


ARTICLE

DOI: 10.1038/s42003-017-0001-7

OPEN

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

Juan Pablo Tosar ^{1,2}, Carlos Rovira³ & Alfonso Cayota^{2,4}

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.

¹Nuclear Research Center, Faculty of Science, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay. ²Functional Genomics Unit, Institut Pasteur de Montevideo, Matajojo 2020, Montevideo 11400, Uruguay. ³Division of Oncology, Department of Clinical Sciences, Lund University Cancer Center, Lund 223 81, Sweden. ⁴Department of Medicine, Faculty of Medicine, Universidad de la República, Av. Italia s/n, Montevideo 11600, Uruguay. Correspondence and requests for materials should be addressed to J.P.T. (email: jptosar@pasteur.edu.uy)

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 manner¹. Argonaute proteins are phylogenetically subdivided into two subclasses, comprising the orthologs of *Arabidopsis* AGO1 and *Drosophila* Piwi (defining AGO and PIWI subfamilies, respectively)². 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^{3, 4}. In 2006, various groups simultaneously reported that murine PIWI proteins MIWI^{5, 6} and MILI⁷ 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 silencing⁸, in the *Drosophila* ovary. This led to the notion that the conserved function of piRNAs is to tackle genome stability-threatening transposable elements in the germline^{9–11}. 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 mechanisms¹².

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 documented¹³. PIWI/piRNA expression was reported in larval salivary glands¹⁴, in the central nervous system of mice¹⁵ and *Aplysia*¹⁶. The role of piRNAs in tumors is also under study, since expression of PIWI-clade proteins was reported in many types of somatic cancer cells¹³. 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 tissues¹⁷. Recently, more than a hundred piRNAs were sequenced in normal human plasma, and some of these were detected at high levels in every sequenced individual¹⁸.

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 tRNA halves from tRNA-Gly were reported to control gene expression in early embryos^{19, 20}. Interestingly, the sequences of these tRNA halves are nearly identical, with only one nucleotide variation in sequence length, to annotated piRNAs (NCBI accession: DQ597916.1 and DQ570956.1).

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 brain¹⁵, somatic cancer¹⁷, and blood¹⁸. 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 artifactual, as reported non-gonadal piRNAs are probably not bona fide piRNAs.

