Published online 1 June 2023

Nucleic Acids Research, 2023, Vol. 51, No. 13 ***e71***
https://doi.org/10.1093/nar/gkad467

ncOrtho: efficient and reliable identification of miRNA orthologs

Felix Langschied[1,*], Matthias S. Leisegang[2,3], Ralf P. Brandes[2,3] and Ingo Ebersberger[1,4,5,*]

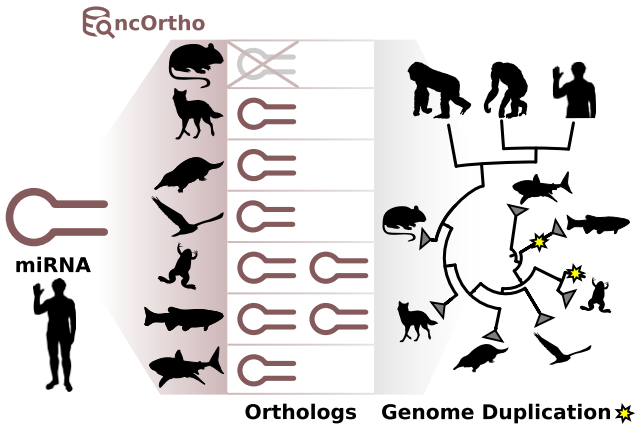
[1]Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt, Germany, [2]Institute for Cardiovascular Physiology, Goethe University, Frankfurt, Germany, [3]German Center of Cardiovascular Research (DZHK), Partner site RheinMain, Frankfurt, Germany, [4]Senckenberg Biodiversity and Climate Research Centre (S-BIK-F), Frankfurt am Main, Germany and [5]LOEWE Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

Received December 06, 2022; Revised May 04, 2023; Editorial Decision May 11, 2023; Accepted May 30, 2023

ABSTRACT

MicroRNAs (miRNAs) are post-transcriptional regulators that finetune gene expression via translational repression or degradation of their target mRNAs. Despite their functional relevance, frameworks for the scalable and accurate detection of miRNA orthologs are missing. Consequently, there is still no comprehensive picture of how miRNAs and their associated regulatory networks have evolved. Here we present ncOrtho, a synteny informed pipeline for the targeted search of miRNA orthologs in unannotated genome sequences. ncOrtho matches miRNA annotations from multi-tissue transcriptomes in precision, while scaling to the analysis of hundreds of custom-selected species. The presence-absence pattern of orthologs to 266 human miRNA families across 402 vertebrate species reveals four bursts of miRNA acquisition, of which the most recent event occurred in the last common ancestor of higher primates. miRNA families are rarely modified or lost, but notable exceptions for both events exist. miRNA co-ortholog numbers faithfully indicate lineage-specific whole genome duplications, and miRNAs are powerful markers for phylogenomic analyses. Their exceptionally low genetic diversity makes them suitable to resolve clades where the phylogenetic signal is blurred by incomplete lineage sorting of ancestral alleles. In summary, ncOrtho allows to routinely consider miRNAs in evolutionary analyses that were thus far reserved to protein-coding genes.

GRAPHICAL ABSTRACT



INTRODUCTION

MicroRNAs (miRNAs) are single-stranded, short non-coding RNAs of ∼22 nucleotides (nt) in length that have essential roles in fine-tuning the expression network of eukaryotic cells (1–3). Canonical miRNAs are transcribed by RNA Polymerase II resulting in primary miRNAs (pri-miRNA) that form at least one distinctive hairpin structure. Pri-miRNA are cleaved by the Microprocessor complex, which contains Drosha and DiGeorge syndrome critical region 8, resulting in microRNA precursors (pre-miRNA) (4). Pre-miRNAs are approx. 55–70 nt long sequences, which are exported into the cytoplasm by Exportin 5 where the terminal loop of the pre-miRNAs is removed to create miRNA duplexes (4,5). This duplex is subsequently resolved, and one of the two strands is loaded into an Argonaut protein to form the mature RNA-induced silencing complex (RISC) (6,7). Target mRNAs are silenced post-transcriptionally

*To whom correspondence should be addressed. Tel: +49 69 798 42112; Fax: +49 69 798 42111; Email: ebersberger@bio.uni-frankfurt.de
Correspondence may also be addressed to Felix Langschied. Tel: +49 69 798 42118; Email: langschied@bio.uni-frankfurt.de

through binding of the RISC, which is induced by a pairing of the 7 nt long seed region in the mature miRNA and a 6–8 nt long target site in the mRNA (8).

miRNA-dependent gene regulation in animals dates at least back to the last common ancestor of the Eumetazoa that lived more than 800 million years ago (9,10). The acquisition of novel miRNA families in the course of evolution has been correlated with body-plan innovations and an increase of overall morphological complexity (11–14). Novel miRNAs can emerge via the repurposing of already transcribed sequences, such as parts of introns or protein-coding exons (15,16). Once these miRNAs are integrated into a regulatory network, purifying selection contributes to their preservation (11). Still, findings that suggest the loss of evolutionary old and hence well integrated miRNA families were presented (17,18), but these were contrasted by the claim that the loss of established miRNAs is exceedingly rare (19,20). It was argued that false-positive miRNA annotations in the miRBase repository (21), which served as the basis of the analyses, paired with a limited sensitivity of the miRNA homolog search created a spurious signal of miRNA loss (22). Meanwhile the manually curated MirGeneDB has put miRNA-based research on a more solid foundation (19). In this database, each miRNA entry is supported by corresponding transcript data and meets stringent annotation criteria. Moreover, MirGeneDB embeds miRNA annotations into an evolutionary context, and version 2.1 of the database provides information about the representation of miRNA genes across 75 species (23). While this increases the specificity of miRNA annotations, MirGeneDB faces two main challenges: first, miRNAs that are expressed at low levels or under specific conditions are prone to be missed. Second, the taxonomic resolution and therefore the evolutionary information content in the data will remain low because the requirement of multi-tissue transcriptomic datasets to support miRNA calls renders the integration of novel taxa into MirGeneDB cost- and labor-intensive. For many protected or rare species, it is even virtually impossible to gather the necessary data, particularly if their export falls under international regulations for species transfer in the context of CITES (https://cites.org/eng/prog/Permit_system).

Scanning genome sequences for the presence of orthologs to miRNAs in MirGeneDB offers a powerful alternative. It allows to propagate miRNA annotation across species independent from the availability of transcriptome data. Genome sequences abound in public databases and can be obtained even with non-invasive sampling (e.g. 24,25). Thus, large and evolutionary diverse taxon collections can be analyzed, which is crucial for improving the signal-to-noise ratio in evolutionary analyses (26). For protein-coding genes, the identification of orthologs and the generation of comprehensive phylogenetic profiles, i.e. the presence/absence pattern of orthologs across large taxon sets, is well established (e.g. 27). In the case of miRNAs, the tools available remain sparse. Individual solutions for genome-wide scans for miRNA homologs have been proposed, but published approaches either lack publicly available software implementations (e.g. 17,28), or are limited to pre-defined taxon sets (e.g. MapMi; 29). To our knowledge, the only currently available method to identify miRNA orthologs in custom assemblies requires the alignment of whole genome sequences (30). While this approach is straightforward, in principle, the computational overhead is immense. For example, an alignment of 242 placental mammalian genomes took two months of computation time on the Amazon Cloud utilizing 260 instances, each equipped with 32 virtual CPUs (31). Moreover, aligning whole genomes of species covering the full diversity of vertebrates is difficult (32). Therefore, only a small fraction of the genome sequences that are currently being generated (e.g. 32–34) will be considered in whole genome alignments.

To close this methodological gap, we have developed ncOrtho, the first software that facilitates a targeted search for miRNA orthologs in individual genome sequences. ncOrtho scales linearly in time with both the number of miRNAs and the number of taxa included in the ortholog search, and therefore enables miRNA ortholog searches in large and customizable genome collections. ncOrtho identifies orthologs with high sensitivity and specificity irrespective of the genome annotation status. The resulting phylogenetic profiles provide the basis for studying miRNA evolution at an unprecedented scale and represent the first step towards projecting regulatory networks of miRNAs from model- to non-model organisms.

MATERIALS AND METHODS

Data

Genome assemblies of 161 mammalian species, 241 non-mammalian vertebrates, and 16 invertebrate species were downloaded from NCBI Refseq Genomes release 207 (33; Supplementary Table S1). Locations and pre-miRNA sequences for all 556 human miRNAs were downloaded from MirGeneDB v2.0. The MirGeneDB nomenclature was derived from miRBase (19,34), and mapping between the nomenclatures is available on the MirGeneDB webpage. Whole genome alignments were downloaded from the 100 vertebrate alignment track in the UCSC Genome Browser (35).

Covariance models

Pairwise orthology assignments between protein-coding genes were identified with OMA standalone v2.4.1 (36) and served as anchors to identify shared syntenic regions between the genomes of humans and each core species. Positional miRNA orthologs were identified in the shared syntenic regions as specified in the main text and were used to generate and train a Covariance Model (CM) for each human miRNA. Sequences in the individual training sets were aligned with R-Coffee (37) and were annotated with secondary structure information calculated by RNAalifold from the Vienna RNA package 2.0 (38). The multiple sequence alignment together with the secondary structure information were then used as input for the Infernal package to train and calibrate the corresponding CM in default settings (39). CM-based searches were performed with the *cmsearch* algorithm provided in the Infernal v1.1 package, and search results were filtered for sequences that achieve at least 50% of the maximally achievable bit score (i.e. the score of the reference miRNA given the CM).

