Non-coding RNA fragments account for the majority of annotated piRNAs expressed in somatic non-gonadal tissues

Juan Pablo Tosar1,2, Carlos Rovira3 & Alfonso Cayota2,4

PIWI-interacting RNAs (piRNAs) are regarded as the guardians of the genome because they tackle genome stability-threatening transposable elements in the germline. Recently, piRNAs were also reported in other types of cells, including mouse brain, malignant and non-malignant somatic tissues, and human plasma. This suggests that piRNA function might be broader than previously expected. Here, we show that different piRNA databases contain a subset of sequences that correspond to piRNA-sized fragments of ncRNAs (rRNAs, tRNAs, YRNAs, snRNAs, and snoRNAs) and intermediates of miRNA biogenesis. We discuss that the biogenesis of these sequences is probably independent of the PIWI pathway, and can therefore be considered contaminants in piRNA databases. Although a minority of annotated piRNAs falls in this category, they account for the vast majority of piRNA expression in somatic non-gonadal tissues. Since ncRNA fragments are ubiquitous and abundant, their confusion with piRNAs strongly impacts the estimation of piRNA expression outside of mammalian gonads.
PIWI-interacting RNAs (piRNAs) are one of the three main classes of regulatory small RNAs, together with small interfering RNAs (siRNAs) and microRNAs (miRNAs). These classes differ in their biogenesis and mode of target regulation, but share some common features such as their ability to guide Argonaute proteins to target nucleic acids in a sequence-dependent manner. Argonaute proteins are phylogenetically subdivided into two subclasses, comprising the orthologs of Arabidopsis AGO1 and Drosophila Piwi (defining AGO and PIWI subfamilies, respectively). While the former are involved in post-transcriptional gene silencing by siRNAs and miRNAs, the biological function of PIWI proteins was initially unclear, although they were shown early on to be essential for germ cell maintenance. In 2006, various groups simultaneously reported that murine PIWI proteins MIWI and MILI bound a novel class of small (26–31 nt) RNAs in the testes, which they termed piRNAs. These piRNAs were encoded in discrete genomic clusters, many of which were present in syntenic genomic regions in humans.

In parallel, Drosophila PIWI proteins were shown to bind repeat-associated siRNAs, a germline-enriched 24–29 nt small RNA family previously known to be involved in transposon silencing, in the Drosophila ovary. This led to the notion that the conserved function of piRNAs is to target genome stability-threatening transposable elements in the germline. Mutations are particularly problematic when affecting germinal cells, and generalized demethylation of genomic DNA upon fertilization in mammals can unleash transposon expression and propagation. To avoid this, a piRNA-based innate immune system operates in the germline, comprising both genetically encoded (primary piRNAs derived from RNA pol II transcription from piRNA clusters) and adaptive (secondary piRNAs produced by ping-pong amplification) resistance mechanisms.

Given the involvement of the piRNA pathway in the germline, it is not surprising that piRNAs were initially cloned and sequenced in mouse testis or Drosophila ovaries. However, a role for PIWI proteins and piRNAs in somatic cells has also been documented. PIWI/piRNA expression was reported in larval salivary glands, in the central nervous system of mice and Aplysia. The role of piRNAs in tumors is also under study, since expression of PIWI-clade proteins was reported in many types of somatic cancer cells. Indeed, a recent analysis of transcriptomic data from The Cancer Genome Atlas identified a variety of somatic piRNAs, which can distinguish tumors from non-malignant tissues. Recently, more than a hundred piRNAs were sequenced in normal human plasma, and some of these were detected at high levels in every sequenced individual.

If piRNAs are expressed in non-germline tissues and are even transported in the bloodstream, one could wonder whether the germline model still stands as the unique environment to study piRNA biology or whether these non-germline piRNAs are true members of this gene family. Recently, sperm-derived RNA halves from tRNA-Gly were reported to control gene expression in early embryos. Interestingly, the sequences of these RNA halves are nearly identical, with only one nucleotide variation in sequence length, to annotated piRNAs (NCBI accession: DQ579796.1 and DQ579796.1).

