Elucidating evolutionary patterns of PIWI/piRNA-mediated transposon and gene regulation

Dissertation
zur Erlangung des Grades
„Doktor der Naturwissenschaften“

am Fachbereich Biologie
der Johannes Gutenberg-Universität Mainz

Daniel Gebert
geboren in Zelinograd am 16. April 1989

Mainz, 2018

Dekan:
1. Berichterstatter:
2. Berichterstatter:

Tag der mündlichen Prüfung:

Inhaltsverzeichnis/Contents

Zusammenfassung

Kleine, nicht-kodierende RNAs (small non-coding RNAs; sRNAs oder sncRNAs) finden sich in allen Domänen der Lebewesen wieder, wo sie eine Vielzahl an Aufgaben wie etwa in der Genregulation, der Transposonrepression und der Virusabwehr übernehmen. Dabei sind sRNAs, wie z.B. microRNAs (miRNAs), small-interfering RNAs (siRNAs) und Piwi-interacting RNAs (piRNAs), grundsätzlich auf Argonautenproteine angewiesen. Als Teil größerer Proteinkomplexe werden Argonauten von der jeweiligen gebundenen sRNA, basierend auf Sequenzkomplementarität, zu ihren Zielnukleinsäuren geführt. Neben dieser grundlegenden Gemeinsamkeit brachte die Evolution ein breites Spektrum an Adaptationen in vielen sRNA-Systemen diverser Entwicklungslinien hervor. Die weitestgehend tierspezifische Klasse der piRNAs, die mit der Argonauten-Untergruppe PIWI assoziiert, umfasst ein System, das durch eine besonders große evolutionäre Plastizität gekennzeichnet ist. piRNAs sind vorwiegend dafür zuständig, die Genomintegrität in der Keimbahn gegen Transposable Elemente (TEs) zu schützen, die in Urkeimzellen von Tieren während der epigenetischen Reprogrammierung reaktiviert werden. Folglich charakterisiert sich das PIWI/piRNA-System als Teil eines fortlaufenden Wettrüstungsprozesses gegen TEs und deren Bestreben, sich zu vermehren und über die Keimbahn des Wirtes in die nächste Generation weitergegeben zu werden.

Die vorliegende Dissertation setzt sich zum Ziel, die allgemeine sRNA-Forschung bioinformatisch zu unterstützen und bisher unbeantwortete Fragestellungen bezüglich des PIWI/piRNA-Systems mit einer komparativen Strategie zu bearbeiten. Nach einer Zusammenfassung des aktuellen Wissensstands über sRNA-basierte TE-Repression in Eukaryoten folgt die Vorstellung des neuentwickelten Programms unitas. Dieses Programm ermöglicht, ausgehend von sRNA-Sequenzierungsdaten, die Annotation einer Vielfalt an sRNAs, darunter miRNAs und deren Isoformen, piRNAs, Phased-siRNAs (phasiRNAs) und tRNA-Fragmente (tRFs) in einer großen Auswahl von Spezies. Die zugrundeliegende Studie des nächsten Kapitels beschreibt die Aktivität des PIWI/piRNA-Systems in somatischen Geweben von Mollusken. Dabei zeigen sich klare Hinweise auf post-transkriptionelle TE-Repression durch piRNAs sowie Zeichen der Anpassung an die Bekämpfung junger Transposons. Daneben beschreibt die Studie ein dynamisches Expressionsmuster von piRNA-produzierenden genomischen Loci während der Auster-Entwicklung, wobei bestimmte Gruppen von Loci als Quelle für piRNAs gegen verschiedene Transposon-Familien dienen. Diese und weitere kürzlich gemachte Entdeckungen verdeutlichen, dass die weitestgehende Keimbahn-Spezifität des PIWI/piRNA-Systems in Wirbeltieren wahrscheinlich eine in der Evolution der Tiere deutlich später erfolgte Anpassung darstellt als zunächst angenommen.

Im nächsten Kapitel wird, ausgehend von einer umfassenden Analyse des piRNA-Transkriptoms aus adulten Testes des Schweins, die post-transkriptionelle Prozessierung von Protein-kodierenden Genen innerhalb eines piRNA-spezifischen Amplifikationsmechanismus, dem sogenannten Ping-Pong-Zyklus, beschrieben. Darüberhinaus wird gezeigt, dass piRNA-Cluster, welche große genomische Bereiche darstellen, die die Mehrzahl an piRNAs in adulten Testes produzieren, Sequenzen von Genen und Pseudogenen beinhalten. Diese sind mögliche Vorlagen für gegensträngige piRNAs, die Zielsequenzen in entsprechenden Elterngenen angreifen können. Schließlich beleuchtet der letzte Teil der Dissertation die Evolution von piRNA-Clustern und den darin vorkommenden Pseudogenen in Primaten sowie die Fähigkeit dieser Loci, Protein-kodierende Gene zu regulieren. Da solche piRNA-Cluster jedoch schnell evolvieren und genregulatorische Potentiale zwischen unterschiedlichen Spezies schwach konserviert sind, scheint es entgegen der Erwartung eher zweifelhaft, dass Pseudogene in piRNA-Clustern eine kritische Funktion besitzen. Die Beobachtung, dass piRNA-Cluster tendenziell in Regionen mit erhöhter Gendichte und GC-Gehalt liegen, was auf offenes und aktives Chromatin hindeutet, scheint eher dafür zu sprechen, dass das Vorhandensein von Pseudogenen nicht mehr als ein Nebenprodukt bei der Entstehung von piRNA-produzierenden Loci darstellt.

Abstract

Small non-coding RNAs (sRNAs or sncRNAs) are present in every domain of life and undertake a great diversity of tasks including gene regulation, transposon repression and antiviral defense. To execute their functions, sRNAs such as microRNAs (miRNAs), small-interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs), all depend on Argonaute proteins that form larger complexes with additional factors. These complexes are then guided by the sRNA to their target nucleic acids based on sequence complementarity. In addition to these commonalities, evolution has given rise to a broad range of adaptations for many sRNA systems in different lineages. The largely animal-specific class of piRNAs, which associates with Piwi clade Argonaute proteins, represents a system with a particularly remarkable degree of evolutionary plasticity. piRNAs are typically known to defend genome integrity in the animal germline against transposable elements (TEs) that become active during the epigenetic reprogramming of early germ cells. Hence the piRNA pathway is involved in an ongoing arms race against TEs attempting to proliferate and transfer to the next generation through host germlines.

The present thesis aims to facilitate sRNA research in general and addresses outstanding questions on the piRNA pathway using a mostly comparative strategy. After summarizing the latest state of art knowledge on sRNA-based TE silencing in eukaryotes, it describes a newly developed software tool called unitas, for the annotation of a variety of sRNAs, including miRNAs and their isoforms, piRNAs, plant-specific phased siRNAs (phasiRNAs), tRNA-derived fragments (tRFs) and other RNAs, in a large selection of species from sRNA sequencing data. The underlying study of the subsequent chapter demonstrates the activity of the PIWI/piRNA pathway in the soma of mollusks, including clear signs of post-transcriptional TE repression and adaptations to targeting young transposons. Additionally, it reveals a dynamic expression pattern of piRNA-producing loci during oyster development, providing sources for piRNAs from different TE families. This work and other recent findings suggest that the near germline-specificity of the piRNA pathway in vertebrates likely represents an adaptation that was acquired later in the evolution of animals.

In the next chapter, an extensive analysis of the piRNA transcriptome of adult porcine testis uncovers post-transcriptional processing of protein-coding genes within a piRNA-specific amplification loop, termed the ping-pong cycle, which is also demonstrated in mouse and human. It further shows that piRNA clusters, which are large genomic loci that produce the majority of piRNAs in adult testis, contain gene and pseudogene sequences that might serve as a source for antisense piRNA that target corresponding parent genes. Finally, the last part of the thesis explores the evolution of piRNA clusters and integrated pseudogenes in primates and their ability to regulate protein-coding genes across species. However, since such piRNA clusters evolve rapidly and the gene targeting capacity by pseudogene-derived piRNAs is weakly maintained among primate species, it seems questionable that pseudogenes in piRNA clusters have a critical function. Based on evidence that piRNA clusters tend to be located in regions with elevated gene-density and higher GC content, both of which are indicators of open and active chromatin, it seems likely that the presence of pseudogenes is merely a byproduct resulting from the generation of piRNA-producing loci.

Introduction

Small non-coding RNAs

The discovery of the mechanism of RNA interference (RNAi) in *Caenorhabditis elegans* by Andrew Fire and Craig Mello twenty years ago (Fire et al. 1998) has revolutionized the life sciences and gave rise to the ever growing field of small non-coding RNA (sRNA) biology. Starting with their observation of the gene-silencing potential of double-stranded RNA (dsRNA), extensive research has revealed the existence of several different types of sRNAs, which build the basis of RNAi. The three major sRNA classes include small-interfering RNAs (siRNA) (Hamilton and Baulcombe 1999, Hammond et al. 2000, Zamore et al. 2000, Bernstein et al. 2001, Elbashir et al. 2001), microRNAs (miRNA) (Reinhart et al. 2000, Pasquinelli et al. 2000, Lagos-Quintana et al. 2001, Lau et al. 2001, Lee and Ambros 2001) and the mostly animal-specific Piwi-interacting RNAs (piRNA) (Aravin et al. 2006, Girard et al. 2006, Grivna et al. 2006, Watanabe et al. 2006), which are involved in processes such as gene-regulation, transposon control and anti-viral defense (Lee et al. 1993, Tabara et al. 1999, Ketting et al. 1999, Reinhart et al. 2000, McCaffrey et al. 2003, Pedersen et al. 2007, Aravin et al. 2007).

Common to all of these sRNA classes is the association to proteins of the evolutionary well-conserved Argonaute family, whose representatives are found in all domains of life (Swarts et al. 2014). Small RNAs bound to Argonaute proteins form the RNA-induced silencing complex (RISC), together with additional factors (Tabara et al. 1999, Hammond et al. 2000, Hammond et al. 2001, Ma et al. 2004, Lingel et al. 2004). The RISC recognizes target transcript sequences complementary to its guide sRNA, whereupon the Argonaute fosters post-transcriptional silencing by mRNA cleavage, decapping, deadenylation or translation blocking (Olsen and Ambros 1999, Martinez et al. 2002, Meister et al. 2004, Liu et al. 2004, Behm-Ansmant et al. 2006, Lu et al. 2009). In some cases, when acting in the nucleus on nascent transcripts, the RISC can recruit additional factors that induce epigenetic silencing (Sigova et al. 2004, Bühler et al. 2006). Argonautes are divided into two sub-families, Argonaute-like proteins that associate with miRNAs and siRNAs (Meister et al. 2004, Baumberger and Baulcombe 2005), and Piwi-like proteins that interact with piRNAs (Aravin et al. 2006, Girard et al. 2006). Another general distinction is that both, miRNAs and siRNA, are produced from double-stranded RNA molecules by the endoribonuclease Dicer (Bernstein et al. 2001), while piRNAs are generated from single-stranded RNA molecules, independently from Dicer (Vagin et al. 2006, Gunawardane et al. 2007, Brennecke et al. 2007, Ipsaro et al. 2012, Nishimasu et al. 2012).

Besides these most prominent sRNAs, additional classes are still being identified. Over the past years, tRNA-derived fragments (tRFs), which were previously seen as tRNA debris, have been recognized as actors in gene regulation and transposon control in many different organisms (Cole et al. 2009, Sharma et al. 2016, Martinez et al. 2017, Shorn et al. 2017). In an intriguing example, small fragments derived from the 3' ends of tRNAs in mouse stem cells have been shown to block the highly conserved primer binding site (PBS) of LTR-retrotransposons, which crucially rely on tRNA priming for reverse transcription (Shorn et al. 2017). Several types of tRFs are distinguished, depending on their origin, including small 5'-/3'-tRFs, larger 5'-/3'-tRNA-halves, internal itRFs and 3'U-tRFs, derived from the 3' end of pre-tRNA molecules (Keam and Hutvagner 2015), some of which have been shown to interact with Argonaute (Haussecker et al. 2010) and Piwi proteins (Keam et al. 2014, Hirano et al. 2014). In plants, additional siRNA kinds were found, including TE-silencing heterochromatic (hc-) siRNAs, natural antisense transcript (NAT-) siRNAs and phased, secondary, siRNAs (phasiRNAs), which are differentiated by their origin and biogenesis (Axtel 2013). For instance, phasiRNAs are generated from long non-coding, as well as protein-coding transcripts in a phased pattern, requiring the involvement of miRNAs as triggers and the siRNA biogenesis machinery in subsequent processing (Fei et al. 2013).

Two sub-types are distinguished, 21 and 24 nucleotide (nt) phasiRNAs, likely functioning in post-transcriptional regulation and in coordinating chromatin modifications, respectively. Their expression pattern differs among plant species, as for instance grasses express phasiRNAs specifically in reproductive tissues, in contrast to *Arabidopsis* (Fei et al. 2013). Conceivably, more small RNA pathways and even greater complexity of RNAi mechanisms will be discovered in the future.

The miRNA and siRNA pathways

Both, miRNAs and siRNAs are present throughout eukaryotes and have similar size ranges of 21-23 and 20-25 nt, respectively (Shabalina and Koonin 2008). In contrast to siRNAs however, many miRNAs are remarkably well conserved across large evolutionary distances (Bentwich et al. 2005). The biogenesis of canonical miRNAs starts with the transcription of distinct miRNA genes by RNA polymerase II (Pol II), resulting in long pri-miRNAs, which fold back to form hairpin structures (Lee et al. 2002, Lee et al. 2004). In animals, such hairpins are taken up by Microprocessor, a complex consisting of the RNA-binding protein Pasha and the endoribonuclease Drosha (Denli et al. 2004), which cuts the stem region of a hairpin with a 2 base pair (bp) offset, creating a stem-loop of about 60 nt that is called pre-miRNA (Lee et al. 2003). After export to the cytoplasm (Yi et al. 2003, Lund et al. 2004), the pre-miRNA is further processed by Dicer (Grishok et al. 2001, Hutvagner et al. 2001), which cuts the stem near the loop, again with a 2 bp offset, leaving a miRNA duplex with 2 nt 3' overhangs on each side (Lee et al. 2003, Zhang et al. 2004). In addition to canonical metazoan miRNAs, there are types of non-canonical miRNAs that do not require Drosha processing. For instance, some introns serve as pre-miRNAs, dubbed mirtrons, for which the ends are defined by the spliceosome (Okamura et al. 2007, Ruby et al. 2007), while some miRNA genes are transcribed as endogenous short-hairpin RNAs (shRNAs) (Babiarz et al. 2008). Generally in plants, miRNA duplexes are created directly from pri-miRNAs by Dicer in the nucleus (Kurihara and Watanabe 2004), following 2'-O-methylation of 3' ends by HEN1, improving stability of plant miRNAs (Yu et al. 2005), and subsequent export to the cytoplasm. In the last step, one RNA molecule of a duplex is finally loaded on an Argonaute protein (Iwasaki et al. 2010) and the RISC is assembled.

During miRNA biogenesis and also afterwards through RNA editing, different isoforms of miRNAs, called isomirs (Morin et al. 2008), are generated to some percentage. For instance, inconsistent selection of strands from miRNA duplexes can result in two functional miRNAs from the same gene with different targets (Ro et al. 2007). Moreover, although cleavage by Drosha and Dicer is very accurate for conserved miRNAs, a certain degree of imprecision can lead to some heterogeneity at 5' and 3' ends (Chiang et al. 2010). Further modifications include the addition of nucleotides at 3' ends, mostly by adenylation and uridylation (Lu et al. 2009, Wyman et al. 2011), which has been shown to affect targeting effectiveness (Burroughs et al. 2010). Additionally, internal modifications, such as through adenine-to-inosine RNA editing, leading to pairing with cytosine instead of uracil, can redirect miRNA targeting (Kawahara et al. 2007).

Once a miRNA is loaded on an Argonaute protein, it is guided to mainly genic target transcripts to induce post-transcriptional silencing either by translation blocking, mRNA degradation, decapping or deadenylation (Olsen and Ambros 1999, Meister et al. 2004). In metazoans, the recognition of target sites, which lie mostly in 3' UTRs, depends decisively on the seed, a region spanning from positions 2-8 from the 5' end of the miRNA (Lewis et al. 2003, Lewis et al. 2005). In contrast, miRNAs of plants, similar to siRNAs in general, require near-perfect complementarity to their binding sites, lying mostly within the coding sequence of their targets (Rhoades et al. 2002, Reinhart et al. 2002).

Unlike miRNAs, siRNAs are not produced from discrete sRNA genes, but instead from a great variety of sources, such as viral RNAs, transposon transcripts and genic mRNAs (Wilkins et al. 2005, Sijen and

Plasterk 2003, Chung et al. 2008, Ghildiyal et al. 2008) in the form of hairpin structures or dsRNA. Plants, fungi and invertebrates use RNA-dependent RNA polymerases to synthesize dsRNA from single stranded TE and gene transcripts, whereas viruses produce double-strands through their own RNA polymerase (Dalmay et al. 2000, Cogoni and Macino 1999, Smardon et al. 2000, Stein et al. 2003). Similar to miRNA biogenesis, double-stranded siRNA precursors are cleaved by Dicer (Bernstein et al. 2001) and bound by Argonaute proteins (Meister et al. 2004, Baumberger and Baulcombe 2005). In plants, siRNAs are methylated at 3' ends by HEN1, just as plant miRNAs (Yang et al. 2006). Besides post-transcriptional regulation, siRNAs can direct transcriptional silencing when binding nascent mRNAs by recruiting factors that induce DNA methylation (Mette et al. 2000) or histone modification (Hall et al. 2002, Volpe et al. 2002, Verdel et al. 2004, Bühler et al. 2006, Fagegaltier et al. 2009).

The piRNA pathway

Piwi-interacting RNAs, typically 24-32 nt long and bound to Piwi-like (PIWI) proteins, are present in animals (Aravin et al. 2006, Girard et al. 2006, Grivna et al. 2006, Watanabe et al. 2006, Vagin et al. 2006, Saito et al. 2006) and ciliates (Fang et al. 2012) and are primarily associated with defending the germline against transposons (Aravin et al. 2007). Different metazoan lineages possess a varying number of PIWI paralogs. In *Drosophila* three paralogs are present, namely Piwi, Aubergine (Aub) and Ago3 (Vagin et al. 2006, Brennecke et al. 2007), however insects in general have very diverse sets of PIWI paralogs (Lewis et al. 2016). Silkworms, for instance, express only two variants, Siwi and BmAgo3 (Kawaoka et al. 2008), similar to the zebrafish, which is equipped with Ziwi (Piwi-like 1 or Piwil1) and Zili (Piwil2) (Houwing et al. 2008). Mammals usually contain four paralogs, Piwil1-4, however, mice apparently lost Piwil3 and only express Miwi (Piwil1), Mili (Piwil2) and Miwi2 (Piwil4) (Aravin et al. 2006, Girard et al. 2006, Carmell et al. 2007). Piwil3 was demonstrated to be most active in the female germline in other mammals (Roovers et al. 2015), whereas Piwil1, 2 and 4 are the predominant PIWI proteins in testes. Accordingly, mice apparently compensate the loss of Piwil3 by employing siRNAs for TE defense that are produced by an oocyte-specific Dicer isoform (Flemr et al. 2013).

The biogenesis of piRNAs proceeds in two different pathways, both being Dicer-independent and requiring single-stranded RNA as source (Vagin et al. 2006, Brennecke et al. 2007, Gunawardane et al. 2007). A large fraction of piRNAs in both vertebrates and invertebrates are produced from discrete genomic loci, called piRNA clusters, ranging from a few thousand bases (kb) to over 100 kb (Aravin et al. 2006, Girard et al. 2006, Grivna et al. 2006, Watanabe et al. 2006, Brennecke et al. 2007). These loci produce large precursor transcripts that are cleaved in the cytoplasm by a mitochondrial single-strand specific endoribonuclease called Zucchini (Zuc/mZuc) in *Drosophila* and PLD6 (mitoPLD) in mammals (Ipsaro et al. 2012, Nishimasu et al. 2012, Watanabe et al. 2011, Huang et al. 2011), thereby initiating the primary piRNA pathway. Generally, other single-stranded RNAs, such as TE transcripts can be fed into Zuc/PLD6 processing (Aravin et al. 2008). The cleaved RNA fragments that are termed piRNA intermediates, are then taken up by PIWI proteins, Piwi and Aub in flies (Gunawardane et al. 2007, Brennecke et al. 2007) or Piwil1 and Piwil2 in vertebrates (Girard et al. 2006, Aravin et al. 2008), which heavily select for 5' uridines (1U), leading to a strong 1U bias in piRNAs bound to Piwi, Aub, but not Ago3 (Brennecke et al. 2007), and further for Piwil1 and Piwil2, but also Piwil4 (Aravin et al. 2006, Girard et al. 2006, Carmell et al. 2007), due to intrinsic preferences for 1U of these PIWI paralogs (Kawaoka et al. 2011, Cora et al. 2014). After binding to PIWI proteins, the piRNA intermediates or pre-piRNAs, being still 35-40 nt long, are trimmed by an exoribonuclease, called Trimmer in silkworm (Kawaoka et al. 2011, Izumi et al. 2016), PARN-1 in *C. elegans* (Tang et al. 2016) and PNLDC1 in mammals (Zhang et al. 2017), whereas flies use the non-homologous exoribonuclease Nibbler (Feltzin et al. 2015). This step determines the final piRNA length, which is distinct for each PIWI paralog. In

mice, piRNA lengths are ~30 nt for Miwi, ~26 nt for Mili and ~28 nt for Miwi2 (Aravin et al. 2006, Girard et al. 2006, Aravin et al. 2008), while in flies piRNA lengths are ~26 nt for Piwi, ~25 nt for Aub and ~24 nt for Ago3 (Brennecke et al. 2007). Finally, after trimming, piRNAs are methylated at the 2'-OH group of the 3' end by the methyltransferase HEN1, enhancing their stability (Horwich et al. 2007, Houwing et al. 2007, Kirino and Mourelatos 2007a, Kirino and Mourelatos 2007b, Ohara et al. 2007, Saito et al. 2007, Kawaoka et al. 2011).

The secondary piRNA pathway starts with the recognition and slicing of reverse complementary target transcripts by the PIWI/piRNA complex in the cytoplasm (Gunawardane et al. 2007, Brennecke et al. 2007, Aravin et al. 2008). In fly ovaries, target slicing by Aub with a 10 nt offset from the 5' end of the guiding piRNA creates a new antisense piRNA precursor, for which the 5' end is defined by the cleavage site. This piRNA precursor is taken up by Ago3 and, similar to the primary pathway, trimmed and methylated, creating a secondary mature piRNA, which first 10 nucleotides at the 5' end are complementary with those of the primary piRNA that originally recognized the target site (Gunawardane et al. 2007, Brennecke et al. 2007). Ago3 can in turn cleave complementary target transcripts, which are taken up by Aub and processed into mature piRNAs, closing an amplification loop, termed ping-pong cycle (Brennecke et al. 2007). Due to the 1U bias of Aub-bound piRNAs, Ago3-bound piRNAs have a tendency for adenine at position 10 (10A), which however is also caused by an intrinsic preference in Aub for adenine at this position (Wang et al. 2014). Both, Aub and Ago3, are localized in a perinuclear, electron dense structure called nuage (Brennecke et al. 2007). In contrast, Piwi predominantly localizes in the nucleus (Cox et al. 2000, Brower-Toland et al. 2007) and does not participate in the ping-pong cycle (Malone et al. 2009) but instead acts on chromatin (Brower-Toland et al. 2007) inducing transcriptional silencing of transposons (Wand and Elgin 2011, Sienski et al. 2012, Rozhkov et al. 2013, Le Thomas et al. 2013). Also, Piwi was shown to act alone in the somatic follicle cells of fly gonads, binding only primary piRNAs (Malone et al. 2009, Li et al. 2009, Saito et al. 2009).

Secondary piRNA biogenesis through the ping-pong cycle was also shown to be present in prenatal mouse testis (Aravin et al. 2007, Aravin et al. 2008). In this case, Mili and Miwi2 are the ping-pong partners, with Mili proteins binding mostly primary piRNAs, while Miwi2 is enriched for secondary piRNAs (Aravin et al. 2008, De Fazio et al. 2011). In the cytoplasm of primordial germ cells, Mili and Miwi2 are localized in two distinct types of granules that lie in close proximity to each other, the pi-body and piP-body, respectively, each containing different sets of interacting proteins (Aravin et al. 2009). Similar to fly Piwi, Miwi2 is also mainly localized within the nucleus inducing transcriptional silencing through DNA methylation by the de novo DNA methyltransferase DNMT3L (Aravin et al. 2008, Kuramochi-Miyagawa et al. 2008). However, it was shown that Mili can sustain an ongoing ping-pong cycle on its own and also induce DNA methylation that is independent from Miwi2 (De Fazio et al. 2011, Manakov et al. 2015). Importantly, in mouse testis, different PIWI paralogs are expressed at different times during germline development and spermatogenesis. Miwi2 is present in early gonocytes after global demethylation and during de novo methylation until shortly after birth (Aravin et al. 2008). Mili expression starts in primordial germ cells and lasts until the round spermatid state of spermatogenesis (Aravin et al. 2006, 2008). Finally, Miwi is detectable from the pachytene stage of meiosis to the haploid round spermatid state (Deng and Lin 2002, Aravin et al. 2006, 2008). Hence in adult mouse testes only Miwi and Mili are active, namely during spermatogenesis. Both paralogs are to a great part localized in large and dense cytoplasmic perinuclear granules, similar to pi- and piP-bodies of prenatal prospermatogonia, called chromatoid bodies that form in late pachytene spermatocytes and last until the post-meiotic round spermatid state (Kotaja et al. 2006, Meikar et al. 2011). Further, it was shown that chromatoid bodies contain other protein components of the piRNA pathway and accumulate mRNAs and piRNAs (Meikar et al. 2011).

Similar to the dynamic expression of PIWI protein paralogs during germline development and spermatogenesis, distinct populations of piRNAs are bound to PIWI proteins at different time points. Pre-pachytene piRNAs are present in primordial germ cells and during spermatogenesis until the pachytene stage of meiosis, upon which they are replaced by pachytene piRNAs (Aravin et al. 2007, 2008). These two piRNA populations have various properties that differentiate them substantially. Firstly, pre-pachytene piRNAs are enriched in TE-derived sequences, wheres pachytene piRNAs have a reduced TE share and are generally more abundant. Secondly, there is little overlap between genomic clusters of both types and while only about half of pre-pachytene piRNAs originate from clusters that are also generally single-stranded, the vast majority of pachytene piRNAs is produced by clusters, which are on average larger and to some part bidirectional (Aravin et al. 2007). In addition pachytene piRNA clusters are controlled by the transcription factor A-MYB, which also initiates the transcription of important PIWI pathway genes, such as Miwi (Li et al. 2013). In contrast to the pre-pachytene piRNAs of the prenatal germ cells, the vast majority of pachytene piRNAs are generated through the primary piRNA pathway, while only a minority is produced by secondary biogenesis (Reuter et al. 2011, Beyret et al. 2012). Also the pachytene piRNAs of both Miwi and Mili were shown to originate from similar sources, from which they are generated for the most part by primary processing.

Besides primary and secondary piRNAs, a tertiary piRNA biogenesis pathway was discovered in *Drosophila* ovaries and mouse testes (Han et al. 2015, Mohn et al. 2015, Homolka et al. 2015). In flies, secondary piRNAs, bound by Ago3, can initiate the generation of primary piRNAs from cleaved TE transcripts that are cut by Zuc in a phased manner into ~26 nt fragments, which are taken up by Piwi after the first fragment is bound by Aub. Similarly in mice, mitoPLD can produce phased Mili-bound piRNAs, which are however trimmed, since mitoPLD cleavage creates slightly larger RNA fragments (Mohn et al. 2015, Yang et al. 2016). Hence the production of phased primary piRNAs represents a conserved mechanism that enhances sequence diversity of piRNA pools.

Beyond TE silencing, the piRNA pathway has been increasingly implicated in further tasks in a variety of species. In *Drosophila*, PIWI proteins and piRNAs are involved in deadenylation and decay of mRNAs in early embryos (Rouget et al. 2010, Barckmann et al. 2015). Similarly in mouse it was shown that Miwi-associated pachytene piRNAs direct broad mRNA elimination during late spermiogenesis (Gou et al. 2014). Fascinatingly, neuronal piRNAs in the sea slug *Aplysia* were found to be involved in controlling memory-related synaptic plasticity by inducing methylation on the promoter of CREB2, a major inhibitor of memory formation (Rajasethupathy et al. 2012). Further, it was demonstrated in the silkworm *Bombyx mori* that one single female-specific piRNA is responsible for sex determination (Kiuchi et al. 2014). These extraordinary specializations yielded by evolution exemplify the extensive flexibility of the piRNA pathway.

Transposable elements

Albeit the fact that many roles for piRNAs beyond transposon silencing were uncovered, the repression of TEs is still regarded as the major function of the piRNA pathway. Transposons are ubiquitously present and often extremely abundant mobile DNA elements that can move from one genomic region to another (McClintock 1950, Kazazian 2004). Thereby they pose a serious threat to genome integrity, e. g. through the disruption of coding genes (Kazazian et al. 1988, Miki et al. 1992) or by causing genomic rearrangements via homologous recombination between non-allelic TE copies (Martin and Lister 1989, Döring et al. 1990). Two major TE classes, RNA transposons (class 1) and DNA transposons (class 2), are distinguished by their mode of transposition, either involving an RNA intermediate or direct DNA excision and genomic reintegration, respectively (Kazazian 2004). The mobility of DNA transposons depends on an enzyme performing both, excision and insertion called

transposase, which autonomous elements encode themselves. Non-autonomous DNA transposons on the other hand, such as the miniature inverted-repeat transposable elements (MITEs), are mobilized by related autonomous families (Feschotte and Mouchés 2000). A hallmark of DNA transposons in general is the presence of terminal inverted repeats (TIRs) at their flanks, which are recognized by transposases to initiate the excision reaction (Feschotte and Pritham 2007). Other characteristics are direct repeats that are left behind as footprints of removed DNA transposons. These repeats are generally formed by TE integrations, which involve staggered DNA breakage and filling resulting gaps, leading to target site duplications (Kazazian 2004). In contrast to DNA transposons, however, RNA transposons, also called retrotransposons, are not mobilized by excision, but instead use reverse transcription of their mRNA to move around the genome and proliferate in the process.

Retrotransposons are grouped into LTR retrotransposons and non-LTR retrotransposons, defined by the presence or absence of long terminal repeats (LTRs) at their flanks (Goodier et al. 2016). LTR retrotransposons are very similar to exogenous retroviruses and encode retroviral genes, such as Gag (group-specific antigen) and Pol (polymerase) that produces reverse transcriptase (RT) and integrase (IN) proteins, as well as Env (envelope) in the case of endogenous retroviruses (ERVs), which however lack extracellular mobility, for instance due to a non-functional Env protein (Lerat and Capy 1999, Stoye 2012). LTRs are crucially involved in retrovirus-like reverse transcription, which is primed by tRNAs at the primer binding site (PBS) near the 5' LTR (Eickbush and Jamburuthugoda 2008).

Non-LTR retrotransposons, including the autonomous long interspersed elements (LINEs) and the non-autonomous short interspersed elements (SINEs), also rely on reverse transcription (Singer 1982), but use a different mechanism (Luan et al. 1993). LINEs are 4-6 kb long and contain a 5' UTR, including a Pol II promoter sequence (Swergold 1990), two open reading frames (ORF1, ORF2) and a short 3' UTR (Scott et al. 1987). ORF1 encodes an RNA-binding protein (Kolosha and Martin 1997; Moran et al. 1996), while ORF2 produces a large protein with endonuclease and reverse transcriptase activities (Mathias et al. 1991), all of which are required for transposition. Together they bind LINE transcripts and after reimport into the nucleus, integrate them into the genome through a process called target primed reverse transcription, where cDNA synthesis proceeds directly at the insertion site (Luan et al. 1993, Cost et al. 2002). The enzymatic machinery of LINEs is also utilized for the mobilization of SINEs and for the generation of processed pseudogenes by retrotransposition of genic mRNAs (Jurka 1997, Kazazian and Moran 1998). SINEs are derived from tRNAs (Singer 1982), 5S ribosomal RNAs (Kapitonov and Jurka 2003) or 7SL RNAs (Ullu and Tschudi 1984) and typically carry an internal Pol III promoter, however they do not encode any proteins, but rely entirely on LINEs to be mobilized. 7SL RNAs-derived SINEs include primate-specific Alu elements and B1 elements that are present in rodent genomes (Ullu and Tschudi 1984). Finally, some retrotransposons are composed of merged segments, such as the hominid specific SVA elements, which consist of an ERV-derived SINE-R sequence, a segment of variable number of tandem repeats (VNTR) and an Alu fragment (Ostertag et al. 2003).

Aims of the thesis

Though many basic characteristics of the respective small RNA pathways are the same in different lineages, small RNA-based machineries in general and the PIWI/piRNA system in particular are very adaptive and show many distinctions between separate phylogenetic groups. While the work on model organisms built the foundation for our knowledge on functions and mechanisms of sRNA pathways and continues to do so, studies in non-model species repeatedly illustrate their evolutionary plasticity and often reveal remarkable adaptations, which deepen our understanding of fundamental features and their evolution.

The first aim of this thesis is the creation of a software tool for the annotation of small RNA sequence data sets in both model and non-model organisms. It should enable the annotation of a variety of sRNA types, such as miRNAs, piRNAs, phasiRNAs, tRFs and generally other non-coding and coding RNAs, some of which are lineage-specific. Further, the tool is intended to be usable by non-experts in regards to bioinformatics background and should therefore depend on as few prerequisites as possible and work mostly automatic through subsequent annotation steps. The ultimate goal is to facilitate small RNA research in a broad range of species and to simplify the necessary bioinformatic analysis.
The second aim is to investigate specific but less well understood features and functions of the PIWI/piRNA pathway in an evolutionary context, including putative activity of PIWIs and piRNAs in the soma. In mammals, piRNAs are predominantly expressed in the male and female germline and in early embryos (Aravin et al. 2006, Roovers et al. 2015). Somatic piRNAs in mammals have been reported (Zheng et al. 2011), but viewed with great skepticism (Ross et al. 2014). Indeed it was confirmed that the majority of these piRNA-like RNAs consists of non-coding RNA fragments (Tosar et al. 2018). Somatic activity of piRNAs is well documented in *Drosophila*, but restricted to the fly Piwi paralog and primary piRNAs in follicle cells (Malone et al. 2009, Li et al. 2009, Saito et al. 2009). Mollusks, like insects, belong to the phylogenetic group of protostomians that split off early in evolution from deuterostomians, which include the vertebrates (Edgecombe et al. 2011). The determination of the presence or absence of somatic piRNAs and their possible characterization in mollusks, specifically the pond snail *Lymnaea stagnalis* and the pacific oyster *Crassostrea gigas*, is expected to yield insights into the basic evolution and origin of the PIWI/piRNA system.
Another still not fully understood function of piRNAs is the regulation of protein-coding genes. A slicer-independent role of PIWI proteins and piRNAs in the removal of genic mRNA was discovered in *Drosophila* embryogenesis (Rouget et al. 2010) and similar observations were made in mouse spermatids, where pachytene piRNAs are involved in the elimination of mRNAs (Guo et al. 2014). Further investigation within this work on a gene-regulatory role of the PIWI/piRNA system in other adult mammal testes, starting with the domestic pig *Sus scrofa*, as it contains the full set of typical mammalian PIWI paralogs, should elucidate the involvement of secondary piRNA processing and possible sources of gene-targeting piRNAs. Moreover, since the vast majority of pachytene piRNAs is produced by piRNA clusters (Aravin et al. 2007), a reconstruction of the evolutionary relationships of these piRNA producing loci and their capacities to generate TE- and gene-targeting piRNAs shall be explored in primates, including the species *Homo sapiens*, *Macaca mulatta*, *Macaca fascicularis*, *Callithrix jacchus*, *Microcebus murinus* and *Loris tardigradus*. Specifically, a proposed role for pseudogene-containing piRNA clusters as regulatory elements of coding genes though the piRNA pathway (Hirano et al. 2014) should be tested with regards to conservation and maintained gene-targeting among species.

Chapter overview

Overall, this thesis contains five chapters that are either published (chs. 1,2,4), accepted for publication (ch. 3) or in preparation to be published (ch. 5) in peer-reviewed journals. The first chapter is a broad review of the current knowledge on small RNA pathways controlling transposon activity in its many variations in eukaryotes. It touches on piRNAs, siRNAs and miRNAs in animals, plants, fungi and ciliates, and discusses the astonishing adaptations of each phylum. Most organisms take special care to protect their germline against transposon proliferation, since it is the ultimate battleground for TEs to make it to the next generation. While metazoans use both piRNAs and siRNAs to silence TEs, plants mainly rely on siRNAs, but with a certain involvement of miRNAs, as just recently was shown. Ciliates on the other hand use piRNAs yet take a fundamentally different route to keep transposons under control. Instead of silencing via transcript degradation or the establishment of suppressive epigenetic

modifications, ciliate piRNAs scan the genome for repetitive DNA stretches in order to completely eliminate TEs from the genome during sexual reproduction. Altogether the review provides an overview on the remarkable diversity of small RNA-based transposon repression systems that have arisen during the evolution of eukaryotes.

Chapter two describes a new convenient software tool, called unitas, that is designed to annotate small RNA sequencing data with very few prerequisites in a wide range of species, uniting a variety of different applications. Specifically, it can be used for annotation of miRNAs and their isoforms, tRNA-derived fragments (tRFs) and other ncRNAs. In animals it helps to identify putative piRNAs, including analyses of 5' overlap rates, nucleotide frequencies and mapping to known piRNA clusters. In plants unitas can be employed to discover phasiRNAs, plant-specific phased siRNAs. Finally, unitas includes useful tools for 3' adapter identification and trimming, as well as for filtering of low complexity reads from datasets. Runs are started from the command line in windows, linux or macOS with a single command, and all additional programs and reference data are downloaded automatically from the relevant databases. Unitas was thoroughly tested with real and artificial data and its performance was compared to existing software tools that exert similar tasks.

In chapter three we show the conserved expression of PIWI genes and piRNAs in somatic tissues in mollusks. Scanning of unannotated genomes and phylogenetic analyses reveal that two PIWI proteins make up the standard repertoire in mollusks and that these are homologous to the vertebrate paralogs Piwil1 and Piwil2, while the Piwil1 gene underwent various duplications during molluskan evolution. The two paralogs, as well as piRNAs, are expressed not only in the germline, but also in a variety of somatic tissues in the pacific oyster and the pond snail. Moreover, the presence of so-called ping-pong signatures in every examined body part, which is the result of ongoing post-transcriptional processing by the secondary PIWI/piRNA pathway, is clear evidence of ubiquitous PIWI activity. The piRNA clusters of both species are enriched for transposon sequences, which are also biased towards specific TE families and younger TE age in general, indicating a role in transposon control in the germline and the soma. In addition, the presence of ping-pong signatures on protein-coding genes in both species suggests a function in gene regulation in mollusks. Interestingly, in the oyster different populations of piRNA clusters with varying sets of TE families are expressed at distinct developmental stages and adult tissues, while most clusters are active in male and female germline and the hemolymph, suggesting sub-functionalization in groups of piRNA-producing loci. Together with similar findings in arthropods (Lewis et al. 2018) and cnidarians (Praher et al. 2017), these results suggest that somatic PIWI and piRNA expression is an ancestral state of metazoans that was mostly lost on the way to vertebrates.

In the fourth chapter we report the identification of gene-targeting piRNAs, as well as piRNA clusters that incorporate protein-coding genes and pseudogenes in pig testes. The underlying study provides an in-depth bioinformatic characterization of the porcine piRNA transcriptome, including TE-associated reads and tRNA-derived sRNAs with piRNA-traits. We then show that gene-derived piRNAs exhibit signs of secondary PIWI/piRNA pathway processing, indicated by ping-pong signatures, in pig, mouse and human, suggesting that gene-targeting by piRNAs is conserved in mammals. Further, porcine piRNA clusters are enriched for genes and pseudogenes, while being depleted of transposons. Where genes or pseudogenes lie in opposite orientation relative to cluster transcription, antisense piRNAs can be produced, which might regulate the expression of these genes or the corresponding parent genes, respectively. Overall, the study provides evidence for a role of the mammalian PIWI/piRNA pathway in post-transcriptional gene regulation and suggests a potential source of gene-targeting piRNAs in gene and pseudogene-containing piRNA clusters.

The last part of the thesis explores the evolution of primate piRNA clusters and therein contained pseudogenes to achieve a deeper understanding about the evolution of mammalian piRNA clusters in

general and the putative gene regulatory role of pseudogene-derived piRNAs. The study shows that only a minority of piRNA clusters is present and active among the examined primates and that even this group of clusters shows distinct expression profiles for each species, while nevertheless producing the majority of piRNAs. It further confirms that homologous piRNA clusters evolve at the same rate as the whole genome on the sequence level, indicating lack of selection pressure. We then show that pseudogenes in reserve orientation relative to cluster transcription in comparison to parallel copies exhibit neither elevated sequence identity to parent genes nor are consistently more abundant within each species or as homologs among species. Moreover, a minority of reversed pseudogenes produces piRNAs, targeting gene transcripts, that are processed in the secondary PIWI/piRNA pathway and this targeting of orthologous genes among species is very weakly conserved. Taken together, these results cast doubt on the idea that pseudogene-derived piRNAs play a major role in PIWI/piRNA-mediated gene regulation. However, this would then raise the question about the reason for the enrichment of pseudogenes in piRNA clusters. In order to address this issue, we examined the genomic environments of piRNA producing loci and show that clusters tend to be located in genomic regions with elevated gene density, which is an indicator of open chromatin (Gilbert et al. 2004), and correlating positively with pseudogene density as well as negatively with transposon age. It is thus conceivable that pseudogene enrichment is merely a by-product in the generation of piRNA clusters that possibly form in genomic segments of active chromatin, instead of representing islands of euchromatin themselves.

1. RNA-based transposon regulation in eukaryotes

Daniel Gebert[1], David Rosenkranz[1]

[1] Institute of Anthropology, Johannes Gutenberg University, Mainz, Germany

This chapter was published as a Review Article in *Wiley Interdisciplinary Reviews: RNA* under the title "RNA-based regulation of transposon expression" (Gebert and Rosenkranz, Wiley Interdiscip Rev RNA 2015 6:687-708).

1.1. Abstract

Throughout the domains of life, transposon activity represents a serious threat to genome integrity and evolution has realized different molecular mechanisms that aim to inhibit the transposition of mobile DNA. Small noncoding RNAs that function as guides for Argonaute effector proteins represent a key feature of so-called RNA interference (RNAi) pathways and specialized RNAi pathways exist to repress transposon activity on the transcriptional and posttranscriptional level. Transposon transcription can be diminished by targeted DNA methylation or chromatin remodeling via repressive Histone modifications. Posttranscriptional transposon silencing bases on degradation of transposon transcripts to prevent either reverse transcription followed by genomic reintegration or translation into proteins that mediate the transposition process. In plants, Argonaute-like proteins guided by short interfering RNAs (siRNAs) are essential for transposon repression on the epigenetic and posttranscriptional level. In the germline of animals, these tasks are often assumed by a second subclass of Argonaute proteins referred to as Piwi-like proteins, which bind a distinct class of small noncoding RNAs named piwi-interacting RNAs (piRNAs). Though the principles of RNAi pathways are essentially the same in all eukaryotic organisms, remarkable differences can be observed even in closely related species reflecting the astonishing plasticity and diversity of these pathways.

1.2. RNAi as a molecular defense against transposons

Transposable elements (TEs), popularly referred to as jumping genes, can be found in virtually all organisms and their representatives are as diverse as their hosts are. As their name suggests, TEs can physically relocate from one genomic locus to another and commonly two major classes of TEs are distinguished according to the mechanism of their transposition. Class 1 TEs, also known as retroposons, propagate via RNA intermediates that are subject to reverse transcription and reintegration into the host genome. In contrast, class 2 TEs, referred to as DNA transposons, are directly excised from one locus and reintegrated into another locus [1,2]. Both classes comprise autonomous and nonautonomous elements. While autonomous elements encode all the enzymes that are necessary for their transposition, nonautonomous elements highjack the enzymatic machinery of autonomous elements. Although TEs are more and more considered as powerful mutagens that have played an essential role in genome evolution, they nevertheless represent a steady threat for genome integrity [3,4]. TE transposition can result in, for example, disruption of functional genes, altered gene expression or aberrant splicing. Most relevant in the evolutionary context, they provide the prerequisite for ectopic recombination resulting in gene duplication, deletion or large-scale rearrangements. In order to ensure genome integrity over evolutionary time scales, species have established molecular defense mechanisms that employ RNA-induced silencing complexes (RISCs) which protect their genomes from TE propagation. RISCs represent dynamic enzymatic machineries that act on their target nucleic acids to promote epigenetic modifications or mRNA decay, a process referred to as RNA interference

(RNAi) [5]. Small noncoding (snc-) RNAs bound to Argonaute effector proteins constitute the functional heart of RISCs and different RNAi pathways can be distinguished depending on both, the class of sncRNA and the involved Argonaute protein. Common to them all is that their target specificity is realized by complementarity with the guiding sncRNA [6]. Argonaute proteins can be subdivided into the Argonaute-like (AGO) and the animal-specific Piwi-like (PIWI) clade and both classes bind different populations of small RNAs. AGO proteins interact with micro (mi-) RNAs and short interfering (si-) RNAs, which are ~20–24 nt in length and are processed from double stranded (ds-) RNA precursors by the RNase III type endonuclease Dicer [7,8]. In contrast PIWI-interacting (pi-) RNAs are typically longer (~24–32 nt) and are processed from single-stranded (ss-) RNA molecules in a Dicer-independent manner. AGO proteins are ubiquitously expressed and function in the regulation of protein-coding genes and, when bound to siRNAs, also in TE silencing. The role of miRNAs in TE suppression is controversial although it is evident that many miRNA genes descend from TEs. PIWI proteins are commonly restricted to the germline and are specialized in TE defense. We will outline the general features and important differences of RNAi pathways that utilize either siRNAs or piRNAs (Figure 1). In the course of this review, we will then take a closer look at the evolutionary realizations of both pathways in all their facets and specific peculiarities.

Do miRNAs control transposon expression?

Although the empirical evidence for a role of miRNAs in TE silencing is disputable, the sheer amount of TE-related miRNAs has led to the speculation that miRNAs may transcriptionally or post-transcriptionally control TE expression similar to siRNAs [26]. Numerous small RNAs that were annotated as miRNAs were shown to originate from TEs, especially from miniature inverted-repeat transposable elements (MITEs) [27–31]. MITE transcripts can form hairpin-like structures similar to miRNA precursors and thus represent putative Dicer substrates. However, it has not been experimentally proven whether biogenesis of MITE-derived small RNAs depends on miRNA pathway factors such as Drosha or DGCR8 and so at least some of them may actually represent siRNAs rather than genuine miRNAs. Either way, their physiological function remains a mystery since putative targets were identified using in silico approaches only. Theoretically, targets could be complementary sequences of DNA transposons, however, DNA transposons are inactive in species were MITE derived RNAs have been described. It was speculated that Dicer processing of fold back TE transcripts efficiently silences TEs that are not accessible for piRNAs and that the resulting small RNAs do not necessarily possess downstream function [32]. In human, Alu targeting miRNAs have been linked to genes with exonized Alu fragments thus displaying a gene regulatory role rather than a function in TE defense [33]. In Arabidopsis, miRNAs were shown to target epigenetically reactivated TEs launching the production of TE-derived siRNAs in a methylation-impaired background [34]. However, the relevance of this mechanism in wildtypes remains unclear. An indirect but yet important role for miRNAs in TE silencing may exist in *Drosophila* follicle cells, where loss of TE-derived piRNAs and TE reactivation was observed upon knock-down of specific miRNAs [35].

1.3. The siRNA pathway at a glance

Both, miRNAs and siRNAs originate from dsRNA precursor molecules that are processed by Dicer [7,8]. miRNAs generally originate from single genes whose transcripts form hairpin-like structures [9], whereas siRNA precursors can have diverse origins. One source of dsRNA are inverted repeats that can arise from the insertion of TEs in opposite directions and, similar to miRNA genes, yield hairpin-like transcripts. Furthermore, dual-strand transcription of one locus can give rise to complementary transcripts and so-called cis-dsRNA. Naturally, complementary transcripts can also originate from

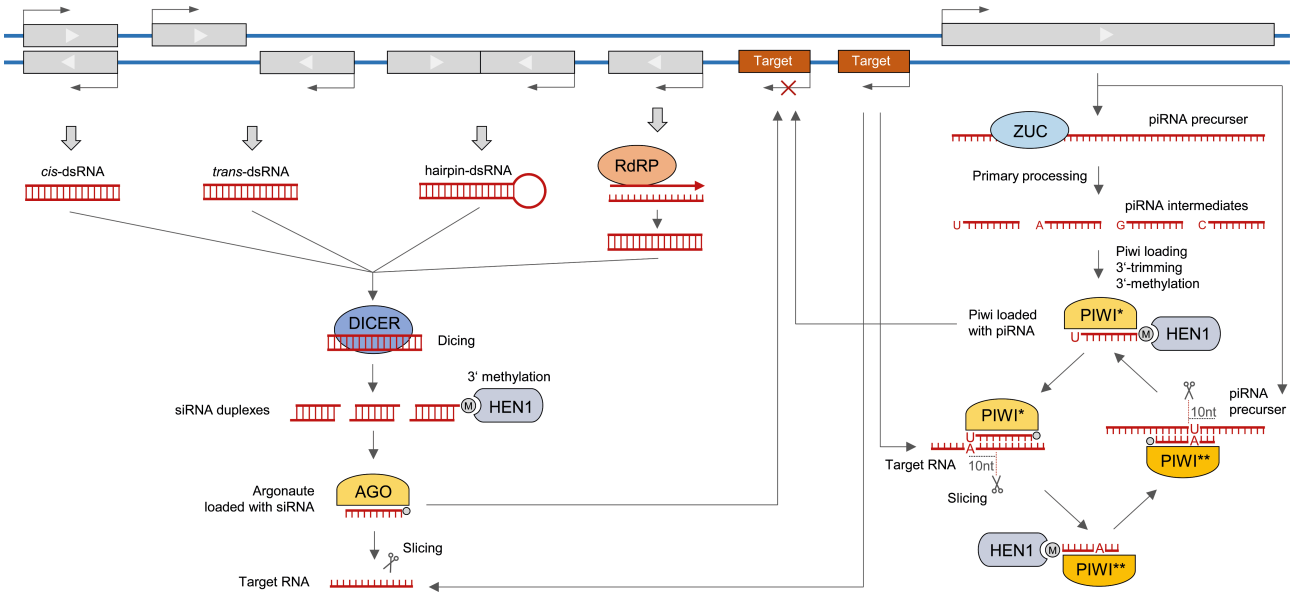


Figure 1 | The short interfering RNA (siRNA) and piwi-interacting RNA (piRNA) pathways at a glance. DNA is indicated in blue, RNA is displayed in red. Asterisks indicate different PIWI paralogs involved in canonical heterotypic ping-pong cycle. Abbreviations: M, methylation; U, uracil; A, adenine; G, guanine, C, cytosine.

distinct loci especially in the presence of interspersed repetitive DNA, yielding so-called trans-dsRNA. Yet another mechanism to produce dsRNA is synthesis of a new RNA strand on an RNA template by RNA-dependent RNA polymerases (RdRPs) that can be found in plants and some animal species but presumably not in vertebrates [10–12]. Once a dsRNA has been 'diced' into siRNA duplexes the so-called passenger strand is removed during RISC assembly, while the other strand is retained in the functional RISC [13,14]. In animals, endogenous siRNAs are loaded onto the AGO-family member Ago2 whereas numerous different AGO proteins can take up siRNAs in plants [15]. As part of the RISC, the loaded AGO proteins are now ready to exert targeted regulatory functions on the transcriptional and posttranscriptional level.

1.4. The piRNA pathway at a glance

piRNAs are generally processed from ssRNA but depending on the molecular mechanism of biogenesis we differentiate two discrete piRNA populations: Those that originate from primary, and those that originate from secondary processing. During primary processing, a ssRNA is sliced into smaller pieces by an endoribonuclease named Zucchini (Zuc) [16,17]. The resulting premature piRNAs are loaded onto PIWI proteins that heavily select for 5' U (1U) fragments [18]. Subsequently the 3' ends of premature piRNAs are trimmed by a yet unknown exonuclease and often 20-O-methylated by Hen1 [19,20]. Similar to the Ago2-siRNA complex, the loaded PIWI protein can now silence targets on the transcriptional level by inducing DNA methylation or Histone modifications. Alternatively, PIWI proteins loaded with primary piRNAs can trigger secondary piRNA biogenesis in a self-sustaining amplification loop pictorially called piRNA ping-pong [21-23]. In this process, PIWI proteins slice target transcripts with a 10-nt offset from the 5' end of the guiding piRNA owing to their RNase H-like activity [24]. The sliced target now serves as substrate for secondary piRNAs. It assembles with another PIWI protein and is subject to trimming and methylation just as primary piRNAs. Secondary piRNAs exhibit a bias for Adenine at position 10 (10A) that not merely results from base pairing with the primary 1U piRNA but rather from an intrinsic preference for 10A of the loaded PIWI protein [25]. The new PIWI-piRNA complex can target the same transcripts that were initially subject to primary processing and, in doing so, produce new piRNAs that resemble the initializing primary

piRNAs. This way the ping-pong loop ensures both, post-transcriptional target silencing and constant supply of new target specific piRNAs.

1.5. Nature's greatest tinkerer comes up with manifold solutions

The key factors of RNAi in plants, animals, and fungi are homologous and thus presumably evolved from ancestral proteins that were already present in the last common ancestor of today living eukaryotes [36]. However, up to 1.6 billion years of independent evolution [37,38] have led to a diversification of RNAi mechanisms that constantly coevolved with distinct transposon repertoires, giving rise to lineage-specific realizations that sometimes appear to share not much more than a common origin. In the following, we will discuss RNAi pathways in different species ranging from plants to mammals illustrating how species utilize their RNAi tools in ever-different variations for the ever-same purpose: Foiling transposon propagation.

1.5.1. RNAi-based TE defense in insects - One goal, many paths

The *Drosophila* genome encodes three PIWI-subclass proteins named Piwi, Aubergine (Aub), and Argonaute 3 (Ago3). *Drosophila* PIWI proteins are expressed in the male and female germline but in *Drosophila* ovaries specialized Piwi pathways were found to act in somatic and germline cells (Figure 2) [39].

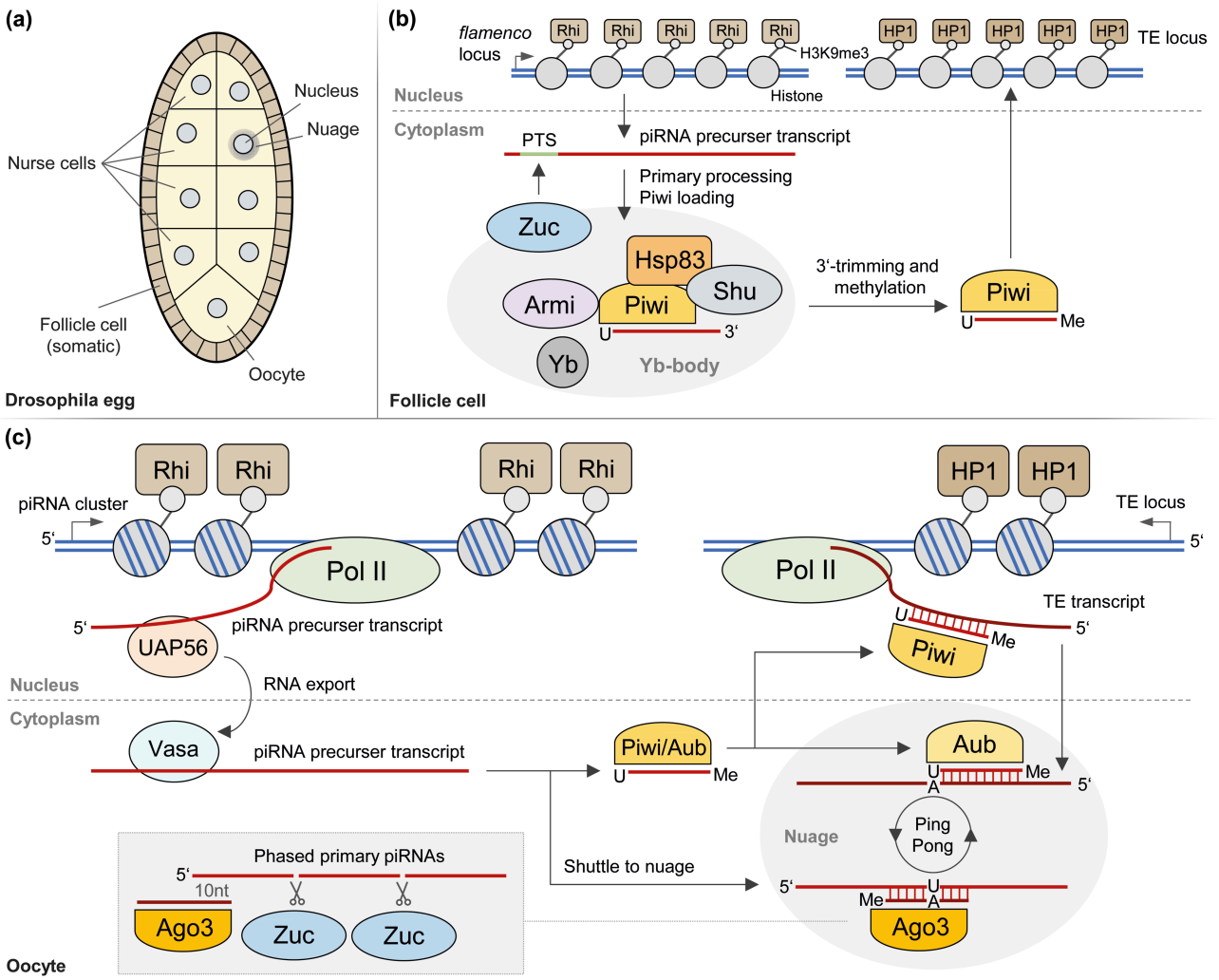


Figure 2 | Piwi pathway in *Drosophila*. (a) Sketch of the *Drosophila* egg. (b) Piwi-mediated transcriptional silencing in somatic follicle cells. (c) The piwi-interacting RNA (piRNA) pathway in the fly's germline. Abbreviations: Rhi, Rhino; Me, methylation; PTS, piRNA trigger sequence.

The somatic piRNA pathway

Somatic follicle cells express solely Piwi and no ping-pong amplification loop was observed in these cells [39]. Instead, piRNAs in follicle cells are exclusively produced from unidirectionally transcribed loci such as the X-chromosomal flamenco locus which contains a 5' structure (PTS, piRNA trigger sequence) that triggers primary processing [40]. flamenco is enriched for dead TE copies, mainly representing retrovirus-like elements from the gypsy family. The insertion direction of these elements suggests selective constraints acting on the flamenco locus favoring insertions that allow the production of antisense piRNAs. Many gypsy family elements in *Drosophila* encode functional envelop genes, and it was supposed that the somatic piRNA pathway of *Drosophila* represents an evolutionary adaptation to prevent germ cell infection with viral particles produced in surrounding follicular cells (Figure 2(a) and (b)) [41].

piRNA loading onto Piwi occurs at cytoplasmic Yb-bodies in dependency of other factors such as Armitage (Armi), Shutdown (Shu), and Heat shock protein 83 (Hsp83) [42–44]. Experiments demonstrate that Piwi but not its Slicer activity is required for proper germ cell development, which suggests that Piwi controls TE expression exclusively on the transcriptional level while being dispensable for ping-pong amplification [45–48]. In line with this, numerous studies report on a nuclear role for Piwi that involves epigenetic TE silencing by induction of H3K9 tri-methylation (H3K9me3) [41,49–56].

The germline piRNA pathway

In *Drosophila* germ cells, nuclear Piwi is accompanied by Ago3 and Aub which are located in germline-specific electron-dense perinuclear ribonucleoprotein particles termed nuage [21,22,57–59]. Nuage granules represent the processing site for germline piRNAs and contain numerous proteins that act in a sophisticated concert to ensure correct nuage assembly and TE repression [45,60–65]. But, although it is evident that nuage is the place where transcripts are processed into piRNAs via the ping-pong loop, we have just started to understand how specific transcripts are selected to feed the PIWI machinery.

Current evidence suggests that the nuclear DEAD box protein UAP56 recognizes the piRNA cluster-associated heterochromatin protein 1 (HP1) variant Rhino and shuttles piRNA cluster transcripts to nuclear pores where they are passed to the cytoplasmic RNA helicase Vasa [66]. Observations in *Bombyx mori* ovary cells suggest that Vasa is part of an Amplifier complex located in nuage that contains two PIWI ping-pong partners, one of them loaded with an antisense piRNA. This amplifier complex promotes loading of premature piRNAs processed from transposon transcripts onto Ago3 and is thus essential for secondary piRNA biogenesis [67]. Furthermore, the Tudor-domain protein Krimper is involved in directing primary piRNAs to be loaded onto Aub as it interacts with unloaded Ago3 and promotes symmetrical Arginine dimethylation, which blocks Ago3 for loading with primary piRNAs [68]. This way, the correct loading of each ping-pong partner with either primary or secondary piRNAs ensures a heterotypic ping-pong cycle with sense- and antisense piRNA pools assorted to a specific PIWI protein. piRNA-guided slicing during the ping-pong cycle can in turn initialize primary biogenesis that results in Zuc-dependent production of phased piRNAs. In this process, the target molecule is sliced consecutively starting from a ping-pong target site, and each downstream cleavage position determines the 3' and 5' end of adjacent (trail-) piRNAs, respectively. This means that one and the same transcript molecule can be subject to primary as well as secondary processing (Figure 2(c)) [40,69,70]. Since phased piRNAs are predominantly loaded onto nuclear acting Piwi, these piRNAs crucially shape the nuclear piRNA repertoire and enforce transcriptional silencing of corresponding TE loci [71]. Besides the presence of a 5' PTS as observed for the flamenco locus, [40] ping-pong-induced primary processing represents the main trigger for primary piRNA biogenesis [71].

The somatic siRNA pathway

Outside the germline, siRNAs represent the major defense line against TEs [72–75]. In *Drosophila*, TE-derived siRNAs originate from loci that are transcribed in both directions. Interestingly, siRNA source loci include the flamenco locus, which demonstrates that one locus can give rise to both piRNAs and siRNAs. The somatic siRNA pathway involves the Dicer paralog Dicer-2 and Ago2, the AGO-family member that predominantly binds TE-derived siRNAs and slices target transcripts [73,74,76,77]. Loading of Ago2 with siRNAs occurs in cytoplasmic foci termed D2 bodies and the observation that in R2D2 knockouts siRNAs are misloaded onto Ago1 suggests that R2D2 is responsible for correct sorting of siRNAs to Ago2 [78,79].

Besides the posttranscriptional silencing of TEs in the cytoplasm by endonucleolytically active Ago2, there is a body of evidence that links Ago2-siRNA complexes with epigenetic silencing of TEs. Dicer-2 mutants were found to exhibit dramatically decreased levels of H3K9 dimethylation (H3K9me2) at repeat associated loci [80]. Furthermore, aberrant H3K9me2/3 patterns and ectopic HP1 localization on chromosomes were observed in animals where nuclear siRNAs were artificially sequestered by the viral RNAi suppressor P19 [81]. There is also evidence that Dicer-2 and Ago2 associate with chromatin and that Ago2 functions in transcriptional repression of specific protein-coding genes [82,83] though it remains to be proven whether the same mechanisms are responsible for TE repression.

1.5.2. RNAi-based TE defense in mammals - The exception in the model

piRNA pathway in the male germline

First and yet still the most insights into mammalian RNAi pathways were obtained from experiments in the mouse model [84–87]. Like *Drosophila*, mice express three PIW family proteins named Miwi (Piwi-like 1 or Piwil1), Mili (Piwil2) and Miwi2 (Piwil4) that probably share a common ancestor with *Drosophila* Aub and arose from subsequent and successive gene duplication events [88]. A remarkable difference compared to the situation in flies and other vertebrates such as zebrafish is the fact that mouse PIWI proteins are expressed mainly during male gametogenesis while being dispensable in the female germline [89–91]. During spermatogenesis, Piwil4 is expressed in the early gonocyte stage, [23] whereas Piwil1 expression starts later in the pachytene stage in meiotic spermatocytes and persists in elongating spermatids [89]. In contrast, Piwil2 expression is fairly long lasting starting in primordial germ cells and persisting until the round spermatid stage (Figure 3(a)) [90]. Due to this dynamic expression pattern, a ping-pong cycle involving Piwil2 and Piwil4 in pre-pachytene germ cells is superseded by a ping-pong cycle involving Piwil2 and Piwil1 in subsequent developmental stages. Most important, the piRNAs that are expressed in the different time windows differ dramatically. piRNA clusters that are transcribed in pre-pachytene stages are mostly mono-directional and enriched for TE copies, as are the resulting piRNAs [23,92]. Later on, transcription from these loci ceases and the transcription factor A-Myb initiates transcription from bidirectional clusters resulting in a piRNA pool that is depleted for TE-related sequences [93].

During early stage spermatogenesis, the PIWI machinery localizes at two different types of nuage granules with Piwil2 associated with pi-bodies and Piwil4 associated with piP-bodies [94]. Both granules contain different additional factors and are often found in close proximity, suggesting these conglomerates to represent the sites of ongoing ping-pong processing [94]. Contrasting the situation in *Drosophila*, TE transcripts represent the source of primary piRNAs that initialize the ping-pong loop [23]. These sense-TE piRNAs are bound predominantly to Piwil2 while Piwil4 gets loaded with antisense-TE piRNAs during ping-pong amplification. In addition to its role as a ping-pong player, Piwil4 localizes to the nucleus and induces epigenetic silencing of TEs via DNA- and histone

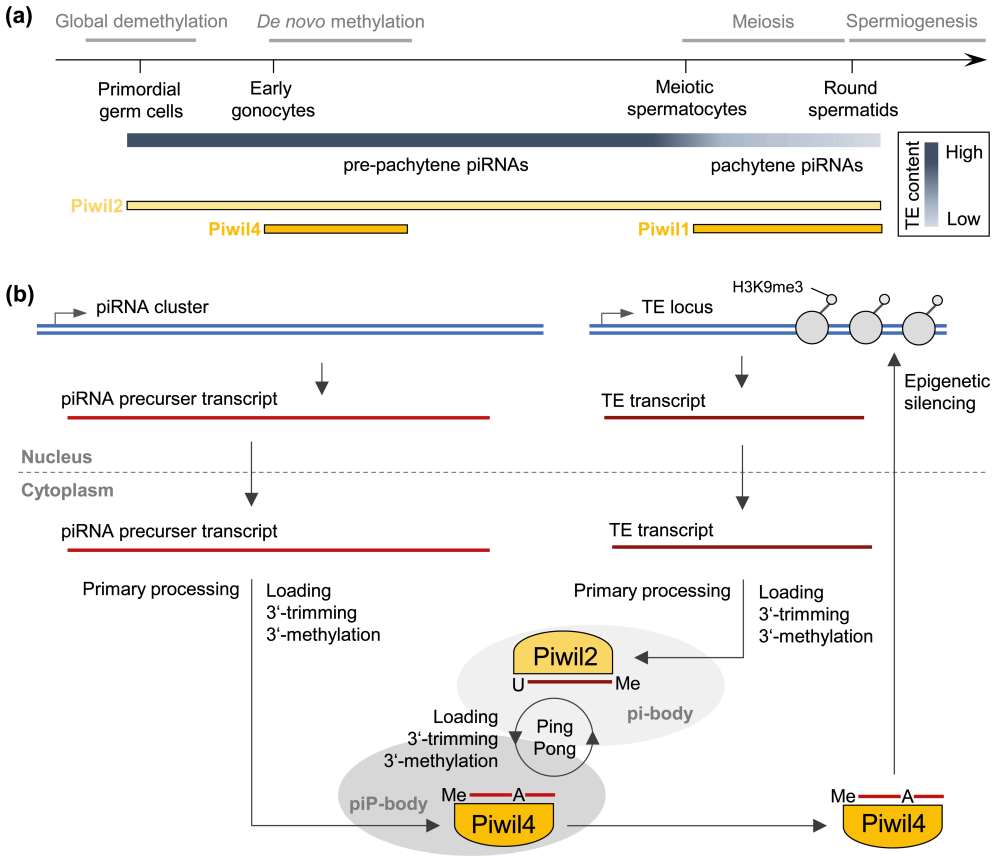


Figure 3 | Piwi pathway in mammals. (a) Expression of Piwi paralogs and piwi-interacting RNA (piRNA) pools during male gametogenesis. (b) Piwi-mediated TE silencing in the male germline during pre-pachytene stages.

