Evolutionary analysis implicates RNA polymerase II pausing and chromatin structure in nematode piRNA biogenesis

Beltran, T.[1,2], Barroso C. [1,2], Birkle, T.Y. [1,2], Stevens, L.[3], Schwartz, H. T.[4], Sternberg, P. W.[4], Fradin, H.[5,6,7], Gunsalus., K.[5,6,7], Piano, F.[5,6,7], Cotton, J.[6], Martínez-Pérez E. [1,2], Blaxter, M.*[3], Sarkies, P.*[1,2]

[1] MRC London Institute of Medical Sciences, Du Cane Road, London W12 0NN, UK

[2] Institute of Clinical Sciences, Imperial College London, Du Cane Road, London, W12 0NN, UK

[3] Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3TF, UK

[4] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA

[5] Department of Biology, New York University, New York, NY 10003, USA

[6] Center for Genomics and Systems Biology, New York University, New York, NY 10003, USA

[7] Center for Genomics and Systems Biology, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

[8] Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom

* correspondence

mark.blaxter@ed.ac.uk

psarkies@imperial.ac.uk

Abstract

Piwi-interacting RNAs (piRNAs) control transposable elements widely across metazoans but have rapidly evolving biogenesis pathways. In *Caenorhabditis elegans*, almost all piRNA loci are found within two 3Mb clusters on Chromosome IV. Each piRNA locus possesses an upstream motif that recruits RNA polymerase II to produce a ~28 nt precursor transcript. Here, we use comparative epigenomics across nematodes to gain insight into piRNA biogenesis. We show that the piRNA upstream motif is derived from core promoter elements controlling snRNA biogenesis. We describe two alternative modes of piRNA organisation in nematodes: in *C. elegans* and closely related nematodes, piRNAs are clustered within repressive H3K27me3 chromatin, whilst in other species, typified by *Pristionchus pacificus*, piRNAs are distributed genome-wide within introns of actively transcribed genes. In both groups, piRNA production depends on downstream sequence signals associated with RNA polymerase II pausing, which synergise with the chromatin environment to control piRNA precursor transcription.

Introduction

piRNAs are 21-30 nucleotide small RNAs that bind to members of the Piwi subfamily of Argonaute proteins. Conserved across animals, their ancestral role appears to be to defend the genome against transposable elements (TEs). piRNAs hybridise to TE-derived RNAs, instigating post-transcriptional and transcriptional silencing of TEs (Rödelsperger et al., 2017; Siomi et al., 2011). In many organisms, piRNAs are essential for fertility, and germ cell development is defective in their absence (Weick and Miska, 2014). piRNA biogenesis has been characterised in arthropods and mammals. In these taxa long precursor RNAs are produced from ~100 genomic loci and processed into mature 26-30 nt piRNAs bound to Piwi proteins. In germ cells, primary piRNA processing occurs as part of a ping-pong cycle that amplifies piRNAs targeted to expressed TEs (Brennecke et al., 2007).

Despite the conservation of the general logic of piRNA pathway function and biogenesis across animals, the piRNA machinery diverges rapidly amongst closely related organisms. Many of the dedicated piRNA biogenesis factors characterised in *Drosophila melanogaster* are not conserved even amongst Diptera, let alone across arthropods (Weick and Miska, 2014). Some aspects of piRNA function are idiosyncratic. For example in the mosquito *Aedes aegypti* piRNAs derived from RNA viruses are found in the gut (Miesen et al., 2015) while in the silkworm *Bombyx mori* piRNAs regulate a specific protein-coding target gene in the sex determination pathway (Kiuchi et al., 2014). How and why piRNA function and biogenesis diverges so rapidly remains unclear.

Nematodes (phylum Nematoda) represent an extreme example of the diversity of the piRNA system. Although piRNAs are conserved within the Rhabditina (Clade V (Blaxter et al., 1998)), the entire piRNA pathway has been lost independently in several lineages across the phylum (Sarkies et al., 2015). Rhabditina contains the model nematode *Caenorhabditis elegans*, in which characterisation of piRNAs is most advanced. As in other metazoans, piRNAs in *C. elegans* (termed 21U, for their typical length and 5' uracil) associate with a Piwi protein, PRG-1, target transposable elements and are important for fertility (Batista et al., 2008; Das et al., 2008). However, the *C. elegans* piRNA system displays intriguing differences to other organisms. First, the ping-pong cycle is not present. Instead, piRNAs recruit RNA dependent RNA polymerases (RdRPs) to target RNAs. RdRP activity produces 22 nt RNAs that have a 5' G (22G RNAs), antisense to targets, which bind to nematode-specific argonautes (WAGO) to bring about silencing (Bagijn et al., 2012). Second, piRNAs are produced from monocistronic genomic loci, the vast majority of which are preceded by a GTTTC consensus motif (the Ruby motif) (Batista et al., 2008; Das et al., 2008; Ruby et al., 2006) that recruits RNA polymerase II (Pol II) to produce a 5’ capped 28 nucleotide precursor transcript (Billi et al., 2013; Gu et al., 2012). Two piRNA clusters on chromosome IV, spanning 2.5 Mb and 3.7 Mb, contain over 90% of piRNA loci (Ruby et al., 2006). Pol II transcription of piRNA precursors requires the nematode-specific pseudokinase PRDE-1 (Billi et al., 2013; Gu et al., 2012; Weick et al., 2014) and the small nuclear RNA (snRNA) activating protein complex (SNAPc) component SNPC-4 (Kasper et al., 2014). Together, SNPC-4 and PRDE-1 have been hypothesised to establish a specific chromatin structure essential for piRNA biogenesis (Kasper et al., 2014).

Several key aspects of nematode piRNA production remain poorly understood. Amongst the outstanding problems are the evolutionary origin of the Ruby motif, how Pol II is controlled such that it transcribes only very short piRNA precursors and how the clustering of piRNA loci into genomic regions contributes to piRNA production. Here, through a comparative analysis of piRNA biogenesis across nematodes, we obtain new insights into these questions. First, we uncover the evolutionary origin of the Ruby motif, showing that nematode piRNA transcription evolved from snRNA transcription. Second, we reveal two distinct modes of piRNA locus organisation in nematodes - either forming high density clusters within H3K27me3 chromatin domains or dispersed throughout the genome within active genes enriched for H3K36me3 chromatin. Third, we discover a novel downstream sequence signal at nematode piRNA loci, which likely promotes RNA Pol II pausing at piRNA loci to generate short piRNA precursors. By using CRISPR-mediated genome editing we confirm that both the surrounding chromatin environment and RNA Pol II pausing sequence signals determine the activity of piRNA loci in nematodes.

Results

The Ruby motif is an ancient piRNA regulatory module

We surveyed a wide taxonomic span of nematodes for components of the piRNA system. All of the nematodes we examined within Rhabditina (Clade V) possessed the Piwi protein PRG-1 (Figure 1), with the exception of *Caenorhabditis plicata*, suggesting a recent loss of the piRNA pathway in this species. Outside of Clade V we only identified Piwi in the closely related free living nematodes *Plectus sambesii* and *Plectus murrayi*. The presence of Piwi in *P. sambesii*, which is ancestral to Rhabditida (which includes Clades III, IV and V), is consistent with the hypothesis that the piRNA pathway was lost independently in Clade III and Clade IV (Figure 1) (Sarkies et al., 2015).

To reveal the evolutionary dynamics of piRNA biogenesis we sequenced small RNAs and mapped them to their genomes in a number of additional species (Figure 1). We sequenced and assembled *de novo* the genomes of *P. sambesii* and *Poikilolaimus oxycercus* and sourced other genomes from ongoing and published projects. In the majority of nematodes with Piwi, including in *P. sambesii* and *P. murrayi*, we found a motif upstream of 21U-RNAs which bore remarkable similarity to the *C. elegans* core Ruby motif (Figure 1). We did not find any motifs upstream of any other type of small RNA. This suggests that the Ruby motif originated at least 350 mya (Figure 1). Surprisingly, there was no evidence of upstream motifs in *Heterorhabditis bacteriophora* despite the presence of Piwi, suggesting that that the Ruby motif has been lost in this species. In all nematodes with Ruby-motif dependent piRNAs transcription initiated two nucleotides upstream of the mature piRNA (Figure S1A), just as in *C. elegans* (Billi et al., 2013; Gu et al., 2012; Weick et al., 2014). We observed large variability in the number of piRNA loci across genomes (Figure 1) that could not be attributed to differences in sequencing depth. The biological and evolutionary correlates of these differences await future study.

The Ruby motif is derived from the nematode SNAPc binding motif

The Ruby motif upstream of piRNAs in plectid nematodes was strongly associated with an additional 5' motif (Figure 1A). The two motifs almost always occurred together (Figure S1B). This extra motif bears a striking similarity to the *C. elegans* SNAPc motifs, predicted from binding sites of the SNAPc subunit SNPC-4 (Kasper et al., 2014; Figure 2A). We predicted SNAPc motifs in *P. sambesii* from regions upstream of Pol II and Pol III-dependent snRNA genes. The *P. sambesii* SNAPc motif associated with snRNA genes and the piRNA motif are highly similar (Figure 2A). We recovered a TATA-box at Pol III-dependent snRNA loci, but not at Pol II snRNAs or piRNA loci, consistent with Pol II transcription of piRNAs (Hung and Stumph, 2011). We predicted SNAPc motifs in a wider sampling of nematodes with sequenced genomes, many of which have lost the piRNA pathway (Sarkies et al., 2015). Alignment of these motifs showed strong conservation of the 5' half of the motif, whilst the 3' half was more divergent. There were strong sequence similarities between the 3' SNAPc motif and the Ruby motif in all nematodes (Figure S2-2), and the positioning of the 3' SNAPc motif relative to snRNA transcription start sites (TSSs) mirrors the positioning of the Ruby motif at piRNA loci (Figure 2A). Comparison of SNAPc motifs from Clade II nematodes strongly suggested that the Ruby motif was derived from the 3' half of an original nematode SNAPc motif and the two motifs diverged from each other since the last common ancestor of Rhabditida (Clades III, IV, V) (Figure 2, Figure S2-1). Altogether, our data suggests that nematode piRNA transcription evolved from snRNA transcription.

Two levels of piRNA clustering in nematodes

piRNA loci controlled by Ruby motifs in *C. elegans* are almost exclusively located in two ~3 Mb regions on Chromosome IV. However, in the chromosomal genome assembly of *Pristionchus pacificus* (Rödelsperger et al., 2017), piRNA loci are distributed relatively evenly across the five autosomes but absent from the X chromosome (Figure 3A). Pristionchus is also in Rhabditina (Clade V) but is a member of Diplogasteromorpha, while *C. elegans* is in the Rhabditomorpha.

Most other nematode genome assemblies are of lower quality than those of *C. elegans* and *P. pacificus*. To analyse species with more fragmented genome assemblies we developed and validated a method to identify piRNA locus-dense contigs (Figure S3A,B). We then asked if it was likely that the clusters were in conserved regions across species, by using 1:1 orthologous protein coding genes to map the contigs to the *C. elegans* genome. We observed two large blocks of piRNA regions that contained orthologues that mapped to *C. elegans* Chromosome IV in the majority of *Caenorhabditis* species and also in *O. tipulae* (blue in Figure 3B). This pattern is consistent with the conservation of a small number of piRNA gene clusters in these species. In other nematodes, including *P. pacificus*, regions with elevated piRNA gene densities contain orthologues that map to several *C. elegans* chromosomes (red in Figure 2B). Across all nematodes, there was a depletion of piRNA loci linked to protein coding genes currently residing on the *C. elegans* X chromosome (Figure 3B). X linkage of sequence scaffolds was confirmed independently in three species: *O. tipulae*, where sex-linked genes have been used to identify the X chromosome (M.A-Felix, personal communication), in *H. contortus*, for which there is a recent chromosomal assembly (James Cotton, unpublished data), and in the chromosomal assembly of *P. pacificus* (Figure S3C).

We quantified piRNA clustering by examining the extent to which piRNA loci were concentrated within genomic regions (Figure 3C). Bimodal regression analysis identified two groups, one with more clustered piRNAs and one with less clustered piRNAs. These groups were congruent with those recovered from mapping to orthologous regions in *C. elegans* (Figure 3D). On the basis of these analyses we propose that there are two distinct modes of organisation of piRNA loci in nematodes, one clustered, similarly to *C. elegans* and one dispersed, similarly to *P. pacificus*.

Clustered and dispersed piRNA loci have different chromatin environments

To further characterise the differences between clustered and non-clustered modes of piRNA gene organisation, we examined the chromatin environment of piRNA loci in *C. elegans* and *P. pacificus*. In the germline, the *C. elegans* genome is organised into mutually exclusive, stable domains of H3K27me3 repressive chromatin ("regulated domains") and H3K36me3 transcriptionally active chromatin ("active domains") (Evans et al., 2016). The majority (95%) of piRNA loci within the clusters overlapped with H3K27me3/regulated domains and only 5 of 9735 21U-RNA sequences were produced from H3K36me3/active domains (Figure 4A). Regulated domains were significantly larger within piRNA cluster regions compared to the rest of the genome (Figure S4A). piRNA loci were 227-fold depleted from active chromatin compared to a uniform distribution across piRNA cluster regions (empirical $p<0.01$). While chromatin domains have been defined based on analyses of the early embryo, they are present in the adult germline and show similar ChIP-seq profiles (Evans et al., 2016) (Figure S4B). Genes that overlapped with piRNA loci showed extremely low or null expression in the germline, while genes depleted of piRNA loci within cluster regions were highly expressed (Wilcox unpaired test $p<2.2e-16$, Figure S4C). Together these data strongly suggest that expression of *C. elegans* piRNAs occurs predominantly from regulated domains enriched in H3K27me3.

The organisation of piRNA loci in *P. pacificus* contrasts with that of *C. elegans*. 88% of piRNA loci were localised within protein-coding genes (1.6-fold enrichment within genes compared to intergenic regions, $p<0.01$ simulation test, Figure S4D). *C. elegans* piRNAs were biased toward intergenic regions (1.34-fold depletion from genes, $p<0.01$ simulation test, Figure S4D). *P. pacificus* piRNA loci are distributed approximately equally in the sense and antisense orientation of the genes in which they are found, but 97% of the intragenic piRNA loci are found within introns (empirical $p<0.01$ relative to random simulation) (Figure S4C). The chromatin environment of piRNA loci in *P. pacificus* is also distinct from *C. elegans.* The *C. elegans* 1:1 orthologues of the *P. pacificus* genes that contained piRNA loci were found almost exclusively in active H3K36me3 chromatin regions (Figure 4B,C) that are highly expressed in the germline (Figure S4C). While we do not have corresponding chromatin state data for *P. pacificus*, these data strongly suggest that its piRNAs are associated with active, rather than repressed, chromatin.

We examined the fraction of Ruby motifs associated with expressed 21U-RNAs in *C. elegans* and *P. pacificus* in different chromatin environments. In *C. elegans*, 74% of motifs in regulated chromatin were expressed, contrasting with less than 2% of motifs in active chromatin (Figure 4D, Figure S4E). In *P. pacificus*, piRNA motifs within genes whose *C. elegans* orthologues were within active domains were 2-fold more likely to produce piRNAs than those associated with *C. elegans* orthologues in regulated domains (Figure 4D, Figure S4E). The fraction of expressed motifs in *P. pacificus* was positively correlated with the expression of the overlapping protein-coding genes (Figure 4E), suggesting that host protein-coding gene expression promotes expression of the 21U-RNAs.

To test the involvement of chromatin structure in piRNA biogenesis directly, we used CRISPR-Cas9 genome editing in *C. elegans.* We selected an endogenous piRNA locus (21UR-3922) and modified the 21UR sequence to generate an artificial piRNA that could be distinguished from all endogenous piRNAs by sequencing (21UR-3922m; Figure 4F). We inserted the artificial piRNA gene into adjacent regulated and active chromatin domains within the piRNA cluster at the centre of Chromosome IV (Figure 4F). When inserted into regulated chromatin the artificial piRNA was expressed at the same level as the endogenous piRNA (Figure 4G,H). However, there was no detectable expression of the mature 21U-RNA when inserted into active chromatin (Figure 4G,H). Thus a repressive chromatin environment is necessary for piRNA production in *C. elegans*.

To generalise these observations we assessed the intergenic and genic proportions and the predicted chromatin environment of piRNA loci across nematodes. Species could be divided into two groups, one with more genic piRNAs associated with *C. elegans* orthologs enriched for active chromatin and one with more intergenic piRNAs associated with *C. elegans* orthologs enriched for regulated chromatin. These categories overlapped strongly with the pattern of clustering. Species with clustered piRNAs had H3K27me3/regulated-enriched piRNAs, whilst less clustered species had H3K36me3/active enriched piRNAs (Figure 4I).

We propose that there are two fundamental modes of piRNA organisation in nematodes. “Caenorhabditis”-type (C-type) piRNAs are organized into dense clusters within repressive chromatin. These sub-clusters are then grouped together into a large “super-cluster”, for example the regions on chromosome IV in *C. elegans*. In contrast, “Pristionchus”-type (P-type) piRNAs are more widely distributed across the genome where they are found within the introns of germline-expressed genes, enriched for H3K36me3.

Chromatin domain organisation of piRNAs is under selection in nematodes

These different modes of piRNA organisation led us to ask how selection is acting on them. Using wild-isolate, genome-wide single nucleotide polymorphism (SNP) data from the *C. elegans* Natural Diversity Resource (Cook et al., 2017), *C. briggsae* (Thomas et al., 2015) and *P. pacificus* (Rödelsperger et al., 2014), we explored the predicted effects of SNPs on Ruby motifs by measuring their effect on match to a position weight matrix of all motifs. The allele frequency spectrum was markedly different between motifs within the piRNA locus clusters and those outside for both *C. elegans* and *C. briggsae* (Figure S5A,B). In *C. briggsae* and *C. elegans*, alleles with a low motif score were much less likely to be the major allele (present in >90% of strains) within the cluster than outside the cluster, implying that piRNAs are under stronger selection within the cluster than outside (p<1e-3 for both, Fisher’s Exact Test, Figure 5A,B). Similarly, both within and outside the cluster, selection was stronger on piRNAs within H3K27me3 domains than H3K36me3 domains (Figure 5C).

In *P. pacificus*, observed SNPs tended to have a smaller effect on the motif score (Figure S5C), perhaps reflecting the lower information content within the motif (Figure 1).

The major allele frequencies of low-scoring motifs did not vary with the density of piRNA loci, consistent with the lack of clusters (Figure 5D). However, in *P. pacificus*, SNPs predicted to disrupt piRNAs were less likely to be the major allele in highly expressed genes compared to lowly expressed genes (Figure 5E). Exactly the opposite trend was observed in *C. elegans* (Figure 5F). Together, these data confirm that the two different modes of piRNA biogenesis are under selection in their respective species.

Comparative analysis implicates RNA Pol II pausing in piRNA production in nematodes

A puzzling feature of piRNA biogenesis in nematodes is how Pol II is regulated to produce the ~30 nt precursor of mature piRNAs (Gu et al., 2012). Our finding that piRNAs can be found in both repressed and active chromatin environments is challenging, as *Pristionchus*-like piRNAs localise to regions of high transcriptional activity expected to favour elongating polymerase. Pol II is known to pause at the transcription start site (TSS) at certain types of eukaryotic protein-coding gene promoters. This results in the production of short-capped RNAs that are rapidly degraded by the nuclear exosome. This has been proposed as the underlying mechanism of piRNA transcription in nematodes (Gu et al., 2012) but so far evidence has been lacking.

