Accurate microRNA annotation of animal genomes using trained covariance models of curated microRNA complements in MirMachine

Sinan Uğur Umu[1], Håvard Trondsen[1], Vanessa M. Paynter[2], Tilo Buschmann[3], Trine B. Rounge[4,5], Kevin J. Peterson[6], Bastian Fromm[2*]

[1] Department of Pathology, Institute of Clinical Medicine, University of Oslo, Norway

[2] The Arctic University Museum of Norway, UiT -The Arctic University of Norway, Tromsø, Norway

[3] Independent Researcher, Leipzig, Germany

[4] Department of Research, Cancer Registry of Norway, Oslo, Norway

[5] Centre for Bioinformatics, Department of Pharmacy, University of Oslo, Norway

[6] Department of Biological Sciences, Dartmouth College, Hanover NH, USA

*- corresponding author: Bastian.Fromm@uit.no

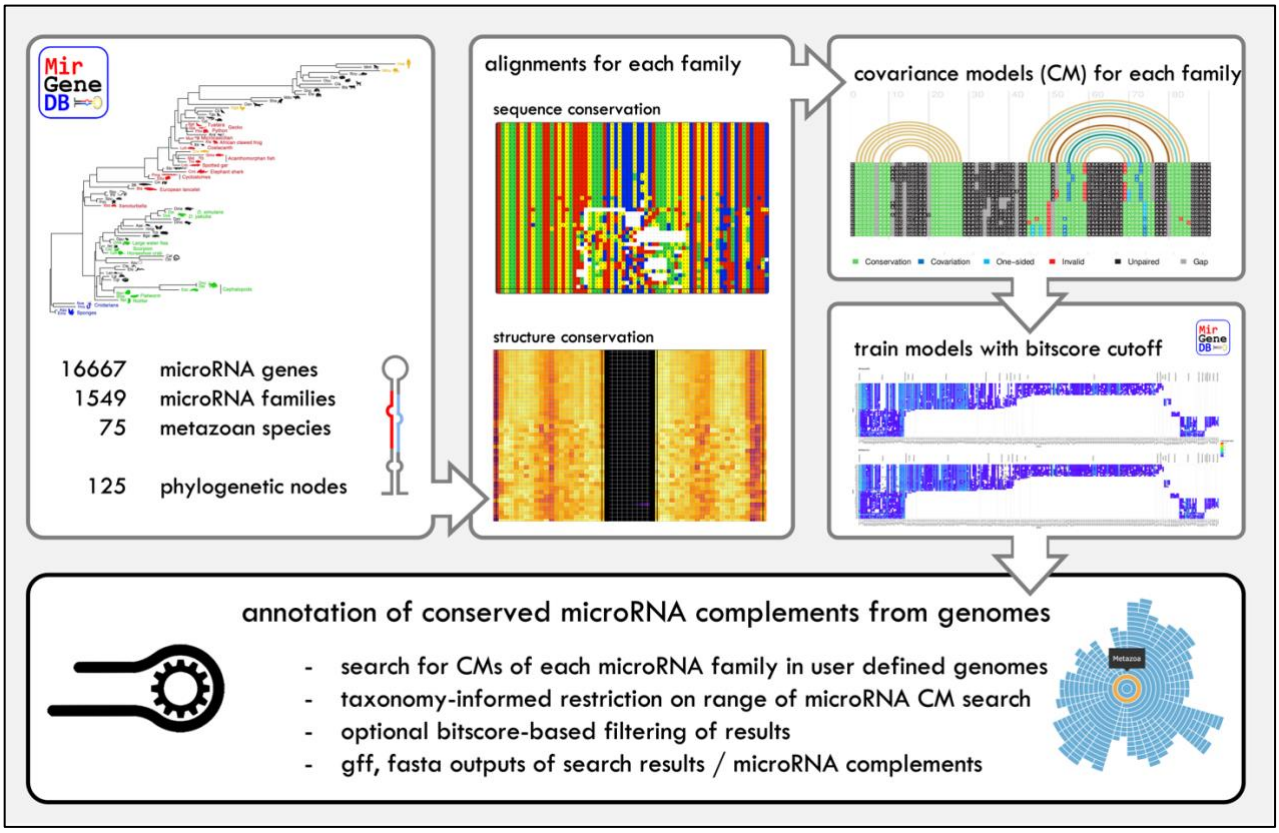
Highlights

- An annotation pipeline using trained covariance models of microRNA families
- Enables massive parallel annotation of microRNA complements of genomes
- MirMachine creates meaningful annotations for very large and extinct genomes
- microRNA score to assess genome assembly completeness

Summary

Understanding the evolution of organismic complexity and the genomic basis of gene-regulation is one of the main challenges in the postgenomic era. While thousands of new genomes are available today, no accurate methods exist to reliably mine those for microRNAs, an important class of post-transcriptional regulators. Currently, their prediction and annotation depend on the availability of transcriptomics data sets and hands-on expert knowledge leading to the large discrepancy between novel genomes made available and the availability of high-quality microRNA complements. Using the more than 16,000 microRNA entries from the manually curated microRNA gene database MirGeneDB, we generated and trained covariance models for each conserved microRNA family. These models are available in MirMachine, our new pipeline for automated annotation of conserved microRNAs. We show that MirMachine can be used to accurately and precisely predict conserved microRNA complements from genome assemblies, correctly identifying the number of paralogues, and by establishing the novel microRNA score, the completeness of assemblies. Built and trained on representative metazoan microRNA complements, we used MirMachine on a wide range of animal species, including those with very large genomes or additional genome duplications and extinct species such as mammoths, where deep small RNA sequencing data will be hard to produce. With accurate predictions of conserved microRNAs, the MirMachine workflow closes a long-persisting gap in the microRNA field that will not only facilitate automated genome annotation pipelines and can serve as a solid foundation for manual curation efforts, but deeper studies on the evolution of genome regulation, even in extinct organisms. MirMachine is freely available (https://github.com/sinanugur/MirMachine) and also implemented as a web application (www.mirmachine.org).

Graphical abstract



Keywords

microRNAs, annotation, machine-learning, evolution

Introduction

MicroRNAs are among the most conserved regulatory elements in animal genomes and have crucial roles in development and disease (Bartel, 2018; Fromm et al., 2015). They have long been proposed as disease biomarkers (Mendell and Olson, 2012; Umu et al., 2022; Wang et al., 2016), phylogenetic markers for studying animal systematics (Tarver et al., 2013, 2018), and for understanding the evolution of complexity in metazoans (Heimberg et al., 2008; Peterson et al., 2009). Currently, however, the annotation and naming of *bona fide* microRNA complements requires assembled genome references, small RNA sequencing (smallRNAseq) data from different tissues and developmental stages, and substantial hands-on curation of the outputs from microRNA prediction tools (Friedländer et al., 2008; Hackenberg et al., 2009; Wheeler et al., 2009). Because these tools were not designed to handle the amount of sequencing data or genome assembly sizes available today and often have high false-positives rates, using them is a tedious process that requires years of training, often extensive computational resources, experience and substantial amounts of time (Fromm et al., 2022a). Especially in larger projects that are not focused on microRNAs, but rather might attempt to annotate them along with other coding and non-coding genes, the required level of attention to detail is often missing which inevitably results in biologically meaningless microRNA results (Fromm et al., 2018, 2019a, 2022b, 2022a; Witwer and Halushka, 2016) as well as thousands of spurious microRNA annotations (Fromm et al., 2015). These shortcomings, coupled with the availability of high-quality and publicly available microRNA annotations suited for comparative genomics studies led to the construction of the curated microRNA gene database MirGeneDB (Fromm et al., 2015, 2020a, 2022c). MirGeneDB version 2.1 (2022) now contains microRNA complements for 75 metazoan species spanning all major metazoan phyla over ~850 million years of animal evolution (Fromm et al., 2022c). Since each gene and family were manually curated in all species in MirGeneDB, highly accurate alignments across this wide span of animal evolution are available that capture a high proportion of the sequence variability for each family. Importantly, each microRNA gene and family is associated with a detailed phylogenetic reconstruction of the evolutionary node of origin and estimated age. This dataset, hence, represents a starting point to better

understand features of microRNAs (Kang et al., 2021) and to generate better tools for the prediction of microRNAs.

Despite MirGeneDB curating a relatively large number of phyla, the number of species currently covered (75 species) is a far cry relative to the thousands of high-quality animal genomes currently available (Hotaling et al., 2021) (Figure 1).

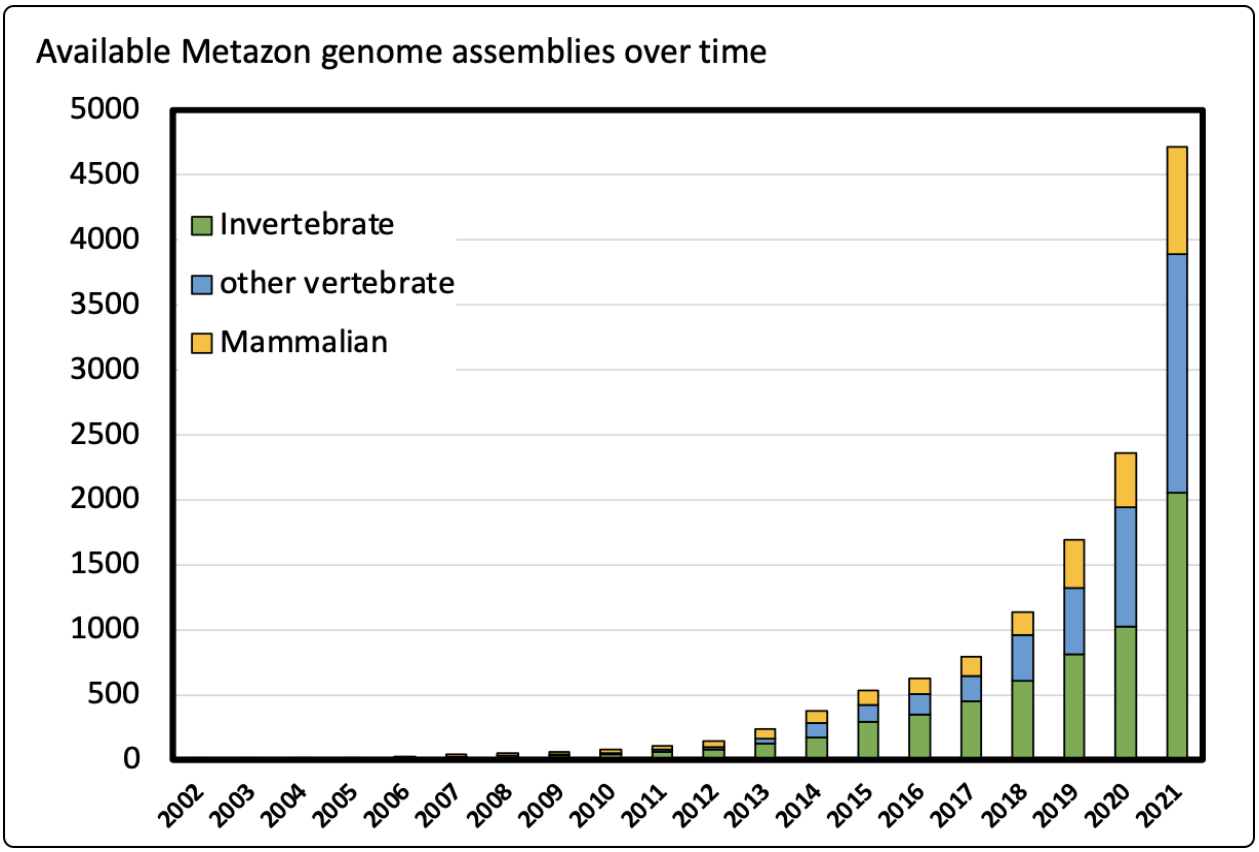


Figure 1: The number of available animal genome assemblies grows exponentially and with more than 4500 currently (2021) available datasets has dramatically grown (Clark et al., 2016).

