MirGeneDB: The metazoan microRNA complement

Bastian Fromm[1,2*], Diana Domanska[3,4], Eirik Høye[2,5], Vladimir Ovchinnikov[6,7], Wenjing Kang[1], Ernesto Aparicio-Puerta[8], Morten Johansen[3], Kjersti Flatmark[2,5,9], Anthony Mathelier[10,11], Eivind Hovig[2,3], Michael Hackenberg[8], Marc R. Friedländer[1] and Kevin J. Peterson[12]

[1]Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden, [2]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway, [3]Center for Bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway, [4]Department of Pathology, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, [5]Institute of Clinical Medicine, University of Oslo, Oslo, Norway, [6]School of Life Sciences, Faculty of Health and Life Sciences, University of Nottingham, UK, [7]Department of Human and Animal Genetics, The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russian Federation, [8]Department of Genetics, Faculty of Sciences, University of Granada, Granada, Spain, [9]Department of Gastroenterological Surgery, The Norwegian Radium Hospital, Oslo University Hospital, Nydalen, Oslo, Norway, [10]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, Oslo, Norway, [11]Department of Cancer Genetics, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital Radiumhospitalet, Oslo, Norway and [12]Department of Biological Sciences, Dartmouth College, Hanover, NH, USA

*corresponding author: Bastian.Fromm@scilifelab.se

MicroRNAs are well defined

- Two 20–26 nt long reads expressed from each of the two arms derived from a hairpin precursor with 2-nt offsets between the 5p and 3p arms [Drosha, Dicer].
- 5-end homogeneity of expression.
- At least 16-nt complementarity between the two arm sequences.
- The loop sequence is at least 8 nt in length; there is no apparent maximum but species with single Dicer proteins typically not exceed ~40 nt

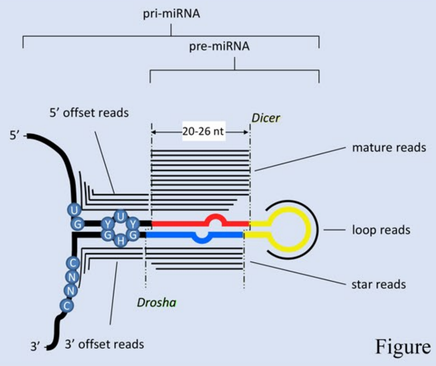


Figure 1

The state of microRNA annotations

- Non-coding RNAs are important players of animal development and diseases
- miRNAs are unique as individual gene sequences can be conserved across all animals
- Because of the fundamental roles miRNAs play, it is important that homologous miRNAs in different species are correctly identified, annotated, and named using consistent criteria against the backdrop of numerous other types of coding and non-coding RNA fragments
- This is not the case for complements of metazoans in miRBase that are extremely heterogeneous and not usable for comparative studies (Figure 2 top)
- Using more than 400 smallRNAseq datasets, we have expanded MirGeneDB (Fromm et al 2015) and annotated in total more than 10,000 miRNA genes from 45 organisms representing nearly every major metazoan group.