(H3K9me3) methylation (Figure 3 (b)) [91,95–97]. In later stages of spermatogenesis, when epigenetic TE silencing is established, the PIWI machinery continues acting on the posttranscriptional level. At this point, primary antisense-TE piRNAs loaded onto Piwil1 and Piwil2 are sufficient for TE silencing which does no longer depend on ping-pong amplification [98,99].

piRNA pathway in the female germline

Based on findings from the mouse model, it was initially assumed that the PIWI machinery is not required for female gametogenesis in mammals, although PIWI proteins and piRNAs were also detected in mouse oocytes [23,100–103]. However, following studies revealed the special status of the mouse-rat lineage, which represent a unique evolutionary realization that is not representative for other mammals.

Contrasting the situation observed in mice, recent experiments revealed the presence of important piRNA pathway components in the female mammalian germline [104]. Interestingly, maturing bovine oocytes express the hitherto enigmatic Piwil3 that underwent pseudogenization on the lineage leading to mouse and rat. In addition, they express piRNAs that are enriched for TE-related sequences thus resembling pre-pachytene testis piRNAs. Together, these results suggest that Piwil3 is critically involved in TE repression during mammalian oogenesis.

An oocyte-specific siRNA pathway in mice

So, how do mice and rats protect their genomes from active TEs during oogenesis without having functional Piwil3? Flemr and colleagues showed that mice express an oocyte-specific Dicer isoform (Dicer^{O}, in contrast to the somatic isoform Dicer^{S}) whose expression is essential for mouse oocyte

development [105]. Dicer^{O} was found to produce more siRNAs from TE coding loci as compared with Dicer^{S} and abolishing Dicer^{O} expression resulted in sterility and increased levels of siRNA targets including TE transcripts from the MT family. Whether the suppression of MT transposons in the mouse female germline is a direct consequence of Dicer^{O} processing or in addition is reinforced by AGO family proteins that are loaded with the resulting MT-derived siRNAs and may induce further silencing on the transcriptional or posttranscriptional level, must be subject for further investigations. However, it seems reasonable to assume that the muridae-specific Dicer^{O} isoform, driven by a MT-C transposon insertion into the Dicer locus, represents an evolutionary adaptation that compensates the loss of Piwil3 – piRNA-dependent TE repression in oocytes.

1.5.3. RNAi-based TE defense in Caenorhabditis elegans - How piRNAs trigger siRNAs

As is the case with many other aspects of *Caenorhabditis (C.) elegans* biology, its RNAi pathways represent very unique evolutionary realizations and the role of the piRNA pathway in TE repression is rather limited. *C. elegans* piRNAs where initially described as 21U RNAs based on their length of 21 nt and the observed bias for 5' U [106]. 21U RNAs do not originate from Zuc-dependent processing of ssRNA. Instead 21U RNAs are encoded by separate genes that accumulate in two piRNA cluster-like regions on chromosome IV [106,107]. Each gene has a Polymerase II promoter that is recognized by Forkhead transcription factors which results in decapping and 2 nt 5' end trimming of the ~26 nt transcripts. The premature transcripts are then loaded onto the *C. elegans* PIWI protein PRG-1 and trimmed at their 3' end to form the functional PRG-1–21U-RNA complex [108–110]. Transcripts of the DNA transposon Tc3 represent the only TE-related PRG-1–21U-RNA target [106,111]. Other targets comprise protein-coding genes that are likewise derepressed in PRG-1 mutants [112]. Once the PRG-1–21U-RNA complex has caught its target, it can recruit an RdRP complex to synthesize a complementary RNA strand. The resulting dsRNA is further processed into secondary siRNAs named 22G-RNAs which are loaded onto worm-specific AGO proteins (WAGO) that can induce epigenetic silencing through repressive histone modifications such as H3K9me3 (Figure 4) [112–115].

While the PIWI/piRNA system triggers the 22G-RNA pathway that suppresses Tc3 elements, silencing of other TEs, such as Tc1, Tc4, Tc5, and Tc7, is independent of initiation by 21U RNAs. Instead, siRNAs that feed the RNAi machinery can be processed from read-through TE transcripts that comprise terminal inverted repeats and thus are capable of forming fold-back double stranded structures. These dsRNA molecules may serve as substrate for the RNaseIII-like enzyme DCR-1 that produces siRNAs that can initiate RNAi pathways directed against these elements [8,116].

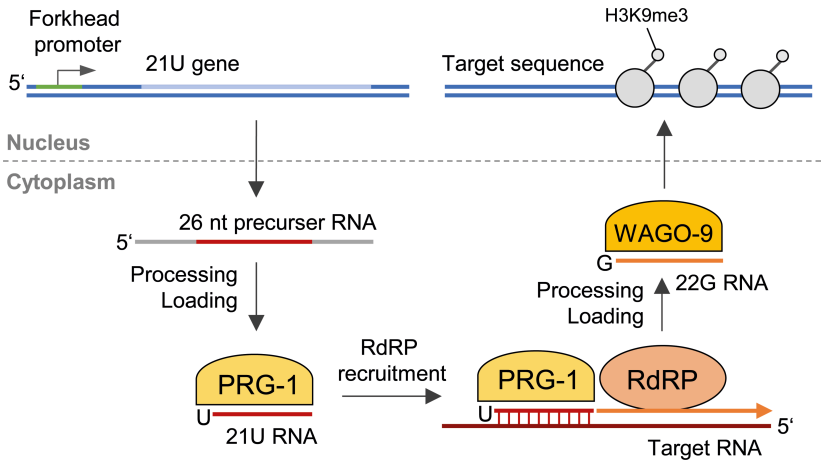


Figure 4 | 21U small RNA pathway in the germline of *Caenorhabditis elegans*.

1.5.4. RNAi-based TE defense in plants - A job for AGOs

In plants, active and evolutionary young TEs are mainly regulated by specialized siRNA pathways [117,118] in both somatic and germline cells, [119] while RNA-independent DNA methylation mechanisms maintain repressive methylation states of deeply silenced TEs [120]. Dynamic regulation of transposon expression primarily involves two major classes of siRNAs, which are distinguishable by their size profile. Typically, 21 – 22 nt siRNAs are involved in post-transcriptional gene silencing (PTGS), whereas 24 nt siRNAs mediate transcriptional gene silencing (TGS) in a pathway termed RNA-directed DNA methylation (RdDM) [121]. Besides different Dicer-like (DCL) and Argonaute proteins, as well as RNA-dependent RNA polymerases (RDR), which represent key players in both pathways, plants have developed the specialized DNA-dependent RNA polymerases Pol IV and Pol V, homologs of Pol II [122] that play crucial roles in RdDM [123].

Plant siRNAs in posttranscriptional silencing

The posttranscriptional silencing pathway of siRNAs in *Arabidopsis*, which functions in TE repression and virus resistance, starts with the synthesis of double stranded RNA (dsRNA) from single-stranded Pol II transcripts by RDR6 [124] and its cofactor SGS3 [125]. The processing of these dsRNAs by DCL2 and DCL4, which mostly act redundantly but yet hierarchically, results in the production of 22 and 21 nt siRNAs, respectively [126]. These siRNAs are subsequently methylated at the 2'-OH of the 3'-terminal nucleotide by HEN1, increasing their stability [127,128]. Finally, AGO1 is loaded with 21 – 22 nt siRNAs, which guide the slicer active Argonaute to its targets to direct RNA degradation [121,129,130].

The RdDM pathway

The canonical RdDM pathway in plants (Figure 5) is initiated by the recruitment of Pol IV to its genomic target loci, which is required for the production of the vast majority of siRNAs in *Arabidopsis* [131]. The chromatin interacting protein SHH1, which binds to methylated H3K9 and unmethylated H3K4, guides Pol IV to a large subset of the most active RdDM-associated siRNA producing loci, sometimes referred to as siRNA clusters [132]. Single-stranded Pol IV transcripts are copied by RDR2

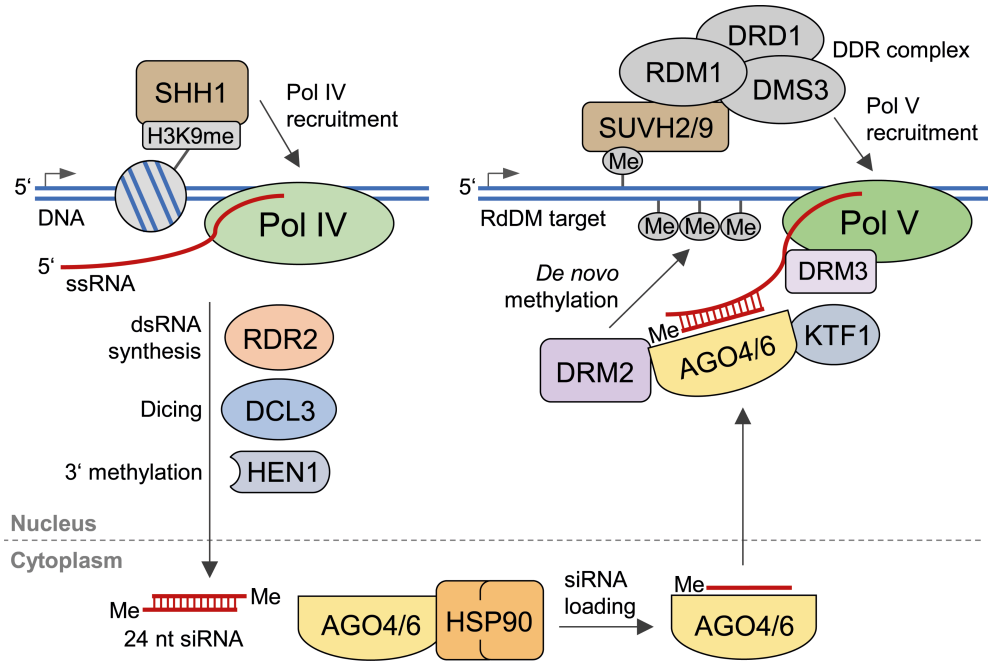


Figure 5 | Canonical RNA-directed DNA methylation (RdDM) in plants mediated by 24 nt short interfering RNAs (siRNAs).

to produce dsRNAs, [133] which are processed into 24 nt siRNA duplexes by DCL3, methylated by HEN1 [127,128] and exported to the cytoplasm, where the loading of one strand onto AGO4 [134,135] is facilitated by HSP90 [136]. Besides AGO4, the paralogous AGO6 seems to play a non-redundant role in RdDM and acts in parallel to AGO4 [137,138], while AGO9 specifically acts in reproductive tissues [139]. The assembled AGO–siRNA complex is reimported to the nucleus [136], where it is guided to genomic target loci through base pairing of associated siRNAs with nascent Pol V transcripts [140], assisted by the adaptor protein KTF1 that binds both Pol V scaffold transcripts and AGO4 [141].

The recruitment of Pol V to RdDM target loci is enabled by the SU(VAR)3 – 9 homologs SUVH2 and SUVH9, which bind to pre-existing methylated DNA and facilitate the chromatin interaction of Pol V [142,143] by associating with the DDR complex, comprising the factors DRD1, DMS3, and RDM1 [117,144] The binding of the AGO-siRNA complex to nascent Pol V transcripts leads to the recruitment of the plant Dnmt3 methyltransferase ortholog DRM2 to establish de novo DNA methylation at target loci, resulting in transcriptional silencing [145]. Another ortholog, DRM3, presumably promotes Pol V transcriptional elongation or assists in the stabilization of Pol V transcripts [146].

Epigenetic repression of plant transposons

Active TEs are epigenetically silenced by RdDM [117,118], which mediates DNA methylation by DRM2 at CG, CHG and CHH sites (H=A, T, or C) [145]. Both Pol IV and Pol V, which produce siRNA precursors and RdDM targets, respectively, target genomic loci that include high-copy repeats and transposons. In *Arabidopsis*, Pol IV transcripts were shown to originate primarily from regions where transposons and other repetitive sequences cluster, [131] while Pol V targets mainly represent promoters and evolutionarily young transposons [117]. Furthermore, Pol V-dependent siRNA-generating loci are associated predominantly with short repetitive sequences in intergenic regions, which are enriched for SINE repeats [147].

Besides RdDM, which establishes de novo methylation, TEs are kept under control by pathways that maintain DNA methylation, as well as histone methylation pathways, mainly targeting H3K9 [148]. The plant homolog of mammalian Dnmt1 MET1, which targets CG sites [149], as well as the plant-specific CMT3 and CMT2, specialized for CHG and CHH site methylation [150] are crucial for sustaining silencing states of TEs. The chromatin remodeler DDM1, being similarly essential for TE silencing, enables DNA methyltransferases to access heterochromatin embedded TEs [120,151,152]. DDM1-dependent pathways and RdDM together mediate nearly all TE methylation and cooperate to inhibit transposition in *Arabidopsis* [120].

When transposons awake

The emergence of a new transposon represents a special situation, as its replication proceeds through Pol II transcripts, which the canonical RdDM pathway normally cannot target directly. The study of a de novo genome invasion by the single-copy LTR retroelement Évadé (EVD), using inbred lineages of hybrid *Arabidopsis* epigenomes, monitored over multiple generations, demonstrated how plants employ their defense mechanisms against a new TE offensive [118]. It showed that an increasing copy number concurs with an accumulation of RDR6- and DCL2/4-dependent 21-22 nt EVD-derived siRNAs, indicating that PTGS forms the first line of defense during a novel TE propagation event. The Gag nucleocapsid protein of EVD protects it against PTGS, but after reaching a fixed threshold of about 40 copies EVD expression is brought to a halt. This coincides with a loss of corresponding 21 – 22 nt siRNAs and an initiated production of AGO4-associated 24 nt siRNA accompanied by increased DNA

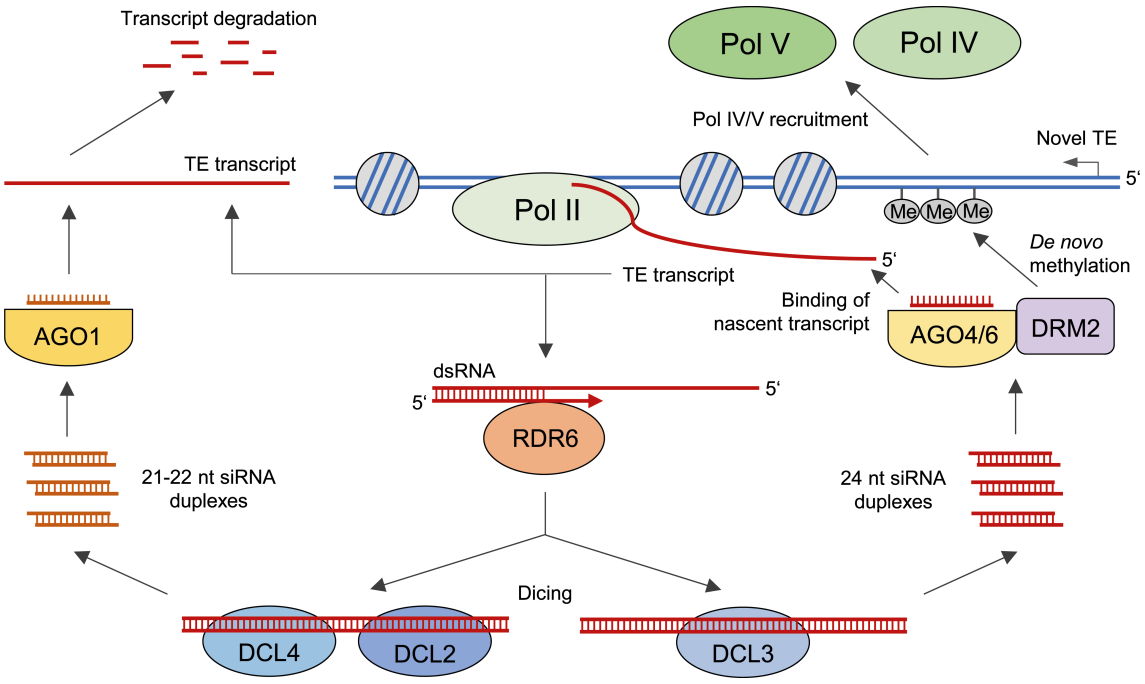


Figure 6 | Transition from PTGS to TGS in plants during a de novo genome invasion by an active TE.

methylation of EVD copies, suggesting transcriptional silencing by the RdDM pathway. In support, another study under similar conditions showed that EVD silencing depends on Pol IV and Pol V [153], which are possibly recruited after de novo methylation of EVD DNA [118].

The transition from PTGS to TGS (Figure 6) appears to be achieved through the processing of RDR6-generated dsRNAs by DCL3 to produce 24 nt siRNAs, which might be enabled by a large increase of TE transcripts that leads to a saturation of the DCL2 and DCL4 machineries of the regular PTGS pathway [118]. This RDR6-dependent RdDM pathway represents a form of noncanonical RdDM that likely functions to recognize Pol II transposon transcripts to trigger epigenetic silencing of novel and active TEs [154]. Indeed, it was shown that Pol II produces scaffold transcripts that recruit AGO4-bound siRNAs through physical interaction to direct DNA methylation at low-copy intergenic loci. Moreover, Pol II transcription recruits Pol IV and Pol V to exert siRNA-mediated RdDM at homologous loci [155], which might be the final step in the transition from PTGS via RDR6-RdDM to canonical Pol IV-dependent RdDM (Figure 6). Alternatively, Pol II-derived 21-22 nt siRNAs bound by AGO6 can initiate RDR6-dependent methylation of novel Pol II-transcribed TEs [137,154], representing another entry point for Pol V, which requires some pre-existing methylation for its recruitment [142].

Interestingly, a recent study additionally showed that some miRNAs can target epigenetically reactivated TEs and trigger RDR6-mediated 21 nt siRNA production in *Arabidopsis* DDM1 mutants [34], resembling the activity of miRNA-dependent 21 nt trans-acting siRNAs (ta-siRNAs), which regulate genes posttranscriptionally in land plants [156,157]. However, an important part in our understanding is still missing, as it is unclear how novel Pol II-transcribed TEs that are unknown to the host plant are distinguished from non-TE transcripts and recognized in a homology-independent manner.

Defending the plant germline

Tight regulation of TEs is especially critical in the germline, where transposition events are carried over to the next generation. While animals have developed the Piwi/piRNA pathway, plants have established specialized mechanisms involving siRNAs and Argonaute proteins to secure genome integrity of their gametes.

Mammals globally erase DNA methylation marks in their germline by epigenetic reprogramming, but in plants DNA methylation is largely retained through sexual reproduction [158,159]. In contrast, in *Arabidopsis* the companion cells of both egg and sperm, central cell and vegetative nucleus, respectively, undergo active DNA demethylation (Figure 7) mediated by the DNA glycosylase DEMETER (DME), which preferentially targets euchromatic TEs [160]. In addition, in the pollen vegetative cell the TE silencing DDM1, as well as many TE targeting siRNAs are downregulated [119]. Consequently, TEs are reactivated and mobilized in pollen, but only in the vegetative cell, leading to the production of 21 nt transposon-derived siRNAs, which accumulate in both vegetative and sperm cells, implying that siRNAs are transferred from companion cell to sperm cells (Figure 7(a)). Supporting this idea, lack of DME in vegetative cells causes reduced RdDM of TEs in sperm [160].

Similarly, in the female gametophyte it was shown that global demethylation of the central cell DNA leads to transposon reactivation and siRNA accumulation in the central cell (Figure 7(b)) and subsequently in the maternal endosperm genome (Figure 7(c)) [161,162]. Further, AGO9, which is expressed in companion cells but not in gametes, has been shown to interact with 24 nt TE-derived siRNAs and being necessary for transposon silencing in the egg cell before fertilization [139]. Finally, endosperm demethylation is coupled with extensive local hypermethylation of siRNA-targeted sequences in the endosperm and 24 nt siRNA-guided de novo methylation of embryo TEs [159,162].

It is therefore believed that companion cells as well as the endosperm, which do not contribute genetic material for the next generation, sacrifice their genome integrity by epigenetic reprogramming to reveal intact TEs and produce corresponding siRNAs, which are transferred to gametes and to the embryo to reinforce transposon methylation in the germline of plants [119,160,162].

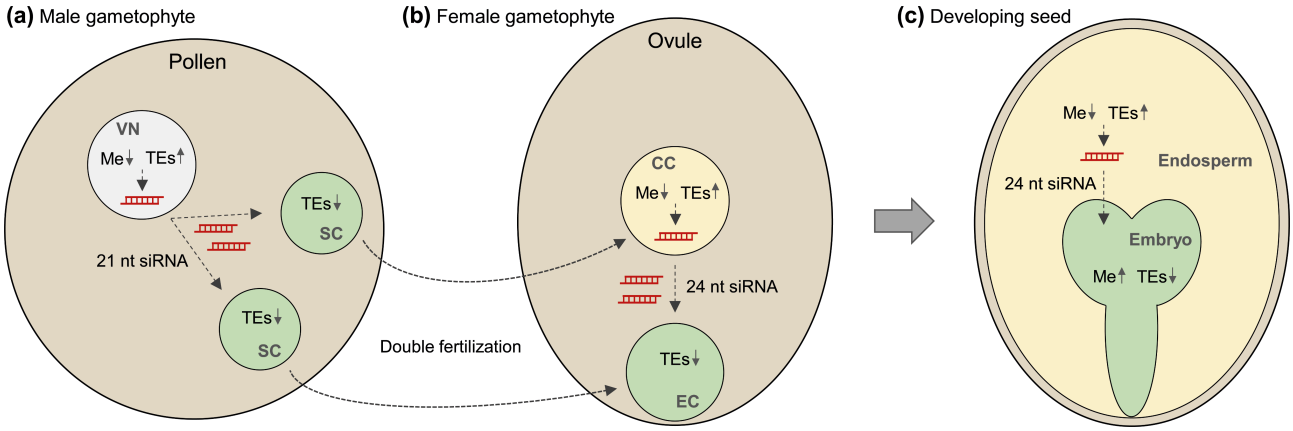


Figure 7 | Transposon silencing in the plant germline. Global DNA demethylation in companion cells and in the endosperm leads to reactivation of TEs and production of short interfering RNAs (siRNAs) that are transferred into gametes and embryo to reinforce TE silencing in germline and developing embryo. Abbreviations: VN, vegetative nucleus; SC, sperm cell; CC, central cell; Me, DNA methylation.

1.5.5 RNAi-based TE defense in fungi - Some can live without

The canonical siRNA pathway

Though a functional RNAi machinery has been lost in some fungi including the model organism *Saccharomyces (S.) cerevisiae*, RNAi components including Dicer, Argonaute, and RdRP proteins are present in many other fungal species [163–166]. In *Saccharomyces castellii* and *Candida albicans* noncanonical Dicer proteins process dsRNA to produce 22 – 23 nt siRNAs that are strongly enriched for LTR- (Ty), LINE-like- (Zorro) and subtelomeric repeat (Y0) sequences [166]. These siRNAs were shown to associate with Ago1 in *S. castellii* and loss of either Dicer or Ago1 results in increased levels

of Y0 and Ty transcripts. dsRNA that serves as Dicer substrate results from bidirectional transcription of one locus yielding paired sense-antisense transcripts or alternatively from partially overlapping transcripts produced from opposite strands. Despite the elevated level of Y0 transcripts and instability of introduced plasmids, Dicer and Ago1 mutants do not show defects in growth, mating, sporulation, or chromosome stability which might reflect the absence of TEs that are capable of active transposition in the genome of *S. castellii*. It remains a mystery why a functional RNAi machinery was retained during evolution in a species without active TEs while being lost in, for example, *S. cerevisiae* where TEs show signs of recent activity. One explanation might be that RNAi components underwent pseudogenization in *S. cerevisae* after TE activity dropped off and that active TEs were subsequently reintroduced by horizontal transfer [167].

The nuclear siRNA pathway

A nuclear processing pathway that can silence TEs by heterochromatin formation has recently been [168] described in *Saccharomyces pombe*. Here, the Mtr4-like protein Mlt1 interacts with Nrl1, which associates with factors involved in pre-mRNA splicing. The Mlt1-Nrl1 complex targets transcripts with cryptic introns resulting in the formation of heterochromatin domains at loci encoding retrotransposons. Deletion of cryptic introns resulted in abolished siRNA production and H3K9me at the according locus. How exactly the RNAi machinery is attracted to target specific transcripts is yet unclear but it has been proposed, that the spliceosome itself recruits RNAi factors. This is supported by the observation, that the splicing machinery interacts with components of the RdRP complex, which is involved in the production of siRNAs and H3K9me at heterochromatin domains [169,170]. A similar mechanism can be found in the yeast *Cryptococcus neoformans* [171] where a so-called SCANR (Spliceosome-Coupled And Nuclear RNAi) complex is essential for biogenesis of Ago1 bound TE-related siRNAs from transcripts that exhibit suboptimal introns. Loss of the RdRP Rdp1 chokes siRNA production, which indicates that selected mRNAs serve as a template for dsRNA production, thus triggering the RNAi machinery.

1.5.6. RNAi-based TE defense in ciliates - Extinguishing the unwanted

So far we have discussed examples that illustrate how TEs can be suppressed either on the transcriptional level by repressive DNA- or Histone modifications that prevent TE transcription, or on the posttranscriptional level by targeted degradation of TE transcripts. In ciliates, we can observe an even more radical mechanism: Repetitive sequences are recognized by small RNAs and completely erased from the genome (Figure 8).

Most ciliates display nuclear dualism with a somatic macronucleus separated from a germline micronucleus [172]. Comparing the germline genome with the somatic genome provides the basis for the selection of junk DNA. During sexual reproduction of *Tetrahymena*, the micronucleus undergoes mitosis followed by meiosis and cross-fertilization to generate zygotic nuclei. While the old macronucleus is degraded, the zygotic nuclei further divide and differentiate to new macro- and micronuclei [173]. During this process it comes to both DNA rearrangement and DNA deletion and more than 6000 genomic sites, together making up ~15% of the *Tetrahymena* genome, are excised from the newly developing macronucleus. These sites represent noncoding single-copy as well as repetitive, transposon-like sequences such as Tlr-1 and Tel1 [174,175]. Before DNA elimination occurs, bidirectional transcription of nongenic loci in the micronucleus was observed and it was suggested that the resulting dsRNA may feed a RNAi pathway that is responsible for the selection of genomic sites to be erased [176]. Indeed, DNA elimination depends on a class of ~26-31 nt small siRNA-like RNAs termed scan (scn-) RNAs that are produced from dsRNA precursors by the Dicer-like enzyme Dcl1

inside the micronucleus. scnRNAs are then exported to the cytoplasm and loaded onto the *Tetrahymena* PIWI homolog Twi1 [174,177–179]. The Twi1-scnRNA complexes localize to the parental macronucleus, where they undergo a yet mysterious selection process that selects complexes that target sites that are present only in the micronucleus this way determining specific sites for elimination [177]. The selected Twi1-scnRNA complexes now translocate from the parental to the newly developing macronucleus where they interact with target chromatin, presumably via nascent transcripts [180]. Latest observations suggest that the imported early scnRNAs can target loci in trans and induce biogenesis of so-called late scnRNAs from these loci resulting in robust DNA elimination of TE-related sequences [181].

It was shown that DNA elimination depends on H3K9me and H3K27me marks as the disruption of the histone methyltransferase Ezl1p that catalyzes these modifications inhibits DNA elimination [182,183]. The greatly reduced H3K9me accumulation in Dcl1 and Twi1 mutants provides strong evidence that the Twi1-scnRNA complex acts upstream of, and is necessary for, heterochromatin formation [178,179,182]. These results indicate that Twi1-scnRNA complexes decoy the methylation machinery that finally marks targets for DNA elimination though the exact mechanisms are unknown yet also little is known about the final process of DNA excision that is supposed to depend on a putative protein complex that comprises a domesticated PiggyBac transposase-like protein and recognizes the epigenetically marked targets [184,185].

A similar mechanism that distinguishes between good and bad DNA in the germline genome acts in *Oxytricha*. But contrasting the situation in *Tetrahymena*, the RNAi system that acts during sexual reproduction in *Oxytricha* marks DNA that is retained in, and not erased from, the somatic genome of the upcoming generation [186].

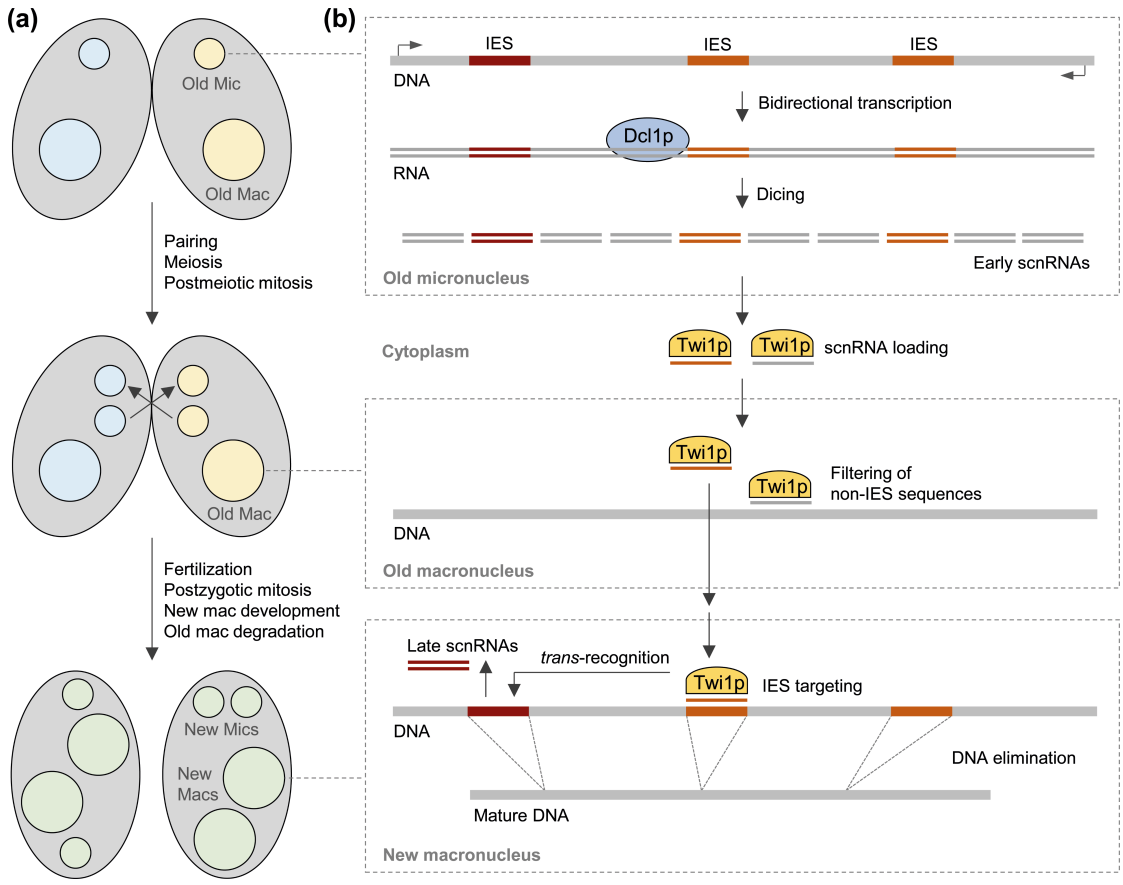


Figure 8 | DNA elimination in *Tetrahymena*. (a) Stages of sexual reproduction of *Tetrahymena*. (b) Twi1p and scnRNA-mediated DNA elimination. Abbreviations: Mic, Micronucleus; Mac, Macronucleus; IES, internal eliminated sequences.

1.6. Conclusion

Most organisms are faced with genomic parasites that reside and proliferate in their genomes. Immobilizing these enemies inside is essential to maintain genome integrity over generations. Molecular defense mechanisms emerged early in evolution and were subject to constant adaptation along phylogenetic branches. Although the variety of evolutionary realizations unrolled in this review is already impressive, we must point out that the addressed organisms represent a rather small subsection in the tree of life. While we focused on eukaryotic RNAi pathways, different mechanisms of immunity from genome invaders and TEs exist in bacteria and archaea. Most notably, we want to mention the CRISPR-Cas system that came to fame in the past years owing to its utilization for genome manipulation in genetic engineering [187,189]. Another example is the Hfq/antisense-RNA mediated down regulation of transposase expression in *Escherichia coli* [190] and the list could be continued. The curious reader will find some recommendations for further reading below. Certainly, the discovery of yet unknown small RNA pathways will not be long in coming, considering (1) the enormous evolutionary plasticity of RNA-based defense mechanisms and (2) the ever-growing number of species that are subject to in-depth investigation including whole-genome- and RNA-seq analysis.

1.7. Declarations

Acknowledgement

This work was supported by grants from the International PhD Program (IPP) coordinated by the Institute of Molecular Biology IMB, Mainz, Germany and the intramural funding program MAIFOR (Johannes Gutenberg University Medical Center). We further acknowledge the members of the 'Forschungsschwerpunkt' (research focus) GeneRED (Gene Regulation in Evolution and Development) for valuable discussion.

Further Reading

- Watanabe T, Nozawa T, Aikawa C, Amano A, Maruyama F, Nakagawa I. CRISPR regulation of intraspecies diversification by limiting IS transposition and intercellular recombination. Genome Biol Evol 2013, 5:1099–1114.
- Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol 2014, 12:36.
- Hickman AB, Dyda F. CRISPR-Cas immunity and mobile DNA: a new superfamily of DNA transposons encoding a Cas1 endonuclease. Mob DNA 2014, 5:23.
- Ross JA, Trussler RS, Black MD, McLellan CR, Haniford DB. Tn5 transposition in Escherichia coli is repressed by Hfq and activated by over-expression of the small non-coding RNA SgrS. Mob DNA 2014, 5:27.
- Ross JA, Ellis MJ, Hossain S, Haniford DB. Hfq restructures RNA-IN and RNA-OUT and facilitates antisense pairing in the Tn10/IS10 system. RNA 2013, 19:670–684.
- Phillips P, Progulske-Fox A, Grieshaber S, Grieshaber N. Expression of porphyromonas gingivalis small RNA in response to hemin availability identified using microarray and RNA-seq analysis. FEMS Microbiol Lett 2014, 351:202–208.
- Waters JL, Salyers AA. The small RNA RteR inhibits transfer of the bacteroides conjugative transposon CTnDOT. J Bacteriol 2012, 194:5228–5236.
- Trigui H, Dudyk P, Sum J, Shuman HA, Faucher SP. Analysis of the transcriptome of Legionella pneumophila hfq mutant reveals a new mobile genetic element. Microbiology 2013, 159:1649–1660.

1.8. References

1. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 2007, 8:973– 982.
2. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems—a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol 2015, 86:90–109.

3. Feschotte C. Opinion: transposable elements and the evolution of regulatory networks. Nat Rev Genet 2008, 9:397–405.
4. Biémont C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. Genetics 2010, 186:1085 – 1093.
5. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet 2014, 15:423 – 437.
6. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet 2009, 10:94–108.
7. Hammond SM, Bernstein E, Beach D, Hannon GJ. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature 2000, 404:293–296.
8. Bernstein E, Caudy AA, Hammond SM, Hannon GJ. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 2001, 409:363 – 366.
9. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004, 116:281 – 297.
10. Sugiyama T, Cam H, Verdel A, Moazed D, Grewal SI. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. Proc Natl Acad Sci USA 2005, 102:152 – 157.
11. Zong J, Yao X, Yin J, Zhang D, Ma H. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: duplications and possible losses before and after the divergence of major eukaryotic groups. Gene 2009, 447:29–39.
12. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell 2009, 136:642–655.
13. Tomari Y, Matranga C, Haley B, Martinez N, Zamore PD. A protein sensor for siRNA asymmetry. Science 2004, 306:1377–1380.
14. Leuschner PJ, Ameres SL, Kueng S, Martinez J. Cleavage of the siRNA passenger strand during RISC assembly in human cells. EMBO Rep 2006, 7:314–320.
15. Meister G. Argonaute proteins: functional insights and emerging roles. Nat Rev Genet 2013, 14:447–459.
16. Ipsaro JJ, Haase AD, Knott SR, Joshua-Tor L, Hannon GJ. The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. Nature 2012, 491:279–283.
17. Nishimasu H, Ishizu H, Saito K, Fukuhara S, Kamatani MK, Bonnefond L, Matsumoto N, Nishizawa T, Nakanaga K, Aoki J, et al. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. Nature 2012, 491:284–287.
18. Cora E, Pandey RR, Xiol J, Taylor J, Sachidanandam R, McCarthy AA, Pillai RS. The MID-PIWI module of Piwi proteins specifies nucleotide- and strand-biases of piRNAs. RNA 2014, 20:773 – 781.
19. Kirino Y, Mourelatos Z. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. RNA 2007, 13:1397 – 1401.
20. Kawaoka S, Izumi N, Katsuma S, Tomari Y. 3' end formation of PIWI-interacting RNAs in vitro. Mol Cell 2011, 43:1015 – 1022.
21. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. Discrete small RNA-generating loci as master regulators of TE activity in Drosophila. Cell 2007, 128:1089 – 1103.
22. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science 2007, 315:1587–1590.
23. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell 2008, 31:785–799.
24. Parker JS, Roe SM, Barford D. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. EMBO J 2004, 23:4727–4737.
25. Wang W, Yoshikawa M, Han BW, Izumi N, Tomari Y, Weng Z, Zamore PD. The initial uridine of primary piRNAs does not create the tenth adenine that is the hallmark of secondary piRNAs. Mol Cell 2014, 56:708 – 716.
26. Shalgi R, Pilpel Y, Oren M. Repression of transposable-elements—a microRNA anti-cancer defense mechanism? Trends Genet 2010, 26:253–259.
27. Smalheiser NR, Torvik VI. Mammalian microRNAs derived from genomic repeats. Trends Genet 2005, 21:322–326.
28. Piriyapongsa J, Jordan IK. A family of human micro-RNA genes from miniature inverted-repeat transposable elements. PLoS One 2007, 2:e203.
29. Piriyapongsa J, Mariño-Ramírez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. Genetics 2007, 176:1323 – 1337.
30. Devor EJ, Peek AS, Lanier W, Samollow PB. Marsupial-specific microRNAs evolved from marsupial-specific transposable elements. Gene 2009, 448:187– 191.
31. Yuan Z, Sun X, Liu H, Xie J. MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. PLoS One 2011, 6:e17666.
32. Rosenkranz D, Rudloff S, Bastuck K, Ketting RF, Zischler H. Tupaia small RNAs provide insights into function and evolution of RNAi-based transposon defense in mammals. RNA 2015, 21:911 – 922.
33. Lehnert S, Van Loo P, Thilakarathne PJ, Marynen P, Verbeke G, Schuit FC. Evidence for co-evolution between human MicroRNAs and Alu-repeats. PLoS One 2009, 4:e4456.
34. Creasey KM, Zhai J, Borges F, Van Ex F, Regulski M, Meyers BC, Martienssen RA. miRNAs trigger wide-spread epigenetically activated siRNAs from transposons in Arabidopsis. Nature 2014, 508:411 – 415.

35. Mugat B, Akkouche A, Serrano V, Armenise C, Li B, Brun C, Fulga TA, Van Vactor D, Pélisson A, Chambeyron S. MicroRNA-dependent transcriptional silencing of transposable elements in Drosophila follicle cells. PLoS Genet 2015, 11:e1005194.
36. Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol 2008, 23:578–587.
37. Wang DY, Kumar S, Hedges SB. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc Biol Sci 1999, 266:163–171.
38. Simpson AG, Roger AJ. The real 'kingdoms' of eukaryotes. Curr Biol 2004, 14:R693 – R696.
39. Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell 2009, 137:522–535.
40. Homolka D, Pandey RR, Goriaux C, Brasset E, Vaury C, Sachidanandam R, Fauvarque MO, Pillai RS. PIWI slicing and RNA elements in precursors instruct directional primary piRNA biogenesis. Cell Rep 2015, 12:418–428.
41. Rozhkov NV, Hammell M, Hannon GJ. Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev 2013, 27:400–412.
42. Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. Genes Dev 2010, 24:2493– 2498.
43. Olivieri D, Sykora MM, Sachidanandam R, Mechtler K, Brennecke J. An in vivo RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in Drosophila. EMBO J 2010, 29:3301– 3317.
44. Olivieri D, Senti KA, Subramanian S, Sachidanandam R, Brennecke J. The cochaperone shutdown defines a group of biogenesis factors essential for all piRNA populations in Drosophila. Mol Cell 2012, 47:954– 969.
45. Sarot E, Payen-Groschêne G, Bucheton A, Pélisson A. Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. Genetics 2004, 166:1313 – 1321.
46. Kalmykova AI, Klenov MS, Gvozdev VA. Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. Nucleic Acids Res 2005, 33:2052–2059.
47. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A distinct small RNA pathway silences selfish genetic elements in the germline. Science 2006, 313:320–324.
48. Darricarrère N, Liu N, Watanabe T, Lin H. Function of Piwi, a nuclear Piwi/Argonaute protein, is independent of its slicer activity. Proc Natl Acad Sci USA 2013, 110:1297–1302.
49. Pal-Bhadra M, Leibovitch BA, Gandhi SG, Chikka MR, Bhadra U, Birchler JA, Elgin SC. Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. Science 2004, 303:669 – 672.
50. Brower-Toland B, Findley SD, Jiang L, Liu L, Yin H, Dus M, Zhou P, Elgin SC, Lin H. Drosophila PIWI associates with chromatin and interacts directly with HP1a. Genes Dev 2007, 21:2300 – 2311.
51. Yin H, Lin H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. Nature 2007, 450:304 – 308.
52. Gu T, Elgin SC. Maternal depletion of Piwi, a component of the RNAi system, impacts heterochromatin formation in Drosophila. PLoS Genet 2013, 9:e1003780.
53. Wang SH, Elgin SC. Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. Proc Natl Acad Sci USA 2011, 108:21164– 21169.
54. Sienski G, Dönertas D, Brennecke J. Transcriptional silencing of TEs by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell 2012, 151:964–980.
55. Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H. A major epigenetic programming mechanism guided by piRNAs. Dev Cell 2013, 24:502 – 516.
56. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev 2013, 27:390 – 399.
57. Cox DN, Chao A, Lin H. piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development 2000, 127:503–514.
58. Harris AN, Macdonald PM. Aubergine encodes a Drosophila polar granule component required for pole cell formation and related to eIF2C. Development 2001, 128:2823 – 2832.
59. Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev 2006, 20:2214 – 2222.
60. Vagin VV, Klenov MS, Kalmykova AI, Stolyarenko AD, Kotelnikov RN, Gvozdev VA. The RNA interference proteins and vasa locus are involved in the silencing of retrotransposons in the female germline of Drosophila melanogaster. RNA Biol 2004, 1:54–58.
61. Savitsky M, Kwon D, Georgiev P, Kalmykova A, Gvozdev V. Telomere elongation is under the control of the RNAi-based mechanism in the Drosophila germline. Genes Dev 2006, 20:345 – 354.
62. Lim AK, Kai T. Unique germ-line organelle, nuage, functions to repress selfish genetic elements in Drosophila melanogaster. Proc Natl Acad Sci USA 2007, 104:6714–6719.
63. Pane A, Wehr K, Schüpbach T. zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. Dev Cell 2007, 12:851–862.

64. Patil VS, Kai T. Repression of retroelements in Drosophila germline via piRNA pathway by the Tudor domain protein Tejas. Curr Biol 2010, 20:724–730.
65. Anand A, Kai T. The tudor domain protein kumo is required to assemble the nuage and to generate germ-line piRNAs in Drosophila. EMBO J 2012, 31:870– 882.
66. Zhang F, Wang J, Xu J, Zhang Z, Koppetsch BS, Schultz N, Vreven T, Meignin C, Davis I, Zamore PD, et al. UAP56 couples piRNA clusters to the perinuclear transposon silencing machinery. Cell 2012, 151:871– 884.
67. Xiol J, Spinelli P, Laussmann MA, Homolka D, Yang Z, Cora E, Couté Y, Conn S, Kadlec J, Sachidanandam R, et al. RNA clamping by Vasa assembles a piRNA amplifier complex on transposon transcripts Cell 2014, 157:1698 – 1711.
68. Sato K, Iwasaki YW, Shibuya A, Carninci P, Tsuchizawa Y, Ishizu H, Siomi MC, Siomi H. Krimper enforces an antisense bias on piRNA pools by binding AGO3 in the Drosophila germline. Mol Cell 2015, 59:553 – 563.
69. Han BW, Wang W, Li C, Weng Z, Zamore PD. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science 2015, 348:817 – 821.
70. Mohn F, Handler D, Brennecke J. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science 2015, 348:812 – 817.
71. Senti KA, Jurczak D, Sachidanandam R, Brennecke J. piRNA-guided slicing of transposon transcripts enforces their transcriptional silencing via specifying the nuclear piRNA repertoire. Genes Dev 2015, 29:1747–1762.
72. Pélisson A, Sarot E, Payen-Groschêne G, Bucheton A. A novel repeat-associated small interfering RNA-mediated silencing pathway downregulates complementary sense gypsy transcripts in somatic cells of the Drosophila ovary. J Virol 2007, 81:1951 – 1960.
73. Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler EL, Zapp ML, Weng Z, et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science 2008, 320:1077–1081.
74. Chung WJ, Okamura K, Martin R, Lai EC. Endogenous RNA interference provides a somatic defense against Drosophila transposons. Curr Biol 2008, 18:795–802.
75. Lau NC, Robine N, Martin R, Chung WJ, Niki Y, Berezikov E, Lai EC. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome Res 2009, 19:1776 – 1785.
76. Kawamura Y, Saito K, Kin T, Ono Y, Asai K, Sunohara T, Okada TN, Siomi MC, Siomi H. Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature 2008, 453:793 – 797.
77. Li W, Prazak L, Chatterjee N, Grüninger S, Krug L, Theodorou D, Dubnau J. Activation of transposable elements during aging and neuronal decline in Drosophila. Nat Neurosci 2013, 16:529 – 531.
78. Liu X, Jiang F, Kalidas S, Smith D, Liu Q. Dicer-2 and R2D2 coordinately bind siRNA to promote assembly of the siRISC complexes. RNA 2006, 12:1514 – 1520.
79. Nishida KM, Miyoshi K, Ogino A, Miyoshi T, Siomi H, Siomi MC. Roles of R2D2, a cytoplasmic D2 body component, in the endogenous siRNA pathway in Drosophila. Mol Cell 2013, 49:680–691.
80. Peng JC, Karpen GH. H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. Nat Cell Biol 2007, 9:25 – 35.
81. Fagegaltier D, Bougé AL, Berry B, Poisot E, Sismeiro O, Coppée JY, Théodore L, Voinnet O, Antoniewski C. The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proc Natl Acad Sci USA 2009, 106:21258 – 21263.
82. Cernilogar FM, Onorati MC, Kothe GO, Burroughs AM, Parsi KM, Breiling A, Lo Sardo F, Saxena A, Miyoshi K, Siomi H, et al. Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. Nature 2011, 480:391–395.
83. Taliaferro JM, Aspden JL, Bradley T, Marwha D, Blanchette M, Rio DC. Two new and distinct roles for Drosophila Argonaute-2 in the nucleus: alternative pre-mRNA splicing and transcriptional repression. Genes Dev 2013, 27:378 – 389.
84. Grivna ST, Beyret E, Wang Z, Lin H. A novel class of small RNAs in mouse spermatogenic cells. Genes Dev 2006, 20:1709–1714.
85. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, et al. Nature 2006, 442:203–207.
86. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 2006, 442:199 – 202.
87. Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, Imai H. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev 2006, 20:1732–1743.
88. Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Hériché JK, Hu Y, Kristiansen K, Li R, et al. TreeFam: 2008 Update. Nucleic Acids Res 2008, 36:D735– D740.
89. Deng W, Lin H. miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell 2002, 2:819 – 830.
90. Kuramochi-Miyagawa S, Kimura T, Ijiri TW, Isobe T, Asada N, Fujita Y, Ikawa M, Iwai N, Okabe M, Deng W, et al. Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Development 2004, 131:839 – 849.

91. Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ. MIWI2 is essential for spermatogenesis and repression of TEs in the mouse male germline. Dev Cell 2007, 12:503 – 514.
92. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. Developmentally regulated piRNA clusters implicate MILI in TE control. Science 2007, 316:744–747.
93. Li XZ, Roy CK, Dong X, Bolcun-Filas E, Wang J, Han BW, Xu J, Moore MJ, Schimenti JC, Weng Z, et al. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell 2013, 50:67 – 81.
94. Aravin AA, van der Heijden GW, Castañeda J, Vagin VV, Hannon GJ, Bortvin A. Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. PLoS Genet 2009, 5:e1000764.
95. Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, et al. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev 2008, 22:908 – 917.
96. De Fazio S, Bartonicek N, Di Giacomo M, Abreu-Goodger C, Sankar A, Funaya C, Antony C, Moreira PN, Enright AJ, O'Carroll D. The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature 2011, 480:259 – 263.
97. Pezic D, Manakov SA, Sachidanandam R, Aravin AA. piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. Genes Dev 2014, 28:1410 – 1428.
98. Reuter M, Berninger P, Chuma S, Shah H, Hosokawa M, Funaya C, Antony C, Sachidanandam R, Pillai RS. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature 2011, 480:264–267.
99. Beyret E, Liu N, Lin H. piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. Cell Res 2012, 22:1429 – 1439.
100. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 2008, 453:534 – 538.
101. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 2008, 453:539 – 543.
102. Ding X, Guan H, Li H. Characterization of a piRNA binding protein Miwi in mouse oocytes. Theriogenology 2013, 79:610.e1 – 615.e1.
103. Lim AK, Lorthongpanich C, Chew TG, Tan CW, Shue YT, Balu S, Gounko N, Kuramochi-Miyagawa S, Matzuk MM, Chuma S, et al. The nuage mediates retrotransposon silencing in mouse primordial ovarian follicles. Development 2013, 140:3819–3825.
104. Roovers EF, Rosenkranz D, Mahdipour M, Han CT, He N, de Sousa C, Lopes SM, van der Westerlaken LA, Zischler H, Butter F, et al. Piwi proteins and piRNAs in mammalian oocytes and early embryos. Cell Rep 2015, 10:2069 – 2082.
105. Flemr M, Malik R, Franke V, Nejepinska J, Sedlacek R, Vlahovicek K, Svoboda P. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. Cell 2013, 155:807–816.
106. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell 2006, 127:1193–1207.
107. Billi AC, Freeberg MA, Day AM, Chun SY, Khivansara V, Kim JK. A conserved upstream motif orchestrates autonomous, germline-enriched expression of Caenorhabditis elegans piRNAs. PLoS Genet 2013, 9:e1003392.
108. Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S, et al. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. Mol Cell 2008, 31:67–78.
109. Wang G, Reinke V. A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. Curr Biol 2008, 18:861–867.
110. Cecere G, Zheng GX, Mansisidor AR, Klymko KE, Grishok A. Promoters recognized by forkhead proteins exist for individual 21U-RNAs. Mol Cell 2012, 47:734–745.
111. Das PP, Bagijn MP, Goldstein LD, Woolford JR, Lehrbach NJ, Sapetschnig A, Buhecha HR, Gilchrist MJ, Howe KL, Stark R, et al. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. Mol Cell 2008, 31:79 – 90.
112. Bagijn MP, Goldstein LD, Sapetschnig A, Weick EM, Bouasker S, Lehrbach NJ, Simard MJ, Miska EA. Function, targets, and evolution of Caenorhabditis elegans piRNAs. Science 2012, 337:574 – 578.
113. Gu W, Shirayama M, Conte D Jr, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ, et al. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the C. elegans germline. Mol Cell 2009, 36:231–244.
114. Lee HC, Gu W, Shirayama M, Youngman E, Conte D Jr, Mello CC. C. elegans piRNAs mediate the genome-wide surveillance of germline transcripts. Cell 2012, 150:78 – 87.
115. Buckley BA, Burkhart KB, Gu SG, Spracklin G, Kershner A, Fritz H, Kimble J, Fire A, Kennedy S. A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. Nature 2012, 489:447 – 451.

116. Sijen T, Plasterk RH. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature 2003, 426:310 – 314.
117. Zhong X, Hale CJ, Law JA, Johnson LM, Feng S, Tu A, Jacobsen SE. DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. Nat Struct Mol Biol 2012, 19:870 – 875.
118. Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O. Reconstructing de novo silencing of an active plant retrotransposon. Nat Genet 2013, 45:1029 – 1039.
119. Slotkin RK, Vaughn M, Borges F, Tanurdzi c M, Becker JD, Feijó JA, Martienssen RA. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell 2009, 136:461 – 472.
120. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 2013, 153:193 – 205.
121. Qi Y, Denli AM, Hannon GJ. Biochemical specialization within Arabidopsis RNA silencing pathways. Mol Cell 2005, 19:421 – 428.
122. Ream TS, Haag JR, Wierzbicki AT, Nicora CD, Norbeck AD, Zhu JK, Hagen G, Guilfoyle TJ, Pasa-Toli c L, Pikaard CS. Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. Mol Cell 2009, 33:192 – 203.
123. Haag JR, Pikaard CS. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. Nat Rev Mol Cell Biol 2011, 12:483 – 492.
124. Dalmay T, Hamilton A, Rudd S, Angell S, Baulcombe DC. An RNA-dependent RNA polymerase gene in Arabidopsis is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. Cell 2000, 101:543 – 553.
125. Mourrain P, Béclin C, Elmayan T, Feuerbach F, Godon C, Morel JB, Jouette D, Lacombe AM, Nikic S, Picault N, et al. Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. Cell 2000, 101:533– 542.
126. Parent JS, Bouteiller N, Elmayan T, Vaucheret H. Respective contributions of Arabidopsis DCL2 and DCL4 to RNA silencing. Plant J 2015, 81:223–232.
127. Li J, Yang Z, Yu B, Liu J, Chen X. Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. Curr Biol 2005, 15:1501– 1507.
128. Yang Z, Ebright YW, Yu B, Chen X. HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 20 OH of the 3' terminal nucleotide. Nucleic Acids Res 2006, 34:667 – 675.
129. Morel JB, Godon C, Mourrain P, Béclin C, Boutet S, Feuerbach F, Proux F, Vaucheret H. Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. Plant Cell 2002, 14:629-639.
130. Baumberger N, Baulcombe DC. Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. Proc Natl Acad Sci USA 2005, 102:11928 – 11933.
131. Zhang X, Henderson IR, Lu C, Green PJ, Jacobsen SE. Role of RNA polymerase IV in plant small RNA metabolism. Proc Natl Acad Sci USA 2007, 104:4536–4541.
132. Law JA, Du J, Hale CJ, Feng S, Krajewski K, Palanca AM, Strahl BD, Patel DJ, Jacobsen SE. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. Nature 2013, 498:385 – 389.
133. Haag JR, Ream TS, Marasco M, Nicora CD, Norbeck AD, Pasa-Tolic L, Pikaard CS. In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. Mol Cell 2012, 48:811–818.
134. Zilberman D, Cao X, Jacobsen SE. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science 2003, 299:716–719.
135. Qi Y, He X, Wang XJ, Kohany O, Jurka J, Hannon GJ. Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. Nature 2006, 443:1008 – 1012.
136. Ye R, Wang W, Iki T, Liu C, Wu Y, Ishikawa M, Zhou X, Qi Y. Cytoplasmic assembly and selective nuclear import of Arabidopsis Argonaute4/siRNA complexes. Mol Cell 2012, 46:859 – 870.
137. McCue AD, Panda K, Nuthikattu S, Choudury SG, Thomas EN, Slotkin RK. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. EMBO J 2015, 34:20–35.
138. Duan CG, Zhang H, Tang K, Zhu X, Qian W, Hou YJ, Wang B, Lang Z, Zhao Y, Wang X, et al. Specific but interdependent functions for Arabidopsis AGO4 and AGO6 in RNA-directed DNA methylation. EMBO J 2015, 34:581 – 592.
139. Olmedo-Monfil V, Durán-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, Grimanelli D, Slotkin RK, Martienssen RA, Vielle-Calzada JP. Control of female gamete formation by a small RNA pathway in Arabidopsis. Nature 2010, 464:628–632.
140. Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nat Genet 2009, 41:630– 634.
141. He XJ, Hsu YF, Zhu S, Wierzbicki AT, Pontes O, Pikaard CS, Liu HL, Wang CS, Jin H, Zhu JK. An effector of RNA-directed DNA methylation in Arabidopsis is an ARGONAUTE4- and RNA-binding protein. Cell 2009, 137:498 – 508.

142. Johnson LM, Du J, Hale CJ, Bischof S, Feng S, Chodavarapu RK, Zhong X, Marson G, Pellegrini M, Segal DJ, et al. SRA/SET domain-containing proteins link RNA polymerase V occupancy to DNA methylation. Nature 2014, 507:124 – 128.
143. Liu ZW, Shao CR, Zhang CJ, Zhou JX, Zhang SW, Li L, Chen S, Huang HW, Cai T, He XJ. The SET domain proteins SUVH2 and SUVH9 are required for Pol V occupancy at RNA-directed DNA methylation loci. PLoS Genet 2014, 10:e1003948.
144. Law JA, Ausin I, Johnson LM, Vashisht AA, Zhu JK, Wohlschlegel JA, Jacobsen SE. A protein complex required for polymerase V transcripts and RNA-directed DNA methylation in Arabidopsis. Curr Biol 2010, 20:951 – 956.
145. Zhong X, Du J, Hale CJ, Gallego-Bartolome J, Feng S, Vashisht AA, Chory J, Wohlschlegel JA, Patel DJ, Jacobsen SE. Molecular mechanism of action of plant DRM de novo DNA methyltransferases. Cell 2014, 157:1050 – 1060.
146. Zhong X, Hale CJ, Nguyen M, Ausin I, Groth M, Hetzel J, Vashisht AA, Henderson IR, Wohlschlegel JA, Jacobsen SE. Domains rearranged methyltransferase3 controls DNA methylation and regulates RNA polymerase V transcript abundance in Arabidopsis. Proc Natl Acad Sci USA 2015, 112:911 – 916.
147. Lee TF, Gurazada SG, Zhai J, Li S, Simon SA, Matzke MA, Chen X, Meyers BC. RNA polymerase V-dependent small RNAs in Arabidopsis originate from small, intergenic loci including most SINE repeats. Epigenetics 2012, 7:781–795.
148. Saze H, Tsugane K, Kanno T, Nishimura T. DNA methylation in plants: relationship to small RNAs and histone modifications, and functions in transposon inactivation. Plant Cell Physiol 2012, 53:766– 784.
149. Kato M, Miura A, Bender J, Jacobsen SE, Kakutani T. Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis. Curr Biol 2003, 13:421 – 426.
150. Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol 2014, 21:64 – 72.
151. Hirochika H, Okamoto H, Kakutani T. Silencing of retrotransposons in Arabidopsis and reactivation by the ddm1 mutation. Plant Cell 2000, 12:357–369.
152. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. Bursts of retrotransposition reproduced in Arabidopsis. Nature 2009, 461:423– 426.
153. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O. Selective epigenetic control of retrotransposition in Arabidopsis. Nature 2009, 461:427–430.
154. Nuthikattu S, McCue AD, Panda K, Fultz D, DeFraia C, Thomas EN, Slotkin RK. The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. Plant Physiol 2013, 162:116–131.
155. Zheng B, Wang Z, Li S, Yu B, Liu JY, Chen X. Intergenic transcription by RNA polymerase II coordinates Pol IV and Pol V in siRNA-directed transcriptional gene silencing in Arabidopsis. Genes Dev 2009, 23:2850–2860.
156. Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crété P. Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. Mol Cell 2004, 16:69–79.
157. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. Genes Dev 2004, 18:2368 – 2379.
158. Feng S, Jacobsen SE, Reik W. Epigenetic reprogramming in plant and animal development. Science 2010, 330:622 – 627.
159. Calarco JP, Borges F, Donoghue MTA, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Jeijó JA, Becker JD, et al. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell 2012, 151:194 – 205.
160. Ibarra CA, Feng X, Schoft VK, Hsieh TF, Uzawa R, Rodrigues JA, Zemach A, Chumak N, Machlicova A, Nishimura T, et al. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. Science 2012, 337:1360 – 1364.
161. Gehring M, Bubb KL, Henikoff S. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. Science 2009, 324:1447–1451.
162. Hsieh T-F, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D. Genome-wide demethylation of Arabidopsis endosperm. Science 2009, 324:1451 – 1454.
163. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, et al. The genome sequence of Schizosaccharomyces pombe. Nature 2002, 415:871 – 880.
164. Nakayashiki H, Kadotani N, Mayama S. Evolution and diversification of RNA silencing proteins in fungi. J Mol Evol 2006, 63:127–135.
165. Laurie JD, Linning R, Bakkeren G. Hallmarks of RNA silencing are found in the smut fungus Ustilago hordei but not in its close relative Ustilago maydis. Curr Genet 2008, 53:49 – 58.
166. Drinnenberg IA, Weinberg DE, Xie KT, Mower JP, Wolfe KH, Fink GR, Bartel DP. RNAi in budding yeast. Science 2009, 326:544 – 550.
167. Carr M, Bensasson D, Bergman CM. Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. PLoS One 2012, 7:e50978.
168. Lee NN, Chalamcharla VR, Reyes-Turcu F, Mehta S, Zofall M, Balachandran V, Dhakshnamoorthy J, Taneja N, Yamanaka S, Zhou M, et al. Mtr4-like protein coordinates nuclear RNA processing for heterochromatin assembly and for telomere maintenance. Cell 2013, 155:1061 – 1074.

169. Bayne EH, Portoso M, Kagansky A, Kos-Braun IC, Urano T, Ekwall K, Alves F, Rappsilber J, Allshire RC. Splicing factors facilitate RNAi-directed silencing in fission yeast. Science 2008, 322:602 – 606.
170. Yamanaka S, Mehta S, Reyes-Turcu FE, Zhuang F, Fuchs RT, Rong Y, Robb GB, Grewal SI. RNAi triggered by specialized machinery silences developmental genes and retrotransposons. Nature 2013, 493:557 – 560.
171. Dumesic PA, Natarajan P, Chen C, Drinnenberg IA, Schiller BJ, Thompson J, Moresco JJ, Yates JR 3rd, Bartel DP, Madhani HD. Stalled spliceosomes are a signal for RNAi-mediated genome defense. Cell 2013, 152:957 – 968.
172. Karrer KM. Tetrahymena genetics: two nuclei are better than one. Methods Cell Biol 2000, 62:127– 186.
173. Orias E. The molecular biology of ciliated protozoa: ciliate conjugation. New York: Academic Press;1986, 45 – 84.
174. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell 2002, 110:689–699.
175. Yao MC, Chao JL. RNA-guided DNA deletion in Tetrahymena: an RNAi-based mechanism for programmed genome rearrangements. Annu Rev Genet 2005, 39:537 – 559.
176. Chalker DL, Yao MC. Nongenic, bidirectional transcription precedes and may promote developmental DNA deletion in Tetrahymena thermophila. Genes Dev 2001, 15:1287–1298.
177. Mochizuki K, Gorovsky MA. Conjugation-specific small RNAs in Tetrahymena have predicted properties of scan (scn) RNAs involved in genome rearrangement. Genes Dev 2004, 18:2068 – 2073.
178. Malone CD, Anderson AM, Motl JA, Rexer CH, Chalker DL. Germ line transcripts are processed by a Dicer-like protein that is essential for developmentally programmed genome rearrangements of Tetrahymena thermophila. Mol Cell Biol 2005, 25:9151 – 9164.
179. Mochizuki K, Gorovsky MA. A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. Genes Dev 2005, 19:77–89.
180. Aronica L, Bednenko J, Noto T, DeSouza LV, Siu KW, Loidl J, Pearlman RE, Gorovsky MA, Mochizuki K. Study of an RNA helicase implicates small RNA-noncoding RNA interactions in programmed DNA elimination in Tetrahymena. Genes Dev 2008, 22:2228 – 2241.
181. Noto T, Kataoka K, Suhren JH, Hayashi A, Woolcock KJ, Gorovsky MA, Mochizuki K. Small-RNA-mediated genome-wide trans-recognition network in Tetrahymena DNA elimination. Mol Cell 2015, 59:229 – 242.
182. Liu Y, Mochizuki K, Gorovsky MA. Histone H3 lysine 9 methylation is required for DNA elimination in developing macronuclei in Tetrahymena. Proc Natl Acad Sci USA 2004, 101:1679 – 1684.
183. Liu Y, Taverna SD, Muratore TL, Shabanowitz J, Hunt DF, Allis CD. RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in Tetrahymena. Genes Dev 2007, 21:1530–1545.
184. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, Bétermier M. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev 2009, 23:2478 – 2483.
185. Cheng CY, Vogt A, Mochizuki K, Yao MC. A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. Mol Biol Cell 2010, 21:1753 – 1762.
186. Fang W, Wang X, Bracht JR, Nowacki M, Landweber LF. Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement. Cell 2012, 151:1243 – 1255.
187. Karginov FV, Hannon GJ. The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol Cell 2010, 37:7-19.
188. Barrangou R, Marraffini LA. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. Mol Cell 2014, 54:234 – 244.
189. Koonin EV, Krupovic M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat Rev Genet 2015, 16:184–192.
190. Ross JA, Wardle SJ, Haniford DB. Tn10/IS10 transposition is downregulated at the level of transposase expression by the RNA-binding protein Hfq. Mol Microbiol 2010, 78:607 – 621.

2. Unitas: The universal tool for annotation of small RNAs

Daniel Gebert[1], Charlotte Hewel[1], David Rosenkranz[1]

[1] Institute of Organismic and Molecular Evolutionary Biology, Anthropology, Johannes Gutenberg University, 55099 Mainz, Germany

This chapter was published as a Software Article in *BMC Genomics* under the title "unitas: the universal tool for annotation of small RNAs" (Gebert et al., BMC Genomics 2017 18:644).

2.1. Abstract

Background: Next generation sequencing is a key technique in small RNA biology research that has led to the discovery of functionally different classes of small non-coding RNAs in the past years. However, reliable annotation of the extensive amounts of small non-coding RNA data produced by high-throughput sequencing is time-consuming and requires robust bioinformatics expertise. Moreover, existing tools have a number of shortcomings including a lack of sensitivity under certain conditions, limited number of supported species or detectable sub-classes of small RNAs.

Results: Here we introduce unitas, an out-of-the-box ready software for complete annotation of small RNA sequence datasets, supporting the wide range of species for which non-coding RNA reference sequences are available in the Ensembl databases (currently more than 800). unitas combines high quality annotation and numerous analysis features in a user-friendly manner. A complete annotation can be started with one simple shell command, making unitas particularly useful for researchers not having access to a bioinformatics facility. Noteworthy, the algorithms implemented in unitas are on par or even outperform comparable existing tools for small RNA annotation that map to publicly available ncRNA databases.

Conclusions: unitas brings together annotation and analysis features that hitherto required the installation of numerous different bioinformatics tools which can pose a challenge for the non-expert user. With this, unitas overcomes the problem of read normalization. Moreover, the high quality of sequence annotation and analysis, paired with the ease of use, make unitas a valuable tool for researchers in all fields connected to small RNA biology.

2.2. Background

Small non-coding (snc-) RNAs are important players in diverse cellular processes, often acting as guide molecules in transcriptional and post-transcriptional gene regulation [1–3]. Micro (mi-) RNAs, short interfering (si-) RNAs and Piwi-interacting (pi-) RNAs constitute their most prominent representatives but the number of described sncRNA classes continuously increases. Moreover, degradation products of larger RNA molecules such as rRNA or tRNA fragments further contribute to sequence heterogeneity of sncRNA transcriptomes [4, 5]. As diverse as their source molecules are the places where sncRNAs can be found within an organism, ranging from nuclear and cytoplasmic localization inside a cell, to extracellular exosomes being released into diverse body fluids [6, 7]. Studying the role of sncRNAs in diverse biological contexts typically involves high-throughput sequencing of sncRNAs derived from total RNA extracts. Subsequent disentangling of the complex composition of such sncRNA transcriptomes is one of the initial steps in sequence data processing and critical for all kinds of downstream analysis. As the use of high throughput sequencing technologies becomes more and more common, while this does not necessarily apply to bioinformatics knowhow, a robust and easy to use solution for reliable annotation of sncRNA sequence datasets is highly desirable.