Very few of these species have been annotated for microRNAs, or have small RNA sequencing data published, thus, comparatively little progress has been made on the suggested microRNA applications (but see (Fromm et al., 2013; Peterson et al., 2021; Wheeler et al., 2009; Zolotarov et al., 2022) for examples using manual curation). This discrepancy persists because, among other things, no reliable *in silico* method currently exists to annotate conserved or species-specific microRNA complements from genomic references only. Despite the availability of computational methods for the search of short RNAs such as microRNAs (Velandia-Huerto et al., 2021) and sophisticated machine-learning based tools for non-coding RNA applications (Amin et al., 2019), there is currently no approach satisfying the demands of high precision, low false discovery rates and minimized computational demand in a fully automated and user-friendly pipeline (Yazbeck et al., 2017). It is a widely acknowledged problem for machine learning

applications in genomics in general that existing tools are based on incomplete models (Sacar et al., 2013; Whalen et al., 2021). This is the case for microRNA families from miRBase (Kalvari et al., 2021). Such models, for instance covariance models (CMs) of individual RNA classes, families or genes, as used in the Rfam database (Kalvari et al., 2021), are technically quite accurate in detection of many non-coding RNA families (Eddy and Durbin, 1994). However, they require high quality alignments from curated RNAs ideally coupled with detailed evolutionary information to distinguish families and genes over evolutionary time that, until recently, did not exist for microRNAs.

Taking advantage of the manually curated and evolutionarily informed microRNA complements of 75 metazoan organisms in MirGeneDB 2.1, we here built and trained high-quality CMs for 1,157 conserved microRNA families and integrated them into a fully automated pipeline for microRNA annotation: MirMachine. We show that MirMachine produces highly accurate microRNA annotations in a time-efficient manner from animal genomes of all classes, including very large and recently duplicated genomes, as well as from genomes of extinct species. Using the example of 88 eutherian genomes, we further show that MirMachine predictions can be summarized in a microRNA score that predicts low contiguity or completeness of genome assemblies. MirMachine is freely available (https://github.com/sinanugur/MirMachine) and also implemented as a user-friendly web application (www.mirmachine.org).

Results

Accurate Covariance models of 508 conserved microRNA families

16,670 microRNA precursor sequences from 75 species were downloaded from MirGeneDB and all variants from the same genes, antisense loci, and species-specific microRNAs (i.e., not conserved in any other species) were removed arriving at a total of 14,953 genes representing 508 families (Figure 2A).

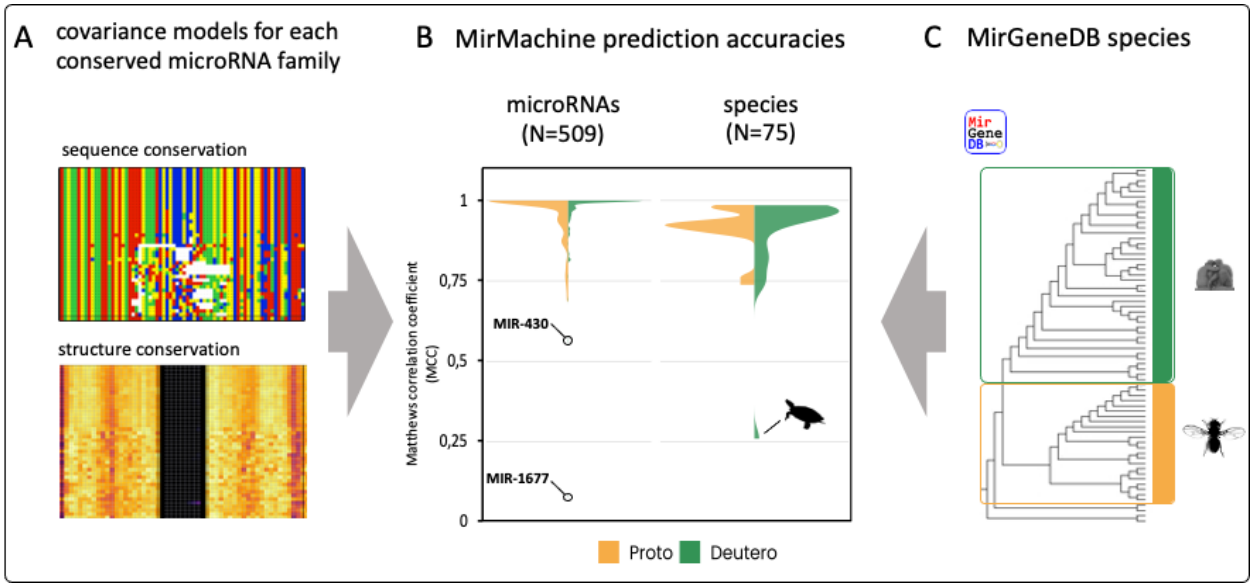
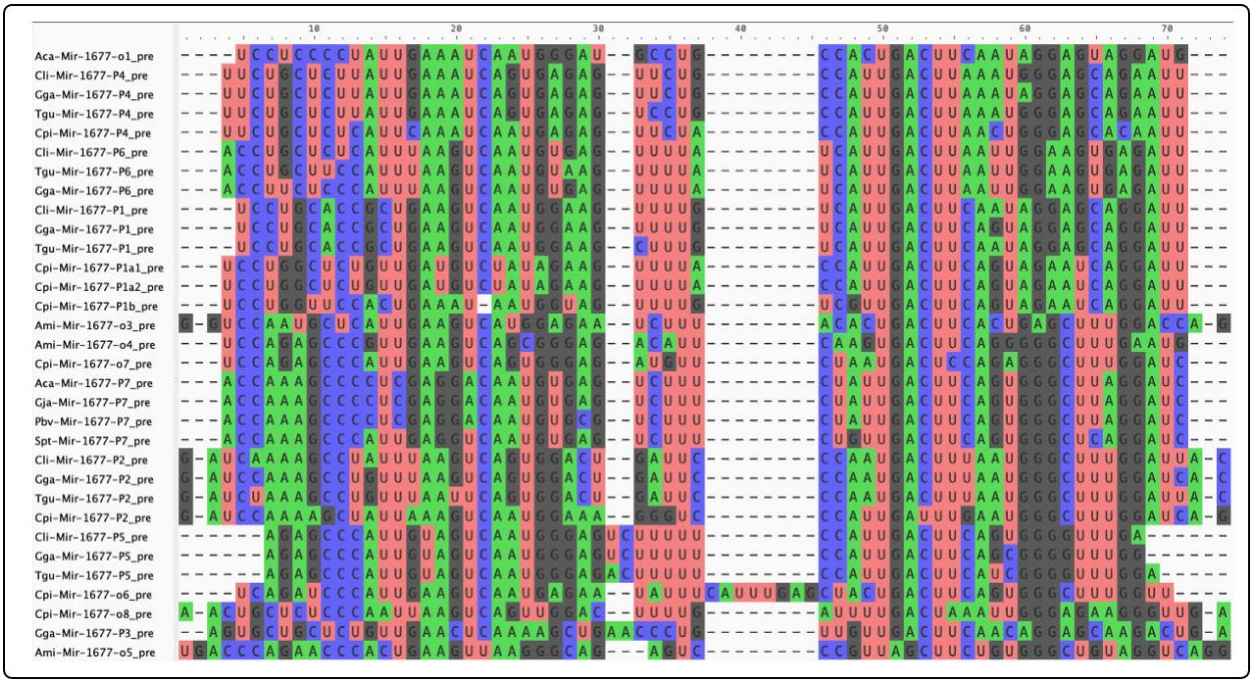


Figure 2: Developing MirMachine covariance models (CMs). A) The MirMachine workflow uses microRNA family-based precursor sequence alignments and structural information to build CMs that B) show very good overall prediction performances when models are run on C) 75 MirGeneDB species using distinct models for protostomes (yellow) and deuterostomes (green) or combined models (not shown).

We then split deuterostome (N=42) and protostome (N=29) representatives and all microRNA genes for each family were aligned, and covariance models (CM) were built (388 microRNA family models for deuterostomes and 143 microRNA family models for protostomes). Using machine-learning, these models were subsequently trained on the full MirGeneDB dataset to derive optimal cutoffs for their prediction. To measure the prediction accuracy of these models we then used the models on all MirGeneDB species comparing the predictions to the actual complements. An overall very high mean prediction accuracy of 0.975 (Matthews Correlation coefficient (MCC)) for combined models, and 0.975 for deuterostomes, and 0.966 for protostome-models, respectively, was found (Figure 2B, left & Figure 2C). Two microRNA families, MIR-430 and MIR-1677 from the deuterostome models, showed substantially lower MCC scores due to a well-known variability within the MIR-430 family (Bazzini et al., 2012; Choi et al., 2007; Giraldez et al., 2006) and a combination of low level of complexity and high variation between orthologues in the Diapsida-specific MIR-1677 (Supplementary Figure 1).



Supplementary Figure 1: Alignment of Mir-1677 genes from MirGeneDB shows low conservation that explains poor performance of MIR-1677 CMs in MirMachine.

Conversely, we observe high mean species accuracies of 0.91 for combined models, 0.92 for deuterostomes and 0.92 for the protostome models (Figure 2B, right). The reason that the turtle (*Chrysemys picta bellii*) has such a low MCC is due to the identification of nearly two thousand likely artifactual hits for MIR-1677.

MirMachine CMs models are not dependent on individual species

To identify potential effects from circular logic of predicting microRNAs of a species that were included to build the query models, we retrained all models for deuterostomes without including human and all protostome models without including the polychaete *Capitella teleta*. We then used the new deuterostome and protostome CMs to predict microRNA complements in human and *C. teleta*, respectively. We found that MCC for *H. sapiens* only very slightly decreased in accuracy from 0.97 to 0.96 highlighting the robustness of MirMachine covariance models in deuterostomes. In protostomes, the effect on MCC was stronger for leaving out *C. teleta* with a decrease from 0.92 to 0.76. Specifically, some families were not found, including the bilaterian families MIR-193, MIR-210, MIR-242, MIR-278, MIR-281, MIR-375, the protostome families MIR-12, MIR-1993 and the lophotrochozoan family MIR-1994, which were still predicted, but fell below a newly defined threshold. This highlights a markable higher sequence divergence within protostomes, which is likely due to the age of the group, the lower number of representative clades, lower number of paralogues and orthologues per family, and a lower number of species in general. The annelid families MIR-1987, MIR-1995, MIR-2000, MIR-2685, MIR-2687, MIR-2689 and MIR-2705 were not searched because no

models were built given the absence of a second annelid species, highlighting the importance of including at least two representative species for each clade in MirGeneDB (Fromm et al., 2022c).

Performance of MirMachine prediction versus MirGeneDB complement

To get a comprehensive understanding of the performance of MirMachine on the microRNA complements of MirGeneDB species, we looked in more detail at the performance of CMs, and their respective cut-offs, for a selection of major microRNA families (N=305) including all gene-copies (N=12,430) (Figure 3). When comparing the MirGeneDB complements (Figure 3A) with the predictions from MirMachine (Figure 3B), similarities were striking and overall differences limited to few families (Figure 3C); indicating either potentially false positives (231) or false negatives (421), respectively (Supplementary File 1). These are of further interest as they either represent missed microRNAs in MirGeneDB, or significant deviations from the general CMs and, hence, possibly incorrectly assigned microRNA paralogues in MirGeneDB.

