scientific reports

Check for updates

OPEN

piRNA-like small RNAs target transposable elements in a Clade IV parasitic nematode

Mona Suleiman[1,6], Asuka Kounosu[2,6], Ben Murcott[1], Mehmet Dayi[2,3], Rebecca Pawluk[1], Akemi Yoshida[4], Mark Viney[5], Taisei Kikuchi[2]✉ & Vicky L. Hunt[1]✉

The small RNA (sRNA) pathways identified in the model organism *Caenorhabditis elegans* are not widely conserved across nematodes. For example, the PIWI pathway and PIWI-interacting RNAs (piRNAs) are involved in regulating and silencing transposable elements (TE) in most animals but have been lost in nematodes outside of the *C. elegans* group (Clade V), and little is known about how nematodes regulate TEs in the absence of the PIWI pathway. Here, we investigated the role of sRNAs in the Clade IV parasitic nematode *Strongyloides ratti* by comparing two genetically identical adult stages (the parasitic female and free-living female). We identified putative small-interfering RNAs, microRNAs and tRNA-derived sRNA fragments that are differentially expressed between the two adult stages. Two classes of sRNAs were predicted to regulate TE activity including (i) a parasite-associated class of 21–22 nt long sRNAs with a 5′ uridine (21-22Us) and a 5′ monophosphate, and (ii) 27 nt long sRNAs with a 5′ guanine/adenine (27GAs) and a 5′ modification. The 21-22Us show striking resemblance to the 21U PIWI-interacting RNAs found in *C. elegans*, including an AT rich upstream sequence, overlapping loci and physical clustering in the genome. Overall, we have shown that an alternative class of sRNAs compensate for the loss of piRNAs and regulate TE activity in nematodes outside of Clade V.

Small RNAs (sRNA) are short non-coding RNAs important for the regulation of gene expression via post-transcriptional gene silencing. They regulate the expression of at least 30% of genes in humans and are associated with chromatin structure, mRNA translation and the regulation of transposable element (TE) activity[1–3]. Three main sRNA classes have been described in eukaryotes; microRNAs (miRNAs), small-interfering RNAs (siRNAs), and piwi-interacting RNA (piRNAs), classified based on their biogenesis, function and interaction with specific Argonaute proteins[4]. The majority of sRNA research has been carried out in model organisms including *Drosophila melanogaster*, *Mus musculus* and *Caenorhabditis elegans*. *C. elegans* belongs to the Clade V nematodes and possess all three classes of sRNA like other model organisms. However, recent studies have shown that sRNA pathways are highly diverged in nematodes and *C. elegans* does not closely represent the sRNAs used by more distantly related nematodes, including parasitic species[5]. For example, the PIWI pathway involved in the production of piRNAs is important in regulating TE activity and has been well characterised in *C. elegans* but has been lost in nematodes outside of Clade V, including *Strongyloides* spp.[6,7]. It remains unclear how nematodes outside of Clade V compensate for the loss of piRNAs to regulate TE activity.

piRNAs were first identified in *D. melanogaster* as PIWI-clade Argonaute interacting sRNAs and have subsequently been discovered in most other animals, including *M. musculus,* humans and *C. elegans* where they regulate TE activity, particularly in the germline[8,9]. TEs are mobile DNA sequences that move around the genome from one location to another, inserting randomly and causing mutations[8]. They play important roles in the evolution of eukaryotic organisms but can have detrimental effects to the genome and require tight regulation[10,11]. Interestingly, while the role of piRNAs is widely conserved across eukaryotes including their role in fertility and protecting the germline from TEs[12–14], the PIWI pathway, piRNA biogenesis and the mechanism of action has diverged between organisms[2]. For example, in *D. melanogaster* and *M. musculus*, piRNAs are often involved in the "ping-pong" cycle, where antisense primary piRNAs initiate an amplification loop to generate secondary

[1]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK. [2]Parasitology, Department of Infectious Dieses, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan. [3]Forestry Vocational School, Duzce University, 81620 Duzce, Turkey. [4]Laboratory of Genomics, Frontier Science Research Center, University of Miyazaki, Miyazaki 889-1692, Japan. [5]Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool L69 7ZB, UK. [6]These authors contributed equally: Mona Suleiman and Asuka Kounosu. ✉email: taisei_kikuchi@med.miyazaki-u.ac.jp; bs1vlh@bath.ac.uk

piRNAs. These piRNAs are 24–30 nucleotides (nt) long with a bias for a 5′ uracil (5′ U) and silence transposons through perfect antisense complementarity to their target sequences[15,16]. This is not the case in *C. elegans* where piRNAs are 21nt long and although have a bias for a 5′ U, they don't require perfect complementarity to their target sequence and are not involved in the ping-pong cycle. Instead, piRNAs in *C. elegans* initiate the synthesis of 22 nt long siRNAs with a 5′ guanine (22G) through RNA-dependent RNA polymerases (RdRPs) that silence complementary target sequences[5,13,14]. The absence of piRNAs in nematodes outside of Clade V leads us to consider if an alternative sRNA pathway can compensate for the loss of piRNAs and regulate TE activity in nematodes in other clades. It is therefore essential that we study nematodes outside of the *C. elegans* Clade, including both parasitic and free-living species, to better understand the diversity of sRNAs involved in regulating TEs and the role of sRNAs in parasitism.

Although both miRNA and siRNA pathways are found in all nematodes studied to date, their roles in gene regulation and mechanisms of silencing are still not fully understood. In contrast to piRNAs, miRNA sequences and the miRNA pathway show greater conservation among animals. miRNAs are a class of sRNA of 20–23 nt in length important for the regulation of protein-coding genes with diverse functions including the differentiation of larval stages and adult development[17,18]. Complementarity of the miRNA and its target mRNA occurs through the seed sequence found in nucleotides 2–8[19,20]. More than 250 miRNAs have been identified in *C. elegans* where each of them can target more than one mRNA and each mRNA can be targeted by more than one miRNA, increasing the complexity of gene regulation. The siRNAs, in comparison, are approximately 21–27 nt in length, and are important in chromatin regulation, transcriptional regulation, RNA degradation and protein modification[21]. Classes of siRNAs can usually be classified by features such as their sequence length, 5′ starting nucleotide, 5′ end modifications[7], and the specific Argonaute protein they are loaded onto in a siRNA pathway[4]. In *C. elegans*, processing of siRNAs by the enzyme Dicer creates primary siRNA with a 5′ monophosphate (5′pN). The primary siRNAs interact with specific Argonaute proteins, depending on their pathway, and create a complex with RdRPs, that uses the target transcript as a template for synthesis of the secondary siRNA, which are not Dicer-processed and typically have a 5′ triphosphate modification[22]. In contrast to miRNAs that can target many mRNA through their small seed sequence, siRNAs require perfect complementarity to their specific target sites[23].

Here, we have investigated the role of sRNAs in the endogenous regulation of genes and TEs in the nematode *Strongyloides ratti,* a well-established laboratory model of nematode parasitism[24]. *Strongyloides* species are gastrointestinal parasites which infect an estimated 600 million people globally causing chronic morbidity and, more rarely, fatal disseminated strongyloidiasis[25]. They also infect animals causing substantial economic loss in livestock practices[26]. The life cycle of *S. ratti* includes genetically identical free-living (FLF) and parasitic (PF) adult female stages. Direct comparison between these two adult stages can uncover genetic features associated with parasitism including differences in sRNA and TE activity. More interestingly, the two adult stages of *S. ratti* employ distinct reproduction modes. In the parasitic generation, only females exist and reproduce via parthenogenesis, whereas the free-living generation reproduces via sexual mating between males and females. A comparison of these two adult stages is therefore useful to understand how TE dynamics and regulation differ between sexual and asexual reproduction.

The *S. ratti* genome has been sequenced and assembled into a highly contiguous reference genome (two autosomes in single scaffolds and the X-chromosome in ten main scaffolds)[27] which enables an accurate genetic analysis of sRNAs and their targets. We have sequenced sRNAs that are expressed in PF and FLF of *S. ratti*. We then classified the sRNAs into classes or subsets of classes of sRNAs and identified those differentially expressed between the PF and FLF. We identified two classes of sRNAs that are predicted to target TEs, that were differentially expressed between the two adult stages. The sRNAs expressed by the parasitic stage shared multiple features in common with piRNAs including similar length (21–22 nt), a 5′ uracil, a 5′ monophosphate, overlapping loci, physical clustering in the genome and an upstream AU-rich sequence; representing the first set of piRNA-like sRNAs outside of Clade V nematodes. We also identified miRNA families more abundant in the parasitic stage and tRNA fragments expressed specifically in the free-living stage, which indicates that specific sRNAs classes may be directly related to parasitism. Understanding the mechanisms associated with these sRNAs can therefore help us understand parasitism, and has the potential to lead to improved disease diagnostics and treatments.

Results

Strongyloides ratti parasitic and free-living adult females express similar proportions of miRNAs and other sRNAs. sRNA expression in genetically identical PF and FLF *S. ratti* was investigated using two library types; (i) enriched for sRNAs with a 5′ monophosphate (5′pN enriched library), or (ii) RppH-treated to increase the cloning efficiency of 5′ polyphosphorylated and 5′ capped sRNAs (5′ modification-independent library). Reads were classified as either miRNAs, or as sRNAs derived from tRNAs (tRFs), rRNA (rsRNA), or as putative siRNAs originating from either protein-coding genes (including CDS and intronic regions), intergenic region or TEs. The most abundantly expressed class of sRNAs identified in the 5′pN enriched library was miRNAs with lengths of 21–23 nt, which made up 17.4% and 11.0% of total PF and FLF reads, respectively. The sRNAs originating from intergenic regions ranging in length between 21 and 24 nt were the second most highly expressed class of sRNA in both the PF and FLF (3.88% and 1.75% of total reads, respectively) (Fig. 1a,b, Supplementary Data 1). Interestingly, 21–22 nt sRNAs originating from CDS and TEs were expressed at higher levels in the PF than the FLF (10.87% and 6.33% of 21–22 nt reads in PF and FLF, respectively). In contrast, tRFs were more highly expressed in the FLF (0.7% and 1.6% in PF and FLF, respectively) (Fig. 1b, Supplementary Data 1). Overall, there were more unique tRFs expressed in the FLF *cf.* PF (1036 and 185 sequences, respectively). tRFs were significantly more abundant (EdgeR, FDR < 0.01) in the FLF compared with the PF 5′pN-enriched libraries, and primarily originated from the central region of the mature tRNA sequences also known as misc-tRFs (Sup-

