

SOFTWARE

Open Access



piRAT: piRNA Annotation Tool for annotating, analyzing, and visualizing piRNAs

Dominik Robak¹ and Guillem Ylla^{1*}

*Correspondence:

Guillem Ylla

guillem.ylla@uj.edu.pl

¹Laboratory of Bioinformatics and Genome Biology, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, 30-387 Kraków, Poland

Abstract

Background In recent years, Piwi-interacting RNAs (piRNAs) have been found to be involved in different biological roles beyond their initially identified role in protecting the germline genome. However, there is a lack of robust computational tools that facilitate their accurate annotation, analysis, and visualization, which would allow large-scale analysis of piRNAs across samples and species.

Results We present piRAT, a piRNA Annotation Tool for annotating both primary piRNAs from piRNA clusters and secondary piRNAs generated via the ping-pong cycle, using small RNA-seq data mapped to a genome. piRAT also performs descriptive analyses and generates comprehensive reports with visualizations to support annotations and provide further insights.

Conclusions In our comparative analyses, piRAT emerged as the most comprehensive tool available in terms of functionalities. It consistently produced higher-quality annotations of piRNA clusters and ping-pong sites and generated the most detailed reports. As piRAT requires no prior annotations or special data preprocessing, it is a versatile choice for any species, particularly for non-model organisms.

Keywords Small RNA, piRNAs, Genome annotation, Transcriptomics

Background

Piwi-interacting RNAs (piRNAs) are a class of small non-coding RNAs, typically 26–33 nucleotides in length depending on the species, first identified in the early 2000s [1]. Similar to other small RNAs, such as micro RNAs (miRNAs) and silencing RNAs (siRNAs), piRNAs bind to Argonaute-family proteins, which mediate their interaction with target transcripts, regulating the target's expression at a transcriptional or post-transcriptional level. In the case of piRNAs, as their name suggests, bind to PIWI proteins [2] a clade belonging to the Argonaute family [3].

Initially, piRNAs were identified in the germ cells of *Drosophila*, and subsequently in the gonads of other animals, where they were shown to target transposable elements, a role for which they became known as guardians of the genome [4]. However, more recent studies have revealed that piRNAs are also abundantly expressed in both gonadal and non-gonadal somatic tissues in many animal species [5–7]. These discoveries have



broadened the focus toward the roles of piRNAs beyond the germline. Indeed, piRNAs have been found to be involved in a variety of biological processes, including sex determination in silkworms [8], multiple human cancers [9–11], and early embryo development [12].

Despite their important roles, a limited number of tools are available for robustly annotating piRNAs, which is a crucial step in their study. This lack of tools is especially detrimental for research on non-model organisms for which no prior small RNA annotations, including piRNA annotations and piRNA databases, exist [13–15]. In this article, we present piRAT, a piRNA Annotation Tool, that leverages small RNA-seq data aligned to a reference genome to annotate piRNA regions from the primary and secondary pathways, on the basis of information regarding piRNA biogenesis, suitable for both model and non-model organisms.

piRNA biogenesis signatures to annotate piRNAs

The piRNA biogenesis is most well understood in the model organism *D. melanogaster*, and although some steps are still under discussion, most animals share similar mechanisms of piRNA biogenesis [16]. It is well established that piRNAs can be generated through two pathways, the primary pathway and the secondary pathway, also known as the ping-pong cycle. Both pathways display specific signatures in the small RNA-seq data that piRAT leverages to annotate bona fide piRNAs loci.

In brief, the primary pathway starts with the transcription of long primary transcripts (up to hundreds of kilobases long) that are subsequently cleaved by the Zucchini protein, releasing small RNA fragments (pre-piRNAs). These RNAs are loaded onto a PIWI protein, followed by 3' trimming and 2-O-methylation, resulting in 26–33 nts long mature primary piRNAs [16]. Based on this knowledge, we can identify regions producing primary transcripts, also known as piRNA clusters, as genomic regions with a high density of small RNA-seq reads of the length corresponding to the piRNAs on the given species.

The secondary pathway, known as the ping-pong cycle, was originally discovered in the *D. melanogaster* germline [17, 18] and subsequently detected in other animals' germline, embryogenesis, and more recently in non-gonadal tissues of some animal lineages [7, 19–21]. In this cycle, piRNAs bind to PIWI (or its paralog in flies Aubergine) and identify their target RNAs based on sequence complementarity. Vasa and Krimper cleave the target at the position corresponding to the complementary 10th nucleotide of the primary piRNA. Then, AGO3 binds to the 5' end of the cleaved RNA and Nibbler trims the 3' end, leaving a ~28nts sequence bound to AGO3, which is then methylated by Hen1, producing a mature secondary piRNA. This mature secondary piRNA, is complementary to the piRNA cluster transcripts, which it slices, producing more primary piRNAs as the one that initiated the cycle (more details in our recent publication [7]).

The presence of the ping-pong cycle can be robustly determined through small RNA-seq data by the identification of the so-called ping-pong signature [22]. This ping-pong signature consists of two small RNA-seq reads of ~28 nts with a 5' to 5' overlap of 10 nts in opposite strands of the genome. Additionally, because primary piRNAs are enriched with a U at the 5' end, secondary piRNAs are enriched by an A at position 10.

Tools for piRNA annotation and analysis

Despite the relevance of piRNAs, few methods exist for robust identification, annotation, classification, and visualization of piRNAs that can work across different animals. Additionally, the few existing tools present significant limitations (Table 1).