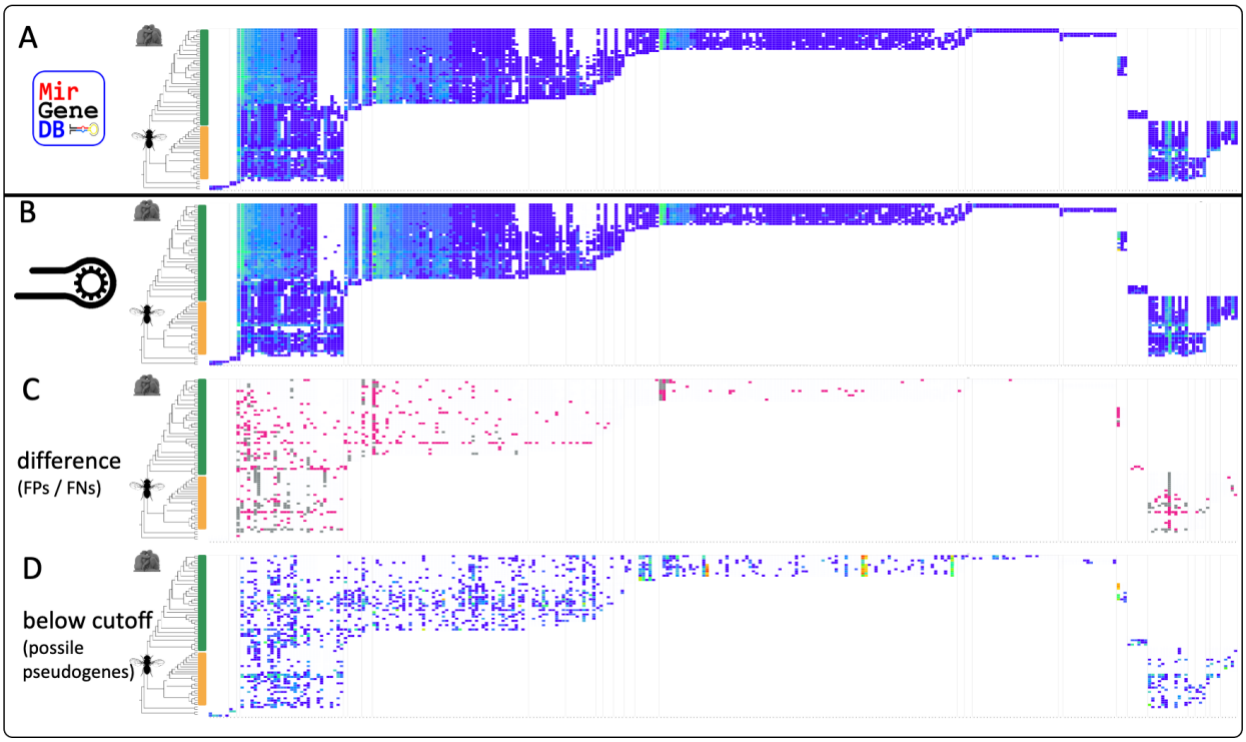


Figure 3: Detailed comparison of MirMachine predictions on 75 MirGeneDB species and 305 representative microRNA families in the form of banner-plots. Columns are microRNA families sorted by phylogenetic origin and rows are species. Heatmap indicates number of paralogues / orthologues per family. A) the currently annotated microRNA complements in MirGeneDB 2.1 (Fromm et al., 2022c). B) MirMachine predictions for the same species and families show very high similarity to A. C) Differences between A and B highlighted as potential false-positives (pink) or false negatives (gray). D) MirMachine predictions below cut-off based on training of CMs on MirGeneDB show a range of potential random predictions and pseudogenes, highlighting the effect of curation & machine learning on models.

Finally, we found a substantial number of low-scoring MirMachine predictions of microRNA families that did not reach the determined cutoff based on trained CMs (Figure 3D) and therefore are not considered *bona fide* microRNAs. However, we found that these also contain pseudogenized microRNA orthologues (or paralogues) exemplified by a hitherto unknown human LET-7 pseudogene that is not found expressed in any MirGeneDB sample (Figure 4). To our knowledge, this is the first report of, and MirMachine the respective tool for, pseudogene-predictions for microRNAs. Pseudogenes, or 'gene-fossils', are potentially very useful to determine the rate of gene duplication and follow the evolution of sequence changes in organisms and might be included in studies studying cause and consequences of duplications on microRNAs (Peterson et al., 2021).

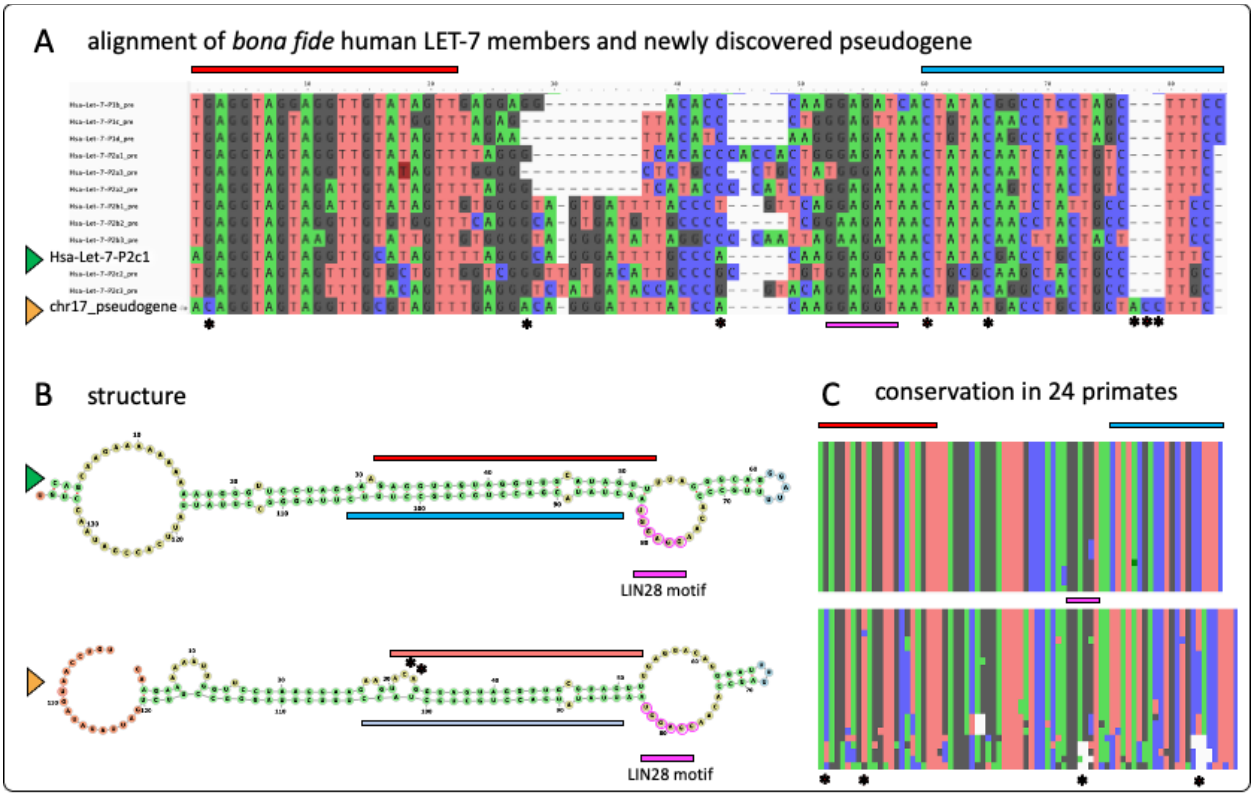


Figure 4: The human Chr.17 LET-7 pseudogene. A) sequence alignment of the currently annotated 12 *bona fide* LET-7 family members in human and the pseudogene candidate discovered by MirMachine. Non-random sequence similarities, including LIN28 binding sites (pink) are apparent with few noteworthy differences (asterisks) such as in position 2 on the 5' end (red box indicates mature annotation, position 2 equals seed-sequence) or a triplet insertion at the 3' end (blue box indicates star sequence annotation) are indications for non-functionality. B) Structural comparison of a representative *bona fide* LET-7 member (Hsa-Let-7-P2c1, green triangle) with the pseudogene (yellow triangle) highlights similarities of pseudogene candidate to *bona fide* microRNA, but points out disruptive nature of nucleotide changes for the structure (asterisks) very likely affecting a potential Drosha processing. C) sequence conservation of *bona fide* Hsa-Let-7-P2c1 (top) and the pseudogene (bottom) in 24 primate genome (ENSEMBL v100) highlights the

sequence conservation of *bona fide* microRNAs from the loop showing some changes, the star (blue) few changes and the mature (red) showing none, while the pseudogene shows many more changes and seems to be enriched in disruptive changes in the mature / seed region.

The microRNA complements of eutherians reveal the microRNA score as simple feature for genome contiguity

Applying MirMachine to a testcase, we downloaded 89 eutherian genomes currently available in Ensembl that are not curated in MirGeneDB and annotated their conserved microRNA complements. Altogether 38,550 genes in 260 families, in about 4,400 CPU hours, were found and showed an overall very high concordance between species (Figure 5A). As expected, Catharrini (pink) and Muridae (light green) specific microRNAs were only found in the respective representatives, but surprisingly, six species (Figure 5, yellow arrows) showed substantial absences of microRNA families. We therefore wondered whether these absences indicate microRNA losses due to biological simplifications (see (Fromm et al., 2013)), proposed random events (Dunn, 2014; Thomson et al., 2014), or whether they might be due to technical reasons (Tarver et al., 2018). Given that the outlier species (Alpaca, Shrew, Hedgehog, Tree shrew, Pika, and Sloth) have no particularly reduced morphology, we reasoned that the source might be technical and recovered N50 contiguity values for all genomes. We found that all six genomes had substantially lower N50 values than all other genomes, indicating that microRNAs might be able to predict completeness of genome assemblies (Figure 5B). Therefore, we next developed a simple microRNA scoring system defined as the percentage of expected conserved microRNA families found from a genome (in this case including 175 microRNA families found in most eutherians according to MirGeneDB (Fromm et al., 2022c), and showed that microRNA scores below 80% correlate with very poor N50 values <10kb and that N50 values of 100kb indicate microRNA scores of 90% and higher (Figure 5C, red and blue lines). A noteworthy exception is the microbat *Myotis lucifugus* with a N50 of 64kb and a microRNA score of 74%, which might be explainable by previously suggested genome evolution mode through loss (Huang et al., 2016; Jebb et al., 2020).