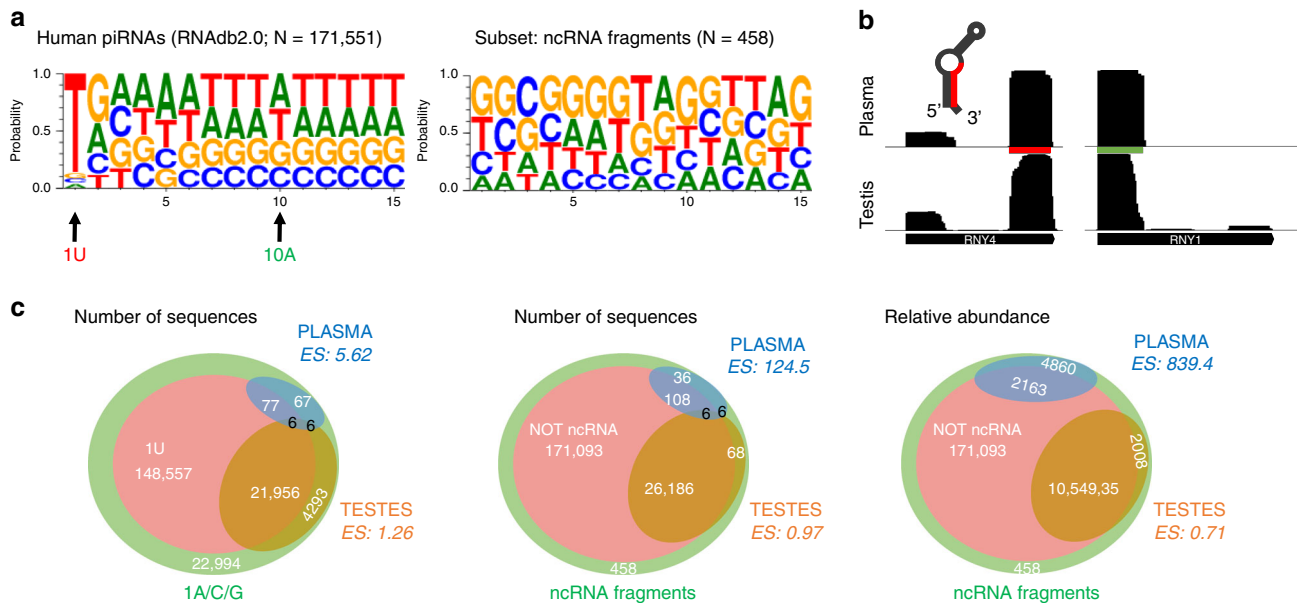


Fig. 1 piRNAs in human plasma are non-coding RNA fragments. **a** Sequence logo showing 1U bias in piRNA sequences in the RNAb 2.0 database, and absence of such signature in the subset of sequences matching other known ncRNAs. For simplicity, only the first 15 bases are shown. **b** Read coverage of YRNAs (RNY4 and RNY1) in human plasma (SRR2496797) and human testis (ERA246774). The red and green bars represent the most and second-most abundant piRNAs found in human plasma in Freedman *et al.*¹⁸. The predicted secondary structure of RNY4 is shown, with the RNY4-derived piRNA highlighted in red. **c** Venn diagrams showing the 171,551 piRNAs in RNAb 2.0, classified by either their 5' start base or their identity to other ncRNAs. The piRNAs described in human plasma (blue) and human testis (orange) are overimposed, and enrichment scores (ES) were calculated as described in methods

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											
RNAdb 2.0	171,551	111	132	18	5	97	2	13	14	392	0.23
piRBase	32,826	42	85	13	0	104	8	12	14	278	0.85
MM = 1											
RNAdb 2.0	171,551	132	478	34	6	114	3	17	17	801	0.47
piRBase	32,826	47	156	19	0	117	15	14	17	385	1.18

MM mismatch allowance, m_tRNA mitochondrial tRNAs, m_rRNA mitochondrial rRNAs

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	GGCTGGTCCGAAGGTAGTGAGTTATCTCAA	1
DQ705026	6,257	snoRNA 2	CTGAAATGAAGAGAATACTCTTGCTGATC	0
DQ555094	3,439	rRNA 28S	TGGGGGGCCCAAGTCTTCTGATCGAGGCCCA	0
DQ719597	2,459	snoRNA 85	GGTCGATGATGAGAGCTTTGTTCTGAGC	0
DQ689086	1,514	snoRNA 27	TGCAATGATGTCATCTTACTACTGAAA	0
DQ540285*	1,433	rRNA 18S	ATCGATGTGGTGCTCCGGAGTTCTTTCGGGC	0
DQ540981	1,360	rRNA 28S	CGGGCCCGCGGTGAAATACCACTACTCTCA	0
DQ720186	849	miR-3102-3p	AGAGCACCCCATTTGGCTACCCAC	0
DQ555093	775	rRNA 28S	TGGGGGGCCCAAGTCTTCTGATCGAGGC	1
DQ540862	639	snoRNA Z12	CCGGGTGATGCGAATCGTAATCTGAGCCGA	0
DQ540284*	635	rRNA 18S	ATCGATGTGGTGCTCCGGAGTTCTTTCGGG	0
DQ541506*	580	rRNA 18S	GATCGATGTGGTGCTCCGGAGTTCTTCT	0
DQ539915	304	CDS: MT_CO1	AACATTTCTGGCCCTTCAGGAATACCA	0
DQ540861	252	snoRNA 104	CCGGGTGATGCGAATCGTAATCTGAGC	0
DQ715526	207	snoRNA 17	CACCAAGATGAGTGGTGCAAATCTGATC	0
DQ543676*	182	rRNA 18S	TCGATGTGGTGCTCCGGAGTTCTTTCGGGC	0
DQ722288	175	snoRNA D81	TTACTTGATGATAGTAAAAGATCTGATG	0
DQ551351	168	CDS: Fth1	TGCTTCAACAGTGCTTGAACGGAACCCGGT	1
DQ550765	118	snRNA U12	TGCGGGATGCCTGGGTGACGCGATCTGCCG	0
DQ708131	115	snoRNA 25	TATCTGTGAGGATAAGTAACTCTGAGG	0

The analysis corresponds to every mouse piRNA presented in Table 1 of Lee *et al.*¹⁵. 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. Column 5 shows the number of reads corresponding to each piRNA in a Miwi RIP-seq study performed in wild-type adult testis (GEO: GSM822760)

