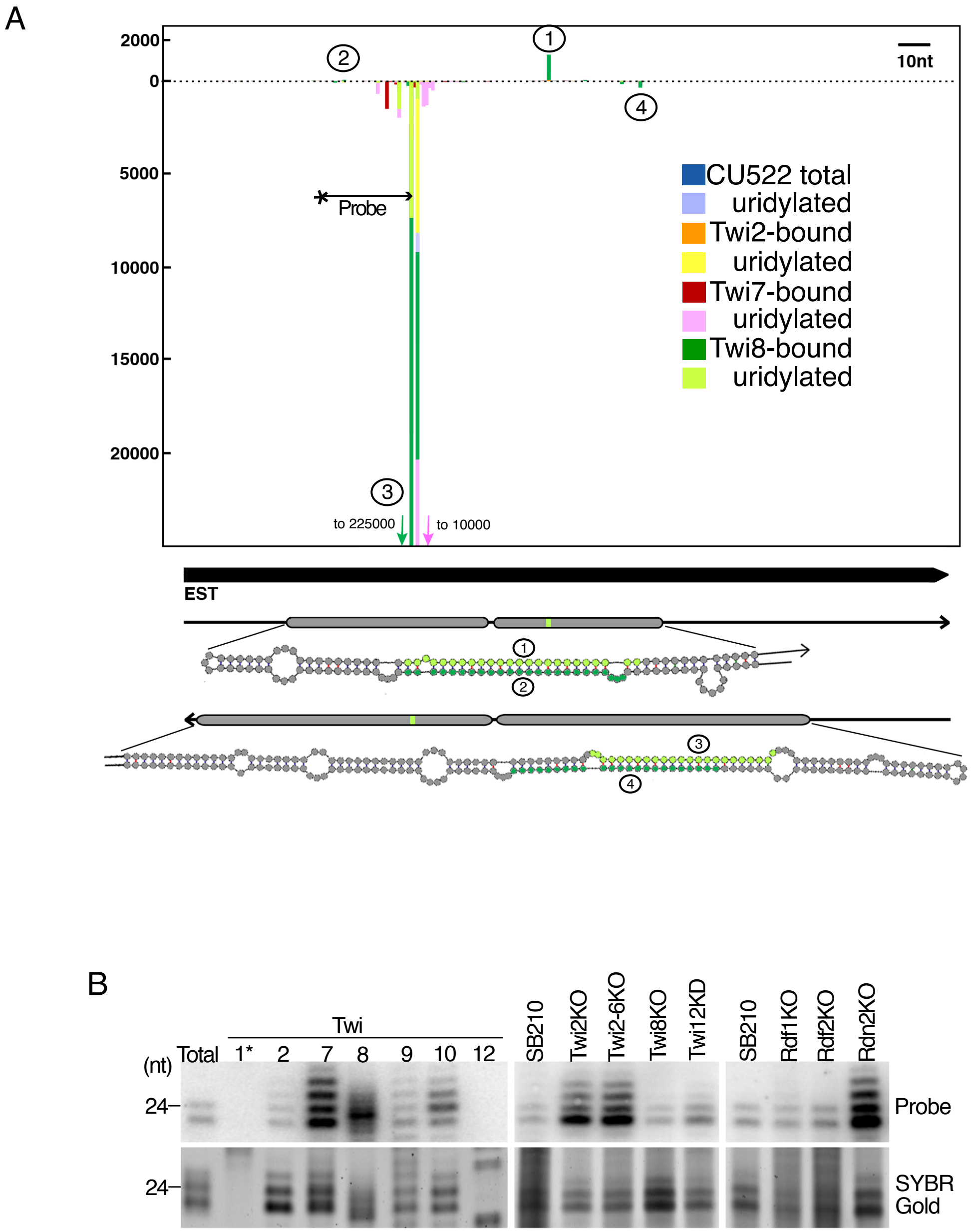


Figure 6, part 2

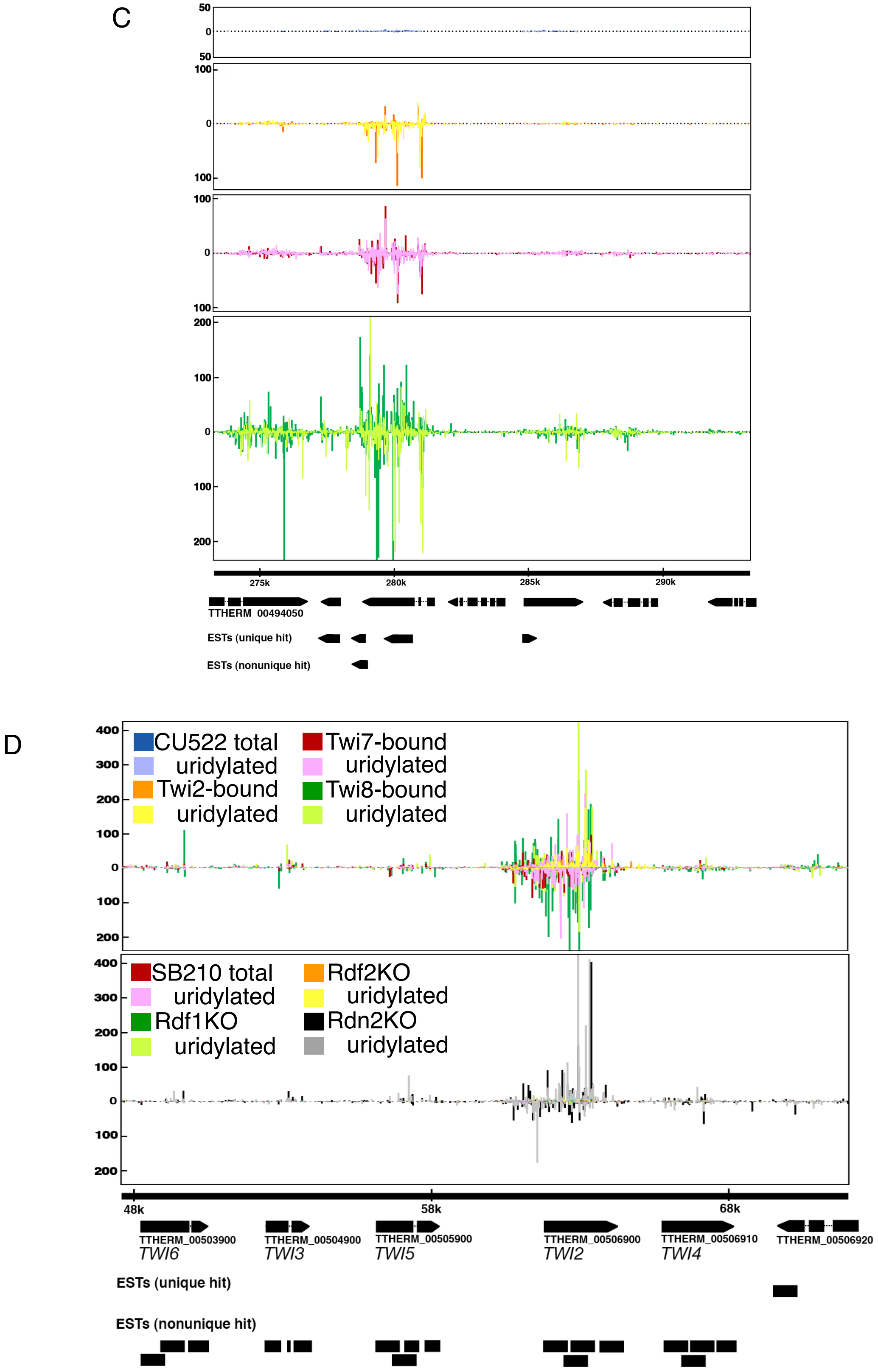


Figure 6, part 3

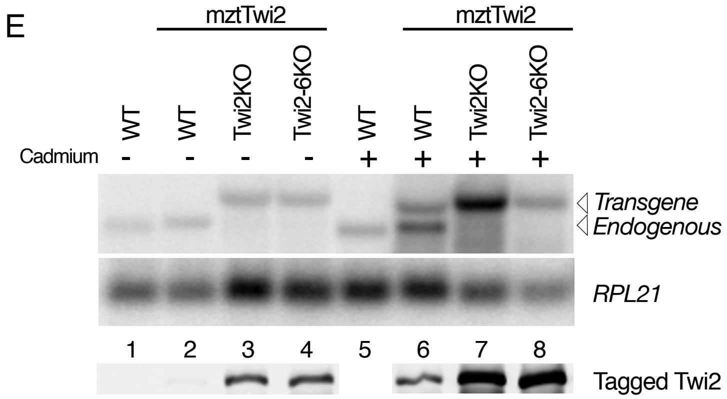


Figure 6. Twi8-asociated sRNA clusters.

(A) Small RNA 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) Small RNA density plots for a representative locus of Twi8-enriched sRNAs with convergently transcribed ORFs and/or ESTs that do not map uniquely.

(D) Small RNA 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. 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. Bottom panel is an immunoblot detecting tagged Twi2; equal loading was confirmed by total protein staining (not shown).

Figure 7

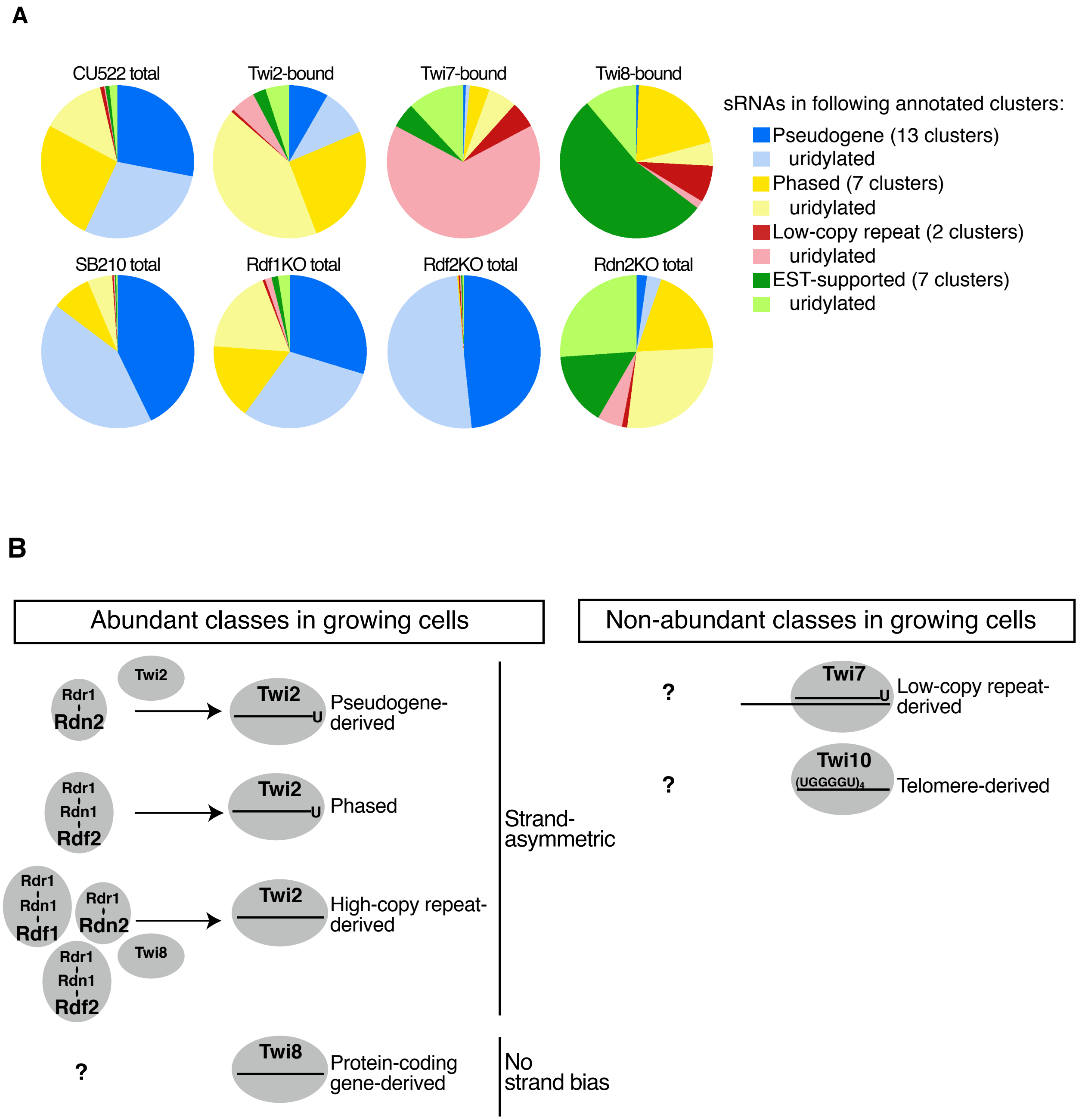
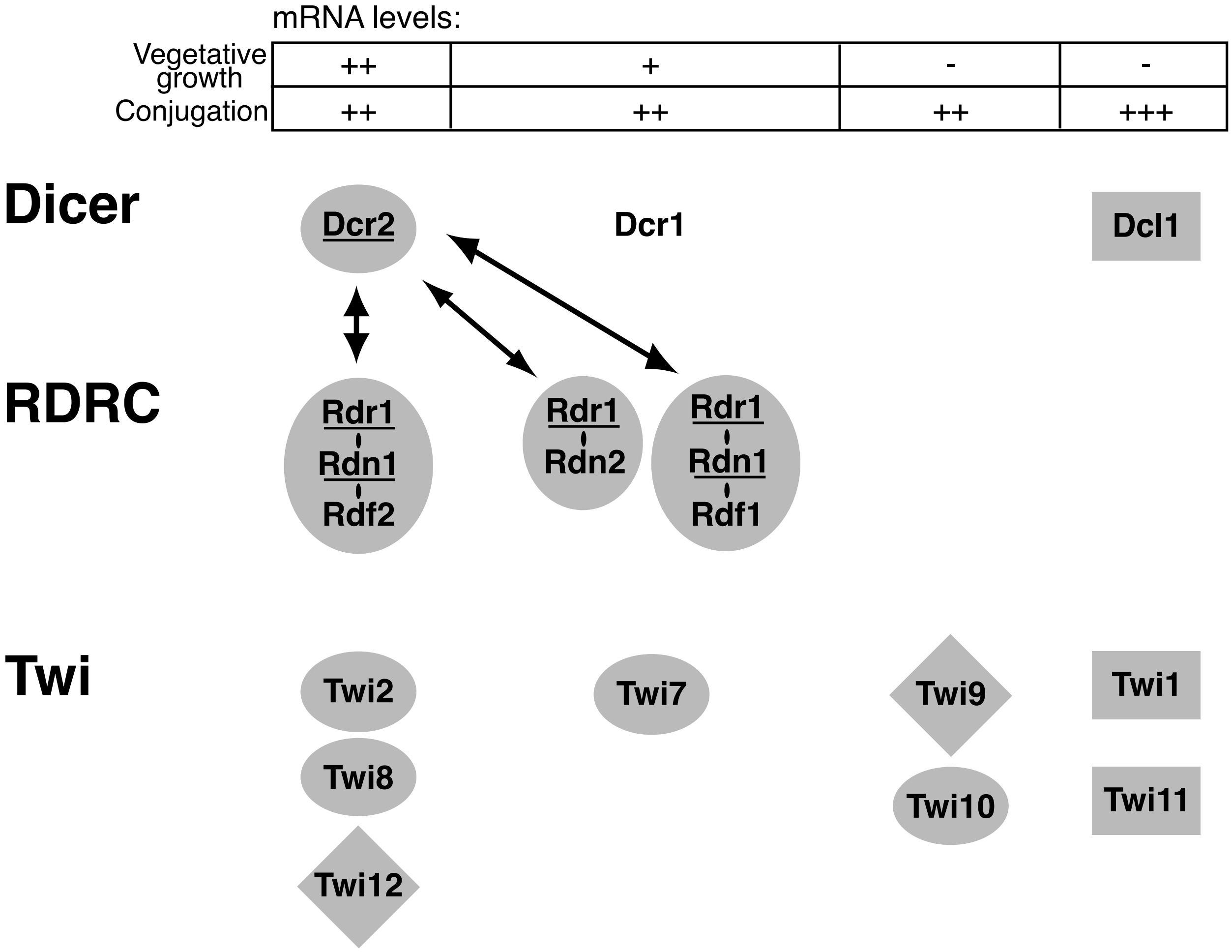


Figure 7. Summary of library representation and sRNA classes.
(A) Representation of abundant and preferentially enriched sRNA classes. Only annotated classes with large enough representation to generate a visible pie slice in at least one library are included; these classes are listed in the legend at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

Supplemental Figure S1

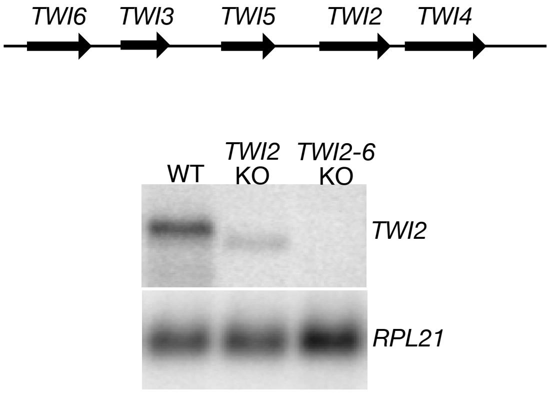


mRNA levels:

| | | | | |
|---|---|---|---|---|
| Vegetative growth | ++ | + | - | - |
| Conjugation | ++ | ++ | ++ | +++ |

Supplemental Figure S1. Illustration of characterized RNA silencing machinery in T. thermophila. Underlining indicates shared components of more than one complex. Circles indicate machinery involved in production of 23-24 nt sRNAs, rectangles indicate machinery involved in production of 27-30 nt sRNAs, and diamonds indicate Twi proteins that bind heterogeneously sized sRNAs. Arrows indicate known physical interactions.

Supplemental Figure S2



Supplemental Figure S2. Northern blot showing TWI2 expression and TWI2 probe cross-hybridization to a slightly smaller mRNA in the TWI2 knockout (KO) strain that is absent with knockout of TWI2-6. RPL21 expression is shown as a loading control. Gene structure at the wild-type TWI2-6 locus is illustrated.

Supplemental Figure S3

A

Identity scores (%)

Similarity scores (%)

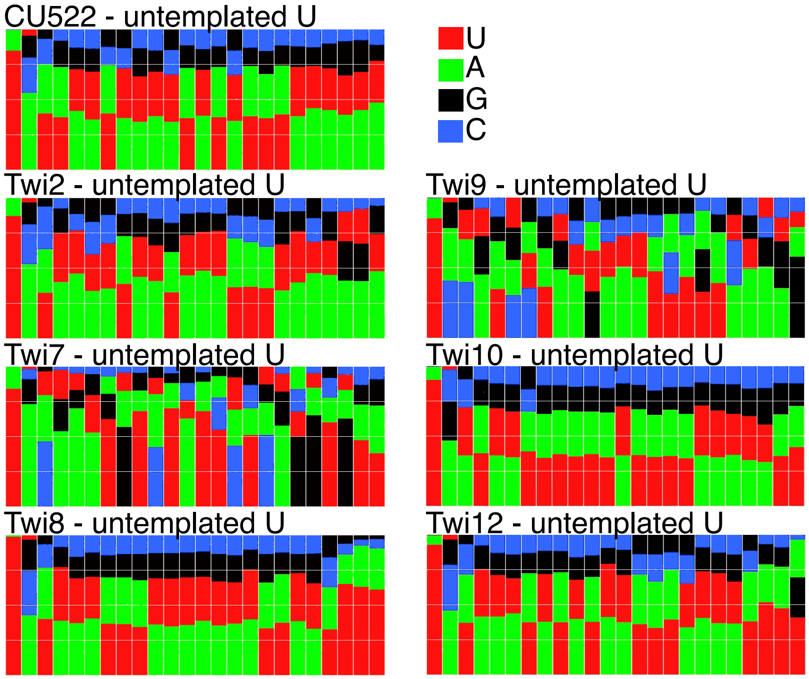
| | Twi1 | Twi2 | Twi4 | Twi7 | Twi8 | Twi9 | Twi10 | Twi11 | Twi12 |
|---|---|---|---|---|---|---|---|---|---|
| Twi1 | | 23 | 23 | 22 | 19 | 20 | 18 | 25 | 16 |
| Twi2 | 44 | | 91 | 63 | 26 | 40 | 29 | 23 | 17 |
| Twi4 | 44 | 96 | | 61 | 26 | 40 | 29 | 23 | 18 |
| Twi7 | 43 | 80 | 79 | | 26 | 39 | 29 | 22 | 17 |
| Twi8 | 36 | 44 | 44 | 43 | | 24 | 19 | 19 | 14 |
| Twi9 | 42 | 60 | 60 | 62 | 43 | | 28 | 20 | 16 |
| Twi10 | 37 | 48 | 48 | 49 | 37 | 47 | | 19 | 14 |
| Twi11 | 45 | 41 | 41 | 41 | 36 | 39 | 34 | | 13 |
| Twi12 | 35 | 34 | 34 | 34 | 29 | 32 | 30 | 31 | |

B

```
  HsAgo2   VIFLGADV   IFYRDGVS   IPAPAYYAHLVAF
  DmPiwi   LMTIGFDI   VFYRDGVS   VPAVCQYAKKLAT
  TtTwi1   TMVVGMDV   IIFRDGVG   TPSAVRYAHTLSN
  TtTwi2   TMIIGIET   IVYRQGLG   VPAAMKYAEKLAK
  TtTwi7   TMIVGIET   IIYRQGLG   VPAVMKYAEKLAK
  TtTwi8   TMIIGTSV   IYLRENIA   YPAQIQYAKKLAK
  TtTwi9   TMIVGIET   IIYKQGQG   VPAALKYAEKLAK
 TtTwi10   TMIIAIAV   IIYRQGLI   IPAQLKYACKLLK
 TtTwi11   TMLVGIDY   IIYRQAAN   IPSILKYAEQQTK
 TtTwi12   CTMIGFSV   IIIRDGIF   SPACVQNAYKLAE
```

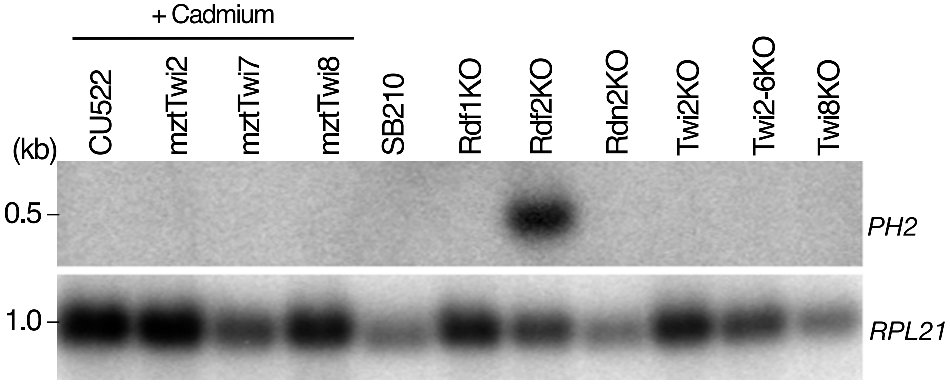
Supplemental Figure S3
(A) Pair-wise amino acid similarity/identity matrix for full-length Twi proteins.
(B) PIWI domain alignment showing residues surrounding the RNaseH-like catalytic triad (D-D-D/H/E/K) thought to be required for slicer activity. PPD proteins with names shaded conserve the motif. Hs, Homo sapiens; Dm, Drosophila melanogaster; Tt, Tetrahymena thermophila.

