PLOS Computational Biology

RESEARCH ARTICLE

PymiRa: A rapid and accurate classification tool for small non-coding RNAs, including microRNAs

Zachary G. L. Scurlock[1], Cinzia G. Scarpini[1], Nicholas Coleman[1,2], Matthew J. Murray[1,3]*, Anton J. Enright[1]*

1 Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, United Kingdom, **2** Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge, United Kingdom, **3** Department of Paediatric Haematology and Oncology, Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge, United Kingdom

* mjm16@cam.ac.uk (MJM); aje39@cam.ac.uk (AJE)

Check for updates

OPEN ACCESS

Citation: Scurlock ZGL, Scarpini CG, Coleman N, Murray MJ, Enright AJ (2026) PymiRa: A rapid and accurate classification tool for small non-coding RNAs, including microRNAs. PLoS Comput Biol 22(3): e1014114. https://doi.org/10.1371/journal.pcbi.1014114

Editor: Adam Ewing, University of Queensland, AUSTRALIA

Received: August 29, 2025

Accepted: March 10, 2026

Published: March 26, 2026

Copyright: © 2026 Scurlock et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The code for the PymiRa tool is available on the GitHub page (https://github.com/ZScurlock/PymiRa). Additionally, the simulated data and analysis scripts to reproduce the findings are also found on the GitHub page. The biological datasets

Abstract

Small non-coding RNAs (sncRNA; < 200 nucleotide length) are of increasing research interest due to their key regulatory roles in a host of fundamental biological processes. For example, microRNAs (miRNAs), a specific class of sncRNAs, regulate gene expression through messenger RNA (mRNA) interactions, and their dysregulation is associated with disease. Classifying sncRNAs is an important bioinformatic task in small RNA-sequencing pipelines. Here we have developed an aligner called PymiRa, written in Python, to identify and quantify miRNAs from FASTA/FASTQ sequencing files. Unlike other approaches, PymiRa utilises a Burrows-Wheeler algorithm to align an input file against a reference hairpin precursor FASTA file derived from miRBase, the online miRNA registry, permitting up to two mismatches at the 3' end of a read. Previous tools used either a Burrows-Wheeler genome alignment or dynamic programming alignment to precursors; we demonstrate that combining both approaches yields improved results and efficiency. Importantly, the PymiRa aligner accounts for 3' post-transcriptional modifications to miRNAs that typically occur. PymiRa is a fast, accurate, and publicly accessible aligner available via GitHub and/or a webserver for sncRNA identification, including miRNAs, enabling accurate counts to be produced as part of a small RNA-sequencing pipeline. PymiRa will undergo relevant revisions over time e.g., with miRBase version updates. The PymiRa aligner will facilitate a deeper biological understanding of the landscape of sncRNA expression in normal physiological conditions and their dysregulation in disease states, including cancer.

Author summary

RNA-sequencing is a popular methodology for studying levels of RNAs in different biological samples, with large amounts of data generated. There is an

used are all publicly available. For the Kim et al., 2016 doi:10.1073/pnas.1602532113 miRNA biogenesis protein knockout (ko) dataset, the samples used were the A) Parental Wild type [available at National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) accession number SRR3174960], B) DROSHA protein knockout (SRR3174962), C) XPO5 (Exportin 5) protein knockout (SRR3174965), and D) DICER protein knockout (SRR3174967). This biological dataset can also found at the Gene Expression Omnibus (GEO) accession number GSE77989. For the Chen et al., 2024 doi:10.1186/s12864-024-10298-y small RNA-sequencing dataset from different developmental stages of Mus musculus testes, data was obtained from the NCBI SRA; BioProject accession number PRJNA849281. For the Oliver et al., 2021 doi:10.1093/plcell/koab280 small RNA-sequencing dataset from Arabidopsis thaliana across several pollen developmental stages; Uninuclear, Binuclear, Trinuclear, and Mature; data was available at the GEO accession number GSE132485. Finally, for the independent Chaves-Solano et al., 2025 doi:10.3389/fcell.2025.1692501 dataset, human placental small RNA-sequencing data was obtained from NCBI SRA accession number SRR35103152).

Funding: This work was funded through the Vice-Chancellor Award Biosciences Doctoral Training Partnership (DTP) PhD Studentship, University of Cambridge, UK, funded by the Doctoral Landscape Awards under UK Research and Innovation (UKRI), with support from the Waldmann fund (Department of Pathology, University of Cambridge, UK) for ZGLS. The funders had no role in study design, data collection and analysis, decision to publish, nor preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

increasing volume of research into small RNAs and how they may be used to detect disease and/or be the targets for new treatments. However, identifying these RNAs accurately and rapidly from sequencing data is challenging. Current methodologies often struggle to correctly identify these RNAs and take much longer to count them. Here we present PymiRa, a rapid, accurate, and accessible tool to identify small RNAs from sequencing experiments. By studying how small RNA levels are different in biological samples, this can help find new ways to detect and treat diseases.

Introduction

Identifying small non-coding (sncRNAs) from sequencing data is an increasingly important bioinformatic task. SncRNA have essential roles modulating cellular processes and regulating gene expression. A well characterised class of sncRNA are microRNAs (miRNAs), which are key post-transcriptional regulators of gene expression; their dysregulation is associated with pathology such as cancer and cardiovascular disease [1]. The potential of miRNAs as diagnostic or prognostic molecules make them compelling biomarker candidates and as a result they are the subject of substantial research interest [2–4]. Accessible tools enabling their rapid and accurate identification and quantification for differential expression analyses are therefore in increasingly high demand.

A common method for identifying miRNAs from next generation sequencing (NGS) data involves aligning short sequencing reads to the entire human genome ('whole genome alignment'), a task that requires large computational resources in terms of both memory and processing power. As sequencing experiments have become more commonplace, annotated small RNA databases have been compiled into centralised resources for researchers to use. One example is miRBase (https://www.mirbase.org/), the public database and registry for storing miRNA sequences from multiple organisms [5]. Consequently, an alternative approach to whole genome alignment is to directly align putative miRNA sequences against e.g., the miRBase database, requiring fewer computational resources in terms of processing, storage, and time. A successful example of this approach is Chimira [6], utilising dynamic programming to align sncRNAs to precursor sequences, and which demonstrated speed and computational benefits over other tools aligning to the entire genome. Unfortunately, however, Chimira is currently unsupported and unavailable to the public.

The Burrows-Wheeler Transformation (BWT) has been commonly utilised in many popular sequence aligners such as the Burrows-Wheeler Aligner (BWA) and Bowtie2 [7,8]. The combination of the lossless compression method of the BWT, coupled with the use of a Ferragina-Manzini index (FM-index), creates a memory-efficient approach for indexing and searching an entire genome [9]. Using a backward search, these algorithms are able to efficiently align sequencing reads against a reference sequence (genome) whilst accounting for mismatches, making them ideal for aligning short reads [7]. Aligning miRNAs can be challenging however, requiring flexibility in

specific 3' regions to account for subsequent sequence modifications. Typically, these occur via adenosine deaminases that act on RNA (ADAR) editing transcripts, as well as other post-transcriptional modifications that typically occur at the 3' end such as terminal uridylation by TUTase enzymes [10,11]. To be effective, alignment tools need to be able to accurately account for such modifications.

Here, we have developed 'PymiRa', a rapid, robust, and accessible sequence aligner for miRNA identification and quantification using a BWT-based algorithm, written in Python. The aligner inputs a sequencing file (FASTA/FASTQ.gz) that identifies and quantifies miRNAs by aligning to a species-specific miRNA hairpin FASTA file derived from miRBase, containing the precursor hairpin sequences mature miRNAs are processed from. PymiRa allows up to two mismatches at the 3' end of a read, as is good practice in other miRNA alignment methodologies [6]. However, PymiRa also offers utility for identification of other sncRNA classes, by e.g., allowing changes to the reference database for alignment, providing a fast and consistent approach to identifying the full landscape of sncRNA expression from sequencing experiments.

To assess the ability of PymiRa to accurately identify and quantify mature miRNAs from RNA-sequencing data, it was tested on both simulated and real biological datasets alongside the commonly used aligners Bowtie2, Chimira, and miRDeep2.

