Functional coherence among miRNA targets: a potential metric for assessing biological signal among target prediction methods in non-model species

Authors:

Christopher W. Wheat[1*]
Rachel A. Steward[1]
Yu Okamura[2,3]
Heiko Vogel[2]
Philipp Lehmann[1,4]
Kevin T. Roberts[1]

[1]Department of Zoology, Stockholm University, Stockholm, Sweden
[2]Department of Insect Symbiosis, Max Planck Institute for Chemical Ecology, Jena, Germany
[3]Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo, Japan
[4] Zoological Institute and Museum, Greifswald University, Greifswald, Germany

* Author for Correspondence: Christopher W. Wheat, Department of Zoology, Stockholm University, Stockholm, Sweden. +46 721 958586.
chris.wheat@zoologis.su.se

Abstract

Although miRNA regulation of protein production is a likely target of adaptive evolution, high false-positive rates in the identification of mRNAs targeted by miRNAs in non-model species' complicates interpretation of recent advances. Here we document the challenges and then outline steps for the community to address these challenges.

Keywords

miRNA, target detection, false-positives, functional coherence, gene set enrichment analysis

One major revelation of the genomics era is that gene regulatory networks (GRNs) exhibit extensive functional coherence, as most transcription factors regulate the transcription of functionally related modules of genes, resulting in co-expressed genes generally comprising coherent developmental and metabolic pathways (Stuart et al. 2003; Wolfe et al. 2005). GRNs are at the core of evolutionary biology studies, since it is the modification of GRNs, as well as their co-option into novel developmental contexts, that is the major axis upon with evolutionary adaptations and novelty arise (Bruce & Patel 2020; Erwin 2021). However, mRNA transcription alone does not determine protein concentrations and hence phenotypes, but rather a diverse set of dynamics, including post-transcriptional and post-translational regulation, significantly modify the transcriptome, forming a key feature of the genotype to phenotype map (Liu et al. 2016; Bartel 2018).

Here we focus upon post-transcriptional regulation via microRNAs (miRNAs), ~22 nucleotides (nt) long RNAs. In most animals, miRNAs are produced after transcription via a series of processes (hairpin formation, cleavage, export to cytoplasm, cleavage), then bound by the Argonaute protein, creating a silencing complex that selectively binds mRNA based upon a short (6-8 nt) sequence seed matching between the miRNA and mRNA, primarily in the 3' UTR region of mRNA transcripts, which then initiates various forms of translation repression (Bartel 2018). Via this post-transcriptional action regulating the mRNA to protein production relationship, miRNAs play an important role in developmental progression and physiological functioning (Bartel 2018; Gebert & MacRae 2019). Numerous studies over the past decade, across both invertebrates and vertebrates, have found significant differential expression of miRNA genes associated with adaptive phenotypes, suggesting that these “sculptors” of the transcriptome play an important role in adaptive evolution (Bartel 2018; Leung & Sharp 2010; Fruciano et al. 2021). However, investigating how such differential expression of miRNA causally leads to adaptive phenotypes necessitates identifying the mRNAs that are targeted by miRNAs, as only this allows researchers to make causal connections between differential miRNA expression, protein expression changes, and ultimately differential reproductive success. Unfortunately, identifying which mRNAs are targeted by which miRNAs remains a complex problem (Bracken et al. 2016).

Based upon insights from model-species (e.g. humans, flies, worms), animals are expected to have 100's of miRNA families (miRNAs that target the same canonical motif in mRNA), each of which can effectively reduce the protein production of 100's genes. In humans these numbers correspond to about 500 miRNAs, 300 of which can be placed into about 170 gene families, with each family on average posttranscriptionally repressing roughly 400 genes (Bartel 2018). From the perspective of a given mRNA sequence, nearly half of fly (~ 40%) and human (> 60%) mRNAs contain conserved miRNA binding targets, with each mRNA on average containing multiple miRNA binding sites (of the same and/or different miRNA families). Thus, across diverse taxa, miRNAs have the potential to sculpt a large faction of the transcriptome.

Genomic core facilities now routinely provide short RNA sequencing, enabling quantitative assessments of miRNA abundance in nearly any taxa. However, identifying the biologically meaningful targets of differentially expressed miRNA remains challenging, despite technological advances. While direct sequencing of the mRNA pool bound by the silencing complex is possible (crosslinking-immunoprecipitation-sequencing, CLIP-seq), a high concentration of cells is required, with results necessarily averaging over the diverse miRNA regulation dynamics among cells lineages. While a single cell approach has just been developed (Sekar et al. 2023), neither technique is able to identify the miRNAs directly involved.

As an initial, or only, foray into miRNA research, many research groups rely upon bioinformatic prediction of miRNA targets in their focal species, for initial interpretation of differential miRNA expression. In animals, miRNA binding to mRNA primarily relies upon 6 to 8 nucleotides of complimentary sequence, referred to as seed pairing. While legions of such short motifs populate the UTR regions of transcriptome, only a small fraction are involved in post-transcriptional repression (Agarwal et al. 2015, 2018; Fridrich et al. 2019). This scenario highlights the inherently challenging nature of target prediction due to the exceptional potential for statistically significant false positives (Fridrich et al. 2019), with the challenge of accurate *in silico* prediction spawning yet another bioinformatics cottage industry (~ 100 different software approaches to date (Fridrich et al. 2019; Kern et al. 2020; Ritchie et al. 2009).

Emerging from diverse efforts in model-species to understand miRNA post-transcriptional regulation comes the robust result that signatures of evolutionary conservation, generated due to consistent purifying selection acting over 10 to 100's of millions of years, provides a powerful means of discriminating functionally important seed regions from other candidates in the dynamically evolving UTR regions of mRNA. Indeed, compared to using only identified motifs in a single species, or in combination with various ways of modeling local thermodynamics, only approaches incorporating evolutionary conservation appear accurate (Friedman et al. 2009; Agarwal et al. 2015), though the field continues to explore additional parameters and approaches (Kern et al. 2020). Of direct relevance to this journal's readership, the prediction tools most commonly employed by the ecology and evolution, non-model species community are those using data from only one species without information on evolutionary conservation, which exhibit false-positives rates approaching 50% or fail to identify true-positives in well verified experiments (e.g. miRanda, RNAhybrid; (Agarwal et al. 2015; Pinzón et al. 2017; Fridrich et al. 2019; Krüger & Rehmsmeier 2006)).

