CENTRO DE INVESTIGACIÓN Y ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

Unidad de Genómica Avanzada

Caracterización de la biogénesis de RNAs pequeños, derivados de elementos transponibles, secretados por *Heligmosomoides bakeri*.

Tesis que presenta

Isaac Martínez Ugalde

Para obtener el grado de

Maestro en Ciencias

En la especialidad de

Biología Integrativa

Director de tesis:

Cei Abreu-Goodger

Irapuato, Guanajuato. Febrero de 2020

CENTRO DE INVESTIGACIÓN Y ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

Advanced Genomics Unit

Characterization of the biogenesis of secreted transposable element derived small RNAs by *Heligmosomoides bakeri*

Thesis presented by:

Isaac Martínez Ugalde

Submitted to obtain the degree of

Master of Science

In the specialty of

Integrative Biology

Thesis director:

Dr. Cei Abreu-Goodger

Irapuato, Guanajuato. February 2020

MASTER THESIS

Author: Isaac Martínez Ugalde

MSc Program: Integrative Biology

Institute: CINVESTAV-Irapuato

Advanced Genomics Unit

Student number: 183880003

E-mail: issac.martinez@cinvestav.mx

Thesis director: Dr. Cei Abreu-Goodger

E-mail: cei.abreu@cinvestav.mx

Advising Committee: Dr. Luis José Delaye Arredondo

Dr. Rafael Montiel Duarte

Acknowledgments

I want to thank the Consejo Nacional de Ciencia y Tecnología (CONACYT) and the Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav) for the facilities and economic support while I carried out my thesis under the program of scholarships and postgraduate studies (CVU): 896776.

This project was also supported by a Human Frontiers Science Program (HFSP) grant RGY0069/2014 (awarded to Amy Buck, Julie Claycomb and Cei Abreu-Goodger) and a SEP-CONAYT Ciencia Básica grant 284884 awarded to Cei Abreu-Goodger.

I also want to thank Julie M. Claycomb and her group at the University of Toronto for receiving me in their lab and for providing access to their data from *prg-1* mutants and PRG-1 immunoprecipitations.

I want to thank my advisor Dr. Cei Abreu, for the advice and support during this project. Finally, I want to thank Dr. Luis Delaye and Dr. Rafael Montiel for the valuable observations and comments on this thesis.

Resumen

Los elementos transponibles (TE por sus siglas en inglés) son usualmente considerados parásitos genómicos. Sin embargo, también son una fuente importante de elementos reguladores, como promotores o RNAs pequeños (sRNAs por sus siglas en inglés). En animales, los sRNAs derivados de TE usualmente están involucrados en regulación endógena. Sin embargo, estos elementos también están involucrados en comunicación extracelular mediada por RNA. El parásito intestinal *Heligmosomoides bakeri*, durante la infección, secreta sRNAs derivados de TE, dentro de microvesículas. Estas microvesículas pueden inmunosuprimir al hospedero. A pesar de que los sRNAs secretados provienen de TE, su biogénesis se desconoce. También se desconoce si clases o familias específicas de TE están asociadas con la producción de los sRNAs secretados. Debido a que los sRNAs derivados de TE han sido estudiados en *Caenorhabditis elegans*, nosotros comparamos su ruta de biogénesis entre *C. elegans* y *H. bakeri*. En *C. elegans* la ruta de los RNAs asociados a PIWI (piRNAs por sus siglas en inglés) está involucrada en la regulación negativa de transcritos de TE. Los piRNAs desencadenan la actividad de RNA polimerasas dependientes de RNA (RdRps por sus siglas en inglés) para producir sRNAs derivados de TE, que son los responsables de la regulación negativa de transcritos de TE. Comparando la expresión de sRNAs en microvesículas y nematodos adultos, encontramos enriquecida la producción de sRNAs secretados en retroelementos LINE, específicamente en las familias LINE/*RTE-BovB* y LINE/*RTE-RTE* y transposones de DNA como RC/*Helitrones* y DNA/*hAT-Tip100*. Inesperadamente, encontramos que en *H. bakeri* la producción de sRNAs a partir de LINE/*RTE-BovB* y RC/*Helitrones* es en sentido opuesto al esperado. Sin embrago, encontramos que estos sRNAs son productos de RdRps, lo cual sugiere que el modelo de predicción de estos elementos transponibles tiene la dirección de cadena invertida. Por otra parte, en *C. elegans,* usando datos de sRNA-seq en mutantes nulos para la proteína PIWI (PRG-1), así como los niveles de expresión de piRNAs y predicciones de sus blancos, encontramos que la producción de sRNAs derivados de TE tiene lugar en regiones cercanas a sitios predichos de unión de piRNA. Usando esta

estrategia establecimos las bases para entender la biogénesis de sRNAs derivados de TE en *H. bakeri.*

Abstract

Transposable elements (TEs) are usually considered genomic parasites. However, they are also an important source of regulatory elements such as promoters or small RNAs (sRNAs). In animals, TE-derived sRNAs are usually involved in endogenous regulation. These elements, however, can also be involved in extracellular RNA communication. The intestinal parasite *Heligmosomoides bakeri* secretes TE-derived sRNAs within microvesicles, during infection. These vesicles can immunosuppress the host. Although secreted sRNAs come from TEs, their biogenesis pathway remains obscure. It is also unknown whether specific classes or families of TEs are associated with the production of secreted sRNAs. Because TE-derived sRNAs have been previously studied in *Caenorhabditis elegans*, we compared their biogenesis pathways between *C. elegans* and *H. bakeri*. In *C. elegans*, the piRNA pathway is involved in downregulating TE transcripts. piRNAs binding triggers RNA dependent RNA polymerases (RdRps) to produce TE-derived sRNAs, that are responsible for downregulating TE transcripts. Comparing the expression of sRNAs in microvesicles and adult nematodes we found an enrichment of secreted sRNAs in LINE retroelements, specifically in LINE/*RTE-BovB* and LINE/*RTE-RTE,* and DNA transposons such as RC/*Helitrons* and DNA/*hAT-Tip100*. Unexpectedly, we found that the production of sRNAs from LINE/*RTE-BovB* and RC/*Helitrons* is from the opposite strand than expected. We found, however, that these sRNAs are RdRp products, suggesting that the prediction model of these transposable elements has the wrong strand. Using sRNA-seq data of *C. elegans PIWI (prg-1)* mutants, as well as the expression levels of piRNAs and piRNA target predictions, we found that TE-derived sRNA production occurs in nearby regions to predicted piRNA binding site. Using this approach, we established the bases to understand the biogenesis pathway of TE-derived sRNAs in *H. bakeri*.

Contents

Introduction 1

Gene regulation by small RNAs 1

Argonaute expansion in nematodes 2

Transposable elements 5

Heligmosomoides bakeri and TE-derived sRNAs 7

Background 8

Hypotheses 11

Objectives 11

Methods 12

Processing of sRNA-seq data 12

Genomic annotation and quantification of ncRNA production. 13

Linear models and siRNA production 14

Differential expression and gene set enrichment analysis 14

piRNA target prediction 15

Validation of piRNA target prediction 16

Results: 17

Genomic comparison of *C. elegans* and *H. bakeri* 17

Small RNA transcriptomic comparison of *C. elegans* and *H. bakeri* 18

Length of genomic regions and sRNA production 21

Inconsistent strand in sRNA production 24

The production of secreted sRNA is enriched in TE classes and families 26

Full length transposons contribute to EV content 33

siRNA production is enriched in DNA TE classes and families in *C. elegans* 35

piRNA Target prediction with pirScan 38

pirFinder development and reimplementation of piRNA target prediction 44

RdRps produce sRNA in nearby regions to piRNA target sites in TEs 45

Discussion 48

Perspectives 51

References 52

Figures

Figure 1. Genomic comparison of *C. elegans* and *H. bakeri*. Different colors represent the proportion in the genome annotated with each category 18

Figure 2. Transcriptomic comparison of sRNAs from polyphosphatase-treated libraries between *C. elegans*, *H. bakeri* adult nematodes and *H. bakeri* EVs 19

Figure 3. Genomic alignments of polyphosphatase-treated sRNAs 22G and 23G, comparing *H. bakeri* adult nematodes.....19

Figure 4. Comparison of sRNA production between 22G sRNAs and 23G sRNAs using genomic positions 20

Figure 5. Genomic alignments of polyphosphate sRNAs of 19-25 nt, comparing *C. elegans*, adult nematodes of *H. bakeri* and EVs 21

Figure 6. Linear regressions comparing the total nucleotides that each category occupies in the genome and the number of sRNAs that are produced 21

Figure 7. Linear regressions comparing the nucleotides which each category use in the genome and the number 19-25 nt sRNAs that are produced, subclassifying categories in sense and antisense.....23

Figure 8. Comparison of sRNA proportion in sense or antisense direction of TE families which have more than 5000 assigned reads.....24

Figure 9. Genome browser visualization of sRNA stacking in 4 different *H. bakeri* LINE/RTE-BovB and RC/Helitrons.....25

Figure 10. Differential expression analysis comparing *H. bakeri* EVs against adult nematodes using just aligned sRNA counts as initial factor of normalization.....26

Figure 11. Expression pattern of all antisense and sense, introns comparing sRNA production between EVs and adult nematodes.....28

Figure 12. MA plot of differential expression analysis comparing *H. bakeri* EVs against adult nematodes after normalization by the expression pattern of introns.....29

Figure 13. MA plot of sense and antisense LINE/*RTE-BovB* and RC/*Helitron* sRNA production during EVs and adult nematodes comparison.....31

Figure 14. MA plot of differential expression analysis comparing *H. bakeri* EVs 5'triP sRNAs against EVs untreated polyP sRNAs.....32

Figure 15. Cumulative length distribution of selected *H. bakeri* TE families.....34

Figure 16. MA plot of differential expression analysis comparing *C. elegans prg-1* mutants against wild type nematodes.....36

Figure 17. Linear correlation of number of *C. elegans* relaxed piRNA prediction targets with pirScan1.0 and the number of nucleotides of the TE target.....39

Figure 18. Lineal correlations comparing pirScan score against the normalized sRNA production.....39

Figure 19. MA plot of differential expression analysis comparing C. elegans prg-1 mutants against wild type nematodes.....40

Figure 20. Expression patterns of piRNAs in *C. elegans prg-1* mutant vs adult nematodes comparison.....42

Figure 21. Differential expression analysis comparing *C. elegans* PRG-1 Ip.....43

Figure 22. MA plot of differential expression analysis in 100 nt windows with respect to pirFinder target position.....45

Figure 23. Cumulative distribution of FDR of gene set enrichment test.....47

Tables

Table 1. sRNAs seq libraries including *C. elegans* and *H. bakeri* experiments used in this work.....12

Table 2. Comparisons of sRNA-seq experiments for differential expression analyses.....14

Table 3. Comparison of TE diversity and genomic proportion between *C. elegans* and *H. bakeri*.....18

Table 4. Gene set enrichment analysis comparing EVs and adult nematodes, using TE class classification.....27

Table 5. Gene set enrichment analysis comparing EVs and adult nematodes, using TE family classification.....27

Table 6. Gene set enrichment analysis after using introns to calculate normalization factors.....29

Table 7. Enrichment of sRNA production in TE families after using introns to calculate normalization factors.....29

Table 8. EV enriched TE families in polyP against untreated sRNA comparison.....33

Table 9. Gene set enrichment analysis of *C. elegans prg-1* mutants against wild type at TE class level.....36

Table 10. Gene set enrichment analysis of *C. elegans prg-1* mutants against wild type nematodes using TE family classification.....37

Table 11. Gene set enrichment analysis at TE class level, in *C. elegans prg-1* mutant vs adult nematodes comparison.....41

Table 12. Gene set enrichment analysis including TE family classification in *C. elegans prg-1* mutant vs adult nematodes comparison.....41

Table 13. Gene set enrichment including TE families related to 100 nt windows with respect to pirFinder target position.....46

Introduction

Gene regulation by small RNAs

Intriguingly, most transcribed RNAs in Metazoans are non-coding, however this does not mean that these RNAs don't have a function. Non-coding RNAs (ncRNAs) such as small RNAs (sRNAs) and long non-coding RNAs (lncRNAs) play an essential role during the development of organisms (Gaiti et al., 2017), by regulating gene expression at the transcriptional and posttranscriptional level (Wilson & Doudna, 2013; Yoon, Abdelmohsen, & Gorospe, 2013).

The most famous class of sRNAs are micro-RNAs (miRNAs), first discovered in *Caenorhabditis elegans* when studying *lin-4* and later *let*-7 (Lee, Feinbaum, & Ambros, 1993; Reinhart et al., 2000). During the early development of *C. elegans*, the miRNA *lin*-4 induces the down regulation of LIN-14, allowing the progression from larval stage 1 (L1) to L2. Similarly, during the last larval stage (L4), the miRNA *let*-7 is responsible of the down regulation of LIN-41, and the expression of *let-7* is essential for establishing the adult stage.

In animals, there are at least 3 main classes of small RNAs that are directly involved in post-transcriptional gene regulation: miRNAs, small interfering RNAs (siRNAs) and PIWI interacting RNA (piRNAs) (Wynant et al., 2017).

Since their discovery, miRNAs have been associated with regulating developmental processes, such as cell determination and differentiation. This relation is not surprising because many miRNAs are expressed in spatiotemporal manner. During the early development of animals, miRNAs can regulate hox genes. For example, the miR-310/313 cluster can modulate the expression of Ultrabithorax (Ubx), a master transcriptional factor which regulates the establishment of the dorsoventral axis (Kaschula et al., 2018). To induce downregulation of transcripts, miRNAs are loaded onto Argonaute proteins, building RNA Induced Silencing Complexes (RISC). The RISC complex is guided mainly by complementarity between the seed region (nt 2-7) of the miRNA and the 3' un-translated region of the target mRNA, as well as partial complementarity of the remaining nucleotides of the miRNA. In animals, miRNA binding induces

deadenylation and degradation of the target mRNA, as well as inhibition of translation (O'Brien et al., 2018).

On the other hand, piRNAs have been related with the down-regulation of transposable element (TE) transcripts. piRNAs were first described in *Drosophila melanogaster*, where these sRNAs were originally classified as repeat-associated siRNAs (rasiRNAs) (Lin & Spradling, 1997). In *D. melanogaster* the depletion of the PIWI argonaute, which loads piRNAs, allows the transposition of retroelements, inducing defects in germline cell divisions and fertility (Czech et al., 2018).

Interestingly, not all metazoans have siRNAs. In contrast to miRNAs and piRNAs, siRNAs are not produced by RNA pol II (Claycomb 2014; Wynant et al., 2017). Their absence in some metazoan clades is related with the independent loss of RNA dependent RNA polymerases (RdRps) which produce siRNAs (Wynant et al., 2017). In animals and plants RdRps use as a template target transcripts, producing double stranded RNA (dsRNA), then these RdRp products are cleaved by Dicer (Pinzón et al., 2019). RdRp/Dicer products which are loaded by Argonaute (Ago) proteins are classified as siRNAs (Hoogstrate et al., 2014). Usually siRNAs recognize RNA targets by full base-pair complementarity and trigger their degradation. In nematodes siRNAs can have different functions, such as germline gene regulation by transposable element silencing and chromatin modifications (Billi et al., 2014). Notably the function of siRNAs depends on their related RdRp and Ago protein. *C. elegans,* for example, has at least 3 different RdRps and their products are loaded onto different Ago proteins (Hoogstrate et al., 2014).

Argonaute expansion in nematodes

Notably, nematodes have developed a wide diversity of sRNA pathways, in part due to the expansion of the Ago protein family (Buck & Blaxter, 2013). As a refence, while humans have 4 Agos, *C. elegans* encodes for 25 (Wynant et al., 2017). Intriguingly, 19 of the 25 *C. elegans* Agos are specific of nematodes and are considered within the Worm Argonautes group (WAGOs). Many of these

WAGOs can have highly specific temporal or spatial expression patterns (Buck & Blaxter, 2013).

It is important to highlight that in *C. elegans* siRNAs are the most abundant group of sRNAs, higher even in expression than miRNAs. These siRNAs can be loaded onto WAGOs to regulate different processes, including epigenetic modification, gene regulation, transposon silencing, environment sensing and transgenerational inheritance (Billi et al., 2014; Buck & Blaxter, 2013; Gu et al., 2009). In the following, I will briefly describe what is known about the function of some WAGOs and their guide siRNAs.

In *C. elegans,* a WAGO named heritable RNAi defective (HRDE-1) has been associated with stress response (Spracklin et al., 2017) . HRDE-1 is expressed in the nucleus of germ line cells. During stress conditions, such as heat, the H3K9 trimethylation (a repressive expression mark) is removed near some genes, such as those encoding heat shock proteins (Hsp). To regulate the expression of these genes, HRDE-1 is essential. It has also been shown that siRNAs related to HRDE-1 can regulate the expression of heat response genes in the germ line (Ni et al., 2016). Interestingly some of HRDE-1 heat response targets are near transposons and the loss of HRDE-1 activity through generations is related with infertility (Ni et al., 2016).

The antagonistic relationship between chromosome-segregation and RNAi deficient 1 (CSR-1) and PIWI related gene 1 (PRG-1) within germ line is essential to regulate gene expression during the early development of *C. elegans* (Youngman & Claycomb, 2014). Interestingly, CSR-1 loads siRNAs which bind to mRNAs of endogenous genes that are essential to development, preventing the binding of piRNAs which induces silencing through mRNA degradation (Shen et al., 2019; Youngman & Claycomb 2014; Zhang et al. 2018).

Although PRG-1 and piRNAs are important in *C. elegans* for regulating mRNAs during early development, in animals the main function of piRNAs is to silence transposable element expression (Czech et al., 2018). In nematodes, piRNAs are 21 nt and generally have an uracil at the 5’ end (they were initially described as 21U-RNAs). These piRNAs are loaded onto PRG-1 (PIWI related gene 1) (Hoogstrate et al., 2014). The biogenesis of piRNAs in *C. elegans* depends of the

“ruby” motif (*GTTTC* consensus motif) that is recognized by RNA pol II (Bagijn et al., 2012). The presence or absence of the ruby motif is used to subclassify piRNAs (type I are ruby motif dependent, while type II are not) (Weick et al., 2014). For type I piRNA biogenesis, piRNA silencing-defective gene 1 (PRDE-1) binds to the ruby motif and recruits RNA pol II to produce short transcripts of 28-29 nt called pre-piRNA (Weick et al., 2014), their end defined by a stop motif that is recognized by RNA pol II (Beltran et al. 2019). These pre-piRNAs have a 5’-cap and a 2’-o-methylation at the 3’ end. The 3’ methylation is produced by the HEN of Nematode 1 (HENN-1) methyltransferase, and stabilizes the piRNA (Zeng et al., 2019). It has also been shown that Twenty One U-RNA biogenesis Fouled Up 6 (TOFU-6), piRNA biogenesis and chromosome segregation (PISC-1) and piRNA-induced silencing-defective (PID-1), form a complex to remove the 5’ cap before PRG-1/piRNA complex formation (Zeng et al., 2019). In addition, PARN-1 RNase is necessary to remove extra nucleotides in the 3’ end (Tang et al., 2016). On the other hand, little is known about the biogenesis of type II piRNAs. It has been shown that the ruby motif and factors such as PRDE-1 and TOFU-6 are unnecessary (Zeng et al., 2019).