Results

Simulated dataset

First, we sought to establish the ability of each aligner (Bowtie2, Chimira, miRDeep2, and PymiRa) to identify and quantify miRNAs across the whole simulated dataset (Materials and Methods). To ensure comprehensive benchmarking, Bowtie2 was run under a number of different conditions [default settings, i.e., standard *(std)*, --very-sensitive-local (*vsl)*, and --local -N1 (*local)*], and, furthermore, coupled with two popular RNA read counting tools, namely HTSeq [12] and FeatureCounts (FC) [13]. For clarity, the known count of actual mature miRNAs from the simulated dataset is subsequently referred to throughout the text as 'Real'. In this analysis, each aligner's counts were normalised to the 'Real' count, where fewer deviations from zero (red line) indicated a more accurate count (Fig 1A). PymiRa demonstrated excellent concordance with the Real count, reflected in a very high Pearson's *r* score (0.982) and with only relatively small deviations. Chimira and miRDeep2 similarly performed well and achieved counts close to the Real count for many miRNAs, reflected by high Pearson's *r* scores of 0.962 and 0.967, respectively. PymiRa's correlation with the Real count was found to be significantly greater than Chimira's ($r = 0.982$ *vs.* $r = 0.962$; $p < 0.001$; Steiger's test); there was however no difference between PymiRa and miRDeep2 ($r = 0.982$ *vs.* $r = 0.967$; $p = 0.9255$; Steiger's test). These three aligners identified the most miRNAs, achieving a minimum of 92% of the Real miRNA count (Table 1A). Bowtie2 demonstrated the largest differences to the Real count of all the aligners assessed, with substantial differences across most miRNAs observed, reflected in a maximum Pearson's *r* score of only 0.84, achieved using *std* settings with FC (Fig 1A). Interestingly, FC appeared to quantify marginally more miRNAs correctly than HTSeq, with improvements in Pearson's *r* scores for Bowtie2 of ~0.04. Additionally, the Bowtie2 condition with the highest miRNA count (Bowtie2_*std*_FC) still represented the lowest number of miRNAs identified from all the aligners, missing >15% of the total (Table 1A). Next, the Root Mean Square Error (RMSE) assessment was performed, which showed that on average, PymiRa's counts were closest to the Real count (228), compared with the counts from the best performing Bowtie2 condition (Bowtie2_*std*_FC), which were on average off by 610 (Table 1A). However, as expected, all aligners struggled to produce accurate counts for miRNAs from the highlighted miR-548 family, which are associated with transposable elements and, accordingly, arise from multiple different genomic loci [15] (Fig 1A and 1B). The close sequence homology shared between miRNA members of the miR-548 family (S1 Table) make accurate counts typically near impossible for any alignment tool.

To determine which of the three Bowtie2 conditions (*std, vsl*, and *local*) should be evaluated further, ternary plots were produced, representing the proportion of three variables, namely the performance of each condition being studied, that

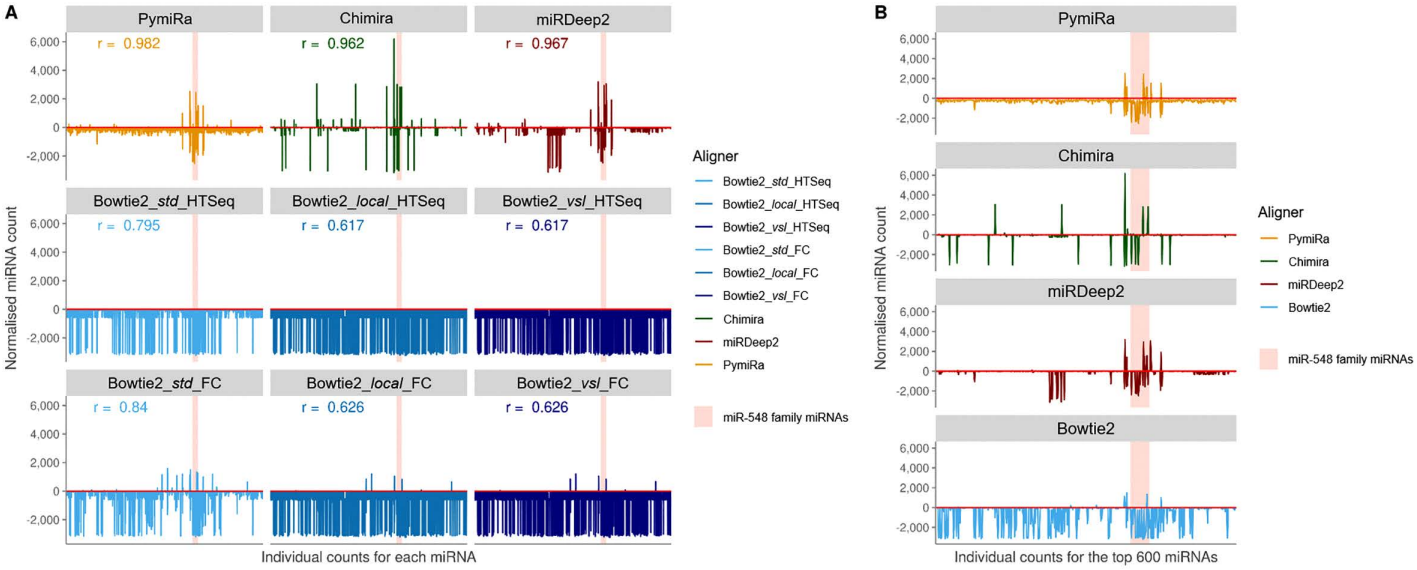

Fig 1. Normalised miRNA counts to the Real (actual) count, showing the differences from each aligner (PymiRa, Chimira, miRDeep2, and Bowtie2) on A) the whole simulated miRNA dataset, and B) the top 600 over-expressed miRNAs from the simulated miRNA dataset. The simulated 10 million read dataset was comprised of three million real miRNA sequences and seven million shuffled control sequences. The red horizontal line at y = 0 indicates zero deviation from the Real counts. Deviations from the red line indicate a difference in each aligner's count for each individual specific miRNA, compared with the respective count obtained from the Real simulated miRNA dataset. Of the real miRNA sequences, 600 of these were randomly selected to be given increased counts to simulate a disease model where there is miRNA dysregulation (i.e., over-expression of specific miRNAs). Pearson's *r* scores were calculated for each aligner comparing all the aligner's counts with the Real counts. Bowtie2 was run with different settings/conditions, combined with either HTSeq or FeatureCounts (FC). The settings used were standard default (*std*), *--very-sensitive-local* (*vsl*) and *--local -N1* (*local*). Bowtie2 (*std*_FC) was used to represent Bowtie2 for evaluating performance for the top 600 miRNAs due to superior performance to other conditions. The average Real count per miRNA was ~ 1,140. The miR-548 family of miRNAs is highlighted, as all aligners struggled to generate accurate counts for these miRNAs, as expected and as described in the text, due to sequence homology (S1 Table).

https://doi.org/10.1371/journal.pcbi.1014114.g001

summed to a constant value (i.e., the total counts of a miRNA) (S1A and S1B Fig). Points located near the centre of the plot showed an approximately equal distribution across each variable (condition) (e.g., ~ 33% each). In contrast, points located along an edge between two variables on the plot indicated similar proportions for those two aligners, with little contribution from the third variable (condition). The cluster of points at ~33% for each variable but also at the Bowtie2_*std* settings showed that the default standard setting (*std)* produced miRNA counts that the other Bowtie2 settings/conditions missed. The outcome was the same across the different quantification tools HTSeq and FC (S1A and S1B Fig). Bowtie2 with default settings combined with FeatureCounts (Bowtie2_*std*_FC) was therefore used to represent Bowtie2 for the remainder of the benchmarking, as it produced the optimal Bowtie2 performance based on the total number of miRNAs aligned (Table 1A), and the Pearson's *r* correlation scores (Fig 1A).

