



Small RNAs Are Implicated in Regulation of Gene and Transposable Element Expression in the Protist *Trichomonas vaginalis*

Sally D. Warring,^{a*} Frances Blow,^a Grace Avecilla,^a Jordan C. Orosco,^a Steven A. Sullivan,^a  Jane M. Carlton^a

^aCenter for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, USA

ABSTRACT *Trichomonas vaginalis* is the causative agent of trichomoniasis, the most prevalent nonviral sexually transmitted infection worldwide. Repetitive elements, including transposable elements (TEs) and virally derived repeats, comprise more than half of the ~160-Mb *T. vaginalis* genome. An intriguing question is how the parasite controls its potentially lethal complement of mobile elements, which can disrupt transcription of protein-coding genes and genome functions. In this study, we generated high-throughput RNA sequencing (RNA-Seq) and small RNA-Seq data sets in triplicate for the *T. vaginalis* G3 reference strain and characterized the mRNA and small RNA populations and their mapping patterns along all six chromosomes. Mapping the RNA-Seq transcripts to the genome revealed that the majority of genes predicted within repetitive elements are not expressed. Interestingly, we identified a novel species of small RNA that maps bidirectionally along the chromosomes and is correlated with reduced protein-coding gene expression and reduced RNA-Seq coverage in repetitive elements. This novel small RNA family may play a regulatory role in gene and repetitive element expression. Our results identify a possible small RNA pathway mechanism by which the parasite regulates expression of genes and TEs and raise intriguing questions as to the role repeats may play in shaping *T. vaginalis* genome evolution and the diversity of small RNA pathways in general.

IMPORTANCE Trichomoniasis, caused by the protozoan *Trichomonas vaginalis*, is the most common nonviral sexually transmitted infection in humans. The millions of cases each year have sequelae that may include complications during pregnancy and increased risk of HIV infection. Given its evident success in this niche, it is paradoxical that *T. vaginalis* harbors in its genome thousands of transposable elements that have the potential to be extremely detrimental to normal genomic function. In many organisms, transposon expression is regulated by the activity of endogenously expressed short (~21 to 35 nucleotides [nt]) small RNA molecules that effect gene silencing by targeting mRNAs for degradation or by recruiting epigenetic silencing machinery to locations in the genome. Our research has identified small RNA molecules correlated with reduced expression of *T. vaginalis* genes and transposons. This suggests that a small RNA pathway is a major contributor to gene expression patterns in the parasite and opens up new avenues for investigation into small RNA biogenesis, function, and diversity.

KEYWORDS *Trichomonas vaginalis*, small RNA, transposable element

The parabasalid protist *Trichomonas vaginalis* colonizes the human urogenital tract and has an estimated global incidence of ~270 million new cases per year, making it the most common nonviral sexually transmitted infection (STI) (1). While *T. vaginalis* infections are often asymptomatic (and typically so in men), they can also cause vaginitis, urethritis, and pelvic inflammatory disease (2) and, importantly, can increase the


Citation Warring SD, Blow F, Avecilla G, Orosco JC, Sullivan SA, Carlton JM. 2021. Small RNAs are implicated in regulation of gene and transposable element expression in the protist *Trichomonas vaginalis*. *mSphere* 6:e01061-20. <https://doi.org/10.1128/mSphere.01061-20>.

Editor Katherine S. Ralston, University of California, Davis

Copyright © 2021 Warring et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Jane M. Carlton, jane.carlton@nyu.edu.

* Present address: Sally D. Warring, Earlham Institute, Norwich Research Park, Norwich, United Kingdom.

 Trichy sexually transmitted parasite *Trichomonas vaginalis* is choc-a-bloc with transposable elements. Ever wondered how it silences them? New paper from @SallyWarring @genome_jane @francesblow1 suggests through a small RNA pathway. Read it here!

Received 20 October 2020

Accepted 8 December 2020

Published 6 January 2021

risk of HIV-1 infection up to 2-fold (3–5). In expectant mothers, infections can result in premature rupture of membranes, low-birth-weight babies, and preterm deliveries (6). Metronidazole and tinidazole are the two U.S. Food and Drug Administration-approved drugs used to treat *T. vaginalis* infection. However, clinical failure of metronidazole currently ranges from ~4% in the United States to 17% in Papua New Guinea (7–9).

Despite the prevalence of the disease and its association with poor pregnancy outcomes and increased HIV-1 risk, there are no established *T. vaginalis* screening, surveillance, or control programs for women or men in the United States, resulting in the disease being considered a “neglected” STI (10). In addition, there are many gaps in our understanding of *T. vaginalis* basic biology, pathogenesis, and molecular mechanisms underlying key clinical phenotypes. While some advances have been made recently (reviewed in references 11 and 12), the complex genome of *T. vaginalis* makes molecular genetic studies challenging. The genome is unusually large for a parasitic protist (13) and contains an extraordinary complement of expanded gene families (14) and repetitive elements, including multiple families of transposable elements (TEs) (15) and virally derived DNA; almost two-thirds of the genome is composed of such sequences (15).

The large burden of TEs and other repetitive elements in *T. vaginalis* has potentially extraordinary consequences for the functioning of the genome. TEs are typically composed of noncoding regions, such as terminal inverted repeats (TIRs), and genes that encode the protein machinery required for their own transposition, such as transposase and integrase genes. *T. vaginalis* contains class 2 DNA transposons, which rely on cut-and-paste mechanisms to replicate and transpose (15, 16), in contrast to class 1 RNA transposons, which move via transcription and RNA intermediates. Examples of TE families identified in *T. vaginalis* include the *Tvmar1* Mariner family. *Tvmar1* TEs have a consensus length of 1,304 bp including a single gene encoding a transposase protein (17). The *Tvmar1* family is present in ~600 copies accounting for ~1 Mb of the genome. The Maverick family, which contains ~5,000 copies each up to 30 kb in length and including as many as 20 genes (15, 17, 18), is the largest family, comprising an astounding ~73 Mb of the genome. Other smaller TE families in the genome include Mutator and Kolobok (19, 20). The massive expansion of in particular Maverick TEs in *T. vaginalis* appears to be recent (15), and our previous work has shown evidence for transposition events, including TE insertion polymorphisms between strains (17, 21).

The abundance of TEs in *T. vaginalis* is extremely unusual among parasitic protists, which tend to have compact genomes (22, 23). Moreover, TE transposition can interrupt genes and regulatory sequences, cause genome rearrangements and duplication, and silence the activity of nearby genes (21, 24, 25). They can also provide novel regulatory sequences for host genes and be a significant source of transcription-regulating signals (26, 27). TEs are usually rare in haploid, asexual organisms such as *T. vaginalis*, since such organisms lack the capacity for genetic exchange and purging of deleterious TE insertions from their genomes (28). Alternately, there may be mechanisms that keep transposition of TEs under control.

In many organisms, the expression of TEs is regulated by the activity of several classes of endogenously expressed small RNAs (sRNAs)—short (~20 to 35 nucleotides [nt]) RNA molecules that effect gene silencing either by targeting mRNAs for degradation or by recruiting epigenetic silencing machinery to specific locations in the genome (29–31). In animals, TEs are targeted by a class of 21- to 35-nt small RNAs called PIWI-interacting RNAs (piRNAs), which are produced from defined genomic loci called piRNA clusters (32–37). In plants and the yeast *Schizosaccharomyces pombe*, TEs are targeted for silencing by ~20- to 24-nt short interfering RNAs (siRNAs) that are produced from long double-stranded RNAs (dsRNAs) (38, 39). While the latter process requires the activity of the RNase III enzyme Dicer, production of piRNAs does not (32, 40, 41). All small RNAs form complexes with Argonaute family proteins as part of effector ribonucleoprotein complexes that carry out silencing (41–43). In protists, Argonaute proteins and Dicer enzymes are involved in the production and activity of small RNAs

which have roles including gene and TE control in *Trypanosoma brucei* (44, 45), retro-transposon and protein-coding gene control in *Entamoeba histolytica* (46), and precise indication of TE and gene excisions in ciliates (47–49).

The enormous burden of TEs and repetitive sequences, and our previous identification of a putative RNase III enzyme and two putative Argonaute proteins (AGO1 and AGO2) encoded in the *T. vaginalis* genome (15), led us to investigate how the expression of TEs and *T. vaginalis* protein-coding genes might be regulated. While two previous studies have investigated the small RNA complement of *T. vaginalis*, both focused on identifying microRNAs (miRNAs; small [~22 nt] noncoding RNA molecules found in plants, animals, and viruses that derive from short hairpins in RNA transcripts) mapping to endogenous *T. vaginalis* protein-coding genes (50, 51). Here, we describe identification of a novel species of small (~34 nt) RNA that is correlated with reduced expression of *T. vaginalis* genes and transposons. We undertook a phylogenetic analysis of the *T. vaginalis* AGO1 and AGO2 proteins, identifying them as most similar to PIWI-like AGO proteins in other organisms, which regulate TEs via piRNA interference (piRNAi). We also identified putative piRNA clusters (regions that generate the sRNAs for sRNA-guided gene silencing by Argonaute proteins) in the *T. vaginalis* genome, indicating that the 34-nt sRNAs are likely piRNAi guides. Combined, these data suggest that a small RNA pathway is a major contributor to gene expression patterns in this sexually transmitted parasite, and they open up new avenues for investigation into small RNA biogenesis, function, and diversity.

RESULTS

***T. vaginalis* AGO proteins cluster in the PIWI-like clade.** We undertook a phylogenetic analysis of the two *T. vaginalis* AGO1 and AGO2 proteins with Argonaute proteins from a range of phylogenetically diverse organisms (Fig. 1; see also Table S1 in the supplemental material). This recovered four previously identified eukaryotic AGO protein clades: *Trypanosoma* AGO-like and PIWI-like (AGO-TRYP and PIWI-TRYP, respectively), *Caenorhabditis elegans* WAGO, AGO-like, and PIWI-like. AGO orthologs from *T. vaginalis* form a monophyletic cluster within the PIWI-like clade, which includes AGO proteins from the ciliates *Tetrahymena thermophila*, *Paramecium tetraurelia*, and *Oxytricha trifallax*, which function in the germ line micronucleus to excise TE sequences by RNAi-mediated programmed DNA elimination (52). The PIWI-like clade also includes AGO proteins from the metazoans *Bombyx mori*, *Drosophila melanogaster*, *Homo sapiens*, and *Mus musculus*, all of which function in piRNAi-mediated suppression of TE activity in the germ line (53). These results indicate that *T. vaginalis* AGO proteins may also function in small RNA or piRNA-guided TE regulation by RNAi.

We also compared annotated functional domains of AGO proteins from the PIWI-like clade that are known to function in TE regulation (Fig. S1). *T. vaginalis* AGO1 Pfam functional domain structure comprised ArgoN, PAZ, and Piwi domains and most closely resembled that of *B. mori* SIWI and *D. melanogaster* Aubergine, rather than AGO proteins from the more closely related protists or ciliates. Despite forming a monophyletic branch with AGO1, *T. vaginalis* AGO2 had a unique Pfam domain architecture compared to AGO1, lacking the ArgoN domain and with a significantly truncated Piwi domain. *B. mori* SIWI and *D. melanogaster* Aubergine both function in piRNAi-mediated TE suppression in the germ line using ~26- to 30-nt piRNA guides (54, 55). Combined, these results suggested that *T. vaginalis* may employ an ancestral piRNAi mechanism mediated by AGO1 to regulate TEs and led us to investigate this further.