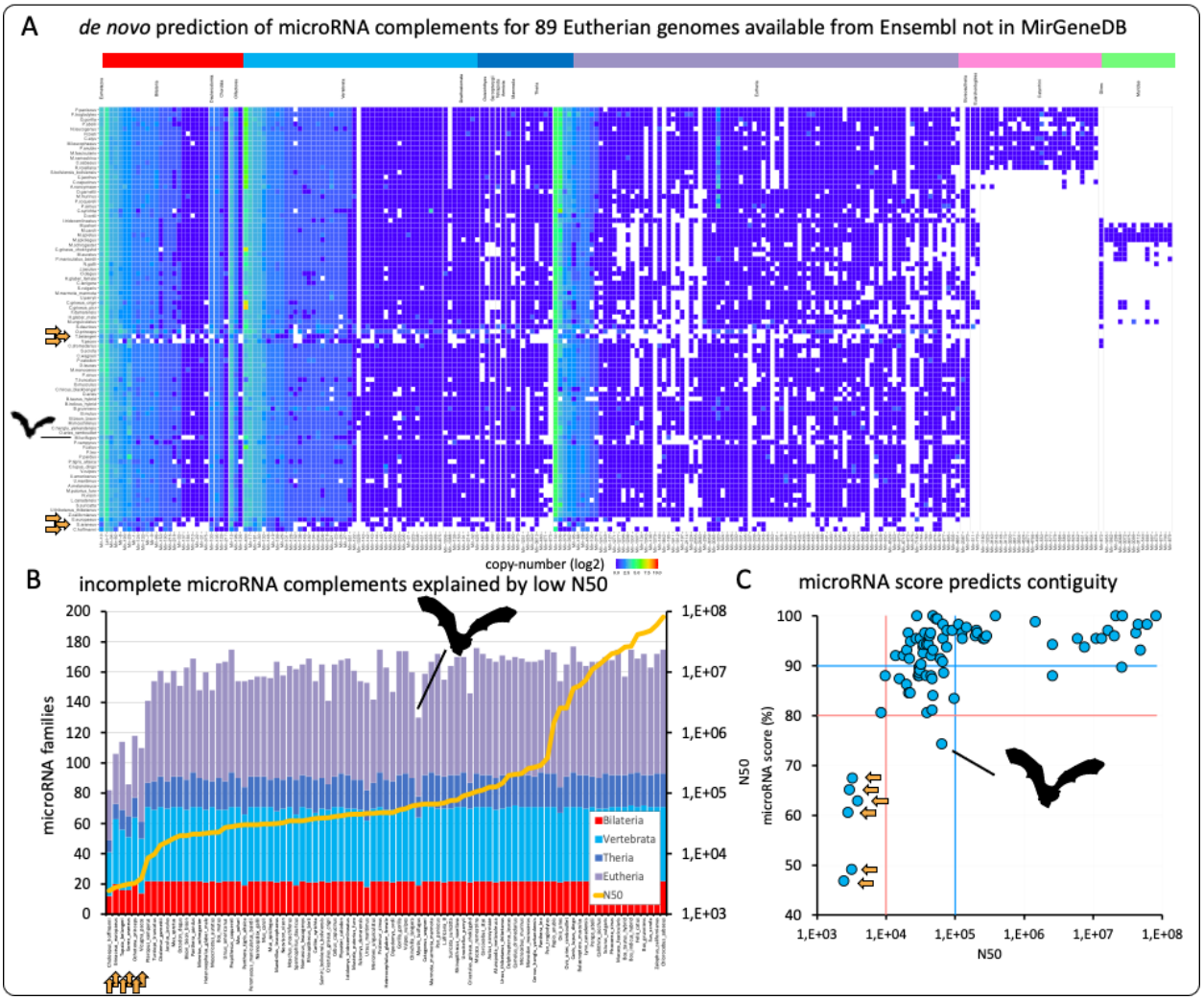


Figure 5: MirMachine predicts conserved microRNA complements of 89 eutherian mammals available on Ensembl and not currently represented in MirGeneDB. A) banner plot of results for MirMachine predictions on 88 eutherian mammalian species for selected range of major microRNA families and genes showed very strong homogeneity of microRNA complements in general and identified a number of clear outliers (yellow arrows, including Alpaca, Shrew, Hedgehog, Tree shrew, Pika, and Sloth). B) Stacked histogram sorted by N50 values). Outlier species (yellow arrows: same as in A)) all have very low N50 values, indicating an artificial absence of these phylogenetically expected microRNA families. C) The microRNA score predicts the assembly contingency and is the proportion of phylogenetically expected microRNA families that are found in respective genomes (here eutherians). microRNA scores below 80% (red horizontal line) tend to have low N50 values (red vertical line indicates N50 below 10,000 nucleotides), while scores above 90% indicate N50 higher than 10,000 nucleotides. Noteworthy exception is the bat *Myotis lucifugus* which might be explainable by previously suggested genome evolution mode through loss (Huang et al., 2016; Jebb et al., 2020).

MirMachine predicts microRNAs from extinct organisms and very large genomes

High quality *in silico* annotation of genomes is particularly important for organisms where no RNA is likely to ever become available. This is the case for species such as mammoths that went extinct millennia or even millions of years ago (but see (Fromm et al., 2019b)). Using available data from extinct and extant elephantids (Palkopoulou et al., 2015, 2018), we ran MirMachine on 16 afrotherian genomes, including the hyrax (*Procavia capensis*) from Ensembl and the tenrec (*Echinops telfairi*) from MirGeneDB, and 14 elephantids including extant savanna elephants (*Loxodonta africana*), forest elephants (*Loxodonta cyclotis*) and asian elephants (*Elephas maximus*) respectively (Figure 6A, green elephantid silhouettes), but also extinct american mastodon (*Mammuthus americanum*), straight-tusked elephants (*Palaeoloxodon antiquus*), columbian mammoth (*Mammuthus columbi*) and the woolly mammoths (*Mammuthus primigenius*) (Figure 6A, red elephantid silhouettes). We find a very high degree of similarities between afrotherians, and striking congruence between extinct and extant species which indicates the high accuracy of the MirMachine workflow. More so we find patterns of microRNA losses that could be phylogenetically informative (Figure 6A, arrows). For instance, we do not find MIR-210 in any of the elephant species, which might be a elephantid specific loss (Figure 6A, pink arrow), we further find that *P. antiquus* and *L.cyclotis* have both lost MIR-1251 (Figure 6A, light blue arrow), and a shared loss of MIR-675 and MIR-1343 (Figure 6A, purple arrows), both supporting previously identified sister group relationships (Palkopoulou et al., 2018).

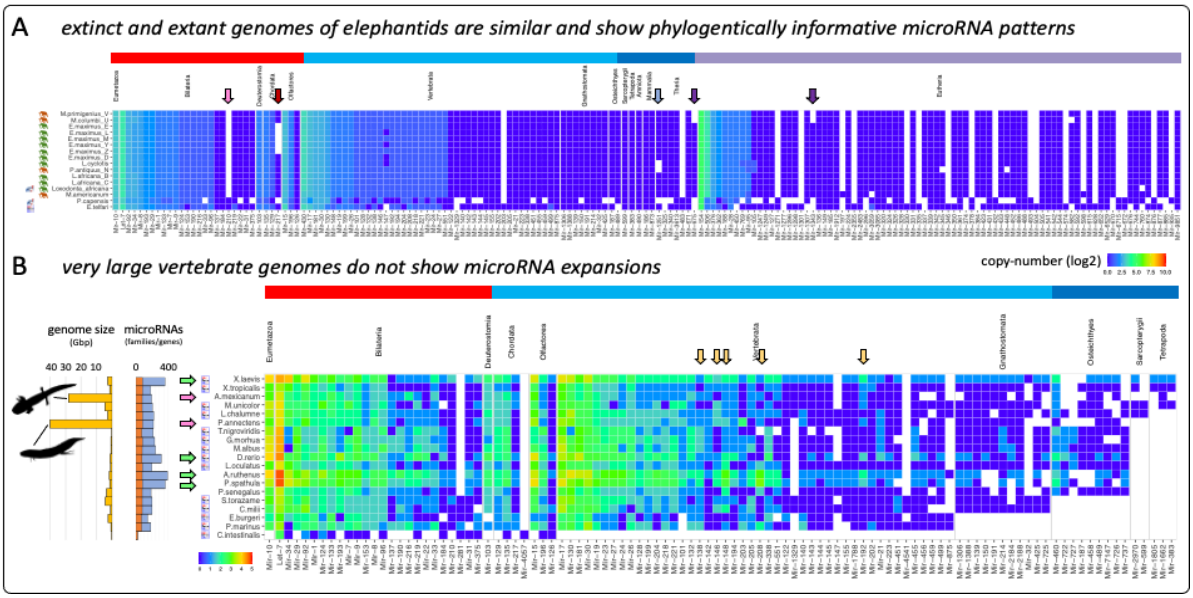


Figure 6: MirMachine enables microRNA complement annotations from extinct and very large genomes. A) MirMachine predictions from afrotherians show no clear differences between extinct and extant genomes, but likely phylogenetically informative losses of microRNA families (colored arrows). B) MirMachine predictions in organisms with extensive genome expansions (pink arrows) show no expansion of microRNAs, but organisms with known genome duplications (green arrows) do. A number of shared microRNA copies in sterlet (*A. ruthenus*) and paddlefish (*P. spatula*) support a common genome duplication event in the last common ancestor of Acipenseriformes (yellow arrows).

A pertaining challenge for microRNA prediction and annotation of extant species, is the occurrence of additional whole genome duplication events and, not necessarily connected, extreme genome expansions. This often leads to computational challenges where identical copies are hard to distinguish based on read-mappings or genomes are simply so large that existing pipelines need extensive computational resources often facing programmatic limits. Therefore, we next investigated the performance of MirMachine in vertebrate species with very large genomes and of known additional rounds of genome duplications. For the first group, we included the axolotl (*Ambystoma mexicanum*) with a genome of 28 Gbp and the african lungfish (*Protopterus annectens*) with a genome of bigger than 40 Gbp into our analysis. For the second group we included the African clawed frosh (*Xenopus laevis*) with known allotetraploid genome (Session et al., 2016) and the zebrafish (*Danio rerio*) from MirGeneDB, the sterlet (*Acipenser ruthenus*) with proposed sturgeon specific genome duplication and occurrence of segmental rediploidization (Du et al., 2020), as well as the american paddlefish (*Polyodon spathula*) with a recently shown genome duplication which was, however, interpreted as sturgeon independent (Cheng et al., 2021). We combined these species with the gray bichir (*Polypterus senegalus*) that has a moderately sized (e.g., human-sized) genome and no unique known genome duplication events, along with 13 other MirGeneDB species representing a range of Olfactores, vertebrates, gnathostomes, Osteichthyes, Sarcopterygii and Tetrapoda representatives (Figure 6B). We find that MirMachine ran very well on all genomes using 32 cores and under 2 hours per species, whereas the lungfish ran the longest (around 3 hours 45 mins). As expected, we find that the size of the genomes do not affect the microRNA complements (Figure 6B, pink arrows), but that organisms with additional whole genome duplications (Figure 6B, green arrows) clear trace of duplications (also see (Peterson et al., 2021)). A curious observation was that sterlet and paddlefish showed very consistent microRNA copy-number patterns, in particular in the retention of additional MIR-138, MIR-146, MIR-148, MIR-192 and MIR-208 copies (Figure 6B, orange arrows) indicating a likely common origin of genome duplication at the last common ancestor (Acipenseriformes), or very similar retention pressure in the more unlikely case of independent duplication. Altogether MirMachine is a suitable tool for the annotation of microRNA complements from extinct and very large genomes alike.

MirMachine models outperform existing Rfam models

In the most recent Rfam update (v. 14) an expanded assembly of microRNA models based on miRBase was released (Kalvari et al., 2021). As mentioned here before, and stated elsewhere, a major concern in microRNA research has been the quality of this online repository of published microRNA candidates (Axtell and Meyers, 2018; Castellano and Stebbing, 2013; Chiang et al., 2010; Fromm et al., 2015, 2020b; Guo et al., 2020; Jones-Rhoades, 2012; Langenberger et al., 2011; Ludwig et al., 2017; Meng et al., 2012;

Tarver et al., 2012; Taylor et al., 2014; Wang and Liu, 2011) with estimates of two out of three false-positive entries. Thus, the database contains more false positives than microRNAs. These are for instance numerous tRNA, rRNA or other fragments, but also incorrectly annotated *bona fide* microRNAs that strongly influence interpretations of data. In addition to the false positives, numerous miRBase annotations are imprecise and have varying precursor annotation forms (with or without flanking regions of varying lengths) and not both arms are annotated, 3' ends are incorrect, and in a few cases even 5' are not correctly annotated which substantially affects target predictions (for details see (Fromm et al., 2015)). Further, it uses an outdated nomenclature which is inconsistent in that members of the same microRNA family are not named the same way making the identification of family members cumbersome. This problem has to a large extent been transferred to Rfam and their microRNA family models in particular (e.g. MIR-95 family member Hsa-Mir-95-P4 (https://mirgenedb.org/show/hsa/Mir-95-P4) with own model https://rnacentral.org/rna/URS0002313758/9606, or MIR-15 member Hsa-Mir-15-P1d https://mirgenedb.org/show/hsa/Mir-15-P1d) with own model (https://rnacentral.org/rna/URS000062BB4A/9606 (see Supplementary File 2). This all has been addressed in the manually curated microRNA gene database MirGeneDB.org (Fromm et al., 2015, 2022c) and MirMachine, respectively.