These observations thereby suggest that our community faces extensive challenges, not only when hypothesizing about the potential range of functional impacts of differentially expressed miRNAs, but when trying to conduct functional validation studies. Currently, it is common to see studies intersecting miRNA expression patterns with RNAseq results, scanning for inverse relationships. Unfortunately, finding meaningful negative correlations between miRNA and mRNA levels is likely to challenging, as the power of such correlations depends upon the number of time points in comparison and the accuracy of identified miRNA-mRNA interactions. Given that each miRNA can have hundreds of predicted targets, we fear that without a substantially large dataset of such comparison across tissues and timepoints, such efforts will always be beset by high false-positive rates. In sum, the aforementioned issues highlight the need for an external means of assessing the accuracy of miRNA target set prediction, especially one that could be used by the non-model species community.

Here we present rational for an external means of assessing the accuracy of miRNA target set prediction. We take as our starting point that the regulatory network of miRNAs is non-random, as miRNA targets are significantly higher than expected in genes having positive regulatory motifs and being highly-connected GRN components, such as transcription factors (Cui et al. 2006; Bracken et al. 2016). Co-expressed miRNAs, whether co-localized or not, have also been found to target specific genes and pathways (Lee et al. 2012; Xu & Wong 2008; Bracken et al. 2016). Additionally, individual miRNA gene families have been found to exhibit functional coherence in the genes they target (Tsang et al. 2010). Indeed, the functional coherence of mRNA targets is itself central to resolving the paradox between the small post-transcriptional effect of miRNAs upon individual genes and the larger phenotypic effects of miRNAs, as miRNA action upon multiple steps of a pathway is expected to culminate in larger phenotypic impacts (Bracken et al. 2016). However, currently little is known about the extent of such functional coherence across miRNA gene families as a whole. Specially, we can find no global scale analyses of the functional coherence of individual miRNA targets in species other than humans within a disease context (Bracken et al. 2016; Gusev 2008), highlighting the lack of a general understanding of how such coherence varies among taxa. Nevertheless, identifying a signature of functional coherence, beyond informing on the miRNA GRN and how it evolves, could provide a biologically informative metric for assessing *de novo* target predictions in novel taxa.

Our work here began with trying to identify the miRNA targets in a novel species, the Green-veined White butterfly *Pieris napi* (Lepidoptera, Pieridae). Ultimately our goal was to identify the miRNAs involved in the different states of diapause progression, but in order to understand patterns of differentially expressed miRNAs, we needed to identify their potential targets in the transcriptome. We present a comparison of different miRNA prediction approaches, finding that only our approach incorporating evolutionary constraint, results in a detectable functional coherence among the targets per miRNA. In order to validate this finding, we present evidence using miRNA target predictions across model and non-model species that animals generally exhibit extensive functional coherence across miRNA gene families. Therefore, functional coherence provides a biologically informative metric for assessing *de novo* target predictions in novel taxa that could greatly facilitate ability of the ecology and evolutionary genomics community to

Commented [PL1]: Sounds a bit defensive. Perhaps it could be written more neutrally?

“finding that only an approach incorporating evolutionary constraint, …”

make logical connections between miRNA to relevant protein expression changes and their eventual phenotypic impacts.

Methods

Samples, processing, miRNA identification

Data generation, from collection to sequencing through to miRNA gene and seed identification was performed previously (Roberts et al., in review). Although readers are directed to this other work for methodological details (Roberts et al., in review), here they are briefly presented for clarity. A total of 73 samples were taken throughout pupal progression (12 timepoints (0, 3, 6 days direct development; 0,3,6,24,114,144,155 days diapause development), for each of 2 tissues (head, abdomen), each with 3-4 biological replicates). After library construction using Illumina small RNA library kits they were sequenced using HiSeq 2500 50SR, generating an average of 6.9 M reads / library. The miRTrace pipeline was used to check data quality (v1.0.1; (Kang et al. 2018)), contamination and taxonomic bias, followed by filtering and adapter removal (Roberts et al., in review). Using miRDeep2 processing scripts (Friedlander et al. 2011), reads greater than 17bp were mapped against the chromosomal level assembly for *P. napi* genome GCA_905231885.1 (Lohse, Hayward, et al. 2021), with miRNAs detected using *Bombyx mori* and *Heliconius melpomene* as reference miRNA sets.

Target identification

miRNA targets were identified using two separate approaches, the first relying primarily upon evolutionary conservation and the second using data from a single species. Our first approach aligned genomes of 6 species of Pieridae using the software Progressive Cactus (Armstrong et al. 2020), each increasing evolutionary distance from our focal species *P. napi*, which was used as the reference (*P. napi* (GCA_905231885.1; (Lohse, Hayward, et al. 2021), *P. rapae* (GCA_905147795.1; (Lohse, Ebdon, et al. 2021))*, P. brassicae* (GCA_905147105.1; (Lohse, Mackintosh, et al. 2021))*, P. macdunnoughii* (Steward et al. 2021). The last two of these 6 genomes were high-quality draft assemblies, using MaSuRCA (Zimin et al. 2017) for genome assembly of Oxford Nanopore Sequencing data and and Illumina short read data for *P. melete*

(PRJEB59056, 376 contigs, 320 Mbp, N50 2.6 Mbp, BUSCO: CS:94.1%, CD:4.4%, F:0.3%, M:1.2% (BUSCO v. 5.5.0 (Manni et al. 2021)*,* n:5286, lepidoptera_odb10), and using Flye ver. 2.7 (Kolmogorov et al. 2019) for *Pontia daplidice* (PRJEB59056, 142 contigs, 223 Mbp, N50 3.6 Mbp, BUSCO: CS:97.7%, CD:0.5%, F:0.3%, M:1.4%, (n:5286, lepidoptera_odb10). The last common ancestor of these species was approximately 23 million years ago (Chazot et al. 2019). We next sought to identify 3'UTR regions that were expressed in the relevant tissue and developmental stage of our miRNA data.

Obtaining accurate 3'UTR annotations is challenging for several reasons. First, the 3'UTR per locus is highly variable, with > 65% of human and *Drosophila* loci producing alternative polyadenelated mRNAs across tissues and development (Derti et al. 2012; Ye et al. 2023; Sanfilippo et al. 2017). This gains relevance as the available genomic annotation of our focal species did not use RNAseq data from diapause relevant tissues for their annotation. Second, methods for predicting 3'UTR regions from DNA alone, or even with RNAseq data, perform with high variability across species and in general, poorly in non-model species (Ye et al. 2023; Bryce-Smith et al. 2023), and though some have tried to directly address this (Huang & Teeling 2017)**,** obtaining meaningful UTR predictions is challenging in novel species. Thus, in order to efficiently move beyond data and bioinformatic limitations, here we deployed a simplified approach for exploring potential 3'UTR regions for our focal species.

