Title: *FishPi*: a bioinformatic prediction tool to link piRNA and transposable elements in zebrafish

Alice M. Godden[1], Benjamin Rix[1], Simone Immler[1]

[1] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom

Corresponding author:

Alice M. Godden Email: alice.godden@uea.ac.uk

Abstract

Background

Piwi-interacting RNAs (piRNA)s are non-coding small RNAs that post-transcriptionally affect gene expression and regulation. Through complementary seed region binding with transposable elements (TEs), piRNAs protect the genome from transposition, and therefore a tool to link piRNAs with complementary TE targets is needed. Tools like *TEsmall* can process sRNA-seq datasets to produce differentially expressed piRNAs and *piRScan* developed for nematodes can link piRNAs and TEs but it requires the user to know the target region of interest and work backwards.

Results

We have therefore developed *FishPi* to predict the pairings between piRNA and TEs. *FishPi* works with individual piRNAs or a list of piRNA sequences in fasta format. The software focuses on the piRNA:TE seed region and analyses reference TEs for piRNA complementarity. TE type is examined, counted and stored to a dictionary, with genomic loci recorded. Any updates to piRNA-TE binding rules, can easily be incorporated by updating the code underlying *FishPi*. *FishPi* provides a graphic interface, using *tkinter*, that requires the user to input piRNA sequences to generate comprehensive reports on piRNA:TE dynamics. *FishPi* can easily be adapted to other genomes opening the interpretation of piRNA functionality to a wide community.

Conclusions

Users will gain insight into genome age and *FishPi* will help further our understanding of the biological role of piRNAs and their interaction with TEs in a similar way that public databases have improved the access to and the understanding of the role of small RNAs.

Keywords
PiRNAs, transposons, evolution, zebrafish

Background

Environmental changes and related biotic and abiotic stressors can disrupt the balance and expression of small RNAs and transposable elements (TEs) in the genome [1]. TEs are parasitic DNA that can cause insertions in the genome, as well as deletions and genomic rearrangements [2]. To counter this activity, Piwi-interacting RNAs (piRNAs), have evolved to repress the action of these repetitive elements [3]. PiRNAs constitute a substantial fraction of small RNAs and are predominantly found in animal germ cells [4]. The main function of piRNAs is to silence TEs and repetitive elements in animal germ cells to preserve the integrity of the germline genome [5] and loss of TE silencing by piRNAs can lead to sterility for example in the nematode *Caenorhabditis elegans* [6]. TEs are parasitic and are not randomly distributed in the genome but have preferred regions for insertion in the genome, including distorted and bent DNA [7,8], or open chromatin regions [8,9]. Large populations of piRNAs are complementary to TE sequences and messenger RNAs (mRNA) [10]. In *C. elegans*, the 2nd to the 7th nucleotide at the 5' end of the piRNA [11] or the 2nd to the 8th nucleotide [12] have both been suggested to be the piRNA:TE seed region. Similarly, in *Mus musculus*, the 2nd to 8th nucleotide was a predicted seed region [13] and in *Drosophila melanogaster*, the first 10 nucleotides at the 5' end of the piRNA are predicted to be the piRNA:TE seed binding site [14]. Finally, in zebrafish and other teleost fish, the first 10 nucleotides at the 5' end of the piRNA are loaded into the Argonaute machinery to produce Ago2-bound piRNAs complementary to TEs [3,15].

TEs are known to be activated through environmental stressors. In *C. elegans* for example, thermal stress has been shown in mutant piRNA biogenesis strains that lack the ability to generate piRNAs. In this work nematodes were heat-treated and displayed transgenerational impairments for up to three subsequent generations: with thermal stress decreasing the rate of piRNA biogenesis. [16,17]. In *Drosophila,* the *mariner-Mos1* transposon is enriched following heat-stress but not following ultraviolet irradiation [18]. TEs are expressed in a tissue-specific manner and can affect transcription and gene expression [19], by generating new genes and

mRNAs through genomic rearrangements and *de novo* mutations, and cause alteration of regulatory networks and even trigger immune responses [7]. Understanding the interaction between piRNAs and their target activities is therefore key to understand the defence mechanisms in the germ line against the effects of environmental conditions.