Next, the top 600 over-expressed miRNA counts from the simulated dataset (Materials and Methods) were normalised against the Real count and plotted for each aligner (Fig 1B). The vast majority of PymiRa's (yellow track) counts were at similar levels to the Real data, with few deviations, resulting in the lowest RMSE of 458 (Table 1B). PymiRa's multi-mapping behaviour was demonstrated by this low-level discrepancy seen across the dataset whilst still providing a very representative count, shown by the very high Pearson's *r* score (Fig 1A). Chimira (green track) and miRDeep2 (red track) also had few deviations but were typically seen at greater amplitude than PymiRa, reflected in higher RMSE's of 577 and 533, respectively. Chimira's multi-mapping methodology of assigning all reads to the single top-ranking alignment was illustrated here by a particular miRNA (hsa-miR-520c-3p), which had over 6,000 more counts than observed in the Real data. MiRDeep2, which also used fractional counting here (*-W* flag), displayed similar characteristics to PymiRa but

Table 1. Performance of the four aligners PymiRa, Chimira, miRDeep2, and Bowtie2 in the 10 million read simulated dataset. A) Number of mature miRNAs identified and their Root Mean Square Error (RMSE). B) RMSE of the miRNAs identified by each aligner in the top 600 'over-expressed' simulated miRNAs. Of the real miRNA sequences, 600 of these were randomly selected to be given increased counts to simulate a disease model where there is miRNA dysregulation (i.e., over-expression of specific miRNAs).

A

Aligner	MiRNA count	Percentage of Real miRNA achieved (%)	Root mean square error (RMSE)
PymiRa	2,759,725	92.0%	228
Chimira	2,974,604	99.2%	294
miRDeep2	2,955,867	98.5%	271
Bowtie2_*std*_HTSeq	2,357,752	78.6%	700
Bowtie2_*local*_HTSeq	1,452,082	48.4%	1,029
Bowtie2_*vsl*_HTSeq	1,452,526	48.4%	1,029
Bowtie2_*std*_FC	2,503,190	83.4%	610
Bowtie2_*local*_FC	1,492,774	49.8%	1,013
Bowtie2_*vsl*_FC	1,493,396	49.8%	1,013

B

Aligner	Root mean square error (RMSE) to the Real count
PymiRa	458
Chimira	577
miRDeep2	533
Bowtie2_*std*_FC	1,013

The dataset contained three million mature miRNA sequences and seven million shuffled sequences using the *uShuffle* software [14] that acted as control sequences. The dataset was also subject to a mismatch and an A/U base insertion probability of 10% to add further noise (see Materials and Methods). The RMSE was calculated using $\sqrt{(\Sigma (Ai - Ri)^2 / n)}$, where A = Aligner count, R = Real count and n = Number of unique miRNAs.

https://doi.org/10.1371/journal.pcbi.1014114.t001

undercounted some specific miRNAs (Fig 1B). Bowtie2 (blue track) failed to align many of the top 600 miRNA counts, with frequent count deviations below the Real count, representing overall low alignment performance. Interestingly, Bowtie2 either achieved similar counts to Real or failed to align the miRNAs completely in an almost 'binary' fashion (Fig 1B) which was reflected in a higher RMSE of 1,013.

As previously demonstrated, the aligners all struggled with identifying accurate miR-548 family miRNA counts, which had been randomly selected in the top 600 over-expressed miRNAs (Materials and Methods), but to different extents. Bowtie2 struggled to assign any reads to multiple miRNAs in the miR-548 family, with Chimira, miRDeep2, and PymiRa all giving closer results to the Real count and achieving similar counts for some miRNAs (S2 Fig).

For further comparison of how each aligner performed compared with the Real count, ternary plots were produced. The points between Real, PymiRa, and Chimira at ~33% for each variable demonstrated that there was strong concordance in miRNA counts. However, there was also a cluster of points at the PymiRa-Real edge, indicating counts identified by the PymiRa aligner and in the Real dataset but not by Chimira. This was similarly found with PymiRa, Real, and miRDeep2, highlighting the better performance of PymiRa over Chimira and miRDeep2 (Fig 2A and 2B). Conversely, there were very few points on the miRDeep2-Real edge, indicating very few instances where PymiRa had a low miRNA count (Fig 2B). Chimira, miRDeep2, and Real all shared similar counts with an intensity of points at ~33% but with groups on the axes both indicating counts that either aligner missed (Fig 2C). However, with all of the Bowtie2 counts, strong lines of points were shown, all indicating miRNAs that Bowtie2 missed but which were counted by the other two variables (aligners) (Fig 2D-2F).

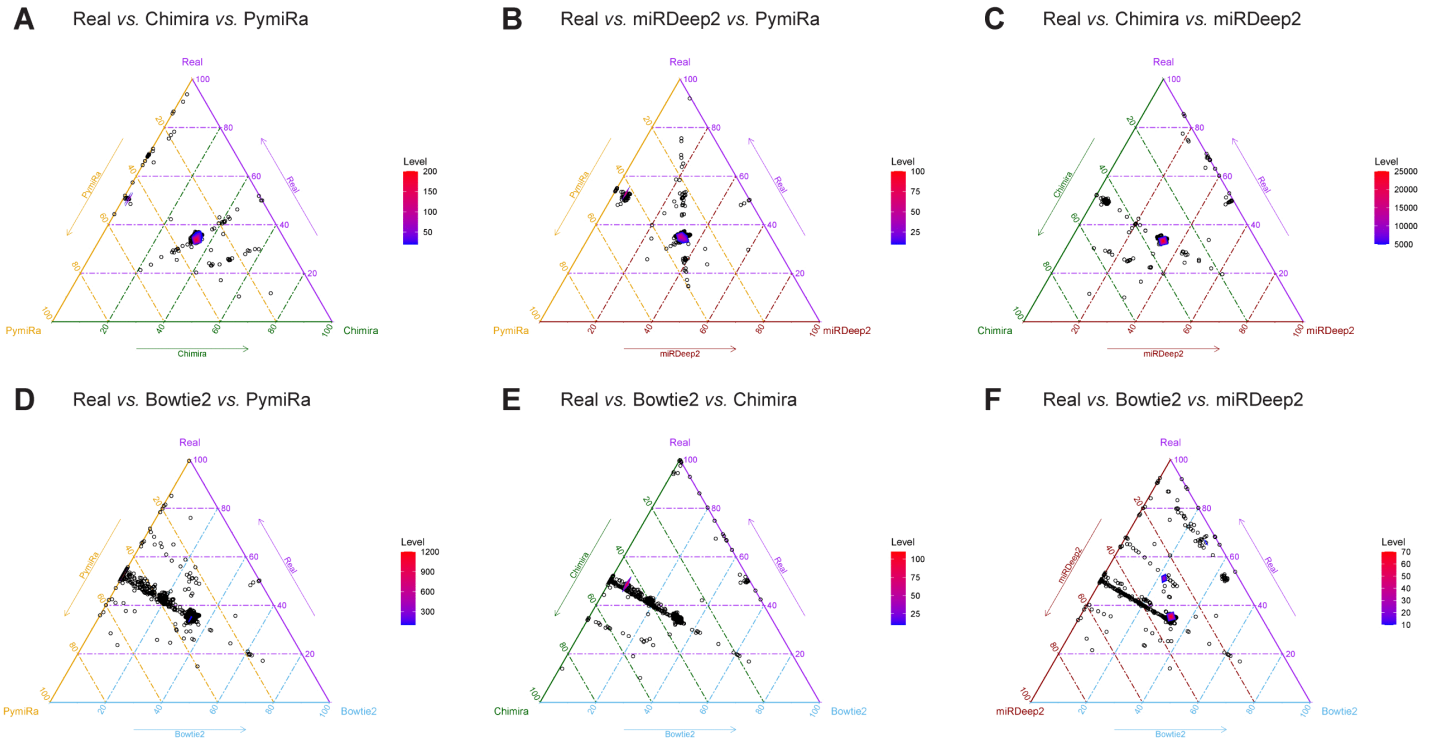

Fig 2. Ternary plots of the whole simulated miRNA dataset comparing the Real counts with the counts achieved by each aligner (PymiRa, Chimira, miRDeep2, and Bowtie2). The numbering on each edge represents the proportion of counts achieved by each variable (aligner). MiRNA counts with an equal proportion were found at ~33%. Points found on an edge between two variables represent miRNA counts achieved by two variables (aligners) with little contribution from the third variable (aligner). A two-dimensional kernel density estimation was calculated for each plot, where a high density of counts was found in an increasing spectrum of blue to red ('Level') to visualise areas of overlapping points. Comparing the Real miRNA counts with **A)** Chimira and PymiRa; **B**) miRDeep2 and PymiRa; **C)** Chimira and miRDeep2; **D)** Bowtie2 (*std*_FC) and PymiRa; **E)** Bowtie2 (*std*_FC) and Chimira; **F)** Bowtie2 (*std*_FC) and miRDeep2. Bowtie2 (*std*_FC) was used to represent Bowtie2 due to superior performance over the other settings/conditions evaluated.

https://doi.org/10.1371/journal.pcbi.1014114.g002

Biological datasets

Next, the top three performing aligners were selected for benchmarking in a biological dataset, based on results from the simulated dataset. Specifically, aligners with the highest Pearson's *r* score (Fig 1A), the highest miRNA counts (Table 1), and the lowest RMSE scores (Table 1) were selected, namely PymiRa, Chimira, and miRDeep2.