We assessed the 3'UTR annotation for the *P. napi* genome and found that it had overpredicted UTR regions (GCA_905231885.1; (Lohse, Hayward, et al. 2021), such that UTR regions routinely overlapped with flanking genes. In addition, at the time of our analyses, the annotation of GCA_905231885.1 available from the Darwin Tree of Life Program relied on an early annotation pipeline that was not optimized for Lepidoptera. Accordingly, we chose to rely upon a *de novo* genome annotation we previously generated (Steward et al., in review). This *de novo* annotation was produced using the BRAKER2 pipeline (v.2.1.5, (Brůna et al. 2020; Hoff et al. 2016; Ter-Hovhannisyan et al. 2008; Stanke et al. 2006, 2008; Lomsadze et al. 2005; Hoff et al. 2019), run in protein mode using Arthropoda OrthoDB (v.10) reference proteins. This annotation contained 123,638 exons, 16,449 genes and was found to contain 98.4% complete

BUSCOs for Lepidoptera_ODB10. Comparisons between this annotation and two accessed from the Darwin Tree of Life revealed the BRAKER2 annotation to be the most complete (i.e. fewest fragmented BUSCOs, a small proportion of single exon genes, and more total estimated transcripts (see Supplementary methods; Table S1, S2 in Steward et al., in review).

Among moths and flies, the majority of 3'UTR regions are expected to be within 1 kb of the stop codon in the terminal coding exon, based upon detailed studies from several *Drosophila* species (Sanfilippo et al. 2017; Wang et al. 2019) and 3'UTR lengths for the an exemplar moth (*Bombyx mori* mean=923 bp, n=27,556) and butterfly (*Heliconius melpomene* mean=600, n=11,770) downloaded from UTRdatabase (Lo Giudice et al. 2023). While alternative UTRs in animals can involve spliced introns, the frequency in 3'UTR regions are lower than 5'UTR, and usually < 10% (Mignone et al. 2002). Based upon these expectations of 3'UTRs, we generated a bed file of likely 3'UTR regions, extending 1kb beyond every stop codon (and containing 9 codons (27 bp) prior to the terminal codon), of every protein isoform. We then assessed whether any of these candidate 3'UTR regions had a significant match via blastn when searched against the assembled transcriptome of an RNAseq dataset. The assembled transcriptome was generated using Trinity (Haas et al. 2013), default parameters, with RNAseq data comprising all of the same tissues and timepoints of our miRNA samples (Pruisscher et al. 2021). Alignments were filtered to only include candidate 3'UTR regions that had at least 70 bp of 3pUTR (filter settings: DNA identity > 90%, e-value < 0.000001, bitscore > 300, alignment length > 100 bp; NCBI BLAST v. 2.2.28+; (Camacho et al. 2009). Coordinates for these post-filtered 3'UTR regions, which we expect to be expressed 3'UTRs, were then used to identify these regions in the *P. napi* genome, then whole genome alignment of all species, followed by the extraction of each expressed 3'UTR region, which were then used as the input for conserved miRNA target identification via targetscan_70.pl, part of TargetScan v.7 (Agarwal et al. 2018). Manipulation of GFF files used bedtools2 (Quinlan & Hall 2010), which was also used to assign nearest coding gene ID to each candidate 3'UTR region, while alignment filtering used maffilter, with default settings unless indicated (remove_duplicates=yes, reference=Pnapi, min_size=6), min_length=50, dist_max=1200; (Dutheil et al. 2014).

The other input file for targetscan_70.pl was the seed sequences for each of the identified miRNA genes, predicted from mirDeep2 (Roberts et al., in review).

For each identified target region, the resulting output provides information on species depth and seed size, which can be used to filter for differing degrees of evolutionary conservation. Species depth indicates the number of species having the identical target sequence in the alignment, ranging from all of the species down to only 2 species. Targets only found in 2 of the 6 species likely identify a region of lower evolutionary constraint compared to targets identical across all species. Seed size of the identified target can vary in size from an 8-mer down to a 6-mer, indicating the length of base pairs of the identified target. Targets shorter in length are more likely to occur by random chance compared to those of longer length. We use this information to explore the quality of targets in later analyses.

Our second approach for miRNA target prediction used only two files as the input for miRAnda (Enright et al. 2003) and RNAhybrid (Krüger & Rehmsmeier 2006). These were the expressed 3’UTR coordinates for *P. napi* and seed sequences for *P. napi*, both of which were described above. Both programs were run on default settings. Thresholds for targets were set at e-value < 0.1 for miRanda, and p-value < 0.1 for RNAhybrid.

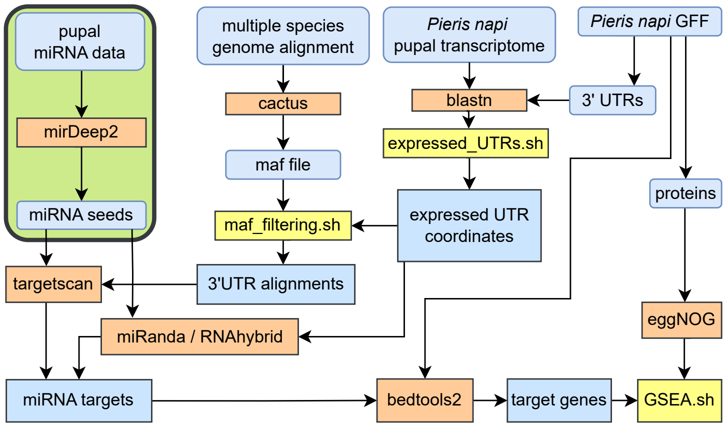

Fig. 1. Flowchart of miRNA target detection in *Pieris napi*, using two methods that lead to gene set enrichment analyses (GSEA). Shown are the data files (blue), various

software programs (orange), and custom bioinformatic scripts (yellow) that were used. Generation of miRNA data through to miRNA seed input file is from previously published work (green enclosed portion of flow chart; Roberts et al., in review). Made using diagrams.net.

Functional coherence via gene set enrichment analysis

Target sets predicted per miRNA family were assessed for their functional coherence via gene set enrichment analysis (GSEA) using the r package topGO v2.46 (Alexa & Rahnenfuhrer 2023), with inputs of GO terms assigned to the coding regions of genes having identified 3'UTR targets. For each GSEA of a miRNA target set, we took the -$\log_{10}$ P-values of the top ten most significant categories, and quantified their distribution as a function of the number of aligned species having identical seed sequences, and for different seed pairing lengths, from 6mer to 8mer.