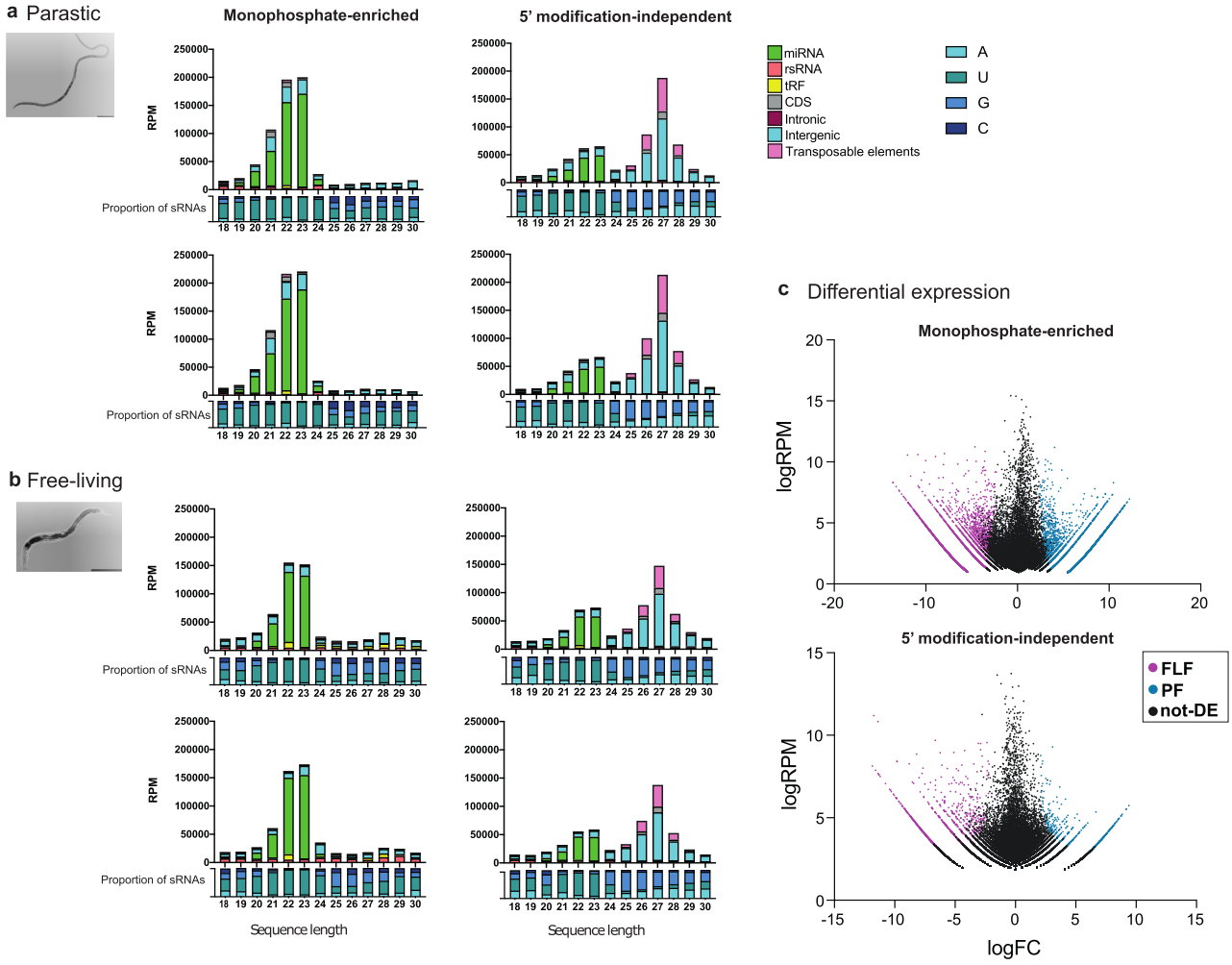

Figure 1. sRNA classification and differential expression. Classification of sRNAs expressed by parasitic female (PF) and free-living female (FLF) stages of *S. ratti* including sRNAs (**a**) enriched for a 5′ monophosphate, or (**b**) 5′ modification-independent sequences which includes sRNAs with a 5′ monophosphate or polyphosphate modifications. Graphs on the top show the classification of sRNAs as either miRNAs, rRNA-derived sRNAS (rsRNA), tRNA-derived sRNAs (tRFs) or as putative siRNAs originating from either protein-coding genes (CDS or intronic regions), intergenic regions or transposable elements (TE). RPM = reads per million. Graphs below the x-axis show the proportion of the first 5′ nucleotide for each length of sRNA. Results from two biological replicates of each condition are shown in the figure. (**c**) Differential expression of sRNAs with a 5′ monophosphate (top), and 5′ modification-independent library (bottom). Differentially expressed sequences are highlighted in pink (FLF-overexpressed) and blue (PF-overexpressed) (FDR of < 0.01, fold change > 2) and sequences that are not differentially expressed are shown in black (logRPM = log reads per million, logFC = log fold change).

plementary Fig. 1). The sRNA sequences expressed in the 5′pN-enriched library predominantly started with a 5′ uracil, consistent with the most common 5′ starting base for miRNAs (Fig. 1a,b).

RppH treatment removes 5′ modifications in sRNA including 5′ triphosphate and other 5′ modifications, thus sRNAs that are observed only in RppH-treated libraries are likely to have 5′ modifications. As expected, we observed similar peaks of 5′pN sRNAs including the miRNAs across both libraries (Fig. 1a,b). In addition, sRNAs between 24 and 30 nt in length were enriched in the 5′ modification-independent libraries indicating that *S. ratti* PF and FLF also express sRNAs with a 5′ modification. The 26–28 nt sRNAs originating from intergenic spaces and TE sequences were the most highly expressed class of sRNA in the 5′ modification-independent libraries for both the PF and FLF. Together, intergenic- and TE-derived sRNAs comprised 65.9% and 41.2% of all sRNAs sequences in the 5′ modification-independent libraries for PF and FLF, respectively, making them the largest set of sRNAs expressed (Fig. 1a,b). Overall, the sRNA expression profiles for FLF and PF in the 5′ modification-independent library were similar i.e. 27 nt sRNAs with a 5′ modification were most highly expressed, followed by 22–23 nt miRNAs with a 5′ monophosphate (Supplementary Data 2). In the 5′ modification-independent library, sRNA sequences between 18 and 23 in length, predominantly started with a 5′ uracil and were classified as miRNAs, and also identified in the 5′pN-enriched library. sRNA sequences between 24 and 30 nt in length predominantly started with either a guanine or adenine at the 5′end (Fig. 1a,b). To identify if specific sRNAs

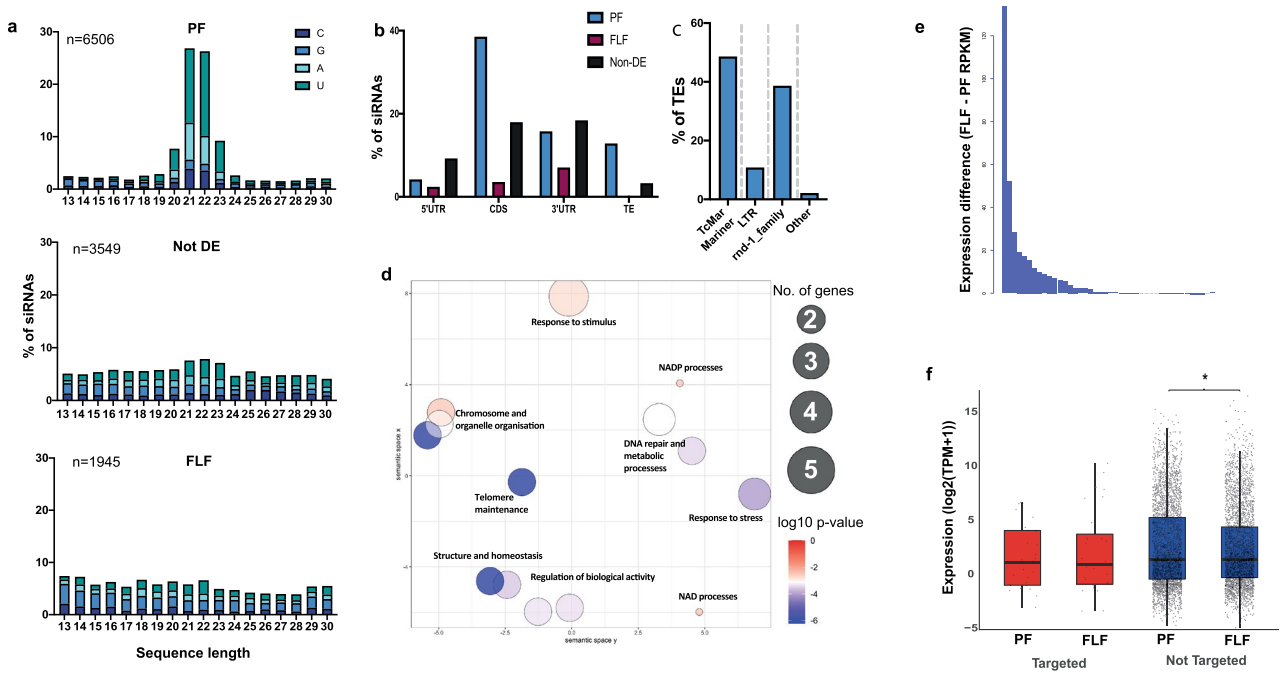

Figure 2. 21–22Us with a 5′ monophosphate associated with parasitism. (**a**) Length distribution and 5′ starting nucleotide of 13–30 nt sRNAs with a 5’ modification originating from either protein-coding genes, intergenic spaces or transposable elements, that are significantly more abundant in the PF (n = 6506 sequences), FLF (n = 1945 sequences) or not DE (n = 3549 sequences). (**b**) Predicted targets of significantly more abundant 21–22Us in the PF, FLF and non-DE based on antisense sequence complementarity. Targets include protein-coding genes (CDS, 5′UTR and 3′UTR) and TE sequences. (**c**) Classification of predicted TE targets of 21–22Us in the PF. Targets include TcMar-Mariner DNA transposon (48.5%), LTR retrotransposons (10.8%) and an unclassified family (38.6%). (**d**) Enriched GO terms based on the biological function (BF) of the 21–22U predicted target genes. Clustering of the GO terms based on similar functions was carried out using REVIGO. Size of circle indicates the number of genes and colour indicates p-value. (**e**) Difference in expression level (RPKM: FLF minus PF) of genes putatively targeted by PF-overexpressed 21–22Us. The targeted genes were more highly expressed in the FLF *cf.* PF stage. (**f**) Expression of TEs either predicted to be targeted or not targeted by PF-overexpressed 21-22Us. *Indicates $p < 0.01$. RPKM = Reads per kilobase million; TPM = Transcripts per kilobase million.

were overexpressed in the PF *cf.* the FLF, a differential expression analysis was carried out for both libraries using edgeR[28]. We found that 22.3% (n = 5584) and 21.9% (n = 5478) of 5′pN sRNA sequences (n = 25,047) were significantly overexpressed in the PF and FLF stages, respectively. In the 5′ modification-independent library, 1.6% (n = 730) and 2.8% (n = 1278) of all sRNA sequences (n = 43,473) were significantly overexpressed in the PF and FLF, respectively (Fig. 1c) (FDR < 0.01, Supplementary Data 3). Together, these results indicate that distinct sets of sRNAs are overexpressed in the PF and the FLF, suggesting they have specific roles in these life cycle stages. Interestingly, some sRNA sequences such as the 21-22Us, described below, were found in the 5′pN enriched library but not the 5′ modification-independent library, highlighting the importance of using multiple library preparation methods to investigate sRNA expression.

21–22U RNAs with a 5′ monophosphate resembling piRNA are associated with the parasitic stage. After miRNAs, sRNAs originating from intergenic spaces, protein-coding genes and TEs were most highly expressed group of sRNAs in the 5′pN-enriched library sequences (Fig. 1a). We further investigated these sRNAs to identify specific classes of sRNAs differentially expressed between the PF and FLF. Analysis of the length and first nucleotide at the 5′ site revealed that 21–22 nt long sRNAs starting with uracil (hereon in referred to as 21–22Us) were the most highly expressed 5′pN sRNA in PF compared with FLF (Fig. 2a). In contrast, sRNA sequences with a 5′ pN either overexpressed in the FLF or not differentially expressed (DE) showed no bias for a particular length or propensity for a particular 5′ base (Fig. 2a). In total, we identified 1887 unique 21–22U sequences overexpressed in the PF, 86 sequences in the FLF and 218 sequences non-differentially expressed (non-DE), respectively. Given the larger number of unique 21–22U sequences and the higher expression levels in the PF, we propose that the 21–22Us are a class of sRNAs with a role in parasitism or a feature associated with the parasitic stage such as parthenogenetic reproduction.