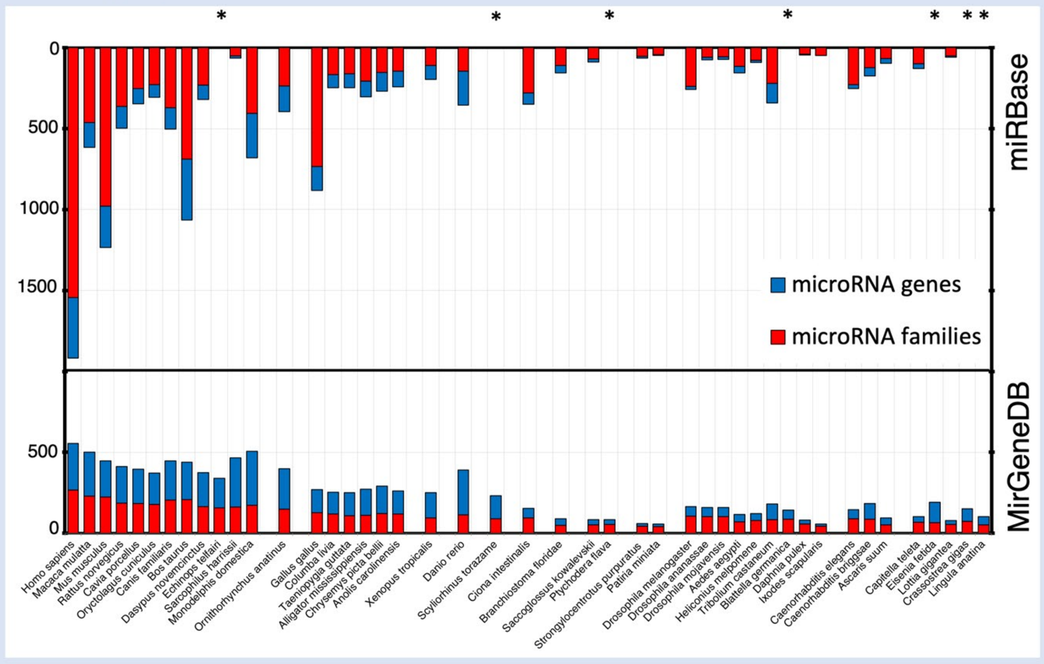
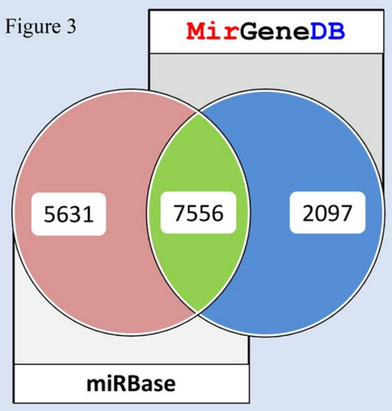


Figure 2

Figure 3



High number of incorrect and missing miRNA annotations in miRBase. The comparison of the microRNA complements of 38 organisms shared between miRBase and MirGeneDB revealed that only 7,556 of the 13,187 entries in miRBase where common with MirGeneDB (green) and 5,631 entries represented *false positive* entries (red). Additional 2,097 miRNA genes that were annotated in MirGeneDB 2.0 for the 38 species were not found in miRBase (= *false negatives*) (blue).

