Cell Genomics, Volume *3*

Supplemental information

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

Sinan Uğur Umu, Vanessa M. Paynter, Håvard Trondsen, Tilo Buschmann, Trine B. Rounge, Kevin J. Peterson, and Bastian Fromm

Figure S1

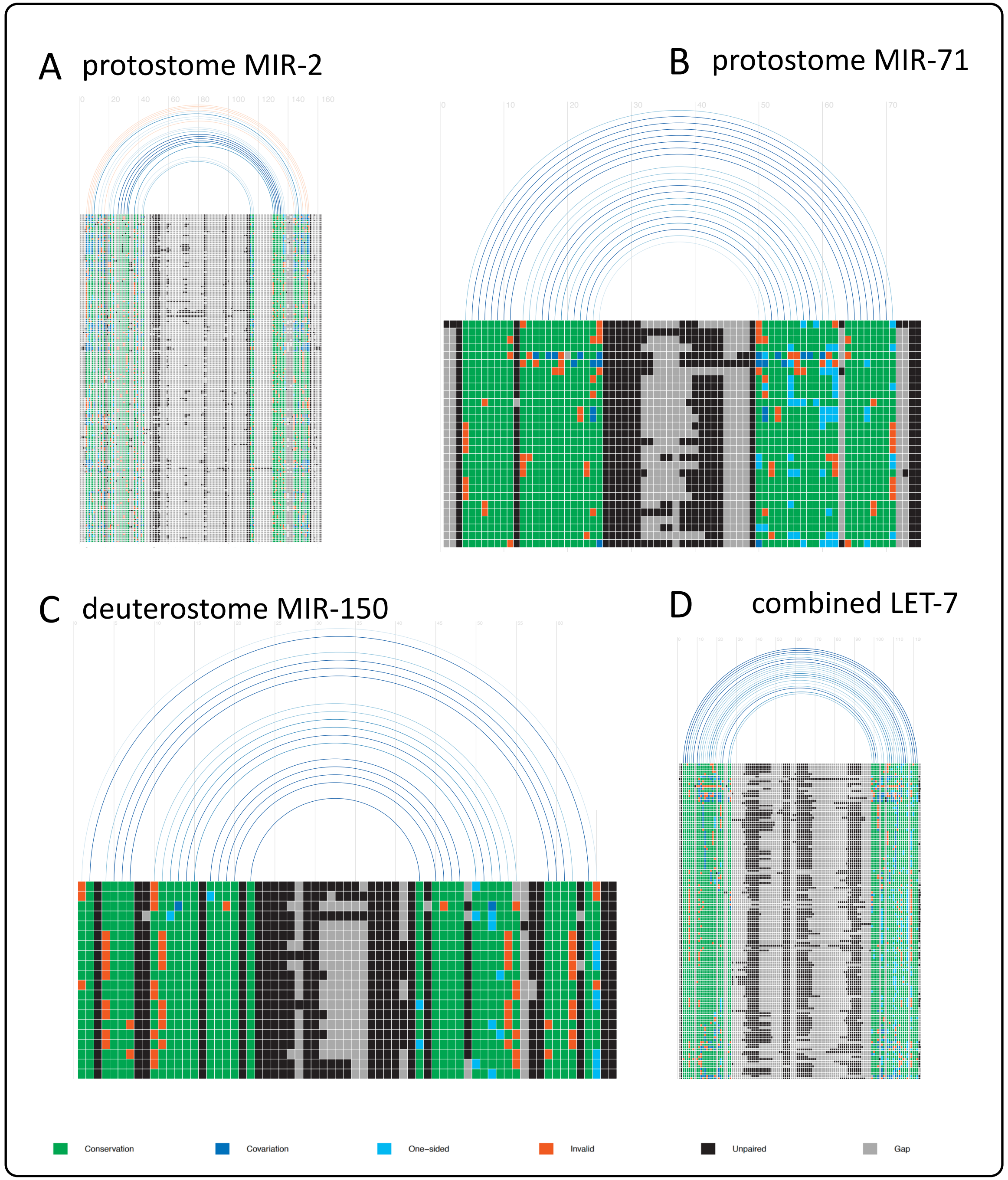
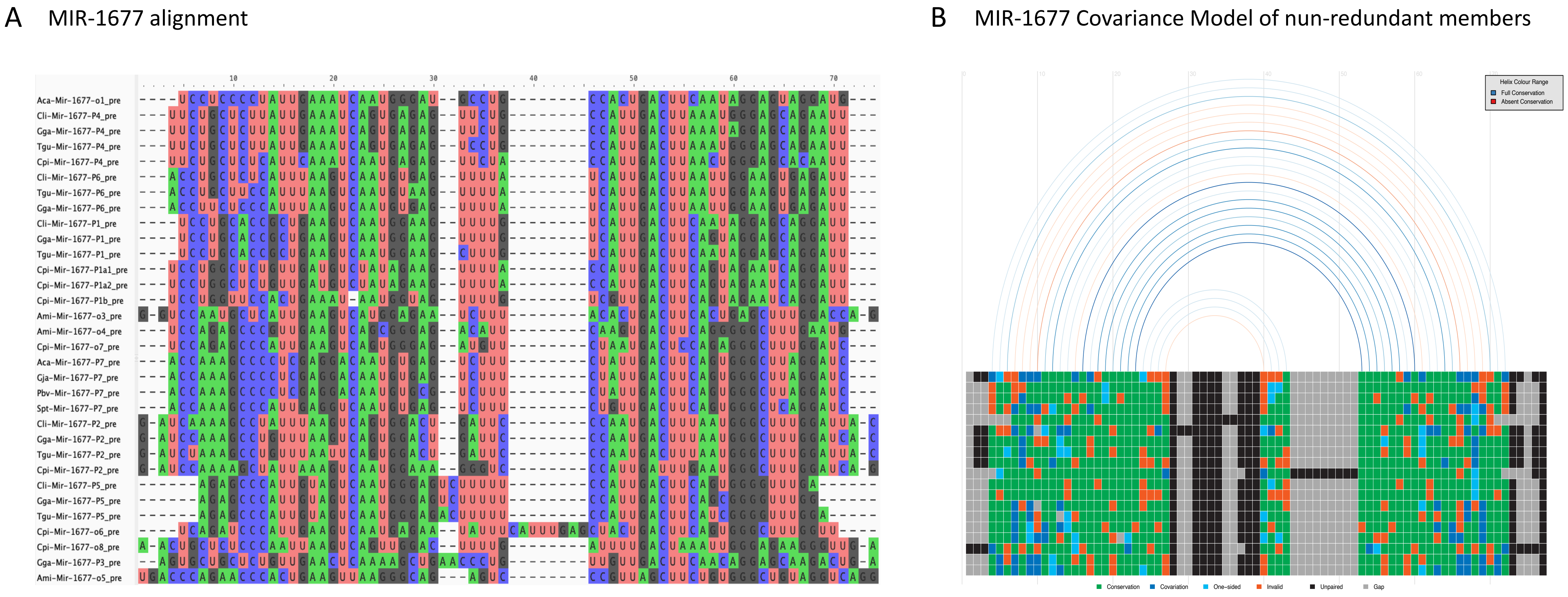