Supplemental Figure S4



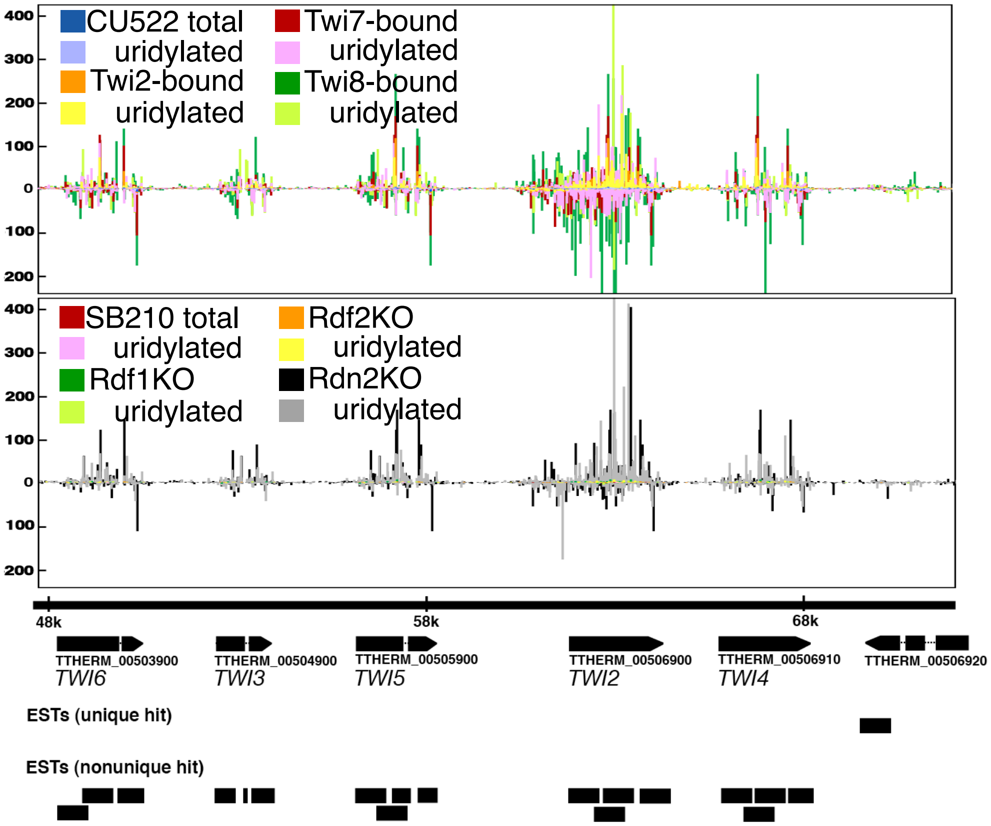
Supplemental Figure S4
Nucleotide frequency plots for 3' uridylated sequences in each library. Note that sample size is small for Twi8, Twi9, and Twi12 libraries.

Supplemental Figure S5



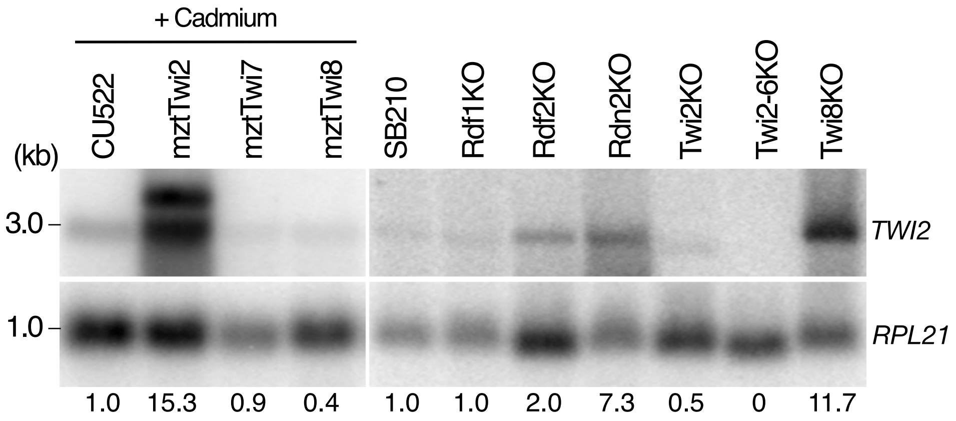
Supplemental Figure S5
Northern blot showing expression of a transcript from phased cluster locus Ph2 (PH2) and RPL21 as a loading control. Strains indicated as mztTwi2, mztTwi7, and mztTwi8 harbor a transgene for cadmium-inducible expression of the Twi protein; cadmium was added to the cultures used for total RNA isolation and to the culture from the matched background strain CU522.

Supplemental Figure S6



Supplemental Figure S6
Small RNA density plots at TWI2-6. Mapping of unique and non-unique sRNA sequences is illustrated without normalization for the number of genome mapping locations (each sRNA is mapped to all possible sites of origin in the genome). For comparison, mapping of only unique-sequence sRNAs is shown in Figure 6D.

Supplemental Figure S7



Supplemental Figure S7
Northern blots showing TWI2 expression in different strain backgrounds. Strains indicated as mztTwi2, mztTwi7, and mztTwi8 harbor a transgene for cadmium-inducible expression of the Twi protein; cadmium was added to the cultures used for total RNA isolation and to the culture from the matched background strain CU522. Quantification of TWI2 mRNA level normalized to RPL21 mRNA level is indicated below the blot with the matched wild-type strain CU522 or SB210 set at 1.0.

# Supplemental Table S1, part 1

| Name | Category | Database location | Scaffold location | Gene prediction |
|---|---|---|---|---|
| IA | Pseudogene/A-rich tract | ChrUn2:20000-30000 | 8254028:20000..30000 | TTHERM_00670140-00670150 |
| IB | Pseudogene/A-rich tract | ChrUn2:6870000-6885000 | 8254600:157364..172364 | TTHERM_00277590-00277610 |
| IC | Pseudogene/A-rich tract | ChrUn2:16045000-16055000 | 8254638:959810..969810 | TTHERM_00192110-00192130 |
| ID | Pseudogene/A-rich tract | ChrUn1:16685000-16689000 | 8254659:1289306..1293306 | TTHERM_00052070 |
| IE | Pseudogene/A-rich tract | ChrUn1:19441000-19445000 | 8254661:30912..34912 | TTHERM_01335320 |
| IIA | Pseudogene/A-rich tract | ChrUn1:36704000-36710000 | 8254557:42963..48963 | TTHERM_01295370 |
| IIB | Pseudogene/A-rich tract | ChrUn1:27546000-27552000 | 8254822:132438..138438 | TTHERM_00874720 |
| IIIA | Pseudogene/A-rich tract | ChrUn2:22586000-22589000 | 8254823:260949..263949 | TTHERM_00548090-00548100 |
| IIIB | Pseudogene/A-rich tract | ChrUn1:26978000-26984000 | 8254678:15820..21820 | TTHERM_01256540-01256560 |
| IVA | Pseudogene/A-rich tract | ChrUn1:40104059-40114000 | 8254572:0..9941 | TTHERM_00887990-00888010 |
| IVB | Pseudogene/A-rich tract | ChrUn1:40415000-40430000 | 8254718:112775..127775 | TTHERM_00809110-00809140 |
| IVC | Pseudogene/A-rich tract | ChrUn1:9602000-9612000 | 8254631:343..10343 | TTHERM_01076810-01076820 |
| IVD | Pseudogene/A-rich tract | ChrUn1:1206000-1209000 | 8254233:1050..4050 | TTHERM_01679240 |
| VA | Pseudogene/A-rich tract | ChrUn1:30108000-30114000 | 8254697:271823..277823 | TTHERM_00083890-00083900 |
| Ph1 | Phased/Hairpin | ChrUn1:7660591-7663091 | 8254617:879000..881500 | none |
| Ph2 | Phased/Hairpin | ChrUn2:27630400-27631100 | 8254370:140130..140830 | none |
| Ph3 | Phased/Hairpin | ChrUn2:40113400-40114500 | 8254594:261431..262531 | TTHERM_00286880 (antisense, half in intron) |
| Ph4 | Phased/Hairpin | ChrUn2:40770000-40773000 | 8254597:2898..5898 | none |
| Ph5 | Phased/Hairpin | ChrUn2:15552000-15553000 | 8254638:466810..467810 | none |
| Ph6 | Phased/Hairpin | ChrUn1:48074000-48076000 | 8254745:491842..493842 | TTHERM_00433990 (antisense, only a portion overlaps) |
| Ph7 | Phased/Hairpin | ChrUn1:44967400-44967900 | 8254590:193617..194117 | none |
| Twi8.1 | EST-supported/Endo-siRNA-like | ChrUn1:36312500-36317500 | 8254552:86737..91737 | EV846396.1 (between TTHERM_00794190 & -200) |
| Twi8.2 | EST-supported/Endo-siRNA-like | ChrUn2:8520000-8535000 | 8254607:352351..367351 | TTHERM_00439140 - 00439190 |
| Twi8.3 | EST-supported/Endo-siRNA-like | ChrUn2:18910369-18931869 | 8254654:47500..69000 | TTHERM_00503900 - 00506910 |
| Twi8.4 | EST-supported/Endo-siRNA-like | ChrUn2:2858000-2868000 | 8254428:44115..54115 | TTHERM_00535270 - 00535340 |
| Twi8.5 | EST-supported/Endo-siRNA-like | ChrUn2:4956700-5300000 | 8254446:0..343300 | part of TTHERM_00492410; TTHERM_00492450, -60; TTHERM_00492680 - 00492790 (part); TTHERM_00492820 - 00492850; TTHERM_00494060 -00494100 |
| Twi8.6 | EST-supported/Endo-siRNA-like | ChrUn2:14195000-14198000 | 8254486:382438..385438 | TTHERM_00558620 |
| Twi8.7 | EST-supported/Endo-siRNA-like | ChrUn1:2680000-2711000 | 8254415:14314..45314 | TTHERM_01205310 - 01205340 |
| Twi8.8 | EST-supported/Endo-siRNA-like | ChrUn1:3660000-3670000 | 8254431:205268..215268 | TTHERM_00312520 - 00312550 |
| Twi8.9 | EST-supported/Endo-siRNA-like | ChrUn2:31200000-31250000 | 8254382:86765..136765 | TTHERM_00947530 - 00947580 |
| Twi8.10 | EST-supported/Endo-siRNA-like | ChrUn2:23314000-23320200 | 8254688:404795..410995 | TTHERM_00295330 - 00295350 |
| Twi8.11 | EST-supported/Endo-siRNA-like | ChrUn2:16140000-16150000 | 8254495:75321..85321 | TTHERM_00564050 - 00564080 |
| Twi8.12 | EST-supported/Endo-siRNA-like | ChrUn2:3350000-3400000 | 8254429:65953..115953 | TTHERM_01043230 - 01043320 |
| Twi7.1 | Low-copy repeat | ChrUn2:21241250-21241400 | 8254811:305564..305714 | TTHERM_00058540 |
| Twi7.2 | Low-copy repeat | ChrUn1:646000-649224 | 8254049:83444..86668 | TTHERM_01145020 |
| _Twi7.2a | Low-copy repeat | ChrUn1:4740000-4740850 | 8254436:177071..177921 | none |
| _Twi7.2b | Low-copy repeat | ChrUn1:31187253-31188000 | 8254848:0..747 | none |
| Rpt1 | High-copy repeat | All over, e.g. ChrUn1:26982000-26992000 | 8254678:19820..29820 | none |

## Supplemental Table S1, part 2

| Name | Gene description | Other comments |
|---|---|---|
| IA | hypothetical proteins | Lee et. al. 2006 8254028/CH445461 cluster |
| IB | hypothetical proteins | Lee et. al. 2006 8254600/CH445618 cluster |
| IC | Proline-rich P65-related, hypothetical proteins | Lee et. al. 2006 8254638/CH445644 cluster |
| ID | hypothetical protein | Lee et. al. 2006 8254659/CH445663 cluster |
| IE | hypothetical protein | Lee et. al. 2006 Non-clustered/8254661/CH445665 |
| IIA | hypothetical protein | Lee et. al. 2006 82545578/CH445585 cluster |
| IIB | hypothetical protein | Lee et. al. 2006 8254822/CH445791 cluster |
| IIIA | hypothetical proteins | Lee et. al. 2006 8254823/CH445792 cluster |
| IIIB | hypothetical proteins | Lee et. al. 2006 8254678/CH445681 cluster |
| IVA | hypothetical proteins | Lee et. al. 2006 8254572/CH445593 cluster |
| IVB | hypothetical proteins | |
| IVC | hypothetical proteins | |
| IVD | hypothetical protein | |
| VA | hypothetical proteins | Lee et. al. 2006 8254697/CH445695 cluster |
| Ph1 | | Lee et. al. 2006 8254617/CH445632 cluster |
| Ph2 | | |
| Ph3 | hypothetical protein | |
| Ph4 | | Lee et. al. 2006 8254597/CH445615 cluster |
| Ph5 | | |
| Ph6 | hypothetical protein | |
| Ph7 | | |
| Twi8.1 | none- EST from TSA library | structured |
| Twi8.2 | 6 predicted: 2 with ESTs from TSA library | converging ORFs and non-unique ESTs |
| Twi8.3 | 5 predicted: Twi2-6 | non-unique ESTs (and downstream ORF is in opp direction |
| Twi8.4 | ~6 predicted genes: 2 make sense sRNAs in wt libraries | converging ORFs |
| Twi8.5 | Ku70/80 beta barrel domain; Peptidyl-prolyl cis-trans isomerase; ribulose-phosphate 3-epimerase; hypoxanthine phosphoribosyltransferase (2); DYH14: inner arm dynein heavy chain; EF hand family; ilsA: isoleucyl-tRNA synthestase; zinc finger, zz type; queuine tRNA-ribosyltransferase | Very large cluster (really many clusters within ~250kb region - not all ORFs across the region make sRNAs over background) |
| Twi8.6 | Atu1 | non-unique ESTs |
| Twi8.7 | Protein Kinase domain; Hypothetical proteins | converging ORFs |
| Twi8.8 | All hypothetical proteins (~4) | converging ORFs |
| Twi8.9 | oxidoreductase, short chain dehydrogenase/reductase family; oxidoreductase, aldo/keto reductase family; NRK7: NIMA-related kinase; hypothetical proteins | converging ORFs |
| Twi8.10 | polyadenylate-binding like; polyadenylate-binding protein 2; hypothetical protein | converging ORFs and non-unique ESTs |
| Twi8.11 | Erv1/Alr family; Dpy-30 motif | converging ORFs |
| Twi8.12 | hypothetical proteins | converging ORFs and non-unique ESTs |
| Twi7.1 | | 2 repeats |
| Twi7.2 | | 4 repeats |
| _Twi7.2a | | 1 repeat (same as Twi7.2) |
| _Twi7.2b | | 2 repeats (same as Twi7.2) |
| Rpt1 | many ESTs | |

Supplemental Table S1
List of sRNA clusters analyzed. Information is provided including name, scaffold and coordinates, annotation, and additional description. Shading indicates the clusters generating abundant sRNAs included in Figure 7A.

# CHAPTER TWO

## A growth-essential *Tetrahymena* Piwi protein carries tRNA fragment cargo

Based on Couvillion et al., Genes & Development 2010

### Abstract

Argonaute/Piwi proteins associate with small RNAs that typically provide sequence specificity for RNP function in gene and genome regulation. Here we show that Twi12, a *Tetrahymena* Piwi protein essential for growth, is loaded with mature tRNA fragments. The tightly bound ~18-22 nt tRNA 3' fragments are biochemically distinct from the tRNA halves produced transiently in response to stress. Notably, the end positions of Twi12-bound tRNA 3' fragments precisely match RNAs detected in total small RNA of mouse embryonic stem cells and human cancer cells. Our studies demonstrate unanticipated evolutionary conservation of mature tRNA processing to tRNA-fragment small RNAs.

## Introduction

New non-coding RNA populations continue to be discovered (Mercer et al., 2009; Siomi and Siomi, 2009), implying that there are biological events of RNA-mediated regulation and RNA processing that remain to be appreciated. Beyond the microRNAs, small interfering RNAs, and germline Piwi-interacting RNAs that guide the specificity of Argonaute/Piwi (Ago/Piwi) RNP function (Tolia and Joshua-Tor, 2007), the search for novel small RNAs (sRNAs) has uncovered fragments of mature tRNAs that are poorly, non-preferentially, or not specifically associated with Ago proteins (Cole et al., 2009; Haussecker et al., 2010; Thompson and Parker, 2009b). Abundant tRNA fragments resulting from conditionally induced cleavage of the anticodon loop were first reported in the ciliated protozoan *Tetrahymena*, and similar tRNA cleavage phenomena have been revealed as a broadly conserved prokaryotic and eukaryotic response to stress or change in developmental state (Garcia-Silva et al., 2010; Lee and Collins, 2005; Thompson and Parker, 2009b). Stress-induced tRNA cleavage involves a nuclease from the RNase T2 or RNase A family in budding yeast or human cells respectively, not an RNase III family Dicer enzyme, and therefore generates a 5' hydroxyl rather than 5' monophosphate product (Fu et al., 2009; Thompson and Parker, 2009a; Yamasaki et al., 2009). Libraries of RNA sequences obtained following Ago/Piwi protein enrichment or size-selection of total RNA typically contain a minor fraction of tRNA fragments, which are expected contamination based on high tRNA abundance and the multiple pathways of tRNA degradation (Phizicky and Alfonzo). Some examples of precursor or mature tRNA processing by Dicer have been reported, one of which occurs by Dicer recognition of an alternative short-hairpin-like fold of the primary transcript (Babiarz et al., 2008; Cole et al., 2009). Recent studies also describe Dicer-independent accumulation of the tRNA 3' trailer (a primary transcript segment between the tRNA 3' end and the transcription termination signal), creating sRNAs proposed to regulate cellular proliferation and/or the homeostasis of RNA silencing (Haussecker et al., 2010; Lee et al., 2009b).

We previously purified tagged versions of eight distinct *Tetrahymena* Piwi-family (Twi) proteins for analysis of sRNAs generated by the *Tetrahymena* Dicer enzymes Dcl1 and Dcr2. Dcl1 produces ~28-29 nt sRNAs that mediate heterochromatin formation and DNA elimination in the sexual cycle of reproduction (Malone et al., 2005; Mochizuki and Gorovsky, 2005), while Dcr2 produces ~23-24 nt sRNAs involved in gene regulation during asexual growth (Couvillion et al., 2009; Howard-Till and Yao, 2006; Lee and Collins, 2006; Lee and Collins, 2007). Of the eight distinct Twi proteins, only Twi12 failed to enrich a profile of sRNAs consistent with the size range for products of Dcl1 or Dcr2 (Couvillion et al., 2009). Surprisingly, among the *TWI* genes expressed in growing cells, only *TWI12* is individually essential (Couvillion et al., 2009). Because the heterogeneously sized sRNAs bound to Twi12 were previously isolated using Twi12 overexpressed in the presence of competing endogenous protein, they were of uncertain physiological specificity. Here we investigate the specificity of Twi12 sRNA loading.

## Results and Discussion

We first established the biological function of N-terminally tagged Twi12 by disruption of endogenous *TWI12* in the presence of a tagged-protein transgene integrated at the non-essential *BTU1* locus (Fig. 1A). The *MTT1* promoter used to drive transgene expression has low basal expression but is rapidly induced to high expression by cadmium addition to media (Shang et al., 2002). The transgene open reading frame encodes Twi12 fused to tandem Protein A domains with an intervening protease cleavage site (ZZtev; Fig. 1A, left). Wild-type cells lacking the integrated transgene were not able to replace the endogenous *TWI12* locus with a neo2 selection cassette, while cells with the transgene could fully replace *TWI12* even when selected in absence of cadmium (Fig. 1A, right). The uninduced level of transgene mRNA expression was only slightly higher than the level of endogenous *TWI12* mRNA, while cadmium induction gave more than 10-fold overexpression (Supplemental Fig. S1). Cells with an uninduced level of transgene expression that also lacked endogenous Twi12 were used to purify Twi12 and associated sRNAs from the rapidly dividing cells of a vegetative growth culture (veg) or from cultures of non-dividing cells harvested after 12 hours of nutrient starvation (st12). Copurified RNA was resolved by denaturing polyacrylamide gel electrophoresis and directly stained with SYBR Gold (Fig. 1B, lanes 1-2). Comparison to total RNA size-enriched for smaller RNAs (<100 nt) revealed that the Twi12-bound sRNAs were distinct from any abundant sRNA population including the constitutively accumulated ~23-24 nt sRNA products of Dcr2 and the starvation-induced ~30-35 nt halves of tRNA cleaved in the anticodon loop (Fig. 1B, lanes 3-4). Parallel mock purifications using wild-type cell extract lacking tagged Twi12 failed to enrich sRNAs (Supplemental Fig. S2).

Twi12-bound sRNAs from vegetative growth were used for library construction and deep sequencing. Surprisingly, the majority of sequences from both Twi12-associated sRNA size classes mapped to tRNA genes (Fig. 1C, with dark or light shading for perfect sequence matches or allowing a single internal mismatch, respectively). Given that conserved post-transcriptional tRNA modifications such as $N^1$-methyl-adenosine at position 58 greatly reduce the efficiency of the reverse transcription step required for library construction (Saikia et al., 2010), the observed predominance of tRNA sequence reads is likely to be a dramatic underestimate. Indeed, we were unable to detect the presence of any non-tRNA sequence read in the Twi12-enriched sRNA pool by direct hybridization (see below; additional data not shown). Read numbers for each tRNA sequence were plotted against gene copy number as an approximation of tRNA abundance, with the caveat that modification differences between tRNAs would bias relative representation of tRNA reads in the library. This analysis suggests a general trend for more abundant tRNAs to generate more tRNA fragment reads (Fig. 1D; Couvillion et al., 2010 Supplemental Table 1 provides an updated and curated inventory of *Tetrahymena thermophila* tRNA gene sequences and copy numbers). We conclude that Twi12 carries sRNAs broadly representative of total cellular tRNA.

To verify the abundance of tRNA fragments as Twi12-associated sRNAs, we used oligonucleotides complementary to each end of several tRNAs to probe blots of Twi12-associated sRNAs. Consistently, Twi12-bound tRNA 5'-end fragments were predominantly ~25-30 nt in length while tRNA 3'-end fragments were predominantly less than 23 nt in length, matching the two size classes of Twi12-bound sRNAs detected by SYBR Gold (Fig. 2A; additional data not shown). These fragment sizes and sRNA sequence data establish that typical Twi12-bound tRNA 3' fragments begin in the T-loop and end at variable positions of the mature tRNA 3’ CCA tail (see below), while typical Twi12-bound tRNA 5’ fragments begin at the mature tRNA 5’ end and end variably in the anticodon stem (illustrated in Fig. 2B).

To investigate differential association of Twi12 with the 5’ versus 3’ tRNA fragments, we used more stringent RNP purification conditions. Under some conditions, Twi12 selectively enriched only the ~18-22 nt sRNAs (Fig. 2C). Furthermore, when Twi12 was purified in the presence of high detergent concentration, with or without prior in vivo crosslinking with formaldehyde, only the ~18-22 nt sRNA population was enriched (Fig. 2D, lanes 2 and 5). The sRNAs of crosslinked but not native complexes were resistant to dissociation by urea (Fig. 2D, compare lanes 2-4 and 5-7), confirming the direct crosslinking and thus close-range physical association of Twi12 and the ~18-22 nt sRNAs. Hybridization analysis confirmed that the sRNAs preferentially enriched with Twi12 after in vivo crosslinking are the tRNA 3’ fragments (Fig. 2E; additional data not shown). These findings suggest that analogous to other Ago/Piwi proteins, Twi12 is initially loaded with an asymmetric RNA duplex such that one guide strand is tightly bound (the strand containing the 3’ side of the acceptor stem) while the other passenger strand is readily released (the strand containing the 5’ side of the acceptor stem).

