MirGeneDB2.0: the curated microRNA Gene Database

Non-coding RNAs (ncRNA), a significant part of the increasingly popular '*dark matter*' of the human genome[1], have gained substantial attention due to their involvement in animal development and human disorders such as cardiovascular diseases and cancer[2]. Although many different types of regulatory ncRNAs have been discovered over the last 25 years, microRNAs (miRNAs) are unique within these as they are the only class of ncRNAs with individual genes sequentially conserved across the animal kingdom[3]. Because of the conserved roles miRNAs play in establishing robustness of gene regulatory networks across Metazoa[4], it is important that homologous miRNAs in different species are correctly identified, annotated, and named using consistent criteria[5] against the backdrop of numerous other types of coding and non-coding RNA fragments[6].

Unlike miRBase[7], which has developed organically through community-wide submissions, and thus does not use consistent annotation or nomenclature criteria[6], MirGeneDB2.0 (http://mirgenedb.org), a manually curated open source miRNA gene database, contains high quality annotations of 7,785 *bona fide* and consistently named miRNAs from 32 species representing major metazoan groups (including many invertebrate and vertebrate model organisms). The number of miRNAs conforming to the annotation criteria is almost four times higher than in miRBase (~2000 for the miRBase 'high confidence' set[7]), and can be considered free of *false positives*. For the expansion of the previous version, we used more than 250 publicly available sequencing datasets (for a total of 4.2 billion reads) derived from at least one representative dataset for each organism (such as whole organisms, organs, tissues or cell-types), which allowed for a consistent and uniform annotation of microRNAomes for each species (Supplementary File, "file_info"; Supplementary Methods)[8]. Existing MirGeneDB.org miRNA complements for human, mouse, chicken and zebrafish were expanded from our initial effort by 65, 49, 28 and 100 genes, respectively

(Supplementary File, table), and annotation-accuracy was further improved using available Cap Analysis of Gene Expression (CAGE) data when available (Supplementary File, "CAGE")[9].

Because miRBase has become increasingly heterogeneous with respect to the number of *bona fide* miRNAs relative to other types of non-coding RNAs, it has considerable variation in the number of miRNAs for closely related groups (Supplementary File, graph miRBase). However, in MirGeneDB, congruent miRNA complements in terms of total miRNA genes and miRNA families were observed in related groups, such as the Vertebrates and arthropods[3,10] (Figure 1). Big differences between miRBase and MirGeneDB2.0 can be observed because miRBase has on the one hand a much larger number of annotated sequences for some of the 23 taxa shared with MirGeneDB2.0 including human, mouse, and chicken, accounting for 4,243 *false positives*, and on the other hand it lacks 22% of all MirGeneDB2.0 genes, accounting for 1,180 *false negatives* (Figure 2, Supplementary File, "overview"). Finally, 31% of the remaining 4,275 miRNAs are incompletely annotated in miRBase, whereas in MirGeneDB2.0 each miRNA has both arms annotated, with a clear distinction made between sequenced reads and predicted reads for each miRNA entry with predictions derived from both considerations of secondary structure and expressed orthologues in other taxa.

The expanded web-interface of MirGeneDB2.0 allows browsing, searching and downloading of miRNA-complements for each organism. Annotations are downloadable as fasta, gff, or bed-files containing distinct sub-annotations for all miRNA components such as precursor (pre), mature, loop, co-mature or star sequences. Unlike miRBase, seed sequences are also identified, and can be searched independently from the rest of the mature sequence. In addition, we included 30-nucleotide flanking regions on both arms for each precursor transcript to generate an extended precursor transcript, which again is downloadable.

MirGeneDB2.0 employs an internally consistent nomenclature system where genes of common descent are assigned the same miRNA family name, allowing for the easy recognition of both orthologues in other species, and paralogues within the same species. This nomenclature system allows for an accurate reconstruction of ancestral miRNA repertoires – both at the family level and at the gene level – that is now provided in MirGeneDB2.0 for all nodes leading to the 32 terminal taxa considered, which allows users to easily assess both gains and losses of miRNA genes through time. However, in order to not increase confusion about the naming of miRNA genes, we continue to provide commonly used miRBase names – if available – in our "*Browse*" section of MirGeneDB2.0 (i.e. http://mirgenedb.org/browse/hsa).