So far, annotation of sncRNA sequence datasets is demanding for various reasons. On the technical side, existing tools cover particular aspects of sequence annotation (e.g. miRNA annotation) which means that complete annotation including all types of sncRNAs requires installation of a set of programs with different dependencies, some of which are restricted to specific operating systems. Illustrating the complexity of the task, a typical annotation process could include the following steps: i) 3′ adapter recognition with Minion [8] or DNApi [9], ii) adapter trimming with e.g. reaper [8] or cutadapt [10], iii) filtering of low complexity sequences with dustmasker [11] or Repeat-Soaker [12], iv) miRNA annotation with Chimira [13], v) annotation of tRNA-derived fragments with tDRmapper [14] or MINTmap [15], vi) annotation of other ncRNA or mRNA fragments with NCBI BLAST and, if applicable, vii) annotation of phased RNAs with PhaseTank [16] or viii) annotation of putative piRNAs by mapping sncRNA sequences to known piRNA producing loci [17].

However, when having established a local annotation pipeline it is almost impossible to correctly normalize the obtained results in case that a given sequence maps to different types of non-coding RNA. Even with a profound bioinformatics expertise, custom annotation is challenging due to the fact, that reference non-coding RNA sequences are stored at different online databases such as Ensembl database, miRBase, GtRNAdb and SILVA rRNA database. Further, mapping sncRNA sequences to reference sequences, once having gathered a complete collection, and subsequent parsing of the obtained results is bedeviled by, e.g., the presence of isomiRs or post-transcriptionally adenylated or uridylated miRNAs.

In order to facilitate and speed-up sncRNA annotation while making the obtained results comparable across different studies, we have developed unitas, a tool for sncRNA sequence annotation that requires not more than a computer with internet connection. Our aim is to provide a maximally convenient tool that runs with an absolute minimum of prerequisites on any popular operating system, making high-quality sequence annotation available for everyone. By providing complete annotation with one tool we intend to tackle the problem of normalization of multiple mapping sequences. In addition, we designed all annotation and analysis algorithms with the aim to overcome a number of limitations of existing tools, in order to make unitas the means of choice compared to a notional pipeline with state-of-the-art tools connected in series. The unitas source code and precompiled executable files are freely available at https://sourceforge.net/projects/unitas/ and http://www.smallrnagroup.uni-mainz.de/software.html.

2.3. Implementation

2.3.1. General requirements

We provide precompiled standalone executable files of unitas for Linux, Mac and Windows systems. Unitas itself is written in Perl and designed to run with an absolute minimum of prerequisites, relying on Perl core modules, or modules which are part of widely used free Perl distributions such as Archive::Extract and LWP::Simple. Perl is commonly preinstalled on Linux and MacOS systems, where users can run the unitas Perl script without any further requirements. Windows users that prefer to run the Perl script rather than the executable file may have to install a free Perl distribution such as ActivePerl or Strawberry Perl. More detailed information and help is available in the unitas documentation. Since unitas uses publicly available online databases for sncRNA annotation, the program needs an internet connection when run for the first time. Later runs can use previously downloaded data. Input files can be sequence files in FASTA or FASTQ format (with or without 3′ adapter sequence), or alternatively map files in SAM or ELAND3 format. Some data analysis features are only available when using map files as input.

2.3.2. Reference sequence data management

Sequence annotation with unitas relies on publicly available reference sequences from Ensembl [18], miRBase [19], GtRNAdb [20], SILVA rRNA database [21] and piRNA cluster database [17] (Fig. 1). Currently, unitas supports 835 different species or strains for which information on ncRNAs is available at least in one of the Ensembl databases. Prior to annotation, unitas downloads a collection of latest reference sequences which are stored in a separate folder on the local machine for subsequent mapping. As availability of reference sequences is crucial, unitas is designed to address possible challenges that can occur during acquisition of that data. Since database URLs often change with new releases or updates of reference sequences, relying on URLs stored inside the programs source code would require frequent updates of the unitas software itself. Therefore, unitas connects to the Mainz University Server (MUS) and loads the latest list of URLs for downloading the required reference sequence data. However, in the event of these URL not being up to date (URLs are updated monthly), unitas ultimately downloads the required sequence data directly from MUS where the datasets are available via stable URLs and are synchronized regularly (Fig. 1).

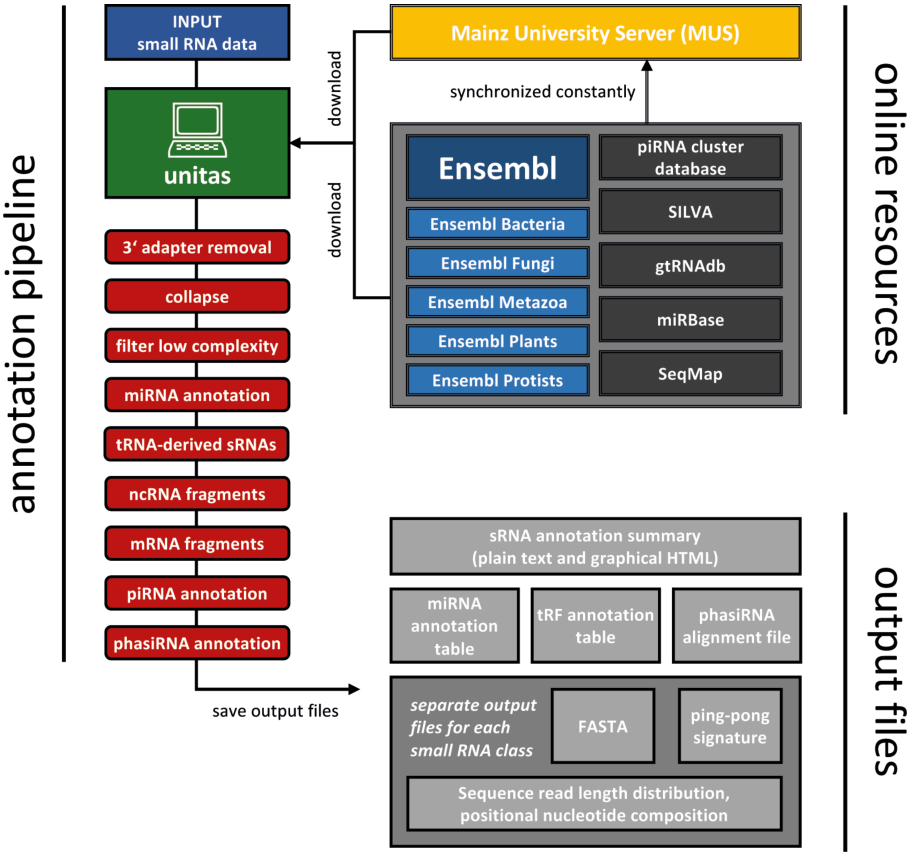


Figure 1 | Architecture of the unitas workflow at a glance.

By default, downloaded sequence datasets are used for subsequent unitas runs without anew downloading by default. Users can also download reference sequence data collections for any supported species at any time and use the downloaded data for later offline runs. The downloaded sequence data can be updated anytime.

2.3.3. Automated 3′ adapter recognition and trimming

Standard cloning protocols for small RNA library preparation prior to high throughput sequencing involve ligation of adapter molecules to both ends of a RNA molecule. During sequencing, the sequencing primer typically hybridizes to the 3′ end of the 5′ adapter, which means that the resulting

sequence read starts with the original small RNA sequence and, given a sufficient read length, ends with the 3′ adapter sequence. However, there exist manifold commercially available 3′ adapters that can be used for library construction. In addition, these adapters can contain different index/barcode sequences. Finally, working groups may even use custom made adapter molecules. Since information on 3′ adapter sequences that were used to generate a small RNA dataset is not always available or at least difficult to find out, we integrated an adapter recognition and trimming module that can be applied using the option '-trim'.

Initially, unitas identifies the most frequently occurring sequence ignoring sequence read positions 1 to 22 (typical length for miRNAs). unitas adjusts the length of the motif, m, to be identified automatically according to the formula m = n - 22 (as long as: $6 \leq m \leq 12$) where n refers to the sequence read length. A first round of adapter trimming is then performed based on the identified motif allowing 2 mismatches for 12 nt motifs, 1 mismatch for motifs ≤11 nt and 0 mismatch for motifs ≤8 nt. If the original motif is not found within a given sequence read, unitas truncates the motif sequentially by one 3′ nt and checks for its occurrence at the very 3′ end of the sequence read until the motif is found or the motif length falls below 6 nt. Following this first round of adapter trimming, unitas checks the positional nucleotide composition of the trimmed sequence reads and will remove further 3′ nucleotide positions in case they exceed a specified nucleotide bias (default = 0.8). It is noteworthy that there may exist scenarios in which unitas will not detect the correct 3′ adapter sequences when using the default settings, particularly in cases with short library read length (≤35 nt) combined with a high amount of reads that share 3′ similarity such as, e.g., tRNA-derived fragments. In these special cases, adapter recognition can be improved by increasing the amount of 5′ positions to be ignored when searching for frequent sequence motifs (option: -trim_ignore_5p [n]).

2.3.4. Filtering low complexity reads

To filter out low complexity reads, unitas employs an advanced version of the duster algorithm from the NGS TOOLBOX [22]. By default, sequence reads with a length fraction $f > 0.75$ being composed of one repetitive sequence motif (default motif length = 1–5 nt) are rejected. Further, sequences with a length fraction

$$f' > f + [(1 - f) * f]$$

being composed of only two specific nucleotides are also rejected.

2.3.5. miRNA annotation

unitas performs miRNA annotation in several consecutive steps. Mature miRNA sequences and miRNA hairpin sequences are downloaded from miRBase and miRNAs are annotated in the following order: i) Canonical miRNAs of the species in question, ii) post-transcriptionally 3′-tailed canonical miRNAs of the species in question, iii) offset miRNAs of the species in question, iv) post-transcriptionally 3′-tailed offset miRNAs of the species in question. Subsequently, this procedure is repeated using miRNA sequence data from all other species included in miR-Base, which is particularly useful for those species with bad miRNA annotation status considering the fact that many miRNA sequences are widely conserved. Since the according output file comprises information on the source species of each matched miRNA gene, unitas users are able to assess the relevance of each match in a case-dependent manner. However, it is important to be aware that this approach will not identify new, unannotated lineage-specific miRNA genes, which can only be identified using de novo prediction tools. Nevertheless, accurate filtering of known miRNA sequences will be helpful for downstream de

novo miRNA prediction. By default, the maximum number of allowed non-template 3′ nucleotides is 2 and the maximum number of allowed internal modifications is 1. In order to map sncRNA sequences to miRNA precursors (or to other ncRNA sequences in later annotation steps), unitas employs the mapping tool SeqMap [23] which not requires prior indexing of reference sequences and allows subsequent analysis of non-template 3′-nucleotides.

2.3.6. ncRNA/mRNA annotation

Following miRNA annotation, sequences that do not correspond to miRNAs are mapped to a species-specific collection of non-coding RNA and cDNA sequences downloaded from Ensembl database [18], Genomic tRNA database [20] and SILVA rRNA database [21]. Read counts of sequences that match different classes of reference sequences equally well are apportioned according to the simple equation:

$$c_{class} = \sum_{i=1}^{n} \frac{r_i}{h_i}$$

where c_{class} refers to the read counts for ncRNA/cDNA class, while n is the total number of non-identical input sequences that map to this class and r_i and h_i refer to read counts and hits to different ncRNA/cDNA classes of input sequence i, respectively. During this process, special attention is payed to sequence reads matching tRNAs since different classes of functional tRNA derived fragments, so-called tRFs, have been described in the recent past [24–30]. unitas classifies these sequences into 5′ tRFs (5′ to D-loop), 5′ tR-halves (5′ to Anticodon-loop), 3′ tRFs (TψC-loop to 3′), 3′ CCA-tRFs (TψC-loop to 3'CCA), 3′ tR-halves (Anticodon-loop to 3′), tRF-1 (3′ end of mature tRNA to oligo-T signal), tRNA-leader (sequence upstream of 5′ ends of mature tRNAs) and misc.-tRFs (miscellaneous tRFs). Worth mentioning, unitas relies on available ncRNA annotation and will not perform de novo prediction of ncRNA genes that e.g. encode tRNAs or rRNAs.

2.3.7. piRNA annotation

Considering the fact that piRNAs are highly diverse and virtually not conserved across different species, piRNA annotation based on sequence is challenging. However, many piRNAs originate from few genomic loci, many of which are annotated in the piRNA cluster database [17]. Providing that information on piRNA clusters is available for the species in question, sequences that were not annotated as (fragment of) any other class of non-coding RNA are mapped to known piRNA producing loci of the respective species. Since almost every nucleotide position within a piRNA precursor transcript can give rise to the 5′ end of a mature piRNA, though there is certainly a bias for 5′-U, this procedure more reliably identifies putative piRNAs compared to the approach of directly mapping sequence reads to annotated piRNAs. Further evidence for the presence of genuine piRNAs can be obtained from sequence read length distribution and positional nucleotide composition which unitas outputs for each class of small RNAs separately. Providing that the input file provided by the user represents a map file, unitas can further screen the map file for the so-called ping-pong signature (using the option -pp), which refers to a bias for 10 nt 5′ overlaps of mapped sequence reads which arises from secondary piRNA biogenesis (ping-pong cycle) and indicates the presence of primary and secondary piRNAs. Screening for a ping-pong signature also includes calculation of a Z-score according to the method described by Zhang and coworkers [31].

2.3.8. phasiRNA annotation

The commonly applied method for identification of phased RNAs bases on calculation of a so-called phase score, P. After consolidation of mapped reads from both strands with an offset of 2 nt for minus strand mapped reads, P results from the following formula:

$$P = ln\left[\left(1 + \sum_{i=1}^{8} k_i\right)^{n-2}\right]$$

in which n refers to the number of phase cycle positions occupied by at least one small RNA read within an eight-cycle window, and k refers to the total number of reads for all small RNAs with consolidated start coordinates in a given phase within an eight-cycle window [32].

Although the given formula yields higher P values with increasing k or n, the weighting between both factors, and finally the decision of which threshold to choose for P is rather arbitrary. We therefore decided to use a different method, which utilizes the binomial distribution to calculate the probability p to observe a defined number (or more) of phased reads within a given sliding window (default = 1 kb) according to the formula:

$$p = 1 - \left(\sum_{k=0}^{j} \binom{n}{k} q^k (1-q)^{n-k}\right)$$

in which j refers to the observed number of reads with length i in a specified phase, n refers to the total number of reads with length i and q is given by 1/i and refers to the probability of a read to be located in a given phase, assuming that a sequence read can map to any position within the sliding window with equal probability. As is the case for calculation of P, reads mapped to different strands are consolidated prior to calculation of p. If the p value of a locus under examination is below the critical value (default = 0.05, with strict Bonferroni correction based on the number of analyzed sliding windows), unitas applies further thresholds to reduce the rate of false positive predictions. By default, the fraction of phased RNAs has to be ≥50% of all mapped reads within a sliding window. Further, the phased reads must map to ≥5 different loci while not more than 90% of the phased reads must derive from one strand. Critical values for each of the mentioned parameters, including p and sliding window size can be adjusted by the user. Prediction of phasiRNAs requires map files (SAM or ELAND3) as input and can be performed with the option '-phasi [n]' where n refers to the length of the phased RNAs.

2.4. Results

We have tested unitas using a number of artificial datasets, real RNA-seq data and combinations of both. A detailed description of the datasets and the methods that were applied to generate them can be found in Additional file 1 (Supplementary Methods).

2.4.1. 3′ adapter identification and trimming

The first steps in the analysis of small RNA data usually involve the removal of sequencing adapters from 3′ ends, for which numerous tools exist. However, this task becomes problematic if the adapter sequence is not known, e.g. if a dataset is deposited without the appropriate information. The number of programs for adapter prediction, in contrast to removal, is rather limited. The only published tools for this purpose are DNApi [9] and Minion from the Kraken package [8], which also contains Reaper for adapter trimming.

To test the efficiency of the 3′ adapter identification and trimming function of unitas, we processed ten randomly chosen datasets from the NCBI Sequence Read Archive and put the performance into comparison to the existing software. Both, unitas and DNApi, reliably predicted the correct adapter sequences in all cases, whereas Minion predicted a false adapter with a slightly deviated sequence for one of the ten libraries (SRA accession: SRR5130142), leading to a considerably reduced efficacy in subsequent read trimming by Reaper (Fig. 2a). Altogether, in eight instances unitas removed more adapter sequences than Reaper, hence resulting in higher quantities of trimmed reads that could be mapped perfectly to the corresponding genome (+9.7% on average, Additional file 2: Table S1).

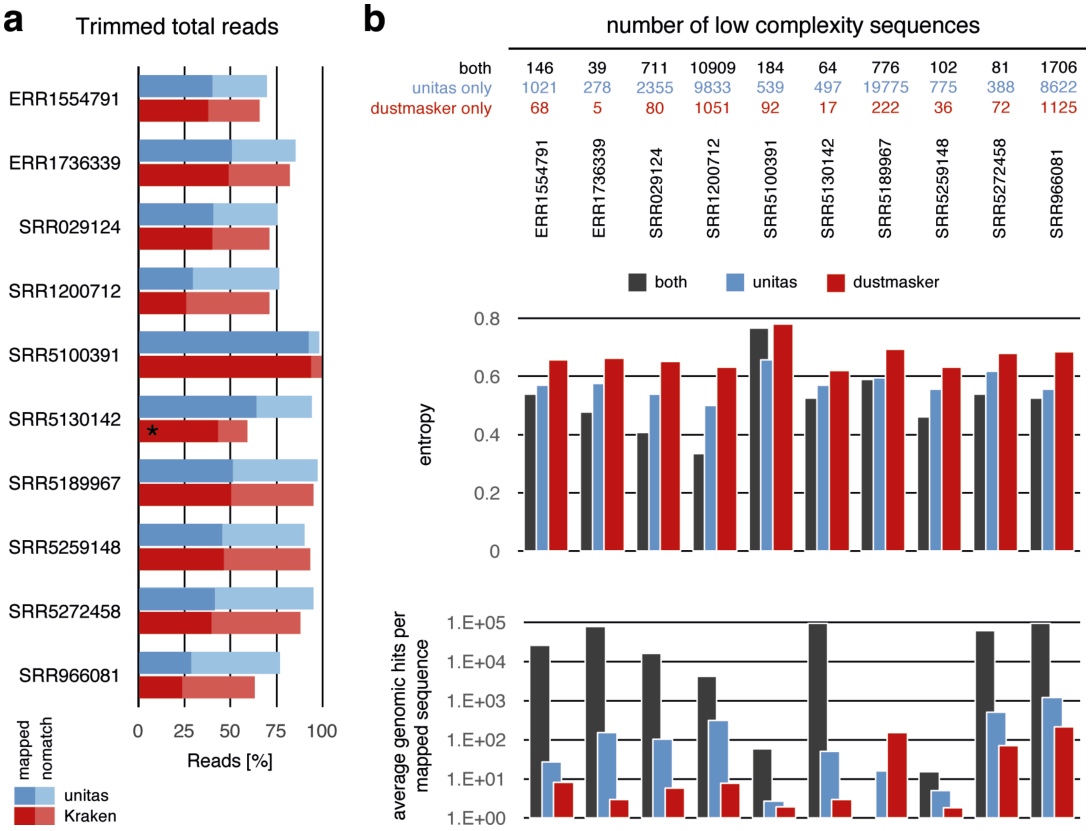


Figure 2 | Adapter trimming and removal of low complexity reads. **a** Efficiency of 3′ adapter identification and trimming by unitas and Kraken tool kit. Asterisk marks the dataset for which adapter prediction by Minion failed. Trimmed reads were mapped to the corresponding genome without mismatches. **b** Removal and analysis of low complexity sequences filtered by unitas and dustmasker.

2.4.2. Removal of low complexity sequences

The presence of low complexity reads can weaken biological signals within RNA-seq datasets and it has been demonstrated, that the correlation between RNA-seq and microarray gene expression data can be improved with strict filtering of sequences that map to genomic regions with low sequence complexity [12]. Since this method relies on the availability of a RepeatMasker annotation for the genome in question, we implemented a low complexity filter upstream of sequence annotation. We compared our filter with dustmasker, a popular tool for masking low complexity regions in DNA sequences, which is part of the NCBI blast+ package [11]. We ran dustmasker on ten adapter-trimmed NGS sequence datasets (see above) with default settings and discarded sequence reads with more than 75% of bases being masked to produce results that are comparable to the results generated by unitas, which by default filters sequences being composed of more than 75% of repetitive sequence motifs.

First, we found that unitas generally filters more sequences (Fig. 2b top), which taken by itself is certainly not indicative to favor one tool over the other since both algorithms can easily be adjusted to filter more or less sequences by changing the corresponding thresholds. Therefore, we in-depth analyzed the complexity of sequences filtered by both tools as well as those sequences that were filtered either by unitas or by dustmasker. We quantified the complexity of filtered sequences based on sequence entropy [33], Wootton-Federhen-complexity [34] and gzip (developed by Jean-loup Gailly and Mark Adler) compression ratio using the program SeqComplex, which was written by Juan Caballero. As expected, sequences filtered by both tools usually exhibit the lowest degree of entropy. Further, sequences filtered only by unitas exhibit a lower degree of entropy compared to sequences filtered only by dustmasker, in spite of the fact that unitas filters more sequences (Fig. 2b middle). According to Wootton-Federhen-complexity and gzip compression ratio, sequences filtered only by unitas also exhibit lower complexity compared to sequences filtered only by dustmasker and even compared to those sequences filtered by both tools (Additional file 3: Table S2).
We further wanted to check whether these rather theoretical assessments can be translated into a biological dimension. To this end we mapped the filtered sequences to the respective genomes and counted the number of genomic hits per sequence, assuming that the amount of information obtained by mapping a specific sequence decreases with a growing number of genomic hits. In line with the previous results, sequences filtered by both tools show the highest number of genomic hits, thus providing the lowest amount of information (Fig. 2b bottom). With one exception, sequences filtered exclusively by unitas map more frequently to the genome compared to sequences filtered exclusively by dustmasker (Fig. 2b bottom). Together, these results demonstrate that unitas filters sequences with low complexity in a more sensitive and more specific manner.

2.4.3. Annotation of miRNAs

Numerous programs for miRNA annotation in small RNA-seq data have been published in the past with varying focuses [35–37]. To compare the performance of unitas on this task we chose Chimira, which is a recent tool with a similar range of functions, primarily aiming at miRNA expression and modification analysis [13]. Chimira is a web-based system, accepting multiple input files in FASTA or FASTQ format at once and supporting 209 genomes so far. Input reads are mapped against miRBase [19] hairpin sequences using BLASTn with two tolerated mismatches, identifying modifications at 3′ and 5′ ends, as well as internal substitutions (single nucleotide polymorphisms, SNPs) and ADAR-dependent editing events in the process. However, in order for an internal modification to be classified as a SNP, an arbitrary value of 70% is applied as a threshold for the ratio of modification counts to overall counts.
For a controlled comparison, we produced an artificial miRNA dataset based on human hairpin sequences from miRBase (release 21), incorporating internal modifications and 3′ tailings. Of overall 466,810 generated reads, unitas identified 99.9% as miRNAs, while Chimira detected only 85.8% of the original set. Moreover, unitas showed higher precision in assigning read counts to respective miRNA genes of origin than Chimira did, indicated by Pearson correlation coefficients of 0.9514 and 0.9146, respectively (Fig. 3a, Additional file 4: Table S3). Furthermore, unitas detected 3′ tailings and internal modifications more reliably, whereas Chimira barely showed the latter type, probably due to the considerably high (70%) threshold for internal modifications (Fig. 3b). It is noteworthy that the test dataset was designed to include all possible combinations of offset-, tailing-, and mismatch-scenarios without any weighting between canonical and non-canonical sequences. Consequently, the differences between unitas and Chimira annotations are typically less marked for real biological datasets (these observations are principally true for annotation of tRNA fragments as well, see below).

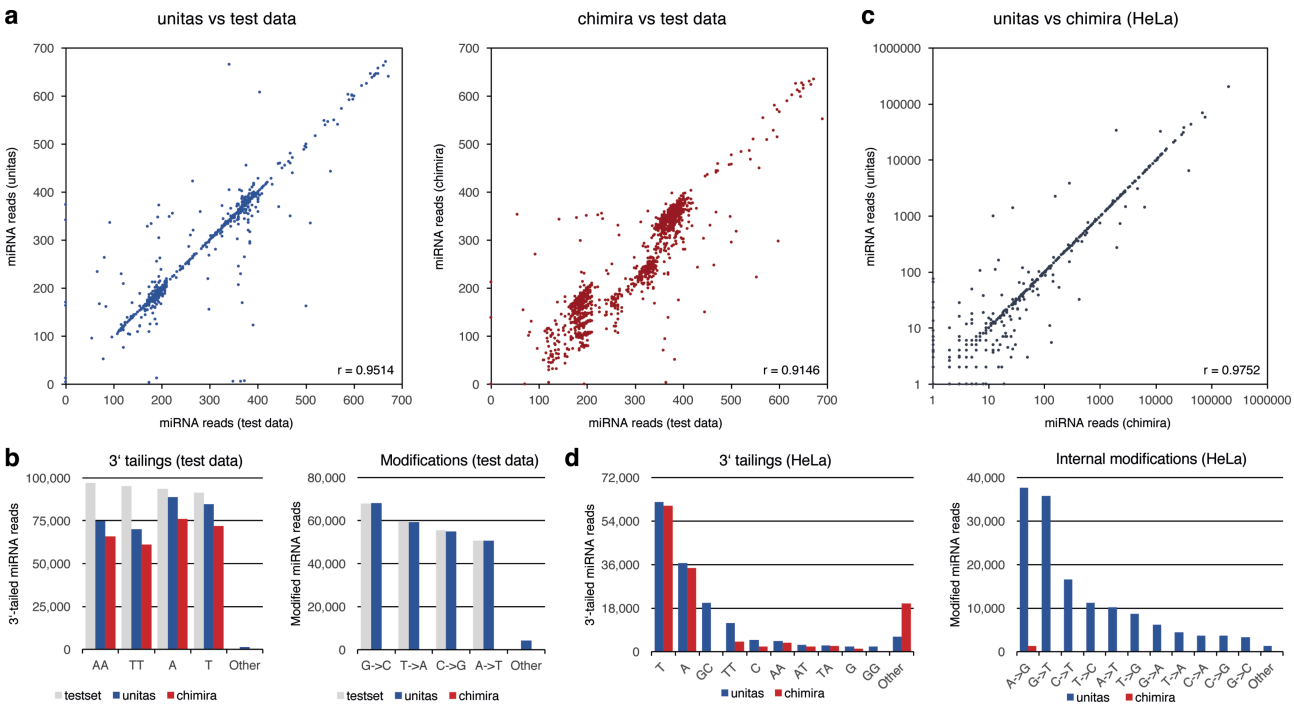


Figure 3 | Analysis of miRNA expression and modification by unitas compared to Chimira. **a** Correlations of reads assigned to different miRNA genes by unitas and Chimira compared to reads profile of artificial dataset. For better comparison of miRNA counts, reads from same miRNA gene were combined. **b** 3′ tailings and internal modifications detected by unitas and Chimira in test data. **c** miRNA expression in HeLa cells determined by unitas compared to Chimira on logarithmic scale. **d** miRNA 3′ tailings and internal modifications quantified by unitas and Chimira in HeLa cells.

Subsequently, both tools were tested on published RNA-seq data obtained from HeLa cells (SRA accession: SRR029124) [38]. The resulting miRNA expression profiles are highly similar, with a Pearson correlation coefficient of 0.9752 (Fig. 3c). While unitas found 961,840 miRNA reads, Chimira called 960,880, which increased to 961,563 if the option '-split counts from paralogs' was selected. The amount of identified uridylation and adenylation events of 3′ ends were largely similar with a slight advantage on the side of unitas, but other tailing patterns were detected to a much lesser degree by Chimira (Fig. 3d). Analogous to the artificial test data, Chimira did not identify internal modifications to a comparable extent as unitas, apart from some amount of ADAR-dependent edits (A-to-G), which was also the most frequent modification detected by unitas.

Since both, unitas and similar computational approaches for miRNA annotation, rely on miRBase, it should be noted that the quality of database annotations in general vary among species, particularly those which are less well studied. Therefore, we point to existing tools designed specifically for the de novo prediction of miRNAs, such as CAP-miRSeq [35] and Oasis [36].

2.4.4. Annotation of tRNA-derived small RNAs

Currently, there are three major tools specific to the identification of tRNA-derived sncRNAs [39]. The tRFfinder of the tRF2Cancer web server package [40] is restricted to the analysis of human samples and considers sequences with lengths between 14 and 32 nt only. This, however, poses a limitation for the detection of longer tRNA-derived sncRNAs like tRNA-halves. For instance, 56% of tRNA-derived reads are larger than 32 nt in a sncRNA dataset of seminal exosomes (SRA accession: SRR1200712) [41].

The second tool, tDRmapper [14], is a command line based set of Perl scripts, which identifies tRNA-derived RNAs (tDRs) from 14 to 40 nt in human and murine samples. Other species can be added manually according to a provided guide and with the help of Perl scripts that depend on bedtools.

Notably, there are some features to the algorithm of tDRmapper that may hamper direct comparison with unitas. First, sequences with 100 reads or less are discarded. Subsequently, so-called primary tDRs are determined by the location, at which more than 50% of all reads mapping to a source tRNA are aligned. For example, the 5′-tRF type is assigned if more than 50% map at the 5′-end of the source tRNA. Moreover, tRFs are defined by length as being smaller than 28 nt and tRNA-halves as 28 nt or larger, regardless of alignment position. Further, a primary tDR is only specified if more than 66% of all reads mapping to the source tRNA map to any position of the considered tDR. Lastly, tDRs are quantified by 'relative abundance', which is calculated by multiplying the percentage of tDR reads that map to its source tRNA and the proportion of reads on the area with the highest read coverage across the source-tRNA. Importantly, the resulting counts are not normalized, meaning there is no fractional assignment for multimapping reads. As the authors themselves point out, this approach may overestimate the relative abundance of a primary tDR.

Finally, another command line based tool called MINTmap was recently developed for the profiling of tRNA fragments from human small RNA-seq data, emphasizing the profiling of both nuclear and mitochondrial tRNA fragments [15]. However, tRFs generated from trailer sequences (tRF-1) and 5′ leader-tRFs are excluded from analysis.

To test the efficiency and accuracy of unitas in the detection of tRNA-derived small RNAs, we produced an artificial dataset based on human tRNAs from the genomic tRNA database, incorporating one mismatch in 50% of sequences. Running with default settings, unitas assigned 92% of reads to tRNAs, which increased to 97% if miRNA detection was skipped (option '-no_miR'). For running tDRmapper on the test data, we disabled the rejection of sequences with less than 101 reads, since this would eliminate the entire input. Both, tDRmapper and MINTmap, deviated considerably from the original dataset in read shares assigned to different tRNA-derived sRNA classes, in contrast to unitas (Fig. 4a). A direct comparison of read counts, however, was not possible due to the previously described quantification method of tDRmapper, which calculates so-called relative abundance. Further, read shares were most precisely assigned to source tRNAs by unitas, indicated by the highest Pearson correlation coefficient (0.9896) among the tested tools (Fig. 4b).

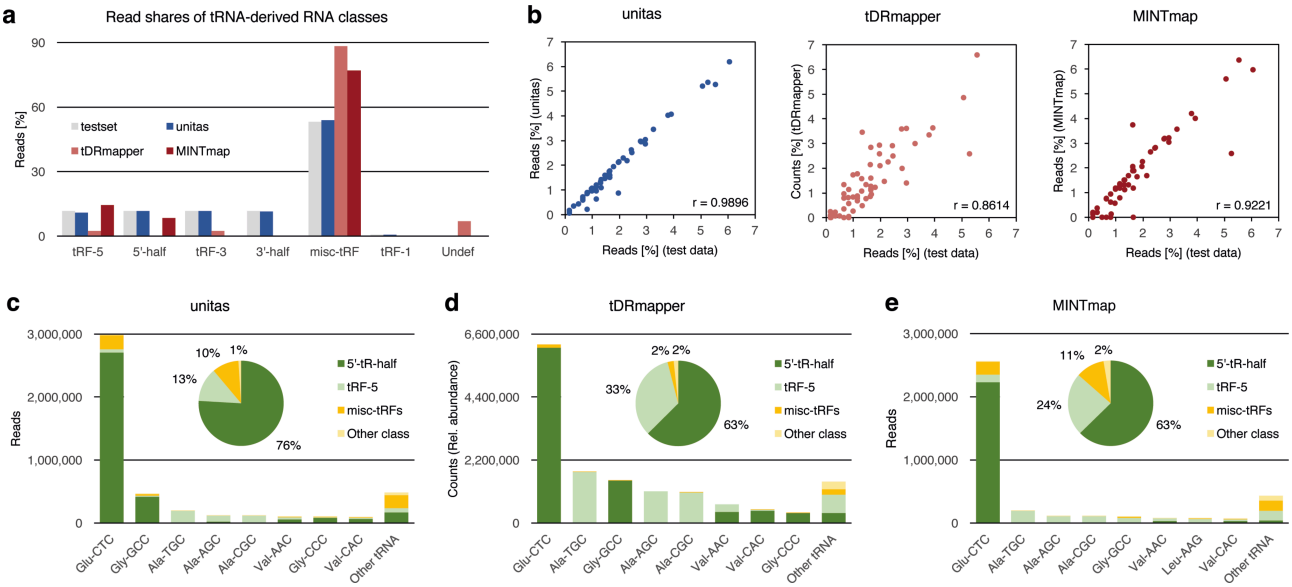


Figure 4 | Detection of tRNA-derived small RNAs by unitas, tDRmapper and MINTmap. **a** Read shares assigned to different tRNA-derived RNA classes of artificial test data by unitas and other tools compared to test set profile. Misc-tRFs (miscellaneous tRFs) are equivalent to internal tRFs. On test data, the elimination of sequences with less than 101 reads by tDRmapper was disabled. **b** Correlation of read proportions allocated to source tRNAs by unitas and other tools with original test data read shares per tRNA. **c** Analysis of RNA-seq data from seminal exosomes by unitas, **d** tDRmapper and **e** MINTmap.

Additionally, unitas, tDRmapper and MINTmap were tested on RNA-seq data of exosomes from seminal fluid (SRR1200712) [41], using default settings. The differences in results between unitas (Fig. 4c) and tDRmapper (Fig. 4d) are largely due to the lack of fractional assignment for multi-mapping reads in the quantification approach of tDRmapper. Analysis by MINTmap yielded results that are largely similar to the output of unitas, but with overall slightly lower read counts and changed order of the source tRNAs with descending read coverage (Fig. 4e). Details on the test dataset and the annotation of tRFs from artificial and biological data with different tools are available in Additional file 5: Table S4 (A-G).

2.4.5. Annotation of phasiRNAs

We tested and compared phasiRNA annotation performance of unitas and PhaseTank, which is currently the only published tool for prediction of phased RNAs [16]. We used artificial datasets with known amounts of phased RNAs (Additional file 6: Table S5) as well as biological small RNA data from panicles of the two rice strains 93–11 and Nipponbare [42] to predict phased RNAs with unitas and PhaseTank using default parameters. All test datasets were collapsed to non-identical sequences, retaining information on read counts for each sequence in the FASTA header. Subsequently, the datasets were formatted to satisfy PhaseTank requirements (special format of FASTA headers) and used as input for PhaseTank (v.1.0) using default settings to search for 21 nt phased RNAs. Subsequently we searched for 24 nt phased RNAs with PhaseTank using the option '-size 24'. To generate input files for unitas, we mapped the test datasets to the human genome (GRCh38) with bowtie1, bowtie2 and STAR using settings that correspond to recommended and widely used settings for mapping of small RNAs with these tools and considering only perfect matches to be in line with PhaseTank default settings. The resulting SAM alignment files were used as input for unitas which was started twice with the option '-phasi 21' or '-phasi 24', respectively, to search for 21 nt and 24 nt phased RNAs.

Using artificial datasets, we found that both tools perform equally well with those datasets that comprise exclusively phased RNAs or have rather low amounts of non-phased RNAs (Fig. 5a). However, PhaseTank drastically loses its sensitivity with an increasing amount of non-phased sequences within a dataset, while the sensitivity of unitas remains unaffected (Fig. 5a). When we assigned a read count value of 10 to each artificial phased RNA sequence, PhaseTank performs approximately as well as unitas, illustrating that sensitivity of PhaseTank not only depends on the number of phased sequences, but also on the number of reads per phased sequence. Consequently, PhaseTank will particularly miss those phasiRNA-producing loci that have a low sequence read coverage (Fig. 5a and b). For neither tool we observed a namable issue with false positive predicted phasiRNAs when running both programs with datasets comprising no phased RNAs (Additional file 7: Table S6).

When searching for 21 nt phased RNAs in biological datasets we noted that unitas identifies slightly more phasiRNAs compared to PhaseTank, while the number of identified clusters was identical (Fig. 5c and d). Overall, the congruency between unitas and PhaseTank results is very high (Fig. 5d). However, when searching for 24 nt phased RNAs in the same datasets we observed remarkable differences with unitas identifying both more phasiRNA sequences and more phasiRNA clusters (Fig. 5e and f). Considering that PhaseTank is less sensitive when the fraction of phased RNAs within a given dataset is low, these results are in line with the fact that the abundance of 24 nt phasiRNAs in rice panicles is several times lower compared to 21 nt phasiRNAs [42]. Accordingly, phasiRNA clusters identified only by unitas have relatively low read coverage, while phasiRNA clusters identified only by PhaseTank were rejected by unitas because of a high strand bias (>95%) of mapped sequence reads.

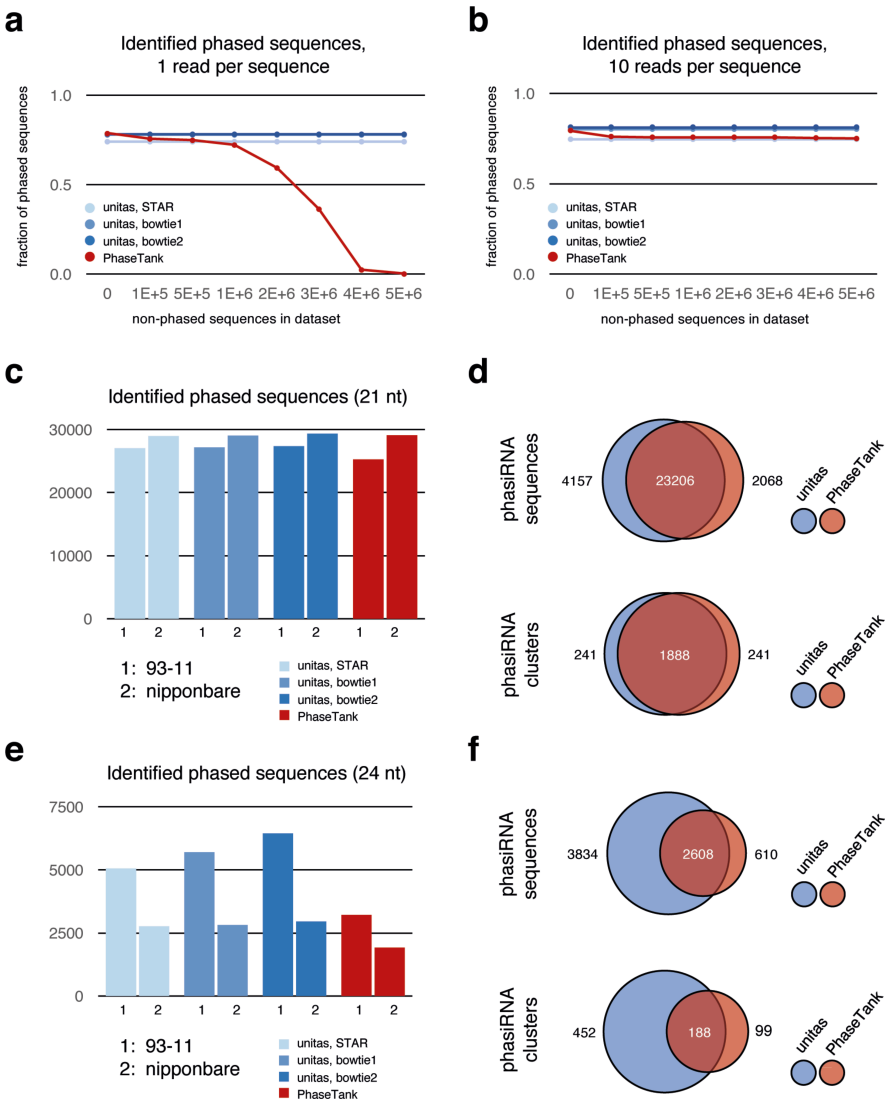


Figure 5 | Identification of phased RNAs with unitas and PhaseTank. **a** Identified phased RNAs in artificial datasets, with one read per phased RNA and growing amount of background sequences. **b** Identified phased RNAs in artificial datasets, with ten reads per phased RNA and growing amount of background sequences. **c** Number of identified 21 nt phasiRNAs in small RNA datasets from rice panicles. **d** Congruency between unitas and PhaseTank results for 21 nt phasiRNA prediction. **e** Number of identified 24 nt phasiRNAs in small RNA datasets from rice panicles. **f** Congruency between unitas and PhaseTank results for 24 nt phasiRNA prediction.

2.4.6. Complete annotation of NGS datasets

To emphasize the broad range of possible applications of unitas on diverse sncRNA-seq datasets, we analyzed three exemplary libraries, which differ in origin and structure (Fig. 6). The sncRNA annotation output of unitas provides a general overview of the small RNA composition, as shown for a dataset produced from HeLa cancer line cells (SRA accession: SRR029124, Fig. 6a) [38]. In this library, the largest fraction of sncRNAs is constituted by miRNAs, which are of growing interest for cancer studies and clinical trials using miRNA profiling for patient diagnosis [43]. Apart from expression profiles and 3′ tailings, unitas offers a convenient description of miRNA modifications per position (Fig. 6b). For target recognition, complementarity of the seed region of a miRNA (positions 2–7) to its target is critical for downstream silencing efficacy. Beyond the seed region, a strong sequence conservation can also be observed at position 8 [44], and finally, miRNA sequences frequently start

with a uridine which was found to promote miRNA loading on Argonaute proteins [45]. According to these functional aspects, the unitas output shows that in HeLa cells the first eight positions from the 5′ end are rarely modified. Internal modifications occur predominantly at distinct positions downstream of the seed region, with A-to-G (A-to-I), known as ADAR edits [46], and G-to-T being the most common, followed by modifications leading to uridine or guanine incorporation (Fig. 6b).

Next, we analyzed a library generated from exosomes of human seminal fluid (SRA accession: SRR1200712) [41]. Notably, this dataset is particularly abundant in tRNA-derived reads, while containing less miRNAs and other annotated reads (Fig. 6c). For the analysis of such sequences, unitas provides a summary of read counts for each tRNA gene, apportioned to classes of tRNA-derived sncRNAs. The majority of tRNA-derived reads in this library is represented by 5′ halves, originating mainly from the tRNAs Glu-CTC and Gly-GCC (Fig. 6d). It has been shown that fragments specifically derived from 5′ ends of tRNA-Gly-GCC in mouse epididymosomes repress genes by regulating the endogenous retroelement MERVL, whereas an RNA interference against the middle or 3′ end of tRNA-Gly-GCC and other tRNAs had no effect on these MERVL-dependent genes [27]. Generally, we found 5′ halves and 5′ tRFs to be the dominant classes, being especially associated with tRNAs exhibiting high read coverage (Fig. 6d). With decreasing coverage, however, tRNAs tend to give rise primarily to internal fragments and to a lesser extent to 3′ halves and 3′ tRFs. This suggests that

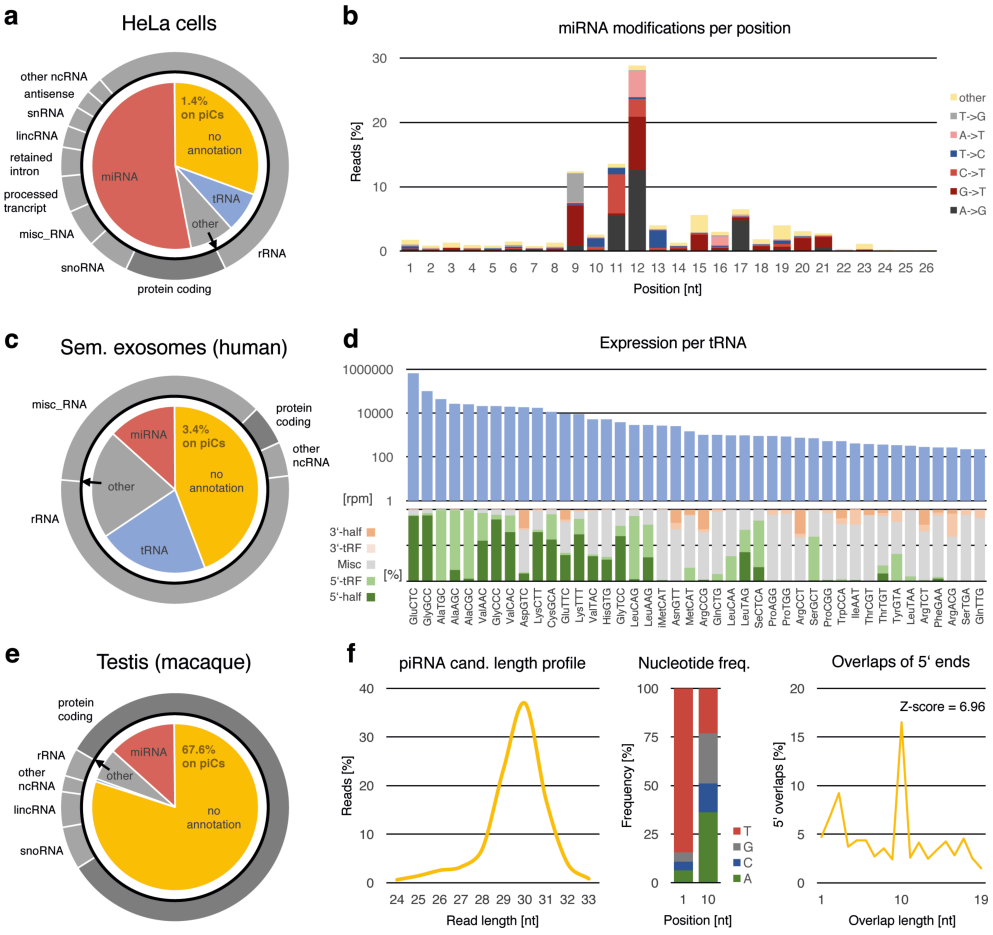


Figure 6 | Exemplary analysis of three small RNA libraries by unitas. **a** General sncRNA annotation of sRNA-seq data from HeLa cells (SRR029124). **b** Overall miRNA modification per position in HeLa cells data. **c** General sncRNA annotation of sRNA-seq data from exosomes of human seminal fluid (SRR1200712). **d** Read coverage of tRNAs as reads per million (rpm) on logarithmic scale and percentages of tRNA-derived sRNA classes per tRNA. **e** General sncRNA annotation of sRNA-seq data from macaque testis (SRR553581). **f** Rates of 5′ overlaps and nucleotide frequencies of piRNA candidate reads, i.e. not annotated reads that map to known piRNA producing loci or piRNA clusters (piCs).

internal fragments, which we describe as miscellaneous tRFs (misc-tRFs), might be rather characterized as random debris of degraded tRNAs than a class of tRNA-derived small RNAs in its own right.
Lastly, we chose a sRNA dataset from macaque testis for analysis (SRA accession: SRR553581) [47]. As expected, the vast majority of testis-expressed sRNAs did not match any class of known non-coding RNA (Fig. 6e). In contrast to the former libraries, unitas found that the bulk of non-annotated reads (67.6%) maps to known piRNA producing loci. Besides the typical length profile (Fig. 6f), these piRNA candidate sequences show a strong bias for uridine at 5′ ends (84.5%) which is typical for primary piRNAs being processed by the endonuclease Zucchini (PLD-6) [48, 49]. Moreover, unitas attests a significant ping-pong signature (Z-score = 6.96), namely a high rate of 10 nt 5′ overlaps of sense and antisense reads, which is a hallmark of secondary piRNA biogenesis via the ping-pong cycle [50–52].

2.5. Discussion

Small RNA biology has become a major field in molecular biology research. In 2016, sequence data from 7271 Illumina sequencing runs with miRNA sequencing strategy, comprising more than 82 billion sequence reads, was uploaded to NCBI's Sequence Read Archive. Assuming total costs of 15 USD per 1 Million clean sequence reads, the total miRNA sequencing value for 2016 amounts to 1.2 million USD. Noteworthy, these numbers only refer to published sequence data and certainly only mirror the tip of the iceberg. Nevertheless, in light of these numbers, even a seemingly trivial improvement of adapter recognition and trimming by unitas yields a surplus value of more than 100,000 USD per year, considering the amount of additionally mapped sequence reads. However, although this is a benefit of unitas that can be descriptively quantified, it clearly reflects only a minor aspect of the overall value of unitas.
Within the field of small RNA biology, miRNAs receive widespread attention owing to their pervasive contribution to gene regulatory processes [53]. However, it is not only mere miRNA expression, but also their post-transcriptional modification that vitally affects miRNA activity. Uridylation of miRNAs is thought to play a role in miRNA stability and possibly marks small RNAs for degradation [54, 55]. Adenylation has recently been linked to clearance of maternal miRNAs in Drosophila eggs [56]. Further, internal modification events of miRNAs (or their precursors) can have wide implications for miRNA biogenesis and function [57–60]. It is therefore of immense importance, to accurately identify post-transcriptional editing events to gain a deeper understanding of miRNA-dependent regulatory processes (Additional file 8). As we have shown, unitas is more sensitive in detecting 3′ tailing events and much more sensitive in detecting internal modifications compared to existing tools. Importantly, unitas not only focuses on well-known adenylation, uridylation and ADAR-dependent A-to-I editing, but also allows to detect all other types of modification events which can greatly facilitate the detection of yet unknown enzymatic editing activity in the future.
tRNA-derived small RNAs have been regarded as simple and non-functional degradation products for a long time. However, strong evidence for diverse functional roles in gene regulation, cancer biology, apoptosis and protein synthesis is mounting [24–30]. Since tRNAs and their precursor transcripts can be processed into functionally distinct types of tRFs, accurate attribution of tRNA-derived small RNAs to the different types of tRFs is important to make functional interpretations. In this regard, unitas shows higher precision than existing tools, while being also more sensitive in overall tRF detection. Notably, the recently published tool for tRF annotation MINTmap [15], which is more sensitive and accurate than the older tDRmapper [14] cannot identify some of the yet rather enigmatic tRFs which have their origin beyond the mature tRNA molecule, namely tRF-1 and 5′ leader-tRFs. Here, unitas can enable researches to elucidate possible functions of these cryptic RNAs by first of all spotting them in sncRNA transcriptomes.

Initially described as trans-acting siRNAs [61], plant specific phasiRNAs are well-characterized actors in posttranscriptional gene silencing [62]. In most cases, phasiRNAs are 21 nt in length, but different pathways that produce 22 nt and 24 nt phasiRNAs have been described as well. Since the latter are by far less abundant, their detection in small RNA transcriptomes is challenging and the current approach underestimates the number of, e.g., phased 24 nt RNAs in small RNA datasets from rice panicles. In contrast, sensitivity of unitas depends far less on the amount of background reads, making it more suitable for the detection of particularly low abundance phasiRNAs and their source loci.

2.6. Conclusion

So far, accurate annotation of sncRNA required a large set of different software tools with a number of additional pre-requisites. While the installation of these bioinformatics tools can pose a challenge for the non-expert user, unitas brings together all annotation and analysis features with an absolute minimum of further requirements. A complete annotation run is finished within a few minutes and can be started with one simple shell command, making its usage very convenient. By facilitating sncRNA annotation and providing in depth analyses that previously was not accessible for the non-expert user, we believe that unitas is a valuable tool for researchers in all fields connected to small RNA biology.

2.7. Declarations

Acknowledgements

We thank Hans Zischler, Hendrik Dörschmann, Sabrina Saurin, Julia Jehn and Julian Kiefer for testing unitas in the course of their projects and providing valuable feedback how to improve this software.

Funding

This project was supported by the Natural and Medical Sciences Research Center of the University Medicine Mainz (NMFZ) and the International PhD Programme coordinated by the Institute of Molecular Biology (IMB), Mainz, funded by the Boehringer Ingelheim Foundation.

Author contributions

DR has designed the unitas software, **DG** and DR coded it. The software was tested by CH, **DG** and DR. In-depth analysis of existing annotation algorithms was done by CH and **DG**. Figures were designed by **DG** and DR. The manuscript was written by **DG** and DR with valuable input of CH. All authors read and approved the final manuscript.

2.8. References

1. Holoch D, Moazed D. RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet. 2015;16:71–84.
2. Gebert D, Rosenkranz D. RNA-based regulation of transposon expression. WIRES RNA. 2015;6:687–708.
3. Gorski SA, Vogel J, Doudna JA. RNA-based recognition and targeting: sowing the seeds of specificity. Nat Rev Mol Cell Biol. 2017;18:215–28.
4. Li Z, Ender C, Meister G, Moore PS, Chang Y, John B. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. Nucleic Acids Res. 2012;40:6787–99.
5. Daugaard I, Hansen TB. Biogenesis and function of ago-associated RNAs. Trends Genet. 2017;33:208–19.
6. Sarkies P, Miska EA. Small RNAs break out: the molecular cell biology of mobile small RNAs. Nat Rev Mol Cell Biol. 2014;15:525–35.
7. Rashed MH, Bayraktar E, Helal GK, Abd-Ellah MF, Amero P, Chavez-Reyes A, et al. Exosomes: from garbage bins to promising therapeutic targets. Int J Mol Sci. 2017;18:E538.
8. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods. 2013;63:41–9.
9. Tsuji J, Weng Z. DNApi: a de novo adapter prediction algorithm for small RNA sequencing data. PLoS One. 2016;11:e0164228.

10. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10–2.
11. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006;13:1028–40.
12. Dozmorov MG, Adrianto I, Giles CB, Glass E, Glenn SB, Montgomery C, et al. Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. BMC Bioinformatics. 2015;16(Suppl 13):S10.
13. Vitsios DM, Enright AJ. Chimira: analysis of small RNA sequencing data and microRNA modifications. Bioinformatics. 2015;31:3365–7.
14. Selitsky SR, Sethupathy P. tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. BMC Bioinformatics. 2015;16:354.
15. Loher P, Telonis AG, Rigoutsos I. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. Sci Rep. 2017;7:41184.
16. Guo Q, Qu X, Jin W. PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades. Bioinformatics. 2015;31:284–6.
17. Rosenkranz D. piRNA cluster database: a web resource for piRNA producing loci. Nucleic Acids Res. 2016;44:D223–30.
18. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res. 2017;45:D635–42.
19. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68–73.
20. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016; 44:D184–9.
21. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
22. Rosenkranz D, Han CT, Roovers EF, Zischler H, Ketting RF. Piwi proteins and piRNAs in mammalian oocytes and early embryos: from sample to sequence. Genom Data. 2015;5:309–13.
23. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008;24:2395–6.
24. Lee YS, Shibata Y, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. 2009;23:2639–49.
25. Sobala A, Hutvagner G. Small RNAs derived from the 5′ end of tRNA can inhibit protein translation in human cells. RNA Biol. 2013;10:553–63.
26. Keam SP, Hutvagner G. tRNA-derived fragments (tRFs): emerging new roles for an ancient RNA in the regulation of gene expression. Lifestyles. 2015;5:1638–51.
27. Sharma U, Conine CC, Shea JM, Boskovic A, Derr AG, Bing XY, et al. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science. 2016;351:391–6.
28. Venkatesh T, Suresh PS, Tsutsumi R. tRFs: miRNAs in disguise. Gene. 2016;579:133–8.
29. Kumar P, Kuscu C, Dutta A. Biogenesis and function of transfer RNA-related fragments (tRFs). Trends Biochem Sci. 2016;41:679–89.
30. Keam SP, Sobala A, Ten Have S, Hutvagner G. tRNA-derived RNA fragments associate with human Multisynthetase complex (MSC) and modulate ribosomal protein translation. J Proteome Res. 2017;16:413–20.
31. Zhang Z, Xu J, Koppetsch BS, Wang J, Tipping C, Ma S, et al. Heterotypic piRNA ping-pong requires qin, a protein with both E3 ligase and Tudor domains. Mol Cell. 2011;44:572–84.
32. Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, et al. Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/ DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. Plant Cell. 2007;19:926–42.
33. Schmitt AO, Herzel H. Estimating the entropy of DNA sequences. J Theor Biol. 1997;188:369–77.
34. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 1996;266:554–71.
35. Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, et al. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. BMC Genomics. 2014;15:423.
36. Capece V, Garcia Vizcaino JC, Vidal R, Rahman RU, Pena Centeno T, Shomroni O, et al. Oasis: online analysis of small RNA deep sequencing data. Bioinformatics. 2015;31:2205–7.
37. Zheng Y, Ji B, Song R, Wang S, Li T, Zhang X, et al. Accurate detection for a wide range of mutation and editing sites of microRNAs from small RNA high-throughput sequencing profiles. Nucleic Acids Res. 2016;44:e123.
38. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and Polyadenylation activates Oncogenes in cancer cells. Cell. 2009;138:673–84.
39. Xu W-L, Yang Y, Wang Y-D, Qu L-H, Zheng L-L. Computational approaches to tRNA-derived small RNAs, Non-Coding RNA, vol. 3; 2017. p. 2.
40. Zheng L, Xu W, Liu S, Sun W, Li J, Wu J, et al. tRF2Cancer: a web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers. Nucleic Acids Res. 2016;44:W185–93.

41. Vojtech L, Woo S, Hughes S, Levy C, Ballweber L, Sauteraud RP, et al. Exosomes in human semen carry a distinctive repertoire of small non-coding RNAs with potential regulatory functions. Nucleic Acids Res. 2014;42:7290–304.
42. SongX,LiP,ZhaiJ,ZhouM,MaL,LiuB,etal.RolesofDCL4andDCL3bin rice phased small RNA biogenesis. Plant J. 2012;69:462–74.
43. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends Mol Med. 2014;20:460–9.
44. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15–20.
45. Seitz H, Tushir JS, Zamore PD. A 5′-uridine amplifies miRNA/miRNA* asymmetry in drosophila by promoting RNA-induced silencing complex formation. Silence. 2011;2:4.
46. Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science. 2007;315:1137–40.
47. Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, et al. Birth and expression evolution of mammalian microRNA genes. Genome Res. 2013;23:34–45.
48. Ipsaro JJ, Haase AD, Knott SR, Joshua-Tor L, Hannon GJ. The structural biochemistry of zucchini implicates it as a nuclease in piRNA biogenesis. Nature. 2012;491:279–83.
49. Nishimasu H, Ishizu H, Saito K, Fukuhara S, Kamatani MK, Bonnefond L, et al. Structure and function of zucchini endoribonuclease in piRNA biogenesis. Nature. 2012;491:284–7.
50. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in drosophila. Cell. 2007;128:1089–103.
51. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, et al. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in drosophila. Science. 2007;315:1587–90.
52. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 2008;31:785–99.
53. Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. Nat Rev Genet. 2015;16:421–33.
54. Li J, Yang Z, Yu B, Liu J, Chen X. Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in Arabidopsis. Curr Biol. 2005;15:1501–7.
55. Ameres SL, Horwich MD, Hung JH, Xu J, Ghildiyal M, Weng Z, et al. Target RNA-directed trimming and tailing of small silencing RNAs. Science. 2010;328:1534–9.
56. Lee M, Choi Y, Kim K, Jin H, Lim J, Nguyen TA, et al. Adenylation of maternally inherited microRNAs by wispy. Mol Cell. 2014;56:696–707.
57. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.
58. Chawla G, Sokol NS. ADAR mediates differential expression of polycistronic microRNAs. Nucleic Acids Res. 2014;42:5245–55.
59. Cui Y, Huang T, Zhang X. RNA editing of microRNA prevents RNA-induced silencing complex recognition of target mRNA. Open Biol. 2015;5:150126.
60. Behm M, Öhman M. RNA editing: a contributor to neuronal dynamics in the mammalian brain. Trends Genet. 2016;32:165–75.
61. Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, et al. Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. Mol Cell. 2004;16:69–79.
62. Fei Q, Xia R, Meyers BC. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. Plant Cell. 2013;25:2400–15.

2.9. Supplement

2.9.1. Building test datasets for miRNA annotation

We used human mature miRNA sequences and miRNA precursor sequences from miRBase to build a miRNA test dataset comprising a specified number of unmodified and modified miRNAs. The modifications comprised all possible combinations of 5' and 3' offsets from -2 to +2 nucleotides and in addition up to 3 nucleotides 5' offset without 3' offset and vice versa. We further added A, AA, U or UU 3'-tailed versions for each canonical and offset miRNA. Finally, for every sequence we obtained this way, we added a counterpart comprising one internal sequence modification at position ten.

miRNA sequences were generated based on their precursor sequences, meaning that every precursor hairpin can yield the same number of miRNA reads, given that an offset modification not extends the resulting miRNA sequence beyond the precursor sequence. Based on the latter, and the fact that some miRNA genes have several (possibly slightly different) genomic copies, the obtained number of reads for different miRNAs varies (Additional file 7: Table S7). Both the test dataset and the Perl script that was used to create the test dataset are freely available at http://www.smallrnagroup.uni-mainz.de/data/UNITAS/resources.html.

2.9.2. Building test datasets for tRNA annotation

We used human tRNA sequences downloaded from Genomic tRNA database to build all types of tRNA fragments from each annotated tRNA sequence. 5'tRFs matched the 5' end of a tRNA and ranged in size from 18 to 22 nt. 3'tRFs matched the 3' end of a tRNA and ranged in size from 20 to 24 nt. Each tRNA was cut into two pieces at positions 32 to 36 to yield 5'-halves and 3'-halves. Miscellaneous tRNA fragments ranging from 18 to 40 nt in size were created using internal tRNA sequence starting from position 8. tRNA trailer sequences were used to generate 18 to 40 nt tRF1s. Both the test dataset and the Perl script that was used to create the test dataset are freely available at http://www.smallrnagroup.uni-mainz.de/data/UNITAS/resources.html.

2.9.3. Building test datasets for phasiRNA annotation

In order to test the sensitivity and accuracy of phasiRNA prediction, we generated a collection of different artificial test datasets comprising those that contain solely phased RNAs, those that contain no phased RNAs and precisely defined mixtures of both. We first generated artificial phasiRNA datasets in silico, applying the following procedure: For each human chromosome including chromosomes X and Y, we quasi-randomly chose 91 loci that served as template for generation of phased small RNA sequences, starting at coordinate 1,000,001. If the 1050 bp downstream sequence did not comprise stretches of N, we generated 100 subsequences representing 50 artificial phased 21 nt RNAs per strand, directly adjacent to each other, with two nucleotides offset for plus strand sequences. For each next locus, we moved 10 kb downstream and generated siRNAs as described above. While keeping 100 artificial small RNAs for the first locus of a chromosome, we randomly rejected an increasing number of artificial siRNAs ending with 10 artificial small RNAs at locus 91 of a given chromosome. This procedure resulted in 125,125 artificial phased siRNAs representing 116,017 non-identical sequences.

To allow for quantification of false-negative as well as false-positive phasiRNA prediction, we prepared datasets containing our artificial phased small RNAs and an increasing number of non-phased sequence reads from human miRNA datasets representing Universal Human Reference RNA (Agilent Technologies, #750700) and human brain total RNA (Life Technologies, #AM6050) [1]. The two human miRNA datasets (SRA accessions: SRR950876 and SRR950878) were downloaded from NCBI's Sequence Read Archive. 3' adapter sequences from human miRNA datasets were clipped screening for TGGAATTCTCGGNx-3' and only sequences ranging from 18 to 40 nt were chosen for further

processing. For the different test datasets, we subsequently added 0, 1E+5, 5E+5, 1E+6, 2E+6, 3E+6, 4E+6 and 5E+6 sequence reads from SRR950876 (test datasets 1-8) or SRR950878 (test datasets 9-16) to the artificial phased small RNAs. We further generated a second collection of test datasets just as described above, but assigning a sequence read count of ten to each artificial phased RNA (test datasets 17-32). We also used both miRNA datasets without adding artificial phased small RNAs to test for false positive phasiRNA prediction (test datasets 33 and 34, Additional file 6: Table S6). Test datasets 1-34 were mapped to the human genome GRCh38 with STAR (command line options: --outSAMstrandField All --outFilterScoreMinOverLread 0 --outFilterMatchNmin 15 --outFilterMatchNminOverLread 0 --outFilterMismatchNoverLmax 0 --alignIntronMax 1) [2], bowtie1 (command line options: -f -v 0 -k 10 -S -t) [3] and bowtie2 (command line options: -- local -p 16 -f -D 20 -R 3 -N 0 -L 8 -i S,1,0.50 -k 10 -t -x) [4], considering only perfect matches by subsequent filtering of SAM map files. The same aligners and settings were used to map sncRNA sequences from rice strains nipponbare and 93-11 [5] to the respective genomes [6, 7]. Test datasets 1-34 and the Perl script that was used to create artificial phased RNAs are freely available at http://www.smallrnagroup.uni-mainz.de/data/UNITAS/resources.html.

2.9.4. References

1. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods. 2014;11:809-815.
2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
3. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
4. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357-359.
5. Song X, Li P, Zhai J, Zhou M, Ma L, Liu B et al. Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J. 2012;69:462-474.
6. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6:4.
7. Rise Database. BGI Shenzhen, China. 2009. http://rise2.genomics.org.cn/page/rice/download.jsp. Accessed 10 Feb 2017.

2.9.5. Supplementary tables

Supplementary tables are available at https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4031-9.

3. PIWIs and piRNAs in the germline and soma of mollusks

Julia Jehn[1]*, Daniel Gebert[1]*, Frank Pipilescu[1]*, Sarah Stern[1], Julian S. T. Kiefer[1], Charlotte Hewel[1], David Rosenkranz[1]

[1] Institute of Organismic and Molecular Evolution, Anthropology, Johannes Gutenberg University Mainz, Anselm-Franz-von-Bentzel-Weg 7, 55099 Mainz, Germany
* These authors contributed equally to this work

This chapter was accepted for publication as a Research Article in *Communications Biology* under the title "PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation" and is in final revision.

3.1. Abstract

PIWI proteins and PIWI-interacting RNAs (piRNAs) suppress transposon activity in animals, thus protecting their genomes from detrimental insertion mutagenesis. Here, we reveal that PIWI genes and piRNAs are ubiquitously expressed in mollusks, similar to the situation in arthropods. We describe lineage specific adaptations of transposon composition in piRNA clusters in the great pond snail and the pacific oyster, likely reflecting differential transposon activity in gastropods and bivalves. We further show that different piRNA clusters with unique transposon composition are dynamically expressed during oyster development. Finally, bioinformatics analyses suggest that different populations of piRNAs presumably bound to different PIWI paralogs participate in homotypic and heterotypic ping-pong amplification loops in a tissue- and sex specific manner. Together with recent findings from other animal species, our results support the idea that somatic piRNA expression represents the ancestral state in metazoans.

3.2. Introduction

In virtually all animals, PIWI proteins protect germ cells from the steady threat of mobile genetic elements, so-called transposons [1,2]. Based on sequence complementarity to their target transcripts, 23-31 nt non-coding RNAs, termed PIWI-interacting (pi-) RNAs, function as guide molecules for PIWI proteins that slice matching targets through their endonuclease activity. Besides post-transcriptional transposon control, PIWI proteins and piRNAs can trigger the establishment of repressive epigenetic DNA or chromatin modifications, thus inducing efficient transposon silencing on the transcriptional level [3-6].

Analyses of piRNA pathways in representatives of many animal taxa have unveiled a great diversity of lineage specific adaptations, challenging the universal validity of insights obtained from model organisms [7-19]. For a long time, PIWI proteins and piRNAs were thought to be dispensable for female germ cell development in mammals until it became clear that the model organisms mouse and rat represent an exception from the mammalian rule in that they employ an oocyte specific Dicer isoform for transposon control instead of Piwil3 which is expressed in the bovine and human female germline [15,20]. Similarly, evidence for a gene regulatory role of piRNAs [14,21-27] and their widespread somatic expression in many animals [19,28-35] have eroded the dogma that the piRNA pathway is restricted to the germline, being exclusively responsible for silencing of transposons. Indeed, it has been shown that piRNAs are essential for regeneration and stem cell maintenance in the flatworm *Schmidtea mediterranea* [28], provide an adaptive immunity against virus infections in *Aedes*

aegypti [36], are responsible for sex determination in *Bombyx mori* [37] and memory-related synaptic plasticity in *Aplysia californica* [38].