Comparative assessment of functional coherence

In order to gain a robust assessment of miRNA functional coherence, with miRNA target sets independent of our work and for model species having higher quality target prediction, we repeated our analyses on the miRNA targets from 4 additional diverse animals. Three datasets were downloaded from TargetScan databases (*Homo sapiens*: TargetScanHuman release 8.0, Predicted_Targets_Info.default_predictions.txt (McGeary et al. 2019); *Mus musculus*: TargetScanMouse release 8.0, Predicted_Targets_Info.default_predictions.txt (McGeary et al. 2019); *Drosophila melanogaster*: TargetScanFly release 7.2, Predicted_Targets_Info.default_predictions.txt, (Agarwal et al. 2018)), while predicted cichlid targets for *Oreochromis niloticus* (Mehta et al. 2022), were provided by Dr. T. Mehta upon request. Note that for each TargetScan species dataset, in order to connect miRNA ID to coding gene ID to GO terms of the latter, for the relevant genome assembly, its GFF annotation was downloaded and protein sequences per ID extracted using gffread from cufflinks-2.2.1 (Trapnell et al. 2010), for which GO annotations were generated using functional annotation via orthology assignment, implemented in the online server eggNOG using default settings (Huerta-Cepas et al. 2019), which was then joined to the miRNA table downloaded from the relevant TargetScan database. An estimate of the evolutionary depth over which 3'UTR alignments were made in order to assess evolutionary constrain was estimated from. Ages for each clades of data upon which miRNA targets were based, i.e. the age of the relevant crown groups (the paraphyletic *Drosophila* genus at 53 MYA (Suvorov et al. 2022); the dataset for *H. sapiens* involved using 84 of 100 species of the UCSC multiz alignment (Agarwal et al. 2015), including all species sister to, *Latimeria chalumnae*, as well as this coelacanth, with their crown age estimated at roughly 400 MYA (Amemiya et al. 2013); the dataset for *M. musculus* only included 52 species of the 60-way multiz alignment of UCSC, and has a similar crown age as *H. sapiens*; the dataset for target *O. niloticus* has a crown age estimated at 10 MYA (Mehta et al. 2022).

Results and Discussion

An extensive miRNA sequencing effort has recently identified 257 miRNAs expressed during pupal development of *P. napi* (236 expressed in head tissue, 207 in the abdomen; Roberts et al., in review). Here we use this data to predict mRNA targets of these miRNAs in *P. napi*. We began by identifying which mRNAs, among all candidate 3'UTR regions in the genome of *P. napi*, were expressed in a tissue matched RNAseq transcriptome assembly. We then identified these 3'UTR regions if mRNA in a multispecies, whole-genome alignment (n=6 species of Pieridae, Lepidoptera) that span nearly 23 million years of divergence (Chazot et al. 2019). The resulting 3'UTR alignment, together with the seed sequences from the identified miRNA genes of *P. napi*, were then used as input for TargetScan v.7, which uses evolutionary conservation in 3'UTRs to predict miRNA targets (Agarwal et al. 2018).

Next, we sought an independent means of quantifying whether these predicted target sets per miRNA gene had more biological meaning than random sets, as critiques of target prediction methods suggest that target sets generate from tools such as miRAanda and RNAhybrid may be dominated by false positives (Fridrich et al. 2019; Pinzón et al. 2017; Krüger & Rehmsmeier 2006). We reasoned that since a general feature of gene regulatory networks (GRN) is their extensive functional coherence of regulated genes, as most transcription factors regulate related modules of genes (Stuart et al. 2003; Wolfe et al. 2005), the same is likely true for the targets of miRNA (see Methods for additional discussion). Functional coherence was quantified using gene set enrichment analysis (GSEA) upon the predicted set of gene targets for each miRNA, using the average significance of the top ten most enriched GO categories as the representative metric.

In order to assess whether there was any functional coherence in our predicted targets, we quantified GSEA of the miRNA target sets using variable levels of evolutionary constraint. TargetScan output provides two axes upon which to vary evolutionary constraint in miRNA target prediction. First, we used differing thresholds of constraint upon the species alignment of the 3'UTR, by varying the number of species for which the seed site was required to be identical. Our lowest evolutionary constraint level required only 2 species to have identical sequences in the alignment for the miRNA seed site (the lowest threshold we could set), while our most stringent required all 6 species to have the same identical sequence for the seed site. Second, there are 5 different sizes of target sites for the seed match region of the 3'UTR, ranging from 6 bp (6mer) to 8 bp (8mer) in length. Requiring target sites to be longer in length is a more stringent requirement. In combination, our most relaxed setting was 6mer for only 2 species in the alignment, while our most constrained was 8mer for all species. In order to assess the relative tradeoff across these axes of constraint in the prediction of miRNA targets, we explored our results extensively (fig. 2 *A,B*). As the stringency increases, via increasing the number of species having target seed or increasing the size of the seed match category, the predicted number of targets per miRNA gene decreases, suggesting there is a biological signal in our target prediction method. While these results are highly variable across miRNA genes (fig. 2*C*), we concluded that a good balance between over-prediction and power was using a 7mer seed match size and higher (termed 7mer-inclusive, which includes all targets from 7mer variants and 8mer) that is present and conserved across all of the aligned species.