The 3’ CCA is added post-transcriptionally to eukaryotic tRNAs. To better define the termini of the tRNA 3' fragments tightly bound to Twi12, we deep-sequenced a library from sRNAs enriched by crosslinking to Twi12 in vivo. This population of tRNA 3' fragments was analyzed for sequence reads that mapped only to the genome-encoded portion of the tRNA or were extended at their 3’ end by addition of C, CC, or CCA (Fig. 2F). About one third of the reads include at least one post-transcriptionally added nucleotide, indicating that the Twi12-associated tRNA fragments derive at least in part from matured tRNAs. A pronounced 5’ sequence signature was evident for the crosslinked Twi12 sRNAs (Fig. 2G). The 5’ ends of the tRNA fragments correspond to cleavage between the thymidine and pseudouridine (ψ) of the TψC motif, giving rise to a strong 5’ UC bias in the sequenced fragments. The overwhelming representation of cytidine at the second position suggests that tightly bound Twi12 sRNAs vary more in their extent of retained 3' CCA than in their 5' end position within the T-loop.

The Twi12-bound tRNA fragments differ from the starvation or other stress-induced tRNA halves (Lee and Collins, 2005) in their constitutive accumulation, shorter length, and lower abundance. Notably, the starvation-induced tRNA halves were not detectably associated with Twi12 (Fig. 1B), even if Twi12 was dramatically overexpressed by cadmium induction (data not shown). To investigate potential biochemical differences between the Twi12-bound and starvation-induced tRNA 3' fragments, we compared the phosphorylation and modification status of the fragment ends. Size-enriched total RNA from starved cells containing predominantly ~30-35 nt tRNA halves and ~23-24 nt products of Dcr2 (Fig. 3A, lane 1) or Twi12-bound sRNAs enriched by affinity purification after in vivo crosslinking (lane 4) were subject to nuclease or chemical treatment. The entire pool of RNAs was visualized by direct staining with SYBR Gold (Fig. 3A) and specific tRNA 3' fragments were detected by blot hybridization (Fig. 3B; additional data not shown). The 5’ monophosphate-dependent nuclease Terminator degraded the ~23-24 nt sRNA products of Dcr2 but not the starvation-induced tRNA 3' halves (Figs. 3A and 3B, lanes 3), consistent with the known specificities of Dicer and the nucleases responsible for stress-induced tRNA cleavage in yeast and mammals (Fu et al., 2009; Thompson and Parker, 2009a; Yamasaki et al., 2009). Unlike the starvation-induced tRNA 3' halves, the Twi12-bound tRNA 3' fragments were degraded by Terminator treatment (Figs. 3A and 3B, lanes 6). This result establishes that the tightly bound Twi12 sRNAs possess a 5’ monophosphate. A substantial fraction of the ~23-24 nt Dcr2 products, ~30-35 nt starvation-induced tRNA 3' halves, and ~18-

22 nt Twi12-bound tRNA 3' fragments were reactive to ß-elimination, indicating the presence of both 2' and 3' hydroxyl groups at the RNA 3' end (Figs. 3A and 3B, lanes 2 and 5).

These assays demonstrate that Twi12-bound tRNA 3' fragments have biochemical features distinct from the starvation-induced tRNA 3' halves (summarized in Fig. 3C), including the presence of a 5' monophosphate characteristic of most Ago/Piwi-bound sRNAs. Consistent with known principles of Ago/Piwi sRNA strand incorporation and displacement (Kawamata et al., 2009; Siomi and Siomi, 2009; Tolia and Joshua-Tor, 2007), we envision that Twi12 binds initially to full-length or endonucleolytically cleaved tRNAs retaining most of the tRNA secondary structure (Fig. 3D). The tRNA fold stacks the acceptor and T-loop stems, presenting a relatively canonical length of sRNA duplex for Ago/Piwi interaction. The cleavage reaction(s) that generate a 5' monophosphate terminus at the ψ of the TψC motif may provide an optimal 5' end for recognition by Twi12. The 3' end of the tightly bound sRNA is typically not a complete CCA, which could reflect endogenous turnover of this sequence prior to Twi12 association or the trimming of a bound tRNA fragment to optimal length for end-protection. The tRNA 5' fragment initially base-paired to the 3' fragment would have less protected ends, allowing nuclease nibbling and ultimately loss of the 5' fragment via nuclease degradation, helicase unwinding, or passive dissociation.

Selective retention of tRNA 3' fragments as the guide strands of Twi12 RNPs predicts that steady-state accumulation of this specific tRNA fragment population should depend on the cellular level of Twi12. Following unsuccessful attempts to engineer inducible *TWI12* genetic depletion, we addressed whether Twi12-associated sRNAs were altered in steady-state accumulation by transient protein overexpression. Cadmium addition results in a rapid increase in expression from the *MTT1* promoter, followed by a decline as cadmium is chelated by induced metallothionein proteins. Transgene-encoded ZZTwi12 was robustly induced from its basal expression level by cadmium addition during vegetative growth (Fig. 4A, lanes 1-4) and then declined in level as cells continued to divide until reaching the maximal density of stationary phase (lanes 5-8). Notably, Twi12 overexpression increased the accumulation of RNAs smaller than the ~23-24 nt sRNA products of Dcr2 (Fig. 4B), which were detected as tRNA 3' fragments by blot hybridization (Fig. 4C). The increased steady-state accumulation of tRNA 3' fragments was not accompanied by an increase in accumulation of the 5' fragments or the stress-induced tRNA halves (Supplemental Fig. S3), consistent with selective stabilization of the 3' fragments in Twi12 RNPs.

What is the biological function of Twi12 RNPs? Twi12-faciliated tRNA turnover may be important in tRNA quality control or in balancing tRNA levels to improve translational fidelity (Dittmar et al., 2006; Drummond and Wilke, 2009; Kramer and Farabaugh, 2007; Phizicky and Alfonzo). Also, although Twi12 does not conserve the active site residues of some Ago/Piwi proteins, the tRNA fragments could provide sequence specificity for target RNA or DNA regulation. Non-abundant sRNA populations notably similar to the Twi12-bound sRNAs have been detected in total RNA of rapidly dividing mouse and human cell lines. In mouse embryonic stem cells, beneath the abundant reads from Dicer processing of a misfolded tRNA precursor, some Dicer-independent tRNA 5' and 3' end fragments are similar in length to the Twi12-bound sRNAs (Babiarz et al., 2008). Also, among the tRNA fragment populations sequenced from total RNA of human liver carcinoma or prostate cancer cell lines, there are tRNA 3' fragments with the same length, 5' end position, 3' CCA residues, and 5' monophosphate terminus (when possible to infer from the cloning method) as the Twi12-bound sRNAs (Kawaji et al., 2008; Lee et al., 2009b). Thus, the unique loading specificity and function of Twi12 may represent a ciliate evolutionary adaptation for handling the products of a conserved tRNA endonucleolytic processing reaction. Whether animal Piwi proteins are also loaded with ~20 nt tRNA 3' fragments remains to be investigated in suitable physiological context.

## Materials and Methods

*Strains and extracts*

Strain construction and culture growth were performed as previously described (Couvillion et al., 2009) in the CU522 strain background. Transgene-encoded Twi12 is expressed from the endogenous *TWI12* open reading frame. Purifications under condition 1 (C1) were performed as previously described (Lee and Collins, 2007). For condition 2 (C2), extract preparation was modified by cell lysis in a volume ~10-fold that of the packet cell pellet, addition of 0.1% Triton X-100, doubled monovalent ion concentration (100 mM NaCl), omission of ß-mercaptoethanol, and clearing by centrifugation at 16,000 x *g* for 15 min. For crosslinking, cells were washed in 1X PBS, resuspended to 5 x $10^5$ cells/mL, and swirled with 0.75% formaldehyde for 10 min. The reaction was quenched by addition of 125 mM Tris-HCl (pH 7.5) for 5 min. Cells were then washed twice in 1X PBS, resuspended to 5 x $10^6$ cells/mL in RIPA (50 mM Tris-HCl at pH 7.5, 1% Igepal, 0.5% sodium deoxycholate, 0.05% SDS, 1 mM EDTA, 150 mM NaCl) plus protease inhibitors and sonicated in 15 sec pulses at 20% amplitude until fully lysed (2 to 2.5 min). Lysate from crosslinked cells was cleared at 16,000 x *g*. After Twi12 binding to IgG agarose, resin was washed 3 times for 10 min each at room temperature in RIPA-HS (RIPA with 500 mM NaCl, 1% sodium deoxycholate, and 0.1% SDS) with the indicated final concentration of urea.

*Nucleic acid analyses*

DNA blots were hybridized with hexamer-primed probes. RNA blots were hybridized with 5' end-labeled 20-22 nt oligonucleotides perfectly complementary to the indicated tRNA with an endpoint at the mature tRNA 5' or 3' end (excluding the 3' CCA). Total RNA was isolated and size-enriched as previously described (Lee and Collins, 2006). Terminator exonuclease (Epicentre) was used based on the manufacturer's protocol, and ß-elimination was performed as described (Couvillion et al., 2009). Blotting to detect the ZZ tag was performed using whole-cell lysates (Couvillion et al., 2009). Library preparation, deep sequencing, and annotation were performed largely as described (Couvillion et al., 2009) with additional consideration for tRNA-specific features (see below). Reads from the sRNA libraries are deposited at Gene Expression Omnibus (accession no. GSEX, data sets GSMX–GSMXX).

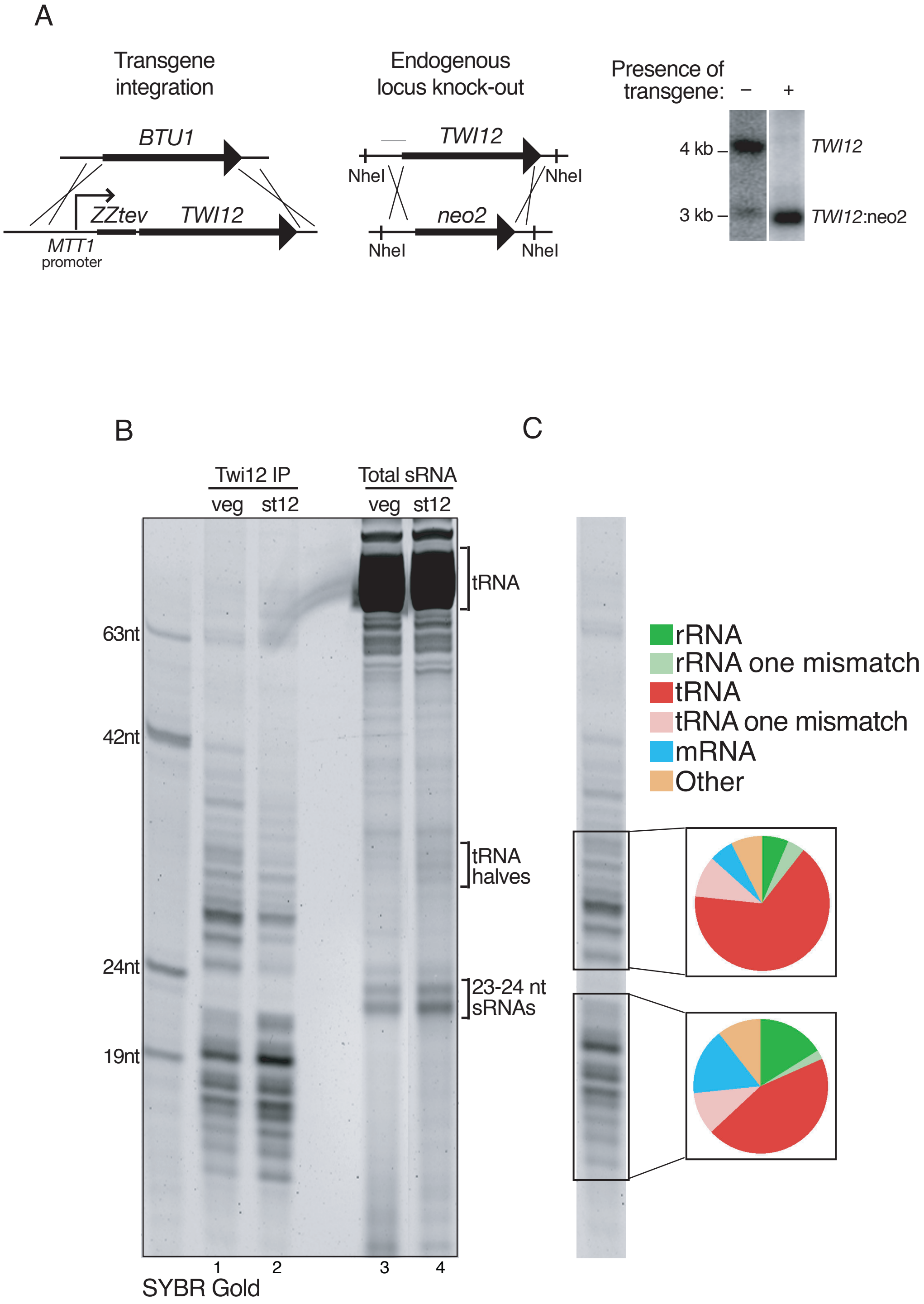
*Mapping and annotation of sequencing reads*

After discovering the prevalence of tRNA fragment reads, a sequence file of mature tRNAs was used for mapping to prevent counting of 3' untemplated C, CC, or CCA as a mismatch. For sRNA annotation in Fig. 1C, reads were mapped first to the chromosome encoding the large ribosomal RNA precursor and to tRNA sequences, then unmatched reads were allowed to map to the rest of the genome. Reads that mapped to tRNA often mapped multiple times but each read was counted only once. The total number of annotated sequences was 2,031,775 for the 15–22 nt size range and 3,362,521 for the 25–34 nt size range. For tRNA isoform assignment in Fig. 1D, reads were first mapped allowing no internal mismatch but allowing for 3' untemplated C, CC, or CCA. If a read matched more than one isoform, it was counted toward totals for each. After removing reads that mapped with no internal mismatch, the procedure was repeated allowing for one internal mismatch.

Genes encoding tRNAs were previously identified using tRNAscan-SE 1.21 (Eisen et al., 2006; Lee and Collins, 2005). The tabulation of tRNA genes was updated for use in this study

(Couvillion et al., 2010 Supplemental Table S1). Most sequences previously annotated as ambiguous were assigned, a full list of mitochondrially encoded tRNAs was included, and non-standard tRNAs were validated (Sec, iMet, and mtMet). Sequences with a Cove score of less than 40 or with an undetermined anticodon using the eukaryotic tRNA model were assigned as pseudogenes.

Figure 1, part 1



# Figure 1, part 2

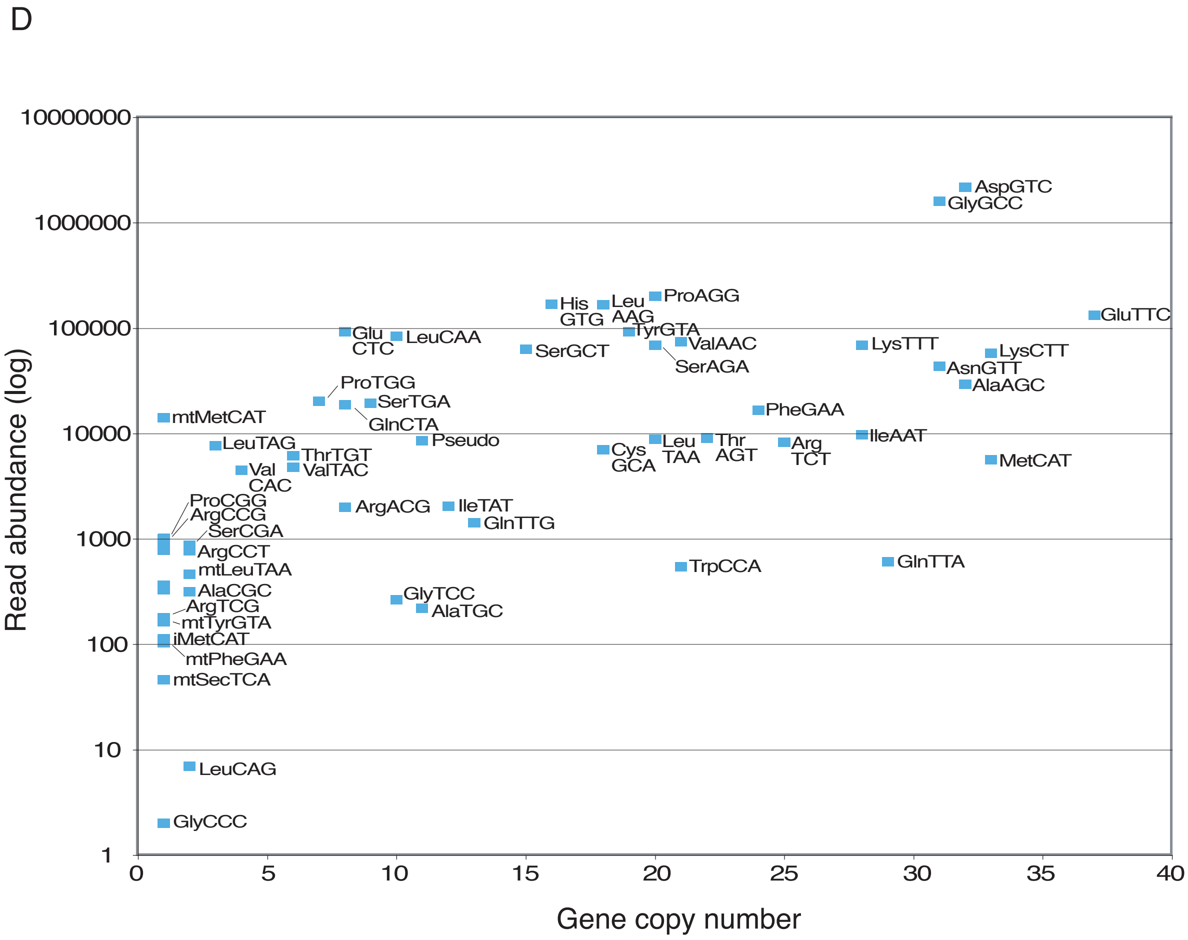


Figure 1. Twi12 association with tRNA fragments.
(A) The schematics show the strategy of ZZTwi12 transgene knock-in and endogenous TWI12 locus knock-out; the latter is evidenced with Southern blot data using NheI-digested genomic DNA at right. The position of the Southern blot probe is indicated in the TWI12 locus schematic with a thin gray line.
(B) RNAs copurified with Twi12 in vegetative growth (veg) or after 12 hours of starvation (st12) were stained directly with SYBR Gold in comparison to size-selected total RNA from the same cultures.
(C) The composition of sRNA library sequencing reads is depicted for reads with either a perfect match to the genome or a single internal mismatch as indicated.
(D) Read numbers for tRNA fragments copurified with Twi12 were plotted relative to tRNA gene copy number as an approximation of full-length tRNA abundance. The labels of some data points at left were removed for clarity: mtGluTCC, SecTCA, ThrCGT, mtHisGTG. Pseudo indicates combined reads from the 11 annotated tRNA pseudogenes; mt indicates mitochondrial.