The 21-22Us target TE-associated protein coding-genes and TE sequences. Based on sequence complementarity, we predicted the targets of 21–22U RNAs overexpressed in the PF. Of the 1887 unique 21–22U sequences, 726 showed perfect sequence complementarity to the coding sequence (38.47%), followed by 296 sequences to the

	piRNA (*C. elegans*)	21–22U (*S. ratti*)	27GA (*S. ratti*)	piRNA (*D. melanogaster*)
Length (nt)	21	21–22	27	24–30
5′ base	U	U	G/A	U
5′ modification	Monophosphate	Monophosphate	Polyphosphate?	Monophosphate
Clustered in genome	Yes	Yes	No	Yes
Overlapping loci	Yes	Yes	No	Yes
Upstream motif	Yes	No	No	No
AU rich downstream sequence	Yes	Yes	No	No

Table 1. Comparison of small RNAs putatively-targeting transposons in *S. ratti* with piRNAs in *C. elegans* and *D. melanogaster.*

3′ UTR (15.69%) and 78 sequences to the 5′ UTR (4.13%) regions of 42 *S. ratti* protein-coding genes (Fig. 2b). We also identified 324 21–22U sequences showing perfect complementarity to TEs (17.17%). Of the 42 protein-coding genes, 13 (30.9%) had predicted functions (Supplementary Data 4), that were directly associated with TE activity including DNA helicase, helitron-like proteins, reverse transcriptase and transposase-encoding genes. In addition, 12 putatively-targeted genes (28.6%) were classed as 'hypothetical' and were not annotated with a function (Supplementary Data 4). To characterise the function of the hypothetical proteins, we grouped them into orthofamiles using Orthofinder2[29] with 18 other nematode species (Supplementary Data 5). Interestingly, we found that ten of the twelve hypothetical genes were *S. ratti*-specific. Only two genes were found in other species, one which belongs to the Mos1 transposase family and the second which did not have a known function in related species. The additional 17 putatively-targeted genes had a variety of functions related to ATP-binding, ubiquitin and the general maintenance of the parasitic nematode (Supplementary Data 4). The putative target genes were significantly enriched (Fishers Exact Test FDR < 0.01) for Gene Ontology (GO) terms associated with chromosome and telomere organisation and maintenance, cellular responses to stress and DNA damage and homeostatic processes (Fig. 2d, Supplementary Data 6). We found that the genes putatively targeted by 21–22Us were expressed at higher levels in the FLF compared with PF suggesting that 21–22Us may repress gene expression (Fig. 2e). In addition to the protein-coding genes, ten miRNA precursor sequences and two tRNA genes also showed perfect complementarity, and therefore likely to be targeted by PF-overexpressed 21–22U RNAs (Supplementary Data 4).

Because many of the protein-coding genes predicted to be targeted were associated with TE-activity, we sought to investigate if 21–22Us are involved in directly targeting and regulating the expression of TE sequences. We first improved the annotation of TE sequences within the *S. ratti* genome (see "Methods", Supplementary Figs. 2–4, Supplementary Data 7) and identified TE sequences that are perfectly antisense to 21–22Us. Our results showed that 12.8% of PF-overexpressed 21–22U RNAs targeted the DNA transposon TcMar-Mariner (48.5% of the TEs targeted), followed by an unannotated class of TEs (38.6% of the TEs targeted) and long terminal repeats (LTR) retrotransposons (10.8% of the TEs targeted) (Fig. 2c, Supplementary Data 8). These results suggest that the main role of the PF-overexpressed 21–22Us is the regulation of TE activity (Table 1), mainly of the DNA transposon family. We then examined the expression of the same TEs putatively targeted by 21–22Us and non-targeted TEs in the PF and the FLF. -Our results have shown that the expression level of TEs putatively targeted by PF-upregulated 21–22Us were similar in the PF and FLF (Fig. 2f, Supplementary Data 9). In comparison, the non-targeted TEs had a significant difference in expression between the PF and FLF (Fig. 2f), indicating that the PF-targeted TEs are being regulated to have a similar expression to the FLF.

21–22Us and their TE targets are physically clustered on the X-chromosome. To examine the distribution of the 21–22Us overexpressed in the PF, we mapped the sequences to the *S. ratti* genome. The 21–22U loci clustered on the second largest scaffold of the X-chromosome, forming two main clusters across both strands of the genome spanning 1 and 2.7 Mb, respectively (Fig. 3a, Table 1). In total, 53.27% of the PF-upregulated 21–22Us mapped to this X-chromosome scaffold. For comparison, we also mapped the comparatively small number of 21–22Us sequences upregulated in the FLF and non-DE datasets to the genome, and these showed no evidence of large clustering (Supplementary Fig. 5). The TEs putatively targeted by 21–22Us in the PF were also predominantly located on the X-chromosome, but from different regions (Fig. 3b). This is in contrast to the distribution of TEs in general which are found throughout the genome (Supplementary Fig. 6). Further investigation of 21–22U loci in the genome revealed the sequences originate from overlapping same-strand clusters (found across both strands) in the genome (Fig. 3c). In total, 88.78% of 21–22Us loci overlapped with at least one other 21–22U sequence on the same strand of the genome, and on average each 21–22U sequence overlapped with 13.9 ± 0.28 (mean ± SE) sequences (Fig. 3c), which is similar to the pattern observed for piRNAs in *C. elegans*[30].

Upregulated 21–22Us in the PF have AU rich upstream sequences. We have determined that the 21–22Us expressed by *S. ratti* PF share similar features with *C. elegans* piRNAs including a similar length and first 5′ nucleotide, targeting of TEs and they originate from overlapping sequences in the genome (Table 1). We further investigated *S. ratti* 21–22Us for features that are associated with *C. elegans* piRNAs, namely, an AT rich upstream sequence and a conserved CTGTTTCA motif upstream of the 21U loci found in type I piRNAs[30,31]. We

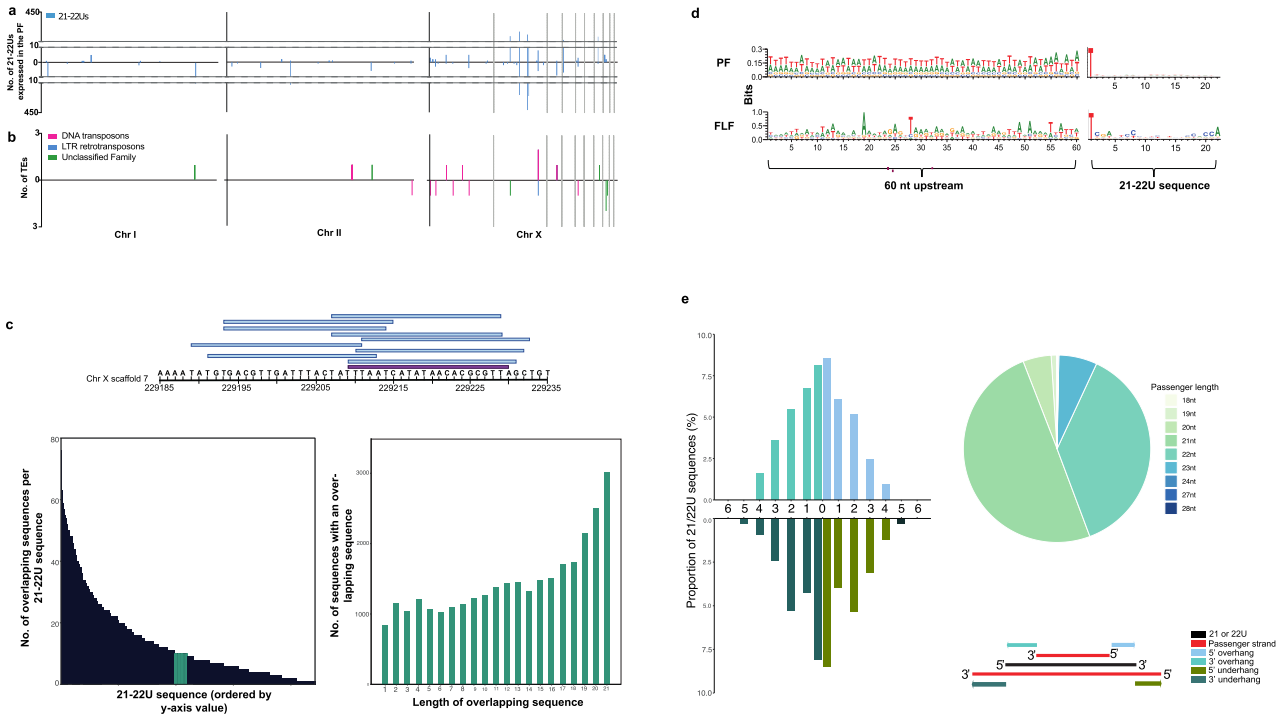

Figure 3. Chromosomal distribution and overlapping patterns of 21-22Us is similar to *C. elegans* piRNAs. (**a**) Distribution of overexpressed 21–22Us across the genome which consists of two autosomes (11.7 Mb chromosomes I and 16.7 Mb chromosome II) and the chromosome X (12.9 Mb made up of 10 scaffolds). (**b**) Distribution of 21–22U predicted TE targets across the genome (n = 20 TE targets). (**c**) Identification of an overlap signature in the 21–22Us. Figure showing an example of a 21U originating from chromosome X and all overlapping 21–22U sequences. Bottom left figure showing the number of 21–22Us that overlap with other 21-22Us (88.78%). Bottom right figure showing the overlap lengths of 21–22Us that overlap with other 21-22U sequences in the genome. (**d**) Sequence logos of PF 21–22U upstream sequences versus the FLF 21–22Us to identify nucleotide richness based on the bits of each nucleotide. (**e**) No dicer-processing signature in PF 21–22Us, based on the predicted overhang of the passenger strand. Passenger strands cleaved by the enzyme Dicer, leave a distinguishable 2–3′ overhang. Pie chart shows the percentage of passenger strands.

identified AT richness in the sequence upstream of *S. ratti* 21–22Us comparable to *C. elegans* piRNAs (Fig. 3d). However, a piRNA-associated motif was not found upstream of the *S. ratti* 21–22U loci.

21–22Us are not Dicer-processed. Classes of sRNA can be characterised by the mechanism used to produce mature sRNA sequences from double stranded RNA precursor sequences. We searched for Dicer-processing signatures in 21–22U sequences, which can be identified by a 3′overhang in sRNA duplexes. However, a Dicer-signature was not observed for 21–22Us and the profile of sRNA duplexes more closely resembled patterns observed for RdRP-processing (Fig. 3e). Furthermore, no evidence of a ping-pong signature was observed for 21–22Us (Table 1), which is usually associated with piRNAs in *D. melanogaster* showing a 10 nt overlap of the 5′ ends of the piRNAs with other sRNA sequences (Supplementary Fig. 7).

Distinct subsets of 27GAs with a 5′ polyphosphate are predicted to target TEs in the PF and FLF. We sought to identify 5′-triphosphated sRNAs (potential RdRP siRNAs) in *S. ratti* by removing the sequences that were present in the 5′pN-enriched libraries from those in the 5′ modification-independent libraries. Expression analysis after the subtraction revealed that the most highly expressed sRNAs in the PF were 27 nt long (hereinafter referred to as 27GAs) (Fig. 4a). A subset of 27GAs was also significantly overexpressed in the FLF (Fig. 4a). In total, we identified 14,292 unique 27GAs including 193 significantly overexpressed in PF and 52 in FLF. Expression of 27GAs was previously reported for mixed-sex free-living adults[7], but here we found that 27GAs are also expressed in the PF, and that distinct subsets of 27GAs are associated with the PF and FLF stages.

