piRAT

Version 0.1.0

- Documentation -

Table of Contents

Overview

piRAT is a Python package used for annotating PIWI-interacting RNAs (piRNAs), a class of small non-coding RNAs involved in genome defense and transposon silencing. It enables direct analysis of Next Generation Sequencing (NGS) data in .bam format without prior pre-processing. piRAT runs out-of-the-box on Linux and Window through docker container and can be launched with a single command.

Features

- Direct analysis of .bam files (no pre-processing required)
- Annotation of primary piRNA clusters and ping-pong amplification signatures
- Multithreading for improved performance
- Multi-file analysis
- Auto-tuning of clustering parameters (range, MinReads, Eps)
- Comprehensive HTML reports with visual summaries

Installation

You can install piRAT directly from the Python Package Index (PyPI):

```
pip install pirat
```

Or, from a local directory containing the setup.py file:

```
pip install .
```

Note: piRAT requires Python 3.8+ and samtools installed in your environment.

Quick Start

Once installed, all you need is your NGS data in .bam format (along with .bai index files – piRAT can generate these if they are missing). To begin analysis:

```
pirat -p <input_path>
```

You will then be prompted to detect the piRNA size range automatically. After detection, clustering parameters are optimized, clusters are identified, ping-pong signatures are detected, and quality metric are computed.

```
----------------piRAT----------------
piRAT 0.1.0
Input path: <input_path>
Output path: <output_path>

You didn't specify the range of size of piRNAs, do you want to automatically find
the range of size of piRNAs? [y/n]

y

Performing analysis of file: <sample_file>.bam...
Done!

Range of the size of piRNAs in bam file is found to be 27 to 28.

Do you want to input your own range based on plot found in
/home/dom/test_pirat/pirat/hm_somatic/? [y/n]:

y

Please input your range in this format: lower_limit,upper_limit

26,32

Adjusted range is from 26 to 32
Finding optimal clustering parameters...

Performing cluster finding with parameters:
        range of size of piRNAs: [26, 32],
        k: 10,
        eps: 720
        variation threshold: 3

Found 30 clusters!

Performing statistical analysis of found clusters...
100%|
          | 2/2 [00:03<00:00,  1.73s/it]

16 clusters out of 30 found clusters are high quality!

Performing final annotation...
piRNA clustering took 258.68 seconds!
```

By default, piRAT uses one thread and processes chromosomes/scaffolds sequentially.

For faster performance, specify more threads:

```
pirat -p <path_to_the_dataset_directory> -t 16
```

Note: piRAT will create an output directory for results. Avoid renaming or deleting files in this directory during analysis. Running piRAT multiple times on the same dataset/output folder will overwrite existing results.

Examples

Run cluster annotation only:

```
pirat -p <path> -m primary
```

Run cluster annotation and generate cluster plots:

```
pirat -p <path> -m primary -d
```

Run ping-pong signature detection:

```
pirat -p <path> -m secondary
```

Run complete analysis (clusters and ping-pong signatures annotation)