Previous genome-wide analyses in *D. melanogaster* and human cell lines identified a sequence signature associated with RNA Pol II pausing characterised by a region with increased T_m relative to the genome-wide background and a region with reduced T_m immediately 3' (Gressel et al., 2017; Nechaev et al., 2010). This architecture is thought to cause RNA Pol II to backtrack from the 3' low T_m region to the 5' high T_m region where RNA-DNA interactions are more stable. Consequently RNA Pol II pausing is observed at the boundary between the high T_m and low T_m regions. We examined T_m signatures around nematode piRNA TSSs. *P. pacificus* piRNA loci displayed a strong match to the pausing-associated sequence signature, such that a marked reduction in T_m relative to the genome-wide background coincided exactly with the predicted 3' ends of piRNA precursors (Figure 6A,B; Figure S6A). Sequence motifs have not previously been identified downstream of *C. elegans* piRNAs (Ruby et al., 2006); nevertheless we found a clear match to the pausing-associated signature at *C. elegans* piRNA loci, although this was less pronounced than in *P. pacificus* (Figure 6A,B; S6A,B).

We then examined the role of the pausing-associated sequence signature on piRNA expression by stratifying piRNA loci according to the strength of this signature. In *P. pacificus*, mature 21U-RNA expression was significantly greater from the loci with the strongest pausing-associated signatures than from the weakest ($p<1e-10$ Wilcoxon unpaired test) (Figure 6C). This effect was less marked but still significant in *C. elegans* (Figure 6C). The GC content of the 21U-RNA is not responsible for this effect, as sequence downstream of the 21U itself alone had a significant impact on mature piRNA abundance (Figure S6C).

To test the effect of the pausing signal on piRNA transcription, we examined piRNA precursors from nuclear short capped RNA sequencing data from a previous study (Weick et al., 2014). piRNA precursors produced from loci with strong pausing signals were significantly shorter and on average two-fold more abundant than those produced from loci with weak pausing signals (Figure 6D). This effect was emphasized in *prg-1* mutants, where mature piRNAs are not present, leading to accumulation of piRNA precursors in sequencing data (Figure 6E, S6D-F). Consistently, the few atypical longer precursors (40-100bp) sequenced by capseq (Gu et. al., 2012) mapped almost exclusively to loci with the weakest pausing signals (Figure 6F, S6D-F).

We modified the sequence downstream of the artificial 21UR-3922m to increase the T_m by 27.5ºC, and inserted this new artificial piRNA (“21UR-3922m-bad terminator”) into the same site within the H3K27me3 domain (Figure 6G). This allowed us to compare of the expression of two piRNA loci that have different pausing-associated sequence signatures but are otherwise identical. The 21U-RNA with the higher downstream Tm was expressed at around 3-fold lower levels than the endogenous piRNA with the same sequence (Figure 6H). Moreover, the piRNA precursor was expressed at ~10-fold lower levels than the endogenous piRNA precursor (Figure 6H). Thus the Tm of the downstream sequence alone is sufficient to affect transcription of the piRNA locus.

We detected similar Pol II pausing-associated sequence signatures across nematodes (Figure 6I, Figure S7-4). In *P. sambesii,* the low T_m region downstream of the 21Us is shifted ~10 nt further downstream compared to the rest of nematodes examined (Figure 6I, Figure S7-4). piRNA precursors are significantly longer in this species compared to *C. elegans* (Figure 6J), consistent with a role for the pausing signal in piRNA transcription termination.

To test the involvement of pausing in piRNA production we examined *C. elegans* mutants carrying a deletion in TFIIS. TFIIS is a general transcription elongation factor that rescues backtracked Pol II complexes by promoting cleavage of the 3’ end of the RNA by Pol II itself (Schweikhard et al., 2014). In *D. melanogaster*, TFIIS deletion leads to an increase in the length of promoter-associated short capped RNAs, indicative of pausing and backtracking of Pol II at the promoter (Nechaev et al., 2010). *C. elegans* mutants carrying a deletion in the orthologue of TFIIS (*T24H10.1* ok2479) showed significantly longer piRNA precursors (Figure 7A, S7-1 A-D) and a modest, but consistent, decrease in mature piRNA abundance (Figure 7B). The effect of the TFIIS deletion is specific for short-capped RNAs, as we did not find differences in length distributions of rRNA degradation fragments (Figure S7-2 A-D) or tailed 22G-RNAs (Figure S7-2 E). To test whether TFIIS promotes the cleavage of piRNA precursors, we examined sequencing data for potential 3’ cleavage fragments, with a 5’ monophosphate mapping 28-38 nucleotides downstream of the piRNA transcription start site. We detected potential cleavage products from around 10% of loci in wild type nematodes, but these products were 2-fold reduced in TFIIS mutants (Figure 7 C,D). These data support a direct role for Pol II pausing in proper termination of piRNA transcription and release of piRNA precursors.

In *C. elegans* piRNA precursors are trimmed by the exonuclease PARN-1, an orthologue of Trimmer from *B. mori* (Tang et al., 2016). An alternative explanation for our findings would be if the sequence signature exerted its effect post-transcriptionally by affecting the trimming of capped piRNA precursors. However, this is unlikely to be the case as capped piRNA precursor length differences between loci with strong and weak pausing signals remain unaffected in the absence of trimming (Figure S7-3B). In addition, the overall length of capped piRNA precursors does not change in *parn-1* (tm869) mutants compared to wild-type nematodes (Figure S7-3C). Finally, the increase in precursor length due to TFIIS deletion is still significant when assessed in a *parn-1* (tm869) background (Figure S7-3D). This data is consistent with cytoplasmic trimming on an uncapped Piwi-bound piRNA precursor, which has been shown in *B. mori* (Izumi et al., 2016). Therefore, nuclear short capped RNAs represent early unprocessed piRNA precursor transcripts.

Altogether, our data strongly suggests that the sequence properties of piRNA loci promote RNA Pol II pausing and are thus important for generation of short piRNA precursors.

Synergy between chromatin environment and pausing-associated sequence signals in nematode piRNA biogenesis

Though clearly important for piRNA biogenesis, the Pol II pausing signature is weaker in *C. elegans* than in *P. pacificus*. To test whether this extended to the remaining C-type and P-type species, we quantified the strength of Pol II pausing-associated sequence signatures at piRNA loci across nematodes (Figure S7-4). Across nematode species the strength of the Pol II pausing-associated sequence signature correlated positively both with the proportion of piRNAs in active chromatin and the proportion of piRNAs within genes (Spearman rho > 0.67 for both, p=1.9e-3 and p=3.89e-6; Figure 8E,F). Thus C-type species, in which piRNA loci are found within H3K27me3 chromatin domains, have relatively weak pausing signals while stronger pausing signals are found in P-type species where piRNA loci are found within H3K36me3 domains.

Discussion

Our analysis of piRNA organisation and biogenesis across nematodes not only illuminates the evolutionary history of the piRNA system in nematodes but also provides unexpected insights into the fundamental mechanism of piRNA production in *C. elegans*.

piRNA transcription evolved from snRNA transcription

The Ruby motif upstream of *C. elegans* piRNAs was derived from the 3' half of the ancestral nematode SNAPc motif upstream of snRNA loci. In *C. elegans*, SNAPc binding sites are globally enriched in piRNA clusters but do not seem to localise at a particular distance from piRNA TSSs (Kasper et al., 2014). Our analysis predicts that SNAPc binds directly upstream of piRNA loci to drive Pol II-dependent transcription of piRNAs in *P. sambesii* and *P. murrayi*. Even in ancestral nematodes that have lost the piRNA pathway, the SNAPc motif upstream of snRNA loci closely resembles the *P. sambesii* piRNA motif. Thus nematode piRNA biogenesis evolved from snRNA transcription. The separation of Ruby and SNAPc motifs in rhabditid nematodes (Clades III, IV, V) may correlate to the appearance of specialised piRNA factors such as PRDE-1, which are not conserved in plectids. PRDE-1 interacts with SNPC4 (Kasper et al., 2014) and may recruit SNAPc to the *C. elegans* Ruby motif in the absence of direct DNA binding.

RNA Pol II pausing as a key feature of piRNA production in nematodes.

Previous identification of piRNA precursors by deep sequencing in *C. elegans* demonstrated piRNA production by RNA polymerase II (Billi et al., 2013; Gu et al., 2012), and it was hypothesized that piRNA precursors might result from promoter-proximal Pol II pausing (Gu et al., 2012). This suggestion was prescient: our analysis of piRNA biogenesis across nematodes reveals that this process is indeed likely to be responsible for the short length of piRNA precursors.

Following transcription initiation, Pol II forms an early elongation complex with limited elongation capacity, requiring further regulation to transition into productive elongation. During early elongation, Pol II is highly susceptible to pausing, which can lead to backtracking by a few nucleotides (Schweikhard et al., 2014). Release of the stalled complex requires cleavage of the 3' end of the transcript, promoted by TFIIS. At many promoters Pol II rarely switches into elongation and is instead released from the template. This process results in accumulation of a short, capped RNAs derived from the promoter of the gene. We suggest that exactly this cycle of backtracking, cleavage and release of RNA results in the production of piRNA precursors. Our model is supported by recent data showing that binding and release of Pol II at transcription start sites and the production of short capped RNA occurs continuously at paused loci (Krebs et al., 2017).

The involvement of Pol II pausing in production of piRNAs in *C. elegans* is consistent with other findings demonstrating production of functional small RNAs from promoter-proximal Pol II. Mammalian Ago-2 binds to small RNAs produced from hairpins derived from paused Pol II at protein-coding gene promoters (Zamudio et al., 2014). In ciliates, Pol II pausing in conjunction with TFIIS activity generates ~25 nt RNAs that are loaded into Piwi proteins (Gruchota et al., 2017). These observations suggest that Pol II pervasive initiation and pausing can be co-opted for the biogenesis of functional small RNAs across eukaryotic genomes, including mammals.

Dynamic evolution of the piRNA pathway: two alternative modes of piRNA organisation and biogenesis across nematodes.

Here we show that two modes of piRNA biogenesis are found in nematodes (Figure 8G,H). in C-type species piRNAs are found clustered together at high density, within regions of H3K27me3-rich chromatin, as in *C. elegans*. *C. elegans* piRNA clusters are therefore better described as "clusters-of-clusters" within H3K27me3 domains separated by H3K36me3 rich regions depleted of piRNAs. Contrastingly, in P-type species piRNAs are found in active chromatin, and are not clustered together, as observed in *P. pacificus*. The piRNA types also differ in their reliance on cis-acting Pol II pausing-associated sequence signatures downstream of the piRNAs: whilst these signatures are relatively weak at piRNA loci in C-type species, they are much more obvious at piRNA loci in P-type species.

piRNA precursors must be transcribed at a sufficient rate whilst ensuring that Pol II does not switch to elongation mode. We suggest that the two modes of piRNA biogenesis in nematodes represent alternative solutions to this problem. The high density of piRNAs in C-type species may serve to attract sufficient Pol II to a region devoid of active genes (Figure 7B). Once present in the H3K27me3-rich chromatin surrounding piRNA loci, Pol II is unlikely to switch to elongation. A contrasting situation occurs in P-type species. Here, we suggest that piRNA promoters "piggyback" on the capacity of protein-coding genes to attract Pol II, and thus a high piRNA density is not required. However, Pol II is prone to switch to elongation within these regions so a stronger pausing-associated sequence is required downstream of the piRNA to ensure correct termination.

This model is consistent with genome-wide studies in a variety of model systems showing that H3K27me3 is anticorrelated with the elongating form of Pol II. It has been reported that H3K27me3 accumulation at promoters in the absence of H2K119 mono-ubiquitination by PRC1, directly promotes short capped RNA production at poised genes (Min et al., 2011). This is particularly interesting because *C. elegans* lacks a clear germline-expressed ortholog of PRC1 and thus likely do not have germline H2B-monoubiquitination in H3K27me3 rich regions (Karakuzu et al., 2009). However, this finding applies to bivalent genes enriched in H3K4me3 in addition to H3K27me3 and whether H3K27me3 alone is sufficient is unknown.

Currently we have insufficient evidence to establish which mode of biogenesis was ancestral. It appears that the mode may switch dynamically. For example *D. pachys* has P-type piRNAs despite being nested within the C-type species *C. elegans* and *O. tipulae*. It is therefore possible that both types of piRNAs use common biogenesis machinery and could coexist within the same organism. In support of this we note that some species are "intermediate" between clustered and non-clustered organisation (Figure 3F), and there is a gradation of both the percentage of piRNA loci in genes and the strength of the pausing-associated sequence signature downstream of the piRNAs (Figure 7E,F).

Overall, the novel aspects of piRNA biogenesis that we describe for *C. elegans* illustrate the utility of using a cross-species approach to investigate mechanistic questions in model organisms. We predict that this approach will be similarly important in addressing fundamental questions of mechanisms for other pathways of epigenetic regulation and genome defence, in part because these systems evolve rapidly across species.

Methods

Nematode culture and RNA extraction

C. elegans, *Oscheius tipulae* and *Diploscapter pachys* were grown at 20°C in standard nematode growth medium (NGM) agar plates feeding on HB101 *E. coli* seeded in LB medium. *P. oxycercus* was grown in the same way at 25°C. *P. sambesii* was grown at 25°C in low salt NGM feeding on HB101 seeded in water. *H. bacteriophora* strain M31e was grown on lipid agar media using *Photorhabdus temperata* TRN16 as a food source. To enrich for germline tissue, *C. elegans*, *O. tipulae* and *H. bacteriophora* nematodes were synchronised by hypochlorite treatment and grown to adulthood. *P. sambesii* adults were selected using the COPAS sorter (Union Biometrica). Nematodes were resuspended in Trizol reagent (Life Technologies) and tissues were disrupted by bead-beating except H. bacteriophora where tissues were disrupted by grinding. RNA was then extracted and precipitated from Trizol according to the manufacturer's instructions.

Genome sequence and assembly of P. oxycercus

Genomic DNA of *P. oxycercus* was prepared using a DNeasy Blood and Tissue kit (Qiagen), according to the manufacturer's instructions. Library preparation and sequencing of P. oxycercus was performed by GATC Biotech.

Quality control of *P. oxycercus* raw genomic data was assessed using Fastqc v0.10.1, and reads were quality trimmed using skewer v0.2.2 (Jiang et al., 2014) with parameters –q 30 –l 51. K-mer plots were generated with kmc, which revealed extensive heterozygosity. A preliminary single-end assembly was generated with Velvet (Zerbino and Birney, 2008) and contaminants were identified using blobtools (Kumar et al., 2013). The data were digitally normalised to 80x coverage to facilitate assembly. Assembly was carried out with SPAdes v3.5.1(Bankevich et al., 2012) correcting the reads with BayesHammer within the SPAdes pipeline. Assembly parameters were –k 21,33,55,77,99 –cov-cutoff auto --careful. *Plectus sambesii* was assembled as described in Rosic. et. al., 2017.

Examination of the resulting distribution of contig read coverages revealed a bimodal distribution in both genomes suggesting heterozygosity. Haploid coverage contigs were collapsed and postprocessed with Redundans (Pryszcz and Gabaldón, 2016), that runs SSPACE3 (Boetzer et al., 2011) and SOAPdenovo Gapcloser internally. The identity percentage identity for the collapse of redundant contigs was chosen to be 90%, since lower values did not result in a significantly greater reduction of the span, and no decrease in the number of KOGs by CEGMA (Parra et al., 2007) was found. RepeatModeler v1.0.8 (Smit and Hubley *RepeatModeller open 1.0*) was used to generate species-specific repeat libraries that were concatenated with a nematode repeat library from RepBase. The combined library was used to mask the genome with RepeatMasker v4.0.7 (Smit, Hubley and Green *RepeatMasker open 4.0*).

To annotate the *P. oxycerca* genome, we generated rRNA-depleted paired-end total RNA-seq libraries. Quality control of the RNA-seq reads was performed with Fastqc and Skewer with parameters –q 30 –l 50. Reads were aligned to the assemblies with STAR v2.5.2

(Dobin et al., 2013) with default parameters except for --twopassMode Basic. Alignment bam files were used for automated annotation using BRAKER v1.9(Hoff et al., 2016).

Small RNA library preparation and sequencing

RNAs were pretreated to modify 5' ends in two ways. For RPPH treatment, 1-2 μg of total RNA was treated with 10 units (2 μl) RNA 5' pyrophosphohydrolase (NEB) for 1 h at 37°C. The treated RNA was phenol-chloroform extracted and ethanol precipitated with sodium acetate and glycogen for 2 days, and resuspended in RNase-free water. For CIP-RPPH treatment, 3-5 μg of total RNA was treated with 4 μl of QuickCIP (Quick dephosphorylation kit, NEB) in a total volume of 40 μl for 90-120 min at 37°C. RNA was phenol-chloroform extracted, precipitated overnight with sodium acetate and glycogen and resuspended in RNase-free water.

Small RNA libraries from treated or untreated RNA were built using the TruSeq small RNA kit from Illumina according to the manufacturer's instructions except for an increase in the number of PCR cycles from 11 to 15. Libraries were eluted in 0.3 M NaCl, ethanol precipitated and quantitated with Qubit and TapeStation. Libraries were pooled in groups of 6 to 12 per lane and sequenced on an Illumina HiSeq2000. The Illumina universal adapter was trimmed from small RNA reads using cutadapt v1.10 and reads were mapped to the corresponding genome assemblies with Bowtie v0.12 (Langmead et al., 2009) with parameters –v 0 –m 1.

Phylogenetic profiling of piRNA pathway genes

Identification of Piwi from nematode genomes was performed essentially as described (Sarkies et al., 2015). Briefly, reciprocal blastp searches were performed on predicted protein sets using *C. elegans* PRG-1 as a query sequence. In species where Piwi is not present, the best hit for PRG-1 is another Argonaute subfamily member. In species where we did not identify Piwi from predicted proteins (such as *C. plicata*), we additionally performed tblastn searches against the genome sequence using *C. elegans* Piwi as a query sequence, to verify the absence of the gene.

De novo motif discovery and piRNA annotation

Sequences 110 bp upstream and 30 bp downstream of 21U mapping sites were extracted and subsets of 2000 sequences were randomly selected to predict motifs de novo using MEME v4.10.1 (Bailey et al., 2009) with parameters –dna –mod zoops –maxsize 2000000 -nmotifs 10. Genome-wide nucleotide content was used as a background model for MEME. The resulting motifs from each subset were compared to assess the consistency of the analyses, and the full dataset of upstream sequences was scanned with the obtained motif position weight matrix (PWM) using a custom Python implementation (scripts available upon request). For each species, the distribution of mapping positions showed a peak at the expected distance of 42 b from the 21U mapping site. The distributions of PWM scores showed bimodal distributions, indicating good separation of true piRNA motifs. These distributions were used to define species-specific threshold to select true positive motifs, which were annotated as piRNA loci. Motifs were used for genome-wide scans without 21U sequence information using the same threshold. In the case of *P. sambesii*, the increased length of the motif resulted in an exceptional specificity of genome-wide scans, so all motif positions were considered “true” piRNA loci. *P. murrayi* piRNAs were defined in the same way using the *P. sambesii* motif. In *H. bacteriophora* upstream sequences displayed no positional bias for GTTTC nor any of the motifs predicted by MEME. We performed motif discovery on other small RNA lengths and 5’ nucleotides as controls, finding no enrichment of motifs. To map early precursor transcripts, we mapped RPPH treated small RNA libraries to piRNA loci identifying reads containing 5’ and 3’ extensions to the 21U-RNA sequence as described previously (Weick et al., 2014).