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 secondary piRNAs generated by the ping-pong cycle) was also evident, albeit with a much weaker signal. To study the overlap between the piRNA sequences contained in these databases and other non-coding RNAs, we aligned each sequence against genomic or mitochondria-encoded tRNAs, rRNAs, snRNAs, snoRNAs, YRNAs, and miRNAs. With a zero mismatch allowance, we found 392 (RNAdb 2.0) and 278 (piRBase) piRNAs whose classification as either piRNAs or ncRNA fragments was ambiguous (Table 1 and Supplementary Data 1). These represent 0.23 or 0.85% of the total number of sequences included in each database, respectively. The subset of ambiguous piRNAs shows no preference for 5' uridine or adenine at position 10 (Fig. 1a). 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.

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. 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.

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

Table 3 human plasma piRNAs are mostly ncRNA fragments

piRNA	N (%)	RPM (mean)	Alternative	Start/type	Sequence
PIR54042 27	40 (100)	2,295.39	RNAY4	67	CCCCCACTGCTAAATTTGACTGGCT
PIR2888 30	40 (100)	1,684.62	RNAY1	1	GGCTGGTCCGAAGGTAGTGAGTTATCTCA
PIR58596 31	40 (100)	1,604.31	MT_rRNA 16S	1	GCTAAACCTAGCCCCAAACCCACTCCACCC
PIR43376 32	40 (100)	327.78	MT_tRNA-Val	1	CAGAGTGTAGCTTAACACAAAGCACCCAAC
PIR57581 31	39 (98)	312.29	MT_tRNA-Ser	1	GAGAAAGCTCACAAGAAGCTGTAACCTCATG
PIR54043 26	40 (100)	59.41	RNAY4	67	CCCCCACTGCTAAATTTGACTGGT
PIR59288 32	33 (82)	57.53	tRNA-AlaCGC	1	GGGGGGTGTAGCTCAGTGGTAGAGCGCGTGC
PIR40304 32	40 (100)	44.88	MT_tRNA-His	32	TGAATCTGACAACAGAGGCTTACGACCCCTT
PIR41574 31	40 (100)	41.07	piRNA	Cluster Chr5	TGAGATGCGGGAGCTCCGGCGCACACACTC
PIR227919 21	40 (100)	34.69	MT_tRNA-Met	1	AGTAAGGTCAGCTAAITTAAG
PIR75448 31	40 (100)	33.04	tRNA-IleAAT	1	GGCCGGTTAGCTCAGTAGGTTAGAGCTTGG
PIR45809 31	34 (85)	28.77	MT_tRNA-Ser	29	TGCCCCCATGTCTAACAAACATGGCTTCTC
PIR57849 32	20 (50)	26.67	CDS: MT_CO2	Antisense	GAGGGCGTGATCATGAAAGGTGATAAGCTCT
PIR57322 27	36 (90)	20.9	CDS: PPP1R3E	Sense, exon	GACAACAACGGCGGCCGTGACTATGC
PIR59786 31	30 (75)	18.06	MT_tRNA-Phe	3' END	GTTTAGACGGGCTCACATCACCCCATAAAC
PIR37665 28	23 (58)	17.91	piRNA	Cluster Chr11	TCCTGTATTTGCCGAATTGGTGTTT
PIR52755 30	35 (88)	17.16	CDS: SPATA31D1	Sense, exon	TGTGCAGAAATTTGGTGCAGTTATAAGAG
PIR55478 30	18 (45)	17.05	CDS: C6orf89	Sense, exon	TTCCAGTGCCGAAGACATTGTGCTGCTGT
PIR33872 31	33 (82)	16.74	CDS: VKORC1L1	Sense, exon	TCAAGGCTAAATCTGCTCATGTCGCCACTG
PIR31112 30	34 (85)	15.99	MT_tRNA-Met	3' END	AAATGTTGGTTATACCTTCCCGTACTAC
PIR49916 28	31 (78)	15.85	piRNA	Cluster Chr6	TGGGAGTGAAATCAGTGTTTAGGACTA
PIR59752 31	33 (82)	13.49	MT_tRNA-Leu	1	GTTAAGATGGCAGAGCCCGGTAATCGCATA
PIR59421 32	27 (68)	12.25	rRNA_28S	4549	GGTTAGTTTTACCCTACTGATGATGTTGT
PIR51124 27	20 (50)	11.73	MT_tRNA-Glu	41	TGGTCGTGGTTGTAGTCCGTGCGAGA
PIR1340 31	33 (82)	10.51	MT_tRNA-Met	5	AGGTCAGCTAAITTAAGCTATCGGGCCATA