This observation led us to study the degree of similarity between piRNAs and ncRNA fragments. To our surprise, we found that a considerable number of human sequences in distinct piRNA databases showed 100% identity to other ncRNAs, and that these ambiguous sequences accounted for the vast majority of the piRNAs described in the mouse brain, somatic cancer, and blood. Furthermore, these sequences do not share hallmarks of PIWI-dependent selection, such as a bias toward uridine at the 5’ end. We also show that the evidence for PIWI association to these ncRNA fragments is scarce in humans. Overall, we suggest that piRNA expression in mammalian non-gonadal cells is greatly overestimated or directly artificial, as reported non-gonadal piRNAs are probably not bona fide piRNAs.
Results

Overlap between piRNA databases and non-coding RNAs. RNAdb 2.0 and piRBase are two compendiums of piRNA sequences extracted from the scientific literature, and currently contain 171,551 and 32,826 human piRNAs, respectively. Analysis of these sequences showed a strong bias for uridine at the first position (1 T in our data set), in accordance with the preferential binding of PIWI proteins to transcripts starting with U (Fig. 1a). A bias toward adenine at position 10 (a hallmark of human piRNAs: (i) cloning and sequencing in human testis, (ii) immunoprecipitation of human PIWI proteins has prevented the identification of this subset as bona fide piRNAs.

Also, their size distribution is biased toward longer lengths (\(P = 0.008\); two-tailed t-test), with 26.6% of the sequences being equal to or higher than 30 nt, in contrast to 13.8% in the whole piRNA database (RNAdb 2.0). This is consistent with the slightly longer length of tRNA halves with respect to canonical piRNAs.

Table 1 Overlap between annotated piRNAs in piRNA databases (RNAdb 2.0 and piRBase) and non-coding RNAs

| Database | # piRNAs | rRNA | tRNA | miRNA | snRNA | snoRNA | YRNA | m_tRNA | m_rRNA | TOTAL | % |
|----------|----------|------|------|-------|-------|--------|------|--------|--------|-------|-----|
| MM = 0   | 171,551  | 111  | 132  | 18    | 5     | 97     | 2    | 13     | 14     | 392   | 0.23|
| RNAdb 2.0| 32,826   | 42   | 85   | 13    | 0     | 104    | 8    | 12     | 14     | 278   | 0.85|
| MM = 1   | 171,551  | 132  | 478  | 34    | 6     | 117    | 3    | 17     | 17     | 801   | 0.47|
| piRBase  | 32,826   | 47   | 156  | 0     | 0     | 117    | 15   | 14     | 17     | 385   | 1.18|

The analysis corresponds to every mouse piRNA presented in Table 1 of Lee et al15. piRNA accession numbers correspond to the NCBI database. Asterisks denote four piRNAs, which were described as belonging to a large piRNA cluster in chromosome 17 in the referenced report. Our alternative annotation is shown in column 3. Underlined bases correspond to mismatches according to our annotation.

Table 2 piRNAs present in the hippocampus of mouse brain are mostly ncRNA fragments

| piRNA | Reads | Alternative | Sequence | Reads: MIWI-IP |
|-------|-------|-------------|----------|---------------|
| DQ541777| 16,130 | RNY1 | GGCCTGTCCGAAGTAGTGAATCTC 1 |
| DQ705026| 6,257  | snoRNA 2 | CTGAAATAGAGAAGAATCTCTTGCTG AC |
| DQ555094| 3,439  | rRNA 28S | TGGGGGGGCAAATCTCTTTGAGGGGCA |
| DQ719597| 2,459  | snoRNA 17 | TTGCGATGTGGTGCTCCGGAGTTCTT |
| DQ689086| 1,514  | snoRNA | TCAGATGTGCTATTACTCTAA |
| DQ540285*| 1,433 | rRNA 185 | AGCCGGCGCCGAGTTAACACTCTCCT |
| DQ540981| 1,360  | rRNA 28S | CCGGGCCGCCGAGTTAACACTCTCCT |
| DQ720186| 849    | miR-3102-3p | AGGACACCCCATGTGACACCA |
| DQ55093| 775    | rRNA 28S | TGGGGGGGGCAATCTCTTTGAGGGGCA |
| DQ540862| 639    | snoRNA Z12 | CCGGGTGATGGAGTTAACACTCTCCT |
| DQ540284*| 635   | rRNA 185 | ATCGATGTGCTCTGAAGTTACCTT |
| DQ541506*| 580   | rRNA 18S | GATCGATGTGCTCTGAAGTTACCTT |
| DQ539915| 304    | CDS: MT_C01 | AACATTTCTCCGCTTTCAAGAATACCA |
| DQ540861| 252    | snoRNA 104 | CCGGGTGATGGAGTTAACACTCTCCT |
| DQ715526| 207    | snoRNA 17 | CACCAAGATGAGTGCGAAATCTGAT |
| DQ543676*| 182   | rRNA 185 | TCGATGTGCTCTGAAGTTACCTT |
| DQ722288| 175    | snoRNA D81 | TTAATTTGATGATGAGTTACCTT |
| DQ551351| 168    | CDS: Fth1 | TGCTATCATCATCATCATCATCAT |
| DQ550765| 118    | snRNA U12 | TGCGGGATGGCCGCGCCGAGTTACCTT |
| DQ708131| 115    | snoRNA 25 | TATCTGGTGGAGTTAACACTCTCCT |