Evaluation of orthology assignment performance

MirGeneDB provided the gold-standard to evaluate orthology assignments by ncOrtho (9). For individual species, MirGeneDB used genome assemblies other than those deposited in NCBI RefSeq Genome. In these cases, we mapped the miRNAs stored in MirGeneDB to the NCBI RefSeq assembly of the corresponding species using BLASTn (cutoff: ≥90% identity and ≥80% query coverage) (40). During the software benchmark, we compared the location of each predicted miRNA ortholog to the genomic location of MirGeneDB entries from the same miRNA family. In case of an overlap, the candidate was accepted as a true positive (TP). Results that cover a genomic region with no matching entry in MirGeneDB were tentatively assigned as false positives (FP). An assignment was considered as false negative (FN) if the corresponding MirGeneDB entry was not identified as an ortholog. All cases in which no ortholog was detected and MirGeneDB has no entry for that family and species were counted as true negatives (TN). We then calculated the sensitivity as TP/(TP + FN), the specificity as TN/(TN + FP), the accuracy as (TP + TN)/(TP + FP + FN + TN), and the *F*1-score as (2 × TP)/(2 * TP + FP + FN).

Alternative approaches of miRNA ortholog identification

Whole Genome alignments: We downloaded the alignment blocks of the 100 Vertebrate alignment from the UCSC Browser FTP-server (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/maf), and extracted the longest alignment block covering each human pre-miRNA locus annotated in MirGeneDB (9). We considered an ortholog candidate to be found in a species if its sequence covered at least 70% of the human miRNA locus excluding gaps. BLASTn search: We used the pre-miRNA sequences from MirGeneDB as input for a local BLASTn search (40) using the 'blastn' task with default parameter settings. The best hit was used as the miRNA ortholog candidate. MapMi: Mature miRNA sequences provided by MirGeneDB were used as input for the MapMi implementation on the EMBL-EBI webserver using the 59 'Ensembl Species' set (41,42).

Population variation analysis

Human single nucleotide polymorphism (SNP) data was retrieved from dbSNP build 153 (43). We considered only data from gnomAD v2.1 Genomes, which contains variants from 15708 human genomes (44), to ensure comparability between the SNP counts for the following seven genomic regions: (i) mature, (ii) star and (iii) loop region of a miRNA, the 30 nt (iv) upstream- and (v) downstream flanking regions that are annotated by MirGeneDB, (vi) all CDS and (vii) all lncRNA genes of the GRCh38 assembly. SNP counts for each region were normalized by their respective length and then multiplied by 1000 resulting in the SNP density per 1kb.

Phylogenetic analyses

The presence/absence patterns of human miRNAs were visualized using PhyloProfile (45) once on the level of individual genes and once on the miRNA family level. The evolutionary emergence of a miRNA family was tentatively dated to the last common ancestor of the two most distantly related species in which an ortholog from the corresponding family was detected. miRNA orthologs were aligned with MUSCLE v3.8.155 (46). The individual alignments were concatenated and alignment columns with more than 50% gaps were removed. ModelFinder (47) identified the TIM3 + F + R5 evolutionary model as the best fitting model for the data, and the maximum likelihood tree was calculated using IQ-TREE v1.6.8 with 1000 ultra-fast bootstrap replicates (48,49). The full pipeline used for reconstructing pre-miRNA sequence trees is implemented into the ncOrtho package. Phylogenetic trees were visualized and annotated with iTOL (50) or the ETE 3 toolkit (51). Competing phylogenetic hypothesis testing was performed with the Approximately Unbiased (AU) test as implemented in IQ-TREE (49).

RESULTS

miRNAs are not yet routinely considered in large-scale evolutionary analyses because scalable approaches for the reliable identification of miRNA orthologs across large and phylogenetically diverse taxon collections were missing. We therefore developed ncOrtho to identify orthologs of miRNA genes in genome assemblies irrespective of their annotation status. Key to this approach is the use of Covariance Models (CMs), which are probabilistic models that integrate the consensus nucleotide sequence with the secondary structure of a functional RNA and facilitate a sensitive and discriminative homolog identification (52,53). The phylogenetic profiles generated by ncOrtho provide detailed insights into the taxonomic distribution, the minimal evolutionary age, and the duplication/deletion history of the analyzed miRNAs.

Algorithm

ncOrtho accepts an individual pre-miRNA (the 'reference miRNA') together with its genomic position in the reference species as input. Additionally, two lists of taxa together with the corresponding genome sequences are required. The first list defines a set of 'core species' that are considered in the training phase of the covariance model (Figure 1A). Core species should be closely related enough for microsynteny to be conserved to an extent that a comprehensive identification of positional miRNA orthologs is possible. Additionally, the last common ancestor of the core species should not be older than the miRNA gene of interest. To facilitate core species selection, we supply a function with which microsynteny conservation can be estimated as part of the ncOrtho package (Supplementary Figure S1). For the core species, the annotation of all protein-coding genes as well as the pairwise orthology assignments to the proteins encoded in the reference species are needed. The second list specifies the target species whose genomes should be scanned for the presence of miRNA orthologs. ncOrtho then follows a two-stage procedure to accomplish the ortholog search (see Figure 1). Initially, a set of high confidence orthologs is compiled from the core species which are then used for CM con-

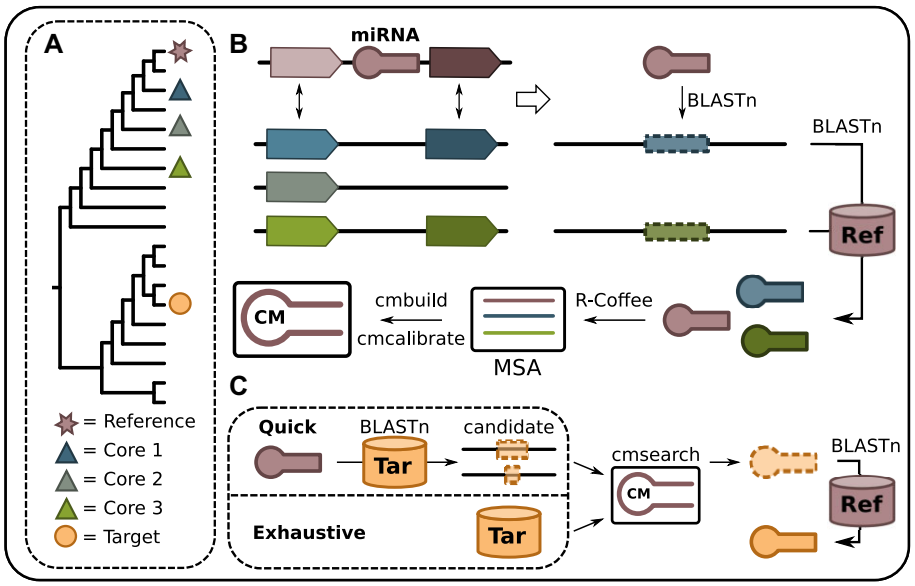


Figure 1. Workflow of ncOrtho. (**A**) Taxon selection. The species that harbors the miRNA in focus is selected as 'reference'. 'Core species' are used to identify positional miRNA orthologs for Covariance Model (CM) training. The CM-based miRNA ortholog search is performed in the 'target' species. (**B**) Covariance Model training. Positional miRNA ortholog candidates are identified via BLASTn in regions of conserved protein-coding gene order. Candidates are confirmed as orthologs if a reverse BLASTn search in the reference genome (Ref) identifies the reference miRNA as best hit. Positional miRNA orthologs are aligned, and the resulting multiple sequence alignment (MSA) is then used for CM training. (**C**) Targeted ortholog search. In 'quick mode', the reference miRNA is used as a query for a BLASTn search against the target genome (Tar) to narrow the search space to candidate regions of ∼2000 bp in length. In the exhaustive mode, the entire genome is used. A cmsearch using the pre-trained CM identifies then putative miRNA orthologs. Candidate orthologs are confirmed if a BLASTn search in the reference genome obtains the reference miRNA as the best hit.

struction and training. Subsequently, the CM is utilized in the search for miRNA orthologs in all target species.

Covariance model training. ncOrtho uses a set of high confidence orthologs to the reference miRNA for CM construction. This training data is compiled by exploiting collinearity of homologous genomic segments (shared synteny) to identify positional miRNA orthologs in the core species. We distinguish two cases: If the miRNA is positioned between two protein-coding genes, we use the flanking genes as syntenic anchors and consider a region in the core species' genome as shared syntenic if no more than k protein-coding genes separate the orthologs of the anchor genes (parameter –mgi; Default: $k = 3$). To account for gene losses or gene annotation artefacts, we provide the option to consider up to n flanking genes as alternative syntenic anchors (parameter –max_anchor_dist; Default: $n = 1$) (Figure 1B). If the miRNA is located within a protein-coding gene, the shared syntenic region in the core species is the genomic locus harboring the ortholog of the protein-coding gene. If no ortholog could be detected, the algorithm proceeds as if the miRNA-gene would be in an intergenic region. Upon the identification of a shared syntenic region, its sequence is extracted and a reciprocal best BLASTn hit approach using the reference pre-miRNA sequence as query identifies a miRNA ortholog, if present (54). This procedure is repeated for all core species, which results in a set of positional orthologs that are used to train a covariance model of the reference miRNA.

Targeted miRNA ortholog search using covariance models. The genome sequence of each target species is scanned for the presence of a miRNA using the corresponding CM (Figure 1C). Candidates represented by CM search hits meeting the score threshold (see methods) are accepted as an ortholog if a BLASTn search against the reference genome using the candidate as query reveals the reference miRNA as the best hit. If two or more candidate orthologs for the same reference miRNA are confirmed by the reverse search, all are kept as co-orthologs. The runtime complexity of CM searches renders scans of entire genomes across larger taxon sets time consuming (Supplementary Figure S2). Therefore, ncOrtho provides the option to reduce the search space by first searching for subsequences in the target genome that share a local sequence similarity to the reference miRNA ('Quick' mode; Figure 1C). Hits are then extended by 1000 nucleotides (nt) up- and downstream, and the resulting candidate regions serve as input for a subsequent refined search using the reference miRNA-specific CM.