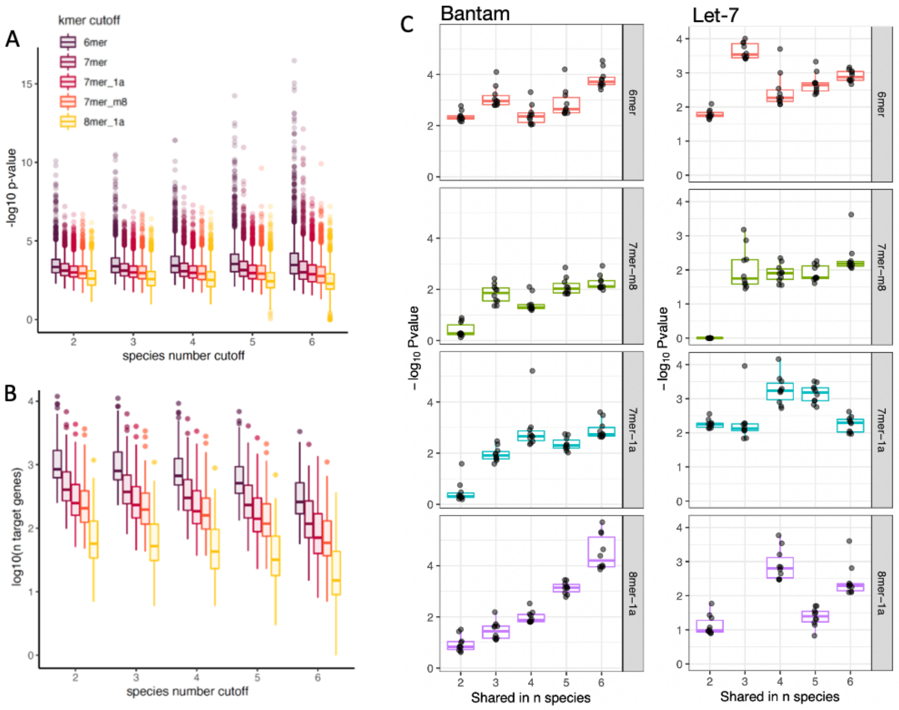

Fig. 2. Assessment of GSEA results across predicted targets per miRNA gene. (*A*) Significance of the top 10 GO terms per target set per miRNA gene (each dot is one term) shown as a boxplot of all results, as a function of the number of species for which seed was identical, for each of 5 different sizes of site type of the seed match (color scale purple to yellow). As the stringency of predicted targets increases from being found only in 2 species to all 6 species, the significance values increase for the smaller seed match sizes (e.g. 6mers increase while 8mers do not). (*B*) Number of targets per miRNA gene (each dot is count for a miRNA gene), across different prediction thresholds of species number and miRNA seed match size (as in A). As the stringency increases, via increasing the number of species having target seed or increasing the size of the seed match category (color scale purple to yellow), the predicted number of targets per miRNA gene decreases (6mer in 2 species is largest set, 8mer in 6 species is the smallest). (*C*) Shown are GSEA results for two miRNA genes (left is Bantam, right

is Let-7), displaying effects of stringency increase on significance of the top 10 GO terms per target set per miRNA gene. These exemplify the range of variation between miRNA genes in their GSEA results, with Bantam exhibiting a strong increase in GSEA P-value as evolutionary constraint is maximized (8mer-1a panel) and Let-7 lacking this trend.

For comparison, we also used single species target prediction methods. Using the 3'UTR regions of *P. napi* and seed sequences of miRNA genes as input, we used the most commonly employed target prediction tool by the ecological and evolutionary genomics community, miRanda (Enright et al. 2003). We additionally employed a second single species tool with the same input data, RNAhybrid (Krüger & Rehmsmeier 2006). In order to compare the predicted targets across these tools, we quantified their relative functional coherence via GSEA using the 7mer-inclusive conservation threshold (described above). As a control, a GSEA was conducted on random sets of gene targets conditional on the set size of the observed miRNA targets, which we used as our background expectation of significance given concerns about GSEA significance thresholds when working with miRNA targets (Bleazard et al. 2015).

The predicted targets of each miRNA from both methods exhibited significant GSEA results, with average P-values for miRanda of 0.0185 and 0.0420 for RNAhybrid (fig. 3*a*). However, GSEA results on sets of randomly drawn genes had P-value distributions that entirely overlapped with the gene set targets predicted by these methods (fig. 3*a*). Thus, GSEA P-value for targets from miRAnda, RNAhybrid, and random draws were lower than nominal P-value significance thresholds (i.e., alpha = 0.05), highlighting two issues. First, these results exemplify previously noted challenges of GSEA when investigating miRNA targets (Bleazard et al. 2015), in that resulting P-values are poorly controlling for diverse many to many relationships, as GSEA were not designed for such relationships. Second, neither miRAnda nor RNAhybrid predicted targets that performed better than random.

In stark contrast to the previous results, miRNA targets predicted using evolutionary conservation via TargetScan exhibited extensive functional coherence (fig.

3*a*), with GSEA P-values much higher than random draws. This result suggests two mutually exclusive explanations. Either *P. napi* has miRNA targets that lack functional coherence, which could explain the miRanda and RNAhybrid results and therefore justify continued use of such tools by the non-model species community, *or* the miRNAs of this butterfly exhibit functional coherence and only biologically meaningful target sets can reveal this pattern. When facing variable results among target prediction methods, studies in the non-model species community commonly intersect results from various target prediction methods, despite this being explicitly discouraged by experts in the miRNA field (Fridrich et al. 2019; Ritchie et al. 2009). To quantify the performance of such an intersection approach, here we assess the overlap of targets from miRAnda and RNAhybrid with respect to target predictions from TargetScan. We find no substantial overlap across these three methods. Further, the level of overlap among methods does not covary with the degree of functional coherence observed in our TargetScan results (fig. 3*b*).

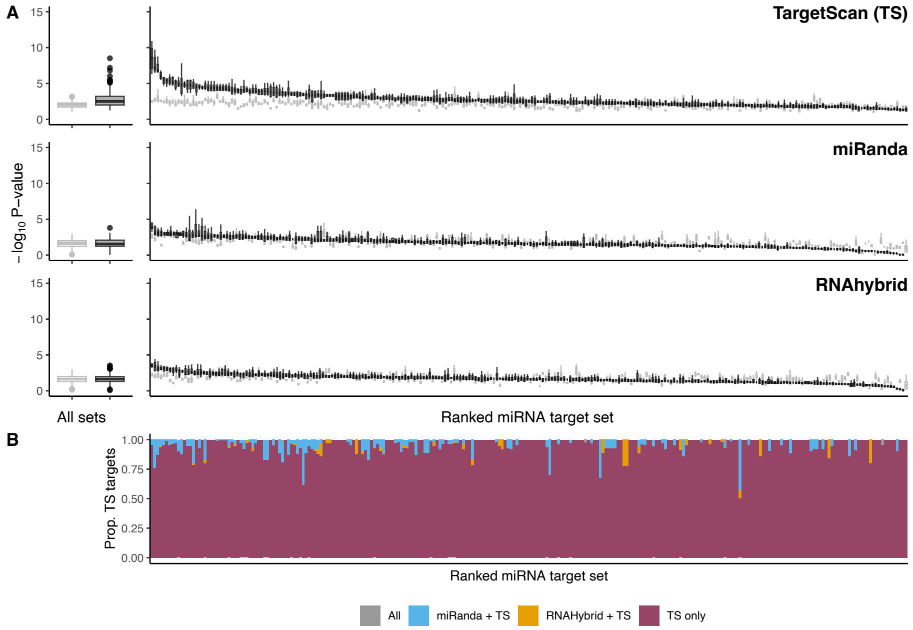

Fig. 3. The functional coherence of miRNA targets across animals measured using gene set enrichment analysis (GSEA). (*A*) Comparison of the functional coherence of miRNA target predictions and their relationships, predicted in the butterfly *Pieris napi*. Gene set enrichment analysis P-values for top 10 GO terms for each miRNA (Y-axis) for targets predicted using Targetscan (top panel), miRanda (middle panel), RNAhybird (lower panel). Left-hand panels summarize median P-values for random (light grey) and predicted (black) miRNA target sets, while right-hand panels show results each miRNA target set. (*B*) Intersection of predicted targets from all three methods in relation to TargetScan results, shown as a proportion. Order of miRNAs along X axis are by mean P-value based upon Targetscan GSEA results.