Repeats are underrepresented in RNA-Seq but not in sRNA-Seq data. Repeats, including many transposable elements (TEs), account for 62.8% of *T. vaginalis* genomic sequence, while protein-coding genes account for 24.6% (Fig. 2A). We found that the majority of high-throughput RNA sequencing (RNA-Seq) reads map to protein-coding genes (69.6%) and intergenic regions (18.3%), while only 12.2% of the RNA-Seq data map to repeats. In contrast, small RNA-Seq (sRNA-Seq) reads map to repeats and protein-coding genes in roughly the same proportion that they are present in the genome (Fig. 2A). Using a statistical method to determine how “transcribed” or “covered”

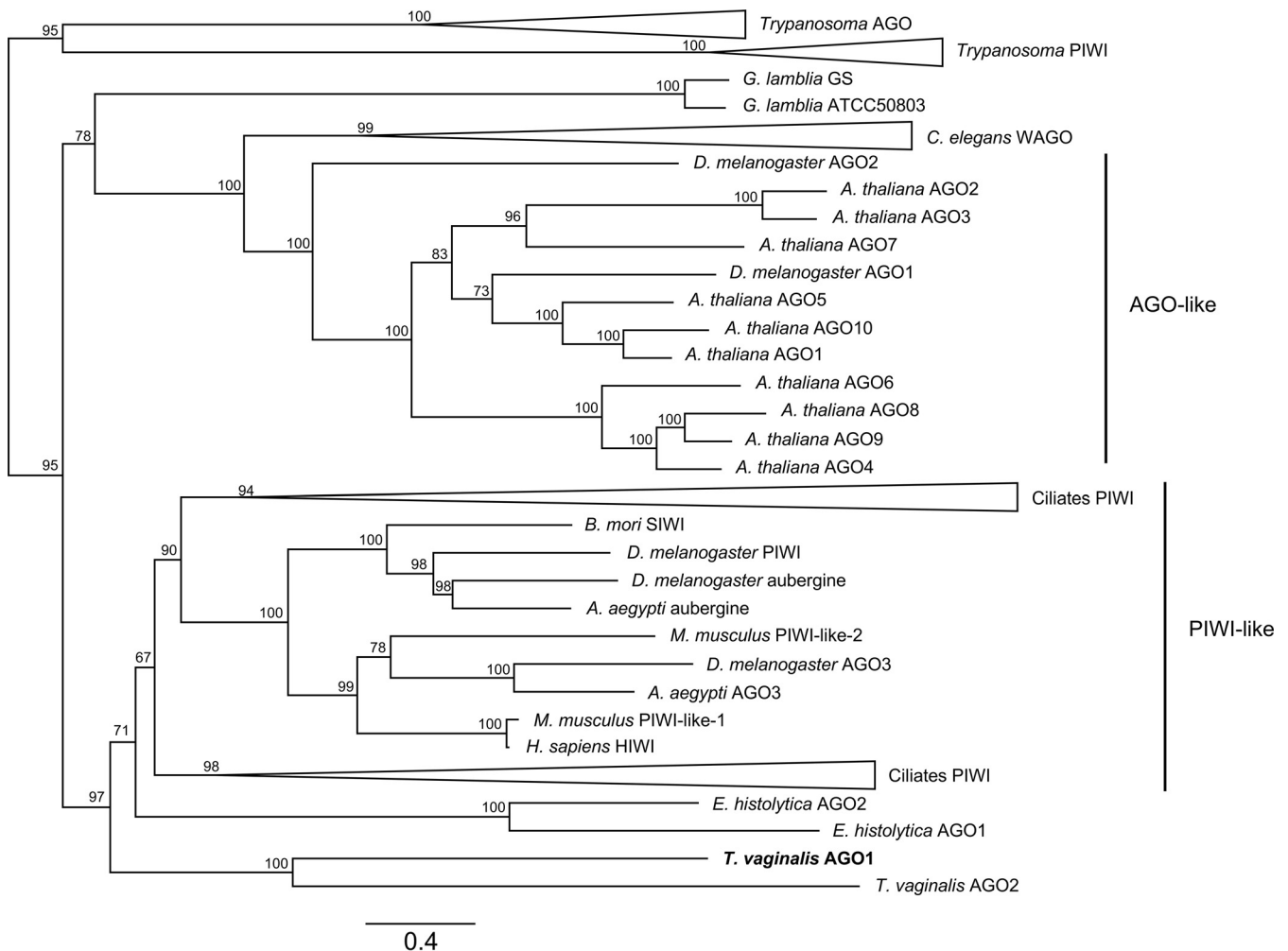


FIG 1 IQ-TREE maximum likelihood phylogeny estimated from full-length amino acid sequences for Argonaute (AGO) proteins. Numbers on nodes correspond to bootstrap support, and the scale bar indicates amino acid substitutions per site. AGO proteins from the following species were included in the phylogenetic analysis (details in Table S1): *Aedes aegypti*, *Arabidopsis thaliana*, *Bombyx mori*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Entamoeba histolytica*, *Giardia lamblia*, *Homo sapiens*, *Leishmania braziliensis*, *Leishmania infantum*, *Leishmania major*, *Mus musculus*, *Oxytricha trifallax*, *Paramecium tetraurelia*, *Tetrahymena thermophila*, *Trichomonas vaginalis*, *Trypanosoma brucei brucei*, *Trypanosoma congolense*, and *Trypanosoma vivax*. Previously defined AGO clades *Trypanosoma* AGO, *Trypanosoma* PIWI, *C. elegans* WAGO, AGO-like, and PIWI-like are indicated. Clades labeled “Ciliates PIWI” are collapsed nodes with PIWI-like Argonaute proteins from the ciliates *O. trifallax*, *P. tetraurelia*, and *T. thermophila*. *T. vaginalis* AGO1 is highlighted in bold.

different regions of the genome are (see Materials and Methods), we found that repeats are depleted in RNA-Seq coverage compared to protein-coding genes, but not in sRNA-Seq coverage (Fig. 2B). In addition, a higher proportion of protein-coding genes are covered by RNA-Seq reads than sRNA-Seq reads, while for repeats the opposite is true: a lower proportion of repeats in the genome were covered by RNA-Seq reads than sRNA-Seq reads (Fig. 2B). We also asked whether genes and repeats differ in magnitude of RNA-Seq and sRNA-Seq coverage by plotting the fragments per kilobase per million (FPKM) and reads per kilobase per million (RPKM) for each gene/repeat individually. We found that protein-coding genes have a higher average RNA-Seq FPKM and a lower average sRNA-Seq RPKM than repeats (Fig. 2C).

The majority of annotated TE families are represented in RNA-Seq and sRNA-Seq data. We next questioned whether members of each of the described TE families was covered by RNA-Seq and sRNA-Seq reads. We found that while the majority (92.8%) of RNA-Seq reads that map to repeats appear to align to “unknown repeats,” the sRNA-Seq reads map to each of the different repeat and TE families in proportions that more closely reflect their presence in the genome, with the Maverick TE family

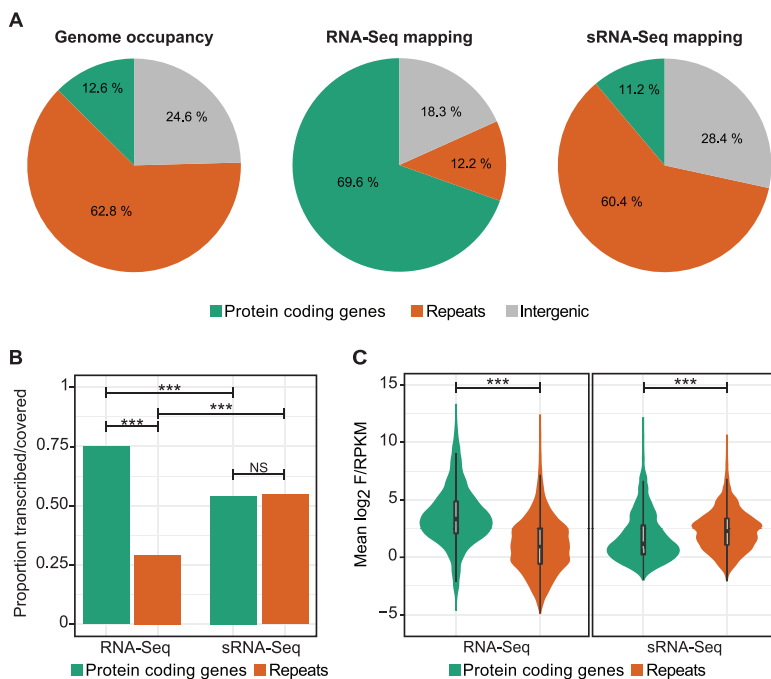


FIG 2 Repeats are depleted in RNA-Seq reads and have higher sRNA-Seq RPKM values. (A) Genome occupancy of genes and repeats and proportions of RNA-Seq and sRNA-Seq libraries aligning to these genomic features. Results are averaged across replicates. (B) Proportion of genes versus repeats that are above the F/RPKM threshold for RNA-Seq and sRNA-Seq reads and are considered transcribed or covered by those data sets. ***, P value < 0.0005 (Fisher's exact test). NS, nonsignificant. (C) \log_2 F/RPKM for genes and repeats averaged across replicates. ***, P value < 0.0005 (two-sided t test).

accounting for 61.6% of all repeats in the genome and 48.8% of the small RNA-Seq reads (Fig. 3A and Table 1). All but 1 (MuDR 7) of the 16 described *T. vaginalis* TE groups have at least one element that was covered by RNA-Seq data, while all of the TE groups have at least one member that was covered by sRNA-Seq reads, and in 11 of the 16 families, 100% of elements are covered by sRNA-Seq reads (Fig. 3B and Table 1). Again, we found that the sRNA-Seq RPKM was greater than the RNA-Seq FPKM for elements in all but one family, Harbinger 1N1 (Fig. 3B and Table 1).

Many TEs contain open reading frames (ORFs) coding for transposases or integrases. For example, using Northern blot methods, we previously identified active transcription of the single ORF that encodes a transposase in Mariner (17). We interrogated expression of the Mariner and Maverick TE ORFs at the mRNA level by determining how many have RNA-Seq FPKM higher than the threshold for coverage described above (and described in Materials and Methods), classifying all those that are above this threshold as putatively expressed at the mRNA level. By this metric, of 38,656 TE ORFs annotated, we found that only 25 were above the threshold and putatively expressed (Table 2).