Among the available tools for piRNA cluster annotation, proTRAC [23], is one of the most widely used. However, while proTRAC specializes in piRNA cluster annotation, it does not annotate ping-pong piRNAs. Furthermore, owing to its probabilistic method, which is based on the assumption of a uniform distribution, proTRAC was found to fail to accurately annotate piRNA clusters in some species [24]. In response to this limitation, another tool called piClust [24] was developed on the basis of the existing Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN) [25].

While piClust was shown to outperform proTRAC in piRNA cluster discovery [24], it has several limitations that make it unsuitable for many purposes. A major limitation is that piClust ran through a web server that accepted a single SAM file with a maximum size of 200 Mb per submission. Additionally, at the time of writing, the piClust server is down. A further drawback of piClust and proTRAC, is the lack of ping-pong annotations, forcing the user to look for alternative methods to identify ping-pong sites.

One of the most recently published tools for piRNA cluster annotation is PILFER [26]. When the authors benchmarked it against proTRAC, PILFER identified fewer clusters but longer [26]. However, PILFER was not designed to work solely with small RNA-seq reads mapped to a genome. In their publication [26], the authors used human small RNA-seq reads that were classified either as canonical piRNAs or potential piRNAs. To identify canonical piRNA reads, they mapped the reads to the human genome and retained those overlapping with piRNAs annotated in piRNABank. Potential piRNAs were defined as 26–33 nucleotide reads that mapped to the genome but were not found in piRNABank and did not overlap with other known non-coding RNAs. The final input for PILFER was a BED file containing the genomic coordinates of both canonical and putative piRNAs. This preprocessing requires extra work and relies heavily on prior piRNA and non-coding RNA annotations, limiting its utility in non-model species.

Our tool, piRAT, addresses the limitations of existing piRNA annotation tools and offers several novel features, including comprehensive graphical outputs that visually support the annotations. Consequently, piRAT stands out for a) identifying piRNAs

Table 1 Feature comparison of the principal piRNA annotation tools and the herein presented piRAT

	Annotation of piRNA clusters	Annotation of ping-pong loci	Input file type	No special preprocessing	Graphical output	Max. number of input files per run	Multi-thread computing
piRAT	✓	✓	.bam	✓	✓	Limited by hardware	✓
proTRAC	✓	✗	.sam	✗	✓	1	✗
piClust	✓	✗	.sam	✗	✗	1	✗
PILFER	✓	✗	.bed	✗	✗	1	✗
PingPongPro	✗	✓	.sam/.bam	✗	✓	1	✗

Comparison includes whether they can annotate piRNA clusters and ping-pong loci, the required type of inputs and their format, the generation of graphical outputs, and the use of multiple CPU threads

from both biogenesis pathways in a single run, b) accepting multiple standard binary alignment map (BAM) files of any size (with limits determined by hardware capabilities), c) not requiring special preprocessing steps or needing external databases or annotations), and d) generating graphical outputs that support the annotations and provide overall descriptive statistics.

Thus, piRAT provides a comprehensive set of features, utilizes standard input and output file formats, and doesn't require any type of genome annotation. Moreover, as demonstrated subsequently, it also yields higher-quality piRNA annotations than existing tools.

Implementation

As input, piRAT requires a directory containing BAM files of small RNA-seq aligned to the same reference genome. The output consists of general feature format (GFF) files with the genomic coordinates of the annotated piRNA clusters and ping-pong signature sites found within the samples (the annotations are performed with the information from pooling all BAM files together to increase the sensitivity, an important factor especially for low-expressed piRNAs). Additionally, piRAT returns comprehensive graphical reports displaying the supporting data for the annotations and other summary statistics.

While piRAT executes with a single command, internally it comprises four distinct modules: the first module determines the piRNA length within the provided dataset; the second module annotates the primary piRNA clusters; the third module identifies the ping-pong loci; and the fourth module generates graphical reports of both the input data and final annotations (Fig. 1). Each of these modules is described in detail below.

piRNA length detection and read variant purging

The piRNAs often range from ~26 to 32 nts, but the specific lengths depend on the species. If the length of piRNAs in a particular species is known, the user can manually input this information into piRAT as a parameter. Alternatively, piRAT predicts the range of piRNA sizes for a given species based on the length of reads displaying 10 nt 5' to 5' overlap, or in the small RNA-seq read length distribution when running in cluster-only mode (Supplementary Methods). Only reads that fall within the specified piRNA lengths, which can be either provided by the user or determined by piRAT, are utilized in subsequent piRNA annotation steps. Reads within this length range are herein referred to as "piRNA-length reads".

In small RNA-seq data, reads mapped to the same genomic locus might be slightly different even if they originate from the same genomic element (miRNA, piRNA, etc.) [27]. This can be due to technical reasons, such as errors during sequencing, or biological phenomena such as 3'-tailing, presence of immature RNA molecules, degradation, polymorphisms, and other post-transcriptional modifications. Often, we observe that for a given miRNA or piRNA, there is one read with a large number of copies and several other reads with slight variation in the extremes (mostly in 3') with many fewer copies.