Figure 2, part 1

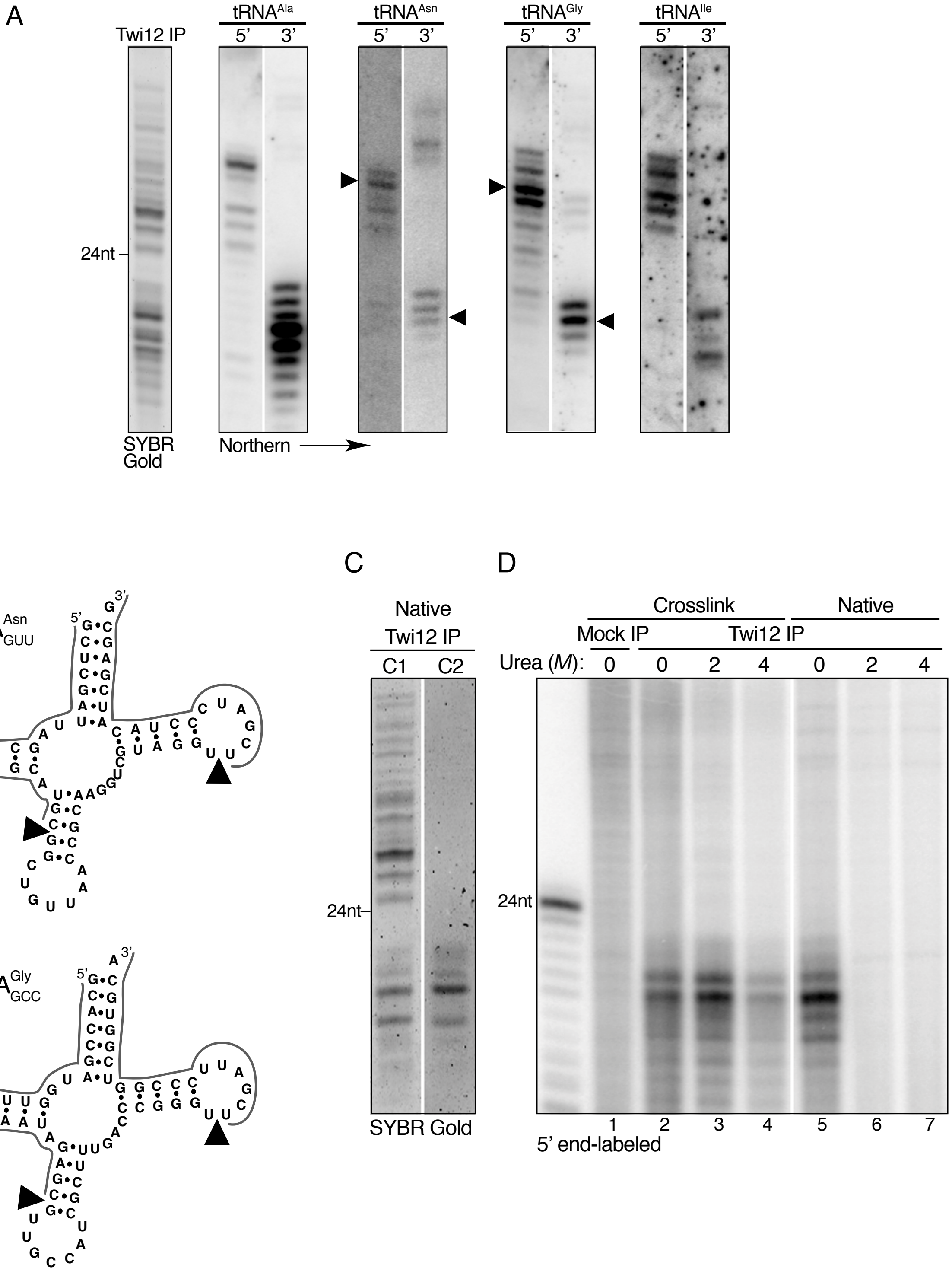


Figure 2, part 2

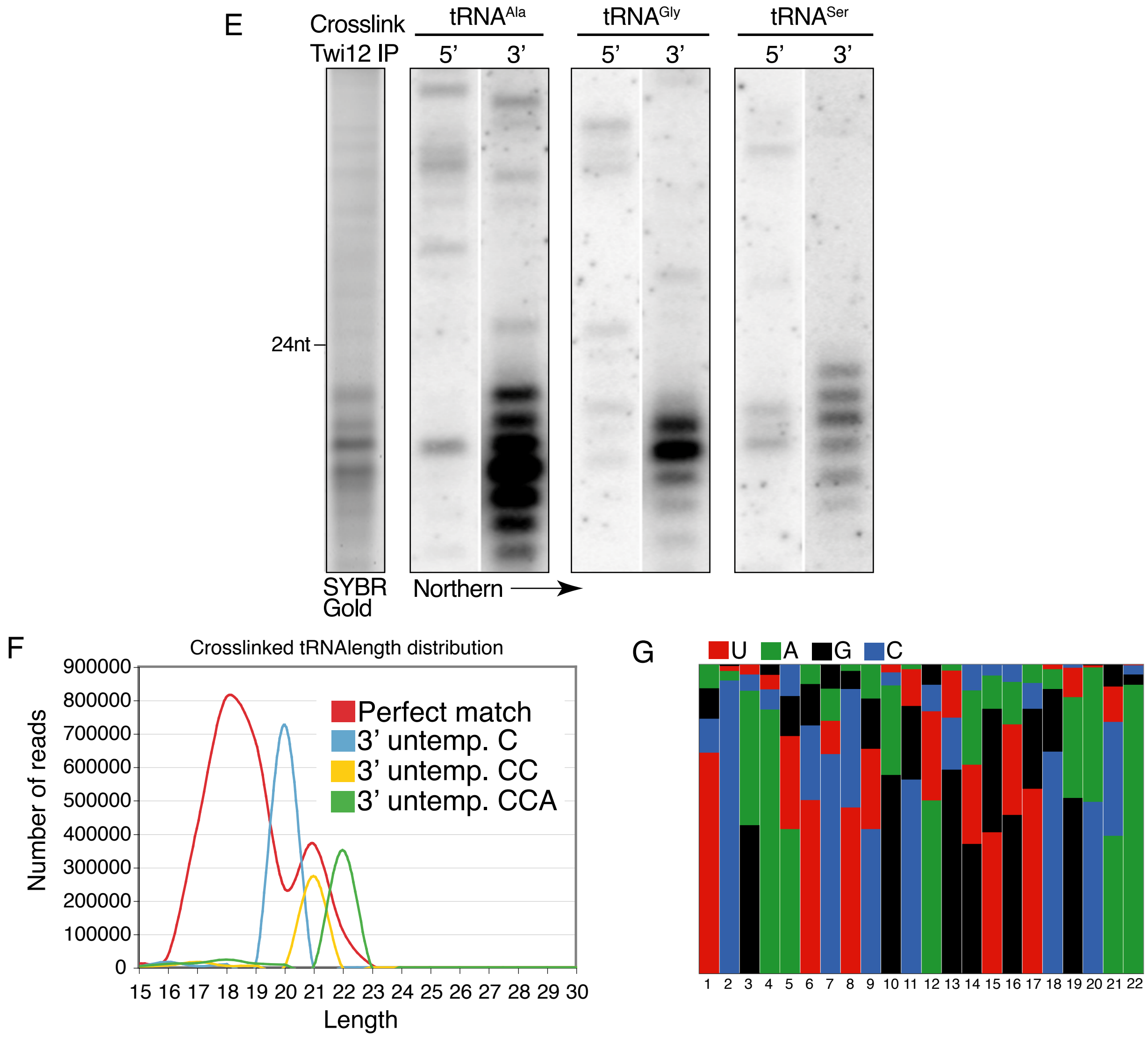


Figure 2. Retention of the tRNA 3' fragment as the tightly bound strand.
(A) Blot hybridization of Twi12-bound sRNAs was performed using probes complementary to the 5' or 3' end of several abundant tRNAs. Direct SYBR Gold staining of Twi12-bound sRNAs is shown at left for reference.
(B) Arrowheads and lines indicate respectively the potential processing positions and detected sRNAs for two of the tRNAs from (A). The 3' CCA nucleotides of a mature tRNA are not illustrated.
(C) Twi12-associated sRNAs were directly stained after RNP isolation using two different gentle, non-denaturing buffer conditions.
(D) Twi12-associated sRNAs were 5' end-labeled after isolation using stringent purification conditions, with or without prior crosslinking; different concentrations of urea were used for washes prior to RNA extraction.
(E,F,G) Twi12-crosslinked sRNAs washed with 2 M urea were used for blot hybridization to detect specific tRNA fragments (E) and for library construction. The length distribution of sequencing reads (F) is shown for 3' fragments that match the genome perfectly or have an untemplated 3' C, CC, or CCA, with the combined nucleotide frequency at each position (G) revealing a 5' sequence signature.

Figure 3

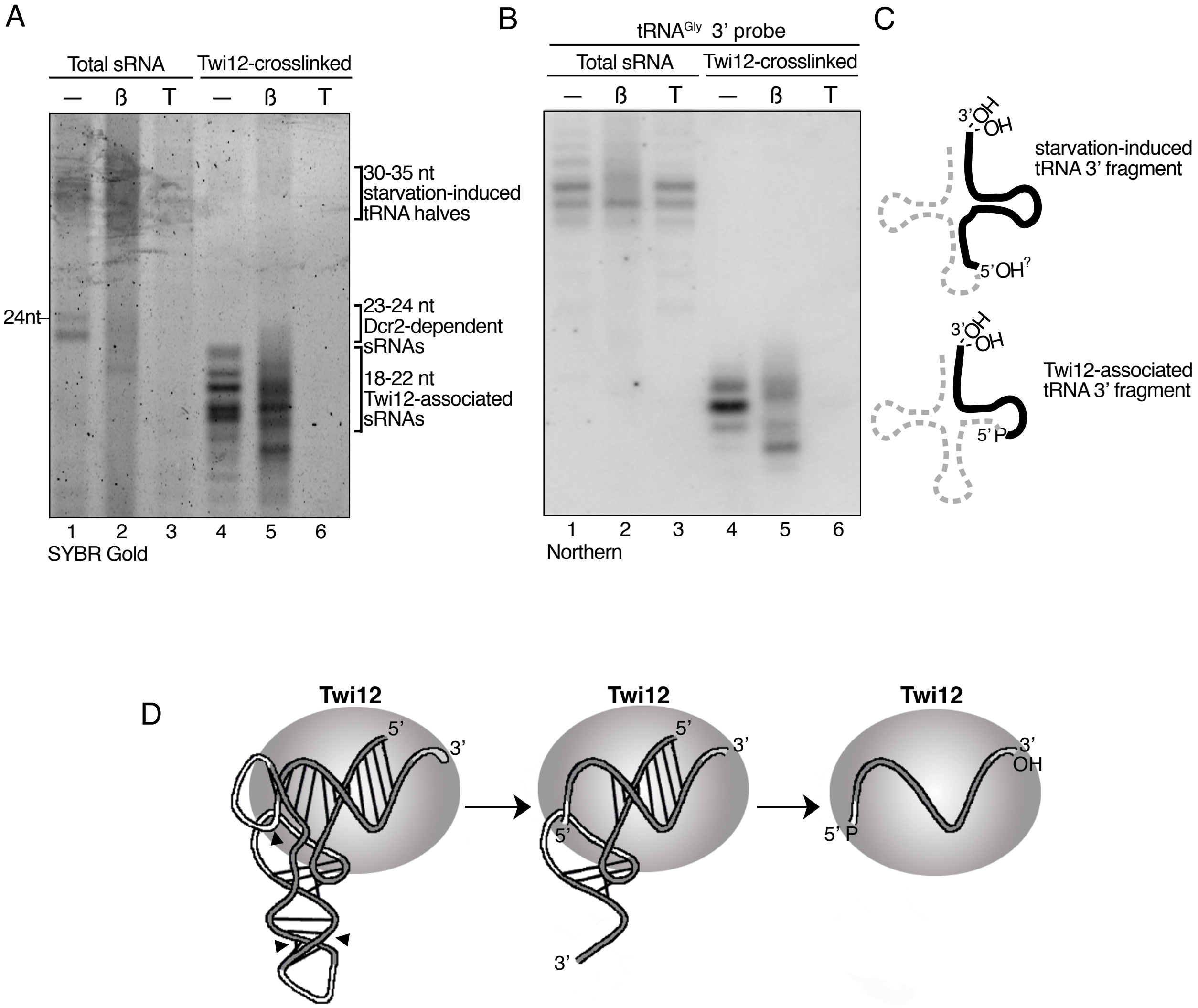


Figure 3. Distinct biochemical features of the Twi12-bound and starvation-induced tRNA 3' fragments.
(A,B,C) Size-enriched total RNA from cells harvested 3 hours after initiation of starvation (lanes 1-3) and crosslinked Twi12 sRNAs from vegetative growth (lanes 4-6) were subject to ß-elimination or Terminator exonuclease treatment followed by direct staining (A) or blot hybridization to detect a specific tRNA 3' fragment (B). The tRNAGly 3' fragment detected by blot hybridization is depicted in (C) as a solid black line against the dashed full-length tRNA.
(D) A speculative model is shown for Twi12 tRNA fragment loading and release. The line drawing is based on previous illustrations of tRNA tertiary structure (Purves et al. 1995).

Figure 4

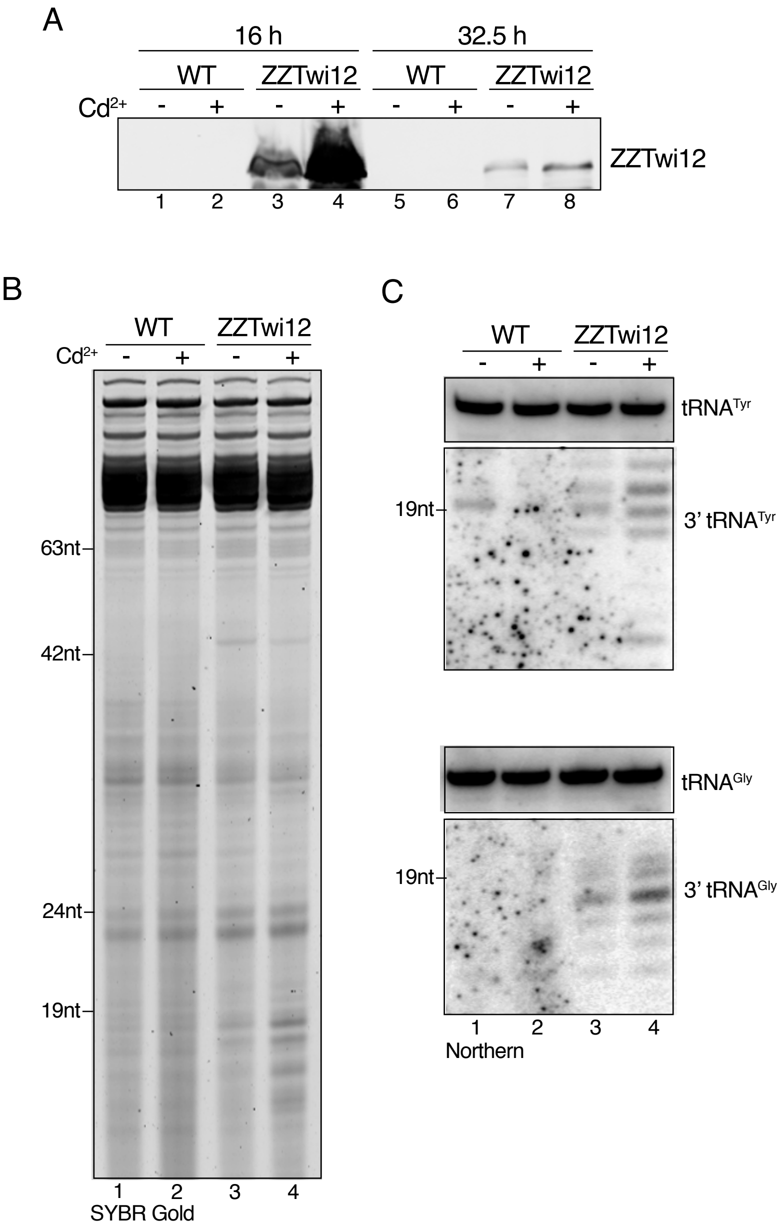
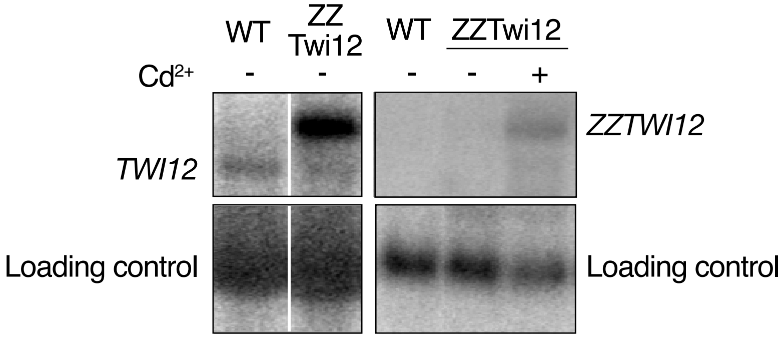


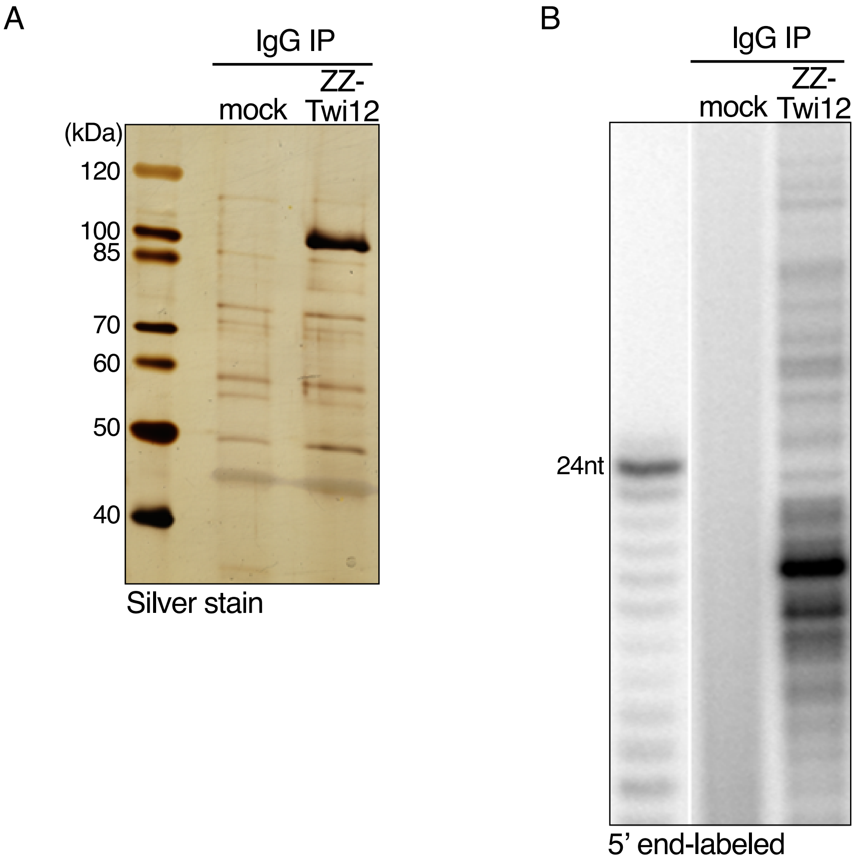
Figure 4. Increased tRNA 3' fragment accumulation upon Twi12 overexpression.
(A) ZZTwi12 accumulation in wild-type or transgene-containing cells grown with or without cadmium (Cd2+) was determined by immunoblot against the epitope tag at the indicated time points of continuous growth. Cadmium was first administered at time zero (0.75 μg/mL) and supplemented again after 12 hours (0.3 μg/mL) and 21.5 hours (0.5 μg/mL), followed by overnight culture to the maximal cell density of stationary phase.
(B,C) Total RNA was harvested and size-enriched from cells in stationary phase (32.5 hours) followed by direct staining (B) or blot hybridization (C). Similar results were observed using samples from cells harvested prior to reaching maximum density (data not shown), but stationary-phase culture sRNAs are shown here to allow comparison of Twi12-bound tRNA 3' fragments and all other tRNA fragment populations including tRNA halves (see Supplemental Fig. S3).

Supplemental Figure S1



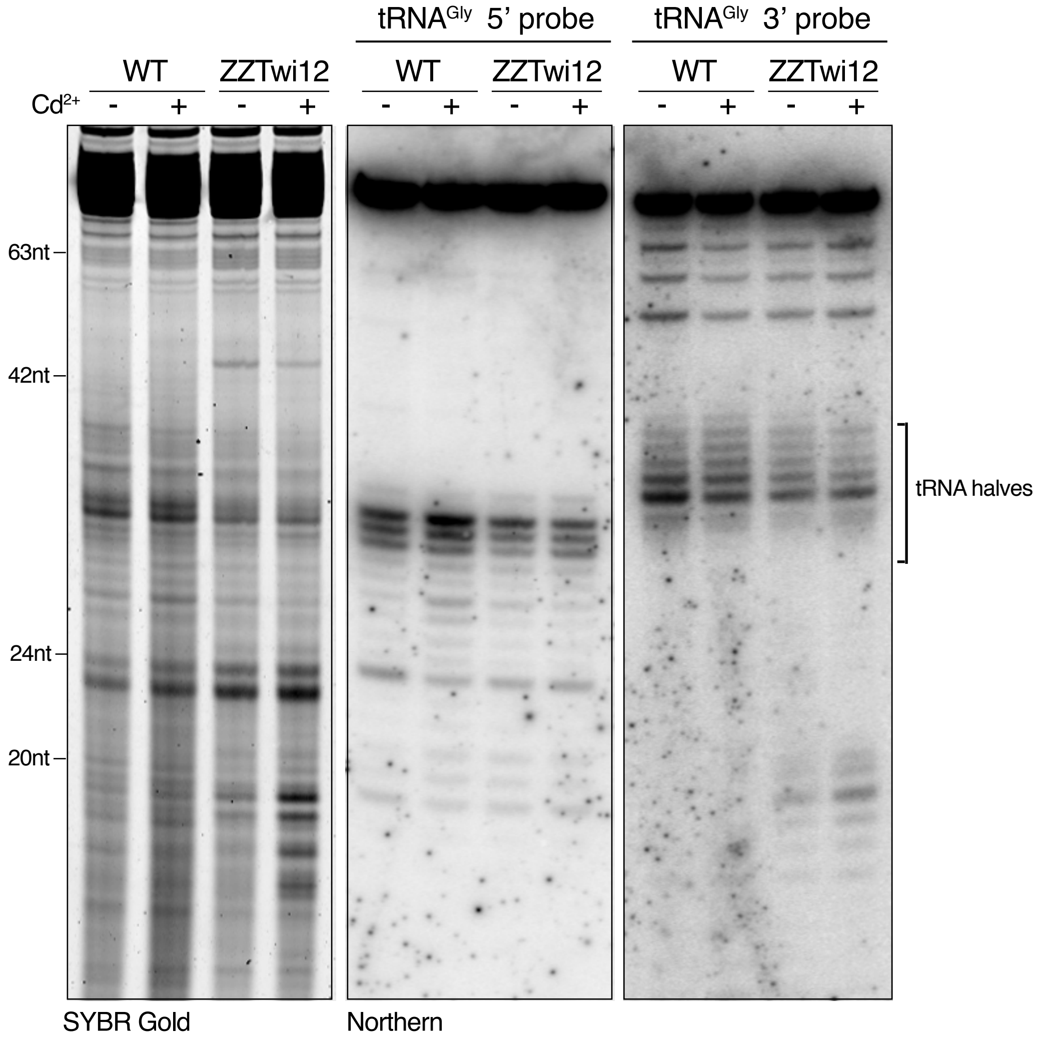
Supplemental Figure S1. Levels of expression for endogenous Twi12 mRNA (from the TWI12 locus) and transgene mRNA encoding ZZTwi12 (ZZTWI12 locus) in the wild-type strain (WT) or the transgene strain with an intact TWI12 locus. Total RNA from growing cells was used for Northern blot hybridization with a probe detecting the Twi12 mRNA open reading frame. Loading controls shown are rRNA (for the first two lanes, which were cropped from the exposure of the same blot) or an mRNA encoding a ribosomal protein (RPL21 locus; the last three lanes are from a separate experiment testing cadmium induction). The uninduced level of transgene mRNA was estimated by quantitation as ~5-fold the level of endogenous Twi12 mRNA in the wild-type strain. However, this ratio may vary across expression conditions due to the different promotersthat drive endogenous versus transgene mRNA expression.

# Supplemental Figure S2



Supplemental Figure S2. Immunopurified Twi12 is specifically associated with RNA.
(A) Silver stained protein content of purifications from the parental strain CU522 (mock) or the Twi12 transgene strain (ZZTwi12) and
(B) 5’ end-labeled associated RNA.

Supplemental Figure S3



Supplemental Figure S3. Neither tRNA halves nor Twi12 passenger-strand 5' tRNA fragments are increased in steady-state accumulation by Twi12 overexpression. The size-enriched RNA samples used in Figure 4 were directly stained (at left) or hybridized to probes detecting the 5' or 3' end of tRNA$^{Gly}$ (at center and right). The stress of high density growth induces the production of tRNA halves less than does starvation, but the tRNA halves produced by anticodon loop cleavage are still readily detectable (as indicated).

# CHAPTER THREE

## A Piwi protein with tRNA fragment cargo complexes with Xrn2 in the nucleus

### Introduction and Summary

Argonaute/Piwi (Ago/Piwi) RNP complexes use a variety of mechanisms to carry out a common function: down-regulation of target expression. Some Ago/Piwis have slicer activity and can directly cleave a target transcript. In other cases, the mechanism used depends on what proteins associate with a given Ago/Piwi. For example, Ago1 in *S. pombe* interacts with a chromodomain protein and also recruits a histone methyltransferase complex to direct heterochromatin formation at centromeres, telomeres, and mating type loci (Bayne et al., 2010; Buhler and Moazed, 2007) Argonautes loaded with miRNAs in animals interact with GW182 proteins which promote translational repression and deadenylation through interaction with poly(A) binding protein (Zekri et al., 2009). Piwi proteins loaded with piRNAs in animal germlines interact with Tudor domain-containing proteins, although the mechanistic consequence of this interaction is still unclear (Siomi et al., 2010).

We recently showed that an essential Piwi protein in *Tetrahymena*, Twi12, binds tRNA fragments from the 3' ends of all tRNAs (Couvillion et al., 2010). We showed that the fragments start at a conserved position in the T loop and end at the mature tRNA 3' terminus. We did not determine whether the fragments carry the canonical tRNA base modifications, which would have implications for target base pairing, nor whether the Twi12 RNP contains other proteins. Twi12 does not contain the slicer motif present in some Ago/Piwi proteins, so as suggested above there may be other proteins that carry out the essential function of the complex through association with the sRNA guide-containing Twi12. In fact, here we show that Twi12 stably interacts with a novel protein, Tan1, and a 5' to 3' exonuclease, Xrn2, in the nucleus. Xrn2 is essential for growth, like Twi12, and we find that when the interaction between the two is compromised, culture growth slows dramatically and aberrant RNAs accumulate. Therefore we have uncovered a novel mechanism used by Ago/Piwi RNP complexes. Future work will aim to define the nucleic acid targets and target sites of the complex.