Mean abundance (RPM) and number of patients (N=40) in which each piRNA was sequenced were extracted from Freedman *et al.*¹⁸. 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	Σ RPM cancer	Alternative	Sequence
FR072386	38,911,332	miR-let-7a-1	TGAGGTAGTAGGTTGTATAGTTTTAGGGTC
FR182987	3,814,379	miR-532-5p	CATGCCCTTGAGTGTAGGACCGT
FR074386	1,093,572	snoRNA 98	ATGCAGTGTGGAACACAATGAACTGAAC
FR140858	1,177,195	miR-106b	TAAAGTGTGACAGTGCAGATAGTGGTCCCTC
FR114004	823,121	snoRNA 1B	TTTCTGTGTGGAATTTGAATATCTGAAA
FR075316	744,575	snoRNA 82	ACCTGATGTTACATTGTAGTGTGCTGATG
FR132879	635,984	snoRNA 58A	CTGCAGTGTGACTTTCTTAGGACACCTTTG
FR064000	407,046	snoRNA 58B	CTGCGATGATGGCATTCTTAGGACACCTTTG
FR163199	301,387	snoRNA 138	CATGATACTGTAACCGCTTTCTGATG
FR043670	268,790	MT_tRNA-Glu	TGGTCGTGGTTGTAGTCCGTGCGAGAA
FR091055	185,930	miR-744-5p	TGCGGGGCTAGGGCTAACGCA
FR190827	201,282	snoRNA 6	AGGGGCTGAATGAAAATGGCTTTCTGAAC
FR090905	168,436	MT_rRNA 16S	GCTAAACCTAGCCCCAAACCCACTCCA
FR016773	172,895	snoRNA 42B	ACTTGTGATGTCTTCAAAGGAACCACTGATG
FR136216	162,693	snoRNA 62A	GGGAGATGAAGAGGACAGTGTAGTGAAGAC
FR004819	123,249	snoRNA 114-1	GACGGTGAATACAGGCTGGAAGTCTGAGGT
FR103462	96,101	snoRNA 98	AAATGCAGTGTGGAACACAATGAACTGAAC
FR089006	95,184	tRNA-Gly	AGCGCCGCTGGTGTAGTGGTATCATGCAAG
FR026913	74,640	snoRNA 89	GAGGAATGATGACAAGAAAAGGCCGAA
FR019019	81,678	snoRNA 107	GTTTATGATGACACAGGACCTTGTCTGAAC

piRNA accession numbers correspond to the Functional RNA database (fRNAdb), as reported by Martinez *et al.*¹⁷. Our alternative annotation is shown in column 3. Underlined bases correspond to mismatches according to our annotation

hippocampus¹⁵. 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 piRNA¹⁵ (DQ541777; mmu-piR-1889) could be caused by inhibition of full-length RNY1, which is the actual target of such oligonucleotides.