Gene-pages for each miRNA gene contain names, orthologues & paralogues, downloadable sequences, structure, and a range of other previously available information including genomic coordinates (i.e. http://mirgenedb.org/show/hsa/Let-7-P1). New features in MirGeneDB2.0 include accurate information on 3' non-templated uridylations, which characterize an important sub-group of miRNAs; information of the presence or absence of the recently discovered sequential motifs (UG, UGUG, CNNC); and the visualization of at least one expression dataset for each gene in each organism. Further, *read-pages* are also provided for each gene (i.e. http://mirgenedb.org/static/graph/hsa/results/Hsa-Let-7-P1.html), which show an overview of read-stacks on the corresponding extended precursor sequence of each *gene-page.* They contain detailed representation of templated and non-templated reads for individual datasets for each gene including reports on miRNA isoforms and downloadable read-mappings.

The establishment of this carefully curated data base of miRNA genes, supplementing existing databases including miRBase, allows for a stable and robust foundation for miRNA studies, in particular studies that rely on cross-species comparisons to explore the roles

miRNAs play in development and disease, as well as the evolution of miRNAs (and animals) themselves.

Note: Supplementary Methods and Supplementary files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank Victor Ambros, David Bartel, Marc Friedländer, Marc Halushka, Andreas Keller, Gianvito Urgese for discussions, Georgios Magklaras for IT support. B.F. has been supported by the South-Eastern Norway Regional Health Authority (Grant No. 2014041). AM has been supported by the Norwegian Research Council, Helse Sør-Øst, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM), which is part of the Nordic European Molecular Biology Laboratory partnership for Molecular Medicine. K.J.P. was supported by NASA-Ames.

AUTHOR CONTRIBUTIONS

BF and KJP conceived MirGeneDB2.0, compiled miRNA complements for all organisms. DD created read-pages and heatmaps. MJ set up the framework and database. MH processed sRNA sequencing data, AM processed and analyzed CAGE data. EHøye created scripts for mature/star annotation. EHovig and KF provided infrastructure and all authors read and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Bastian Fromm 1, Diana Domanska 2, Michael Hackenberg 3, Anthony Mathelier 4,5, Eirik Høye 1, Morten Johansen 6, Eivind Hovig 1,2,6, Kjersti Flatmark 1,7,8 & Kevin J. Peterson 9.

1 - Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Nydalen, N-0424 Oslo, Norway, 2 - Department of Informatics, University of Oslo, Blindern, N-0318 Oslo, Norway, 3 - Department of Genetics, Faculty of Sciences, University of Granada, Granada, 1s8071, Spain, 4 - Centre for Molecular

Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, N-0318 Oslo, Norway, 5 - Department of Cancer Genetics, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital Radiumhospitalet, N-0424 Oslo, Norway, 6 - Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, N-0424 Oslo, Norway, 7 - Department of Gastroenterological Surgery, The Norwegian Radium Hospital, Oslo University Hospital, Nydalen, N-0424 Oslo, Norway, 8 - Institute of Clinical Medicine, University of Oslo, Blindern, N-0318 Oslo, Norway, 9 - Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03755, U.S..

e-mail: BastianFromm@gmail.com or Kevin.J.Peterson@dartmouth.edu

References

1 Blaxter, M. Genetics. Revealing the dark matter of the genome. *Science* **330**, 1758-1759, doi:10.1126/science.1200700 (2010).

2 Esteller, M. Non-coding RNAs in human disease. *Nature reviews. Genetics* **12**, 861-874, doi:10.1038/nrg3074 (2011).

3 Wheeler, B. *et al.* The deep evolution of metazoan microRNAs. *Evolution & development* **11**, 50 - 68 (2009).

4 Ebert, M. S. & Sharp, P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149**, 515-524, doi:10.1016/j.cell.2012.04.005 (2012).

5 Ambros, V. A uniform system for microRNA annotation. *Rna* **9**, 277-279, doi:10.1261/rna.2183803 (2003).

6 Tosar, J. P., Rovira, C. & Cayota, A. Non-coding RNA fragments account for the majority of annotated piRNAs expressed in somatic non-gonadal tissues. *Communications Biology* **1**, 2, doi:10.1038/s42003-017-0001-7 (2018).