In order to discriminate between the two aforementioned explanations, we next quantified functional coherence using four published miRNA target sets. Across diverse metazoans, from arthropods to vertebrates, we found extensive functional coherence across many miRNAs (fig. 4). In each species, a large faction of predicted miRNAs exhibited a significantly greater functional coherence than background. Importantly, all of these previously published target sets were generated using the TargetScan framework, using phylogenetic conservatism of miRNA binding sites as a core identification criteria (Friedman et al. 2009; Agarwal et al. 2015). Common to all species is a substantial variation among miRNA gene sets in their functional coherence (the left vs right side of the P-value ranked distribution of miRNA genes). Whether this variation arises due to unequal coherence across miRNAs, variation in the functional annotation of relevant targets, poorly annotated 3'UTRs, or other factors warrants attention. However, the extensive functional coherence seen across nearly all miRNA genes in *H. sapiens* suggests such variation likely arises due to factors other than unequal coherence among the target sets of miRNA genes. Among these diverse metazoans, the lower functional coherence observed in these cichlids likely arises due to the young age of the clade analyzed (~ 10 million years), as this necessarily results in a lower power via phylogenetic conservatism. Highlighting the need and challenges of bioinformatic target assessment in young clades, this clade of cichlids is an exemplar of adaptive radiations, having generated > 2000 species in the 10 million years, making observations of their massive reorganization of the miRNA GRN incredibly intriguing for evolutionary study (Mehta et al. 2022).

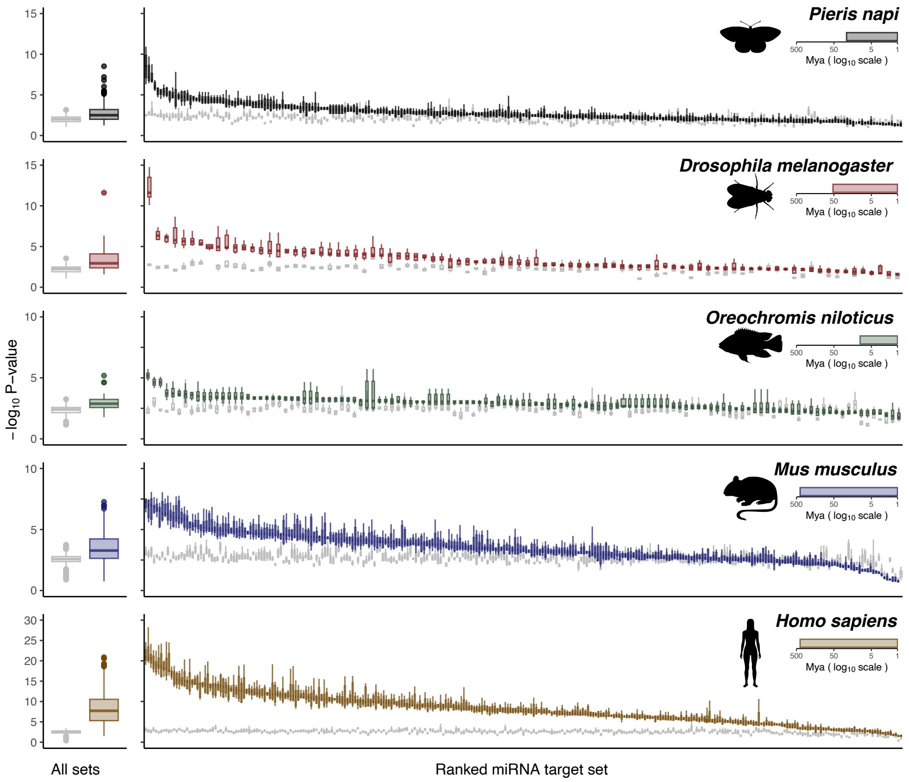

Fig. 4. Functional coherence of miRNA targets across animals measured using gene set enrichment analysis (GSEA). Left-hand boxplots summarize median P-values for random and predicted miRNA target sets, while right-hand boxplots show P-values for the top 10 enriched GO terms per per miRNA gene, ordered by median GSEA P-value within each species. Results from predicted targets are colored while results from randomly selected genes are shown in gray. Inset horizontal bars indicate crown age (million years) of the species used to generate miRNA target predictions. Results from *P. napi* (fig. 3a) are presented here, allowing for direct comparison with four divergent taxa whose published datasets were generated using TargetScan.

Conclusions

Functional coherence in the targets of miRNA genes appears to be common in the tree of life. Using this observation, together with an in-depth study of miRNA targets in a non-model species, our finding of no biological signal among the miRNA targets produced by miRanda and RNAhybrid predictions is consistent with previous findings and warnings of their low precision (Fridrich et al. 2019; Agarwal et al. 2015, 2018; Pinzón et al. 2017; Ritchie et al. 2009). We conclude that a substantial body of research may benefit from revising hypotheses based upon miRNA expression patterns, when those hypotheses relied upon miRNA target prediction lacking measures of evolutionary conservation.

Much remains to be discovered about the role the miRNAs play in adaptive evolution and there has never better time for investigating the role of miRNA posttranslational repression in novel species. An ever-increasing diversity of high-quality genomes provides an unprecedented opportunity for exploiting evolutionary conservation via recent advances in miRNA target prediction (Agarwal et al. 2018). We note however that target predictions are merely another set of hypotheses. Since most miRNAseq studies are also coupled with RNAseq, we further note that correlations between increased miRNA expression and decreases in putative mRNA target expression are also hypotheses fraught with a potential for high false-positives, given the diverse patterns of expression in such datasets coupled with generally few sets of diverse sampling points. Finally, while identified miRNA function in model species can certainly aid hypothesis formulation of miRNA impacts, such relies upon increasingly tenuous assumptions of evolutionarily conserved function (Rusin 2023).