In the zebrafish (*Danio rerio*) genome, small RNAs post-transcriptionally regulate gene expression and regulation [20]. Piwi proteins are predominantly expressed in the animal germ line and associate with piRNAs. The loss of Ziwi (zebrafish Piwi) leads to apoptosis of germ cells in embryonic development [3]. Ziwi and Zili zebrafish Piwi proteins provide a piRNA amplification system that is highly conserved. TEs represent over 50% of the zebrafish genome [21] and it is therefore key to improve our understanding of the mechanisms that evolved to allow an organism to regulate their activity. TEs have over 2000 families in the zebrafish and exploit the host's genome by dynamic transposition events. The mechanisms underlying this transposition can be divided into two categories: retrotransposons, these TEs synthesise an RNA intermediate that can insert itself into the host genome [22]. In contrast, DNA transposons cut-and-paste themselves into the host genome [7,22,23].

We therefore developed the software FishPi to provide a means to map zebrafish piRNAs against complementary TEs and link them to understand their biological association [10]. FishPi facilitates research into the function of piRNAs and accelerates analyses in the same way as microRNA databases like Targetscan have supported small RNA research [24]. FishPi provides a user-friendly graphic interface (Figure 1), which is based on the latest zebrafish reference genome and TE sequences and generates a comprehensive output that can be exported as a report including complementary TE counts, family level annotation, chromosomal and counts plots, with the option to export the full list of TEs and their sequences that are complementary to the piRNA sequence fed in by the user. The user can enter individual piRNAs or lists of piRNAs in fasta format. For zebrafish and other model organisms, a list of piRNA sequences are available at piRBase

(http://bigdata.ibp.ac.cn/piRBase/browse.php), [4] or piRNAclusterDB (https://www.smallrnagroup.uni-mainz.de/piRNAclusterDB/), [25].

Implementation

FishPi is a python-based programme and uses the following packages and versions: Python v3.11, Tkinter (included in Python v3.11 [26]), Pillow 10.0.1 [27], and Matplotlib 3.8.0 [28]. The code for the programme was written and tested using PyCharm Community Edition 2023.2.1 and is platform-independent, as it is designed to be executed from the Linux command line; however, the software can be run and used in PyCharm IDE. The software initially defines functions to assay piRNA TE complementarity, creating a dictionary to count the TEs by class. The families assigned to each class were checked using dfam.org (Table 1), [29].

To generate a fasta file of the TE sequences to use in place of the zebrafish reference genome GRCz11 TE sequences, a TE.bed file with clade was created by using the UCSC table genome browser [30] as follows: Vertebrate, group: Variation and Repeats, genome: zebrafish, assembly: May 2017 GRCz11, track: RepeatMasker, table: rmsk. To extract DNA sequences from the reference genome based on the co-ordinates supplied in the TE.bed file, the following command lines were entered:

```
bedtools getfasta -s -name -fi Danio_rerio.GRCz11.dna.primary_assembly.fa -fo
GRCz11.teseqs.use.fasta -bed GRCz11.teannotation.bed
```

The following command line serves to mask the sequences in the reference genome:

```
bedtools maskfasta -fi Danio_rerio.GRCz11.dna.primary_assembly.fa -fo
GRCz11.masked.fasta -bed GRCz11.teannotation.bed
```

To create the TE-merged reference genome [31]:

```
cat GRCz11.masked.fasta GRCz11.teseqs.1.fasta > GRCz11.teseqs.fishpi.fasta
```

This preparatory work was based on the pipeline described for PopoolationTE2 [32]. The Ensembl reference genome GRCz11 v107 was used in this project and can be retrieved here: https://ftp.ensembl.org/pub/release-107/fasta/danio_rerio/dna/ "Danio_rerio.GRCz11.dna.primary_assembly.fa.gz" [33]. These instructions can be adjusted for other reference genomes and allow the user to run the analyses on any available genome version.