Benchmarking ncOrtho orthology assignments

Performance. To evaluate ncOrtho, we used the manually curated miRNA database MirGeneDB as a gold standard (19). We identified orthologs to 556 miRNA genes representing 266 miRNA families. *Homo sapiens* served as the reference species, and *Macaca mulatta, Gorilla gorilla*, *Pongo abelii* and *Nomascus leucogenys* were used as core species. 242 miRNA genes were located within, and 314 between protein-coding genes, and there was no difference in the core ortholog set sizes between the two groups (Supplementary Figure S3). For the benchmark, we concentrated on 20 vertebrate species that were represented both in MirGeneDB 2.0 and in the RefSeq partition of the

NCBI genome database. We further included 16 invertebrate animals to sound out the sensitivity limits of ncOrtho in miRNA families that are conserved across the Bilateria (9). Table 1 shows that sensitivity and specificity of ncOrtho in Quick mode are very high across all analyzed vertebrates. Using this mode, ncOrtho identifies orthologs with a median runtime of 0.7 s per miRNA and species (Supplementary Figure S4). In the invertebrates, orthologs are identified with near perfect specificity but the sensitivity drops substantially. Repeating the ortholog search in 'Exhaustive' mode (see Figure 1C) had almost no effect on the sensitivity, but the run time increased by two orders of magnitude (Table 1, Supplementary Figure S2). Thus, the search heuristic used in the Quick mode of ncOrtho does not account for the sensitivity drop. On closer inspection, we noticed that the missing of invertebrate miRNA orthologs coincides with changes in the loop and star region that are specific to the invertebrate orthologs (Supplementary Figure S5).

ncOrtho orthologs without MirGeneDB confirmation. ncOrtho identifies miRNA orthologs with an overall high accuracy, and the mature human miRNAs align with <3 mismatches to 99% of all identified orthologs (Supplementary Figure S6A). Additionally, there was no difference in the predicted secondary structure between miRNA orthologs identified by ncOrtho and those deposited in MirGeneDB (Supplementary Figure S6B). This indicates that ncOrtho fulfills the criteria for miRNA ortholog annotation defined by the miRNA community (34). Still, on first sight a specificity of 0.90 in the vertebrate set of target species suggests that the false positive rate can be improved. Interestingly, the fraction of orthologs predicted solely by ncOrtho is the highest in rhesus macaque (Figure 2A). Humans and rhesus have a genome-wide average pairwise sequence identity of 93% (55), which should render the identification of miRNA orthologs straightforward. We therefore used this species as an example to show that our findings can be at least partly explained by missing data in MirGeneDB.

A parsimonious interpretation of the phylogenetic profiles of human miRNA orthologs as provided by MirGeneDB v2.0 suggested two independent losses of miRNA families in rhesus and 55 gains on the human lineage (Figure 2B). Complementing the profiles with orthologs exclusively predicted by ncOrtho reduced the number of human-specific families that are confined to humans to 0, and many of the ncOrtho-only miRNA orthologs share the seed sequence with the human miRNA (Figure 2B). Moreover, the updated version 2.1 of MirGeneDB now includes 16 miRNA families from 8 species that were predicted by us but were not represented in MirGeneDB v2.0, for example Mir-6715 or Mir-1912 (Supplementary Figure S7). Thus, many ncOrtho-exclusive orthologs, which were initially considered as false positive assignments in our benchmark, represent genuine orthologs that are not represented in MirGeneDB v2.0. The specificity of ncOrtho is therefore substantially higher than 0.91. However, there are also cases where ncOrtho extends the phylogenetic profile with entries that are not covered by MirGeneDB. For example, Mir-1287 is annotated as human-specific even in MirGeneDB v2.1, but ncOrtho identified orthologs with conserved seed sequences in species as distant as the armadillo (*D. novemcinctus;* Figure 2C). Consulting the 100 vertebrate whole genome alignment (35) revealed that the orthologs identified by ncOrtho reside in a genomic region that is conserved across the vertebrates.

Comparison of ncOrtho with other tools. To the best of our knowledge, only three tools have been developed that allow the identification of miRNA homologs in genome assemblies. Of these, we did not consider miRNAminer (56) because it does not support batch requests for multiple miRNAs and/or species. A second tool, miROrtho (28), is no longer publicly available. MapMi (29) detects putative miRNA homologs in a fixed set of 59 species, of which 18 are also represented in MirGeneDB 2.0. We further compared the performance of ncOrtho to the identification of miRNA orthologs using the 100 Vertebrate whole genome alignment (see Supplementary Figure S8 for further details) and using a naïve BLASTn search. Applying the same benchmark criteria as described above (see Table 1), revealed that ncOrtho outperforms the alternative methods by far (Table 2). Only BLASTn was superior in terms of sensitivity, but this comes at the cost of a specificity of only 0.37.

Runtime analysis. The run time of the ncOrtho search scales linear with both the number of reference miRNAs and the number of target species (Supplementary Figure S9). The ortholog search of 556 human miRNAs took 5 h and 23 min with 4 CPUs in rhesus (*Macaca mulatta*). ncOrtho evaluated 8490 genomic regions resulting in 1222 ortholog candidates that entered the final validation. The same search using mouse as target species finished after 53 min (Supplementary Figure S2). Here, numbers drop to only 2481 genomic regions and 638 ortholog candidates, which explains the shorter run time. Note that much of the surplus of candidates in rhesus results from few miRNAs that are in repeat rich regions. For example, the ortholog search for Mir-1271 alone requires the evaluation of 2498 genomic regions in rhesus. The median search time for orthologs to 556 human miRNAs across all 402 vertebrate species was 23 min when using 4 CPUs (59 min CPU time).

Phylogenetic profile of human miRNA orthologs across 402 vertebrates

Investigating the evolutionary trajectory of miRNAs has so far been hindered by a limited taxon sampling in publicly available data repositories of miRNA orthologs. In many cases, systematic groups up to the order level are represented by only one or at most a few representatives. When analyzing missing miRNAs, this makes it hard to differentiate between noise, e.g. due to incomplete data, and a true miRNA loss. To obtain a better resolved picture of miRNA evolution, we extended the taxon sampling for the miRNA ortholog search to represent 402 vertebrate genomes. The resulting phylogenetic profiles summarized on the miRNA family level are shown in Figure 3 and the gene-level resolution is shown in Supplementary Figure S10. The genomic locations and sequences of all ncOrtho orthology assignments are provided in Supplementary Table S2.

Table 1. Performance of ncOrtho. Median runtime per species

Taxon set	Setting	Accuracy	*F*1-score	Specificity	Sensitivity	Runtime (median)
Vertebrates	Quick	0.93	0.93	0.90	0.97	27 min
	Exhaustive	0.93	0.93	0.90	0.97	35 h 57 min
Invertebrates	Quick	0.93	0.24	0.99	0.14	7 min
	Exhaustive	0.93	0.26	0.99	0.15	6 h 13 min