Similar to miRNAs, piRNAs also have specific targeting rules (Bagijn et al., 2012). Using synthetic piRNAs and GFP reporters, Zhang and co-workers determined piRNA targeting rules in *C. elegans* (Zhang et al., 2018). They used synthetic piRNAs with different numbers and positions of mismatches with respect to the reporter transcript to identify how piRNAs can bind to a transcript. They found that during piRNA-transcript interaction, as with miRNAs, the first nucleotide of the piRNA does not require complementary. There is also a high complementarity from the second to seventh nucleotide (seed region), however a few non-Watson-Crick interactions, specifically GU wobbles, are allowed. Finally, in the whole interaction a maximum of 6 mismatches are allowed (Wu et al., 2018; Zhang et al., 2018).

After the interaction of a piRNA to a TE transcript, RNA dependent RNA polymerases such as RNA-dependent RNA polymerase Family -1 (RRF-1) or Enhancer of Glp-1 (EGO-1) are recruited (Ashe et al., 2012; Bagijn et al., 2012; Shen et al., 2019; Youngman & Claycomb, 2014). RdRps then polymerize 22Gs antisense to the TE transcripts in the vicinity to the location of piRNA binding

(Ashe et al., 2012; Bagijn et al., 2012). Importantly, and in contrast to many other sRNAs, the 22Gs have a triphosphate at their 5' end, a consequence of being the direct product of an RNA-dependent RNA polymerase. Secondary Ago proteins then load the 22Gs, using them as guides to bind to other TE transcripts with the same sequence, inducing their degradation(Bagijn et al., 2012).

Transposable elements

Transposons or transposable elements (TEs) are DNA sequences whose replication does not require genome replication. These elements have been described in bacteria, plants, fungi and animals (Platt et al., 2018). TEs were first described by Barbara McClintock in her work *The Origin And Behavior Of Mutable Loci In Maize* (McClintock, 1950). Although she proposed TEs as regulatory elements, TEs are currently classified more often as genomic parasites (Chuong et al., 2017).

TEs can be classified in two main groups: type I TEs and type II TEs (Chuong et al., 2017). Type I TEs are retrotransposons and can also be subclassified as LINE (Long interspersed nuclear element), SINE (Short interspersed nuclear element) and LTR (Long Terminal Repeats). Retrotransposons usually depend on an RNA intermediary for their transposition, using reverse transcriptase proteins to produce DNA. Conversely, type II TEs are DNA transposons, these TEs use a transposase to cut out their sequence in the genome and jump in a new location (Bourque et al., 2018).

LTRs are usually composed of 2 genes: *gag* and *pol*, and these genes are flanked by long terminal repeats of 700 to 5,000 bases (Chuong et al., 2017). The LTR genes encode an integrase and a reverse transcriptase. Both strands of LTRs have RNA pol II binding sites, and for this reason both strands can be transcribed, however a single strand is generally used in specific cell types (Tan et al., 2016). LTRs also contain a tRNA-binding site, LTRs use tRNAs, which bind to this site as primers for reverse transcriptase and the start of polymerization. After the first DNA strand is produced by the reverse transcriptase, RNase H is essential to remove the RNA strand so the other DNA strand can be made by the reverse

transcriptase. Finally, the integrase is used to insert the dsDNA in a new location of the genomic DNA (Finnegan, 2012).

LINE elements have a main 5' RNA pol II promotor. Interestingly, some LINE elements can also have an antisense promotor (Chuong et al., 2017). Most LINE elements have two ORFs, which usually encode an endonuclease with an RNase H domain and a reverse transcriptase. In contrast to LTRs, LINEs do not have a tRNA-binding site to start polymerization by reverse transcriptase (Finnegan, 2012). For replication and transposition LINEs use Target-Primed Reverse Transcription (TPRT), in which the endonuclease creates a double-strand break in the target position of DNA. This break allows the freed 3′-OH of the target site to be used to start polymerization of the LINE, and homologous recombination finalizes the transposition (Kazazian, 2004).

The third class of retroelements, SINEs, are non-autonomous. These TEs are usually overlapped with other sequences such as LINEs, tRNAs or rRNAs (Chuong et al., 2017) that can help drive their transcription. SINEs however can contain RNA pol II promotors which can influence the expression of nearby genes (Finnegan, 2012).

Finally, DNA transposons do not depend on an RNA intermediate. They instead use a transposase to directly move from one location to another in the genome. They encode a transposase that recognizes and cuts the terminal inverted repeats of the transposon (Chuong et al., 2017). This enzyme also makes a cleavage in the new locus, causing short gaps that are repaired, producing target sequence duplication (TSD) (Bourque et al., 2018).

Today, there is much evidence that supports the original idea of McClintock. In fact, TEs are an important source of regulatory elements such as ncRNAs, including sRNAs and lncRNA (Cho, 2018). TEs can also function as promotors or enhancers of endogenous genes (Kidwell & Lisch, 2000). It is important to highlight that TE misregulation can cause genome instability (Ayarpadikannan & Kim, 2014). However, both plants and animals have developed different mechanisms based on sRNAs to regulate the expression of these elements (Cho, 2018; Czech et al., 2018). As mentioned before, many animals use piRNAs while

plants use an analogous class of sRNAs called epigenetically activated small interfering RNAs (easiRNAs).

The expression of TEs can also influence processes such as development and adaptation. During early embryo development in humans, for example, retroelements such as LINE L1 directly influence cell pluripotency (Blanco-Jimenez, & García-Pérez 2015). The expression of this retroelement is involved in the regulation of the timing of specific cell differentiation such as cells in trophectoderm, which gives rise to the placenta (Garcia-Perez, et al., 2016). In other animals, such as *C. elegans* during the 4-16 cells embryo stage, SINEs are highly expressed in AB cells. These cells are the precursors of neurons, pharynx and epidermis. Ansaloni and co-workers suggest that the expression of SINEs is related with cell specification in these tissues (Ansaloni et al., 2019). However, the specific function of SINEs during early development of *C. elegans* remains unclear.

Interestingly, the fungal parasite *Botrytis cinerea,* which infects a wide variety of plants, takes advantage of its TEs during parasitism. *B. cinerea* can deliver siRNAs derived from TEs to its host, that are capable of silencing defense genes (Weiberg et al., 2013). In fact, one of these siRNAs, named Bc-siR37, is produced from an LTR. This suggests that TEs are an important part of the parasitic lifestyle of this fungus. It is interesting to think how TEs are been used as a mechanism of extracellular communication, and how these elements are shaping a strategy of immunomodulation.

Heligmosomoides bakeri and TE-derived sRNAs

Nematodes are considered one of the most successful groups of parasites, due to the huge diversity of hosts that they can invade. The success of their lifestyle is related with their skills to evade the immune system of their host (Cooper & Eleftherianos, 2016). These parasites have developed a wide diversity of mechanisms and strategies to avoid the immune system. Mammalian parasites, for example, can secrete anti-inflammatory molecules such as proteins, peptides and extracellular vesicles (EVs) (Maizels & McSorley, 2016).

It has been shown that the intestinal parasite of mice *Heligmosomoides bakeri*, secretes EVs during infection, and these EVs can immunosuppress the host *in vivo* (Buck et al., 2014). These EVs contain an Argonaute protein and small RNAs such as miRNAs, yRNAs and siRNAs (Buck et al., 2014; Chow et al., 2019). Unexpectedly, EVs seem to be a common mechanism of parasitic nematodes to induce host immunosuppression. Human parasites such as *Brugia malayi* and *Onchocerca volvulus*, also secrete immunomodulatory EVs (Quintana et al. 2015; Zamanian et al. 2015).

Notably siRNAs are the most abundant class of sRNAs within the EVs that *H. bakeri* secretes (Chow et al., 2019). Most of these siRNAs map to TEs such as DNA transposons, retrotransposons and novel repeats (Chow et al., 2019).

In relation with the secreted sRNAs derived from TEs, it is important to mention that nearly 45% of the *H. bakeri* genome is composed of TEs (Chow et al., 2019).Using this as a reference, we are looking to clarify if there is a general contribution of all genomic TEs or if specific TEs are mostly responsible of TE-derived sRNA production. Even though most of the secreted sRNAs map to TEs, the biogenesis of these sRNAs has not been studied.

Background

Most of the secreted TE-derived sRNAs of *H. bakeri* have a length between 22-23 nucleotides (nt). Additionally, the first nucleotide at the 5’ end of these sRNAs is biased to a guanine (Chow et al., 2019). In nematodes, the first nucleotide of sRNAs is relevant due to the affinity of Agos to certain classes of sRNAs(Hoogstrate et al., 2014). It is important to highlight that within secreted EVs there is also an Ago protein called exWAGO (extracellular worm Ago) (Chow et al., 2019). Coincidently, in the free living nematode *C. elegans*, the most abundant class of sRNAs also are of 22 nt and have a guanine at their 5’ end (Billi et al., 2014). These sRNAs are named 22G-siRNAs or 22Gs. Notably, the SAGO-1, SAGO-2 and PPW-1 Ago proteins load 22Gs in *C. elegans*, and are orthologs of exWAGO (Chow et al., 2019). It is important to mention that all these

proteins are expressed in the intestine (Seroussi & Claycomb, *unpublished*). Secreted EVs are likely produced as well in the intestine (Buck et al., 2014).

In *C. elegans*, the 22Gs are considered secondary siRNAs, because their biogenesis depends on other classes of sRNAs acting as a primary silencing trigger. There are at least 3 different primary siRNAs in *C. elegans*, that can trigger 22G siRNA biogenesis: 26G-siRNAs (26 nt and a guanine at the 5' end), siRNAs from bidirectional transcripts, and piRNAs (Ashe et al., 2012; Billi et al., 2014; Gu et al., 2009). siRNAs can be produced from different classes of transcripts depending on the trigger and their related Ago, for example piRNAs and PRG-1 trigger the biogenesis of siRNAs from TEs.

With respect to TEs in *C. elegans,* DNA transposons, specifically members of the *Tc-Mar-Mariner* superfamily, are believed to be the only active TEs within the *C. elegans* genome (Billi et al., 2014). However, new evidence using single cell RNA-seq, has revealed that retroelements are also expressed. As mentioned before, Ansaloni and co-workers followed the expression of transposable elements during the early development of *C. elegans* (Ansaloni et al., 2019). They found that the expression of DNA transposons such as *Tc-Mar-Mariner* do not depend on a specific cellular or developmental context. However, retroelements such as SINEs, LINEs and LTRs are expressed only in specific cellular linages and during specific developmental stages.

Although piRNAs were thought to be just expressed in the germline (Bagijn et al., 2012), new evidence has shown the expression of piRNAs and other essential factors such as PRDE-1 and PRG-1 in other cells. Using GFP reporters, Kim and co-workers showed for the first time the expression of PRG-1 within neurons and that depletion of PRG-1 can cause defects in axon-regeneration (Kim et al., 2018). Also, in other non-nematode models such as *Aedes aegypti* or *Drosophila spp.*, piRNAs are also expressed in the intestine (Mukherjee et al., 2019). Together, these evidences open questions related to the importance of TE silencing by piRNAs in somatic cells.

In *H. bakeri,* retrotransposons and novel repeat elements are enriched in the production of secreted TE-derived sRNAs (Chow et al., 2019). However, it is unclear if specific TEs produce most of the secreted TE-derived sRNAs.

Although not all nematodes have piRNAs (Beltran et al., 2019; Sarkies et al., 2015), *H. bakeri* does have all the essential components for the type I piRNA biogenesis pathway, including ruby motifs, PRD-1, stop pol II motif and PRG-1 (Beltran et al., 2019). Together, these evidences suggest that piRNAs could trigger the biogenesis of secreted TE-derived sRNAs in *H. bakeri*.

In this thesis, we performed a comparative analysis between TE-derived sRNAs in *C. elegans* and *H. bakeri* to understand how secreted TE-derived sRNAs might be produced, and to understand if the production of TE-derived sRNAs is enriched in specific TE classes or families.

Hypotheses

1. The production of secreted sRNAs by *Heligmosomoides bakeri* is enriched in specific TE families.
2. The biogenesis of secreted TE-derived sRNAs by *H. bakeri* depends on the piRNA pathway.

Objectives

1. Determine if the production of secreted TE-derived sRNAs by *H. bakeri* is enriched in specific TE families.
2. Determine if piRNAs can trigger the biogenesis of secreted TE-derived sRNAs by *H. bakeri*.

Methods

Processing of sRNA-seq data

We used 25 sRNA-seq libraries to compare the TE-derived sRNA production between *C. elegans* and *H. bakeri* (Table 1) (Bagijn et al., 2012; Chow et al., 2019; Seroussi & Claycomb *unpublished*). Most libraries were prepared with polyphosphatase treatment (**Table 1**, column Treatment), which means that polyphosphatase was used to remove 5' tri- and di-phosphates, leaving a 5' mono-phosphate to allow adapter ligation (Bagijn et al., 2012; Chow et al., 2019). It is important to highlight that this treatment does not affect 5' monophosphate sRNAs. No treatment classification means that these libraries were not treated with polyphosphatase. Finally, piRNA 21UR-1349 libraries have a (GFP)–histone H2B construct with a binding site for this piRNA, as a part of experiments described by Bagijn and co-workers (Bagijn et al., 2012). In the last column (Number of reads), we have the summation of reads of all replicates.

Table 1. sRNAs seq libraries including *C. elegans* and *H. bakeri* experiments used in this work.

sRNA-seq	Treatment	Number of replicates	Total number of reads
H. bakeri adult	polyphosphatase	3	42,557,710
H. bakeri adult	no treatment	3	67,563,111
H. bakeri EVs	polyphosphatase	2	28,401,822
H. bakeri EVs	no treatment	2	11,064,217
C. elegans wild type	polyphosphatase	4	96,068,369
C. elegans *prg-1* mutant	polyphosphatase	4	45,082,526
C. elegans PRG-1 IP	polyphosphatase	2	24,536,893
C. elegans input	polyphosphatase	2	17,496,523
C. elegans *prg-1* sensor	piRNA 21UR-1349	1	22,962,449
C. elegans sensor	piRNA 21UR-1349	2	33,739,610

We first used Reaper and Cutadapt to remove the Illumina small RNA adapter sequences. Then, for mapping all reads to the reference genomes, we used ShortStack, with the parameters --mismatches 2 --mmap u --bowtie_m 500 --

ranmax 500 --bowtie_cores 20. As reference genomes we used c_elegans.PRJNA13758.WS254.genomic.fa.gz for *C. elegans* and heligmosomoides_polygyrus.PRJEB15396.WBPS14.genomic.fa for *H. bakeri*, from https://parasite.wormbase.org.

Genomic annotation and quantification of ncRNA production

In order to assign the production of ncRNAs to categories in both genomes in a consistent manner, we built custom genome annotations. Because we don't want to deal with overlaps between sequences, we developed a program which uses the setdiff function from the GenomicRanges R package (Lawrence et al., 2013), to remove overlaps. This program uses a category hierarchy to assign the intersect of two or more annotations. Our hierarchy prioritizes known producers of small RNAs, then TEs, exons and introns, in the last part of the hierarchy there are unknown elements in the genome. When we find an overlap in the annotation, we use the setdiff function to obtain the non-intersected portion and the intersect. We remove the intersection from the category with less hierarchical priority and we assign the intersected positions to the sequence with highest hierarchical priority.

Specifically for TEs, for both nematodes we used TE predictions produced with RepeatMasker and RepeatModeler (Chow et al., 2019), as TE annotations. Using models of repetitive elements within Dfam and RepBase data bases, RepeatMasker identifies and classifies TE sequences at family or class level. On the other hand, RepeatModeler is a *de novo* repetitive element identification tool, and searches for new models of repetitive sequences across each genome (http://www.repeatmasker.org). We use a sub classification to prioritize TEs, this classification is based on the number of sRNAs that align to each family independently of overlaps with other sequences (highest priority goes to the TE class with the most unambiguously aligned sRNAs). We also annotated the antisense regions for all classifications to evaluate the alignments of sRNAs on both strands. Finally, in each annotation we have individual and non-overlapped positions for each sequence, so as genomic categories we have sRNA families, TE families, exons, introns and Novel repeat elements. The level of sRNA

production in each genomic region was obtained using the findOverlaps function from GenomicRanges R package (Lawrence et al., 2013). In order to avoid overlaps, we count according to the overlap of the central nucleotide of each read, using the resize function, with parameter fix="center" from the GenomicRanges R package.

Linear models and siRNA production

In order to understand if the length in nucleotides of each TE class or family biased the production of sRNAs, we used linear models. We compared the total summed length in nucleotides of each TE class or family against the number of sRNAs that align to each category. Based on a confidence interval, we classified TE classes and families as over or under producers of TE-derived sRNAs.

Differential expression and gene set enrichment analysis

For differential expression analysis, we used the edgeR R package (Robinson et al., 2009). We use the counts of sRNAs produced with resize::fix="center" and countOverlaps(query = genome, subject = reads, minoverlap=1, type="within", ignore.strand=FALSE) in each sRNA-seq experiment to perform the comparisons. We performed 6 differential expression analyses (see **Table 2**).

Table 2. Comparisons of sRNA-seq experiments for differential expression analyses. Log_2FC column represents the magnitude of change between compared conditions that we used as cutoff values to distinguish between expressed sequences in one or other condition of the comparison. FDR column represents the cutoff value that we used as false discovery rate.

Species	Comparison	Log_2FC	FDR
C. elegans	*C. elegans* prg-1 mutants vs wild type nematodes	2	.05
	C. elegans IP vs input	2	.05
	C. elegans prg-1 mutant 100 nt widows vs wild type 100 nt	2	.05
H. bakeri	EVs vs adult nematodes	1.5	.05
	EVs polyphosphate vs EVs monophosphate	1	.05

For each differential expression analysis, we estimated the trended dispersion, in order to do not overestimate individual dispersions, fit general linear models, and used the glmTreat to perform the comparisons. We used glmTreat instead of glmQLFTest because this allowed us to identify more biologically relevant sequences, by requiring expression changes to be significantly greater to a predefined log_2FC cutoff, instead of simply significantly different from 0 (Mccarthy & Smyth, 2009).

In order to find enriched categories within the deferential producers of sRNAs we used the *fry* function of the Limma R package (Ritchie et al., 2015). This function allows to evaluate the enrichment of predefined categories. *Fry* is a type of "self-contained" gene set test and evaluates whether each gene category is enriched in one or the other direction. The null hypothesis of *fry* is that none of the genes of a category is differential expressed. After using linear models based on the expression values for differentially expressed sequences, *fry* uses the residual values to establish the direction of the category (Wu et al., 2010).

piRNA target prediction

piRNA target prediction was performed using the lists of *C. elegans* and *H. bakeri* piRNAs published by (Beltran et al., 2019). As potential piRNA targets we used all TEs in the genomes, predicted with RepeatModeller and RepeatMasker as described by Chow and co-workers (Chow et al., 2019). The piRNA target prediction initially was performed using a custom version of pirScan (Zhang et al., 2018). For this analysis, we allowed a maximum of 6 mismatches in the interaction piRNA-TE, 2 mismatches in the seed region, and a maximum of 3 GU mismatches in the non-seed region. As a control we used the last version of pirScan (http://cosbi4.ee.ncku.edu.tw/pirScan/).

To improve piRNA target prediction we developed a pipeline called pirFinder. This pipeline uses the same piRNA targeting rules to produce a score for each piRNA-TE interaction. pirFinder allows a maximum of 6 mismatches in the whole interaction, no more than 2 GU mismatches in the seed region (bases 2-7), up to 3 non-GU mismatches outside the seed and up to 2 GU mismatches outside the seed (Zhang et al., 2018).