7 Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **42**, D68-73, doi:10.1093/nar/gkt1181 (2014).

8 Fromm, B. *et al.* A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annual review of genetics* **49**, 213-242, doi:10.1146/annurev-genet-120213-092023 (2015).

9 de Rie, D. *et al.* An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nature biotechnology* **35**, 872-878, doi:10.1038/nbt.3947 (2017).

10 Tarver, J. E. *et al.* miRNAs: small genes with big potential in metazoan phylogenetics. *Molecular biology and evolution* **30**, 2369-2382, doi:10.1093/molbev/mst133 (2013).

Figures

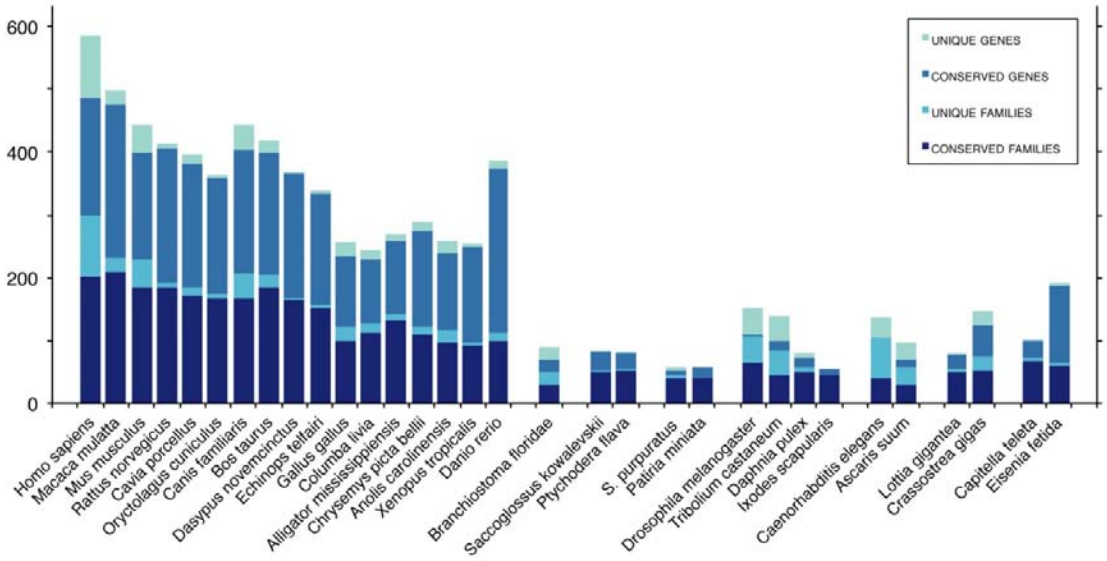


Figure 1: High consistency of conserved miRNA gene and family numbers in closely related groups in MirGeneDB2.0 can be observed for groups with more than two representatives. High variation in gene-numbers for *Danio* and *Eisenia* (double asterisks) are explainable by genome-duplication events within that particular monophyletic group (vertebrates and annelids, respectively), while high numbers of unique /novel genes and families in *Homo*, *Mus*, *Canis*, *Drosophila*, *Tribolium* and *Caenorhabditis* might be explainable by the significantly higher number of studies and/or the relatively higher number of absolute small RNA reads on these organisms (single asterisks).