## Results and Discussion

Previous studies of Twi12 and bound RNA used a strain that overexpresses tagged Twi12 (Couvillion et al., 2010). During the course of creating a *Tetrahymena* strain that expresses tagged Twi12 from the endogenous locus we found an in-frame ATG upstream of the predicted start codon. This ATG is included in the transcript (data not shown), and is therefore likely used as the start codon. Accordingly, we have been able to detect endogenous expression of the longer form of the Twi12 protein, but have not been able to detect the shorter version in a wild type strain. For this reason we made new strains in which the longer version of Twi12 is N-terminally tagged with two Protein A domains, a Tobacco Etch Virus cleavage site, and three copies of the FLAG peptide: ZZ-TEV-3XFLAG (ZZF). We have renamed the short version Twi12$^{S}$ and Twi12 now designates the full-length protein (see Fig. 1A for a schematic of the Twi12 proteins referred to here). For overexpression, the tagged *TWI12* or *TWI12$^{S}$* transgene is expressed from the cadmium-inducible *MTT1* promoter at the non-essential *BTU1* locus. *MTT1*-driven ZZFTwi12 (mZZFTwi12) binds ~18- 22-nt tRNA 3' fragments, as previously reported for *MTT1*-driven ZZTwi12$^{S}$ (mZZTwi12$^{S}$) (Fig. 1B, deep sequencing data not shown, and Couvillion et al., 2010).

Mature tRNAs are decorated by base modifications that are important for folding and stability of tRNAs or translation or translation fidelity (Phizicky and Alfonzo, 2010). Evidence suggests that Twi12-bound tRNA fragments are derived from mature tRNAs (Couvillion et al., 2010), but the base modification status of the fragments has not previously been tested. Although specific *Tetrahymena* tRNA base modifications are not known, several modified positions are conserved across tRNAs and across eukaryotic species. Two-dimensional thin layer chromotography of ZZFTwi12-bound 18- to 22-nt sRNAs showed the presence of the base modifications pseudouridine (Ψ) and 1-methyl adenosine (m$^{1}$A), conserved modifications found at the 3' end of tRNAs (Fig. 2A, middle panel). Dihydrouridine (D) and thymidine (T) on the other hand are not detectable in bulk ZZFTwi12-bound 18- to 22-nt sRNAs, although they are in full-length tRNAs (Fig. 2A, left panel and see Fig. 2B). None of the conserved tRNA base modifications are detectable in total 23- to 24-nt sRNAs, which are produced by Dicer (Fig. 2A, right panel). Interestingly, the m$^{1}$A is predicted to be located at the fourth position in the Twi12-bound sRNA, within a potential "seed sequence" for target base pairing. This bulky base modification would affect canonical base pairing, making targets difficult to predict.

The Twi12-tRNA fragment RNP may have a function downstream of assembly that serves to guide the complex to a nucleic acid target. Alternatively Twi12 may function only to alter the tRNA pool or affect translation, and the RNP complex once assembled may be dissociated to liberate free Twi12. To gain insight into the function of Twi12, we looked for interacting proteins. MudPIT mass spectrometry analysis after a two-step immunopurification (IP) of ZZFTwi12 driven by its endogenous promoter (eZZFTwi12) (Fig. 3A) identified a 12.3 kD protein with no homologs that we named Tan1 (Twi-associated novel 1), and a 126 kD 5' to 3' monophosphate-dependent exonuclease (Fig. 3B). The *Tetrahymena* genome encodes three predicted 5' to 3' exoribonucleases. Two do not cluster with either the Xrn1 or Xrn2 subclass, and are cytoplasmic (Doug Chalker, personal communication), and one is most similar to Xrn2/Rat1 (Supplemental Fig. S1). Xrn2/Rat1 homologs are generally nuclear, whereas Xrn1 homologs are generally cytoplasmic, so the interaction of Twi12 with Xrn2 was surprising because we previously reported ZZTwi12$^{S}$ to be cytoplasmic (Couvillion et al., 2009). However, indirect immunofluoresence (IF) of endogenously tagged ZZFTwi12, Tan1FZZ, and Xrn2FZZ revealed each of them to be nuclear (Fig. 3C). IF analysis of *MTT1*-driven ZZFTwi12 (mZZFTwi12) showed mostly nuclear localization when transgene transcription was not induced, and cytoplasmic localization when transcription was induced by addition of cadmium (Supplemental Fig. S2A). Therefore it seems overexpression can cause altered apparent localization of nuclear proteins. *MTT1*-driven ZZTwi12$^S$ (mZZTwi12$^S$) showed cytoplasmic localization whether transgene transcription was uninduced or induced. However, it is not clear whether the cytoplasmic localization of the shortened protein implicates the N terminus in nuclear localization because basal expression of mZZTwi12$^S$ is higher than basal expression of mZZFTwi12 (Supplemental Fig. S2B). Current work will resolve this question.

To test whether Tan1 and Xrn2 are essential for growth we attempted to replace their loci with a drug resistance cassette (neo2*)* on all chromosomes in the somatic macronucleus by phenotypic assortment. *TAN1* can be fully replaced by neo2 (Supplemental Fig. S3A and additional RT-PCR data not shown). Therefore it is not essential. *XRN2* can be only partially replaced (Supplemental Fig. S3B), suggesting that it is essential. Xrn2/Rat1 homologs are known to be important for many nuclear processes including rRNA maturation (Geerlings et al., 2000; Henry et al., 1994; Wang and Pestov, 2011), RNA polymerase II termination (Luo et al., 2006; West et al., 2004), telomere elongation (Luke et al., 2008), and degradation of hypomodified tRNA (Chernyakov et al., 2008). For this reason we suspected Twi12 to be only one of many Xrn2 interactors and thought it likely that Xrn2 was involved in loading Twi12 by trimming tRNA fragment intermediates. Therefore we first focused analysis on Tan1.

IP of endogenously tagged Tan1 (eTan1FZZ) co-purifies Twi12 (Fig. 4A), as well as Twi12-bound sRNA (Fig. 4B) verifying their interaction. However, knockout of Tan1 does not affect binding of overexpressed ZZFTwi12 (mZZFTwi12) to 3' tRNA fragments, accumulation of tRNA fragments in a strain overexpressing ZZFTwi12 (mZZFTWI12), or nuclear localization of mZZFTwi12 or eXrn2FZZ (data not shown). Although Tan1 has no conserved domains of known function, we hypothesized that it may be a ribonuclease because it is small, basic, and interacts with an Ago/Piwi protein with no slicer motif. To test this we expressed recombinant Tan1 in *E. coli* to use in nuclease assays (Fig. 5A). When given *Tetrahymena* total RNA, recombinant Tan1 has detectable activity on only one RNA species under the conditions used: 5.8S rRNA (Fig. 5B). It is unclear whether this is nuclease activity or some other activity that slightly increases the mobility of a portion of 5.8S rRNA in denaturing gel electrophoresis.

We next used Tan1FZZ from *Tetrahymena* for in vitro nuclease assays. Not surprisingly, we detected a 5' monophosphate-dependent nuclease activity, which is consistent with co-purification of Xrn2 (Fig. 6A). As expected, ZZFTwi12 also co-purifies Xrn2 activity (Fig. 6A and 6B, lane 2). *S. pombe* Rat1 interacts with an activating partner, Rai1, which allows Rat1 to degrade RNAs with stable secondary structure more effectively, and also possesses pyrophosphohydrolase activity, converting 5' triphosphate into 5' monophosphate (Xiang et al., 2009). To test whether Tan1 may play a role in preparing substrates for Xrn2, directly enhancing Xrn2 activity, or linking Xrn2 to Twi12, we immunoprecipitated ZZFTwi12 from a strain lacking Tan1. No detectable change in Xrn2 activity is observed in the absence of Tan1 (Fig. 6B, lane 3). Thus, we have not yet been able to define any activity or find any function for Tan1.

Twi12$^S$ is truncated by 17 amino acids on the N terminus compared to Twi12, and starts at an in-frame internal ATG. Strikingly, IP of mZZTwi12$^S$ co-purifies a severely reduced Xrn2 activity (Fig. 6B, lane 4). It is unclear whether this reflects a reduced ability of Xrn2 to interact with Twi12$^S$ because the N terminus contributes to the interaction, or because the *MTT1*-driven ZZTwi12$^S$ localizes to the cytoplasm. Current work will resolve this question. No matter the reason for the reduced interaction, interestingly, a strain expressing *MTT1*-driven ZZTwi12$^{S}$ in which the endogenous locus has been deleted (mZZTWI12$^{S}$; eTWI12Δ) has a slow growth phenotype. On the other hand, the analogous strain expressing full-length Twi12 (mZZFTWI12; eTWI12Δ), which does co-purify Xrn2 activity, has no growth phenotype (Fig. 7A). This suggests that interaction with Xrn2 may be essential for Twi12 function.

In addition to slow growth we have also noticed increased accumulation of apparently aberrant RNA species in the strain mZZTWI12$^{S}$; eTWI12Δ compared to both eZZFTwi12 and mZZFTWI12; eTWI12Δ. These include a telomeric repeat transcript, a short form of a non-coding RNA of unknown function that is complementary to a Twi7-bound 24-nt sRNA, and a macronuclear high-copy repeat transcript (Fig. 7B, right panels and data not shown). RNAs of defined function however, such as full-length tRNA and pre-tRNA, do not accumulate differentially in these strains (Fig. 7B, middle panels). Current studies aim to confirm and extend these preliminary observations. It is possible that this RNA accumulation phenotype is a result of compromised interaction of Xrn2 with mZZTwi12$^{S}$. If this is the case, it will be important to determine whether the accumulation results from a lack of Xrn2 activity on these RNA species.

To further investigate the role of Xrn2 in the Twi12 complex we immunopurified eXrn2FZZ. To our surprise, the only stably associated proteins under the IP conditions used that are detectable by silver staining are Twi12 and Tan1 (Fig. 8A). Xrn2 might therefore only act in the Twi12 complex, and the essential function of Xrn2 might be the same as the essential function of Twi12: the function of the complex. It is also possible that Xrn2 acts alone or interacts transiently with other proteins separate from the Twi12 complex (Fig. 8B). Current and future experiments will address this question (see below).

## Conclusions and Future Directions

We have uncovered a novel nuclear Piwi protein-containing complex composed of tRNA fragments, Twi12, Tan1, and Xrn2. Tan1 is not essential and no knockout phenotype has yet been identified. Twi12 and Xrn2 are both essential and preliminary data presented here suggests their essential functions are related. Xrn2 may be important for facilitating loading of Twi12 with sRNAs, carrying out the catalytic activity of the complex guided by the Twi12-bound sRNA, or both. It is also tempting to speculate that Twi12 may be required for some or all of the conserved nuclear functions of Xrn2 in *Tetrahymena*.

Several experiments are currently underway to validate the data presented here. First, since it is unclear whether mZZTwi12$^S$ localizes to the cytoplasm because of N-terminal truncation or because of overexpression, we are creating a strain that expresses endogenously tagged Twi12$^S$ (eZZFTwi12$^S$). If eZZFTwi12$^S$ is nuclear, we can also use this strain to test whether the interaction with Xrn2 depends on the N terminus of Twi12. We have also repeated the Northern blotting analysis for aberrant RNAs in mZZFTWI12$^S$; eTWI12Δ and will be checking for accumulation of additional aberrant species such as rRNA processing intermediates and antisense transcripts. We also included in this analysis RNA from *TAN1* knockout, *XRN2* knockdown, and *TWI12* knockdown strains. *XRN2* and *TWI12* can only be knocked down to ~50% using phenotypic assortment (Supplemental Fig. S3 and data not shown), so it is not surprising that we have not detected aberrant RNA accumulation in these strains (data not shown). We plan to create inducible knockout strains to study *XRN2* and *TWI12* knockout phenotypes. The open reading frame of each will be knocked into the *MTT3* locus under the control of the *MTT3* promoter. This promoter is cadmium inducible and has very little if any basal expression at its endogenous locus. Then, in the presence of cadmium, the endogenous *XRN2* or *TWI12* open reading frame will be replaced with a drug resistance cassette by phenotypic assortment. When cadmium is removed from the media transcription from the *MTT3* locus will stop. Protein levels will be monitored by immunoblot, as timing will depend on the half-life of each protein.

Direct identification of RNA targets will provide information about the function and mechanism of the Twi12 complex. We are using photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) to identify RNA targets of eZZFTwi12. This method involves incorporating 4-thiouridine (4SU) into transcripts in growing cells then using 365 nm light to crosslink *in vivo*. A unique feature of PAR-CLIP as compared to other CLIP strategies is that it allows identification of the site of crosslinking within a target (Hafner et al., 2010). We have found that *Tetrahymena* will only take up 4SU to detectable levels when grown in synthetic media lacking uridine, and supplemented with 100 μM 4SU (data not shown). The experiment is currently in progress, and initial steps have been successful (Fig. 9). Targets found using PAR-CLIP of ZZFTwi12 and Xrn2FZZ will be compared to find the extent of overlapping substrates.

PAR-CLIP may clearly identify target sites that could be complementary to base-modified 3' tRNA fragments. However, if it does not, there is no evidence that sRNA binding is required for the function of Twi12. It is also possible that sRNA binding has a role other than as a guide. For example, it may be required for nuclear import. *Tetrahymena* Twi1 is not imported into the nucleus until it is loaded with a single stranded RNA to make a mature functional complex (Noto et al., 2010). Alternatively, sRNA binding to Twi12 could induce a conformational change important for complex activity. To test the requirement for sRNA binding we are constructing a Twi12 mutant in which a conserved tyrosine residue important for 5' phosphate binding and base stacking with the first nucleotide of the sRNA (Ma et al., 2005) is mutated to a glutamate residue. This substitution has been shown to prevent efficient sRNA binding for human Ago2 (Rudel et al., 2011).

Figure 1

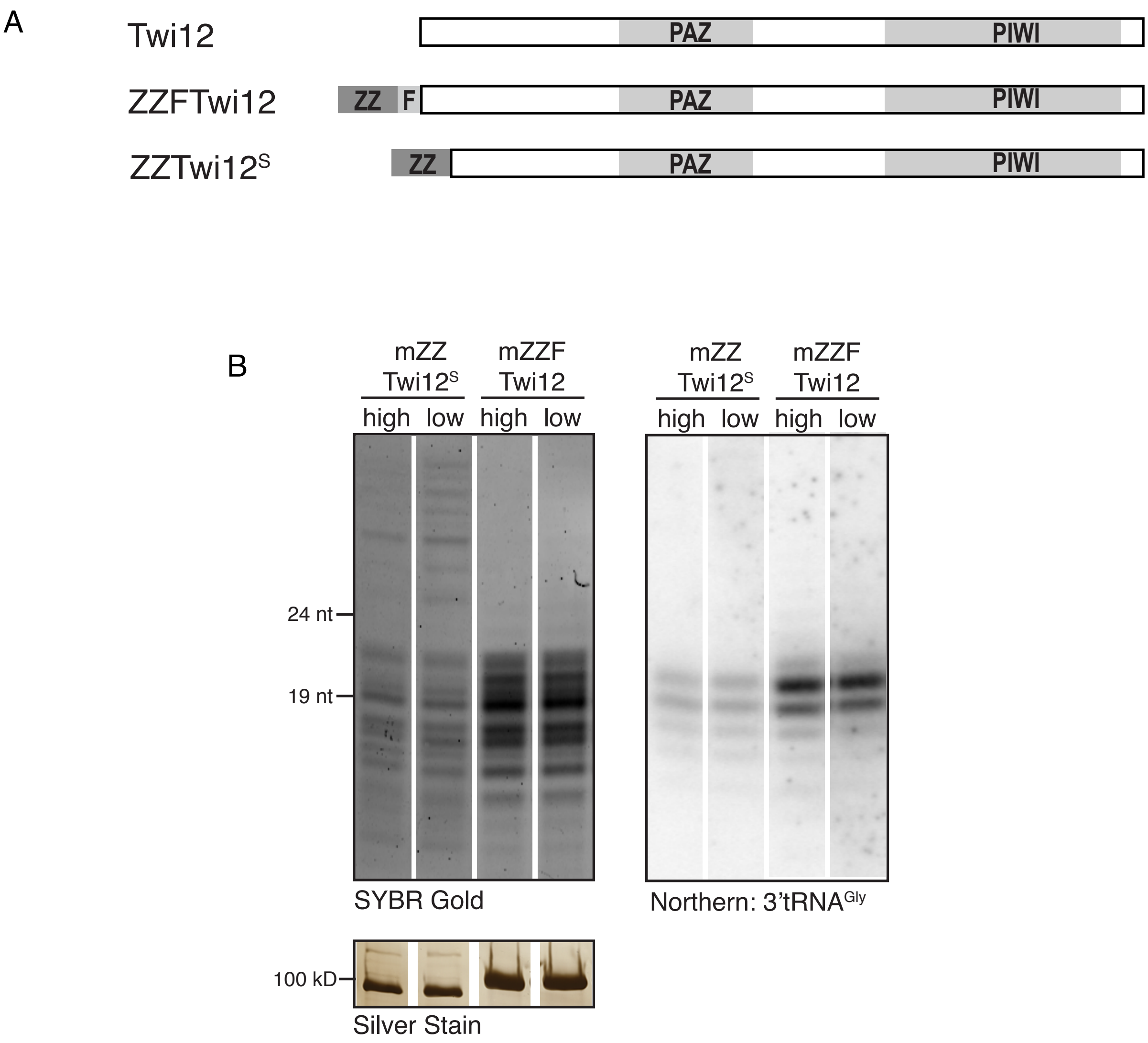


Figure 1. Full-length Twi12 binds 3' tRNA fragments
(A) Schematic of Twi12 and the fusion proteins used and referred to in this report. Twi12$^S$ is truncated by 17 amino acids on the N terminus. ZZ: two Protein A domains followed by a TEV cleavage site. F: three copies of the FLAG epitope.
(B) Comparison of small RNAs that co-precipitate with MTT1-driven ZZTwi12$^S$ (mZZTwi12$^S$) and MTT1-driven ZZFTwi12 (mZZFTwi12) under two different conditions after IgG IP and Tev elution. High: binding in 100 mM NaCl, 0.1% TritonX-100. Low: binding in 50 mM NaCl, no TritonX-100.

Figure 2

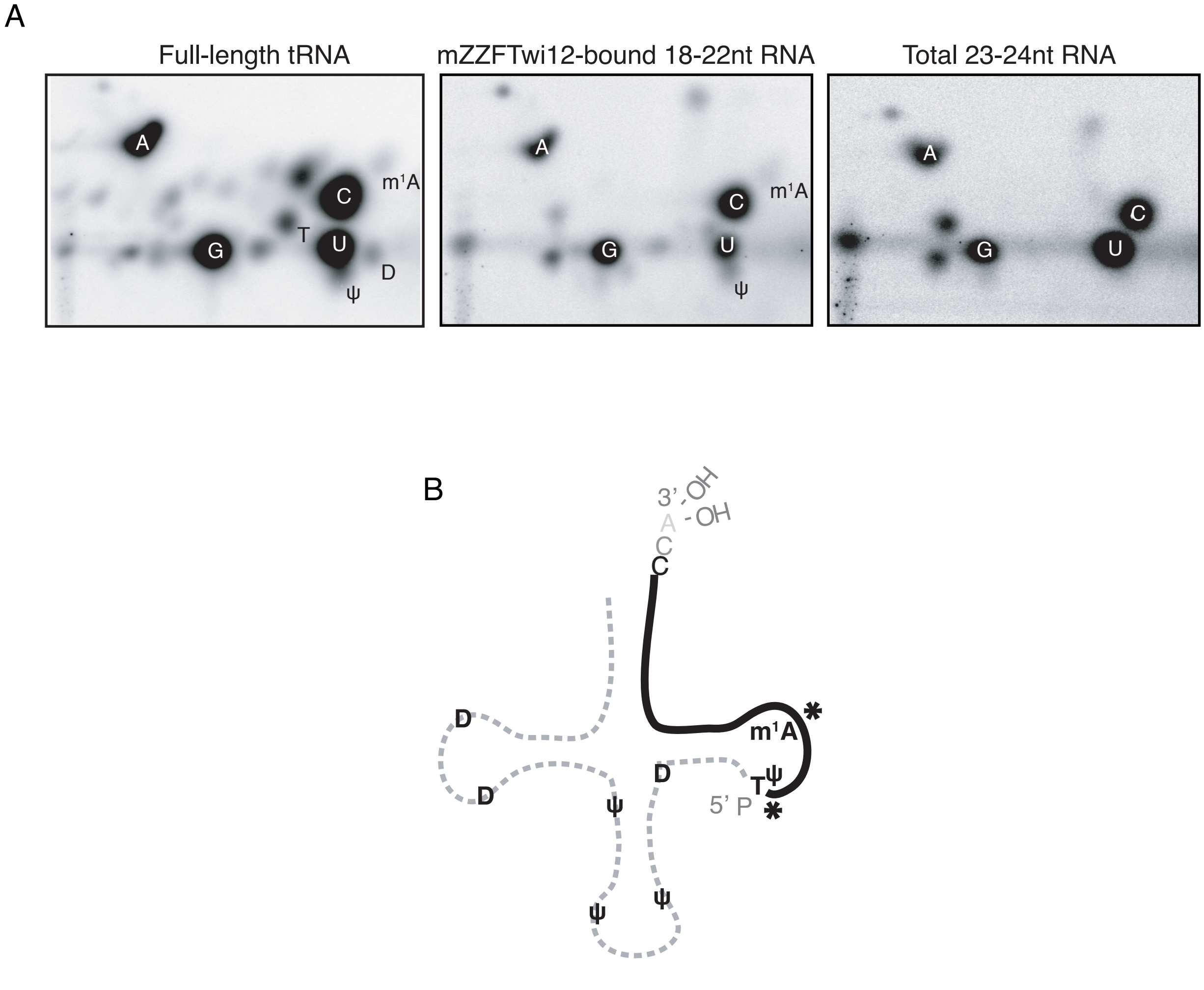
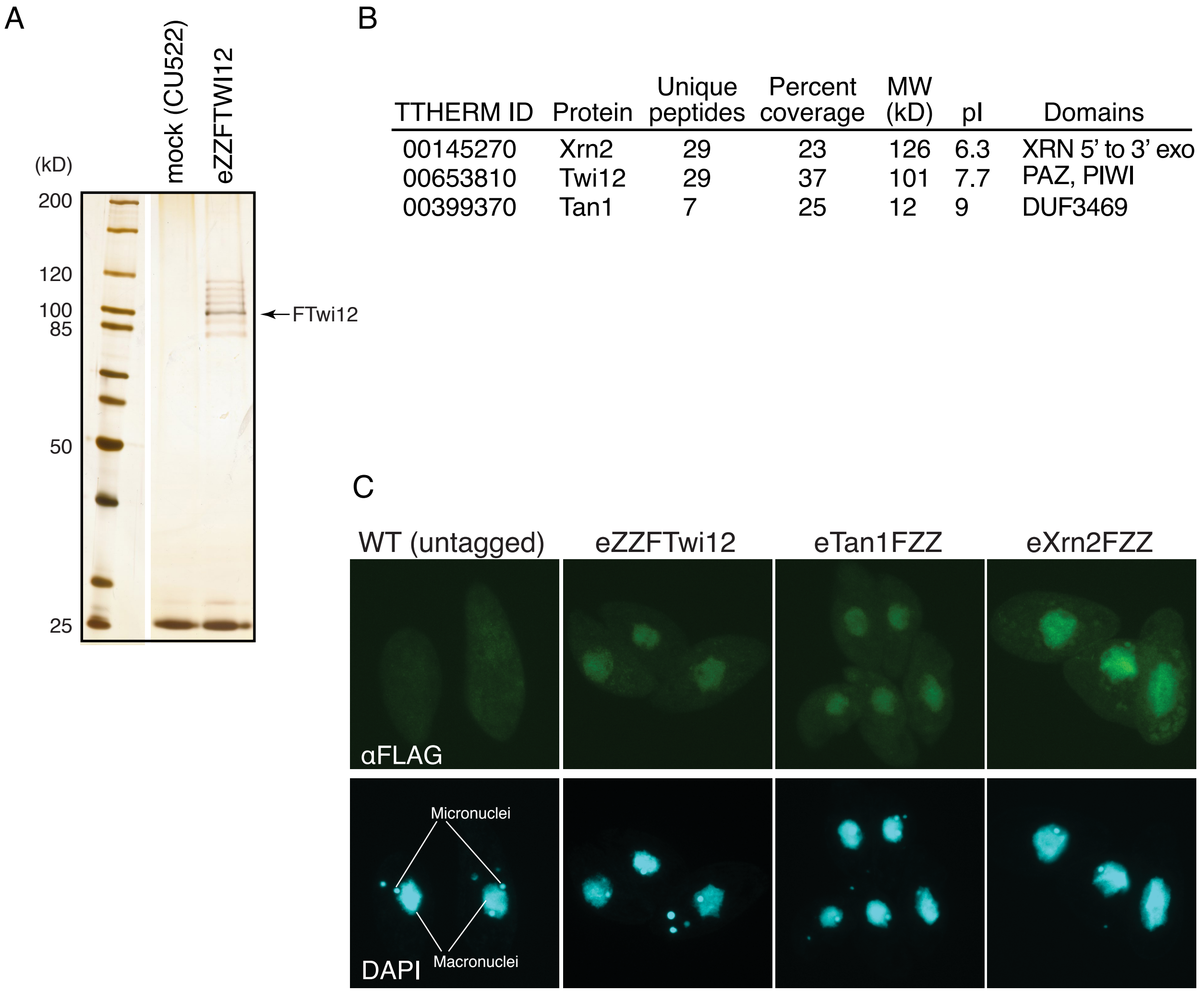


Figure 2. Twi12-bound sRNAs contain base modifications.
(A) Two-dimensional thin layer chromatography analysis of post-labeled 5' monophosphate nucleosides from full-length tRNA, mZZFTwi12-bound sRNA, or total 23- to 24-nt sRNA. Conserved identifyable bases are labeled. First dimension is isobutyric acid: ammonium hydroxide: water, 66:1:33. Second dimension is 0.1 M sodium phosphate buffer pH 6.8: ammonium sulfate: 1-propanol, 100:60:2.
(B) Diagram showing a generic tRNA with the positions of the conserved base modifications labeled in A. Solid black line: Twi12-bound fragment.