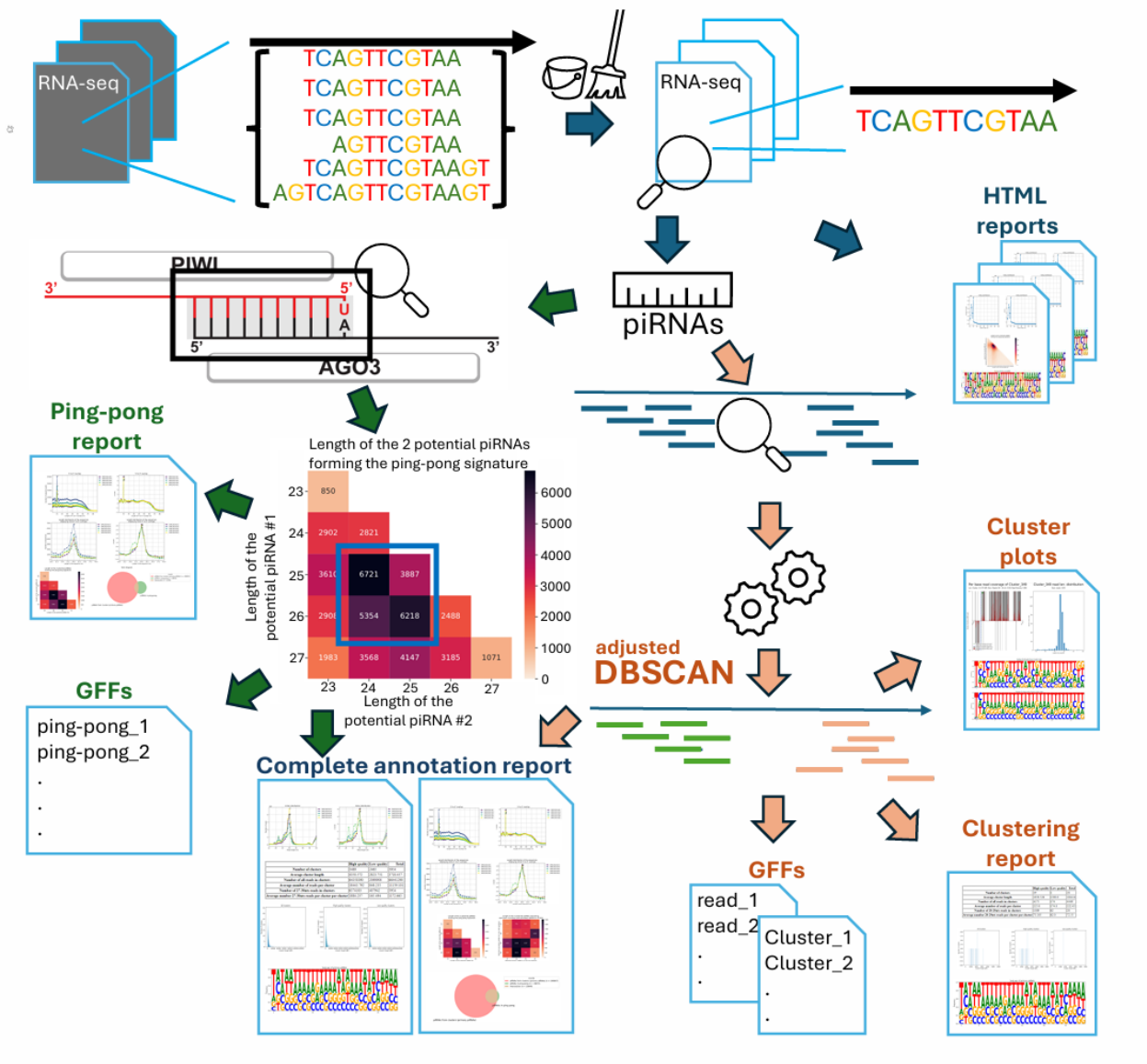
```
pirat -p <path> -m both
```

Whole analysis with predefined parameters

```
pirat -p <input_path> -o <output_path> -t 16 -a -d -r 28,30 -k 10 -e 800
```

- -a: Automatic mode (no interactive prompts)
- -d: Generate cluster plots
- -r: piRNA size range
- -k Minimum reads for cluster seed (one of clustering parameters)
- -e: Maximum clustering distance (eps)

How piRAT works



When you run piRAT, it performs the following steps:

Input Validation:

It scans the provided directory for .bam files and their corresponding .bai index files. If an index is missing, piRAT will attempt to generate it automatically using samtools. If this process fails, an error is raised indicating that given file is corrupted.

Pipeline Modules:

piRAT has three main pipelines:

- Sample-wise analysis: Scans each file for ping-pong signature existence.
- Clustering: Annotates primary piRNA clusters.
- Ping-pong signatures: Annotates read pairs displaying ping-pong signatures.

You can specify mode using the -m (or –module) flag:

- primary: Performs only piRNA cluster detection.
- secondary: Conduct sample-wise analysis and final ping-pong signature annotation.
- both (default): Runs both pipelines sequentially.