Figure S1: Graphical representation of CMs of four representative microRNA familie, related to figure 2a A) MIR-2, B) MIR-71, C) MIR-150, D) LET-7. Conserved base pairs are colored in green. Blue indicates a compensatory mutation relative to the green pairs (dark blue for a double-sided mutation, light blue for a one-sided mutation). Non-canonical paired bases are red, non-base-pairing bases are black. Graphical representations of all CMs used by MirMachine can be found on github (10.5281/zenodo.7897616 & https://github.com/sinanugur/MirMachine-supplementary/tree/main/CM_figures).

Figure S2

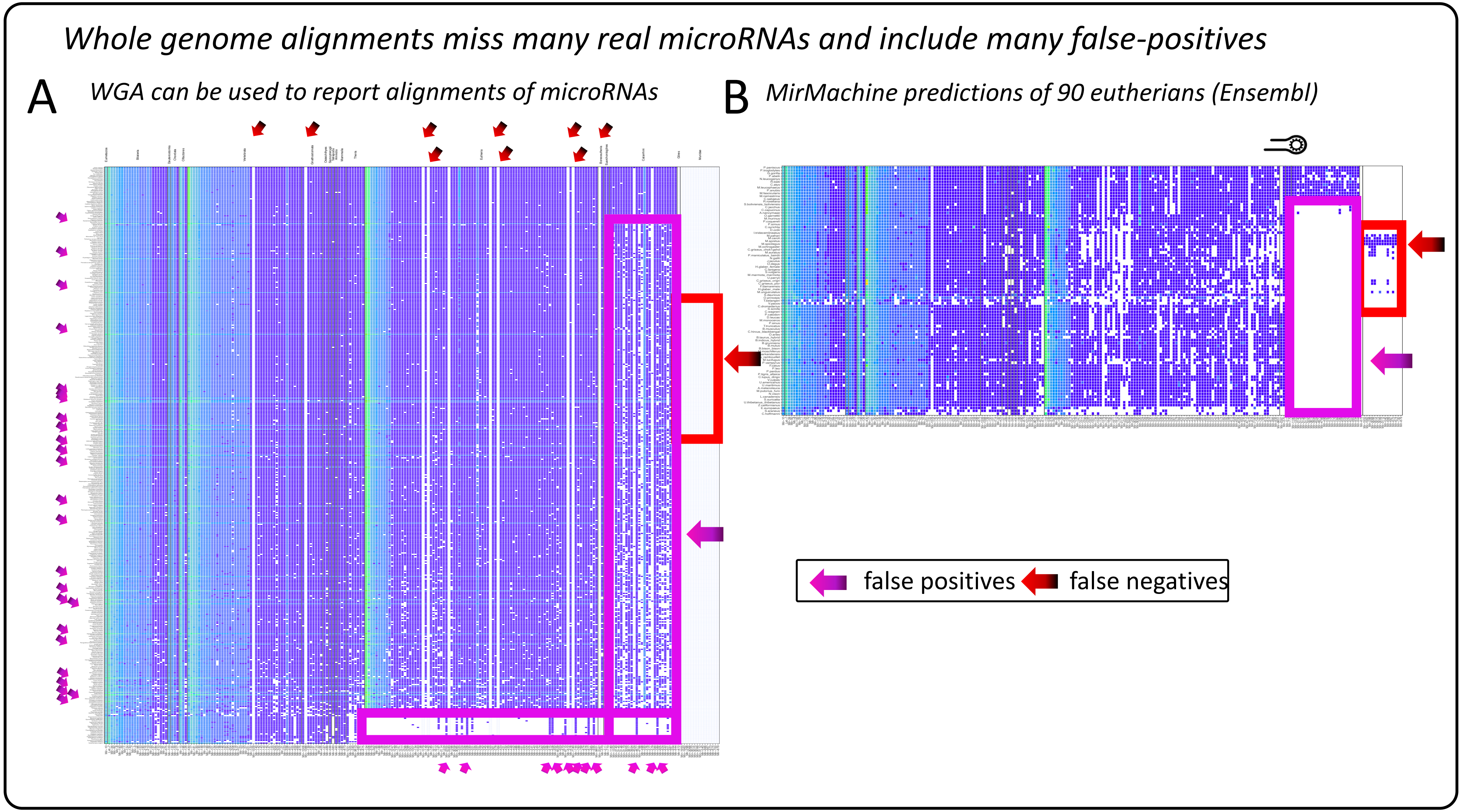


A MIR-1677 alignment

B MIR-1677 Covariance Model of nun-redundant members

Figure S2: MIR-1677 is a highly deviating microRNA family, related to figure 2b. A) Alignment of MIR-1677 genes from MirGeneDB shows low conservation that explains poor performance of B) MIR-1677 CMs in MirMachine.