Annotation of snRNA genes and prediction of SNAPc motifs

We annotated SNAPc-dependent non-coding RNA loci across nematode genomes using Infernal v1.1.2 (Nawrocki and Eddy, 2013). RFAM alignments for U1, U2, U3, U4, U4atac, U5, U7, U11 and U12 snRNAs (Pol II-dependent), and U6 snRNA, U6atac snRNA, 7SK RNA and MRP RNA (Pol III-dependent) (see Supplementary tables for RFAM IDs) were used to build and calibrate Infernal covariance models with default parameters. Nematode genomes were searched with the trained models with default parameters. The 100 bp upstream of significant hits (E-value >1e-3) were extracted for motif prediction with MEME v4.10.2 (Bailey et al., 2009) with parameters –dna –mod zoops –nmotifs 10.

piRNA clustering analysis and comparison to the C. elegans genome

We identified piRNA-enriched contigs by applying a binomial test against a null hypothesis of even distribution of piRNAs across the genome. Contigs were ranked based on the p-value of the enrichment, and a cumulative plot of piRNA counts by span was generated. This approach yielded a good separation between piRNA-enriched and piRNA-depleted contigs and allowed us to define the span covered by piRNAs for each species. We validated this method using a fragmented *de novo* assembly of the *C. elegans* genome, where we successfully recovered the genomic regions corresponding to the *C. elegans* piRNA clusters. We applied the same method to the genome of *O. tipulae*, for which there is a good sequence-anchored linkage map. The top 10 most enriched contigs were found in two blocks mapping to chromosomes *Ot*IV and *Ot*V, consistent with the presence of clusters similar to *C. elegans*. To quantify the extent of clustering we calculated the significance of the enrichment of each contig compared to the null hypothesis of a uniform distribution of piRNAs across the genome using a chi-squared test and ordered by decreasing significance of enrichment. We then calculated the cumulative frequency of total loci and identified the span required to reach 95% of the total loci. We then plotted this value against the total number of piRNAs to give the plot in Figure 3. Multimodal regression analysis was performed using the Mixtools package in R.

For each species, we defined gene orthology to *C. elegans* with Orthofinder v0.6.1(Emms and Kelly, 2015) running mcl 14.137. 1:1 orthologs were filtered from the final dataset (Supplementary tables, *all species-piRNA organisation*), and the *C. elegans* orthologs of genes localised in piRNA contigs for each species were identified.

Analysis of the genomic organisation of piRNAs

We used Integrative Genome Viewer to visualise *C. elegans* piRNA positions relative to genes and a variety of chromatin tracks from the modENCODE project. The modENCODE accession numbers for H3K27me3 and H3K36me3 in early embryos and germline-containing adults are 6254, 6256, 5163 and 5165. The overlap of *C. elegans* piRNAs with *C. elegans* chromatin domains in early embryos and L3 larvae (Evans et al., 2016) was calculated with Bedtools v2.25.0 (Quinlan and Hall, 2010). To estimate chromatin environment of piRNA loci in other species, 1:1 orthologs were selected as described above, and annotated for chromatin state based on the state of the *C. elegans* ortholog. We tested the association of piRNA-containing genes with particular chromatin domain types with a Fisher's exact test. We defined the "chromatin enrichment score" as the log$_2$ of the odds ratio of a 2x2 table of genes with and without piRNA locus association in active versus regulated chromatin. The germ line expression of these groups of genes in *C. elegans* was assessed using RNA-seq from dissected germ line tissue (Ortiz et al., 2014), and differences in expression were tested with a Wilcoxon rank sum test.

The localisation of piRNAs relative to genes and intronic regions in each genome was calculated using bedtools intersect v2.25.0. We randomly simulated piRNA positions genome-wide to generate null distributions, and simulated piRNA positions within their contig of origin to account for potential regional biases in those genomes We defined a "genicness score" for each species as the genic/intergenic ratio of piRNA locus positioning in the real dataset divided by the median of the 100 genic/intergenic ratios of simulated datasets.

Expression of artificial piRNA loci

We used CRISPR-Cas9 to insert an artificial piRNA locus with a 21U-RNA sequence absent from the *C. elegans* genome. We based the artificial piRNA on the endogenous piRNA 21UR-3922 (IV:15671231-15671251) and edited the mature piRNA sequence such that it had no match to the *C. elegans* genome (TCGGATCGGGTCATACCGGAT > TCGGATCGGGTCATACGCGAT), leaving the upstream and downstream parts the same (mature 21-RNA sequence in capital letters):

aaaaaattttgtaatgtttcacatttaccataaaattgtcctaatttaaaactgaattgaTCGGATCGGGTCATACGCGATttaaaacattaaatgtgtat

This construct was inserted into a regulated domain (IV:5583983), or into the first intron of *htp-1* within active chromatin (IV:5391471) by injecting preformed Cas9-gRNA complex along with the appropriate repair oligo (Table 1), with *dpy-10* as a coinjection marker (Paix et al., 2015). Correct insertions were verified by PCR (Table 1) and Sanger sequencing.

We modified the region immediately downstream of the mature piRNA sequence to increase its GC content, and thus the melting temperature of the region, and inserted this new locus (21UR-3922m "bad terminator" into the same location within the regulated domain (IV:5583983).

Two strains carrying each insertion representing independent insertion events were generated. The strains were synchronised by hypochlorite treatment and grown from L1 for 72 h for RNA isolation, and RPPH libraries were prepared as described above. We recovered 21U-RNA and piRNA precursor counts for each piRNA locus. We used DEseq2 to estimate library size factors and identify differentially expressed 21U-RNAs. Additionally, we normalised the mature piRNA and precursor read counts of the artificial locus relative to those of the endogenous locus.

Repair oligo htp1(active chromatin)	CGGATTCCATTGATGATAAGGTTTTTTTCATGGTTAAAAAATTTTGTAATGTTTCACATTTACCATAAAATTGTCCTAATTTAAAACTGAATTGATCGGATCGGGTCATACGCGATTTAAAACATTAAATGTGTATTTATGGCCATTGTAATTTTTTTCATCTTTAGAAAT
Repair oligo regulated chromatin	AACATAAACAGTCTTTGAATTCACTATTTTTGCGTAAAAAATTTTGTAATGTTTCACATTTACCATAAAATTGTCCTAATTTAAAACTGAATTGATCGGATCGGGTCATACGCGATTTAAAACATTAAATGTGTATTATTGGAATTGAAGTTAAAAAATTTTGAAAAACGA
Repair oligo regulated chromatin “bad terminator”	AACATAAACAGTCTTTGAATTCACTATTTTTGCGTAAAAAATTTTGTAATGTTTCACATTTACCATAAAATTGTCCTAATTTAAAACTGAATTGATCGGATCGGGTCATACGCGATCGACACCAGCGGGACTGTGTATTATTGGAATTGAAGTTAAAAAATTTTGAAAAACGA
gRNA htp-1	AGGTTTTTTTCATGGTTTTA
gRNA regulated chromatin	ATTCACTATTTTTGCGTTAT
Htp-1 PCR primer F	TGATCGGATCGGGTCATACG
Htp-1 PCR primer R	TCCAATCGAAAATTAGCCGTGA
Regulated chromatin PCR primer F	CAATTTCTATGGAGGCATTTGTCC
Regulated chromatin PCR primer R	CCACAATTCAAATTAGGAACCCG

Effect of SNPs on piRNA motif scores

We downloaded genomic variation data for *C. elegans* from the *C. elegans* natural diversity resource (CeNDR, https://www.elegansvariation.org/,(Cook et al., 2017). *C. briggsae* SNPs were taken from (Thomas et al., 2015). *P. pacificus* SNPs were downloaded from the *Pristionchus* variome site (http://www.pristionchus.org/variome/, Rödelsperger et al., 2014). To establish the impact of SNPs on piRNA motifs, we identified Ruby motifs using the specific position weight matrices (PWM) and selected motifs with a score >=500 log odds ratio and recalculated the motif scores after substituting any SNPs. For piRNA loci with a log odds ratio >1000, we identified SNPs with a predicted effect of >=400 change in log odds score (*C. elegans*, *C. briggsae*) or >=200 for *P. pacificus* (due to lower motif information content). We determined whether SNPs affecting piRNAs were major alleles (present in >90% of strains) or minor alleles (present in <10% of strains). We explored their association with annotated piRNA clusters, chromatin domains and genes using Bedtools for data integration.

T_m analysis of piRNA loci and their effects in piRNA transcription

Sequences 200 bp upstream and downstream of piRNA motifs were extracted using the Bedtools module getfasta v2.25.0, and predicted DNA-DNA melting temperature (T_m) profiles were calculated using EMBOSS dan v6.6.0 with a window size of 9 nucleotides, a window shift of 1 nucleotide and default DNA and salt concentrations. Stratification of piRNA loci was carried out by calculating the strength of the T_m profile as the positive difference of T_m values to background at the high T_m region (TSS to +19), plus the negative difference of T_m values to background at the low T_m region (+20 to +40). Background was defined as the average T_m at -200 to -100 and +100 to +200 from the TSS.

Quantification of T_m profiles across genomes is challenging given the differences in GC across genomes that result in differences in background nucleotide composition around piRNAs. We calculated this metric in two ways: first, we used a Bayesian Integral Log-Odds approach (BILD) (Yu et al., 2015), using the genome-wide nucleotide composition as background model. This analysis gives a measure of entropy of the sequence at each position, and captures the piRNA termination region as a peak in entropy. We calculated the strength of termination as the area of this peak. Second, we calculated the difference between the maximum T_m value of the high GC content region and the minimum T_m value of the low GC content region.

To compare expression of piRNAs from loci with different pausing signature strengths we used two independent methods to normalize library size. We first normalised piRNA counts to total non-structural mapped reads (non-rRNA and non-tRNA). Second, we fitted a linear model to miRNA counts across pairs of libraries, and inferred size factors as the estimated value of the slope +/- 1.96 times the standard error of the estimate.

For analysis of piRNA precursors, reads with starts mapping exactly 2 nt upstream of annotated 21Us were selected. For analysis of TFIIS-derived 3' cleavage products fragments, all 21U reads were removed to avoid any potential interference from 21U reads coming from annotated and unannotated piRNA loci. Uniquely mapping reads longer than 10 bases mapping at positions +28 to +38 to the TSS were counted in each replicate and normalised to total non-structural mapped reads. The total number of unique sequences was quantified and normalised to the total number of unique non-structural mapped reads.

We tested the difference in piRNA precursor length at CIP-RPPH libraries by a Wilcoxon rank sum test.

We used the length distributions of rRNA degradation products (reads mapping sense to rRNA) as a control to rule out variability across libraries due to size selection during library preparation. To compare the lengths of N precursors from library 1 and M precursors from library 2, we sampled N and M rRNA structural reads longer than 23 bases from each library, and calculated a Wilcoxon rank sum test p-value. We repeated the procedure 1000 times to calculate a null distribution of p-values. We noted that differences in sequencing depth can bias the length distribution of unique piRNA precursor sequences, since longer precursors are less abundant. To control for this, we generated 10000 subsamples of 2500 precursor sequences from N2-TFIIS library pairs (3 biological replicates), by weighted sampling of piRNA precursors according to their abundance (number of reads). We generated a distribution of p-values by a one-tailed Wilcoxon rank sum test (TFIIS>N2) and compared this distribution to a null distribution of p-values generated by comparing pairs of subsets of precursors sampled from the N2 library for each biological replicate.

We analysed the parn-1 (tm869) mutant data as described above, with exception of a sample size increase to 5000 piRNA precursor sequences, due to increased sequencing depth in these samples resulting in the detection of a higher number of unique precursor sequences.

Data availability

Small RNA sequencing data has been submitted to the SRA Study SRP117954: Evolution of piRNAs in nematodes
The *P. oxycercus* and *P. sambesii* genomes have been uploaded to http://caenorhabditis.org/.

Figure Legends

Figure 1. Overview of piRNA evolution in nematodes.

Assesment of Piwi protein presence/absence, Ruby motif presence/absence, and number of piRNA loci with upstream motifs and unless otherwise indicated, 21U-RNA reads detected by high-throughput sequencing of small RNAs. Nematode phylogeny was taken from Koutsovolos and Blaxter, 2016.

Figure 2. The nematode piRNA upstream motif evolved from the SNAPc binding motifs.

A. Ruby and SNPC-4 motifs in *C. elegans* and *Plectus sambesii*.

B. Evolution of SNPC-4 motifs defined upstream of annotated snRNA loci across nematodes with and without piRNAs.

Figure 3. Two modes of piRNA locus organization in nematodes

A. Genomic distribution of piRNAs in *P. pacificus* (dispersed piRNAs) and *C. elegans* (clustered piRNAs).

B. Mapping of piRNA regions in multiple nematode species to the *C. elegans* genome. Odds ratio (OR) and Fisher's exact test p-value for X chromosome depletion is tabulated on the right hand side.

C. Cumulative number of piRNAs against cumulative span of contigs ranked by the significance of piRNA enrichment. Examples of clustered species (*C. elegans*, *C. brenneri*) and non-clustered species (*P. pacificus*, *P. murrayi*).

D. Discrimination plot of nematode genomes based on the number of piRNAs and the span of piRNA regions. Species were classified as clustered (blue dots) or non-clustered (red dots) on the basis of mapping of piRNA regions to *C. elegans* in B. Best-fit lines are from bimodal regression analysis using all points as input.

Figure 4. piRNAs in *C. elegans* and *P. pacificus* occupy different chromatin environments.

A. *C. elegans* piRNA cluster regions are organised in multiple subclusters of piRNAs associated with H3K27me3 chromatin and interrupted by H3K36me3 chromatin domains depleted of piRNAs.

B. *P. pacificus* piRNA loci are found within the introns of genes within H3K36me3 chromatin domains.

C. Germ line expression of genes overlapping with piRNAs in *C. elegans* and *P. pacificus*.

D. Correlation between gene expression and piRNA expression in *C. elegans* and *P. pacificus*.

E. The fraction of Ruby motifs expressed in different chromatin environments in *C. elegans* and *P. pacificus*.

F. Design and generation of artificial piRNA loci by genome-editing.

G. Expression of an artificial piRNA placed in either regulated or active chromatin compared to the endogenous locus.

H. Two fundamental modes of piRNA organisation in nematodes. Discrimination plot based on genic/intergenic proportions of piRNAs and mappings of 1:1 orthologs to *C. elegans* chromatin domains. Species are coloured according to the designation in Figure 3.

Figure 5. Selection at piRNAs is dependent on their chromatin environment.

A-B. Fraction of high-scoring Ruby motifs that are the major allele (present in >90% of strains) in *C. elegans* (A) and *C. briggsae* (B).

C. Frequency of the major allele of high-scoring Ruby motifs across different chromatin environments in *C. elegans*.

D. Fraction of high-scoring Ruby motifs that are the major allele in regions of low and high piRNA density in *P. pacificus.*

E. Fraction of Ruby motifs that are the major allele in genes stratified according to expression in *C. elegans* and *P. pacificus*.

Figure 6. Transcription of short capped piRNA precursors in nematodes occurs through pausing of RNA polymerase II.

A-B. Melting temperature profiles around piRNA transcription start sites in *C. elegans* and *P. pacificus*.

C. Relationship between strength of pausing-associated sequence signature and mature piRNA levels.

D-E. Length and abundance of piRNA precursors produced from loci with strong (top 10%) or weak (bottom 10%) pausing-associated signatures in *C. elegans* assessed by high-throughput sequencing of capped RNAs less than 50bp in size. D shows WT and E shows *prg-1* mutants.

F. Longer nuclear Capped RNAs (40-60bp) emanating from piRNA loci with strong (top 10%) or weak (bottom 10%) pausing-associated sequence signatures assessed by high-throughput sequencing (compare with D and E).

G. Design and generation of artificial piRNA loci with good and bad terminator sequences.

H. Expression of the artificial piRNA loci with good and bad terminators.

I. Location of the low melting temperature region downstream of piRNA loci. The distance to the centre of the low T_m "valley" from the piRNA transcription start site is shown.

J. piRNA precursor length in *C. elegans* and *Plectus sambesii.*

Figure 7. Chromatin and RNA Pol II pausing together control nematode piRNA biogenesis.

A. piRNA precursor length in TFIIS mutants compared to wild type.

B. Mature piRNAs in TFIIS mutants compared to wild type.

C. 3’ degradation fragments of piRNA precursors in TFIIS compared to WT *C. elegans*. 3’ degradation fragments are defined as 5’ monophosphate small RNAs whose 5’ end maps between 28 and 38 downstream of the piRNA transcription start site.

D. Model for how Pol II pausing contributes to piRNA biogenesis. The AT-rich region downstream of the 21U sequence causes Pol II to destabilise and stall, leading to backtracking and TFIIS-mediated cleavage of the 3’ ends of nascent RNAs. This leads to a decrease in piRNA precursor length and an increase in the abundance of 3’ degradation fragments, and influences 21U levels by altering Pol II dynamics at piRNA promoters.E-F. Relationship between the average strength of pausing signal and enrichment of piRNAs in active chromatin (A) and enrichment of piRNAs in genic regions (B) across different nematode species.

G. Model for piRNA transcription in C-type and P-type nematodes.

H. Widespread phylogenetic distribution of active-genic and regulated-intergenic piRNAs.

Supplementary Figure Legends

Figure S1. Further analysis of piRNA transcription (relates to Figure 1)

A. Transcription initiation at piRNA loci relative to mature 21U sequences across nematodes. 5’ extensions of mapped piRNA precursor transcripts are shown.

B. Co-occurrence of GTTTC and AACCCGTC motifs in *Plectus sambesii* and *Plectus murrayi* but not in *C. elegans*.

C. *Plectus sambesii* piRNA motifs and SNPC-4 binding sites upstream of U6 snRNAs are the same element. Conserved positions of this motif with the *C. elegans* SNPC-4 binding motif are marked with asterisks.

Figure S2-1. SNAPc and Ruby motif evolution across nematodes (Relates to Figure 2)

A. Sequence similarities between the 3’ SNAPc and Ruby motifs in several nematode species.

Figure S2-2. Full tree showing the evolution of the SNAPc motif in nematodes (Relates to Figure 2)

Figure S3. Validation of piRNA clustering analysis methods (Relates to Figure 3)

A. Dot plot showing the alignment of the top 100 contigs enriched in piRNAs in a short-read *C. elegans* genome assembly (see Supplementary tables, *Data sources*) to the *C. elegans* reference genome. These map exclusively to *C. elegans* piRNA clusters in IV.

B. Identification of scaffolds enriched in piRNAs in the *Oscheius tipulae* linkage map.

C. Chromosomal enrichment of piRNAs in *O. tipulae, P. pacificus and H. contortus.*

Figure S4. Regulation of piRNA biogenesis by the chromatin environment (Relates to Figure 4)

A. Size of regulated and active domains within *C. elegans* piRNA regions compared to the rest of the genome.

B. Examples of H3K27me3 and H3K36me3 profiles at piRNA cluster regions in the early embryo and the young adult stage.

C. Germline gene expression of genes containing piRNAs in *C. elegans* or the *C. elegans* orthologues of *P. pacificus* genes containing piRNAs.

D. Comparison of intergenic/genic piRNA proportions to simulated positions.

E. Summary table of annotated piRNA loci and scanned Ruby motifs and their chromatin domain localisations in *C. elegans* and *P. pacificus*.

Figure S5. piRNA loci are under selection in both C-type and P-type species (Relates to Figure 5)

A-B. Allele frequency distribution of SNPs predicted to alter piRNA motif strength by >400 in *C. briggsae* and *C. elegans*.

C. Distribution of the predicted changes in motif scores as a result of all SNPs mapping to piRNA loci in *C. briggsae*, *C. elegans* and *P. pacificus*.

D. Enrichment of high scoring motifs within piRNA clusters in *C. elegans* and *C. briggsae* using the lower threshold of a >200 change in strength, compared to the difference observed between the 10 highest density contigs and the 10 lowest density contigs in *P. pacificus*.

Figure S6. RNA Pol II pausing signals are important for piRNA biogenesis (Relates to Figure 6)

A. Entropy sequence logos of piRNA loci in *C. elegans* and *P. pacificus* using either genomic GC content or equiprobable nucleotide frequency as the background model.

B. Nucleotide frequency logos of piRNA loci in *C. elegans* and *P. pacificus*.

C. Effect of the pausing signal downstream of the 21U sequence on piRNA abundance.

D-F. Dissection of the effect of different parts of the pausing-associated sequence signal on piRNA precursor length and abundance.