PirFinder uses bwa 0.7.12 (Li & Durbin, 2010) to create an index of TEs and then to align each piRNAs in our piRNA database to each TE. For the prediction of piRNA targets we used, for *C. elegans,* 10,096 type I piRNAs and 4,572 piRNAs for *H. bakeri* (Beltran et al., 2019). It is important to note that the online version of pirScan uses 15,364 type l and 2,485 type ll piRNAs, to perform piRNA target predictions in *C. elegans*. Additional piRNAs in pirScan data base could produce differences in the number of predictions comparing pirScan against pirFinder.

For indexing pirFinder uses bwa index -a bwtsw $GENE_FA, for the alignment we used bwa aln -t 8 -n 6 -N -o 0 $BWA_INDEX $PIRNA_FA > $sai and samtools faidx $GENE_FA > $fai. BWA_INDEX is the indexed positions of the input sequence, PIRNA_FA is a multifasta file which contains the piRNA sequences to align and finally GENE_FA is the input sequence in fasta format. After the alignment, sam2tsv (Pierre, 2015) is used to build a table with the position of mismatches: java -jar ~/software/jvarkit/dist/sam2tsv.jar -r $GENE_FA > output. Finally, we process the output of sam2tsv in a custom R script with the Biostrings R package (Pagès et al., 2019). This script finds the GU and non-GU mismatch positions. The score of pirFinder is based on scale from 10 to 0 as pirScan. If they are no mismatches the score is 10. Each GU mismatch within the seed of the piRNA subtracts 1.5, while GU mismatches and non-GU mismatches out of the seed subtract 1.5 and 2 respectively (Wu et al., 2018).

In order to predict piRNA targets in both genomes we used RepeatModeller and RepeatMasker TE predictions as described by Chow and co-workers (Chow et al., 2019). We filtered the original data set of TEs to keep only those longer than 50nt.

Validation of piRNA target prediction

After running pirFinder, we used the GenomicRanges R package to get the coordinates of the piRNA interaction within each TE. After obtaining the coordinates we used the GenomicRanges (Lawrence et al., 2013) resize function with the parameters fix = "center", width=100, this allowed us to get a window of interaction between piRNAs and TEs.

We evaluated the production of TE-derived sRNAs within each window using the GenomicRanges countOverlaps function using each library of sRNA-seq. In order to determine if piRNAs trigger the production of TE-derived sRNAs, we performed a differential expression analysis with the counts of these windows, comparing *prg-1* mutants against wild type nematodes. Because more than one piRNA can bind with the same TE in the same window we used the expression of each piRNA in the PRG-1 IP library as a tiebreaker criterion, preferring the one with highest counts. With this approach we finally get one piRNA per window.

With these differential expression results, we also performed gene set enrichment analysis using *fry* as described above. With this we aimed to identify TE classes or families in which piRNAs truly trigger the production of TE derived sRNA.

Results:

Genomic comparison of *C. elegans* and *H. bakeri*

We started by comparing the number of nucleotides that each genomic category represents in the genome, for example exons, introns, TEs etc. (**Figure 1**). We found that the *H. bakeri* genome is nearly 7 times bigger than the *C. elegans* genome. The main difference between both genomes is the number of nucleotides that repetitive elements (such as DNA transposons, retro elements and novel repeats) represent in the genome. To give you an idea, retroelements (LINES, SINES and LTRs) represents 1.3% of the *C. elegans* genome, which means 1,307,772 bases, while in *H. bakeri* these elements represents 15% of the genome, or 104,917,538 bases.

Although in *H. bakeri* TE use more bases in the genome, *C. elegans* has a slightly higher number of different TE families (**Table 3**). It is important to highlight that DNA transposons are the most abundant class of TEs in *C. elegans*, while in *H. bakeri* LINEs and novel repeats seems to be the most abundant classes (**Figure 1**). This is important, because by chance the biggest categories can produce more TE-derived sRNAs (see below).

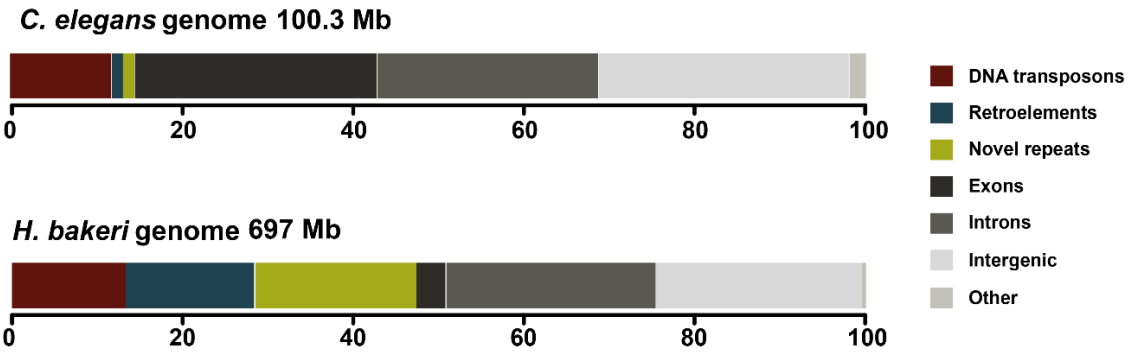

Figure 1. Genomic comparison of *C. elegans* and *H. bakeri*. Different colors represent the proportion in the genome annotated with each category. Both genomes are scaled to 100%.

Table 3. Comparison of TE diversity and genomic proportion between *C. elegans* and *H. bakeri.*

	TE families	Proportion of the Genome	Number of nucleotides	Shared families
C. elegans	104	13%	13,282,817	64
H. bakeri	94	47%	329,373,957	64

Small RNA transcriptomic comparison of *C. elegans* and *H. bakeri*

After the genomic comparison, we analyzed the sRNA-seq libraries. We first focused on the polyphosphatase-treated libraries of adult nematodes of *C. elegans*, adult nematodes of *H. bakeri* and EVs **(Figure 2**). This comparison considers the length distribution of the sequenced sRNAs and the first nucleotide at the 5' end. At least in *C. elegans* these characteristics are related with the specificity of Argonaute proteins which load these sRNAs. We found in *C. elegans*, as reported previously (Stricklin et al., 2005), that 22G sRNAs are the most abundant class of sRNAs. In *H. bakeri*, however, 22G and 23G sRNAs are the most abundant classes of sRNAs (see **Figure 2**). Nevertheless, when we

independently align *H. bakeri* 22G and 23G, we found that these RNAs come in almost the same proportion from similar genomic categories in EVs (**Figure 3**).

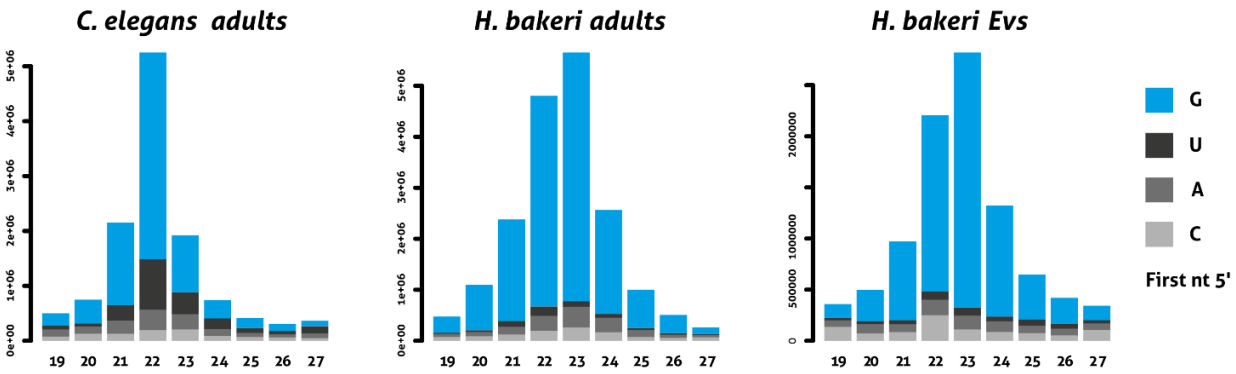

Figure 2. Transcriptomic comparison of sRNAs from polyphosphatase-treated libraries between *C. elegans*, *H. bakeri* adult nematodes and *H. bakeri* EVs. Different colors in each bar represent the first nt at the 5' end. X axis represent the length distribution of sRNAs, and Y axis represent the frequency.

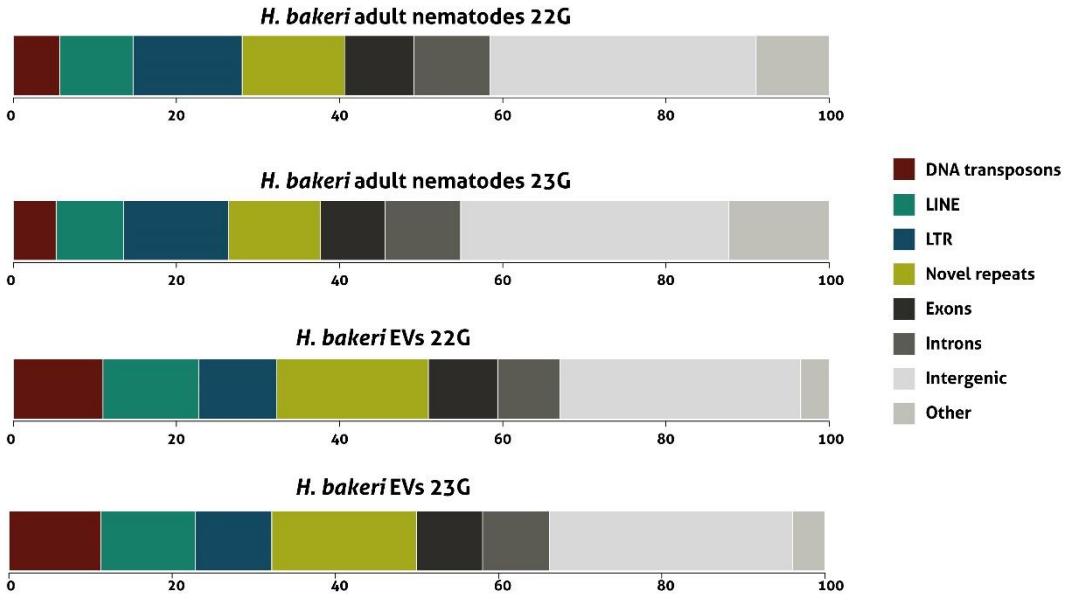

Figure 3. Genomic alignments of polyphosphatase-treated sRNAs 22G and 23G, comparing *H. bakeri* adult nematodes. Different colors in each bar represent the genomic category from which sRNAs come from.

Although the proportion of genomic categories from which 22G and 23G are produced seems very similar, in adult nematodes, there are other categories

which include for example rRNA or tRNA which are more different. Notably when we compare the production of 22G vs 23G in EVs, the proportions look almost equal. This reflects that the population of these sRNAs in adults is more heterogeneous.

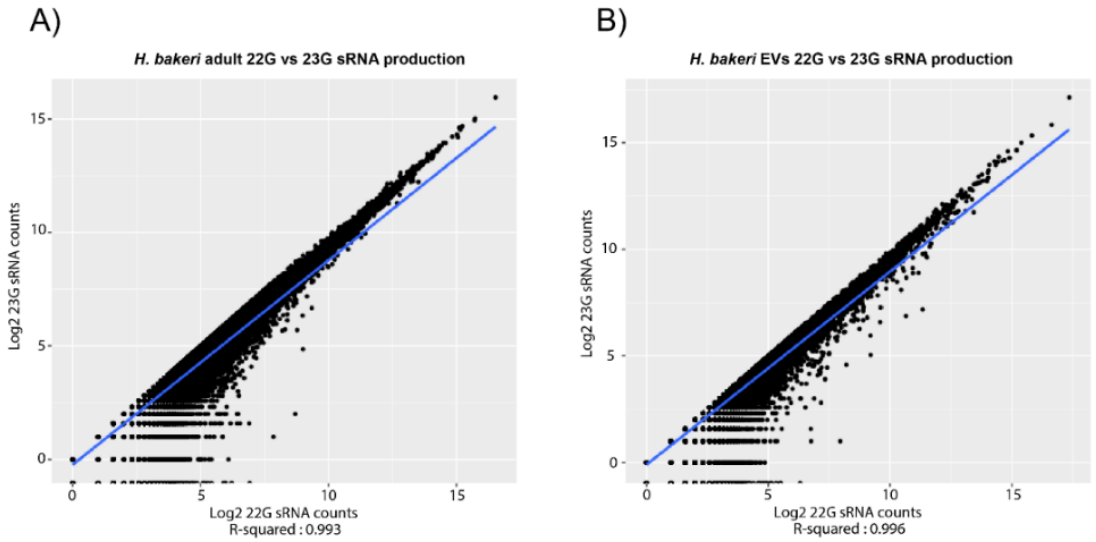

Figure 4. Comparison of sRNA production between 22G sRNAs and 23G sRNAs using genomic positions. A) 22G against 23G sRNAs using the mean expression value of sRNAs in 3 adult nematode sRNA libraries. B) Comparison of 22G and 23G sRNA production using the mean expression of sRNAs in 2 EVs sRNA libraries. Each dot represents all annotation of the genome in *H. bakeri* genome with at least 1 CPM.

To quantitively confirm that 22G and 23G counts are highly correlated, we compared their counts in every annotation of the genome (**Figure 4**). We thus decided to use all reads between 19-25 nt, to compare adult nematodes of *C. elegans*, adult nematodes of *H. bakeri* and EVs. In *C. elegans*, most of the sRNAs come from exons, and just a small proportion comes from TEs, while in *H. bakeri* adults at least 42% of sRNAs come from TEs such as DNA transposons, LINEs, LTRs and novel repeats. EVs, however, exhibit an increase in sRNA production from novel repeats and LINEs, with respect to the adults (**Figure 5**). Comparing *C. elegans* and *H. bakeri* adults, the differences in the categories which produce these sRNAs may be due to the number of nucleotides that these categories represent in the genome.

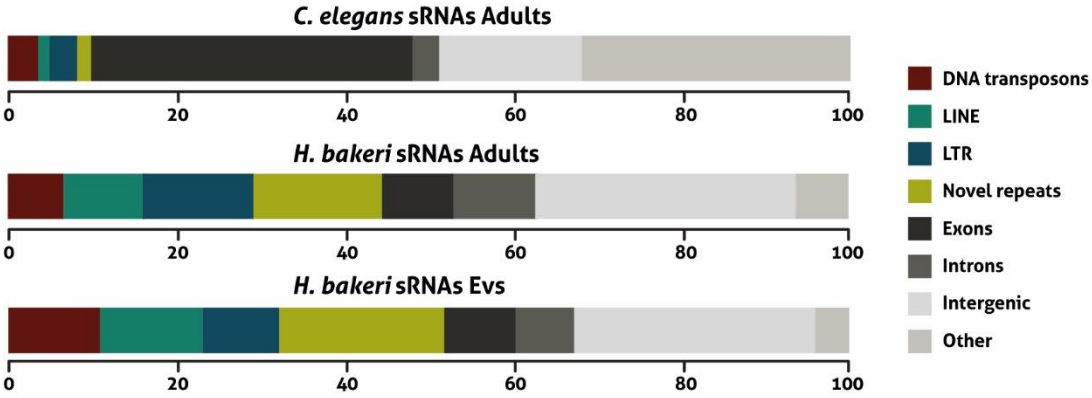

Figure 5. Genomic alignments of polyphosphate sRNAs of 19-25 nt, comparing *C. elegans*, adult nematodes of *H. bakeri* and EVs. Different colors in each bar represent the annotated genomic regions.

Length of genomic regions and sRNA production

In order to answer if categories with more nucleotides produce more sRNAs, we performed a linear regression comparing the number of nucleotides against the number of sRNAs that are produced from these categories (**Figure 6**).

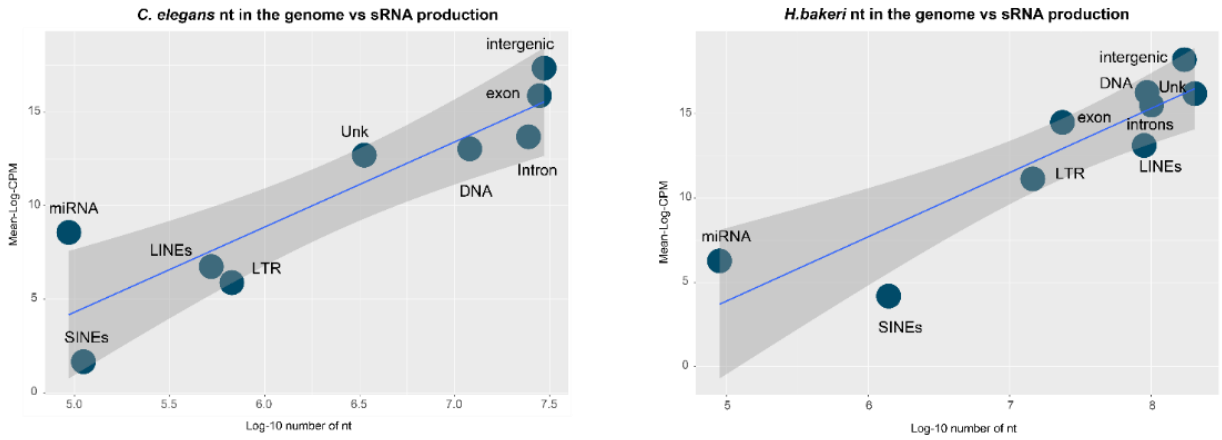

Figure 6. Linear regressions comparing the total nucleotides that each category occupies in the genome and the number of sRNAs that are produced.

As expected, categories with more nucleotides produce more sRNAs. Also, we found that some categories such as miRNAs at least in *C. elegans* produce more

sRNAs than expected by their number of nucleotides. We are, however, more interested in the production of TE-derived sRNAs. As described by Youngman & Claycomb (2014), these sRNAs are produced antisense to transcripts, because of this we decided to sub classify our sRNAs and genomic categories in sense and antisense and fit the linear models again (**Figure 7**). These analyses reveal that at the class level, LTRs in antisense produce even more sRNAs than expected in both nematodes (**Figure 7 A, C & E**). At the family level, we found many TEs that produce more sRNAs than expected by their number of nucleotides. With respect to LTRs, just *Pao* and *Gypsy* families are shared between both adult nematodes as producing even more sRNAs than expected (**Figure 7 B & D**). Although LTR activity in *C. elegans* is controversial (RuvkunG, 2019), *Pao* and *Gypsy* elements are some of the most dispersed TE families in animals (De La Chaux & Wagner, 2011). Also, in both nematodes these families represent the largest proportion of LTRs. As reported previously the DNA/*TcMar* superfamily is an active producer of sRNAs in *C. elegans* (Bagijn et al., 2012), and this is confirmed in our results. In *H. bakeri*, however, although this superfamily is present in the genome, the production of sRNAs is not enriched with respect to the number of nucleotides occupied in the genome. There are however other DNA families, such as DNA/*hAT-Tip100* elements which are producing even more sRNAs than expected in *H. bakeri* (**Figure 7 D & F**). This family has the highest residual value, which means that this family is an over-producer of sRNAs. Using linear model approach with EVs sRNAs, we found almost the same distribution as in *H. bakeri* adult nematodes (**Figure 7 D**).