We argue that treating all unique small RNA-seq reads as distinct RNA molecules artificially inflates the number of annotated piRNAs. To mitigate this issue, piRAT selects a single "representative read" for each locus where multiple piRNA-length reads are mapped. The representative read is defined as the highest copy number among those mapped reads that exhibit up to three nucleotide differences at the start and end

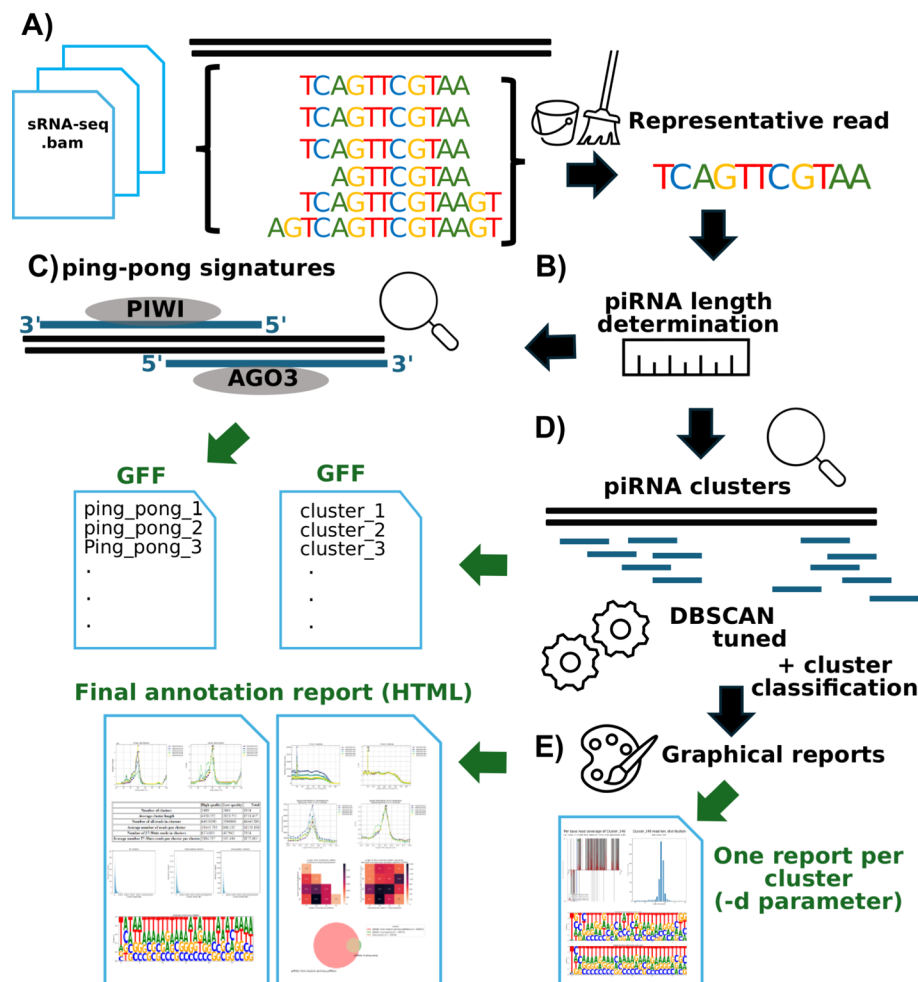


Fig. 1 Schematic workflow of piRAT. **A** Starting with small RNA-seq datasets mapped to a reference genome, piRAT first cleans the data by selecting a single representative read per locus. **B** Then, it determines the piRNA length in the dataset (if not provided by the user), and proceeds with the **C** annotation of ping-pong loci using the characteristic ping-pong signature, and **D** piRNA clusters via a tuned version of DBSCAN followed by a cluster classification step. Finally, **E** piRAT generates extensive graphical outputs in the final report, and optionally, one report per cluster with plots showing its supporting data. The outputs (green arrows) are GFF files with annotations and HTML files with reports

positions. The variation threshold was set to three nucleotides by default, based on our analysis, which revealed that beyond three, the number of ping-pong signatures plateaus (Supplementary Figure S1). Users can manually adjust this threshold using the parameter (--variation-threshold, -vt), with a minimum value of zero, which would force piRAT to include all mapped reads with unique start and end coordinates.

Primary piRNA cluster annotation

In the primary piRNA biogenesis pathway, long precursor RNAs are cleaved into shorter fragments, bound to PIWI proteins, and trimmed, producing mature piRNAs. When small RNA-seq reads from these piRNAs are mapped to the genome, the reads originating from the same precursor RNA align to the genomic region of the precursor, typically referred to as a primary piRNA cluster. Annotating these piRNA clusters is the standard method for identifying piRNAs produced through the primary pathway [23, 24, 26].

To identify piRNA clusters, piRAT, similarly to piClust, is based on the DBSCAN algorithm [25], although with some modifications explained below and in Supplementary Methods with more detail. DBSCAN requires two parameters k and ϵ , with the k parameter being the k -th nearest neighbor used to compute the distance, and ϵ defining the maximum distance between two points (in this case, two small RNA-seq reads) for one to be considered within the neighborhood of the other. If the k -th neighbor is within the ϵ distance, the two points are connected in a density-reachable manner, and hence belong to the same cluster (more details in Supplementary Methods). Both k and ϵ , can be provided by the user. If not specified, piRAT estimates the optimal values for these parameters as described in Supplementary Methods.

Since piRAT's cluster detection is strand-specific, an additional step is performed to merge clusters into bidirectional or dual-stranded clusters. In this step, piRAT measures the distance from each cluster to its neighbors in the complementary strand, and if the distance between the 3' ends is within the ϵ value used during clustering, they are merged into a bidirectional cluster. If the length of overlap is greater than 50% of the shorter cluster, they are merged into a dual-strand cluster (Supplementary Methods).