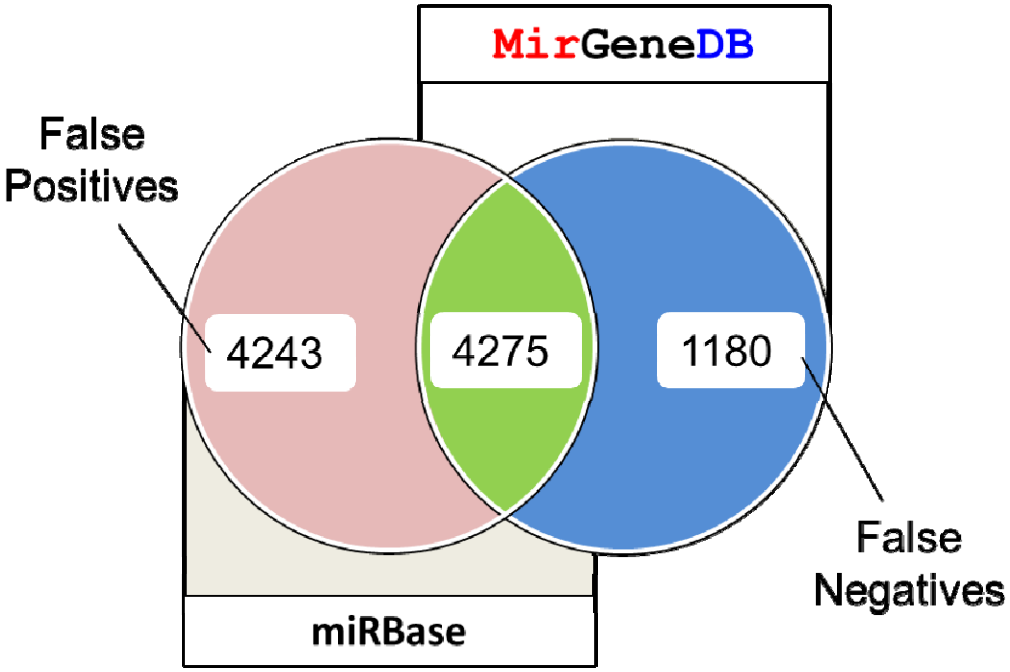


Figure 2: High number of incorrect and missing miRNA annotations in miRBase as compared to MirGeneDB. A comparison of the microRNA complements of 23 organisms shared between miRBase and MirGeneDB revealed that only 4,275 of the 8,531 entries in miRBase are shared with MirGeneDB (green). An additional 4,243 miRBase entries represent false positives (red), miRNAs found in miRBase that do not satisfy standard annotation criteria, whereas 1,180 MirGeneDB entries represent false negatives (blue), miRNAs that are present in these taxa that are not currently annotated in miRBase.

Supplementary Methods

Features of miRNAs

In the last two decades the small non-coding RNA field has significantly expanded beyond snRNAs & snoRNAs[1] to include piRNAs[2], siRNAs[3] , novel small RNAs derived from known non-coding RNAs including tRNAs[4] and rRNAs[5] and, of course, microRNAs (miRNAs)[6-9]. Each of these types of smallRNAs is characterized by a distinctive suite of characteristics and a unique evolutionary history. MiRNAs can be distinguished from other small genomically encoded RNA families by a set of unique features as described earlier[10,11]: The presence of two 20-26nt long reads that are expressed from each of the two arms derived from a stable hairpin precursor is essential to assess whether or not Drosha and Dicer were involved in the processing. Since the ends of canonical miRNA reads are generated enzymatically, the 5' ends of the reads are homogeneous (>90%). The hairpin precursor shows imperfect complementarity and base pairs in at least 16 of the ~ 22 nucleotides. The 5p and 3p reads are offset by 2 nucleotides on both ends due to the sequential processing of the miRNA transcript by Drosha and Dicer to generate the mature ~22 nucleotide read(s). In some cases, the Drosha offset is only offset by 1 templated nucleotide, but in these cases the 3' end of the 3p arm is monouridilyated[12,13]. The length of the loop is at least 8 nucleotides long; there is no apparent maximum in loop length, even in organisms possessing only a single Dicer gene, contra our earlier statement[11], even though most taxa like vertebrates with single Dicer genes never show loop lengths greater than ~40 nucleotides.

There are other features of miRNAs, in particular structural and evolutionary signatures that allow them to be further distinguished from other small RNAs. The mature miRNA sequence usually starts with A or U, and is often mismatched with the complementary arm, which seems to facilitate arm selection by Argonaute (at least in mammals)[11,14,15]. Nucleotide

positions 2 through 8 of the mature sequence (the "seed") are strongly conserved through evolution, as are positions 13-16 (the 3' complementary region)[11,16]. Recently it was demonstrated that processing motifs are often (but not always) present in the primary miRNA transcript including a UG motif 14 nucleotides upstream of the 5p arm, a UGU motif at the 3' end of the 5p arm, and a CNNC motif 17 nucleotides downstream of the 3p arm[17-19].