Results & Discussion

FishPi successfully links piRNAs and TEs and is a helpful tool for anyone interested in studying the role of TEs and sRNAs in shaping the germline genome. FishPi is designed to be used by those with minimal coding experience. A detailed walk-through use and installation is listed in the readme on the git repository here: https://github.com/alicegodden/fishpi . The simple graphic user interface offers a point and click tool to rapidly gain insight into many piRNA sequences. Each time a new piRNA sequence is provided, the programme resets itself before providing the new results. As more research into piRNA:TE seed sites evolve, FishPi can be adapted and updated to reflect changes and updates to the reference genome. Alternatively, users can generate and use other reference TE genome files and use FishPi for other model systems. This makes FishPi particularly powerful for users looking at non-model systems. The software has a graphic user interface for FishPi (Figure 1), and the outputs can be exported as hi-resolution bar-charts and chromosomal plots presenting raw count data of the different target TE categories for presentation (Figure 1, Supplementary Files 2 & 3). Additionally, users can export a .csv file with all complementary TE mRNA sequences, along with genomic co-ordinates, allowing the user to conduct further downstream analyses and explore each class of TE individually (Supplementary File 1). Benchmarking data can be found in Supplementary File 4 for optimal running of the software. FishPi's advantage over other tools is that it can be easily updated and modified as more research into piRNA:TE seeds are

validated and will be fully customisable for future genome releases for other understudied species including fish and other model systems.

Conclusions

By analysing piRNA complementarity to a TE, we can infer recent genomic expansion but also adaptive evolution. For DNA class TEs, we can find footprints of evolutionary activity, as DNA TEs are often associated with past transposition events. In the zebrafish genome, DNA TEs comprise 40%, with retrotransposons comprising 10% zebrafish genome [23]. Currently available tools to link piRNA and TEs are limited and require knowledge of known targets or are restricted to a pre-defined genome and species due to piRNA binding rules [11-13,34]. Further analysis with RNA-seq data and tools such as RepeatMasker and RepeatLandscape would complement this by Kimura distance analysis [35]. Additionally, use of the Retroseq pipeline can highlight non-reference insertion mutations by TEs from high-coverage DNA-seq data and help predict TE activity [36]. In the future, other adaptations of the software could be made to look at other small RNAs. The software allows users to rapidly confirm predictions of complementary TEs from specific piRNAs. FishPi software is freely available and is distributed under GPL-3.0 licence.

Availability and requirements

Project name: FishPi

Project home page: https://github.com/alicegodden/fishpi/tree/fishpi

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python v3.11

License: GPL-3.0

Any restrictions to use by non-academics: GPL-3.0 license needed

Declarations

Ethics

No ethics approvals were required for this research project.

Consent for publication

All authors consent to publication.

Availability of data and materials

FishPi is available on the Git repository here: https://github.com/alicegodden/fishpi/tree/fishpi

Competing interests

Authors declare no financial or competing interests.

Funding

This project was funded by a Consolidator Grant from the European Research Council to SI (SELECTHAPLOID – 101).

Author's contributions

AMG conceived the idea for the project, AMG and BR wrote the software, BR tested the software, AMG wrote the manuscript and SI contributed to the writing and provided guidance throughout.

Acknowledgements

AMG would like to acknowledge Prof. Oliver Buckley from University of East Anglia for software publication and licensing advice.

List of abbreviations

mRNA- Messenger RNA

PiRNA- Piwi-Interacting RNA

TE- Transposable element

References

1 Godden, A. M. & Immler, S. The potential role of the mobile and non-coding genomes in adaptive response. *Trends Genet* (2022). https://doi.org/10.1016/j.tig.2022.08.006

2 Kofler, R. piRNA Clusters Need a Minimum Size to Control Transposable Element Invasions. *Genome Biol Evol* **12**, 736-749 (2020). https://doi.org/10.1093/gbe/evaa064

3 Houwing, S. *et al.* A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**, 69-82 (2007). https://doi.org/10.1016/j.cell.2007.03.026