To assess the capability of quantifying miRNAs in a biological dataset, FASTQ files from Kim *et al.*, 2016 were used [16]. This small RNA-sequencing dataset was selected as it comprised miRNA biogenesis protein knockout (*ko*) experiments; namely DROSHA, Exportin5 (XPO5), and DICER. This approach allowed for a qualitative and quantitative evaluation of the performance of each aligner. Due to different miRNA biogenesis mechanisms being selectively affected, the levels of particular miRNAs varied across the different protein knockout experiments e.g., DICER-independent miRNAs could be assessed in the DICER *ko* experiment [17]. After adapter trimming, quality filtering, and the removal of reads outside of the 15–50 nucleotide (nt) range with Cutadapt, miRNAs were identified and quantified by each aligner. For each condition, ternary plots were made with each aligner on an edge to assess the similarity of mature miRNA counts.

Specifically in the Parental (non-*ko*) sample, a cluster was found at ~33% for each aligner, indicating similar counts for these miRNAs (Fig 3A). For all three *ko* conditions studied, there were miRNAs for which any two aligners had more similar counts to each other than the third aligner, seen as lines across the middle of each plot (Fig 3). Importantly, points

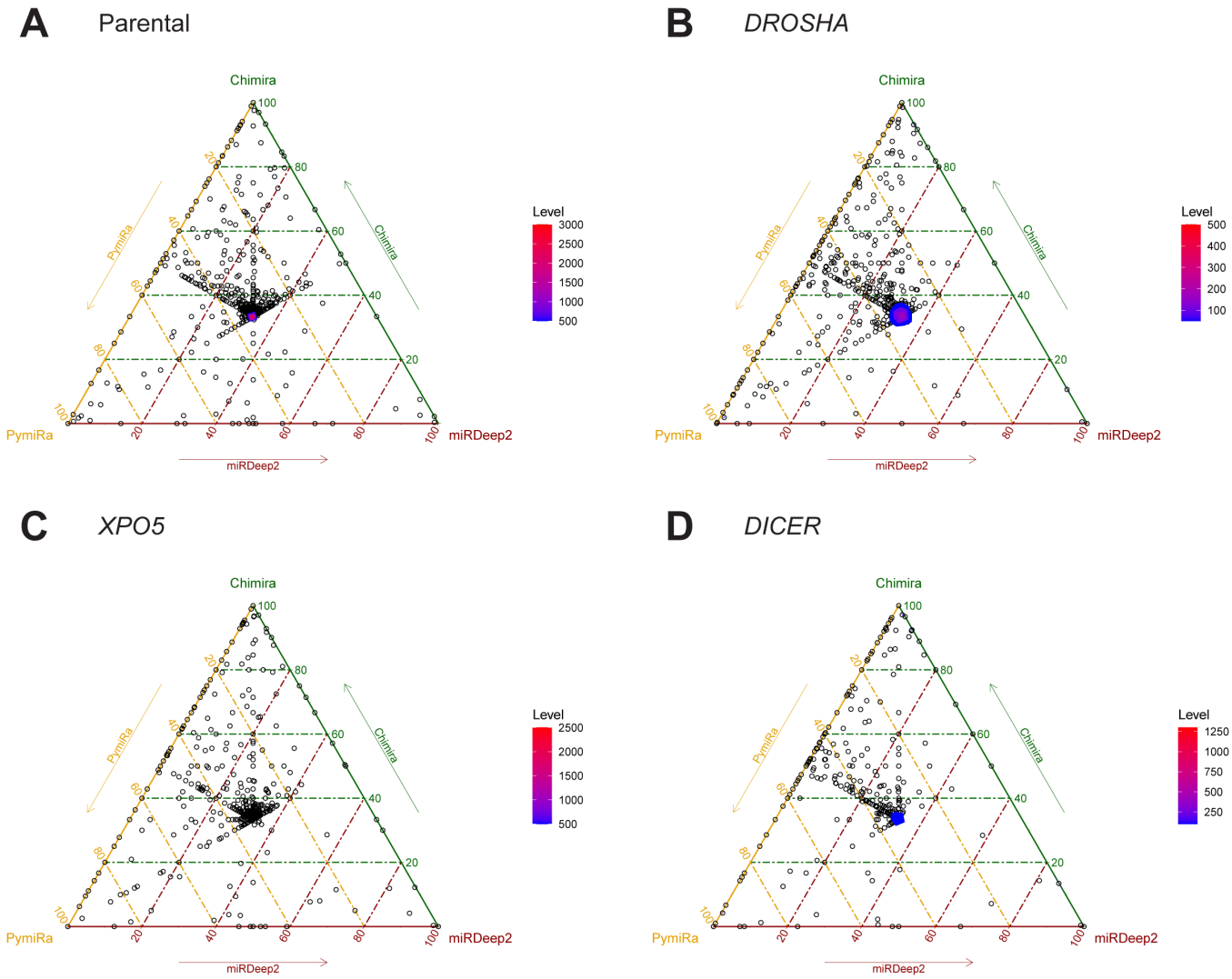

Fig 3. Ternary plots of the miRNA counts obtained from a biological dataset by each aligner (PymiRa, Chimira, and miRDeep2). For this work, the Kim *et al*., 2016 miRNA biogenesis protein knockout (*ko*) dataset was utilised. The samples used were the **A)** Parental Wild type [available at National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) accession number SRR3174960], **B)** DROSHA protein knockout (SRR3174962), **C)** XPO5 (Exportin 5) protein knockout (SRR3174965), and **D)** DICER protein knockout (SRR3174967). The numbering on each edge represents the proportion of counts achieved by each variable (aligner). MiRNA counts with an equal proportion were found at ~33% (i.e., in the centre of the plot). Points found on an edge between two variables represent miRNA counts achieved by two variables (aligners) with little contribution from the third variable (aligner). A two-dimensional kernel density estimation was calculated for each plot, where a high density of counts is found in an increasing spectrum of blue to red ('Level') to visualise areas of overlapping points. This original biological dataset can also be found at the Gene Expression Omnibus (GEO), accession number GSE77989.

https://doi.org/10.1371/journal.pcbi.1014114.g003

were also found directly at the periphery, primarily on the PymiRa-Chimira edge, indicating that both PymiRa and Chimira identified miRNAs that miRDeep2 failed to map in *ko* datasets. This was observed across all three *ko* experiments, indicating that PymiRa and Chimira more accurately identified miRNA counts from real biological data.

To further illustrate the accuracy of PymiRa in miRNA alignment, we investigated the fold changes of specific miRNAs from different *ko* conditions and compared our findings with those generated by Kim *et al*., 2016 [16]. The fold change between the DROSHA *ko* and Parental samples was calculated (S3 Fig). Following the same normalisation methodology

described in Kim *et al*, using endogenous hsa-miR-320a-3p, the highly similar miRNA fold changes (Pearson's $r = 0.994$) between the DROSHA *ko* and Parental samples, as observed in the original manuscript, further demonstrated the accuracy of PymiRa, despite increased miRNA counts, highlighting PymiRa's ability to maintain the proportions of miRNAs observed (Table 2). The reproducibility of the fold changes using PymiRa provided further confirmation of its ability to accurately identify subtle changes in miRNA expression/counts in biological datasets.