Despite the likely more than seventy thousand living molluskan species [39] there exist only a few functional descriptions of PIWI proteins or piRNAs for this taxon based on experiments in the sea slug *Aplysia californica* [38], the Farrer's scallop *Chlamys farreri* [40] and in the dog whelk *Nucella lapillus* [41]. Importantly, Waldron and coworkers recently showed that piRNA-like small RNAs matching virus and transposon sequences are somatically expressed in *Nucella lapillus*. However, the available data does not allow to draw any conclusions on whether this represents a conserved or lineage-specific feature of the PIWI/piRNA system within mollusks. In order to further elucidate the evolution of the PIWI/piRNA system in mollusks, we have reconstructed the evolution of PIWI genes in this phylum based on 11 sequenced genomes showing that Piwil1 and Piwil2 are conserved in mollusks. We perform quantitative real-time PCR experiments to analyze the expression patterns of the identified PIWI paralogs across a representative set of tissues from the great pond snail *Lymnaea stagnalis* (*L. stagnalis*) and the pacific oyster *Crassostrea gigas* (*C. gigas*). We apply high-throughput sequencing of small RNAs from *L. stagnalis* to verify the presence of piRNAs in germline and muscle tissue. We further reanalyze published small RNA sequence data from *C. gigas* to characterize the dynamic expression of piRNAs from distinct piRNA clusters during oyster development. Finally, we use bioinformatics approaches to show that different piRNA populations and PIWI paralogs participate in the ping-pong amplification loop in a tissue- and sex specific manner.

3.3. Results

3.3.1. The molluskan PIWI gene repertoire

Many PIWI gene tree reconstructions have been published in the past years, however they do not provide a coherent picture regarding the evolution of PIWI genes in early bilaterians. Thus, we first wanted to characterize the PIWI protein equipment of sequenced mollusks to infer the ancestral molluskan state and subsequent evolution of PIWI paralogs within the molluskan clade. To this end, we used available PIWI protein sequence data from six molluskan species (*Biomphalaria glabrata*, *Aplysia californica*, *Crassostrea gigas*, *Crassostrea virginica*, *Mizuhopecten yessoensis*, *Octopus bimaculoides*) and further manually annotated PIWI genes based on five publicly available but not yet (sufficiently) annotated genomes (*Lymnaea stagnalis*, *Radix auricularia*, *Lottia gigantea*, *Bathymodiolus platifrons*, *Pinctada martensii*). We found that the PIWI family members Piwil1 and Piwil2 are conserved in mollusks and are orthologous to Piwil1 and Piwil2 in vertebrates, suggesting a duplication event in an early bilaterian ancestor prior to the split of protostomes and deuterostomes. According to our results and in consistency with a number of previously published gene trees, *Drosophila* AGO3 shares a common ancestral gene with Piwil2 clade members [18,42-44]. However, the insect-specific PIWI genes Piwi and Aubergine, the latter one resulting from a duplication event in dipteran flies [44,45], do not group with the Piwil1 clade (Figure 1a). It is worth mentioning in this context that different rates of sequence evolution, selective regimes and gene turnover for Argonaute subfamilies make it difficult to infer their ancient evolutionary history, which is mirrored by numerous published but contradicting PIWI gene trees, none of which correctly mirrors the phylogenetic relationship of the included species. Consequently, the presented gene tree reconstruction aims to provide a reliable reconstruction of molluskan PIWI gene evolution while the deeper topology should be considered with caution.

While we did not observe further gene duplication events within the molluskan Piwil2 clade, several duplication events are present in the Piwil1 clade resulting in two Piwil1 paralogs in *Bathymodiolus platifrons* and even three Piwil1 paralogs in *Lymnaea stagnalis* and *Radix auricularia*. Generally, PIWI gene duplication events are in line with the previously described erratic evolution of PIWI family genes in

arthropods [19,44-46]. Noteworthily, it was also a successive duplication of Piwil1 on the eutherian lineage that gave rise to Piwil3 (with subsequent loss on the murine lineage) and Piwil4 [47,48] (Figure 1a).

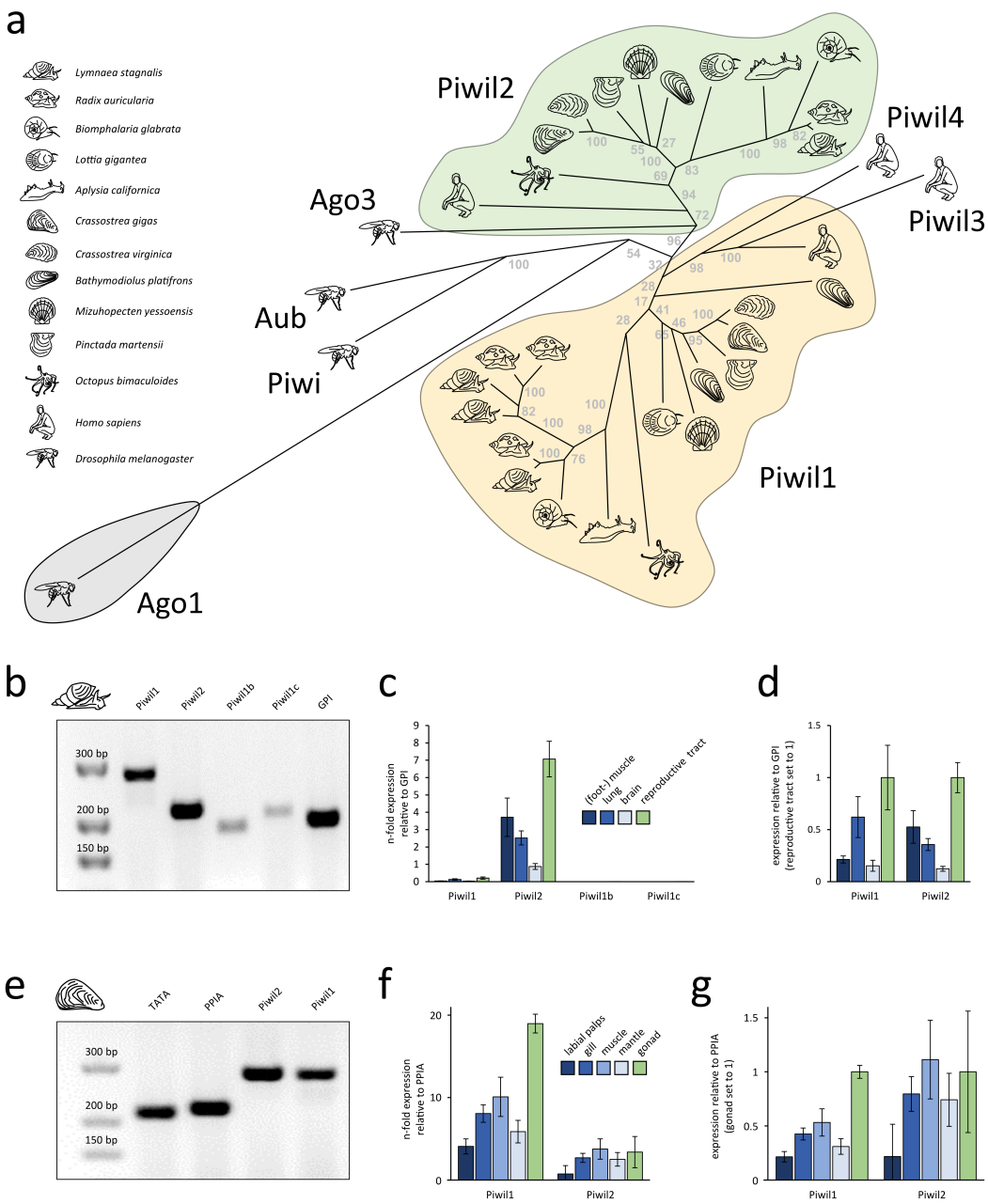


Figure 1 | Evolution and expression of PIWI genes in mollusks. (a) PIWI gene tree reconstruction of molluskan PIWI genes. (b) Control PCR with PIWI paralog specific primers and *L. stagnalis* cDNA from the reproductive tract. (c) RT-qPCR results for PIWI paralog expression in different tissues of *L. stagnalis*, measured as n-fold expression of the housekeeping gene GPI. Error bars indicate standard deviation. (d) PIWI paralog expression in different tissues of *L. stagnalis*, normalized by the expression of the housekeeping gene GPI, values from reproductive tract set to 1. Error bars indicate standard deviation. (e) Control PCR with PIWI paralog specific primers and *C. gigas* cDNA from the adductor muscle. (f) RT-qPCR results for PIWI paralog expression in different tissues of *C. gigas*, measured as n-fold expression of the housekeeping gene PPIA. Error bars indicate standard deviation. (g) PIWI paralog expression in different tissues of *C. gigas*, normalized by the expression of the housekeeping gene PPIA, values from male gonad set to 1. Error bars indicate standard deviation.

3.3.2. Expression of PIWI genes in L. stagnalis and C. gigas

To investigate the expression of PIWI genes in mollusks we chose two representative species, the pacific oyster *Crassostrea gigas* (*C. gigas*, Bivalvia) showing no Piwil1 duplication, and the great pond snail *Lymnaea stagnalis* (*L. stagnalis*, Gastropoda), featuring three predicted Piwil1 paralogs (Figure 1a). We performed quantitative real-time PCR (qPCR) for each PIWI paralog on a representative set of tissues from both species.

For the great pond snail *L. stagnalis* we measured PIWI expression on the mRNA level in the hermaphroditic reproductive tract, comprising both male and female gametes, foot muscle, lung and brain. Significant expression was detectable for Piwil1 and particularly Piwil2, while the Piwil1 duplicates Piwil1b and Piwil1c were only expressed at very low levels (Figure 1b, 1c and Supplementary Figure 1) suggesting a spatiotemporal sub-functionalization. As expected, we observed the highest expression of Piwil1 and Piwil2 in the reproductive tract. However, both genes were significantly expressed in the other analyzed tissues as well, reaching 62%, 21% and 15% of germline expression for Piwil1 in muscle, lung and brain respectively, and 36%, 53% and 12% of germline expression for Piwil2 in muscle, lung and brain, respectively (Figure 1d).

For the dioecious pacific oyster *C. gigas*, PIWI mRNA expression was measured in the male gonad, labial palps, gill, adductor muscle and mantle. We detected significant expression of Piwil1 and Piwil2 across all analyzed tissues, particularly in gonadal tissue (Figure 1e and 1f), confirming data on Piwil1 expression in the Hong Kong Oyster *Crassostrea honkongensis* [49]. In relation to gonadal expression, Piwil1 and Piwil2 were expressed in levels ranging from 21% (Piwil1 in labial palps) to 111% (Piwil2 in adductor muscle, Figure 1g). The observed expression patterns suggest that a functional PIWI machinery acting in the soma and the germline is conserved in mollusks. Considering the somatic expression of PIWI proteins and piRNAs in many arthropod species [19], it is parsimonious to assume that somatic PIWI/piRNA expression represents the ancestral state that was established in an early protostomian ancestor.

3.3.3. piRNAs in L. stagnalis muscle and reproductive tract

In order to characterize molluskan piRNAs, we sequenced small RNA transcriptomes from *L. stagnalis* extracted from the hermaphroditic reproductive tract and (foot-) muscle, since muscle tissue was found to exhibit the highest somatic PIWI expression in both *L. stagnalis* and *C. gigas*. Importantly, we want to clarify that we will use the term piRNA *bona fide*, without formal evidence for physical interaction with PIWI proteins but based on the evidence provided in the following.

The sequence read length profiles for both tissues show a maximum for 21 nt RNAs, with a considerable amount of 22 nt RNAs being present in the muscle, but not in the reproductive tract. We further observed a smaller fraction of RNAs in the range of 24-29 nt in both samples (Figure 2a). Annotation of sRNA sequences with unitas [50] revealed a similar proportion of different sRNA classes in each tissue type, with miRNAs accounting for 47% and 53% of reads in the reproductive tract and muscle, respectively (Figure 2b, Supplementary Table 1). Interestingly, we found a substantial difference in the abundance of tRNA fragments (tRFs). In both samples, 21 nt RNAs derived from the 3' end of tRNAs (3' tRFs, particularly from tRNA-Gly-TCC) constitute the vast majority of tRNA fragments. However, the share of 3' tRFs in the reproductive tract is considerably higher compared to muscle (17% and 10%, respectively, Supplementary Table 1). Recently, 3' tRFs were found to silence Long Terminal Repeat (LTR) retrotransposons in mouse stem cells by targeting their functionally essential and highly conserved primer binding sites [51]. The remarkable amount of 3' tRFs in the analyzed samples supports the idea proposed by Schorn and coworkers who assume that this

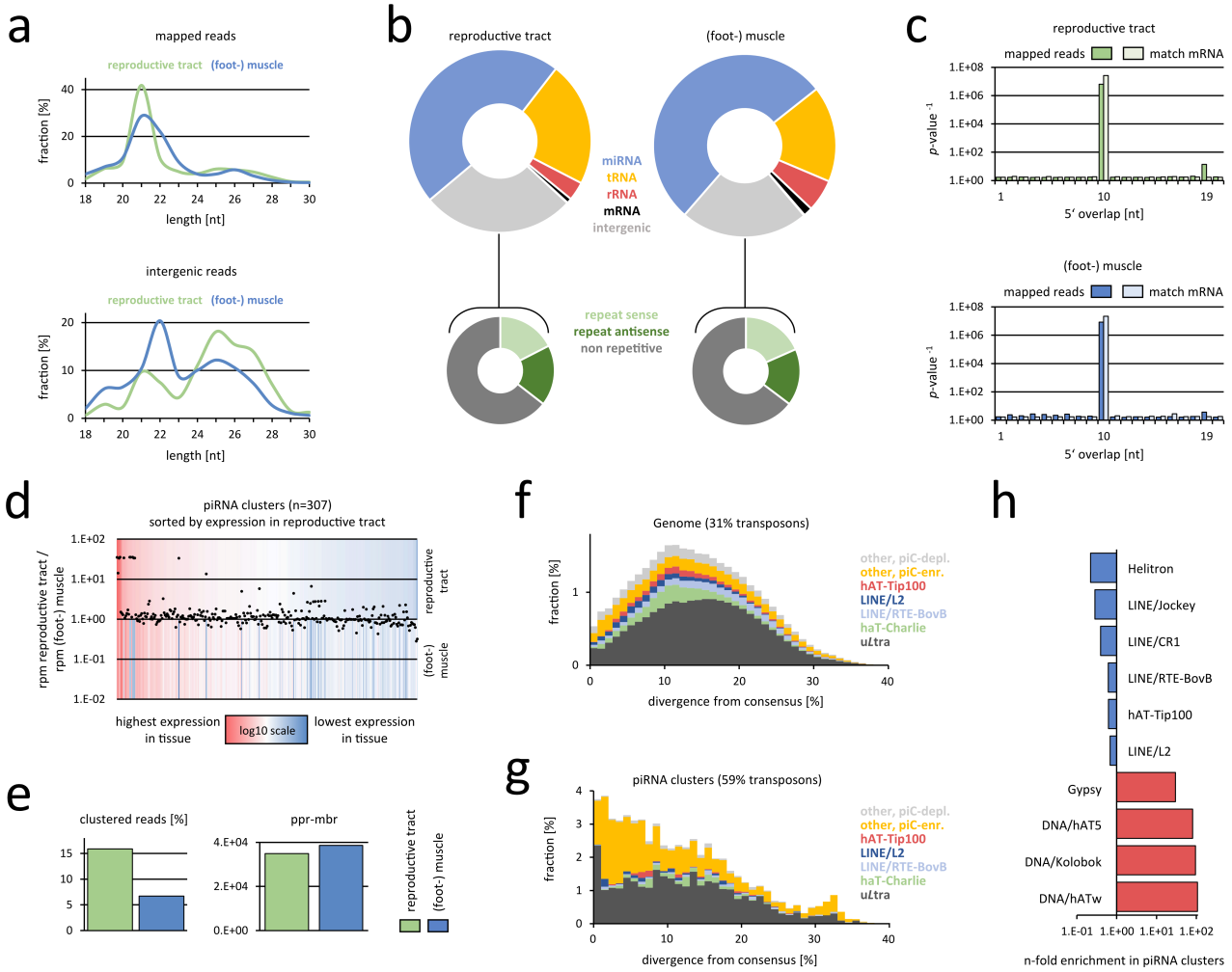


Figure 2 | Characterization of small RNAs from *L. stagnalis* (foot-) muscle and reproductive tract. (a) Sequence read length distribution of mapped (top) and unannotated (intergenic) reads (bottom). (b) Results from mall RNA annotation with unitas (top) and transposon content of intergenic reads (bottom). (c) Ping-pong signature. P-values are deduced from the corresponding Z-scores. P-values for all reads and reads that match mRNA are shown. (d) Differential expression of 307 predicted piRNA clusters. Colors refer to expression relative to highest/lowest expression within one tissue. Dots indicate n-fold expression of a given cluster in reproductive tract relative to muscle. (e) Amount of clustered reads and ping-pong reads per million bootstrapped reads (ppr-mbr). (f) Representation of transposons in the genome of *L. stagnalis*, plotted by divergence [%] from transposon consensus. (g) Representation of transposons within piRNA clusters of *L. stagnalis*, plotted by divergence [%] from transposon consensus. (h) Prominent transposons that are enriched or depleted in *L. stagnalis* piRNA clusters.

mechanism could be highly conserved across different species, providing an innate immunity against LTR propagation.

Focusing on putative piRNAs, we analyzed the fraction of sequence reads that did not match to any other class of non-coding RNA nor mRNA. This dark matter of intergenic sRNAs comprises 27% and 23% of sequence reads in the reproductive tract and in muscle, respectively, and is enriched for transposon sequences, suggesting a role in transposon control (Figure 2b). Analyses of their sequence read length distribution revealed a prominent class of 22 nt molecules in muscle and to a lesser extend in the reproductive tract, suggesting that transposon defense in *L. stagnalis* involves 22 nt siRNAs in addition to piRNAs (Figure 2a). To verify the presence of piRNAs, we checked for the so-called ping-pong signature (bias for 10 bp 5' overlap of mapped sequence reads), which is a hallmark of secondary piRNA biogenesis and requires the catalytic activity - and thus expression - of PIWI proteins [52]. Remarkably, we detected a significant ping-pong signature in both, the reproductive tract and muscle (Figure 2c), suggesting active PIWI/piRNA-dependent transposon silencing in the germline and in the

soma. In addition, a ping-pong signature can also be observed for sequence reads that match protein coding genes, indicating piRNA-dependent gene regulation (Figure 2c).
Next, we used proTRAC [53] to identify 308 piRNA producing loci in the reproductive tract, and 246 piRNA producing loci in muscle tissue. Merging of independently annotated contiguous (<10 kb distance) or overlapping piRNA producing loci revealed a total of 307 distinct piRNA clusters in *L. stagnalis*, covering 0.27% of the genome (Figure 2d, Supplementary Table 2). More precisely, all piRNA producing loci identified in muscle tissue correspond to predicted piRNA clusters based on piRNAs from the reproductive tract, which illustrates that piRNAs in muscle originate from the same set of piRNA clusters compared to the reproductive tract. Nonetheless, there exist 12 clusters whose expression is 14- to 36-fold higher in the reproductive tract compared to muscle tissue, while no clusters show muscle-specific expression to a comparable extent. We found that 15.9% of sequence reads from the reproductive tract map to piRNA clusters, while only 6.7% of sequence reads from muscle do so, indicating rather moderate production of primary piRNAs in the soma compared to the germline (Figure 2e). Besides the presence of primary piRNAs, we found that the number of piRNAs that participate in ping-pong-amplification (measured as ping-pong reads per million bootstrapped reads, ppr-mbr) is slightly higher in muscle (~39k ppr-mbr) compared to the situation in the reproductive tract (~35k ppr-mbr), suggesting higher amounts of secondary piRNAs and emphasizing the functional importance of somatic PIWI/piRNA expression (Figure 2e). In line with the transposon-suppressive role of piRNAs, the identified piRNA clusters show a 2-fold enrichment for transposon sequences compared to the whole genome situation (59% and 31%, respectively, Figure 2f and 2g), whereas only 1.7% of piRNA cluster sequence represents protein coding sequence. Interestingly, the transposon composition in piRNA clusters does not at all reflect the transposon landscape of the genome. Instead, piRNA clusters are enriched for Gypsy retrotransposons and particularly DNA transposons such as Kolobok, hAT5 or hATw showing up to 108-fold enrichment in piRNA clusters (Figs 2g and 2h). This non-random distribution suggests a selective regime that favors insertion events of transposons with low divergence from their consensus sequence, likely representing evolutionary young and active elements.

3.3.4. Ubiquitous and dynamic expression of piRNAs in C. gigas

Based on our observation that PIWI genes and piRNAs are expressed in the soma and the germline of *L. stagnalis*, we reanalyzed previously published small RNA datasets from *C. gigas* that were used to investigate the dynamic expression of miRNAs during oyster development without further examination of a putative piRNA fraction [54] (NCBI Sequence Read Archive Project ID SRP007591). We annotated *C. gigas* sRNAs from the male and female gonad, different developmental stages ranging from the egg to juvenile, and a representative set of somatic tissues from adult animals (Supplementary Table 2). In all datasets, particularly in gonadal tissues, eggs and early embryo stages but also in hemolymph we detected a large amount of sequence reads that did not match to any known ncRNA class but was instead enriched for transposon sequences. The transposon-matching sub-fraction itself was enriched for antisense sequences (Supplementary Table 2). Analogous to the procedure applied for the *L. stagnalis* datasets, we verified the presence of primary and secondary piRNAs by analyzing the ping-pong signature of each dataset. Remarkably, we detected a significant ping-pong signature across all analyzed datasets (Figure 3a, Supplementary Figure 2), but also found that the number of ping-pong reads (measured as ppr-mbr) differs considerably depending on the tissue and developmental stage (Figure 3a, Supplementary Figure 3). Noteworthily, as is the case with *L. stagnalis*, a ping-pong signature is also detectable when taking only those reads into account that match protein coding sequences, suggesting a relevant and conserved role of the PIWI/piRNA pathway in post-transcriptional

regulation of protein coding genes in gonads, egg, blastula, digestive gland and hemolymph (Supplementary Table 3). We further used sequences without ncRNA annotation to predict piRNA clusters with proTRAC and checked whether we can observe a differential expression of specific piRNA clusters in time and space (Figure 3a).

In contrast to the situation in *L. stagnalis*, we found that different genomic loci are responsible for production of primary piRNAs in the germline and in the soma, but also during different

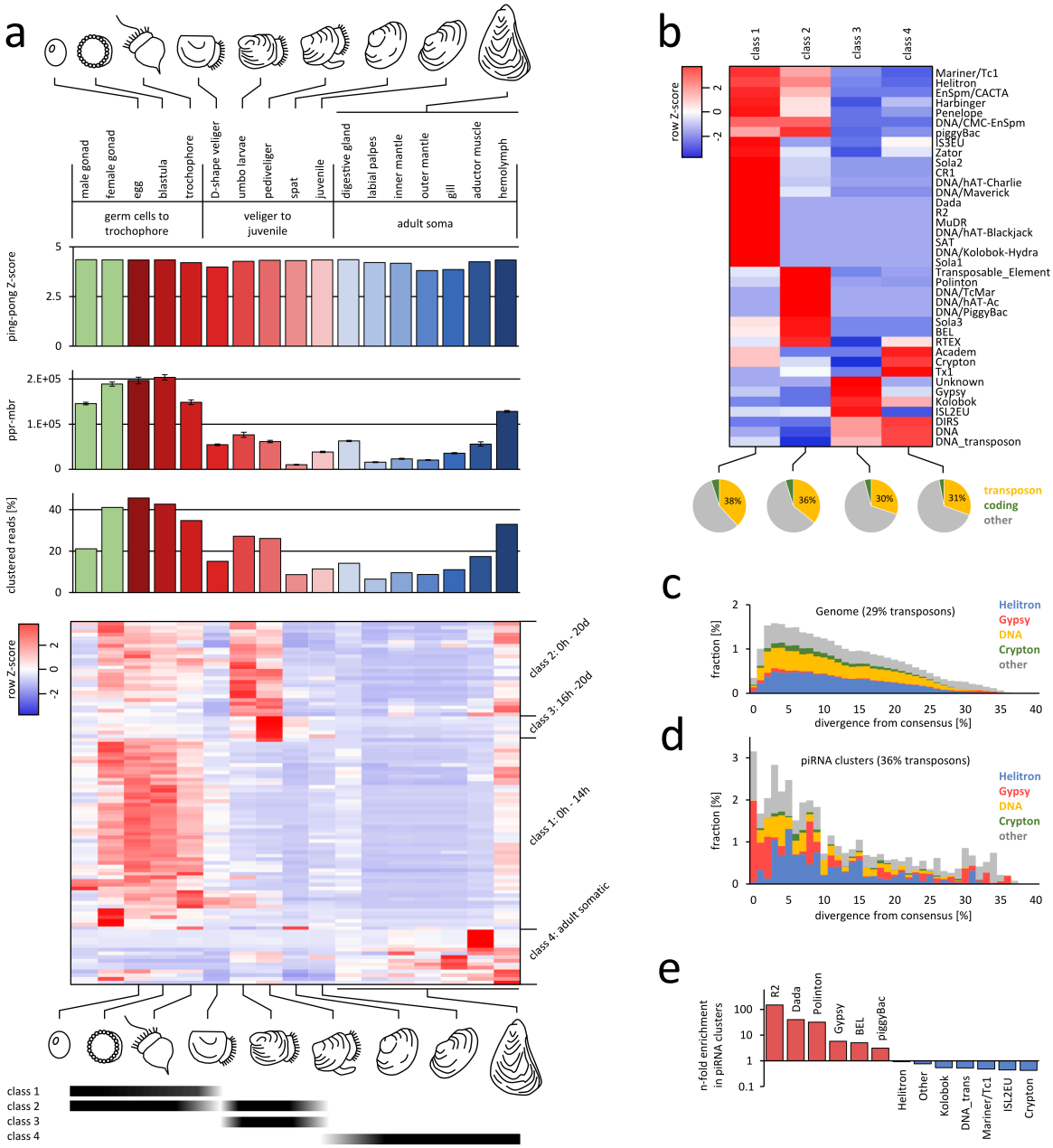


Figure 3 | Characterization of small RNAs and piRNA clusters from different *C. gigas* samples. (a) Sequence reads without annotation produce a significant ping-pong signature (top row of bars, only Z-scores for 10 bp 5' overlap are shown). The number of ping-pong reads per million bootstrapped reads (middle row of bars), and the number of clustered reads (bottom row of bars) differs considerably across the samples. Heatmap shows the differential expression of the top 100 piRNA clusters in terms of maximum rpm coverage. Different classes of piRNA clusters are expressed during oyster development and in adult somatic tissues (bottom). (b) Transposon composition of piRNA clusters belonging to four different classes. (c) Representation of transposons in the genome of *C. gigas*, plotted by divergence [%] from transposon consensus. (d) Representation of transposons within piRNA clusters of *C. gigas*, plotted by divergence [%] from transposon consensus. (e) Prominent transposons that are enriched or depleted in *C. gigas* piRNA clusters. Error bars indicate standard deviation.

developmental stages, which is similar to the situation in the sea anemone *Nematostella vectensis* [18] and the German cockroach *Blatella germanica* [55]. A clustering approach based on average linkage [56] revealed four distinct groups of piRNA clusters which we named class 1-4 piRNA clusters (Figure 3a). Class 1 piRNA clusters are active in the adult germline (male and female) and in the early embryo until the D-shaped veliger stage where larvae are approximately 14 hours old. The same applies to class 2 piRNA clusters, however, following the D-shape veliger stage, class 1 piRNA clusters become inactive, while class 2 piRNA clusters remain active and class 3 piRNA clusters start piRNA production. Both, class 2 and class 3 piRNA cluster activity is measurable until the juvenile stage, where oysters are approximately 20 days old. In somatic tissues of adult oysters, class 4 piRNA clusters represent the main source of primary piRNAs (Figure 3a, bottom). Interestingly, all four classes of piRNA clusters are active in hemocytes, which also feature the highest amount of clustered reads, and ping-pong reads compared to other somatic tissues. This might reflect the presence of stem cells within the hemocyte cell population, which are subject to complex differentiation processes [57,58].

Interestingly, the four classes of piRNA clusters differ considerably regarding the overall transposon content as well as the specific transposon composition (Figure 3b-3d). Class 1 and class 2 piRNA clusters are generally enriched for transposon sequences showing 38% and 36% transposon derived sequences, respectively, compared to a genomic transposon content of 29%. The surprisingly high accumulation of young (as deduced from the divergence from their consensus) Gypsy elements in piRNA clusters, suggests a strong selection for Gypsy element insertions, probably as a consequence of Gypsy activity in *C. gigas*. Noteworthily, the accumulation of young transposons in molluskan piRNA clusters sharply contrasts the situation in *Drosophila* and human, where older transposons are more abundant in piRNA producing loci [59,60] Considering transposons that are generally enriched in piRNA clusters, we found that R2 retrotransposons (149-fold enrichment in piRNA clusters) and Dada DNA transposons (40-fold enrichment in piRNA clusters) are most abundant in class 1 piRNA clusters (Figure 3e). In contrast, Polinton DNA transposons (32-fold enrichment in piRNA clusters) and BEL retrotransposons (5-fold enrichment in piRNA clusters) are most abundant in class 2 piRNA clusters. Different from class 1 and class 2 piRNA clusters, class 3 and class 4 piRNA clusters display only slight transposon enrichment (30% and 31%, respectively). Noteworthily, high copy number Gypsy retrotransposons (5-fold enrichment in piRNA clusters) are most abundant in class 3 piRNA clusters, while Academ, Crypton and Tx1 transposons are most abundant in class 4 piRNA clusters.

The fact that different piRNA clusters are expressed in the germline (class 1 and class 2) and in adult somatic tissues (class 4) of *C. gigas* contrasts with the situation in *L. stagnalis*, where identical piRNA producing loci are active in the germline and in the soma. Moreover, we can observe considerable differences in the transposon composition of piRNA clusters in the two species, which likely reflect a divergent transposon activity in gastropods and bivalves, resulting in varying selective constraints on the different phylogenetic lineages.

3.3.5. Homotypic and heterotypic ping-pong amplification

The ping-pong amplification loop describes a process that is responsible for the post-transcriptional silencing of transposable elements [52]. In *Drosophila* and mouse, this process typically involves two PIWI paralogs (heterotypic ping-pong), one loaded with antisense piRNAs targeting transposon transcripts, and the other loaded with sense piRNAs targeting piRNA cluster transcripts, which contain transposon sequences in antisense orientation [61,62]. Likely for steric reasons, premature piRNAs loaded onto the different PIWI paralogs are more or less rigorously trimmed at their 3' ends. This is why piRNA populations bound to different PIWI paralogs not only differ regarding the amount of sense- and antisense-transposon sequences, but also in their sequence length profiles [52,63,64]. In

addition to the heterotypic ping-pong amplification, homotypic ping-pong has been shown to occur in qin mutant flies (Aub:Aub, [65]), and wildtype prenatal mouse testis (Miwi2:Miwi2, Mili:Mili, [62]).
Since the typical molluskan genome encodes two ubiquitously expressed PIWI paralogs, Piwil1 and Piwil2, we asked whether we can provide evidence for the participation of distinct piRNA populations and PIWI paralogs in the ping-pong cycle. We conducted a bioinformatics approach under the premise that Piwil1- and Piwil2-bound piRNAs exhibit different length profiles, which is the case for the corresponding mouse homologs Piwil1 (Miwi) that preferentially binds 29/30 nt piRNAs, and Piwil2 (Mili) which preferentially binds 26/27 nt piRNAs [66]. A similar, yet not equally pronounced, difference between Piwil1 (Ziwi) and Piwil2 (Zili) -bound piRNAs also exists in zebrafish, suggesting the evolutionary conservation of this pattern [8]. We analyzed pairs of mapped *C. gigas* and *L. stagnalis* sequence reads that showed a 10 bp 5' overlap (ping-pong pairs), with respect to the sequence length of

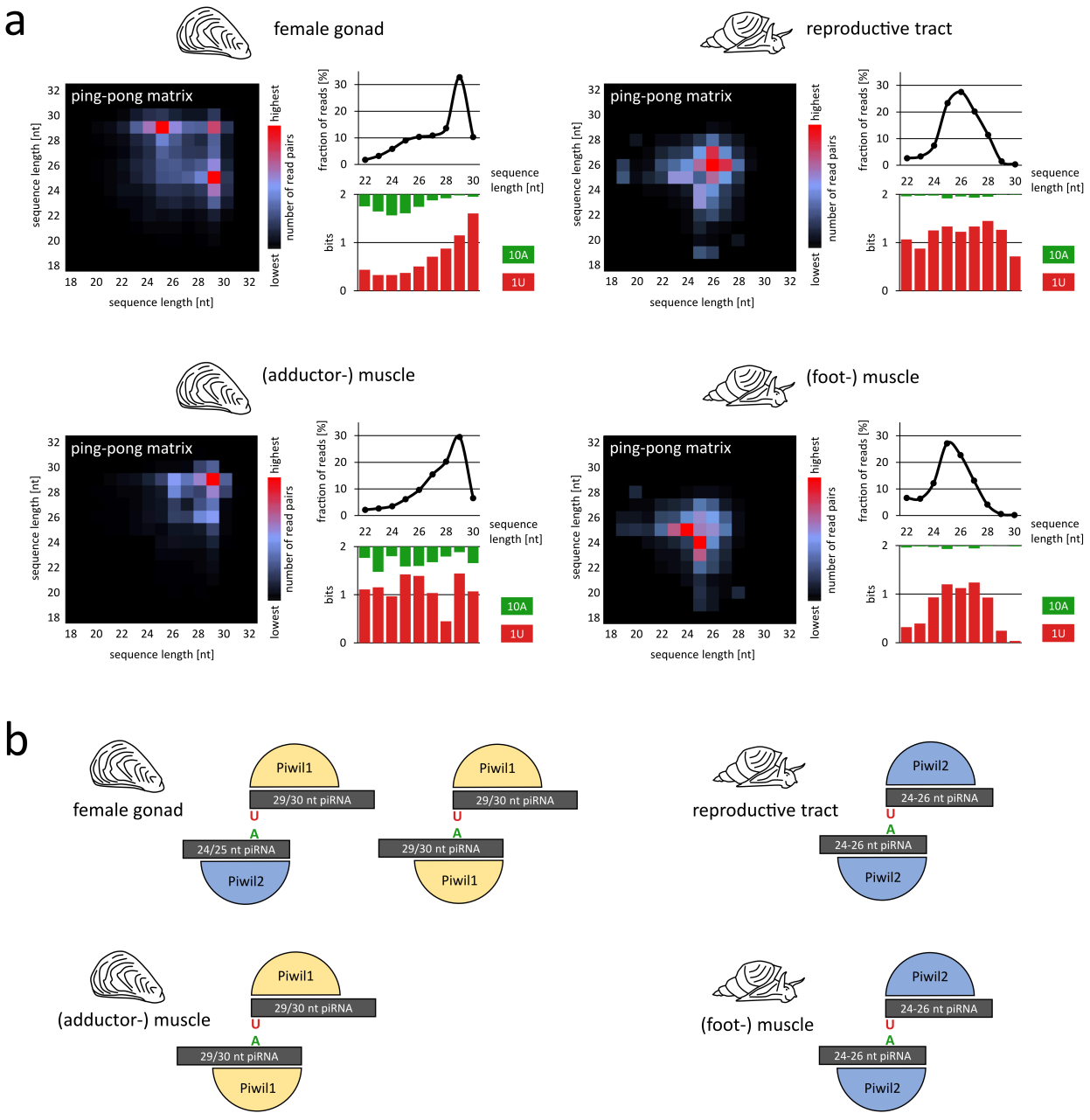


Figure 4 | Analysis of piRNAs that participate in the ping-pong amplification loop. (a) Ping-pong matrices illustrate frequent length-combinations of ping-pong pairs (sequences with 10 bp 5' overlap). Sequence read length distribution and 1U/10A bias [bits] for ping-pong sequences are shown. (b) Proposed model of ping-pong amplification in the germline and muscle of *C. gigas* and *L. stagnalis*.

each ping-pong partner (Figure 4, Supplementary Figure 4). In the female gonad of *C. gigas*, most ping-pong pairs combine piRNAs with a length of 25 nt and 29 nt (Figure 4a), suggesting heterotypic Piwil1-Piwil2-dependent ping-pong amplification as depicted in Figure 4b. In support of this, 29 nt piRNAs, presumably bound to Piwil1, are heavily biased for a 5' uridine (a hallmark of primary piRNAs), whereas 25 nt piRNAs, presumably bound to Piwil2, show a stronger bias for an adenine at position 10 (typical for secondary piRNAs). In contrast, ping-pong pairs in *C. gigas* muscle predominantly combine two 29 nt piRNAs, suggesting homotypic, Piwil1-dependent ping-pong amplification (Figure 4b). Generally, the observed patterns of ping-pong pairs are very diverse across the different samples, for instance displaying heterotypic ping-pong in the digestive gland and homotypic Piwil2-dependent ping-pong in hemolymph cells (Supplementary Figure 4).

Since the expression of Piwil1 compared to Piwil2 is considerably lower in *L. stagnalis*, we were curious to check whether the corresponding ping-pong pairs might reflect this fact. Indeed, 26/26 nt pairs (homotypic, Piwil2-dependent ping-pong) represent the majority of ping-pong pairs in the reproductive tract (Figure 4a). In addition, homotypic Piwil2-dependent ping-pong amplification with 24/25 nt ping-pong pairs is also dominant in the *L. stagnalis* muscle (Figure 4b). However, we also observed differences in ping-pong patterns that do not correlate with the measured mRNA levels of Piwil1 and Piwil2. For example, our data suggests homotypic Piwil2-dependent ping-pong amplification in the oyster gill but homotypic Piwil1-dependent ping-pong amplification in the oyster muscle (Supplementary Figure 4), while both tissues display a very similar expression of both Piwi paralogs on the mRNA level (Figure 1f). Thus, we assume that factors other than mere PIWI expression critically influence characteristics of the ping-pong amplification loop.

Moreover, we clearly cannot rule out the possibility that binding preferences of PIWI paralogs have changed on the molluskan lineage and are different from those observed in fly, fish and mouse. This could mean that length profiles of piRNAs associated to each of the molluskan PIWI paralogs might be exactly reciprocal compared to our presumption. One could even speculate that both PIWI paralogs may bind the whole range of piRNAs, which is not possible to disprove without performing corresponding co-Immunoprecipitation experiments. However, based on the presence of piRNA populations with different length profiles (Figure 2a), their representation in ping-pong pairs together with the differences in their amount of 1U and 10A reads (Figure 4a), we believe that the above made interpretations are a reasonable and parsimonious interpretation of the data at hand, yet not the only possible one.

3.4. Discussion

Our results reveal that mollusks utilize the PIWI/piRNA pathway as a defense against transposable elements in the germline and in the soma, which corresponds to the situation in arthropods and therefore suggests somatic PIWI/piRNA expression to represent a plesiomorphic protostomian character state. In fact, available data from deeper branching metazoans such as poriferans and cnidarians supports the view that this system was established in the soma even long before the split of protostomes and deuterostomes [7,18,41]. In addition, based on the observation that a substantial fraction of arthropod and mollusk piRNAs targets messenger RNAs producing the generic ping-pong signature, it seems likely that the last common ancestor of arthropods and mollusks applied the PIWI/piRNA pathway also for post-transcriptional regulation of protein coding genes. Recently, the Xenacoelomorpha phylum, a group of marine worms that were previously thought to belong to the Platyhelminthes clade, was found to represent the sister group of Nephrozoa which comprise protostomes and deuterostomes [67,68]. Presently, piRNAs for this outgroup are not characterized but having such data would doubtlessly provide valuable insights and allow to draw conclusions regarding

the function of the PIWI/piRNA system in the last common ancestor of all bilaterians, particularly with respect to an ancestral gene-regulatory role. Especially with regard to the latter, functional studies in non-model organisms are urgently needed since the pure bioinformatical evidence for piRNA-dependent processing of protein coding genes does not give any information on its factual biological relevance this process might have in different species. In vertebrates, somatic PIWI/piRNA expression appears to have faded away and reports on somatically expressed piRNAs in mammals are often considered with skepticism for good reasons [69]. However, remnants of the former somatic expression might have outlasted to fulfill special functions in specific cells and/or in narrowly defined timespans of development or cell differentiation in the one or the other clade. In any case, we should be aware that experiments with *Drosophila* and mouse will not tell us everything that is worth knowing about the PIWI/piRNA pathway.

3.5. Material and Methods

3.5.1. Piwi gene annotation and tree reconstruction

In order to reconstruct the phylogenetic relations of mollusk Piwi proteins, we first searched for Piwi genes in species with an available genome sequence that lack proper annotation (*Lymnaea stagnalis*, *Radix auricularia*, *Lottia gigantea*, *Bathymodiolus platifrons*, *Pinctada martensii*). To this end, we scanned the relevant genomes for sequences that are homologous to annotated Piwi paralogs of the pacific oyster (EKC35279 and EKC29295) by aligning translated DNA sequences using tblastx (v2.7.1+, [70]). Neighboring hits with a distance smaller than 10 kb were grouped as exons of distinct gene loci. Only groups containing the overall best hits for a given locus were retained. Finally, the predicted gene sequences were checked for presence of PIWI and PAZ domains using NCBI conserved domain database [71]. Similarly, for Piwi expression analysis by qPCR in the pond snail, we identified the housekeeping gene GPI (glucose-6-phosphate isomerase) by comparison with the human ortholog (ARJ36701).

The predicted and annotated Piwi protein sequences of the 11 available molluskan species together with PIWI paralogs of human (Piwil1-4) and fly (Ago3, Piwi, Aub), as well as fly argonaute Ago1 were aligned using MUSCLE (v.3.8.31, [72]). Subsequently, the resulting protein alignment was curated with Gblocks (v.0.91b), allowing smaller final blocks with gap positions and less strict flanking positions. Using ModelGenerator (v.0.85, [73]) we determined LG+G+F [74] to be the best-fitting model of substitution for our data. The curated alignment (Supplementary Data 1) was then used for phylogenetic tree reconstruction with PhyML (v3.1, [75]) applying approximate likelihood-ratio test (SH-like) and LG substitution model, including empirical gamma distribution (G) and character frequencies (F). Support values were generated by bootstrap with 100 replicates.

3.5.2. qPCR

Experiments were performed on commercially available *C. gigas* animals from the western French Atlantic coast (lle d'Oleron) and captured wild living *L. stagnalis* animals from South-western Germany (Heppenheim). To estimate the expression of the Piwil homologs in several tissues of *L. stagnalis* and *C. gigas* we performed qPCR with cDNA synthesized from the total RNA fraction of these tissues. Total RNA was isolated with TriReagent and the polyadenylated transcriptome was reversely transcribed with SuperScript IV using the RT-primer 5'-CGAATTCTAGAGCTCGAGGCAGGCGA-CATGT25VN-3'. Primers amplifying ~ 200 bp long products of the respective Piwil homologs and housekeeping genes were designed with the NCBI tool primer-BLAST on basis of the *L. stagnalis* genome assembly GCA_900036025.1 v1.0 and the *C. gigas* genome assembly GCA_000297895.1 oyster_v9. To prevent amplification of residual genomic DNA, primers were designed to be exon-

junction spanning or to span at least several intronic regions. The respective biological replicates were analyzed as technical duplicates on a Corbett Rotor-Gene 6000 real-time PCR cycler and the copy numbers of the genes of interest were quantified by standard curves of the individual primer pair amplicons. For each cDNA sample the calculated PIWI copy numbers were relativized by the calculated copy numbers of the housekeeping genes to calibrate for variabilities in sample preparation. These n-fold expression values were finally used to calculate the mean and standard deviation of the replicates. For an improved visualization, the n-fold expression values of each Piwi homolog are additionally displayed as a percentage of the respective gonad value.

3.5.3. Small RNA extraction and sequencing

We extracted total RNA from *L. stagnalis* reproductive tract (incl. ovotestis, oviduct, spermatheca, spermiduct, prostate, uterus, vagina, vas deferens) and foot muscle, and total RNA from *C. gigas* adductor muscle and gonadal tissue with TriReagent according to the manufacturer's instructions. For each species we sampled two different individuals per tissue. The small RNA fractions of each obtained total RNA sample were sequenced at BGI, Hong Kong, on a BGISEQ-500 unit. Small RNA sequence datasets for *L. stagnalis* and *C. gigas* are deposited at NCBI's Sequence Read Archive (SRA) and can be accessed under the SRA project IDs SRP130729 and SRP130745. We further used previously published small RNA sequence data from *C. gigas* [54] to analyze piRNA expression and characteristics with respect to different developmental stages.

3.5.4. Repeat annotation

We performed *de novo* prediction of repetitive elements in the genome of *L. stagnalis* with RepeatScout (v. 1.0.5, [76]). Predicted repetitive elements were classified with RepeatClassifier which is part of the RepeatModeler (v. 1.0.11) package. Transposons that failed to be classified based on known transposons from other species are referred to as unclassified *Lymnaea*-specific transposons (uLtra). The resulting repeat sequences, as well as a complete collection of currently available molluskan repeat sequences from RepBase [77] were used as reference sequences for repeat masking of the *L. stagnalis* and *C. gigas* genomes with RepeatMasker (v. 4.0.7) using the cross_match search engine and the option -s for most sensitive masking. Annotated repeats in the RepeatMasker output were analyzed with respect to transposon families and divergence from their consensus sequence using the Perl script TE_landscape.pl. Analysis was conducted with the entire repeat dataset as well as with repeats localized in predicted piRNA clusters. TE_landscape.pl is freely available at https://sourceforge.net/projects/protrac/files/tools/.

3.5.5. Gene annotation

We performed *de novo* gene annotation of the *L. stagnalis* genome assembly gLs_1.0 [78] using the MAKER genome annotation pipeline (v.2.31.8) in order to identify sRNAs that match protein-coding sequences [79]. Initially, we masked the *L. stagnalis* genome with WindowMasker [80] using default settings including the duster option to mask low complexity regions. Then, we used available molluskan cDNA data from Ensembl database (release 92) and available mRNA and protein data from *L. stagnalis* deposited at NCBI (Effective April 25, 2018) as input for MAKER. MAKER output files for separate scaffolds were merged using the Perl script mergeMAKERoutput.pl which is freely available at https://sourceforge.net/projects/protrac/files/tools/. The complete genome annotation in GFF3 format and a corresponding mRNA sequence file in FASTA format are available as Supplementary Data 2 and Supplementary Data 3.

3.5.6. Processing and annotation of small RNA sequence data

Small RNA sequence datasets were collapsed to non-identical sequences, retaining information on sequence read counts using the Perl script *collapse*. Sequences >36nt were rejected using the Perl script *length-filter*. Finally, low complexity sequences were filtered using the Perl script *duster* with default parameters. All Perl scripts mentioned are part of the NGS toolbox [81].

We then applied a customized mapping strategy of the remaining small RNA sequence reads based on the consideration that our datasets presumably contain considerable amounts of transposon-derived piRNAs as well as post-transcriptionally edited (e.g. A-to-I) or tailed miRNAs and piRNAs. Genomic mapping was performed with SeqMap [82] using the option /output_all_matches and allowing up to three mismatches. The obtained alignments were further filtered using the Perl script seqmap_filter.pl that is freely available at https://sourceforge.net/projects/protrac/files/tools/. For the final alignments we allowed up to two non-template 3' nucleotides and up to one internal mismatch. For each sequence, we only considered the best alignments in terms of mismatch counts, but did not reject alignments with equal quality in case of multiple mapping sequences. Sequences that did not produce at least one valid alignment to the reference genome were rejected.

To improve small RNA sequence annotation, we performed *de novo* tRNA, rRNA and miRNA prediction based on the available reference genome assemblies gLs_1.0 (*L. stagnalis*) and GCA_000297895.1 oyster_v9 (*C. gigas*). tRNA annotation was performed with a local copy of tRNAscan (v.1.3.1, [83]). Only tRNAs with less than 5% N's were taken for further analysis. rRNA sequences were predicted using a local copy of RNAmmer (v.1.2, [84]) and hmmer (v.2.2g, [85]). Both tools were run with default parameters. We pooled small RNA sequence reads from different replicates and tissues for each species separately to perform miRNA *de novo* prediction with ShortStack (v.3.8.4, [86]) using default parameters. The predicted tRNA, rRNA and miRNA precursor sequences, as well as previously published miRNA precursor sequences [54,87,88], were used as additional reference sequences for small non-coding RNA annotation with unitas (v.1.4.6, [50]) which was run with the option -riborase. For *L. stagnalis*, we also included predicted cDNA data based on MAKER annotation (see above). sRNA sequences that did not match to any ncRNA or mRNA of *C. gigas* or *L. stagnalis* were blasted against NCBI nucleotide collection (nr) to search for possible contaminants of parasitic species. Sequences that produced better alignments to genomes of species that possibly parasitized the sampled individuals (*Dicrocoelium*, *Legionella*, *Panagrellus*, *Thelazia*, *Trichobilharzia*) were considered as contaminants and not used for downstream analyses.

3.5.7. piRNA cluster identification

Sequences that did not produce a match to known non-coding RNAs were considered as putative piRNAs and were used for prediction of piRNA clusters with proTRAC (v. 2.4.0, [53]) applying default settings. piRNA clusters were predicted for each dataset and species separately. The resulting piRNA cluster predictions for each species were condensed, merging clusters with less than 10 kb distance from each other using the Perl script merge_clusters which is freely available at https://sourceforge.net/projects/protrac/files/tools/. To preclude false positive annotation of e.g. tRNA or rRNA genes as piRNA clusters, we validated predicted piRNA clusters by analyzing sRNA reads that mapped to them with respect to their relation to mRNA or other ncRNA classes (Supplementary Figure 5a). To further check whether piRNA cluster calling may under- or overestimate the number of primary piRNAs in our datasets, we performed an arithmetical approach to estimate the fraction of genuine primary piRNAs based on the fraction of 5' U reads in annotated and non-annotated reads with 24-29 nt length which yields results very close to the number of clustered reads (Supplementary Methods, Supplementary Figure 5b). We calculated the sequence read coverage [rpm] for each of the

resulting piRNA clusters per dataset. For *C. gigas* piRNA clusters, a heat map for the top 100 piRNA clusters in terms of maximum rpm coverage (accounting for 64% of summed rpm values) was constructed with Heatmapper [56] applying Pearson distance and average linkage clustering. Finally, predicted piRNA clusters were analyzed with respect to their repeat and gene content using the Perl script piC_content.pl which is freely available at https://sourceforge.net/projects/protrac/files/tools/.

3.5.8. Ping-pong quantification

In order to compare ping-pong signatures across multiple datasets with different sequencing depth, we constructed a software tool, PPmeter (v.0.4), that creates bootstrap pseudo-replicates from original datasets and subsequently analyzes the ping-pong signature and number of ping-pong sequence reads of each pseudo-replicate (default: 100 pseudo-replicates each comprising one million sequence reads). The obtained parameters 'ping-pong score per million bootstrapped reads' (pps-mbr) and 'ping-pong reads per million bootstrapped reads' (ppr-mbr) can be used for quantification and direct comparison of ping-pong activity in different small RNA datasets. The software is freely available at http://www.smallRNAgroup.uni-mainz.de/software.html and https://sourceforge.net/projects/protrac/files/tools/.

3.5.9. Data availability

Sequence data have been uploaded to NCBI's Sequence Read Archive and can be accessed via the accessions SRP130729 and SRP130745.

3.5.10. Code availability

Source code of software that has been written for data processing and analysis is freely available at https://sourceforge.net/projects/protrac/files/tools/.

3.6. Declarations

Acknowledgements

We thank Sacha Heerschop, Julia Schumacher, Isabel Fast and Hans Zischler for helpful comments and discussion. Thanks go also to the Mark Helm group for kindly providing chemicals. This work was supported by the International PhD Program (IPP) coordinated by the Institute of Molecular Biology IMB, Mainz, Germany, funded by the Boehringer Ingelheim Foundation.

Author contributions

JJ and JSTK performed total RNA extraction and qPCR experiments. JJ analyzed qPCR data and prepared the corresponding figures. **DG** identified PIWI paralogs in sequenced but unannotated molluskan genomes. **DG** performed PIWI gene tree reconstruction and prepared the corresponding figure. FP was responsible for farming and dissection of *L. stagnalis* animals. FP and SS performed RNA extraction for subsequent sRNA sequencing. CH, FP and SS performed de novo miRNA annotation based on the obtained sRNA data. **DG** performed bioinformatics analysis of piRNA clusters and prepared the corresponding figures. DR analyzed sRNA data and developed Perl scripts for data analysis. **DG**, JJ and DR wrote the manuscript. JSTK, FP and CH provided valuable input for corrections and improvements of the manuscript.

The authors declare no competing interests.

3.7. References

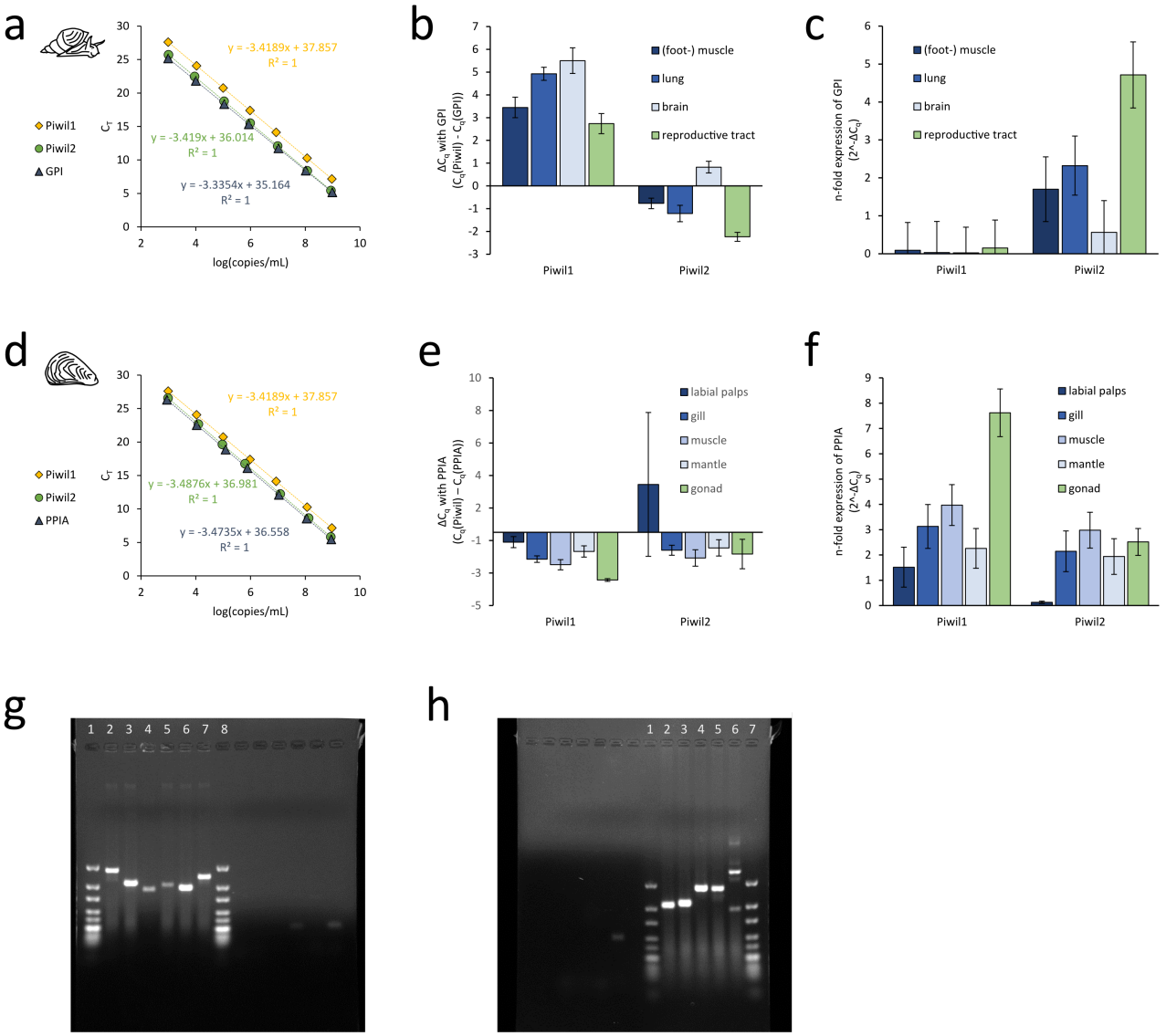
1. Thomson, T. & Lin, H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annu Rev Cell Dev Biol. 25, 355-376 (2009).
2. Iwasaki, Y. W., Siomi, M.C. & Siomi, H. PIWI-Interacting RNA: Its Biogenesis and Functions. Annu Rev Biochem. 84, 405-433 (2015).
3. Reuter, M. et al. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature. 480, 264-267 (2011).
4. Di Giacomo, M. et al. Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis. Mol Cell. 50, 601-608 (2013).
5. Pezic, D., Manakov, S. A., Sachidanandam, R. & Aravin, A. A. piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. Genes Dev. 28, 1410-1428 (2014).
6. Manakov, S. A. et al. MIWI2 and MILI Have Differential Effects on piRNA Biogenesis and DNA Methylation. Cell Rep. 12, 1234-1243 (2015).
7. Grimson, A. et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature. 455, 1193-1197 (2008).
8. Houwing, S., Berezikov, E. & Ketting, R. F. Zili is required for germ cell differentiation and meiosis in zebrafish. EMBO J. 27, 2702-2011 (2008).
9. Das, P. P. et al. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. Mol Cell. 31, 79-90 (2008).
10. Li, X. Z. et al. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell. 50, 67-81 (2013).
11. Lim, R. S., Anand, A., Nishimiya-Fujisawa, C., Kobayashi, S. & Kai, T. Analysis of Hydra PIWI proteins and piRNAs uncover early evolutionary origins of the piRNA pathway. Dev Biol. 386, 237-251 (2014).
12. Ha, H. et al. A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. BMC Genomics. 15, 545; 10.1186/1471-2164-15-545 (2014).
13. Hirano, T. et al. Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. RNA. 20, 1223-1237 (2014).
14. Gebert, D., Ketting, R. F., Zischler, H., Rosenkranz, D. piRNAs from Pig Testis Provide Evidence for a Conserved Role of the Piwi Pathway in Post-Transcriptional Gene Regulation in Mammals. PLoS One. 10, e0124860; 10.1371/journal.pone.0124860 (2015).
15. Roovers, E. F. et al. Piwi proteins and piRNAs in mammalian oocytes and early embryos. Cell Rep. 10, 2069-2082 (2015).
16. Rosenkranz, D., Rudloff, S., Bastuck, K., Ketting, R. F., Zischler, H. Tupaia small RNAs provide insights into function and evolution of RNAi-based transposon defense in mammals. RNA. 21, 911-922 (2015).
17. Madison-Villar, M. J., Sun, C., Lau, N. C., Settles, M. L., Mueller, R. L. Small RNAs from a Big Genome: The piRNA Pathway and Transposable Elements in the Salamander Species Desmognathus fuscus. J Mol Evol. 83, 126-136 (2016).
18. Praher, D. et al. Characterization of the piRNA pathway during development of the sea anemone Nematostella vectensis. RNA Biol. 14, 1727-1741 (2017).
19. Lewis, S. H. et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol. 2, 174-181 (2018).
20. Flemr, M. et al. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. Cell. 155, 807-816 (2013).
21. Zhang, P. et al. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. Cell Res. 25, 193-207 (2015).
22. Russell, S. et al. Bovine piRNA-like RNAs are associated with both transposable elements and mRNAs. Reproduction. 153, 305-318 (2017).
23. Rouget, C. et al. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. Nature. 467, 1128–32 (2010).
24. Gou, L.-T. et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. 24, 680–700 (2014).
25. Watanabe, T. & Lin, H. Posttranscriptional Regulation of Gene Expression by Piwi Proteins and piRNAs. Mol Cell. 56, 18–27 (2014).
26. Barckmann, B. et al. Aubergine iCLIP Reveals piRNA-Dependent Decay of mRNAs Involved in Germ Cell Development in Germ Cell Development in the Early Embryo. Cell Rep. 12, 1205–1216 (2015).
27. Rojas-Ríos, P., Chartier, A., Pierson, S. & Simonelig, M. Aubergine and piRNAs promote germline stem cell self-renewal by repressing the proto-oncogene Cbl. EMBO J. 36, 3194– 3211 (2017).
28. Palakodeti, D., Smielewska, M., Lu, Y. C., Yeo, G. W. & Graveley, B. R. The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. RNA. 14, 1174-1186 (2008).
29. Perrat, P. N. et al. Transposition-driven genomic heterogeneity in the Drosophila brain. Science. 340, 91-95 (2013).
30. Nandi, S. et al. Roles for small noncoding RNAs in silencing of retrotransposons in the mammalian brain. Proc Natl Acad Sci U S A. 113, 12697-12702 (2016).

31. Jones, B. C. et al. A somatic piRNA pathway in the Drosophila fat body ensures metabolic homeostasis and normal lifespan. Nat Commun. 7, 13856; 10.1038/ncomms13856 (2016).
32. Teixeira, F. K. et al. piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. Nature. 552, 268-272 (2017).
33. Ross, R. J., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. Nature. 505, 353-359 (2014).
34. Juliano, C. E. et al. PIWI proteins and PIWI-interacting RNAs function in Hydra somatic stem cells. Proc Natl Acad Sci U S A. 111, 337–342 (2013).
35. Funayama, N., Nakatsukasa, Ã. M., Mohri, K., Masuda, Y., & Agata, K. Piwi expression in archeocytes and choanocytes in demosponges: insights into the stem cell system in demosponges. Evol Dev. 12, 275–287 (2010).
36. Miesen, P., Girardi, E. & van Rij, R. P. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. Nucleic Acids Res. 43, 6545-6556 (2015).
37. Kiuchi, T. et al. A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature. 509, 633-636 (2014).
38. Rajasethupathy, P. et al. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. Cell. 149, 693-707 (2012).
39. Rosenberg, G. A New Critical Estimate of Named Species-Level Diversity of the Recent Mollusca. American Malacological Bulletin. 32, 308-322 (2014).
40. Ma, X. et al. Piwi1 is essential for gametogenesis in mollusk Chlamys farreri. PeerJ. 5, e3412; 10.7717/peerj.3412 (2017).
41. Waldron, F. M., Stone, G. N. & Obbard, D. J. Metagenomic sequencing suggests a diversity of RNA interference-likeesponses to viruses across multicellular eukaryotes. BioRxiv https://www.biorxiv.org/content/early/2018/03/20/166488 (2018).
42. Zhou, X., Liao, Z., Jia, Q., Cheng, L. & Li, F. Identification and characterization of Piwi subfamily in insects. Biochem Biophys Res Commun. 362, 126-131 (2007).
43. Schurko, A. M., Logsdon, J. M. Jr. & Eads, B. D. Meiosis genes in Daphnia pulex and the role of parthenogenesis in genome evolution. BMC Evol Biol. 9, 78; 10.1186/1471-2148-9-78 (2009).
44. Lewis, S. H., Salmela, H. & Obbard, D. J. Duplication and Diversification of Dipteran Argonaute Genes, and the Evolutionary Divergence of Piwi and Aubergine. Genome Biol Evol. 8, 507-518 (2016).
45. Kerner, P., Degnan, S. M., Marchand, L., Degnan, B. M. & Vervoort, M. Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge Amphimedon queenslandica. Mol Biol Evol. 28, 2289-2303 (2011).
46. Dowling, D. et al. Phylogenetic Origin and Diversification of RNAi Pathway Genes in Insects. Genome Biol Evol. 8, 3784-3793 (2016).
47. Sasaki, T., Shiohama, A., Minoshima, S. & Shimizu, N. Identification of eight members of the Argonaute family in the human genome. Genomics. 82, 323-330 (2003).
48. Murchison, E. P. et al. Conservation of small RNA pathways in platypus. Genome Res. 18, 995-1004 (2008).
49. Tong, Y. et al. Transcriptomics Analysis of Crassostrea hongkongensis for the Discovery of Reproduction-Related Genes. PLoS One. 10, e0134280; 10.1371/journal.pone.0134280 (2015).
50. Gebert, D., Hewel, C. & Rosenkranz, D. unitas: the universal tool for annotation of small RNAs. BMC Genomics. 18, 644; 10.1186/s12864-017-4031-9 (2017).
51. Schorn, A. J., Gutbrod, M. J., LeBlanc, C., Martienssen, R. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell. 170, 61-71 (2017).
52. Czech, B., Hannon, G. J. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci. 41, 324-337 (2016).
53. Rosenkranz, D. & Zischler, H. proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics. 13, 5; 10.1186/1471-2105-13-5 (2012).
54. Xu, F. et al. Identification of conserved and novel microRNAs in the Pacific oyster Crassostrea gigas by deep sequencing. PLoS One. 9, e104371; 10.1371/journal.pone.0104371 (2014).
55. Llonga, N., Ylla, G., Bau, J., Belles, X., & Piulachs, M. Diversity of piRNA expression patterns during the ontogeny of the German cockroach. J Exp Zool B Mol Dev Evol. 10.1002/jez.b.22815 (2018).
56. Babicki, S. et al. Heatmapper: web-enabled heat mapping for all. Nucleic Acids Res. 44, W147-153 (2016).
57. Fisher, W. S. Structure and Functions of Oyster Hemocytes in Immunity in Invertebrates (ed. Brehélin, M.) 25-35 (Springer, 1986).
58. Lau, Y. T., Sussman, L., Pales Espinosa, E., Katalay, S., Allam, B. Characterization of hemocytes from different body fluids of the eastern oyster Crassostrea virginica. Fish Shellfish Immunol. 71, 372-379 (2017).
59. Senti, K.A., Jurczak, D., Sachidanandam, R., Brennecke, J. piRNA-guided slicing of transposon transcripts enforces their transcriptional silencing via specifying the nuclear piRNA repertoire. Genes Dev. 29, 1747-1762 (2015).
60. Gainetdinov, I., Skvortsova, Y., Kondratieva, S., Funikov, S., Azhikina, T. Two modes of targeting transposable elements by piRNA pathway in human testis. RNA 23, 1614-1625 (2017).
61. Brennecke, J. et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 128, 1089-1103 (2007).

62. Aravin, A. A. et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 31, 785-799 (2008).
63. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science. 318, 761-764 (2007).
64. Kawaoka, S., Izumi, N., Katsuma, S. & Tomari, Y. 3' end formation of PIWI-interacting RNAs in vitro. Mol Cell. 43, 1015-1022 (2011).
65. Zhang, Z. et al. Heterotypic piRNA Ping-Pong requires qin, a protein with both E3 ligase and Tudor domains. Mol Cell. 44, 572-584 (2011).
66. Vourekas, A. et al. Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nat Struct Mol Biol. 19, 773-781 (2012).
67. Cannon, J. T. et al. Xenacoelomorpha is the sister group to Nephrozoa. Nature. 530, 89–93 (2016).
68. Rouse, G. W., Wilson, N. G., Carvajal, J. I., & Vrijenhoek, R. C. New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. Nature. 530, 94–97 (2016).
69. Tosar, J. P., Rovira, C., & Cayota, A. Non-coding RNA fragments account for the majority of annotated piRNAs expressed in somatic non-gonadal tissues. Communications Biology, 1, 2; 10.1038/s42003-017-0001-7 (2018).
70. Camacho, C. et al. BLAST+: Architecture and applications. BMC Bioinformatics. 10, 421; 10.1186/1471-2105-10-421 (2009).
71. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43, D222-226 (2015).
72. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5, 113; 10.1186/1471-2105-5-113 (2004).
73. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., & McInerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol, 6, 29; 10.1186/1471-2148-6-29 (2006).
74. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. Mol Biol Evol. 25, 1307–1320 (2008).
75. Guindon, S., Delsuc, F., Dufayard, J. F., Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537, 113-137 (2009).
76. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics. 21, 351-358 (2005).
77. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6, 11; 10.1186/s13100-015-0041-9 (2015).
78. Davison, A. et al. Formin Is Associated with Left-Right Asymmetry in the Pond Snail and the Frog. Curr Biol. 26, 654-660 (2016).
79. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188-196 (2008).
80. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 22,134-141 (2006).
81. Rosenkranz, D., Han, C. T., Roovers, E. F., Zischler, H., Ketting, R. F. Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. Genom Data. 5, 309-313 (2015).
82. Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 24, 2395-2396 (2008).
83. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 44, W54-57 (2016).
84. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100-3108 (2007).
85. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 11, 431; 10.1186/1471-2105-11-431 (2010).
86. Axtell, M. J. ShortStack: comprehensive annotation and quantification of small RNA genes. RNA. 19, 740-751 (2013).
87. Zhou, Z. et al. The identification and characteristics of immune-related microRNAs in haemocytes of oyster Crassostrea gigas. PLoS One. 9, e88397; 10.1371/journal.pone.0088397 (2014).
88. Zhao, X., Yu, H., Kong, L., Liu, S. & Li, Q. High throughput sequencing of small RNAs transcriptomes in two Crassostrea oysters identifies microRNAs involved in osmotic stress response. Sci Rep. 6, 22687; 10.1038/srep22687 (2016).