Figure S7-1. Loss of function of TFIIS modulates piRNA biogenesis in *C. elegans* (Relates to Figure 7)

A-D. Comparison of unique and total piRNA precursor reads length distributions in N2 and TFIIS mutant *C. elegans*. The length distribution of rRNA degradation fragment unique and total reads was plotted as a control. P-values for the increase in length were calculated using a Wilcoxon rank sum test, and were compared against a null distribution of p-values calculated by sampling sets of a matching number of unique or total reads from the rRNA length distributions (sampling reads >23 nt). In TFIIS-N2 comparisons, the real p-value is much lower than the calculated null distribution, while in comparisons between two independent N2 libraries, the real p-value is greater, showing that the differences in precursor length are not a result of a global shift in read length in the libraries due to technical variation.

E. Length distribution of tailed 22G-RNAs in N2 and TFIIS mutants.

Figure S7-2. Further analysis of piRNAs in TFIIS mutants controlling for sequencing depth (Relates to Figure 7)

A. Length distribution of piRNA precursors in subsets of 2500 precursors sampled according to relative abundance in N2 and tfiis(ok2749) mutants.

B. P-value distributions of a Wilcoxon rank sum test for a difference in length, in 10,000 N2 vs TFIIS random subsets (blue) and N2-N2 random subsets (pink).

Figure S7-3. Analysis of piRNA precursors in the absence of trimming

A. Length distribution of piRNA precursors in subsets of 5000 precursors sampled according to relative abundance in N2 compared to parn-1(tm869) mutants.

B. piRNA precursor length distributions comparing the top and bottom 10% of piRNA loci stratified according to the strength of the pausing-associated sequence signature.

C. Length distribution of piRNA precursors in subsets of 5000 precursors sampled according to relative abundance in parn-1(tm869) mutants compared to parn-1(tm869) tfiis(ok2749) double mutants.

Figure S7-4. Entropy measurement of the strength of RNA Pol II pausing-associated sequence signatures across nematodes (Relates to Figure 8)

A-B. Sequence logos downstream of piRNA loci in several nematodes.
C. Melting temperature profiles and Bayesian Integral Log-Odds (BILD) entropy of the sequence around piRNA loci across nematodes. The strength of the pausing signature was calculated as the difference between the maximum value of the high GC content peak and the lowest value of the high AT content region from the Tm profiles, and alternatively as the area under the peak downstream of the 21U relative to background from the entropy profiles.
D. Relationship between the different entropy measures used (Spearman $r > 0.9$).
E. Relationship between entropy measures used and GC content (Spearman $r < 0.1$).

Supplementary tables

Table S1. Data sources

Table S2. piRNA organisation across nematode genomes

Table S3. piRNAs and chromatin domains in *C. elegans* and *P. pacificus*

Table S4. Mature piRNA abundance and pausing-associated sequence signature scores at piRNA loci.

Bibliography

Bagijn, M.P., Goldstein, L.D., Sapetschnig, A., Weick, E.M., Bouasker, S., Lehrbach, N.J., Simard, M.J., and Miska, E.A. (2012). Function, Targets, and Evolution of Caenorhabditis elegans piRNAs. Science *337*, 574–578.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research *37*, W202–W208.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology *19*, 455–477.

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

Billi, A.C., Freeberg, M.A., Day, A.M., Chun, S.Y., Khivansara, V., and Kim, J.K. (2013). A Conserved Upstream Motif Orchestrates Autonomous, Germline-Enriched Expression of Caenorhabditis elegans piRNAs. PLoS Genet *9*, e1003392.

Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., Vanfleteren, J.R., Mackey, L.Y., Dorris, M., Frisse, L.M., et al. (1998). A molecular evolutionary framework for the phylum Nematoda. Nature *392*, 71–75.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell *128*, 1089–1103.

Cook, D.E., Zdraljevic, S., Roberts, J.P., and Andersen, E.C. (2017). CeNDR, the Caenorhabditis elegans natural diversity resource. Nucleic Acids Research *45*, D650–D657.

Das, P.P., Bagijn, M.P., Goldstein, L.D., Woolford, J.R., Lehrbach, N.J., Sapetschnig, A., Buhecha, H.R., Gilchrist, M.J., Howe, K.L., Stark, R., et al. (2008). Piwi and piRNAs Act Upstream of an Endogenous siRNA Pathway to Suppress Tc3 Transposon Mobility in the

Caenorhabditis elegans Germline. Molecular Cell *31*, 79–90.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology *16*, 157.

Evans, K.J., Huang, N., Stempor, P., Chesney, M.A., Down, T.A., and Ahringer, J. (2016). Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. Proceedings of the National Academy of Sciences.

Gressel, S., Schwalb, B., Decker, T.M., Qin, W., Leonhardt, H., Eick, D., and Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. eLife *6*.

Gruchota, J., Denby Wilkes, C., Arnaiz, O., Sperling, L., and Nowak, J.K. (2017). A meiosis-specific Spt5 homolog involved in non-coding transcription. Nucleic Acids Research *45*, 4722–4732.

Gu, W., Lee, H.-C., Chaves, D., Youngman, E.M., Pazour, G.J., Conte, D., Jr, and Mello, C.C. (2012). CapSeq and CIP-TAP Identify Pol II Start Sites and Reveal Capped SmallRNAs as C. elegans piRNA Precursors. Cell *151*, 1488–1500.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics *32*, 767–769.

Hung, K.-H., and Stumph, W.E. (2011). Regulation of snRNA gene expression by the Drosophila melanogastersmall nuclear RNA activating protein complex (DmSNAPc). Critical Reviews in Biochemistry and Molecular Biology *46*, 11–26.

Izumi, N., Shoji, K., Sakaguchi, Y., Honda, S., Kirino, Y., Suzuki, T., Katsuma, S., and Tomari, Y. (2016). Identification and Functional Analysis of the Pre-piRNA 3. Cell *164*, 962–973.

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics *15*, 182.

Karakuzu, O., Wang, D.P., and Cameron, S. (2009). MIG-32 and SPAT-3A are PRC1 homologs that control neuronal migration in Caenorhabditis elegans. Development *136*, 943–953.

Kasper, D.M., Wang, G., Gardner, K.E., Johnstone, T.G., and Reinke, V. (2014). The C. elegans SNAPc component SNPC-4 coats piRNA domains and is globally required for piRNA abundance. Developmental Cell *31*, 145–158.

Kiuchi, T., Koga, H., Kawamoto, M., Shoji, K., Sakai, H., Arai, Y., Ishihara, G., Kawaoka, S., Sugano, S., Shimada, T., et al. (2014). A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature *509*, 633–636.

Krebs, A.R., Imanci, D., Hoerner, L., Gaidatzis, D., Burger, L., and Schübeler, D. (2017). Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. Molecular Cell 1–17.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet *4*, 237.

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

Miesen, P., Girardi, E., and van Rij, R.P. (2015). Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. Nucleic Acids Research *43*, 6545–6556.

Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes & Development *25*, 742–754.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

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

Ortiz, M.A., Noble, D., Sorokin, E.P., and Kimble, J. (2014). A new dataset of spermatogenic vs. oogenic transcriptomes in the nematode Caenorhabditis elegans. G3 (Bethesda) *4*, 1765–1772.

Paix, A., Folkmann, A., Rasoloson, D., and Seydoux, G. (2015). High Efficiency, Homology-Directed Genome Editing in Caenorhabditis elegans Using CRISPR-Cas9 Ribonucleoprotein Complexes. Genetics *201*, 47–54.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.

Pryszcz, L.P., and Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Research *44*, e113.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rödelsperger, C., Meyer, J.M., Prabh, N., Lanz, C., Bemm, F., and Sommer, R.J. (2017). Single-Molecule Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode Model Organism Pristionchus pacificus. CellReports *21*, 834–844.

Rödelsperger, C., Neher, R.A., Weller, A.M., Eberhardt, G., Witte, H., Mayer, W.E., Dieterich, C., and Sommer, R.J. (2014). Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data. Genetics *196*, 1153–1165.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in C. elegans. Cell *127*, 1193–1207.

Sarkies, P., SELKIRK, M.E., Jones, J.T., Blok, V., Boothby, T., Goldstein, B., Hanelt, B., Ardila-Garcia, A., Fast, N.M., Schiffer, P.M., et al. (2015). Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages. Plos Biol *13*, e1002061.

Schweikhard, V., Meng, C., Murakami, K., Kaplan, C.D., Kornberg, R.D., and Block, S.M. (2014). Transcription factors TFIIF and TFIIS promote transcript elongation by RNA polymerase II by synergistic and independent mechanisms. Proceedings of the National Academy of Sciences *111*, 6642–6647.

Siomi, M.C., Sato, K., Pezic, D., and Aravin, A.A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nature Publishing Group *12*, 246–258.

Tang, W., Tu, S., Lee, H.-C., Weng, Z., and Mello, C.C. (2016). The RNase PARN-1 Trims piRNA 3' Ends to Promote Transcriptome Surveillance in C. elegans. Cell *164*, 974–984.

Thomas, C.G., Wang, W., Jovelin, R., Ghosh, R., Lomasko, T., Trinh, Q., Kruglyak, L., Stein, L.D., and Cutter, A.D. (2015). Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. Genome Research *25*, 667–678.

Weick, E.M., Sarkies, P., Silva, N., Chen, R.A., Moss, S.M.M., Cording, A.C., Ahringer, J., Martinez-Perez, E., and Miska, E.A. (2014). PRDE-1 is a nuclear factor essential for the biogenesis of Ruby motif-dependent piRNAs in C. elegans. Genes & Development *28*, 783–796.

Weick, E.-M., and Miska, E.A. (2014). piRNAs: from biogenesis to function. Development *141*, 3458–3471.

Yu, Y.-K., Capra, J.A., Stojmirović, A., Landsman, D., and Altschul, S.F. (2015). Log-odds sequence logos. Bioinformatics *31*, 324–331.

Zamudio, J.R., Kelly, T.J., and Sharp, P.A. (2014). Argonaute-bound small RNAs from promoter-proximal RNA polymerase II. Cell *156*, 920–934.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research *18*, 821–829.

Figure 1

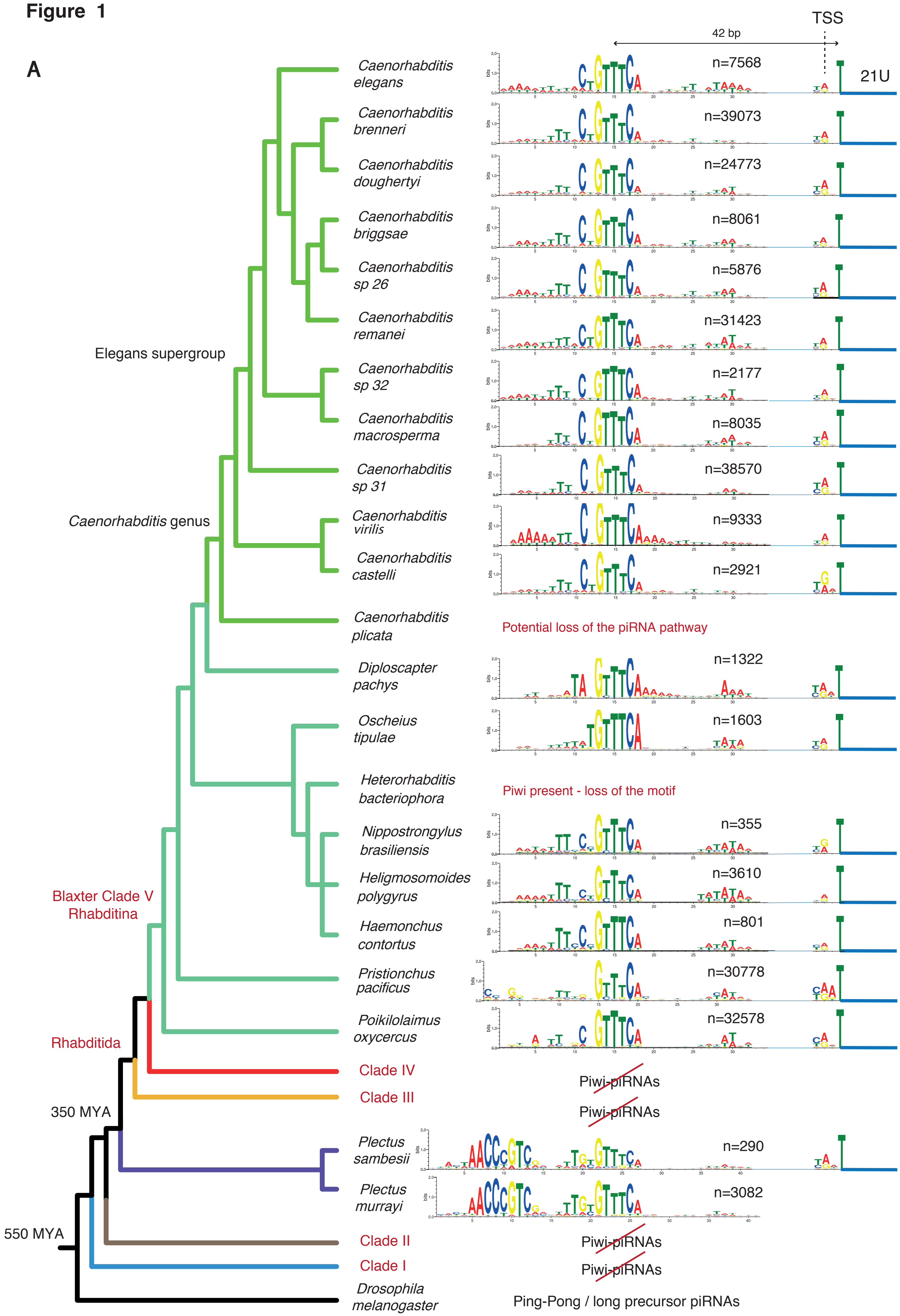

A
TSS
42 bp
21U
Caenorhabditis elegans
n=7568
Caenorhabditis brenneri
n=39073
Caenorhabditis doughertyi
n=24773
Caenorhabditis briggsae
n=8061
Caenorhabditis sp 26
n=5876
Caenorhabditis remanei
n=31423
Elegans supergroup
Caenorhabditis sp 32
n=2177
Caenorhabditis macrosperma
n=8035
Caenorhabditis sp 31
n=38570
Caenorhabditis genus
Caenorhabditis virilis
n=9333
Caenorhabditis castelli
n=2921
Caenorhabditis plicata
Potential loss of the piRNA pathway
Diploscapter pachys
n=1322
Oscheius tipulae
n=1603
Heterorhabditis bacteriophora
Piwi present - loss of the motif
Nippostrongylus brasiliensis
n=355
Heligmosomoides polygyrus
n=3610
Blaxter Clade V Rhabditina
Haemonchus contortus
n=801
Pristionchus pacificus
n=30778
Poikilolaimus oxycercus
n=32578
Rhabditida
Clade IV
Piwi-piRNAs
Clade III
Piwi-piRNAs
350 MYA
Plectus sambesii
n=290
Plectus murrayi
n=3082
550 MYA
Clade II
Piwi-piRNAs
Clade I
Piwi-piRNAs
Drosophila melanogaster
Ping-Pong / long precursor piRNAs
bits
2.0
1.0
0.0
5
10
15
20
25
30
35
40

Figure 2

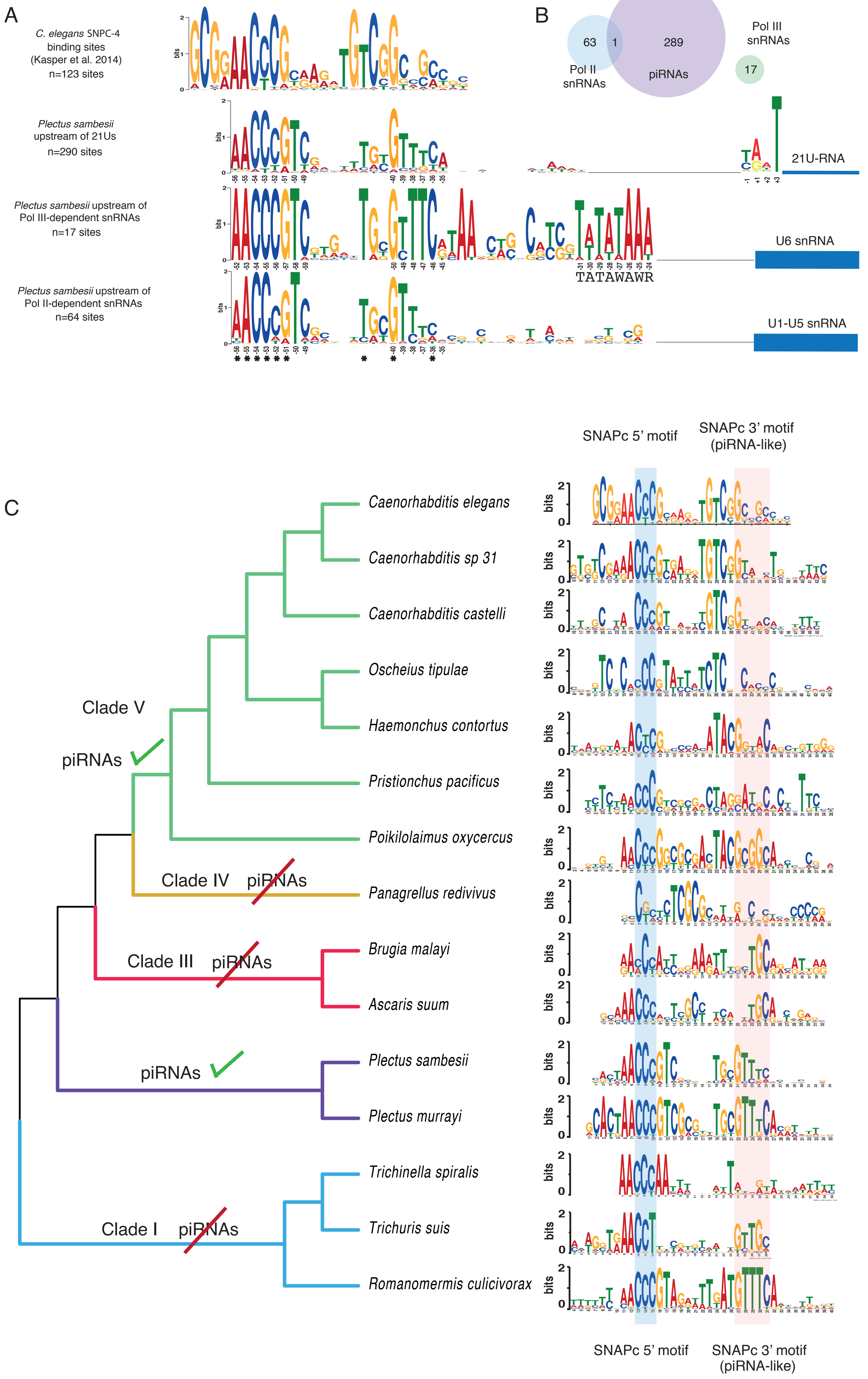

A
C. elegans SNPC-4 binding sites (Kasper et al. 2014) n=123 sites
Plectus sambesii upstream of 21Us n=290 sites
Plectus sambesii upstream of Pol III-dependent snRNAs n=17 sites
Plectus sambesii upstream of Pol II-dependent snRNAs n=64 sites
TATAWAWR
21U-RNA
U6 snRNA
U1-U5 snRNA
bits
B
63
1
289
17
Pol II snRNAs
piRNAs
Pol III snRNAs
C
SNAPc 5' motif
SNAPc 3' motif (piRNA-like)
Clade V
piRNAs ✓
Clade IV
piRNAs
Clade III
piRNAs
piRNAs ✓
Clade I
piRNAs
Caenorhabditis elegans
Caenorhabditis sp 31
Caenorhabditis castelli
Oscheius tipulae
Haemonchus contortus
Pristionchus pacificus
Poikilolaimus oxycercus
Panagrellus redivivus
Brugia malayi
Ascaris suum
Plectus sambesii
Plectus murrayi
Trichinella spiralis
Trichuris suis
Romanomermis culicivorax
SNAPc 5' motif
SNAPc 3' motif (piRNA-like)