To account for the possibility that RNA-Seq (and sRNA-Seq) reads were being distributed across the ~600 annotated Mariner elements and thus leading to misleadingly low expression level results, we summed both sRNA-Seq and RNA-Seq reads across every *TvMar1* element ± 1 kb upstream and downstream. Figure 4 shows these mapping results, split by strand relative to the orientation of each *Tvmar1* transposase ORF. This revealed that sRNA-Seq reads map unevenly along both strands of the *Tvmar1* element, while the RNA-Seq signal comes from small mapping regions that overlap at the 5' and 3' ends of the TE coordinates. It appears that the Mariner elements are not covered by RNA-Seq reads and that sRNA-Seq reads notably map where the RNA-Seq reads are not (Fig. 4).

TABLE 1 Numbers of protein-coding genes and repeats expressed in RNA-Seq data and covered by sRNA-Seq data

| Feature type | No. annotated in the genome | No. expressed in RNA-Seq data | No. covered by sRNA-Seq data |
|----------------------|-----------------------------|-------------------------------|------------------------------|
| Protein-coding genes | 19,917 | 15,036 | 10,925 |
| Repeats (total) | 50,382 | 13,878 | 27,549 |
| Harbinger.1 | 7 | 6 | 7 |
| Harbinger.1N1 | 44 | 44 | 33 |
| Harbinger.2 | 2 | 2 | 2 |
| hAT.1 | 27 | 20 | 27 |
| hAT.2 | 12 | 12 | 12 |
| hAT.3N1 | 142 | 99 | 141 |
| Kolobok.3 | 56 | 8 | 56 |
| Kolobok.4 | 18 | 18 | 18 |
| Kolobok.5 | 9 | 9 | 9 |
| Maverick | 4,808 | 507 | 4,731 |
| MuDr.1 | 28 | 7 | 28 |
| MuDR.5 | 69 | 3 | 69 |
| MuDR.7 | 27 | 0 | 27 |
| MuDRx.1 | 2 | 2 | 2 |
| P.1N1 | 164 | 82 | 93 |
| Repeat unknown | 43,572 | 13,022 | 21,602 |
| <i>Tvmar1</i> | 600 | 25 | 600 |
| TE family unknown | 795 | 12 | 92 |

remained intact (Fig. 5C). Relative migration analysis of the band (R_f) showed the midpoint at 34 nt, exactly matching the size distribution from our sequencing data. A Northern blot using a DNA probe identical to the 1.3-kb consensus sequence of the *Tvmar1* TE family (MAR1) showed a band between the 30- and 40-nt size markers congruent with the band observed on our RNA gels and in our sequencing libraries (Fig. 5D).

We analyzed the nucleotide diversity of the 5' base in our small RNAs, categorizing them by length. We found that the small RNAs between the lengths of 25 and 37 nt have a slight 5' U bias, which is strongest in the 33- and 34-nt sequences, where ~50% begin with a U (Fig. 5E). Plotted along the length of the 34-nt sequences, the bias is observed in the 5' base only, with the additional bases reflecting the base composition of the *T. vaginalis* genome more closely (~67% AT [Fig. 5F]).

Proximity to repeats may be correlated with *T. vaginalis* protein-coding gene expression and sRNA-Seq coverage. Our previous studies using quantitative real-time PCR showed that the presence of a *Tvmar1* element close to a protein-coding gene is associated with a decrease or lack of expression (21). To explore this further, we plotted the RNA-Seq FPKM and the sRNA-Seq RPKM for every gene versus its distance from the nearest repeat. We found a small but significant trend showing that RNA-Seq FPKM in genes is positively correlated with increasing distance from the nearest repeat and that the sRNA-Seq RPKM is negatively correlated with increasing distance from the nearest repeat (Fig. 6A).

Given this result, and the finding that very few Maverick and *Tvmar1* family genes are expressed, and the observed sRNA-Seq and RNA-Seq mapping pattern for *Tvmar1* (i.e., sRNA-Seq reads mapping to the genome where RNA-Seq reads do not), we asked whether these small RNAs were associated with silencing. To determine this, we categorized genes as "transcribed" or "silent" according to whether they had an RNA-Seq FPKM above or below the described threshold (see Materials and Methods and above) and repeats as "with coverage" or "without coverage" according to whether they had an RNA-Seq FPKM above or below the described threshold. We found that sRNA-Seq RPKM is higher in protein-coding genes and repeats that are silent/without RNA-Seq coverage (Fig. 6B). However, when we compared the proportions of transcribed/with RNA-Seq protein-coding genes and repeats with sRNA-Seq reads mapped, we found

TABLE 2 Number of *Tvmar1* and Maverick TE family genes analyzed, expressed in RNA-Seq data and covered by sRNA-Seq data

| TE family | Gene annotation | No. analyzed | No. expressed in RNA-Seq | No. covered in sRNA-Seq |
|-----------|---|--------------|--------------------------|-------------------------|
| Maverick | c-integrase | 3,551 | 3 | 3,188 |
| | conserved_hypothetical_protein | 8,742 | 3 | 5,308 |
| | DNA_polymerase_type_B_organellar_and_viral | 1,491 | 6 | 1,069 |
| | hypothetical_protein | 6,950 | 3 | 2,802 |
| | Kil-A_N-terminal_domain_protein_1 | 3,183 | 1 | 2,546 |
| | Mav1.6_DNA_primase_domain_protein | 2,716 | 3 | 2,113 |
| | protein_similar_to_RAD50_ATPase | 2,485 | 0 | 2,267 |
| | protein_similar_to_viral_structural_protein_S1_structurally_related_to_reovirus_core_protein_PRD1_capsid_protein_and_phage_tail_fiber_protein | 3,907 | 1 | 3,104 |
| | protein_similar_to_viral_structural_protein_S2_and_structurally_related_to_PRD1_capsid_protein | 2,934 | 0 | 2,623 |
| | protein_similar_to_viral_structural_protein_S3_and_structurally_similar_to_PRD1_capsid_protein | 1,047 | 0 | 802 |
| | protein_structurally_similar_to_glycosyltransferase | 1,051 | 1 | 740 |
| Mariner | mariner_transposase | 599 | 4 | 598 |
| Total | | 38,656 | 25 | 27,160 |

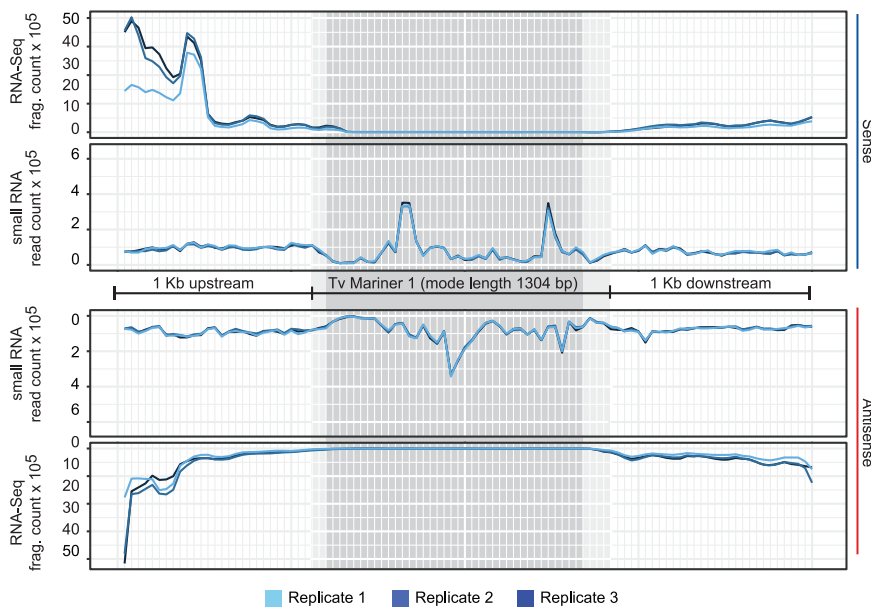


FIG 4 RNA-Seq and sRNA-Seq alignments across all *Tvmar1* family members. All 599 copies of *Tvmar1* ± 1 kb upstream and downstream were divided into 100 bins showing summed read counts in each bin. The dark gray region indicates bins where 100% of *Tvmar1* elements are represented, and the light gray region represents bins where $\geq 90\%$ of *Tvmar1* elements are represented, due to the varying lengths of *Tvmar1* elements.

that a higher proportion of transcribed genes have sRNA-Seq mapping than silent genes, whereas the opposite relationship exists for repeats (Fig. 6C). We attributed this to the presence of *T. vaginalis* gene mRNA degradation products in the sRNA-Seq libraries discussed below and consider it to be an artifact.

Antisense-mapping small RNAs are associated with reduced gene expression.

Next, we investigated the strandedness and read length of the sRNA-Seq reads across transcribed/with RNA-Seq coverage and silent/without RNA-Seq read coverage in protein-coding genes and repeats and compared these to reads mapping to intergenic regions. We found that the sRNA-Seq reads map about equally to the forward and reverse strands of intergenic regions and repeats. However, silent protein-coding genes have slightly elevated levels of antisense sRNA-Seq reads (55%), and expressed protein-coding genes have $>80\%$ of their sRNA-Seq reads mapping to the sense strand (Fig. 7A). To investigate the small RNA strandedness of each gene and the proportion of sRNA-Seq reads mapping to the reverse strand of each TE/repeat (the reverse strand of the contig was used, as many repeats have unknown orientation). Expressed genes were found to be dominated by sense-mapping sRNA-Seq reads, and in all other cases, most repeats and TEs were found to have a 50% distribution of sRNA-Seq reads mapping to each strand, with this pattern observed most strongly in genes and repeats that do not have RNA-Seq coverage.

We next plotted the length distribution of reads in each category, split by mapping strand, and found that in all but one case, mapped reads have a strong mode length of 34 nt (Fig. 7A). The exception is sense-mapping sRNA-Seq reads that map to expressed genes, for which the 34-nt peak is much shallower, with a greater proportion of the reads at all other lengths (Fig. 7A). As a control, we plotted the strandedness and length distribution of the sRNA reads mapping to *Tvmar1* elements, because the orientation of members of the *Tvmar1* family is known, unlike for the other repeat families. This revealed a 50% sense/antisense distribution, with reads mapping to both strands having a strong modal peak at 34 nt (Fig. S3). In addition, we replotted base composition plots for sRNA-Seq reads aligning to all genomic feature types and split by strand.