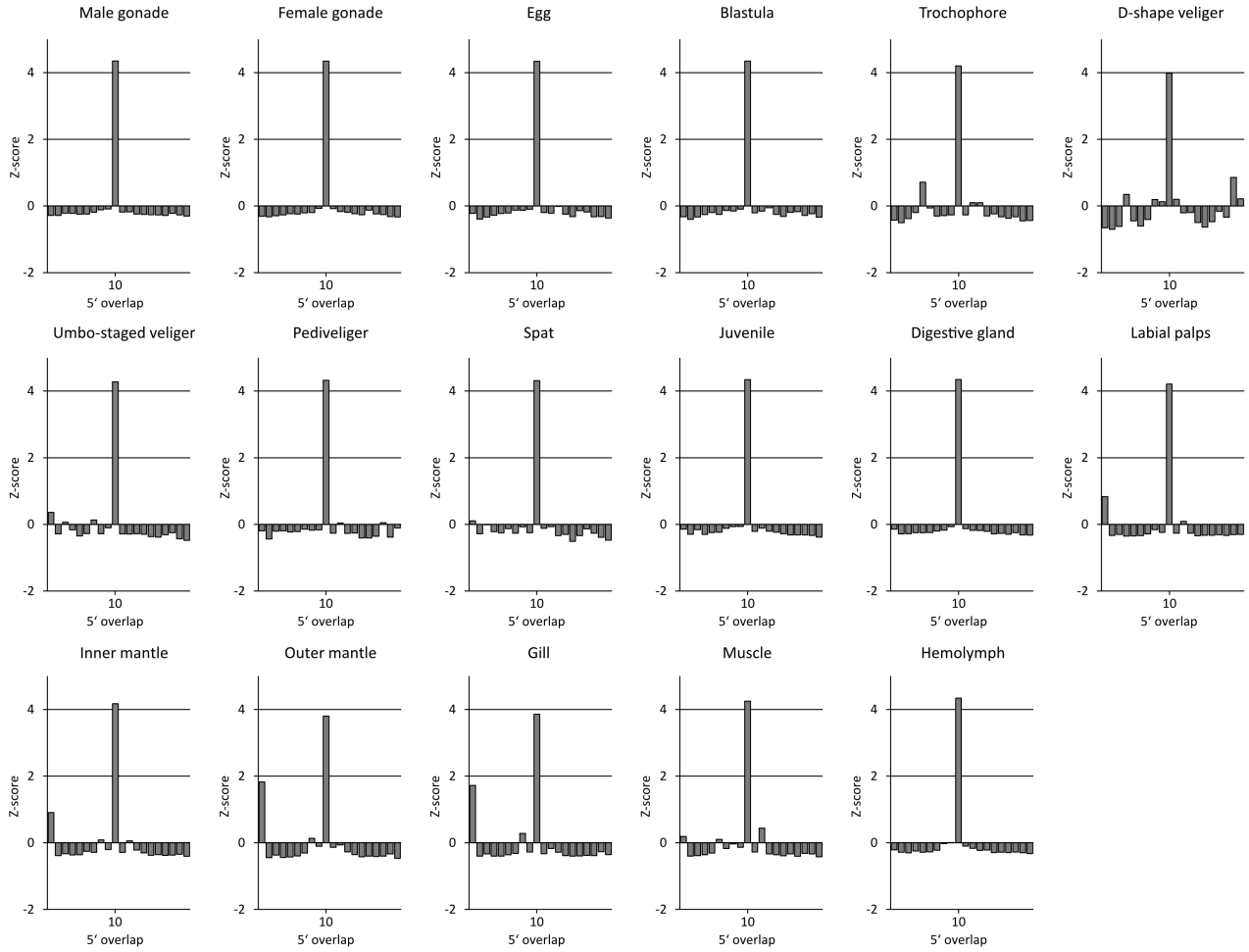
3.8. Supplement

Supplementary figure 1



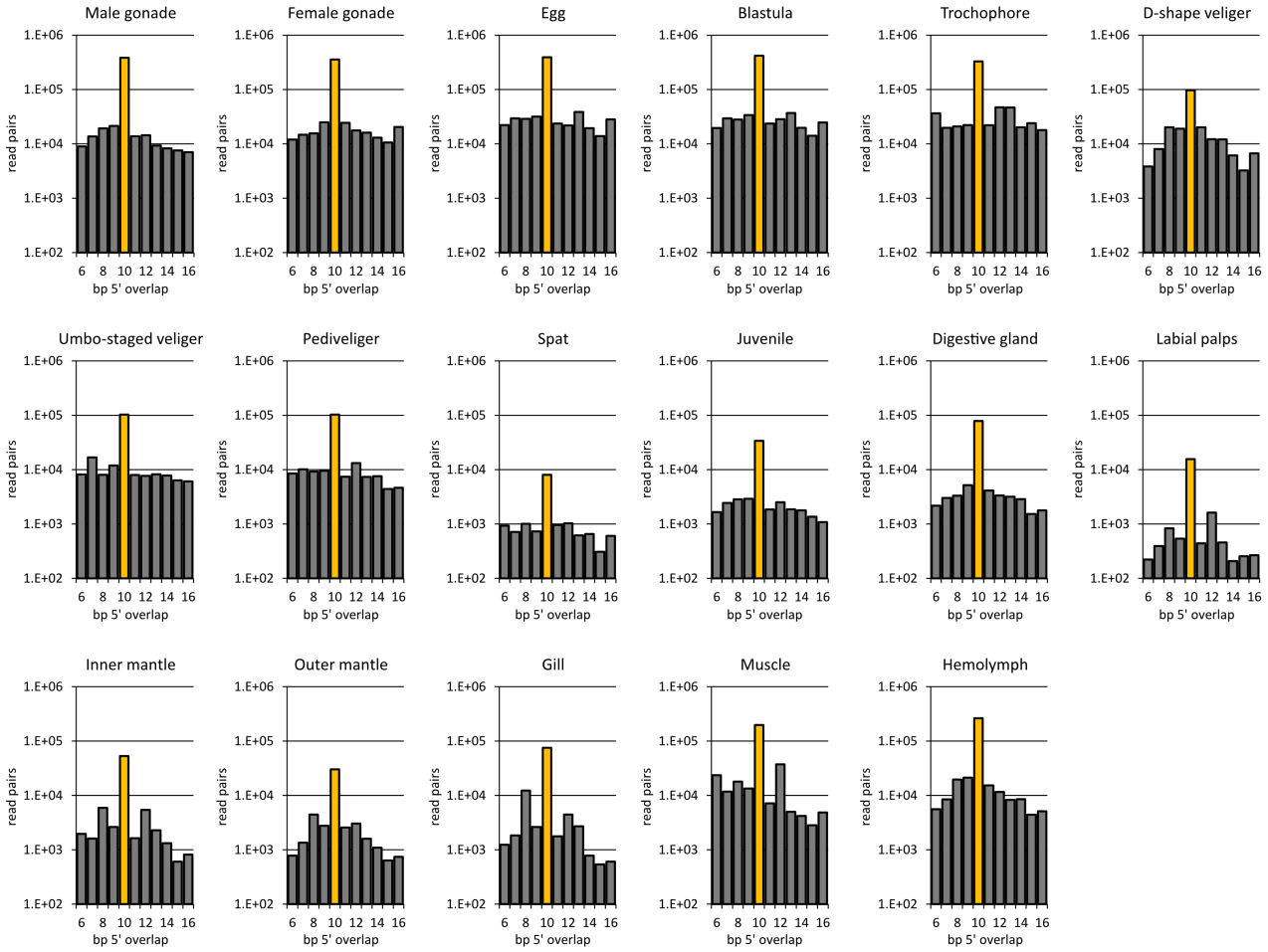
Supplementary figure 1 | RT-qPCR quantification of PIWI paralog expression using the standard curve method and the Cq method. (a) Standard curves used for absolute quantification of the PIWI homolog transcripts and the transcripts of the housekeeping gene glucose phosphate isomerase (GPI) in different tissues of *L. stagnalis*. The GPI copy numbers of each sample were used to calibrate the copy numbers of the PIWI paralogs for variabilities in sample preparation as shown in Figure 1c and 1d. (b) Relative expression of the PIWI paralog transcripts in *L. stagnalis* as calculated by the Cq method, where Cq = Cq (PIWI paralog) - Cq (GPI). Higher Cq values represent lower PIWI expression. (c) Relative expression of the PIWI paralog transcripts in *L. stagnalis* as determined by the Cq method. (d) Standard curves used for absolute quantification of the PIWI homolog transcripts and the transcripts of the housekeeping gene peptidylprolyl isomerase A (PPIA) in different tissues of *C. gigas*. The PPIA copy numbers of each sample were used to calibrate the copy numbers of the PIWI paralogs for variabilities in sample preparation as shown in Figure 1f and 1g. (e) Relative expression of the PIWI paralog transcripts in *C. gigas* as calculated by the Cq method, where Cq = Cq (PIWI paralog) - Cq (PPIA). Higher Cq values represent lower PIWI expression. Error bars indicate standard deviation. (f) Relative expression of the PIWI paralog transcripts in *C. gigas* as determined by the Cq method. (g) Control PCR with PIWI paralog specific primers and *L. stagnalis* cDNA from the reproductive tract. Complete gel from figure 1b. Probes in lanes from 1-8 are: Ultra Low Range DNA Ladder (Thermo Scientific), Piwil1 amplicon, Piwil2 amplicon, Piwil1b amplicon, Piwil1c amplicon, GPI amplicon, EMC7 amplicon, Ultra Low Range DNA Ladder. (h) Control PCR with PIWI paralog specific primers and *C. gigas* cDNA from adductor muscle. Complete gel from figure 1e. Probes in lanes from 1-7 are: Ultra Low Range DNA Ladder (Thermo Scientific), Piwil1 amplicon (primers target both annotated Piwil1 splice isoforms), Piwil1 amplicon (primers target the 18-exon Piwil1 splice isoforms), Piwil2b amplicon, PPIA amplicon, TATA amplicon, Ultra Low Range DNA Ladder.

Supplementary figure 2



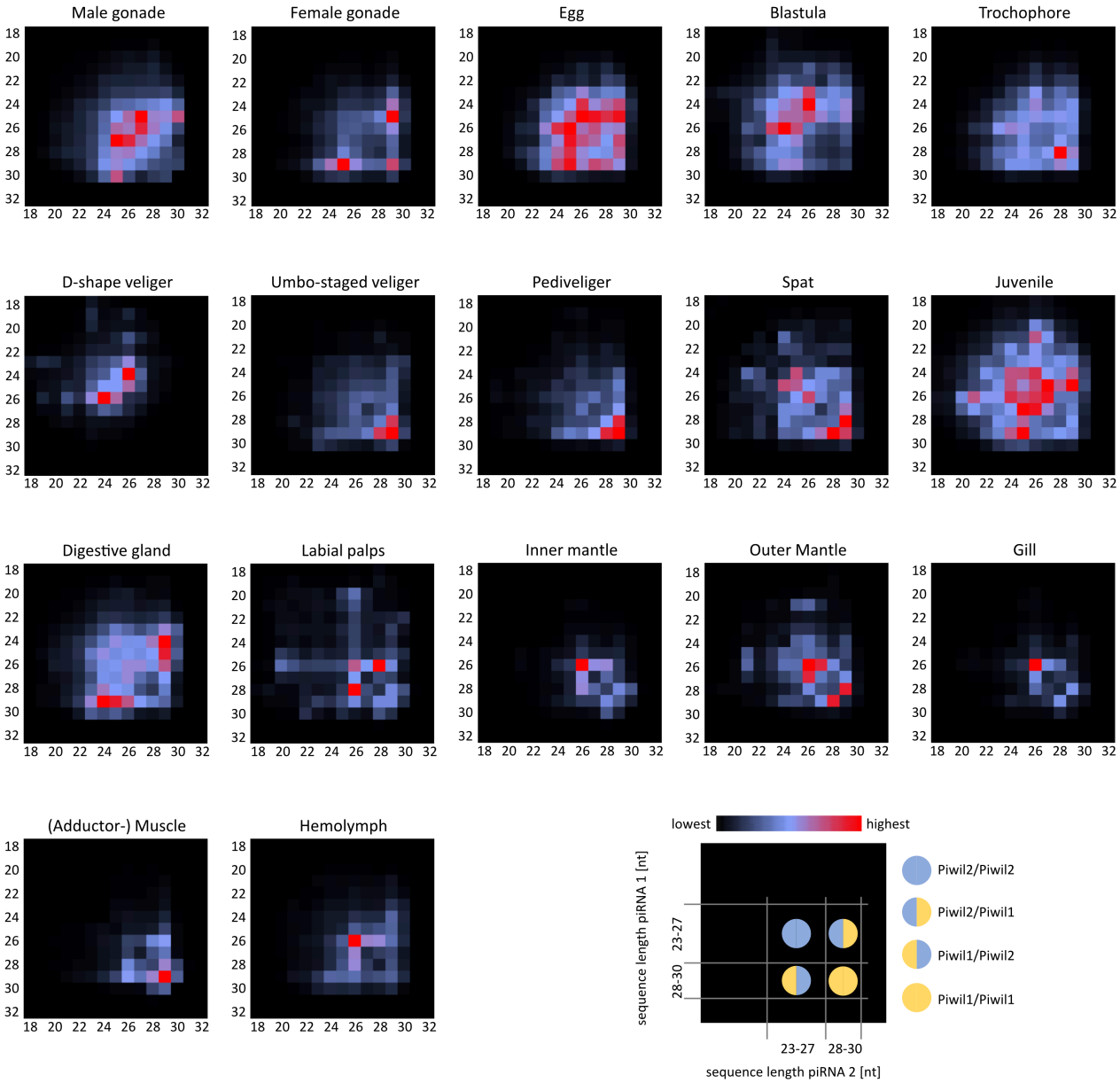
Supplementary figure 2 | Ping-pong signature of small RNAs from different C. gigas samples. Z-scores for specific 5' overlaps for each sample are shown.

Supplementary figure 3



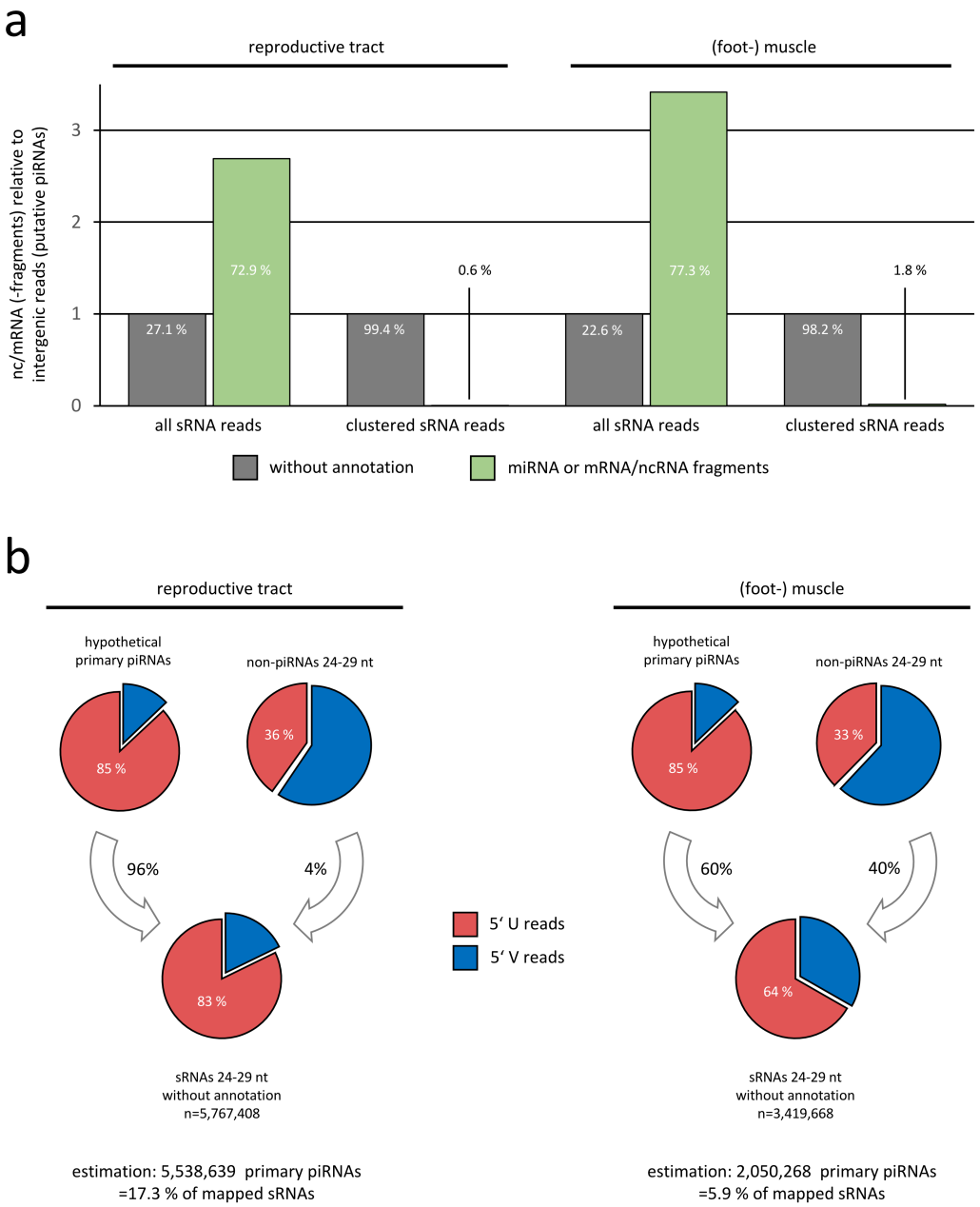
Supplementary figure 3 | Ping-pong read pairs per million bootstrapped reads from different *C. gigas* samples. Graphs depict the average number of sequence read pairs with a specific 5' overlap for 100 pseudo-replicates (PR) per dataset with one million reads per PR. The value for read pairs with 10 nt overlap (yellow) can serve as a measurement for the intensity of ping-pong amplification and is directly comparable across different datasets.

Supplementary figure 4



Supplementary figure 4 | Ping-pong matrices for small RNA from different *C. gigas* samples. Frequent length-combinations of ping-pong pairs (sequences with 10 bp 5' overlap) are indicated in red. x-axis and y-axis refer to sequence read length of the two sequences of a ping-pong pair [nt].

Supplementary figure 5



Supplementary figure 5 | Characterization of clustered reads and alternative estimation of primary piRNA amount. (a) Fraction of reads that are either miRNAs or mRNA/ncRNA fragments regarding all sRNA reads and sRNA reads that map to piRNA clusters. (b) Estimation of the amount of primary piRNAs within the fraction of sRNA reads without annotation based on observed 1U content in 24-29 nt reads without annotation and annotated 24-29 nt reads.

Supplementary table 1

type	reproductive tract	muscle	reproductive tract [%]	muscle [%]
total mapped	32054605	34775495	100.00	100.00
miRNA	14668478	17804121	45.76	51.20
rRNA	1040112	1913871	3.24	5.50
tRNA	6952690	5702772	21.69	16.40
5'tR-halves	55077	5342	0.17	0.02
5'tRFs	753080	1447500	2.35	4.16
3'tR-halves	5081	2247	0.02	0.01
3'tRFs	5457457	3428816	17.03	9.86
3'CCA-tRFs	8954	6464	0.03	0.02
misc-tRFs	673041	812404	2.10	2.34
protein_coding	271084	537846	0.85	1.55
snRNA	21435	42597	0.07	0.12
SRP_RNA	1082	315	0.00	0.00
snoRNA	342	166	0.00	0.00
ncRNA	132	5	0.00	0.00
RNase_MRP_RNA	13	9	0.00	0.00
RNase_P_RNA	17	3	0.00	0.00
scaRNA	2	0	0.00	0.00
vault_RNA	2	0	0.00	0.00
no annotation	9099216	8773790	28.39	25.23
putative parasitic contamination	580323	1178049	1.81	3.39
repeat sense	1488944	1397442	4.65	4.02
repeat antisense	1534980	1282555	4.79	3.69
non-repeat	5494969	4915745	17.14	14.14

Supplementary Table 1 | Annotation of small RNAs from *L. stagnalis* with unitas.

Supplementary table 2

scaffold	start	end	scaffold	start	end	scaffold	start	end	scaffold	start	end	scaffold	start	end
6	412333	420415	1100	70214	79017	2757	5060	9694	7142	41	5723	16165	13	4621
34	132008	136991	1142	121048	126593	2824	8007	56920	7229	15123	22966	16330	2749	10684
34	162144	171023	1174	50323	61018	2853	1	7970	7344	8395	17009	16543	3000	10538
39	250660	259018	1208	43015	49434	2938	15014	22977	7396	4008	13024	16651	1071	9995
45	210392	214604	1214	11008	76919	2948	14162	22156	7453	8058	15962	17349	2449	9941
55	249001	258022	1223	100034	107882	3055	45005	52273	7594	17121	22562	17457	1011	7608
73	35007	43014	1231	38181	46859	3101	5006	14001	7666	98	11961	17490	21	9710
91	250445	256182	1252	6001	17016	3130	270	6382	7694	11442	19604	17829	4538	5621
110	26228	34336	1303	56069	63355	3277	38699	39938	7935	162	6457	18332	816	6786
122	128	6829	1306	65064	73848	3305	11155	20016	8111	9730	17808	18681	2033	9140
133	220110	224903	1342	31163	39945	3328	42282	50013	8276	156	5021	18708	19	6008
138	101151	108874	1364	95134	106994	3464	3264	9991	8355	5067	13658	18983	59	5026
195	8068	112861	1410	25201	33770	3536	51039	56025	8378	13216	21915	19490	9	5363
208	221187	228722	1413	68476	74163	3551	15122	23879	8538	297	8804	19820	10	8560
212	80457	88931	1440	26023	35004	3599	18153	27007	8581	13030	22838	19856	22	3558
216	100013	179954	1456	91238	96989	3615	11021	14026	8628	11079	20013	19903	3	4779
228	68025	76909	1470	107223	109063	3634	38009	56999	8803	19473	23908	19963	737	7313
228	175147	183968	1474	89002	97620	3640	43220	49970	8859	12015	23701	20403	4	5994
232	6521	13938	1529	79016	89799	3712	11015	41018	8979	9091	21956	20988	4260	7309
236	84075	91018	1567	9003	15697	3720	17377	28010	9171	18101	22814	21073	3957	7635
304	146931	154968	1614	51099	60018	3944	332	15380	9280	16134	20428	21726	8	7160
310	91012	109015	1657	96000	100987	4011	16292	24775	9285	2	21352	22035	133	7448
311	3013	11008	1690	76005	81520	4046	4016	26982	9308	19017	22375	22217	2651	7430
318	1	25734	1697	10025	49019	4058	46074	51839	9409	9019	15980	22556	2833	7314
326	207379	213612	1698	37003	44967	4090	5980	11458	9471	3	4995	22750	4202	7159
334	16099	24906	1729	2025	11023	4116	3416	12020	9552	3004	12736	23447	318	6894
342	160004	167900	1729	23042	44993	4163	8094	10515	9625	2001	13798	23553	18	6778
371	94031	102698	1761	50307	58993	4315	2087	12024	9763	2569	6957	23732	1	6728
404	139001	172952	1874	52027	60615	4376	10004	19018	10040	13	7781	25718	47	2785
426	182036	188345	1874	81024	89992	4437	17358	23990	10169	9	16822	26008	38	5309
436	20002	29012	1880	13629	21942	4467	10047	16978	10414	6072	19546	26149	4	5622
441	108030	158944	1896	31053	36682	4615	26004	36016	10480	10067	18879	26446	111	5534
468	173136	178981	1901	58063	66572	4673	6002	14950	10570	6711	13698	26818	1	5431
513	4207	13478	1909	43421	45022	4675	32080	40813	10654	12025	18986	26930	9	4880
517	49099	53740	1911	69191	73916	4687	412	11645	10669	17447	18975	27535	90	4997
535	48465	143979	1929	48019	53022	4704	36088	44795	10723	74	3798	27794	99	5211
537	154478	161962	1931	75780	82699	4736	22404	29008	10726	126	7828	29610	11	4925
543	64232	70533	1951	46108	55007	4768	30015	38950	10894	14	4126	33017	7	4677
560	102370	108861	1953	58004	78986	4784	9014	18020	10960	8045	18225	33294	2	4675
563	83044	90874	1991	10002	39378	4800	83	8945	11045	7072	14208	33382	1	4570
567	138670	147015	2079	68164	75592	4846	23402	30420	11059	7148	9846	34360	151	4466
569	121358	129958	2089	8009	31015	4931	38001	43843	11602	1	4921	35274	1	4293
579	2	7008	2173	65274	73701	5086	7128	16017	11884	5450	15105	35519	27	4422
604	57100	63764	2245	31	5253	5179	25947	35018	12024	18	9640	36012	295	4275
617	156025	161376	2247	27015	31946	5408	27027	35900	12181	12169	15948	36769	6	3324
630	61504	65874	2280	37002	62003	5436	13	5009	12352	10934	12597	38402	5	3539
659	1076	9918	2290	10001	19024	5871	2596	10699	12465	405	5523	38694	3	3634
664	49109	54818	2290	29038	37303	5887	6016	14970	12754	4093	11987	39390	7	3508
724	101007	108009	2325	13388	15752	5888	2	5593	12945	6018	14662	40915	206	3251
809	39004	47004	2408	26004	34943	6091	31382	35514	12959	8179	14657	41699	100	3203
830	628	6999	2414	9596	16019	6104	9070	18006	12960	2	4908	42172	40	3161
849	4042	12804	2443	31220	39958	6341	6080	14840	12968	8019	14634	44934	3	2959
853	107023	115863	2443	62129	69501	6392	28001	33681	13221	5263	12870	50003	3	1937
925	68020	73098	2446	30030	37960	6509	6095	13986	13296	10	9023	50420	231	1861
952	115073	123768	2458	38	13988	6838	21099	29575	13541	4002	13019	54154	74	1563
964	59086	64954	2469	34007	52016	6856	10067	15752	13596	60	9000	55759	3	1446
964	88008	96757	2487	27129	36018	6868	13070	14170	13612	8368	13780	60834	2	1248
971	10109	15011	2504	19352	27016	6904	17030	25942	13650	9582	13629	67420	14	1084
974	101196	109967	2504	49220	64022	6969	20017	28006	14828	4217	12004	68864	1	1058
1010	16666	24957	2516	9	6007	7000	29	5008	14922	1	8973			
1034	19002	24983	2556	16001	31017	7007	15133	23567	15010	15	4608			
1036	82008	88728	2625	5	5788	7127	15078	23986	15501	2	7013			

Supplementary Table 2 | Location of predicted piRNA clusters in *L. stagnalis*.

Supplementary table 3

type	male gonad	female gonad	egg	blastula	trochophore	D-shape veliger	Umbo-staged veliger	pediveliger	spat	juvenile	digestive gland	labial palps	inner mantle	outer mantle	gill	muscle	hemolymph
miRNA	1750589	235904	131343	490238	2447283	5454840	4074084	3451640	308870	2883310	3256489	4610791	4115836	3840707	3947487	2831842	2251774
rRNA	278571	136879	38225	59889	1207499	140692	65630	206196	1905155	808739	926410	299517	351470	1131787	656076	383393	133959
tRNA	42281	12877	6561	12743	51324	42842	71093	102552	98330	81350	86718	38098	55057	138852	80557	65008	305407
no annotation	4650764	5359557	5654057	5458360	5160122	1287421	1673130	1573734	662346	1169143	1688554	549725	539093	601556	708366	867161	4839829
repeat (sense)	991438	731730	972825	984860	1182123	327134	328575	289195	61046	332841	323065	61468	68096	84411	117016	101064	636675
repeat (antisense)	1411635	1815815	2121549	2129309	1805968	496982	547068	371832	60740	320661	394822	55348	71818	74194	76224	124948	1398057
protein coding	836511	404553	560097	537516	849040	226431	290251	247450	148820	339142	276266	66147	80364	138219	134306	98305	337002
SRP RNA	3627	1083	142	219	2336	742	948	1399	2615	11208	18299	5297	8505	24683	16380	6030	742
snRNA	1115	649	1199	1446	9124	1447	2744	3327	10059	2171	2902	953	1242	2636	2175	556	4179
snoRNA	274	61	72	82	1266	168	512	625	1954	572	461	220	295	553	372	127	303
ncRNA	84	18	8	12	156	27	121	151	256	174	151	75	76	154	88	49	154
vault_RNA	51	22	27	24	29	6	150	186	135	61	52	33	18	29	272	3	37
RNase MRP RNA	13	8	6	1	37	5	36	17	61	33	20	12	17	23	15	5	14
nontranslating CDS	104	113	57	59	128	25	26	23	8	44	40	8	5	11	9	5	50
RNase P RNA	29	5	0	0	38	7	13	7	38	15	32	11	21	26	29	15	18
scaRNA	3	2	0	1	9	2	7	7	10	4	4	2	1	4	3	2	1
lncRNA	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
total mapped	9967090	8699276	9486167	9674760	12716483	7978771	7054387	6248342	3260443	5949467	6974286	5687704	5291914	6037843	5739376	4478512	9908199

Supplementary Table 3 | Annotation of small RNAs from different *C. gigas* samples with unitas. Reads without annotation or reads that match transposon (repeat) sequences represent putative piRNAs. Read counts of multiple mapping sequences were fractionated accordingly. Values are rounded, which explains possible discrepancies with the total number of mapped reads.

Supplementary table 4

5' overlap [nt]	male gonad	female gonad	egg	blastula	trochophore	D-shape veliger	Umbo-staged veliger	pediveliger	spat	juvenile	digestive gland	labial palps	inner mantle	outer mantle	gill	muscle	hemolymph
1	741	201	12711	58374	19108	15174	350038	2965	0	377	279	2	1	0	1	3	139
2	364	525	571	1172	452	61	979	39	0	98	418	1	0	0	0	0	366
3	538	263	843	1348	1992	561	43179	1321	1	80	273	6	1	5	1	14	73
4	927	454	1584	2955	3190	3989	15734	365	1	60	257	0	0	2	1	0	101
5	1302	577	2047	1751	1234	104	7181	156	0	212	279	2	1	0	0	1	112
6	398	668	726	759	668	219	2026	184	1	84	306	1	0	0	0	3	90
7	1962	3310	12088	6555	10252	497	961	74	0	92	753	0	1	1	1	0	1604
8	2161	670	2111	2474	2301	635	4496	87	1	222	244	0	1	0	0	2	442
9	2128	755	2278	2421	977	597	418	10	3	76	478	0	0	1	0	0	10169
10	65665	22281	44756	57158	26768	7908	4253	378	5	4536	50170	22	10	16	24	9	167909
11	1355	687	1409	1263	1738	153	72	9	0	25	268	0	0	0	1	2	279
12	1262	634	1730	7732	1299	1418	816	154	1	257	297	0	0	1	0	4	624
13	900	5312	2765	2857	5098	520	18017	633	1	85	4646	2	5	11	2	1	9229
14	5589	3337	5842	19404	10693	13444	2184	758	1	150	59	0	0	0	0	1	58
15	2964	1386	2208	5560	2747	3068	2869	529	1	215	58	0	0	0	1	3	92
16	2471	589	723	676	466	91	278	65	8	30	66	1	0	6	1	29	174
17	418	345	419	432	281	85	45	81	9	33	19	0	0	1	0	0	27
18	979	1000	760	577	1207	143	131	36	12	41	24	1	0	0	1	1	63
19	962	262	250	298	152	61	84	19	7	13	27	8	3	17	115	1	162
20	1398	310	664	591	378	160	184	32	10	73	23	0	0	0	0	0	35

Supplementary Table 4 | 5' overlap of mRNA-matching reads from different *C. gigas* samples.

Supplementary table 5

scaffold	start	end	scaffold	start	end	scaffold	start	end	scaffold	start	end	scaffold	start	end
C22996	8	2013	1267	488231	496058	1836	1017250	1025999	37788	42009	46966	481	190008	213675
C24160	15	2599	1288	19019	24548	1842	57328	63393	37880	5	10908	481	1176843	1180844
C24292	4	2660	13	156000	163503	1843	207193	215896	37880	30010	38490	486	43	5008
C24888	37	2943	1301	49002	58983	1851	32621	37085	37944	7041	48883	487	438034	446963
C26128	27	3672	1307	21023	28672	1855	43004	52018	38292	39184	47969	489	587053	593532
C26536	346	2597	1307	318877	345938	1856	77322	82243	38306	29433	41004	489	514129	518872
C26582	92	4021	1307	256015	265773	1863	14686	21574	38354	12034	16935	50	12273	15241
C27694	55	4886	1315	516072	524150	1867	103000	109021	38492	3	35026	501	923096	931841
C28006	19	5145	1316	156025	162821	1867	148855	152479	38492	47036	56026	501	761480	768489
C28026	9	5168	1328	13001	22026	188	6120	14460	38514	1019	9878	501	1340524	1349811
C28510	84	5491	1328	308139	320228	1883	122225	127023	38546	13027	20599	507	120634	124984
C29184	345	6309	133	95180	103744	1885	179022	187944	386	271566	274572	507	149063	184995
C29298	130	6237	1343	77139	108014	1891	95128	104018	38650	6029	13691	520	31003	64528
C29400	15	6024	1345	193032	201541	1897	382002	405027	38650	34779	42997	520	118003	129265
C29410	228	4811	1347	16049	21380	191	745045	751000	38672	54038	58995	520	238248	246755
C29682	1	4380	1350	32827	35519	192	22281	30874	387	39041	51015	520	143533	148027
C29766	81	5527	1359	266016	274943	20	217254	224801	387	68013	75876	521	141012	146229
C30332	4	3289	1382	212012	219011	204	879026	887763	38736	1	4982	524	247896	249342
C30848	14	4944	1388	119652	123732	205	6041	10745	38784	8504	17480	524	338171	347984
C30988	150	5006	1403	571161	577972	207	1016	20165	38900	42049	49763	53	396008	400959
C31380	1007	9369	1409	340002	341541	211	3003	8980	38978	36015	56756	534	62026	70000
C31488	31	5973	141	140253	148513	215	122821	129826	39196	48122	67059	535	285147	292847
C31644	375	5641	142	901435	905676	22	930188	936933	39220	57173	65961	535	555006	564028
C31806	12	5743	142	1174129	1181017	221	219	6690	39220	38080	47013	557	206808	212777
C31910	48	8750	142	1206001	1248018	221	48798	109810	39240	50045	59020	557	519003	529958
C32228	7	5956	1427	38666	45819	222	43182	47190	39268	48055	57027	563	467484	469525
C32276	7245	11192	1427	69705	89015	222	108197	114808	393	749256	755945	568	29675	37711
C32578	674	9013	1436	252	3418	222	146035	154983	393	1185061	1198004	575	89402	98807
C32614	1852	7874	1436	126002	141858	227	63914	69563	39380	54024	62933	578	317055	320887
C32778	1	11851	1440	55092	61830	232	101770	105863	394	83040	108015	579	186005	193897
C32838	7095	12702	1442	70094	77317	237	349014	355001	39470	16581	24897	581	176037	201988
C32888	3006	12019	1448	188339	193887	240	30463	41734	39508	51000	62502	583	245049	271022
C33378	5083	13933	1449	20879	22805	240	60140	83740	39526	29184	46044	591	100550	107009
C33534	5386	10248	1449	159029	172904	24066	27	2486	39580	1486	5903	591	400569	405010
C33708	12	8521	145	798161	811983	243	43228	51008	396	182792	187277	593	552011	569966
C33952	1082	5622	146	892063	899982	247	507186	515852	39620	69064	74479	599	69014	74625
C34120	3097	11735	1462	8332	16375	266	280539	287991	39642	14033	22770	6	20481	26976
C34402	5047	11943	1480	74195	85050	266	333117	340689	39724	26223	30992	608	49732	51923
C34436	10	13023	1480	127076	164310	267	102027	128755	39858	74071	80324	616	22155	27807
C34488	3011	12029	1491	43232	47797	267	139025	148019	39858	3129	13822	626	578198	585917
C35006	66	5996	150	500690	502460	267	211214	218770	40010	26	3671	632	94032	101559
C35760	1032	9910	1507	5737	15564	3	780052	788981	40010	31026	39774	664	489124	507001
C35844	21325	28131	1512	99630	120916	301	871006	878518	40010	69406	70867	664	377404	385602
C35952	19327	27999	152	255030	281991	30818	2861	8351	40154	16212	34958	664	533272	551024
C36038	14004	22334	152	355010	363364	30826	2695	7553	40246	14073	31016	664	648839	655457
101	427099	428845	1525	358040	369017	30854	140	4839	40346	25006	29698	680	23938	28768
1016	307225	317000	1525	269100	292988	309	132009	137022	40440	9135	16279	683	50186	56717
1016	430838	436768	1525	206248	212261	313	627909	634008	40476	25336	31982	689	354031	358658
1018	429463	439731	1525	304015	338025	31382	196	8741	40634	10292	14086	70	583121	586591
1018	460024	478006	1532	366178	367822	317	199152	211027	40744	91004	98582	705	325698	330573
1018	505453	514022	1533	224036	232931	31756	2023	7698	408	293120	298948	709	95235	105602
102	480103	483357	1535	93262	102021	31920	2017	7570	41086	16176	23011	709	118390	126911
1020	36273	44022	1570	29007	39881	321	40168	46418	41470	66000	73346	716	343013	351912
1024	99268	107798	1579	82018	88992	321	634002	640192	415	73058	78297	72	410343	414965
1024	1053694	1093463	1597	97847	103207	33050	1010	10849	41560	77097	78960	721	26224	34827
1035	122233	131979	160	445005	451131	331	24366	28134	41588	83826	90868	723	217643	222434
1053	103797	107539	160	486014	496762	33414	98	14211	41594	43004	51694	733	29004	37882
1053	361062	365277	1603	53205	58961	33958	7030	15955	41598	56346	107899	748	134250	142481
1064	110842	116025	1605	768230	776286	34	266020	270972	416	134351	139768	751	96009	117826
1066	129024	133941	1610	151746	156491	34144	4008	13022	416	58087	66018	753	95783	108483
1070	71140	80336	1627	42	4769	34278	4119	12419	41648	2033	21024	759	95091	104012
1073	1	1443	1630	202074	208599	343	185466	189510	417	62059	67472	773	169067	178456
1077	343085	349360	1636	4526	6944	343	612137	629640	41786	52022	63169	773	235017	267007
1077	319540	329955	1643	100904	105945	34350	42	4788	41812	34283	36942	775	126002	135021
1080	166093	174889	1647	193438	201717	348	470075	485025	41834	63210	75621	789	337035	348959
1086	508714	514327	1651	188072	193747	35014	12001	18902	419	1252257	1260900	789	368099	383990
1097	43218	72863	1654	52478	57751	351	278737	317657	419	1369042	1377989	790	216042	225005
1100	92053	100912	1665	13759	21826	351	184054	220989	42062	84853	86936	810	52007	56889
1123	332345	338861	1671	110593	115678	35136	4034	14502	42172	56000	63557	825	33043	41967
113	568387	576670	1679	38380	43353	35194	144	8012	42604	150065	153663	834	21558	29980
1131	246002	254043	168	994392	1000630	35230	18019	23325	42648	160091	172020	852	47320	54928
1131	213001	221998	1700	49044	66018	35358	423	23012	42656	148031	157016	854	162403	167146
1132	494581	495659	1703	805573	813007	35366	2	1523	42880	63035	71981	86	1438141	1446022
1132	366001	373810	1703	169004	248065	354	750201	754393	42948	151230	159758	867	216511	222148
1138	170130	178988	1703	130221	135012	35438	18485	24412	43	119828	125959	874	16173	22991
1140	6320	15781	1704	22105	31997	35466	9028	14890	43	79503	96961	88	194071	199005
1143	25056	33919	1710	84010	89706	35640	134	4452	43076	66000	75000	90	115055	149925
1144	365183	374118	1710	541074	563334	35988	15	25905	43182	129001	144010	90	161019	169985
1149	678744	680089	1711	229357	242989	35994	3	25972	43210	25044	31906	90	185453	190467
1154	391042	395919	1711	255035	263961	36	1013	8973	43242	12013	21002	907	19852	37024
1154	854095	858324	1722	752	9924	36196	114	3250	43296	209502	216889	908	36015	44461
1154	302044	308339	1725	46352	52874	36390	6007	12000	433	975034	983722	91	15093	22012
1159	81185	87853	1738	5003	10004	36390	24004	32899	433	1254247	1270015	911	135201	144026
1164	71443	75972	1741	144387	148880	36398	25013	31986	43302	56181	64832	919	35002	69099
1164	223339	224826	1745	13629	16209	364	281102	288919	43446	129696	136749	933	470113	478165
117	55716	66790	1749	109124	117020	364	315019	323814	43634	1	5011	934	155143	162008
1173	111027	118199	1759	47304	53941	365	292959	297911	43634	137002	151814	948	312540	315892
1174	1067000	1080024	1762	1206	6020	36806	21727	23824	43662	22004	27023	951	26020	35820
1186	177205	185915	1773	213192	221707	36898	12021	22438	43686	154021	162937	959	215	4803
1186	196032	203700	1773	264014	272607	36916	5075	10550	43818	250974	257023	964	144017	152766
1187	44907	52011	1788	75084	90007	37128	33722	40680	43842	305083	311583	967	98099	105291
1188	77040	82948	1794	513055	520735	37252	16029	24922	43844	97513	102715	967	127001	152960
1188	156204	177011	1794	473170	481997	374	199529	204874	44054	235559	238028	968	142030	148780
1199	20343	44790	1800	438074	445029	374	317259	325605	442	79697	82663	968	70048	78003
1209	68473	146246	1806	133007	141196	37400	4750	12774	444	51	2551	971	37605	44947
1209	189707	192926	1815	196615	202886	37532	7001	21016	447	259280	267162	986	37127	45885
1237	115012	123961	1816	117026	125983	377	404749	411757	447	364042	381596	99	308266	312492
1257	433	3502	1832	185176	207012	37772	4047	47835	448	339127	351989	998	3207	9776
1259	32275	40443	1833	92009	96947	37788	72	3752	461	513694	521677			

Supplementary Table 5 | Location of predicted piRNA clusters in *C. gigas*.

Supplementary Methods

An Arithmetical estimation of the amount of primary piRNAs

In order to support our interpretation that the number of primary piRNAs in our datasets corresponds to the number of sequence reads mapped to predicted piRNA clusters, we estimated the amount of primary piRNAs in *Lymnaea stagnalis* reproductive tract and muscle tissue with an arithmetical approach which is independent of piRNA cluster annotation. The approach is based on the assumption, that the fraction of sequence reads that do not match any other class of non-coding RNA such as rRNAs or tRNAs, contains genuine piRNAs, in addition to other RNAs that passed the annotation procedure due to either poor genome annotation, intra-species variability, post-transcriptional RNA modification or sequencing errors. To reasonably asses the amount of primary piRNAs in this mixed RNA population, we only considered molecules with a sequence length ranging from 24 nt to 29 nt as putative piRNAs, based on sequence read length profiles (Fig. 2A) and calculated ping-pong matrices (Fig. 4A and S3). Under the presumption that 85% of primary piRNAs start with a U [1,2], we argue that any deviation from 85% 5'-U reads is caused by RNA molecules other than piRNAs which can also have a U at position 1 (1U). We estimated the fraction of 1U reads in the non-piRNA population on the basis of 1U content of annotated reads in the 24-29 nt size range, ignoring tRNA fragments which were shown to preferentially bind PIWI proteins [3,4]. Using the equations below,

$$a + b = 1$$
$$a * 1U_a + b * 1U_b = 1U_{observed}$$

where a is the fraction of piRNAs and b is the fraction of non-piRNAs, $1U_{observed}$ describes the fraction of 1U reads in 24-29 nt RNAs without annotation (0.831 and 0.643 in reproductive tract an muscle, respectively), $1U_a$ describes the fraction of 1U reads in piRNAs (presumed to be 0.85) and $1U_b$ describes the fraction of 1U reads in non-piRNAs (0.359 and 0.333 for reproductive tract and muscle, respectively), we estimate the fraction of primary piRNAs to be 96,0% and 60,0% of all 24-29 nt reads without annotation in reproductive tract and muscle, respectively. These fractions account for 17.3% and 5.9% of total mapped reads in the corresponding datasets and are almost identical to the fraction of reads that map to predicted piRNA clusters, supporting the validity of piRNA cluster calling results (Supplementary Figure 5b).

Supplementary References

1. Aravin, A. A. et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 442, 203-207 (2006).
2. Brennecke, J. et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 128, 1089-1103 (2007).
3. Keam, S. P. et al. The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. Nucleic Acids Res. 42, 8984-8995 (2014).
4. Honda, S. et al. The biogenesis pathway of tRNA-derived piRNAs in Bombyx germ cells. Nucleic Acids Res. 45, 9108-9120 (2017).

4. Regulation of protein-coding genes by piRNAs in the pig

Daniel Gebert[1], René Ketting[2], Hans Zischler[1], David Rosenkranz[1]

[1] Institute of Anthropology, Johannes Gutenberg-University, Mainz, Germany
[2] Institute of Molecular Biology IMB, Mainz, Germany

This chapter was published as a Research Article in *PLOS One* under the title "piRNAs from Pig Testis Provide Evidence for a Conserved Role of the Piwi Pathway in Post-Transcriptional Gene Regulation in Mammals" (Gebert et al., PLOS One 2015 10:e0124860). The experimental part, as well as preliminary analyses were conducted during the Master thesis of DG.

4.1. Abstract

Piwi-interacting (pi-) RNAs guide germline-expressed Piwi proteins in order to suppress the activity of transposable elements (TEs). But notably, the majority of pachytene piRNAs in mammalian testes is not related to TEs. This raises the question of whether the Piwi/piRNA pathway exerts functions beyond TE silencing. Although gene-derived piRNAs were described many times, a possible gene-regulatory function was doubted due to the absence of antisense piRNAs. Here we sequenced and analyzed piRNAs expressed in the adult testis of the pig, as this taxon possesses the full set of mammalian Piwi paralogs while their spermatozoa are marked by an extreme fitness due to selective breeding. We provide an exhaustive characterization of porcine piRNAs and genomic piRNA clusters. Moreover, we reveal that both sense and antisense piRNAs derive from protein-coding genes, while exhibiting features that clearly show that they originate from the Piwi/piRNA-mediated post-transcriptional silencing pathway, commonly referred to as ping-pong cycle. We further show that the majority of identified piRNA clusters in the porcine genome spans exonic sequences of protein-coding genes or pseudogenes, which reveals a mechanism by which primary antisense piRNAs directed against mRNA can be generated. Our data provide evidence that spliced mRNAs, derived from such loci, are not only targeted by piRNAs but are also subject to ping-pong cycle processing. Finally, we demonstrate that homologous genes are targeted and processed by piRNAs in pig, mouse and human. Altogether, this strongly suggests a conserved role for the mammalian Piwi/piRNA pathway in post-transcriptional regulation of protein-coding genes, which did not receive much attention so far.

4.2. Introduction

Small non-coding RNAs (sncRNAs or sRNAs) are involved in many cellular processes such as gene regulation, transposon repression and antiviral defense, which they realize by the principle of RNA interference [1]. To fulfill their functions all types of sRNA are dependent on Argonaute proteins, for which they act as guides that recognize targets based on sequence complementarity. Piwi-interacting RNAs (piRNAs, ~24–32 nt in length) represent a class of sRNAs that associate with Piwi clade Argonaute proteins, of which different species possess a varying number of paralogs [2–6].

The majority of piRNAs is organized in large genomic clusters, distributed throughout the genome at defined loci, ranging from 1–100 kb in size [3–6]. Further, piRNAs are characterized by a strong bias for uracil at the 5‘ end position (1U) and a preference for adenine at position ten (10A) for secondary piRNAs (see below). Finally, they are typically longer than miRNAs and siRNAs while displaying a broader size-distribution which is likely caused by the piRNA-specific 3’ end processing by exonucleases. These traits are the result of the biogenesis mechanisms of piRNAs which include two pathways [7–9]. In primary biogenesis, piRNAs are generated through the processing of long precursor transcripts into piRNA intermediates, which are loaded onto Piwi proteins that heavily select for 1U

fragments [10], followed by 3' trimming and 2'-O-methylation of the 3' end by the methyltransferase Hen1 [11,12]. In secondary biogenesis, also known as ping-pong amplification loop, Piwi proteins loaded with primary piRNAs, target complementary transcripts, which are cleaved with a 10 nt offset from the 5' end of the guiding primary piRNA to generate secondary piRNAs. Owing to this offset and the 1U bias of primary piRNAs, the resulting secondary piRNAs preferentially contain an adenine at the tenth position [7–9]. This 10 nt 5' overlap of primary and secondary piRNAs is commonly referred to as ping-pong signature.

One of the major functions of the Piwi/piRNA pathway is the repression of transposable elements (TEs or transposons). Piwi proteins are primarily expressed in germ cells, regarding mammals especially during spermatogenesis [3,9]. In the course of spermatogenesis, genome wide demethylation, as part of the epigenetic reprogramming, leads to a reactivation of TEs [9,13]. In both mouse and fruit fly, mutations of Piwi proteins result in derepression of TEs in the germline leading to male sterility [14–17]. Similarly, deficiency of murine piRNA clusters results in an increased activity of TEs, emphasizing the importance of piRNAs in transposon silencing [18]. Accordingly, piRNA clusters are commonly perceived as transposon traps that acquire the capability of producing piRNAs directed against particular TEs as soon as the TE by chance jumps into such a locus [19].

Despite their important role in repressing transposon activity, in mouse only meiotically (pre-pachytene) expressed piRNAs are enriched for TE-related sequences, in contrast to pachytene piRNAs, of which only about 17% are TE-derived [3,4,14]. This led to the presumption that piRNAs might fulfill other functions besides TE silencing. Indeed, several studies in fruit fly suggested a role for piRNAs in regulation of protein-coding genes, including Stellate, vasa [20,21], Fasciclin 3 [22], and nanos [23]. Hints for a gene-regulatory function of the Piwi pathway in mammals have also been obtained in mouse [24,25], but neither the underlying mechanism, nor the discrete function has become clear so far.

The majority of mammalian species, including humans, possess a standard set of four paralogous Piwi proteins [26], while the bulk of research on mammalian Piwi/piRNA biology was conducted in mouse or rat, which express only three Piwi paralogs. In that sense, mice and rats might represent an exceptional realization of Piwi/piRNA biology. Hence, to investigate the nature of piRNAs in the mammalian germline in a context that resembles the regular condition with respect to Piwi protein equipment, we sequenced and analyzed testis expressed sRNAs of the pig, a species expressing all four mammalian Piwi paralogs. The pig is particularly interesting in the context of Piwi/piRNA biology, considering the unique TE landscape of the porcine genome, comprising e.g. active tRNA-derived short interspersed elements (SINEs) and pig-specific endogenous retroviruses (ERVs), while at the same time having a considerably lower share of TE sequences compared to other mammals [27]. Besides, porcine spermatozoa are known to exhibit extreme fitness due to domestic breeding and sexual selection in promiscuous mating systems resulting in sperm competition. Adding to the previous initial characterizations of porcine piRNAs [28,29], we focused on both, possible new aspects of the TE silencing function, as well as potential roles in the regulation of non-TE targets. Our present study strongly indicates that the mammalian Piwi/piRNA system is involved in post-transcriptional gene regulation and that piRNA clusters, which occupy a central role in this process, might be more dynamic and adaptable than previously thought.

4.3. Methods

4.3.1. Ethics statement

This study did not require approval by an ethics committee. Biological samples were obtained under current law from a licensed provider (Georg-August-University Göttingen, Animal Breeding and Genetics, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany).

4.3.2. Preparation of sRNA libraries

Testis tissue was taken from an adult boar (*Sus scrofa domestica*) and stored at -80°C. Total RNA was extracted directly from testis tissue using TRI Reagent (Ambion) according to the manufacturer's instructions. The employment of 50 mg of tissue resulted in an RNA yield of approximately 140 µg. Total RNA was applied to a urea-based denaturing polyacrylamide gel (10%) together with the GeneRuler Ultra Low Range DNA Ladder and run for 20 minutes (1200 V, 50 mA, 60W). The 20–35 nt fraction was excised from the gel and resolved in 30 µl water using Amicon's Ultrafree-MC and Ultra-0.5 3K centrifugal devices according to the manufacturer's instructions.

We portioned the obtained sRNA sample into two fractions and conducted sodium periodate treatment followed by ß-elimination with one of the two fractions according to the method applied by Rajasethupathy and colleagues with minor adjustments regarding the sample volumes [30]. A 5'-diphosphorylated and 3'-blocked RNA adapter (5'-rAppCTGTAGGCACCATCAATddC-3', Integrated DNA Technologies) was directionally ligated to the 3' end of periodate treated and untreated sRNA samples in absence of ATP using New England Biolabs T4 RNA Ligase 1 according to the following reaction mixture: 43 µl sRNA sample, 6 µl of 100% DMSO, 6 µl 10x NEB ligation buffer, 2 µl 3' RNA adapter, 2 µl T4 RNA ligase (10 U/µl) and 1 µl of RiboLock RNase Inhibitor (Thermo Scientific). The mixture was incubated at room temperature for 2 hours. For separation of sRNA molecules linked to a 3' adapter we conducted acid phenol chloroform (Life Technologies) extraction and ethanol precipitation followed by separation of molecules ranging from 40 to 55 nt in length using polyacrylamide gel electrophoresis with subsequent gel extraction as described above.

A second RNA adapter carrying a 4 nt sequence tag and lacking a 5'-phosphate was ligated to the periodate treated (5'-GACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA-3') and untreated (5'-GACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGAUCGAA-3') sRNA samples in presence of ATP using New England Biolabs T4 RNA Ligase 1 according to the following reaction mixture: 36 µl sRNA sample, 3 µl RNA adapter, 6 µl 100% DMSO, 6 µl NEB 10x ligation buffer, 6 µl 10mM ATP, 2 µl T4 RNA ligase (10 U/µl), 1 µl RiboLock. The mixture was incubated at 37°C for 30 minutes.

The ligation reaction was stopped and RNA was purified by acid phenol chloroform extraction and ethanol precipitation and dissolved in water. Following cDNA synthesis using Superscript III Reverse Transcriptase (Life Technologies), the sample was PCR amplified (forward primer for periodate treated sample: 5'-ACATGGACTGAAGGAGTAGA-3', forward primer for untreated sample: 5'-ACATGG-ACTGAAGGAGATCG-3', reverse primer for both samples: 5'-ATTGATGGTGCCTACAG-3') and ethanol precipitated. Both tagged samples were high throughput sequenced in parallel on an Illumina HiSeq 2000 system.

4.3.3. Bioinformatic data processing and analysis

First, 5' adapter and 3' adapter sequences were clipped from NGS raw sequences and reads were allocated to periodate treated sRNA and untreated sRNA datasets based on the differentially tagged 5' adapter. Considering a putative contamination by non-piRNA sequences, reads ranging from 18 to 34 nt in length were mapped in sense orientation to available ncRNA sequences from Ensembl database (release 77), miRBase [31] and the Genomic tRNA Database [32] using SeqMap [33] (version 1.0.12) to sort out sequences resembling microRNAs (miRNA) or fragments of other ncRNA types such as miRNA precursors, snRNA, snoRNA, rRNA and tRNA. Sequences that did not produce a match to any known ncRNA, thus representing putative piRNAs, were mapped to the genome of Sus scrofa (Sscrofa10.2.75) using SeqMap, taking only perfect matches into account.

To determine the amount of sRNA sequences related to TEs, the porcine genome was masked using RepeatMasker software and porcine transposon sequence data from Repbase [34]. The quantity of reads mapping to TEs was normalized for each sequence by the total number of genomic hits it produced. The identification and analysis of piRNA clusters was performed using the tracking and analysis software proTRAC [35] (version 2.0.2), searching for clusters with a minimum size of 10 kb, applying a sliding window size of 1 kb and an increment of 0.1 kb.

In order to identify cDNA sequences that exhibit a ping-pong signature, thus representing putative piRNA targets, we mapped sRNA sequences to annotated cDNA (Ensembl release 77). We applied a coverage threshold of 10 mapped sequence reads (counts were normalized by the number of hits per sequence) per one million mapped sequence reads to ensure comparability across the different probes that comprised different total numbers of sequence reads. The principals of this computational approach are described in Antoniewski 2014 [36].

To search for conserved cDNA targets of piRNAs in mammals, we applied identical procedures to human and mouse testis expressed sRNA datasets that are deposited at NCBI's sequence read archive (SRA) under the accessions SRX271415, SRX271416 and SRX271417 for human sRNAs [37] and SRX154530 for mouse sRNAs [38]. We considered a ping-pong signature to be evident if the peak referring to the 10 nt overlap was at least 2-fold higher compared to the next highest peak. Generally, z-scores for ping-pong signatures were calculated according to the method applied by Zhang and coworkers [39]. The identified genes were subjected to GO term enrichment analysis [40], applying a p-value threshold of $p = 0.05$. To verify that the numbers of homologous genes targeted in different species is higher than expected by chance, we randomly sampled genes from two species according to the number of observed piRNA target genes (one million draws). We calculated expected values (E(X)) for the number of homologs that are present in both random sets based on the observed cross match. P-values correspond to the frequency of observed cases with a cross match equal or higher than observed for the original data set. The applied Perl scripts are available upon request.

4.3.4. Data deposition

The complete sequence dataset is available at NCBI's SRA under the following accessions: BioProject ID: PRJNA267635, Experiment: SRX761355, Run: SRR1654828.

4.4. Results

4.4.1. Annotation of porcine sRNAs

Overall 13,596,939 raw sequence reads were obtained from sequencing of periodate treated porcine testis RNA and subjected to several filtering and initial processing steps. Of 12,508,703 reads within the size range of 18–34 nt, 1,502,807 reads (12.0%) could be classified as miRNAs (0.09%) or fragments of other ncRNA species such as tRNA (11.6%), rRNA (0.09%), snoRNA (0.06%) and snRNA (0.01%), leaving 11,005,896 reads, comprising 3,226,011 non-identical sequences that represent putative piRNAs. A fraction of 7,219,711 reads, originating from 928,481 non-identical sequences, mapped perfectly to the genome of Sus scrofa, producing 24,579,193 genomic hits.

The mapped sRNAs show a roughly Gaussian length distribution, ranging mainly from 24– 33 nt with a peak at 30 nt (Fig 1A). More than 99% of all reads fall into the typical size range of mammalian piRNAs (24–32 nt) and the vast majority (91%) of sequence reads maps to one of 142 predicted piRNA clusters (see below). The mapped sRNA sequences exhibit a strong ping-pong signature, meaning a strong preference (z-score = 44.4) for 10 nt 5' overlaps between sequences mapping to the sense and antisense strands of the genome, which is a hallmark of piRNAs, attributable to their specific biogenesis mechanism during the ping-pong cycle (Fig 1B, S1 File).

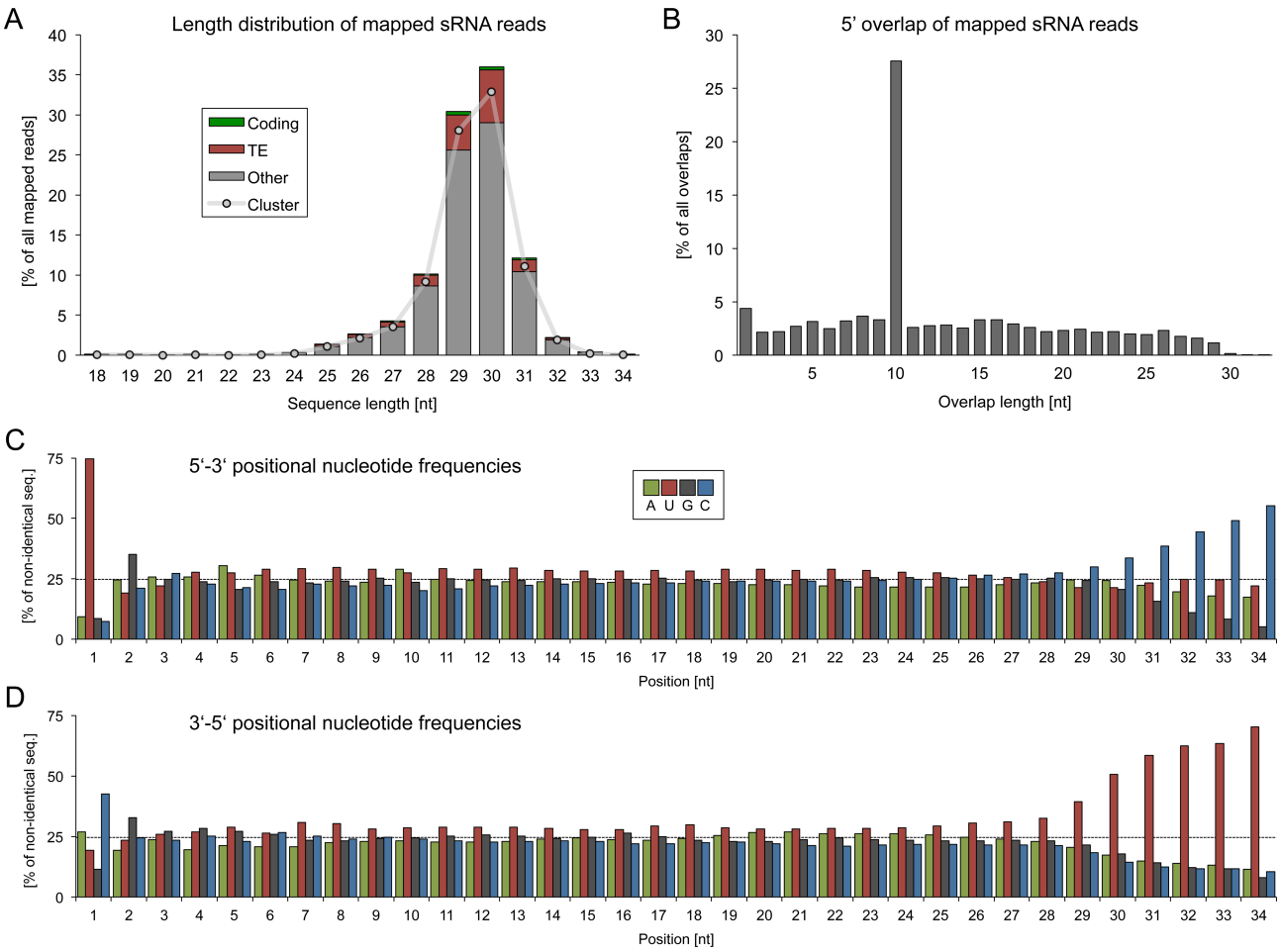


Figure 1 | Basic characterization of putative piRNAs in porcine testes. (A) Length distribution of small RNAs. The mapped sRNA reads show an approximately Gaussian length distribution, ranging mostly from 24 to 33 nt with a peak at 30 nt. The majority of each size fraction maps to predicted piRNA clusters. (B) 5' overlap of sRNAs. Sense and antisense sRNA reads produce a high rate of 10 nt 5' overlaps. (C) Positional nucleotide frequencies starting from 5' end. (D) Positional nucleotide frequencies starting from 3' end.

Another characteristic trait of piRNAs is constituted by a strong bias for uracil at the 5' end and a consequential preference for adenine at the tenth position for secondary piRNAs. Nearly 75% of non-identical sequences start with a uracil, while adenine is only slightly enriched at position ten (29.2%), which suggests that the bulk of porcine pachytene piRNAs originates from primary processing (Fig 1C). Furthermore, we observed a bias for cytosine at the 3' terminus (42.6%) and for guanine at the second position of both the 5' (34.5%) and 3' ends (32.9%) (Fig 1C and 1D).
Together, though we do not provide formal evidence for binding of these sRNAs to Piwi proteins, the overall characteristics of the analyzed sRNA dataset (size distribution, nucleotide composition, genomic clustering, ping-pong-signature) are in compliance with the typical piRNA traits and indicate a very low degree of contamination by non-piRNA sequences.

4.4.2. TE-derived piRNAs

Transposon silencing is considered as the main function of piRNAs, hence the mapped piRNAs were screened for sequences that target genomic loci annotated as TEs. Overall 14.0% of total mapped reads (representing 16.3% of non-identical sequences) match transposon sequences (Fig 2A), of which SINEs contribute the largest proportion (5.9%), followed by LTR retrotransposons (4.0%), LINEs (3.5%) and DNA transposons (0.6%) (Fig 2B). Quantity and composition of TE-related piRNAs contrast the overall genomic situation with a total of 32.6% corresponding to TEs (Fig 2A), where the

largest TE fraction is represented by LINEs (15.6%), followed by SINEs (12.2%), LTR transposons (3.5%) and DNA transposons (1.4%) (Fig 2B).

Though piRNAs generally map to TEs in both orientations and the overall amount of sense and antisense piRNAs is roughly equal, the sense/antisense ratio differs considerably for different transposon families (Fig 2B). While tRNA-derived SINEs like the abundant PRE elements show a strong bias for sense piRNAs, ERV1 elements exhibit a strong bias for antisense piRNAs. Since the majority of TE-related piRNAs originate from piRNA clusters, we assumed that these differences might result from insertional strand bias. Therefore we checked the insertion direction of TEs relative to the transcribed piRNA cluster strand. Indeed, we found that the insertion direction correlates well with, and thus can explain the different sense/antisense piRNA ratios for the most prominent TE classes, namely tRNA-derived SINEs, L1 and ERV1 (S1 Fig), which comprise more than three quarters of all TE-derived piRNA reads.