The analysis corresponds to every mouse piRNA presented in Table 1 of Lee et al15. piRNA accession numbers correspond to the NCBI database. Asterisks denote four piRNAs, which were described as belonging to a large piRNA cluster in chromosome 17 in the referenced report. Our alternative annotation is shown in column 3. Underlined bases correspond to mismatches according to our annotation.

The vast majority of reported non-gonadal piRNAs in humans. Nevertheless, contaminating or at least ambiguous piRNAs account for less than 1% of the total number of sequences in the analyzed databases. This might be considered negligible, if it was not for the fact that this subset represents the vast majority of reported non-gonadal piRNAs in humans. Altogether, these observations question the classification of this subset as bona fide piRNAs.

An inspection of the data sources used to create these databases shows that, in contrast to flies and mice, 97% of the unique sequences contained in piRBase are derived from only one study. This study is one of the seminal reports in which piRNAs were first described5. In this study, three criteria were used to annotate human piRNAs: (i) cloning and sequencing in human testis, (ii) size in the 25–32 nt range, and (iii) lack of similarity to other known ncRNAs. Thus, piRNA annotation was based on sequencing of a size-selected small RNA library, without direct evidence of PIWI interaction and, as a consequence, human piRNA databases might contain PIWI pathway-independent RNAs. In other model organisms, RIP-seq and CLIP-seq data are available, but lack of highly specific antibodies for the immunoprecipitation of human PIWI proteins has prevented such studies in humans.

Reported somatic non-gonadal piRNAs are ncRNA fragments. One of the first reports on mammalian piRNA expression outside of the gonads described a subset of piRNAs expressed in mouse
Importantly, co-immunoprecipitation with the murine PIWI protein MIWI was confirmed. In situ hybridization in cultured neurons showed signal from one of these piRNAs in the dendritic compartment, and its antisense suppression suggested a role in dendritic spine morphogenesis. However, we found that all the most abundant piRNAs described in this study were also fragments of YRNAs, C/D box snoRNAs, rRNAs, and even miRNAs (Table 2). One may ask to what extent the biological effects observed upon LNA-based inhibition of the most abundant brain piRNA (DQ541777; mmu-piR-1889) could be caused by inhibition of full-length RNY1, which is the actual target of such oligonucleotides.