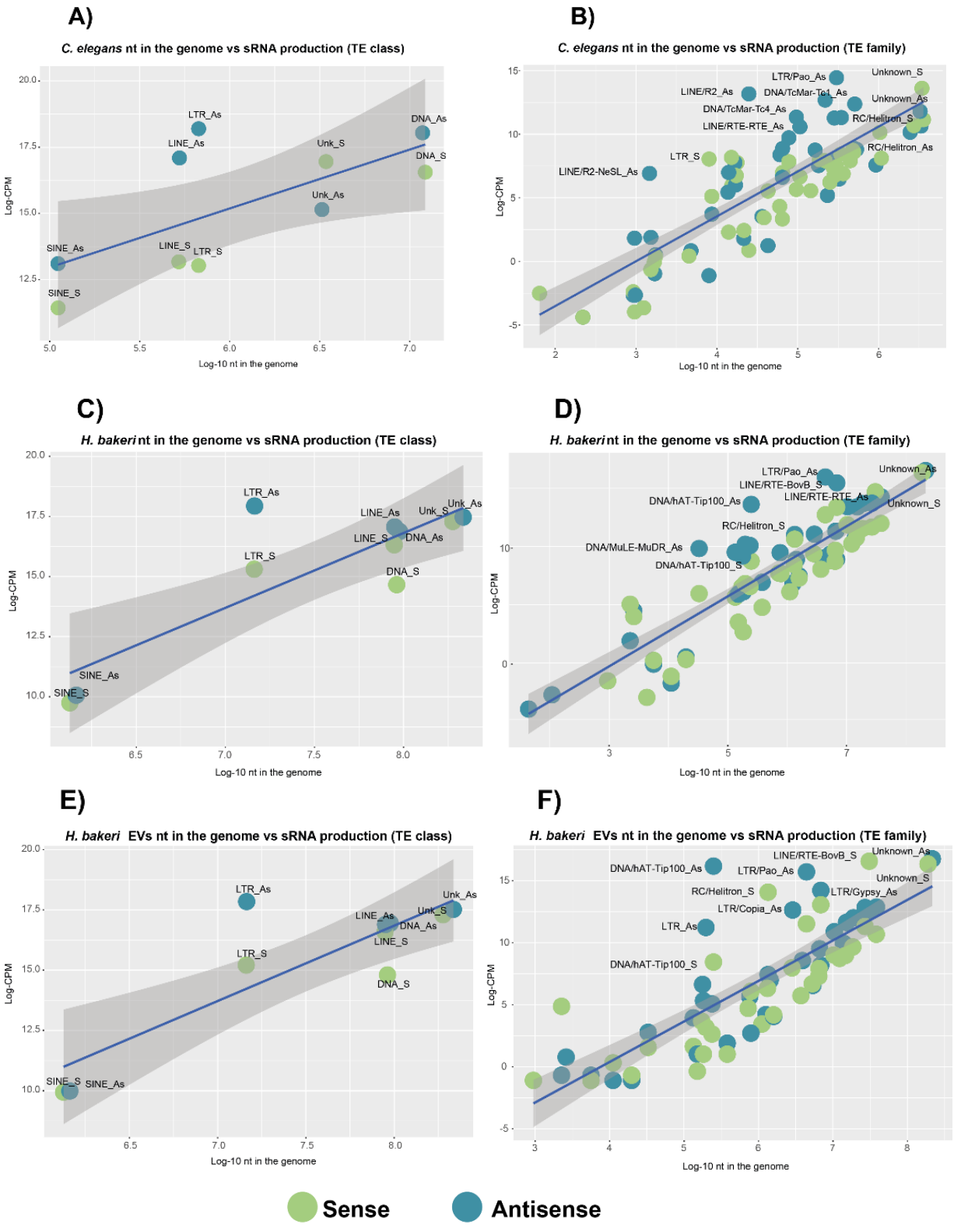

Figure 7. Linear regressions comparing the nucleotides which each category use in the genome and the number 19-25 nt sRNAs that are produced, subclassifying categories in sense and antisense. A, C and E represent *C. elegans*, *H. bakeri* and EVs, linear regressions at TE class level respectively. B, D and F show *C. elegans*, *H. bakeri* and EVs, linear regressions at TE family level. Green color represents sense categories while blue represents antisense categories.

Inconsistent strand in sRNA production

Unexpectedly, some families such as *RC-Helitrons* (present in both nematodes) and *LINE/RTE-BovB* (exclusive family of *H. bakeri*) produce more sRNAs in sense than in antisense in *H. bakeri* (**Figure 8**). This result could have two explanations, these TE families may produce sRNAs by bidirectional transcription, or it is also possible that the TE motif in RepeatMasker prediction has the inverse direction. If these sRNAs are products of bidirectional transcription we expect almost the same proportion of sRNAs in both directions, while if the motif has the inverted direction, we expect a bias towards the sense strand. To answer this question, we first used the IGV browser to visualize how sRNAs are produced from individual loci (**Figure 9**). *RC-Helitrons* (in both nematodes) and LINE/*RTE-BovB* (just in *H. bakeri*) exhibit sRNA production biased towards the sense direction of the annotation. These results suggest that the motifs for both families have the wrong direction.

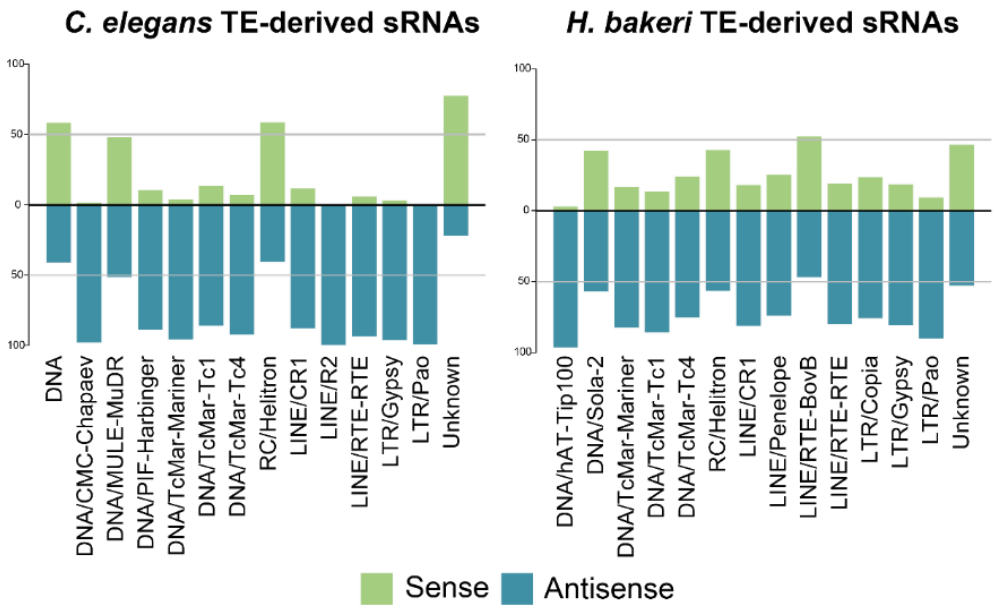

Figure 8. Comparison of sRNA proportion in sense or antisense direction of TE families which have more than 5000 assigned reads. Up direction of the bar's represents sRNA production while down direction represents sRNA production in antisense.

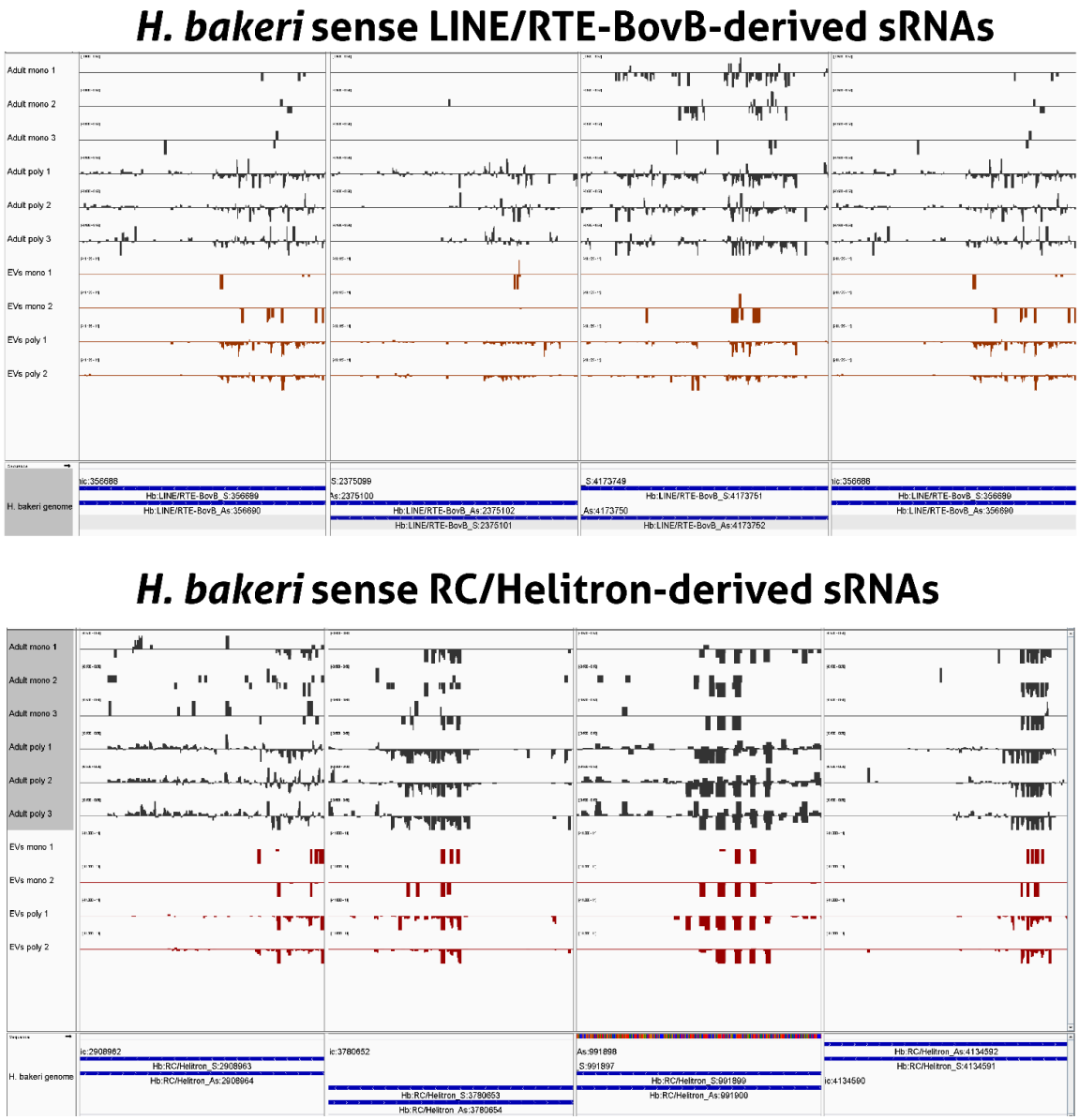

Figure 9. Genome browser visualization of sRNA stacking in 4 different *H. bakeri* LINE/RTE-BovB and RC/Helitrons. Dark gray bars represent sRNAs in adult nematodes while red bars represent sRNA in EVs. For each figure the first 3 and last 2 rows represents polyphosphate sRNAs while the others represent monophosphate sRNAs.

It seems that the TE-derived sRNAs within vesicles are a subset of the sRNA population in the adult. We decided to go one step further, in order to understand if specific TEs classes or families are selective in the production of secreted TE-derived sRNAs, so we decided to perform differential expression analysis.

The production of secreted sRNA is enriched in TE classes and families

In order to determine if specific genomic elements were responsible for EV content, we compared sRNA production between EVs and adult nematodes of *H. bakeri* using the polyphosphatase treated libraries (see Methods). With this comparison we obtained 4,760 regions enriched in EVs with respect to adults. The MA plot reveals however that with this comparison we are identifying more accurately differential expressed sequences in adults (**Figure 10**). So, we decided to perform a gene set enrichment test (for more details see Methods), to identify groups of sequences enriched in adult nematodes, in order to use their expression values to calculate more appropriate normalization factors using edgeR (**Table 4**). With this approach we can center the expression and more accurately capture differential expressed sequences specific to EVs.

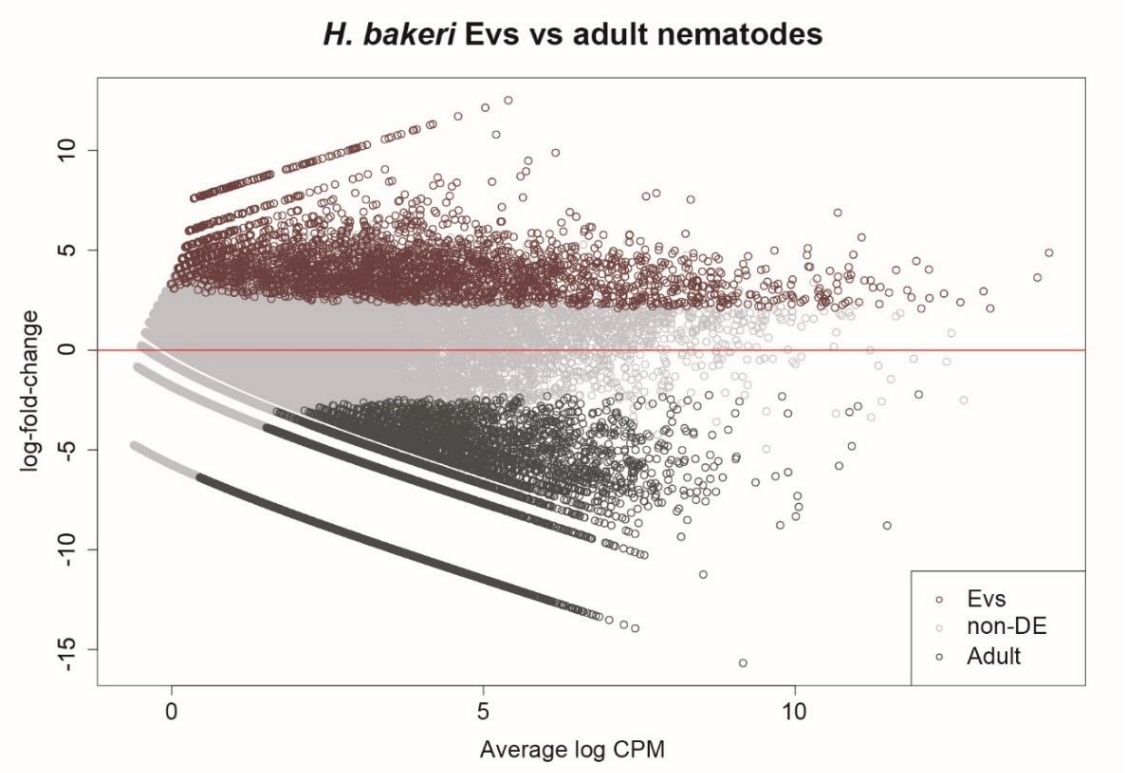

Figure 10. Differential expression analysis comparing *H. bakeri* EVs against adult nematodes using just aligned sRNA counts as initial factor of normalization.

Table 4. Gene set enrichment analysis comparing EVs and adult nematodes, using TE class classification. Down direction represents the enrichment of a TE classes in adult nematodes.

	Direction	FDR
DNA	Down	<0.0001
Exon	Down	<0.0001
Unks	Down	<0.0001
LINE	Down	0.0002
LTR	Down	0.0004
SINE	Down	0.0053

At the level of TE class, gene set enrichment analysis showed that all classes are enriched in adult nematodes, these results strengthen the idea that we need another approach to get information about EVs.

Table 5. Gene set enrichment analysis comparing EVs and adult nematodes, using TE family classification. Down direction represents the enrichment of a TE classes in adult nematodes.

	Direction	FDR
introns_As	Down	<0.0001
introns_S	Down	<0.0001
Unknown_S	Down	0.0001
DNA/MuLE-MuDR_As	Down	0.0001
DNA/PiggyBac_S	Down	0.0001
LINE/Penelope_As	Down	0.0001
exons_As	Down	0.0001
LTR/Gypsy_As	Down	0.0002
DNA/TcMar-Tc4_As	Down	0.0003
LINE/Penelope_S	Down	0.0004
LINE/CR1_As	Down	0.0008
DNA/TcMar-Tc1_As	Down	0.0008
LTR/Pao_As	Down	0.0018

Based on FDR values of the gene set enrichment test we found both introns in sense and antisense as enriched sequences in adult nematodes (**Table 5**). Because both categories exhibit a similar expression pattern (**Figure 11 A and B**) we decided to use the whole category to recalculate edgeR normalization factors.

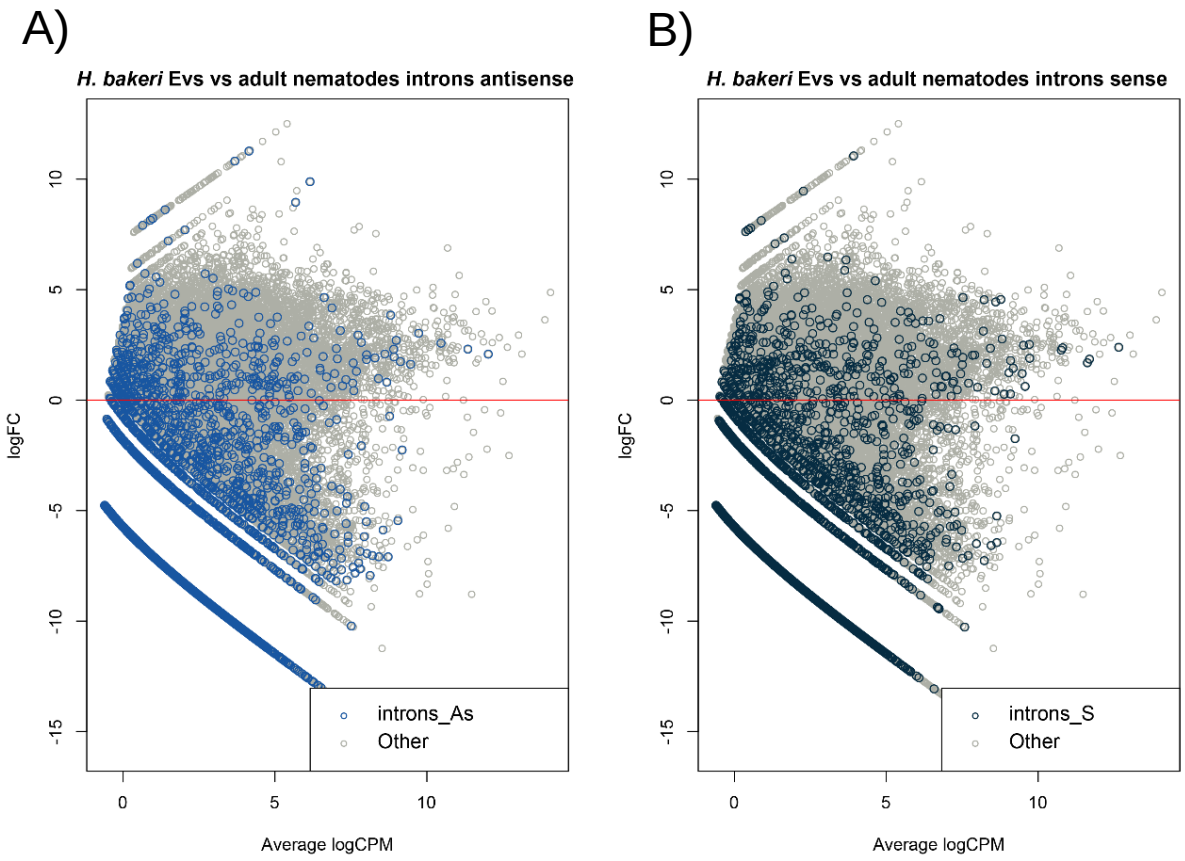

Figure 11. Expression pattern of all antisense (A) and sense (B), introns comparing sRNA production between EVs and adult nematodes.