Figure 3

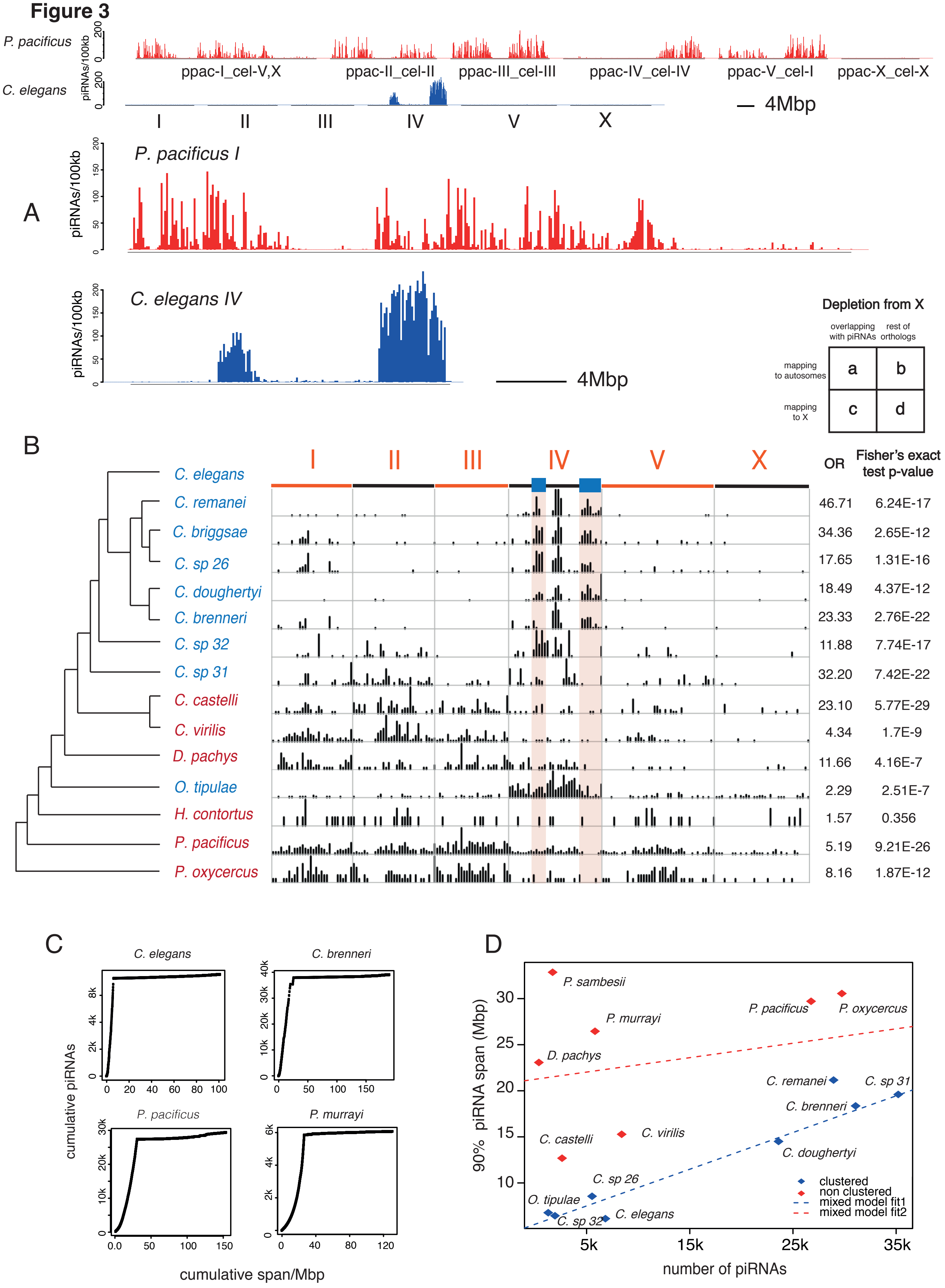

P. pacificus
C. elegans
piRNAs/100kb
200
0
ppac-I_cel-V,X
ppac-II_cel-II
ppac-III_cel-III
ppac-IV_cel-IV
ppac-V_cel-I
ppac-X_cel-X
4Mbp
I
II
III
IV
V
X
A
P. pacificus I
piRNAs/100kb
0
50
100
150
200
C. elegans IV
4Mbp
Depletion from X
overlapping with piRNAs
rest of orthologs
mapping to autosomes
mapping to X
a
b
c
d
B
I
II
III
IV
V
X
OR
Fisher's exact test p-value
C. elegans
C. remanei 46.71 6.24E-17
C. briggsae 34.36 2.65E-12
C. sp 26 17.65 1.31E-16
C. doughertyi 18.49 4.37E-12
C. brenneri 23.33 2.76E-22
C. sp 32 11.88 7.74E-17
C. sp 31 32.20 7.42E-22
C. castelli 23.10 5.77E-29
C. virilis 4.34 1.7E-9
D. pachys 11.66 4.16E-7
O. tipulae 2.29 2.51E-7
H. contortus 1.57 0.356
P. pacificus 5.19 9.21E-26
P. oxycercus 8.16 1.87E-12
C
C. elegans
C. brenneri
P. pacificus
P. murrayi
cumulative piRNAs
cumulative span/Mbp
0
2k
4k
8k
0
20
40
60
80
100
0
10k
20k
30k
40k
0
50
100
150
0
10k
20k
30k
0.0
50
100
150
0
2k
4k
6k
0
40
80
120
D
90% piRNA span (Mbp)
number of piRNAs
10
15
20
25
30
5k
15k
25k
35k
P. sambesii
P. murrayi
P. pacificus
P. oxycercus
D. pachys
C. remanei
C. sp 31
C. brenneri
C. castelli
C. virilis
C. doughertyi
C. sp 26
O. tipulae
C. sp 32
C. elegans
clustered
non clustered
mixed model fit1
mixed model fit2

Figure 4

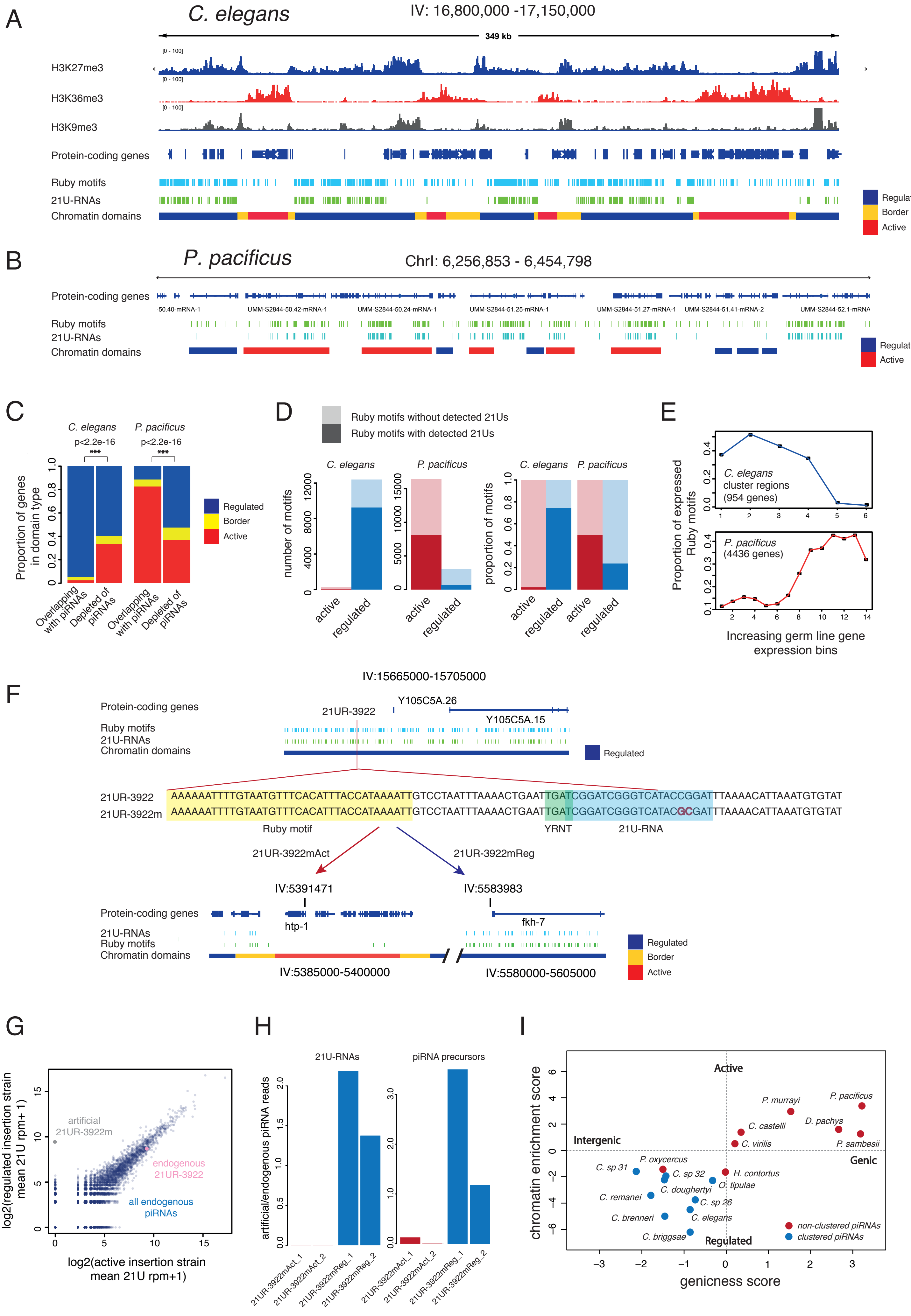

A
C. elegans
IV: 16,800,000 -17,150,000
349 kb
H3K27me3
H3K36me3
H3K9me3
[0 - 100]
Protein-coding genes
Ruby motifs
21U-RNAs
Chromatin domains
Regulated
Border
Active
B
P. pacificus
ChrI: 6,256,853 - 6,454,798
Protein-coding genes
-50.40-mRNA-1
UMM-S2844-50.42-mRNA-1
UMM-S2844-50.24-mRNA-1
UMM-S2844-51.25-mRNA-1
UMM-S2844-51.27-mRNA-1
UMM-S2844-51.41-mRNA-2
UMM-S2844-52.1-mRNA
Ruby motifs
21U-RNAs
Chromatin domains
Regulated
Active
C
C. elegans
P. pacificus
p<2.2e-16
***
Proportion of genes in domain type
0 0.2 0.4 0.6 0.8 1.0
Overlapping with piRNAs
Depleted of piRNAs
Regulated
Border
Active
D
Ruby motifs without detected 21Us
Ruby motifs with detected 21Us
number of motifs
0 4000 8000 12000
0 5000 10000 15000
proportion of motifs
0.0 0.2 0.4 0.6 0.8 1.0
active
regulated
E
Proportion of expressed Ruby motifs
C. elegans cluster regions (954 genes)
P. pacificus (4436 genes)
Increasing germ line gene expression bins
F
IV:15665000-15705000
Protein-coding genes
Ruby motifs
21U-RNAs
Chromatin domains
21UR-3922
Y105C5A.26
Y105C5A.15
Regulated
21UR-3922 AAAAAATTTTGTAATGTTTCACATTTACCATAAAATTGTCCTAATTTAAAACTGAATTGATCGGATCGGGTCATACCGGATTTAAAACATTAAATGTGTAT
21UR-3922m AAAAAATTTTGTAATGTTTCACATTTACCATAAAATTGTCCTAATTTAAAACTGAATTGATCGGATCGGGTCATACGCGATTTAAAACATTAAATGTGTAT
Ruby motif
YRNT
21U-RNA
21UR-3922mAct
21UR-3922mReg
IV:5391471
IV:5583983
htp-1
fkh-7
21U-RNAs
Ruby motifs
IV:5385000-5400000
IV:5580000-5605000
Regulated
Border
Active
G
log2(regulated insertion strain mean 21U rpm+ 1)
log2(active insertion strain mean 21U rpm+1)
artificial 21UR-3922m
endogenous 21UR-3922
all endogenous piRNAs
H
21U-RNAs
piRNA precursors
artificial/endogenous piRNA reads
21UR-3922mAct_1
21UR-3922mAct_2
21UR-3922mReg_1
21UR-3922mReg_2
I
Active
Intergenic
Genic
Regulated
chromatin enrichment score
genicness score
P. murrayi
P. pacificus
D. pachys
C. castelli
C. virilis
P. sambesii
P. oxycercus
C. sp 31
C. sp 32
H. contortus
O. tipulae
C. doughertyi
C. remanei
C. sp 26
C. brenneri
C. elegans
C. briggsae
non-clustered piRNAs
clustered piRNAs

Figure 5

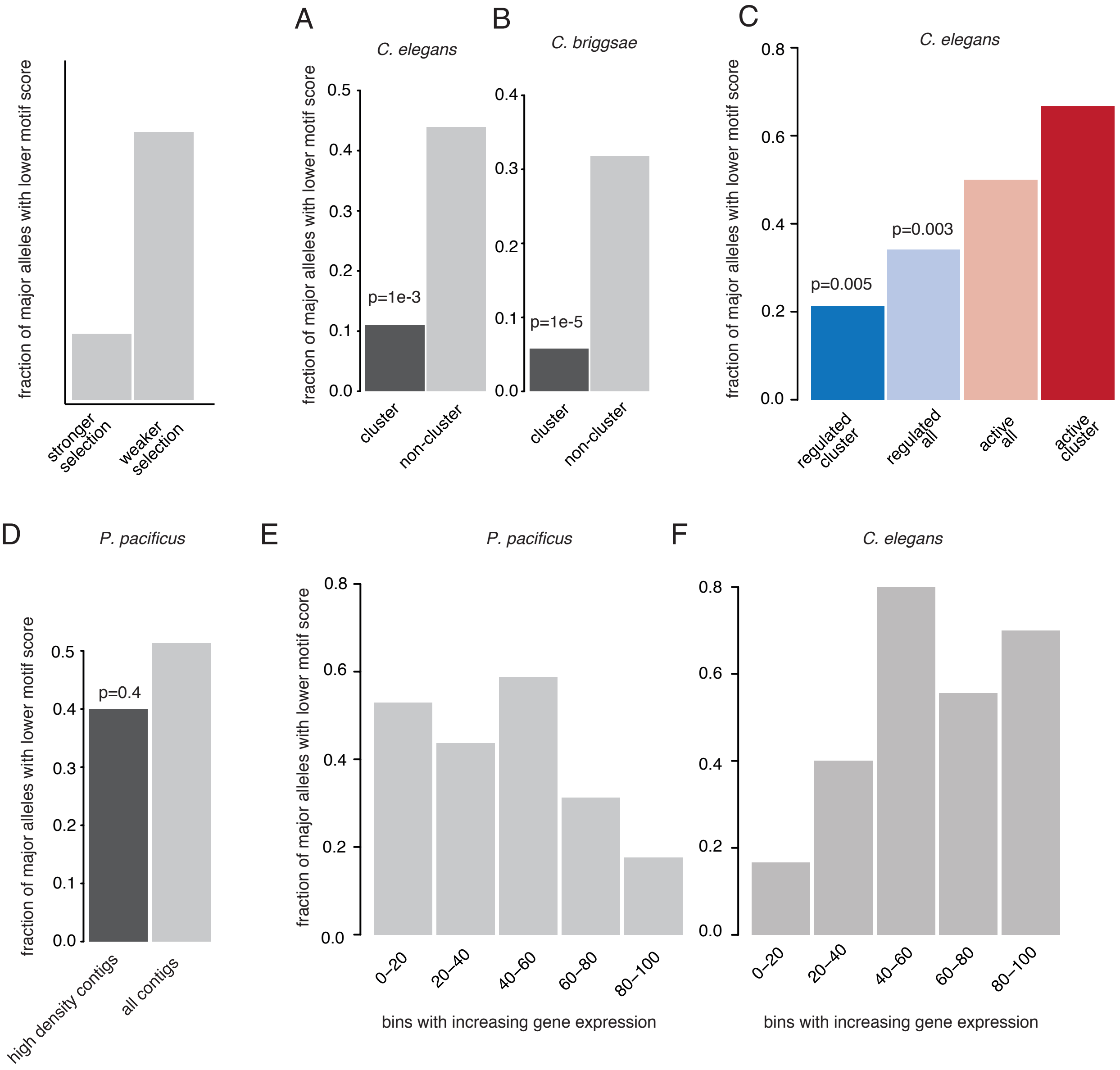

fraction of major alleles with lower motif score
stronger selection
weaker selection
A
C. elegans
fraction of major alleles with lower motif score
0.5
0.4
0.3
0.2
0.1
0.0
p=1e-3
cluster
non-cluster
B
C. briggsae
0.4
0.3
0.2
0.1
0.0
p=1e-5
cluster
non-cluster
C
C. elegans
fraction of major alleles with lower motif score
0.8
0.6
0.4
0.2
0.0
p=0.005
p=0.003
regulated cluster
regulated all
active all
active cluster
D
P. pacificus
fraction of major alleles with lower motif score
0.5
0.4
0.3
0.2
0.1
0.0
p=0.4
high density contigs
all contigs
E
P. pacificus
fraction of major alleles with lower motif score
0.8
0.6
0.4
0.2
0.0
0–20
20–40
40–60
60–80
80–100
bins with increasing gene expression
F
C. elegans
0.8
0.6
0.4
0.2
0.0
0–20
20–40
40–60
60–80
80–100
bins with increasing gene expression