**Table 3** Human plasma piRNAs are mostly ncRNA fragments

| piRNA     | N (%) | RPM (mean) | Alternative | Start/type | Sequence                                      |
|-----------|-------|------------|-------------|------------|-----------------------------------------------|
| PIR54042  | 27    | 2,295.39   | RNAY4       | 67         | CCCCCACTGCTAATTTTGACTGCT                     |
| PIR2888   | 30    | 1,684.62   | RNAY1       | 1          | GGGTGTCCGAAGTGGTATGATGTTACTCC                |
| PIR58596  | 31    | 1,604.31   | MT_rRNA-165 | 1          | GCTAAACCTAGCCCAAAACCCACTCCAC                |
| PIR43376  | 32    | 327.78     | MT_rRNA-Val | 1          | CAGAGTGTGCTTTACAAACGACGACCCACCA            |
| PIR57581  | 39    | 312.29     | MT_rRNA-Ser | 1          | GAGAAGCTCAAAAGAATGCTAATCTAGT                |
| PIR54043  | 26    | 59.41      | RNAY4       | 67         | CCCCCACTGCTAATTGTTACTGCT                   |
| PIR59288  | 32    | 57.53      | rRNA-AlaCGC | 1          | GGGGTTGTAGCTAGTGAGAGCCGCGTG                 |
| PIR40304  | 31    | 44.88      | MT_rRNA-His | 32         | TAACTGTCACAAAGACGCTTACGCTACCCCCCC          |
| PIR41574  | 31    | 41.07      | piRNA-Cluster Chr5 | 1 | TGGATGCGCGAGATGCTCCGCCACACACCC         |

Mean abundance (RPM) and number of patients (N = 40) in which each piRNA was sequenced were extracted from Freedman et al. The analysis includes every piRNA in the cited study with an abundance ≥ 10 RPM. piRNA accession numbers correspond to the RNAdb 2.0 database. Our alternative annotation is shown in columns 4-5. Underlined bases correspond to mismatches according to our annotation.

**Table 4** The top 20 (most abundant) piRNAs found in human tumors, after analysis of The Cancer Genome Atlas, are miRNAs or ncRNA fragments

| piRNA     | ∑ RPM cancer | Alternative | Sequence                                      |
|-----------|--------------|-------------|-----------------------------------------------|
| FR072386  | 38,911,332   | mir-let-7a-1 | TGGAGTGTAGTTGTATAGTTATTTAGGGTTC             |
| FR182987  | 3,814,379    | mir-532-5p  | CATGCCTTGGATGAGGACGT                      |
| FR074386  | 1,093,572    | snoRNA 98   | AGTGAGTGGAGGACAGTACGTGA                    |
| FR140858  | 1,177,195    | mir-106b    | TAAAGTGCTGACAGTGTGAGTTATAAGG               |
| FR133872  | 33 (82)      | piRNA       | TCCAGTGCCAGAATGCTGCTGAGTGTGTTCACTG         |
| FR31112   | 34 (85)      | MT_rRNA-Met | 3’ END AACGGTGTTATTACCCCTGTACATTAC         |
| FR49916   | 31 (78)      | piRNA       | TGGGAAGTGAATCGTGTTAGTGAATGCTA             |
| FR95795   | 31 (75)      | MT_rRNA-Phe | 3’ END GACAAACAGGCGGCGTATGCTCCA            |
| FR37665   | 23 (15)      | piRNA       | TCCGTTATGCGAGTGTGAGGG                    |
| FR52755   | 35 (88)      | CNSPATA3D1  | Sense, exon                                |
| FR55478   | 18 (45)      | CNSC6orf99  | Sense, exon                                |
| FR33872   | 33 (82)      | CNSVORC1L1  | Sense, exon                                |
| FR31112   | 34 (85)      | MT_rRNA-Met | 3’ END AAGTGTGTTATTACCCCTGTACATTAC         |
| FR49916   | 31 (78)      | piRNA       | TGGGAAGTGAATCGTGTTAGTGAATGCTA             |
| FR95795   | 31 (75)      | MT_rRNA-Leu | 1                                         |
| FR59421   | 27 (68)      | rRNA-28S    | 4549                                      |
| FR51124   | 20 (50)      | MT_rRNA-Glu | 41                                        |
| FR1340    | 31 (33)      | MT_rRNA-Met | 5                                         |

piRNA accession numbers correspond to the Functional RNA database (fRNAdb), as reported by Martinez et al. Our alternative annotation is shown in column 3. Underlined bases correspond to mismatches according to our annotation.