27GA RNAs target TE-associated genes. The target sequences of 27GAs were predicted based on antisense complementarity. Of the 193 and 52 27GAs significantly overexpressed in the PF and FLF, the majority of 27GAs were predicted to target the coding sequence (46.63% and 67.31%) followed by the 3′UTR (15.03% and 11.54%) and 5′UTR (10.36% and 9.62%) of protein-coding genes (Fig. 4b). A large number of 27GAs were also predicted to directly target TEs in the PF and FLF (58.03% and 17.31%). A similar pattern was observed for the 14,047 non-DE 27GAs, which were predicted to target the coding sequence (55.64%), followed by the 3′UTR (17.41%), the 5′UTR (8.73%) and TEs (36.61%) (Fig. 4b). In total, the PF-overexpressed 27GAs were predicted to target

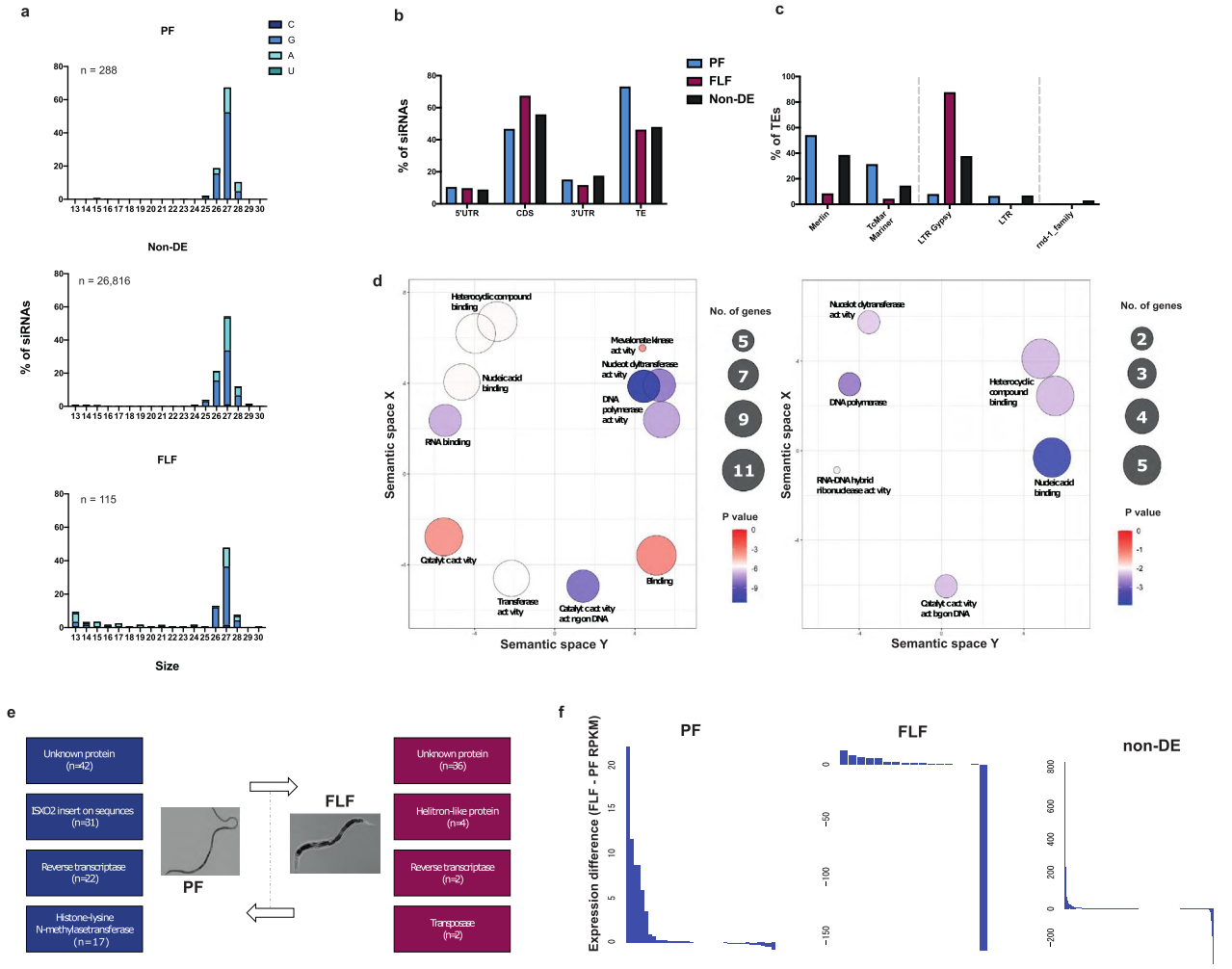

Figure 4:. 27GA 5′-polyphosphate sRNAs overexpressed at the two adult stages. To identify sRNAs with a 5′ modification, sRNA sequences found in a 5pN library were filtered out from sRNAs identified in the 5′ modification-independent library. (**a**) Length distribution and 5′ starting nucleotide of 13–30 nt sRNAs with a 5′ modification, overexpressed in the PF (n = 288 sequences), FLF (n = 115 sequences) or those not DE (n = 26,816), excluding miRNAs, tRFs and rsRNA sequences. 27GAs are the most abundantly expressed sequence. (**b**) Putative targets of the 27GAs in the PF, FLF and non-DE. Target protein-coding genes (CDS, 5′UTR and 3′UTR) and TEs were predicted based on antisense sequence complementarity. (**c**) Classes of TE antisense to PF-overexpressed, FLF-overexpressed and non-DE 27GAs. (**d**) Enriched GO terms of the genes predicted to be targeted by PF-overexpressed 27GAs (left) or FLF-overexpressed 27GAs (right). Clustering of each GO term (biological function) based on similar functions was carried out using REVIGO. Size of circle indicates the number of genes and colour indicates p-value. (**e**) Predicted protein function of putative target genes and the number (n) of 27GAs. (**f**) Difference in expression level (RPKM: FLF minus PF) of genes putatively targeted by PF-overexpressed, FLF-overexpressed and non-DE 27GAs. RPKM = Reads per kilobase million.

40 protein-coding genes, FLF-overexpressed 27GAs targeted 17 protein-coding genes and non-DE 27GAs targeted 452 protein-coding genes. No potential targets were identified for rRNA, tRNA and miRNA precursor sequences. The predicted target genes across all three subsets, PF-overexpressed, FLF-overexpressed and non-DE, were predominantly associated with TE activity including transposase, reverse transcriptase, helicase, helitron integrase-coding genes (Supplementary Data 10). The PF-overexpressed, FLF-overexpressed and non-DE predicted gene targets were enriched for similar Biological Processes (BP) and Molecular Function (MF) GO terms (Supplementary Data 11) including GO terms associated with DNA integration and biosynthesis, RNA binding and DNA polymerase activity (Fig. 4d, Supplementary Data 11).

However, the specific genes predicted to be targeted by either the PF-overexpressed or the FLF-ocerexpressed 27GAs were different. PF-overexpressed 27GAs were predicted to target genes coding for transposase insertion sequence XO2 (31 sRNAs for11 genes, compared to only one gene by FLF- overexpressed 27GAs) known to mediate transposition, and reverse transcriptase related genes (22 sRNAs for 7 genes) (Fig. 4e). A large proportion of genes presumably targeted by the PF-overexpressed and FLF-overexpressed 27GAs coded for 'hypothetical' proteins (PF: 42 27GAs for 13 genes, FLF: 36 27GAs for 7 genes; Supplementary Data 10). The PF-overexpressed

27GAs are predicted to target genes belonging to two orthofamilies. Eight genes belong to an orthofamily comprising genes encoding ISXO2 transposase family and three genes belong to an orthofamliy comprising Mos1 transposase genes, both of which are related to TE activity. The FL-overexpressed 27GAs putatively targeted one gene belonging to an orthofamily coding for genes related to nucleic acid binding activity (Supplementary Data 5). Together, these results suggest the role of 27GAs in TE regulation, and that specific subclasses of TEs are targeted at the two different adult stages. A notable difference between the sets of target genes is that the PF-overexpressed, but not FLF-overexpressed 27GAs, target four histone-lysine N methyltransferase-coding genes, all located on the X-chromosome. Expression level of genes putatively targeted by PF-presumable 27GAs were lower in the PF compared with FLF. The non-DE 27GA-putatively-targeted genes were expressed at similar levels in the PF and FLF (Fig. 4f).

27GAs target TE sequences. The 27GAs previously identified in the FLF were predicted to target TE sequences[7]. Our analysis above further suggests that 27GAs expressed in the PF and FLF are targeting protein-coding genes associated with TE activity. Here, we also investigated if PF-expressed 27GAs directly target TE sequences (Fig. 4b, Supplementary Data 12). We identified 27GAs that aligned perfectly antisense to class I and class II TEs in the *S. ratti* genome and found that a distinct set of TEs were presumably targeted by the PF and FLF. Our results showed that 73.06% (n = 141) of PF-overexpressed 27GAs were predicted to target 42 TE sequences, many of which were DNA transposons. Of the PF-overexpressed 27GAs for TEs, 53.9% were antisense to the DNA transposon from the Merlin family and a further 31.2% putatively targeted TcMar-Mariner. In addition, 7.8% and 6.4% putatively targeted the retrotransposons LTR gypsy family and LTRs of unannotated families, respectively (Fig. 4c). In comparison, 46.15% (n = 24) of the FL-overexpressed 27GAs and 47.83% (n = 6719) of the non-DE 27GAs were also predicted to target TEs, but the overall proportion of these sequences that were predicted to target TEs was lower compared to PF-overexpressed TEs. Unlike the 27GAs overexpressed in the PF, FLF-overexpressed 27GAs predicted targets were predominantly the class I retrotransposons LTR gypsy (87.5% of the TEs targeted), followed by the DNA transposons Merlin (8.3% of the TEs targeted) and TcMar-Mariner (4.2% of the TEs targeted). In the non-DE 27GAs several TE families were predicted to be targeted including Merlin (38.4%), LTR gypsy (37.5%), TcMar-Mariner (14.4%), LTRs belonging to an unannotated subfamily (6.6%), unclassified TE families (2.8%) and LTR copia (0.08%) (Fig. 4c). Together, these results indicate that 27GAs in the PF are important in targeting, and presumably regulating, the activity of TEs within the DNA transposon family, in comparison to the FL-overexpressed and non-DE 27GAs which also target and regulate retrotransposons (Table 1).

To investigate the role of TEs further, we compared the expression of TE sequences that were predicted to be targeted vs. not targeted by the different subsets of 27GAs (PF-overexpressed, FLF-overexpressed and non-DE) (Supplementary Fig. 9). We found that the TE sequences that are predicted to be targeted by either PF-overexpressed or FLF-overexpressed 27GAs were expressed at similar levels to non-targeted TEs (Supplementary Data 9). The TEs predicted to be targeted by non-DE 27GAs were expressed at lower levels compared to the non-targeted TEs. A significant difference was found in the TE expression level between PF and FLF for predicted TE targets. However, the difference in expression level was not clearly directional and included TEs that were expressed at either higher or lower levels in the PF (Supplementary Fig. 9), demonstrating the diversity in TE expression levels, and presumably their regulation, in the two adult stages.