Figure S3



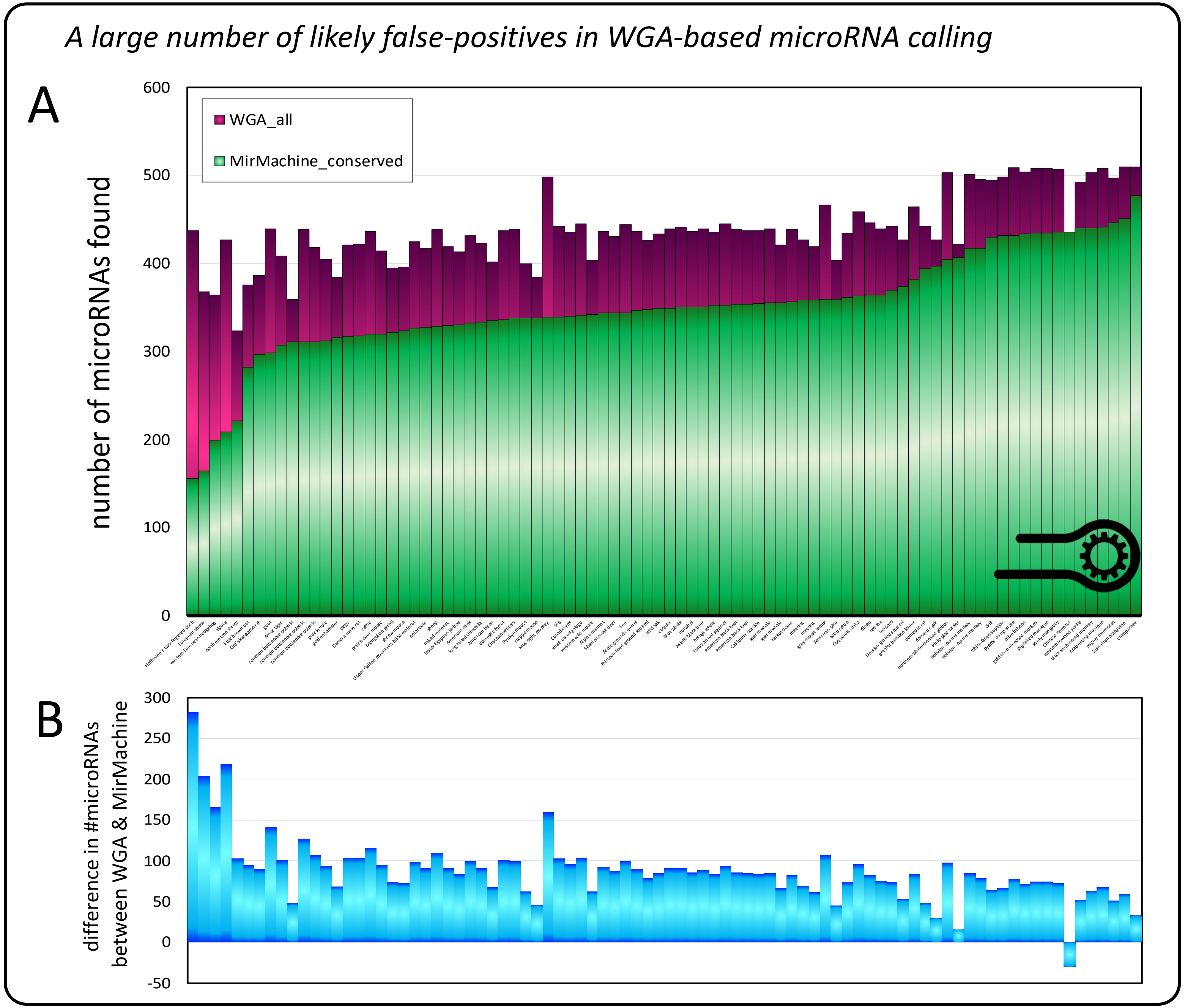
Whole genome alignments miss many real microRNAs and include many false-positives

A *WGA can be used to report alignments of microRNAs*

B *MirMachine predictions of 90 eutherians (Ensembl)*

Figure S3: Whole genome alignments miss many real microRNAs and include many false-positives, related to figure 5a. When comparing overall performance of (A) alignments reported for each of the 470 mammalian species, the overall impression is that many microRNA loci in human are aligned in a majority of mammalian genomes. However, when comparing to the MirMachine output (B), a number of bona fide microRNA families are not reported (red arrows) due to their absence in the human reference (red box: murid microRNA families). Additionally, a high number families and genes that are not expected (pink boxes) given the phylogenetic level of the species (i.e. not Eutherian, not Catharrini) is reported, which seems unlikely to be correct. This also goes for very high number of copies in a number of species (pink arrows left site of A) that would indicate genome duplication, which have not been reported, and likely are false calls.

Figure S4



A large number of likely false-positives in WGA-based microRNA calling

Figure S4: A large number of likely false-positives in WGA-based microRNA calling, related to figure 5a. Comparison of the subset of species from the 470 MULTIZ WGA (A – pink) and our Ensembl based 90 eutherians analysis (A –green). On average, more than 90 false positives are found per genome using WGA (B).