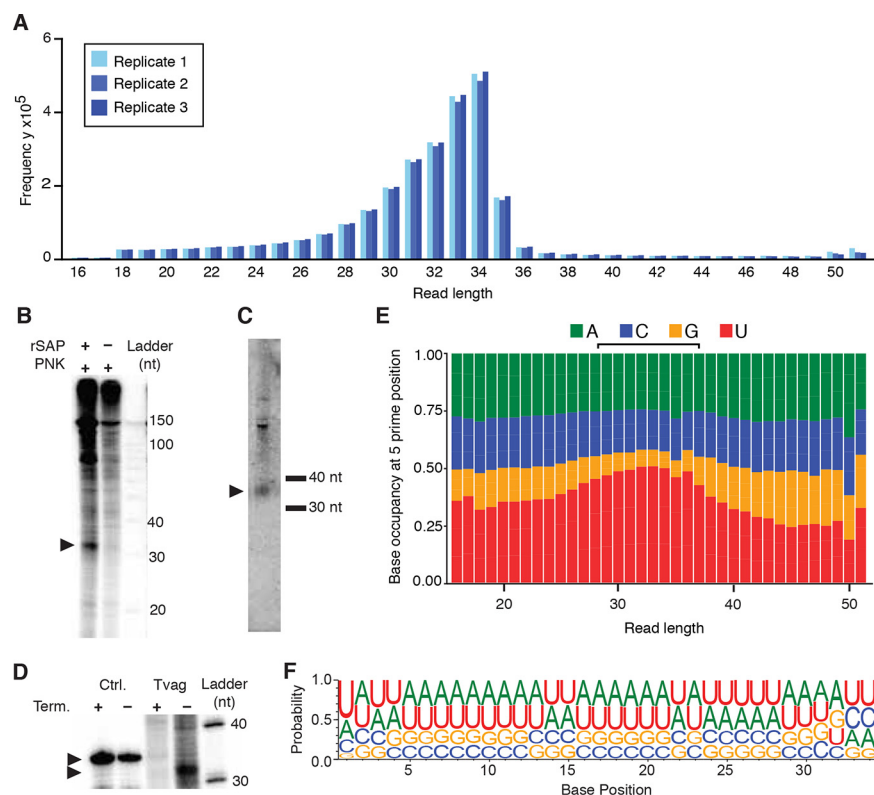


FIG 5 Features of the *T. vaginalis* sRNA-Seq reads. (A) Length distribution plot of the unique sRNA reads. (B) Phosphorimage of labeled *T. vaginalis* total RNA separated on a 15% polyacrylamide gel. One sample was treated with shrimp alkaline phosphatase (rSAP) prior to 5' end labeling with polynucleotide kinase (PNK). Arrowheads indicate bands at ~34 nt. (C) Phosphorimage of ~34-nt band from labeled *T. vaginalis* total RNA and synthetic 34-nt RNA oligonucleotide, each treated and not treated with Terminator 5'-phosphate-dependent exonuclease. The R_f for the 34-nt oligonucleotide was calculated to be 37, and the R_f for the *T. vaginalis* band was calculated to be 34 using the Ambion Decade marker system as a reference. (D) Phosphorimage of a Northern blot of unlabeled *T. vaginalis* total RNA separated on a 15% polyacrylamide gel with custom 30- and 40-nt RNA size markers. (E) Nucleotide distribution at the 5' base in small RNAs of different read lengths (18 to 48 nt). (F) Nucleotide composition at each base along the 34-nt small RNAs. For panels E and F, data from the three small RNA-Seq replicates were summed.

We observed the 5' U bias in reads aligning to all features except for the reads aligning to the sense strand of expressed genes (Fig. S4). Finally, we plotted the RNA-Seq FPKM versus the sRNA-Seq RPKM for genes and repeats and found a positive correlation between increasing FPKM and RPKM in the case of genes and the opposite relationship, i.e., increasing sRNA-Seq reads correlated with decreasing RNA-Seq reads, for repeats (Fig. 7B). We included the proportion of antisense/reverse-mapping sRNA-Seq reads in this analysis and noticed again that many of the transcribed genes for which RNA-Seq and sRNA-Seq were positively correlated had most of their sRNA-Seq reads mapping to the sense strand (Fig. 7B). When we plotted RNA-Seq FPKM versus the proportion of antisense sRNA-Seq reads for each gene, we found that increased proportion of antisense reads was correlated with decreased RNA-Seq FPKM for genes (Fig. 7C).

These findings lead us to hypothesize that the antisense- or bidirectionally mapping ~34-nt small RNAs are correlated with gene and repeat silencing, while the sRNA-Seq reads mapping to the sense strand of expressed genes are most likely a mix of small RNAs and library contaminants produced by mRNA degradation.

Putative piRNA clusters identify 34-nt sRNAs as likely piRNAi guides. We used the software ShortStack to identify regions in the *T. vaginalis* genome that fit the description of piRNA clusters and may generate the sRNAs for sRNA-guided gene

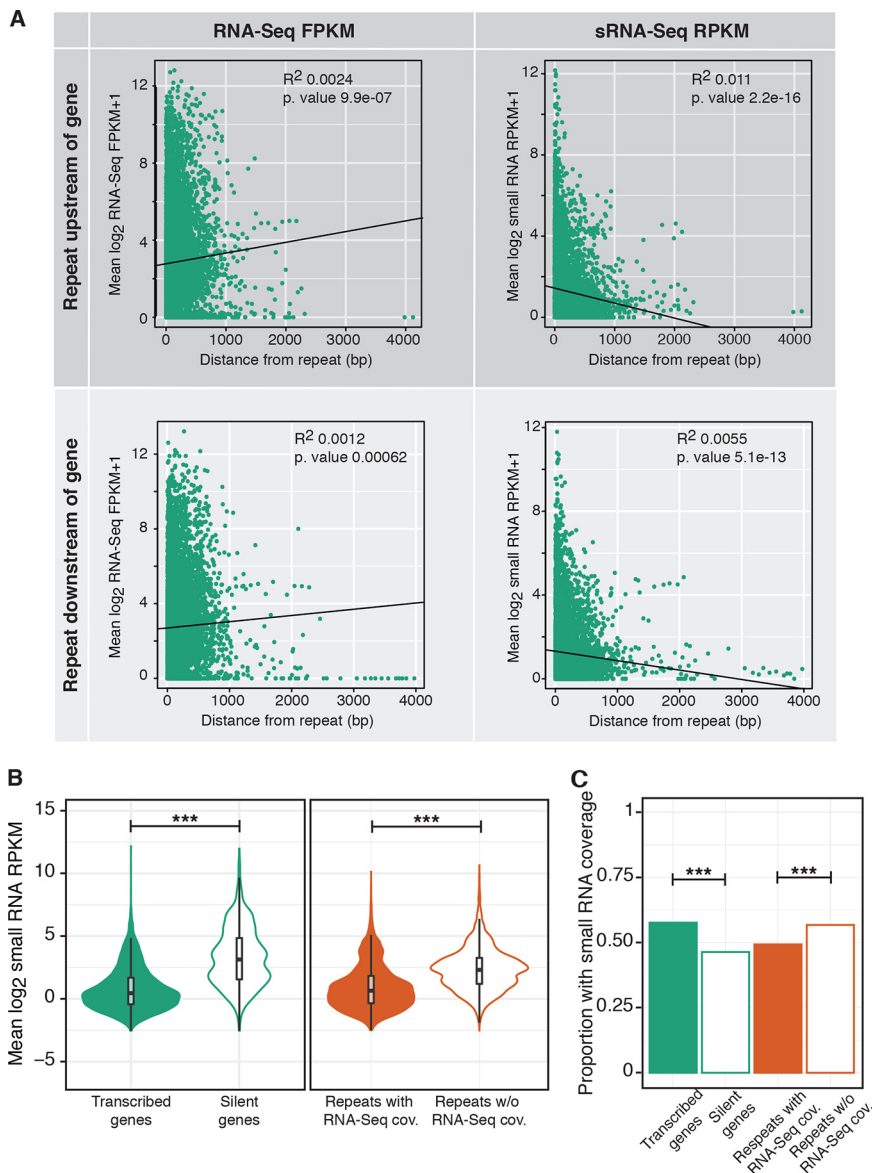


FIG 6 Correlation between gene RNA-Seq or small RNA-Seq and distance from the nearest repeat. (A) Scatterplots showing \log_2 RNA-Seq FPKM or sRNA-Seq RPKM averaged across replicates and plotted against increasing distance from the nearest repeat when the nearest repeat is upstream or downstream of the gene. (B) \log_2 small RNA-Seq RPKM for genes and repeats that are either expressed/covered by RNA-Seq given the threshold. ***, P value < 0.0005 (two-sided t test). (C) Proportion of expressed versus silent genes and covered versus not covered by RNA-Seq repeats that are above the RPKM threshold for small RNA-Seq reads. ***, P value < 0.0005 (Fisher's exact test). cov., coverage.

silencing by PIWI-like Argonaute proteins (Fig. S5 and Table S2). In particular, there were high densities of bidirectionally mapping sRNAs within a 10-Mb region on chromosome IV (14 to 24 Mb) (Fig. S5), which lead us to identify this as a candidate piRNA-generating locus.

DISCUSSION

T. vaginalis has an extraordinarily large genome for a parasitic protist, a trait thought to have arisen by the recent expansion of thousands of repetitive elements (15, 18). Small RNAs are common regulators of repetitive elements in many organisms (32–36, 38, 39, 44, 45), including unicellular eukaryotes. For example, *E. histolytica* generates sRNAs, the most abundant of which are ~27 nt long and have 5'-polyphosphate