4 Wang, J. *et al.* piRBase: a comprehensive database of piRNA sequences. *Nucleic Acids Res* **47**, D175-D180 (2019). https://doi.org/10.1093/nar/gky1043

5 Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103 (2007). https://doi.org/10.1016/j.cell.2007.01.043

6 Spichal, M. *et al.* Germ granule dysfunction is a hallmark and mirror of Piwi mutant sterility. *Nat Commun* **12**, 1420 (2021). https://doi.org/10.1038/s41467-021-21635-0

7 Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol* **19**, 199 (2018). https://doi.org/10.1186/s13059-018-1577-z

8 Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* **18**, 292-308 (2017). https://doi.org/10.1038/nrg.2017.7

9 Gebrie, A. Transposable elements as essential elements in the control of gene expression. *Mob DNA* **14**, 9 (2023). https://doi.org/10.1186/s13100-023-00297-3

10 Houwing, S., Berezikov, E. & Ketting, R. F. Zili is required for germ cell differentiation and meiosis in zebrafish. *EMBO J* **27**, 2702-2711 (2008). https://doi.org/10.1038/emboj.2008.204

11 Zhang, D. *et al.* The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes. *Science* **359**, 587-592 (2018). https://doi.org/10.1126/science.aao2840

12 Shen, E. Z. *et al.* Identification of piRNA Binding Sites Reveals the Argonaute Regulatory Landscape of the C. elegans Germline. *Cell* **172**, 937-951 e918 (2018). https://doi.org/10.1016/j.cell.2018.02.002

13 Gou, L. T. *et al.* Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell Res* **24**, 680-700 (2014). https://doi.org/10.1038/cr.2014.41

14 Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761-764 (2007). https://doi.org/10.1126/science.1146484

15 La, Y. *et al.* Identification and Characterization of Piwi-Interacting RNAs for Early Testicular Development in Yak. *Int J Mol Sci* **23** (2022). https://doi.org/10.3390/ijms232012320

16 Belicard, T., Jareosettasin, P. & Sarkies, P. The piRNA pathway responds to environmental signals to establish intergenerational adaptation to stress. *BMC Biol* **16**, 103 (2018). https://doi.org/10.1186/s12915-018-0571-y

17 Kurhanewicz, N. A., Dinwiddie, D., Bush, Z. D. & Libuda, D. E. Elevated Temperatures Cause Transposon-Associated DNA Damage in C. elegans Spermatocytes. *Curr Biol* **30**, 5007-5017 e5004 (2020). https://doi.org/10.1016/j.cub.2020.09.050

18 Jardim, S. S., Schuch, A. P., Pereira, C. M. & Loreto, E. L. Effects of heat and UV radiation on the mobilization of transposon mariner-Mos1. *Cell Stress Chaperones* **20**, 843-851 (2015). https://doi.org/10.1007/s12192-015-0611-2

19 He, J. *et al.* Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. *Nat Commun* **12**, 1456 (2021). https://doi.org/10.1038/s41467-021-21808-x

20 Wei, C., Salichos, L., Wittgrove, C. M., Rokas, A. & Patton, J. G. Transcriptome-wide analysis of small RNA expression in early zebrafish development. *RNA* **18**, 915-929 (2012). https://doi.org/10.1261/rna.029090.111

21 Huang, C. R., Burns, K. H. & Boeke, J. D. Active transposition in genomes. *Annu Rev Genet* **46**, 651-675 (2012). https://doi.org/10.1146/annurev-genet-110711-155616

22 Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491-500 (1985). https://doi.org/10.1016/0092-8674(85)90197-7

23 Chang, N. C., Rovira, Q., Wells, J., Feschotte, C. & Vaquerizas, J. M. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res* **32**, 1408-1423 (2022). https://doi.org/10.1101/gr.275655.121

24 Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4** (2015). https://doi.org/10.7554/eLife.05005

25 Rosenkranz, D., Zischler, H. & Gebert, D. piRNAclusterDB 2.0: update and expansion of the piRNA cluster database. *Nucleic Acids Res* **50**, D259-D264 (2022). https://doi.org/10.1093/nar/gkab622