Figure S5

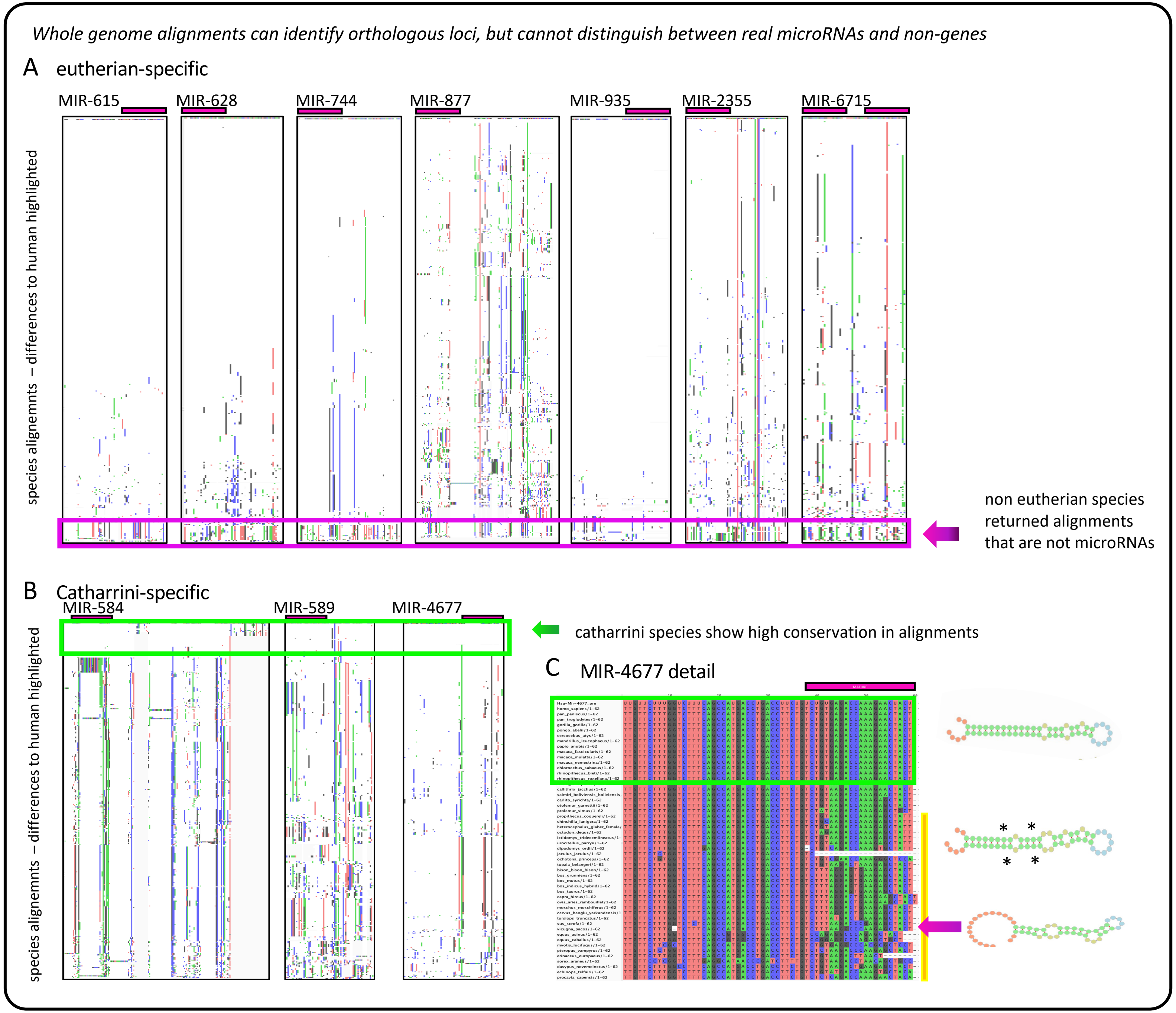