Next, we wanted to confirm PymiRa's ability in aligning miRNAs from different non-human species for which a reference miRBase precursor hairpin dataset was available i.e., *Mus musculus* (mouse) and *Arabidopsis thaliana* (thale cress). Small RNA-sequencing data was obtained from Chen *et al.,* 2024, who profiled small RNAs from different developmental stages of *Mus musculus* testes [18], available at the NCBI SRA; BioProject accession number PRJNA849281. After adapter trimming and quality controlling the reads using FastQC and Cutadapt (see Materials and Methods), PymiRa was run to identify miRNAs from mice on two sequencing samples, one from 3-week-old (pubertal) testis and one from 11-week-old (adult) testis. PymiRa achieved an excellent correlation to the Chen *et al.,* 2024 datasets, with very high Pearson's *r* scores of 0.98 and 1.0 for the 3-week-old and 11-week-old data, respectively, across the top 25 most abundant miRNAs (S4A and S4B Fig). PymiRa accurately recreated the top 25 miRNA counts for both datasets as ranked by abundance from Chen *et al.,* 2024 (3-week-old: $rho = 0.93$; $p < 0.005$; 11-week-old: $rho = 0.97$; $p < 0.005$; Spearman's rank correlation) showing its utility in identifying miRNAs from other species (Tables A and B in S2 Table).

Small RNA-sequencing samples were also obtained from Oliver *et al.,* 2021, where the authors profiled miRNAs from *Arabidopsis thaliana* across several pollen developmental stages; Uninuclear, Binuclear, Trinuclear, and Mature [19]; available at the GEO accession number GSE132485. After adapter trimming and quality controlling the reads using FastQC and Cutadapt, PymiRa identified miRNAs and achieved an excellent correlation to the counts obtained by Oliver *et al.,* 2021, with Pearson's *r* scores ranging from 0.92 to 0.99 (S5A-S5D Fig). PymiRa was also able to accurately detect the top 25 ranking miRNAs by abundance across these positions (Uninuclear: $rho = 0.84$; $p < 0.005$; Binuclear: $rho = 0.88$; $p < 0.005$; Trinuclear: $rho = 0.91$; $p < 0.005$; Mature: $rho = 0.83$; $p < 0.005$; *Spearman's rank correlation)* (Tables A-D in S3 Table), further exemplifying PymiRa's capabilities in sncRNA detection across species.

Quantitative benchmarking

For any bioinformatic algorithm/tool to be widely usable and functional, both rapid and accurate miRNA identification and counting is essential. To ensure that PymiRa performed at a comparable computational time when compared with the other alignment pipelines being investigated, the time to align different FASTA files and quantify

Table 2. Top miRNA fold changes (>0.2) and miRNA counts between the DROSHA knockout (*ko)* and Parental samples found by Kim *et al.,* 2016 (left) and by using PymiRa (right).

Kim *et al*., 2016				PymiRa		
MiRNA	**DROSHA *ko***	**Parental**	**FC (DROSHA *ko*/Parental)**	**DROSHA *ko***	**Parental**	**FC (DROSHA *ko*/Parental)**
hsa-miR-7706-3p	1,160	437	2.654	1,634	615	2.656
hsa-miR-877-5p	2,055	1,708	1.203	3,356	2,902	1.156
hsa-miR-3615-3p	959	851	1.127	1,737	1,481	1.172
hsa-miR-320a-3p	5,925	5,925	1.000	9,092	9,092	1.000
hsa-miR-484-3p	1,703	1,801	0.946	2,613	3,050	0.856
hsa-miR-320b-3p	454	963	0.471	1,428	859	0.601

The fold change (FC) was calculated by first normalising the miRNA counts by levels of the endogenous hsa-miR-320a-3p (as conducted by Kim *et al.,* 2016) and then dividing the specific miRNA counts found in the DROSHA *ko* sample by those found in the Parental sample. Of note, only six miRNAs were identified with a fold change of >0.2, as expected, given the DROSHA ko condition, affecting global miRNA biogenesis.

https://doi.org/10.1371/journal.pcbi.1014114.t002

miRNAs was benchmarked. The time recorded was the total time elapsed, including the alignment and counting of the reads. For quantitative purposes only, the best-performing Bowtie2 variant (Bowtie2_*std*_FC) was also included in this analysis.

For the 10 million read simulated dataset, the fastest was Bowtie2 *std*_FC, which was completed in <100 seconds. However, when considering the number of miRNAs aligned, Bowtie2 aligned far fewer miRNAs as a result, with PymiRa identifying 256,535 more miRNAs (a 10% increase) than Bowtie2 in a total of 127 seconds (Fig 4A). MiRDeep2 and Chimira both aligned 196,142 and 214,879 more miRNAs than PymiRa, respectively, but subsequently took 21.4 and 124.2 seconds longer. Despite the differences in the number of miRNAs aligned, PymiRa had both the highest Pearson's *r* score and lowest RMSE from the simulated datasets, providing confidence that PymiRa was still able to produce representative counts (Fig 1A and Table 1). For the human biological dataset across the majority of Kim *et al.,* 2016 samples, PymiRa ran fastest of the aligners whilst still providing representative miRNA counts. MiRDeep2 and Bowtie2 took similar amounts of time to PymiRa but both missed miRNAs that PymiRa identified, highlighting PymiRa as a competitive aligner, capable of accurate identification of miRNAs at a rapid rate. Comparatively, Chimira was slower than the other aligners assessed, but was consistent in identifying miRNAs (Fig 4).

Finally, PymiRa is made freely publicly accessible via a webserver, available at https://www.pymira.co.uk. FASTA/FASTQ files may be uploaded and miRNA counts, log files and an alignment summary are generated for download. It is also available on GitHub for easy installation, only requiring Python with very few dependencies.

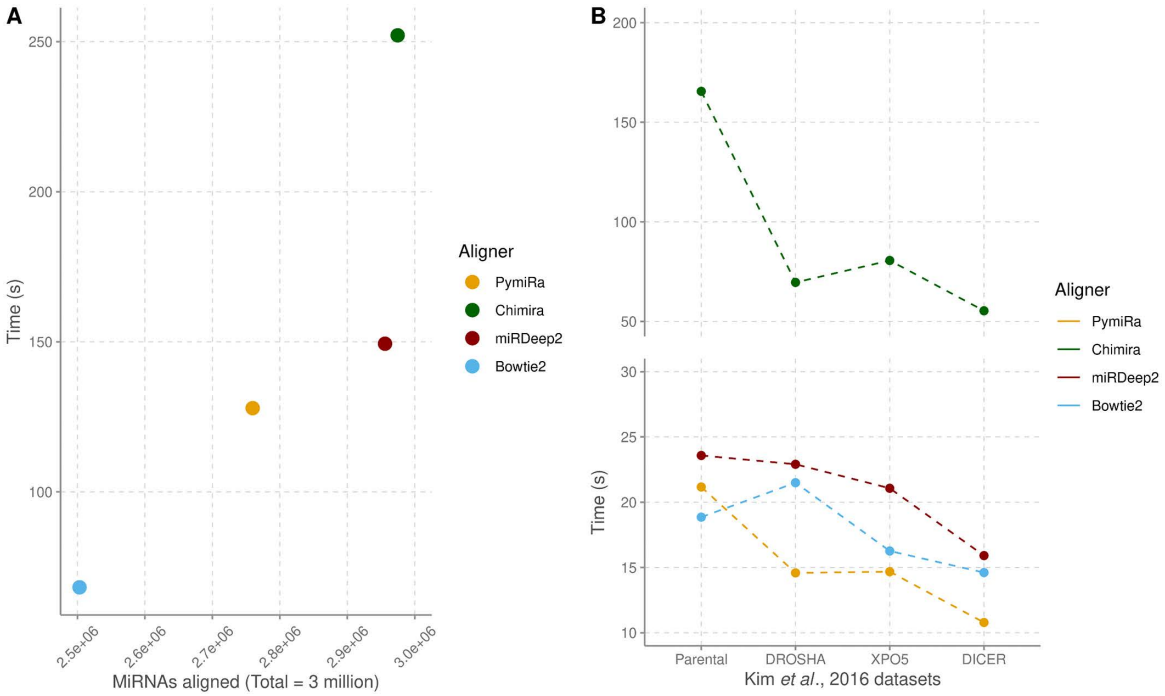

Fig 4. Quantitative benchmarking of the speed and accuracy of each aligner (PymiRa, Chimira, miRDeep2, and Bowtie2) on the simulated and biological miRNA datasets. A) Total time taken (speed) and the number of miRNAs aligned (accuracy) using different aligners on the simulated 10 million read dataset. **B)** The average time taken (speed) from three repeats ($n = 3$) to align the data from different miRNA biogenesis knockout experiments from the Kim *et al.,* 2016 biological dataset using the different aligners. Bowtie2 (*std*_FC) was used to represent Bowtie2 due to superior performance to the other settings/conditions tested.

https://doi.org/10.1371/journal.pcbi.1014114.g004

Discussion

PymiRa is a rapid, accurate, and accessible aligner for analysing small non-coding RNA (sncRNA) sequencing data, here utilised for miRNA identification and quantification from sequencing data. We have shown its effectiveness on both simulated and biological human and non-human datasets and how overall, it outperforms existing tools in speed, accuracy of alignment, and accessibility.