27GAs are clustered within the X-chromosome. We investigated the genomic distribution of 27GAs by mapping the 27GAs sequences to the *S. ratti* genome. Similar to the 21–22Us, the 27GAs were predominantly located on the X-chromosome (Fig. 5). However, unlike 21–22U RNAs that were clustered in one particular region of the X-chromosome (Fig. 2d), the distribution of the 27GAs spanned across most of the X-chromosome scaffolds (Fig. 5). Of the PF-overexpressed 27GAs, 21.24% were located on the 3rd largest X-chromosome scaffold, followed by 12.95% on the second largest X-chromosome. In comparison, the largest cluster of 15.38% of FLF-overexpressed 27GAs were located on the opposite strand of the 6th largest scaffold of the X chromosome spanning the first 100 kbp. The non-DE 27GAs were found throughout the X-chromosome as well as chromosome I and II (Fig. 5).

Subsets of miRNAs are differentially expressed by PF and FLF stages. Using Mirdeep2[32], we identified a total of 158 miRNAs, including 103 and 94 miRNAs expressed across both replicates of either the PF or FLF stage, respectively. The majority of miRNAs we identified had previously been reported in miRBase v22[33], however, we identified four novel miRNAs (Supplementary Data 13). Then, we identified miRNA sequences that were differentially expressed between the PF or FLF stages. Nine and six miRNA sequences were significantly overexpressed in the PF and FLF stages, respectively (edgeR FDR < 0.01, Fig. 6a, Supplementary Data 13).

We categorised the *S. ratti* miRNAs into families based on their seed sequences and identified a total of 92 seed families (Supplementary Data 14). Comparison of the expression levels based on seed revealed that 23 and 17 seed sequences were significantly overexpressed in PF and FLF, respectively (Fig. 6b, Supplementary Data 14). The seed families that were differentially expressed comprised between 1 and 21 miRNA sequences and included both miRNA families conserved across other species and uncharacterised seed families. The miRNAs in the seed family with the most members (seed sequence UUGCGAC) were predominantly overexpressed in the PF and may therefore target a specific set of mRNAs important in parasitism. The UUGCGAC miRNA family was not found in seven other nematode species where data was available on miRBase (*Ascaris suum, Brugia malayi, C. elegans, Haemonchus contortus, Heligmosomoides polygyrus, Pristionchus pacificus* and *Panagrellus redivivus*) indicating that it is likely to be a *Strongyloides*-specific family.

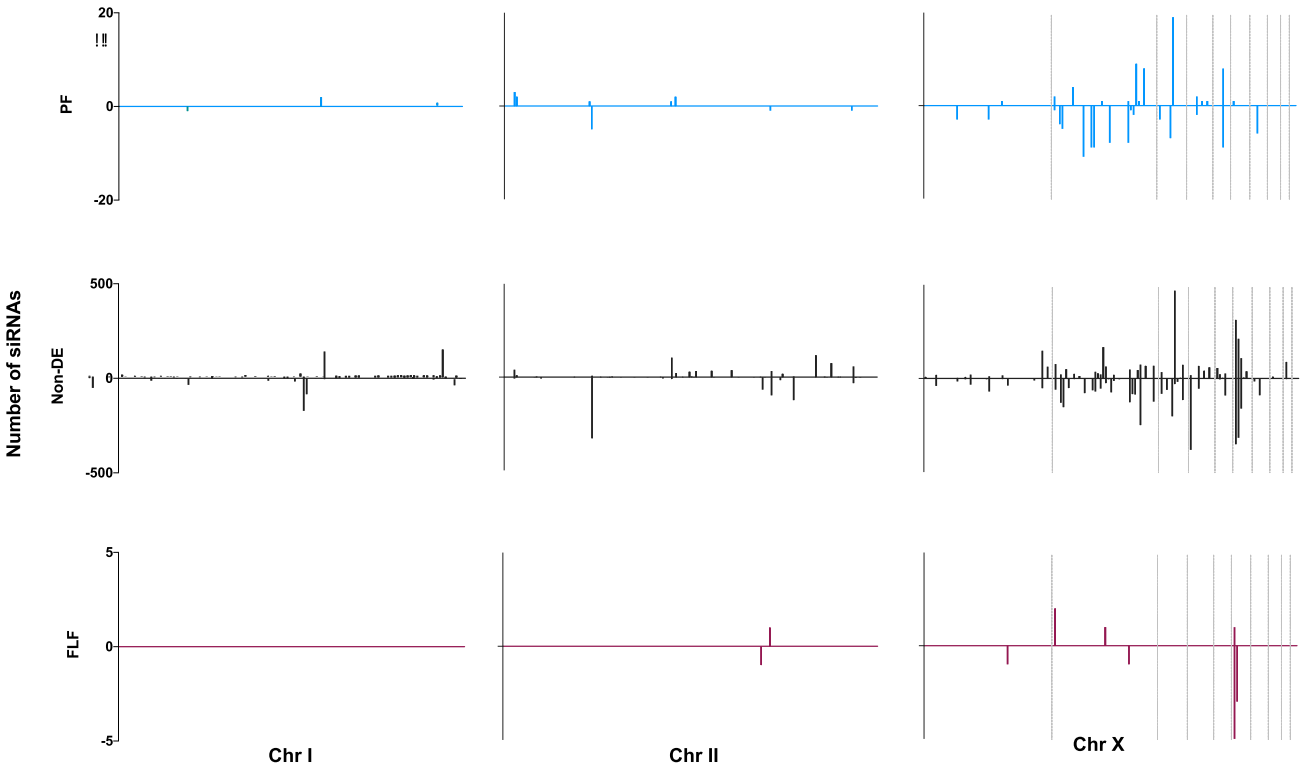

Figure 5. Chromosomal distribution of overexpressed 27GAs in the PF (n = 193 sequences), FLF (n = 52 sequences) and non-DE (n = 14,047 sequences) within I (11.7 Mb), II (16.7 Mb) and the X chromosome (13 Mb), identified using sequence complementarity.

Discussion

We have investigated sRNAs in the gastrointestinal parasite *S. ratti* by directly comparing sRNAs expressed in genetically identical PF and FLF stages. We have identified two distinct classes of sRNAs, characteristic of siRNAs, predicted to target TEs including a (i) piRNA-like 21–22Us with a 5′pN, highly associated with the PF stage, and (ii) 27GAs with a 5′ modification, which have distinct subsets of sequences overexpressed in the PF and FLF. We have also identified miRNA sequences and miRNA families based on their seed sequence, that are differentially expressed in the PF and FLF. We propose that sRNAs expressed at higher levels in the PF are either directly related to parasitism or related to a feature associated with the parasitic generation like parthenogenetic reproduction.

Parasitism-associated 21–22U sRNAs are predicted to target TEs and resemble piRNAs. We identified 1887 unique 21–22Us significantly overexpressed in the PF. We postulate that these 21–22Us are particularly related to the *S. ratti* adult parasitic stage because high expression levels of this class of sRNAs were not observed in FLFs in this study or in other life cycle stages previously investigated[7]. We predicted the sequences that were targeted by 21–22Us based on perfect antisense complementarity. Collectively, our results strongly support that 21–22Us are targeting TEs. sRNAs that regulate TEs are usually most highly expressed in the germline in *C. elegans* and other animals[9,14], and it is therefore likely that the TEs and 21–22Us expressed here are from the PF germline cells. Interestingly, the TEs presumably targeted by PF-overexpressed 21–22Us were expressed at similar levels in the PF and FLF and compared to TEs not targeted by 21–22Us. If we assume, based on evidence in other animals, that the expression of TEs targeted by sRNAs are repressed[34], then it is likely that 21–22Us could be acting to repress the expression of a subset of highly expressed TEs back to the 'normal' levels observed in FLF and non-targeted TEs. It is important to note here that the analysis of TE transcript activity was based on polyA-selected RNAseq data and therefore is only informative about polyadenylated TEs e.g. retrotransposons with a polyA sequence and the transposase component of DNA transposons.

In *C. elegans, D. melanogaster* and *M. musculus,* piRNAs have a key role in regulating and silencing TE activity[6,7,35]. Given the similarity in size (21–22 nt), 5′ nt (uracil), 5′ monophosphate, no Dicer-processing signature and their predicted targeting of TEs, we investigated if other features associated with piRNAs were also common to *S. ratti* 21–22Us. The piRNAs found in *C. elegans* originate from large genomic clusters that give rise to short 21U piRNAs. In *C. elegans,* piRNAs can be further divided into two groups: type I 21U RNAs that make up 95% of total piRNAs and the less abundant type II 21U RNAs[31,35,36]. Type I piRNAs are transcribed from thousands of AT rich loci that accumulate within two large clusters on chromosome IV and have a conserved upstream 'CTGTTTCA' Ruby motif[31,37]. These clusters mainly originate from introns and intergenic regions that overlap with other 21Us[30,31]. In contrast, Type II 21Us are distributed throughout the genome and have no upstream motif[30,36]. *S. ratti* 21–22Us resemble *C. elegans* type I piRNAs, because they (i) are clustered in a specific