Finally, piRAT classifies clusters as high- or low-quality based on specific biogenesis criteria. Using all small RNA-seq reads mapped within an annotated cluster, piRAT assigns a high-quality classification if two conditions are met: (1) more than 50% of the reads start with uridine (U, in the RNA-seq data), and (2) over 50% of the reads mapped within the cluster are of the defined piRNA length. Both thresholds can be adjusted by the user.

The cluster annotations are provided to the user in three GFF files: one containing all clusters, one for high-quality clusters, and one for low-quality clusters. Although both high- and low-quality clusters are returned, we recommend using only the high-quality clusters, as these are used in the subsequent benchmarking analyses. Additional output files are also generated, including information such as the genomic coordinates of all putative piRNAs within the clusters.

Ping-pong piRNA annotation

To identify piRNAs generated through the ping-pong cycle, piRAT utilizes the well-established "ping-pong signature" [22]. This signature is characterized by pairs of piRNA-length small RNA-seq reads mapped to opposite strands of the genome, with a 5' to 5' overlap of 10 nucleotides. piRAT detects this ping-pong signature using the "representative reads" of piRNA length. The resulting ping-pong site annotations are provided in a GFF file containing the coordinates of these two pairs of potential piRNAs. This GFF can be used downstream to quantify the expression of each specific ping-pong pair across samples and can also be visualized in genome browsers such as IGV.

Graphical reports

The graphical outputs generated by piRAT are designed to provide support for the annotations and provide an overview of the characteristics of the annotated piRNAs in a given dataset.

When piRAT runs with the active graphical output for the clusters (parameter $-d$), it generates several plots for each cluster, including a bar plot showing the per-nucleotide read coverage along the cluster, the read length distribution, and two sequence logos

representing all reads within the cluster and the other displaying only the representative reads. Although only piRNA-length representative reads are used for annotating clusters, all reads mapped to the cluster are included in these plots to provide comprehensive evidence of cluster quality. Since generating these graphical outputs increases the runtime and hardware requirements when hundreds of clusters are annotated, this option is disabled by default.

In addition to the individual plots for each cluster, at the end of the run, piRAT generates the “piRAT Final Report”, which includes the following plots and tables (Fig. 2):

Initial Read Length Distribution Plot: Shows the distribution of small RNA-seq reads (20–40nts) across samples, with the x-axis representing read length and the y-axis the read count (Fig. 2A). A z-score normalized version is also provided to compare samples with differing library sizes. The presence and height of a peak around 26–33 nt, provide clues about the relative abundance of piRNAs versus other small RNAs in the samples and length of the piRNAs in the given species.

Table of piRNA Cluster Statistics: This table summarizes key statistics for high-quality, low-quality, and total annotated piRNA clusters (Fig. 2B).

Cluster Length Distribution: These three plots illustrate the length distributions of high-quality, low-quality, and all clusters (Fig. 2C–D).

Primary piRNA Sequence Logo: This plot displays the sequence logo of piRNA-length reads inside the annotated clusters. Given the known enrichment of uracil at the first position of primary piRNAs, bona fide piRNA clusters should exhibit a clear bias toward T in this position (Fig. 2G).

5' to 5' Overlaps plot: It displays the number of read pairs for each 5' to 5' overlap value. A clear peak at 10 nts overlap strongly indicates the presence of ping-pong piRNAs in the given sample (Fig. 2E). A z-score normalized version of the plot is also provided for comparing the intensity of the ping-pong signature across samples.

Length distribution of ping-pong signature sequence pairs: This plot displays the length distributions of the read pairs exhibiting a 10nt 5' to 5' overlap on opposite strands. It is enriched for true ping-pong piRNAs, providing clearer insights into the actual lengths of piRNAs involved in the ping-pong cycle for each sample. A z-score normalized version is also included. Representative reads within the piRNA-length range and showing this overlap are considered piRNAs participating in the ping-pong cycle (Fig. 2F).

Heatmap of Potential piRNA Pairs Lengths: This heatmap shows the frequencies of the lengths of the two piRNA displaying ping-pong signatures (Fig. 2H). This allows the user to determine whether there is homotypic ping-pong (the two piRNAs are of the same length) or heterotypic ping-pong (between piRNAs of different lengths).

Heatmap of Potential piRNA Pair Lengths Split by 1 T and 10A: This heatmap differs from the previous heatmap in that the putative primary piRNAs (displaying 1 T) and the putative secondary piRNA (10A) are on different axes (Fig. 2I). In the case of heterotypic ping-pong, this plot allows extrapolating what piRNA lengths are generated from the primary and secondary pathways.

Venn diagram: Displays the number of piRNAs in clusters and ping-pong sites, and the intersection shows the number of piRNAs found in a cluster that also participate in the ping-pong in other loci (Fig. 2J).