Perhaps the most important way forward for the non-model species community seeking to connect miRNA expression changes with adaptive phenotypes will be via harnessing of emerging gene manipulation technologies in the testing of functional hypotheses (Gudmunds et al. 2022). While the diverse many-to-many relationships inherent in the miRNA GRN necessitate careful design and interpretation of such experiments (Bartel 2018), these also offer unique opportunities. For example, consider a scenario where many independent miRNA genes target the same seed sequence within mRNA. While KO of all such miRNA genes could be lethal, knock out of one, several, or many genes within such a gene family could effectively titrate phenotypic effects. Additionally, advances in single cell sequencing of RNA could greatly advance insights (Sekar et al. 2023), especially in the assessment of miRNA interactions with mRNA GRNs across diverse tissues and developmental courses.

In conclusion, numerous studies across diverse taxa document differential expression of miRNAs suggestive of a potentially important role in adaptive evolutionary phenotypes. However, much work remains to be conducted in order to establish such genotype to phenotype connections. Here, by drawing attention to the challenges of *de novo* miRNA target prediction, we hope that more biologically meaningful hypotheses will emerge that can then be tested by modification of miRNA genes or their target sites, much as mRNA based hypotheses are now routinely explored via CRE and coding region manipulations (Gudmunds et al. 2022).

Acknowledgements

This work greatly benefitied from discussion with Marc Frilander, Emilio Sanchez, and members of the Wheatlab. We also thank the Society for Molecular Biology and Evolution for allowing C.W.W. to present this work at, and receive extensive feedback from attendees, at the SMBE annual meeting in 2023. This work was supported by the Carl Tryggers Stiftelse (grant no. CTS20-242), the Swedish Research Council (2015-04218, 2017-04386, 2019-03441), and the Knut and Alice Wallenberg Foundation (grant number 2012.0058).

Author contributions

C.W.W. performed all the bioinformatic analyses involved in the generation of miRNA targets using TargetScan. R.S. provided R code for generating systematic GSEA for all miRNA gene families and plotting the results. P.E. and K.R. ran the miRanda and RNAhybrid analyses. C.W.W. and K.R. conceived of the study, with input from R.S. C.W.W. wrote the manuscript with feedback from K.R. and the other coauthors. Y.O. and H.V. provided two genomes for analyses. All authors approve of the manuscript.

Data Availability Statement

Scripts

Targets

Intermediate files

References

Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs Izaurralde, E, editor. eLife. 4:e05005. doi: 10.7554/eLife.05005.

Agarwal V, Subtelny AO, Thiru P, Ulitsky I, Bartel DP. 2018. Predicting microRNA targeting efficacy in Drosophila. Genome Biol. 19:152. doi: 10.1186/s13059-018-1504-3.

Alexa A, Rahnenfuhrer J. 2023. topGO: Enrichment Analysis for Gene Ontology. doi: 10.18129/B9.bioc.topGO.

Amemiya CT et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature. 496:311–316. doi: 10.1038/nature12027.

Armstrong J et al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 587:246–251. doi: 10.1038/s41586-020-2871-y.

Bartel DP. 2018. Metazoan MicroRNAs. Cell. 173:20–51. doi: 10.1016/j.cell.2018.03.006.

Bleazard T, Lamb JA, Griffiths-Jones S. 2015. Bias in microRNA functional enrichment analysis. Bioinformatics. 31:1592–1598. doi: 10.1093/bioinformatics/btv023.

Bracken CP, Scott HS, Goodall GJ. 2016. A network-biology perspective of microRNA function and dysfunction in cancer. Nat. Rev. Genet. 17:719–732. doi: 10.1038/nrg.2016.134.

Bruce HS, Patel NH. 2020. Knockout of crustacean leg patterning genes suggests that insect wings and body walls evolved from ancient leg segments. Nat. Ecol. Evol. 4:1703–1712. doi: 10.1038/s41559-020-01349-0.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2020. BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. bioRxiv. 2020.08.10.245134. doi: 10.1101/2020.08.10.245134.

Bryce-Smith S et al. 2023. *Extensible benchmarking of methods that identify and quantify polyadenylation sites from RNA-seq data*. Bioinformatics doi: 10.1101/2023.06.23.546284.

Camacho C et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421. doi: 10.1186/1471-2105-10-421.

Chazot N et al. 2019. Priors and Posteriors in Bayesian Timing of Divergence Analyses: The Age of Butterflies Revisited. Syst. Biol. 68:797–813. doi: 10.1093/sysbio/syz002.

Cui Q, Yu Z, Purisima EO, Wang E. 2006. Principles of microRNA regulation of a human cellular signaling network. Mol. Syst. Biol. 2:46. doi: 10.1038/msb4100089.

Derti A et al. 2012. A quantitative atlas of polyadenylation in five mammals. Genome Res. 22:1173–1183. doi: 10.1101/gr.132563.111.

Dutheil JY, Gaillard S, Stukenbrock EH. 2014. MafFilter: a highly flexible and extensible multiple genome alignment files processor. BMC Genomics. 15:np. doi: 10.1186/1471-2164-15-53.

Enright AJ et al. 2003. MicroRNA targets in Drosophila. Genome Biol. 5:R1. doi: 10.1186/gb-2003-5-1-r1.

Erwin DH. 2021. A conceptual framework of evolutionary novelty and innovation. Biol. Rev. 96:1–15. doi: https://doi.org/10.1111/brv.12643.

Fridrich A, Hazan Y, Moran Y. 2019. Too Many False Targets for MicroRNAs: Challenges and Pitfalls in Prediction of miRNA Targets and Their Gene Ontology in Model and Non-model Organisms. BioEssays. 41:1800169. doi: 10.1002/bies.201800169.

Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2011. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 40:37–52. doi: 10.1093/nar/gkr688.

Friedman RC, Farh KK-H, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19:92–105. doi: 10.1101/gr.082701.108.

Fruciano C, Franchini P, Jones JC. 2021. Capturing the rapidly evolving study of adaptation. J. Evol. Biol. 34:856–865. doi: 10.1111/jeb.13871.

Gudmunds E, Wheat CW, Khila A, Husby A. 2022. Functional genomic tools for emerging model species. Trends Ecol. Evol. 37:1104–1115. doi: 10.1016/j.tree.2022.07.004.

Gusev Y. 2008. Computational methods for analysis of cellular functions and pathways collectively targeted by differentially expressed microRNA. Methods. 44:61–72. doi: 10.1016/j.ymeth.2007.10.005.

Haas BJ et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512. doi: 10.1038/nprot.2013.084.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. Bioinformatics. 32:767–769. doi: 10.1093/bioinformatics/btv661.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. In: Gene Prediction: Methods and Protocols. Kollmar, M, editor. Methods in Molecular Biology Springer: New York, NY pp. 65–95. doi: 10.1007/978-1-4939-9173-0_5.