Figure 3



| TTHERM ID | Protein | Unique peptides | Percent coverage | MW (kD) | pI | Domains |
|---|---|---|---|---|---|---|
| 00145270 | Xrn2 | 29 | 23 | 126 | 6.3 | XRN 5' to 3' exo |
| 00653810 | Twi12 | 29 | 37 | 101 | 7.7 | PAZ, PIWI |
| 00399370 | Tan1 | 7 | 25 | 12 | 9 | DUF3469 |

Figure 3. Twi12 interacts with Tan1 and Xrn2 in the nucleus.
(A) Silver stained SDS-gel showing proteins purified with two steps: IgG IP, Tev elution, αFLAG IP, 8 M urea elution.
(B) eZZFTwi12-associated proteins identified by mass spectrometry. It is unclear which silver-stained band corresponds to Xrn2, and Tan1 is would have ran off the gel bottom in A.
(C) Immunofluorescence showing localization of each of the endogenously tagged proteins identified by mass spectrometry. WT is a the background strain without a tagged protein.

Figure 4

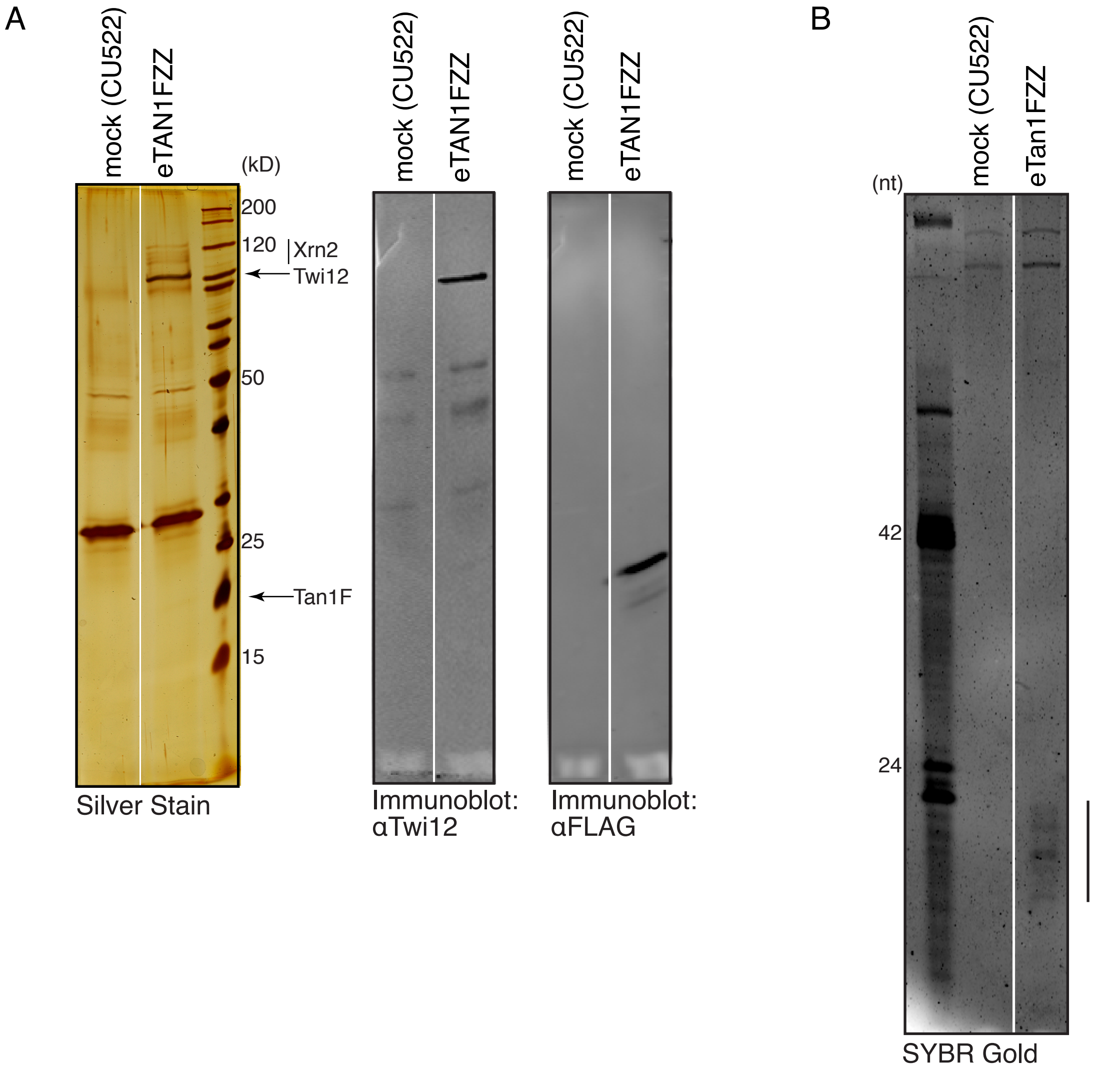


Figure 4. Tan1FZZ co-purifies Twi12.
(A) Left panel: silver stained SDS-gel showing proteins precipitated by IgG IP (Tev eluted) in the strain expressing endogenously tagged Tan1 (eTAN1FZZ). Right panels: immunoblots verifying the identity of the bands in the silver-stained gel. Note that Tan1 does not silver stain well.
(B) SYBR Gold-stained urea-gel showing the RNA co-precipitated with eTan1FZZ (line).

Figure 5

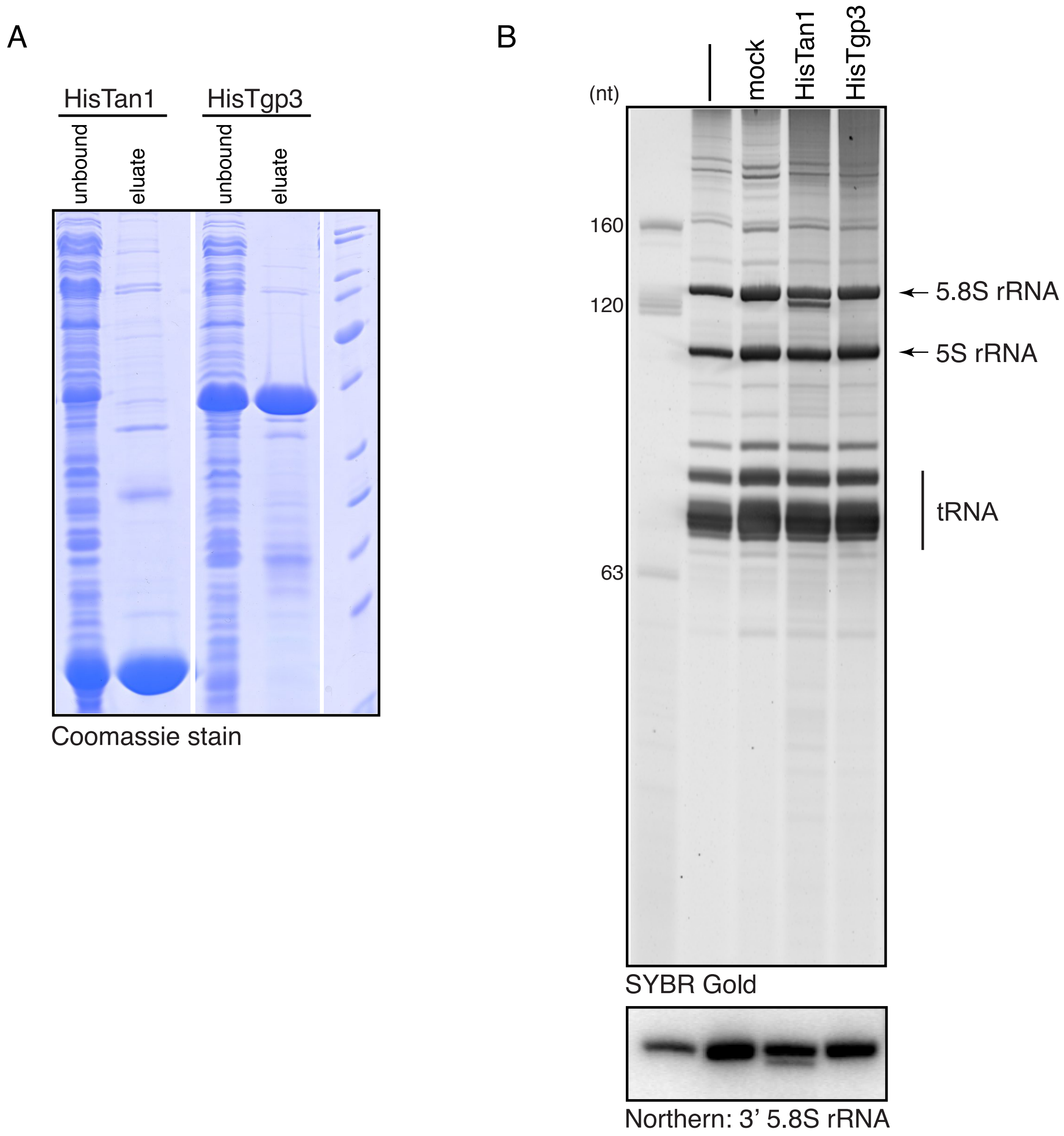


Figure 5. Recombinant Tan1 alters 5.8S rRNA in vitro.
(A) Coomassie stained SDS-PAGE showing HisTan1 or HisTgp3 (a negative control) purified from *E. coli*.
(B) In vitro activity assay in which the eluates shown in A were mixed with 4 $\mu$g of *Tetrahymena* total RNA in buffer containing 10 mM KCl and 1 mM $MgCl_2$ buffered in 20 mM Tris pH 7.5. The reaction was allowed to proceed at 30°C for 45 minutes.

Figure 6

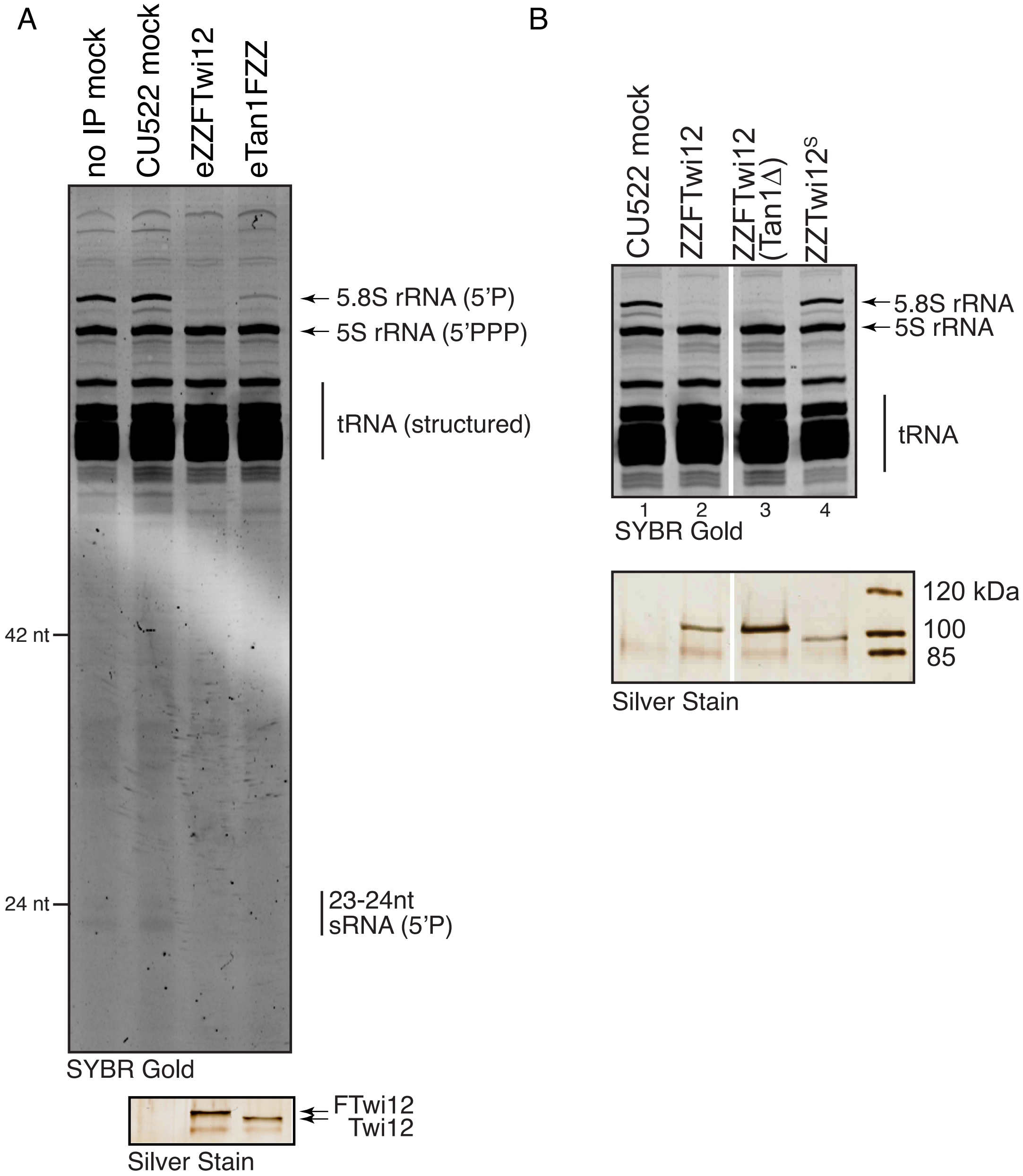


Figure 6. *Tetrahymena* Tan1FZZ and ZZFTwi12, but not ZZTwi12$^S$, co-purify Xrn2 activity.
(A) In vitro activity assay in which the eluates from a one-step IgG IP (Tev eluted) were mixed with 300 ng of *Tetrahymena* sRNA in buffer containing 50 mM KCl and 1 mM $MgCl_2$ buffered in 10 mM Tris, pH 7.5. The reaction was allowed to proceed at 30°C for 60 minutes. FTwi12 or Twi12 co-precipitated by Tan1 shown in lower panel (Tan1 and Xrn2 are not detectable in the silver stained gel).
(B) Assay as in A.

Figure 7

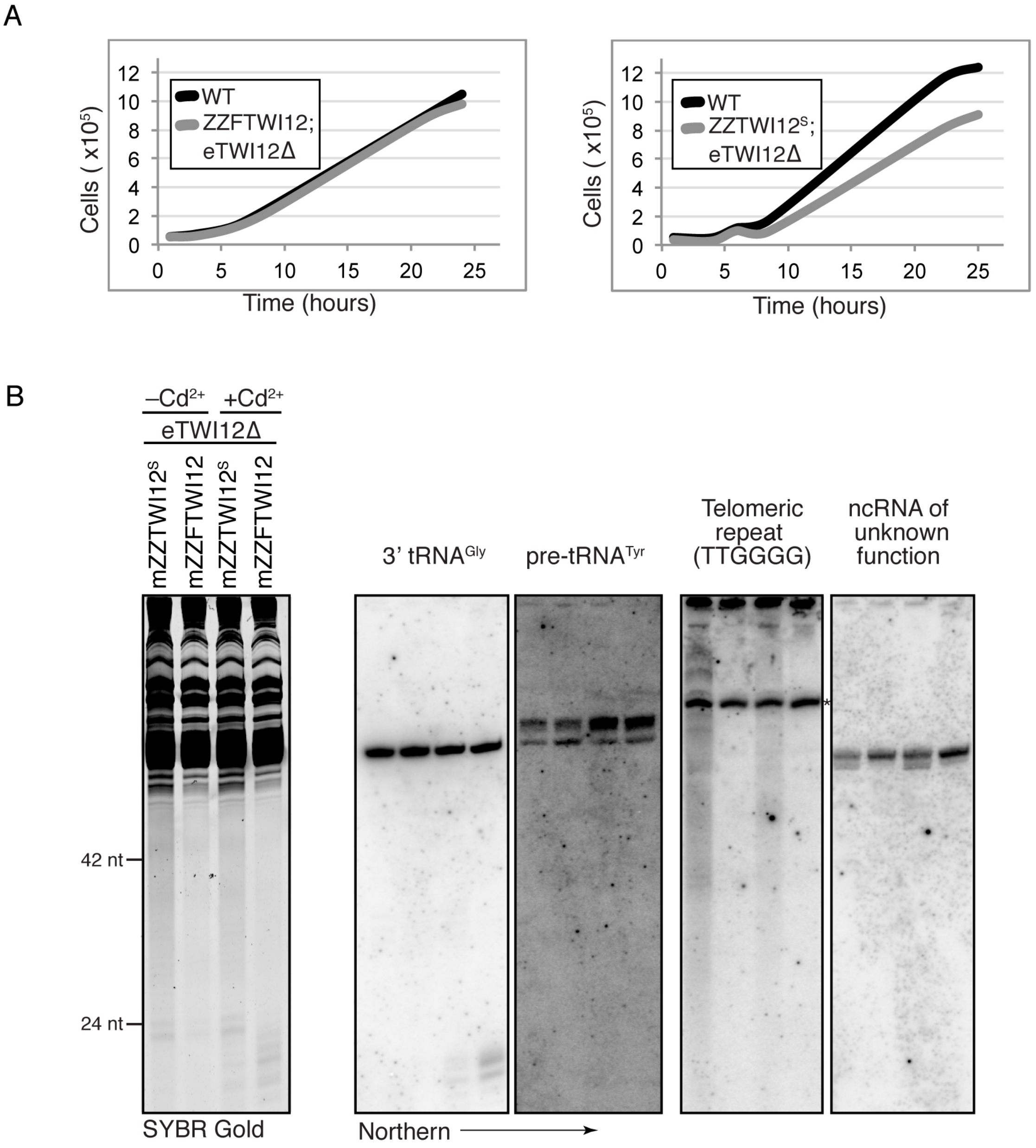


Figure 7. ZZTwi12$^S$ does not fully complement endogenous Twi12 knockout.
(A) Growth curves for which cells were starved overnight then inoculated at 0.5 x 10$^5$ cells per mL.
(B) Five $\mu$g total RNA from the strains indicated was resolved by urea-PAGE and SYBR Gold stained (left) and probed for the RNA species indicated (right). Asterisk indicates cross-hybridization of the $(C_4A_2)_3$ probe with 5S rRNA.

Figure 8

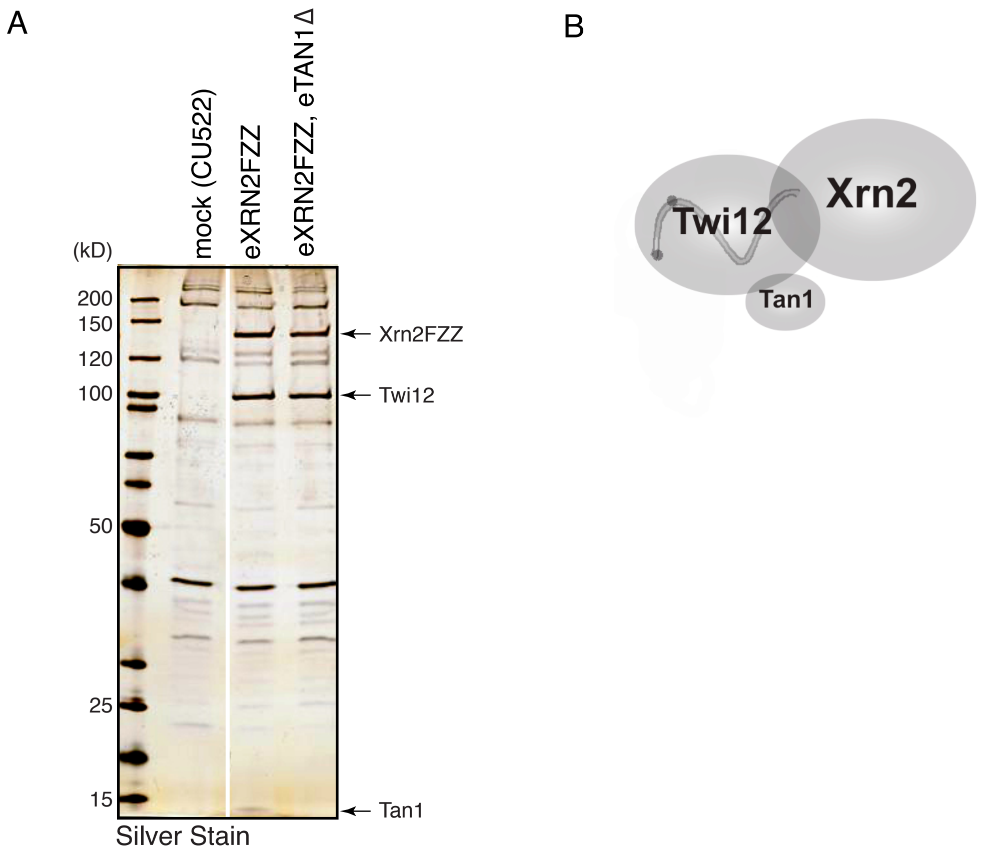


Figure 8. Xrn2 does not stably associate with protein partners other than tTwi12 and Tan1.
(A) Silver stained SDS-gel showing proteins purified by one-step αFLAG IP (eluted with FLAG peptide) from the strains indicated.
(B) Model of Twi12-Xrn2-Tan1 complex. Twi12 and/or Xrn2 may also have functions separate from the complex.