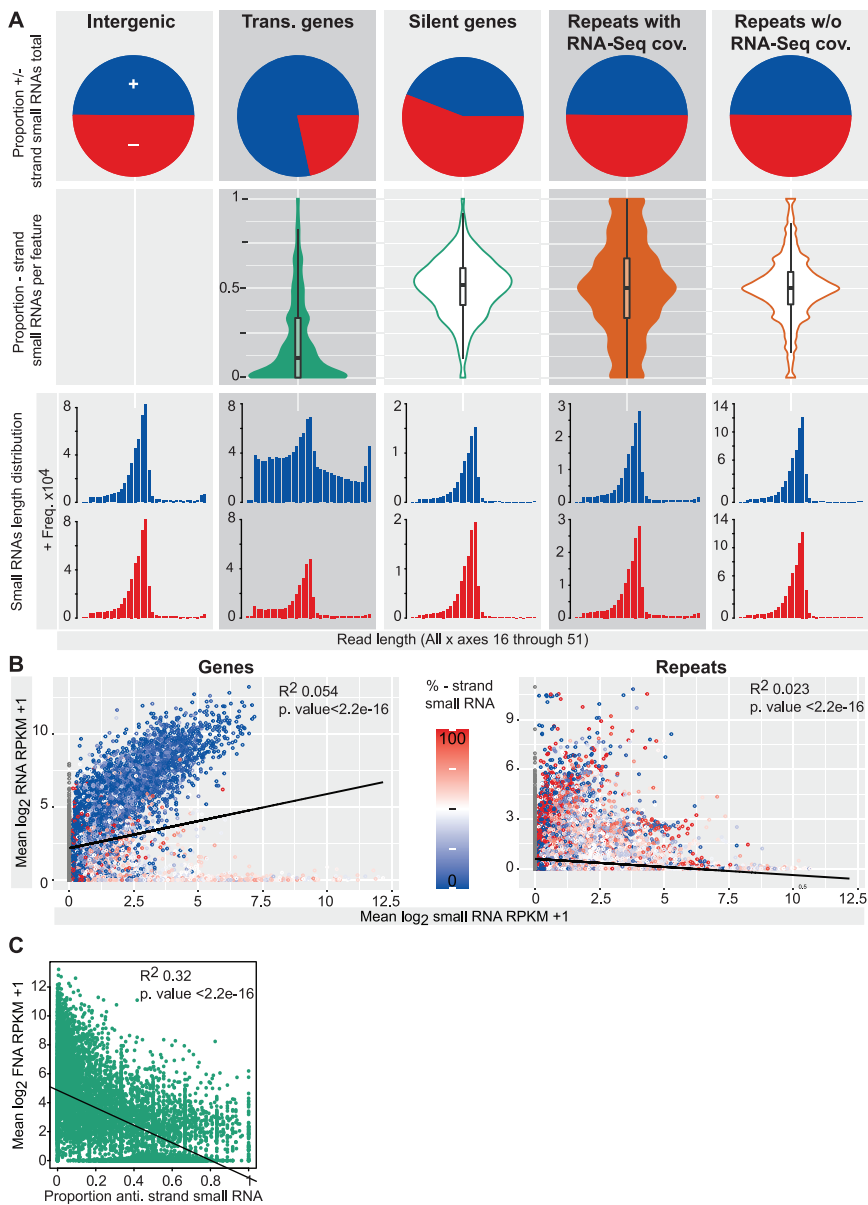


FIG 7 Small RNAs map bidirectionally. (A) Overall proportion of sRNA-Seq reads aligning to genomic features mapped to positive (blue) and negative (red) strands, split by representation in RNA-Seq data. Proportion of antisense/minus-strand mapping small RNA-Seq reads for each gene or repeat, split by representation in RNA-Seq data. Read length distributions for sRNA-Seq reads mapping to genes, repeats split by representation in RNA-Seq data, and intergenic regions are shown. (B) Scatterplots showing log₂ RNA-Seq FPKM versus sRNA-Seq RPKM averaged across replicates and colored by proportion of antisense/reverse-strand-mapping small RNA reads. (C) Scatterplot showing log₂ RNA-Seq FPKM versus proportion of antisense sRNA-Seq reads for genes. Trans., transcribed; anti., antisense; rev., reverse; +, forward/sense strand; -, reverse/antisense strand.

termini (58). Ciliates, including *Tetrahymena thermophila*, *Paramecium tetraurelia*, and *Oxytricha trifallax*, encode PIWI-like Agos, which use a population of ~25- to 30-nt long small cytoplasmic RNAs (scRNAs) to guide TE excision from germ line DNA in the macronucleus (59), both of which are reminiscent of the TE regulatory piRNA pathway in basal (60) and higher (37) metazoans. Here, we present evidence that *T. vaginalis* may employ an ancestral piRNA mechanism mediated by Argonaute protein(s) to regulate its repeats. Our small RNA sequencing results revealed a population of bidirectionally mapping small RNAs with a mode length of ~34 nt, in the range of piRNAs which

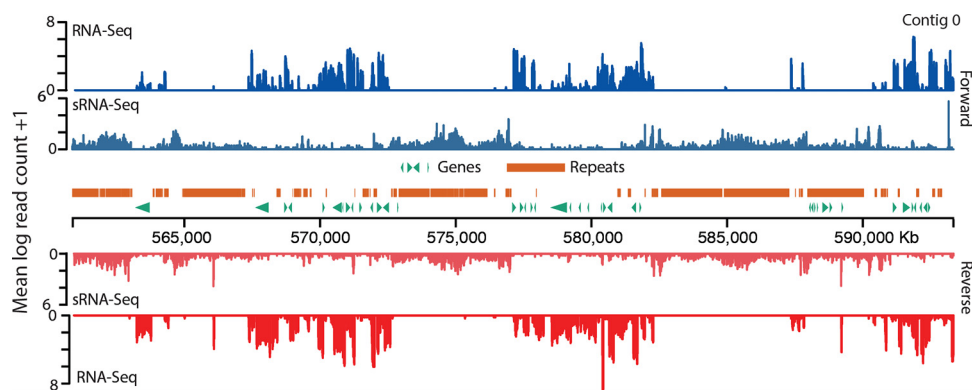


FIG 8 A 300-kb section of chromosome IV plotted to illustrate spatial distribution of sRNAs relative to RNA-Seq reads, protein-coding genes, and repeats. A section of chromosome IV was chosen by generating the first position using a random-number generator and plotting from this position to +300 kb. Read counts were calculated in 100-bp windows.

are 21 to 35 nt (reviewed in reference 37). We confirmed the length of the *T. vaginalis* small RNAs in a total RNA gel and by Northern blotting using a *Tvmar1* element as a probe. In addition, the ~34-nt small RNA population has features characteristic of piRNAs in other organisms, such as a 5'-phosphate group and a 5' U bias (36, 61, 62). Two other lines of evidence agree with our 34-nt sRNA population corresponding to piRNAs: first, the *T. vaginalis* genome encodes two Argonaute orthologs, both of which more closely resembled PIWI-like (piRNA-guided) than AGO-like (miRNA and siRNA-guided) Argonaute proteins (37) in phylogenetic and functional domain analysis, and second, we detected an ~10-Mb region of the *T. vaginalis* genome with characteristics of piRNA clusters, which are required for the generation of functional piRNAs (54). We found that the population of small RNAs maps bidirectionally to the genome, often associating with features and regions that are not transcribed as mRNA and/or are associated with reduced mRNA expression. For example, visualization of the spatial distribution of sRNAs relative to RNA-Seq reads, protein-coding genes, and repeats on a 300-kb region of *T. vaginalis* chromosome IV illustrates our findings (Fig. 8 and Fig. S6); we note that the sRNA-Seq reads cluster in regions that contain repeats and/or have low RNA-Seq mapping. Our conclusion is that this population of ~34-nt bidirectionally mapping small RNAs is likely part of a regulatory mechanism that reduces the transcription of features, particularly repeats and TEs but also protein-coding genes, to which they map. Due to the repetitive nature of the *T. vaginalis* genome, the mapping patterns of the small RNAs do not necessarily indicate their genomic origins. Indeed, we hypothesize that the small RNAs are produced from a few genomic loci, such as the putative piRNA clusters discussed above, and can target multiple additional loci with which they share high sequence similarity. Experiments to test this hypothesis are ongoing in our laboratory.

Our findings are significant for several reasons. First, while these small RNAs may not be the only regulators of expression (for example, *T. vaginalis* miRNAs [which regulate gene expression by RNAi] have previously been identified [50, 63], and other complementary factors, such as promoters and heterochromatin formation, are likely to be involved [64]), the exceptional burden of TEs in the genome, and the demonstrated importance of TE regulation by piRNAs (65), means that they may be an important component of gene regulation in the parasite. Second, this finding complements previous studies in our lab which have shown that TE insertion site polymorphisms exist between different *T. vaginalis* strains and are associated with changes in the expression of proximal protein-coding genes (21). In the work presented here, we observed a trend for reduced sRNA-Seq RPKM and increased RNA-Seq FPKM for protein-coding genes with increased distance from repeats. It is possible that the recruitment of small RNAs to repetitive loci and associated silencing of targeted genes and TEs may be one

mechanism which influences the expression of those nearby genes, possibly through epigenetic changes that alter the chromatin state, as has been shown in *Arabidopsis* (66).

We detected the presence of mRNA degradation products in the small RNA-Seq data, which is a confounding factor for our analyses. Therefore, a primary goal for further elucidation of this possible mechanism of gene and TE regulation is to determine the subcellular localization and piRNA binding propensity of the two *T. vaginalis* Argonaute proteins. The differences in functional domains encoded by the two genes indicate that they likely perform different functions. Experiments to generate antibodies against the *T. vaginalis* AGO1 and AGO2 are ongoing in our laboratory. If differential localization and piRNA binding propensity of the proteins are identified, this may alter either of the two previously characterized methods of piRNA-guided expression regulation: DNA modification (e.g., DNA methylation or alteration of histone marks in the nucleus) (67, 68) and posttranscriptional mRNA silencing (e.g., mRNA cleavage in the cytoplasm) (61, 69, 70). Another priority is to determine the mechanism by which regulatory sRNAs are generated, in particular, whether this requires action of the Dicer ortholog that is encoded in the *T. vaginalis* genome or whether this is a Dicer-independent process, as in metazoan piRNA pathways (37).

In summary, we have identified a novel species of small RNA molecule expressed in *T. vaginalis* parasites grown under standard laboratory conditions. These small RNAs are correlated with reduced expression of protein-coding genes and repeats at the mRNA level. This finding raises the possibility that a small RNA pathway is a major contributor to gene and TE expression patterns in this parasite's genome, opening up new avenues for further investigation into the nature and function of the *T. vaginalis* small RNAs and the diversity of small RNA biogenesis, structure, and function on a wider scale. This mechanism presents an opportunity for harnessing such a system to control gene expression in *T. vaginalis* in a laboratory setting.

MATERIALS AND METHODS

***T. vaginalis* genome sequence and TE annotation.** We used our new genome assembly and preliminary annotation of all six chromosomes of the *T. vaginalis* reference strain G3, which has been submitted to the eukaryotic pathogen genomics database TrichDB (<http://trichdb.org/trichdb/>). Additional identification and annotation of the Maverick family, the largest family of TEs in the genome, and other TE families and repeats were undertaken. Briefly, predicted protein sequences of the 18 ORFs encoding putative proteins of ≥ 50 amino acid residues in Maverick Tv1.1 (the longest of 14 canonical *T. vaginalis* Maverick elements [18]), plus a DNA primase domain protein predicted from an ORF of Maverick Tv1.6, were used in a BLASTx (E value $\leq 1e-03$) (71) search of the *T. vaginalis* G3 genome assembly. Different subclasses of Mavericks can be identified on the basis of subclass-specific ORF order (15; S. Sullivan, personal observation). Inverted repeat coordinates in the assembly output by Inverted Repeats Finder (72) were added to the BLASTx output, and manual inspection revealed blocks of Maverick sequence ranging from full-length "canonical" elements (containing characteristic numbers and orders of ORFs, flanked by terminal inverted repeats) to elements that appeared to have undergone end-to-end fusion, to nested elements, to fragmentary elements. After merging fused, overlapping, and nested coordinates, Maverick blocks of ≥ 150 nt were used in this work.

Consensus TE sequences from Repbase (August 2018, v23.07) were used in BLASTn queries to identify complete and nearly complete non-Maverick transposable elements in the G3 assembly. Additional non-Maverick transposable elements were identified using the Extensive *de novo* TE Annotator (EDTA) pipeline (73). The pipeline comprises a collection of *de novo* TE identification and homology/structure-based annotation programs. LTR_Finder, LTRharvest, and LTR_retriever were used to identify the type 1 transposon family long terminal repeats, GenericRepeatFinder and TIR_learner were used for the identification of type 2 transposons, and HelitronScanner was used for the identification of the Helitron transposon family. EDTA was run under default settings without a prior TE or coding sequence library. RepeatModeler (RepeatModeler Open-1.0, 2008 to 2015 [<http://www.repeatmasker.org/>]), a more general repeat annotation program that uses RECON and RepeatScout, was used at the end of the pipeline for the identification of any remaining unannotated transposon families using default parameters.