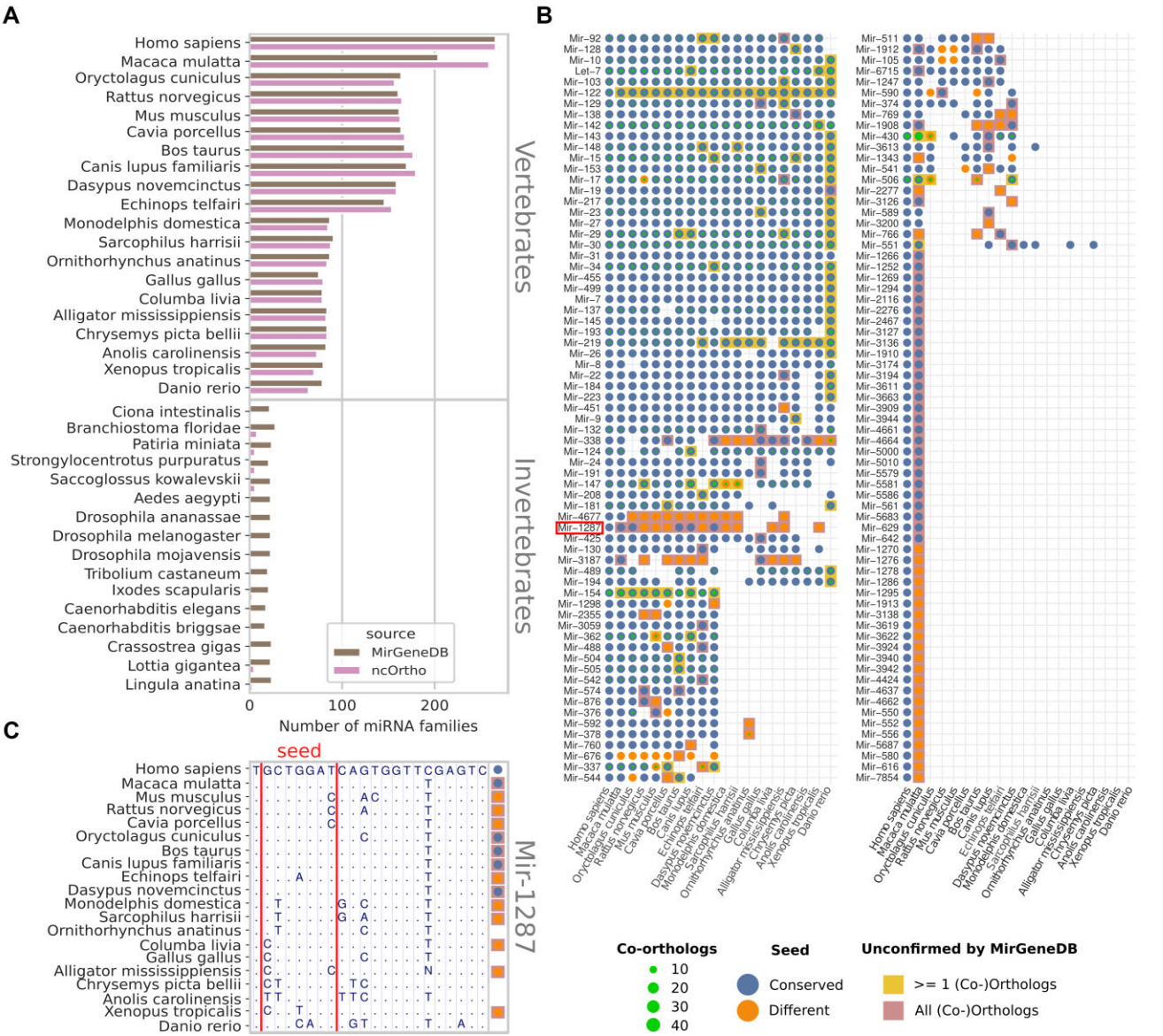


Figure 2. Orthology-based phylogenetic profiles of human miRNAs. (**A**) Human miRNA families represented in 35 species according to MirGeneDB 2.0 and ncOrtho, respectively. (**B**) Phylogenetic profiles for 138 miRNA families with one or more orthologs unconfirmed by MirGeneDB across 20 vertebrate species. The presence of a miRNA ortholog in a species is indicated by a dot. Dot color: blue – seed sequence is identical to the human miRNA; orange – seed sequence differs. The green inlay indicates the presence of co-orthologs, and the inlay diameter is proportional to the co-ortholog numbers summed over all family members. Cell color indicates the fraction of (co-)orthologs that are supported by MirGeneDB: white – all; yellow – some; red – none. A 20-species alignment of the genomic region harboring the mature Mir-1287 (highlighted in red) is shown in (**C**). Only nucleotides that differ from the human reference are specified. The seed region of the miRNA is marked with red lines, and a dot next to a sequence indicates that an ortholog to Mir-1287 was detected in this species where the color coding follows (B). Note, the ortholog assignment uses the full pre-miRNA sequence and not only the section shown in the alignment.

Table 2. Performance of ncOrtho and of three alternative approaches in the identification of human miRNA orthologs across 18 vertebrate species. The highest score is highlighted in bold format

Tool	Accuracy	*F*1-score	Specificity	Sensitivity
ncOrtho	**0.93**	**0.93**	**0.91**	0.95
MapMi	0.55	0.50	0.64	0.46
WGA	0.80	0.78	0.86	0.71
BLASTn	0.66	0.73	0.37	**0.97**

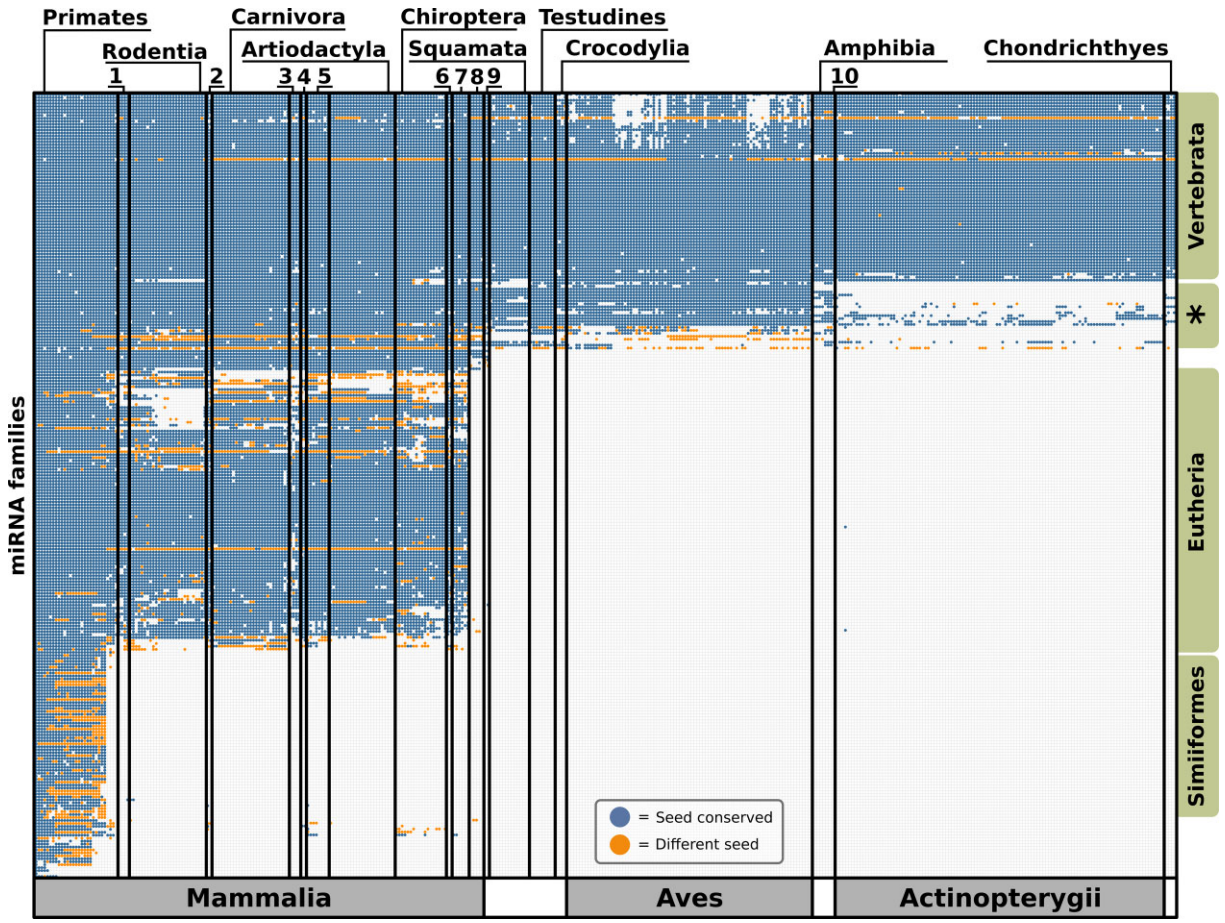


Figure 3. Phylogenetic profiles of 266 human miRNA families across 402 vertebrates. The representation of a miRNA family (rows) by at least one ortholog in a species (column) is indicated by a dot. Species are ordered according to increasing taxonomic distance to humans. Green labels indicate the taxonomic distribution of the corresponding miRNA families. The asterisk indicates human miRNAs whose orthologs are either confined to the *Sarcopterygii* or occur only sporadically in more distantly related taxa. 1 = Lagomorpha, 2 = Scandentia/Dermoptera, 3 = Eulipotyphla, 4 = Pholidota, 5 = Perissodactyla, 6 = Xenarthra, 7 = Afrotheria, 8 = Metatheria, 9 = Monotremata, 10 = *Latimeria chalumnae*.

the evolutionary age of human miRNA families. The profiles reveal that the 266 human miRNA families can be broadly distinguished into four phylostratigraphic layers (53; see Supplementary Table S3). 53 families are confined to the higher primates (Simiiformes) and 86 families are represented only in the Eutheria. The emergence of 71 families predates the diversification of vertebrates. Eighteen miRNA families are preferentially found in the sarcopterygians (tetrapods and coelacanth). However, some of these families are also sporadically present in representatives of the earlier branching lineages, which makes their evolutionary age harder to assess. We next overlaid the phylogenetic profiles of the miRNA families with information about the conservation of the human seed sequence (see Figure 3). In the evolutionarily older miRNAs, the seed sequences are overall highly conserved. This picture changes substantially for the miRNA families that emerged in the last common ancestor of the higher primates. Here, seed sequence changes are commonly observed, and in many cases a single substitution suffices to explain the differences between the individual primate lineages (Supplementary Figure S11).

Loss of miRNA families. Our results show that the loss of miRNA families is rare. The entire data matrix in Figure 3 has only 12.1% missing data, i.e. a miRNA family was not detected in a species that diversified after the evolutionary emergence of the family (Supplementary Table S3; see Supplementary Figure S10 for a gene level resolution). In most cases, these are sporadic absences of miRNA families

in individual species. Without further manual curation, this is best interpreted as noise. However, a few notable exceptions exist where the absence of a miRNA family is consistently observed in several species within a systematic group, which results in the characteristic 'windows' in the phylogenetic profiles (see Figure 3). We can differentiate two main scenarios. The joint loss of several miRNA families in the rodents is one prominent example for a concerted loss of several miRNA families in the last common ancestor of a monophyletic group of species. The situation is different for the missing miRNA families in birds. Integrating the presence/absence pattern of miRNA families with the evolutionary relationships of the birds indicates multiple and independent losses of the same miRNA family on individual evolutionary lineages (Supplementary Figure S12).