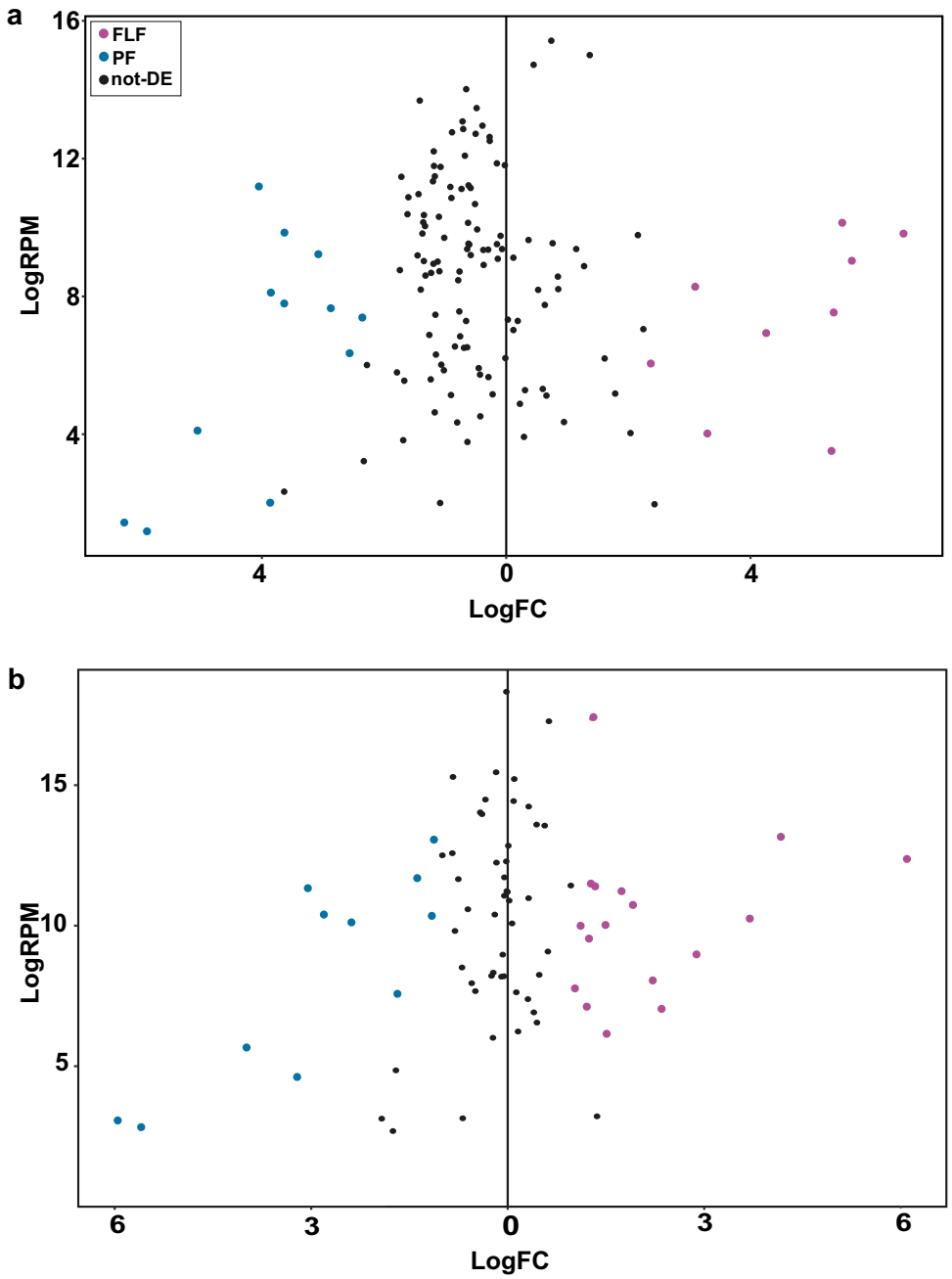

Figure 6. Differential expression of (**a**) miRNA sequences showing miRNAs significantly overexpressed in the PF (n = 12 sequences), FLF (n = 9 sequences) (FDR < 0.01) and (**b**) seed sequences overexpressed in the PF (n = 23) and FLF (n = 17). miRNAs with the seed sequence UUGCGAC were predominately overexpressed in the PF.

region of the genome, though they are located on the X-chromosome rather than an autosome, (ii) have an AT rich upstream sequence, (iii) share a similar, same-strand overlapping pattern in the genome for the 21–22U loci. Like *C. elegans* piRNAs, the number of nucleotides overlapped are varied, but most commonly, 21–22Us overlapped by 20–21 nucleotides with other 21–22Us. Interestingly, the only validated target of the *C. elegans* piRNA pathway is the DNA transposon family, Tc3[13]. We have shown that *S. ratti* 21-22Us also predominantly target DNA transposons. Together our data suggest that *S. ratti* 21–22Us may be the equivalent to the *C. elegans* piRNAs and processed through a similar manner.

An intriguing difference between *S. ratti* 21–22Us and *C. elegans* 21Us, is that *S. ratti* 21–22Us share characteristics associated with piRNAs in *D. melanogaster* not found in *C. elegans*. *D. melanogaster* piRNAs are transcribed from piRNA clusters made up of transposons or repeat elements and are characterised as repeat associated siRNAs (rasiRNAs)[15,38,39]. These rasiRNAs highly target transposons through complementary base

pairing and are required for the normal development of the germline[16,40]. A similar pattern was observed in *S. ratti* 21–22Us. Although we have shown that 21–22Us originate from clusters found on the X-chromosome, some of these clusters are found in intergenic regions, however, like *D. melanogaster,* 21–22Us also originate from TEs, as well as predominately targeting TEs clustered on the X-chromosome through perfect antisense complementarity. This indicates that 21–22Us may be derived from TE regions that are required to be regulated and silenced, similar to the piRNAs of *D. melanogaster.*

Collectively, our results suggest that *S. ratti* 21–22Us share many similarities with the piRNA class of sRNAs, a pathway which is assumed to have been lost in nematodes outside of clade V[5]. It is possible that the 21–22Us may have originated from an ancestral PIWI pathway which has now lost key components such as the PIWI Argonautes, and the pathway has subsequently diverged in *S. ratti*. However, the lack of an upstream Ruby motif that is found for *C. elegans* piRNAs or the ping pong signature in *D. melanogaster* suggests that this is either not the case or the 21–22Us derived from piRNA substantially a long time ago. Another possibility is that many of the features common to piRNA and 21–22Us are universally advantageous to TE targeting by sRNAs and they have evolved independently. Furthermore, a lack of PIWI Argonaute-coding genes in *S. ratti* suggests that an alternative pathway is associated with 21–22Us. Indeed, an alternative class of sRNA for silencing TEs has been proposed for other nematodes lacking the PIWI pathway[5]. However, studies on the alternative classes of TE-targeting sRNAs in nematodes have not reported similarities with piRNAs, as is the case for the 21–22Us in this study. Previously, we have identified a group of Argonaute-coding genes closely related to *C. elegans* WAGOs that are significantly upregulated in the PF compared with the FLF stages in four *Strongyloides* species investigated[6], and we speculate that based on their expression patterns these could be associated with 21–22Us in *S. ratti*. Further work is required to identify if these Argonaute proteins are associated with 21–22Us and to identify other components of this pathway.

27GAs are predicted to regulate TEs in both the parasitic and free-living adult stages. We identified a second class of sRNAs, the 27GAs, that also putatively target and regulate the expression of TEs. In contrast to the 21–22Us, 27GAs were expressed across both stages, and were only observed in the 5′ modification-independent library, suggesting that they possess a polyphosphate or capped 5′ modification. The 27GAs have previously been described in adult free-living *S. ratti* worms (mixed female and male)[7] but this is the first time they have been identified in the PF. In addition to the 27GAs that were expressed at similar levels in both stages, smaller subsets of 27GAs were overexpressed in the PF or FLF. Although 27GAs in both the PF and FLF were predicted to target TEs, there was a clear difference in the classes of TEs targeted. The PF-overexpressed 27GAs were predicted to be mainly targeted TEs from the class II DNA transposons, namely Merlin and TcMar-Mariner, whereas the FLF-overexpressed 27GAs were predicted to target predominantly the class I retrotransposon LTR gypsy. However, it is not clear why there is a difference between the type of TEs targeted in the PF and FLF. Protein-coding genes presumably targeted by the 27GAs were associated predominantly with TE activity but some are with other biological processes. For example, the PF-overexpressed 27GAs were predicted to target histone-lysine-*n*-methyltransferase-coding genes. These proteins are involved in histone modification and play a role in chromatin structure and gene expression[41]. Inhibiting histone methyltransferase has been shown to stop the life cycle of *Schistosoma mansoni*, suggesting that the role of histone methylation is extremely important in parasites where differentiation of the life cycle occurs within a host[41]. There could therefore be a specific role of 27GAs in regulating histone modification that is specific to the PF stage.

Unlike the 21–22Us, the 27GAs share few similarities with piRNAs. Instead the 27GAs are more similar to the secondary 22G siRNAs in *C. elegans* which have a 5′ triphosphate, a 5′ guanine (5′G) bias and are processed by RdRP[42,43]. The 27GAs, showed no evidence for Dicer-processing further supporting that they are likely to be processed in a similar manner to *C. elegans* 22Gs RdRPs.

sRNAs similar to secondary 22Gs in *C. elegans* have also been reported in other nematodes. For example, the clade I–III nematodes produce 22G sRNAs processed by RdRPs, which mainly target TEs in the absence of piRNAs[5,44]. RdRP orthologues related to the processing of 22G siRNAs in *C. elegans* are also present in *Strongyloides*[7], indicating that the RdRP pathway is active.

Why is TE regulation important in the parasitic stage? Overall, our results have shown that expression of sRNAs that are predicted to target TEs in *S. ratti* are expressed at higher levels in the PF *cf.* FLF, including the 21–22Us and 27GAs. This suggests that the regulation of TE activity is higher in the PF, raising the question, *why is the regulation of TE higher in the PF stage?* It is widely accepted that TE activity is often associated with the germline[45,46]. Most TE insertions are considered to be detrimental to the genome integrity and can lead to disruption of genes or regulatory regions[11]. However, TE activity may also be beneficial and can be related to genome rearrangement, regulation of genes and chromosome stability[11,47]. For example, LINE-like retrotransposons are highly important and required for the maintenance of telomeres in *Drosophila,* where the telomerase enzyme is missing[48]. From our study, it is unclear if the higher level of TE regulation observed in PF is beneficial to the *S. ratti* genome and this requires further investigation. The increase in TE regulation in the PF could reflect the stressful environment that *S. ratti* is exposed to in the rat host, and the increased sRNA activity is a direct response to regulate the activity of increased TE activity. It is also possible that differences in expression level between PF and FLF represent differences in the proportion of germline cells. For example, if PFs have a larger proportion of germline cells compared with FLF, this could be observed as higher level of expression of germline-related transcripts e.g. TEs or sRNAs. We observed DAPI-stained worms and counted nuclei in germline and somatic cells in a whole body of PF and FLF and found that the proportion of germline cells were larger in PF than in FLF (Supplementary Data 16). While this may partially explain the differences in expression levels, it does not appear to be a reasonable explanation for the predominance of a particular sRNA at either stage.

In addition to differences in lifestyle i.e. parasitic stages inside the host vs. free-living stages outside of the host, PF and FLF also differ in their reproductive strategy; the PF reproduces through mitotic parthenogenesis and the FLF reproduces through sexual reproduction[49]. The differences observed between TE activity and the subsequent TE regulation by sRNAs could reflect differences in embryogenesis and development. Previous studies have suggested that dynamics of TEs differ between sexual and asexual organisms[50,51]. In sexual reproduction, TEs can proliferate and jump from one genomic background to another as well as remove detrimental TE insertions. In asexually reproducing organisms, on the other hand, it is more difficult to remove detrimental TEs and TEs can persist to be active across generations without the presence of regulators[50]. Therefore, the parasite-associated class of 21–22Us and 27GAs in *S. ratti* possibly have a role in regulating TEs throughout the asexual reproduction to prevent TEs that may be deleterious for the genome[51].

TEs have also been associated with a role in cis regulation of genes involved in sexual development[51]. The 21–22Us, 27GAs and the TEs that they likely targeted are predominantly located on the X-chromosome (this is not the case for TEs in general). TEs are often associated with sex chromosomes and in some cases these TEs have a role in regulating sex-chromosomal genes[51]. With the results presented in this study and the presence of two genetically identical adult stages with different reproductive strategies, *S. ratti* offers a particularly interesting study platform to investigate the role of TEs and sRNAs in reproduction, sexual development and the regulation of sex-chromosomal genes. Lastly, TEs are fast evolving sequences subject to higher mutation rates compared with other sequences in the genome. It is therefore important that a regulatory system can respond and adapt rapidly to variation in these sequences to accurately regulate active TEs. In this study we have looked at a single time point in the PF and FLF stages. The 27GAs more abundant in the PF or FLF represent a response to the TEs that are upregulated in that snapshot of time. However, the TEs that are active in the genome at any one time may vary and the 27GAs would subsequently vary in response to these TE sequences. Interestingly there are clearly distinct trends between the TEs and TE-associated genes presumably targeted by the PF and FLF and this could mean that particular classes or TE are more likely to be active in either stage.

Other classes of sRNA are differentially expressed. We identified a distinct set of miRNAs and miRNA families differentially expressed in the PF and FLF. Unlike siRNAs which require perfect or near-perfect complementarity to their target, miRNAs have just a small seed sequence of ~ 7 nucleotides and bioinformatic prediction of miRNA targets is thus prone to false positives and was therefore not addressed here. We have previously identified protein-coding genes upregulated in the PF *cf.* FLF stage that have a putative role in parasitism including genes that are physically clustered in the genome[6,27,52]. The differentially expressed miRNAs may be involved in regulating these 'parasitism genes' and further lab-based approaches are required to investigate this.