Recognition and utilization of clear and, for the most part mechanistically well understood, criteria for the annotation of miRNAs allows the delineation of *bona fide* miRNAs from the myriad small RNAs generated in eukaryotic cells, providing deeper and more significant insights into their function, possible mis-regulation, and evolution.

Data processing

Publicly available smallRNA sequencing data of whole organisms, healthy organs, tissue or cell-isolates was downloaded from European Genome-phenome Archive (EGA), the Sequence Read Archive (SRA) and the the Gene Expression Omnibus (GEO) respectively (see Supplementary File, “file_info”). For download and processing we used the latest version of sRNAbench[20]. Corresponding files were automatically downloaded and converted into fastq files. All datasets were consistently processed with the following parameters: 3' adapter sequences were automatically identified and trimmed using sRNAbench (detection of at least 10nt of the adapter allowing 1 mismatch)[20]; reads within length of 18 and 27 nts were retained and collapsed for mapping employing fastx-toolkit, and custom perl scripts. Collapsed reads were mapped to miRBase complements and MirGeneDB[11] using bowtie1.2[21], requiring an 18 nucleotide seed sequence of zero mismatches to avoid cross-mapping. All mappings were transformed to bam-files using SAMtools[22].

Confident miRNA complement annotations for 32 metazoan taxa

Similar to our previous efforts for the four vertebrates of MirGeneDB1.0[11], we analyzed all miRBase "miRNA" entries for the 23 available taxa. We included, however, at least one smallRNAseq dataset to assess the status of the unique miRNA features for each miRNA individually. For this task, we used an improved version of MirMiner to visualize read-mappings and structures, which also allowed us to predict previously missing genes and the miRNA complements for the nine organisms which are not found in miRBase[16].

Because these criteria can only be applied in miRNAs processed by the canonical pathway (i.e. they are processed by Drosha and Dicer respectively), and literally no non-canonical miRNA is conserved beyond very close relatives, we have only considered canonical miRNAs. One exception we made was the non-canonical erythroid miRNA Mir-451 that is a very important regulator of erythroid development and highly conserved in all vertebrates[23] (http://mirgenedb.org/browse/ALL?family=MIR-451).

Mature arm annotations

Mature and star, and Co-mature status of miRNAs was assigned by assessing the expression of 5' and 3' arms overall available datasets, respectively. Only if one arm was expressed more than twofold higher as the other mature status was assigned, else Co-mature status was given (Supplementary markdown). In the few cases were arms were not both expressed we used information from orthologous genes in related organisms and assigned predicted status of mature /star based on the expression ratios in the corresponding datasets of the related species.

Refinement of pre-miRNA 3'end annotation with CAGE data

Human annotation

We downloaded the hg38 bigwig files associated to all ENCODE CAGE experiments from the ENCODE data portal (see https://www.encodeproject.org/metadata/type=Experiment&assay_slims=Transcription&assay_title=CAGE&assembly=GRCh38&files.file_type=bigWig/metadata.tsv). We merged the data from all the experiments and converted the files in the BED format. Computation and plotting of the distribution of CAGE tags around the 3' end of pre-miRNAs annotated in MirGeneDB were performed using the deepTools v. 2.4.0[24,25] (Supplementary Figure 1a). As described in previous studies[26,27] we observed a peak for CAGE tags 1 nt downstream (i.e. the +1 nt) of the pre-miRNA 3' ends (Supplementary Figure 1). We considered for manual curation the pre-miRNAs showing a higher number of CAGE tags at positions 0 or +2 with respect to the annotated pre-miRNA 3' end, which could correspond to a 1 nt off misannotation (see http://fantom.gsc.riken.jp/zenbu/gLyphs/#config=ufw7Z_rvFF5juG_FZhbOD for an example). After manual curation through the Zenbu genome browser[28], we corrected the 3' end position for pre-miRNAs Hsa-Mir-145, Hsa-Let-7-P12, Hsa-Let-7-P7 (Supplementary File, "CAGE").

Zebrafish annotation

We applied the same methodology to the CAGE data obtained from 12 developmental stages of embryogenesis in zebrafish[27]. Bigwig files of CAGE tags mapping were retrieved using the CAGEr R package[29]. Data from all developmental stages were merged to analyze the distribution of CAGE tags around pre-miRNA 3' ends (Supplementary Figure 1b). After manual curation, we updated the 3' end position of the pre-miRNAs Dre-Mir-153-P1a and Dre-Let-7-P6 (Supplementary File, "CAGE").