The experiments utilising the simulated small RNA-sequencing dataset were effective in establishing a set of miRNAs of known quantity, and provided a robust method of assessing alignment accuracy. PymiRa, Chimira, and miRDeep2 all aligned more miRNAs, with greater correlations to the Real count, than achieved with Bowtie2. Of note, Bowtie2 with default standard settings (*std*) was the best performing Bowtie2 condition, despite allowing up to one mismatch in the 'seed' region in the *local* setting, confirming that miRNA identification is not a straightforward alignment task.

PymiRa had the lowest RMSE to the Real count from all the aligners tested, both in the entire simulated dataset and looking at the top 600 over-expressed subset, highlighting its ability to accurately detect miRNAs over other aligners. As expected, all aligners struggled to accurately identify miRNAs from the highly homologous miR-548 family (S1 Table). However, the counts of miRNAs from highly homologous clusters/families should typically be interpreted with caution due to such fundamental difficulties in accurate identification and counting.

PymiRa was effective at aligning miRNAs from biological data in terms of overall speed and accuracy, identifying miRNAs that Chimira and miRDeep2 alone were unable to (Fig 3). There was some consensus in miRNA counts between the three aligners, with some high intensity clusters at the centre for some knockout conditions. However in the Parental (wild-type) condition, there was more disparity, with only a small cluster at ~ 33% and many counts highlighting miRNAs not detected by miRDeep2.

With Chimira currently being unsupported and inaccessible to end users, and miRDeep2's primary function being to discover novel miRNAs, there is a need for a dedicated rapid and accurate, aligner for miRNA identification/quantification, capable of being run on a laptop computer for the sncRNA community. PymiRa's accessibility is an advantage over other tools, with built-in rapid reference generation and quantification from a single command. Additionally, PymiRa is not restricted to miRNA detection, unlike other tools. Users are able to align to a reference FASTA/FASTQ file of their choice with minimal setup time and with no additional steps required, providing a simplistic approach to sncRNA alignment. Further, the ability to run alignments with the webserver quickly to miRBase and other sncRNA databases such as GtRNAdb for detecting transfer RNAs (tRNAs), as is currently available, is a major advantage over other alignment services [20].

With small RNA-sequencing increasingly being undertaken, even small decreases in computational run time will be advantageous for analysing large datasets. PymiRa was consistently one of the fastest aligners whilst not compromising on accuracy, due to its ability in allowing up to two mismatches at the 3' end of read sequences, meaning PymiRa is a fundamentally useful addition for the sncRNA community. PymiRa has thus been designed for a wide range of downstream applications such as rapid differential expression analysis and biomarker discovery, aggregating -5p/-3p arm-level signals to capture biologically relevant miRNA expression changes. Of note, whilst it is broadly applicable to sequencing data analyses, it is not designed for specialised niche applications such as isomiR discovery.

In summary, we present PymiRa as a rapid, accurate, and accessible classification tool for sncRNAs, including miRNAs, available through a webserver. PymiRa will facilitate a deeper biological understanding of the landscape of sncRNA expression in normal physiological conditions and their dysregulation in disease states, including cancer.

Materials and Methods

PymiRa is available on the webserver at https://www.pymira.co.uk or available on our GitHub repository at https://github.com/ZScurlock/PymiRa for easy install using *pip*. Installation instructions and examples of the input/output files are available on the GitHub repository. PymiRa was designed to allow the rapid quantification of sncRNAs without the need for a high-performance computing cluster. However, depending on the input sample size, it may require greater random access

memory (RAM) specifications. PymiRa was created and tested on a Unix machine with up to 32Gb of RAM and eight processors.

Input/Usage notes

PymiRa accepts either FASTA or FASTQ (which can be gzipped) files as input. These must be trimmed of any adapter/barcode sequences and quality controlled. All reads should be of similar length to the RNA size expected and at least 15nt in length i.e., between 15–50 nts for mature miRNAs. PymiRa can be run with the minimum of *--input_file* (file to be aligned) *--ref_file (*reference to be aligned to) *--out_path* (basename of output files). The output for PymiRa is a counts file ({BASENAME}_pymira_counts.txt), a log file ({BASENAME}_pymira_log.json), and an alignment summary ({BASENAME}_pymira_alignment_summary.json).

PymiRa has an optional *--mirna* flag specific for aligning miRNAs. This aims to prevent alignments through the middle of a miRNA hairpin sequence [i.e., through the hairpin (stem) loop] to reduce the possibility of capturing degradation products and isomiRs. IsomiRs are defined as sequences with additional nucleotide/sequence variants compared with their archetype miRNA [21]. When isomiRs have been explicitly identified in other studies, only approximately 5% of these (primarily 5' variants) altered the 5' miRNA seed region which determines binding specificity, and were thus understood to have distinct functional effects from their canonical miRNA. This observation supports PymiRa's use of recording aggregated, -5p/-3p arm-level miRNA abundances [21]. Of note, this approach will not therefore completely rule out the possibility of a small number of potential isomiRs being counted as canonical miRNAs. If a valid alignment is found, -5p/-3p miRNA notation is assigned, depending on alignment to the 5' or 3' end of the reference, respectively. Additionally, PymiRa includes a multi-processing flag (*--num*_proc) and the option to specify the number of allowed mismatches at the 3' end (*--mismatches*_3p). On the webserver, the user can upload their samples and choose which reference database to align against. PymiRa outputs a counts table of the chosen sncRNAs identified, a log file (.json) recording the alignment of each read, as well as an alignment summary.json file.

Alignment workflow

PymiRa's implementation was initially inspired by Cody Glickmann's GitHub repository on using Burrows-Wheeler Transformation (BWT) within Python (https://github.com/glickmac/Burrows_Wheeler_in_Python). To create PymiRa, we have heavily modified and extended this approach to generate an efficient Ferragina-Manzini (FM) index encoding the suffix array to conduct the backward search the algorithm relies upon. PymiRa creates this for the reference file selected (miRNA hairpin precursor), providing a fast lookup for alignment.

PymiRa initially attempts to perfectly align each read in the input file to the reference, without any initial tolerance for mismatches. If a mismatch does occur at the 3' end of a read (last 45% of the read length), it attempts to realign each read, but allowing for up to two mismatches in this region (S6 Fig). This is specifically set to 45% to maximise the length of the 3' fragment to allow for any mismatches to occur and to prevent spurious matches to other regions, especially as small RNAs are already short in sequence. The number of mismatches permitted in this region can be altered within the '*mismatches_3p*' parameter function of PymiRa. Any reads that do not fulfill these criteria are discarded. For reads that align to more than one precursor sequence (*n*), counts are assigned fractionally with equal weight amongst all matches (*1*/*n*).

