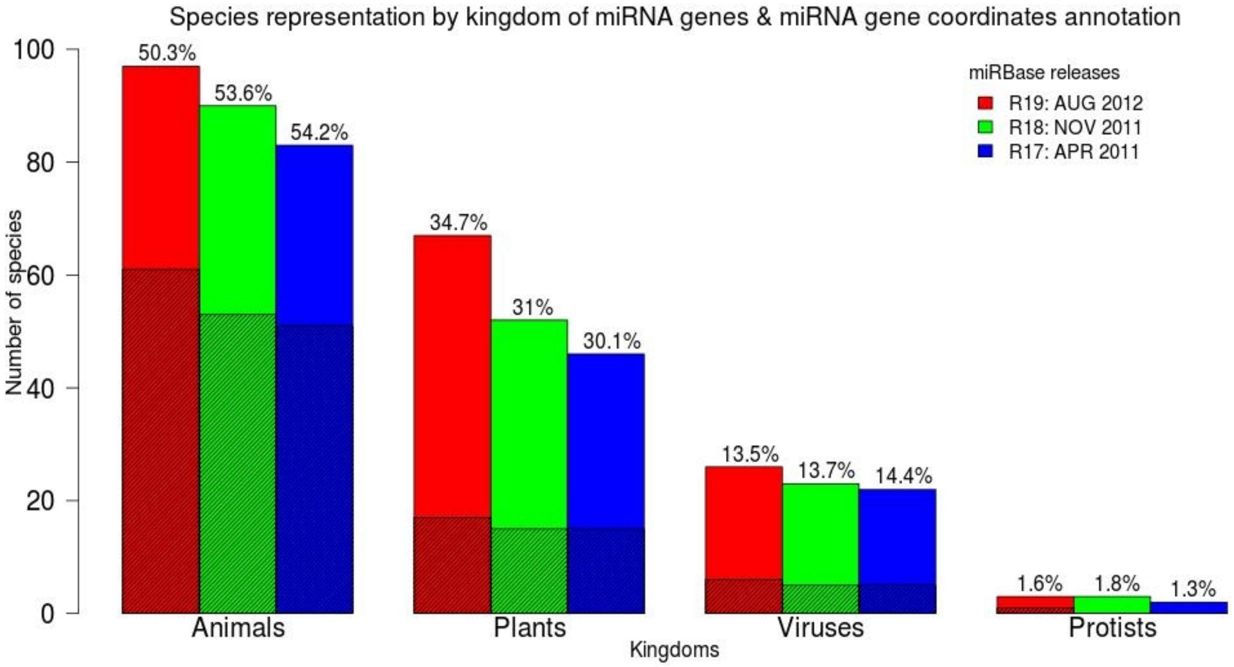
SUPPLEMENTARY FIGURES AND TABLES

Exploration of miRNA families for hypotheses generation

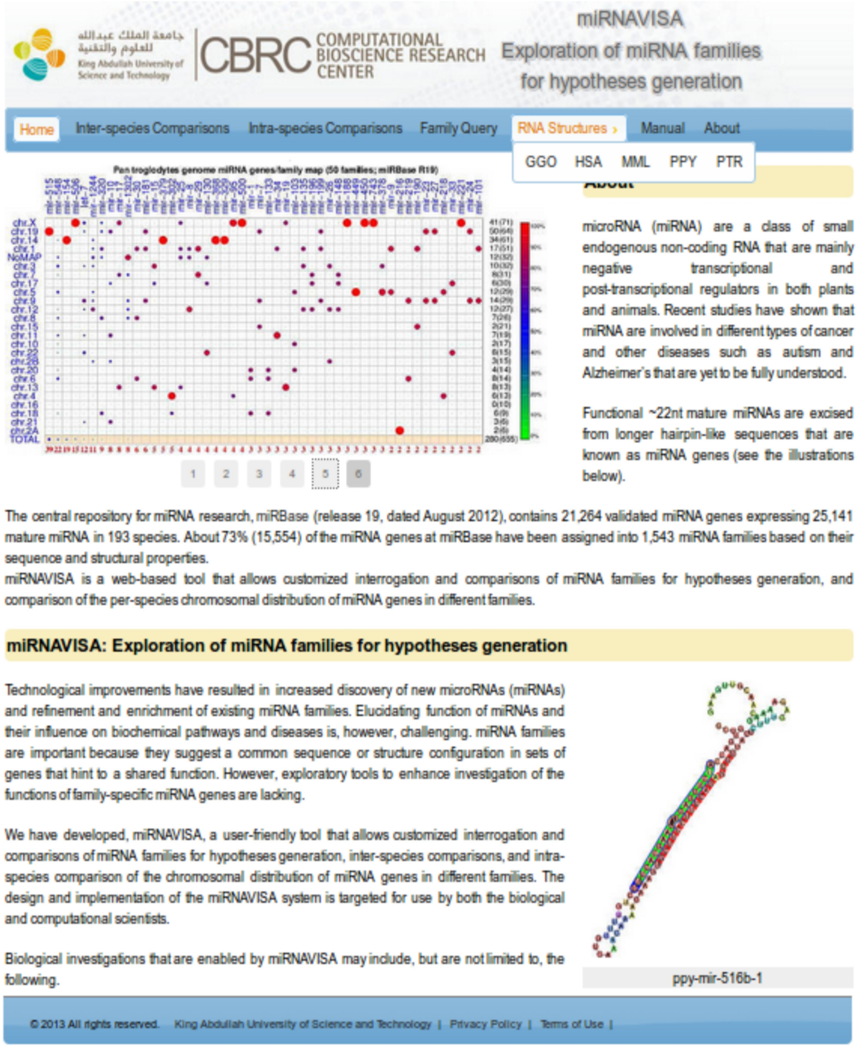
Timothy K. K. Kamanu, Aleksandar Radovanovic, John A. C. Archer, and Vladimir B. Bajic

King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Thuwal, Kingdom of Saudi Arabia

SUPPLEMENTARY FIGURES



Supplementary Figure S1: Species coverage of miRNA genes by kingdoms in the miRBase databases showing the continuing annotation of miRNA gene coordinates. The red, green and blue bars represent miRBase R19, R18 and R17 databases which span 193, 168 and 153 species, respectively (see Fig. 1). The relative magnitude of the superimposed-shaded bars indicates the number of species where the chromosomal co-ordinates of miRNA gene have been determined. The lack of chromosomal coordinates does not imply that miRNA genes are not real.



About

microRNA (miRNA) are a class of small endogenous non-coding RNA that are mainly negative transcriptional and post-transcriptional regulators in both plants and animals. Recent studies have shown that miRNA are involved in different types of cancer and other diseases such as autism and Alzheimer's that are yet to be fully understood.

Functional ~22nt mature miRNAs are excised from longer hairpin-like sequences that are known as miRNA genes (see the illustrations below).

The central repository for miRNA research, miRBase (release 19, dated August 2012), contains 21,264 validated miRNA genes expressing 25,141 mature miRNA in 193 species. About 73% (15,554) of the miRNA genes at miRBase have been assigned into 1,543 miRNA families based on their sequence and structural properties.
miRNAVISA is a web-based tool that allows customized interrogation and comparisons of miRNA families for hypotheses generation, and comparison of the per-species chromosomal distribution of miRNA genes in different families.