Figure 6

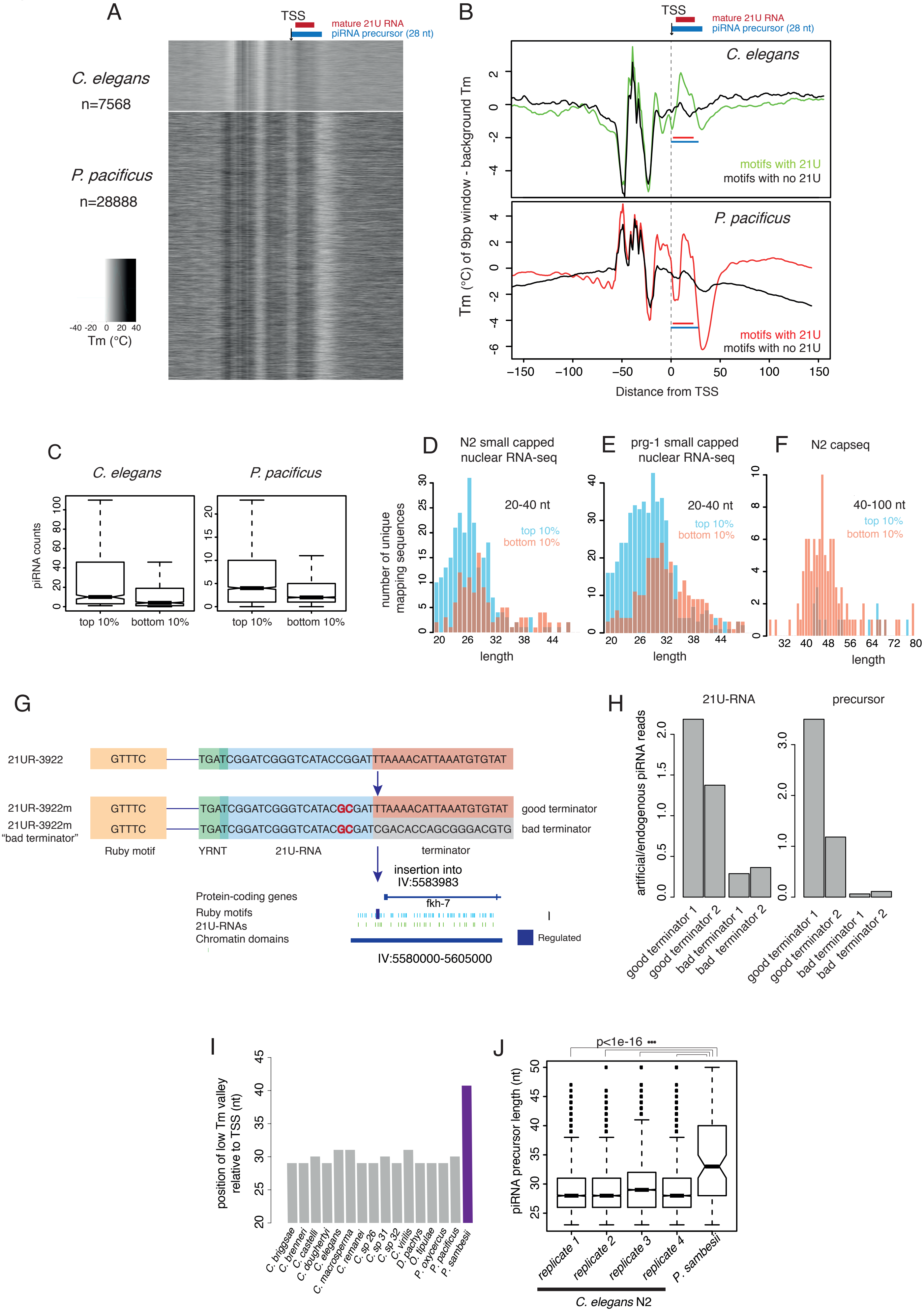

A
TSS
mature 21U RNA
piRNA precursor (28 nt)
C. elegans
n=7568
P. pacificus
n=28888
-40 -20 0 20 40
Tm (°C)
B
TSS
mature 21U RNA
piRNA precursor (28 nt)
C. elegans
motifs with 21U
motifs with no 21U
P. pacificus
motifs with 21U
motifs with no 21U
Tm (°C) of 9bp window - background Tm
Distance from TSS
−150 −100 −50 0 50 100 150
C
C. elegans
P. pacificus
piRNA counts
top 10%
bottom 10%
top 10%
bottom 10%
D
N2 small capped nuclear RNA-seq
20-40 nt
top 10%
bottom 10%
number of unique mapping sequences
length
E
prg-1 small capped nuclear RNA-seq
20-40 nt
top 10%
bottom 10%
length
F
N2 capseq
40-100 nt
top 10%
bottom 10%
length
G
21UR-3922
GTTTC
TGATCGGATCGGGTCATACCGGATTTAAAACATTAAATGTGTAT
21UR-3922m
GTTTC
TGATCGGATCGGGTCATACGCGATTTAAAACATTAAATGTGTAT
good terminator
21UR-3922m "bad terminator"
GTTTC
TGATCGGATCGGGTCATACGCGATCGACACCAGCGGGACGTG
bad terminator
Ruby motif
YRNT
21U-RNA
terminator
insertion into IV:5583983
Protein-coding genes
Ruby motifs
21U-RNAs
Chromatin domains
fkh-7
Regulated
IV:5580000-5605000
H
21U-RNA
precursor
artificial/endogenous piRNA reads
good terminator 1
good terminator 2
bad terminator 1
bad terminator 2
good terminator 1
good terminator 2
bad terminator 1
bad terminator 2
I
position of low Tm valley relative to TSS (nt)
C. briggsae
C. brenneri
C. castelli
C. doughertyi
C. elegans
C. macrosperma
C. remanei
C. sp 26
C. sp 31
C. sp 32
C. virilis
D. pachys
O. tipulae
P. oxycercus
P. pacificus
P. sambesii
J
p<1e-16 ***
piRNA precursor length (nt)
replicate 1
replicate 2
replicate 3
replicate 4
P. sambesi
C. elegans N2

Figure 7

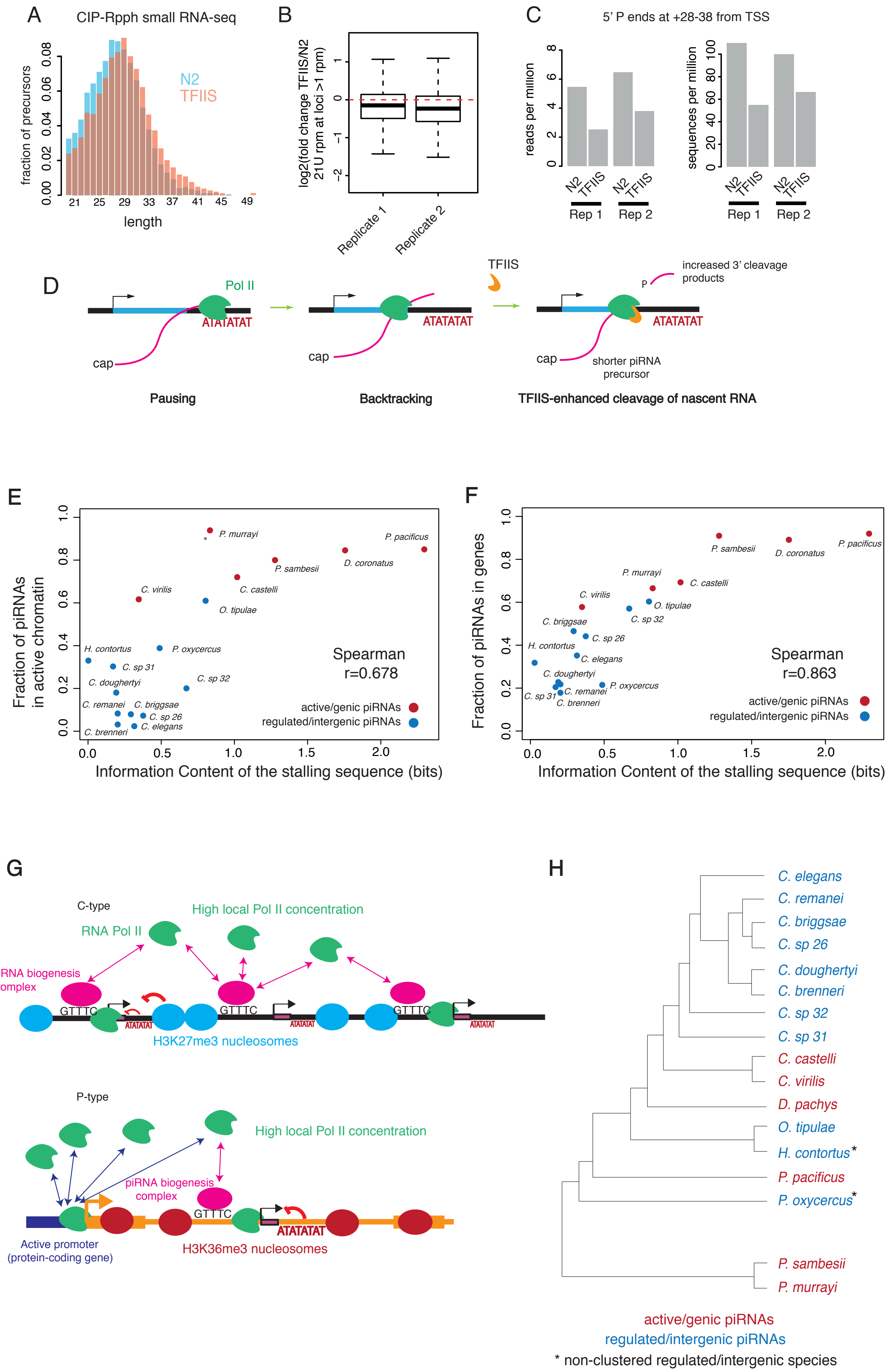

A
CIP-Rpph small RNA-seq
N2
TFIIS
fraction of precursors
0.00 0.02 0.04 0.06 0.08
21 25 29 33 37 41 45 49
length
B
log2(fold change TFIIS/N2
21U rpm at loci >1 rpm)
−2 −1 0 1
Replicate 1
Replicate 2
C
5' P ends at +28-38 from TSS
reads per million
0 2 4 6 8
sequences per million
0 20 40 60 80 100
N2
TFIIS
Rep 1
Rep 2
D
Pol II
ATATATAT
cap
TFIIS
P
increased 3' cleavage products
shorter piRNA precursor
Pausing
Backtracking
TFIIS-enhanced cleavage of nascent RNA
E
Fraction of piRNAs in active chromatin
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.5 1.0 1.5 2.0
Information Content of the stalling sequence (bits)
P. murrayi
*
P. pacificus
P. sambesii
D. coronatus
C. castelli
C. virilis
O. tipulae
H. contortus
P. oxycercus
C. sp 31
C. sp 32
C. doughertyi
C. remanei
C. briggsae
C. sp 26
C. elegans
C. brenneri
Spearman
r=0.678
active/genic piRNAs ●
regulated/intergenic piRNAs ●
F
Fraction of piRNAs in genes
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.5 1.0 1.5 2.0
Information Content of the stalling sequence (bits)
P. sambesii
D. coronatus
P. pacificus
P. murrayi
C. castelli
C. virilis
O. tipulae
C. sp 32
C. briggsae
C. sp 26
H. contortus
C. elegans
C. doughertyi
C. sp 31
C. remanei
P. oxycercus
C. brenneri
Spearman
r=0.863
active/genic piRNAs ●
regulated/intergenic piRNAs ●
G
C-type
RNA Pol II
High local Pol II concentration
piRNA biogenesis complex
GTTTC
ATATATAT
H3K27me3 nucleosomes
P-type
High local Pol II concentration
piRNA biogenesis complex
GTTTC
ATATATAT
Active promoter (protein-coding gene)
H3K36me3 nucleosomes
H
C. elegans
C. remanei
C. briggsae
C. sp 26
C. doughertyi
C. brenneri
C. sp 32
C. sp 31
C. castelli
C. virilis
D. pachys
O. tipulae
H. contortus*
P. pacificus
P. oxycercus*
P. sambesii
P. murrayi
active/genic piRNAs
regulated/intergenic piRNAs
* non-clustered regulated/intergenic species

Figure S1

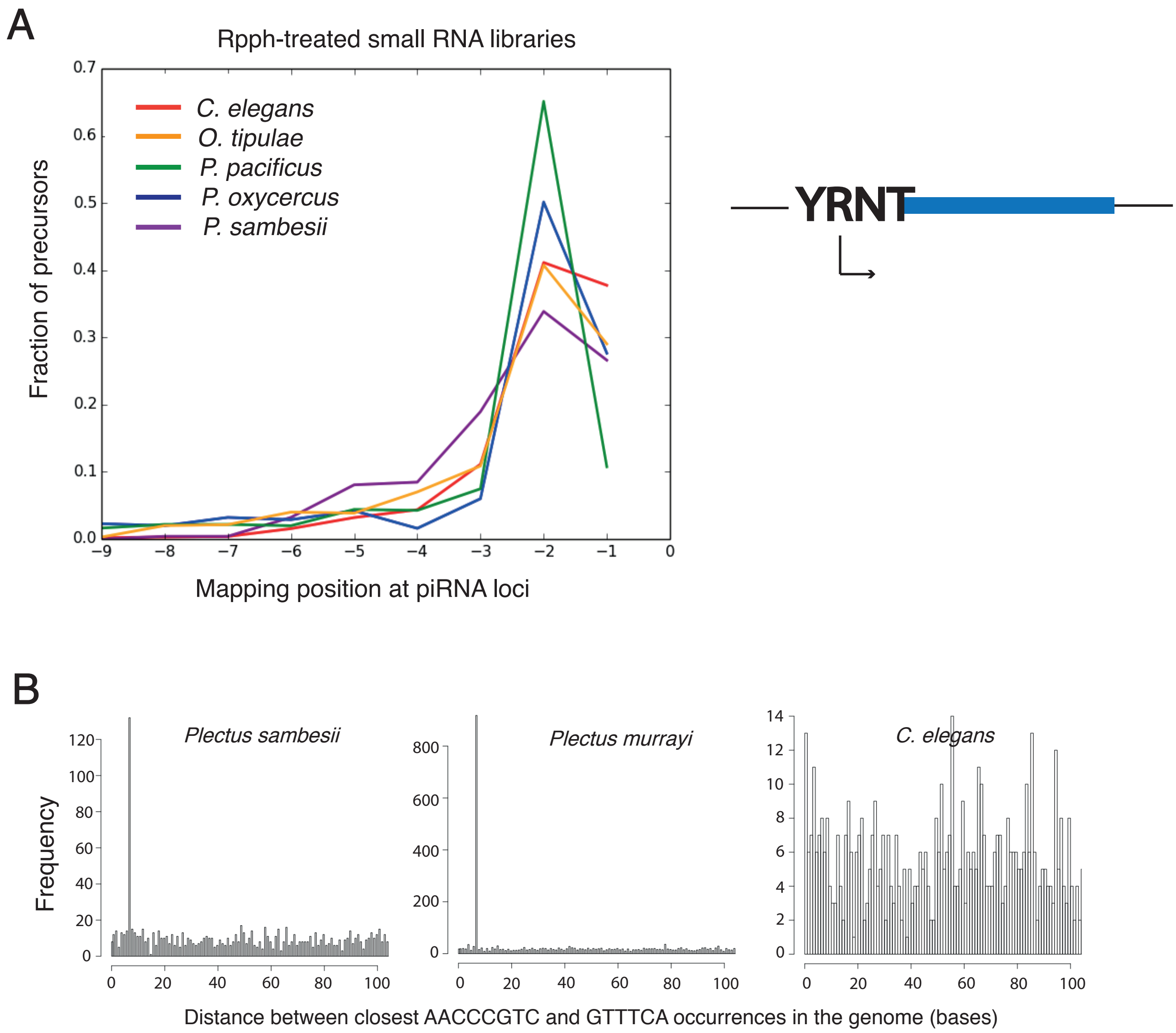

A
Rpph-treated small RNA libraries
Fraction of precursors
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
C. elegans
O. tipulae
P. pacificus
P. oxycercus
P. sambesii
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
Mapping position at piRNA loci
YRNT
B
Frequency
Plectus sambesii
0
20
40
60
80
100
120
Plectus murrayi
0
200
400
600
800
C. elegans
0
2
4
6
8
10
12
14
0
20
40
60
80
100
Distance between closest AACCCGTC and GTTTCA occurrences in the genome (bases)

Figure S2-1

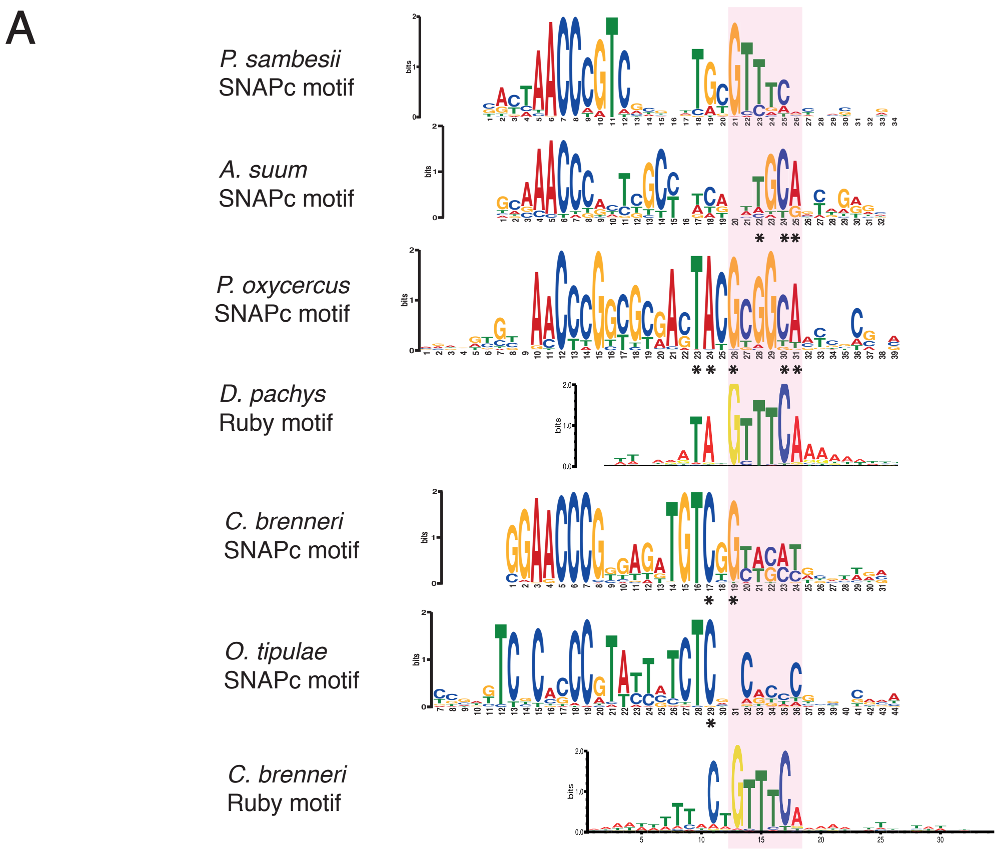

A
P. sambesii
SNAPc motif
A. suum
SNAPc motif
P. oxycercus
SNAPc motif
D. pachys
Ruby motif
C. brenneri
SNAPc motif
O. tipulae
SNAPc motif
C. brenneri
Ruby motif
bits

Figure S2-2

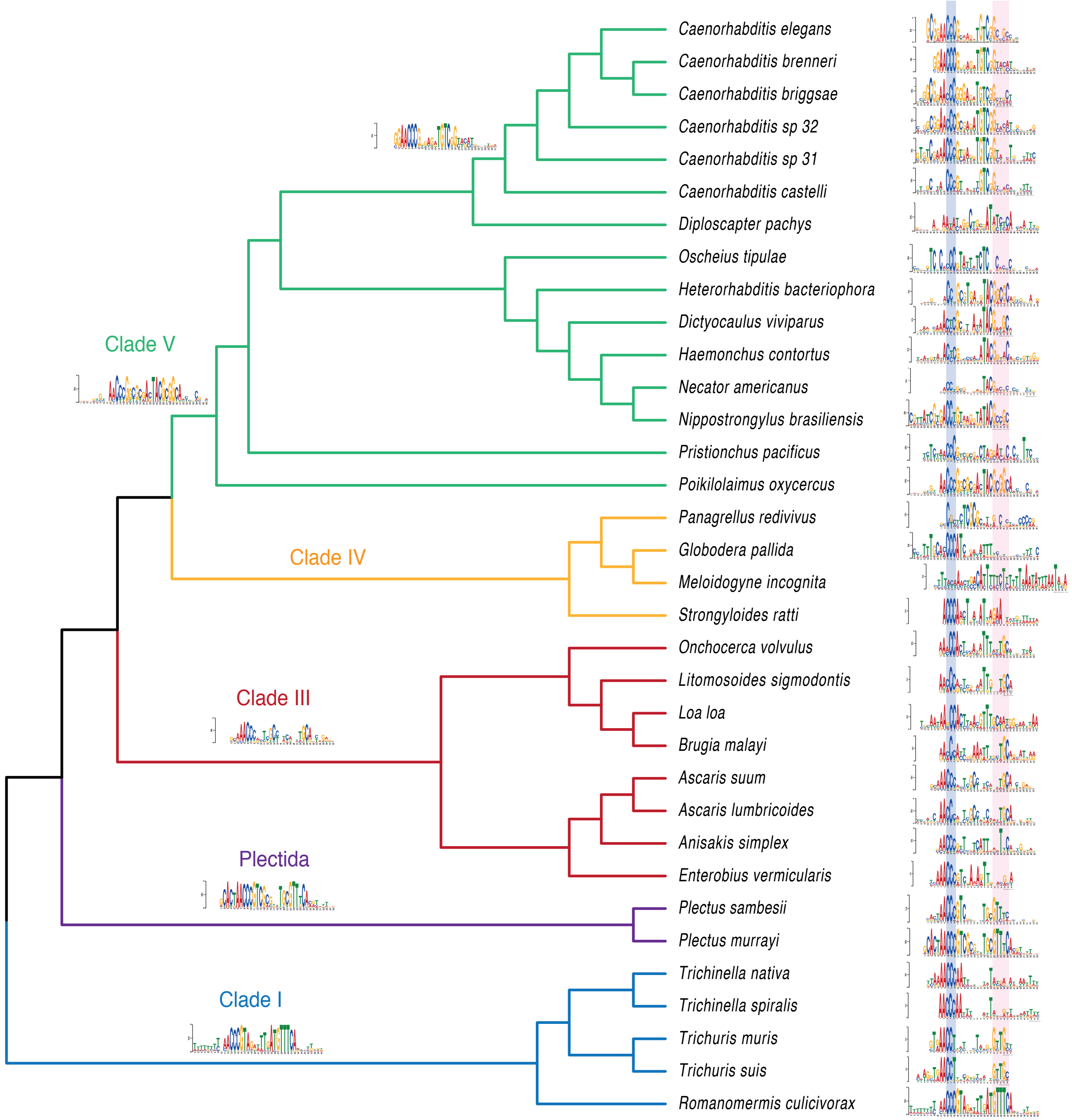

Caenorhabditis elegans
Caenorhabditis brenneri
Caenorhabditis briggsae
Caenorhabditis sp 32
Caenorhabditis sp 31
Caenorhabditis castelli
Diploscapter pachys
Oscheius tipulae
Heterorhabditis bacteriophora
Dictyocaulus viviparus
Haemonchus contortus
Necator americanus
Nippostrongylus brasiliensis
Pristionchus pacificus
Poikilolaimus oxycercus
Panagrellus redivivus
Globodera pallida
Meloidogyne incognita
Strongyloides ratti
Onchocerca volvulus
Litomosoides sigmodontis
Loa loa
Brugia malayi
Ascaris suum
Ascaris lumbricoides
Anisakis simplex
Enterobius vermicularis
Plectus sambesii
Plectus murrayi
Trichinella nativa
Trichinella spiralis
Trichuris muris
Trichuris suis
Romanomermis culicivorax
Clade V
Clade IV
Clade III
Plectida
Clade I