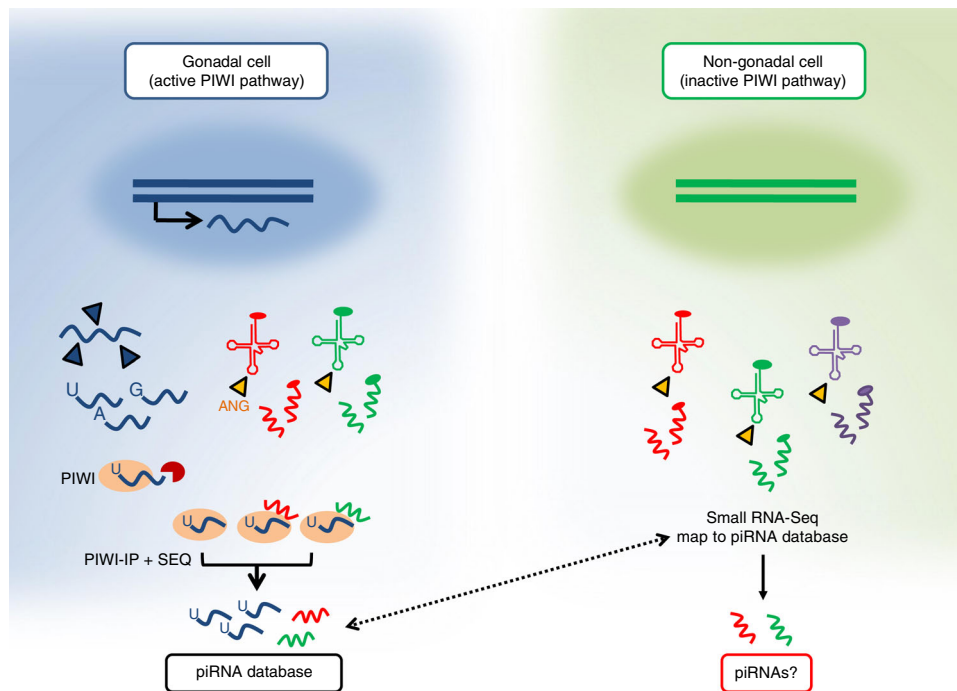


Fig. 2 Presence of ncRNA fragments in piRNA databases accounts for piRNA expression in cells where the piRNA pathway is off. A gonadal somatic or germinal cell (blue) produces PIWI pathway-dependent piRNAs (blue RNAs; only the primary piRNA pathway is represented), which are allocated in the 5' binding pocket of a PIWI-clade protein (orange). ncRNA fragments (red, green) are also present in PIWI immunoprecipitates, but they are not necessarily bona fide piRNAs. When sequencing data are mapped against a piRNA database, ncRNA fragments will be thought as piRNAs, even in the absence of a functional piRNA pathway

A recent survey of circulating small RNAs present in human blood plasma¹⁸ 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 ≥ 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 testis²¹. 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^{22, 23}. 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 space²⁴. 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 cells¹⁷. 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 ≥ 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 misclassification 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^{18, 25, 26} 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 genome²⁷. According to their biogenesis, piRNAs can be further classified as genome-encoded primary piRNAs, ping-pong generated secondary piRNAs, and even phased tertiary piRNAs²⁸. 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^{29, 30}. Although some of these papers relied on RIP-seq for piRNA identification⁷, 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 ncRNAs⁵. Furthermore, as there is still a lack of suitable antibodies for selective immunoprecipitation of human PIWI proteins³¹, 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 be this extrapolated to other tissues? For instance, if a tRNA-derived fragment³² 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 (Fig. 2). 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-IP¹⁵. 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 line³¹. However, this result should be interpreted carefully, as the authors found a very

strong correlation ($r=0.91$) 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 testis¹², suggesting a possible contamination with AGO-clade bound RNAs. Secondly, 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 *mili* knockout animals¹², 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 testis¹², 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 database³³. 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 should be noted 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 17¹⁵ (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 tRNAs³⁴. 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 solely 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 repositories³⁵. Analogously, the curation of piRNA databases will enable the study of hypothetical piRNA

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, Fasta files containing the complete list of human piRNAs were downloaded from either RNAdb 2.0²⁹ or piRBase³⁶, and mapped to ad hoc references containing human genomic and mitochondrial rRNAs (downloaded from NCBI), tRNAs (downloaded from the Genomic tRNA Database, GtRNAdb, and the mitochondria tRNA database, mitotRNAdb), small nuclear RNAs, small nucleolar 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 studies reporting piRNAs in human blood¹⁸ and human testes²¹ were directly extracted from their supplementary Materials section, and the sequences were compared to a reference file containing all piRNAs annotated in RNAdb 2.0 for which we could find 100% identity to ncRNAs. In the case of human piRNAs present in the Cancer Genome Atlas¹⁷, we computed the number of normalized reads for each putative piRNA across cancer samples, and analyzed the 20 most frequently detected piRNAs in cancer. For piRNAs present in mouse hippocampus¹⁵, we performed Blast alignments against the NCBI collection of non-redundant nucleotide sequences of mice, and looked for ncRNAs with 100% identity to the putative brain piRNAs.