We also observed a difference in the expression levels of sRNAs originating from mature tRNAs known as tRNA derived sRNA fragments (tRF)[53,54]. Mature tRNAs can be cleaved to produce several classes of tRF classified as 5′ and 3′ tRFs, 5′ and 3′ tR-halves, 3′ CCA-tRF and internal-tRF. We analysed tRFs found in the PF and FLF, and found that there is an increased expression level of internal-tRFs in the FLF. The tRFs have important biological roles in the translation of genes as well as gene regulation through the interaction with proteins and mRNAs[55]. Their expression has been shown to be upregulated during stress and starvation in *Trypanosoma brucei*[56], which suggests that increased expression in the FLF may be due to the stressful conditions of the environment. tRFs have not been well characterised and relatively little is known about this class of sRNA. More work in this area is required to elucidate the role of tRFs. The library preparation methods used here enrich for RNA molecules with a 5′ monophosphate which are likely to represent 'true' sRNAs, however, we cannot rule out that some of the reads represent degraded products of tRNA or other longer RNA molecules.

Conclusion

This is the first report of sRNAs expressed in the PF stage of *S. ratti*. Most parasitic nematodes do not have genetically identical parasitic and free-living adult stages, and *S. ratti* therefore offers an almost unique opportunity to identify sRNAs specifically associated with parasitism, and to investigate sRNA-mediated targeting of TEs in a parasitic nematode. We directly compared the sRNAs expressed in the PF and FLF and key findings from this work are an identification of a novel family of 21–22U piRNA-like sRNAs in the parasitic stage of *Strongyloides* and differential expression of 27GAs in the two adult stages. These putative siRNAs originate from the X-chromosome and were predicted to target X-chromosome associated TEs. TE-targeting sRNAs were particularly evident in the PF, suggesting increased levels of TE regulatory activity associated with the parasitic stage.

Materials and methods

Collection of *S. ratti* and sequencing. *Animals.* Wistar male rats aged 4–6 weeks were used to maintain *S. ratti* (strain ED321) by serial passage injections of 1000 iL3 prepared by faecal culture (in 23 °C for 5 days) as described by Hino et al.[57]. All animal experiments in this study were performed under the applicable laws and guidelines for the care and use of laboratory animals, as specified in the Fundamental Guidelines for Proper Conduct of Animal Experiment and Related Activities in Academic Research Institutions under the jurisdiction of the Ministry of Education, Culture, Sports, Science and Technology, Japan, 2006. All studies are reported in accordance with the ARRIVE guidelines (https://arriveguidelines.org/).

Collection of PF and FLF stages. Approximately 3000 iL3s in PBS were injected subcutaneously into rats. To collect PF, rats were sacrificed on 6 days post infection (dpi) and small intestines were collected. The small intestines were developed longitudinally, washed twice with prewarmed (37 °C) PBS and incubated in PBS at 37 °C for 2 h to isolate PF. PF were washed with PBS before proceeding RNA extraction. To obtain FLF, faeces collected from infected animals (8–10 dpi) was incubated at 23 °C for 3 days using 2% agar plates. PF and FLF were transferred

individually to a new tube containing TRIzol (ThermoFisher) using a needle picker after quick wash by PBS. The worm samples were snap frozen in liquid nitrogen and stored at − 80 °C until required.

Extraction of RNA and library preparation. Total RNA was extracted from ~ 100 PF or ~ 50 FLF using TRIzol, according to the manufacturer's instructions (ThermoFisher). Small RNA libraries were constructed from 50 ng of total RNA using QIAseq miRNA Library Kit (Qiagen) according to manufacturer's instruction. For the 5′ independent-phosphate library construction, the QIAseq miRNA Library construction protocol was modified to include RNA 5′ Pyrophosphohydrolase treatment (RppH) to remove 5′ phosphates. Briefly, 50 ng of total RNA was processed up to 3′ adapter ligation step according to the manufacturer's instruction (Qiagen). The adapter-ligated RNA was treated with 5U of RppH (New England Biolabs) at 37 °C for 30 min followed by heat inactivation at 65 °C for 20 min, incubated at 4 °C for 5 min and proceed immediately to the 5′ ligation protocol. Sequencing was carried out using Illumina MiSeq with MiSeq reagent kit v3-150 (Illumina). A summary of samples sequenced and total read counts for each sample is summarised in Supplementary Data 15.

Identification and analysis of sRNAs. *Processing of raw reads.* Fastq files were trimmed to remove adaptor sequences using umi-tools[58].

Mapping of reads to the genome. The *S. ratti* genome (bioproject PRJEB125) was downloaded from WormBase ParaSite and trimmed reads were mapped to the *S. ratti* genome with BBtools[59] using default settings.

Sequence length distribution. Sequences were filtered based on their length using the Next generation library (NGS) Toolbox perl script TBr2_length-filter.pl[60].

Identification of microRNA sequences. miRDeep2[32] was used to identify known and novel miRNA sequences. Trimmed read data from all replicates and samples generated from the monophosphate-enriched library was combined and run with miRDeep2 to identify miRNAs using default settings and using known precursor and mature reference *S. ratti* sequences as references downloaded from the miRBase database (106 precursor sequences and 208 mature sequences based on Refs.[7,61]). Mature miRNA sequences available in the miRBase database (Release 21) for ten other nematode species (*Ascaris suum, Brugia malayi, Caenorhabditis brenneri, Caenorhabditis briggsae, Caenorhabditis elegans, Caenorhabditis remanei*, *Haemonchus contortus, Heligmosomoides polygyrus, Pristionchus pacificus, Pangarellus redivivus*) used as input for mature sequences from related species. miRNA sequences were quantified using the quantifier.pl script from miRDeep2. An estimated signal-to-noise value of 10 was used as a cut-off value and only miRNAs scoring above this threshold were used for further analysis. The resulting list of miRNA sequences was used for classification of miRNA sequences, described below.

Annotation of TE sequences. To identify the TE sequences, we constructed de-novo repeat library for *S. ratti* using RepeatModeler2 and then passed the library RepeatMasker 4.1.1[62]. LTR retrotransposons were further annotated by LTR_harvest[63] and LTR_digest[64].

mRNA and TE expression. mRNAseq expression data for the PF and FLF was obtained from Hunt et al.[27]. To assess TE expression, RNAseq reads were aligned to the *S. ratti* genome using STAR[65] with options (–outFilterMultiMapNmax 10 –winAnchorMultiMapNmax 50). TEcounts[66] then ran in *multi* mode to generate counts for TEs while optimising for multimapping events. Differentially expressed TEs were identified using edgeR with thresholds of count per million (CPM) > 1 in at least 2 samples, FDR < 0.01 and fold change > 2.

Classification of sRNA sequences. The Unitas script (version 1.7.0)[67] was customised to use for a non-model organism (using the – species × parameter and custom reference databases), to classify sRNA sequences. The protein-coding genes (CDS and introns were separated), rRNA and tRNA sequences for *S. ratti* (downloaded from WormBase version WS277), TE sequences and *S. ratti* miRNA sequences (see above) were used as reference databases (using the option -refseq). Sequences were first filtered to remove low complexity reads and were then identified as miRNAs based on mature and precursor reference sequences. Sequences that were not identified as miRNAs were aligned to other reference categories. At this step if a sequence aligns in more than one reference category, they were labelled as multi-mapped. Sequences that were not classified into any of the categories outlined above were assigned as sRNAs from intergenic regions. First 5′ nucleotide data for sRNA sequences was also obtained from Unitas.

Classification of sRNAs by length and 5′ nucleotide. The NGS_Toolbox perl script TBr2_length-filter.pl[60] was used to separate sRNA reads by size[68]. To ensure that there were no repeated sequence reads between the two libraries, sequences that were found in both the 5′ pN and 5′ modification-independent libraries were removed from the 5′ modification-independent data to generate a set of sequences that had a unique 5′ polyphosphate or capped modification. Seqkit (version 0.13.2)[69] was used to find duplicated sequences between the fasta files.

Differential expression of sRNA. The edgeR (Bioconductor version 3.11) R package (version 4.0.2)[28] was used to carry out a differential expression analysis. Only reads with more than 2 CPM in at least two samples were retained. Significant values included only those differentially expressed sRNAs with an FDR of < 0.01 and a fold change of > 2.

sRNA clusters on chromosomes. To identify clusters of sRNAs in the genome, sRNAs were mapped to the genome using Bowtie2 version 2.4.1[70]. The sam output file was converted to bam using Samtools (version 2.1)[71] followed by Bedtools (version 2.28.0)[72] to obtain a bed file.

Target site prediction. To predict potential target sequences of putative siRNAs, only reads that were found in both replicates were used. The reverse complement of sRNAs was first identified using seqkit[69]. Sequences were then mapped to the *S. ratti* coding sequence (CDS), 250 nt upstream (predicted 5′UTR region) and 500 nt downstream (3′UTR) from the CDS sequence obtained from WormBase ParaSite, using Bowtie2 (version 2.4.1), allowing for up to one mismatches *-N 0* and *-norc* to prevent alignment of the reverse complement of the sequence. Reads were also mapped using the same method to TE sequences.

Protein family prediction. Predicted function of the proteins that genes coded to was updated using BLAST[73] and InterProScan[74] (Supplementary Data 16).

Gene ontology. topGO (version 2.42.0)[75] on R was used for GO enrichment analysis**.** GO terms associated with each gene were obtained from WormBase parasite. REVIGO[76] was used to cluster GO terms according to their GO function and the REVIGO output script was customised to also report the number of sRNAs that targeted each GO term.

Sequence logo. To identify the nucleotide richness and the presence of any conserved motifs, WebLogo[77] was used to find sequence motif logos. To identify upstream sequence motifs, 21–22Us were first mapped to the *S. ratti* reference genome using bowtie2[70]. Flanking sequences were extracted using Bedtools.

Dicer-signature analysis. The stepRNA tool (version 1.0.3) (https://pypi.org/project/stepRNA/) was used to search for Dicer-processing signatures using default settings. Identical sRNA sequences were collapsed and used as input.

Same strand analysis. sRNA sequences were aligned to the genome using Bowtie2 with multimapping enabled (-f -a -N 0 –no-1 mm-upfront –norc). Identification of overlapping sequences was achieved with Bedtools intersect[72] using *-s* and *-wo* parameters.

Statistical analyses. All statistical analyses were carried out in R studio version 3.6.3[78].

Microscopy. Worms were fixed in an ice-cold fixation buffer (3 ml glacial acetic acid and 100 ml methanol) for 5 min, washed three times by SSCT (0.3 M sodium chloride-0.03 M sodium citrate with 0.2% Tween-20). Specimen of worms were then transferred to glass-bottom dishes with 5 µg/ml DAPI in Vectashield (Vector Laboratories) and imaged on BZ-X800 microscopic system (KYENCE) with Z-stacks (per 0.5 µm) to acquire the whole worm at different overlapping field-of-views. We extracted the individual optical sections of each Z-stack as full resolution TIFF files and nuclei count was performed using ImageJ software[79].

Data availability

All sequence data from the genome projects have been deposited at DDBJ/ENA/GenBank under BioProject accession PRJDB13088. All relevant data are available from the authors.