Simulated and biological dataset testing

A simulated ten million read dataset was created to accurately benchmark the aligners. This dataset was constructed from three million real miRNA sequences and seven million control sequences, described below. Firstly, for the real miRNA sequences, 1.5 million were randomly selected from all 1,917 mature miRNA sequences with *Homo sapiens (hsa-)* annotation from the current version of miRBase (v22.1) (S7A Fig). The remaining 1.5 million reads were distributed between a random subset of 600 miRNAs to simulate 'upregulated' miRNA expression, which may be observed in disease states

including cancer, for example the miR-371~373 and miR-302/367 clusters in malignant germ cell tumours and the miR-155/210 cluster in diffuse large B-cell lymphomas [22,23]. The seven million control sequences were shuffled sequences from miRBase using the sequence shuffling software *uShuffle,* providing similarly sized reads with the same nucleotide composition but in a uniquely different order to real miRNAs [14]. To add further noise to the simulated dataset, both real and shuffled sequences were subject to a mismatch and an A/U base insertion probability of 10%, mimicking adenosine deaminases that act on RNA (ADAR) edits [i.e., post-transcriptional modifications where adenosine (A) nucleotides in RNA are converted to inosine (I) nucleotides], that typically occur at the 3' end of a miRNA [10]. This simulated approach was favoured over randomly generating sequence reads to maintain the proportion of nucleotides found in real miRNAs and to provide a greater challenge for the aligners by creating a more realistic small RNA-sequencing dataset (S7 Fig). To validate the effectiveness of the simulated control dataset, we experimented on an independent human sncRNA dataset (NCBI SRA accession number SRR35103152) isolated from human placenta, from Chaves-Solano *et al.,* 2025 [24]. After quality controlling and trimming of any adapters, the dataset consisted of ~18 million reads, of which PymiRa identified ~5.5 million (5,491,337) miRNAs (28.9% of all reads). This dataset would have contained multiple types of small RNAs as well as degradation products. To highlight the effectiveness of our approach, we used *uShuffle* to shuffle the bases of the filtered FASTQ file. After shuffling, PymiRa identified only 12,238 miRNAs (0.06% of all reads), representing a 99.78% decrease in miRNA counts (S4 Table). This vast reduction in miRNA count from the shuffled reads gave us confidence that our simulated dataset was a sufficient and appropriate control. Pearson's *r* scores were calculated for each aligner from their miRNA counts on the simulated dataset and compared with the Real count using Steiger's test from the 'cocor' R package, with $p < 0.05$ considered significant [25].

For the human biological dataset, FASTQ files from Kim *et al*., 2016 were used to further evaluate PymiRa [16]. They were trimmed of adapters and quality controlled using Cutadapt to remove any reads outside the range of 15–50 nt. These data were selected as they created multiple small RNA-sequencing datasets, each containing a different miRNA biogenesis protein knockout (*ko*); namely DROSHA, Exportin5 (XPO5), and DICER. This allowed for a quantitative and qualitative evaluation of each aligner, as the levels of particular miRNAs vary across the different protein knockouts due to different biogenesis mechanisms, e.g., DICER-independent miRNAs [17]. PymiRa was then run using the *--mirna* flag to obtain mature miRNA counts using a human-filtered miRNA precursor hairpin FASTA file from miRBase.

The *Mus musculus* and *Arabidopsis thaliana* small RNA-sequencing datasets used were from Chen *et al.,* 2024 and Oliver *et al.,* 2021 respectively [18,19]. For both datasets, FASTQ files were trimmed and quality controlled using Cutadapt to remove any reads outside the range of 15–50 nt. MiRNA precursor hairpin reference FASTA files were filtered for sequences from their respective species from miRBase. The *Mus musculus* miRNA counts were run on PymiRa with the *--mirna* flag to obtain mature miRNA counts. However, as the abundance data from Oliver *et al.,* 2021 was for miRNA reference hairpins, rather than mature miRNA, PymiRa was run without any optional flags to obtain counts with the same nomenclature. For both datasets, the top 25 miRNA counts in each dataset were manually consolidated between the original paper counts and the PymiRa counts to allow for an accurate comparison.

Chimira (BLAST-based), Bowtie2, and miRDeep2 (BWT-based) were chosen for comparison as popular aligners and algorithms used in small RNA-sequencing methodologies. For consistency, all aligners ran with parallelisation settings set to eight cores. Chimira was run with default settings and with modified base detection turned off. To comprehensively benchmark Bowtie2, different settings were used: standard default (*std*), *--very-sensitive-local* (*vsl*), and *--local -N1* (*local*). A whole genome index was created for Bowtie2 using the GRCh38 primary assembly FASTA file available from the Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/). To obtain miRNA counts from a Bowtie2 BAM file, the popular RNA read counting tools HTSeq and FeatureCounts (FC) were used alongside a GFF file from miRBase [12,13]. For miRDeep2 benchmarking, first collapsed reads of a FASTA file were generated using the *mapper.pl* module (*-c -m -o 8 -s*). Then the *quantifier.pl* module was used by supplying both a FASTA file of mature human miRNAs and a human precursor hairpin reference (-d -W) from miRBase. For both Bowtie2 and

miRDeep2, a bash script was used to combine multiple commands and to more accurately record the total time elapsed (alignment/prior processing and counting). All counts were then consolidated to a single naming format using a custom Python script. This was necessary as PymiRa and Chimira can both assign -5p/-3p miRNA notation based upon alignment to the hairpin precursor, even if currently there was no annotation on the miRBase GFF file that Bowtie2 used.

Supporting information

S1 Fig. Ternary plots of the whole simulated miRNA dataset comparing the miRNA count from different Bowtie2 conditions (Bowtie2_*std*, Bowtie2_*local* and Bowtie2_*vsl*) using HTSeq and FeatureCounts (FC). The numbering on each edge represents the proportion of counts achieved by each variable (condition). MiRNA counts with an equal proportion were found at ~33%. Points found on an edge between two variables represent miRNA counts achieved by two variables (conditions), with little contribution from the third variable (condition). A two-dimensional kernel density estimation was calculated for each plot, where a high density of counts was found in an increasing spectrum of blue to red ('Level') to visualise areas of overlapping points. Comparing the miRNA counts of A) Bowtie2_*std*, Bowtie2_*local* and Bowtie2_*vsl* with HTSeq and B) Bowtie2_*std*, Bowtie2_*local* and Bowtie2_*vsl* with FeatureCounts (FC).
(TIF)

S2 Fig. The raw miRNA count of the homologous miR-548 family from the overall top 600 miRNAs from the simulated 10 million read dataset (S1 Table). All aligners (PymiRa, Chimira, miRDeep2, and Bowtie2_*std* FC) struggled to accurately count these miRNAs due to their close homology and repeated sequences. For these 73 miR-548 family miRNAs, the Real count was ~ 3,000.
(TIF)

S3 Fig. Comparison of the top miRNA fold changes (>0.2) found by Kim *et al*., 2016 and PymiRa. The top miRNA fold changes were calculated from the DROSHA knockout (*ko*) and Parental samples (raw data found in Table 2). Pearson's *r* scores were calculated for the fold changes found by Kim *et al*., 2016 and by using PymiRa. Of note, only six miRNAs were identified with a fold change of >0.2, as expected, given the DROSHA *ko* condition, affecting global miRNA biogenesis.
(TIF)

S4 Fig. Comparison of *Mus musculus* miRNA counts found by Chen *et al*., 2024 and PymiRa. The samples used were from A) 3-week-old (pubertal) (SRR19795711) and B) 11-week-old (adult) (SRR19795707) testes tissue from mice. The original dataset can be found using the BioProject accession number PRJNA849281.
(TIF)

S5 Fig. Comparison of *Arabidopsis thaliana* miRNA counts found by Oliver *et al*., 2021 and PymiRa. The samples used were from A) Uninuclear (SRR9261642), B) Binuclear (SRR9261645), C) Trinuclear (SRR9261647) and D) Mature (SRR9261648) stages of pollen development. The original dataset can be found using Gene Expression Omnibus (GEO) accession number GSM3869843.
(TIF)

S6 Fig. PymiRa workflow for aligning miRNAs with up to two mismatches at the 3' end of a read. PymiRa initially inputs a FASTA/FASTQ file. A Burrows-Wheeler Transformation (BWT) reference is generated from a miRBase precursor sequence FASTA file (v22.1) by default. For each read in the input file, PymiRa attempts to first align it against the reference with no tolerance for any mismatches. However, if a mismatch is found within the 3' end of a read, PymiRa attempts to realign it, allowing up to two mismatches, but only at the 3' end of the read. Any reads with mismatches at the 5' end, or with three or more mismatches, are discarded.
(TIF)