**Table 4** The top 20 (most abundant) piRNAs found in human tumors, after analysis of The Cancer Genome Atlas, are miRNAs or ncRNA fragments

| piRNA     | ∑ RPM cancer | Alternative | Sequence                                      |
|-----------|--------------|-------------|-----------------------------------------------|
| FR072386  | 38,911,332   | mir-let-7a-1 | TGGAGTGTAGTTGTATAGTTATTTAGGGTTC             |
| FR182987  | 3,814,379    | mir-532-5p  | CATGCCTTGGATGAGGACGT                      |
| FR074386  | 1,093,572    | snoRNA 98   | AGTGAGTGGAGGACAGTACGTGA                    |
| FR140858  | 1,177,195    | mir-106b    | TAAAGTGCTGACAGTGTGAGTTATAAGG               |
| FR133872  | 33 (82)      | piRNA       | TCCAGTGCCAGAATGCTGCTGAGTGTGTTCACTG         |
| FR31112   | 34 (85)      | MT_rRNA-Met | 3’ END AACGGTGTTATTACCCCTGTACATTAC         |
| FR49916   | 31 (78)      | piRNA       | TGGGAAGTGAATCGTGTTAGTGAATGCTA             |
| FR95795   | 31 (75)      | MT_rRNA-Phe | 3’ END GACAAACAGGCGGCGTATGCTCCA            |
| FR37665   | 23 (15)      | piRNA       | TCCGTTATGCGAGTGTGAGGG                    |
| FR52755   | 35 (88)      | CNSPATA3D1  | Sense, exon                                |
| FR55478   | 18 (45)      | CNSC6orf99  | Sense, exon                                |
| FR33872   | 33 (82)      | CNSVORC1L1  | Sense, exon                                |
| FR31112   | 34 (85)      | MT_rRNA-Met | 3’ END AAGTGTGTTATTACCCCTGTACATTAC         |
| FR49916   | 31 (78)      | piRNA       | TGGGAAGTGAATCGTGTTAGTGAATGCTA             |
| FR95795   | 31 (75)      | MT_rRNA-Leu | 1                                         |
| FR59421   | 27 (68)      | rRNA-28S    | 4549                                      |
| FR51124   | 20 (50)      | MT_rRNA-Glu | 41                                        |
| FR1340    | 31 (33)      | MT_rRNA-Met | 5                                         |

piRNA accession numbers correspond to the Functional RNA database (fRNAdb), as reported by Martinez et al. Our alternative annotation is shown in column 3. Underlined bases correspond to mismatches according to our annotation.

***hippocampus***15. Importantly, co-immunoprecipitation with the murine PIWI protein MIWI was confirmed. In situ hybridization in cultured neurons showed signal from one of these piRNAs in the dendritic compartment, and its antisense suppression suggested a role in dendritic spine morphogenesis. However, we found that all the most abundant piRNAs described in this study were also fragments of YRNAs, C/D box snoRNAs, rRNAs, and even miRNAs (Table 2). One may ask to what extent the biological effects observed upon LNA-based inhibition of the most abundant brain piRNA (DQ541777; mmu-piR-1889) could be caused by inhibition of full-length RNY1, which is the actual target of such oligonucleotides.
A recent survey of circulating small RNAs present in human blood plasma revealed a total of 144 piRNAs, many of which were abundant and detectable in every sequenced individual. We took a closer look at these piRNAs, and found that 100% with average expression $\geq 50$ RPM and 68% with a 10 RPM cutoff were in fact ncRNA fragments (Table 3). To better illustrate this point, we divided piRNA reads in RNAdb2.0 (which was the database used in this study) as those starting or not starting with a uridine (148,557 vs. 22,994 sequences, respectively) or those mapping or not mapped to ncRNAs (392 vs. 171,159 sequences, respectively).

As a positive control, we analyzed piRNAs in normal human testis. As expected, the distribution of piRNAs in testis matched the distribution of piRNAs in the database (Fig. 1b). This is evident either by using the number of sequences (Fig. 1b, left and middle) or by considering their expression (Fig. 1b, right). In contrast, piRNAs in human plasma were highly enriched in sequences not starting with uridine (ES: enrichment score $> 5$) and, more strikingly, sequences mapping to ncRNAs (ES $> 150$ in number of sequences, and ES $> 900$ when considering relative abundances).