Huang Z, Teeling EC. 2017. ExUTR: a novel pipeline for large-scale prediction of 3′-UTR sequences from NGS data. BMC Genomics. 18:847. doi: 10.1186/s12864-017-4241-1.

Huerta-Cepas J et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47:D309–D314. doi: 10.1093/nar/gky1085.

Kang W et al. 2018. miRTrace reveals the organismal origins of microRNA sequencing data. Genome Biol. 19:213. doi: 10.1186/s13059-018-1588-9.

Kern F et al. 2020. What's the target: understanding two decades of in silico microRNA-target prediction. Brief. Bioinform. 21:1999–2010. doi: 10.1093/bib/bbz111.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37:540–546. doi: 10.1038/s41587-019-0072-8.

Krüger J, Rehmsmeier M. 2006. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res. 34:W451–W454. doi: 10.1093/nar/gkl243.

Lee SY, Sohn K-A, Kim JH. 2012. MicroRNA-centric measurement improves functional enrichment analysis of co-expressed and differentially expressed microRNA clusters. BMC Genomics. 13:S17. doi: 10.1186/1471-2164-13-S7-S17.

Leung AKL, Sharp PA. 2010. MicroRNA Functions in Stress Responses. Mol. Cell. 40:205–215. doi: 10.1016/j.molcel.2010.09.027.

Liu Y, Beyer A, Aebersold R. 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 165:535–550. doi: 10.1016/j.cell.2016.03.014.

Lo Giudice C et al. 2023. UTRdb 2.0: a comprehensive, expert curated catalog of eukaryotic mRNAs untranslated regions. Nucleic Acids Res. 51:D337–D344. doi: 10.1093/nar/gkac1016.

Lohse K, Mackintosh A, et al. 2021. The genome sequence of the large white, Pieris brassicae (Linnaeus, 1758). Wellcome Open Res. 6:262. doi: 10.12688/wellcomeopenres.17274.1.

Lohse K, Ebdon S, et al. 2021. The genome sequence of the small white, Pieris rapae (Linnaeus, 1758). Wellcome Open Res. 6:273. doi: 10.12688/wellcomeopenres.17288.1.

Lohse K, Hayward A, et al. 2021. The genome sequences of the male and female green-veined white, Pieris napi (Linnaeus, 1758). Wellcome Open Res. 6:288. doi: 10.12688/wellcomeopenres.17277.1.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494–6506. doi: 10.1093/nar/gki937.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. 38:4647–4654. doi: 10.1093/molbev/msab199.

McGeary SE et al. 2019. The biochemical basis of microRNA targeting efficacy. Science. 366:eaav1741. doi: 10.1126/science.aav1741.

Mehta TK et al. 2022. Evolution of miRNA-Binding Sites and Regulatory Networks in Cichlids Parsch, J, editor. Mol. Biol. Evol. 39:msac146. doi: 10.1093/molbev/msac146.

Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. Genome Biol. 3:reviews0004.1. doi: 10.1186/gb-2002-3-3-reviews0004.

Pinzón N et al. 2017. microRNA target prediction programs predict many false positives. Genome Res. 27:234–245. doi: 10.1101/gr.205146.116.

Pruisscher P, Lehmann P, Nylin S, Gotthard K, Wheat CW. 2021. Extensive transcriptomic profiling of pupal diapause in a butterfly reveals a dynamic phenotype. Mol. Ecol. n/a. doi: 10.1111/mec.16304.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842. doi: 10.1093/bioinformatics/btq033.

Ritchie W, Flamant S, Rasko JEJ. 2009. Predicting microRNA targets and functions: traps for the unwary. Nat. Methods. 6:397–398. doi: 10.1038/nmeth0609-397.

Rusin LY. 2023. Evolution of homology: From archetype towards a holistic concept of cell type. J. Morphol. 284:e21569. doi: 10.1002/jmor.21569.

Sanfilippo P, Wen J, Lai EC. 2017. Landscape and evolution of tissue-specific alternative polyadenylation across Drosophila species. Genome Biol. 18:229. doi: 10.1186/s13059-017-1358-0.

Sekar V et al. 2023. Detection of transcriptome-wide microRNA–target interactions in single cells with agoTRIBE. Nat. Biotechnol. 1–7. doi: 10.1038/s41587-023-01951-0.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24:637–644. doi: 10.1093/bioinformatics/btn013.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 7:62. doi: 10.1186/1471-2105-7-62.

Steward RA, Okamura Y, Boggs CL, Vogel H, Wheat CW. 2021. The Genome of the Margined White Butterfly (Pieris macdunnoughi): Sex Chromosome Insights and the Power of Polishing with PoolSeq Data Lavrov, D, editor. Genome Biol. Evol. 13:evab053. doi: 10.1093/gbe/evab053.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science. 302:249–255. doi: 10.1126/science.1087447.

Suvorov A et al. 2022. Widespread introgression across a phylogeny of 155 Drosophila genomes. Curr. Biol. 32:111-123.e5. doi: 10.1016/j.cub.2021.10.052.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 18:1979–1990. doi: 10.1101/gr.081612.108.

Trapnell C et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28:511–515. doi: 10.1038/nbt.1621.

Tsang JS, Ebert MS, van Oudenaarden A. 2010. Genome-wide Dissection of MicroRNA Functions and Cotargeting Networks Using Gene Set Signatures. Mol. Cell. 38:140–153. doi: 10.1016/j.molcel.2010.03.007.

Wang W et al. 2019. Evolutionary and functional implications of 3′ untranslated region length of mRNAs by comprehensive investigation among four taxonomically diverse metazoan species. Genes Genomics. 41:747–755. doi: 10.1007/s13258-019-00808-8.

Wolfe CJ, Kohane IS, Butte AJ. 2005. Systematic survey reveals general applicability of ‘guilt-by-association’ within gene coexpression networks. BMC Bioinformatics. 6:227. doi: 10.1186/1471-2105-6-227.

Xu J, Wong C. 2008. A computational screen for mouse signaling pathways targeted by microRNA clusters. RNA. 14:1276–1283. doi: 10.1261/rna.997708.

Ye W, Lian Q, Ye C, Wu X. 2023. A Survey on Methods for Predicting Polyadenylation Sites from DNA Sequences, Bulk RNA-seq, and Single-cell RNA-seq. Genomics Proteomics Bioinformatics. 21:67–83. doi: 10.1016/j.gpb.2022.09.005.

Zimin AV et al. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii* , a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 27:787–792. doi: 10.1101/gr.213405.116.