Whole genome duplications extend the miRNA repertoire. Lineage-specific duplications of a gene results in two copies that are both co-orthologous to the corresponding gene in a species that branched of prior to the duplication event (57). The number of co-orthologs is therefore correlated with the number of successive lineage-specific duplications. ncOrtho detected miRNAs represented by two or more co-orthologs in almost all taxa (Figure 4; see Supplementary Figure S10 for the full taxon set). However, the fraction of miRNAs with co-orthologs varies, in parts, substantially between species. Within the tetrapods and coelacanth (Sarcopterygii), co-orthologs are overall rare with one exception: in the African clawed frog (*Xenopus laevis*) about three quarters of the represented human miRNA genes have two co-orthologs. The ray-finned fish (Actinopterygii) are substantially more diverse. Co-orthologs are rare in the early branching lineages, except for the Acipenseriformes represented by the sterlet (*Acipenser ruthenus*) and the paddle fish (*Polyodon spathula*). Throughout the teleosts, the fraction of human miRNAs represented by two or more co-orthologs are up to 15 times higher compared to the tetrapods. On most lineages, human miRNA genes are represented by two co-orthologs, but three or four co-orthologs are common (>10% for either class) in both the Salmoniformes and the Cypriniformes. Reconciling the lineage-specific increase of co-ortholog numbers with existing evidence for polyploidization or whole genome duplications (WGD) results in a perfect match (see Figure 4). Two primordial diploid frog species likely hybridized giving rise to the allotetraploid *X. laevis* (58). Polyploidizations were reported for the Acipenseriformes (59,60), a teleost-specific whole genome duplication was proposed (3R-Hypothesis; 61-62), and one additional round of WGD (4R) occurred in the Salmoniformes and in the Cypriniformes, respectively (63–65).

Phylogenomics with miRNA genes

The targeted search for miRNA orthologs across vertebrate diversity has reconstructed the evolutionary trajectory of human miRNAs at an unprecedented scale and resolution. Individual studies based on limited data have explored the use of miRNAs for the reverse approach, where patterns of sequence change between miRNA orthologs should inform about the evolutionary histories of the species they reside in (e.g. 22,61,66). We next investigated the potential of miRNAs for resolving evolutionary relationships in greater detail. Phylogenetic markers should meet two criteria: they should be rarely lost to warrant taxon-gene matrices with little missing data. Moreover, their within-species diversity should be low over evolutionary time scales to reduce the risk that incomplete lineage sorting (ILS) of ancestral alleles blurs the phylogenetic signal generated by speciation events (67). Since miRNAs fulfill the first requirement (see Figure 3) we next investigated their diversity.

Genetic diversity of miRNA genes. We determined the frequency of SNPs in human miRNA genes separately for the mature, star and loop regions and compared this to the SNP frequency in flanking regions of the miRNA, in protein-coding sequences, and in long non-coding RNAs. This revealed a significantly lower frequency in the miRNA regions compared to the other regions in the human genome (*t*-test, *P*-value $< 10^{-8}$; Figure 5A), with the mature miRNA having the lowest diversity among all. Moreover, the represented variants have a lower minor allele frequency (Figure 5B). Both observations are in line with constraints on the evolutionary change of miRNAs that are imposed by miRNA secondary structure formation and by their function. The sequence of the mature miRNA must remain conserved since it mediates the pairing to the target mRNA (8). This constraint extends to the star sequence as it must form a stem–loop during pre-miRNA hairpin formation although individual mismatches are admissible (34). A reduced diversity compared to other genes is therefore a feature that most likely applies to miRNAs in general, irrespective of the species they reside in.

Vertebrate tree of life reconstructed with pre-miRNA sequences. We have shown that miRNAs are rarely lost, and that their genetic diversity is lowest among all investigated genomic regions. The orthology-assignments of ncOrtho for 556 human miRNAs across 402 vertebrate species comprises therefore an unprecedented data basis for miRNA-based phylogenomics. Figure 6 shows the maximum likelihood tree that is based on a supermatrix compiled from this data. *Petromyzon marinus* was not considered in this analysis because <20% of human miRNAs were represented by an ortholog in this species. Additionally, the tarsier (*Carlito syrichta*) was excluded because this species could not be stably placed in the tree (see Supplementary Figure S14). The resulting tree topology reproduces the accepted branching patterns of all major vertebrate groups (Figure 6A). This underlines that the phylogenetic signal in concatenated miRNA genes is sufficient to resolve deep splits in the vertebrate phylogeny accurately and unambiguously.

The mammalian subtree is shown in Figures 6B. Deep splits in this tree are well resolved, and monophyletic Xenarthra and Afrotheria are placed as the earlies branching lineage within the Eutheria ($ML_{BS} = 100$). While this resembles the Atlantogenata hypothesis (see also 68), one of the two competing hypotheses (Xenarthra branched off first) cannot be rejected with this data (p-AU = 0.06; 69). Within the Eutheria, the tree is well resolved on the order level, and here in particular for clades in which ILS is known or at least suspected to interfere with the species tree re-

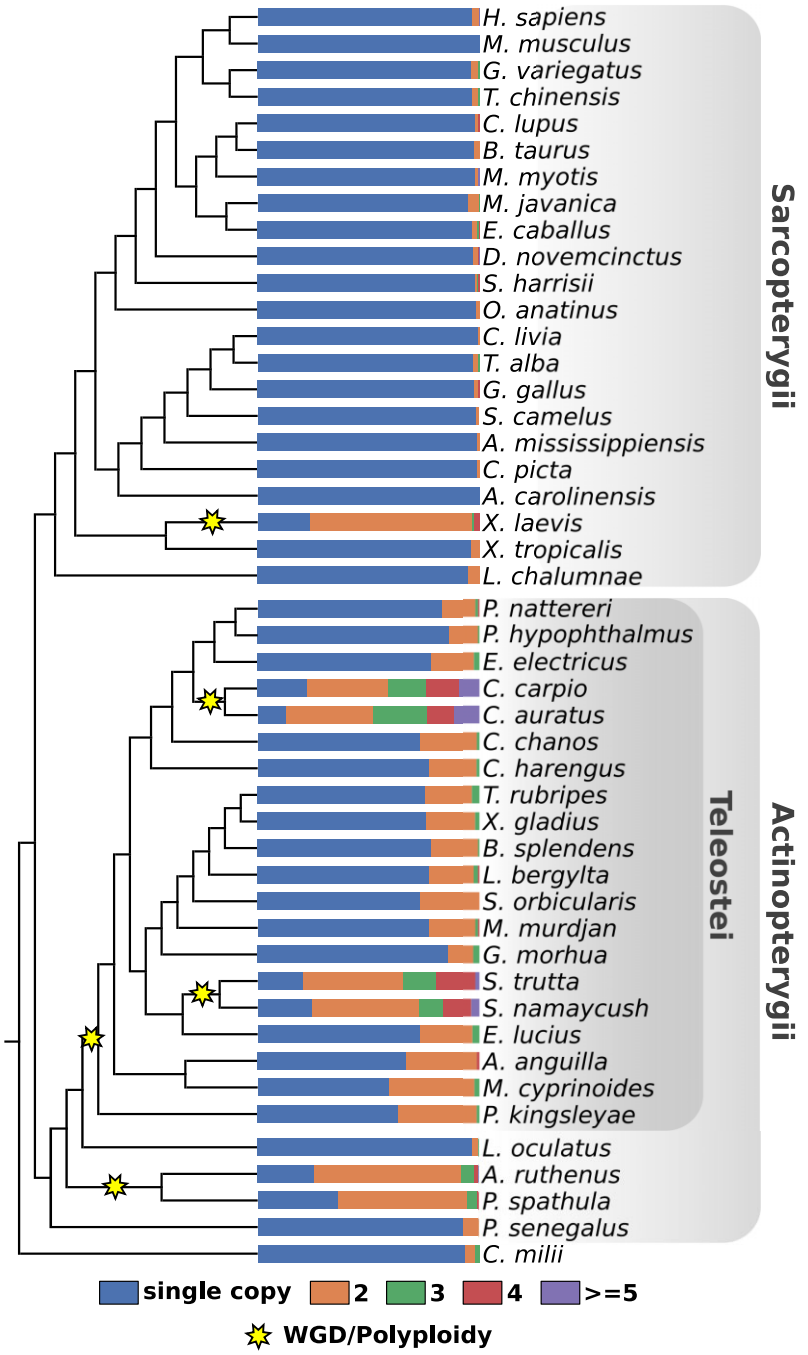


Figure 4. miRNA co-ortholog counts flag whole genome duplications. For each species, the fraction of human miRNAs that are represented by a single ortholog, or by two or more co-orthologs is shown. Yellow star indicates whole genome duplication events (WGD) or polyploidization (see main text). Cypriniformes are represented by *C. carpio* and *C. auratus*, and Salmoniformes are represented by *S. trutta*, and *S. namaycush*. Results for the full taxon set are shown in Supplementary Figure S13.

construction (68,70–73) (Figure 6B, Supplementary Figure S16). The hamster-like (Hystricomorpha) and mouse-like (Myomorpha) rodents are placed in a monophyletic clade to the exclusion of the squirrel-like (Sciuromorpha) rodents ($ML_{BS} = 94\%$; Supplementary Figure S16B). Within the Caniformia, we find support for the Ursoidea and Musteloidea families as their respective closest relatives to the exclusion of the Pinnipedia (Bootstrap support – 97%, Supplementary Figure S16C). However, for both cases, alternative topologies (see 73,74) could not be rejected (p-AU of 0.69 and 0.57, respectively). The situation is different for more recent splits. Maximal bootstrap support was obtained for monophyletic chimpanzees and bonobos as the closest living relatives of humans. Within the new