In order to search for evidence of ongoing TE repression via the ping-pong cycle, we analyzed the 5' overlaps of sense and antisense piRNAs mapped to TE sequences (Fig 2C and 2D). Though we observed a marked ping-pong signature (z-score = 17.3) for TE-related piRNAs, indicating Piwi-

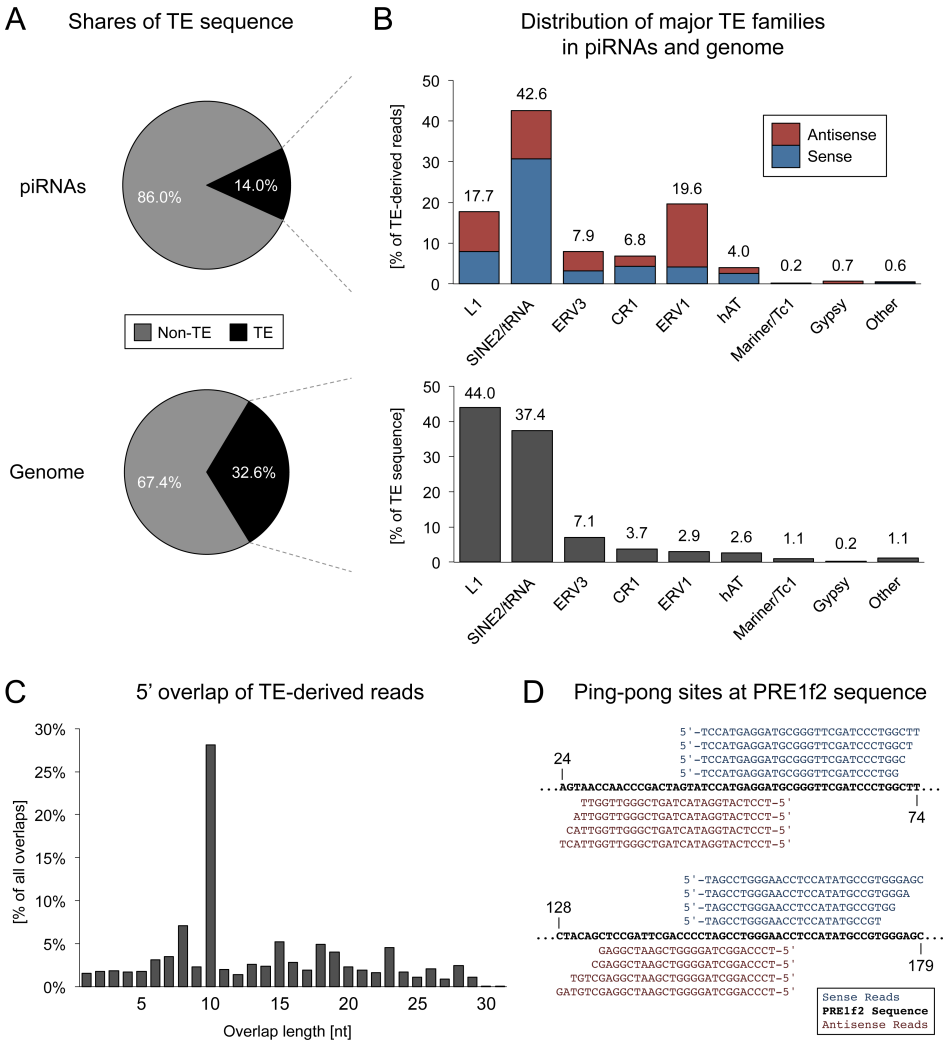


Figure 2 | Transposon-derived piRNAs. (A) Shares of TE sequences in mapped piRNA reads and in the porcine genome. (B) Representation of TE families within the sequences of piRNA reads in sense and antisense direction compared to the genomic TE family distribution. (C) 5' overlap of TE-derived piRNAs. (D) Mapping of piRNA reads to the sequence of a PRE1f2 element, a member of the tRNA-derived SINE subfamily of porcine repetitive elements. Regions from positions 24–79 nt and 128–179 nt are shown as exemplary target sites of ping-pong processing.

dependent processing, both sense and antisense piRNAs show a strong 1U bias (84% and 87%, respectively), and only a slight elevation for 10A (38%) can be observed for antisense TE reads. This is in line with previous findings from mouse [9], where 1U-biased primary piRNAs generated from TE transcripts target piRNA cluster transcripts resulting in secondary 10A-biased antisense piRNA. Together, our data suggest that a noticeable fraction of antisense piRNAs originates from secondary processing while still most pachytene piRNAs are generated via the primary processing mechanism.

4.4.3. Gene-derived piRNAs

Gene-derived piRNAs were previously observed in diverse species but were generally considered to represent a byproduct derived from mRNAs that accidentally fall into the clutches of the Piwi/piRNA pathway, mainly because only sense piRNAs could be found. To investigate a potential impact of piRNA function on protein-coding genes, mapped piRNA reads were initially screened for sequences mapping to annotated coding DNA (cDNA). In total 1.8% of mapped reads, representing 9.4% of non-identical sequences, produce perfect matches to porcine cDNA. Intriguingly, when focusing on protein-coding genes we found that 7.6% of piRNA reads map to intronic sequences in sense (3.0%) and antisense (4.6%) orientation, which apparently cannot be explained by processing of spliced mRNA. Further, 1.6% map to exonic regions in sense (1.24%) and antisense (0.32%) orientation and 0.02% of piRNA reads match pseudogenes mainly in sense direction (Fig 3A).

To determine whether sRNAs that mapped to exonic sequences of protein-coding genes represent degraded mRNA or resemble genuine piRNAs, the according sRNA reads were examined for piRNA characteristics. Both sense and antisense reads, which all range between 24 and 32 nt, show a strong bias for 1U (82.6% and 70.9%, respectively), while only sense sequences exhibit a marginal preference for 10A (28.2%) as compared to antisense reads (21.3%). Furthermore, piRNA reads that mapped to 115 genes exhibit a marked ping-pong signature (z-score = 22.7, Fig 3B). In addition, the length distribution of both sense and antisense cDNA-matching piRNAs reveals the presence of at least two different piRNA populations and thus the participation of different Piwi paralogs in the generation of gene-derived piRNAs (Fig 3C). Generally, piRNAs map to specific gene transcripts in a very similar fashion as compared to TE transcripts with clear signs of ping-pong-mediated amplification, which implies that mRNA is not only subject for primary processing, but can also be targeted by primary piRNAs and processed into secondary piRNAs (Fig 3D).

In order to check whether this pattern can be found in additional species, we performed the same analysis on available mouse and human sRNA and cDNA datasets. Remarkably, we observed a large amount of cDNA-matching sequences producing a clear ping-pong signature that is mainly concentrated on 185 (ping-pong-z-score = 41.2) and 424 (ping-pong-z-score = 13.4) different genes in mouse and human, respectively (Fig 3B; S2 File). Moreover, targeting of a number of gene transcripts appears to be conserved over evolutionary timescales. For instance, we noticed high piRNA coverage and ping-pong signatures on several members of the NUT (Nuclear protein in testis) gene family (NUTM2A, NUTM2B, NUTM2D, NUTM2E) for porcine as well as human piRNAs. Furthermore, ping-pong signatures were also detectable on transcripts of Histone H2A genes for all three datasets, though the read coverage is considerably lower as compared to NUT gene transcripts (S2 File). Altogether, pig and human share 15 homologous target genes ($p = 0.0050$, $E(X) = 1.2188$), while 7 homologs are targeted in both pig and mouse ($p = 0.0241$, $E(X) = 1.0236$), which are significant numbers compared to a random overlap between non-related, randomly selected genes. Hence, targeting of homologous gene transcripts across distantly related species suggests that the Piwi/piRNA system snatches mRNAs not in a random fashion but rather implies a specific biological function.

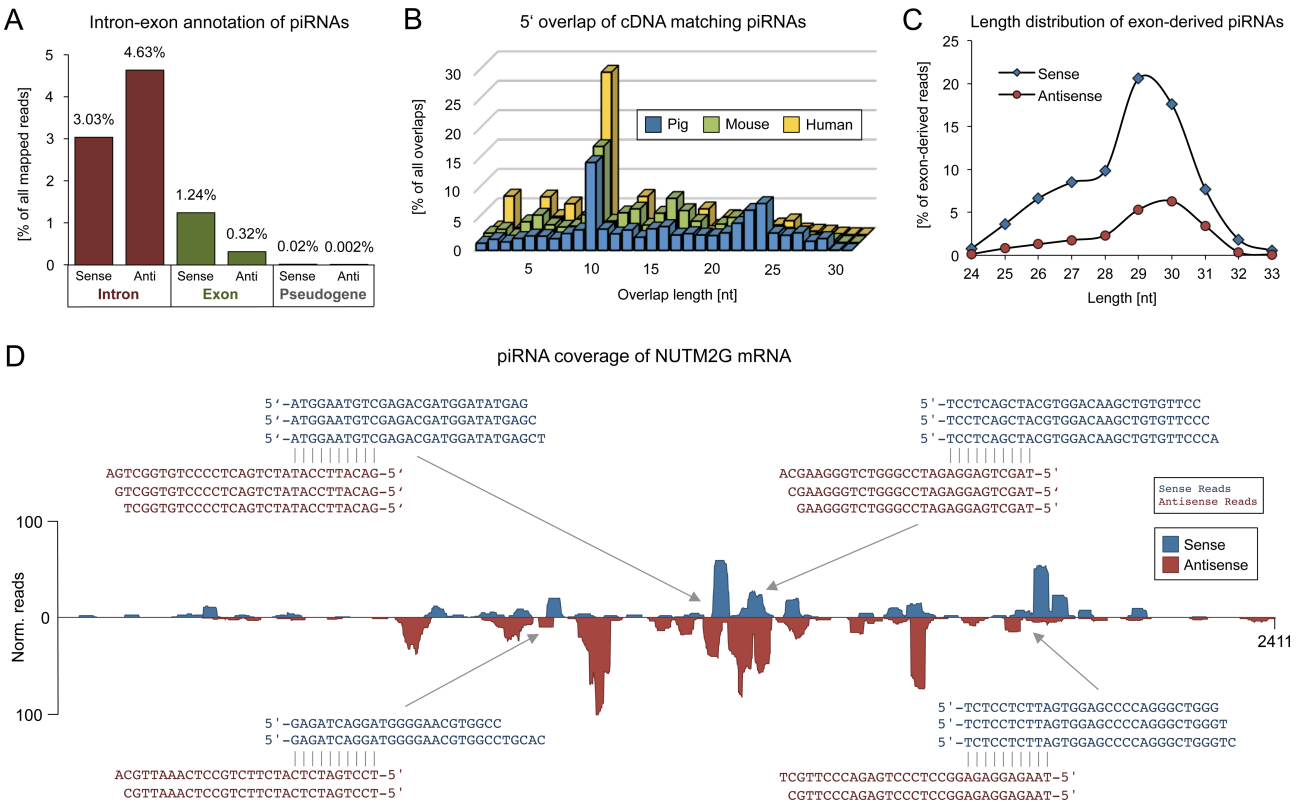


Figure 3 | Gene-derived piRNAs. (A) Portions of piRNA reads mapping to introns, exons and pseudogenes in the porcine genome. (B) 5' overlap of testis piRNAs from pig, mouse and human, mapping to corresponding annotated cDNA. In all three species a high rate of 10 nt 5' overlaps is detectable. (C) Length distribution of sense and antisense exon-derived piRNAs. (D) Mapping of piRNA reads to the mRNA sequence of the protein-coding gene NUTM2G. Exemplary sites with 10 nt 5' overlap between sense and antisense piRNA reads are indicated by arrows.

In addition, we noticed that most conserved target genes represent factors with nuclear localization that interact with DNA. Therefore we performed a GO term enrichment analysis [40] for all identified human and mouse targets with respect to the cellular component and the molecular function (S3 File, porcine data not available). Indeed, we found a significant association with the term nucleus for both human and mouse targets ($p = 0.0008$, $p = 0.0055$, respectively) compared to a non-significant association with the term cytoplasm ($p = 1$, $p = 0.2503$, respectively). Regarding the molecular function of targets we observed that human as well as mouse targets are significantly associated with the term nucleic acid binding ($p = 0.0018$, $p = 0.0163$). Together, these results suggest that post-transcriptional gene regulation by the Piwi/piRNA system mainly concerns nuclear factors with DNA binding activity.

4.4.4. tRNA-derived sRNAs

The by far largest proportion of sRNA reads that has been annotated as known ncRNA is represented by sequences that map perfectly to tRNAs and that are known as tRNA related fragments (tRFs [41], Fig 4A). Interestingly, the identified tRFs share striking similarities with piRNAs.

First, the sequence length distribution of tRNA-derived sRNAs ranges mainly from 29 to 32 nt which corresponds to the typical size of mammalian piRNAs, though we note that the length profile of the tRNA-related reads is much sharper, possibly indicating differences in biogenesis (Fig 4B).

Second, comparison with our control sRNA library without a sodium periodate treatment step reveals a less marked enrichment of tRNA-derived sequences (6.6% vs. 11.6%) while the share of other ncRNAs such as miRNAs, snRNAs, snoRNAs and rRNAs is increased (1.3% vs. 0.4%) (Fig 4A). This suggests that tRNA-derived sRNA sequences are not eliminated by sodium periodate treatment,

presumably because of a modification at their 3' end that protects them from degradation like 2'-O-methylation in case of piRNAs.

Third, tRNA-derived sequences are not randomly distributed among the various tRNA types, but rather derive mainly from the 5' ends of five tRNA types, namely Asp-GTC, Glu-TTC, Glu-CTC, Gly-CCC, and Gly-GCC, altogether accounting for 98% of all tRNA-derived sRNA reads (Fig 4C–4E). As a consequence, about 90% of tRNA-derived reads start with a uracil. In contrast, this share reaches only 77% for the non-oxidized library with a multiple of tRNA-derived sequences that do not match the 5' end of a tRNA (S2 Fig), which indicates the presence of random tRNA degradation products that are efficiently eliminated by periodate treatment. As opposed to 5'-end-derived reads (99.56%), only a minor share (0.01%) maps to the 3'-ends of tRNAs. In the light of the different length profiles of 5' tRFs (18–33 nt) and 3' tRFs (18–22 nt) [41] we suppose that the observed bias is most likely introduced by the applied cloning procedure that favors molecules larger than 24 nt.

Nonetheless, tRNA-derived sRNAs also exhibit features that clearly separate them from regular piRNAs. Interestingly, while protein-coding loci are not targeted at all, 73.5% (1,066,063 reads; 1647 non-identical sequences) of all tRNA-derived sRNA reads that map to the genome match genomic TE copies in sense (99.9%) according to RepeatMasker annotation. Not surprising, these almost exclusively represent tRNA-derived SINEs (99.1%). Finally, the share of tRNA-derived sRNA reads antisense to tRNA sequences is similarly marginal (0.002%) and a ping-pong signature is not detectable.

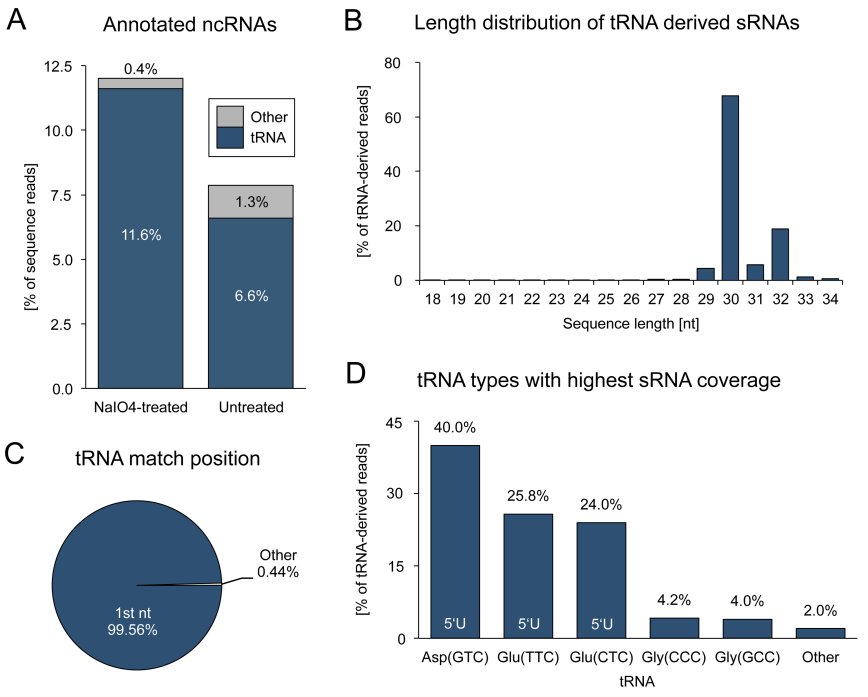


Figure 4 | tRNA-derived small RNAs. (A) Fractions of sRNAs that were annotated as known ncRNA in a sodium periodate treated and untreated sRNA library. (B) Length distribution of tRNA-derived sRNA reads. (C) Positions on tRNAs matched by 5' ends of sRNA reads. (D) Shares of sRNA reads mapping to distinct tRNAs. (E) Alignment of tRNA sequences and their most abundant matching sRNAs (numbers refer to read counts).

4.4.5. Identification and characterization of piRNA clusters

Using proTRAC [35], overall 142 piRNA clusters larger than 10 kb were identified, of which 114 are unidirectional and 28 are bidirectional, altogether comprising 3.8 Mb (S4 File). These piRNA clusters are unevenly distributed across the genome, but can be found on every chromosome except for chromosomes 16 and Y (S3 Fig). The majority of total mapped sRNA reads (91%) and mapped non-identical sequences (63%) falls into the identified piRNA clusters.

In depth analyses of the distribution of transposon classes and families in piRNA clusters compared to the genomic situation revealed interesting differences in TE composition. ERV1 and ERV2 elements are highly overrepresented in piRNA clusters (9.1% and 0.7%) as compared to their total genomic amount (3.0% and 0.3%) (Fig 5A). At the same time, ERV1 and ERV2 elements exhibit the lowest average sequence divergence to their consensus compared to other TE classes, which implicates younger propagation events and recent activity of these elements. On the other side, CR1, L1, Mariner/Tc1, other DNA transposons and other Non-LTR elements are underrepresented in piRNA clusters, while showing a tendency for increased sequence divergence, typical for older transposon copies.

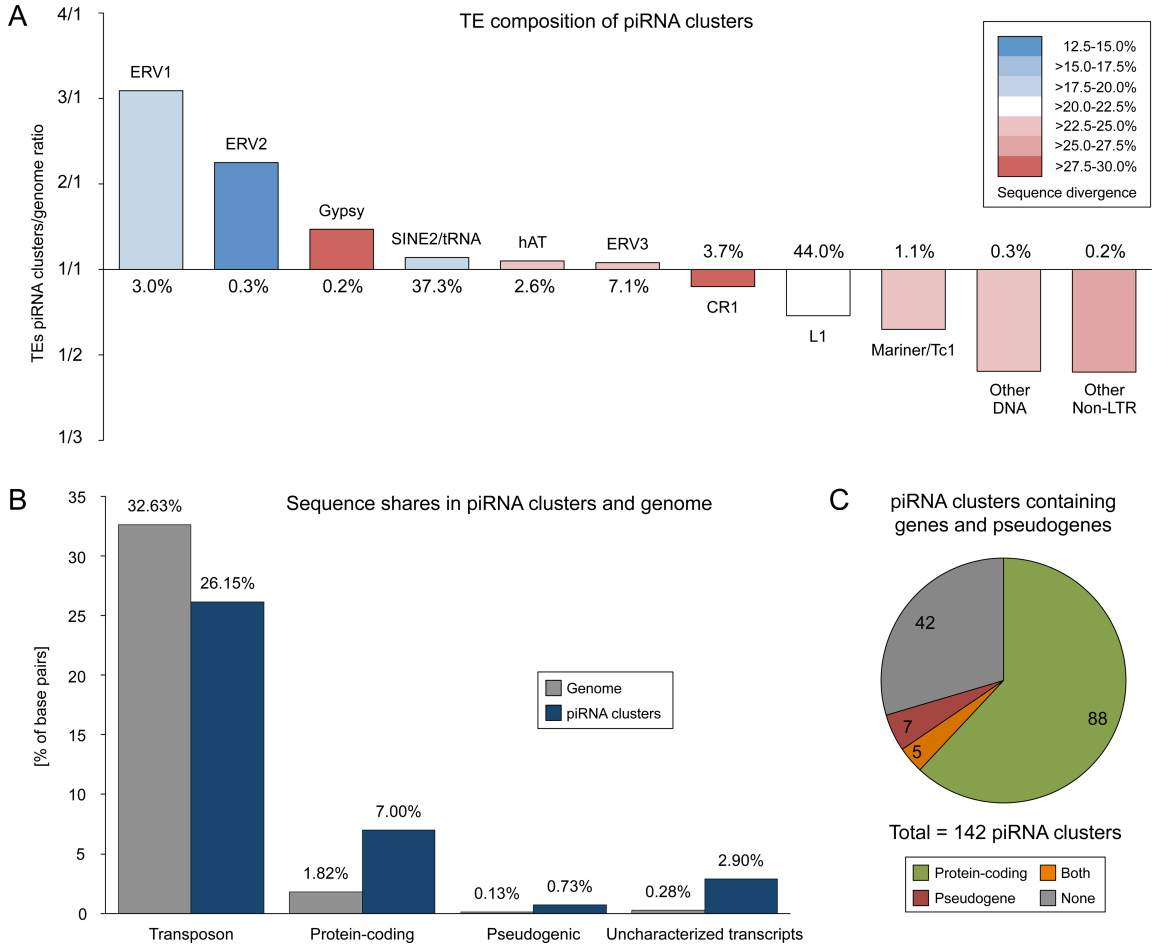


Figure 5 | Sequence characterization of piRNA clusters. (A) TE composition of predicted piRNA clusters compared to the genomic sequence of the pig. Percentages represent the share of a TE group in the genome. A ratio above 1 indicates an enrichment of a TE group in piRNA clusters, while a ratio below 1 indicates the depletion of a TE group in piRNA clusters. Different colors express the sequence divergence of a TE group to its consensus. (B) Sequence shares of TEs, protein-coding genes, pseudogenes, and uncharacterized transcribed sequences within piRNA clusters compared to the whole genome of the pig. (C) Number of piRNA clusters containing sequences of protein-coding genes, pseudogenes or both within the same piRNA cluster.

Notably, although piRNA clusters are apparently enriched for young TEs, the overall amount of transposon sequences within piRNA clusters is considerably reduced (26.2%) as compared to the whole genome (32.6%) (Fig 5B). In contrast, exonic sequences of both protein-coding genes (7.0%) and pseudogenes (0.73%) are highly enriched. Moreover, uncharacterized transcribed sequences are drastically increased in piRNA clusters (2.9%).

Overall 93 of the 142 identified piRNA clusters contain exonic sequences of protein-coding genes, while 12 contain pseudogene sequences (Fig 5C, S5 File). Only a minority of 42 piRNA clusters contains neither. We checked whether predicted piRNA clusters that span exonic sequences may simply correspond to mRNAs that are subject to primary piRNA processing. In this case we would expect piRNAs to map exons in sense orientation while no piRNAs should match to the according intronic regions. Indeed we could verify this pattern for 69 predicted piRNA clusters comprising exonic sequences that lie in sense direction of the predicted piRNA cluster and that are not producing antisense piRNAs. Since the exon-matching piRNAs also generally exhibit a high 1U rate we assume these loci to represent genes whose transcripts are processed to primary piRNAs without subsequent ping-pong amplification.

Intriguingly, 62 predicted piRNA clusters comprising both, mono- and bidirectional clusters, cover protein-coding genes in opposite orientation with regards to the predicted transcription directionality of the piRNA cluster. Further, 8 out of 12 pseudogenes within piRNA clusters are oriented in antisense direction relative to the main strand of the piRNA cluster.

While piRNA reads mapping to the main strand, which corresponds to the putative primary piRNA cluster transcript, are distributed across the entire piRNA cluster sequence, piRNAs matching the

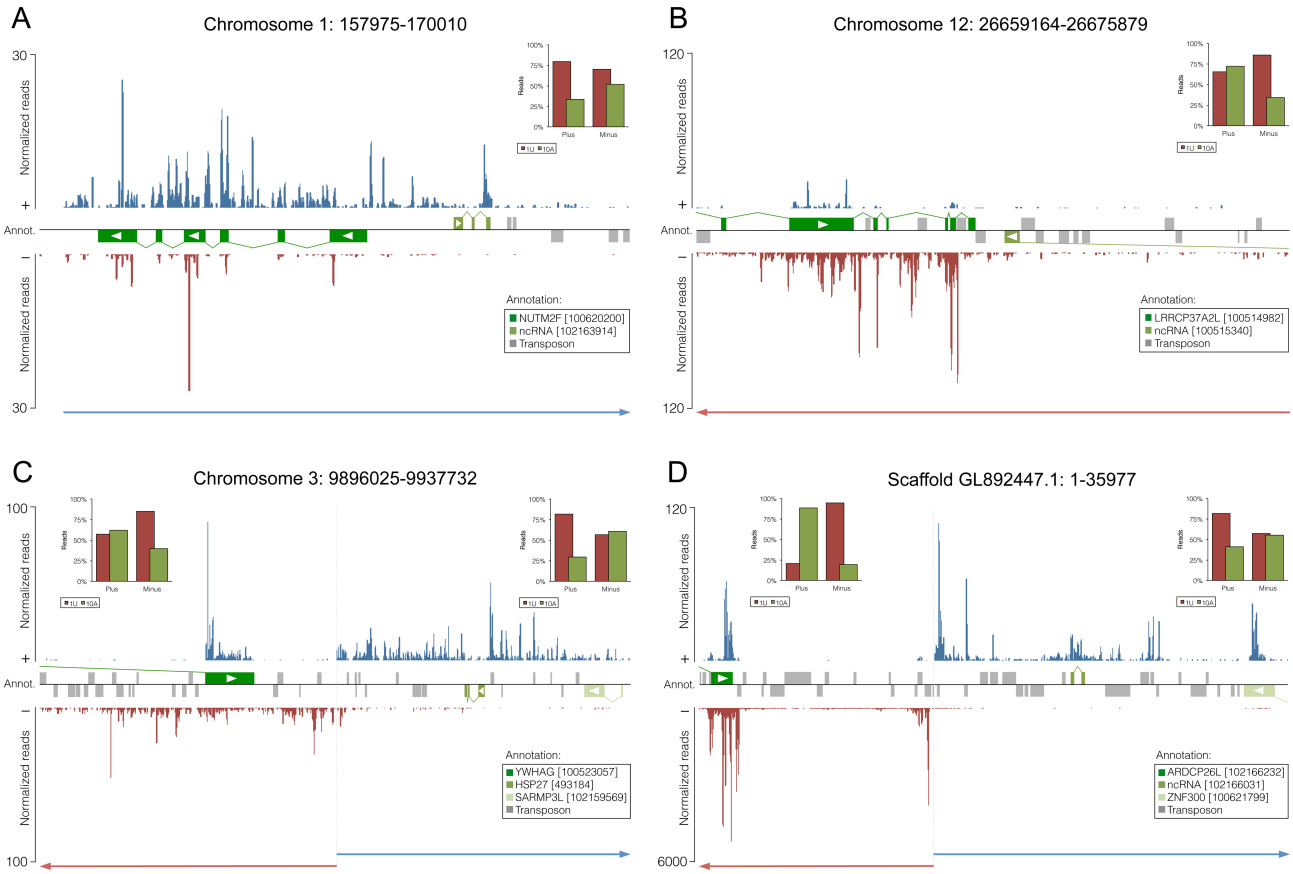


Figure 6 | piRNA clusters containing protein-coding genes or pseudogenes. Mapping of piRNA reads on plus and minus strands of piRNA cluster sequences combined with RefSeq (NCBI) annotation of transcribed sequences and RepeatMasker annotation of TEs. NCBI GeneIDs for transcribed sequences are stated in brackets. Directions of transcription for RefSeq sequences are indicated by white arrows.

opposite strand are largely restricted to the exonic regions of the corresponding overlapping gene (Fig 6A–6D and 7A). Notably, the latter generally exhibit a reduced 1U rate but an increased 10A rate as compared to main strand reads. These data strongly suggest that primary antisense piRNAs produced from these loci are targeting spliced transcripts of genes that are transcribed from the opposite strand, and that this targeting is followed by secondary piRNA biogenesis (Fig 7B).

Overall 24% of the piRNA reads that match porcine cDNA sequences originate from predicted piRNA clusters. Interestingly, cDNA-matching piRNA reads that lie outside of piRNA clusters are strongly biased towards sense sequences (88%) indicating mainly primary processing of the according transcripts. In contrast, cDNA-derived reads that can be assigned to piRNA clusters exhibit a nearly balanced ratio of sense versus antisense reads (55% and 45%, respectively). This points to a central role of piRNA clusters in the processing of specific protein-coding gene transcripts within the ping-pong cycle of the Piwi/piRNA pathway.

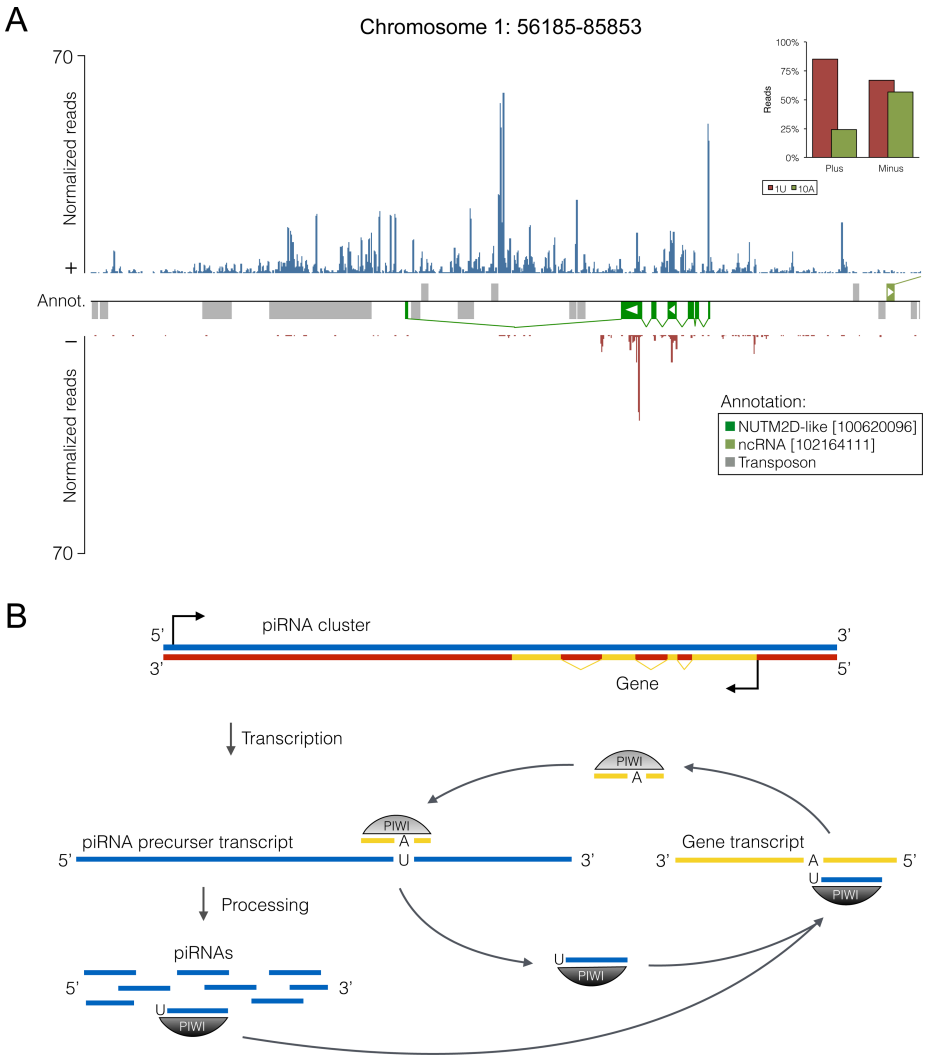


Figure 7 | Model of post-transcriptional regulation of protein-coding genes by the Piwi/piRNA pathway. (A) piRNA cluster containing a protein-coding gene. (B) Hypothetical model of post-transcriptional gene regulation mediated by piRNA clusters, based on data of porcine piRNAs. piRNA clusters containing sequences of genes or pseudogenes in reverse orientation relative to the cluster directionality can presumably produce primary piRNAs complementary to spliced mRNA, which can direct the decay of such transcripts and produce secondary piRNAs within the ping-pong amplification loop.

4.5. Discussion

Studies on model organisms like *Drosophila* and mouse have been highly informative relating to the functions and the molecular mechanisms of the Piwi/piRNA pathway. However, these organisms do not reflect the equipment of Piwi paralogs in most mammals, including human [26]. In this respect, the pig with its full set of four mammalian Piwi paralogs is more comparable to humans. Furthermore, the availability of a high quality porcine genome assembly combined with a thorough annotation of porcine TEs, along with powerful molecular biological tools [42,43] render the pig a suitable model for Piwi/piRNA research. Our extensive characterization of the porcine piRNA transcriptome represents the initial step on the way to understand piRNA function in the pig and to obtain a broader knowledge of the Piwi/piRNA pathway in mammals.

Considering their main features, porcine piRNAs closely reflect previously described characteristics of mammalian piRNAs. The length distribution of pig piRNAs ranges mainly from 24–32 nt, though the majority of 20–25 nt sized sRNAs was also found to exhibit typical piRNA characteristics and could be mapped to predicted piRNA clusters, thus most likely representing genuine piRNAs rather than non-oxidized siRNAs. Further, porcine piRNAs expressed in the adult testis show a strong bias for 1U and only a marginal bias for 10A, suggesting that the bulk originates from primary processing, while only a small fraction results from ping-pong amplification. However, while previous studies on porcine piRNAs did not report any ping-pong signatures [28,29], they are apparent in our data, clearly demonstrating that ping-pong-mediated silencing is active also in the adult germline. Moreover, though a recent study on porcine piRNAs reported the absence of ping-pong signatures [29], we could on the contrary validate our findings (ping-pong-z-score = 8.7) using the data produced by Kowalczykiewicz and colleagues (NCBI Gene Expression Omnibus (GEO); accession number GSE57414). Interestingly, we could also show a ping-pong signature in the corresponding sRNA dataset obtained from pig ovaries (ping-pong-z-score = 28.5), though on a very low level (S6 File). In line with this, piRNA expression and ping-pong-signatures in the female germline were also very recently described in human, macaque and bovine piRNA populations by Roovers and coworkers [44].

The analyzed piRNAs in our study, isolated from whole testes and thus representing a mixture of piRNAs from all germ cell stages (pachytene and pre-pachytene), are clearly depleted of TE-related sequences compared to the total genomic amount of TEs in pig. This is in line with findings from the mouse model in that only meiotically (pre-pachytene) expressed piRNAs are enriched for TE-related sequences and participate in the ping-pong cycle to repress TEs that become active during global de- and re-methylation in spermatogenesis [9,13]. This gives rise to the question whether piRNAs, especially in pachytene stages may be involved in functions beyond TE silencing.

4.5.1. tRNA-derived sRNAs with piRNA characteristics

piRNAs exhibit a methylation of the 2'-hydroxyl group at their 3' end and are therefore protected from sodium periodate-mediated ß-elimination [45,46]. RNA molecules lacking this modification are removed during library preparation [47]. Comparing oxidized and non-oxidized libraries, we noted an enrichment of tRNA-related sRNAs after periodate treatment, while sRNAs related to other ncRNA types almost completely disappeared. This suggests that they may also carry a 3' methylation similar to piRNAs that prevents their decay. Indeed, methylation of tRNA nucleotides is a common phenomenon and 2'-O-methylation of nucleotides 30 to 32 is described for tRNAs of many mammalian species [48], although data on porcine tRNAs is lacking.

Another interesting characteristic is that nearly all tRNA-derived sRNAs originate from the 5' ends of only five different tRNA types, with the majority of them starting with a uracil. tRFs [41], such as the 5' tRNA halves that we describe here, along other types of short fragments of tRNAs like 5' tRFs, 3'

tRFs and 3' tRNA halves have been previously found in many different species [49–52,38]. Presumably, 5' tRNA halves are produced by a conserved stress response mechanism in eukaryotes [53] and play a role in translational regulation [54], as well as impact the siRNA pathway by inhibiting Dicer activity [55]. Some 5' tRFs have been shown to be produced by Dicer, bound by Argonaute proteins and further to carry blocked 2' hydroxyl termini [56]. With regard to their biological role, 5' tRFs have been implicated in gene regulation [57], e.g. by inhibition of protein translation, which does not require complementary base pairing [58]. Also, tRFs have been reported very recently to be present in male and female gonads of the pig [29], although the composition of tRNA types differed notably from our results.

Recently, the Piwil1 homolog Marwi of the common marmoset has been found to bind considerable amounts of tRNA-derived sRNAs, which exhibit very similar characteristics as described here [59]. Furthermore, various tRFs associate with the human Piwil2 homolog Hiwi2 [60] and the *Tetrahymena* Piwi Twil2 [61]. In addition, short tRNA sequences have been previously described as piRNAs in several organisms such as rat, human [4], mouse [62] and hamster [63].

We speculate that generally all tRNAs should be subject to a processing mechanism that yields 5' tRFs but that Piwi proteins are loaded only with 1U fragments that a priori carry a 3' methylation as do the corresponding tRNAs. Therefore, we hypothesize that the described tRNA-derived sRNAs literally represent piRNAs in that they interact with Piwi proteins. However, since we did neither detect a ping-pong signature nor identified putative complementary target transcripts, their biological role, if any, may be limited to functions that are not related to the Piwi pathway.

4.5.2. Repression of transposable elements

Silencing of transposons is regarded as the major task of piRNAs in the animal germline [19] and a considerable amount of porcine piRNA sequences indeed maps to TE sequences. Consistent with the fact that the share of TE sequences in the porcine genome is lower than reported for other mammalian genomes [27], the proportion of TE-derived piRNAs is likewise reduced with respect to other species.

The elevated shares of piRNAs mapping to tRNA-derived SINEs and especially to ERVs compared to the genomic amount of these elements might reflect a recent activity of these transposon classes in the porcine genome. Indeed, ERV1 elements have been found to show hints of recent activity on the pig lineage and an increased insertion rate at pig specific evolutionary breakpoint regions [27], while tRNAGlu-derived SINEs, a cetartiodactyl specific TE superfamily [64], have been found to be overrepresented in cetartiodactyl evolutionary breakpoint regions [65]. What further supports the presumption of a recent activity is the fact that ERV1, ERV3 and tRNA-derived SINEs show the least sequence divergence to their consensus compared to other TE classes, pointing to a younger age and more recent activity. These TEs, foremost ERV1, are also enriched in the predicted piRNA clusters identified here. This suggests not only that the Piwi/piRNA system is highly adaptable, but it also might indicate that piRNA clusters can act more dynamically and/or selectively than commonly thought.

Hypothetically, new piRNA clusters might emerge at sites with a high rate of recent integrations of active TEs. On the other hand, since piRNA clusters represent transcriptionally highly active regions in the genome, non-inert TEs might more likely integrate into such regions than into sites that have a more closed chromatin structure. Contrasting this intuitive assumption, piRNA clusters are not enriched for TEs, but on the contrary are poorer of TE sequences compared to the remaining genome. Apparently there must be either an efficient TE insertion avoidance mechanism or alternatively natural selection against the accumulation of TEs into piRNA clusters which could explain the general bias towards non-TE sequences.

4.5.3. Regulation of protein-coding genes

The first identification of piRNAs derived from protein-coding genes dates back to the initial description of piRNAs [4–6], but a regulatory role was not considered even in following studies [14] due to a lack of antisense piRNAs. A later report showed that the 3' untranslated regions (3' UTRs) of a set of mRNAs in murine testes are processed into primary piRNAs, while no secondary piRNAs or signs of ping-pong processing could be observed [24]. Indeed, we confirm that the mapping density (reads per kb) of porcine piRNAs on cDNA is highest on 3' UTRs, which however can be partly explained by the fact, that 3' UTRs are enriched for TE sequences compared to 5' UTR and coding sequence, though the share of TE-related piRNA reads mapping to 5'- and 3' UTRs does not differ substantially (S4 Fig).

In this study we found that both sense and antisense piRNAs map to exonic sequences of protein-coding genes, showing marked ping-pong signatures resulting from sense and antisense reads derived from mRNA sequences of a large number of genes. Moreover, the length distribution of exon-derived piRNAs indicates the participation of different Piwi paralogs in their generation. Together, this suggests that gene transcripts are processed into piRNAs within the ping-pong cycle.

A central role for this process, as known for TEs, seems to be occupied by piRNA clusters. piRNAs mapping to both strands at exonic regions of piRNA clusters that span genes in reverse direction, as well as their opposing 1U and 10A rates suggest that piRNAs antisense to the corresponding gene are produced in primary biogenesis from large cluster transcripts. These primary piRNAs can in turn guide the piRNA-induced silencing complex (piRISC) machinery to target mRNAs that enter the ping-pong cycle to generate secondary sense piRNAs (Fig 7B). In support of this model, the majority of antisense gene-related reads derives from piRNA clusters, although only a quarter of all gene-derived reads can be assigned to piRNA clusters. Overall, these observations reveal a mechanism by which antisense piRNAs are produced to direct mRNA processing and exert Piwi-mediated post-transcriptional regulation on protein-coding genes.

Finally, the fact that specific genes are targeted not only in pig but also in human and mouse suggests a conserved biological function during eutherian divergence. In support of this, GO term enrichment analysis revealed that targeted genes mainly represent factors with nuclear localization and DNA binding activity, suggesting their involvement in transcriptional regulation and chromatin modification. These results strengthen findings from a previous study on porcine piRNAs that revealed similar patterns regarding possible piRNA target genes but lacks a quest for ping-pong signatures [28].

Whether the processing of gene transcripts by the Piwi/piRNA pathway, foremost within the ping-pong cycle, has a significant effect on transcription levels yet has to be investigated. However, it is likewise conceivable that target genes are not extensively silenced, but rather experience a fine-tuning of their expression. The specific role of targeted transcripts in spermatogenesis is yet unresolved. Though many of the highly targeted transcripts in human such as DNM1P46, GOLGA2P11, NPAP1P6 or FBXO25 are exclusively or mainly expressed in testis according to Expression Atlas data [66], evidence for an involvement in spermatogenesis is generally lacking. One exception is the NPAP1 gene (alias c15orf2) which has been linked to spermatogenesis and male infertility in human [67].

Our findings line up into a range of results from previous studies on mammalian piRNAs and reinforce the idea that piRNAs are involved in post-transcriptional gene regulation. Recently, it has been demonstrated that pachytene piRNAs direct mRNA elimination during late spermatogenesis in mouse [25]. Importantly, a very recent study [68] led to observations similar to ours regarding ping-pong-mediated mRNA processing in mouse testis. It further showed that the proper turnover of certain key piRNA targets seems to be essential for sperm formation, strengthening the concept of an important role for the Piwi pathway in the regulation of protein-coding genes.

Moreover, analyses of testis expressed piRNAs from the common marmoset also showed that pseudogenes are located in piRNA clusters and tend to be in reverse orientation relative to piRNA cluster directionalities [59]. However, these pseudogenic regions were only covered by piRNAs on one strand, whereas one would expect signs of a ping-pong signature if these piRNAs would participate in Piwi-mediated silencing of the corresponding genes. Going back to the initial description of testis expressed piRNAs in mouse, protein-coding genes have been found to overlap with piRNA cluster sequences, though possible gene regulatory functions were ruled out because of a lack of gene-derived antisense piRNAs [6]. Nevertheless, the existence of piRNA clusters containing gene or pseudogene sequences is not pig specific, but likely a widespread phenomenon.

Interestingly, antisense transcripts for NUTM2A (lncRNA), NUTM2B (lncRNA) and NUTM2D (ncRNA) and other target genes are predicted for human according to the HAVANA genome annotation. In addition, though only very few porcine lncRNAs are annotated, sRNA reads derived from such sequences show clear piRNA characteristics, such as a marked ping-pong signature and 1U and 10A bias (S5 Fig). Concordantly, putative piRNAs have been recently found to map to lncRNA sequences in humans [37]. This suggests that (long) non-coding RNAs are processed into primary piRNAs or alternatively represent primary piRNA cluster transcripts, which appears to be rather a matter of definition.

In summary, the enrichment of protein-coding gene sequences together with the evidence for their ping-pong-mediated post-transcriptional processing, and the presence of rather young transposon classes accompanied by an overall reduced amount of transposons in piRNA clusters challenge the model of passive transposon traps. Extending this traditional view, we consider it possible that piRNA clusters might specifically arise at genomic loci whose transcripts (protein-coding or not) require control by the Piwi/piRNA system, yielding a beneficial, positively selectable mechanism for the host organism. Clearly, this hypothesis has to be further addressed in the future.

4.6. Declarations

Acknowledgments

Thanks go to Bertram Brenig (Georg-August-University Göttingen, Institute of Veterinary Medicine) for providing pig testis material and Holger Herlyn for helpful discussions. We further thank Christine Driller, Sacha Heerschop, Julia Schumacher and Dana Thiele for helpful comments and discussion.

Author Contributions

Conceived and designed the experiments: DR HZ. Performed the experiments: DG. Analyzed the data: **DG** DR. Contributed reagents/materials/analysis tools: DR. Wrote the paper: **DG** DR RFK HZ.

4.7. References

1. Ketting RF. The many faces of RNAi. Dev Cell. 2011; 20: 148–161.
2. Lin H, Spradling AC. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. Development. 1997; 124: 2463–2476.
3. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006; 442: 203–207.
4. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006; 442: 199–202.
5. Grivna ST, Beyret E, Wang Z, Lin H. A novel class of small RNAs in mouse spermatogenic cells. Genes Dev. 2006; 20: 1709–1714.
6. Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, et al. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev. 2006; 20: 1732–1743.

7. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 2007; 128: 1089–1103.
8. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, et al. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science. 2007; 315: 1587– 1590.
9. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 2008; 31: 785–799.
10. Cora E, Pandey RR, Xiol J, Taylor J, Sachidanandam R, McCarthy AA, et al. The MID-PIWI module of Piwi proteins specifies nucleotide- and strand-biases of piRNAs. RNA. 2014; 20: 773–781.
11. Kawaoka S, Izumi N, Katsuma S, Tomari Y. 3' end formation of PIWI-interacting RNAs in vitro. Mol Cell. 2011; 43: 1015–1022.
12. Kirino Y, Mourelatos Z. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. RNA. 2007; 13: 1397–1401.
13. Smallwood SA, Kelsey G. De novo DNA methylation: a germ cell perspective. Trends Genet. 2012; 28: 33–42.
14. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science. 2007; 316: 744–747.
15. Carmell MA, Girard A, van de Kant HJG, Bourc'his D, Bestor TH, de Rooij DG, et al. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell. 2007; 12: 503–514.
16. Vagin VV, Klenov MS, Kalmykova AI, Stolyarenko AD, Kotelnikov RN, Gvozdev VA. The RNA Interference Proteins and Vasa Locus are Involved in the Silencing of Retrotransposons in the Female Germline of Drosophila melanogaster. RNA Biol. 2004; 1: 54–58.
17. Kalmykova AI, Klenov MS, Gvozdev VA. Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. Nucleic Acids Res. 2005; 33: 2052–2059.
18. Xu M, You Y, Hunsicker P, Hori T, Small C, Griswold MD, et al. Mice deficient for a small cluster of Piwiinteracting RNAs implicate Piwi-interacting RNAs in transposon control. Biol Reprod. 2008; 79: 51–57.
19. Malone CD, Hannon GJ. Small RNAs as guardians of the genome. Cell. 2009; 136: 656–668.
20. Nishida KM, Saito K, Mori T, Kawamura Y, Nagami-Okada T, Inagaki S, et al. Gene silencing mechanisms mediated by Aubergine–piRNA complexes in Drosophila male gonad. RNA. 2007; 13: 1911– 1922.
21. Nagao A, Mituyama T, Huang H, Chen D, Siomi MC, Siomi H. Biogenesis pathways of piRNAs loaded onto AGO3 in the Drosophila testis. RNA. 2010; 16: 2503–2515.
22. Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, Sakota E, et al. A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. Nature. 2009; 461: 1296–1299.
23. Rouget C, Papin C, Boureux A, Meunier AC, Franco B, Robine N, et al. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. Nature. 2010; 467: 1128–1132.
24. Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, et al. A broadly conserved pathway generates 3'UTR-directed primary piRNAs. Curr Biol. 2009; 19: 2066–2076.
25. Gou LT, Dai P, Yang JH, Xue Y, Hu YP, Zhou Y, et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. 2014; 24: 680–700.
26. Seto AG, Kingston RE, Lau NC. The coming of age for Piwi proteins. Mol Cell. 2007; 26: 603–609.
27. Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 2012; 491: 393–398.
28. Liu G, Lei B, Li Y, Tong K, Ding Y, Luo L, et al. Discovery of potential piRNAs from next generation sequences of the sexually mature porcine testes. PLoS One. 2012; 7: e34770.
29. Kowalczykiewicz D, Swiercz A, Handschuh L, Le niak K, Figlerowicz M, Wrzesinski J. Characterization of Sus scrofa Small Non-Coding RNAs Present in Both Female and Male Gonads. PLoS One. 2014; 9: e113249.
30. Rajasethupathy P, Antonov I, Sheridan R, Frey S, Sander C, Tuschl T, et al. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. Cell. 2012; 149: 693–707.
31. Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res. 2004; 32: D109–D111.
32. Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. 2009; 37: D93–D97.
33. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008; 24: 2395–2396.
34. Jurka J. Repbase Update a database and an electronic journal of repetitive elements. Trends Genet. 2000; 16: 418–420.
35. Rosenkranz D, Zischler H. proTRAC—a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics. 2012; 13: 5.
36. Antoniewski C. Computing siRNA and piRNA overlap signatures. Methods Mol Biol. 2014; 1173: 135–46.
37. Ha H, Song J, Wang S, Kapusta A, Feschotte C, Chen KC, et al. A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. BMC Genomics. 2014; 15: 545.
38. Peng H, Shi J, Zhang Y, Zhang H, Liao S, Li W, et al. A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm. Cell Res. 2012; 22: 1609–1612.

39. Zhang Z, Xu J, Koppetsch BS, Wang J, Tipping C, Ma S, et al. Heterotypic piRNA Ping-Pong Requires Qin, a Protein with Both E3 ligase and Tudor Domains. Mol Cell. 2011; 44: 572–584.
40. The Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25: 25–29.
41. Gebetsberger J, Polacek N. Slicing tRNAs to boost functional ncRNA diversity. RNA Biol. 2013; 10: 1798–1806.
42. Bendixen E, Danielsen M, Larsen K, Bendixen C. Advances in porcine genomics and proteomics—a toolbox for developing the pig as a model organism for molecular biomedical research. Brief Funct Genomics. 2010; 9: 208–219.
43. Hai T, Teng F, Guo R, Li W, Zhou Q. One-step generation of knockout pigs by zygote injection of CRISPR/Cas system. Cell Res. 2014; 24: 372–375.
44. Roovers EF, Rosenkranz D, Mahdipour M, Han CT, He N, Chuva de Sousa Lopes SM, et al. Piwi proteins and piRNAs in mammalian oocytes and early embryos. Cell Rep. 2015.
45. Kirino Y, Mourelatos Z. Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. Nat Struct Mol Biol. 2007; 14: 347–348.
46. Ohara T, Sakaguchi Y, Suzuki T, Ueda H, Miyauchi K, Suzuki T. The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated. Nat Struct Mol Biol. 2007; 14: 349–350.
47. Behm-Ansmant I, Helm M, Motorin Y. Use of specific chemical reagents for detection of modified nucleotides in RNA. J Nucleic Acids. 2011; 2011: 408053.
48. Machnicka MA, Milanowska K, Osman Oglou O, Purta E, Kurkowska M, Olchowik A, et al. MODOMICS: a database of RNA modification pathways—2013 update. Nucleic Acids Res. 2013; 41: D262– D267.
49. Lee SR, Collins K. Starvation-induced cleavage of the tRNA anticodon loop in Tetrahymena thermophila. J Biol Chem. 2005; 280: 42744–42749.
50. Calabrese JM, Seila AC, Yeo GW, Sharp PA. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. Proc Natl Acad Sci USA. 2007; 104: 18097–18102.
51. Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev. 2008; 22: 2773–2785.
52. Kawaji H, Nakamura M, Takahashi Y, Sandelin A, Katayama S, Fukuda S, et al. Hidden layers of human small RNAs. BMC Genomics. 2008; 9: 157.
53. Thompson DM, Lu C, Green PJ, Parker R. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA. 2008; 14: 2095–2103.
54. Ivanov P, Emara MM, Villen J, Gygi SP, Anderson P. Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation. Mol Cell. 2011; 43: 613–623.
55. Durdevic Z, Mobin M, Hanna K, Lyko F, Schaefer M. The RNA methyltransferase dnmt2 is required for efficient dicer-2-dependent siRNA pathway activity in Drosophila. Cell Rep. 2013; 4: 931–937.
56. Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JWS, et al. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA. 2009; 15: 2147–2160.
57. Pederson T. Regulatory RNAs derived from transfer RNA? RNA. 2010; 16: 1865–1869.
58. Sobala A, Hutvagner G. Small RNAs derived from the 5' end of tRNA can inhibit protein translation in human cells. RNA Biol. 2013; 10: 553–563.
59. Hirano T, Iwasaki YW, Lin ZYC, Imamura M, Seki MN, Sasaki E, et al. Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. RNA. 2014; 20: 1–15.
60. Keam SP, Young PE, McCorkindale AL, Dang THY, Clancy JL, Humphreys DT, et al. The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. Nucleic Acids Res. 2014; 42: 8984–8995.
61. Couvillion MT, Sachidanandam R, Collins K. A growth-essential Tetrahymena Piwi protein carries tRNA fragment cargo. Genes Dev. 2010; 24: 2742–2747.
62. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008; 453: 539– 543.
63. Gerstl MP, Hackl M, Graf AB, Borth N, Grillari J. Prediction of transcribed PIWI-interacting RNAs from CHO RNAseq data. J Biotechnol. 2013; 166: 51–57.
64. Shimamura M, Abe H, Nikaido M, Ohshima K, Okada N. Genealogy of families of SINEs in cetaceans and artiodactyls: the presence of a huge superfamily of tRNA (Glu)-derived families of SINEs. Mol Biol Evol. 1999; 16: 1046–1060.
65. The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC. The Genome Sequence of Taurine: A Window to Ruminant Biology and Evolution. Science. 2009; 324: 522– 528.
66. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, et al. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res. 2013; 42: D926–D932.
67. Färber C, Gross S, Neesen J, Buiting K, Horsthemke B. Identification of a testis-specific gene (C15orf2) in the Prader-Willi syndrome region on chromosome 15. Genomics. 2000; 65(2): 174–183
68. Zhang P, Kang JY, Gou LT, Wang J, Xue Y, Skogerboe G, et al. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. Cell Res. 2015; 25: 193–207.

4.8. Supplement

4.8.1. Supplementary figures

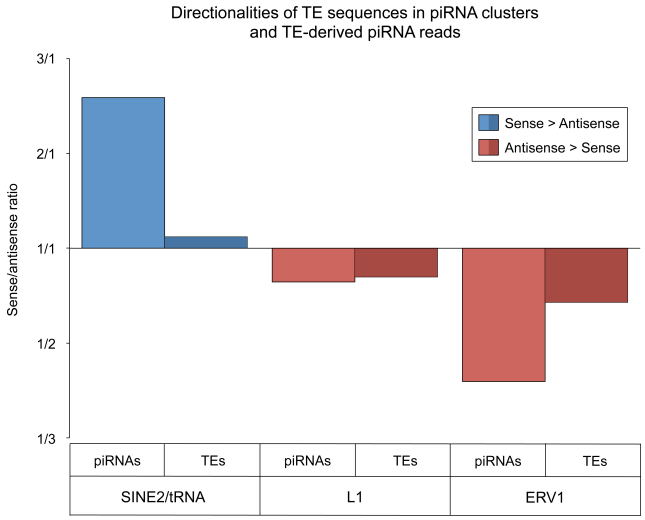


Figure S1 | Directionalities of TE sequences in piRNA clusters and TE-derived piRNA reads. Correlation between insertion bias of TE copies and strand bias of TE-related piRNAs for the TE classes with highest read coverage, tRNA-derived SINEs, L1, and ERV1.

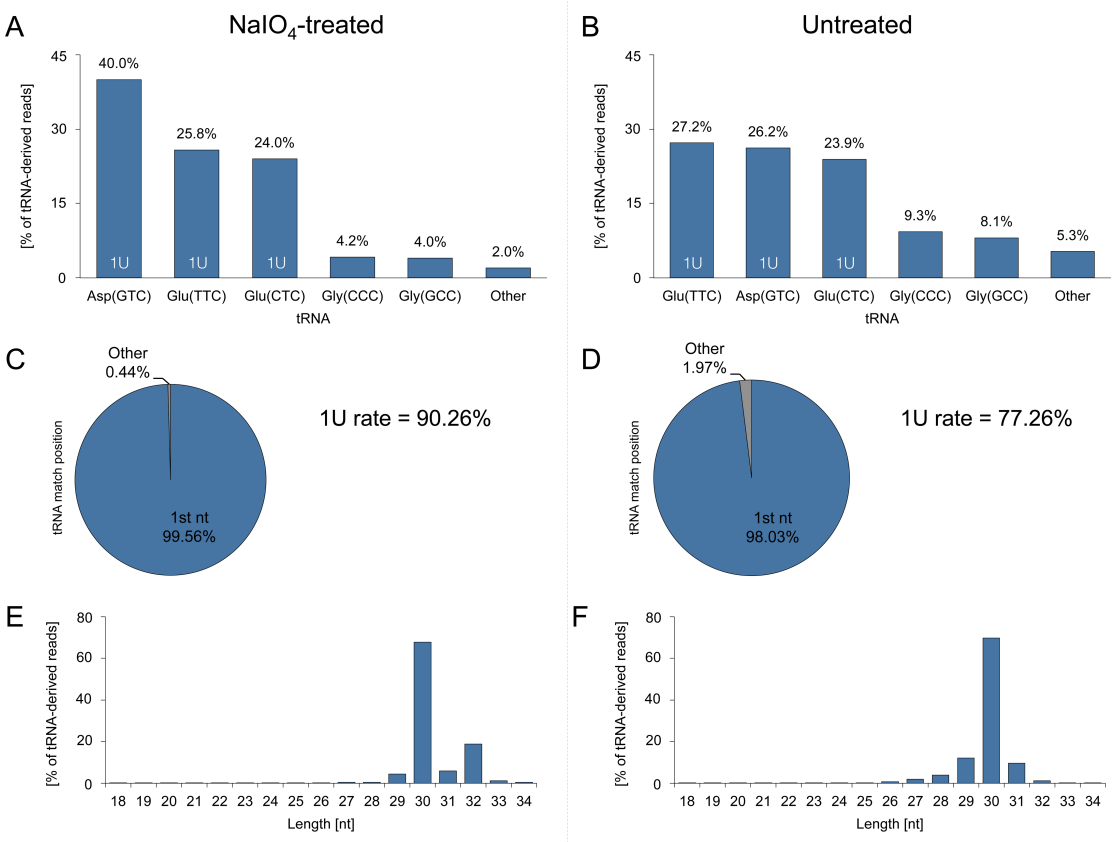


Figure S2 | Periodate treatment of tRNA-derived small RNAs. Comparison of tRNA-derived sRNAs from NaIO4-treated and untreated libraries. (A) and (B) Shares of sRNA reads mapping to distinct tRNAs. tRNAs that possess a 5' uracil are marked with 1U. (C) and (D) Positions on tRNAs matched by 5' ends of sRNA reads and 1U rates of tRNA-derived reads. (E) and (F) Length Distribution of tRNA-derived sRNA reads.

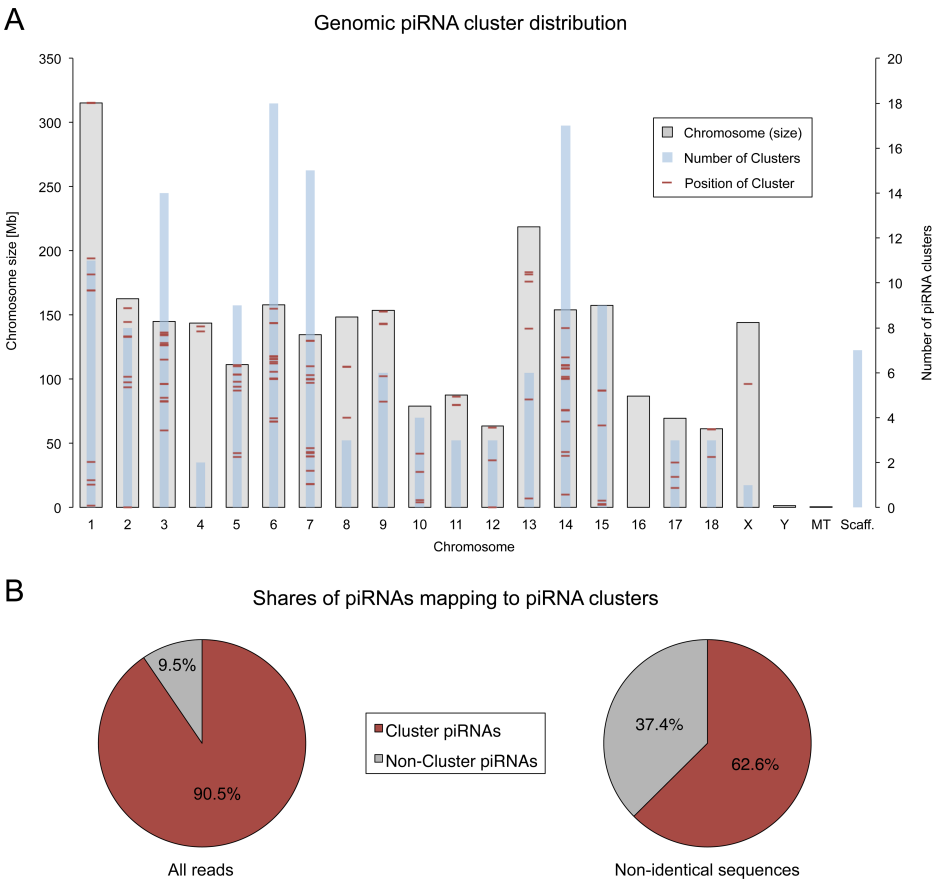


Figure S3 | Distribution of piRNA clusters and piRNAs in the porcine genome. Comparison of tRNA-derived sRNAs from NaIO4-treated and untreated libraries. (A) and (B) Shares of sRNA reads mapping to distinct tRNAs. tRNAs that possess a 5' uracil are marked with 1U. (C) and (D) Positions on tRNAs matched by 5' ends of sRNA reads and 1U rates of tRNA-derived reads. (E) and (F) Length Distribution of tRNA-derived sRNA reads.

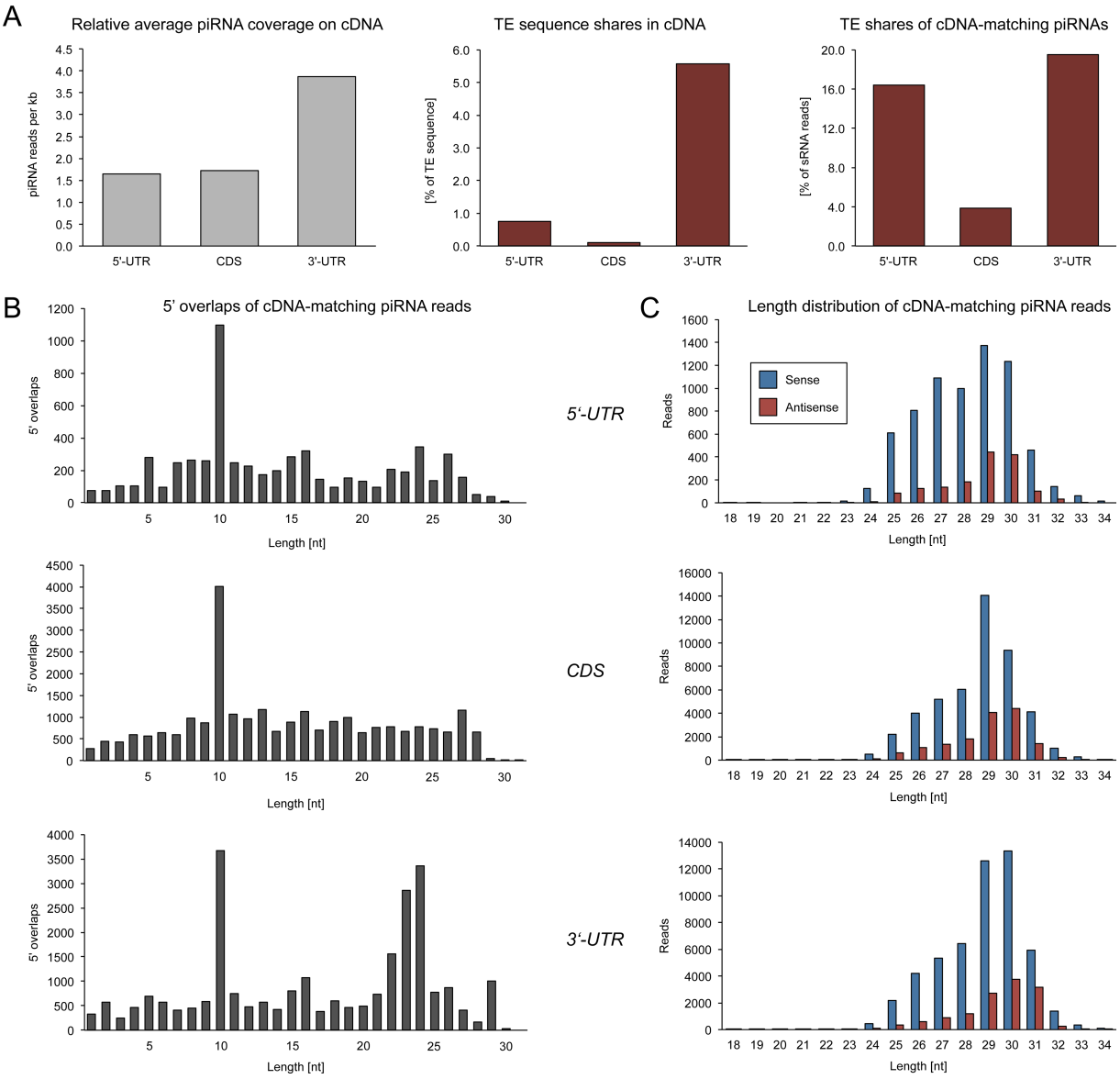


Figure S4 | Small RNA reads derived from 5'UTR, CDS and 3'UTR. (A) Relationship between piRNA mapping bias and TE enrichment of 3'UTRs. (B) 5' overlaps of piRNAs derived from 5'UTRs, CDS and 3'UTRs. (C) Length distributions of piRNAs derived from 5'UTRs, CDS and 3'UTRs.

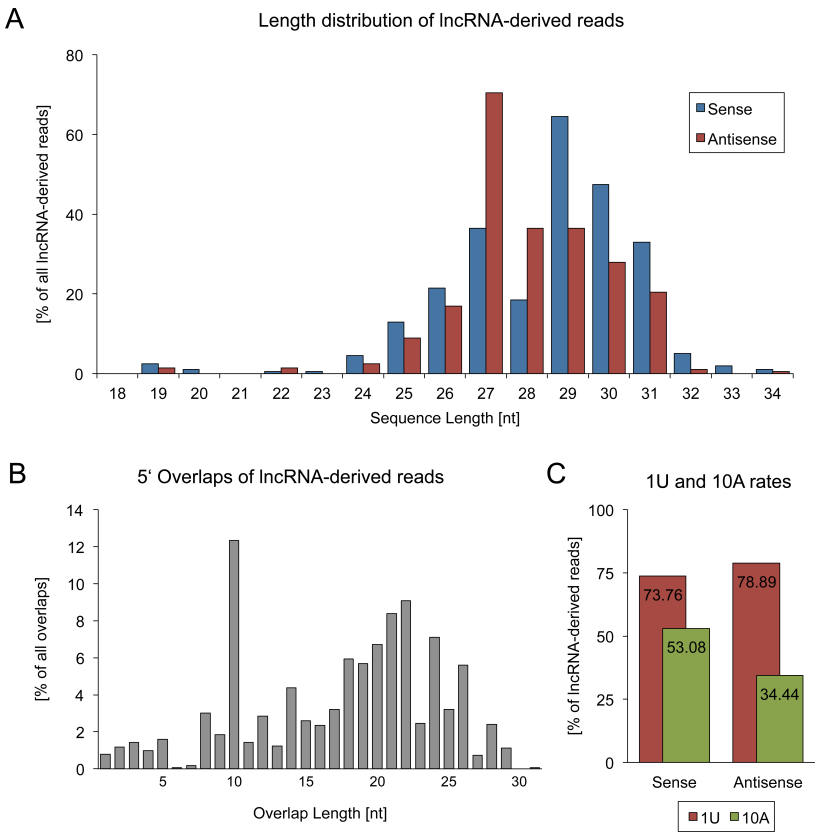


Figure S5 | Characterization of sRNA reads derived from annotated porcine lncRNAs. (A) Length distribution of sense and antisense sRNA reads. (B) 5' overlaps of sRNA reads. (C) 1U and 10 A rates of sense and antisense sRNA reads.

4.8.2. Supplementary files

Supplementary files are available at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124860.

5. Evolution of piRNA clusters and pseudogenes in primates

Daniel Gebert[1], Hans Zischler[1], David Rosenkranz[1]

[1] Institute of Organismic and Molecular Evolutionary Biology, Anthropology, Johannes Gutenberg University, 55099 Mainz, Germany

This chapter is as yet not published in a peer-reviewed journal.