Sequence logo of ping-pong piRNAs: It contains the sequences of all representative reads of the detected ping-pong loci (Fig. 2K). Primary piRNAs typically enriched for U

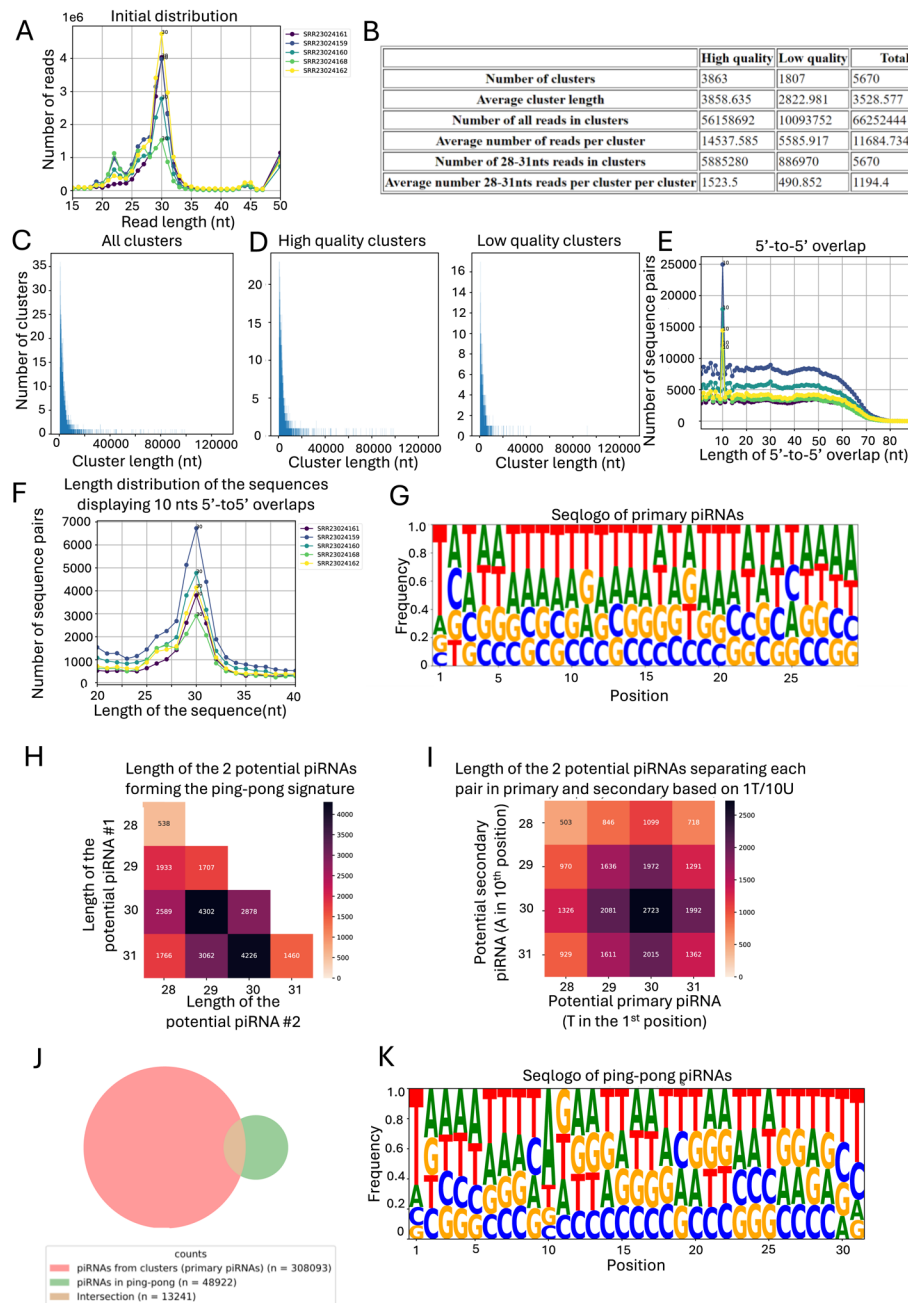


Fig. 2 Example plots of the graphical output of piRAT running on 5 small RNA-seq datasets. **A** Read length distribution in each dataset. **B** Summary of the annotated clusters. **C** Distribution of all the annotated clusters, as well as only for the **D** high-quality and low-quality ones. **E** 5'-to-5' overlaps in each of the datasets. **F** Read length distribution of those read pairs that have 10nts overlap 5' to 5'. **G** Seqlogo of the reads belonging to clusters. **H** Heatmap of the piRNA length of the read pairs displaying 10 nts 5' to 5', as well as **I** separating the pairs based on the ones starting with T and those with A at the 10th position. **J** Venn diagram displaying the overlap of piRNAs found in clusters as well as in ping-pong sites. **K** SeqLogo of all the ping-pong piRNAs

in the first position, whereas secondary piRNAs are enriched for A in the tenth position. If ping-pong piRNAs have been correctly identified, this sequence logo should show a higher frequency of T in the first position and a higher frequency of A at the tenth position.

Results

We benchmarked piRAT with publicly available small RNA-seq datasets from testes, ovaries, and non-gonadal tissues from human (*Homo sapiens*), mouse (*Mus musculus*), and fly (*Drosophila melanogaster*) (Supplementary Table S1). Male and female gonadal tissues were expected to exhibit piRNA expression, whereas somatic tissues served as negative controls, as piRNA expression in mammals is restricted primarily to germ cells. In flies, non-gonadal piRNA expression is possible albeit much lower than in gonads [7].

Ping-pong signature annotation benchmark

We compared piRAT’s ping-pong signature annotations to those of PingPongPro [28] with the same small RNA-seq datasets from human, mouse, and fly testes and ovaries (Table 2). While piRAT provides coordinates for pairs of piRNAs displaying the signature along with comprehensive graphical plots, PingPongPro returns only the regions where the 10 nt overlap occurs. In this comparison, piRAT identified approximately 42–67% of the signatures reported by PingPongPro.