We still considered the possibility that these ncRNA fragments could be genuine piRNAs secreted to the bloodstream. Of note, the most abundant sequences were fragments of YRNAs (none of them starting with uridine), which could also be detected when analyzing independent plasma sequencing studies (NCBI small read archive: SRR2496797). For instance, the top-ranked plasma piRNA (PIR54042 27; hsa-piR-33043) is a sequence derived from the 3' end of RNY4 (Table 3). Importantly, sequences mapping to the 5' end of the same precursor were also detectable in plasma, despite the fact that no piRNAs are annotated in this region (Fig. 1c). This profile is similar in samples from different tissues, but variations in the length and extremes of RNY4 3' fragments are typically not consistent with PIWI-dependent processing, as they resemble a pattern of exonucleolytic processing from their 5' end, incompatible with current models of piRNA biogenesis.

More probably, fragments of YRNAs in these data sets reflect the most stable and sequencing-prone degradation intermediates of full-length YRNAs, which can be secreted to the extracellular space. Alternatively, 5' and 3' fragments from RNY4 could be a result of processing by Dicer, since they correspond to a pair of complementary sequences with protruding ends and a double-stranded core of exactly 21 nucleotides (Supplementary Fig. 1). This reinforces the view that the biogenesis of the piRNAs under study is probably PIWI pathway-independent.

We also analyzed another recent study describing piRNAs in human cancer cells. Here, the authors analyzed transcriptomic data from more than 500 normal tissues and over 5,000 tumor samples from The Cancer Genome Atlas, and discovered 273 and 522 somatic non-malignant and malignant piRNAs, respectively. However, our reanalysis showed that the most abundant piRNAs in cancer cells were either miRNA pathway by-products or ncRNA fragments (Table 4). The top-ranked piRNAs corresponded to miRNAs let-7a-5p and miR-532-5p, which is not expected by chance ($P = 8 \times 10^{-9}$) since the overlap between piRNA databases and miRBase is rather small (Table 1). The rest of the sequences were mostly fragments of C/D box snoRNAs, mitochondrial tRNAs, and rRNAs. Even though we were surprised to see sequences $\geq 30$ nucleotides corresponding to miRNAs (which are typically 22–23 nt), these showed 100% identity with the pre-miRNAs across the entire sequence length, reinforcing their missclassification as piRNAs.
Discussion

The fact that most extragonadal piRNAs in mouse and humans belong to an ambiguous piRNA subset suggests that piRNA expression outside of the gonads is infrequent in mammals. While not excluding the possibility of active piRNA pathways in non-gonadal tissues, detection of somatic bona fide piRNAs might be affected by a subset of highly abundant ncRNA fragments (and even miRNAs), which are reported as piRNAs.

The question is whether the classification of certain ncRNA fragments as piRNAs is erroneous or not. The answer directly impacts on the likelihood of piRNA expression outside of mammalian gonads. The detection of piRNAs circulating in blood plasma is particularly interesting, and could have an impact on liquid biopsy-based diagnosis. To answer this question, it would be necessary to stress the criteria used for piRNA definition. Accepted properties of piRNAs are their length (24–32 nt), bias toward uridine at 5′, 2′-o-methylation of their 3′ end, and clustering of their coding sequences in the genome. According to their biogenesis, piRNAs can be further classified as genome-encoded primary piRNAs, ping-pong generated secondary piRNAs, and even phased tertiary piRNAs. However, strictly speaking, piRNAs are the small RNAs physically bound and functionally related to PIWI proteins. Thus, a piRNA might not satisfy any of the previous characteristics, but still be a piRNA if capable of specific interaction in the 5′ binding pocket of a PIWI-clade protein. Now the question is, can we affirm that for every sequence deposited in piRNA databases?

At least in humans and mouse, most piRNA databases have grown based on data from the articles which described piRNAs for the first time. Although some of these papers relied on RIP-seq for piRNA identification, others annotated candidate piRNAs based on their sequencing abundance in testis, size in the 25–32 nt range, and lack of similarity to other known ncRNAs. Furthermore, as there is still a lack of suitable antibodies for selective immunoprecipitation of human PIWI proteins, human piRNAs were entirely cataloged from size-selected sequencing of gonad RNA rather than RIP-seq studies. So, the first conclusion is that direct evidence of PIWI interaction is not available for every sequence present in piRNA databases, especially in humans.