After this new normalization and using the same cutoff values as before, we identified 7,272 sequences enriched in EVs, an increase of ~53% (**Figure 12**). With this new approach that allowed us to better focus on the EV-enriched genomic regions, we use the *fry* gene set enrichment test to answer if specific TE classes or families are enriched in EVs or adults. At the class level, based on FDR values we found LINE-derived sRNAs as enriched within EVs (**Table 6**). Although, initial comparisons (see **Table 4**) of the sRNA content in adults and EVs suggested an increase of TE classes in adult nematodes with respect to EVs. Using a more appropriate normalization strategy, LINEs seem to be the only class that is statistically enriched within EVs (see **Table 6**).

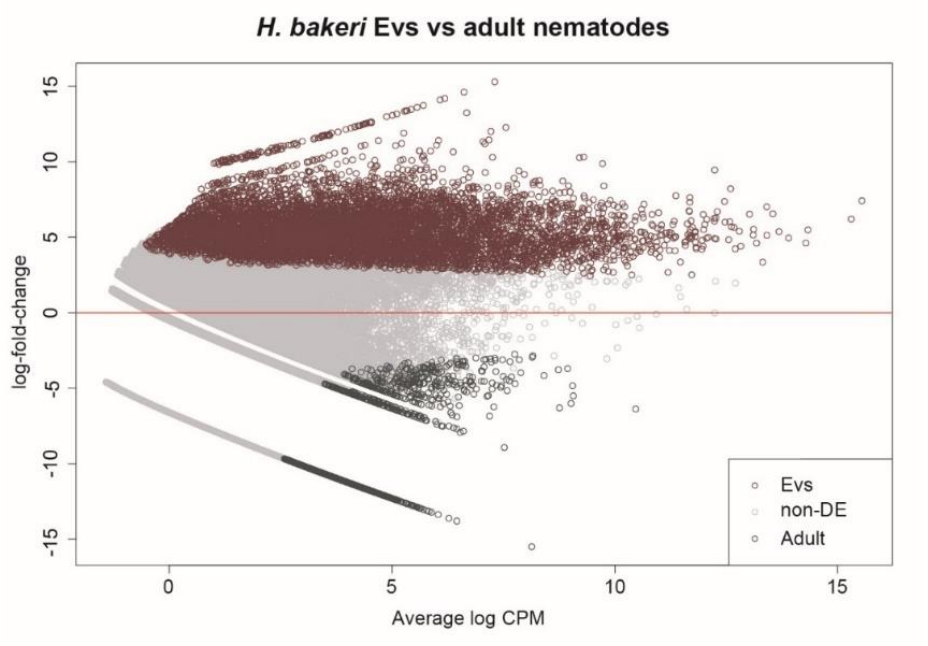

Figure 12. MA plot of differential expression analysis comparing *H. bakeri* EVs against adult nematodes after normalization by the expression pattern of introns.

Table 6. Gene set enrichment analysis after using introns to calculate normalization factors. Down direction represents the enrichment of a TE classes in adult nematodes. Up direction represents the enrichment of TE classes in EVs.

	Direction	FDR
LINE	Up	0.158
LTR	Down	0.158
DNA	Down	0.185
Unks	Down	0.186
SINE	Down	0.383

Table 7. Enrichment of sRNA production in TE families after using introns to calculate normalization factors. Up direction represents the enrichment of a TE families in EVs.

	Direction	FDR
LINE/RTE-BovB_S	Up	0.029
DNA/hAT-Tip100_As	Up	0.029
LINE/RTE-RTE_As	Up	0.115
RC/Helitron_S	Up	0.135

We next performed the same analysis at the TE family level. Based on FDR <0.2 we found four enriched TE families as producers of sRNAs (**Table 7**). We found sense LINE/*RTE-BovB* and RC/*Helitron*, and antisense elements of the superfamily LINE/*RTE-RTE* and DNA/*hAT-Tip100* TEs. It is important to mention that regardless of the TE class or family enrichment, a variety of individual genomic regions which do not necessary belong to enriched families are the source of the most abundant sRNAs within EVs (see **Table 8**). In fact, the most abundant sRNAs within EVs are produced from Unknown elements (see **Table 8**). In contrast to Chow et al. (2019), we found that the whole class of Unknown elements are not enriched in EVs, but it seems that just a few genomic regions produce the majority of the sRNAs assigned to this class.

Table 8. Top 15 EVs TE-derived sRNA producers, using the average of CPM in two EVs libraries ordered by the LogCPM.

	aveLogCPM
Hb:Unknown_As:178627	15.25
Hb:Unknown_S:178629	14.98
Hb:DNA/hAT-Tip100_As:1775745	13.93
Hb:LINE/RTE-BovB_S:1775746	13.36
Hb:DNA/hAT-Tip100_As:2533766	13.26
Hb:DNA/hAT-Tip100_As:4201113	13.08
Hb:LTR/Gypsy_As:1721053	12.79
Hb:DNA/hAT-Tip100_As:634152	12.79
Hb:DNA/hAT-Tip100_As:206819	12.73
Hb:DNA/hAT-Tip100_As:2488294	12.38
Hb:RC/Helitron_S:1985031	12.29
Hb:RC/Helitron_S:4134591	12.28
Hb:Unknown_As:178633	12.27
Hb:DNA/hAT-Tip100_As:4334041	12.17
Hb:RC/Helitron_S:991899	12.13

As mentioned in the Introduction, TE-derived sRNAs should have an antisense direction if their biogenesis depends on RdRps. Intriguingly, gene set enrichment test reveals sense LINE/*RTE-BovB* and RC/*Helitron* as enriched producers of secreted sRNAs (see **Table 7**). We compared the expression pattern of sense and antisense elements in both TE-families. MA plots and gene set enrichment analysis confirm the bias in the direction of sRNA production. It is also important to note that for LINE/*RTE-BovB* and RC/*Helitron* families, antisense elements

seem to be less expressed in general, and enriched in adult nematodes (**Figure 13**).

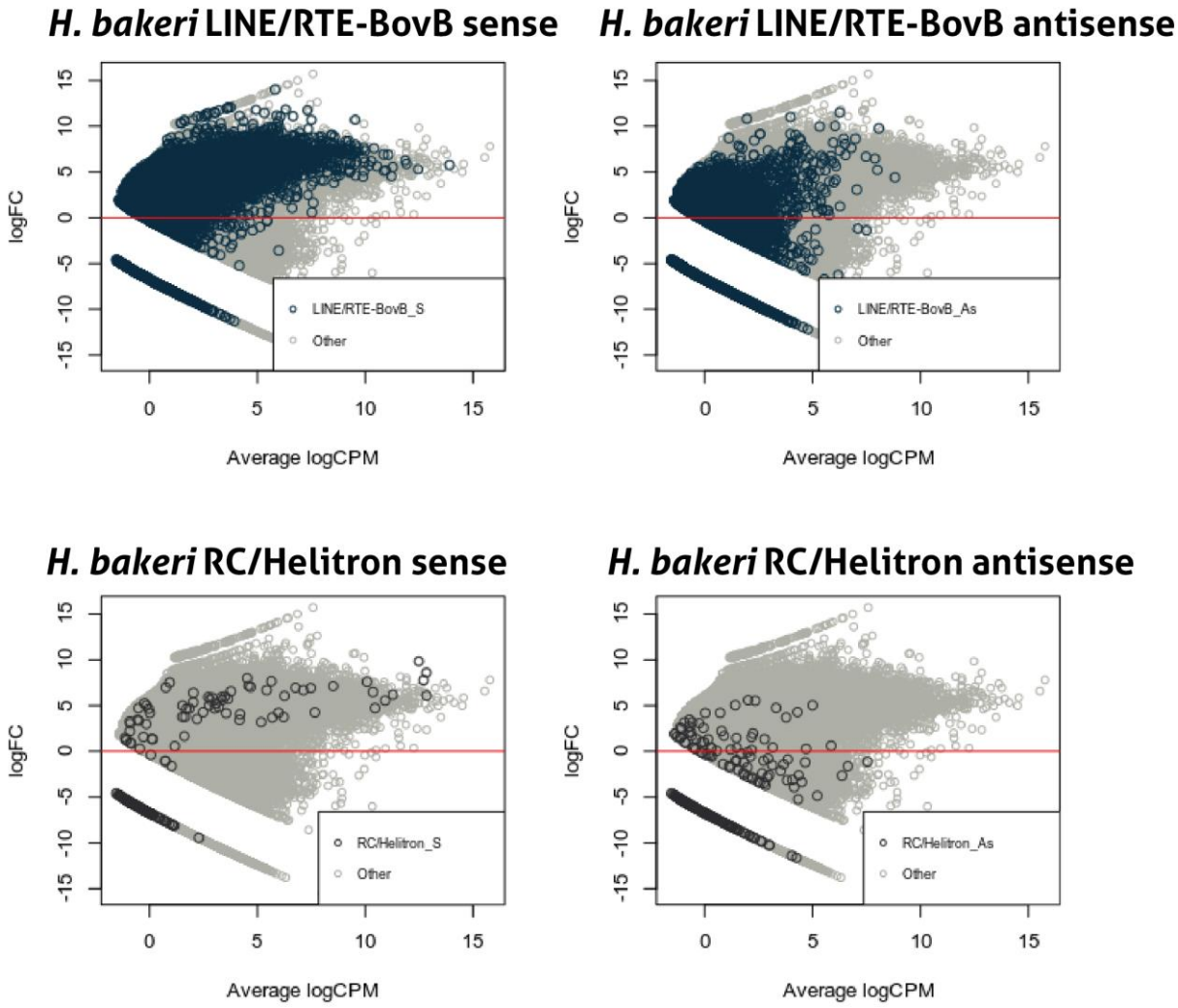

Figure 13. MA plot of sense and antisense LINE/*RTE-BovB* and RC/*Helitron* sRNA production during EVs and adult nematodes comparison.

Because of the unexpected strand of LINE/*RTE-BovB* and RC/*Helitron* derived sRNA production, we decided to try to better understand the biogenesis of these sRNAs. If the production of LINE/*RTE-BovB* and RC/*Helitron* derived sRNA truly depend on RdRps they would be antisense to the transcript, and the sRNAs would have a 5' triphosphate (triP). We decided to perform differential expression and gene set enrichment analyses comparing polyphosphatase-treated (polyP) libraries (required for detecting 5' triP sRNAs) against untreated libraries from EVs. If LINE/*RTE-BovB* and RC/*Helitron* sRNAs are biased to the polyP condition

it would indicate that their biogenesis depends on RdRps. Since EV untreated sequenced sRNAs are a subset of the polyP sRNAs we decided to use more relaxed cutoff values during differential expression analysis. In addition, because rRNAs are not RdRp products, we used rRNA to calculate normalizing factors. We used as cutoff values log_2FC significantly greater than 1 and FDR < 0.05. With these parameters we found 12,544 producers of sRNAs enriched in the polyP condition (**Figure 14**). Furthermore, gene set enrichment analysis confirmed the bias of LINE/*RTE-BovB* and RC/*Helitron* sRNAs on their sense strand towards the polyP condition (**Table 9**). These results indicate that LINE/*RTE-BovB* and RC/*Helitron* sRNAs annotated on the sense strand are RdRp products. With this evidence we conclude that at least in *H. bakeri* the production of sRNAs do not correspond to the model direction in RepeatMasker prediction.

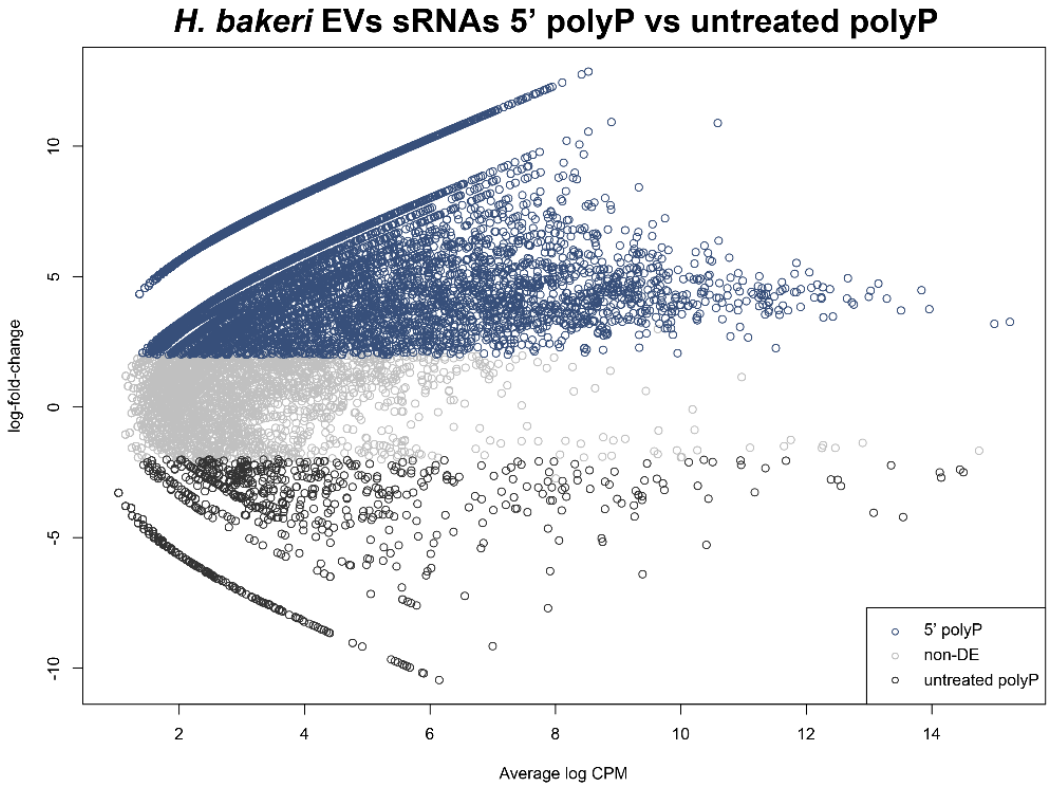

Figure 14. MA plot of differential expression analysis comparing *H. bakeri* EVs 5'triP sRNAs against EVs untreated polyP sRNAs.

With respect to DNA transposons we found enriched antisense DNA/*hAT-Tip100* elements within vesicles in comparison with adult nematodes. With polyP against untreated differential expression analysis, we also found enriched DNA/*hAT-Tip100* sRNAs in polyP. These results reveal DNA/*hAT-Tip100* sRNAs also as RdRp products. As a control, we expected that sense miRNAs are enriched in the untreated condition, and the gene set enrichment test confirmed this assumption (**Table 9**).

Table 9. EV enriched TE families in polyP against untreated sRNA comparison. Down direction represents the enrichment of a TE families in 5' polyP. Up direction represents the enrichment of TE families in untreated polyP.

	Direction	FDR
LINE/RTE-BovB_S	Up	0.027
Unknown_As	Up	0.027
Unknown_S	Up	0.027
LTR/Pao_As	Up	0.027
exons_As	Up	0.027
introns_As	Up	0.027
LTR/Gypsy_As	Up	0.027
DNA/TcMar-Mariner_As	Up	0.027
LINE/RTE-RTE_As	Up	0.027
DNA/hAT-Tip100_As	Up	0.027
LINE/CR1_As	Up	0.027
rRNA_S	Down	0.027
miRNA_S	Down	0.027
introns_S	Up	0.032
exons_S	Up	0.032
RC/Helitron_S	Up	0.037
DNA/TcMar-Tc1_As	Up	0.037

Full length transposons contribute to EV content

Regarding TEs that contribute substantially to vesicle content, we were curious if these elements are full length transposons (more likely to be functional) or only fragments. We plotted the cumulative length distribution of these TEs (**Figure 15**).It has been reported that LINE/*RTE-BovB* elements have a length of 3100 to 3200 nt, depending on an insertion of a SINE element (Dunemann & Wasmuth, 2018). We found that nearly 30% of the LINE/*RTE-BovB* elements that are

enriched in EVs sRNA production have a length of 3100 nt. RC/*Helitrons* on the other hand seem to have a bimodal distribution (**Figure 15 A**), it could be because RC/*Helitrons* are subclassified into two groups, autonomous elements (2000-4500 nt) and non-autonomous (100-1500 nt) (Touati et al., 2018).

Using as an example LINE/*RTE-BovB*, we found that TEs that are the source of secreted TE-derived sRNAs are bigger than the TEs of this family which are enriched in adult nematodes. In addition, producers of secreted TE-derived sRNAs seems to be the biggest LINE/*RTE-BovB* within this TE family in the genome (**Figure 15 B**). These result reveal that an important fraction of TEs which produce secreted TE-derived sRNAs are full length transposons.

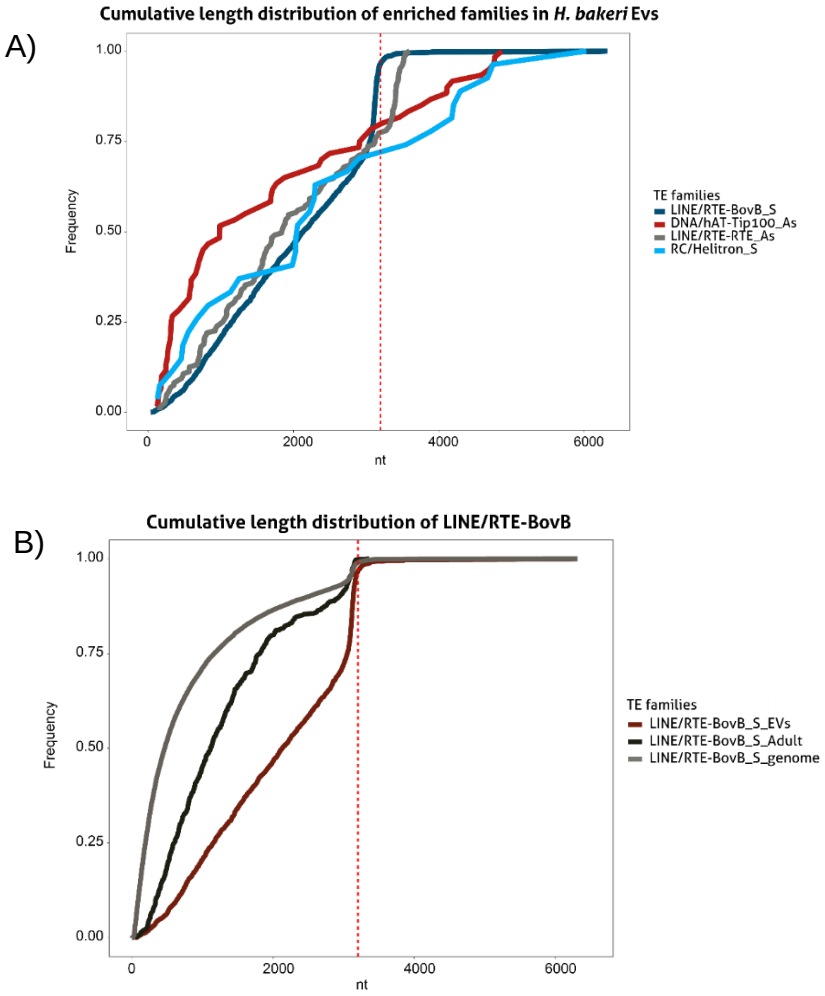

Figure 15. Cumulative length distribution of selected *H. bakeri* TE families. A) Comparison of cumulative length distribution between TE families enriched in EVs. B) Cumulative length distribution of sense LINE/*RTE-BovB*, comparing TEs enriched in EVs, adult nematodes or the full genomic annotations.

siRNA production is enriched in DNA TE in *C. elegans*

Because TE-derived sRNA biogenesis has been previously studied in *C. elegans,* we decided to use this species as a model of TE-derived sRNA production. It has been reported that piRNAs trigger TE-derived sRNA biogenesis at least in *C. elegans* (Ashe et al., 2012; Bagijn et al., 2012). There is however a lack of information at genome wide scale about how piRNAs act as triggers of TE-derived sRNAs. To understand how piRNAs trigger TE derived sRNA biogenesis we used a public sRNA-seq data set of polyphosphatase-treated sRNAs in *prg-1* mutants (GEO GSE37433). This data set contains 5 sRNA-seq libraries, of which 2 have a genetic construct to sense the function of piRNA 21UR-1349 (more detail in Methods and Bagijn et al., 2012). We compared *prg-1* mutants against wild type nematodes to look for TEs in which biogenesis of their related TE-derived sRNAs depends on PRG-1 and piRNAs. In order to have enough replicates for differential expression analysis, we decide to use sensor libraries considering a batch effect in the experimental design. We compared *prg-1* mutants including 1 sensor library against wildtype nematodes which also include 2 sensor libraries. We found 2,911 sRNA producers in wild type nematodes that are reduced in *prg-1* mutants, suggesting these sRNAs depend on PRG-1 for their biogenesis (**Figure 16**).