Water flea annotation

CAGE data for *Daphnia pule*x derived from three developmental states were retrieved from GEO (GSE80141)[30]. We followed the same steps as described above for human and zebrafish CAGE data but did not find any 3' end annotation of pre-miRNAs to update.

Supplementary References

1 Matera, A. G., Terns, R. M. & Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature reviews. Molecular cell biology* **8**, 209-220, doi:10.1038/nrm2124 (2007).

2 Lau, N. C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313**, 363-367, doi:10.1126/science.1130164 (2006).

3 Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950-952 (1999).

4 Goodarzi, H. *et al.* Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. *Cell* **161**, 790-802, doi:10.1016/j.cell.2015.02.053 (2015).

5 Chak, L. L., Mohammed, J., Lai, E. C., Tucker-Kellogg, G. & Okamura, K. A deeply conserved, noncanonical miRNA hosted by ribosomal DNA. *Rna* **21**, 375-384, doi:10.1261/rna.049098.114 (2015).

6 Lee, R. C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862-864 (2001).

7 Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).

8 Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).

9 Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-858 (2001).

10 Ambros, V. A uniform system for microRNA annotation. *Rna* **9**, 277-279, doi:10.1261/rna.2183803 (2003).

11 Fromm, B. *et al.* A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annual review of genetics* **49**, 213-242, doi:10.1146/annurev-genet-120213-092023 (2015).

12 Kim, B. *et al.* TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms. *The EMBO journal* **34**, 1801-1815, doi:10.15252/embj.201590931 (2015).

13 Kim, Y. K., Kim, B. & Kim, V. N. Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E1881-1889, doi:10.1073/pnas.1602532113 (2016).

14 Suzuki, H. I. *et al.* Small-RNA asymmetry is directly driven by mammalian Argonautes. *Nature structural & molecular biology* **22**, 512-521, doi:10.1038/nsmb.3050 (2015).

15 Schirle, N. T., Sheu-Gruttadauria, J. & MacRae, I. J. Structural basis for microRNA targeting. *Science* **346**, 608-613, doi:10.1126/science.1258040 (2014).

16 Wheeler, B. *et al.* The deep evolution of metazoan microRNAs. *Evolution & development* **11**, 50 - 68 (2009).

17 Nguyen, T. A. *et al.* Functional Anatomy of the Human Microprocessor. *Cell* **161**, 1374-1387, doi:10.1016/j.cell.2015.05.010 (2015).

18 Fang, W. & Bartel, D. P. The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. *Molecular cell* **60**, 131-145, doi:10.1016/j.molcel.2015.08.015 (2015).

19 Auyeung, V. C., Ulitsky, I., McGeary, S. E. & Bartel, D. P. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* **152**, 844-858, doi:10.1016/j.cell.2013.01.031 (2013).

20 Rueda, A. *et al.* sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res* **43**, W467-473, doi:10.1093/nar/gkv555 (2015).

21 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

22 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

23 Jee, D. *et al.* Dual Strategies for Argonaute2-Mediated Biogenesis of Erythroid miRNAs Underlie Conserved Requirements for Slicing in Mammals. *Molecular cell* **69**, 265-278 e266, doi:10.1016/j.molcel.2017.12.027 (2018).

24 Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187-191, doi:10.1093/nar/gku365 (2014).