Parasite strains and *in vitro* culture. *T. vaginalis* strain G3, the genome reference strain commonly used in research, isolated from Kent, United Kingdom, in 1963, was used for all experiments in this study (15, 74). Parasites were cultured in modified Diamond's medium (75) supplemented with 10% horse serum, penicillin and streptomycin (Invitrogen), and iron solution composed of ferrous ammonium sulfate and sulfosalicylic acid (Fisher Scientific), as described previously (76).

RNA isolation, RNA-Seq, and small RNA-Seq library preparation. *T. vaginalis* strain G3 was grown in triplicate overnight in 15-ml sealed tubes seeded with 2×10^6 parasites total. Total RNA was isolated

using the Qiagen RNeasy minikit, including a column DNase treatment using the Qiagen RNase-Free DNase kit. Polyadenylated RNA was purified from 5 μ g of total RNA using the Dynabeads mRNA DIRECT purification kit. All experiments were carried out in triplicate. First-strand synthesis was performed by mixing the entire fraction of isolated poly(A)⁺ RNA (8 μ l) with 0.5 μ l of random primers (3 μ g/ μ l; Invitrogen), 10 mM dithiothreitol (DTT), and 0.25 μ l of anti-RNase (15 to 30 U/ μ l; Ambion) in 1 \times first-strand synthesis buffer (5 \times ; Invitrogen), with incubation at 65°C for 3 min to remove RNA secondary structures, and then placed on ice. A total of 0.5 μ l of SuperScript III enzyme (200 U/ μ l; Invitrogen) and deoxynucleoside triphosphates (dNTPs) to a final concentration of 0.125 mM were added to the mixture, and reverse transcription was carried out using the following incubations: 25°C for 10 min, 42°C for 50 min, and 70°C for 15 min. The resulting cDNA/RNA hybrid was purified from the mix using Agencourt RNAClean XP beads according to the manufacturer's instructions. Second-strand synthesis was carried out by mixing the purified cDNA/RNA hybrid with 1 μ l of dUTP mix (10 mM; Roche), 0.5 μ l of RNase H (2 U/ μ l; Invitrogen), and 1 μ l of DNA polymerase I (5 to 10 U/ μ l; Invitrogen) in 1 \times NEBuffer 2 with 2.5 mM DTT. This mixture was incubated at 16°C for 2.5 h. The resulting double-stranded cDNA was purified using Agencourt AMPure XP beads according to the manufacturer's instructions. The cDNA was end repaired by mixing 5 μ l of T4 DNA polymerase (3 U/ μ l; New England BioLabs, Inc. [NEB]), 2 μ l of Klenow DNA polymerase (3 to 9 U/ μ l; Invitrogen), and 5 μ l of T4 polynucleotide kinase (10 U/ μ l; NEB) in 1 \times T4 DNA ligase buffer with 10 mM ATP (NEB) with 0.4 mM dNTPs. The mixture was incubated at room temperature for 30 min. The end-repaired cDNA was purified using Agencourt AMPure XP beads according to the manufacturer's instructions. The purified end-repaired cDNA was then taken through A-tailing, adapter ligation, and PCR enrichment using the Illumina TruSeq stranded mRNA sample preparation kit, with different barcodes for each sample. The libraries were pooled and sequenced on an Illumina HiSeq 2500 with 101 cycles, paired-end reads, and multiplexing.

For small RNA-Seq (sRNA-Seq), total RNA was isolated from overnight cultures using the *mirVana* kit (Ambion) according to the manufacturer's instructions. Approximately 1 μ g of total RNA from each replicate was taken through library preparation using the Illumina TruSeq small RNA sample preparation kit, with different barcodes for each sample. The libraries were pooled and sequenced on an Illumina HiSeq 2500 with 50 cycles, single-end reads, and multiplexing.

Northern blotting. Total RNA was extracted from 15 ml of overnight *T. vaginalis* cultures using TRIzol. Smaller (<200 bp) RNAs were enriched using the *mirVana* kit, and 15 μ g of this small RNA-enriched RNA was separated on a 15% polyacrylamide gel, transferred to a nitrocellulose membrane, and RNA cross-linked to the membrane using 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC). A radioactive probe was prepared by cloning the *Tvmar1* (TVAG_TE_DS113512_1) consensus sequence into a TOPO TA Cloning vector (Invitrogen), amplified from *T. vaginalis* G3 genomic DNA using the following primers: forward primer sequence 5'-GAAATCTGTCGTTAGATCTTCG-3' and reverse primer sequence 5'-ATTAATATTTGAGCTTGTGCAC-3', with an amplicon size of 4,121 bp. After propagation of the cloned insert in TOP10 electrocompetent *Escherichia coli*, the probe was amplified from the vector using primers that bind to the ends of the *Tvmar1* element (5'-GCACAGCGCTCTATATGAGACT-3' and 5'-GCACAAACCTGAATACTGCG-3'), producing a 1,304-bp probe. A random primer DNA labeling system (Invitrogen) was used to label the probe using [γ ³²-P]ATP. Size markers were custom 30-nt (*Tvmar1_30_S*; 5'-GAGAUGACAAAGAUACCAGUACAACAGUC-3') and 40-nt (*Tvmar1_40_S*; 5'-AAUGAAUCAUAAAGAAAA CAUCCUCGCUUGCAAAAAA-3') oligonucleotides complementary to the probe DNA.

Small RNA gels. A total of 1 μ g of total RNA was dephosphorylated using shrimp alkaline phosphatase (New England BioLabs, Inc.), and the reaction was stopped by heat inactivation (65°C for 10 min). Radioactive labeling of RNA was achieved by the addition of polynucleotide kinase in the presence of 0.5 μ l of [γ ³²-P]ATP and incubation at 37°C for 1 h. Equal volumes of RNA loading buffer were added, and 1 μ l of the sample was run on a 15% polyacrylamide gel for 1 to 1.5 h, including 1 μ l of Ambion Decade Marker end labeled with [γ ³²-P]ATP as a ladder. The gel was exposed to a phosphorimaging screen for 10 min and developed on a Typhoon FLA 9000 laser scanner (GE Healthcare Life Sciences). The length of the ~34-nt band was determined by extrapolation using the *R_v* plot method.

For determination of the 5' end of the small RNAs, a synthetic 34-nt RNA oligonucleotide (5'-AUC GCG CAC AAC AUC GAG GAC GGC AGC GUG CAG C -3'; subset of the green fluorescent protein [GFP] gene sequence) with no 5' or 3' modifications was made (Integrated DNA Technologies, Inc.). Approximately 10 μ g each of total RNA and synthetic 34-nt oligonucleotide were treated with Terminator 5'-phosphate-dependent exonuclease (Epicentre), which digests RNA that has a 5'-monophosphate end, according to the manufacturer's instructions. The reaction was stopped by phenol extraction, and RNA was recovered by ethanol precipitation. RNA dephosphorylation, end labeling, and visualization from a polyacrylamide gel were performed as described above.

Phylogenetic and domain analysis of *T. vaginalis* Argonaute proteins. Amino acid sequences of Argonaute proteins from a phylogenetically diverse range of organisms were downloaded from the NCBI website (<https://www.ncbi.nlm.nih.gov/>; accessed 24 March 2020). Amino acid sequences were aligned with ClustalW version 2.0 (77), yielding a total of 1,434 informative sites, and the best-fitting evolutionary model was determined to be LG+G4 using ModelFinder (78). A maximum likelihood phylogeny was inferred using IQ-TREE (79) (accessed 30 June 2020), and node support was evaluated with 1,000 ultrafast bootstraps (80). Pfam functional domains were annotated for a subset of Argonaute protein amino acid sequences from the phylogenetic analysis that clustered within the PIWI-like clade using MotifFinder and the Pfam protein database using an E value cutoff of <0.0001 (<https://www.genome.jp/tools/motif/>; accessed 26 March 2020).

Sequencing data filtering and cleanup. RNA-Seq and sRNA-Seq data were quality filtered and adapters trimmed using TrimGalore version 0.4.4 (81), with a quality Phred score cutoff of 20. Trimmed

TABLE 3 Sequencing library statistics, showing numbers of sequencing read pairs generated in the RNA-Seq and reads in the sRNA-Seq data sets retained through each data filtering and cleanup stage

| Parameter | RNA-Seq 1 | RNA-Seq 2 | RNA-Seq 3 | sRNA-Seq 1 | sRNA-Seq 2 | sRNA-Seq 3 |
|---|-----------------|------------|------------|------------|------------|------------|
| Total raw sequences | 12,776,898 | 17,628,836 | 12,274,083 | 56,823,267 | 47,747,619 | 56,887,075 |
| Total after adapter trimming and quality filtering | 12,599,522 | 17,422,595 | 12,139,283 | 49,656,557 | 43,287,940 | 52,949,241 |
| Total aligning to rRNA | 402,662 | 541,876 | 339,097 | 8,195,308 | 6,806,737 | 7,580,078 |
| Total aligning to tRNA | 405 | 283 | 186 | 29,840,556 | 25,763,072 | 33,558,897 |
| Total aligning to genome (rRNA/tRNA excluded) | 11,922,079 | 16,605,308 | 11,619,118 | 7,274,576 | 6,838,352 | 6,944,101 |
| Total unique reads aligning to genome (rRNA, tRNA excluded) | NA ^a | NA | NA | 2,661,226 | 2,565,682 | 2,677,716 |

^aNA, not applicable.

RNA-Seq pairs were discarded if one read was shorter than 35 bp; trimmed sRNA-Seq reads shorter than 18 bp were also discarded. Reads containing homopolymer (A's or T's) and reads containing N's were removed using Cutadapt version 1.16 (82). Bowtie2 version 2.3.4.3 (83) was used to align the subsequent sets of reads to all the rRNA and tRNA loci from the *T. vaginalis* genome, retaining only unaligned reads and read pairs. Finally, the sRNA-Seq data sets were collapsed to unique sequences as per best practices for genomes that contain multiple repeats (84) using the clumpify tool from the BBMap package version 37.48 (85) and used in all subsequent analyses. The number of reads remaining after each filtering step is shown in Table 3.

RNA-Seq and sRNA-Seq bioinformatics analysis and terminology. The filtered RNA-Seq and unique filtered sRNA-Seq replicates were aligned to the G3 reference genome sequence using Bowtie2 version 2.3.4.3 (83), using default end-to-end mode allowing for a maximum fragment length of 1,300 bp for the RNA-Seq and default single-end mode for the sRNA-Seq. Under these conditions only one mapping locus is returned for each read, including when a read maps to more than one genomic locus; for the sRNA-Seq reads, ~75% of each library mapped to more than one genomic locus. Reads mapping to genomic features were counted using HTSeq-count version 0.9.1 (86) with the minimum MAPQ set at 0, allowing for reads mapping to many locations in the genome to still be counted. The data were further analyzed using Samtools version 1.9, Bedtools version 2.27.1 (87), and custom R and Unix scripts. The sRNA-Seq sequence logos were made using WebLogo 3.5.0 (88, 89).