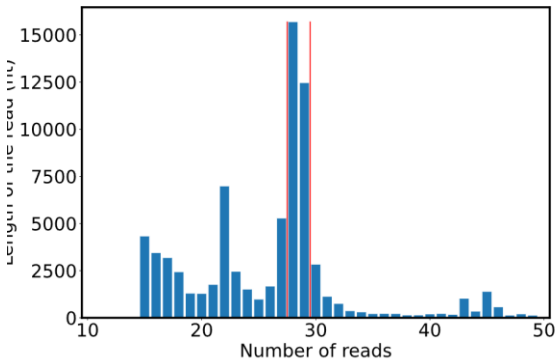
Detailed pipelines steps

Primary

Detection of size of piRNA reads in the dataset file (only clustering version)

piRAT randomly selects 100,000 mapped reads from each input file and computes a read length distribution. It then identifies the most abundant read length within the typical piRNA size range (26-32 nt). If adjacent lengths also show frequencies above a weighted mean threshold, they are included in the detected piRNA range.

The result is visualized in a histogram, which the user can review and optionally override piRNA size range in the analysis.



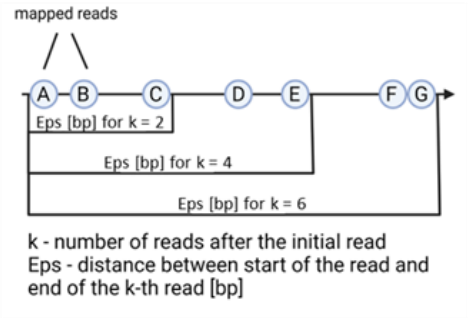
Pre-processing of reads

To mitigate sequencing errors and mapping artifacts, piRAT uses a viariation threshold (default: 3). For each read, it considers alternate mappings within ± 3 nt from the original start and stop positions. Among these candidates, the most abundant version is selected to represent that read during downstream analysis.

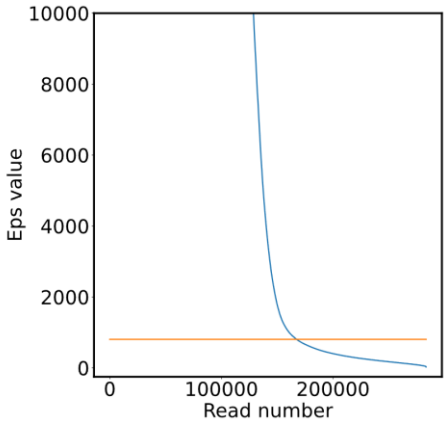
Detection of clustering parameters

piRAT uses a modified version of the DBSCAN clustering algorithm, optimized for piRNA data. It automatically determines two key parameters:

1. Eps: the maximum distance between k-th reads in a cluster
2. k (MinReads) – the minimum number of reads within Eps to define a cluster core



For k-values in {4, 6, 8, 10}, piRAT computes k-th neighbor distanced (filtered at <= 10,000 nt). These distances are sorted and segmented to find a "knee" point where the slop flattens – indicating a shift from high to low density. The Eps value at this elbow is selected for clustering.



To choose the best k, eps parameter pair, piRAT assesses whether the elbow region (curve under the found Eps value) is sufficiently flat. A "flat" region is defined as one where at least 50% of the elbow region have a slope change greater than -0.1.

If multiple k values satisfy this condition, piRAT selects the highest k, favoring denser clusters.

Clustering and DBSCAN modifications

In real-world sequencing data, coverage within a piRNA-producing region may not be uniform. Gaps in read density can occur due to sequencing errors, uneven library prep, or

mapping artifacts. As a result, biologically continuous piRNA clusters may appear fragmented when processed by traditional density-based algorithms like DBSCAN.

To account for these inconsistencies, piRAT modifies DBSCAN to tolerate short discontinuities in read coverage. A read that is withing Eps of a cluster core but doesn't meet the MinReads requirement is not immediately excluded. Instead, piRAT evaluates the neighborhood structure and read density continuity to all such reads to bridge otherwise disconnected but biologically unified regions. This ensured that sequencing artifacts do not artificially split true piRNA clusters.

Uni-strand, dual-strand, and bi-directional clusters classification

After strand-specific clustering, piRAT classifies cluster relationships:

- Dual-strand clusters: If clusters on opposite strand overlap by at least 50% of the smaller one or are fully nested, they are merged.
- Bi-directional clusters: If clusters are within Eps but not overlapping enough, the are marked as interacting.
- Uni-strand clusters: All other clusters.

Statistical analysis of annotated clusters

Annotated clusters are classified in two categories, low-quality and high-quality clusters, based on piRNA biogenesis information, such as the frequency of thymine (T) as the first nucleotide in the piRNA reads and the length of the reads mapped within an annotated cluster.