Nevertheless, it is still possible that some of the ncRNA fragments present in size-selected small RNA libraries of human testis could be bound to PIWI-clade proteins. If that was the case, they should be regarded as piRNAs. But can this be extrapolated to other tissues? For instance, if a tRNA-derived fragment interacts with a PIWI protein in the gonads, but is also abundant in a tissue where PIWI proteins are not expressed, would it still be a piRNA in both cases? One disadvantage of using a biogenesis-independent definition of piRNAs is that it makes piRNAs a context-dependent attribute, rather than an intrinsic property of a sequence. Thus, a tRNA fragment should be considered a piRNA if it interacts with PIWI, but the same sequence should not be considered a piRNA in other contexts, when their existence is unrelated to the PIWI pathway. But this is omitted when mapping somatic small RNA sequencing data to piRNA databases, which actually contain a compendium of small RNAs in the gonads.

The third and more complex issue is that PIWI co-immunoprecipitation should be a necessary but not sufficient condition to claim the presence of bona fide piRNAs. Although we showed that the most abundant piRNAs in mouse hippocampus were ncRNA fragments, we should recognize that the authors did evaluate the presence of these piRNAs in MIWI-IP. Furthermore, tRNA fragments were co-IP with anti-flag antibodies after expressing a flagged version of the human PIWI protein Hiwi2 in a breast cancer somatic cell line. However, this result should be interpreted carefully, as the authors found a very strong correlation between the tRNA fragments found in Hiwi2-IP and the whole-cell extracts. We would have expected some degree of selection for specific tRNA fragments (such as those starting with uridine, for example).

Relying on PIWI-IP for defining piRNAs can be problematic. In the first place, very specific antibodies are needed. We have found a number of miRNAs after analyzing data from MILI-IP coming from 10 days post-partum (dpp) mouse testis, suggesting a possible contamination with AGO-clade bound RNAs. Second, there will usually be a background of RNA fragments stuck to the surface of a PIWI/piRNA complex in any immunoprecipitate, with the contaminants not being truly engaged with the PIWI protein in a biologically meaningful manner. Abundant intracellular RNAs of a similar size (e.g., ncRNA fragments) are risky. By analyzing data from prioritized KO mice, we have observed that the tRNA fragments that are abundant in MILI-IP do not rely on MILI for neither their biogenesis nor their intracellular stability (Supplementary Fig. 2). Importantly, at 10-dpp MILI is the only PIWI-clade protein expressed in mouse testis, discarding association of tRNA fragments with other PIWI-clade proteins. In contrast, transposable element-targeting piRNAs were decreased as expected in MILI KO mice. In our opinion, this distinguishes bona fide piRNAs from frequent contaminants in the piRNA size range.

Overall, we have identified that a subset of ncRNA fragments and miRNAs contaminate most human piRNA databases, and that even though the amount of dubious piRNAs is rather low (usually below 1% of the total), this can be problematic when studying somatic piRNA expression. In these types of studies, we strongly encourage a deep analysis of the hits obtained after mapping to a piRNA database, paying particular attention to other possible hits in the genome. We have noted that most of the problematic or ambiguous piRNAs described herein are not included in the piRNA cluster database. This is remarkable, as this database uses small RNA deep-sequencing data as an input, but then uses the genomic coordinates, length distribution, and positional nucleotide composition of mapped reads to define putative piRNA clusters. Thus, many problematic or ambiguous piRNAs are removed when applying more stringent criteria for piRNA definition, such as genomic context and localization. Nevertheless, it is still important that most “somatic piRNAs” map multiple times in the genome (as a consequence of their sequence identity to tRNAs, rRNAs, and YRNAs) and can show some degree of clustering due to the genomic arrangement of the genes encoding these ncRNAs, or because they map the same ncRNA gene at different positions. An example is the four mouse piRNAs reported to cluster in chromosome 17 (Table 2). Here, the putative cluster is a consequence of the sequences aligning to the same gene (18S rDNA).