Mir Gene DB

Scan me

an evolutionarily informed nomenclature

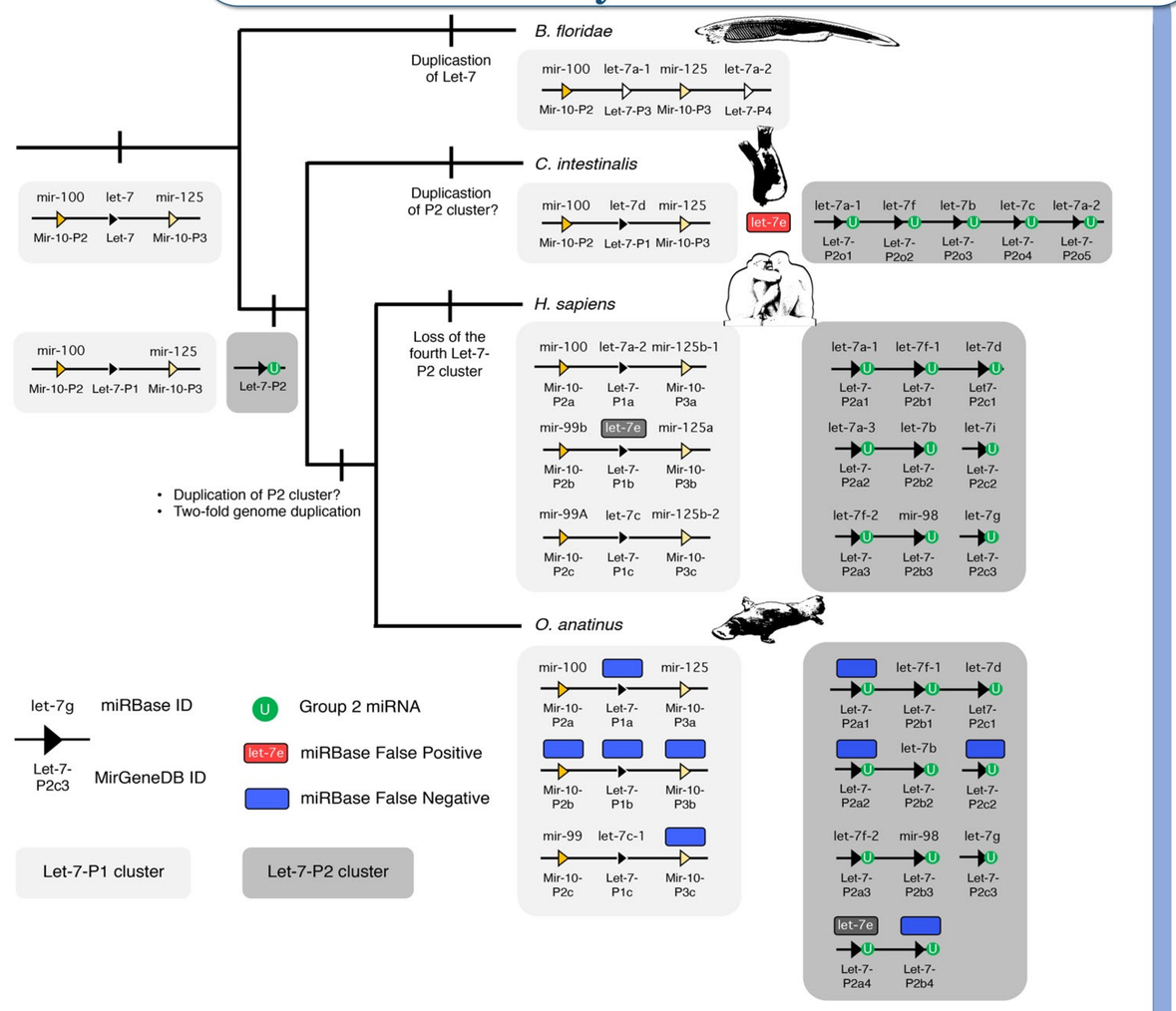


Figure 4: Nomenclature comparison between MirGeneDB and miRBase for representative chordate Let-7s. Shown is the accepted topology for the three major subgroups of chordates, and for each taxon, a (unscaled) representation of the genomic organization of its Let-7 genes/sequences. MirGeneDB names are shown below each of the loci symbols, and the miRBase sequence names are above.