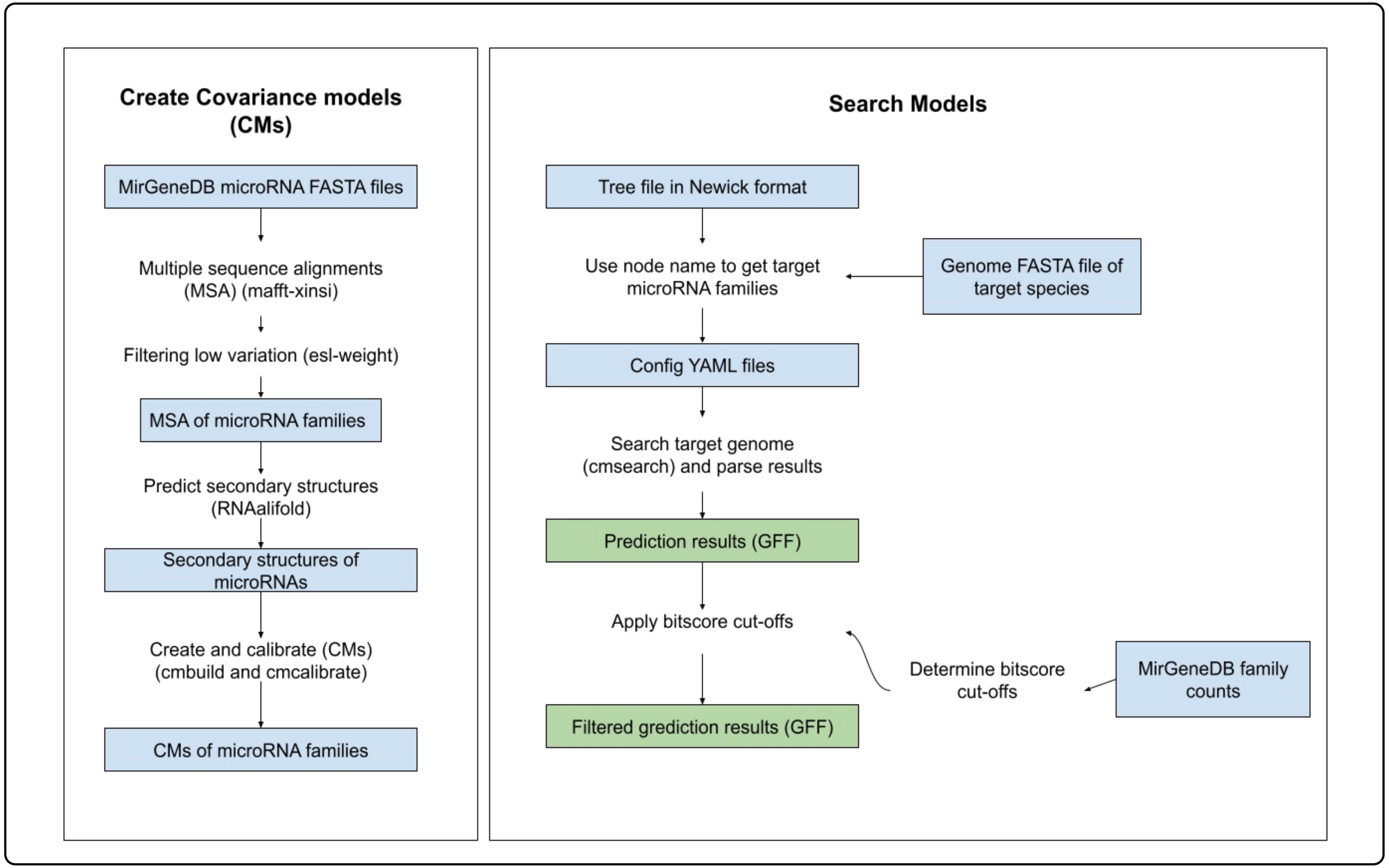