26 Foundation, P. S. Python Language reference, vesion 3.11. https://python.org/downloads/release/python-3110/ (2022). https://doi.org/https://python.org/downloads/release/python-3110/

27 Clark, A. https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf (readthedocs, 2015).

28 Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* **9**, 90-95 (2007). https://doi.org/10.1109/MCSE.2007.55

29 Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**, 2 (2021). https://doi.org/10.1186/s13100-020-00230-y

30 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496 (2004). https://doi.org/10.1093/nar/gkh103

31 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010). https://doi.org/10.1093/bioinformatics/btq033

32 Kofler, R., Gomez-Sanchez, D. & Schlotterer, C. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Mol Biol Evol* **33**, 2759-2764 (2016). https://doi.org/10.1093/molbev/msw137

33 Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res* **50**, D988-D995 (2022). https://doi.org/10.1093/nar/gkab1049

34 Wu, W. S. *et al.* pirScan: a webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. *Nucleic Acids Res* **46**, W43-W48 (2018). https://doi.org/10.1093/nar/gky277

35 Smit, A. F., Hubley, R. & Green, P. RepeatMasker software (2013). https://doi.org/http://www.repeatmasker.org

36 Keane, T. M., Wong, K. & Adams, D. J. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**, 389-390 (2013). https://doi.org/10.1093/bioinformatics/bts697

Figures and Tables

Table 1- Transposable element families assigned to transposable element Class (verified using dfam.org).

Class	Transposable element family
DNA	hAT, Tc1, Tc-Mar, Harbinger, Enspm, Kolobok, Merlin, Crypton, PiggyBac, Dada, Zatar, Ginger, TDR, Polinton, Maverick, Acrobat, Looper, TZF, Angel, Mariner
LTR	Gypsy, DIRS, Ngaro, ERV, Pao, Copia, BEL, HERV, Bhikari
LINE	L1, L2, L1-Tx1, Rex-Babar, RTE, Penelope, Keno, Rex
SINE	Alu, tRNA-V-RTE
RC	Helitron
Satellite	BRSATI, MOSAT

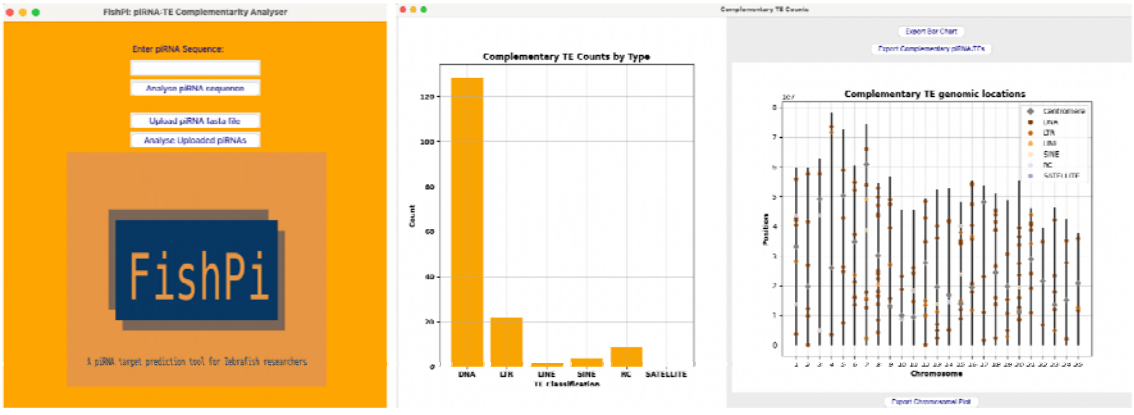

Figure 1- The graphic user interface of FishPi with and pop-out window with exportable results for dre-piRNA-53095 5'- CGGATAAGCATATTGCCCTAGCAACATC - 3'. The buttons included allow users to export the plots at hi-resolution for presentation and publication. Additionally, the user may export a .csv file list of all complementary TE mRNAs for further downstream analysis.