Improved web interface of MirGeneDB

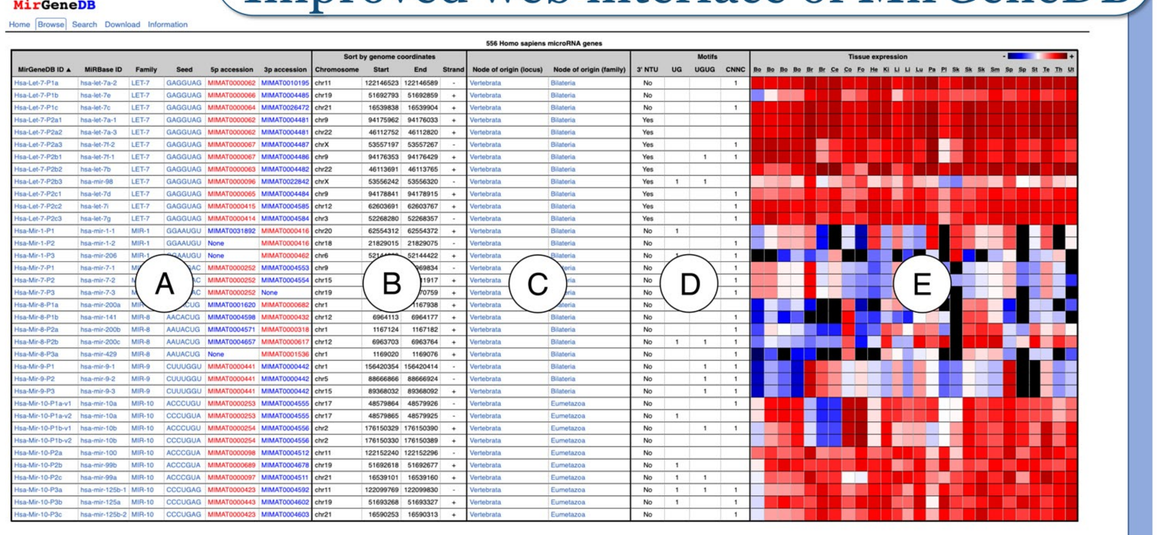
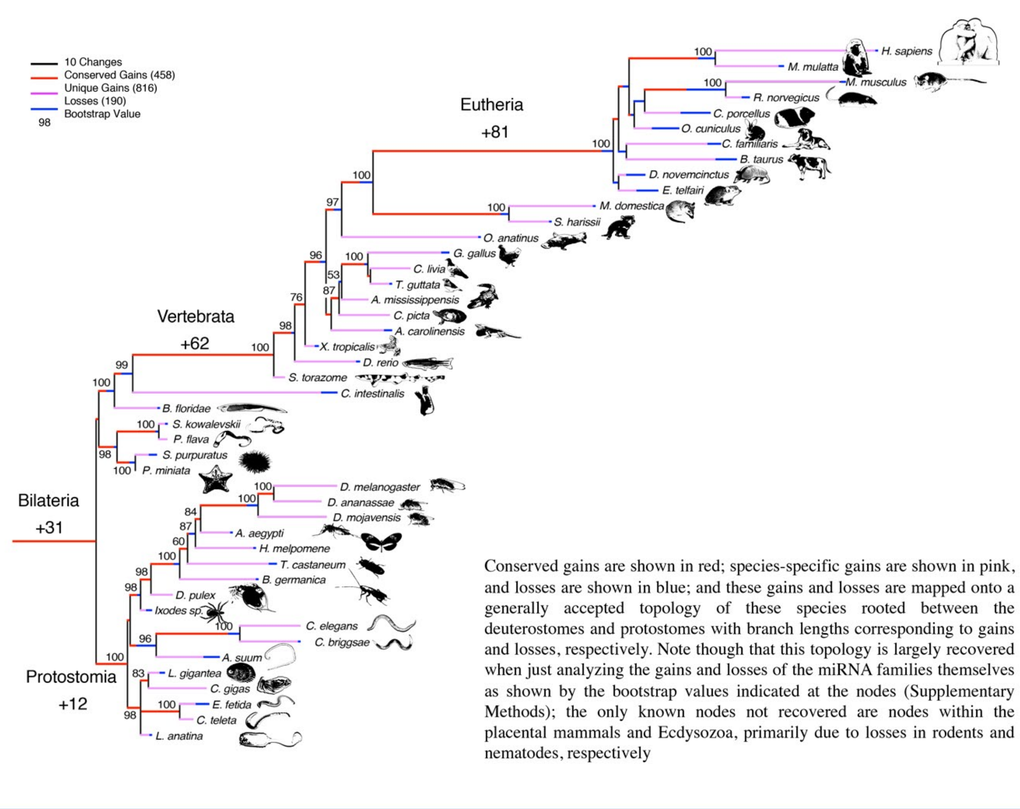


Figure 5: For each species in MirGeneDB an overview browse page exists that lists all genes. For each gene the following information is provided and sortable: hyperlinked names (both MirGeneDB ID and miRBase ID linking to MirGeneDB and miRBase, respectively), family- and seed- assignments, and arm preference (A), genomic coordinates (B); inferred phylogenetic origin of both the gene locus and family (C); information on the presence or absence of 3 NTU's and sequence motifs (D); and a normalized heatmap for available datasets (E).

Comparative genomics of 45 microRNA complements

Figure 6

Figure 7: The evolution of the 1275 microRNA families across the 45 metazoan species currently annotated in MirGeneDB



Conserved gains are shown in red; species-specific gains are shown in pink, and losses are shown in blue; and these gains and losses are mapped onto a generally accepted topology of these species rooted between the deuterostomes and protostomes with branch lengths corresponding to gains and losses, respectively. Note though that this topology is largely recovered when just analyzing the gains and losses of the miRNA families themselves as shown by the bootstrap values indicated at the nodes (Supplementary Methods); the only known nodes not recovered are nodes within the placental mammals and Ecdysozoa, primarily due to losses in rodents and nematodes, respectively

SciLifeLab
NCMM Centre for Molecular Medicine Norway Nordic EMBL partnership for Molecular Medicine
elixir NORWAY
friedländer group
Dartmouth
NASA
erc
Vetenskapsrådet