miRNAVISA: Exploration of miRNA families for hypotheses generation

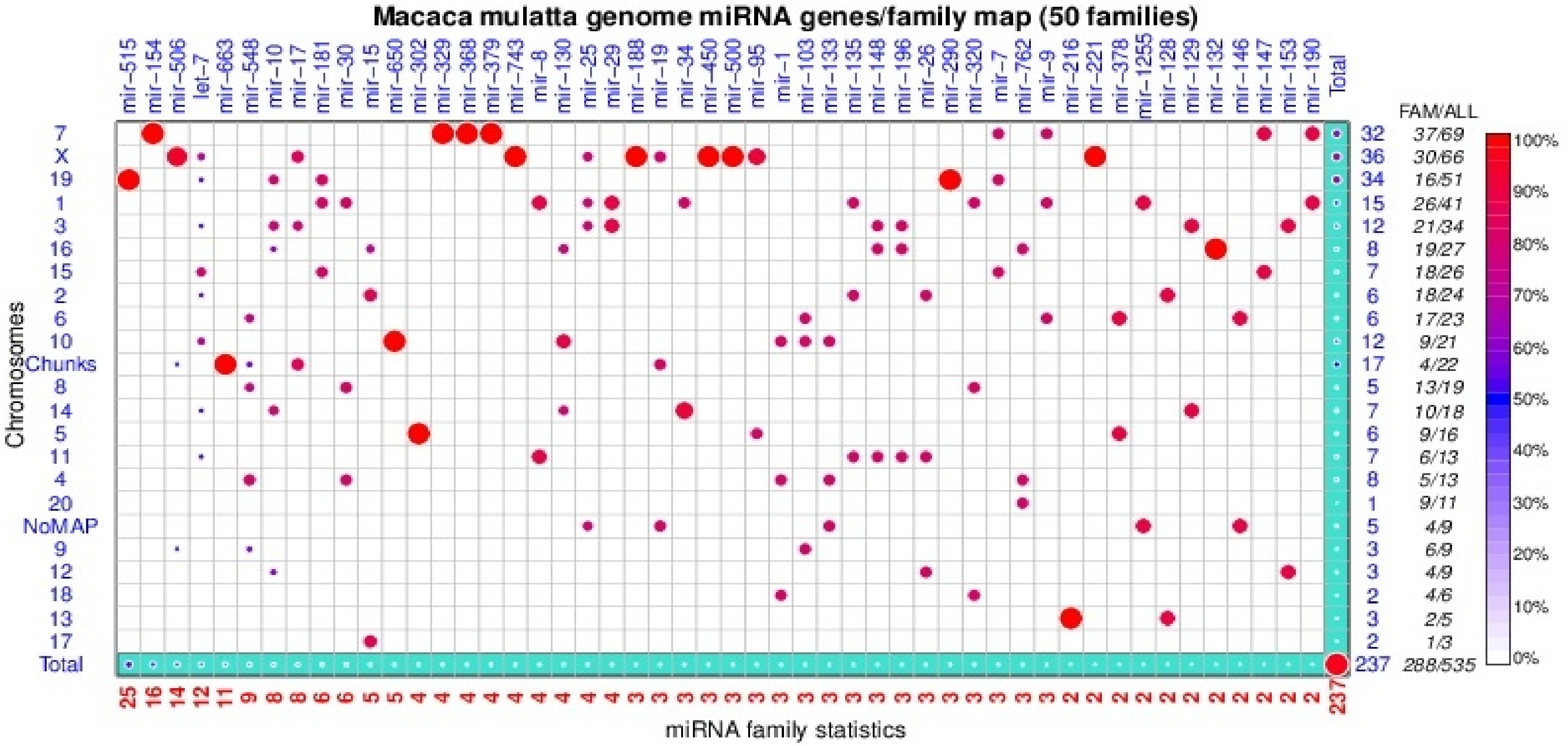
Technological improvements have resulted in increased discovery of new microRNAs (miRNAs) and refinement and enrichment of existing miRNA families. Elucidating function of miRNAs and their influence on biochemical pathways and diseases is, however, challenging. miRNA families are important because they suggest a common sequence or structure configuration in sets of genes that hint to a shared function. However, exploratory tools to enhance investigation of the functions of family-specific miRNA genes are lacking.

We have developed, miRNAVISA, a user-friendly tool that allows customized interrogation and comparisons of miRNA families for hypotheses generation, inter-species comparisons, and intra-species comparison of the chromosomal distribution of miRNA genes in different families. The design and implementation of the miRNAVISA system is targeted for use by both the biological and computational scientists.