Received: 15 March 2022; Accepted: 3 June 2022
Published online: 16 June 2022

References

1. Zhang, C. Novel functions for small RNA molecules. *Curr. Opin. Mol. Ther.* **11**, 641 (2009).
2. Britton, C., Laing, R. & Devaney, E. Small RNAs in parasitic nematodes—Forms and functions. *Parasitology* https://doi.org/10.1017/S0031182019001689 (2020).
3. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* https://doi.org/10.1016/j.cell.2004.12.035 (2005).
4. Billi, A. C., Fischer, S. E. J. & Kim, J. K. Endogenous RNAi pathways in *C. elegans*. *WormBook* https://doi.org/10.1895/wormbook.1.170.1 (2014).
5. Sarkies, P. *et al.* Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages. *PLoS Biol.* **13**, 1–20 (2015).
6. Hunt, V. L., Hino, A., Yoshida, A. & Kikuchi, T. Comparative transcriptomics gives insights into the evolution of parasitism in *Strongyloides nematodes* at the genus, subclade and species level. *Sci. Rep.* https://doi.org/10.1038/s41598-018-23514-z (2018).
7. Holz, A. & Streit, A. Gain and loss of small RNA classes-characterization of small rnas in the parasitic nematode family strongyloididae. *Genome Biol. Evol.* https://doi.org/10.1093/gbe/evx197 (2017).
8. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: The vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* **12**, 246–258 (2011).
9. Tóth, K. F., Pezic, D., Stuwe, E. & Webster, A. The pirna pathway guards the germline genome against transposable elements. *Adv. Exp. Med. Biol.* https://doi.org/10.1007/978-94-017-7417-8_4 (2016).
10. Bergthorsson, U. *et al.* Long-term experimental evolution reveals purifying selection on piRNA-mediated control of transposable element expression. *BMC Biol.* **18**, 1–17 (2020).
11. Szitenberg, A. *et al.* Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements. *Genome Biol. Evol.* **8**, 2964–2978 (2016).

12. Simon, M. *et al.* Reduced insulin/IGF-1 signaling restores germ cell immortality to *Caenorhabditis elegans* Piwi mutants. *Cell Rep.* **7**, 762–773 (2014).
13. Batista, P. J. *et al.* PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol. Cell* https://doi.org/10.1016/j.molcel.2008.06.002 (2008).
14. Das, P. P. *et al.* Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. *Mol. Cell* **31**, 79–90 (2008).
15. Aravin, A. A. *et al.* The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* https://doi.org/10.1016/S1534-5807(03)00228-4 (2003).
16. Svendsen, J. M. & Montgomery, T. A. piRNA rules of engagement. *Dev. Cell* https://doi.org/10.1016/j.devcel.2018.03.006 (2018).
17. Zheng, Y., Cai, X. & Bradley, J. E. MicroRNAs in parasites and parasite infection. *RNA Biol.* https://doi.org/10.4161/rna.23716 (2013).
18. Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* https://doi.org/10.1038/nrm2632 (2009).
19. Kehl, T. *et al.* About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget* https://doi.org/10.18632/oncotarget.22363 (2017).
20. Farazi, T. A., Juranek, S. A. & Tuschl, T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* https://doi.org/10.1242/dev.005629 (2008).
21. Britton, C. *et al.* Application of small RNA technology for improved control of parasitic helminths. *Vet. Parasitol.* https://doi.org/10.1016/j.vetpar.2015.06.003 (2015).
22. Chapman, E. J. & Carrington, J. C. Specialization and evolution of endogenous small RNA pathways. *Nat. Rev. Genet.* https://doi.org/10.1038/nrg2179 (2007).
23. Zhang, C. & Ruvkun, G. New insights into siRNA amplification and RNAi. *RNA Biol.* https://doi.org/10.4161/rna.21246 (2012).
24. Viney, M. & Kikuchi, T. *Strongyloides ratti* and *S. venezuelensis*—Rodent models of Strongyloides infection. *Parasitology* https://doi.org/10.1017/S0031182016000020 (2017).
25. Buonfrate, D. *et al.* The global prevalence of strongyloides stercoralis infection. *Pathogens* https://doi.org/10.3390/pathogens9060468 (2020).
26. Wolstenholme, A. J., Fairweather, I., Prichard, R., von Samson-Himmelstjerna, G. & Sangster, N. C. Drug resistance in veterinary helminths. *Trends Parasitol.* **20**, 469–476 (2004).
27. Hunt, V. L. *et al.* The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nat. Genet.* https://doi.org/10.1038/ng.3495 (2016).
28. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btp616 (2009).
29. Emms, D. M. & Kelly, S. OrthoFinder2: Fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv* (2018).
30. Kato, M., de Lencastre, A., Pincus, Z. & Slack, F. J. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol.* **10**, R54 (2009).
31. Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
32. Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W. & Rajewsky, N. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).
33. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. MiRBase: From microRNA sequences to function. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gky1141 (2019).
34. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* https://doi.org/10.1126/science.1146484 (2007).
35. Izumi, N. & Tomari, Y. Diversity of the piRNA pathway for nonself silencing: Worm-specific piRNA biogenesis factors. *Genes Dev.* https://doi.org/10.1101/gad.241323.114 (2014).
36. Gu, W. *et al.* CapSeq and CIP-TAP identify pol ii start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151**, 1488–1500 (2012).
37. Cecere, G., Zheng, G. X. Y., Mansisidor, A. R., Klymko, K. E. & Grishok, A. Promoters recognized by forkhead proteins exist for individual 21U-RNAs. *Mol. Cell* **47**, 734–745 (2012).
38. Sato, K., Siomi, M. C. & Nagata, S. The piRNA pathway in Drosophila ovarian germ and somatic cells. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* https://doi.org/10.2183/pjab.96.003 (2020).
39. Saito, K. *et al.* Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes Dev.* https://doi.org/10.1101/gad.1454806 (2006).
40. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* https://doi.org/10.1016/j.cell.2007.01.043 (2007).
41. Roquis, D. *et al.* Histone methylation changes are required for life cycle progression in the human parasite *Schistosoma mansoni*. *PLoS Pathog.* https://doi.org/10.1371/journal.ppat.1007066 (2018).
42. Asikainen, S., Heikkinen, L., Wong, G. & Storvik, M. Functional characterization of endogenous siRNA target genes in *Caenorhabditis elegans*. *BMC Genom.* https://doi.org/10.1186/1471-2164-9-270 (2008).
43. Blumenfeld, A. L. & Jose, A. M. Reproducible features of small RNAs in *C. elegans* reveal NU RNAs and provide insights into 22G RNAs and 26G RNAs. *RNA* https://doi.org/10.1261/rna.054551.115 (2016).
44. Zagoskin, M., Wang, J., Neff, A. T., Veronezi, G. M. B. & Davis, R. E. Nematode small RNA pathways in the absence of piRNAs. *bioRxiv* https://doi.org/10.1101/2021.07.23.453445 (2021).
45. Collins, J., Saari, B. & Anderson, P. Activation of a transposable element in the germ line but not the soma of *Caenorhabditis elegans*. *Nature* https://doi.org/10.1038/328726a0 (1988).
46. Aravin, A. A. *et al.* Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.* https://doi.org/10.1016/S0960-9822(01)00299-8 (2001).
47. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol.* https://doi.org/10.1186/s13059-018-1577-z (2018).
48. Villasante, A. *et al.* Drosophila telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. *Genome Res.* https://doi.org/10.1101/gr.6365107 (2007).
49. Viney, M. E. Exploiting the life cycle of *Strongyloides ratti*. *Parasitol. Today* https://doi.org/10.1016/S0169-4758(99)01452-0 (1999).
50. Nowell, R. W. *et al.* Evolutionary dynamics of transposable elements in bdelloid rotifers. *Elife* **10**, e63194 (2021).
51. Dechaud, C., Volff, J. N., Schartl, M. & Naville, M. Sex and the TEs: Transposable elements in sexual development and function in animals. *Mob. DNA* https://doi.org/10.1186/s13100-019-0185-0 (2019).
52. Hunt, V. L., Tsai, I. J., Selkirk, M. E. & Viney, M. The genome of Strongyloides spp. gives insights into protein families with a putative role in nematode parasitism. *Parasitology* https://doi.org/10.1017/S0031182016001554 (2017).
53. Li, S., Xu, Z. & Sheng, J. tRNA-derived small RNA: A novel regulatory small non-coding RNA. *Genes* https://doi.org/10.3390/genes9050246 (2018).
54. Saikia, M. & Hatzoglou, M. The many virtues of tRNA-derived stress-induced RNAs (tiRNAs): Discovering novel mechanisms of stress response and effect on human health. *J. Biol. Chem.* https://doi.org/10.1074/jbc.R115.694661 (2015).

55. Raina, M. & Ibba, M. TRNAs as regulators of biological processes. *Front. Genet.* https://doi.org/10.3389/fgene.2014.00171 (2014).
56. Fricker, R. *et al.* A tRNA half modulates translation as stress response in *Trypanosoma brucei*. *Nat. Commun.* https://doi.org/10.1038/s41467-018-07949-6 (2019).
57. Hino, A., *et al.* Karyotype and reproduction mode of the rodent parasite Strongyloides venezuelensis. *Parasitology* **141**(13) , 1736–1745 (2014).
58. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* https://doi.org/10.1101/gr.209601.116 (2017).
59. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner. Joint Genome Institute, Department of Energy* (2014) https://doi.org/10.1186/1471-2105-13-238.
60. Rosenkranz, D., Han, C.-T., Roovers, E. F., Zischler, H. & Ketting, R. F. Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. *Genom. Data* **5**, 309–313 (2015).
61. Ahmed, R. *et al.* Conserved miRNAs are candidate post-transcriptional regulators of developmental arrest in free-living and parasitic nematodes. *Genome Biol. Evol.* **5**, 1246–1260 (2013).
62. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 378 (2020).
63. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 1–14 (2008).
64. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37**, 7002–7013 (2009).
65. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
66. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
67. Gebert, D., Hewel, C. & Rosenkranz, D. unitas: The universal tool for annotation of small RNAs. *BMC Genom.* **18**, 644 (2017).
68. Panagopoulos, I., Gorunova, L., Bjerkehagen, B. & Heim, S. The 'grep' command but not FusionMap, FusionFinder or ChimeraScan captures the CIC-DUX4 fusion gene from whole transcriptome sequencing data on a small round cell tumor with t(4;19)(q35;q13). *PLoS One* https://doi.org/10.1371/journal.pone.0099439 (2014).
69. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* https://doi.org/10.1371/journal.pone.0163962 (2016).
70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
71. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
72. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
74. Zdobnov, E. M. & Apweiler, R. InterProScan—An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* https://doi.org/10.1093/bioinformatics/17.9.847 (2001).
75. Alexa, A. & Rahnenfuhrer, J. topGO: topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0. *R Top. Doc.* (2010).
76. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One* https://doi.org/10.1371/journal.pone.0021800 (2011).
77. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res.* https://doi.org/10.1101/gr.849004 (2004).
78. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2019).
79. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

Acknowledgements

Genome data analyses were partly performed using the DDBJ supercomputer system. We thank Simo Sun for assistance and comments.

Author contributions

M.V., V.L.H. and T.K. conceived the study. M.S. V.L.H. and T.K. wrote the manuscript with inputs from others. A.K. and A.Y. prepared biological samples and performed sequencing. M.S., M.D., B.M., R.P., performed informatics analyses.

Funding

VLH was funded by an Elizabeth Blackwell Institute fellowship, a Japanese Society for the Promotion of Science Fellowship (PE16024) and a Wellcome Trust Sir Henry Dale Fellowship (211227/Z/18/Z). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. MS was funded by a URSA University of Bath PhD studentship. TK was funded by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers 19H03212 and 17KT0013, and JST CREST Grant Number JPMJCR18S7.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-14247-1.

Correspondence and requests for materials should be addressed to T.K. or V.L.H.