We generated a statistical method to classify what proportion of genes and repeats were transcribed or covered by RNA-Seq and/or covered by sRNA-Seq reads, using FPKM (fragments per kilobase per million mapped fragments, where a fragment represents the two paired-end reads of an RNA-seq fragment) for RNA-Seq data, and RPKM (reads per kilobase per million mapped reads, where a read represents a single ended sRNA-seq read) for sRNA-Seq data. Genes are transcribed and repeats covered for RNA-Seq if they have a \log_2 RNA-Seq FPKM value greater than or equal to a threshold. Genes and repeats are covered for sRNA-Seq if they have a \log_2 sRNA-Seq RPKM value greater than or equal to a threshold. In both cases the threshold is the mean \log_2 RNA-Seq FPKM or sRNA-Seq RPKM minus 2 standard deviations, calculated for all genes and repeats having an RNA-Seq FPKM/sRNA-Seq RPKM ratio of ≥ 0 , and calculated independently in each biological replicate (Fig. S7). This threshold corresponds to RNA-Seq FPKM thresholds of 0.021, 0.014, and 0.022 for RNA-Seq replicates 1, 2, and 3, respectively, and small RNA-Seq RPKM thresholds of 0.162, 0.166, and 0.168 for small RNA-Seq replicates 1, 2, and 3, respectively. This threshold was used because the mechanism by which the small RNAs act is not yet known, and the repetitive nature of the *T. vaginalis* genome means that short reads can be spread out over multiple loci in the genome. A list of all genomic features, genes, and repeats and a description of whether they are expressed and/or covered by sRNA-Seq reads is presented in Table S3. For a gene to be transcribed or a repeat to be covered by RNA-Seq data, or for either to be covered by sRNA-Seq data, that gene or repeat must be equal to or above this threshold in all three biological replicates.

Candidate piRNA loci were identified using ShortStack (90) version 3.8.5 with the options -nohp, a Dicer range of 20 to 35 nt, and the unique weighting mode to place multimapped reads (91). RPM values were calculated by ShortStack.

Data availability. Whole-genome RNA-seq data for *T. vaginalis* strain G3 in triplicate have been deposited in NCBI's Sequence Read Archive under accession no. [SRX1122976](#), [SRX1122977](#), and [SRX1122978](#) and small RNA-seq data for *T. vaginalis* strain G3 under BioProject identifier [PRJNA647375](#).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.2 MB.

FIG S2, PDF file, 0.4 MB.

FIG S3, PDF file, 0.1 MB.

FIG S4, PDF file, 0.2 MB.

FIG S5, PDF file, 0.1 MB.

FIG S6, PDF file, 1 MB.

FIG S7, PDF file, 0.1 MB.

TABLE S1, DOCX file, 0.02 MB.

TABLE S2, XLS file, 4.9 MB.

TABLE S3, XLS file, 8.1 MB.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R21AI149449. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We thank the Center for Genomics and Systems Biology Genomics Core, in particular Tara Rock for help and consultation during library preparation and sequencing. We also thank Laura Landweber and Jaspreet Khurana at Columbia University for discussions on methods and data analysis and Duncan Smith and lab members at NYU for help and equipment to run polyacrylamide gels and Northern blots.

Sally D. Warring, conceptualization, investigation, formal analysis, validation, visualization, writing – original draft; Frances Blow, conceptualization, investigation, formal analysis, visualization, writing – review & editing; Grace Avecilla, formal analysis; Jordan C. Orosco, formal analysis; Steven A. Sullivan, data curation, formal analysis, writing – review & editing; Jane M. Carlton, conceptualization, funding acquisition, methodology, project administration, writing – review & editing.

REFERENCES

- World Health Organization. 2011. Prevalence and incidence of selected sexually transmitted infections, Chlamydia trachomatis, Neisseria gonorrhoeae, syphilis and Trichomonas vaginalis: methods and results used by WHO to generate 2005 estimates. WHO Press, Geneva, Switzerland.
- Kissinger P. 2015. Trichomonas vaginalis: a review of epidemiologic, clinical and treatment issues. BMC Infect Dis 15:307. <https://doi.org/10.1186/s12879-015-1055-0>.
- Johnston VJ, Mabey DC. 2008. Global epidemiology and control of Trichomonas vaginalis. Curr Opin Infect Dis 21:56–64. <https://doi.org/10.1097/QCO.0b013e3282f3d999>.
- McClelland RS, Sangare L, Hassan WM, Lavreys L, Mandaliya K, Kiarie J, Ndinya-Achola J, Jaoko W, Baeten JM. 2007. Infection with *Trichomonas vaginalis* increases the risk of HIV-1 acquisition. J Infect Dis 195:698–702. <https://doi.org/10.1086/511278>.
- Bowden FJ, Garnett GP. 2000. Trichomonas vaginalis epidemiology: parameterising and analysing a model of treatment interventions. Sex Transm Infect 76:248–256. <https://doi.org/10.1136/sti.76.4.248>.
- Cotch MF, Pastorek JG, II, Nugent RP, Hillier SL, Gibbs RS, Martin DH, Eschenbach DA, Edelman R, Carey JC, Regan JA, Krohn MA, Klebanoff MA, Rao AV, Rhoads GG. 1997. *Trichomonas vaginalis* associated with low birth weight and preterm delivery. The Vaginal Infections and Prematurity Study Group. Sex Transm Dis 24:353–360. <https://doi.org/10.1097/00007435-199707000-00008>.
- Watt L, Jennison RF. 1960. Clinical evaluation of metronidazole. A new systemic trichomonicide. Br Med J 2:902–905. <https://doi.org/10.1136/bmj.2.5203.902>.
- Upcroft JA, Dunn LA, Wal T, Tabrizi S, Delgadillo-Correa MG, Johnson PJ, Garland S, Siba P, Upcroft P. 2009. Metronidazole resistance in Trichomonas vaginalis from highland women in Papua New Guinea. Sex Health 6:334–338. <https://doi.org/10.1071/SH09011>.
- Kirkcaldy RD, Augostini P, Asbel LE, Bernstein KT, Kerani RP, Mettenbrink CJ, Pathela P, Schwebke JR, Secor WE, Workowski KA, Davis D, Braxton J, Weinstock HS. 2012. Trichomonas vaginalis antimicrobial drug resistance in 6 US cities, STD Surveillance Network, 2009–2010. Emerg Infect Dis 18:939–943. <https://doi.org/10.3201/eid1806.111590>.
- Muzny CA. 2018. Why does Trichomonas vaginalis continue to be a “neglected” sexually transmitted infection? Clin Infect Dis 67:218–220. <https://doi.org/10.1093/cid/ciy085>.
- Hirt RP, Sherrard J. 2015. Trichomonas vaginalis origins, molecular pathobiology and clinical considerations. Curr Opin Infect Dis 28:72–79. <https://doi.org/10.1097/QCO.0000000000000128>.
- Mercer F, Johnson PJ. 2018. Trichomonas vaginalis: pathogenesis, symbiont interactions, and host cell immune responses. Trends Parasitol 34:683–693. <https://doi.org/10.1016/j.pt.2018.05.006>.
- Zubáková Z, Cimbůrek Z, Tachezy J. 2008. Comparative analysis of trichomonad genome sizes and karyotypes. Mol Biochem Parasitol 161:49–54. <https://doi.org/10.1016/j.molbiopara.2008.06.004>.
- Barratt J, Gough R, Stark D, Ellis J. 2016. Bulky trichomonad genomes: encoding a Swiss army knife. Trends Parasitol 32:783–797. <https://doi.org/10.1016/j.pt.2016.05.014>.
- Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, Sicheritz-Ponten T, Noel CJ, Dacks JB, Foster PG, Simillion C, Van de Peer Y, Miranda-Saavedra D, Barton GJ, Westrop GD, Müller S, Dessi D, Fiori PL, Ren Q, Paulsen I, Zhang H, Bastida-Corcuera FD, Simoes-Barbosa A, Brown MT, Hayes RD, Mukherjee M, Okumura CY, Schneider R, Smith AJ, Vanacova S, Villalvazo M, Haas BJ, Perteu M, Feldblyum TV, Utterback TR, Shu C-L, Osoegawa K, de Jong PJ, Hrdy I, Horvathova L, Zubacova Z, Dolezal P, Malik S-B, Logsdon JM, Henze K, Gupta A, et al. 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science 315:207–212. <https://doi.org/10.1126/science.1132894>.
- Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405. <https://doi.org/10.1038/nrg2337>.
- Silva JC, Bastida F, Bidwell SL, Johnson PJ, Carlton JM. 2005. A potentially functional *mariner* transposable element in the protist *Trichomonas vaginalis*. Mol Biol Evol 22:126–134. <https://doi.org/10.1093/molbev/msh260>.
- Pritham EJ, Putliwala T, Feschotte C. 2007. *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390:3–17. <https://doi.org/10.1016/j.gene.2006.08.008>.
- Meng Q, Chen K, Ma L, Hu S, Yu J. 2011. A systematic identification of *Kolobok* superfamily transposons in *Trichomonas vaginalis* and sequence analysis on related transposases. J Genet Genomics 38:63–70. <https://doi.org/10.1016/j.jcg.2011.01.003>.
- Lopes FR, Silva JC, Benchimol M, Costa GG, Pereira GA, Carareto CM. 2009. The protist *Trichomonas vaginalis* harbors multiple lineages of transcriptionally active *Mutator*-like elements. BMC Genomics 10:330. <https://doi.org/10.1186/1471-2164-10-330>.
- Bradic M, Warring SD, Low V, Carlton JM. 2014. The *Tc1/mariner* transposable element family shapes genetic variation and gene expression in the protist *Trichomonas vaginalis*. Mob DNA 5:12. <https://doi.org/10.1186/1759-8753-5-12>.
- Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. Trends Ecol Evol 15:95–99. [https://doi.org/10.1016/s0169-5347\(99\)01817-0](https://doi.org/10.1016/s0169-5347(99)01817-0).
- Pritham EJ, Feschotte C, Wessler SR. 2005. Unexpected diversity and