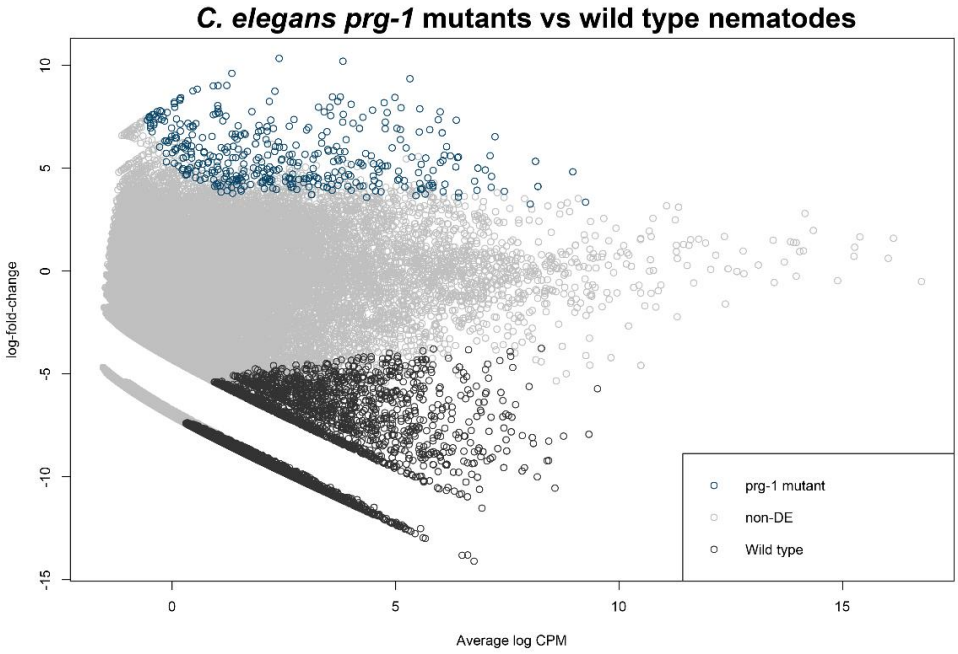

Figure 16. MA plot of differential expression analysis comparing *C. elegans prg-1* mutants against wild type nematodes, including piRNA 21UR-1349 sensor sRNA libraries.

We also performed gene set enrichment analysis, in order to find enriched TE classes and families that depend on PRG-1 (**Table 10** & **Table 11**).

Table 10. Gene set enrichment analysis of *C. elegans prg-1* mutants against wild type at TE class level. Down direction represents the enrichment of a TE classes in wild type nematodes. Up direction represents the enrichment of TE classes in *prg-1* mutants.

	Direction	FDR
DNA	Down	0.018
Unks	Down	0.031
LINE	Down	0.031
LTR	Down	0.045
Exon	Down	0.155
SINE	Up	0.264

Table 11. Gene set enrichment analysis of *C. elegans prg-1* mutants against wild type nematodes using TE family classification. Down direction represents the enrichment of a TE classes in wild type nematodes.

	Direction	FDR
DNA/TcMar-Pogo_As	Down	0.018
DNA_As	Down	0.018
DNA/TcMar-Mariner_As	Down	0.018
introns_S	Down	0.019
DNA/TcMar-Tc1?_As	Down	0.028
DNA/MULE-MuDR_As	Down	0.028
DNA/hAT_As	Down	0.028
LTR/Pao_As	Down	0.030
piRNA_S	Down	0.030
DNA/TcMar-Tc1_As	Down	0.032
DNA/CMC-Chapaev_As	Down	0.047
DNA/PiggyBac_As	Down	0.057
DNA/TcMar-Tc2_As	Down	0.067
DNA/PiggyBac?_As	Down	0.074
DNA/TcMar-Tc4_As	Down	0.074

As reported previously (Bagijn et al., 2012), we found DNA/*TcMar-Mariner* superfamily, including DNA/*TcMar-Tc1* and DNA/*TcMar-Tc2* transposons as producers of TE-derived sRNAs triggered by piRNAs. As a control we expected sense piRNAs in wild type (Down direction), because they depend on PRG-1 (see **Table 11**). Gene set enrichment analysis also reveals exon siRNAs triggered by piRNAs (see **Table 10**). It has been described in *C. elegans* that piRNAs not only trigger TE-derived sRNA production. In fact, CLASH (Cross-linking, Ligation and Sequencing of Hybrids) experiments suggest that exons are also targets of PRG-1 and piRNAs (Shen et al., 2019). As expected, our enrichment analysis also reveals reads mapping to the sense strand of piRNAs associated to the wild type condition. This makes sense, because piRNAs will not be stable in absence of PRG-1. This observation also made us ask if these piRNAs are the triggers of TE-derived sRNA production.

piRNA Target prediction with pirScan

We decided to predict piRNA target sites to answer if piRNAs are the triggers of TE-derived sRNA production. We used pirScan1.0 (Zhang et al., 2018) to predict the targets of differential expressed piRNAs in wild type nematodes after our differential expression analysis. Based on piRNA targeting rules, pirScan provides a targeting score (see Methods). Using the piRNA score and the expression pattern of differential expressed piRNAs we asked if piRNAs are triggering the TE-derived sRNA biogenesis. As output pirScan produces two different prediction sets based on mismatches in the seed region (stringent and relaxed predictions). If there are a maximum of 2 mismatches GU in the seed region the predictions are classified as stringent predictions, while if there are a maximum of 2 GU mismatches as well as one non-GU mismatch in the seed they are classified as relaxed (Zhang et al. 2018). In addition, we decided to remove from relaxed predictions, those that was predicted as stringent too.

We performed a pirScan piRNA target prediction using 153,631 TEs in *C. elegans* genome, pirScan predicted 851,481 relaxed and 18,759 stringent piRNA target positions. Our predictions showed that more than one piRNA can bind to the same target. We found however that classes of TEs with more nucleotides by chance can have more piRNA target sites (**Figure 17**). In relation with pirScan score, we decided to normalize it, adding each score of each TE and then we divide the result by the number of nucleotides of TE. Using the normalized scores of differential expressed piRNAs we asked if better normalized scores are related with an increase in TE-derived sRNA production. Also, we used as a negative control non-differential expressed piRNAs. We failed to find a correlation between normalized score and the TE-derived sRNA production (**Figure 18**).

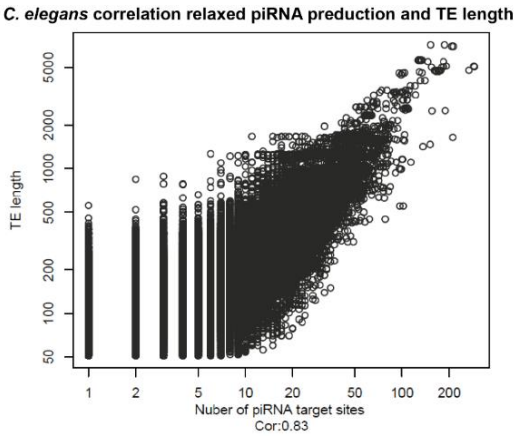

Figure 17 Linear correlation of number of *C. elegans* relaxed piRNA prediction targets with pirScan1.0 and the number of nucleotides of the TE target. Cor represents the Pearson correlation between TE length and number of piRNA targets.

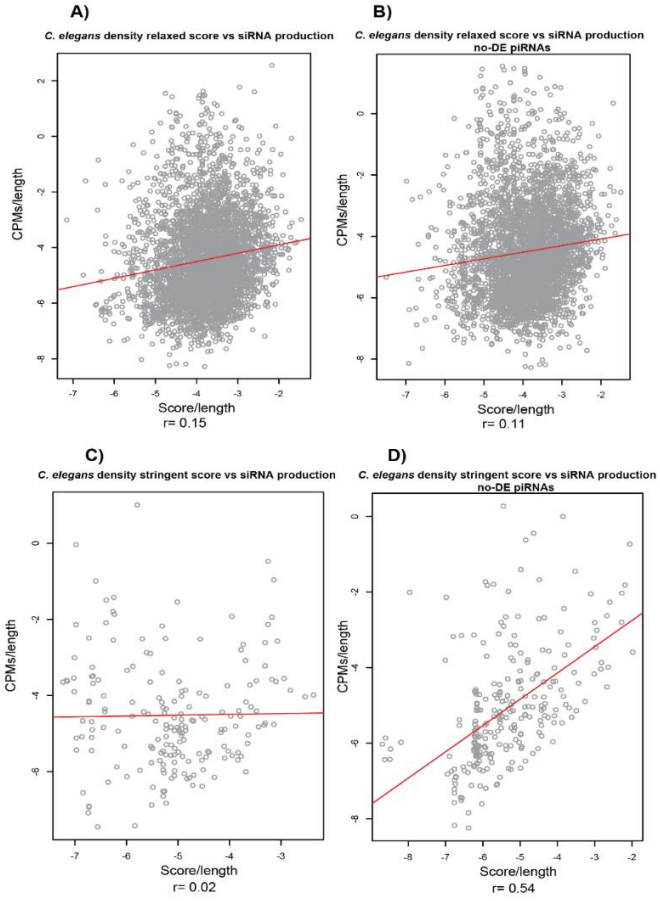

Figure 18. Lineal correlations comparing pirScan score against the normalized sRNA production. A and B shows the comparison using relaxed score of pirScan, between differential expressed piRNAs in *prg-1* mutant vs wild type analysis and non-differential expressed. C and D represents the same correlation using stringent pirScan scores. A and C were calculated using the predictions of piRNAs with $Log_2FC < 0$ and FDR < .2 as cutoff values. C and D were calculated with piRNA predictions with $Log_2FC > 0$ and FDR < .2.

At this point we didn't know if we failed to find the true piRNAs which trigger TE-derived sRNA production in our differential expression analysis or if we failed in piRNA target prediction. We thought that sensor mutants maybe added extra noise in our comparisons, so we decided to improve our analysis using a new data set of *prg-1* mutants. We used a set of 3 *C. elegans prg-1* mutant libraries without an extra genetic background (Seroussi & Claycomb *unpublished*). We performed differential expression and gene set enrichment analyses on these new data, comparing *prg-1* mutants against wildtype nematodes. We found 19,320 differential producers of sRNAs in wildtype nematodes with respect to *prg-1* mutants (**Figure 19**). This represents a substantial increase over the 2911 regions found using the previous dataset (**Figure 16**).

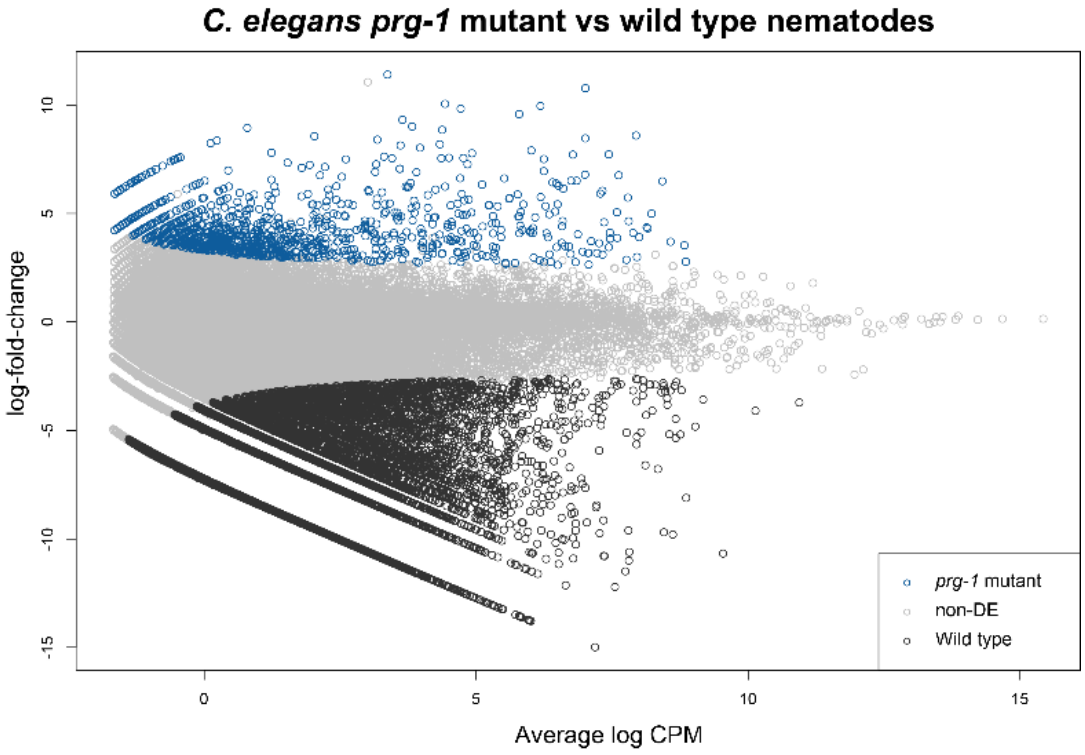

Figure 19. MA plot of differential expression analysis comparing *C. elegans prg-1* mutants against wild type nematodes in the unpublished dataset (Seroussi & Claycomb unpublished).

We found antisense DNA/*TcMar-Tc1*, DNA/*TcMar-Tc4* and DNA/*MULE-MuDR* sRNAs enriched in wild type nematodes (**Table 13 & Table 14**). In contrast we detected sense RC/*Helitrons*. As described in our *H. bakeri* analysis, this result is expected if we consider the misannotation of RC/*Helitrons* in RepeatMasker predictions. We also found some piRNAs enriched in wild type nematodes,

however it seems that not all piRNAs depend on the PRG-1 argonaute (**Figure 20 A**). It is important to mention that in *C. elegans* not all piRNAs depend on PRG-1, however it could be an effect of the genetic background of the sensor libraries.

Table 13. Gene set enrichment analysis at TE class level, in *C. elegans prg-1* mutant vs adult nematodes comparison. Down direction represents the enrichment of a TE classes in wild type nematodes.

	Direction	FDR
Unks	Down	0.0002
DNA	Down	0.0002
LINE	Down	0.0002
LTR	Down	0.0003
SINE	Down	0.1020

Table 14. Gene set enrichment analysis including TE family classification in *C. elegans prg-1* mutant vs adult nematodes comparison. Down direction represents the enrichment of a TE classes in wild type nematodes.

	Direction	FDR
piRNA_S	Down	0.0001
introns_As	Down	0.0001
introns_S	Down	0.0001
Unknown_S	Down	0.0001
DNA_As	Down	0.0001
DNA/TcMar-Tc1_As	Down	0.0001
DNA/TcMar-Tc4_As	Down	0.0001
DNA_S	Down	0.0002
Unknown_As	Down	0.0002
exons_As	Down	0.0010
LINE/RTE-RTE_As	Down	0.0010
RC/Helitron_S	Down	0.0013

Using this dataset, we also detected type II piRNAs enriched in adults, even though not all type II piRNAs seem to depend on PRG-1 (**Figure 20 B**). One

explanation of these results is that other Argonaute proteins such as CSR-1 and C04F12 can bind piRNAs (Seroussi & Claycomb *unpublished*).

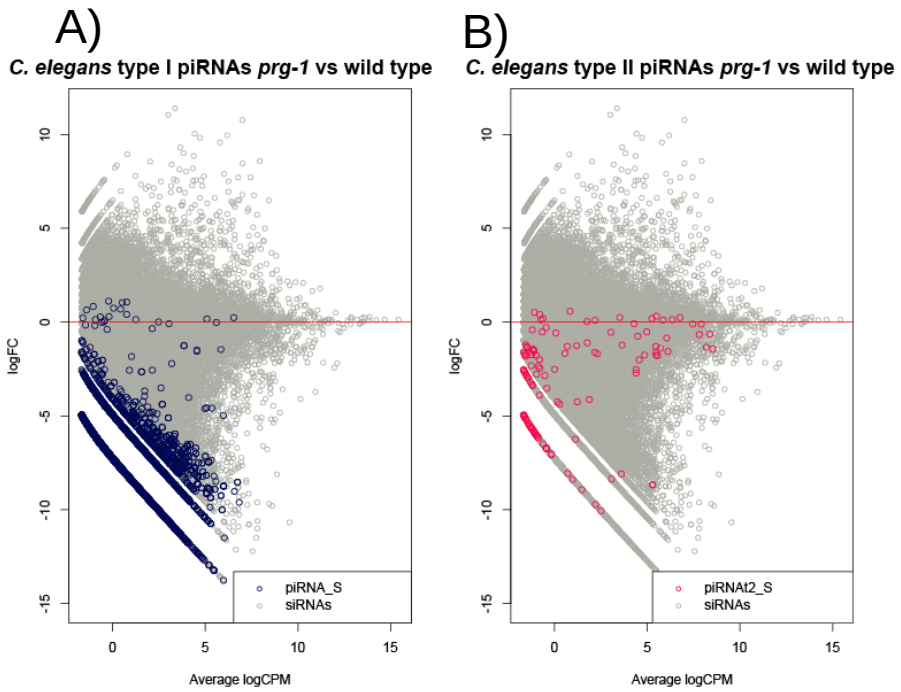

Figure 20. Expression patterns of piRNAs in *C. elegans prg-1* mutant vs adult nematodes comparison. A represents type I piRNAs expression pattern while B show type II piRNAs expression.

In our previous analysis we used the expression pattern of differential expressed piRNAs as a criterion of selection of piRNAs to perform correlations. Because our previous analysis failed, we used the expression pattern of piRNAs in immunoprecipitation (IP) experiment of PRG-1. We compared IP sRNA against the control sRNAs to get the sequences that are preferentially loaded onto PRG-1 (**Figure 21**). We detected 14,616 sRNA regions that are enriched in the PRG-1 IP. When we then compare these to the differentially expressed piRNAs in *prg-1* mutants against wild type nematodes, we get an intersection of 5,275 piRNAs. These represent high-confidence functional piRNAs, confirmed by two independent experiments measuring different characteristics of piRNAs: binding to PRG-1 and expression loss in a *prg-1* mutant background.