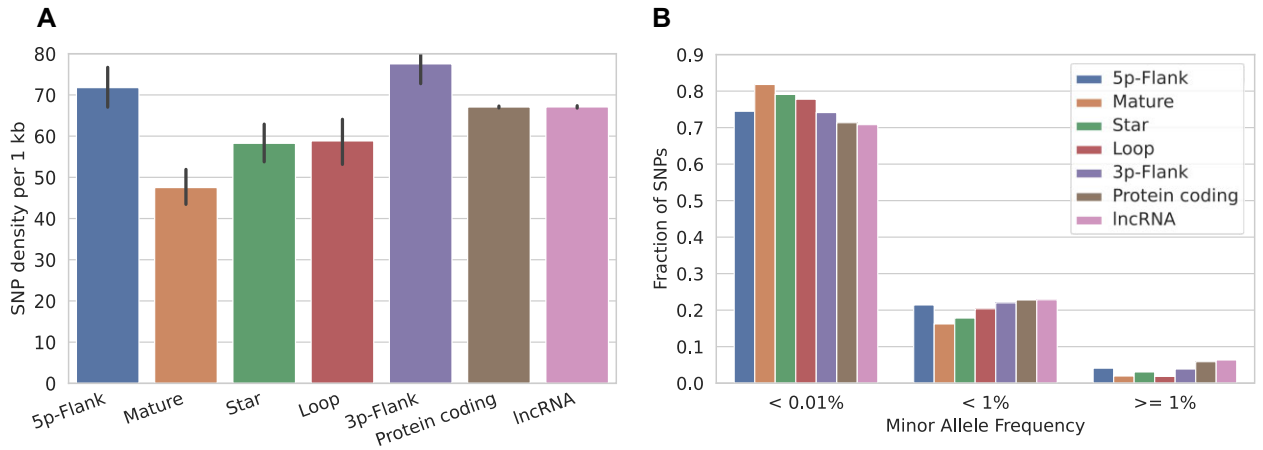


Figure 5. Population-level variation of different regions in the human genome. (**A**) Mean and variance of the SNP density for various regions in the human genome. 'mature', 'star' and 'loop' denote the corresponding parts of a miRNA gene. 5p- and 3p-Flank represents the 30 nt flanking region of miRNA genes on either side. (**B**) Mean fraction of SNPs from (A) with rare (<0.01%), uncommon (<1%), and common (≥1%) minor allele frequencies.

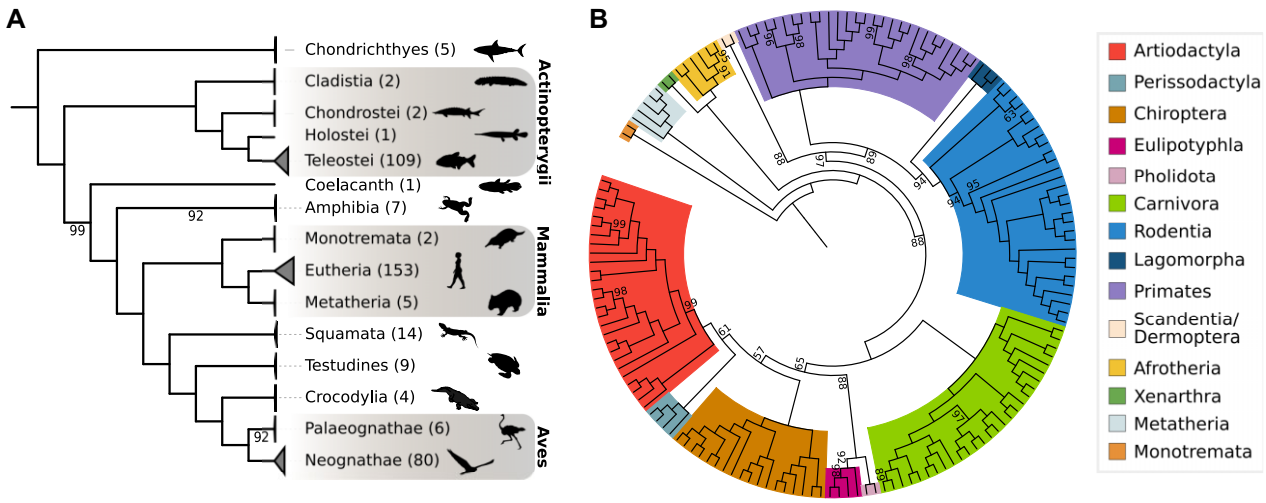


Figure 6. Maximum likelihood phylogeny of the vertebrates using pre-miRNA sequences. The two trees were reconstructed from a miRNA supermatrix comprising 400 species and 12 618 alignment columns. Branch labels denote percent bootstrap support, and only values <100 are provided. (**A**) The backbone phylogeny of the vertebrates rooted with the Chondrichthyes as outgroup. The number of species represented in the collapsed clades are given in parentheses. The fully expanded tree is given as Supplementary Figure 15. (**B**) shows the mammalian subtree.

world monkeys, the data supports monophyletic *Callithrix* and *Aotus* instead of placing *Aotus* in the clade formed by *Saimiri*, *Sapajus*, and *Cebus* as it was suggested by 75. For both clades, the competing hypotheses could be rejected (p-AU of 0.004 and 0.005, respectively; Supplementary Figures S16 and E).

DISCUSSION

The accurate detection of orthologs across large taxon collections is one cornerstone of comparative genomics. It forms the foundation for tracing genes, their functions, and the corresponding functional networks across species and through time (76). ncOrtho allows to extend such analyses routinely to miRNA genes by using covariance models (CMs) in the targeted search for orthologs. Several methods attempted to extend such analyses to miRNA genes (28,29,56). Common to all approaches is the use of an unidirectional sequence similarity-based search for identifying miRNA homolog candidates, followed by a candidate curation using secondary structure- and sequence conservation filters. These tools are either no longer available or suffer from a high false-positive and false-negative rate (see Table 2). ncOrtho is the first tool to identify miRNA orthologs using a reciprocal sequence similarity search. The secondary structure constraints and primary sequence consensus of a miRNA family are captured in the CMs that are used in the ortholog search (52). The performance of these models essentially depends on the quality of the training data, and ncOrtho addresses this issue by exploiting the conservation of gene order to identify positional orthologs (77) that are then used for model training. While this warrants a highly reliable training data, it also implies that, in

principle, whole genome alignments may be directly used for the identification of miRNA orthologs (30). However, this hinges on the condition that the genomic position of a miRNA genes does not change. Additionally, the phylogenetic distances between species under study must not be too large, because otherwise sequences become too diverged for aligning anything but protein coding exons (78). Eventually, the generation of whole genome alignments is computationally highly demanding, which hinders the inclusion of novel species and therefore leaves the data basis of such approaches considerably rigid. More specifically, for large collections of assemblies that are phylogenetically as diverse as the vertebrates, whole-genome alignments are not feasible (32). ncOrtho, in turn is a highly flexible framework for the accurate identification of miRNA orthologs across extensive custom taxon sets, which include species as diverse as humans and sharks who last shared a common ancestor about 600 million years ago (79).

Orthology assignment is an evolutionary reconstruction problem, and as such a ground truth does not exist. Therefore, the benchmarking of orthology assignments for protein-coding genes relies on a standardized framework that allows to assess and compare the performance of individual ortholog search tools (80). Since there is no such framework for miRNA genes yet, we have used the miRNA families that are represented in the manually curated MirGeneDB as a bona fide gold standard. Due to its stringent annotation criteria, this database was excellent for assessing the sensitivity of ncOrtho. The attempt to benchmark the specificity of the orthology assignments indicated, however, that MirGeneDB is likely not fully comprehensive and lacks an unknown but probably non-negligible number of miRNA orthologs. (23). Obviously, this does not rule out that ncOrtho may also identify miRNA orthologs that are no longer functional or makes individual spurious orthology assignments. Such cases may be characterized by multiple independent variations as for example seen in Mir-1287 (Figure 2C). Here, a thorough integration of miRNA orthology assignments based on ncOrtho paired with a subsequent curation via targeted search for these orthologs in RNAseq data would allow to trace miRNAs across species with the highest confidence. This combined approach allows to direct the miRNA search in transcriptomes to predicted, yet missing miRNAs. This would increase the probability to also detect those miRNAs that are only lowly expressed or that are expressed only under certain conditions.

ncOrtho performs a targeted ortholog search resulting in orthology assignments for pairs of species. This has several advantages. The ortholog search can use any reference sequence to start the ortholog search. This makes it independent from pre-compiled catalogs of miRNA genes and allows the tracing of both novel mirRNA genes but also the analysis of miRNAs specific to taxa that are not represented in the public miRNA databases. In the same context, ncOrtho uses a reciprocal hit criterion for the ortholog identification, instead of relying on pre-computed bit-score thresholds (81). This is particularly important, when the training data for computing the bit score thresholds does not cover the full diversity of the miRNAs. The linear scaling in time and CPU usage facilitates an ortholog search across taxon set sizes that are too resource-demanding for more complex search algorithms (82). Lastly, ncOrtho can identify orthologs in target species independent of any a-priori gene annotation. This aspect is particularly important because the annotation of miRNA genes thus far depends on the availability of deep transcriptomic sequencing of small transcripts (21). Consequently, there are substantially varying levels of miRNA annotation quality between species as a result of a dataset-availability bias (18).

While the annotation status of the target genomes does not impact the performance of ncOrtho, the assembly quality does. The number of false negative orthology assignments of genome-based searches is bound to increase when using low-coverage genome assemblies (83). In line with this hypothesis, several evolutionarily old miRNA families show a patchy presence-absence pattern of bird orthologs which could only be explained by the same miRNA being lost multiple times independently (see Figure 3 and Supplementary Figure S12). Among all analyzed vertebrates, this is a unique observation, which hints towards issues with the assembly qualities for these species. However, with the increasing number of chromosomally complete reference assemblies (84–86), the issue of genome completeness can be expected to play only a subordinate role in the future.