To analyze the distribution of sequencing reads mapping to human YRNAs in testes and plasma, we extracted Fastq files from the NCBI small Read Archive (SRR2496797) or the European Nucleotide Archive (ERA246774), clipped adaptors, mapped the clipped reads to the human genome (hg19) using Bowtie2, and created a BigWig file from the output, using a bin size of 1, and used the UCSC Genome Browser for visualization. To address the enrichment of putative blood piRNAs in sequences not starting with uridine or matching ncRNAs, we defined a parameter called the Enrichment Score (ES). For the calculation, we divided human piRNAs present in RNAdb 2.0 (reference) in two mutually exclusive categories (e.g., starting [A] or not starting [B] with uridine; identical [A] or not identical [B] to other non-coding RNAs). We then took the 144 piRNAs described in blood plasma¹⁸ (query), and performed the same categorization. The ES was calculated as the quotient between the number of sequences in the query in categories B and A, divided by the quotient of the number of sequences in the reference in categories B and A. Thus, an ES higher than 1 shows that there is a higher number of sequences in category B in the query than what expected from the distribution in the reference. Alternatively, the sum of the sequencing reads in each category was used, instead of the total number of sequences.

Data availability. All relevant data not present within the manuscript or supplementary files are available from the authors upon request.

Received: 8 September 2017 Accepted: 13 October 2017

Published online: 22 January 2018

References

- Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* **10**, 94–108 (2009).
- Carmell, M. A. et al. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.* **16**, 2733–2742 (2002).
- Cox, D. N. et al. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev.* **12**, 3715–3727 (1998).
- Lin, H. & Spradling, A. C. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development* **124**, 2463–2476 (1997).
- Girard, A. et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199–202 (2006).
- Grivna, S. T. et al. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.* **20**, 1709–1714 (2006).
- Aravin, A. et al. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203–207 (2006).
- Vagin, V. V. et al. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320–324 (2006).
- Brennecke, J. et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).
- Gunawardane, L. S. et al. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* **315**, 1587–1590 (2007).

- Aravin, A. A. et al. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**, 744–747 (2007).
- 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).
- Ross, R. J., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. *Nature* **505**, 353–359 (2014).
- Brower-Toland, B. et al. *Drosophila* PIWI associates with chromatin and interacts directly with HP1a. *Genes Dev.* **21**, 2300–2311 (2007).
- Lee, E. J. et al. Identification of piRNAs in the central nervous system. *RNA* **17**, 1090–1099 (2011).
- Rajasethupathy, P. et al. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell* **149**, 693–707 (2012).
- Martinez, V. D. 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).
- 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).
- Chen, Q. et al. Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science* **351**, 397–400 (2016).
- Sharma, U. et al. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science* **351**, 391–396 (2016).
- Yang, 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).
- 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).
- Czech, B. & Hannon, G. J. A happy 3' ending to the piRNA maturation story. *Cell* **164**, 838–840 (2016).
- Buck, A. H. et al. Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nat. Commun.* **5**, 5488, <https://doi.org/10.1038/ncomms6488> (2014).
- Yeri, A. et al. Total extracellular small RNA Profiles from plasma, saliva, and urine of healthy subjects. *Sci. Rep.* **7**, 44061, <https://doi.org/10.1038/srep44061> (2017).
- Iliev, R. et al. Expression levels of PIWI-interacting RNA, piR-823, are deregulated in tumor tissue, blood serum and urine of patients with renal cell carcinoma. *Anticancer Res.* **36**, 6419–6423 (2016).
- Zuo, L. et al. piRNAs and their functions in the brain. *Int. J. Hum. Genet.* **16**, 53–60 (2016).
- Siomi, H. & Siomi, M. C. RNA Phased piRNAs tackle transposons. *Science* **348**, 756–757 (2015).
- Pang, K. C. et al. RNAdb 2.0--an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.* **35**, D178–D182, <https://doi.org/10.1093/nar/gkl926> (2007).
- Lakshmi, S. & Agrawal, S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.* **36**, D173–D177, <https://doi.org/10.1093/nar/gkm696> (2008).
- 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).
- Anderson, P. & Ivanov, P. tRNA fragments in human health and disease. *FEBS Lett.* **588**, 4297–4304 (2014).
- Rosenkranz, D. piRNA cluster database: a web resource for piRNA producing loci. *Nucleic Acids Res.* **44**, D223–D230, <https://doi.org/10.1093/nar/gkv1265> (2016).
- Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73, <https://doi.org/10.1093/nar/gkt1181> (2014).
- 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).
- Zhang, P. et al. piRBase: a web resource assisting piRNA functional study. *Database* **2014**, bau110, <https://doi.org/10.1093/database/bau110> (2014).

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. A.C. and C.R. discussed results, proposed new experiments/analysis, and made major contributions to the submitted version of the manuscript.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s42003-017-0001-7>.

Competing interests: The authors declare that they have no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018