- differential success of DNA transposons in four species of *Entamoeba* protozoans. *Mol Biol Evol* 22:1751–1763. <https://doi.org/10.1093/molbev/msi169>.
24. Hedges DJ, Deininger PL. 2007. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* 616:46–59. <https://doi.org/10.1016/j.mrfmmm.2006.11.021>.
 25. Callinan PA, Batzer MA. 2006. Retrotransposable elements and human disease. *Genome Dyn* 1:104–115. <https://doi.org/10.1159/000092503>.
 26. Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* 46:21–42. <https://doi.org/10.1146/annurev-genet-110711-155621>.
 27. Thornburg BG, Gotea V, Makalowski W. 2006. Transposable elements as a significant source of transcription regulating signals. *Gene* 365:104–110. <https://doi.org/10.1016/j.gene.2005.09.036>.
 28. Dolgin ES, Charlesworth B. 2006. The fate of transposable elements in asexual populations. *Genetics* 174:817–827. <https://doi.org/10.1534/genetics.106.060434>.
 29. Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297. [https://doi.org/10.1016/s0092-8674\(04\)00045-5](https://doi.org/10.1016/s0092-8674(04)00045-5).
 30. Bologna NG, Voinnet O. 2014. The diversity, biogenesis, and activities of endogenous silencing small RNAs in *Arabidopsis*. *Annu Rev Plant Biol* 65:473–503. <https://doi.org/10.1146/annurev-arplant-050213-035728>.
 31. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391:806–811. <https://doi.org/10.1038/35888>.
 32. Girard A, Hannon GJ. 2008. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol* 18:136–148. <https://doi.org/10.1016/j.tcb.2008.01.004>.
 33. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Philippov DV, Blaser H, Raz E, Moens CB, Plasterk RH, Hannon GJ, Draper BW, Ketting RF. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell* 129:69–82. <https://doi.org/10.1016/j.cell.2007.03.026>.
 34. Kalmykova AI, Klenov MS, Gvozdev VA. 2005. Argonaute protein PIWI controls mobilization of retrotransposons in the *Drosophila* male germline. *Nucleic Acids Res* 33:2052–2059. <https://doi.org/10.1093/nar/gki323>.
 35. Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* 137:522–535. <https://doi.org/10.1016/j.cell.2009.03.040>.
 36. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* 313:320–324. <https://doi.org/10.1126/science.1129333>.
 37. Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* 20:89–108. <https://doi.org/10.1038/s41576-018-0073-3>.
 38. Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol* 2:E104. <https://doi.org/10.1371/journal.pbio.0020104>.
 39. Reinhart BJ, Bartel DP. 2002. Small RNAs correspond to centromere heterochromatic repeats. *Science* 297:1831. <https://doi.org/10.1126/science.1077183>.
 40. Provost P, Silverstein RA, Dishart D, Walfridsson J, Djupedal I, Kniola B, Wright A, Samuelsson B, Radmark O, Ekwall K. 2002. Dicer is required for chromosome segregation and gene silencing in fission yeast cells. *Proc Natl Acad Sci U S A* 99:16648–16653. <https://doi.org/10.1073/pnas.212633199>.
 41. Bernstein E, Caudy AA, Hammond SM, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409:363–366. <https://doi.org/10.1038/35053110>.
 42. Tuschl T, Zamore PD, Lehmann R, Bartel DP, Sharp PA. 1999. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev* 13:3191–3197. <https://doi.org/10.1101/gad.13.24.3191>.
 43. Hammond SM, Bernstein E, Beach D, Hannon GJ. 2000. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* 404:293–296. <https://doi.org/10.1038/35005107>.
 44. Djikeng A, Shi H, Tschudi C, Ullu E. 2001. RNA interference in *Trypanosoma brucei*: cloning of small interfering RNAs provides evidence for retroposon-derived 24–26-nucleotide RNAs. *RNA* 7:1522–1530.
 45. Shi H, Djikeng A, Tschudi C, Ullu E. 2004. Argonaute protein in the early divergent eukaryote *Trypanosoma brucei*: control of small interfering RNA accumulation and retroposon transcript abundance. *Mol Cell Biol* 24:420–427. <https://doi.org/10.1128/mcb.24.1.420-427.2004>.
 46. Zhang H, Ehrenkaufer GM, Hall N, Singh U. 2013. Small RNA pyrosequencing in the protozoan parasite *Entamoeba histolytica* reveals strain-specific small RNAs that target virulence genes. *BMC Genomics* 14:53. <https://doi.org/10.1186/1471-2164-14-53>.
 47. Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF. 2012. Piwi-interacting RNAs protect DNA against loss during *Oxytricha* genome rearrangement. *Cell* 151:1243–1255. <https://doi.org/10.1016/j.cell.2012.10.045>.
 48. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. 2002. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* 110:689–699. [https://doi.org/10.1016/s0092-8674\(02\)00909-1](https://doi.org/10.1016/s0092-8674(02)00909-1).
 49. Mochizuki K, Gorovsky MA. 2004. Small RNAs in genome rearrangement in *Tetrahymena*. *Curr Opin Genet Dev* 14:181–187. <https://doi.org/10.1016/j.gde.2004.01.004>.
 50. Chen XS, Collins LJ, Biggs PJ, Penny D. 2009. High throughput genome-wide survey of small RNAs from the parasitic protists *Giardia intestinalis* and *Trichomonas vaginalis*. *Genome Biol Evol* 1:165–175. <https://doi.org/10.1093/gbe/evp017>.
 51. Huang PJ, Lin WC, Chen SC, Lin YH, Sun CH, Lyu PC, Tang P. 2012. Identification of putative miRNAs from the deep-branching unicellular flagellates. *Genomics* 99:101–107. <https://doi.org/10.1016/j.ygeno.2011.11.002>.
 52. Wang J, Davis RE. 2014. Programmed DNA elimination in multicellular organisms. *Curr Opin Genet Dev* 27:26–34. <https://doi.org/10.1016/j.gde.2014.03.012>.
 53. Czech B, Hannon GJ. 2016. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends Biochem Sci* 41:324–337. <https://doi.org/10.1016/j.tibs.2015.12.008>.
 54. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128:1089–1103. <https://doi.org/10.1016/j.cell.2007.01.043>.
 55. Matsumoto N, Nishimasu H, Sakakibara K, Nishida KM, Hirano T, Ishitani R, Siomi H, Siomi MC, Nureki O. 2016. Crystal structure of silkworm PIWI-clade Argonaute Siwi bound to piRNA. *Cell* 167:484–497.e9. <https://doi.org/10.1016/j.cell.2016.09.002>.
 56. O'Donnell KA, Boeke JD. 2007. Mighty Piwis defend the germline against genome intruders. *Cell* 129:37–44. <https://doi.org/10.1016/j.cell.2007.03.028>.
 57. Zhuang F, Fuchs RT, Robb GB. 2012. Small RNA expression profiling by high-throughput sequencing: implications of enzymatic manipulation. *J Nucleic Acids* 2012:360358. <https://doi.org/10.1155/2012/360358>.
 58. Zhang H, Ehrenkaufer GM, Pompey JM, Hackney JA, Singh U. 2008. Small RNAs with 5'-polyphosphate termini associate with a Piwi-related protein and regulate gene expression in the single-celled eukaryote *Entamoeba histolytica*. *PLoS Pathog* 4:e1000219. <https://doi.org/10.1371/journal.ppat.1000219>.
 59. Schoeberl UE, Mochizuki K. 2011. Keeping the soma free of transposons: programmed DNA elimination in ciliates. *J Biol Chem* 286:37045–37052. <https://doi.org/10.1074/jbc.R111.276964>.
 60. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455:1193–1197. <https://doi.org/10.1038/nature07415>.
 61. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* 315:1587–1590. <https://doi.org/10.1126/science.1140494>.
 62. Stein CB, Genzor P, Mitra S, Elchert AR, Ipsaro JJ, Benner L, Sobti S, Su Y, Hammell M, Joshua-Tor L, Haase AD. 2019. Decoding the 5' nucleotide bias of PIWI-interacting RNAs. *Nat Commun* 10:828. <https://doi.org/10.1038/s41467-019-08803-z>.
 63. Lin WC, Li SC, Lin WC, Shin JW, Hu SN, Yu XM, Huang TY, Chen SC, Chen HC, Chen SJ, Huang PJ, Gan RR, Chiu CH, Tang P. 2009. Identification of microRNA in the protist *Trichomonas vaginalis*. *Genomics* 93:487–493. <https://doi.org/10.1016/j.ygeno.2009.01.004>.
 64. Lizarraga A, O'Brien ZK, Boulias K, Roach L, Greer EL, Johnson PJ, Strobl-Mazzulla PH, de Miguel N. 2020. Adenine DNA methylation, 3D genome organization, and gene expression in the parasite *Trichomonas vaginalis*. *Proc Natl Acad Sci U S A* 117:13033–13043. <https://doi.org/10.1073/pnas.1917286117>.
 65. Aravin AA, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318:761–764. <https://doi.org/10.1126/science.1146484>.
 66. Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence

- between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci U S A* 108:2322–2327. <https://doi.org/10.1073/pnas.1018222108>.
67. Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* 12:503–514. <https://doi.org/10.1016/j.devcel.2007.03.001>.
 68. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Toth KF. 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev* 27:390–399. <https://doi.org/10.1101/gad.209841.112>.
 69. Nishida KM, Saito K, Mori T, Kawamura Y, Nagami-Okada T, Inagaki S, Siomi H, Siomi MC. 2007. Gene silencing mechanisms mediated by Aubergine piRNA complexes in *Drosophila* male gonad. *RNA* 13:1911–1922. <https://doi.org/10.1261/ma.744307>.
 70. Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev* 20:2214–2222. <https://doi.org/10.1101/gad.1454806>.
 71. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
 72. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res* 14:1861–1869. <https://doi.org/10.1101/gr.2542904>.
 73. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, Jiang N, Hirsch CN, Hufford MB. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* 20:275. <https://doi.org/10.1186/s13059-019-1905-y>.
 74. Coombs GH, Clackson TE. 1983. Antitrichomonal activity of compounds that affect DNA and its repair. *J Antimicrob Chemother* 11:191–194. <https://doi.org/10.1093/jac/11.2.191>.
 75. Diamond LS. 1957. The establishment of various trichomonads of animals and man in axenic cultures. *J Parasitol* 43:488–490. <https://doi.org/10.2307/3274682>.
 76. Conrad M, Zubacova Z, Dunn LA, Upcroft J, Sullivan SA, Tachezy J, Carlton JM. 2011. Microsatellite polymorphism in the sexually transmitted human pathogen *Trichomonas vaginalis* indicates a genetically diverse parasite. *Mol Biochem Parasitol* 175:30–38. <https://doi.org/10.1016/j.molbiopara.2010.08.006>.
 77. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948. <https://doi.org/10.1093/bioinformatics/btm404>.
 78. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589. <https://doi.org/10.1038/nmeth.4285>.
 79. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* 44:W232–W235. <https://doi.org/10.1093/nar/gkw256>.
 80. Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* 30:1188–1195. <https://doi.org/10.1093/molbev/mst024>.
 81. Krueger F. 2012. Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
 82. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17:10–12. <https://doi.org/10.14806/ej.17.1.200>.
 83. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>.
 84. Deschamps-Francoeur G, Simoneau J, Scott MS. 2020. Handling multi-mapped reads in RNA-seq. *Comput Struct Biotechnol J* 18:1569–1576. <https://doi.org/10.1016/j.csbj.2020.06.014>.
 85. Bushnell B. 2014. BBMap short read aligner, and other bioinformatic tools. <https://sourceforge.net/projects/bbmap/>.
 86. Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–169. <https://doi.org/10.1093/bioinformatics/btu638>.
 87. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842. <https://doi.org/10.1093/bioinformatics/btq033>.
 88. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* 14:1188–1190. <https://doi.org/10.1101/gr.849004>.
 89. Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18:6097–6100. <https://doi.org/10.1093/nar/18.20.6097>.
 90. Axtell MJ. 2013. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 19:740–751. <https://doi.org/10.1261/ma.035279.112>.
 91. Johnson NR, Yeoh JM, Coruh C, Axtell MJ. 2016. Improved placement of multi-mapping small RNAs. *G3 (Bethesda)* 6:2103–2111. <https://doi.org/10.1534/g3.116.030452>.