Regardless, we tested the performance of 523 Rfam microRNA models, that we curated to be of animal origin, on the 75 MirGeneDB species and found that 36,931 microRNAs were predicted (compared to 16,913 MirMachine and the 15,846 microRNA annotations in MGDB 2.1). Given that the number of conserved microRNA families is a focus of MirGeneDB and very unlikely to be expanded in the future (Fromm et al., 2022a), this much higher number of predictions suggests that Rfam predictions contain thousands of FPs. We further looked for performance of highly conserved families (see materials and methods). Rfam models had MCCs of 0.96, 0.94, 0.96 and 0.89 for microRNA families LET-7, MIR-1, MIR-196 and MIR-71 respectively. The same family performances for MirMachine were 0.97, 0.98, 0.97, 0.97. Thus, as expected, Rfam model had comparable performance for these correctly assigned, and deeply conserved families, but performed poorly for incorrectly assigned microRNAs.

MirMachine functions & options

All models (total, protostome and deuterostome) were implemented into the standalone MirMachine workflow which is available under https://github.com/sinanugur/MirMachine, and the web app www.mirmachine.org. MirMachine also contains the curated "node of origin" information from MirGeneDB that can be used to limit the microRNA gene search to phylogenetically expected microRNA families, substantially reducing the search space and shortening the necessary run-time. Several other options, such as the search for single families (e.g. "LET-7") or families of a particular node (e.g. "Bilateria") are available,

too. In the web app, genome accession numbers can be provided avoiding the need for down- and upload circles.

Discussion

The existence of thousands of animal genome assemblies is massively mismatched by the availability of annotations of important gene-regulatory elements such as microRNAs. Here, we have presented MirMachine as an important first step to overcome this discrepancy, the need for small RNA sequencing data or extensive expert manual curation. The unique combination of well-established covariance model approaches trained on manually curated and phylogenetically informed microRNA family models built from more than 16,000 microRNAs of 75 metazoan species makes MirMachine very sensitive to detect paralogues of a family in a given organism (low false-negative rate) and very robust against wrong predictions (low false-positive rates). MirMachine's ability to accurately predict full conserved microRNA complements from genome assemblies, as exemplified by our analysis of nearly 90 eutherian genomes from Ensembl, will not only enable large comparative microRNA studies and automated genome annotation for microRNAs, but also showed the potential of microRNAs for the assessment of genome assembly completeness (Figure 5). Because of the near-hierarchical evolution of microRNAs, they have a very strong potential not only as taxonomic markers as used in e.g. miRTrace (Kang et al., 2018) or sRNAbench (Aparicio-Puerta et al., 2022), but to also outperform approaches that are based on protein-coding genes such as BUSCO. Those heavily rely on the correct identification of orthologues and paralogues of protein-coding genes, which are much more variable than microRNAs and are therefore often incomplete, and, hence, cannot be used to accurately assess or measure rates of loss. By comparing N50 values and a herein established microRNA score, we have shown that microRNA complements predicted by MirMachine are suited to assess genome completeness and contiguity. This might have wide-reaching consequences for future applications as a microRNA score could be a standard measure for genome annotation pipelines.

We have also shown that it is possible to use MirMachine's 'below cutoff' predictions for the study of pseudogenes, which could enable better understanding of dosage-level regulation or gene- and genome duplication events, in general (Peterson et al., 2021). Using several so far uncharted vertebrate genomes of either extreme size (axolotl, lungfish) and comparing them to smaller, but secondarily duplicated genomes, we could show that MirMachine works on such large genomes and confirm that the size of assemblies does not matter for the number of microRNAs, but that genome duplication events do. By directly comparing the outputs of MirMachine counts for microRNA paralogues in sterlet and paddlefish, we found patterns of microRNA duplicates that support a common genome duplication of the two species.

Finally, we employed MirMachine on extinct species genomes' and could show that besides similarity to extant representatives, several absences / losses of microRNAs were

observed within the elephantids that suggest a phylogenetic signal. These findings are exciting as they might give clues on the genome regulation differences in organisms, where actual RNA will be hard or impossible to get by. Importantly, at this stage, we have not yet made sequence-based comparisons of the microRNAs between any of the species. This is an untapped area for future development.

MirMachine currently provides predictions as community standard file formats GFF or FASTA that are named by family and coordinates, but not according to their possible paralogue or orthologue nomenclature (Fromm et al., 2015). This is due to the fact that the required syntenic information is often not available and not currently analyzed by our pipeline. Furthermore, MirMachine does not predict species specific microRNAs which can play crucial roles in evolution (Zolotarov et al., 2022). MirMachine predictions are a solid foundation for future smallRNAseq driven annotation efforts of novel microRNAs and synteny-supported annotation of paralogues and orthologues.

There are a number of tools to predict novel microRNAs from genomes that are all not based on curated references and, hence, might be of limited value (see (Saçar Demirci et al., 2017; Stegmayer et al., 2019). We are striving to address those issues in the future and would like to stress, in the meantime and in general, that manual curation is a crucial step that should never be disregarded, even though MirMachine heavily reduces the need for extensive and week-long efforts.

The decision to create protostome and deuterostome specific microRNA family models can be seen as a first step toward group-specific microRNA gene-family models that might increase the accuracy of MirMachine further in the future. Variability of model performance based on evolutionary age of families has not been studied here, but the addition of more taxa to MirGeneDB will be an invaluable improvement for group-specific microRNA family prediction and paralogue-specific modeling of microRNAs. Another important area of possible expansion clearly are plant microRNAs, that currently suffer from multiple non-overlapping available databases and potentially stronger curation problems than observed in animals (see (Fromm et al., 2020b; Taylor et al., 2017)).

MirMachine is freely available as a standalone tool or web application. It enables even non-microRNA experts to annotate conserved microRNA complements regardless of the availability of small RNA sequencing data. Thus, it has a strong potential to close the ever-increasing gap between existing high-quality genomes (Formenti et al., 2022; Lewin et al., 2018) and their microRNA annotations. A possible addition of MirMachine into the standard genome annotation pipelines of Refseq and Ensembl is currently discussed. The availability of thousands of metazoan genomes and their microRNA annotations will pave the way toward the promise of microRNAs and a true postgenomic era.

STAR★Methods

Creation of high-quality CMs

MicroRNA precursor sequences were downloaded from MirGeneDB as FASTA files. We then aligned each microRNA family using the *mafft* v7.475 aligner (mafft-xinsi) (Katoh et al., 2019) and created multiple sequence alignments (MSAa) of microRNA families. We filtered out identical or highly similar sequences using the *esl-weight* v0.48 tool (-f --idf 0.90 --rna) from HMMER package (Wheeler and Eddy, 2013). The secondary structures of the MSAs were predicted by RNAalifold v2.4.17 (-r --noPS) (Lorenz et al., 2011). Lastly, CMs for each microRNA family were generated and calibrated using Infernal (Nawrocki and Eddy, 2013). We used the same workflow to create deuterostome and protostome specific CMs.

Determining accuracy of MirMachine predictions

First, we used the *cmsearch* function of Infernal to predict microRNA regions. In this study, true positives (TPs) are correctly predicted microRNA families and false positives (FPs) are false predictions. False negatives (FNs) refer to microRNA annotations available in MirGeneDB but not predicted by MirMachine. Using MirGeneDB and MirMachine, we extracted all true positives, false positives, and false negative predictions. We can calculate an approximation to the Matthews correlation coefficient (MCC) by using the geometric mean of sensitivity and precision.

A standard *cmsearch* run reports bit score value of each prediction, which is a statistical indicator measuring the quality of an alignment score. We determined an optimal bit score value for each microRNA family to maximize MCC scores. We then filtered any MirMachine hits lower than the optimal cut-off points. We reported MCC values (and other metrics) before and after filtering.

Benchmarking MirMachine models

We retrained MirMachine CM models by excluding two species: *Homo sapiens* and *Capitella teleta* and compared MirMachine performance on these species. Another benchmarking was done using Rfam models. We downloaded all microRNA models (523 in total) from the Rfam database (v 14). We predicted microRNA families using Rfam models and compared their model performance with MirMachine on selected families (e.g. LET-7, MIR-1, MIR-71, MIR-196). These families were selected because they are highly conserved and contain low false-positives or false negatives in Rfam. We also reported the total number of microRNA predictions done by both methods.

WebApplication implementation

We implemented the web application using a software stack primarily composed of Django, React and Nginx. The application wraps the MirMachine CLI tool to provide a simpler, interactive interface for users. It is hosted at the Norwegian Research and Education Cloud (NREC), utilizing their sHPC (shared High Performance Computing) resources (Trondsen, 2022). MirMachine is available at https://mirmachine.org.

Available Genome Assemblies

Lists of reference genomes of invertebrates, vertebrate mammalians and other vertebrates were downloaded from NCBI GenBank on 1/24/2022 (Clark et al., 2016). Analysis of yearly submitted reference genomes was conducted using Python and customized scripts.

Acknowledgement

B.F. is supported by the Tromsø Research foundation (Tromsø forskningsstiftelse, TFS) [20_SG_BF 'MIRevolution'] and the UiT Aurora Outstanding program 2020-2022. S.U.U and T.B.R were supported by the Research Council of Norway under the Program Human Biobanks and Health Data (grant numbers 229621/H10 and 248791/H10). We thank Wenjing Kang for help with establishing the banner plots, Eirik Høye for help with the structure heatmaps. We would like to thank Fergal Martin and Leanne Haggerty (Ensembl), Terence Murphy (Refseq), Mark Blaxter (Darwin Tree of Life), Blake Sweeney (RNAcentral, Rfam) for discussion on the integration of MirMachine into their services and useful comments. We would like to acknowledge Torbjørn Rognes and Eivind Hovig for administrative help and we are grateful to Norwegian Research and Education Cloud (NREC) for hosting MirMachine.org.

References

Amin, N., McGrath, A., and Chen, Y.-P.P. (2019). Evaluation of deep learning in non-coding RNA classification. Nature Machine Intelligence *1*, 246–256. .

Aparicio-Puerta, E., Gómez-Martín, C., Giannoukakos, S., Medina, J.M., Scheepbouwer, C., García-Moreno, A., Carmona-Saez, P., Fromm, B., Pegtel, M., Keller, A., et al. (2022). sRNAbench and sRNAtoolbox 2022 update: accurate miRNA and sncRNA profiling for model and non-model organisms. Nucleic Acids Res. https://doi.org/10.1093/nar/gkac363.

Axtell, M.J., and Meyers, B.C. (2018). Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. Plant Cell *30*, 272–284. .

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20–51. .

Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science *336*, 233–237. .

Castellano, L., and Stebbing, J. (2013). Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. Nucleic Acids Res. *41*, 3339–3351. .

Cheng, P., Huang, Y., Lv, Y., Du, H., Ruan, Z., Li, C., Ye, H., Zhang, H., Wu, J., Wang, C., et al. (2021). The American Paddlefish Genome Provides Novel Insights into Chromosomal Evolution and Bone Mineralization in Early Vertebrates. Mol. Biol. Evol. *38*, 1595–1607. .

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev. *24*, 992–1009. .

Choi, W.-Y., Giraldez, A.J., and Schier, A.F. (2007). Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. Science *318*, 271–274. .

Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2016). GenBank. Nucleic Acids Res. *44*, D67-72. .