Figure S3

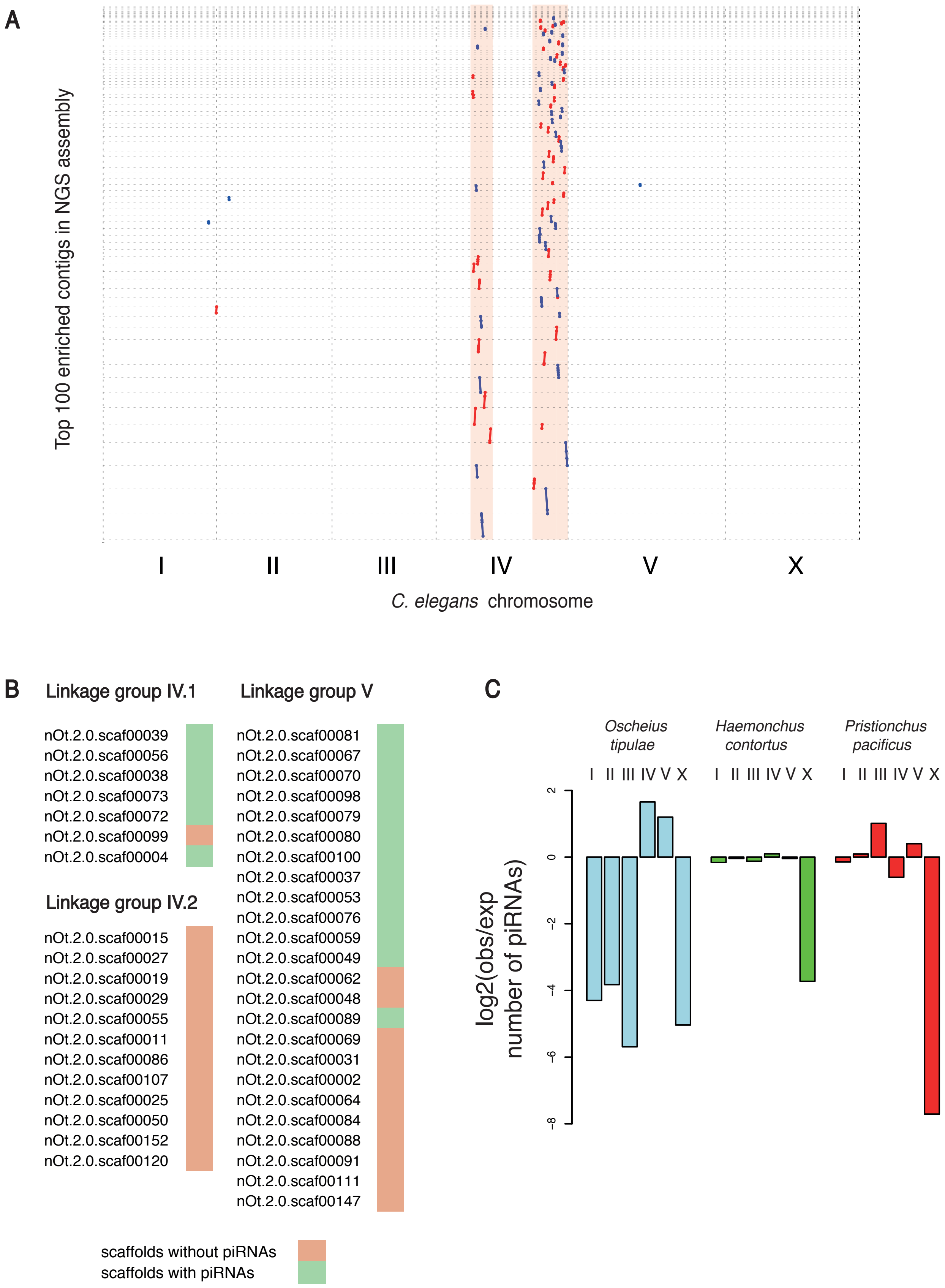
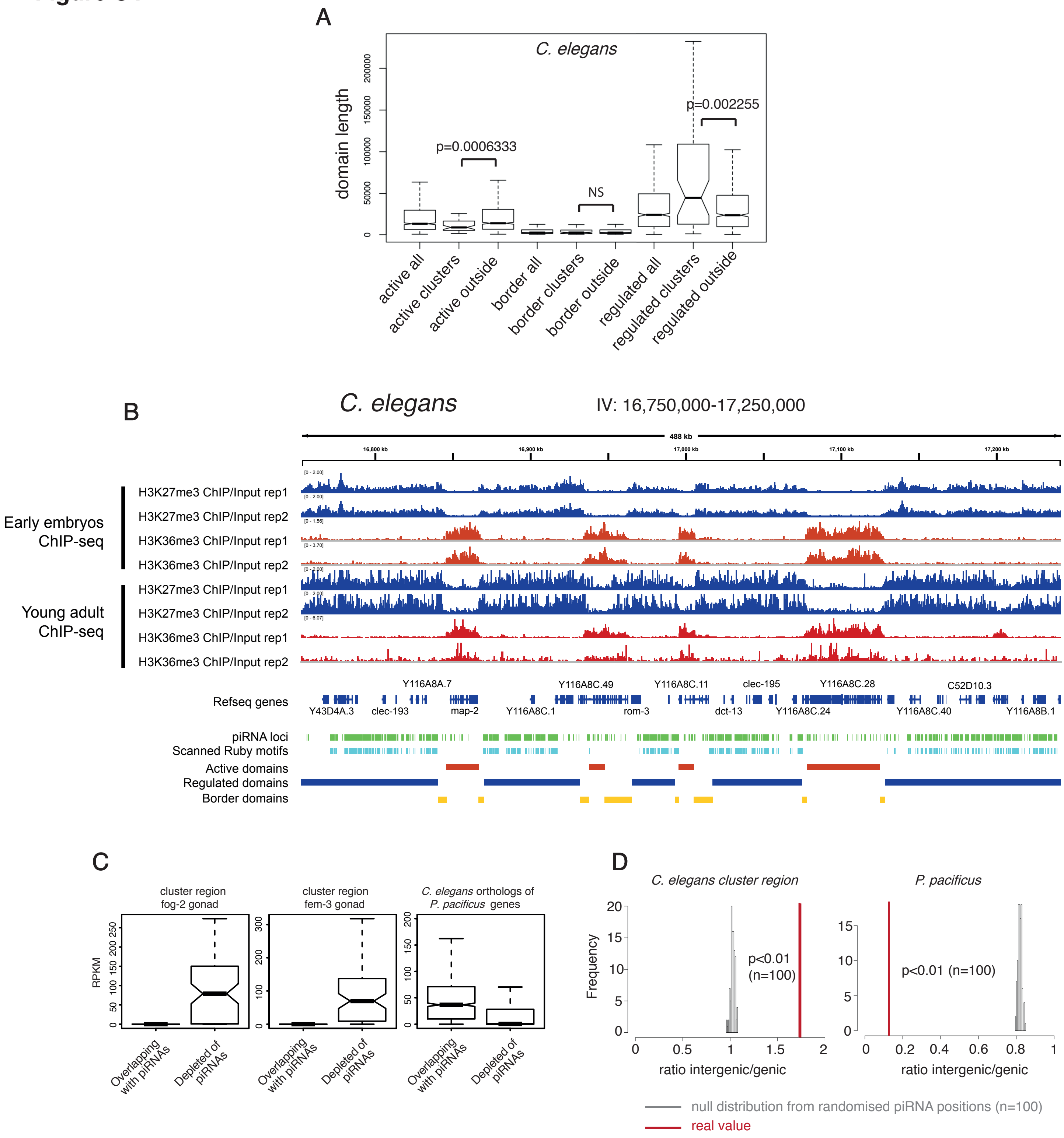

A
Top 100 enriched contigs in NGS assembly
I
II
III
IV
V
X
C. elegans chromosome
B
Linkage group IV.1
nOt.2.0.scaf00039 🟩
nOt.2.0.scaf00056 🟩
nOt.2.0.scaf00038 🟩
nOt.2.0.scaf00073 🟩
nOt.2.0.scaf00072 🟩
nOt.2.0.scaf00099 🟧
nOt.2.0.scaf00004 🟩
Linkage group IV.2
nOt.2.0.scaf00015 🟧
nOt.2.0.scaf00027 🟧
nOt.2.0.scaf00019 🟧
nOt.2.0.scaf00029 🟧
nOt.2.0.scaf00055 🟧
nOt.2.0.scaf00011 🟧
nOt.2.0.scaf00086 🟧
nOt.2.0.scaf00107 🟧
nOt.2.0.scaf00025 🟧
nOt.2.0.scaf00050 🟧
nOt.2.0.scaf00152 🟧
nOt.2.0.scaf00120 🟧
Linkage group V
nOt.2.0.scaf00081 🟩
nOt.2.0.scaf00067 🟩
nOt.2.0.scaf00070 🟩
nOt.2.0.scaf00098 🟩
nOt.2.0.scaf00079 🟩
nOt.2.0.scaf00080 🟩
nOt.2.0.scaf00100 🟩
nOt.2.0.scaf00037 🟩
nOt.2.0.scaf00053 🟩
nOt.2.0.scaf00076 🟩
nOt.2.0.scaf00059 🟩
nOt.2.0.scaf00049 🟩
nOt.2.0.scaf00062 🟧
nOt.2.0.scaf00048 🟧
nOt.2.0.scaf00089 🟩
nOt.2.0.scaf00069 🟧
nOt.2.0.scaf00031 🟧
nOt.2.0.scaf00002 🟧
nOt.2.0.scaf00064 🟧
nOt.2.0.scaf00084 🟧
nOt.2.0.scaf00088 🟧
nOt.2.0.scaf00091 🟧
nOt.2.0.scaf00111 🟧
nOt.2.0.scaf00147 🟧
scaffolds without piRNAs 🟧
scaffolds with piRNAs 🟩
C
Oscheius tipulae
Haemonchus contortus
Pristionchus pacificus
I II III IV V X
I II III IV V X
I II III IV V X
log2(obs/exp number of piRNAs)
2
0
−2
−4
−6
−8

Figure S4



E

		Active	Regulated	
C. elegans	**within clusters**			
	piRNA loci (motif+21U)	5 (231.48)	9230 (1.21)	(fold enrichment)
	scanned Ruby motifs	256 (6.25)	12414 (1.16)	(fold depletion)
	% motifs expressed	1.95	74.35	**38.07**-fold more likely in regulated chromatin
	outside clusters			
	piRNA loci (motif+21U)	41 (2.38)	216 (1.46)	
	scanned Ruby motifs	12891 (1.14)	15222 (1.12)	**4.44**-fold more likely in regulated chromatin
	% motifs expressed	0.32	1.42	
P. pacificus	piRNA loci (motif+21U)	8126	705	**2.08**-fold more likely in active chromatin
	scanned Ruby motifs	16385	2982	
	% motifs expressed	0.49594141	0.236418511	

Figure S5

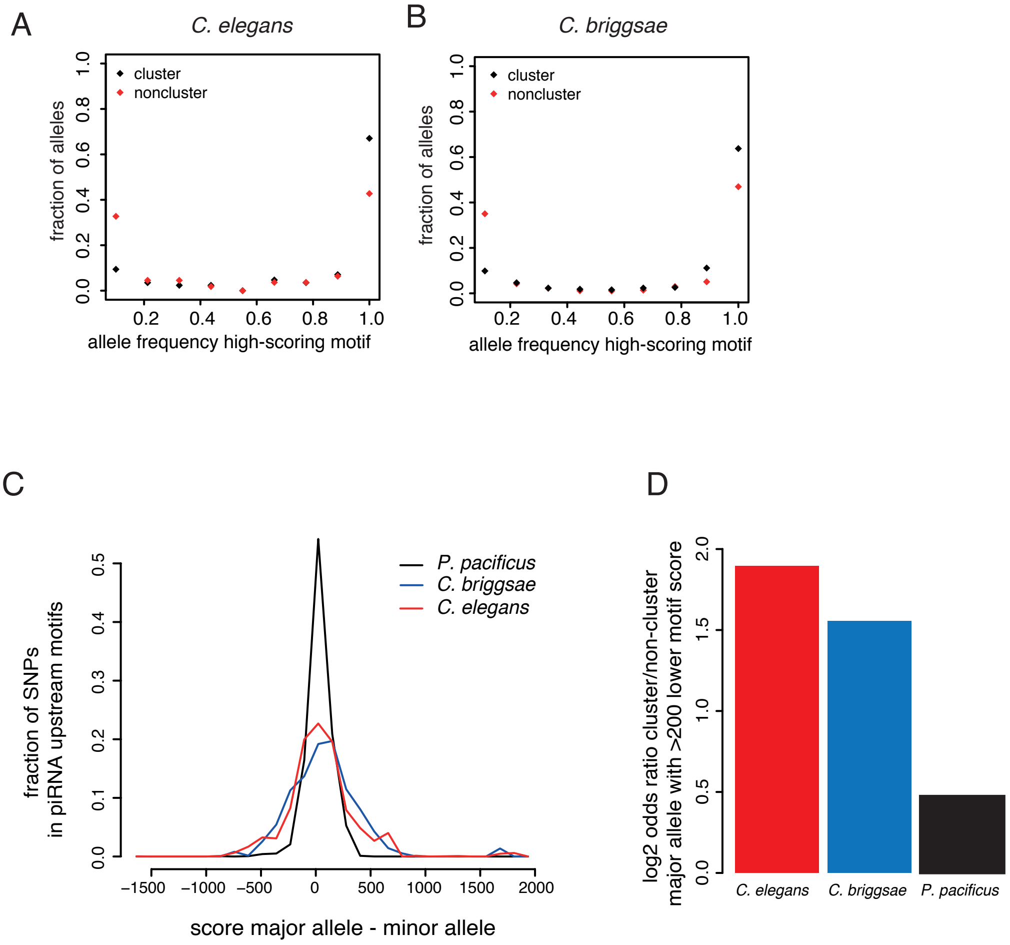

A
C. elegans
cluster
noncluster
fraction of alleles
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
allele frequency high-scoring motif
B
C. briggsae
cluster
noncluster
fraction of alleles
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
allele frequency high-scoring motif
C
fraction of SNPs
in piRNA upstream motifs
0.0
0.1
0.2
0.3
0.4
0.5
P. pacificus
C. briggsae
C. elegans
−1500
−1000
−500
0
500
1000
1500
2000
score major allele - minor allele
D
log2 odds ratio cluster/non-cluster
major allele with >200 lower motif score
0.0
0.5
1.0
1.5
2.0
C. elegans
C. briggsae
P. pacificus

Figure S6

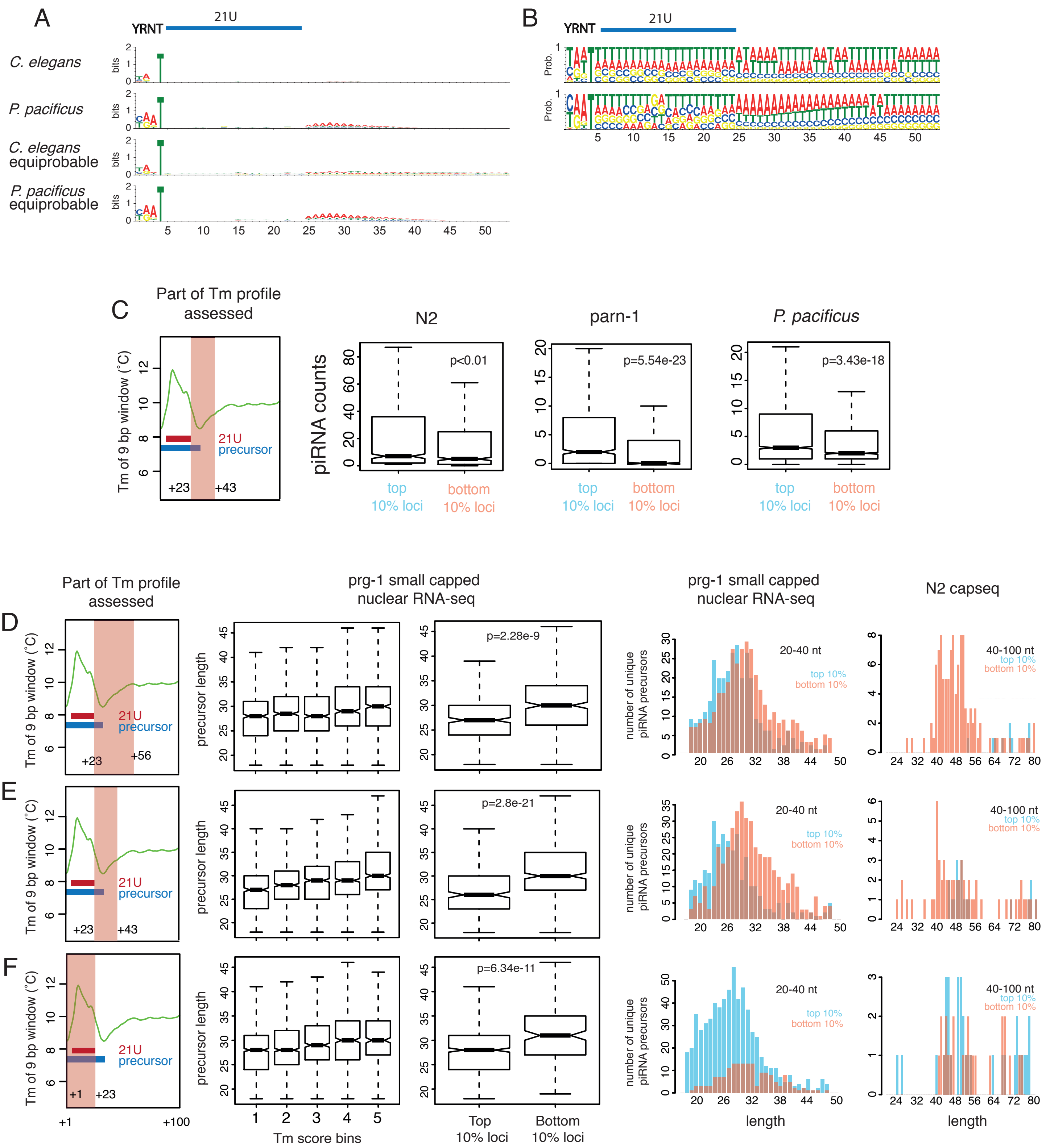

A
YRNT
21U
C. elegans
P. pacificus
C. elegans equiprobable
P. pacificus equiprobable
bits
0
1
2
5
10
15
20
25
30
35
40
45
50
B
YRNT
21U
Prob.
1
5
10
15
20
25
30
35
40
45
50
C
Part of Tm profile assessed
Tm of 9 bp window (˚C)
6
8
10
12
21U
precursor
+23
+43
N2
parn-1
P. pacificus
piRNA counts
p<0.01
p=5.54e-23
p=3.43e-18
0
20
40
60
80
0
5
10
15
20
top 10% loci
bottom 10% loci
Part of Tm profile assessed
prg-1 small capped nuclear RNA-seq
prg-1 small capped nuclear RNA-seq
N2 capseq
D
+23
+56
E
+23
+43
F
+1
+23
+1
+100
precursor length
20
25
30
35
40
45
p=2.28e-9
p=2.8e-21
p=6.34e-11
1
2
3
4
5
Tm score bins
Top 10% loci
Bottom 10% loci
number of unique piRNA precursors
20-40 nt
top 10%
bottom 10%
40-100 nt
top 10%
bottom 10%
20
26
32
38
44
50
24
32
40
48
56
64
72
80
length
length