S7 Fig. Simulated dataset methodology flowchart. A) Three million real mature miRNA sequences were generated from miRBase (v22.1) with a proportion of them randomly selected to have increased counts to simulate 'upregulation' in a true biological dataset. B) The remaining seven million reads were shuffled from mature miRNA sequences using *uShuffle* to create distinct control sequences with the same nucleotide composition as real miRNA sequences.
(TIF)

S1 Table. MiR-548 family miRNAs have highly homologous sequences. Typically, they start with the 'AAA' sequence (bold) commonly found at the 5' end of the miRNA and the 'UUUUG' sequence (bold) commonly found at the 3' end [5]. From the current version of miRBase (v22.1), the total number of miRNAs in the miR-548 family is 81. The unique number of miRNAs from the miR-548 family found in the simulated dataset was 73. They are associated with transposable elements and arise from multiple different genomic loci [15].
(DOCX)

S2 Table. Top 25 *Mus musculus* miRNA counts and abundance ranking from PymiRa and the original Chen *et al*., 2024 small RNA sequencing dataset. The samples used were from A) 3-week-old (pubertal) (SRR19795711) and B) 11-week-old (adult) (SRR19795707) testes tissue from mice. The original dataset can be found using the BioProject accession number PRJNA849281.
(DOCX)

S3 Table. Top 25 *Arabidopsis thaliana* miRNA counts and abundance ranking from PymiRa and the original Oliver *et al*., 2021 small RNA sequencing dataset. The samples used were from A) Uninuclear (SRR9261642), B) Binuclear (SRR9261645), C) Trinuclear (SRR9261647), and D) Mature (SRR9261648) stages of pollen development. The original dataset can be found using Gene Expression Omnibus (GEO) accession number GSM3869843.
(DOCX)

S4 Table. Effect of using *uShuffle* on miRNA detection from background of small RNA-sequencing reads. The sample used was from a small RNA sequencing experiment profiling miRNAs from amniotic tissues (SRR35103152, from Chaves-Solano *et al.,*2025). The FASTQ file was shuffled using *uShuffle* to retain dinucleotide pairs. After shuffling, the number of detected miRNAs from PymiRa was decreased by 99.78%.
(DOCX)

Acknowledgments

We thank Pjotr van der Jagt, Department of Plant Sciences, University of Cambridge for technical assistance with improving the overall computational efficiency of the PymiRa algorithm.

Author contributions

Conceptualization: Zachary G.L. Scurlock, Cinzia G. Scarpini, Nicholas Coleman, Matthew J. Murray, Anton J. Enright.

Formal analysis: Zachary G.L. Scurlock.

Investigation: Zachary G.L. Scurlock.

Methodology: Zachary G.L. Scurlock, Matthew J. Murray, Anton J. Enright.

Project administration: Matthew J. Murray, Anton J. Enright.

Software: Zachary G.L. Scurlock.

Supervision: Cinzia G. Scarpini, Nicholas Coleman, Matthew J. Murray, Anton J. Enright.

Writing – original draft: Zachary G.L. Scurlock, Cinzia G. Scarpini, Nicholas Coleman, Matthew J. Murray, Anton J. Enright.

Writing – review & editing: Zachary G.L. Scurlock, Cinzia G. Scarpini, Nicholas Coleman, Matthew J. Murray, Anton J. Enright.

References

1. Condrat CE, Thompson DC, Barbu MG, Bugnar OL, Boboc A, Cretoiu D, et al. miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. Cells. 2020;9(2):276. https://doi.org/10.3390/cells9020276 PMID: 31979244
2. Zen K, Zhang C-Y. Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers. Med Res Rev. 2012;32(2):326–48. https://doi.org/10.1002/med.20215 PMID: 22383180
3. Fankhauser CD, Nuño MM, Murray MJ, Frazier L, Bagrodia A. Circulating MicroRNAs for Detection of Germ Cell Tumours: A Narrative Review. Eur Urol Focus. 2022;8(3):660–2. https://doi.org/10.1016/j.euf.2022.04.008 PMID: 35537936
4. Murray MJ, Coleman N. Can circulating microRNAs solve clinical dilemmas in testicular germ cell malignancy?. Nat Rev Urol. 2019;16(9):505–6. https://doi.org/10.1038/s41585-019-0214-2 PMID: 31296945
5. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155–62. https://doi.org/10.1093/nar/gky1141 PMID: 30423142
6. Vitsios DM, Enright AJ. Chimira: analysis of small RNA sequencing data and microRNA modifications. Bioinformatics. 2015;31(20):3365–7. https://doi.org/10.1093/bioinformatics/btv380 PMID: 26093149
7. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168
8. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. https://doi.org/10.1186/gb-2009-10-3-r25 PMID: 19261174
9. Ferragina P, Manzini G. Opportunistic data structures with applications. In: Proceedings 41st Annual Symposium on Foundations of Computer Science. 390–8. https://doi.org/10.1109/sfcs.2000.892127
10. Blow MJ, Grocock RJ, van Dongen S. RNA editing of human microRNAs. Genome Biol. 2006;7:R27.
11. Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs. Nat Rev Mol Cell Biol. 2016;17(2):83–96. https://doi.org/10.1038/nrm.2015.4 PMID: 26648264
12. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9. https://doi.org/10.1093/bioinformatics/btu638 PMID: 25260700
13. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677
14. Jiang M, Anderson J, Gillespie J, Mayne M. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics. 2008;9:192. https://doi.org/10.1186/1471-2105-9-192 PMID: 18405375
15. Piriyapongsa J, Jordan IK. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS One. 2007;2(2):e203. https://doi.org/10.1371/journal.pone.0000203 PMID: 17301878
16. Kim Y-K, Kim B, Kim VN. Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. Proc Natl Acad Sci U S A. 2016;113(13):E1881-9. https://doi.org/10.1073/pnas.1602532113 PMID: 26976605
17. Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. Nature. 2007;448(7149):83–6. https://doi.org/10.1038/nature05983 PMID: 17589500
18. Chen A, Ji C, Li C, Brand-Saberi B, Zhang S. Multiple transcriptome analyses reveal mouse testis developmental dynamics. BMC Genomics. 2024;25(1):395. https://doi.org/10.1186/s12864-024-10298-y PMID: 38649810
19. Oliver C, Annacondia ML, Wang Z, Jullien PE, Slotkin RK, Köhler C, et al. The miRNome function transitions from regulating developmental genes to transposable elements during pollen maturation. Plant Cell. 2022;34(2):784–801. https://doi.org/10.1093/plcell/koab280 PMID: 34755870
20. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016;44(D1):D184-9. https://doi.org/10.1093/nar/gkv1309 PMID: 26673694
21. Haseeb A, Makki MS, Khan NM, Ahmad I, Haqqi TM. Deep sequencing and analyses of miRNAs, isomiRs and miRNA induced silencing complex (miRISC)-associated miRNome in primary human chondrocytes. Sci Rep. 2017;7(1):15178. https://doi.org/10.1038/s41598-017-15388-4 PMID: 29123165
22. Lawrie CH, Gal S, Dunlop HM, Pushkaran B, Liggins AP, Pulford K, et al. Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. Br J Haematol. 2008;141(5):672–5. https://doi.org/10.1111/j.1365-2141.2008.07077.x PMID: 18318758
23. Palmer RD, Murray MJ, Saini HK, van Dongen S, Abreu-Goodger C, Muralidhar B, et al. Malignant germ cell tumors display common microRNA profiles resulting in global changes in expression of messenger RNA targets. Cancer Res. 2010;70(7):2911–23. https://doi.org/10.1158/0008-5472.CAN-09-3301 PMID: 20332240
24. Chaves-Solano N, Kau-Strebinger S, Oesterreicher J, Pultar M, Holnthoner W, Grillari J, et al. Distinct miRNA profiles in human amniotic tissue and its vesicular and non-vesicular secretome. Front Cell Dev Biol. 2025;13:1692501. https://doi.org/10.3389/fcell.2025.1692501 PMID: 41234360
25. Diedenhofen B, Musch J. cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One. 2015;10(3):e0121945. https://doi.org/10.1371/journal.pone.0121945 PMID: 25835001