Du, K., Stöck, M., Kneitz, S., Klopp, C., Woltering, J.M., Adolfi, M.C., Feron, R., Prokopov, D., Makunin, A., Kichigin, I., et al. (2020). The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. Nat Ecol Evol *4*, 841–852. .

Dunn, C.W. (2014). Reconsidering the phylogenetic utility of miRNA in animals. Proc. Natl. Acad. Sci. U. S. A. *111*, 12576–12577. .

Eddy, S.R., and Durbin, R. (1994). RNA sequence analysis using covariance models.

Nucleic Acids Res. *22*, 2079–2088. .

Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C., Crottini, A., Godoy, J.A., Höglund, J., et al. (2022). The era of reference genomes in conservation genomics. Trends Ecol. Evol. *37*, 197–202. .

Friedländer, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nat. Biotechnol. *26*, 407–415. .

Fromm, B., Worren, M.M., Hahn, C., Hovig, E., and Bachmann, L. (2013). Substantial loss of conserved and gain of novel MicroRNA families in flatworms. Mol. Biol. Evol. *30*, 2619–2628. .

Fromm, B., Billipp, T., Peck, L.E., Johansen, M., Tarver, J.E., King, B.L., Newcomb, J.M., Sempere, L.F., Flatmark, K., Hovig, E., et al. (2015). A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. Annu. Rev. Genet. *49*, 213–242. .

Fromm, B., Tosar, J.P., Yu, L., Halushka, M.K., and Witwer, K.W. (2018). miR-21-5p and miR-30a-5p are identical in human and bovine, have similar isomiR distribution, and cannot be used to identify xenomiR uptake from cow milk.

Fromm, B., Kang, W., Rovira, C., Cayota, A., Witwer, K., Friedländer, M.R., and Tosar, J.P. (2019a). Plant microRNAs in human sera are likely contaminants. J. Nutr. Biochem. *65*, 139–140. .

Fromm, B., Tarbier, M., Smith, O., Dalén, L., Gilbert, M.T.P., and Friedländer, M.R. (2019b). Ancient microRNA profiles of a 14,300-year-old canid are taxonomically informative and give glimpses into gene regulation from the Pleistocene.

Fromm, B., Domanska, D., Høye, E., Ovchinnikov, V., Kang, W., Aparicio-Puerta, E., Johansen, M., Flatmark, K., Mathelier, A., Hovig, E., et al. (2020a). MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Res. *48*, D132–D141. .

Fromm, B., Keller, A., Yang, X., Friedlander, M.R., Peterson, K.J., and Griffiths-Jones, S. (2020b). Quo vadis microRNAs? Trends Genet. *36*, 461–463. .

Fromm, B., Zhong, X., Tarbier, M., Friedlander, M.R., and Hackenberg, M. (2022a). The limits of human microRNA annotation have been met. RNA https://doi.org/10.1261/rna.079098.122.

Fromm, B., Patil, A.H., and Halushka, M.K. (2022b). A Novel Circulating MicroRNA for the Detection of Acute Myocarditis. N. Engl. J. Med. *387*, 1240. .

Fromm, B., Høye, E., Domanska, D., Zhong, X., Aparicio-Puerta, E., Ovchinnikov, V., Umu, S.U., Chabot, P.J., Kang, W., Aslanzadeh, M., et al. (2022c). MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res. *50*, D204–D210. .

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75–79. .

Guo, Z., Kuang, Z., Wang, Y., Zhao, Y., Tao, Y., Cheng, C., Yang, J., Lu, X., Hao, C., Wang, T., et al. (2020). PmiREN: a comprehensive encyclopedia of plant miRNAs. Nucleic Acids Res. *48*, D1114–D1121. .

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M., and Aransay, A.M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res. *37*, W68-76. .

Heimberg, A.M., Sempere, L.F., Moy, V.N., Donoghue, P.C.J., and Peterson, K.J. (2008). MicroRNAs and the advent of vertebrate morphological complexity. Proc. Natl. Acad. Sci. U. S. A. *105*, 2946–2950. .

Hotaling, S., Kelley, J.L., and Frandsen, P.B. (2021). Toward a genome sequence for every animal: Where are we now? Proc. Natl. Acad. Sci. U. S. A. *118*. https://doi.org/10.1073/pnas.2109019118.

Huang, Z., Jebb, D., and Teeling, E.C. (2016). Blood miRNomes and transcriptomes reveal novel longevity mechanisms in the long-lived bat, Myotis myotis. BMC Genomics *17*, 906. .

Jebb, D., Huang, Z., Pippel, M., Hughes, G.M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L.S., Skirmuntt, E.C., Katzourakis, A., et al. (2020). Six reference-quality genomes reveal evolution of bat adaptations. Nature *583*, 578–584. .

Jones-Rhoades, M.W. (2012). Conservation and divergence in plant microRNAs. Plant Mol. Biol. *80*, 3–16. .

Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. *49*, D192–D200. .

Kang, W., Eldfjell, Y., Fromm, B., Estivill, X., Biryukova, I., and Friedländer, M.R. (2018). miRTrace reveals the organismal origins of microRNA sequencing data. Genome Biol. *19*, 213. .

Kang, W., Fromm, B., Houben, A.J., Høye, E., Bezdan, D., Arnan, C., Thrane, K., Asp, M., Johnson, R., Biryukova, I., et al. (2021). MapToCleave: High-throughput profiling of microRNA biogenesis in living cells. Cell Rep. *37*, 110015. .

Katoh, K., Rozewicki, J., and Yamada, K.D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. *20*, 1160–1166. .

Langenberger, D., Bartschat, S., Hertel, J., Hoffmann, S., Tafer, H., and Stadler, P.F.

(2011). MicroRNA or Not MicroRNA? In Advances in Bioinformatics and Computational Biology, (Springer Berlin Heidelberg), pp. 1–9.

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proc. Natl. Acad. Sci. U. S. A. *115*, 4325–4333. .

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26. .

Ludwig, N., Becker, M., Schumann, T., Speer, T., Fehlmann, T., Keller, A., and Meese, E. (2017). Bias in recent miRBase annotations potentially associated with RNA quality issues. Sci. Rep. *7*, 5162. .

Mendell, J.T., and Olson, E.N. (2012). MicroRNAs in stress signaling and human disease. Cell *148*, 1172–1187. .

Meng, Y., Shao, C., Wang, H., and Chen, M. (2012). Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants. RNA Biol. *9*, 249–253. .

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935. .

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., Omrak, A., Vartanyan, S., Poinar, H., Götherström, A., et al. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. Curr. Biol. *25*, 1395–1400. .

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A.M., To, T.-H., Kortschak, R.D., et al. (2018). A comprehensive genomic history of extinct and living elephants. Proc. Natl. Acad. Sci. U. S. A. *115*, E2566–E2574. .

Peterson, K.J., Dietrich, M.R., and McPeek, M.A. (2009). MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. Bioessays *31*, 736–747. .

Peterson, K.J., Beavan, A., Chabot, P.J., McPeek, M.A., Pisani, D., Fromm, B., and Simakov, O. (2021). microRNAs as Indicators into the Causes and Consequences of Whole Genome Duplication Events. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msab344.

Sacar, M.D., Hamzeiy, H., and Allmer, J. (2013). Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? J. Integr. Bioinform. *10*, 215. .

Saçar Demirci, M.D., Baumbach, J., and Allmer, J. (2017). On the performance of pre-microRNA detection algorithms. Nat. Commun. *8*, 330. .

Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., et al. (2016). Genome evolution in the allotetraploid frog Xenopus laevis. Nature *538*, 336–343. .

Stegmayer, G., Di Persia, L.E., Rubiolo, M., Gerard, M., Pividori, M., Yones, C., Bugnon, L.A., Rodriguez, T., Raad, J., and Milone, D.H. (2019). Predicting novel microRNA: a comprehensive comparison of machine learning approaches. Brief. Bioinform. *20*, 1607–1620. .

Tarver, J.E., Donoghue, P.C., and Peterson, K.J. (2012). Do miRNAs have a deep evolutionary history? Bioessays *34*, 857–866. .

Tarver, J.E., Sperling, E.A., Nailor, A., Heimberg, A.M., Robinson, J.M., King, B.L., Pisani, D., Donoghue, P.C.J., and Peterson, K.J. (2013). miRNAs: small genes with big potential in metazoan phylogenetics. Mol. Biol. Evol. *30*, 2369–2382. .

Tarver, J.E., Taylor, R.S., Puttick, M.N., Lloyd, G.T., Pett, W., Fromm, B., Schirrmeister, B.E., Pisani, D., Peterson, K.J., and Donoghue, P.C.J. (2018). Well-Annotated microRNAomes Do Not Evidence Pervasive miRNA Loss. Genome Biol. Evol. *10*, 1457–1470. .

Taylor, R.S., Tarver, J.E., Hiscock, S.J., and Donoghue, P.C. (2014). Evolutionary history of plant microRNAs. Trends Plant Sci. https://doi.org/10.1016/j.tplants.2013.11.008.

Taylor, R.S., Tarver, J.E., Foroozani, A., and Donoghue, P.C.J. (2017). MicroRNA annotation of plant genomes- Do it right or not at all. Bioessays *39*, 1600113. .

Thomson, R.C., Plachetzki, D.C., Mahler, D.L., and Moore, B.R. (2014). A critical appraisal of the use of microRNA data in phylogenetics. Proc. Natl. Acad. Sci. U. S. A. *111*, E3659-68. .

Trondsen, H.T. (2022). A web application for MirMachine, a MicroRNA annotation tool.

Umu, S.U., Langseth, H., Zuber, V., Helland, Å., Lyle, R., and Rounge, T.B. (2022). Serum RNAs can predict lung cancer up to 10 years prior to diagnosis. Elife *11*. https://doi.org/10.7554/eLife.71035.

Velandia-Huerto, C.A., Fallmann, J., and Stadler, P.F. (2021). miRNAture—Computational Detection of microRNA Candidates. Genes *12*, 348. .

Wang, X., and Liu, X.S. (2011). Systematic Curation of miRBase Annotation Using Integrated Small RNA High-Throughput Sequencing Data for C. elegans and Drosophila. Front. Genet. *2*, 25. .

Wang, J., Chen, J., and Sen, S. (2016). MicroRNA as Biomarkers and Diagnostics. J. Cell. Physiol. *231*, 25–30. .

Whalen, S., Schreiber, J., Noble, W.S., and Pollard, K.S. (2021). Navigating the pitfalls

of applying machine learning in genomics. Nat. Rev. Genet. https://doi.org/10.1038/s41576-021-00434-9.