This discrepancy arises from piRAT’s stringent criteria to reduce false positives. While PingPongPro utilizes all unique reads, piRAT considers only reads of the specified piRNA length and utilizes a single representative read per locus. Therefore, comparing the ping-pong signautres annotated by both tools, shows a large proportion of unique piRAT ping-pong loci overlapping with multiple PingPongPro loci (Supplementary Figures S2 & S3).

piRNA cluster identification with real data

For piRNA clustering annotation, we compared piRAT to the other popular tools, proTRAC, and PILFER, using the same datasets of human, mouse, and fly from gonadal and non-gonadal tissue without any specific preprocessing (Table 3). Note that we were unable to include piClust in our benchmark due to the input file limit and the availability of the web server.

piRAT consistently annotates a greater number of high-quality clusters than proTRAC and PILFER in gonadal tissues (Table 3). Considering that in our negative controls (the mammalian somatic tissues), no high-quality clusters are annotated by piRAT, the increased number of high-quality clusters annotated by piRAT in gonadal tissues is unlikely to be due to false positives. While proTRAC also didn’t annotate any clusters in the mammalian non-gonadal tissues, PILFER found 5 in humans and 95 in mice

Table 2 Number of ping-pong signatures identified by piRAT and PingPongPro in small RNA-seq data from the testes and ovaries of humans, mice, and flies, and the percentage of overlapping signatures between tools

	piRAT	PingPongPro	% piRAT in pingPongPro	% pingPongPro in piRAT
<i>Human</i>				
Testes	60,457	174,441	100%	58.32%
Ovaries	9314	13,641	100%	56.65%
<i>Mouse</i>				
Testes	22,385	22,603	100%	42.93%
Ovaries	14,036	43,781	100%	48.74%
<i>Fly</i>				
Testes	52,881	72,219	92%	67.36%
Ovaries	139,824	227,094	100%	61.57%

Table 3 Number of predicted clusters in each small RNA-seq datasets from the testes, ovaries, and somatic tissues of humans, mice, and flies

Tissue	piRAT high-quality	piRAT low-quality	proTRAC	PILFER
<i>Human</i>				
Testes	22,552	1332	174	174
Ovaries	3978	1157	159	244
Somatic	0	28	0	5
<i>Mouse</i>				
Testes	4237	3400	207	377
Ovaries	6591	4182	8	56
Somatic	0	59	0	95
<i>Fly</i>				
Testes	2630	3038	6	135
Ovaries	4430	395	5	1426
Somatic	9	1807	0	47

(Table 3). The high-quality clusters detected in *D. melanogaster*, are not necessarily false positives, given that non-gonadal piRNAs can occur in this species [7] and that further explorations of the 9 clusters piRAT detected revealed clear signs of bona fide piRNA clusters (Supplementary Figure S4).

As an indicator of the cluster's quality, we calculated the proportion of reads within the clusters annotated from each tool that are putative piRNAs (those of 26–32 nts that start with T), assuming that real piRNA clusters should contain most of the reads having these characteristics (Supplementary Table S2). Here, piRAT high-quality clusters contained between 43.79 and 65.91% of reads with piRNA characteristics, similar to proTRAC (51.50–69.07%), and higher than PILFER (0.97–47.78%).

As an indicator of the completeness of the piRNA clusters of each tool, we calculated the proportion of putative piRNA reads (those between 26 and 32 nts that start with T) within the RNA-seq datasets that map inside the clusters. Highly complete piRNA cluster annotations should contain a large fraction of these reads, assuming that most reads of these characteristics should belong to primary piRNAs and therefore should be found inside clusters (Supplementary Table S3). In this metric, piRAT consistently scored higher, with its high-quality clusters containing between 33.90 and 93.04% of the reads of piRNA characteristics, than proTRAC (0.24–53.12%) and PILFER (0.51–56.53%).

Furthermore, we checked whether the three canonical and well-known piRNA clusters of *Drosophila* had been correctly annotated by piRAT. These were the flamenco, 42AB, and 20A clusters [29], which indeed were annotated by piRAT as high-quality clusters using testes and ovaries data, but not in somatic tissue (Supplementary Figure S5).

Overall, piRAT annotated more clusters across species in gonadal tissues without showing signs of false positives, as it did not detect clusters in mammalian somatic tissues. piRAT high-quality clusters contain a high percentage of small RNA-seq reads with the piRNA characteristics, and additionally, these clusters explain a greater proportion of small RNA-seq reads with piRNA characteristics.

Benchmarking against references

Due to the lack of a definitive ground truth for piRNA cluster annotations, we compared the performance of piRNA annotation tools against two different sources. First, we compare the cluster to those from mouse and human, annotated in Girard et al. [2], which is

Table 4 Percentage of nucleotides of piRNA clusters from piRNadb and Girard et al. [2], that are covered by clusters annotated by piRAT high-quality clusters, proTRAC, and PILFER in humans, mice, and flies (fly annotations only available from piRNadb) using small RNA-seq datasets from ovaries, testes, and somatic tissues

	piRNadb			Girard et al. [2]		
	piRAT-HQ	proTRAC	PILFER	piRAT-HQ	proTRAC	PILFER
<i>Human</i>						
Testes	93.65%	36.91%	54.50%	66.40%	25.95%	37.87%
Ovaries	33.57%	5.54%	1.58%	16.42%	2.29%	3.04%
Somatic	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
<i>Mouse</i>						
Testes	6.85%	4.64%	22.88%	82.97%	75.60%	80.39%
Ovaries	3.17%	0.00%	4.94%	5.61%	0.00%	6.77%
Somatic	0.00%	0.00%	2.04%	0.00%	0.00%	0.96%
<i>Fly</i>						
Testes	62.90%	0.00%	6.90%			
Ovaries	81.96%	4.30%	95.80%			
Somatic	0.22%	0.00%	3.61%			

the same dataset used by the authors of piClust for their benchmarks [24]. The second benchmark was done using piRNadb (<https://www.pirnadb.org>) [30], which contains clusters identified using the method proposed by Lau et al. [31] and using described piRNAs from multiple works.