It would also be worthwhile to extend these considerations to the miRNA field, although bona fide miRNAs are easier to distinguish based on characteristic sequence patterns, which should correspond to reasonable hairpin precursors. Furthermore, miRBase routinely checks and filters submissions for fragments of rRNAs and tRNAs. Consequently, the overlap between miRBase and ncRNAs was much narrower than in the case of piRNA databases. In a more general view, we would like to argue that only mapping sequencing data to a given reference (e.g., a piRNA database) should not be considered sufficient proof to claim the expression of a given RNA family, especially when the classification of mapped sequences is ambiguous. In the miRNA field, curated databases with more stringent inclusion criteria (e.g., MirGeneDB) have served to overcome problems arising from the many false positives present in primary repositories. Analogously, the curation of piRNA databases will enable the study of hypothetical piRNAs
expression outside of mammalian gonads without the interference of piRNA-sized ncRNA fragments.

Methods

Bioinformatic analysis. To study the overlap between piRNA databases and ncRNAs, Fastq files containing the complete list of human piRNAs were downloaded from either RNAdb 2.031 or piRBase36, and mapped to ad hoc references containing human genomic and mitochondrial RNAs (downloaded from NCBI), tRNAs (downloaded from the Genomic RNA Database, GRInADB, and the mitochondria tRNA database, mitoRNADB), small nuclear RNAs, small nuclear RNAs_YRNAs (all downloaded from NCBI), and miRNAs (downloaded from miRBase). Mapping was performed with the Lastz program contained in the Galaxy Project package, using a seed hit of 19 bp, and returning alignments that covered at least 94 % of the length of each sequence, and showed at least 94 % sequence identity (i.e., a maximum of one mismatch for sequences less than 31 nt).

Data from piRNAs-containing piRNA clusters were first draft of the manuscript. A.C. and R. discussed results, proposed new experiments/analysis, and made major contributions to the submitted version of the manuscript.

Acknowledgements

J.P.T. and A.C. are members of the National System of Researchers (ANII, Uruguay) and The Program for the Development of Basic Sciences (PEDECIBA, Uruguay).

Author contributions

J.P.T. conceived the work, performed bioinformatic analysis, and wrote the first draft of the manuscript.

References

1. Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. Trends Biochem. Sci. 31, 5488, https://doi.org/10.1038/ncomms6488 (2014).
2. Czech, B. & Hannon, G. J. A happy 3′-end to the piRNA maturation story. Cell 164, 838–840 (2016).
3. Aravin, A. A. et al. Developmentally regulated piRNA clusters implicate Mili in transposon control. Science 316, 744–747 (2007).
4. 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).
5. Ross, R. I., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. Nature 505, 353–359 (2014).
6. Grivna, S. T. et al. A novel class of small RNAs in mouse spermatogenic cells. Proc. Natl. Acad. Sci. U.S.A. 104, 324–329 (2007).
7. 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).
8. Leseq, Q. et al. MicroRNA and piRNA profiles in normal human testis detected by next generation sequencing. PLoS ONE 8, e66809, https://doi.org/10.1371/journal.pone.0066809 (2013).
9. Chen, Q. et al. Sperm sRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science 351, 397–400 (2016).
10. Sharma, U. et al. Biogenesis and function of RNA fragments during sperm maturation and fertilization in mammals. Science 351, 391–396 (2016).
11. Aravin, A. A. et al. Unique somatic and malignant expression patterns implicate PIWI-interacting RNAs in cancer-type specific biology. Sci. Rep. 5, 10423, https://doi.org/10.1038/srep10423 (2015).
12. Freedman, J. E. et al. Diverse human extracellular RNAs are widely detected in human plasma. Nat. Commun. 7, 11106, https://doi.org/10.1038/ncomms11106 (2016).
13. Gore, B. et al. The pathology of extracellular RNA in cancer. Nat. Rev. Cancer 16, 273–286 (2016).
14. Fromm, B. et al. A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. Nucleic Acids Res. 40, D115–D121, https://doi.org/10.1093/nar/gkv1082 (2012).
15. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating and confident microRNAs using deep sequencing data. Nucleic Acids Res. 42, D68–D73, https://doi.org/10.1093/nar/gkt1181 (2014).
16. Fromm, B. et al. A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. Annu. Rev. Genet. 49, 213–242 (2015).
17. Zhang, P. et al. piBase: a web resource assisting piRNA functional study. Database 2014, bau110, https://doi.org/10.1093/database/bau111 (2014).