Figure 9

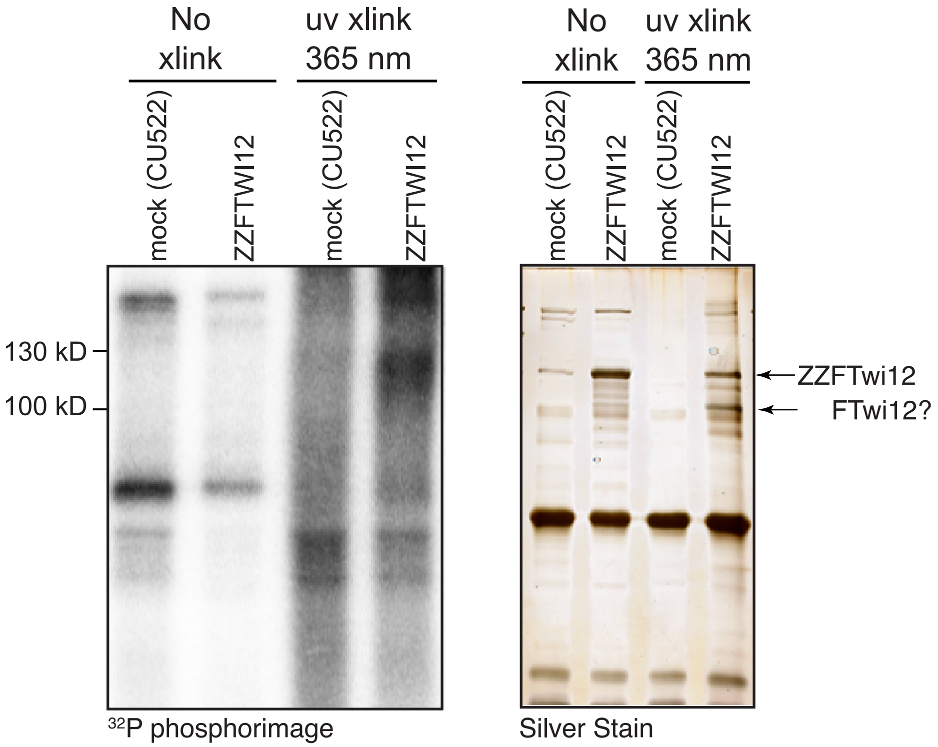
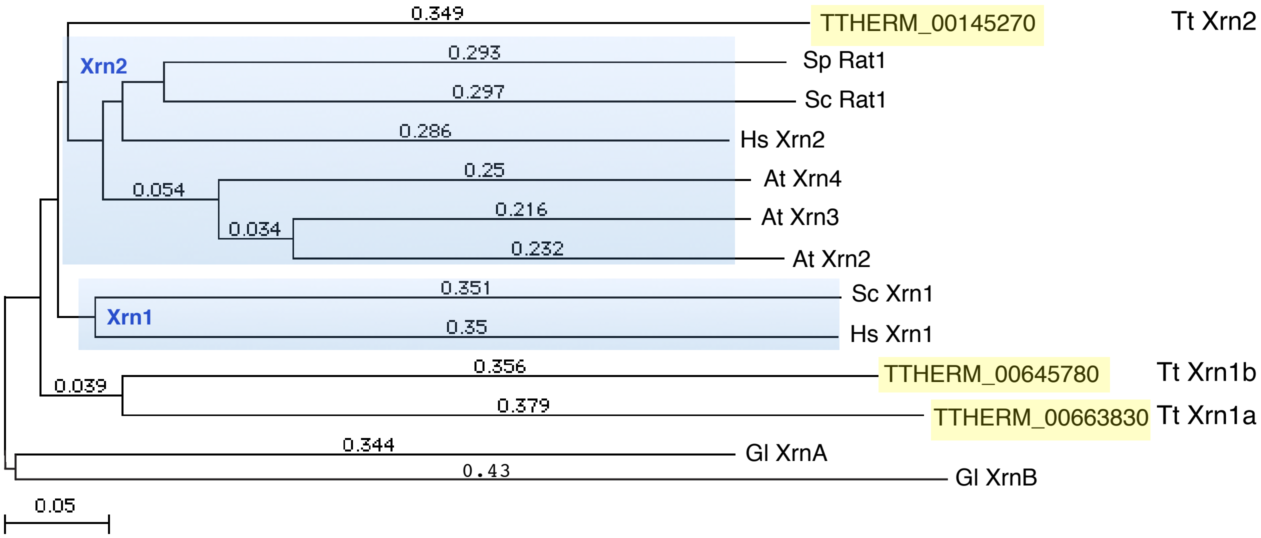


Figure 9. PAR-CLIP method progress.
Left panel: Phosphorimage of SDS-gel resolving 5’-$^{32}$P-labeled RNA crosslinked to FLAG-precipitated (SDS eluted) ZZFTwi12 in cells grown in 4SU. Right panel: matching silver stained SDS-gel.

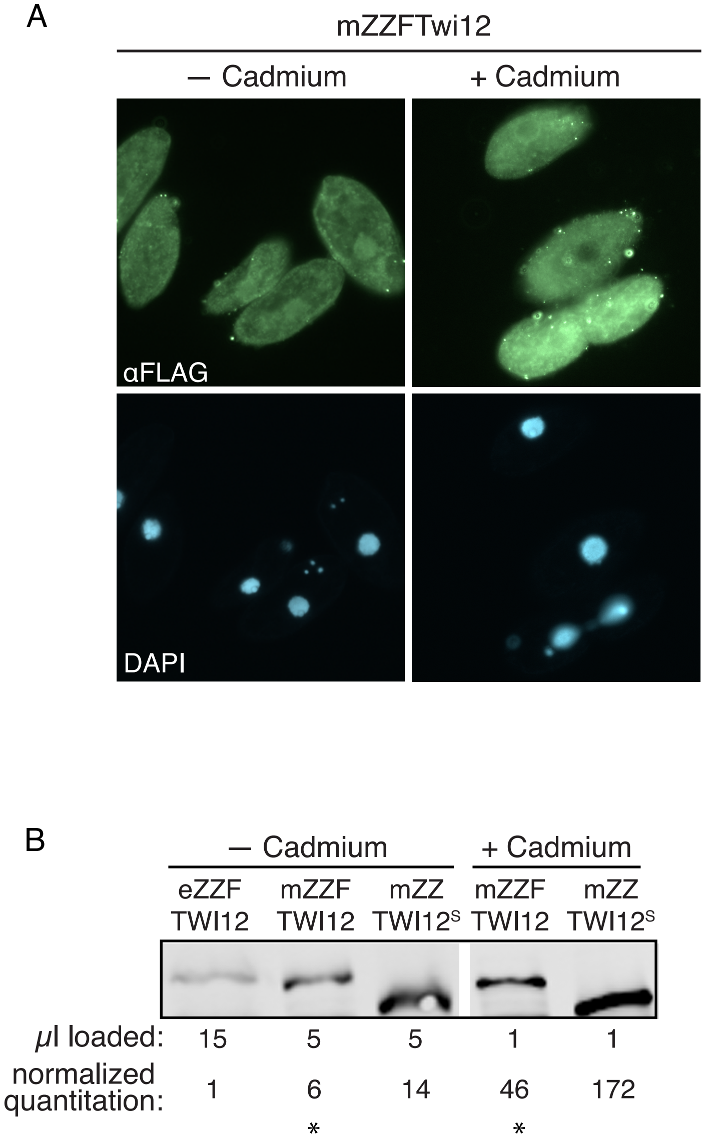
## Supplemental Figure S1

ClustalW (v1.4) Multiple Alignment Parameters:
Open Gap Penalty = 2.0 ; Extend Gap Penalty = 0.1 ; Delay Divergent = 40%
Gap Distance = 8; Similarity Matrix = blosum



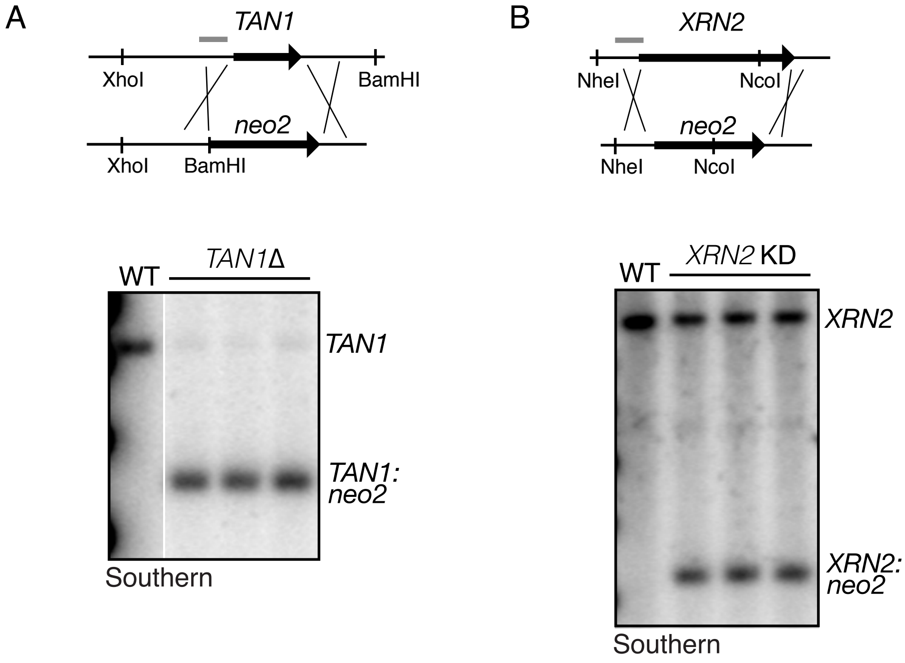
Supplemental Figure S1. The *Tetrhahymena* genome encodes three XRN family proteins. Multiple alignment guide tree of XRN family proteins using translated predicted open reading frames for the *Tetrahymena* proteins. Alignments were performed using ClustalW (v1.4) in the MacVector 8.0 software, with the parameters shown above. Scale bar and numbers represent genetic distance. The names assigned to the *Tetrahymena* Xrn proteins based on this alignment are given to the right. Abbreviations: At: *Arabidopsis thaliana*; Gl: *Giardia lambia*; Hs: *Homo sapiens*; Sc: *Saccaromyces cerevisiae*; Sp: *Schizosaccoaromyces pombe*; Tt: *Tetrahymena thermophila*.

Supplemental Figure S2



Supplemental Figure S2. Overexpressed Twi12 mislocalizes to the cytoplasm.
(A) Immunofluorescence showing localization of mZZFTwi12 in cells grown in complete synthetic media with and without 0.1 $\mu$g/mL cadmium chloride (for 3 hours).
(B) Immunoblot showing expression of tagged Twi12 in the strains indicated with and without 0.1 $\mu$g/mL cadmium chloride (for 3 hours). Amount loaded is from whole cell lysates of equal cell equivalent concentrations. Asterisks indicate strain and conditions shown in A.

Supplemental Figure S3



Supplemental Figure S3. *TAN1* can be fully replaced by neo2, whereas *XRN2* cannot be.
(A) Schematic showing *TAN1* locus replacement strategy and enzymes used in Southern blot analysis. The gray line indicates the position of the probe used. Southern blot below.
(B) Same as in A for *XRN2*.

# CHAPTER FOUR

## Biochemical Approaches for *Tetrahymena thermophila*: Nucleic acid and protein detection, subcellular fractionation, extract preparation, and affinity purification

*Tetrahymena thermophila* is an excellent model organism for the study of conserved eukaryotic biological processes, as exemplified by the discoveries of telomeres (Blackburn and Gall, 1978), telomerase (Greider and Blackburn, 1985), self-splicing RNA (Zaug and Cech, 1986), and the function of histone acetylation (Brownell et al., 1996). Recently, our lab described a novel class of small RNAs (sRNAs) (Lee and Collins, 2006). These were later found to be Piwi-bound sRNAs (piRNAs) and represented the first report of sRNAs with characteristics like those of the now well-known animal piRNAs. Subsequently, a flood of studies have uncovered some of the functions of piRNAs in animal germlines. My thesis research has focused on elucidating the roles of *Tetrahymena* Piwi (Twi) proteins and their bound sRNAs. Unique features of *Tetrahymena* have allowed in-depth characterization of conserved RNA silencing machinery, as presented in chapters one, two, and three. Here I present an instructional chapter describing genetic manipulations and biochemical methods useful for analysis of protein or ribonucleoprotein (RNP) complexes in *Tetrahymena*.

### *Introduction*

Epitope tagging is a powerful approach used to investigate the role(s) of a protein of interest by helping to elucidate its localization, protein interactors, nucleic acid targets or partners, and its activity. Successful purification results in a mixture of the protein of interest and any associated proteins and nucleic acids, as well as nonspecific background proteins and nucleic acids, all as a function of the conditions used. This chapter will discuss some of the common biochemical methods for detecting protein, RNA, and DNA in *Tetrahymena thermophila,* as well as guidelines and considerations for designing an affinity purification experiment, beginning with construction of a tagged transgene for incorporation into the *T. thermophila* genome.

### *General considerations*

Although there can be advantages to using specific antibodies for immunoprecipitation, it is much faster, easier, and better experimentally controlled to make a strain expressing a tagged fusion protein when there are not specific antibodies readily available. Tagging allows for more consistency between affinity purification experiments, which is especially convenient when comparing across a family of related proteins (Couvillion et al., 2009). We most commonly use a tandem epitope tag consisting of two Protein A domains (ZZ), a Tobacco Etch Virus (TEV) protease cleavage site, and triple FLAG peptide (N-terminal ZZtev3XF or C-terminal 3XFtevZZ) (Lee et al., 2009a; Min and Collins, 2009), although other tags have also been used successfully (Lee and Collins, 2007; Mochizuki et al., 2002; Yu and Gorovsky, 2000). Since *T. thermophila* has unique codon usage biases, codon-optimized tags are also available, although we have not directly compared codon-optimized and non-codon-optimized versions. Another important consideration, which is only relevant when designing N-terminally tagged fusion proteins, is the start codon context (Salim et al., 2008). A start codon can become highly disfavored if placed right after a GC-rich restriction site for cloning.

After choosing a tag, the next consideration is genomic locus for integration. Here we will discuss only options for somatic macronucleus integration. For more general background about protein tagging in *T. thermophila* see Yu and Gorovsky, 2000. The endogenous locus is a natural choice, with the drug resistance cassette placed upstream of the gene for an N-terminal tag or downstream for a C-terminal tag. The most commonly used resistance cassettes are those from the *neo* series, which confer resistance to paromomycin (Gaertig et al., 1994a; Mochizuki, 2008). We have recently modified the *bsr1* cassette by replacement of the histone H4-I (*HHF1*) promoter with the *MTT1* promoter to create *bsr2,* which is analogous to *neo3* (*MTT1* promoter/resistance gene/*BTU2* 3'UTR) but confers resistance to blasticidin. Expression from the *HHF1* promoter can be too high to drive complete assortment, whereas basal expression from the *MTT1* promoter (without induction with cadmium) is low enough to do so (unpublished data). Enough endogenous genomic DNA must be cloned between the tag and the resistance cassette to include the promoter and 5' UTR or the 3' UTR, respectively. Average UTR lengths are about 150 bp (Coyne et al., 2008), but for a gene that is not well-annotated, 3' RACE or 5' RACE (or reverse transcription (RT) and PCR with primer walking) should be used to identify UTRs, depending on where the tag will be placed. However, complete promoter regions are hard to define by rapid analysis, and there is always a risk that insertion of the drug resistance cassette could disturb the promoter activity of a gene. Another risk is the possibility that the tag will be separated from the resistance cassette and lost due to unwanted recombination. Nonetheless, we have been successful using this strategy when the tag is not deleterious for essential gene function. Chromosomal assortment of polyclonal lines can be easily monitored during selection using whole-cell PCR (when it works): boil fewer than 15 cells in 20 µl sterile filtered water for 5 minutes then add PCR mix and perform PCR as usual. Another method has recently been established for N-terminal tagging, using the Cre/loxP system (Busch et al., 2010).

Other options include using an ectopic promoter and 5' UTR or 3' UTR when integrating the tag at the endogenous locus to avoid the possibility of unwanted recombination, or integrating a complete transgene at an ectopic locus (Lee et al., 2009a; Min and Collins, 2009). For overexpression, we commonly integrate an open reading frame (ORF) under control of the cadmium-inducible *MTT1* promoter (Shang et al., 2002) at the *BTU1* gene of the strain CU522, which allows for selection against the endogenous locus by taxol resistance without requiring a drug resistance cassette in the transgene vector (Gaertig et al., 1994b). For induction, cadmium is added to a final concentration in the range 0.1 to 1.0 µg/mL. Exact cadmium concentration will depend on the desired level of overexpression and the choice media; richer medias require more cadmium. For example, starving cells (cultured in 10 mM Tris) reach maximum expression from *MTT1* at a lower cadmium concentration than growing cells (Shang et al., 2002). Another variable is the iron source in the media. If Sequestrene® is used for the iron source, it will chelate some of the cadmium, requiring higher concentrations. *T. thermophila* has four additional metallothionein genes, two others most highly induced by cadmium and two induced by copper (Boldrin et al., 2006; Diaz et al., 2007). We have successfully used the promoters from *MTT1*, *MTT2*, *MTT3*, and *MTT5* to drive transgene expression.

Protein overexpression can be a useful strategy to make readily detectable amounts of protein for immunofluorescence or small-scale affinity purification. However, we have recently noted that even the relatively minor overexpression caused by basal transcription from the *MTT1* promoter at the *BTU1* locus can result in altered apparent localization of typically nuclear proteins to the cytoplasm (unpublished data). An alternative approach is to use the *MTT1* promoter at its endogenous locus where the basal transcription is much lower (Shang et al., 2002).

*Whole-cell protein and nucleic acid isolation, detection, and quantification*

Once a transgene construct is transformed, a variety of methods can be used to test for incorporation into the genome, assortment, and expression. Whole-cell PCR, as discussed above, is the fastest way to check for incorporation of DNA at the targeted locus. However, it does not work reliably in our hands. Alternatively, genomic DNA can first be isolated (scale as desired, mix gently but *thoroughly* throughout): collect 2.5 x $10^5$ cells and concentrate to 50 µl in 10 mM Tris pH 7.5. Resuspend in residual Tris and add 200 µl 60°C lysis buffer (10 mM Tris pH 7.5, 0.5 M EDTA, 1% SDS, pH adjusted to 9.5 at 60°C). Add 2 volumes water and incubate at 60°C at least 1 hour. Cool to room temperature, add Proteinase K to 50 µg/mL, and incubate at 37°C overnight. Extract with one volume phenol/chloroform/isoamyl alcohol (PCI), and precipitate with one tenth volume sodium acetate, pH 5.2 and one volume isopropanol. Wash pellet in 70% ethanol, and resuspend in 75 µl 1x TE. Add RNase A to 0.8 µg/µl and incubate at 37°C for 30 minutes. PCI extract, precipitate, and wash as before. This method yields ~15 µg total (macronuclear and micronuclear) genomic DNA per 2.5 x $10^5$ cells. Of note, the AT-rich composition of the *T. thermophila* genome generally necessitates using long primers for reliable PCR, where $T_m$ ~55°C using the formula $T_m$ = 64.9°C + [41°C (number of G + C - 16.4) / N], where N is the length of the primer.

For a quantitative assessment of assortment, Southern blotting is used (Malone et al., 2005). If the method for DNA isolation does not specifically purify macronuclei, a wild-type locus-sized restriction fragment may be detectable even in a strain in which the macronuclear gene copies are fully replaced. This derives from the wild-type diploid micronucleus, which is mostly at 4N since it replicates so early relative to the macronucleus and thus can be up to only ~10 times less abundant than the locus restriction fragment from the polyploid macronucleus. If it is ambiguous whether the endogenous locus has been fully replaced, RT-PCR should be used to check for any remaining endogenous transcript expression. We have noticed that some loci do not seem to assort (or back-assort) as quickly as others (unpublished observation). This means even after allowing clonal populations to back-assort for the standard 14 days with rapid doubling, the remaining wild-type locus-sized genomic locus restriction fragment may still be hard to discern, even if assortment is incomplete. Therefore we have made it a common practice to test strains that appear fully assorted by RT-PCR for the mRNA, which is only produced from the macronuclear gene locus.

Transgene product expression levels can be monitored using whole-cell western blotting. Collect 2 - 4 x $10^5$ cells, depending on the expected protein expression level. Rinse cells in 10 mM Tris pH 7.5 and concentrate to 30 µl. Add protease inhibitors (see the following section for details) then SDS-PAGE loading buffer to a final concentration of 1.5x (5x stock is 0.3 M Tris pH 6.8, 10% SDS, 20% β-mercaptoethanol, 50% glycerol). Boil immediately for 5 minutes and freeze or load gel.

Transcript mRNA levels can be monitored using Northern blotting. Typically 20 - 30 μg total RNA (2 - 5 x $10^5$ growing cell equivalents when isolated using TRIzol® reagent) in a 2 mm x 7 mm x 10 mm well volume is sufficient to detect low to moderately abundant transcripts. For very low abundance transcripts, a poly(A)+ enrichment step can be included to enrich for mRNAs. For a detailed description of Northern blotting and all other RNA methods see "RNA: a laboratory manual" (Rio et al., 2011).

A major focus in our lab is the study of ribonucleoprotein (RNP) complexes. Therefore, sensitive detection of small, non-coding RNAs is vital. Small RNAs (sRNAs) are detected using a modified northern blotting protocol, with or without a filtration-based sRNA enrichment step (Lee and Collins, 2006). For abundant sRNAs like small rRNAs and high copy number tRNAs, less than 1 μg total RNA is sufficient for visualization by SYBR Gold and northern blotting. For low abundance small RNAs like Piwi-bound RNAs, up to 5 μg enriched sRNAs may be needed. This is typically obtained from 200 - 500 μg of total RNA. For very low abundance species modifications to the northern blotting procedure can be used to improve the hybridization sensitivity (Pall et al., 2007).

It may be of interest to differentiate nuclear from cytoplasmic complexes using subcellular fractionation before applying the techniques discussed above, and/or prior to affinity purification, discussed below. Robust methods for isolation of macronuclei, micronuclei, developing macronuclei (anlagen), and nucleoli have been developed (Allis and Dennison, 1982; Gocke et al., 1978; Gorovsky et al., 1975), which we have used prior to affinity purification or immunofluorescence.