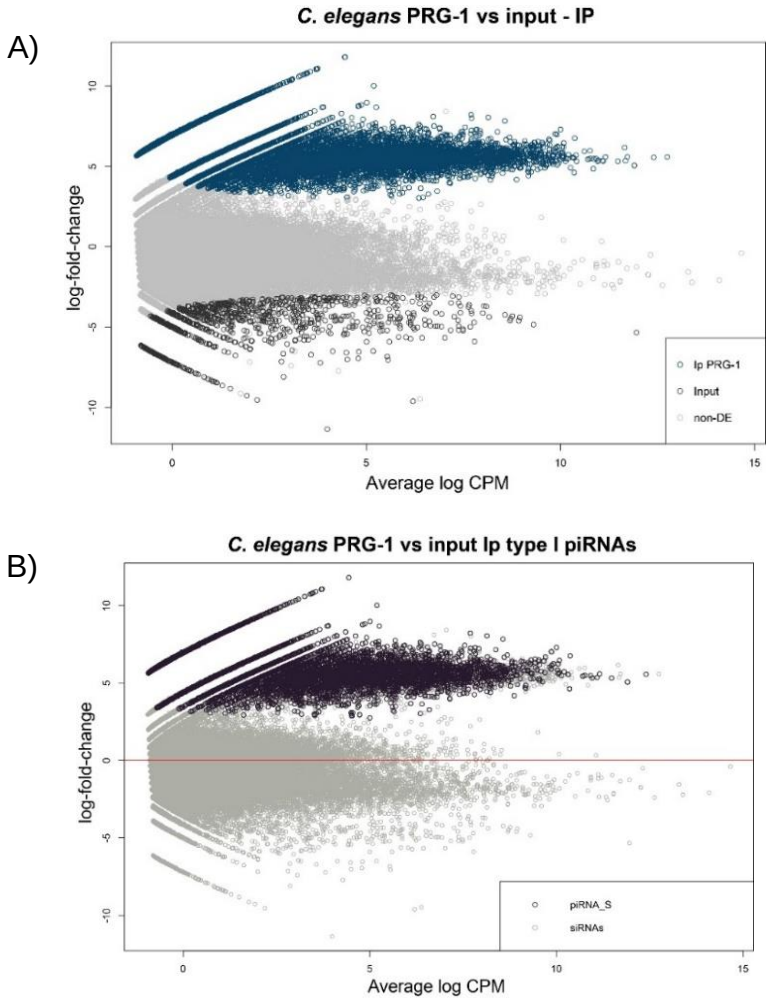

Figure 21. Differential expression analysis comparing *C. elegans* PRG-1 Ip. A) MA plot comparing sRNAs loads onto PRG-1 against non-immunoprecipitated sRNAs. B) Expression pattern of piRNAs in PRG-1 IP experiments.

In our previous attempt to correlate piRNA targeting and TE-derived sRNA production we used a normalized score as a metric of targeting. However as reported by Ashe et al. (2012) and Zhang et al. (2018), the biogenesis of sRNAs triggered by piRNAs starts within 100 nt in respect to the piRNA binding position. Using the addition of all pirScan scores we omitted positional information about how piRNAs trigger sRNA production. In order to include this information in our analysis, we decided to count the number of sRNAs that are produced in a window of 100 nt with respect to the center of the binding site of each piRNA in the PRG-1 IP experiment.

Before starting this new analysis, we decided to compare piRNA target prediction between pirScan1.0 and the last web tool version of pirScan (Wu et al., 2018). The comparison reveals that pirScan1.0 are producing the same score as the web tool version, however pirScan1.0 also produces many additional false positive results. This is related with a misidentification of GU wobbles in alignments, and also an improvement in piRNA targeting rules in the pirScan web tool version as described by W. S. Wu et al. (2018).

pirFinder development and reimplementation of piRNA target prediction

When we started using pirScan1.0, we expected to perform piRNA target prediction in both nematodes. However, the comparison with the web tool version of pirScan showed false positive results. We could instead use the web tool version of pirScan, however this tool has some limitations for our purposes. First, we couldn't predict piRNA target sites in *H. bakeri*, due to the absence of piRNAs for this species in the database of the web tool. Second, the web tool only allows searching for piRNA target sites in one sequence at a time, impeding any large scale piRNA target prediction attempt.

The limitations of pirScan encouraged us to develop a new tool to perform piRNA target predictions. As described in Methods, using reported piRNA targeting rules we developed pirFinder. Comparing pirFinder predictions against the web tool version of pirScan, we obtained the same score and predictions. It's important to highlight that both programs use different piRNA datasets, while pirScan uses WormBase and custom type II piRNA annotation (17,849 piRNAs), pirFinder uses just the type I piRNAs (10,096), described by Beltran and co-workers (Beltran et al., 2019).

Using pirFinder we predicted 593,486 piRNA binding sites in the *C. elegans* TEs and exons, however when we ask for sRNA production in a window of 100 nt around the binding site we found 380,241 unique regions that are producers of sRNAs. It is also important to highlight that just 9,999 of 10,096 piRNAs in our data set have a predicted TE or exon target. For piRNA targeting validation

however, we decided to use just differential expressed piRNAs in IP experiment (**Figure 21 B**).

RdRps produce sRNA in nearby regions to piRNA target sites in TEs

In order to validate piRNA target predictions we decided to compare the sRNA production of each 100 nt window centered on all possible piRNA targets, between *prg-1* mutants and wild type nematodes. With this approach we can distinguish in which window piRNAs truly trigger sRNA production. To avoid overlaps where different piRNAs were predicted to target the same region, we selected the most expressed piRNA according to the IP experiment. Differential expression analysis of sRNAs using the window approach reveals 5,856 windows in which piRNAs trigger significant sRNA production with respect to the *prg-1* mutant condition (**Figure 21**).

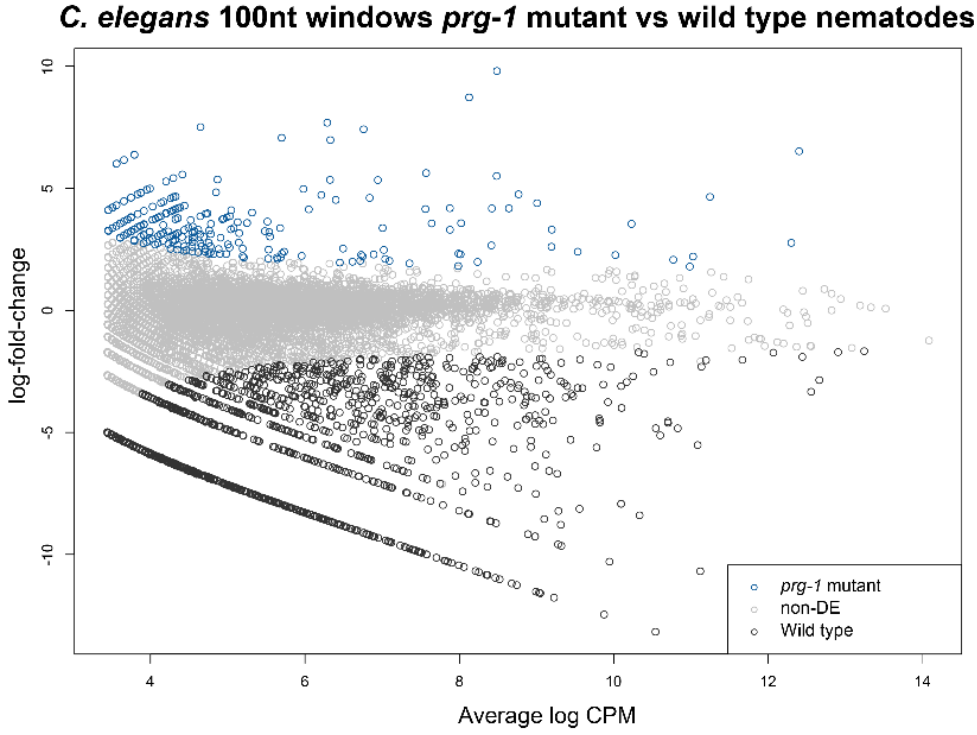

Figure 22. MA plot of differential expression analysis in 100 nt windows with respect to pirFinder target position, comparing sRNA production in *C. elegans prg-1* mutants against wild type nematodes comparison.

In this same experiment, gene set enrichment analysis showed that DNA transposons are the main target of piRNAs in comparison with other TE families (**Table 15**). As reported previous we found that piRNAs trigger sRNA production in DNA/*TcMar-Mariner* superfamily members, including DNA/*TcMar-Tc1*, DNA/*TcMar-Tc4* and DNA/*TcMar-Tc2*. In addition, we also found that DNA/*hAT* and RC/*Helitron* are regulated by piRNAs. Interestingly, although the expression and piRNA regulation of retroelements in *C. elegans* is controversial, we found that piRNAs trigger sRNA production in LTR/*Pao*, LTR/*Gypsy* and LINE/*CR1*, LINE/*R2*, LINE/*RTE-RTE*. We also found that some SINE families such as SINE/*tRNA-RTE* are regulated by piRNAs, but this may depend on the transcription activity of related genes.

Table 15. Gene set enrichment including TE families related to 100 nt windows with respect to pirFinder target position. Down direction means that there is an enrichment of sRNA production in wild type condition with respect to *prg-1* mutants.

	Direction	FDR
DNA_PiggyBac_As	Down	0.00013
DNA_TcMar-Tc4_As	Down	0.00013
DNA_hAT_As	Down	0.00013
DNA_TcMar_Tc1_As	Down	0.00013
LINE/RTE-RTE_As	Down	0.00013
exons_As	Down	0.00013
Unknown_As	Down	0.00013
DNA_MULE-MuDR_As	Down	0.00013
DNA_TcMar-Pogo_As	Down	0.00013
Unknown_S	Down	0.00013
LTR_Gypsy_As	Down	0.00013
RC_Helitron_S	Down	0.00018
SINE_tRNA-RTE_As	Down	0.00105
DNA_TcMar-Tc2_As	Down	0.00112
LTR_Pao_As	Down	0.00166

With these results we were curious if the score of the interaction of piRNAs can predict the production of sRNAs. To board this, we subclassified our piRNA target

predictions based on pirFinder score. We used predictions with score greater or equal to 1 and predictions with score equal to 0 to get new 100nt windows. Using these groups of windows and the counts of sRNAs in *prg-1* mutants and wild type nematodes, we performed individual gene set enrichment analysis of TE families related with these windows. We used the FDR of the enrichment test as an indirect metric to answer if the kind of interaction of piRNAs influence the production of sRNAs and therefore the enrichment of TE families (**Figure 23**). Using the Wilcoxon Rank Sum test, we found that the distribution of FDR values in both analyses are statistically different (see **Figure 23**), also we found lower FDR values in the group with score greater or equal to 1. This result reveals that piRNA targets predicted with a better score do tend to be more associated with TE silencing via sRNA production.

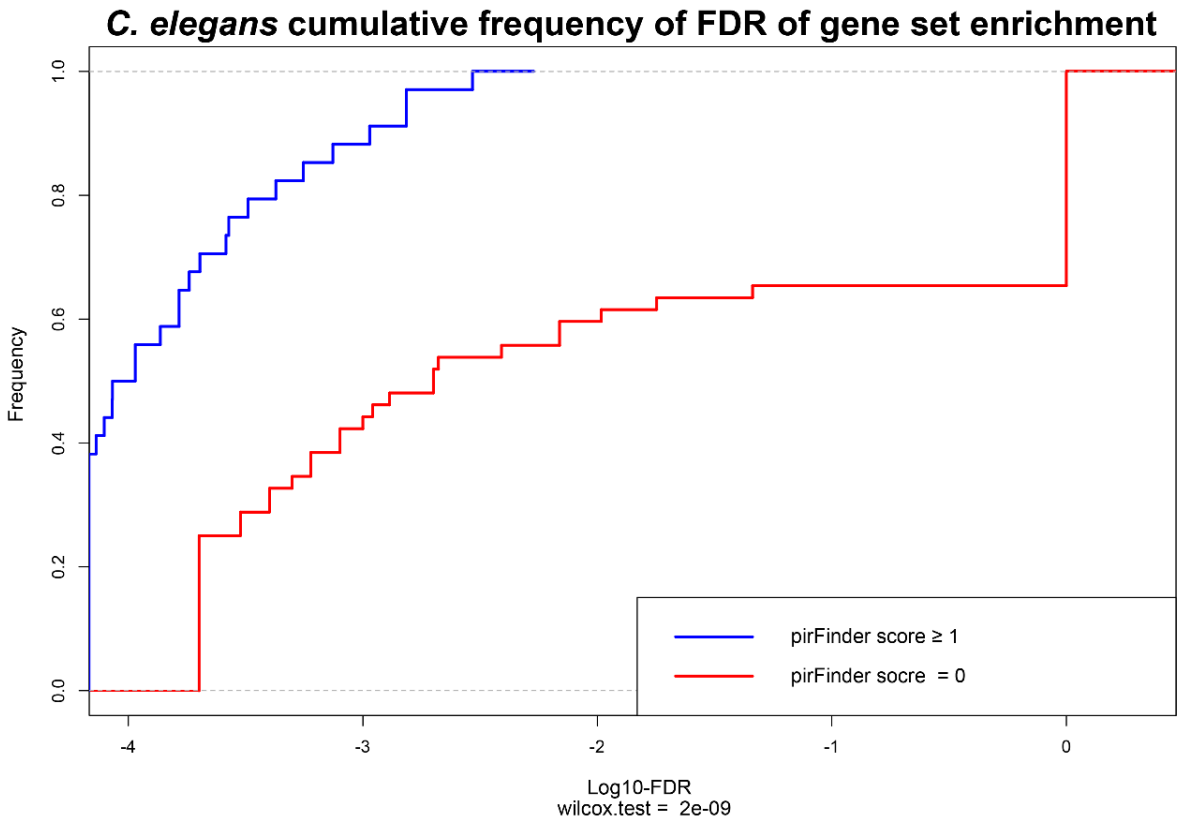

Figure 23. Cumulative distribution of FDR of gene set enrichment test. Y axis represent the Log_{10} of FDR values. The red line represents the cumulative frequency of FDR values after enrichment test of TE families using windows with a pirFinder score equal to 0. The blue line represents the cumulative frequency of FDR after enrichment test using windows with pirFinder score greater or equal to 1. The two distributions are significantly different, according to a Wilcox Rank Sum test ($p = 2x10^{-09}$).

The window approach allowed us to identify in *C. elegans* TE families in which sRNAs production depends on piRNAs. Together our results reveal that not just DNA transposons and exons produce sRNAs triggered by piRNAs. Interestingly, we found that in *C. elegans* piRNAs silence some of the same TE families which are producing sRNAs inside *H. bakeri* EVs (DNA/*hAT* superfamily, LINE/*RTE-RTE* and RC/*Helitron*). This result suggests that the production of secreted TE-derived sRNAs in *H. bakeri* could also depend on piRNAs. Finally we found that the score of piRNA interaction can partially predict the production of sRNAs, as reported previously (Zhang et al., 2018).

Discussion

TE-derived sRNAs are usually important for the regulation of endogenous TE-transcripts. There are, however, many challenges to understand their function in parasitism. Here, we ask which TE classes or families are responsible for the production of secreted TE derived sRNAs in the intestinal parasite *H. bakeri*. Using *C. elegans* as a model we also examine how are TE-derived sRNAs produced.

Although *C. elegans* and *H. bakeri* belongs to the same clade within Blaxter classification (Blaxter et al., 1998) they contain outstanding differences in genomic composition. One of the most important differences between these nematodes is the genomic proportions and diversity of TEs. Despite just 15% of the *C. elegans* genome containing TEs, it has a higher number of TE families in comparison to *H. bakeri*. The reason of the high diversity but low TE copies may be related with the success in TE silencing in *C. elegans*, also it is important to bear in mind that other important difference is that *C. elegans* genome is much more annotated than *H. bakeri*. There is evidence that suggests that the clustering of type I piRNAs in chromosome 4 and the epigenetic environment is essential to control TE transcription and also transposition in *C. elegans* (Beltran et al., 2019; Shen et al., 2019; Kofler, 2019). In *H. bakeri* piRNAs do not appear

to be clustered (Beltran et al., 2019), however, it remains to be explored if this is related with the burst of some TE families and the increase in genome size.

With respect to secreted TE-derived sRNAs, we found selectivity in EVs packaging. We showed that LINE retroelements are enriched among the sRNAs associated to EVs. It is important to highlight that LINE-derived sRNAs are not the most abundant sRNAs within EVs, however they are more specific to EVs with respect to the sRNA population in the adult nematode. In addition, we clarify that not all LINE families are enriched in EVs. Interestingly, just members of the LINE/*RTE-RTE* superfamily including LINE/*RTE-BovB* are enriched in EVs. Also, despite DNA transposons not being enriched as a class, we found DNA/*hAT-Tip100* and RC/*Helitrons* as enriched families in secreted TE-derived sRNA production. It is important to mention that irrespective of the enrichment of classes or families of TEs, the most abundant sRNAs within EVs are produced from novel repeats, with one cluster being particularly important as noted previously (Chow et al., 2019). It remains however as an open question: which TE class or family do these novel repeats belong to?

We showed that at least 27% of the LINE/*RTE-BovB* responsible for the sRNAs enriched in EVs, are full length TEs. It is interestingly that other parasitic nematodes that are phylogenetically closely related to *H. bakeri,* such as *Haemonchus contortus* and *Onchocerca ochengi* have also have full-length active LINE/*RTE-BovB* (Dunemann & Wasmuth, 2018). In addition, other closely related parasites such as *Necator americanus* and *Nippostrongylus brasiliensis* contain full length active LINE/*RTE-RTE* elements (Dunemann & Wasmuth, 2018). It has been shown that at least *H. contortus, N. americanus* and *N. basiliensis* have a ortholog to exWAGO (Chow et al., 2019). Further work is required to understand if LINE elements specifically from *RTE-RTE* superfamily are used as virulence factors by this group of parasitic nematodes. It is also an exciting and unexplored matter how evolutionary forces act over retroelements in the genome of these parasites.

One of the most interesting observations with respect to LINE/*RTE-BovB* is that these elements can be horizontally transferred between parasitic nematodes and their hosts (Dunemann & Wasmuth, 2018; Ivancevic et al., 2018). It is also fascinating that elements such as RC/*Helitrons* and DNA/*hAT* transposons can

be horizontal transferred (Ivancevic et al., 2018). Due to the apparent relevance of these TE families in the *H. bakeri* lifestyle, according to the results presented in this thesis, the evolutionary history of these elements will be an important matter to approach in the future.

Regarding *C. elegans,* we found that not just DNA transposons, specifically members of the *Tc-Mar* superfamily, are active transposons (Laricchia, et al., 2017). In addition to Ansaloni and co-workers, we found TE-derived sRNAs related to LINE elements, which is an important evidence of the activity of LINEs. It has been reported that there is a tissue and developmental stage dependent expression of TEs in *C. elegans* (Ansaloni et al., 2019). Interestingly LINE elements are expressed in E cell precursors and E cells, which give rise to intestinal tissue (Ansaloni et al., 2019). This is highly relevant because it has been proposed that the production of secreted EVs in *H. bakeri* takes place in the intestine (Buck et al., 2014; Chow et al., 2019). There is also evidence that reveals that SAGO-1, SAGO-2 and PPW-1 (orthologous argonautes in *C. elegnas* with respect to exWAGO) are expressed in the intestine (Seroussi & Claycomb, *unpublished*). Together, these observations open the question if the LINE-derived sRNAs secreted by *H. bakeri* are produced only in the adult stage or if these sRNAs are produced in previous developmental stages. Further analysis such as single cell RNA-seq are required to understand if, as in *C. elegans,* the expression of LINE elements is specific to certain tissues and developmental stages.