We computed the percentage of bases in the reference clusters (from Girard et al. [2] and piRNadb) that overlapped with the annotated clusters from piRAT, proTRAC, and PILFER (Table 4). In these comparisons, piRAT high-quality clusters in most cases consistently covered a greater percentage of the reference clusters than did proTRAC and PILFER, indicating higher sensitivity.

Computing performance

While running the benchmarks, we monitored the running time and the maximum RAM usage on a high-performance computing cluster node using a single thread, as well as 26 threads in piRAT (Supplementary Table S4). Running with a single thread piRAT takes the longest. However, the piRAT running time can be significantly shortened via multiple threads. For example, for the four datasets of mouse testes for piRAT it took ~ 29 h, proTRAC ~ 3 h, and PILFER ~ 19 h, but piRAT running time was decreased to ~ 2 h using 26 threads.

To further assess scalability on consumer hardware, we evaluated piRAT performance across varying thread counts (1, 4, 8, 16, 32) on a desktop PC equipped with an AMD Ryzen 9 9950X processor (16 cores, 32 threads) and 64 GB RAM, using human and mouse ovaries datasets. Execution time decreased between 3.5 and 4.7-fold with 32 threads compared to one, while memory usage remained stable (~ 2 Gb) regardless of thread count (Supplementary Figure S6).

Notably, piRAT was designed to run on both high-performance computing clusters and consumer desktops with more limited resources. This is achieved through on-demand data access: during statistical analysis, piRAT queries each BAM file for reads from specific genomic regions rather than loading entire datasets into memory. Users can further tune resource usage through adjustable parameters, making piRAT accessible for researchers without dedicated computational infrastructure.

Conclusions

On the basis of our comparison, piRAT emerged as the most complete tool in terms of functionalities and supporting information that it provides. Compared to alternative tools, benchmarking consistently showed that piRAT delivered higher-quality annotations of piRNA clusters and ping-pong sites. Because it is easy to use, doesn't need prior annotations or special preprocessing, and provides high-quality annotations and reports, we believe it is an excellent choice for any species, but especially for non-model organisms.

Abbreviations

piRNA	Piwi-interating RNA
miRNA	Micro RNA
siRNA	Silencing RNA
BAM	Binary alignment map
GFF	General feature format

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-026-06380-9>.

Additional file1 (DOCX 4649 kb)

Additional file2 (DOCX 754 kb)

Additional file3 (DOCX 27 kb)

Acknowledgements

We thank all the lab members who have been involved in technical discussions at different stages of the development. We especially thank Krystian Komenda for being the first user of piRAT, testing the new features, and providing valuable feedback on it.

Author contributions

DR wrote the piRAT's code, performed analysis, and prepared the figures. GY conceived and led the project and wrote the manuscript. All authors approved the final version of the manuscript.

Funding

This work has been supported by the Polish National Science Centre, under the Sonata-17 grant [UMO-2021/43/D/NZ2/03228]. Furthermore, the Laboratory of Bioinformatics and Genome Biology received structural support by the Faculty of Biochemistry, Biophysics, and Biotechnology at Jagiellonian University (Poland) under the Strategic Programme Excellence Initiative (PRA BioS), and from the Federation of European Biochemical Societies Excellence Award 2023.

Data availability

piRAT is available on GitHub (<https://github.com/ylla-lab/piRAT>), implemented in Python 3, and released under MIT license. It is also installable through pip, conda, and available as a Docker container (<https://hub.docker.com/u/domrob>). The Python implementation runs natively on Linux and in WSL in Windows, and the Docker image can run on Linux, Windows, and MacOS. The small RNA-seq datasets used in this article were retrieved from NCBI-SRA (<https://www.ncbi.nlm.nih.gov/sra>), and the accession identifiers of all datasets are described in Supplementary Table S1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 13 August 2025 / Accepted: 14 January 2026