5.1. Abstract

PIWI proteins and their guiding Piwi-interacting (pi-) RNAs direct the silencing of target nucleic acids in the animal germline, working on transcriptional and post-transcriptional levels. While in primordial male germ cells of mammals so-called pre-pachytene piRNAs are involved in extensive silencing of transposable elements (TEs), pachytene piRNAs that are active from the pachytene meiotic phase of spermatogenesis have additionally been shown to act in post-transcriptional regulation of protein-coding genes. The bulk of pachytene piRNAs is produced from large genomic loci, named piRNA clusters, that harbor many different TE fragments, which serve as the source for TE-targeting piRNAs. Recently, the presence of reversed pseudogene copies within piRNA clusters lead to the idea that piRNAs derived from such pseudogenes might direct regulation of their parent genes. Here, we examine primate piRNA clusters and therein contained pseudogenes in a comparative approach in order to gain a deeper understanding about the evolution of mammalian piRNA clusters in general and the putative gene regulatory role of pseudogene-derived piRNAs. Initially, we provide a broad analysis of the evolutionary relationships of piRNA clusters and their differential activity among six primate species. Subsequently, we show that pseudogenes in reserve orientation relative to cluster transcription do not show signs of selection pressure, compared to pseudogenes in parallel orientation. Further, the fact that only a minority of reversed pseudogenes produces piRNA-targeting gene transcripts that are processed within the PIWI/piRNA pathway and a weak conservation of targeting of homologous genes among species, suggest only a minor impact on gene regulation. Finally, possibly serving as an alternative explanation for the general enrichment of pseudogenes in piRNA clusters, we report that piRNA producing loci themselves tend to be located in gene-dense regions of the genome, indicating open and active chromatin, but also correlating with pseudogene abundance. Hence, the occurrence of pseudogenes in piRNA clusters might be regarded as a by-product of cluster generation.

5.2. Introduction

Piwi-interacting RNAs (piRNAs) represent a class of small non-coding RNAs (sRNAs) in animals that associate with Piwi clade Argonaute proteins (PIWIs). Together they form the core of the piRNA-induced silencing complex (piRISC), which selects target RNAs by sequence complementarity for regulation on the transcriptional and post-transcriptional level (Gebert and Rosenkranz 2015). In mammals, piRNAs have a size range of ~24-32 nucleotides (nt) and are largely germline-specific. The biogenesis of piRNAs ensues within two pathways, resulting in primary and secondary piRNAs (Czech and Hannon 2016). Primary piRNAs are generated by the mitochondrial endoribonuclease PLD6 (Ipsaro et al. 2012, Nishimasu et al. 2012) from larger single stranded RNA molecules, such as piRNA precursors that are transcribed from a few large loci in the genome, named piRNA clusters (Aravin et al. 2006, Girard et al. 2006, Grivna et al. 2006, Watanabe et al. 2006). PIWI proteins loaded with primary piRNAs that are heavily selected for 5' uracil (1U) (Cora et al. 2014) can then enter the so-

called ping-pong cycle that produces secondary piRNAs from reverse complementary transcripts, which are cleaved with a 10 nt offset from the 5' end of the guiding piRNA and bound by another PIWI protein (Brennecke et al. 2007, Gunawardane et al. 2007, Aravin et al. 2008). These secondary piRNAs exhibit a bias for adenine at position 10 (10A) that stems from an intrinsic preference for 10A of the loaded PIWI proteins (Wang et al. 2014). The ping-pong cycle results in post-transcriptional repression of the target and a self-sustaining amplification of sense and antisense piRNAs.

Mammals typically possess four PIWI paralogs, Piwi-like 1-4 (Piwil1-4) (Aravin et al. 2006, Girard et al. 2006, Carmell et al. 2007, Roovers et al. 2015), each of which binds piRNAs of a distinct size range, being a result of 3' end trimming of slightly larger pre-piRNAs by the exoribonuclease PNLDC1 after loading onto PIWI proteins (Zhang et al. 2017). In mouse testis, the paralogs Piwil1, Piwil2 and Piwil4 bind piRNAs with sizes around 30 nt, 26 nt and 28 nt, respectively (Aravin et al. 2006, Girard et al. 2006, Aravin et al. 2008). Two distinct sequential populations of piRNAs exist in mammalian testis, namely pre-pachytene piRNAs, which interact with Piwil2 and Piwil4 in primordial germ cells and pachytene piRNAs that are bound by Piwil2 and Piwil1 and are present from the pachytene meiotic stage of spermatogenesis until the round spermatid stage (Aravin et al. 2006, 2007, 2008).

Pre-pachytene piRNAs, which are enriched for transposon sequences, direct the post-transcriptional and transcriptional repression of transposable elements (TEs) in early gonocytes during epigenetic reprogramming, which is accompanied by extensive ping-pong cycle amplification (Aravin et al. 2007, 2008, De Fazio et al. 2011). Pachytene piRNAs, on the other hand, are depleted of TE-derived sequences and are mostly generated in primary biogenesis from large pachytene-specific piRNA clusters (Aravin et al. 2007, Beyret et al. 2012). While being required for post-transcriptional TE repression (Reuter et al. 2011), pachytene piRNAs were also shown to play a role in gene regulation, involving ping-pong cycle processing (Gou et al. 2014, Zhang et al. 2015, Goh et al. 2015, Gebert et al. 2015).

Pseudogene-containing piRNA clusters have been suggested to be an important source of gene-targeting antisense piRNAs (Hirano et al. 2014, Watanabe et al. 2015, Gebert et al. 2015, Pantano et al. 2015). Generally, while some piRNA-producing loci are active across many species (Chirn et al. 2015), piRNA clusters typically evolve rapidly on a large scale (Assis and Kondrashov 2009). This raises the question of whether pseudogene-containing piRNA clusters are maintained throughout evolution to retain their ability to target genes, which would indicate the significance of pseudogene-dependent PIWI-mediated gene regulation. In this work we study the evolution of primate piRNA clusters and the conservation of therein contained pseudogenes and their capacity to target coding-genes across species to elucidate putative gene-regulatory roles of pseudogene-derived piRNAs.

5.3. Methods

5.3.1. Small RNA datasets and basic analysis

Testis-expressed small RNA transcriptome datasets from haplorhine primates were downloaded from NCBI's sequence read archive (SRA), including samples from *Homo sapiens* (SRR835325), *Macaca mulatta* (SRR116839), *Macaca fascicularis* (SRR1755243) and *Callithrix jacchus* (SRR1041905), while datasets from the strepsirhine primate species *Microcebus murinus* (SRR606735) and *Loris tardigradus* (SRR606744) were previously generated in-house. For comparisons within species, additional datasets for *H. sapiens* (SRR835324) and *M. mulatta* (SRR553581) were obtained from the SRA (ncbi.nlm.nih.gov/sra).

Adapter clipping, filtering of low complexity reads and removal of annotated ncRNAs was achieved with unitas (v1.4.6) (Gebert et al. 2017), using default settings. Subsequently, the cleaned sRNA reads were mapped to the corresponding genomic sequences (GRChg38, rheMac8, macFas5, calJac5, micMur3) with the tool sRNAmapper (v1.0) (Rosenkranz et al. 2015b), retaining only the best matches

(option '-a best'). Since there is no sequenced genome available for *Loris tardigradus*, the genome of the closest relative at hand, *Otolemur garnettii* (otoGar3), was used instead. Genome sequences were obtained from the UCSC genome server (hgdownload.cse.ucsc.edu/downloads.html). Basic analyses of sRNA datasets, aimed at the inspection of piRNA characteristics, such as read length distribution, positional nucleotide composition and rates of 5' overlap lengths, were performed using ngs toolbox (Rosenkranz et al. 2015b). The analysis of ping-pong partners was carried out, as the majority of the following analyses, using in-house perl (v5) scripts (Table S5).

5.3.2. Prediction of piRNA clusters

For in-silico prediction of piRNA clusters (piCs), we used proTRAC (v2.4.0) (Rosenkranz and Zischler 2012), where two different approaches were used for each species, using a strict and less strict set of options. First, piRNA clusters were predicted with a minimum cluster size of 5000 base pairs (5 kb) (option '-clsize 5000'), a p-value for minimum read density of 0.01 (option '-pdens 0.01'), a minimum fraction of normalized reads that have 1T (1U) or 10A of 0.75 (option '-1Tor10A 0.75') and rejecting loci if the top 1% of reads account for more than 90% of the normalized piRNA cluster read counts (option '-distr 1-90'). In a less stringent procedure, we changed the options to a minimum cluster size of 2.5 kb (option '-clsize 2500'), a p-value for minimum read density of 0.05 (option '-pdens 0.05') and a minimum fraction of normalized reads that have 1T (1U) or 10A of 0.5 (option '-1Tor10A 0.5'). Further settings that depart from the default include a minimal fraction of hits with 1T (U) and 10A of 0.33 (option '-1Tand10A 0.33') and a minimal fraction of hits on the main strand of 0.5 (option '-clstrand 0.5'). Generally, proTRAC input included a file containing mapped reads, generated by sRNAmapper, the corresponding genome sequence file, a repeatmasker annotation file and a GTF gene annotation file. Repeatmasker files were obtained from the UCSC genome server (hgdownload.cse.ucsc.edu/downloads.html) and GTF files were taken from Ensembl (ensembl.org/info/data/ftp/index.html). Finally, neighboring clusters with a distance of less than 10 kb were merged. For comparison of piRNA clusters between individuals of the same species, genomic locations and read densities (reads/kb) were extracted from proTRAC output generated with strict options and with less stringent settings.

5.3.3. Identification of homologous piRNA clusters

The bioinformatic procedure for the identification of homologous piRNA clusters between primate species was divided into three main subsequent steps, based on loci predicted with strict proTRAC options. First, information on flanking genes up and downstream of piRNA clusters in the genome of the query species was gathered (Figure 1A). Specifically, exons of the ten neighboring protein-coding genes on each flank of piRNA clusters were localized using GFF gene annotation data, obtained from NCBI's Genome resource (ncbi.nlm.nih.gov/genome), and extracted from the genomic sequence.

The next step constituted the search for the corresponding syntenic regions (Figure 1B). To this end, we scanned the repeatmasked subject species genomes, acquired from the UCSC genome server (hgdownload.cse.ucsc.edu/downloads.html), for sequences homologous to the flanking gene exons of the respective query species, using the blastn command line tool from the NCBI BLAST+ suite (v2.7.1+) (Camacho et al. 2009). Neighboring blast hits were grouped to contiguous gene loci, which in turn were divided into putative syntenic flanks. The most probable syntenic regions were selected according to the number of homologous genes and their sequence similarity to the query genes. Regions with less than four homologous genes were rejected.

Finally, if a syntenic region was found, we screened it for sequence homology to the respective query species piRNA cluster (Figure 1C), using the discontiguous-megablast algorithm (blastn run with

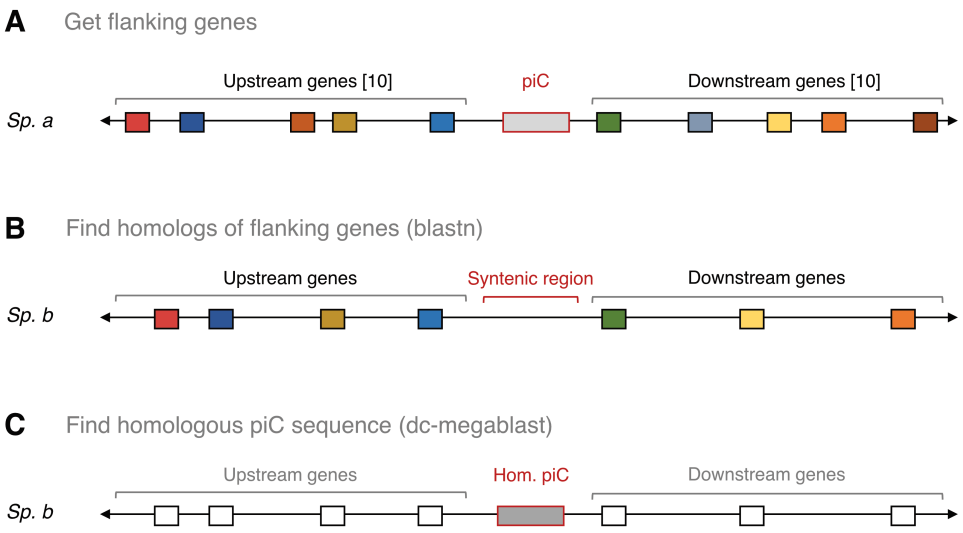


Figure 1 | Bioinformatic procedure for the identification of homologous piRNA clusters. (A) Identification of flanking genes upstream and downstream of query piRNA cluster in species a. (B) Search for homologous sequences of flanking genes in genome of species b to find syntenic region. (C) Seek homologous piRNA cluster sequence in syntenic region.

option '-task dc-megablast'), since the sequence conservation of piRNA clusters is expected to be lower compared to protein-coding genes (Assis and Kondrashov 2009). The resulting blast hits were then sorted, grouped and ranked according to alignment length, genomic region size and query coverage. Hit groups falling below thresholds for query coverage (5%), alignment length (1.5kb) or relative size to query (15%) were discarded.

In order to reconstruct the evolutionary relationships of homologous piRNA clusters among the six primate species examined in this study, we employed an iterative algorithm that combined pairs of homologous loci to chains between species. First, all homology pairs were transferred into a matrix with the respective species in columns and associated loci sets in rows. Subsequently, rows with identical or overlapping genomic coordinates were merged. This step was repeated until each locus was uniquely represented and all redundancies were resolved.

5.3.4. Analysis of homologous piRNA clusters

For the analysis of sequence conservation and presence/absence status of piRNA cluster loci between species, we extracted the relevant information from blast alignment data and compared mean identities and total shares of loci for which a homologous sequence was found for each combination of species. Additionally, mean sequence similarities of exonic sequences between species were obtained using discontiguous-megablast on CDS files from NCBI's genome resource (ncbi.nlm.nih.gov/genome), extracting identities from alignments of gene homologs. The same approach was used to get sequence similarities of genomic sequence, based on comparison between masked chromosomes homologous to human chromosome 1. To inspect which homologous piRNA cluster loci were actually expressed, we checked if an identified homologous locus was predicted as a piRNA cluster by proTRAC in a less strict mode. To determine how these properties change over evolutionary times, the corresponding data were sorted by the time that had passed since the split of the respective species, which is not always undisputed. The time distance between the two Macaque species *M. mulatta* and *M. fascicularis* was set to 1 million years (Li et al. 2009). Further, the Split of hominoidea and cercopithecoidea is estimated at 25 million years ago (mya) (Stevens et al. 2013), while catarrhine and platyrrhine primates are thought to have split 40 mya (Shumaker and Beck 2003). Finally, haplorhines and strepsirhines diverged about 65

mya (Birx 2006), while within the strepsirhines, lemuriformes and lorisiformes split about 58 mya (Masters et al. 2012).

Subsequently, differential expression analyses of homologous piRNA clusters between different species were performed using hierarchical clustering, average linkage and Pearson distance, including the generation of expression heatmaps and dendrograms, with the r-package gplots. The piRNA cluster loci were grouped into clusters that are active in each species, loci that are present in each species but not necessarily expressed and loci that are not found in each genome. Read counts (reads per million, rpm) were extracted from proTRAC output and plotted as contributions to the pool of cluster-derived piRNA reads. Lastly, TE divergence percentages for each group were extracted from Repeatmasker output from the UCSC genome server (hgdownload.cse.ucsc.edu/downloads.html) and plotted as mean TE divergence for each species. Statistical testing for TE divergence was performed using the paired Wilcoxon-Mann-Whitney test, which is included in the R (v3.4.3) and Rstudio (1.1.414) packages.

5.3.5. Prediction of pseudogenes

Since the quality of available pseudogene annotations varies substantially among species, e. g. for GFF data from NCBI (Figure S2), a custom pseudogene prediction routine was applied (Figure 2), based on the method used by Gerstein and colleagues (Zhang et al. 2006, Sisu et al. 2014). The procedure begins with the search for sequences with similarity to known protein-coding genes in the repeatmasked genome of the respective species, using the discontiguous-megablast algorithm (blastn run with option '-task dc-megablast') (Camacho et al. 2009) with CDS data, obtained from NCBI's genome resource (ncbi.nlm.nih.gov/genome), as query sequences. Based on GFF gene annotation data from NCBI, blast hits that overlap with exons of protein-coding genes were discarded (Figure 2A).

Next, the remaining blast hits were processed to construct possible pseudogene-like units (Figure 2B). To this end, overlapping hits were merged to form larger entities, which in turn were combined with adjacent hits to assemble pseudogene units if the genomic distance did not exceed a certain threshold. The allowed gap length threshold was calculated for each putative pseudogene/parent combination as the 1.5-fold of the largest parent gene intron size, while it was not allowed to fall below 30 kb.

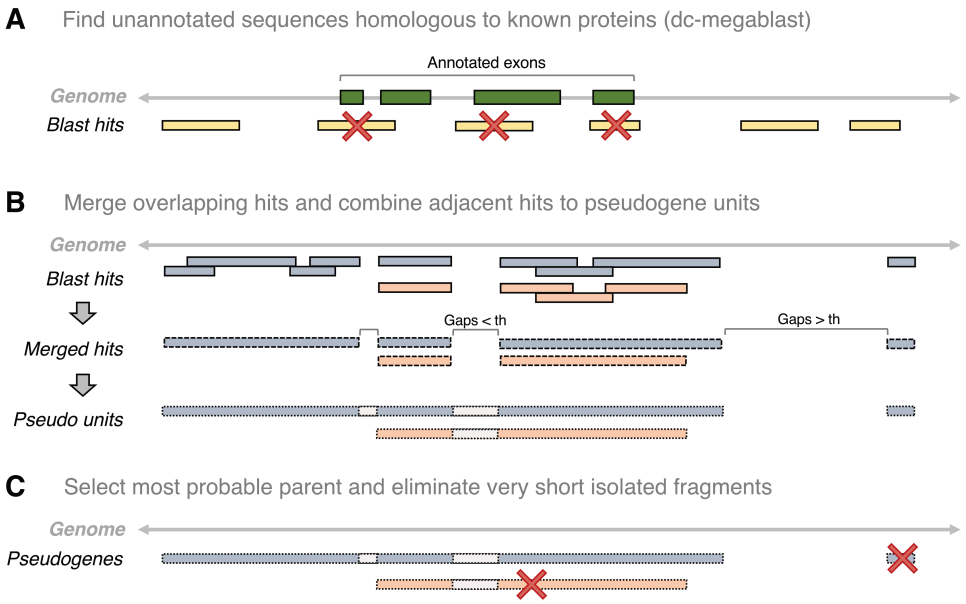


Figure 2 | Bioinformatic procedure for the prediction of pseudogenes in whole genomes. (A) Search for unannotated genomic sequences with similarity to known protein-coding genes with discontiguous megablast. (B) Merging of overlapping blast hits and grouping of merged hits to putative pseudogene units. (C) Selection of best pseudogene units.

In the third step, the most probable parent genes for the presumed pseudogene loci were selected, based on sequence identity, the best e-value of the original blast hits and the overall query coverage. Additionally, short isolated fragments (<300 bp length or <10% query coverage) were discarded. Lastly, the predicted pseudogene units were classified as processed or unprocessed pseudogenes, depending on their number of pseudo-exons compared to the number of exons of their parent genes and the overall query coverage. Specifically, if the number of predicted pseudo-exons was half the number of expected pseudo-exons (coverage fraction times number of parent exons) or less, it was categorized as a processed pseudogene.

5.3.6. Analysis of piRNA cluster pseudogenes and identification of homologs

Pseudogenes in piRNA clusters were located using custom genomic prediction, which was compared to NCBI GFF annotation (ncbi.nlm.nih.gov/genome) (Figure S2). In subsequent analyses, sequence identities of pseudogenes to parent genes, information on orientation with respect to directions of piRNA cluster transcription, as well as shares of processed and unprocessed pseudogenes were extracted from blast alignment output and our custom pseudogene annotation. Statistical testing was performed using the unpaired Wilcoxon-Mann-Whitney test, which is included in the R-package.

To determine which pseudogenes are present throughout homologous piRNA clusters across species, for each pseudogene sequence that is located in a cluster locus a similar sequence was searched for in any homologous locus that was previously identified, using discontiguous-megablast (blastn run with option '-task dc-megablast') and filtering out short total alignments (<150 bp) and hits with coverage below the threshold (<30% query coverage). Any such lineages of homologous pseudogenes that overlapped were merged.

5.3.7. Prediction of piRNA target genes

In order to identify piRNA targets among protein coding genes, clean reads with a length between 24 and 32 nt were mapped to the coding subset of known cDNA sequences, obtained from Ensembl (ensembl.org/info/data/ftp/index.html), in each species, using seqmap (Jiang and Wong 2008). While two mismatches were allowed during mapping, the output was subsequently filtered to permit two mismatches in antisense but none in sense orientation. This data was then used to get information on sense and antisense coverage on cDNA, as well as on the presence or absence of ping-pong signatures. A coverage threshold of 5 reads per kilo base per million mapped reads (RPKM) per gene was applied. A significant ping-pong signature was declared being present if the largest number of overlaps was unambiguously 10 nt long and in addition if the z-score for 10 nt long overlaps compared to the background (1-9 nt and 11-20 nt overlaps) was greater than $z=2.3264$, corresponding to a p-value of less than $p=0.01$ (Zhang et al. 2011).

To find potential gene targets of antisense piRNAs derived from pseudogenes, reads that match the opposite strands of reversed pseudogenic regions in piRNA clusters were mapped to the coding subset of known cDNA sequences with seqmap (Jiang and Wong 2008), allowing two mismatches. The target genes identified in this manner were then checked for presence of ping-pong signatures. Subsequently, ping-pong targets, as well as genes with general piRNA coverage, were compared among different species to find homologous genes, using data on gene homology extracted from Ensembl Biomart (ensembl.org/biomart) (Kinsella et al. 2011). Target genes were used as input for Go-term enrichment analysis using the gene ontology web tool (geneontology.org/page/go-enrichment-analysis) (Ashburner et al. 2000, The Gene Ontology Consortium 2017). Reference genes were extracted from testis-expression data that were accessed from the EMBL-EBI Expression Atlas database (ebi.ac.uk/gxa) (Petryszak et al. 2016).

5.3.8. Analysis of genomic environments of piRNA clusters

For the analysis of the genomic environment of piRNA clusters, we divided the respective genome into windows of 1 million base pairs (Mb) and used repeatmasker output (hgdownload.cse.ucsc.edu/downloads.html) and GFF gene annotation data (ncbi.nlm.nih.gov/genome) to get the frequency for each repeat family, as well as for pseudogenes and genes per Mb. Centromeric regions, of which location information of the respective genome was obtained from the UCSC genome browser server (hgdownload.cse.ucsc.edu/downloads.html), were excluded from the analysis. Further, piRNA clusters were grouped by their internal gene and pseudogene content, based on GFF gene annotation, resulting in populations of loci containing no coding genes and loci containing neither coding genes nor pseudogenes. Additionally, the GC content of complete genomes and of piRNA clusters was calculated using unmasked sequences, ignoring ambiguous bases. Statistical testing was performed using the unpaired Wilcoxon-Mann-Whitney test, which is included in the standard R-package.

5.3.9. Code and data availability

Perl and R scripts used for analyses in this study (Table S5), as well as other relevant files are available at GitHub (github.com/d-gebert/primate-pic-evo).

5.4. Results and Discussion

5.4.1. Basic analyses of sRNA datasets

Prior to prediction of piRNA clusters, we performed basic analyses on the six sRNA datasets, upon which this study is based (Figure S1). Unifying characteristics of piRNAs, such as a size range between 24 and 32 nt (Figure S1A), 1U/10A biases (Figure S1B) and ping-pong signatures (Figure S1C) were observed in each case. The shares of reads that have ping-pong partners is low, being typical for pachytene piRNAs (Reuter et al. 2011), ranging from 5% to 12% of 24-32 nt non-identical reads (Figure S1C). Further, one can infer from local peaks in the read length distributions that both Piwil2 and Piwil1 are present and binding piRNAs of ~26/27 and ~29/30 nt, respectively (Figure S1A), as known from mice (Aravin et al. 2008). Analyzing length combinations of ping-pong reads shows that the majority of ping-pong partner reads consists of pairs with lengths of ~26 and ~30 nt or ~30 nt both, suggesting that ping-pong occurs primarily between Piwil1 and Piwil2 or among Piwil1 proteins, but much less between Piwil2 proteins (Figure S1D). Though we note that the sRNAs were not co-immunoprecipitated from PIWI proteins, hence strictly representing piRNA-like RNAs, we will refer to these sequences as piRNAs, due to the strong evidence for piRNA traits.

5.4.2. Comparability of predicted piRNA clusters among individuals and species

Using proTRAC (Rosenkranz and Zischler 2012) with a strict set of options, we identified a varying number of piRNA clusters, ranging from 171 in *L. tardigradus* to 608 in *M. fascicularis* (Figure 3A). The majority of reads falls into clusters in every dataset except for *L. tardigradus* (Figure S1A), which is likely due to the usage of the *O. garnettii* genome, since a matching reference genome does not yet exist. In each species the share of reads produced by piRNA clusters follows a Pareto distribution, meaning that a small number of clusters is responsible for the majority of piRNA reads, while the bulk of clusters produces relatively few reads (Figure 3B). Since proTRAC most critically relies on read density and locus size, the applied thresholds inevitably lead to sharp cutoffs in the long tails of the distributions, which might have the effect that the comparability between different samples can be problematic.

To test whether piRNA clusters predicted by proTRAC are comparable between individuals of the same species and ultimately between different species, we checked the amount of overlap of identified loci based on different sRNA samples. Indeed if piRNA clusters are predicted in a stringent manner

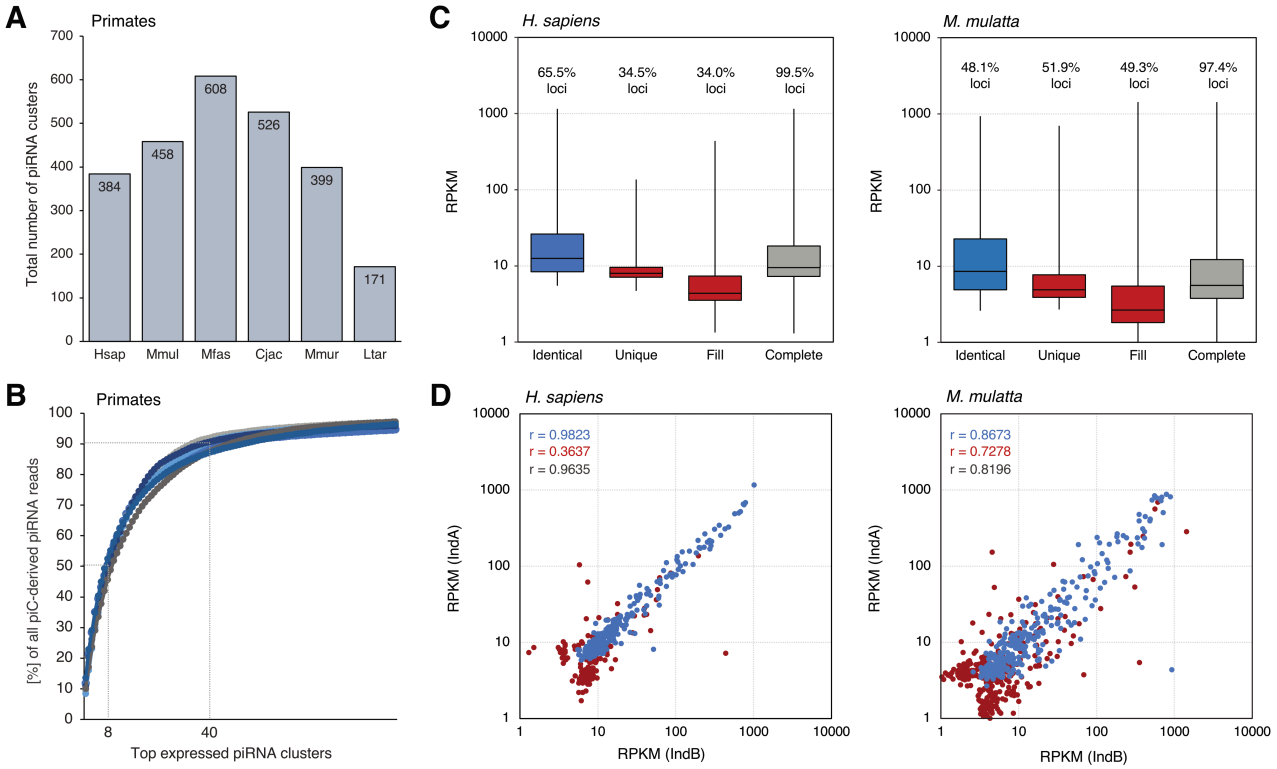


Figure 3 | Comparison of predicted piRNA clusters between species and individuals. (A) Total number of piRNA cluster loci predicted for each species with strict threshold options. (B) Cumulative distribution of read shares produced by top 100 expressed piRNA clusters for each species. (C) Shares of piRNA cluster loci predicted from samples of two individuals of the same species (*H. sapiens* and *M. mulatta*) and their read densities (RPKM; reads per kilo base per million mapped reads). Identical: loci that are found in both samples with strict prediction. Unique: loci that are found only in one the other sample with strict prediction. Fill: loci predicted with less strict threshold options that are identical to ‚unique' loci in the other sample. Complete: Combination of all piRNA cluster loci expressed in each individual, including complementing loci predicted with less strict thresholds. (D) Correlation of read densities (RPKM) of piRNA clusters from two individuals (IndA/B) of the same species (*H. sapiens* and *M. mulatta*).

from samples of two individuals from the same species, e. g. *H. sapiens*, only 65.5% of loci are identical, while 34.5% seem to be unique to the respective individual (Figure 3C). However, using piRNA clusters predicted with less strict options to find loci that are identical to those more strictly predicted that appear to be unique, overall 99.5% of piRNA clusters were found to be expressed in both individuals. Similarly, when comparing two specimen of *M. mulatta*, only 48.1% of loci overlap, which can be increased to 97.4% in the same manner. As expected, the seemingly unique loci are predominantly shifted towards the lower end of the read density (RPKM; reads per kilo base per million mapped reads) spectrum relative to those that have identical equivalents in each individual, while loci that were predicted with less strict options to fill the missing counterparts fall mostly below that range (Figure 3C). Apart from read density, thresholds for cluster size and minimum fraction of reads with 1T (1U) or 10A have a similar, though less marked effect for the prediction of piRNA clusters with the respective properties that come close to these thresholds. Overall, the expression rate of piRNA clusters, represented by read density, highly correlates between two individuals of the same species, supported by Pearson correlation coefficients of 0.96 for *H. sapiens* and 0.82 for *M. mulatta* (Figure 3D). Together these results show that piRNA cluster expression is mostly consistent and comparable between individuals of the same species if loci predicted with less stringent thresholds are added to seemingly sample-specific loci predicted with more strict options to moderate the effects of threshold cutoffs. As a consequence, this consistency that is gained through our approach within species makes the comparison between different species more reliable.

5.4.3. Presence and activity of homologous piRNA clusters across primates

To determine the proportion of piRNA clusters that are shared among the primate species examined in this study, we used an approach based on synteny and sequence similarity. Syntenic regions could be found for the vast majority of piRNA clusters, ranging from 97.7% to 100%, depending on the combination of species (Figure 4A). In contrast, the rate of homologous piRNA cluster loci present between species drops substantially the more distantly two species are related, ranging from 93.2% for *M. mulatta* and *M. fascicularis* to 50.6% for *C. jacchus* and *L. tardigradus*/*O. garnettii* (Figure 4B). Further, the proportion of loci that actively produce piRNAs drops even more distinctly (Figure 4C). While nearly all homologous piRNA cluster loci are expressed between *M. mulatta* and *M. fascicularis*, merely 21.8% of clusters in *C. jacchus* are also active in *L. tardigradus*, which represent only 43% of identified homologous loci.

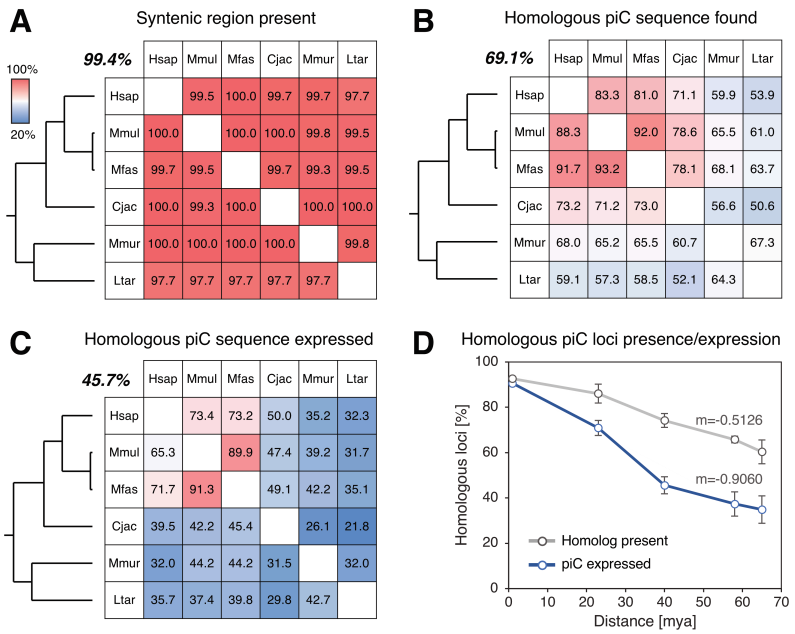


Figure 4 | Presence and expression activity of homologous piRNA cluster loci among primate species. (A) Rates of loci for which syntenic regions could be found. (B) Rates of loci for which homologous sequences could be found. (C) Rates of homologous loci which are expressed. (D) Rates of presence of homologous loci and stably expressed piRNA clusters over evolutionary time distances. Trees show phylogenetic relationships. Bold numbers indicate mean percentages.

Plotted by the evolutionary time distance that separate the here analyzed species (Figure 4D), the difference between the rate by which the share of homologous piRNA cluster loci drops with increased evolutionary distance is lower compared to the share of expressed loci, which is indicated by the slopes (m) of the linear trend estimation of -0.51 and -0.91, respectively. However, the relations are not linear but approximately follow an inverted logistic S-curve. Altogether, for 707 loci homologs were found in every genome, while only 156 clusters are actually expressed across all species. Noteworthily, a previous study described the expression of a core set of 77 piRNA producing loci that are found throughout eutherians (Chirn et al. 2015). 45 of these 77 loci overlap with our 156 homologous piRNA clusters.

Our findings suggest that primate piRNA clusters tend to be located at genomic regions that are lineage-specific, being acquired more or less recently on the evolutionary time scale. Indeed, it was previously shown in a study of mouse and rat piRNA clusters that their genomic contexts are very unstable, since many rodent clusters lie within regions that underwent major rearrangements, including

insertions, deletions and inversions (Assis and Kondrashov 2009). The large discrepancy between presence of homologous loci and their actual activity as piRNA clusters indicates that many loci either lost their piRNA producing activity after their emergence or gained it later after evolutionary partition.
The sequence evolution of piRNA clusters for which homologous loci could be found (Figure 5A) is very similar to the general sequence divergence over time that is observed in the genome as a whole (Figure 5B) and is in stark contrast to the relatively slow change of coding-gene sequences (Figure 5C). Comparable to the whole genome, piRNA cluster loci show a near linear decrease in sequence identity over evolutionary times at a roughly doubled rate compared to coding-genes (Figure 5D), indicating lack of selection pressure on piRNA cluster sequences. This is in line with previous findings which suggested that the small-scale evolution of clusters proceeds at rates typical for mammalian genomes (Assis and Kondrashov 2009). Lastly, we wondered whether those loci that are consistently expressed in every species might show elevated rates of similarity, but no consequent substantial shift in any direction could be observed (mean change: -0.33%; standard deviation: 0.3%).

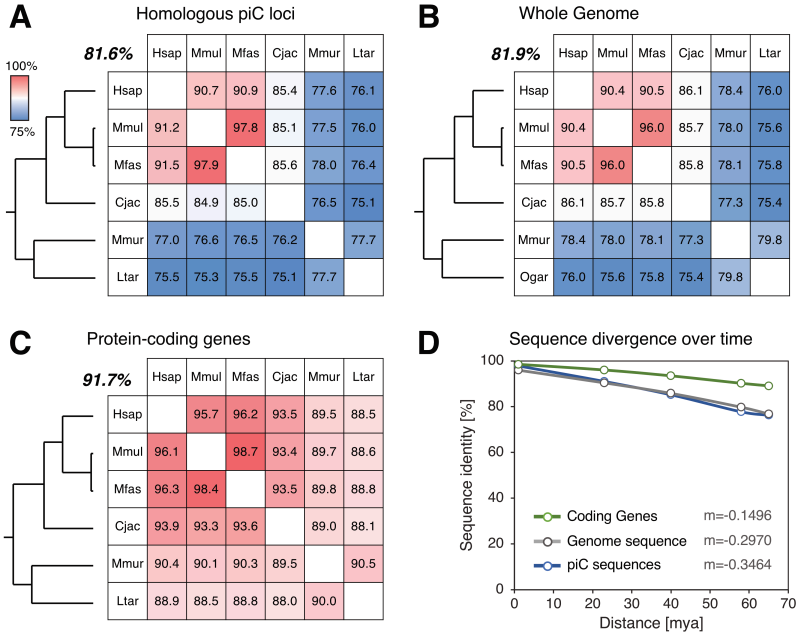


Figure 5 | Sequence evolution of homologous piRNA cluster loci among primate species. (A) Sequence identity of homologous piRNA cluster loci. (B) Sequence identity of genomic sequence. (C) Sequence identity of exon sequences of protein-coding genes. (D) Sequence identity over evolutionary time distances. Trees show phylogenetic relationships. Bold numbers indicate mean percentages.

5.4.4. Expression of homologous piRNA clusters

Next, we analyzed the differential expression of homologous piRNA clusters across species. Loci that are expressed in all species (Figure 6A) were examined separately from those that are present in all six genomes, but do not necessarily produce piRNAs (Figure 6B). In both cases the expression profiles are very specific for each species, supported by hierarchically clustered dendrograms (Figure 6A,B; left), which recapitulate the phylogenetic relations of the six primates in a remarkably accurate way (Figure 6A,B; top).
We then checked the contribution of piRNA clusters with different presence and activity states to the global pool of piRNAs per species. We distinguished clusters that are present and expressed in each species (~156/sp.), loci that are found in each genome but not expressed in every species (~277/sp.)

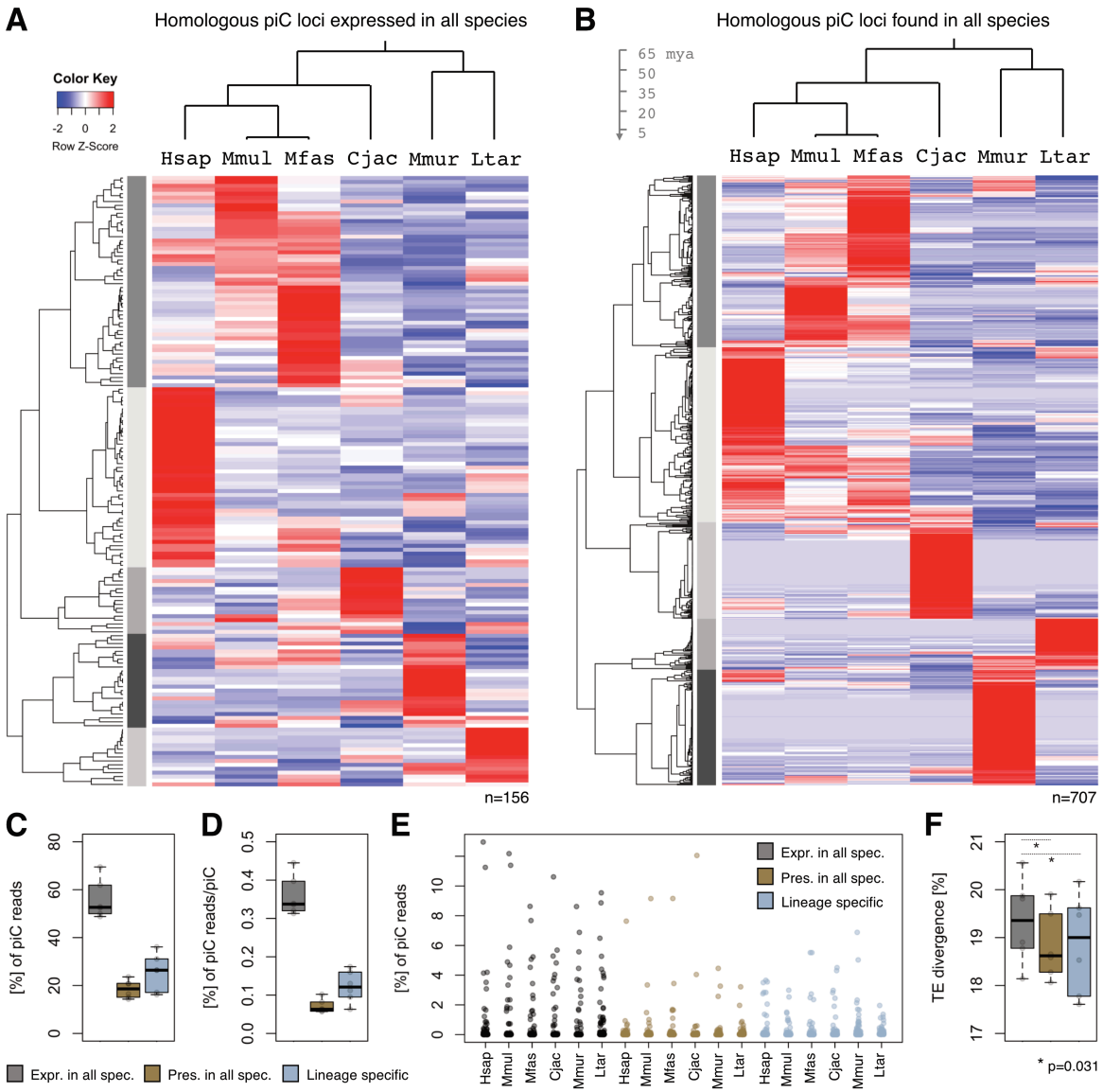


Figure 6 | Expression of homologous piRNA clusters across primate species. (A) Differential expression of homologous loci that are consistently expressed in each species. (B) Differential expression of homologous loci that are present in each species. Non-expressed loci have an expression value of 0. (C) Combined shares of cluster-derived piRNA reads per species from clusters that are present and expressed in each species, clusters that have homologs in each genome but not expressed in each species and clusters that that do not have homologs in each genome. (D) Shares of cluster-derived piRNA reads per expressed cluster. (E) Shares of reads contributed by each cluster to the total pool of cluster-derived piRNA reads per species. (F) Mean sequence divergences from consensus of transposons in piRNA clusters. Same order and key as C,D,E.

and those that that do not have homologs in each genome (~222/sp.). The piRNA clusters that are present and active in all species contribute the majority of reads, ranging from 50 to 70% of all cluster-derived reads, despite constituting the smallest group (Figure 6C). The second group, loci present across species but not ubiquitously expressed, provides 14-24% of reads, while lineage-specific loci contribute slightly larger shares of 16-36%. These relations stay consistent if the mean shares per expressed locus in each group are examined (Figure 6D). On average, each across-species active piRNA cluster contributes 0.36% of reads, in contrast to 0.07% and 0.12% in the second and third group, respectively. However, actual expression rates differ immensely among piRNA clusters (Figure 6E). The observed differences of total read shares are mainly due to a varying number of large contributors, since most clusters yield only a small fraction of piRNA reads.

Finally, we wondered whether the three discussed groups of clusters, having different evolutionary histories, might show differences in transposon age. Indeed, the mean TE divergence from consensus in ubiquitously present and expressed piRNA clusters is significantly higher than in the remaining groups (Figure 6F), suggesting younger transposon age in the latter.

Taken together, these results show that the relatively small faction of homologous and active piRNA clusters already exhibit distinct expression profiles among primates. Even the closely related macaque species show a beginning deviation in this respect. Nevertheless, piRNA clusters that are consistently expressed throughout evolution are the major source of piRNA reads across primates, while lineage-specific loci are important contributors to the total piRNA pool. Moreover, the fact that loci that are present or expressed in a lineage-specific manner harbor on average younger TEs suggests that these groups of clusters represent lineage-specific adaptations to newer transposons.

5.4.5. Characterization of pseudogenes in piRNA clusters

In order to get a deeper understanding about possible shared characteristics of pseudogenes that lie in piRNA clusters, we set out to determine their basic properties in each species. However, since the quality of available pseudogene annotations varies considerably among species (Figure S2), we applied a custom routine to predict pseudogenes for whole genomes. To verify the validity of our annotation method, we compared our results for piRNA cluster regions to GFF gene annotation data from NCBI (Figure S2). We could predict on average 2.4 times the number of annotated pseudogenes, including 78% of annotated sequences, showing the effectiveness of our approach.

Since reverse orientation of pseudogene sequences with regards to piRNA cluster directionality is a prerequisite for the generation of gene-targeting antisense piRNAs, we checked the shares for each

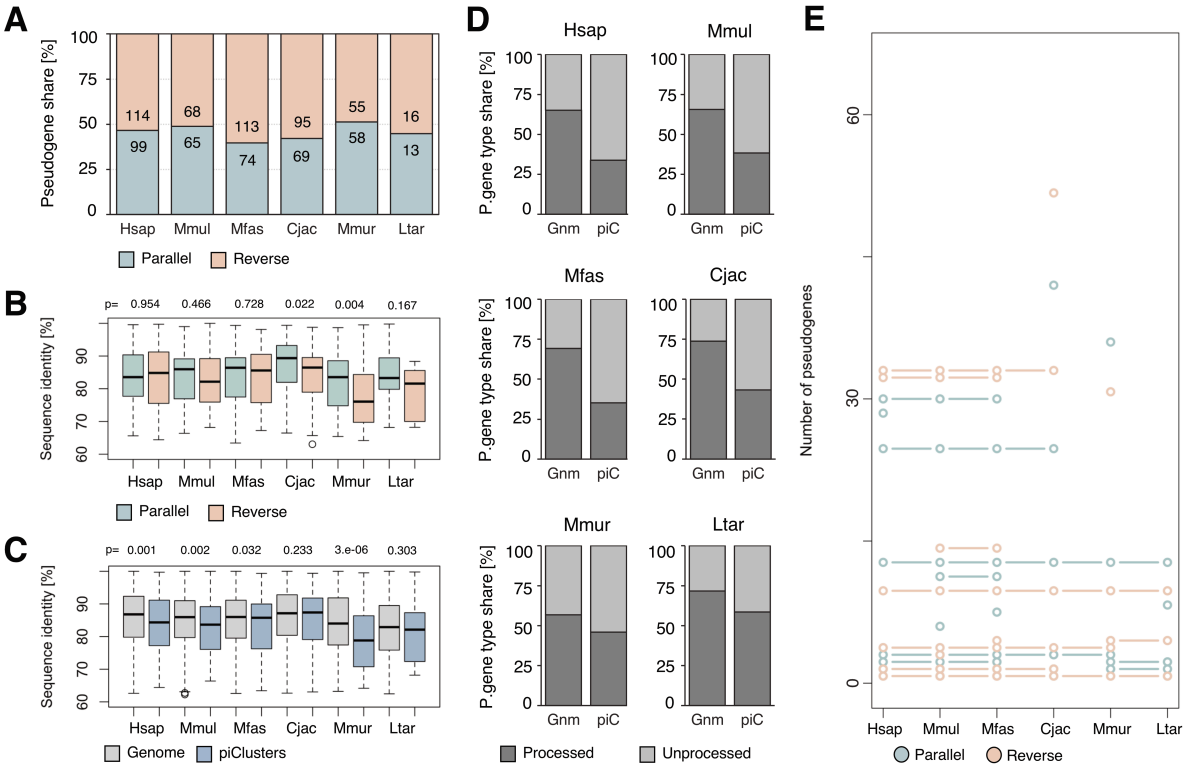


Figure 7 | Characterization of pseudogenes in piRNA clusters. (A) Number of pseudogenes sorted by parallel or reverse orientation relative to piRNA cluster directionality. (B) Sequence identities of pseudogenes in parallel compared to reverse orientation. (C) Sequence identities of pseudogenes in piRNA clusters compared to whole genome. (D) Shares of processed and unprocessed pseudogenes in whole genome (Gnm) and all piRNA clusters (piC). (E) Pseudogene homologs in homologous piRNA cluster loci in parallel and reverse orientation shared among species.

direction. We noticed a slight bias for reverse orientation of pseudogenes, which however is not consistent across species (Figure 7A). We next checked whether reverse pseudogenes are more similar to their parent genes, since a high degree of sequence identity is required for piRNA target recognition (Reuter et al 2011, Huang et al. 2013), which would be expected if pseudogene-dependent gene targeting by piRNAs is beneficial. However, no elevated sequence similarity of reverse pseudogenes compared to those in parallel orientation could be observed (Figure 7B). The same analysis, comparing all pseudogenes in piRNA clusters and in the whole genome, shows no clear consistent pattern, though in three species a statistically relevant tendency towards lower sequence similarity of pseudogenes in clusters to their parent genes can be observed (Figure 7C). It is conceivable that an unwanted interference with normal gene regulation by pseudogene-derived piRNAs might result in increased sequence evolution of the corresponding pseudogenes as a means to escape piRNA targeting.
It was suggested that piRNA clusters may gain the ability to target coding genes through the integration of gene transcripts by retrotransposition, resulting in the formation of processed pseudogenes (Hirano et al. 2014, Gebert et al. 2015). Our analysis of pseudogene types shows that while processed pseudogenes vastly outnumber unprocessed copies in primate genomes, which is in line with previous studies (Sisu et al. 2014), this relation is consistently shifted towards unprocessed pseudogenes in piRNA clusters (Figure 7D). This indicates that retrotransposition is likely not a main contributor for the incorporation of pseudogenes into piRNA clusters. Since it has been shown in rodents that many piRNA clusters originate through duplication by ectopic recombination (Assis and Kondrashov 2009), it could be speculated that genes which accidentally overlap with clusters might get duplicated with the piRNA producing locus and then undergo pseudogenization.
Another prediction, based on the assumption that pseudogene-dependent gene targeting by piRNAs provides an evolutionary benefit, is a higher retention rate of pseudogenes in piRNA clusters in reverse orientation than in parallel. While this is tendentially the case within haplorhines and catarrhines, the opposite is true for pseudogenes being present in homologous piRNA clusters across all six primate species (Figure 7E).
Overall these findings show that in general pseudogenes in piRNA clusters do not exhibit the traits that would be predicted if pseudogene-derived piRNAs were widely used for regulation of coding genes. The bias towards unprocessed pseudogenes in clusters as compared to the whole genome situation, which is the only consistent observation across species, indicates that rather than retrotransposition into existing clusters, duplicated genes become part of piRNA clusters during their emergence.

5.4.6. Gene targeting by pseudogene-derived piRNAs

Following the basic characterization of pseudogenes, we examined the gene-targeting capacities of piRNA cluster-overlapping pseudogenes. To this end, the portion of protein-coding genes that are potentially targeted by pseudogene-dependent piRNAs, where we generally allowed two mismatches, was set in relation to all target genes that show a significant ping-pong signature, again permitting up to two mismatches for antisense reads (Figure 8A). Overall, on average the minority of genes targeted by pseudogene-derived antisense piRNAs showed a ping-pong signature, since this was observed for merely 31% of cases. Further, only small fractions of on average 7% of all ping-pong genes in each species were targeted by pseudogene-derived antisense piRNAs.
Thus, since the targeting of coding genes by piRNAs derived from pseudogenes lying in piRNA clusters can not explain the vast majority of cases of ping-pong coverage on gene transcripts, other mechanisms that initiate processing by the secondary piRNA pathway on protein-coding genes likely play a far greater role. Nevertheless, the fact that still a part of the genes that are potentially targeted by

pseudogene-dependent piRNAs indeed display a ping-pong signature shows that some of these piRNAs likely have the expected capability to lead gene transcripts into the ping-pong cycle.

Next, we quantified the amounts of reversed pseudogenes in piRNA clusters that produce genic antisense piRNAs in general and those that in addition target ping-pong genes (Figure 8B). We found that on average 60% of pseudogenic sequences located in clusters give rise to piRNAs that potentially target coding genes. However, in only 38% of cases on average pseudogene-derived piRNAs aim at genes with ping-pong signatures. This indicates that the majority of pseudogenes in clusters is likely ineffective with regards to the triggering of gene transcript processing through the secondary piRNA pathway.

Examining the evolutionary relationships of ping-pong genes in general among primates, we found that for the vast majority of genes ping-pong targeting is lineage specific, while few homologs are targeted in multiple species (Figure 8C). Merely two homologous genes exhibit ping-pong coverage in all four representatives of the study's major primate groups, namely hominoidea (*H. sapiens*), cercopithecoidea (*M. mulatta*), haplorhines (*C. jacchus*) and strepsirhines (*M. murinus*). Restricting this analysis to ping-pong genes that are targeted by pseudogene-derived antisense piRNAs yields markedly less overlap between target gene homologs. Not a single homologous target is shared among four species and only one ping-pong gene is present in three species, namely human, macaque and marmoset (Figure 8D). Even when expanding the circle of potential homologous targets to genes showing general piRNA coverage above 5 RPKM, regardless of a presence of ping-pong signatures, the amount of orthologous targets remains very limited (Figure 8E).

Together, these results suggest that the PIWI/piRNA pathway triggered by pseudogene-derived antisense piRNAs, is either extremely flexible or otherwise of lesser significance for the regulation of genes. Importantly, it was shown in mice that the knockdown of a specific piRNA cluster containing a pseudogene did not lead to a detectable phenotypic effect, although the mRNA expression level of the corresponding parent gene did in fact change (Watanabe et al. 2015). Thus, it appears likely that, while the presence of pseudogenes in piRNA clusters in reverse orientation has the potential to affect gene targeting, the consequences on the regulation of these genes is not as pronounced as to have an actual physiological effect and hence to be maintained over evolutionary times. This, however, raises the question of why mammalian piRNA clusters are enriched for pseudogene sequences in the first place.

Considering gene-targeting by piRNAs on a global scale, we found that while the total amount of homologous ping-pong genes is rather low (Figure 8C), the number of gene homologs with piRNA coverage above 5 RPKM throughout species in general, with overall 1428, is considerably higher (Figure 8F). Gene Ontology analysis with this gene set indicates enrichment in a variety of functions, localizations and processes, including spermatogenesis, translation regulation, mRNA processing and oxidative phosphorylation (Table S1-3). Generally, on average a majority of 75.8% of genic reads derive from sense strands in each species. Since longer 3'-UTRs can harbor more TE sequences as potential piRNA target sites, we tested whether there is a relationship between piRNA read coverage and 3'-UTR length in human. Although we found no correlation within target genes regarding RPKM ($r=0.0005$), target genes in general have on average longer 3'-UTRs than non-targets (means: 2027 and 1478 bp; medians: 1346 and 858 bp). Wondering what differentiates ping-pong genes from other genes with piRNA coverage, we checked whether there is a relationship between read coverage and the probability for showing a ping-pong signature. Indeed, we found a strong correlation for increased shares of ping-pong targeting among genes within greater RPKM ranges in each species (Figure 8G). This suggests that the ping-pong cycle might play a role in regulating some genes that are consequently more heavily processed. It was shown earlier that gene regulation by pachytene piRNAs in mammals involves the ping-pong cycle at least to some extend (Zhang et al. 2015, Goh et al. 2015, Gebert et al. 2015).

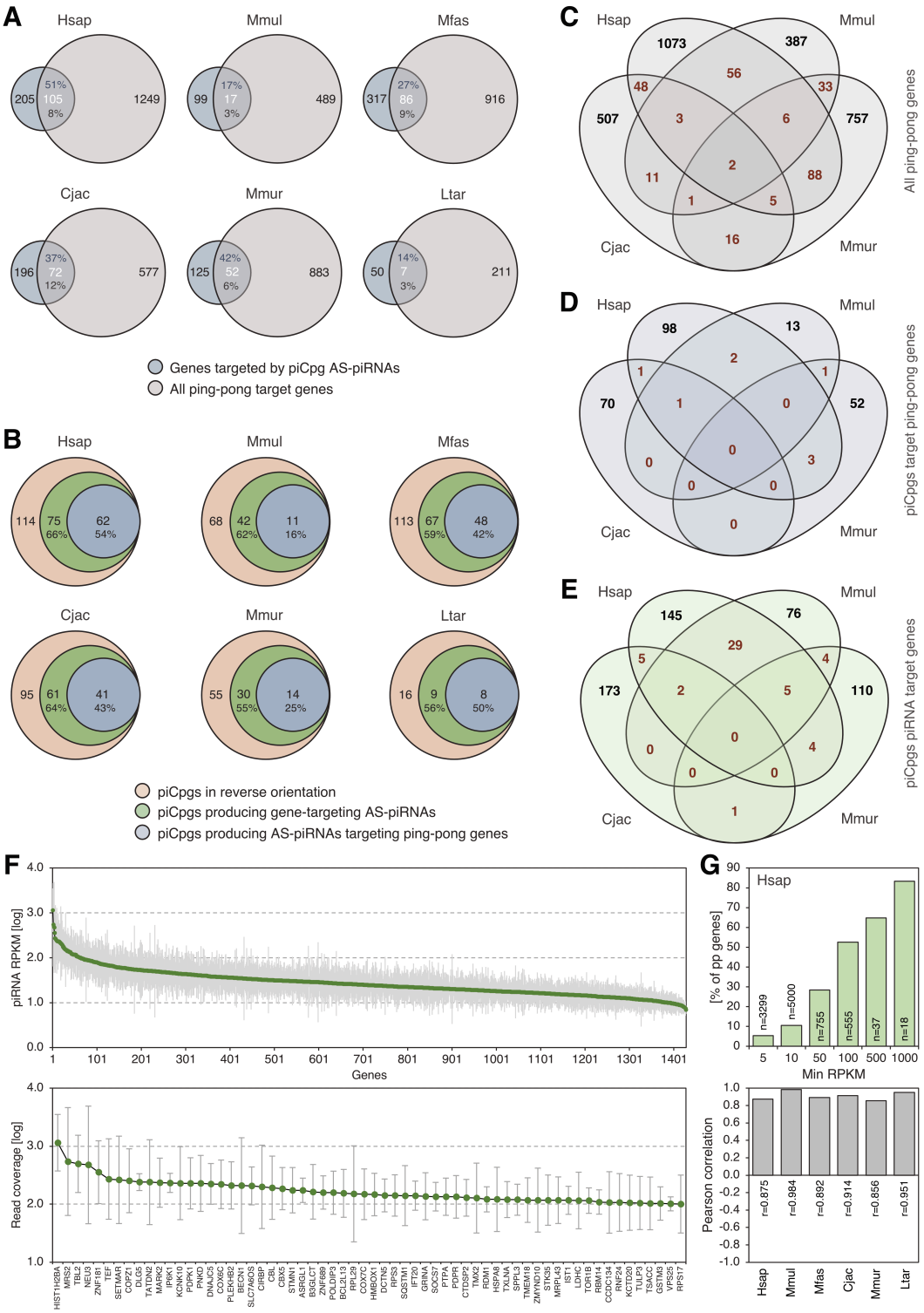


Figure 8 | Targeting of protein-coding genes by pseudogene-dependent piRNAs. (A) Shares of genes targeted by pseudogene-derived antisense (AS) piRNAs that exhibit ping-pong signatures. (B) Amounts of reversed pseudogenes in piRNA clusters (piCpgs) that produce genic antisense piRNAs and those producing genic antisense piRNAs targeting ping-pong genes. (C) Homology of ping-pong target genes among human, macaque, marmoset and mouse lemur. (D) Homology of ping-pong target genes targeted by pseudogene-derived antisense piRNAs among human, macaque, marmoset and mouse lemur. (E) Homologous genes with piRNA coverage, targeted by pseudogene-derived antisense piRNAs, among human, macaque, marmoset and mouse lemur. (F) Homologous genes with piRNA coverage of at least 5 RPKM (reads per kilo base per million reads) in all six primate species (log RPKM means and standard deviations). Bottom: Genes with mean >2 log RPKM. (G) Top: Relationship of RPKM range (e.g. 5: RPKM ≥5, <10) and share of ping-pong genes in human. Bottom: Pearson correlations of RPKM range and share of ping-pong genes (s. top) for each species.

Taken together, the considerably large set of genes that exhibit piRNA coverage in all analyzed species indicates a conserved mechanism for PIWI-mediated gene regulation which however is independent of pseudogene-derived piRNAs. It was shown that 3'-UTRs exhibit the greatest sense piRNA read density on coding genes in diverse metazoan lineages (Robine et al. 2009, Ha et al. 2014) and later it was demonstrated that TE sequences that reside in 3'-UTRs can be targeted by piRNAs, which presumably leads to mRNA decay (Watanabe et al. 2015). Another study showed that the piRNA production from some genes, partly overlapping with our set of homologous genes (29 out of 57 genes; Table S4), is conserved in many eutherians (Chirn et al. 2015). One of these genes, namely CBL, was recently demonstrated, among others, to be repressed by Aub-bound piRNAs in the germline of *Drosophila* through translational repression by binding at 5'- and 3'-UTRs, especially at TE insertion sites (Barckmann et al. 2015, Rojas-Ríos et al. 2017). Thus, some genes are apparently targeted in a highly conserved manner. Moreover, it was shown in mice that pachytene piRNAs induce broad mRNA elimination in mouse elongating spermatids by recruiting the deadenylase CAF1 upon recognition of target sites, which are mainly located in 3'-UTRs (Gou et al. 2014). Hence more than pseudogene-derived sequences, TE-associated piRNAs are likely the major regulators for PIWI/piRNA processing of protein-coding genes.

5.4.7. The genomic environments of piRNA clusters

As the evidence for a significant role of pseudogene-dependent piRNAs in gene regulation seems not convincing, we looked for potential alternative explanations for the enrichment of pseudogenes in piRNA clusters. Therefore, we turned our attention to the genomic environment of piRNA clusters. We scanned the primate genomes with a resolution of 1 Mb to obtain information on gene and pseudogene density, shares of different TE families and total sequence divergence of TEs, initially focusing on human (Figure 9). First, we noticed that piRNA clusters often seem to be located in gene rich regions, as seen for instance on human chromosome 6 (Figure 9A). Within a particular gene dense region, it contains, among others, one of the largest and most strongly expressed piRNA clusters across all six analyzed primate species (Figure 9A, arrow), and being also present and active in tree shrew and mouse (Rosenkranz et al. 2015a, Goh et al. 2015). Analyzing the complete human genome, but ignoring centromeric regions, we found that piRNA clusters indeed show a significant tendency to be located in genomic regions with elevated gene density, compared to the whole genome (Figure 9B). This holds also true if solely loci containing neither genes nor pseudogenes, hence being completely intergenic, are considered, though the contexts of these clusters are lower in gene density. The latter fact might be expected, since the probability for containing genes increases with higher gene abundance. Moreover, there is no statistically relevant difference between intergenic loci and pseudogene-containing piRNA clusters (Figure 9B).

Several factors correlate with gene density. Considering transposons in general, there is a significant negative correlation of -0.486 between TE divergence and gene density, suggesting that younger transposons are enriched in gene-rich regions (Figure 9C). Correspondingly, both, the primate-specific Alu elements (Kriegs et al. 2007), as well as the hominid-specific SVA family elements (Wang et al. 2005) tend to be more abundant in gene-rich regions of the human genome, supported by correlation coefficients of 0.645 and 0.331, respectively. On the other hand, the share of L1 elements tends to be increased in gene-poorer segments, based on a negative correlation of -0.348. This pattern, particularly of Alu and L1 transposons with respect to gene-density was already noticed in the first analysis of the human genome sequence (Lander et al. 2001). Moreover, unsurprisingly, gene density is also correlated with pseudogene abundance per Mb, indicated by a correlation coefficient of 0.347.

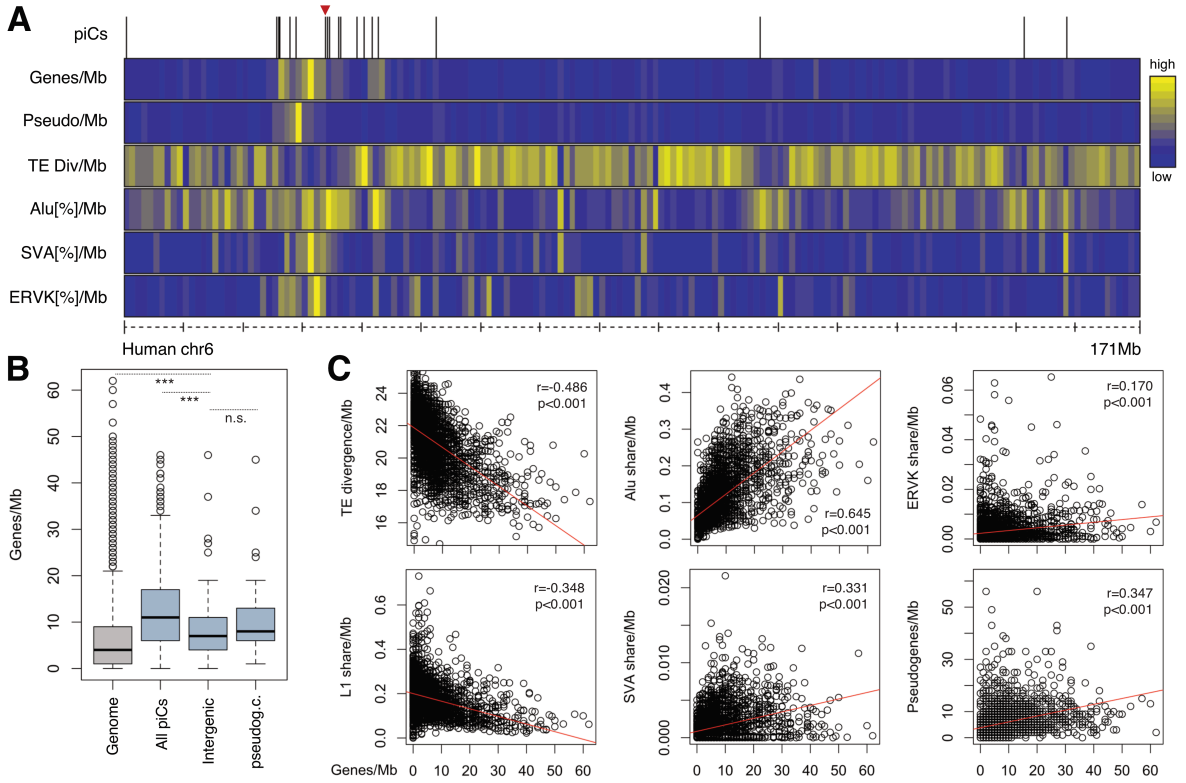


Figure 9 | Genomic environments of piRNA clusters in human. (A) Heatmap showing human chromosome 6 in 1 Mb slices. piCs: piRNA cluster locations; Genes/Mb: Gene density; Pseudo/Mb: Pseudogene density; TE div/Mb: Mean TE divergence per Mb; TE[%]/Mb: Total shares of Alu, SVA and ERVK elements per Mb. The large and highly expressed piRNA cluster at chr6:33,863,000-33,927,000 is marked by an arrow. (B) Gene densities of 1 Mb genomic slices that contain piRNA clusters compared to the whole genome in human. piCs: All human piRNA clusters; Intergenic: piRNA clusters that do not contain coding genes or pseudogenes. pseudog.c.: piRNA clusters that contain pseudogene sequence; p-values: 6.286e-05 (***), 6.798e-05 (***), 0.1349 (n.s.). (C) Correlations of TE divergence, shares of TE families (Alu, L1, SVA) and pseudogene abundance with gene density (genes/Mb) in the human genome.

Expanding the preceding analysis to non-human primates, the results stay consistent. In all six primates, intergenic piRNA clusters show a significant tendency to be located in regions with higher gene density relative to the average of the whole genome (Figure 10A). Further we noticed that regions in which (intergenic) piRNA clusters are located show elevated percentages of guanine and cytosine (GC) bases (Figure 10B). Also, the GC content of (intergenic) piRNA clusters themselves is on average higher than the genome-wide rate across species (Figure 10C). Gene density is known to be correlated with open chromatin structure (Gilbert et al. 2004) and GC rich regions tend to indicate a more active chromatin conformation (Dekker 2007). Additionally, the correlation between genomic gene and pseudogene densities, was confirmed for all species (Figure 10D).