The phylogenetic profiles of 556 human miRNAs representing 266 miRNA families across 402 vertebrate species represent the highest resolving analysis of miRNA evolution to date. The resulting presence/absence patterns are consistent with previous studies that describe a 'burst-wise' acquisition of novel miRNA families in the Eutheria and Amniota (13,87,88). Here, we could pinpoint another surge of miRNA innovation to the higher primates (Simiiformes), which has been previously ascribed to either the primates or the old-world monkeys (17,20). This differentiation is important for reconstructing the genetic basis of primate diversification. But it is also essential for applications that, for example, determine the organismal origin of small RNA-seq samples (89). Next to lineage-specific gains of miRNA families, ncOrtho allows to trace likely changes in the regulatory network of miRNAs due to either miRNA loss or seed change. In the context of protein-coding genes, concerted lineage-specific loss is often interpreted as an indicator for a functional integration of the affected genes (26,90–91). Seed-pairing is the most important determinant for the targeting specificity of miRNAs (8). Notably, we find evidence for pronounced lineage-specific changes in the seed sequence, preferentially for evolutionarily young miRNAs that emerged in the last common ancestor of the higher primates. It is tempting to speculate that these seed changes contribute to the re-wiring of still flexible regulatory networks by changing the spectrum of target genes. We can now identify such events with a resolution that is unprecedented for miRNAs, which lays the foundation for investigating their relevance for re-wiring the regulatory network of miRNAs in the future.

Next to the functional implications in the context of gene regulation, changes in miRNAs allow to trace also evolutionary events on a genome-wide level. miRNAs have recently been used to investigate whole genome duplication (WGD) events in transcriptomic data from comparatively small taxon sets (92,93). Here, we have shown that co-orthologs detected by ncOrtho faithfully trace whole

genome duplications across diverse collections of genome assemblies. This makes it possible to rapidly scan the plethora of vertebrate genomes that will emerge from the ongoing sequencing initiatives for indications of whole genome duplications (84,85).

miRNA sequences have been previously proposed as good phylogenetic markers (19,20), and miRNAs have been used in individual and small-scale phylogenomic studies to shed light on the evolution of individual vertebrate lineages (e.g. 68,94). Here, we could show that miRNAs fulfill two main requirements for phylogenetic markers: They are very rarely lost, and they display among all investigated regions in the human genome the lowest genetic diversity. The latter finding complements previous results that attested miRNAs a low diversity in humans, mice and pigs, but did not put these numbers in relation to other regions in the human genome (95–98). With the help of ncOrtho, the compilation of miRNA-based phylogenomic datasets is now straightforward, which allows to establish miRNAs as an alternative marker to protein-coding genes. We could show that the phylogenetic signal in miRNAs suffices to resolve even deep splits in the vertebrate tree unambiguously and with high confidence. For two recently diverged clades whose resolution was hindered by incomplete lineage sorting, we found unequivocal support for one topology. This lends support to the view that miRNAs are indeed suitable for overcoming the effect of ILS. Despite these encouraging results, trees computed from miRNA alignments also need to be treated with a grain of salt. In line with a previous report (68), our miRNA-based phylogeny of the mammals agrees with the Atlantogenatha hypothesis. However, a topology test reveals that the alternative Xenarthra hypothesis does not explain the data significantly worse. This can indicate either a limited phylogenetic signal in the data, or alternatively that hybridization blurred the phylogenetic signal at the onset of eutherian diversification (99–101). In the light of our results, incomplete lineage sorting becomes a less likely explanation.

In summary, miRNAs are essential regulators of gene expression and their tracing across a comprehensive taxon collection is bound to shed light on the emergence and evolution of the underlying regulatory networks. At the same time, miRNAs are highly informative with respect to the evolution of the species they reside in. With the help of ncOrtho the potential of miRNAs for addressing both questions can now be tapped on a comprehensive scale independent of pre-compiled databases, and thus throughout the eukaryotic tree of life.

DATA AVAILABILITY

All data and software used in this study is open source. The ncOrtho algorithm is available on GitHub (https://github.com/BIONF/ncortho) or FigShare (https://doi.org/10.6084/m9.figshare.21679568.v1). The human miRNA orthologs are available in the Supplementary Data where we also supply a list of all genome assemblies which were used for the ortholog search.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank all researchers for making annotated genome sequences available to the public domain, and Mirko Brüggemann and Andreas Blaumeiser for initial work on the ncOrtho code.

FUNDING

Alfons und Gertrud Kassel-Stiftung; Research Funding Program Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE) of the State of Hessen, Research Center for Translational Biodiversity Genomics (TBG) (to I.E.). Funding for open access charge: GRADE Center IQbio.

Conflict of interest statement. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1. Cech,T.R. and Steitz,J.A. (2014) The noncoding RNA revolution - trashing old rules to forge new ones. *Cell*, **157**, 77–94.
2. Pauli,A., Rinn,J.L. and Schier,A.F. (2011) Non-coding RNAs as regulators of embryogenesis. *Nat. Rev. Genet.*, **12**, 136–149.
3. Peng,Y. and Croce,C.M. (2016) The role of MicroRNAs in human cancer. *Sign. Transduct. Tar. Ther.*, **1**, 15004.
4. Bartel,D.P. (2018) Metazoan MicroRNAs. *Cell*, **173**, 20–51.
5. Leisegang,M.S., Martin,R., Ramírez,A.S. and Bohnsack,M.T. (2012) Exportin T and Exportin 5: tRNA and miRNA biogenesis – and beyond. *Biol. Chem.*, **393**, 599–604.
6. Kawamata,T. and Tomari,Y. (2010) Making RISC. *Trends Biochem. Sci.*, **35**, 368–376.
7. Gebert,L.F.R. and MacRae,I.J. (2019) Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.*, **20**, 21–37.
8. Agarwal,V., Bell,G.W., Nam,J.-W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
9. Fromm,B., Domanska,D., Høye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2020) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, **48**, D132–D141.
10. Kumar,S., Suleski,M., Craig,J.M., Kasprowicz,A.E., Sanderford,M., Li,M., Stecher,G. and Hedges,S.B. (2022) TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.*, **39**, msac174.
11. Meunier,J., Lemoine,F., Soumillon,M., Liechti,A., Weier,M., Guschanski,K., Hu,H., Khaitovich,P. and Kaessmann,H. (2013) Birth and expression evolution of mammalian microRNA genes. *Genome Res.*, **23**, 34–45.
12. Deline,B., Greenwood,J.M., Clark,J.W., Puttick,M.N., Peterson,K.J. and Donoghue,P.C.J. (2018) Evolution of metazoan morphological disparity. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E8909–E8918.
13. Sempere,L.F., Cole,C.N., Mcpeek,M.A. and Peterson,K.J. (2006) The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J. Exp. Zool Part B. Mol. Dev. Evol.*, **306B**, 575–588.
14. 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. *Sci. Adv.*, **8**, eadd9938.
15. Marco,A., Ninova,M., Ronshaugen,M. and Griffiths-Jones,S. (2013) Clusters of microRNAs emerge by new hairpins in existing transcripts. *Nucleic Acids Res.*, **41**, 7745–7752.
16. Campo-Paysaa,F., Sémon,M., Cameron,R.A., Peterson,K.J. and Schubert,M. (2011) microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evol. Dev.*, **13**, 15–27.
17. Hertel,J. and Stadler,P.F. (2015) The expansion of animal MicroRNA families revisited. *Life*, **5**, 905–920.

18. 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–E3668.
19. 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.
20. 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.
21. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
22. 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.
23. 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.* (2022) MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res.*, **50**, D204–D210.
24. Czypionka,T., Krugman,T., Altmüller,J., Blaustein,L., Steinfartz,S., Templeton,A.R. and Nolte,A.W. (2015) Ecological transcriptomics – a non-lethal sampling approach for endangered fire salamanders. *Methods Ecol. Evol.*, **6**, 1417–1425.
25. Peralta,D.M., Ibañez,E.A., Lucero,S., Cappozzo,H.L. and Túnez,J.I. (2020) A new minimally invasive and inexpensive sampling method for genetic studies in pinnipeds. *Mammal Res.*, **65**, 11–18.
26. Linard,B., Ebersberger,I., McGlynn,S.E., Glover,N., Mochizuki,T., Patricio,M., Lecompte,O., Nevers,Y., Thomas,P.D., Gabaldón,T. *et al.* (2021) Ten years of collaborative progress in the quest for orthologs. *Mol. Biol. Evol.*, **38**, 3033–3045.
27. Birikmen,M., Bohnsack,K.E., Tran,V., Somayaji,S., Bohnsack,M.T. and Ebersberger,I. (2021) Tracing eukaryotic ribosome biogenesis factors into the archaeal domain sheds light on the evolution of functional complexity. *Front. Microbiol.*, **12**, 739000.
28. Gerlach,D., Kriventseva,E.V., Rahman,N., Vejnar,C.E. and Zdobnov,E.M. (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.*, **37**, D111–D117.
29. Guerra-Assunção,J.A. and Enright,A.J. (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinf*, **11**, 133.
30. 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.
31. Armstrong,J., Hickey,G., Diekhans,M., Fiddes,I.T., Novak,A.M., Deran,A., Fang,Q., Xie,D., Feng,S., Stiller,J. *et al.* (2020) Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**, 246–251.
32. Rhie,A., McCarthy,S.A., Fedrigo,O., Damas,J., Formenti,G., Koren,S., Uliano-Silva,M., Chow,W., Fungtammasan,A., Kim,J. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.
33. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
34. Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S., Marshall,M. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
35. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
36. Altenhoff,A.M., Levy,J., Zarowiecki,M., Tomiczek,B., Warwick Vesztrocy,A., Dalquen,D.A., Müller,S., Telford,M.J., Glover,N.M., Dylus,D. *et al.* (2019) OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.*, **29**, 1152–1163.
37. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52–e52.
38. 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.
39. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
40. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
41. Guerra-Assunção,J.A. and Enright,A.J. (2012) Large-scale analysis of microRNA evolution. *BMC Genomics [Electronic Resource]*, **13**, 218.
42. Madeira,F., Park,Y.M., Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
43. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
44. Wang,Q., Pierce-Hoffman,E., Cummings,B.B., Alföldi,J., Francioli,L.C., Gauthier,L.D., Hill,A.J., O'Donnell-Luria,A.H., Armean,I.M., Banks,E. *et al.* (2020) Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.*, **11**, 2539.
45. Tran,N.-V., Greshake Tzovaras,B. and Ebersberger,I. (2018) PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. *Bioinformatics*, **34**, 3041–3043.
46. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
47. Kalyaanamoorthy,S., Minh,B.Q., Wong,T.K.F., Haeseler,A. and Jermiin,L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, **14**, 587–589.
48. Hoang,D.T., Chernomor,O., von Haeseler,A., Minh,B.Q. and Vinh,L.S. (2018) UFBoot2: improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.*, **35**, 518–522.
49. Nguyen,L.-T., Schmidt,H.A., von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
50. Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
51. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.
52. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
53. 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.
54. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
55. Gibbs,R.A., Rogers,J., Katze,M.G., Bumgarner,R., Weinstock,G.M., Mardis,E.R., Remington,K.A., Strausberg,R.L., Venter,J.C., Wilson,R.K. *et al.* (2007) Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science*, **316**, 222–234.
56. Artzi,S., Kiezun,A. and Shomron,N. (2008) miRNAminer: a tool for homologous microRNA gene search. *BMC Bioinf*, **9**, 39.
57. Remm,M., Storm,C.E.V. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons11Edited by F. Cohen. *J Mol. Biol.*, **314**, 1041–1052.
58. 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.
59. 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.