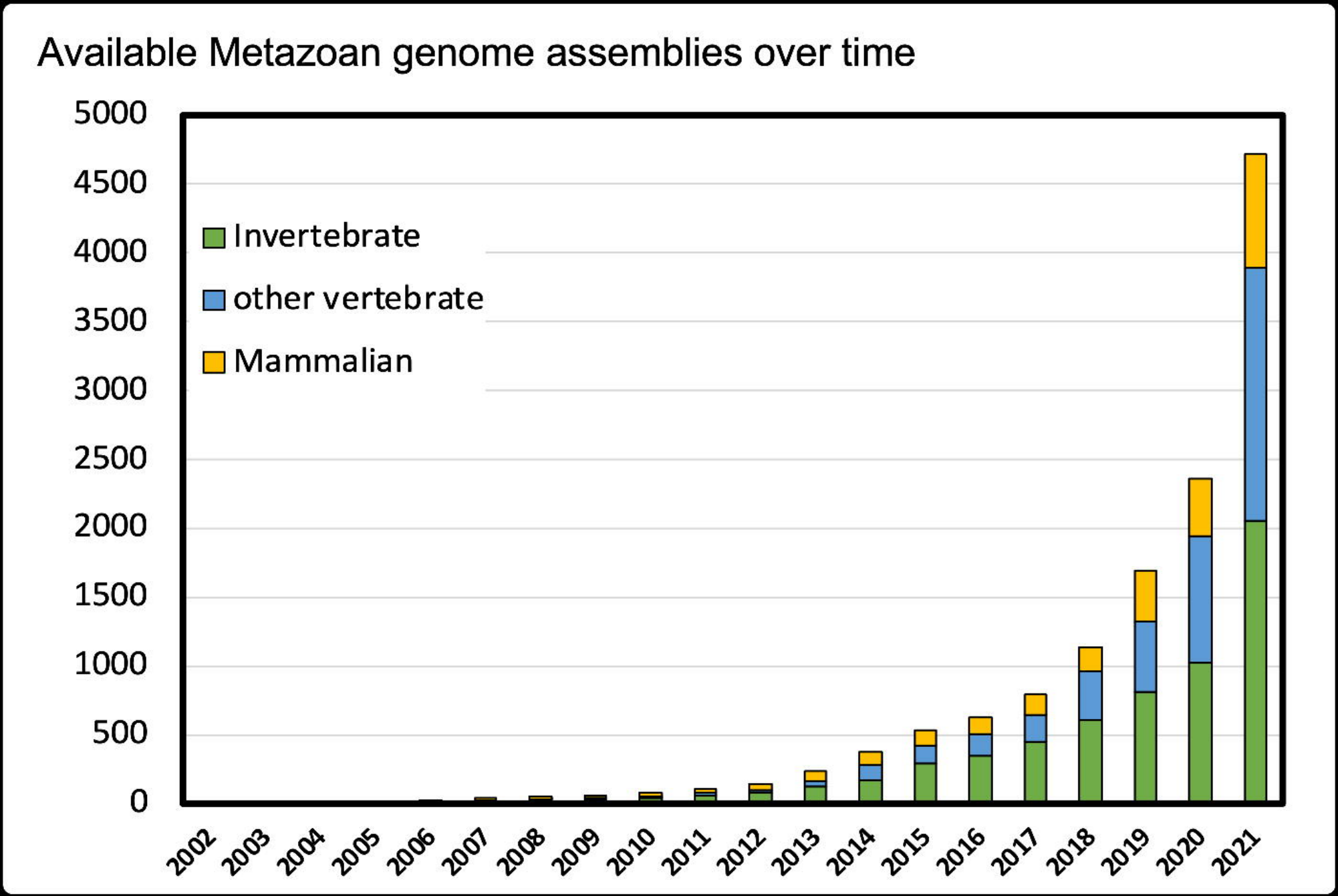
Wheeler, T.J., and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. Bioinformatics *29*, 2487–2489. .

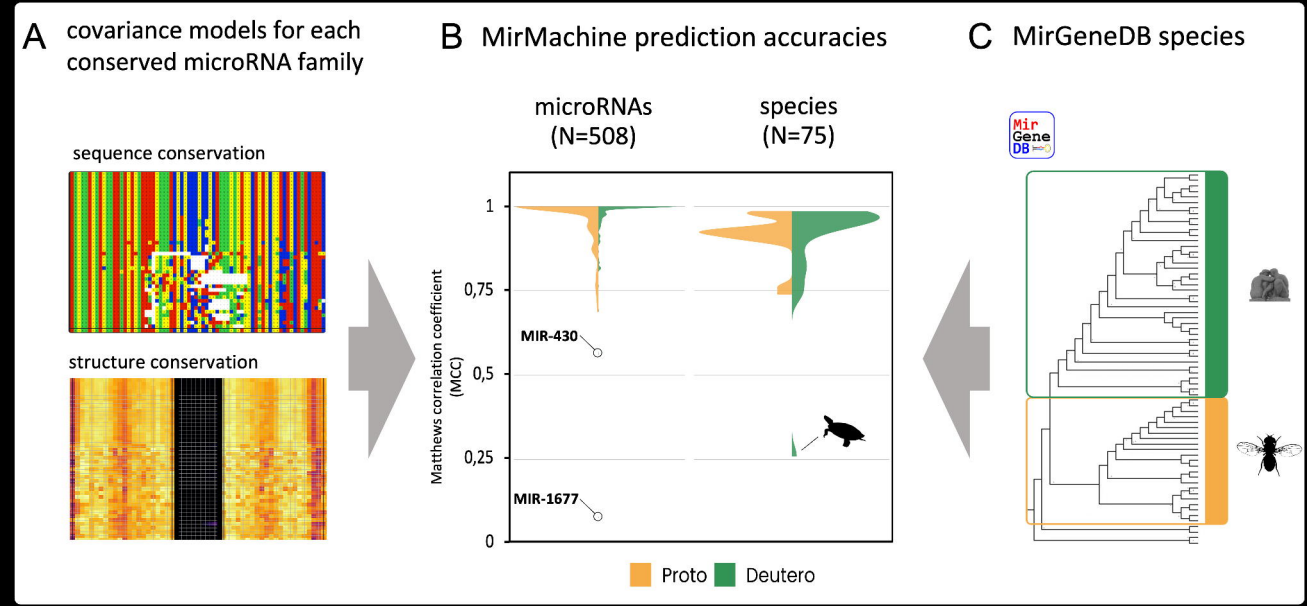
Wheeler, B.M., Heimberg, A.M., Moy, V.N., Sperling, E.A., Holstein, T.W., Heber, S., and Peterson, K.J. (2009). The deep evolution of metazoan microRNAs. Evol. Dev. *11*, 50–68. .

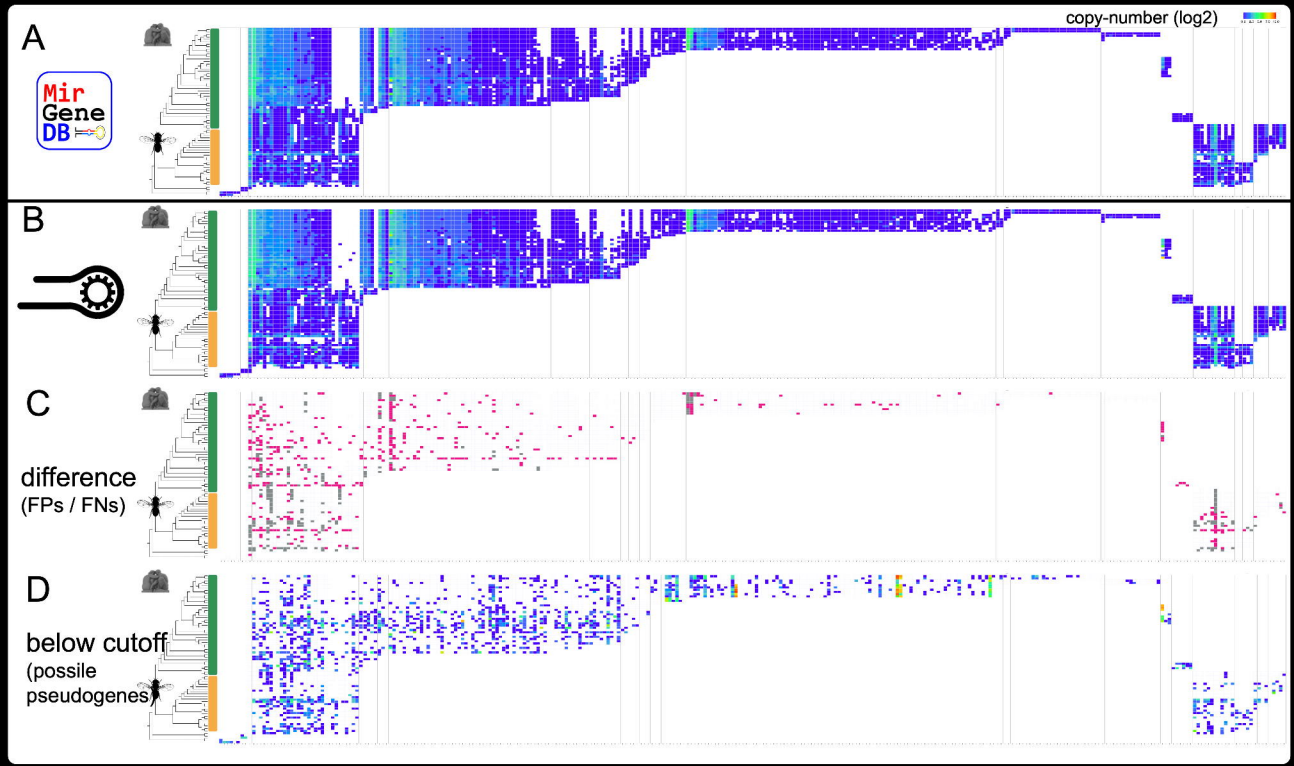
Witwer, K.W., and Halushka, M.K. (2016). Toward the promise of microRNAs - Enhancing reproducibility and rigor in microRNA research. RNA Biol. *13*, 1103–1116. .

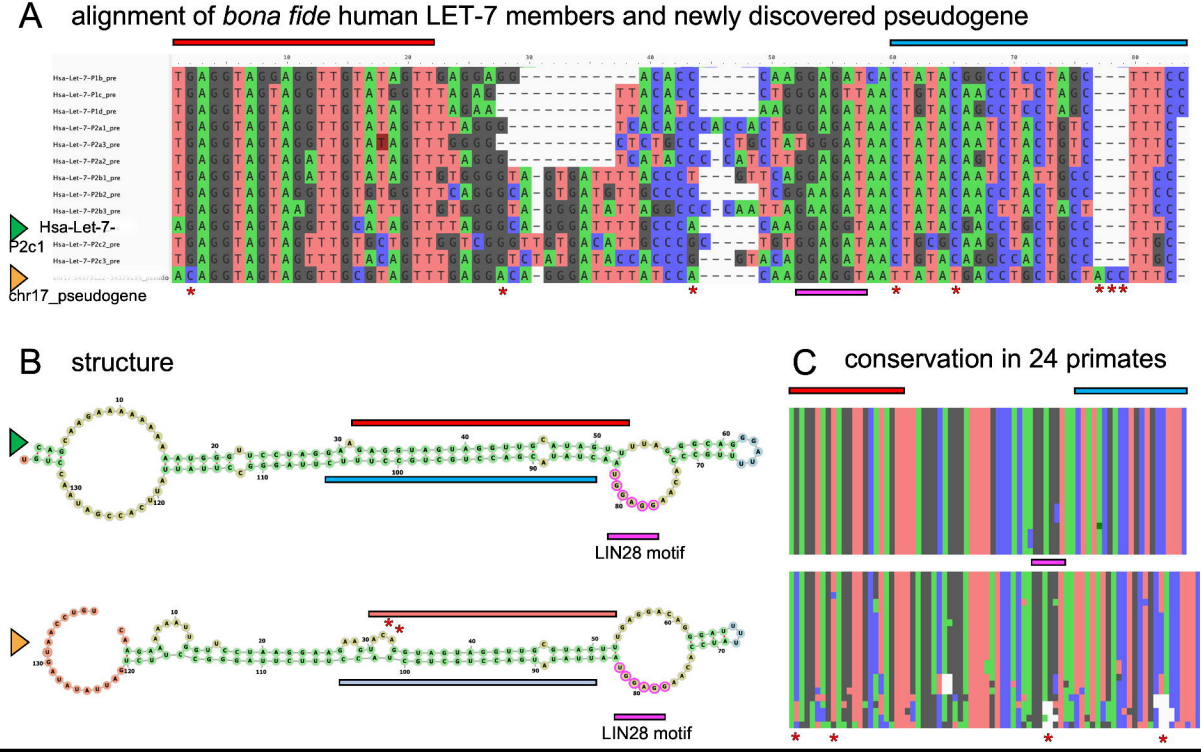
Yazbeck, A.M., Tout, K.R., Stadler, P.F., and Hertel, J. (2017). Towards a Consistent, Quantitative Evaluation of MicroRNA Evolution. J. Integr. Bioinform. *14*. https://doi.org/10.1515/jib-2016-0013.