Next we analyzed whether the respective positive and negative correlations of Alu and L1 element abundance with gene density leads to a bias of cluster localization with regards to shares of Alu and L1 transposons. Indeed we found that piRNA clusters show a significant tendency for regions with higher share of Alu elements, relative to the whole genome, while the opposite is true for L1 transposons, though less distinctly (Figure 10E). Correspondingly, piRNA clusters are depleted of L1 and enriched for Alu elements across primate species (Figure 10F). Similar ratios were observed in other mammals, such as pig, where however tRNA-derived SINEs, instead of 7SL-derived SINEs (Alus), are enriched in clusters (Gebert et al. 2015). Lastly, piRNA clusters show a significant bias for regions with lower average TE divergence, relative to the whole genome (Figure 10G), which is an indication of younger transposon age and hence more recent transposition.

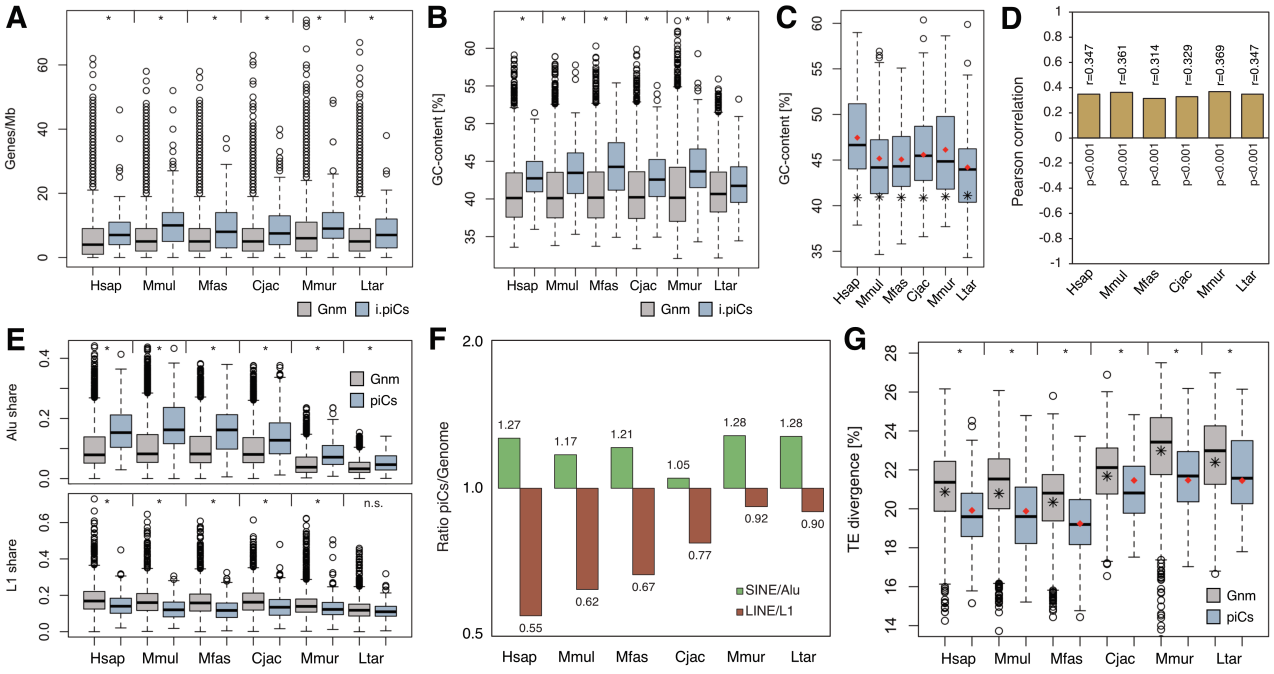


Figure 10 | Genomic environments of piRNA clusters in primates. *: p<0.01; n.s.: p>0.05. (A) Gene densities of 1 Mb genomic slices that contain intergenic piRNA clusters compared to the whole genome. i.piC: intergenic piRNA clusters; Gnm: whole genome. *p-values: 6.3e-05, 1.2e-09, 3.5e-06, 2.2e-07, 6.6e-09, 8.6e-3. (B) GC contents of 1 Mb genomic slices that contain intergenic piRNA clusters compared to the whole genome. *p-values: 1.4e-10, 1.7e-13, <2.2e-16, 1.1e-12, <2.2e-16, 2.3e-3. (C) GC contents of intergenic piRNA clusters. Total means of all intergenic piRNA cluster sequences per species are indicated by red diamond shaped points. Means of whole genomes are shown by star shaped points. (D) Pearson correlations between genomic gene density and pseudogene density. (E) Total Alu/L1 shares of 1 Mb genomic slices that contain piRNA clusters compared to the whole genome. *p-values (Alu): <2.2e-16, <2.2e-16, <2.2e-16, <2.2e-16, <2.2e-16, 2.4e-09. *p-values (L1): <2.2e-16, <2.2e-16, <2.2e-16, <2.2e-16, 4.9e-08, 2.4e-09, 0.099 (n.s.). (F) Ratios of Alu/L1 sequences shares between piRNA clusters and genomic sequence. (G) Mean TE divergences of 1 Mb genomic slices that contain piRNA clusters compared to the whole genome. Total means of all piRNA cluster sequences per species are indicated by red diamond shaped points. Means of whole genomes are shown by star shaped points. *p-values: <2.2e-16, <2.2e-16, <2.2e-16, <2.2e-16, <2.2e-16, <2.2e-16, 2.2e-09.

Taken together, these results suggest that primate piRNA clusters are more likely to inhabit more active regions of the genome with a more open chromatin structure. While it is known that *Drosophila* piRNA clusters exhibit heterochromatic features (Brennecke et al. 2007), it has been demonstrated in BmN4 cells that piRNA clusters of the silkworm are enriched with euchromatic epigenetic marks, foremost H3K4me3 and H3K4me2 (Kawaoka et al. 2013). Our results indicate that in primates, rather than representing islands of open chromatin within heterochromatic regions, piRNA clusters are embedded in active regions of the genome that are more likely to contain newer transposon copies but which also contain pseudogene sequences in higher abundance. This might for some part explain the enrichment of pseudogenes in mammalian piRNA clusters.

5.5. Conclusion

Pseudogenes that are located in piRNA clusters in reverse orientation have been suggested to be an important source of pachytene antisense piRNAs that direct regulation of parent genes (Hirano et al. 2014, Gebert et al. 2015). However, due to a lack of evidence for selection and very weak conservation of targeting of homologous genes, our study indicates that the presence of pseudogenes in piRNA-producing loci might be rather a product of chance, since piRNA clusters tend to be located in regions with elevated gene density, and does not have a significant impact on gene regulation by pachytene piRNAs. Instead, another mechanism, such as piRNA-targeting of transposon sequences in 3'-UTRs

(Watanabe et al. 2015, Zhang et al. 2015), is more likely to represent the main mode of gene regulation in late mammalian spermatogenesis. However, further research is needed to validate this claim. One would expect, for instance, that certain TE sequences in the 3'-UTRs of some coding genes become evolutionary fixed and conserved to ensure faithful regulation by piRNAs, whereas other TE insertions would likely have a negative effect, due to unwanted interference with normal gene expression. The examination of such signs of selection and conservation of TEs in 3'-UTRs would help to understand the mechanisms and the evolution of piRNA-mediated gene regulation.

5.6. Declarations

Acknowledgements

We would like to thank René Ketting and Mark Helm for valuable and fruitful discussions during the course of this project. Further thanks go to Julia Schumacher, Sacha Heerschop and Isabel Fast for helpful comments. This work was supported by the International PhD Programme (IPP) coordinated by the Institute of Molecular Biology IMB, Mainz, Germany, funded by the Boehringer Ingelheim Foundation.

Author contributions

DR, HZ and **DG** conceived the study. **DG** performed all analyses and coded bioinformatics software. **DG** wrote the manuscript.

5.7. References

- Aravin, A. A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 442, 203-207.
- Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., et al. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 31, 785-799.
- Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., Hannon, G. J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science. 316, 744-747.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., et al. (2000). Gene ontology: Tool for the unification of biology. Nat Genet. 25, 25-29.
- Assis, R., Kondrashov, A. S. (2009). Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. Proc Natl Acad Sci USA. 106, 7079-7082.
- Barckmann, B., Pierson, S., Dufourt, J., Papin, C., Armenise, C., et al. (2015). Aubergine iCLIP Reveals piRNA-Dependent Decay of mRNAs Involved in Germ Cell Development in the Early Embryo. Cell Rep. 12, 1205-1216.
- Beyret, E., Liu, N., & Lin, H. (2012). PiRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. Cell Res. 22, 1429-1439.
- Birx, H. J. (2006). Encyclopedia of Anthropology. SAGE Publications, Inc.
- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., et al. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 128, 1089-103.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. BMC Bioinformatics. 10, 421.
- Carmell, M. A, Girard, A., van de Kant, H. J. G., Bourc'his, D., Bestor, T. H., et al. (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell. 12, 503-514.
- Chirn, G., Rahman, R., Sytnikova, Y. A., Matts, J. A., Zeng, M., et al. (2015). Conserved piRNA Expression from a Distinct Set of piRNA Cluster Loci in Eutherian Mammals. PLoS Genet. 11, e1005652.
- Cora, E., Pandey, R. R., Xiol, J., Taylor, J., Sachidanandam, R., et al. (2014). The MID-PIWI module of Piwi proteins specifies nucleotide- and strand-biases of piRNAs. RNA. 20, 773-781.
- Czech, B., Hannon, G. J. (2016). One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci. 41, 324-337.
- De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., et al. (2011). The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature. 480, 259-263.
- Dekker, J. (2007). GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. 8, R116.

- Gebert, D., Ketting, R. F., Zischler, H., Rosenkranz, D. (2015). piRNAs from pig testis provide evidence for a conserved role of the Piwi pathway in posttranscriptional gene regulation in mammals. PLoS One. 10, e0124860.
- Gebert, D., Rosenkranz, D. (2015). RNA-based regulation of transposon expression. Wiley Interdiscip Rev RNA. 6, 687-708.
- Gebert, D., Hewel, C., Rosenkranz, D. (2017). Unitas: The universal tool for annotation of small RNAs. BMC Genomics. 18, 644.
- Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P., et al. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell. 118, 555-566.
- Girard, A., Sachidanandam, R., Hannon, G. J., Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 442, 199-202.
- Goh, W. S. S., Falciatori, I., Tam, O. H., Burgess, R., Meikar, O., et al. (2015). piRNA-directed cleavage of meiotic transcripts regulates spermatogenesis. Genes Dev. 29, 1032-1044.
- Gou, L.-T., Dai, P., Yang, J.-H., Xue, Y., Hu, Y.-P., et al. (2014). Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. 24, 680-700.
- Grivna, S. T., Beyret, E., Wang, Z., Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev. 20, 1709-14.
- Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., et al. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science. 315, 1587-1590.
- Ha, H., Song, J., Wang, S., Kapusta, A., Feschotte, C., et al. (2014). A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. BMC Genomics. 15, 545.
- Hirano, T., Iwasaki, Y. W., Lin, Z. Y.-C., Imamura, M., Seki, N. M., et al. (2014). Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. RNA. 20, 1223-1237.
- Huang, X. A., Yin, H., Sweeney, S., Raha, D., Snyder, M., et al. (2013). A major epigenetic programming mechanism guided by piRNAs. Dev Cell. 24, 502-16.
- Ipsaro, J., Haase, A., Knott, S. (2012). The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. Nature. 491, 279-283.
- Jiang, H., Wong, W. H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 24, 2395-2396.
- Kawaoka, S., Hara, K., Shoji, K., Kobayashi, M., Shimada, T., et al. (2013). The comprehensive epigenome map of piRNA clusters. Nucleic Acids Res. 41, 1581-1590.
- Kinsella, R. J., Ka, A., Spudich, G., Almeida-King, J., Staines, D., et al. (2011). Original article Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011, bar030.
- Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., Schmitz, J. (2007). Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends Genet. 23, 158-161.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., et al. (2001). Initial sequencing and analysis of the human genome. Nature. 409, 860-921.
- Li, J., Han, K., Xing, J., Kim, H., Rogers, J., et al. (2009). Phylogeny of the macaques (Cercopithecidae : Macaca) based on Alu elements. Gene. 448, 242-249.
- Masters, J., Gamba, M., Génin, F. (2012). Leaping Ahead: Advances in Prosimian Biology. Springer.
- Nishimasu, H., Ishizu, H., Saito, K., Fukuhara, S., Kamatani, M. K., et al. (2012). Structure and function of Zucchini endoribonuclease in piRNA biogenesis. Nature. 491, 284-287.
- Pantano, L., Jodar, M., Bak, M., Ballescà, J. L., Tommerup, N., et al. (2015). The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. RNA. 21, 1085-1095.
- Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., et al. (2016). Expression Atlas update - an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 44, D746-752.
- Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., et al. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature. 480, 264-267.
- Robine, N., Lau, N. C., Balla, S., Jin, Z., Okamura, K., et al. (2009). A broadly conserved pathway generates 3'UTR-directed primary piRNAs. Curr Biol. 19, 2066-2076.
- Rojas-Ríos, P., Chartier, A., Pierson, S., Simonelig, M. (2017). Aubergine and piRNAs promote germline stem cell self-renewal by repressing the proto-oncogene Cbl. EMBO J. 36, 3194-3211.
- Roovers, E. F., Rosenkranz, D., Mahdipour, M., Han, C. T., He, N., et al. (2015). Piwi proteins and piRNAs in mammalian oocytes and early embryos. Cell Rep. 10, 2069-2082.
- Rosenkranz, D., Han, C. T., Roovers, E. F., Zischler, H., Ketting, R. F. (2015b). Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. Genom Data. 5, 309-313.
- Rosenkranz, D., Rudloff, S., Bastuck, K., Ketting, R. F., Zischler, H. (2015a). Tupaia small RNAs provide insights into function and evolution of RNAi-based transposon defense in mammals. RNA. 21, 911-922.
- Rosenkranz, D., Zischler, H. (2012). proTRAC--a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics. 13, 5.
- Shumaker, R. W., Beck, B. B. (2003). Primates in Question. Smithsonian Institution Press.

- Sisu, C., Pei, B., Leng, J., Frankish, A., Zhang, Y., et al. (2014). Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci U S A. 111, 13361-13366.
- Stevens, N. J., Seiffert, E. R., Connor, P. M. O., Roberts, E. M., Schmitz, M. D., et al. (2013). Divergence between Old World monkeys and apes. Nature. 497, 611–614.
- The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 45, D331–338.
- Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., et al. (2005). SVA Elements: A Hominid-specific Retroposon Family. J Mol Biol. 354, 994-1007.
- Wang, W., Yoshikawa, M., Han, B. W., Izumi, N., Tomari, Y., et al. (2014). The initial uridine of primary piRNAs does not create the tenth adenine that is the hallmark of secondary piRNAs. Mol Cell. 56, 708–716.
- Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., et al. (2006). Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev. 20, 1732-1743.
- Watanabe, T., Cheng, E., Zhong, M., Lin, H. (2015). Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. Genome Res. 25, 368-380.
- Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P. M., et al. (2006). PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics. 22, 1437-1439.
- Zhang, Y., Guo, R., Cui, Y., Zhu, Z., Zhang, Y., et al. (2017). An essential role for PNLDC1 in piRNA 3′ end trimming and male fertility in mice. Cell Res. 27, 1392-1396.
- Zhang, P., Kang, J., Gou, L., Wang, J., Xue, Y., et al. (2015). MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. Cell Res. 25, 193–207.
- Zhang Z, Xu J, Koppetsch BS, Wang J, Tipping C, Ma S, et al. (2011). Heterotypic piRNA Ping-Pong Requires Qin, a Protein with Both E3 ligase and Tudor Domains. Mol Cell. 44, 572-584.

5.8. Supplement

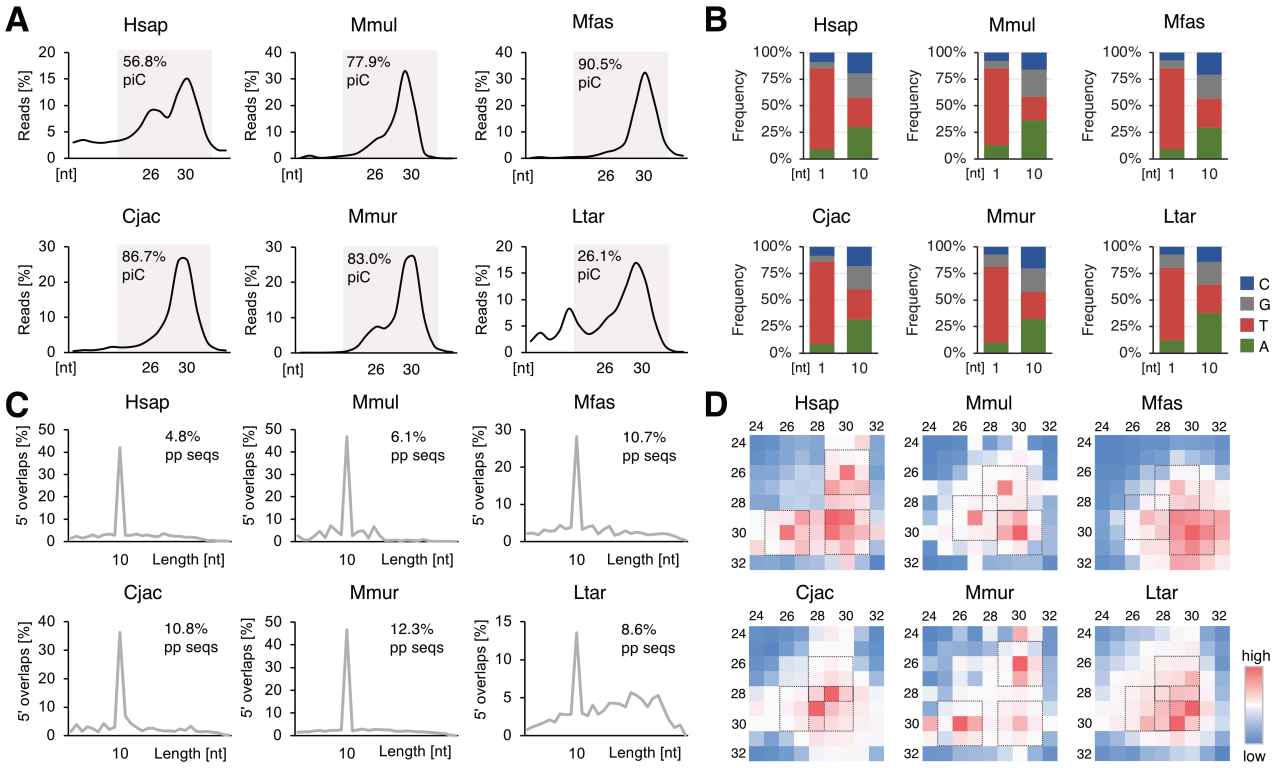


Figure S1 | Basic analysis of primate testis sRNA transcriptome datasets for piRNA traits. (A) Read length distributions. Grey areas show typical mammalian piRNA size range of 24-32 nt. Percentage indicates shares of clustered 24-32 nt reads on genome, predicted by proTRAC. (B) Nucleotide frequencies in mapped reads of positions 1 and 10 starting from 5' end. (C) 5' overlaps of sense and antisense reads on genomic sequence. Percentage shows share of non-redundant sequences with ping-pong partner reads. (D) Matrices for frequencies of read length combinations in ping-pong pairs (Pairs of reads with 10 nt 5' overlaps). Squares mark inferred size range of piRNAs bound to specific PIWI proteins.

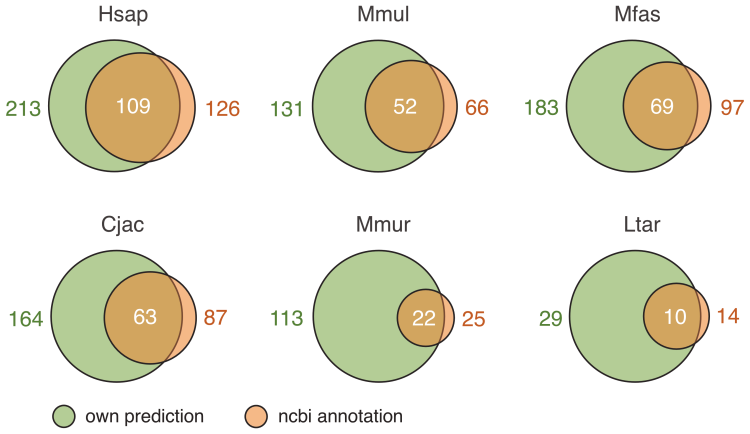


Figure S2 | Comparison of ncbi annotation and custom pseudogene prediction in piRNA cluster sequences. Counts of ncbi annotated and predicted pseudogenes with number of identical pseudogenes.

PANTHER GO-Slim Biological Process	REFLIST (18235)	Genes (1459)	Genes (expected)	Genes (fold Enrichment)	Genes (raw P-value)	Genes (FDR)
oxidative phosphorylation (GO:0006119)	45	12	3.6	3.33	8.10E-04	5.20E-03
mRNA 3'-end processing (GO:0031124)	34	9	2.72	3.31	3.71E-03	1.71E-02
protein folding (GO:0006457)	85	22	6.80	3.23	9.96E-06	1.16E-04
protein methylation (GO:0006479)	49	12	3.92	3.06	1.52E-03	8.81E-03
spermatogenesis (GO:0007283)	59	14	4.72	2.97	8.29E-04	5.19E-03
tRNA metabolic process (GO:0006399)	111	22	8.88	2.48	3.37E-04	2.41E-03
RNA catabolic process (GO:0006401)	66	13	5.28	2.46	7.4E-03	2.95E-02
respiratory electron transport chain (GO:0022904)	102	19	8.16	2.33	2.38E-03	1.21E-02
rRNA metabolic process (GO:0016072)	107	19	8.56	2.22	2.98E-03	1.43E-02
mRNA processing (GO:0006397)	238	42	19.0	2.21	1.43E-05	1.51E-04
translation (GO:0006412)	193	34	15.44	2.20	9.42E-05	7.93E-04
protein targeting (GO:0006605)	167	29	13.36	2.17	3.09E-04	2.36E-03
mRNA splicing, via spliceosome (GO:0000398)	173	28	13.84	2.02	1.06E-03	6.48E-03
RNA splicing, via transesterification reactions (GO:0000375)	151	23	12.08	1.90	7.9E-03	3.11E-02
generation of precursor metabolites and energy (GO:0006091)	166	24	13.28	1.81	1.12E-02	4.14E-02
protein metabolic process (GO:0019538)	1434	198	114.74	1.73	1.59E-12	7.77E-11
chromatin organization (GO:0006325)	254	35	20.32	1.72	4.31E-03	1.88E-02
proteolysis (GO:0006508)	428	54	34.24	1.58	2.60E-03	1.27E-02
protein localization (GO:0008104)	503	63	40.25	1.6	1.09E-03	6.48E-03
RNA metabolic process (GO:0016070)	1488	176	119.06	1.48	9.1E-07	1.31E-05
intracellular protein transport (GO:0006886)	670	78	53.61	1.46	2.1E-03	1.13E-02
organelle organization (GO:0006996)	1149	132	91.93	1.44	9E-05	7.54E-04
cellular protein modification process (GO:0006464)	759	87	60.73	1.43	1.95E-03	1.08E-02
cellular component biogenesis (GO:0044085)	730	83	58.41	1.42	3.2E-03	1.49E-02
response to stress (GO:0006950)	528	60	42.25	1.42	1.29E-02	4.55E-02
catabolic process (GO:0009056)	1117	126	89.37	1.41	2.7E-04	2.11E-03
protein transport (GO:0015031)	709	79	56.73	1.39	5.46E-03	2.26E-02
primary metabolic process (GO:0044238)	4428	484	354.29	1.37	2.48E-13	1.51E-11
nucleobase-containing compound metabolic process (GO:0006139)	2644	287	211.55	1.36	2.82E-07	4.6E-06
biosynthetic process (GO:0009058)	1677	177	134.18	1.32	3.30E-04	2.44E-03
metabolic process (GO:0008152)	5502	579	440.22	1.32	1.24E-13	1.01E-11
cellular component organization or biogenesis (GO:0071840)	1924	199	153.94	1.29	3.74E-04	2.61E-03
cellular component organization (GO:0016043)	1795	179	143.62	1.25	3.7E-03	1.7E-02
nitrogen compound metabolic process (GO:0006807)	2404	231	192.35	1.20	5.12E-03	2.19E-02
regulation of biological process (GO:0050789)	1974	124	157.94	0.79	5.42E-03	2.28E-02

Table S1 | GO term analysis: Panther GO slim biological process. Test: Fisher's Exact with FDR (false discovery rate) multiple test correction. Genes: 1428 orthologous genes with piRNA coverage in all six species. RefList: Testis-expressed genes.

PANTHER GO-Slim Molecular Function	REFLIST (18235)	Genes (1459)	Genes (expected)	Genes (fold Enrichment)	Genes (raw P-value)	Genes (FDR)
translation regulator activity (GO:0045182)	77	17	6.2	2.76	4.95E-04	1.18E-02
mRNA binding (GO:0003729)	134	27	10.72	2.52	5.57E-05	1.77E-03
structural constituent of ribosome (GO:0003735)	121	23	9.68	2.38	5.61E-04	1.07E-02
RNA binding (GO:0003723)	370	58	29.60	1.96	8.81E-06	3.35E-04
catalytic activity (GO:0003824)	4006	377	320.52	1.18	7.51E-04	1.30E-02

Table S2 | GO term analysis: Panther GO slim molecular function. Test: Fisher's Exact with FDR (false discovery rate) multiple test correction. Genes: 1428 orthologous genes with piRNA coverage in all six species. RefList: Testis-expressed genes.

PANTHER GO-Slim Cellular Component	REFLIST (18235)	Genes (1459)	Genes (expected)	Genes (fold Enrichment)	Genes (raw P-value)	Genes (FDR)
mitochondrial inner membrane (GO:0005743)	107	21	8.6	2.45	5.10E-04	2.04E-03
ribosome (GO:0005840)	154	28	12.32	2.27	2.72E-04	1.16E-03
ribonucleoprotein complex (GO:0030529)	413	67	33.04	2.03	5.34E-07	3.42E-06
nuclear outer membrane-endoplasmic reticulum membrane network (GO:0042175)	228	33	18.24	1.81	2.65E-03	8.92E-03
cytosol (GO:0005829)	496	70	39.69	1.76	2.05E-05	1.01E-04
endoplasmic reticulum (GO:0005783)	400	56	32.00	1.75	1.86E-04	8.49E-04
mitochondrion (GO:0005739)	382	51	30.56	1.67	1.1E-03	4.03E-03
macromolecular complex (GO:0032991)	1977	248	158.18	1.57	1.56E-11	1.66E-10
cytoplasm (GO:0005737)	3037	370	242.99	1.52	7.35E-16	2.35E-14
nucleoplasm (GO:0005654)	379	45	30.3	1.48	1.45E-02	4.03E-02
protein complex (GO:0043234)	1630	192	130.42	1.47	3.57E-07	2.5E-06
nucleus (GO:0005634)	1857	214	148.58	1.44	3.1E-07	2.47E-06
organelle (GO:0043226)	3718	416	297.48	1.40	1.50E-12	1.93E-11
intracellular (GO:0005622)	5019	556	401.6	1.38	4.48E-17	2.87E-15
cell part (GO:0044464)	5256	565	420.54	1.34	7.3E-15	1.56E-13

Table S3 | GO term analysis: Panther GO slim cellular component. Test: Fisher's Exact with FDR (false discovery rate) multiple test correction. Genes: 1428 orthologous genes with piRNA coverage in all six species. RefList: Testis-expressed genes.

ASB1	DCAF7	MDM4	POLH	TRIM44
ATXN1L	FAM53B	MIEF1	PRKAB2	UBAP1
CBL	FBXL18	NR6A1	SETX	WASF2
CBX5	GID8	OTUD3	SLC9A8	WIPF2
CCDC117	GOSR2	PDPK1	TBL2	ZHX3
CDS2	KIF24	PDPR	TEF	

Table S4 | List of coding genes with piRNA coverage (≥5 RPKM) in every primate species of this study and across eutherian species in the study of Chirn et al. 2015.

Script	Description	Method section
get_pp_partners.pl	Count reads with ping-pong partners and make matrix of read counts for each partner length combination	Basic analyses
merge_pic_loci.pl	Merge piRNA cluster loci with a distance less than 10 kb	piRNA cluster prediction
compare_strict_pic.pl	Compare piRNA clusters predicted with strict settings within the same species	piRNA cluster prediction
compare_all_pics.pl	Compare piRNA clusters predicted with strict settings within the same species, including less strictly predicted loci	piRNA cluster prediction
find_pic_hom_loci.pl	Identify homologous piRNA clusters in another species by finding syntenic regions and searching for sequence similarity with blastn/dc-megablast	Homologous piRNA cluster identification
get_hom_loc_lineages.pl	Combine pairs of homologous piRNA cluster loci to get lineages of homologous loci across several species	Homologous piRNA cluster identification
get_hom_loc_identities.pl	Extract information on rates of identified syntenic regions, found homologous loci, expressed piRNA clusters and sequence similarities of homologous loci	Homologous piRNA cluster analyses
blast_genomes.pl	Call dc-megablast on all combinations of given (repeatmasked) genome files	Homologous piRNA cluster analyses
blast_cds.pl	Call dc-megablast on all combinations of given cds files	Homologous piRNA cluster analyses
get_genome_identities.pl	Extract information on sequence similarities between genomes	Homologous piRNA cluster analyses
get_cds_identities.pl	Extract information on sequence similarities between cds of genes	Homologous piRNA cluster analyses
extract_all_pic_loci.pl	Extract information on all identified piRNA clusters	Homologous piRNA cluster analyses
ids_to_hom_loc_lineages.pl	Add piRNA cluster ids assigned by proTRAC to lineages of homologous loci and extract cluster expression rates	Homologous piRNA cluster analyses
get_hom_loc_TE_divs.pl	Get TE divergences for piRNA cluster regions from repeatmasker files	Homologous piRNA cluster analyses
heatmapper.R	Create expression heatmap with dendrogram for homologous piRNA clusters using hierarchical clustering, average linkage and pearson distance	Homologous piRNA cluster analyses
predict_pseudogenes.pl	Search for sequences similar to cds using dc-megablast in regions not annotated as genes and combine close hits to predict pseudogenes	Pseudogene prediction
get_pic_pseudogenes.pl	Extract pseudogenes in piRNA cluster regions from genome-wide annotation	Pseudogene analyses
compare_predictions.pl	Compare pseudogene prediction and ncbi annotation within cluster loci	Pseudogene analyses
get_pseudogene_info.pl	Extract information on orientation, type and parents of pseudogenes in the whole genome and within cluster loci	Pseudogene analyses
find_pic_hom_pseudogenes.pl	Search for homologous pseudogene sequences in homologous piRNA cluster loci across species with dc-megablast	Pseudogene analyses
get_pir_target_genes.pl	Use output file generated by seqmap, mapping piRNA reads on cdna, to find general targets and ping-pong genes	Target gene analyses
get_pic_target_genes.pl	Map antisense piRNA reads produced from pseudogenes in piRNA clusters on cdna using seqmap to find general targets and ping-pong genes	Target gene analyses
get_pir_target_orthologs.pl	Find orthologous genes targeted by piRNAs in general	Target gene analyses
get_pic_target_orthologs.pl	Find orthologous genes targeted by pseudogene-derived piRNAs	Target gene analyses
get_pir_target_3utrs.pl	Get 3'-UTR lengths and compare to piRNA coverage	Target gene analyses
get_pic_environments.pl	Scan whole genome with 1 Mb window and get information on TE shares, gene density, pseudogene density, GC content and piRNA cluster locations	Analysis of genomic environments
get_gc_share.pl	Calculate total GC share of genome/piRNA clusters	Analysis of genomic environments

Table S5 | Main Perl and R scripts developed and used in this study (github.com/d-gebert/primate-pic-evo).

Conclusion

The goal of this thesis is to facilitate small RNA research and to elucidate evolutionary changes that provide functional insights into the piRNA pathway. The sRNA annotation tool unitas is designed to be used in a great variety of species and for researchers that lack a strong bioinformatics background. Hopefully this will make annotation of sRNAs from large sequence data easier and more accessible, especially for analyses in non-model species. As new findings on sRNA pathways emerge, these new insights should be incorporated into bioinformatic tools and therefore unitas has to be constantly improved and updated in the future.

The work on molluskan PIWI proteins and piRNAs reveals the activity of the piRNA pathway in the soma, which suggests, together with other recent findings in arthropod (Lewis et al. 2018) and cnidarian species (Praher et al. 2017), that a strong germline-specificity in vertebrates represents an evolutionally adaptive state. This study also shows that different sets of piRNA clusters can be dynamically expressed during development, while the highest activity is detected in gonads and early developmental stages, which already indicates a specialization of germline and embryonic piRNA clusters. This pattern should be further examined to unveil the detailed sub-functionalization of clusters. It would also be interesting to analyze further animal groups, especially from basal phyla, which might reveal other specializations and would help to test the notion that ubiquitous activity of the PIWI/piRNA pathway represents the ancestral state in animals. Moreover, rigorous analysis of potential somatic piRNAs in mammals is needed to clarify the activity status in mammalian somatic cells, since previous attempts are regarded as flawed (Ross et al. 2014, Tosar et al. 2018).

Furthermore, the studies on mammalian germline piRNAs provide evidence that, in addition to TEs, protein-coding genes are regulated by the PIWI/piRNA complex. However, despite initial assumptions, the work on piRNA cluster evolution in primates suggests that pseudogene-containing piRNA clusters seem not to be a major source of gene-targeting piRNAs that would trigger the secondary pathway. Hence other mechanisms, such as the recognition of transposon insertions or other target motives in the 3'-UTRs of genic mRNAs, are likely responsible for the targeting of the majority of coding genes in mammalian testes (Robine et al. 2009, Ha et al. 2014, Watanabe et al. 2015). However, definitive evidence for the exact selection mechanism of RNA molecules as piRNA precursors, whether it be genic mRNA, transposon RNA or piRNA cluster transcripts, is currently still incomplete. In particular, the inspection of TE insertions in 3'-UTRs across species would give insights into their role in piRNA-mediated gene regulation. The analyses here undertaken in primates further revealed the evolutionary relationships of piRNA clusters within this group that emerged about 65 million years ago (Birx 2006), and shed light on sequence evolution and the rate by which cluster loci are retained and active. Yet it is still difficult to state what the exact forces are that potentially drive cluster evolution and ultimately what circumstances are required to transform a genomic locus into a new piRNA cluster. In this context the distinct expression profiles of piRNA clusters in different primate species could hide some hints that point to some mechanisms of adaptation that might have been overlooked so far. Therefore detailed and extensive analysis of differentially expressed piRNA clusters across species could yield some interesting insights into the evolution of piRNA-producing loci.

References

- Aravin, A. A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 442, 203-207.
- Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., et al. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 31, 785-799.
- Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., Hannon, G. J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science. 316, 744-747.
- Aravin, A. A., van der Heijden, G. W., Castañeda, J., Vagin, V. V., Hannon, G. J., et al. (2009). Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. PLoS Genet. 5, e1000764.
- Axtell, M.J. (2013). Classification and comparison of small RNAs from plants. Annu. Rev. Plant Biol. 64, 137-159.
- Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P., Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev. 22, 2773-2785.
- Barckmann, B., Pierson, S., Dufourt, J., Papin, C., Armenise, C., et al. (2015). Aubergine iCLIP Reveals piRNA-Dependent Decay of mRNAs Involved in Germ Cell Development in the Early Embryo. Cell Rep. 12, 1205-1216.
- Baumberger, N., Baulcombe, D. C. (2005) Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. Proc Natl Acad Sci USA. 102, 11928-11933.
- Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., et al. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev. 20, 1885-1898.
- Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. Nature Genet. 37, 766-770.
- Bernstein, E., Caudy, A. A., Hammond, S. M., Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature. 409, 363-366.
- Beyret, E., Liu, N., Lin, H. (2012). piRNA biogenesis during adult spermatogenesis in mice is independent of the ping-pong mechanism. Cell Res. 22, 1429-1439.
- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., et al. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 128, 1089-103.
- Brower-Toland, B., Findley, S. D., Jiang, L., Liu, L., Yin, H., et al. (2007). Drosophila PIWI associates with chromatin and interacts directly with HP1a. Genes Dev. 21, 2300-2311.
- Bühler, M., Verdel, A., Moazed, D. (2006). Tethering RITS to a Nascent Transcript Initiates RNAi- and Heterochromatin-Dependent Gene Silencing. Cell. 125, 873-886.
- Burroughs, A. M., Ando, Y., de Hoon, M. J., Tomaru, Y., Nishibu, T., et al. (2010). A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. Genome Res. 20, 1398-1410.
- Carmell, M. A, Girard, A., van de Kant, H. J. G., Bourc'his, D., Bestor, T. H., et al (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell. 12, 503-14.
- Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., et al. (2010). Mammalian microRNAs: Experimental evaluation of novel and previously annotated genes. Genes Dev. 24, 992-1009.
- Chung, W. J., Okamura, K., Martin, R., Lai, E. C. (2008). Endogenous RNA Interference Provides a Somatic Defense against Drosophila Transposons. Curr Biol. 18, 795-802.
- Cogoni, C., Macino, G. (1999). Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase. Nature. 399, 166-169.
- Cole, C., Sobala, A., Lu, C., Thatcher, S. R., Bowman, A., et al. (2009). Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA. 15, 2147-60.
- Cost, G. J., Feng, Q., Jacquier, A., Boeke, J.D. (2002). Human L1 element target-primed reverse transcription in vitro. EMBO J. 21, 5899-5910.
- Cox, D. N., Chao, A., Lin, H. (2000). piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development. 127, 503-514.
- Dalmay, T., Hamilton, A., Rudd, S., Angell, S., Baulcombe, D. C. (2000). An RNA-dependent RNA polymerase gene in Arabidopsis is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. Cell. 101, 543-553.
- De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., et al. (2011). The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature. 480, 259-263.
- Deng, W., Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell. 2, 819-830.
- Denli, A. M., Tops, B. B. J., Plasterk, R. H. A., Ketting, R. F., Hannon, G. J. (2004). Processing of primary microRNAs by the microprocessor complex. Nature. 432, 231-235.
- Döring, H. P., Pahl, I., Durany, M. (1990). Chromosomal rearrangements caused by the aberrant transposition of double Ds elements are formed by Ds and adjacent non-Ds sequences. Mol Gen Genet. 224, 40-48.

- Edgecombe, G. D., Giribet, G., Dunn, C. W., Hejnol, A., Kristensen, R. M., et al. (2011). Higher-level metazoan relationships: recent progress and remaining questions. Org Divers Evol 11: 151-172.
- Eickbush, T. H., Jamburuthugoda, V. K. (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. 134, 221-234.
- Elbashir, S. M., Lendeckel, W., Tuschl, T. (2001). RNA interference is mediated by 21 and 22 nt RNAs. Genes Dev. 15, 188-200.
- Fagegaltier, D., Bougé, A. L., Berry, B., Poisot, E., Sismeiro, O., et al. (2009). The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proc Natl Acad Sci USA. 106, 21258-21263.
- Fang, W., Wang, X., Bracht, J. R., Nowacki, M., Landweber, L. F. (2012). Piwi-interacting RNAs protect DNA against loss during Oxytricha genome rearrangement. Cell. 151, 1243-1255.
- Fei, Q., Xia, R., Meyers, B. C. (2013). Phased, Secondary, Small Interfering RNAs in Posttranscriptional Regulatory Networks. Plant Cell. 25, 2400-2415.
- Feltzin, V. L., Khaladkar, M., Abe, M., Parisi, M., Hendriks, G. J., et al. (2015). The exonuclease Nibbler regulates age-associated traits and modulates piRNA length in Drosophila. Aging Cell. 14, 443-452.
- Feschotte, C., Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41, 331-368.
- Feschotte, C., Mouchès, C. (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. Mol Biol Evol. 17, 730-737.
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., et al. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 391, 806-811.
- Flemr, M., Malik, R., Franke, V., Nejepinska, J., Sedlacek, R., et al. (2013). A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. Cell. 155, 807-16.
- Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., et al. (2008). Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science. 320, 1077-1081.
- Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P., et al. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell. 118, 555-566.
- Girard, A., Sachidanandam, R., Hannon, G. J., Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 442, 199-202.
- Goodier, J. L. (2016). Restricting retrotransposons: a review. Mob DNA. 7, 16.
- Gou, L.-T., Dai, P., Yang, J.-H., Xue, Y., Hu, Y.-P., et al. (2014). Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. 24, 680-700.
- Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., et al. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. Cell. 106, 23-34.
- Grivna, S. T., Beyret, E., Wang, Z., Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev. 20, 1709-1714.
- Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., et al. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science. 315, 1587-1590.
- Hall, I. M., Shankaranarayana, G. D., Noma, K., Ayoub, N., Cohen, A., et al. (2002). Establishment and maintenance of a heterochromatin domain. Science. 297, 2232-2237.
- Hamilton, A. J., Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science. 286, 950-952.
- Hammond, S. M., Bernstein, E., Beach, D., Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature. 404, 293-296.
- Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R., Hannon, G. J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. Science. 293, 1146-1150.
- Han, B. W., Wang, W., Li, C., Weng, Z., Zamore, P. D. (2015). piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science. 348, 817-821.
- Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., et al. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA. 16, 673–695.
- Hirano, T., Iwasaki, Y. W., Lin, Z. Y.-C., Imamura, M., Seki, N. M., et al. (2014). Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. RNA. 20, 1223-1237.
- Homolka, D., Pandey, R. R., Goriaux, C., Brasset, E., Vaury, C., et al. (2015). PIWI Slicing and RNA Elements in Precursors Instruct Directional Primary piRNA Biogenesis. Cell Rep. 12, 418-428
- Horwich, M. D., Li, C., Matranga, C., Vagin, V., Farley, G., et al. (2007). The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC. Curr Biol. 17, 1265-1272.
- Houwing, S., Berezikov, E., Ketting, R. F. (2008). Zili is required for germ cell differentiation and meiosis in zebrafish. EMBO J. 27, 2702-2711.
- Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., et al. (2007). A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. Cell. 129, 69–82.

- Huang, H., Gao, Q., Peng, X., Choi, S. Y., Sarma, K., et al. (2011). piRNA-Associated Germline Nuage Formation and Spermatogenesis Require MitoPLD Profusogenic Mitochondrial-Surface Lipid Signaling. Developmental Cell. 20, 376-387.
- Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., et al. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science. 293, 834-838.
- Ipsaro, J., Haase, A., Knott, S. (2012). The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. Nature. 491, 279-283.
- Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., et al. (2010). Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. Mol Cell. 39, 292-299.
- Izumi, N., Shoji, K., Sakaguchi, Y., Honda, S., Kirino, Y., et al. (2016). Identification and Functional Analysis of the Pre-piRNA 3′ Trimmer in Silkworms. Cell. 164, 962-973.
- Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA. 94, 1872-1877.
- Kapitonov, V. V., Jurka J. (2003). A novel class of SINE elements derived from 5S rRNA. Mol Biol Evol. 20, 694-702.
- Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, et al. (2007). Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science. 315, 1137-1140
- Kawaoka, S., Izumi, N., Katsuma, S., Tomari, Y. (2011). 3' end formation of PIWI-interacting RNAs in vitro. Mol Cell.43, 1015–1022.
- Kawaoka, S., Minami, K., Katsuma, S., Mita, K., Shimada, T. (2008). Developmentally synchronized expression of two Bombyx mori Piwi subfamily genes, SIWI and BmAGO3 in germ-line cells. Biochem. Biophys Res Commun. 367, 755-760.
- Kazazian, H. H. Jr., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., et al. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature. 332, 164-166.
- Kazazian, H. H. Jr. (2004). Mobile elements: drivers of genome evolution. Science. 303, 1626-1632.
- Kazazian, H. H. Jr., Moran J. V. (1998). The impact of L1 retrotransposition on the human genome. Nat Genet. 19, 19-24.
- Keam, S., Hutvagner, G. (2015). tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. Life. 5, 1638-1651.
- Kirino, Y., Mourelatos, Z. (2007a). Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. Nat Struct Mol Biol. 14, 347-348.
- Kirino, Y., Mourelatos, Z. (2007b). The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. RNA. 13, 1397-1401.
- Kiuchi, T., Koga, H., Kawamoto, M., Shoji, K., Sakai, H., et al. (2014). A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature. 509, 633–636.
- Kolosha, V. O., Martin, S. L. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc Natl Acad Sci USA. 94, 10155-10160.
- Kotaja, N., Bhattacharyya, S. N., Jaskiewicz, L., Kimmins, S., Parvinen, M., et al. (2006). The chromatoid body of male germ cells: Similarity with processing bodies and presence of Dicer and microRNA pathway components. Proc Natl Acad Sci USA. 103, 2647-2652.
- Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., et al. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev. 22, 908-917.
- Kurihara, Y., Watanabe, Y. (2004). From The Cover: Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. Proc Natl Acad Sci USA. 101, 12753-12758.
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science. 294, 853-858.
- Lau, N. C., Lim, L. P., Weinstein, E. G., Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science. 294, 858-862.
- Le Thomas, A., Rogers, A. K., Webster, A., Marinov, G. K., Liao, S. E., et al. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev. 27, 390-399.
- Lee, R. C., Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science. 294, 862-864.
- Lee, R. C., Feinbaum, R. L., Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 75, 843-854.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature. 425, 415-419.
- Lee, Y., Jeon, K., Lee, J.-T., Kim, S., Kim, V. N. (2002). MicroRNA maturation: stepwisee processing and subcellular localization. EMBO J, 21, 4663-4670.
- Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., et al. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J, 23, 4051-4060.
- Lerat, E., Capy, P. (1999). Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol. 16, 1198-1207.

- Lewis, B. P., Burge, C. B., Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 120, 15-20.
- Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., Burge, C. B. (2003). Prediction of Mammalian MicroRNA Targets. Cell. 115, 787-798.
- Lewis, S. H., Quarles, K. A., Yang, Y., Tanguy, M., Frézal, L. et al. (2018). Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. Nat Ecol Evol. 2, 174-181.
- Lewis, S. H., Salmela, H., Obbard, D. J. (2016). Duplication and diversification of dipteran argonaute genes, and the evolutionary divergence of Piwi and Aubergine. Genome Biol Evol. 8, 507-518.
- Li, C., Vagin, V. V., Lee, S., Xu, J., Ma, S., et al. (2009). Collapse of Germline piRNAs in the Absence of Argonaute3 Reveals Somatic piRNAs in Flies. Cell. 137, 509-521.
- Lingel, A., Simon, B., Izaurralde, E., Sattler, M. (2004). Nucleic acid 3′-end recognition by the Argonaute2 PAZ domain. Nat Struct Mol Biol. 11, 576-577.
- Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., et al. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science. 305, 1437-1441.
- Lu, S., Sun, Y. H., Chiang, V. L. (2009). Adenylation of plant miRNAs. Nucleic Acids Res. 37, 1878-1885.
- Luan, D. D., Korman M. H., Jakubczak J. L., Eickbush T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 72, 595-605.
- Lund, E., Güttinger, S., Calado, A., Dahlberg, J.E., Kutay, U. (2004). Nuclear export of microRNA precursors. Science. 303, 95-98.
- Ma, J. B., Ye, K., Patel, D. J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. Nature. 429, 318-322.
- Malone, C. D., Brennecke, J., Dus, M., Stark, A., McCombie, W. R., et al. (2009). Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. Cell. 137, 522-535.
- Manakov, S. A., Pezic, D., Marinov, G. K., Pastor, W. A., Sachidanandam, R., et al. (2015). MIWI2 and MILI Have Differential Effects on piRNA Biogenesis and DNA Methylation. Cell Reports, 12, 1234-1243.
- Martin C, Lister C. (1989). Genome juggling by transposons: Tam3-induced rearrangements in Antirrhinum majus. Dev Genet. 10, 438-451.
- Martinez, G., Choudury, S. G., Slotkin, R. K. (2017). TRNA-derived small RNAs target transposable element transcripts. Nucleic Acids Res. 45, 5142-5152.
- Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell. 110, 563-574.
- Mathias, S. L., Scott, A. F., Kazazian, H. H. Jr., Boeke, J. D., Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science. 254, 1808-1810.
- McCaffrey, A. P., Nakai, H., Pandey, K., Huang, Z., Salazar, F. H., et al. (2003). Inhibition of hepatitis B virus in mice by RNA interference. Nat Biotechnol. 21, 639-644.
- McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA. 36, 344-355.
- Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., et al. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell. 15, 185-197.
- Mette, M. F., Aufsatz, W., Van der Winden, J., Matzke, M. A., Matzke, A. J. M. (2000). Transcriptional silencing and promoter methylation triggered by double-stranded RNA. EMBO J, 19, 5194-5201.
- Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., et al. (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. 52, 643-645.
- Mohn, F., Handler, D., Brennecke, J. (2015). piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science. 348, 812-817.
- Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., et al. (1996). High frequency retrotransposition in cultured mammalian cells. Cell. 87, 917-927.
- Morin, R. D., Connor, M. D. O., Griffith, M., Kuchenbauer, F., Delaney, A., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res. 18, 610-621.
- Nishimasu H., Ishizu H., Saito K., Fukuhara S., Kamatani M.K., et al. (2012). Structure and function of Zucchini endoribonuclease in piRNA biogenesis. Nature. 491, 284-287.
- Ohara, T., Sakaguchi, Y., Suzuki, T., Ueda, H., Miyauchi, K., et al. (2007). The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated. Nat Struct Mol Biol. 14, 349-350.
- Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., Lai, E. C. (2007). The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. Cell. 130, 89-100.
- Olsen, P. H., Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol. 216, 671-680.
- Ostertag, E. M., Goodier, J. L., Zhang, Y., Kazazian, H. H. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet. 73, 1444-1451.
- Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., et al. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature. 408, 86-89.

- Pedersen, I. M., Cheng, G., Wieland, S., Volinia, S., Croce, C. M., et al. (2007). Interferon modulation of cellular microRNAs as an antiviral mechanism. Nature. 449, 919-922.
- Praher, D., Zimmermann, B., Genikhovich, G., Columbus-Shenkar, Y., Modepalli, V. et al. (2017). Characterization of the piRNA pathway during development of the sea anemone Nematostella vectensis. RNA Biol. 14, 1727-1741.
- Rajasethupathy, P., Antonov, I., Sheridan, R., Frey, S., Sander, C., et al. (2012). A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. Cell. 149, 693-707.
- Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelll, A. E., Bettlnger, J. C., et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 403, 901-906.
- Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., Bartel, D. P. (2002). MicroRNAs in plants. Genes Dev. 16, 1616-1626.
- Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., et al. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature. 480, 264-267.
- Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., et al. (2002). Prediction of Plant MicroRNA Targets. Cell. 110, 513-520.
- Rivas, F. V., Tolia, N. H., Song, J. J., Aragon, J. P., Liu, J., et al. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. Nat Struct Mol Biol. 12, 340-349.
- Ro, S., Park, C., Young, D., Sanders, K. M., Yan, W. (2007). Tissue-dependent paired expression of miRNAs. Nucleic Acids Res. 35, 5944-5953.
- Ross, R. J., Weiner, M. M., Lin, H. (2014). PIWI proteins and PIWI-interacting RNAs in the soma. Nature. 505, 353-359.
- Rouget, C., Papin, C., Boureux, A., Meunier, A.-C., Franco, B., et al. (2010). Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. Nature. 467, 1128-1132.
- Rozhkov, N. V., Hammell, M., Hannon, G. J. (2013). Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev. 27, 400-412.
- Ruby, J. G., Jan, C. H., Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature. 448, 83-86.
- Saito, K., Inagaki, S., Mituyama, T., Kawamura, Y., Ono, Y., et al. (2009). A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. Nature. 461, 1296-1299.
- Saito, K., Nishida, K. M., Mori, T., Kawamura, Y., Miyoshi, K., et al. (2006). Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev. 20, 2214-2222.
- Saito, K., Sakaguchi, Y., Suzuki, T., Suzuki, T., Siomi, H., et al. (2007). Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3' ends. Genes Dev. 21, 1603-1608.
- Schorn, A. J., Gutbrod, M. J., LeBlanc, C., Martienssen, R. (2017). LTR-Retrotransposon Control by tRNA-Derived Small RNAs. Cell. 170, 61-71.
- Scott, A. F., Schmeckpeper, B. J., Abdelrazik, M., Comey, C. T., O'Hara, B., et al. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics. 1, 113-125.
- Shabalina, S. A., Koonin, E. V. (2008). Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol. 23, 578-587.
- Sharma, U., Conine, C. C., Shea, J. M., Boskovic, A., Derr, A. G., et al. (2016). Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science. 351, 391-396.
- Sienski, G., Dönertas, D., Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell. 151, 964-980.
- Sigova, A., Rhind, N., Zamore, P. D. (2004). A single Argonaute protein mediates both transcriptional and posttranscriptional silencing in Schizosaccharomyces pombe. Genes Dev. 18, 2359-2367.
- Sijen, T., Plasterk, R. H. (2003). Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature. 426, 310-314.
- Singer, M. F. (1982). SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell. 28, 433-434.
- Smardon, A., Spoerke, J. M., Stacey, S. C., Klein, M. E., Mackin, N., et al. (2000). EGO-1 is related to RNA-directed RNA polymerase and functions in germline development and RNA interference in C. elegans. Curr Biol. 10, 169-178.
- Stein, P., Svoboda, P., Anger, M., Schultz, R. M. (2003). RNAi: mammalian oocytes do it without RNA-dependent RNA polymerase. RNA. 9, 187-192.
- Stoye, J. P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol. 10, 395-406.
- Swergold, G. D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol. 10, 6718-6729.
- Tabara, H., Sarkissian, M., Kelly, W. G., Fleenor, J., Grishok, A., et al. (1999). The rde-1 gene, RNA interference, and transposon silencing in C. elegans. Cell. 99, 123-132.
- 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, 974-984.

- Tosar, J. P., Rovira, C., Cayota, A. (2018). Non-coding RNA fragments account for the majority of annotated piRNAs expressed in somatic non-gonadal tissues. Commun Biol. 1, 2.
- Ullu, E., C. Tschudi. (1984). Alu sequences are processed 7SL RNA genes. Nature. 312, 171-172.
- Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., et al. (2006). A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. Science. 313, 320-324.
- Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., et al. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. Science. 303, 672-676.
- Volpe, T., Kidner, C., Hall, I., Teng, G., Grewal, S., et al. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science. 297, 1833-1837.
- Wang, S. H., Elgin, S. C. R. (2011). Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. Proc Natl Acad Sci USA. 108, 21164-21169.
- Wang, W., Yoshikawa, M., Han, B. W., Izumi, N., Tomari, Y., et al. (2014). The initial uridine of primary piRNAs does not create the tenth adenine that is the hallmark of secondary piRNAs. Mol Cell. 56, 708-716.
- Watanabe, T., Chuma, S., Yamamoto, Y., Kuramochi-Miyagawa, S., Totoki, Y., et al. (2011). MITOPLD Is a Mitochondrial Protein Essential for Nuage Formation and piRNA Biogenesis in the Mouse Germline. Dev Cell. 20, 364-375.
- Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., et al. (2006). Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev. 20, 1732-1743.
- Wilkins, C., Dishongh, R., Moore, S. C., Whitt, M. A., Chow, M., et al. (2005). RNA interference is an antiviral defence mechanism in Caenorhabditis elegans. Nature. 436, 1044-1047.
- Yan, Z., Hu, H. Y., Jiang, X., Maierhofer, V., Neb, E., et al. (2011). Widespread expression of piRNA-like molecules in somatic tissues. Nucleic Acids Res. 39, 6596-6607.
- Yang, Z., Ebright, Y. W., Yu, B., Chen, X. (2006). HEN1 recognizes 21–24 nt small RNA duplexes and deposits a methyl group onto the 2' OH of the 3' terminal nucleotide. Nucleic Acids Res. 34, 667-675.
- Yang, Z., Chen, K. M., Pandey, R. R., Homolka, D., Reuter, M., et al. (2016). PIWI Slicing and EXD1 Drive Biogenesis of Nuclear piRNAs from Cytosolic Targets of the Mouse piRNA Pathway. Mol Cell. 61, 138-152.
- Yi, R., Qin, Y., Macara, I. G., Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. 17, 3011-3016.
- Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., et al. (2005). Methylation as a crucial step in plant microRNA biogenesis. Science. 307, 932-935.
- Zamore, P. D., Tuschl, T., Sharp, P. A., Bartel, D. P. (2000). RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell. 101, 25-33.
- Zhang, Y., Guo, R., Cui, Y., Zhu, Z., Zhang, Y., et al. (2017). An essential role for PNLDC1 in piRNA 3′ end trimming and male fertility in mice. Cell Res. 27, 1392-1396.
- Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E., Filipowicz, W. (2004). Single processing center models for human Dicer and bacterial RNase III. Cell. 118, 57-68.

Autorenbeiträge

RNA-based transposon regulation in eukaryotes (Kapitel 1)
Das Manuskript wurde von Daniel Gebert und David Rosenkranz gemeinsam verfasst. Die Recherche der Quellen bezüglich sRNA-Systemen in Tieren wurde von David Rosenkranz übernommen, während der Schwerpunkt bei der Recherche von Daniel Gebert bei sRNAs in Pflanzen lag. Das Erstellen aller Abbildungen erfolgte durch Daniel Gebert, teilweise mit Anregungen von David Rosenkranz, wobei das Schaubild 1 von David Rosenkranz konzipiert wurde.

Unitas: The universal tool for annotation of small RNAs (Kapitel 2)
Idee und Konzept für die Software stammen von David Rosenkranz. Anregungen wurden von Daniel Gebert eingebracht. Die Software wurde von David Rosenkranz und Daniel Gebert programmiert. Tests und Vergleiche mit anderen Programmen erfolgten durch Daniel Gebert, David Rosenkranz und Charlotte Hewel. Testdaten wurden von David Rosenkranz erzeugt. Die Abbildungen 1 und 5 wurden von David Rosenkranz, Abbildungen 3, 4 und 6 von Daniel Gebert und Abbildung 2 wurde gemeinsam von Daniel Gebert und David Rosenkranz erstellt. Das Manuskript wurde von David Rosenkranz und Daniel Gebert, mit Beiträgen von Charlotte Hewel, verfasst.

PIWIs and piRNAs in the germline and soma of mollusks (Kapitel 3)
Die Studie wurde von David Rosenkranz konzipiert. Gewebeentnahmen, sowie RNA-Extraktion und Vorbereitungen zur sRNA-Sequenzierung wurden von Frank Pipilescu und Sarah Stern durchgeführt. Daniel Gebert identifizierte PIWI-Gene und Haushaltsgene für qPCRs in nicht-annotierten Genomen. Desweiteren war Daniel Gebert für die phylogenetischen Analysen von PIWI-Proteinen der Mollusken zuständig. Primerdesign und qPCRs erfolgten durch Julia Jehn und Julian Kiefer. Die grundlegenden bioinformatischen Auswertungen der sRNA-Daten erfolgte durch David Rosenkranz. Charlotte Hewel, Frank Pipilescu and Sarah Stern führten die de novo miRNA-Annotation durch. Analysen der piRNA-Cluster und der Transposon-, sowie Genzusammensetzungen von Cluster-Loci und Genomen wurden von Daniel Gebert durchgeführt. Die Auswertungen bezüglich des Ping-Pong-Zyklus erfolgten durch David Rosenkranz. Abbildungen 1, 2 und 3 wurden von David Rosenkranz und Daniel Gebert, Abbildung 4 von David Rosenkranz erstellt. Das Manuskript wurde von David Rosenkranz, Daniel Gebert und Julia Jehn, mit Beiträgen von Frank Pipilescu und Charlotte Hewel, verfasst.

Regulation of protein-coding genes by piRNAs in the pig (Kapitel 4)
Die Studie wurde von David Rosenkranz konzipiert. Alle Experimente, sowie vorläufige Auswertungen der sRNA-Daten wurden durch Daniel Gebert während seiner Masterarbeit (2014) durchgeführt. Alle endgültigen Analysen, die Interpretation der Ergebnisse und die Erstellung der Abbildungen erfolgten durch Daniel Gebert. Das Manuskript wurde von Daniel Gebert verfasst und durch Beiträge von David Rosenkranz, René Ketting und Hans Zischler ergänzt.

Evolution of piRNA clusters and pseudogenes in primates (Kapitel 5)
Hans Zischler, David Rosenkranz und Daniel Gebert haben die Studie gemeinsam konzipiert. Die Auswertung der Daten, sowie das Programmieren der nicht-publizierten Analyse-Software wurden von Daniel Gebert durchgeführt. Die Interpretation der Ergebnisse und das Erstellen aller Abbildungen erfolgte durch Daniel Gebert. Das Manuskript wurde von Daniel Gebert verfasst.

Author contributions

RNA-based transposon regulation in eukaryotes (Chapter 1)
The manuscript was written by Daniel Gebert and David Rosenkranz. The research of sources on sRNA systems in animals was performed by David Rosenkranz, while Daniel Gebert focused on sRNA pathways in plants. All figures were prepared by Daniel Gebert with important contributions by David Rosenkranz. Figure 1 was designed by David Rosenkranz.

Unitas: The universal tool for annotation of small RNAs (Chapter 2)
David Rosenkranz developed the idea and design for the software, which include suggestions by Daniel Gebert. The tool was programmed by David Rosenkranz and Daniel Gebert. Tests and comparisons to other software tools were performed by Daniel Gebert, David Rosenkranz and Charlotte Hewel. Test data was generated by David Rosenkranz. Figures 1 and 5 were prepared by David Rosenkranz, figures 3, 4 and 6 by Daniel Gebert and figure 2 by David Rosenkranz and Daniel Gebert. The manuscript was written by David Rosenkranz and Daniel Gebert with contributions by Charlotte Hewel.

PIWIs and piRNAs in the germline and soma of mollusks (Chapter 3)
The study was designed by David Rosenkranz. Preparations for sRNA sequencing were performed by Frank Pipilescu and Sarah Stern. Daniel Gebert identified PIWI genes and housekeeping genes for qPCRs in unannotated genomes. Further, Daniel Gebert carried out the phylogenetic analyses of molluskan PIWI proteins. Primer design and qPCRs were performed by Julia Jehn and Julian Kiefer. Basic bioinformatic analyses of sRNA data were carried out by David Rosenkranz. Charlotte Hewel, Frank Pipilescu and Sarah Stern performed de novo miRNA annotation. Analyses of piRNA clusters and the transposon and gene content were conducted by Daniel Gebert. Investigations regarding the ping-pong cycle were undertaken by David Rosenkranz. Figures 1, 2 and 3 were prepared by David Rosenkranz and Daniel Gebert, figure 4 was made by David Rosenkranz. The manuscript was written by David Rosenkranz, Daniel Gebert and Julia Jehn, with contributions by Frank Pipilescu and Charlotte Hewel.

Regulation of protein-coding genes by piRNAs in the pig (Chapter 4)
The study was designed by David Rosenkranz. All experiments, as well as the preliminary analyses of sRNA data were conducted by Daniel Gebert during his Master thesis project (2014). All final analyses, interpretation of the results and preparation of the figures were carried out by Daniel Gebert. The manuscript was written by Daniel Gebert, with contributions by David Rosenkranz, René Ketting and Hans Zischler.

Evolution of piRNA clusters and pseudogenes in primates (Chapter 5)
Hans Zischler, David Rosenkranz and Daniel Gebert conceived the study. Daniel Gebert performed all analyses and coded the unpublished bioinformatics software. The interpretation of the results and the preparation of the figures was carried out by Daniel Gebert. Daniel Gebert wrote the manuscript.

Danksagung

Eidesstattliche Versicherung

Hiermit versichere ich, Daniel Gebert, dass ich die vorgelegte Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Ich habe oder hatte die jetzt als Dissertation vorgelegte Arbeit nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Auch habe oder hatte ich die vorgelegte Dissertation oder Teile der Arbeit nicht als Dissertation bei einer anderen Fakultät oder einem anderen Fachbereich eingereicht.

Mainz, 10.08.2018

_____

(Daniel Gebert)

Lebenslauf

Daniel Gebert
Hintere Talstraße 16
55130 Mainz
Telefon: 0049 151 42312615
E-Mail: gebert@uni-mainz.de
Staatsangehörigkeit: Deutsch

Bildungsgang

Johannes Gutenberg-Universität Mainz | Jul. 2014-Aug. 2018
Doktorand, Biologie (Dr. rer. nat.)
Institut für Organismische und Molekulare Evolutionsbiologie, Anthropologie
Stipendium: International PhD Programme, koordiniert vom Institut für Molekulare Biologie Mainz

Johannes Gutenberg-Universität Mainz | Okt. 2011-Aug. 2014
Master of Science, Biologie – Schwerpunkt auf Neurogenetik und Evolutionäre Anthropologie

University of Glasgow | Sep. 2012-Feb. 2013
Fächer: Molekulare Zellbiologie und Molekulargenetik

Johannes Gutenberg-Universität Mainz | Okt. 2008-Mär. 2012
Bachelor of Science, Biologie – Schwerpunkt auf Molekulargenetik

Gauß-Gymnasium Worms | 1999-2008
Abitur

Konferenzbeiträge

Präsentation: EMBO Workshop „Multiple functions of piRNAs and PIWI proteins“, Montpellier, April 2016
Poster: Annual Meeting of the Society for Molecular Biology & Evolution, Wien, Juli 2015

Publikationen

- Jehn J*, **Gebert D***, Pipilescu F*, Stern S, Kiefer JST, Hewel C, Rosenkranz D. PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation. *Commun Biol* 2018, *accepted*. (*equal contribution)
- **Gebert D**, Hewel C, Rosenkranz D. unitas: the universal tool for annotation of small RNAs. *BMC Genomics* 2017 18:644.
- Fast I, Hewel C, Wester L, Schumacher J, **Gebert D**, Zischler H, Berger C, Rosenkranz D. Temperature-responsive miRNAs in Drosophila orchestrate adaptation to different ambient temperatures. *RNA* 2017 23:1352-1364.
- **Gebert D**, Rosenkranz D. RNA-based regulation of transposon expression. *Wiley Interdiscip Rev RNA* 2015 6:687-708.
- **Gebert D**, Ketting RF, Zischler H, Rosenkranz D. piRNAs from pig testis provide evidence for a conserved role of the Piwi pathway in post-transcriptional gene regulation in mammals. *PLOS One* 2015 10:e0124860.

Curriculum vitae

Daniel Gebert
Hintere Talstraße 16
55130 Mainz
Phone: 0049 151 42312615
E-Mail: gebert@uni-mainz.de
Citizenship: German

Education

Johannes Gutenberg-Universität Mainz | Jul. 2014-Aug. 2018
Doctoral student, Biology (Dr. rer. nat.)
At: Institute of Organismic and Molecular Evolutionary Biology, Anthropology
Fellowship: International PhD Programme, coordinated by the Institute of Molecular Biology Mainz

Johannes Gutenberg-Universität Mainz | Oct. 2011-Aug. 2014
Master of Science, Biology – Focus on Neurogenetics and Evolutionary Anthropology

University of Glasgow | Sep. 2012-Feb. 2013
Subjects: Molecular Cell Biology and Molecular Genetics

Johannes Gutenberg-Universität Mainz | Oct. 2008-Mar. 2012
Bachelor of Science, Biology – Focus on Molecular Genetics

Gauß-Gymnasium Worms | 1999-2008
Abitur

Conference contributions

Presentation: EMBO Workshop „Multiple functions of piRNAs and PIWI proteins", Montpellier, April 2016
Poster: Annual Meeting of the Society for Molecular Biology & Evolution, Vienna, July 2015

Publications

- Jehn J*, **Gebert D***, Pipilescu F*, Stern S, Kiefer JST, Hewel C, Rosenkranz D. PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation. *Commun Biol* 2018, *accepted*. (*equal contribution)
- **Gebert D**, Hewel C, Rosenkranz D. unitas: the universal tool for annotation of small RNAs. *BMC Genomics* 2017 18:644.
- Fast I, Hewel C, Wester L, Schumacher J, **Gebert D**, Zischler H, Berger C, Rosenkranz D. Temperature-responsive miRNAs in Drosophila orchestrate adaptation to different ambient temperatures. *RNA* 2017 23:1352-1364.
- **Gebert D**, Rosenkranz D. RNA-based regulation of transposon expression. *Wiley Interdiscip Rev RNA* 2015 6:687-708.
- **Gebert D**, Ketting RF, Zischler H, Rosenkranz D. piRNAs from pig testis provide evidence for a conserved role of the Piwi pathway in post-transcriptional gene regulation in mammals. *PLOS One* 2015 10:e0124860.