60. Fofanov,M.V., Prokopov,D.Y., Kuhl,H., Schartl,M. and Trifonov,V.A. (2020) Evolution of MicroRNA Biogenesis Genes in the Sterlet (Acipenser ruthenus) and Other Polyploid Vertebrates. *Int. J.Mol. Sci.*, **21**, 9562.
61. Glasauer,S.M.K. and Neuhauss,S.C.F. (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics*, **289**, 1045–1060.
62. Meyer,A. and Van de Peer,Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, **27**, 937–945.
63. Chen,Z., Omori,Y., Koren,S., Shirokiya,T., Kuroda,T., Miyamoto,A., Wada,H., Fujiyama,A., Toyoda,A., Zhang,S. *et al.* (2019) De novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication. *Sci. Adv.*, **5**, eaav0547–eaav0547.
64. Lien,S., Koop,B.F., Sandve,S.R., Miller,J.R., Kent,M.P., Nome,T., Hvidsten,T.R., Leong,J.S., Minkley,D.R., Zimin,A. *et al.* (2016) The Atlantic salmon genome provides insights into rediploidization. *Nature*, **533**, 200–205.
65. Xu,P., Xu,J., Liu,G., Chen,L., Zhou,Z., Peng,W., Jiang,Y., Zhao,Z., Jia,Z., Sun,Y. *et al.* (2019) The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nat. Commun.*, **10**, 4625.
66. Hoegg,S., Brinkmann,H., Taylor,J.S. and Meyer,A. (2004) Phylogenetic Timing of the Fish-Specific Genome Duplication Correlates with the Diversification of Teleost Fish. *J. Mol. Evol.*, **59**, 190–203.
67. Schrempf,D., Minh,B.Q., von Haeseler,A. and Kosiol,C. (2019) Polymorphism-aware species trees with advanced mutation models, bootstrap, and rate heterogeneity. *Mol. Biol. Evol.*, **36**, 1294–1301.
68. Tarver,J.E., Dos Reis,M., Mirarab,S., Moran,R.J., Parker,S., O'Reilly,J.E., King,B.L., O'Connell,M.J., Asher,R.J., Warnow,T. *et al.* (2016) The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol.*, **8**, 330–344.
69. Shimodaira,H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.*, **51**, 492–508.
70. Ebersberger,I., Galgoczy,P., Taudien,S., Taenzer,S., Platzer,M. and von Haeseler,A. (2007) Mapping human genetic ancestry. *Mol. Biol. Evol.*, **24**, 2266–2276.
71. Wang,X., Lim,B.K., Ting,N., Hu,J., Liang,Y., Roos,C. and Yu,L. (2019) Reconstructing the phylogeny of new world monkeys (platyrrhini): evidence from multiple non-coding loci. *Curr. Zool.*, **65**, 579–588.
72. Churakov,G., Sadasivuni,M.K., Rosenbloom,K.R., Huchon,D., Brosius,J. and Schmitz,J. (2010) Rodent evolution: back to the root.*Mol. Biol. Evol.*, **27**, 1315–1326.
73. Doronina,L., Churakov,G., Shi,J., Brosius,J., Baertsch,R., Clawson,H. and Schmitz,J. (2015) Exploring massive incomplete lineage sorting in arctoids (Laurasiatheria, Carnivora). *Mol. Biol. Evol.*, **32**, 3194–3204.
74. Swanson,M.T., Oliveros,C.H. and Esselstyn,J.A. (2019) A phylogenomic rodent tree reveals the repeated evolution of masseter architectures. *Proc. Biol. Sci.*, **286**, 20190672.
75. Vanderpool,D., Minh,B.Q., Lanfear,R., Hughes,D., Murali,S., Harris,R.A., Raveendran,M., Muzny,D.M., Hibbins,M.S., Williamson,R.J. *et al.* (2020) Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biol.*, **18**, e3000954.
76. Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.
77. Dewey,C.N. (2011) Positional orthology: putting genomic evolutionary relationships into context. *Brief. Bioinform.*, **12**, 401–412.
78. Sharma,V., Elghafari,A. and Hiller,M. (2016) Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. *Nucleic Acids Res.*, **44**, e103–e103.
79. Hedges,S.B., Marin,J., Suleski,M., Paymer,M. and Kumar,S. (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.*, **32**, 835–845.
80. Nevers,Y., Jones,T.E.M., Jyothi,D., Yates,B., Ferret,M., Portell-Silva,L., Codo,L., Cosentino,S., Marcet-Houben,M., Vlasova,A. *et al.* (2022) The quest for orthologs orthology benchmark service in 2022. *Nucleic Acids Res.*, **50**, W623–W632.
81. Manni,M., Berkeley,M.R., Seppey,M. and Zdobnov,E.M. (2021) BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.*, **1**, e323.
82. Altenhoff,A.M., Glover,N.M. and Dessimoz,C. (2019) Inferring orthology and paralogy. In Anisimova,M. (ed.) *Evolutionary Genomics: Statistical and Computational Methods*. Springer New York, New York, NY, pp. 149–175.
83. Milinkovitch,M.C., Helaers,R., Depiereux,E., Tzika,A.C. and Gabaldón,T. (2010) 2x genomes–depth does matter. *Genome Biol.*, **11**, R16–R16.
84. Blaxter,M.L. (2022) Sequence locally, think globally: the Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2115642118.
85. 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.
86. 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.
87. Berezikov,E. (2011) Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genet.*, **12**, 846–860.
88. Students of Bioinformatics Computer Labs 2004 and 2005, Hertel,J., Lindemeyer,M., Missal,K., Fried,C., Tanzer,A., Flamm,C., Hofacker,I.L. and Stadler,P.F.(2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics [Electronic Resource]*, **7**, 25.
89. 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.
90. Kensche,P.R., van Noort,V., Dutilh,B.E. and Huynen,M.A. (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J. R. Soc. Interface*, **5**, 151–170.
91. Stupp,D., Sharon,E., Bloch,I., Zitnik,M., Zuk,O. and Tabach,Y. (2021) Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.*, **12**, 6454.
92. Desvignes,T., Sydes,J., Montfort,J., Bobe,J. and Postlethwait,J.H. (2021) Evolution after whole-genome duplication: teleost MicroRNAs. *Mol. Biol. Evol.*, **38**, 3308–3331.
93. Peterson,K.J., Beavan,A., Chabot,P.J., McPeek,M.A., Pisani,D., Fromm,B. and Simakov,O. (2022) MicroRNAs as indicators into the causes and consequences of whole-genome duplication events. *Mol. Biol. Evol.*, **39**, msab344.
94. Kenny,N.J., Sin,Y.W., Hayward,A., Paps,J., Chu,K.H. and Hui,J.H.L. (2015) The phylogenetic utility and functional constraint of microRNA flanking sequences. *Proc. Biol. Sci.*, **282**, 20142983.
95. Omariba,G., Xu,F., Wang,M., Li,K., Zhou,Y. and Xiao,J. (2020) Genome-wide analysis of MicroRNA-related single nucleotide polymorphisms (SNPs) in mouse genome. *Sci. Rep.*, **10**, 5789.
96. Gong,J., Tong,Y., Zhang,H.-M., Wang,K., Hu,T., Shan,G., Sun,J. and Guo,A.-Y. (2012) Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. *Hum. Mutat.*, **33**, 254–263.
97. Pawlina-Tyszko,K., Semik-Gurgul,E., Gurgul,A., Oczkowicz,M., Szmatoła,T. and Bugno-Poniewierska,M. (2021) Application of the targeted sequencing approach reveals the single nucleotide polymorphism (SNP) repertoire in microRNA genes in the pig genome. *Sci. Rep.*, **11**, 9848.
98. Mármol-Sánchez,E., Luigi-Sierra,M.G., Castelló,A., Guan,D., Quintanilla,R., Tonda,R. and Amills,M. (2021) Variability in porcine microRNA genes and its association with mRNA expression and lipid phenotypes. *Genet. Sel. Evol.*, **53**, 43.
99. Mallet,J., Besansky,N. and Hahn,M.W. (2016) How reticulated are species?*Bioessays*, **38**, 140–149.
100. Szöllősi,G.J., Tannier,E., Daubin,V. and Boussau,B. (2015) The inference of gene trees with species trees. *Syst. Biol.*, **64**, e42–e62.
101. Hallström,B.M. and Janke,A. (2010) Mammalian evolution may not be strictly bifurcating. *Mol. Biol. Evol.*, **27**, 2804–2816.