	High quality	Low quality	Total
Number of clusters	40	3	43
Average cluster length	5148.0	1734.667	4909.86
Number of all reads in clusters	20349	329	20678
Average number of reads per cluster	508.725	109.667	480.884
Number of 28-29nts reads in clusters	5861	152	43
Average number 28-29nts reads per cluster per cluster	146.525	50.667	139.837

piRAT classifies cluster as high-quality if:

- At least 50% of reads fall within the identified piRNA length range.
- At least 50% of reads begin with thymine (T), consistent with piRNA biogenesis.

These clusters are saved in *high_quality_clusters_out.gff*, while others are logged in *low_quality_clusters_out.gff*.

HTML report generation

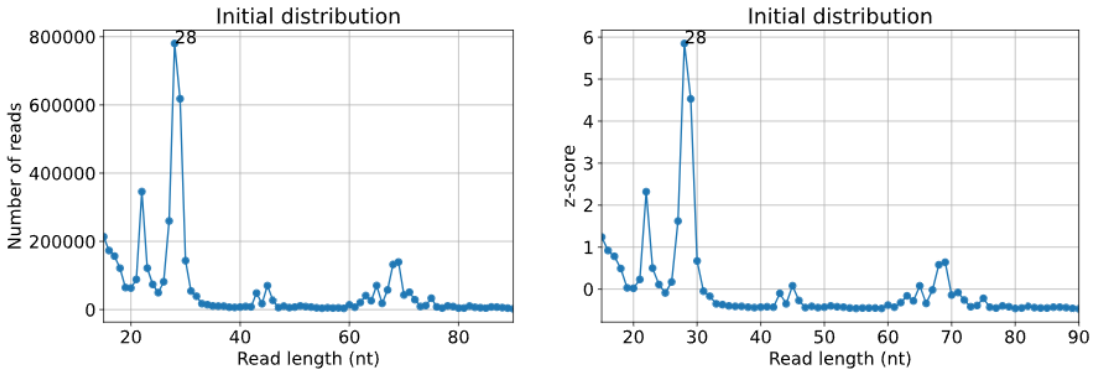
piRAT produces a comprehensive HTML report of the analysis of each files, which includes:

- Analysis timestamp and system info
- Input/output paths and dataset metadata
- Clustering configuration
- Statistics of found clusters
- Clusters length distribution
- SeqLogo of reads belonging to cluster lociSecondary

Sample-wise analysis

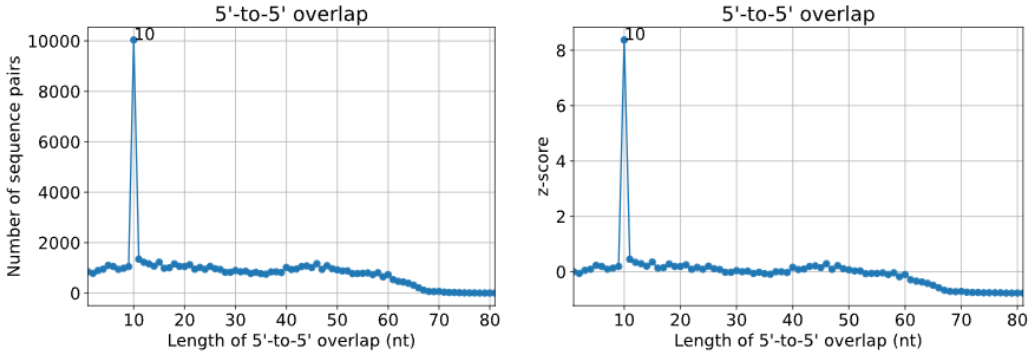
Read length distribution generation

All reads length is loaded to generate distribution of read lengths inside given file.