Published online: 31 January 2026

References

1. Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr Biol*. 2001;11:1017–27. [https://doi.org/10.1016/S0960-9822\(01\)00299-8](https://doi.org/10.1016/S0960-9822(01)00299-8).
2. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*. 2006;442:199–202. <https://doi.org/10.1038/nature04917>.
3. Carmell MA, Xuan Z, Zhang MQ, Hannon GJ. The argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev*. 2002;16:2733–42. <https://doi.org/10.1101/gad.1026102>.
4. Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*. 2007;318:761–4. <https://doi.org/10.1126/science.1146484>.
5. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. *Nat Ecol Evol*. 2018;2:174–81. <https://doi.org/10.1038/s41559-017-0403-4>.
6. Llonga N, Ylla G, Bau J, Belles X, Piulachs MD. Diversity of piRNA expression patterns during the ontogeny of the German cockroach. *J Exp Zool B Mol Dev Evol*. 2018;330:288–95. <https://doi.org/10.1002/jez.b.22815>.
7. Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, et al. A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature*. 2014;509:633–6. <https://doi.org/10.1038/nature13315>.
8. Liu Y, Dou M, Song X, Dong Y, Liu S, Liu H, et al. The emerging role of the piRNA/piwi complex in cancer. *Mol Cancer*. 2019;18:123. <https://doi.org/10.1186/s12943-019-1052-9>.
9. Weng W, Li H, Goel A. Piwi-interacting RNAs (piRNAs) and cancer: emerging biological concepts and potential clinical implications. *Biochimica et Biophysica Acta (BBA) Rev Cancer*. 2019;1871:160–9. <https://doi.org/10.1016/j.bbcan.2018.12.005>.
10. Chen S, Ben S, Xin J, Li S, Zheng R, Wang H, et al. The biogenesis and biological function of PIWI-interacting RNA in cancer. *J Hematol Oncol*. 2021;14:93. <https://doi.org/10.1186/s13045-021-01104-3>.
11. Ramat A, Simonelig M. Functions of PIWI proteins in gene regulation: new arrows added to the piRNA quiver. *Trends Genet*. 2021;37:188–200. <https://doi.org/10.1016/j.tig.2020.08.011>.
12. Montañés JC, Rojano C, Ylla G, Piulachs MD, Maestro JL. siRNA enrichment in Argonaute 2-depleted *Blattella germanica*. *Biochimica et Biophysica Acta (BBA) Gene Regul Mech*. 2021;1864:194704. <https://doi.org/10.1016/j.bbagrm.2021.194704>.
13. Mito T, Ishimaru Y, Watanabe T, Nakamura T, Ylla G, Noji S, et al. Cricket: the third domesticated insect. In: Current topics in developmental biology. Academic Press; 2022. p. 291–306. <https://doi.org/10.1016/bs.ctdb.2022.02.003>.
14. Nakamura T, Ylla G, Extavour CG. Genomics and genome editing techniques of crickets, an emerging model insect for biology and food science. *Curr Opin Insect Sci*. 2022;50:100881. <https://doi.org/10.1016/j.cois.2022.100881>.
15. Gainetdinov I, Colpan C, Arif A, Cecchini K, Zamore PD. A single mechanism of biogenesis, initiated and directed by PIWI proteins, explains piRNA production in most animals. *Mol Cell*. 2018;71:775–790.e5. <https://doi.org/10.1016/j.molcel.2018.08.007>.
16. 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.
17. 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. <https://doi.org/10.1126/science.1140494>.
18. Houwing S, Berezikov E, Ketting RF. Zili is required for germ cell differentiation and meiosis in zebrafish. *EMBO J*. 2008;27:2702–11. <https://doi.org/10.1038/emboj.2008.204>.
19. Kawaoka S, Hayashi N, Katsuma S, Kishino H, Kohara Y, Mita K, et al. *Bombyx* small RNAs: genomic defense system against transposons in the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol*. 2008;38:1058–65. <https://doi.org/10.1016/j.ibmb.2008.03.007>.
20. Kawaoka S, Arai Y, Kadota K, Suzuki Y, Hara K, Sugano S, et al. Zygotic amplification of secondary piRNAs during silkworm embryogenesis. *RNA*. 2011;17:1401–7. <https://doi.org/10.1261/ma.2709411>.
21. Czech B, Hannon GJ. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends Biochem Sci*. 2016;41:324–37. <https://doi.org/10.1016/j.tibs.2015.12.008>.
22. Rosenkranz D, Zischler H. proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics*. 2012;13:5. <https://doi.org/10.1186/1471-2105-13-5>.
23. Jung I, Park JC, Kim S. PiClust: a density based piRNA clustering algorithm. *Comput Biol Chem*. 2014;50:60–7. <https://doi.org/10.1016/j.compbiolchem.2014.01.008>.
24. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd international conference on knowledge discovery and data mining. 1996. pp. 226–31.
25. Ray R, Pandey P. piRNA analysis framework from small RNA-Seq data by a novel cluster prediction tool—PILFER. *Genomics*. 2018;110:355–65. <https://doi.org/10.1016/j.jgeno.2017.12.005>.
26. Raabe CA, Tang T-H, Brosius J, Rozhdestvensky TS. Biases in small RNA deep sequencing data. *Nucleic Acids Res*. 2014;42:1414–26. <https://doi.org/10.1093/nar/gkt1021>.
27. Uhrig S, Klein H. PingPongPro: a tool for the detection of piRNA-mediated transposon-silencing in small RNA-Seq data. *Bioinformatics*. 2019;35:335–6. <https://doi.org/10.1093/bioinformatics/bty578>.
28. Gebert D, Neubert LK, Lloyd C, Gui J, Lehmann R, Teixeira FK. Large *Drosophila* germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Mol Cell*. 2021;81:3965–3978.e5. <https://doi.org/10.1016/j.molcel.2021.07.011>.
29. Piuco R, Galante PAF. piRNAdb: A piwi-interacting RNA database. 2021: 2021.09.21.461238. Available from: <https://doi.org/10.1101/2021.09.21.461238>.
30. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, et al. Characterization of the piRNA complex from rat testes. *Science*. 2006. <https://doi.org/10.1126/science.1130164>.

31. Yamashita T, Komenda K, Miłodrowski R, Robak D, Szrajer S, Gaczorek T, et al. Non-gonadal expression of piRNAs is widespread across Arthropoda. *FEBS Lett.* 2025;599:3–18. <https://doi.org/10.1002/1873-3468.15023>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.