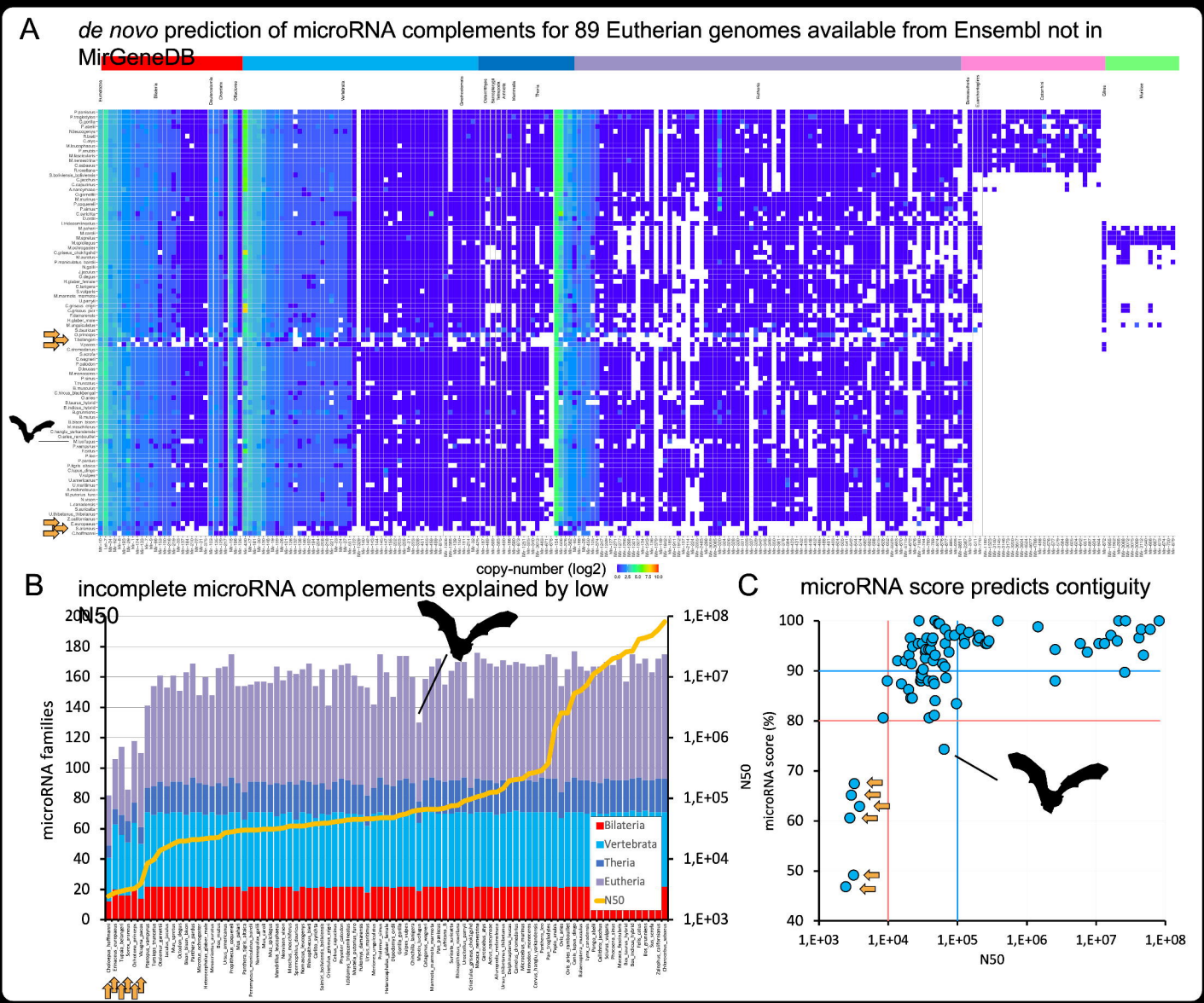
Zolotarov, G., Fromm, B., Legnini, I., Ayoub, S., Polese, G., Maselli, V., Chabot, P.J., Vinther, J., Styfhals, R., Seuntjens, E., et al. (2022). MicroRNAs are deeply linked to the emergence of the complex octopus brain.

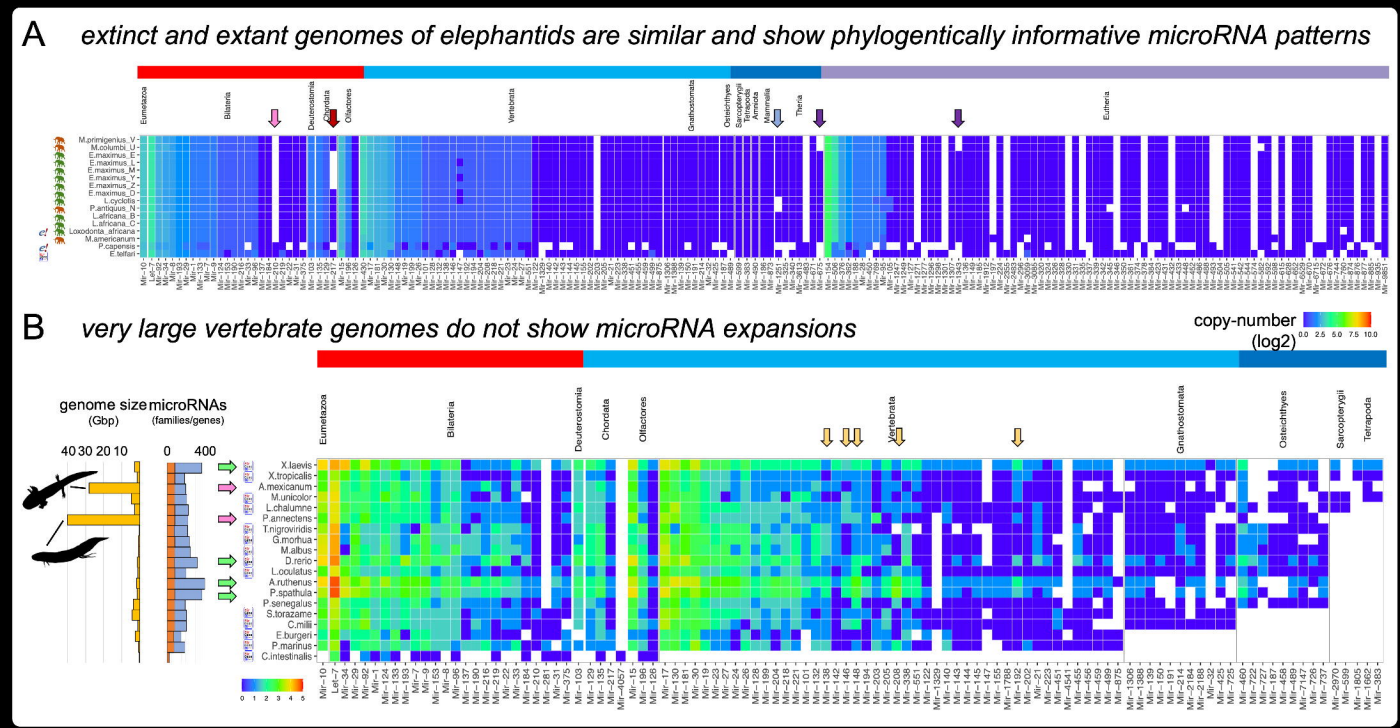

Available Metazoan genome assemblies over time
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Invertebrate
other vertebrate
Mammalian
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

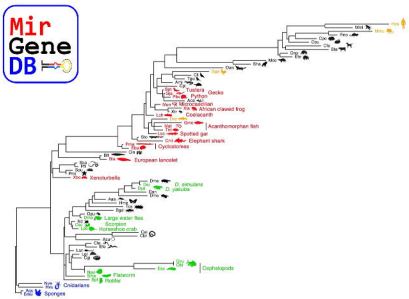

A
covariance models for each conserved microRNA family
sequence conservation
structure conservation
B
MirMachine prediction accuracies
microRNAs (N=508)
species (N=75)
Matthews correlation coefficient (MCC)
1
0,75
0,5
0,25
0
MIR-430
MIR-1677
Proto
Deutero
C
MirGeneDB species


copy-number (log2)
A
Mir
Gene
DB
B
C
difference
(FPs / FNs)
D
below cutoff
(possile
pseudogenes)


A
alignment of bona fide human LET-7 members and newly discovered pseudogene
Hsa-Let-7-P2c1
chr17_pseudogene
B
structure
LIN28 motif
LIN28 motif
C
conservation in 24 primates


A
de novo prediction of microRNA complements for 89 Eutherian genomes available from Ensembl not in MirGeneDB
copy-number (log2)
B
incomplete microRNA complements explained by low N50
microRNA families
N50
Bilateria
Vertebrata
Theria
Eutheria
N50
C
microRNA score predicts contiguity
microRNA score (%)
N50


A
extinct and extant genomes of elephantids are similar and show phylogentically informative microRNA patterns
Eumetazoa
Bilateria
Deuterostomia
Chordata
Olfactores
Vertebrata
Gnathostomata
Osteichthyes
Sarcopterygii
Tetrapoda
Mammalia
Theria
Eutheria
M.primigenius_V
M.columbi_U
E.maximus_E
E.maximus_L
E.maximus_M
E.maximus_Y
E.maximus_Z
E.maximus_O
L.cyclotis
P.antiquus_N
L.africana_B
L.africana_C
Loxodonta_africana
M.americanum
P.capensis
E.telfari
B
very large vertebrate genomes do not show microRNA expansions
copy-number (log2)
0.0 2.5 5.0 7.5 10.0
genome size (Gbp)
microRNAs (families/genes)
40 30 20 10
0
400
X.laevis
X.tropicalis
A.mexicanum
M.unicolor
L.chalumne
P.annectens
T.nigroviridis
G.morhua
M.albus
D.rerio
L.oculatus
A.ruthenus
P.spathula
P.senegalus
S.torazame
C.milii
E.burgeri
P.marinus
C.intestinalis
0 1 2 3 4 5

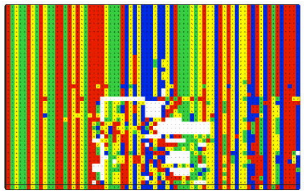


16667 microRNA genes
1549 microRNA families
75 metazoan species

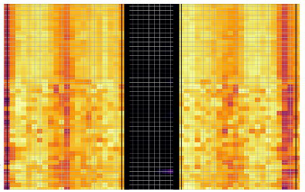
125 phylogenetic nodes

alignments for each family

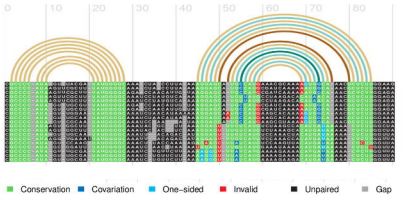
sequence conservation



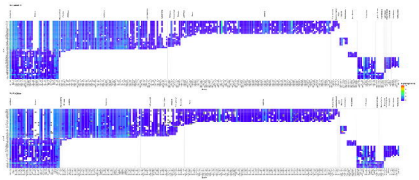
structure conservation



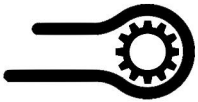
covariance models (CM) for each family



train models with bitscore cutoff



annotation of conserved microRNA complements from genomes



- search for CMs of each microRNA family in user defined genomes
- taxonomy-informed restriction on range of microRNA CM search
- optional bitscore-based filtering of results
- gff, fasta outputs of search results / microRNA complements