Figure S5: While genome alignments can identify candidates of orthologues loci, they cannot distinguish between real microRNAs and non-microRNA loci, related to figure 5a. Alignments of identified microRNA loci show strong variation especially in loci of species unknown to have the corresponding microRNA. Examples shown for A) eutherian-specific and B) Catharrini-specific microRNA families in non-eutherian and non-Catharrini species shows that, while having alignment reported, there are substantial differences indicating that these are either 1) incorrect alignments or 2) that aligned loci do not contain microRNA genes. In C) (MIR-4677 detail) clear differences in nucleotide composition shows the effect of these sequences on the actual structure of the putative microRNAs clearly ruling out a processing as microRNA. A&B) Each plot highlights the differences to the human reference (white = 100% conserved sites)

Figure S6



Create Covariance models (CMs)
MirGeneDB microRNA FASTA files
Multiple sequence alignments (MSA) (mafft-xinsi)
Filtering low variation (esl-weight)
MSA of microRNA families
Predict secondary structures (RNAalifold)
Secondary structures of microRNAs
Create and calibrate (CMs) (cmbuild and cmcalibrate)
CMs of microRNA families
Search Models
Tree file in Newick format
Use node name to get target microRNA families
Genome FASTA file of target species
Config YAML files
Search target genome (cmsearch) and parse results
Prediction results (GFF)
Apply bitscore cut-offs
Determine bitscore cut-offs
MirGeneDB family counts
Filtered gredction results (GFF)

Figure S6. A summary of MirMachine workflow: high-quality CMs were generated using Infernal based on MirGeneDB v2.1 microRNA families, related to figure 2. Bitscore cut-offs were determined using MirGeneDB to maximize MCC scores. We use the cutoffs to filter out low quality predictions.