In this work we also developed a piRNA target prediction tool. It is important to mention that using pirFinder we can use different piRNA data sets, allowing to infer piRNA targets in other nematodes not just in *C. elegans*. Although it has been reported that only nematodes in clade V have piRNAs (Beltran et al., 2019; Sarkies et al., 2015), it is an interesting question to answer which parasites within clade V use piRNAs to produce TE-derived sRNAs, and also if these parasites use TE-derived sRNAs as virulence factors.

Using *prg-1* mutants we found evidence that piRNAs trigger the biogenesis of TE-derived sRNAs in *C. elegans*. As previous reported, our results showed that the piRNA pathway is not just related to TE regulation, but is also involved in the regulation of endogenous genes (Shen, et al., 2018). It has been described,

however, that piRNAs are not the only triggers of exon-derived 22G sRNA production. Pathways such as CSR-1 and 26G can also trigger the production of these sRNAs (Claycomb et al., 2009). With respect to TE-derived sRNAs we found that piRNAs trigger LINE/*RTE-RTE*, DNA/*hAT* and RC/*Helitron* derived sRNAs. These results suggest that the sRNA production of these families in *H. bakeri* can also be triggered by piRNAs. We can't however prove this because we lack experimental evidence such as a mutant strain of *H. bakeri* for PRG-1.

The results of this work open interesting questions about the evolution and regulation of TEs, specifically in parasitic nematodes. Although there are many challenges related with this matter, here we established a general strategy to gain a deeper understanding of the specificity and biogenesis of TE-derived sRNAs.

Perspectives

- Improve the *H. bakeri* genome annotation, specifically with respect to TEs.
- Look for genes, within full length LINE/*RTE-BovB* which produce secreted TE-derived sRNAs.
- Determine if LINE/*RTE-BovB*, RC/*Helitrons*, DNA/*hAT-Tip100* and LINE/*RTE-RTE* are transcribed in adult nematodes, and if they are, determine the direction of their transcription.
- Trace the evolutionary history of LINE/*RTE-BovB* in parasitic nematodes, specifically within the strongylomorpha group
- Determine if as in *H. bakeri*, LINE TEs are the source of TE derived sRNAs in other parasitic nematodes.
- Analyze how evolutionary forces act on TEs which are the source of secreted sRNAs by *H. bakeri.*
- Use IP sRNA-seq of PRG-1 in *H. bakeri* to determine the expression pattern of piRNAs, in order to use the expression of piRNAs and TE-derived sRNA levels to generate a machine learning algorithm to reliably predict piRNA targets, and infer if piRNAs trigger the production of TE-derived sRNAs in this nematode.

References

Ansaloni, F., Scarpato, M., Di Schiavi, E., Gustincich, S., & Sanges, R. (2019). Exploratory analysis of transposable elements expression in the C. elegans early embryo. *BMC Bioinformatics*, *20*(S9), 484. https://doi.org/10.1186/s12859-019-3088-7

Ashe, A., Sapetschnig, A., Weick, E. M., Mitchell, J., Bagijn, M. P., Cording, A. C., … Miska, E. A. (2012). PiRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. *Cell*, *150*(1), 88–99. https://doi.org/10.1016/j.cell.2012.06.018

Ayarpadikannan, S., & Kim, H.-S. (2014). The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. *Genomics & Informatics*, *12*(3), 98. https://doi.org/10.5808/gi.2014.12.3.98

Bagijn, M. P., Goldstein, L. D., Sapetschnig, A., Weick, E., Bouasker, S., Lehrbach, N. J., … Miska, E. a. (2012). Function, Targets, and Evolution of Caenorhabditis elegans piRNAs. *Science*, *337*(August), 574–577. https://doi.org/10.1126/science.1220952

Beltran, T., C., B., Birkle, T. Y., Stevens, L., Schwartz, H. T., Sternberg, P. W., … Sarkies, P. (2018). Evolutionary analysis implicates RNA polymerase II pausing and chromatin structure in nematode piRNA biogenesis. *BioRxiv*.

Beltran, Toni, Barroso, C., Birkle, T. Y., Stevens, L., Schwartz, H. T., Sternberg, P. W., … Sarkies, P. (2019). Comparative Epigenomics Reveals that RNA Polymerase II Pausing and Chromatin Domain Organization Control Nematode piRNA Biogenesis. *Developmental Cell*, *48*(6), 793-810.e6. https://doi.org/10.1016/j.devcel.2018.12.026

Billi, A. C., Fischer, S. E. J., & Kim, J. K. (2014). Endogenous RNAi pathways in C. elegans. *WormBook : The Online Review of C. Elegans Biology*, 1–49. https://doi.org/10.1895/wormbook.1.170.1

Blaxter, M. L., De Ley, P., Garey, J. R., Llu, L. X., Scheldeman, P., Vierstraete, A., … Thomas, W. K. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature*, *392*(6671), 71–75. https://doi.org/10.1038/32160

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., … Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Buck, A. H., & Blaxter, M. (2013). Functional diversification of Argonautes in nematodes: an expanding universe. *Biochemical Society Transactions*, *41*(4), 881–886. https://doi.org/10.1042/BST20130086

Buck, A. H., Coakley, G., Simbari, F., McSorley, H. J., Quintana, J. F., Le Bihan, T., … Maizels, R. M. (2014). Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity.

Nature Communications, *5*, 1–11. https://doi.org/10.1038/ncomms6488

Cho, J. (2018). Transposon-derived non-coding RNAs and their function in plants. *Frontiers in Plant Science*, *9*(May), 1–6. https://doi.org/10.3389/fpls.2018.00600

Chow, F. W. N., Koutsovoulos, G., Ovando-Vázquez, C., Neophytou, K., Bermúdez-Barrientos, J. R., Laetsch, D. R., … Buck, A. H. (2019). Secretion of an Argonaute protein by a parasitic nematode and the evolution of its siRNA guides. *Nucleic Acids Research*, *47*(7), 3594–3606. https://doi.org/10.1093/nar/gkz142

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, *18*(2), 71–86. https://doi.org/10.1038/nrg.2016.139

Claycomb, J. M. (2014). Ancient endo-siRNA pathways reveal new tricks. *Current Biology*, *24*(15), R703–R715. https://doi.org/10.1016/j.cub.2014.06.009

Claycomb, J. M., Batista, P. J., Pang, K. M., Gu, W., Vasale, J. J., van Wolfswinkel, J. C., … Mello, C. C. (2009). The Argonaute CSR-1 and Its 22G-RNA Cofactors Are Required for Holocentric Chromosome Segregation. *Cell*, *139*(1), 123–134. https://doi.org/10.1016/j.cell.2009.09.014

Cooper, D., & Eleftherianos, I. (2016). Parasitic Nematode Immunomodulatory Strategies : Recent Advances and Perspectives. https://doi.org/10.3390/pathogens5030058

Czech, B., Munafò, M., Ciabrelli, F., Eastwood, E. L., Fabry, M. H., Kneuss, E., & Hannon, G. J. (2018). piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annual Review of Genetics*, *52*(1), 131–157. https://doi.org/10.1146/annurev-genet-120417-031441

De La Chaux, N., & Wagner, A. (2011). BEL/Pao retrotransposons in metazoan genomes. *BMC Evolutionary Biology*, *11*(1). https://doi.org/10.1186/1471-2148-11-154

Dunemann, S. M., & Wasmuth, J. (2018). A nematode retrotransposon in the common shrew: horizontal transfer between parasite and host. *BioRxiv*, 424424. https://doi.org/10.1101/424424

En-Zhi Shen, Hao Chen2, Ahmet R. Ozturk, Shikui Tu, Masaki Shirayama, W., & Tang, Yue-He Ding, Si-Yuan Dai, Zhiping Weng, and C. C. M. (2019). Identification of piRNA binding sites reveals the Argonaute regulatory landscape of the C. elegans germline. *Physiology & Behavior*, *176*(3), 139–148. https://doi.org/10.1016/j.physbeh.2017.03.040

Finnegan, D. J. (2012). Retrotransposons. *Current Biology*, *22*(11), 432–437. https://doi.org/10.1016/j.cub.2012.04.025

Gaiti, F., Calcino, A. D., Tanurdžić, M., & Degnan, B. M. (2017). *Origin and evolution of the metazoan non-coding regulatory genome*. *Developmental Biology* (Vol. 427). Elsevier. https://doi.org/10.1016/j.ydbio.2016.11.013

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. *Development (Cambridge)*, *143*(22), 4101–4114. https://doi.org/10.1242/dev.132639

Gu, W., Shirayama, M., Conte, D., Vasale, J., Batista, P. J., Claycomb, J. M., … Mello, C. C. (2009). Distinct Argonaute-Mediated 22G-RNA Pathways Direct Genome Surveillance in the C. elegans Germline. *Molecular Cell*, *36*(2), 231–244. https://doi.org/10.1016/j.molcel.2009.09.020

Höck, J., & Meister, G. (2008). The Argonaute protein family. *Genome Biology*, *9*(2). https://doi.org/10.1186/gb-2008-9-2-210

Hoogstrate, S. W., Volkers, R. J., Sterken, M. G., Kammenga, J. E., & Snoek, L. B. (2014). Nematode endogenous small RNA pathways. *Worm*, *3*(1), e28234. https://doi.org/10.4161/worm.28234

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., & Adelson, D. L. (2018). Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biology*, *19*(1), 1–13. https://doi.org/10.1186/s13059-018-1456-7

Kaschula, R., Pinho, S., & Alonso, C. R. (2018). Microrna-dependent regulation of hox gene expression sculpts fine-grain morphological patterns in a drosophila appendage. *Development (Cambridge)*, *145*(20). https://doi.org/10.1242/dev.161133

Kazazian, H. H. (2004). Mobile Elements: Drivers of Genome Evolution. *Science*, *303*(5664), 1626–1632. https://doi.org/10.1126/science.1089670

Kidwell, M. G., & Lisch, D. R. (2000). Transposable elements and host genome evolution. *Trends in Ecology and Evolution*, *15*(3), 95–99. https://doi.org/10.1016/S0169-5347(99)01817-0

Kim, K. W., Tang, N. H., Andrusiak, M. G., Wu, Z., Chisholm, A. D., & Jin, Y. (2018). A Neuronal piRNA Pathway Inhibits Axon Regeneration in C. elegans. *Neuron*, *97*(3), 511-519.e6. https://doi.org/10.1016/j.neuron.2018.01.014

Kofler, R. (2019). Dynamics of transposable element invasions with piRNA clusters. *Molecular Biology and Evolution*, *36*(7), 1457–1472. https://doi.org/10.1093/molbev/msz079

Laricchia, K. M., Zdraljevic, S., Cook, D. E., & Andersen, E. C. (2017). Natural Variation in the Distribution and Abundance of Transposable Elements Across the Caenorhabditis elegans Species. *Molecular Biology and Evolution*, *34*(9), 2187–2202. https://doi.org/10.1093/molbev/msx155

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., … Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), 1–10. https://doi.org/10.1371/journal.pcbi.1003118

Lee, R., Feinbaum, R., & Ambros, V. (1993). The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to lin-14. *Cell*, *116*(116), 843–854.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with

Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Lin, H., & Spradling, A. C. (1997). A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. *Development*, *124*(12), 2463–2476.

Macia, A., Blanco-Jimenez, E., & García-Pérez, J. L. (2015). Retrotransposons in pluripotent cells: Impact and new roles in cellular plasticity. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1849*(4), 417–426. https://doi.org/10.1016/j.bbagrm.2014.07.007

Maizels, R. M., & McSorley, H. J. (2016). Regulation of the host immune system by helminth parasites. *Journal of Allergy and Clinical Immunology*, *138*(3), 666–675. https://doi.org/10.1016/j.jaci.2016.07.007

Mccarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, *25*(6), 765–771. https://doi.org/10.1093/bioinformatics/btp053

McClintock, B. (1950). THE ORIGIN AND BEHAVIOR OF MUTABLE LOCI IN MAIZE. *Genetics*, 344–355.

Mukherjee, D., Das, S., Begum, F., Mal, S., & Ray, U. (2019). The mosquito immune system and the life of dengue virus: What we know and do not know. *Pathogens*, *8*(2). https://doi.org/10.3390/pathogens8020077

Ni, J. Z., Kalinava, N., Chen, E., Huang, A., Trinh, T., & Gu, S. G. (2016). A transgenerational role of the germline nuclear RNAi pathway in repressing heat stress-induced transcriptional activation in C. elegans. *Epigenetics and Chromatin*, *9*(1), 1–15. https://doi.org/10.1186/s13072-016-0052-x

O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, *9*(AUG), 1–12. https://doi.org/10.3389/fendo.2018.00402

Pagès H, Aboyoun P, Gentleman R, D. S. (2019). Biostrings: Efficient manipulation of biological strings. *R Package Version 2.54.0.*, (1), 1–6.

Pierre, L. (2015). JVarkit: java-based utilities for Bioinformatics. *Figshare*, 2–5. https://doi.org/10.6084/m9.figshare.1425030.v1

Pinzón, N., Bertrand, S., Subirana, L., Busseau, I., Escrivá, H., & Seitz, H. (2019). Functional lability of RNA-dependent RNA polymerases in animals. *PLoS Genetics*, *15*(2), 1–25. https://doi.org/10.1371/journal.pgen.1007915

Platt, R. N., Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research*, *26*(1–2), 25–43. https://doi.org/10.1007/s10577-017-9570-z

Quintana, J. F., Makepeace, B. L., Babayan, S. A., Ivens, A., Pfarr, K. M., Blaxter, M., … Buck, A. H. (2015). Extracellular Onchocerca-derived small RNAs in host nodules and blood. *Parasites and Vectors*, *8*(1), 1–11. https://doi.org/10.1186/s13071-015-0656-1

Reinhart, B., FJ, S., M, B., AE, P., JC, B., AE, R., GB, R. (2000). The 21-

nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, *403*, 901–906.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

RuvkunG, F. S. and. (2019). Caenorhabditis elegans ADAR editing and the ERI-6/7/MOV10 RNAi pathway silence endogenous viral elements and LTR retrotransposons. *Bioarxiv*, (12). https://doi.org/10.7868/s0016675817120025

Sarkies, P., Selkirk, M. E., Jones, J. T., Blok, V., Boothby, T., Goldstein, B., … Miska, E. A. (2015). Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages. *PLoS Biology*, *13*(2), 1–20. https://doi.org/10.1371/journal.pbio.1002061

Spracklin, G., Fields, B., Wan, G., Becker, D., Wallig, A., Shukla, A., & Kennedy, S. (2017). The RNAi Inheritance Machinery of Caenorhabditis elegans. *Genetics*, *206*(July), 1403–1416. https://doi.org/10.1534/genetics.116.198812/-/DC1.1

Stricklin, S. L., Griffiths-jones, S., & Eddy, S. R. (2005). WormBook: C. elegans noncoding RNA genes. *The C. Elegans Research Community, WormBook.* https://doi.org/10.1895/wormbook.1.1.1

Tan, S., Cardoso-Moreira, M., Shi, W., Zhang, D., Huang, J., Mao, Y., … Zhang, Y. E. (2016). LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Research*, *26*(12), 1663–1675. https://doi.org/10.1101/gr.204925.116

Tang, W., Tu, S., Lee, H. C., Weng, Z., & Mello, C. C. (2016). The RNase PARN-1 Trims piRNA 3′ Ends to Promote Transcriptome Surveillance in C. elegans. *Cell*, *164*(5), 974–984. https://doi.org/10.1016/j.cell.2016.02.008

Touati, R., Messaoudi, I., ElloumiOueslati, A., & Lachiri, Z. (2018). Classification of helitron's types in the c.Elegans genome based on features extracted from wavelet transform and SVM methods. *BIOINFORMATICS 2018 - 9th International Conference on Bioinformatics Models, Methods and Algorithms, Proceedings; Part of 11th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2018*, *3*(Biostec), 127–134. https://doi.org/10.5220/0006631001270134

Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., … Jin, H. (2013). Fungal Small RNAs Suppress Plant Immunity by Hijacking Host. *Science (New York, N.Y.)*, *342*(6154), 118–123. https://doi.org/10.1126/science.1239705.Fungal

Weick, E. M., Sarkies, P., Silva, N., Chen, R. A., Moss, S. M. M., Cording, A. C.,

… Miska, E. A. (2014). PRDE-1 is a nuclear factor essential for the biogenesis of Ruby motif-dependent piRNAs in C. elegans. *Genes and Development*, *28*(7), 783–796. https://doi.org/10.1101/gad.238105.114

Wilson, R. C., & Doudna, J. A. (2013). Molecular Mechanisms of RNA Interference. *Annual Review of Biophysics*, *42*(1), 217–239. https://doi.org/10.1146/annurev-biophys-083012-130404

Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M. L., Visvader, J. E., & Smyth, G. K. (2010). ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics*, *26*(17), 2176–2182. https://doi.org/10.1093/bioinformatics/btq401

Wu, W. S., Huang, W. C., Brown, J. S., Zhang, D., Song, X., Chen, H., … Lee, H. C. (2018a). PirScan: A webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. *Nucleic Acids Research*, *46*(W1), W43–W48. https://doi.org/10.1093/nar/gky277

Wu, W. S., Huang, W. C., Brown, J. S., Zhang, D., Song, X., Chen, H., … Lee, H. C. (2018b). PirScan: A webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. *Nucleic Acids Research*, *46*(W1), W43–W48. https://doi.org/10.1093/nar/gky277

Wynant, N., Santos, D., & Vanden Broeck, J. (2017). The evolution of animal Argonautes: Evidence for the absence of antiviral AGO Argonautes in vertebrates. *Scientific Reports*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-08043-5

Yoon, J. H., Abdelmohsen, K., & Gorospe, M. (2013). Posttranscriptional gene regulation by long noncoding RNA. *Journal of Molecular Biology*, *425*(19), 3723–3730. https://doi.org/10.1016/j.jmb.2012.11.024

Youngman, E. M., & Claycomb, J. M. (2014). From early lessons to new frontiers: The worm as a treasure trove of small RNA biology. *Frontiers in Genetics*, *5*(NOV), 1–13. https://doi.org/10.3389/fgene.2014.00416

Zamanian, M., Fraser, L. M., Agbedanu, P. N., Harischandra, H., Moorhead, A. R., Day, T. A., … Kimber, M. J. (2015). Release of Small RNA-containing Exosome-like Vesicles from the Human Filarial Parasite Brugia malayi. *PLoS Neglected Tropical Diseases*, *9*(9), 1–23. https://doi.org/10.1371/journal.pntd.0004069

Zeng, C., Weng, C., Wang, X., Yan, Y. H., Li, W. J., Xu, D., … Guang, S. (2019). Functional Proteomics Identifies a PICS Complex Required for piRNA Maturation and Chromosome Segregation. *Cell Reports*, *27*(12), 3561-3572.e3. https://doi.org/10.1016/j.celrep.2019.05.076

Zhang, D., Tu, S., Stubna, M., Wu, W. S., Huang, W. C., Weng, Z., & Lee, H. C. (2018). The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes. *Science*, *359*(6375), 587–592. https://doi.org/10.1126/science.aao2840