5'-to-5' overlap of reads analysis

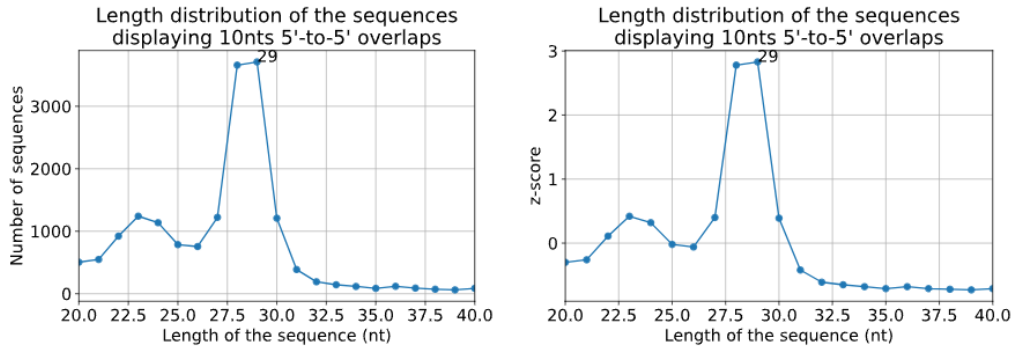
Among all reads, the length of overlap between two reads is calculated, and representative plot is generated



Note: A characteristic peak at 10 nucleotides in the overelap histogram is a hallmark of ping-pong amplification, indicating abundance of piRNA reads in the file.

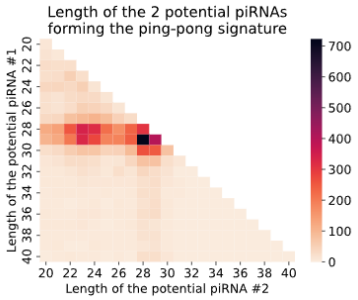
Length read distribution of reads displaying 10 nts 5'-to-5' overlaps

Among read pairs that exhibit a 10 nt 5'-to-5' overlap, piRAT analyzes the length distribution of the reads (within range of 20-40 nt). This helps to differentiate true ping-pong pairs from background noise and confirms the size of piRNA reads.



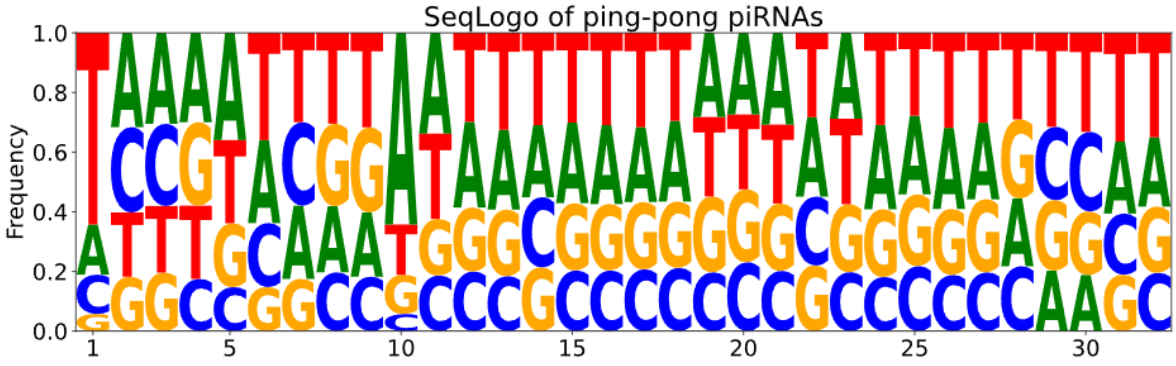
Heatmap generation of length of pairs of ping-pong reads interacting with each other

A heatmap is generated for all read pairs with a 10 nt 5'-to-5' overlap, showing the frequency of each length combination.



SeqLogo of ping-pong piRNAs generation

piRAt generates a SeqLogo from all potential ping-pong piRNA reads.



The resulting motif should display:

- Strong enrichment for Thymine (T) at position 1 – a hallmark of primary piRNAs.
- Adenosine (A) at position 10 – a signature of ping-pong biogenses.

HTML report generation

piRAT produces a comprehensive HTML report of the analysis of each files, which includes:

- Analysis timestamp and system info
- Input/output paths and dataset metadata
- Read length distribution
- 5'-to-5' overlap length distribution
- Length distribution of reads displaying 10 nts 5'-to-5' overlaps
- Heatmap of length distribution of ping-pong pairs
- SeqLogo of found ping-pong piRNAs

Ping-pong signatures annotation

Detection of size of piRNA reads in the dataset file

After gathering length distributions from all samples, piRAT determines a piRNA size range. Reads that contribute to 10 nt 5'-to-5' overlaps and show z-score above 0.5 (after averaging of all files) are chose as a piRNA size range.