Figure S7-1

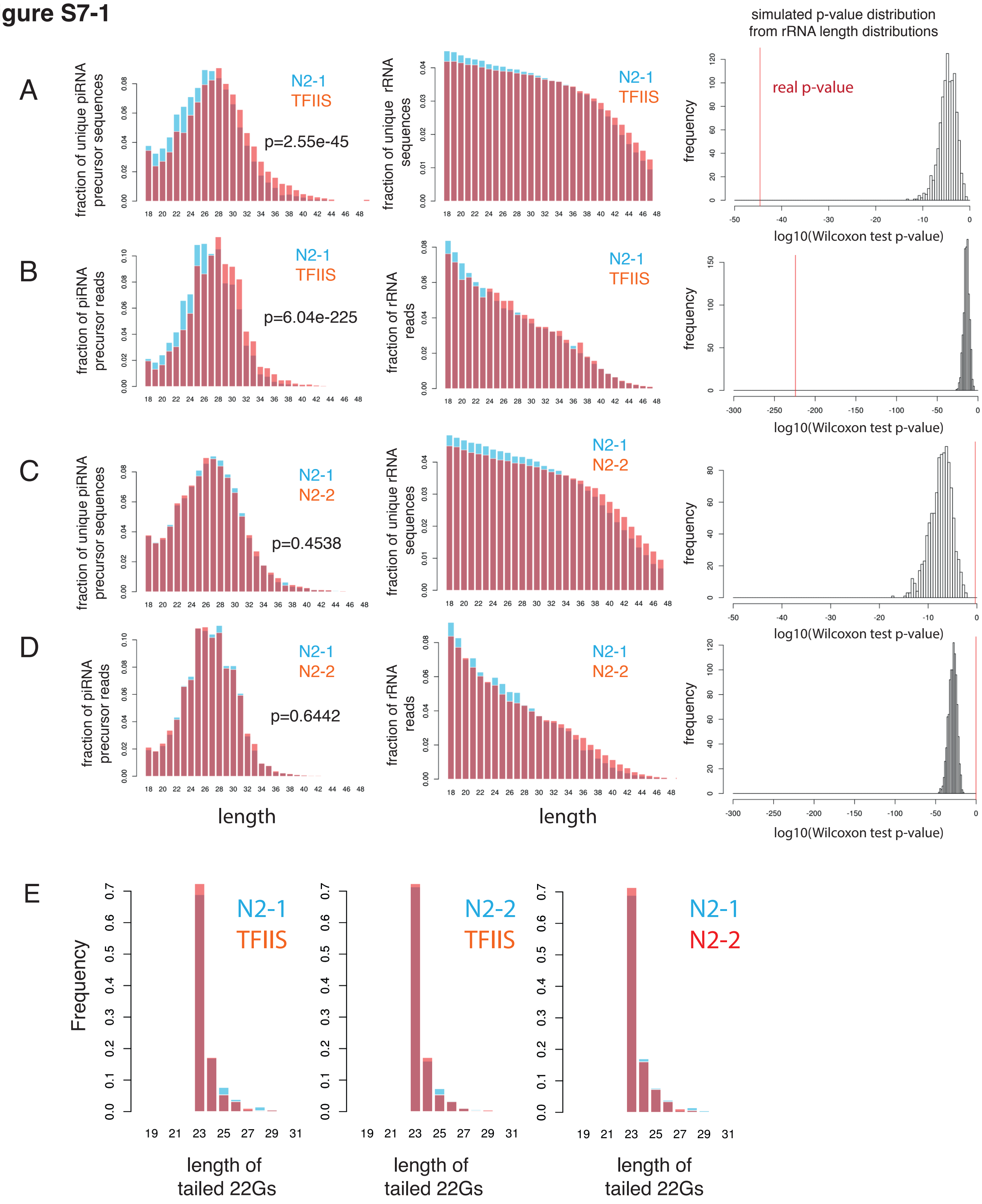

simulated p-value distribution
from rRNA length distributions
A
fraction of unique piRNA precursor sequences
N2-1
TFIIS
p=2.55e-45
fraction of unique rRNA sequences
N2-1
TFIIS
frequency
real p-value
log10(Wilcoxon test p-value)
B
fraction of piRNA precursor reads
N2-1
TFIIS
p=6.04e-225
fraction of rRNA reads
N2-1
TFIIS
frequency
log10(Wilcoxon test p-value)
C
fraction of unique piRNA precursor sequences
N2-1
N2-2
p=0.4538
fraction of unique rRNA sequences
N2-1
N2-2
frequency
log10(Wilcoxon test p-value)
D
fraction of piRNA precursor reads
N2-1
N2-2
p=0.6442
fraction of rRNA reads
N2-1
N2-2
frequency
log10(Wilcoxon test p-value)
length
length
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
E
Frequency
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
N2-1
TFIIS
N2-2
TFIIS
N2-1
N2-2
19 21 23 25 27 29 31
length of tailed 22Gs
length of tailed 22Gs
length of tailed 22Gs

Figure S7-2

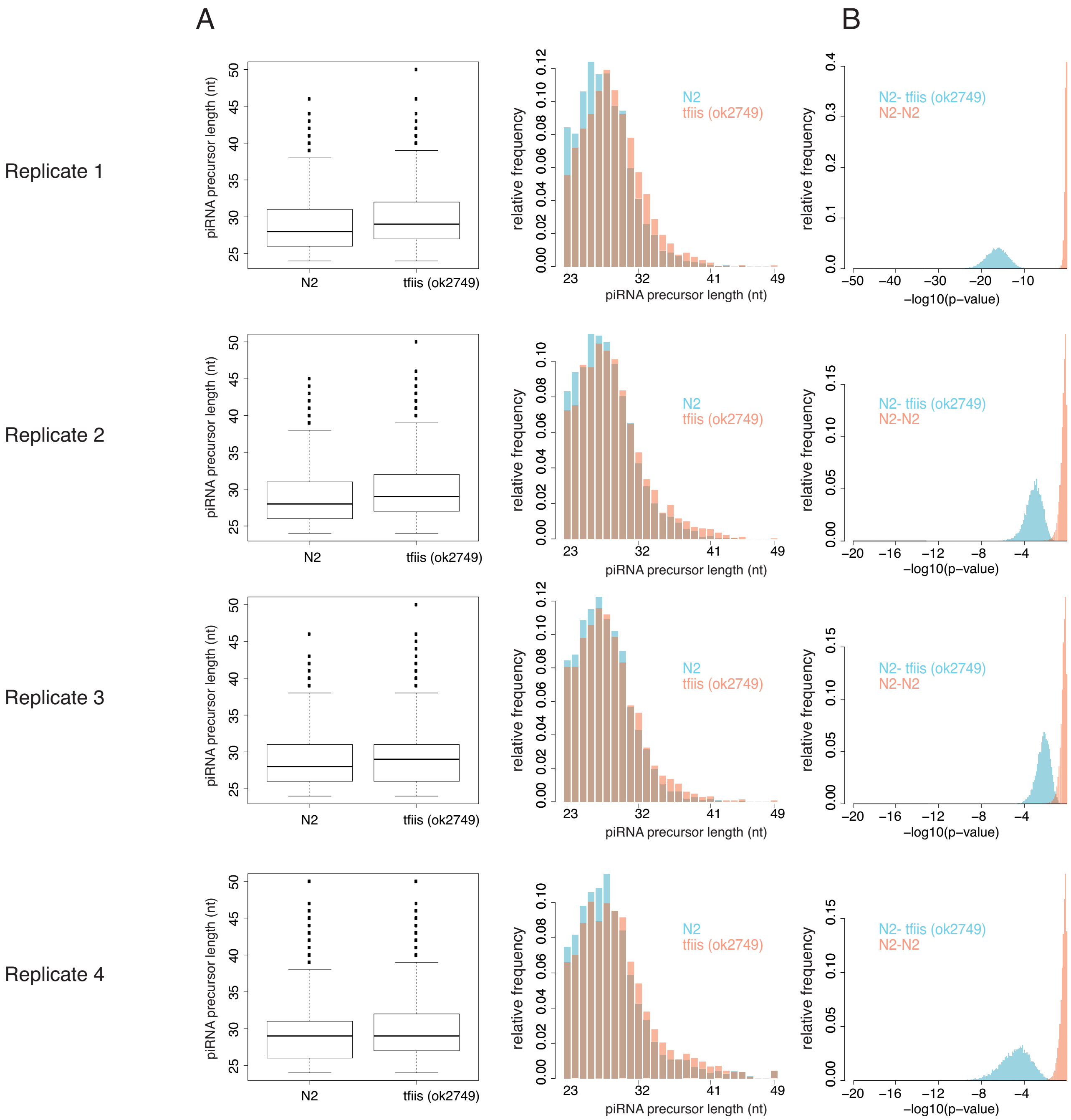

A
B
Replicate 1
Replicate 2
Replicate 3
Replicate 4
piRNA precursor length (nt)
25
30
35
40
45
50
N2
tfiis (ok2749)
relative frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
23
32
41
49
piRNA precursor length (nt)
N2
tfiis (ok2749)
0.0
0.1
0.2
0.3
0.4
−50
−40
−30
−20
−10
−log10(p−value)
N2- tfiis (ok2749)
N2-N2
0.00
0.05
0.10
0.15
−20
−16
−12
−8
−4

Figure S7-3

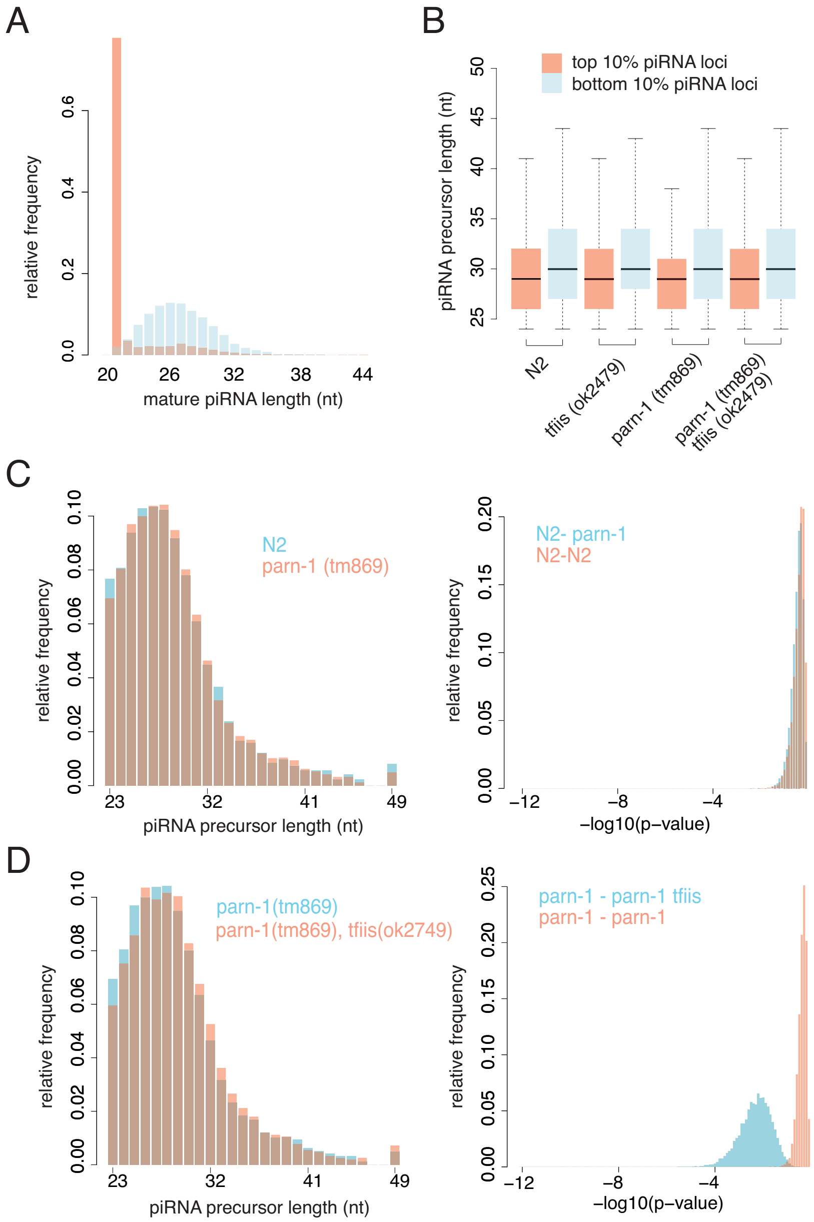
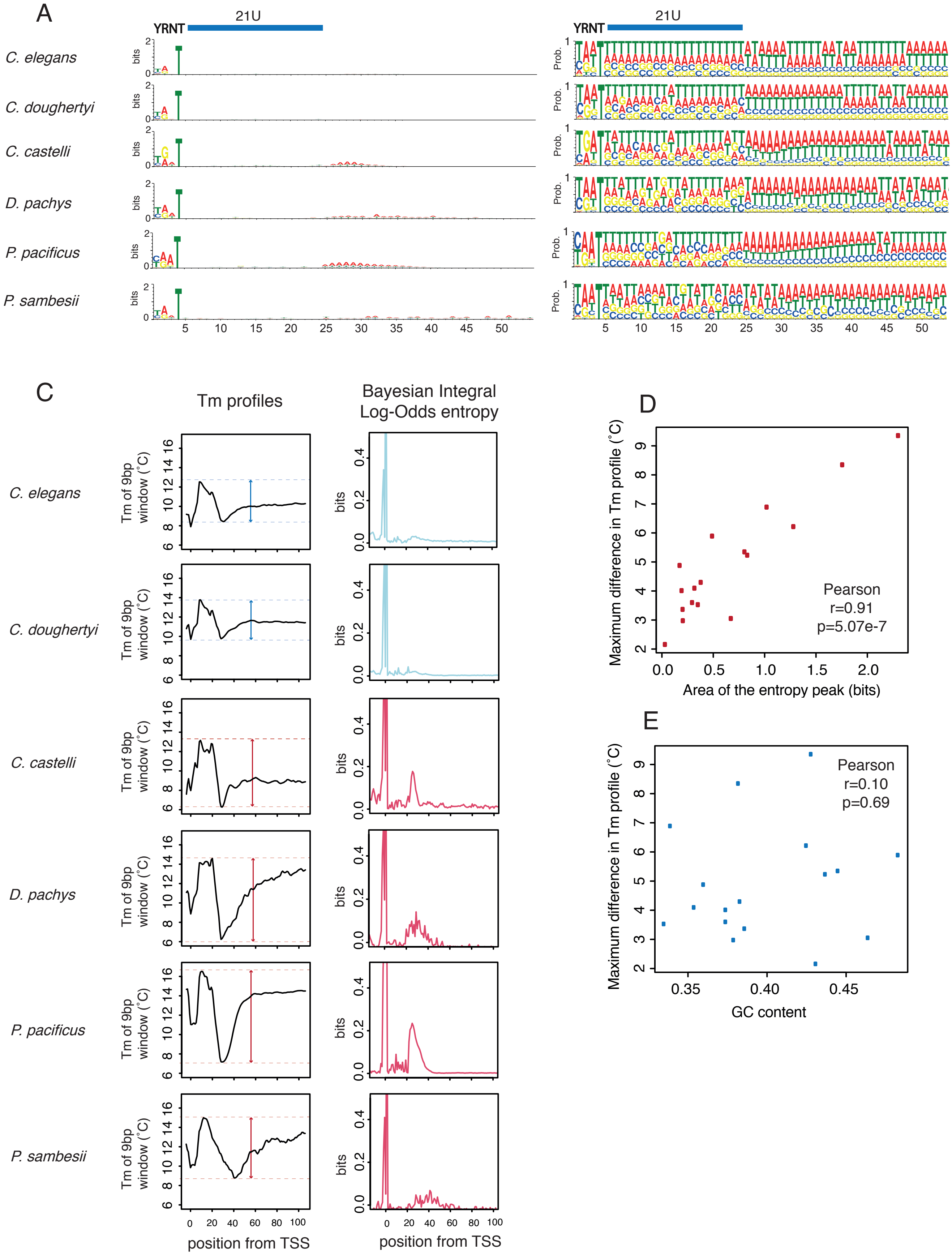

A
relative frequency
0.0
0.2
0.4
0.6
20
26
32
38
44
mature piRNA length (nt)
B
piRNA precursor length (nt)
25
30
35
40
45
50
top 10% piRNA loci
bottom 10% piRNA loci
N2
tfiis (ok2479)
parn-1 (tm869)
parn-1 (tm869)
tfiis (ok2479)
C
relative frequency
0.00
0.02
0.04
0.06
0.08
0.10
N2
parn-1 (tm869)
23
32
41
49
piRNA precursor length (nt)
relative frequency
0.00
0.05
0.10
0.15
0.20
N2- parn-1
N2-N2
−12
−8
−4
−log10(p−value)
D
relative frequency
0.00
0.02
0.04
0.06
0.08
0.10
parn-1(tm869)
parn-1(tm869), tfiis(ok2749)
23
32
41
49
piRNA precursor length (nt)
relative frequency
0.00
0.05
0.10
0.15
0.20
0.25
parn-1 - parn-1 tfiis
parn-1 - parn-1
−12
−8
−4
−log10(p−value)

Figure S7-4


A
YRNT
21U
C. elegans
C. doughertyi
C. castelli
D. pachys
P. pacificus
P. sambesii
bits
Prob.
5 10 15 20 25 30 35 40 45 50
C
Tm profiles
Bayesian Integral Log-Odds entropy
Tm of 9bp window (°C)
6 8 10 12 14 16
bits
0.0 0.2 0.4
0 20 40 60 80 100
position from TSS
D
Maximum difference in Tm profile (°C)
Area of the entropy peak (bits)
0.0 0.5 1.0 1.5 2.0
Pearson
r=0.91
p=5.07e-7
E
Maximum difference in Tm profile (°C)
GC content
0.35 0.40 0.45
Pearson
r=0.10
p=0.69