*Lysate preparation and affinity purification*

1. Grow cells. Grow 10 mL to 1 L cultures of the tagged protein strain and a control wild-type untagged protein strain (mock) to log phase or desired life cycle stage. All efforts should be made to treat cultures similarly between experiments to be compared, including choice of media, cell density, amount and duration of cadmium or other treatment, etc.
2. Collect cells. Spin 1,500 x g for 3 minutes (in a braked rotor), or longer for faster-swimming starving cells, and wash in 10 mM Tris pH 7.5 (if necessary to remove rich media contaminants that can increase proteolysis). Alternatively wash in PBS or Dryl's (Dryl, 1959) for downstream applications not compatible with Tris. Note: cells may undergo osmotic stress in 1X PBS.
3. Optional in vivo crosslinking, for example with formaldehyde or UV (Dedon et al., 1991).
4. Lyse cells. Lysis conditions will vary depending on desired purification stringency, protein complex characteristics, and whether a crosslinking step was included. Always lyse at 4°C to reduce protease and nuclease activities, usually at 3-5 x $10^6$ cell equivalents/mL for 10-15 minutes. Conditions will have to be empirically determined for each new experiment. Guidelines follow.
   - <u>Native (no crosslink)</u>: 20 mM Tris pH 7.5, 0.05 - 1.0 M NaCl, 10% glycerol, 0.1 - 0.2% Igepal®, 0.1 - 0.2% Triton X-100 (optional to more efficiently lyse nuclei). Add protease inhibitors fresh just before use. We use a mammalian protease inhibitor cocktail

(Sigma) as well as a final concentration of 0.1 mM phenylmethylsulfonyl fluoride (PMSF) made as a 0.1 M stock in isopropanol. Note that PMSF is inactivated in aqueous solutions, so add it fresh. Keep the protease inhibitor addition to less than 1/500th total volume to minimize the concentration of organic solvents, unless the purification is done under denaturing conditions.

Optional:

1 mM $MgCl_2$: Will stabilize some protein-RNA interactions, but will also allow $Mg^{2+}$ - dependent enzymes be catalytically active and release their RNA substrates.

1 mM EDTA: A protease and RNase inhibitor and can stall enzymes on RNA substrates by $Mg^{2+}$ chelation.

Reducing agent: 0.5 - 1 mM DTT or 10 - 20 mM β-mercaptoethanol.

- Denaturing (crosslink step included): Radioimmunoprecipitation assay (RIPA) buffer (Harlow and Lane, 1988) is commonly used, with Triton X-100 in place of Igepal® for ChIP applications. Igepal® may inhibit DNA shearing (unpublished observation). Resuspend to 5 x $10^6$ cells/mL.
  Disrupt cells by sonication.

5. Clear lysate. Spin at 16,000 x g for 15 minutes for a crude lysate or 100,000 x g for 1 hour, which will remove ribosomes and other large, macromolecular complexes from the lysate.
6. Optional: flash freeze lysate and store -80°C. This may reduce yield in some cases.
7. Binding. Add prewashed antibody-conjugated beads to cleared lysate. A good starting point is 4 μl 50% bead slurry/mL lysate. Larger bead volumes can increase recovery, but can also disproportionately increase background. The amount used depends on binding capacity of beads and concentration of lysate. Bind at room temperature 1-2 hours or at 4°C 1.5 hours to overnight.
8. Wash beads. Wash 4 - 6 times in at least 20 bead-slurry volumes 5 minutes at room temperature or 4°C. Wash buffer should be similar to binding buffer but is often higher stringency; for example, higher detergent concentration or additional detergents, higher salt concentration, and addition of urea to 2 M (or higher in the case of prior crosslinking).
9. Elute protein/RNP complex. Tag-specific elution results in the lowest background and recovery of native (possibly functional) complexes that can be used for activity assays or a second step of immunoprecipitation. Examples are TEV protease and 3XFLAG peptide. Other options include denaturation by SDS, urea, or low pH (0.1 M glycine pH 2.7).

*Detection of recovered complexes*

When beginning an affinity purification experiment, recovery of the protein and/or RNA of interest should be tracked throughout the procedure using immunoblotting or Coomassie staining and northern blotting or SYBR Gold staining. To direct optimization, complexes should be tracked in whole cells, lysate (supernatant), unbound lysate, eluate, and purification resin (beads) pre- and post-elution. Pellets can also be tracked but are often difficult to resuspend uniformly.

After optimization, generally only a fraction of the eluate is sacrificed for SDS-PAGE and silver staining to detect the tagged protein and specific co-purifying proteins versus background from the mock purification. The remainder is then used for activity assays, identification of proteins by immunoblot or mass spectrometry, or analysis of co-purified nucleic acids. To extract DNA, add equal volume of phenol/chloroform/isoamyl alcohol (PCI). To extract RNA, add equal volume of PCI or 10X volume of TRIzol® reagent. Caution: remember to reverse crosslink (if relevant and possible) or treat with proteinase K before extracting DNA or RNA. DNA and RNA can be differentiated by treatment with DNase or RNase.

Co-purified nucleic acids can be detected by PCR, RT-PCR, or northern blotting if the sequence is known. A valuable application of affinity purification is in the identification of unknown nucleic acid species associated with a protein. For example, after native affinity purification, stably bound sRNAs can be purified, ligated to adaptors, and reverse transcribed to make a cDNA library for high throughput sequencing (sRNA-seq) (Couvillion et al., 2009). Similarly, longer RNAs can be immunoprecipitated and prepared for sequencing (RIP-seq) (Zhao et al., 2010). After crosslinking and denaturing purification, DNA from chromatin immunoprecipitation can be prepared for sequencing (ChIP-seq) (Park, 2009). A variety of methods including high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) (Licatalosi et al., 2008) and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) (Hafner et al., 2010) have been developed to improve recovery of RNA that only transiently interacts with a protein.

# REFERENCES

Akbergenov R., Si-Ammour A., Blevins T., Amin I., Kutter C., Vanderschuren H., Zhang P., Gruissem W., Meins F., Jr., Hohn T., Pooggin M.M. (2006) Molecular characterization of geminivirus-derived small RNAs in different plant species. Nucleic Acids Res 34:462-71.

Allen E., Xie Z., Gustafson A.M., Carrington J.C. (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell 121:207-21.

Allis C.D., Dennison D.K. (1982) Identification and purification of young macronuclear anlagen from conjugating cells of Tetrahymena thermophila. Dev Biol 93:519-33.

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 26:5007-19.

Atkinson G.C., Baldauf S.L., Hauryliuk V. (2008) Evolution of nonstop, no-go and nonsense-mediated mRNA decay and their termination factor-derived components. BMC Evol Biol 8:290.

Azzalin C.M., 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 J.E., Ruby J.G., Wang Y., Bartel D.P., Blelloch R. (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev. 22:2773-85.

Bayne E.H., White S.A., Kagansky A., Bijos D.A., Sanchez-Pulido L., Hoe K.L., Kim D.U., Park H.O., Ponting C.P., Rappsilber J., Allshire R.C. (2010) Stc1: a critical link between RNAi and chromatin modification required for heterochromatin integrity. Cell 140:666-77.

Blackburn E.H., Gall J.G. (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. J Mol Biol 120:33-53.

Boldrin F., Santovito G., Gaertig J., Wloga D., Cassidy-Hanley D., Clark T.G., Piccinni E. (2006) Metallothionein gene from Tetrahymena thermophila with a copper-inducible-repressible promoter. Eukaryot Cell 5:422-5.

Brennecke J., Aravin A.A., Stark A., Dus M., Kellis M., Sachidanandam R., Hannon G.J. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128:1089-103.

Brownell J.E., Zhou J., Ranalli T., Kobayashi R., Edmondson D.G., Roth S.Y., Allis C.D. (1996) Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell 84:843-51.

Buhler M., Moazed D. (2007) Transcription and RNAi in heterochromatic gene silencing. Nat Struct Mol Biol 14:1041-1048.

Busch C.J., Vogt A., Mochizuki K. (2010) Establishment of a Cre/loxP recombination system for N-terminal epitope tagging of genes in Tetrahymena. BMC Microbiol 10:191.

Carthew R.W., Sontheimer E.J. (2009) Origins and Mechanisms of miRNAs and siRNAs. Cell 136:642-55.

Cerutti H., Casas-Mollano J.A. (2006) On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet 50:81-99.

Chalker D.L. (2008) Dynamic nuclear reorganization during genome remodeling of Tetrahymena. Biochim Biophys Acta 1783:2130-6.

Chen X. (2007) A marked end. Nat Struct Mol Biol 14:259-60.

Chernyakov I., Whipple J.M., Kotelawala L., Grayhack E.J., Phizicky E.M. (2008) Degradation of several hypomodified mature tRNA species in Saccharomyces cerevisiae is mediated by Met22 and the 5'-3' exonucleases Rat1 and Xrn1. Genes Dev 22:1369-80.

Cole C., Sobala A., Lu C., Thatcher S.R., Bowman A., Brown J.W., Green P.J., Barton G.J., Hutvagner G. (2009) Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA 15:2147-60.

Couvillion M.T., Sachidanandam R., Collins K. (2010) A growth-essential Tetrahymena Piwi protein carries tRNA fragment cargo. Genes Dev 24:2742-7.

Couvillion M.T., Lee S.R., Hogstad B., Malone C.D., Tonkin L.A., Sachidanandam R., Hannon G.J., Collins K. (2009) Sequence, biogenesis, and function of diverse small RNA classes bound to the Piwi family proteins of *Tetrahymena thermophila*. Genes Dev. 23:2016-32.

Coyne R.S., Thiagarajan M., Jones K.M., Wortman J.R., Tallon L.J., Haas B.J., Cassidy-Hanley D.M., Wiley E.A., Smith J.J., Collins K., Lee S.R., Couvillion M.T., Liu Y., Garg J., Pearlman R.E., Hamilton E.P., Orias E., Eisen J.A., Methe B.A. (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.

Dedon P.C., Soults J.A., Allis C.D., Gorovsky M.A. (1991) A simplified formaldehyde fixation and immunoprecipitation technique for studying protein-DNA interactions. Anal Biochem 197:83-90.

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 274:16590-4.

Diaz S., Amaro F., Rico D., Campos V., Benitez L., Martin-Gonzalez A., Hamilton E.P., Orias E., Gutierrez J.C. (2007) Tetrahymena metallothioneins fall into two discrete subfamilies. PLoS One 2:e291.

Dittmar K.A., Goodenbour J.M., Pan T. (2006) Tissue-specific differences in human transfer RNA expression. PLoS Genet 2:e221.

Drummond D.A., Wilke C.O. (2009) The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet 10:715-24.

Dryl S. (1959) Antigenic Transformation in *Paramecium aurelia* after homologous antiserum treatment during autogamy and conjugation. J Protozool 6.

Duchaine T.F., Wohlschlegel J.A., Kennedy S., Bei Y., Conte D., Jr., Pang K., Brownell D.R., Harding S., Mitani S., Ruvkun G., Yates J.R., 3rd, Mello C.C. (2006) Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways. Cell 124:343-54.

Eisen J.A., Coyne R.S., Wu M., Wu D., Thiagarajan M., Wortman J.R., Badger J.H., Ren Q., Amedeo P., Jones K.M., Tallon L.J., Delcher A.L., Salzberg S.L., Silva J.C., Haas B.J., Majoros W.H., Farzad M., Carlton J.M., Smith R.K., Jr., Garg J., Pearlman R.E., Karrer K.M., Sun L., Manning G., Elde N.C., Turkewitz A.P., Asai D.J., Wilkes D.E., Wang Y., Cai H., Collins K., Stewart B.A., Lee S.R., Wilamowska K., Weinberg Z., Ruzzo W.L., Wloga D., Gaertig J., Frankel J., Tsao C.C., Gorovsky M.A., Keeling P.J., Waller R.F., Patron N.J., Cherry J.M., Stover N.A., Krieger C.J., del Toro C., Ryder H.F., Williamson

S.C., Barbeau R.A., Hamilton E.P., Orias E. (2006) Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol 4:e286.

Farazi T.A., Juranek S.A., Tuschl T. (2008) The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. Development 135:1201-14.

Fu H., Feng J., Liu Q., Sun F., Tie Y., Zhu J., Xing R., Sun Z., Zheng X. (2009) Stress induces tRNA cleavage by angiogenin in mammalian cells. FEBS Lett. 583:437-42.

Gaertig J., Gu L., Hai B., Gorovsky M.A. (1994a) High frequency vector-mediated transformation and gene replacement in Tetrahymena. Nucleic Acids Res 22:5391-8.

Gaertig J., Thatcher T.H., Gu L., Gorovsky M.A. (1994b) Electroporation-mediated replacement of a positively and negatively selectable beta-tubulin gene in Tetrahymena thermophila. Proc Natl Acad Sci U S A 91:4549-53.

Garcia-Silva M.R., Frugier M., Tosar J.P., Correa-Dominguez A., Ronalte-Alves L., Parodi-Talice A., Rovira C., Robello C., Goldenberg S., Cayota A. (2010) A population of tRNA-derived small RNAs is actively produced in *Trypanosoma cruzi* and recruited to specific cytoplasmic granules. Mol. Biochem. Parasitol.

Geerlings T.H., Vos J.C., Raue H.A. (2000) The final step in the formation of 25S rRNA in Saccharomyces cerevisiae is performed by 5'-->3' exonucleases. RNA 6:1698-703.

Ghildiyal M., Zamore P.D. (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10:94-108.

Gocke E., Leer J.C., Nielsen O.F., Westergaard O. (1978) Transcriptional properties of nucleoli isolated from Tetrahymena. Nucleic Acids Res 5:3993-4006.

Gorovsky M.A., Yao M.C., Keevert J.B., Pleger G.L. (1975) Isolation of micro- and macronuclei of Tetrahymena pyriformis. Methods Cell Biol 9:311-27.

Greider C.W., Blackburn E.H. (1985) Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. Cell 43:405-13.

Hafner M., Landthaler M., Burger L., Khorshid M., Hausser J., Berninger P., Rothballer A., Ascano M., Jr., Jungkamp A.C., Munschauer M., Ulrich A., Wardle G.S., Dewell S., Zavolan M., Tuschl T. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141:129-41.

Harlow E., Lane D.A. (1988) Antibodies: A Laboratory Manual Cold Spring Harbor Laboratory, New York.

Haussecker D., Huang Y., Lau A., Parameswaran P., Fire A.Z., Kay M.A. (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA 16.

Henry Y., Wood H., Morrissey J.P., Petfalski E., Kearsey S., Tollervey D. (1994) The 5' end of yeast 5.8S rRNA is generated by exonucleases from an upstream cleavage site. EMBO J 13:2452-63.

Hock J., Meister G. (2008) The Argonaute protein family. Genome Biol 9:210.

Howard-Till R.A., Yao M.C. (2006) Induction of gene silencing by hairpin RNA expression in Tetrahymena thermophila reveals a second small RNA pathway. Mol Cell Biol 26:8731-42.

Jinek M., Doudna J.A. (2009) A three-dimensional view of the molecular machinery of RNA interference. Nature 457:405-12.

Katiyar-Agarwal S., Gao S., Vivian-Smith A., Jin H. (2007) A novel class of bacteria-induced small RNAs in Arabidopsis. Genes Dev 21:3123-34.

Katoh T., Sakaguchi Y., Miyauchi K., Suzuki T., Kashiwabara S., Baba T. (2009) Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. Genes Dev 23:433-8.

Kawaji H., Nakamura M., Takahashi Y., Sandelin A., Katayama S., Fukuda S., Daub C.O., Kai C., Kawai J., Yasuda J., Carninci P., Hayashizaki Y. (2008) Hidden layers of human small RNAs. BMC Genomics 9:157.

Kawamata T., Seitz H., Tomari Y. (2009) Structural determinants of miRNAs for RISC loading and slicer-independent unwinding. Nat. Struct. Mol. Biol. 16:953-60.

Kim V.N., Han J., Siomi M.C. (2009) Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol 10:126-39.

Klattenhoff C., Theurkauf W. (2008) Biogenesis and germline functions of piRNAs. Development 135:3-9

Kramer E.B., Farabaugh P.J. (2007) The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA 13:87-96.

Kurth H.M., Mochizuki K. (2009) 2'-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in Tetrahymena. RNA 15:675-85.

Lee S.R., Collins K. (2005) Starvation-induced cleavage of the tRNA anticodon loop in *Tetrahymena thermophila*. J. Biol. Chem. 280:42744-9.

Lee S.R., Collins K. (2006) Two classes of endogenous small RNAs in Tetrahymena thermophila. Genes Dev 20:28-33.

Lee S.R., Collins K. (2007) Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs. Nat Struct Mol Biol 14:604-10.

Lee S.R., Talsky K.B., Collins K. (2009a) A single RNA-dependent RNA polymerase assembles with mutually exclusive nucleotidyl transferase subunits to direct different pathways of small RNA biogenesis. RNA 15:1363-74.

Lee Y.S., Shibata Y., Malhotra A., Dutta A. (2009b) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. 23:2639-49.

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 15:1501-7.

Licatalosi D.D., Mele A., Fak J.J., Ule J., Kayikci M., Chi S.W., Clark T.A., Schweitzer A.C., Blume J.E., Wang X., Darnell J.C., Darnell R.B. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456:464-9

Liu Y., Mochizuki K., Gorovsky M.A. (2004) Histone H3 lysine 9 methylation is required for DNA elimination in developing macronuclei in Tetrahymena. Proc Natl Acad Sci U S A 101:1679-84.

Luke B., Panza A., Redon S., Iglesias N., Li Z., Lingner J. (2008) The Rat1p 5' to 3' exonuclease degrades telomeric repeat-containing RNA and promotes telomere elongation in Saccharomyces cerevisiae. Mol Cell 32:465-77.

Luo W., Johnson A.W., Bentley D.L. (2006) The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. Genes Dev 20:954-65.

Ma J.B., Yuan Y.R., Meister G., Pei Y., Tuschl T., Patel D.J. (2005) Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature 434:666-70.

Malone C.D., Hannon G.J. (2009) Small RNAs as guardians of the genome. Cell 136:656-68.

Malone C.D., Anderson A.M., Motl J.A., Rexer C.H., Chalker D.L. (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. 25:9151-64.

Mercer T.R., Dinger M.E., Mattick J.S. (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155-9.

Miao W., Xiong J., Bowen J., Wang W., Liu Y., Braguinets O., Grigull J., Pearlman R.E., Orias E., Gorovsky M.A. (2009) Microarray analyses of gene expression during the Tetrahymena thermophila life cycle. PLoS One 4:e4429.

Min B., Collins K. (2009) An RPA-related sequence-specific DNA-binding subunit of telomerase holoenzyme is required for elongation processivity and telomere maintenance. Mol Cell 36:609-19.

Mochizuki K. (2008) High efficiency transformation of Tetrahymena using a codon-optimized neomycin resistance gene. Gene 425:79-83.

Mochizuki K., Gorovsky M.A. (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 N.A., Fujisawa T., Gorovsky M.A. (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell 110:689-99

Molnar A., Schwach F., Studholme D.J., Thuenemann E.C., Baulcombe D.C. (2007) miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. Nature 447:1126-9.

Nilsen T.W. (2008) Endo-siRNAs: yet another layer of complexity in RNA silencing. Nat Struct Mol Biol 15:546-8.

Noto T., Kurth H.M., Kataoka K., Aronica L., DeSouza L.V., Siu K.W., Pearlman R.E., Gorovsky M.A., Mochizuki K. (2010) The Tetrahymena argonaute-binding protein Giw1p directs a mature argonaute-siRNA complex to the nucleus. Cell 140:692-703.

Pak J., Fire A. (2007) Distinct populations of primary and secondary effectors during RNAi in C. elegans. Science 315:241-4.

Pall G.S., Codony-Servat C., Byrne J., Ritchie L., Hamilton A. (2007) Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. Nucleic Acids Res 35:e60.

Park P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669-80.

Phizicky E.M., Alfonzo J.D. (2010) Do all modifications benefit all tRNAs? FEBS Lett 584:265-71.

Purves W.K., al. e. (1988) Life: The Science of Biology. 4th Edition ed. WH Freeman

Rio D.C., Hannon G.J., Ares M., Jr., Nilsen T.W. (2011) RNA: A Laboratory Manual Cold Spring Harbor Press, New York.

Rudel S., Wang Y., Lenobel R., Korner R., Hsiao H.H., Urlaub H., Patel D., Meister G. (2011) Phosphorylation of human Argonaute proteins affects small RNA binding. Nucleic Acids Res 39:2330-2343.

Saikia M., Fu Y., Pavon-Eternod M., He C., Pan T. (2010) Genome-wide analysis of N1-methyl-adenosine modification in human tRNAs. RNA 16:1317-27.

Salim H.M., Ring K.L., Cavalcanti A.R. (2008) Patterns of codon usage in two ciliates that reassign the genetic code: Tetrahymena thermophila and Paramecium tetraurelia. Protist 159:283-98.

Schoeftner S., Blasco M.A. (2008) Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat Cell Biol 10:228-36.

Seto A.G., Kingston R.E., Lau N.C. (2007) The coming of age for Piwi proteins. Mol Cell 26:603-9.

Shang Y., Song X., Bowen J., Corstanje R., Gao Y., Gaertig J., Gorovsky M.A. (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. USA 99:3734-9.

Sijen T., Steiner F.A., Thijssen K.L., Plasterk R.H. (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science 315:244-7.

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

Siomi M.C., Mannen T., Siomi H. (2010) How does the royal family of Tudor rule the PIWI-interacting RNA pathway? Genes Dev 24:636-46.

Taverna S.D., Coyne R.S., Allis C.D. (2002) Methylation of histone h3 at lysine 9 targets programmed DNA elimination in tetrahymena. Cell 110:701-11.

Thompson D.M., Parker R. (2009a) The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in *Saccharomyces cerevisiae*. J. Cell Biol. 185:43-50.

Thompson D.M., Parker R. (2009b) Stressing out over tRNA cleavage. Cell 138:215-9.

Tolia N.H., Joshua-Tor L. (2007) Slicer and the argonautes. Nat. Chem. Biol. 3:36-43.

Ullu E., Lujan H.D., Tschudi C. (2005) Small sense and antisense RNAs derived from a telomeric retroposon family in Giardia intestinalis. Eukaryot Cell 4:1155-7.

Vazquez F., Vaucheret H., Rajagopalan R., Lepers C., Gasciolli V., Mallory A.C., Hilbert J.L., Bartel D.P., Crete P. (2004) Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. Mol Cell 16:69-79.

Wang M., Pestov D.G. (2011) 5'-end surveillance by Xrn2 acts as a shared mechanism for mammalian pre-rRNA maturation and decay. Nucleic Acids Res 39:1811-22.

West S., Gromak N., Proudfoot N.J. (2004) Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature 432:522-5.

Witkin K.L., Collins K. (2004) Holoenzyme proteins required for the physiological assembly and activity of telomerase. Genes Dev 18:1107-18.

Witkin K.L., Prathapam R., Collins K. (2007) Positive and negative regulation of Tetrahymena telomerase holoenzyme. Mol Cell Biol 27:2074-83

Xiang S., Cooper-Morgan A., Jiao X., Kiledjian M., Manley J.L., Tong L. (2009) Structure and function of the 5'-->3' exoribonuclease Rat1 and its activating partner Rai1. Nature 458:784-8.

Yamasaki S., Ivanov P., Hu G.F., Anderson P. (2009) Angiogenin cleaves tRNA and promotes stress-induced translational repression. J. Cell Biol. 185:35-42.

Yu L., Gorovsky M.A. (2000) Protein tagging in Tetrahymena. Methods Cell Biol 62:549-59.

Zaug A.J., Cech T.R. (1986) The Tetrahymena intervening sequence ribonucleic acid enzyme is a phosphotransferase and an acid phosphatase. Biochemistry 25:4478-82.

Zekri L., Huntzinger E., Heimstadt S., Izaurralde E. (2009) The silencing domain of GW182 interacts with PABPC1 to promote translational repression and degradation of microRNA targets and is required for target release. Mol Cell Biol 29:6220-31.

Zhang H., Ehrenkaufer G.M., Pompey J.M., Hackney J.A., 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.

Zhao J., Ohsumi T.K., Kung J.T., Ogawa Y., Grau D.J., Sarma K., Song J.J., Kingston R.E., Borowsky M., Lee J.T. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell 40:939-53.

Zhao T., Li G., Mi S., Li S., Hannon G.J., Wang X.J., Qi Y. (2007) A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. Genes Dev 21:1190-203.