25 Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

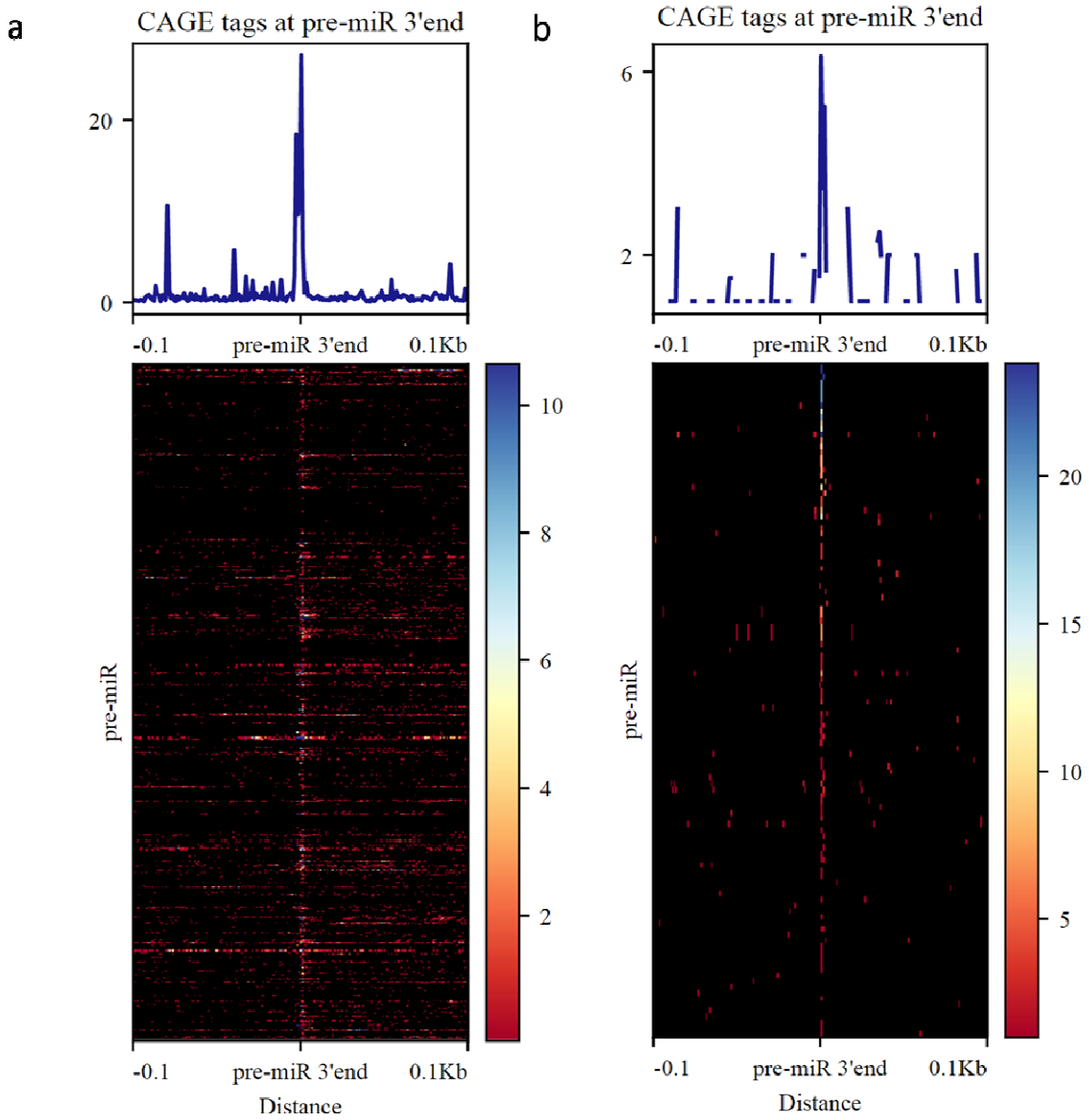
26 de Rie, D. *et al.* An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* **35**, 872-878, doi:10.1038/nbt.3947 (2017).

27 Nepal, C. *et al.* Transcriptional, post-transcriptional and chromatin-associated regulation of pri-miRNAs, pre-miRNAs and moRNAs. *Nucleic Acids Res.* **44**, 3070-3081, doi:10.1093/nar/gkv1354 (2016).

28 Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat. Biotechnol.* **32**, 217-219, doi:10.1038/nbt.2840 (2014).

29 Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51, doi:10.1093/nar/gkv054 (2015).

30 Raborn, R. T., Spitze, K., Brendel, V. P. & Lynch, M. Promoter Architecture and Sex-Specific Gene Expression in Daphnia pulex. *Genetics* **204**, 593-612, doi:10.1534/genetics.116.193334 (2016).



Supplementary Figure 1: The distribution of CAGE tags around the 3' end of pre-miRNAs annotated in MirGeneDB for a) human and b) zebrafish shows a clear peak for CAGE tags 1 nt downstream (i.e. the +1 nt) of the pre-miRNA 3' ends as described before[26,27].