Pre-processing of reads

As defined in Pre-processing of reads

Ping-pong signature detection

Ping-pong signature detection (10 nt 5'-to-5' ends) is searched among reads of defined/found range.

Analysis of found ping-pong piRNA reads

From found probable ping-pong reads, two heatmap are generated displaying

- Length of the two potential piRNAS forming the ping-pong signature
- Length of the two potential piRNAs, where each pair is separated for primary and secondary piRNA based on Thymine on first position and Adenozine (Uridine) on 10th position

Additionaly, SeqLogo of found ping-pong piRNAs is generated

HTML report generation

After analysis, html report is generated with included information about date, input path, analyzed file, read length distribution of files, 5'-to-5' overlap length distribution of files,

length distribution of reads displaying 10 nts 5'-to-5' overlaps of files, heatmaps of length pairs of ping-pong reads interacting with each other, and SeqLogo of found ping-pong piRNAs

Both

Apart from steps defined in two previous points, piRAT analyzes reads found in primary and secondary pathway together, displaying interaction through:

Venn diagram of primary and secondary pathway piRNAs

In combined mode, piRAT cross-validates ping-pong read pairs against cluster annotations. For each read in a ping-pong pair, it cheks whether it originates from a previously detected piRNA cluster. This integration reinforces condifence in the read's identity - ~ 50% of ping-pong reads should overlap clusters, reflecting canonical piRNA biogenesis from genomic hotspost.

HTML report generation

piRAT produces a comprehensive HTML report of the analysis, which includes:

- Analysis timestamp and system info
- Input/output paths and dataset metadata
- Cluster configuration and detected ranges
- Quality stats and length distribution of found clusters
- Heatmaps of detected ping-pong piRNAs
- Primary/secondary SeqLogos and Venn diagrams

Command line options

-p OR --path	Path to the directory with input data. Example: data/ (Required)
-o OR --output-path	Path to the directory for output files. Example: results/
-k OR --minreads	Clustering parameter defining minimum reads for a group of reads to be defined as a cluster core. Type: Integer
-e OR --eps	Clustering parameter of maximum distance between k-th reads to be considered continuous. Type: Integer
-r OR --range_of_size	Range of size of piRNAs. Format: min,max (e.g, 26,32)
-t OR --threads	Number of threads to use. Type: Integer

-m OR --module	Which module to run: primary, secondary, or both. (Default: both)
-v OR --variation_threshold	Variation threshold for cleansing reads. Type: Integer.
-a	Flag to run the analysis without prompting the user (User will be prompted only if there are any abnormalities, e.g. detected piRNA range is outside of typical piRNA size scope (26-32).
-d	Flag to generate cluster plots.
--plot_iter	Number of plots generated per iteration. Higher number speeds up the generation but requires more RAM. Default: 16. Type: Integer.
--version	Display version information and exit.

Troubleshooting

Corrupted files

If at least one of the files is corrupted (failed to be indexed), piRAT will stop running and display:

```
[E::sam_index_build3] SAM file "<sample_file>.bam" not BGZF compressed

Failed to index BAM file <sample_file>.bam in <input_path>: 'samtools returned with error 1: stdout=, stderr=samtools index: failed to create index for <sample_file>.bam"\n'
```

In this case, remove corrupted file and rerun the analysis.

RAM usage

While we aimed at piRAT to not be resource heavy, amount of resources mainly depends on the dataset (number of clusters, and their length, density).

In case you are running out of available RAM, the best solution would be to decrease number of threads (analysis is split by scaffolds/clusters for each thread, so with this change, number of scaffolds/clusters analyzed simultaneously will be lower, decreasing use of system memory). Another step in the clustering part of the analysis, would be to lower –plot iter (statistical analysis is dividing clusters on threads, where analysis of each cluster includes loading reads from original files).

Citation

When you use piRAT for piRNA annotation, please cite:

Dominik Robak, Guillem Ylla (2025). piRAT: piRNA Annotation Tool for annotation, analyzing, and visualizing piRNAs.

Contact

If you have any questions or comments or find any bugs in the software, please do not hesitate to contact us:

Guillem Ylla

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

Email: guillem.ylla@uj.edu.pl

Web: Laboratory of Bioinformatics and Genome Biology - Ylla Lab

Dominik Robak

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

Email: dominikrobak03@gmail.com