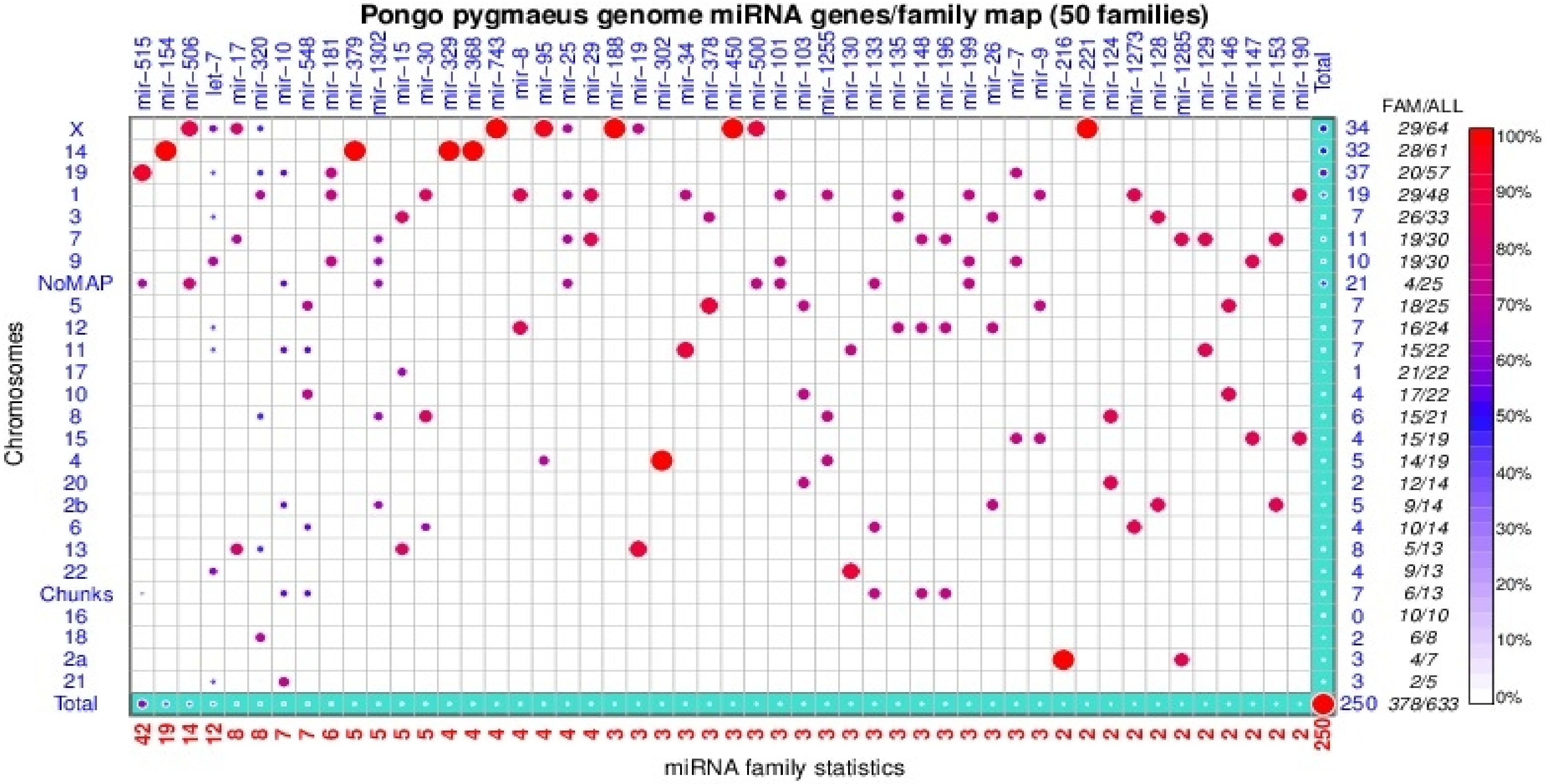
Biological investigations that are enabled by miRNAVISA may include, but are not limited to, the following.

© 2013 All rights reserved. King Abdullah University of Science and Technology | Privacy Policy | Terms of Use |

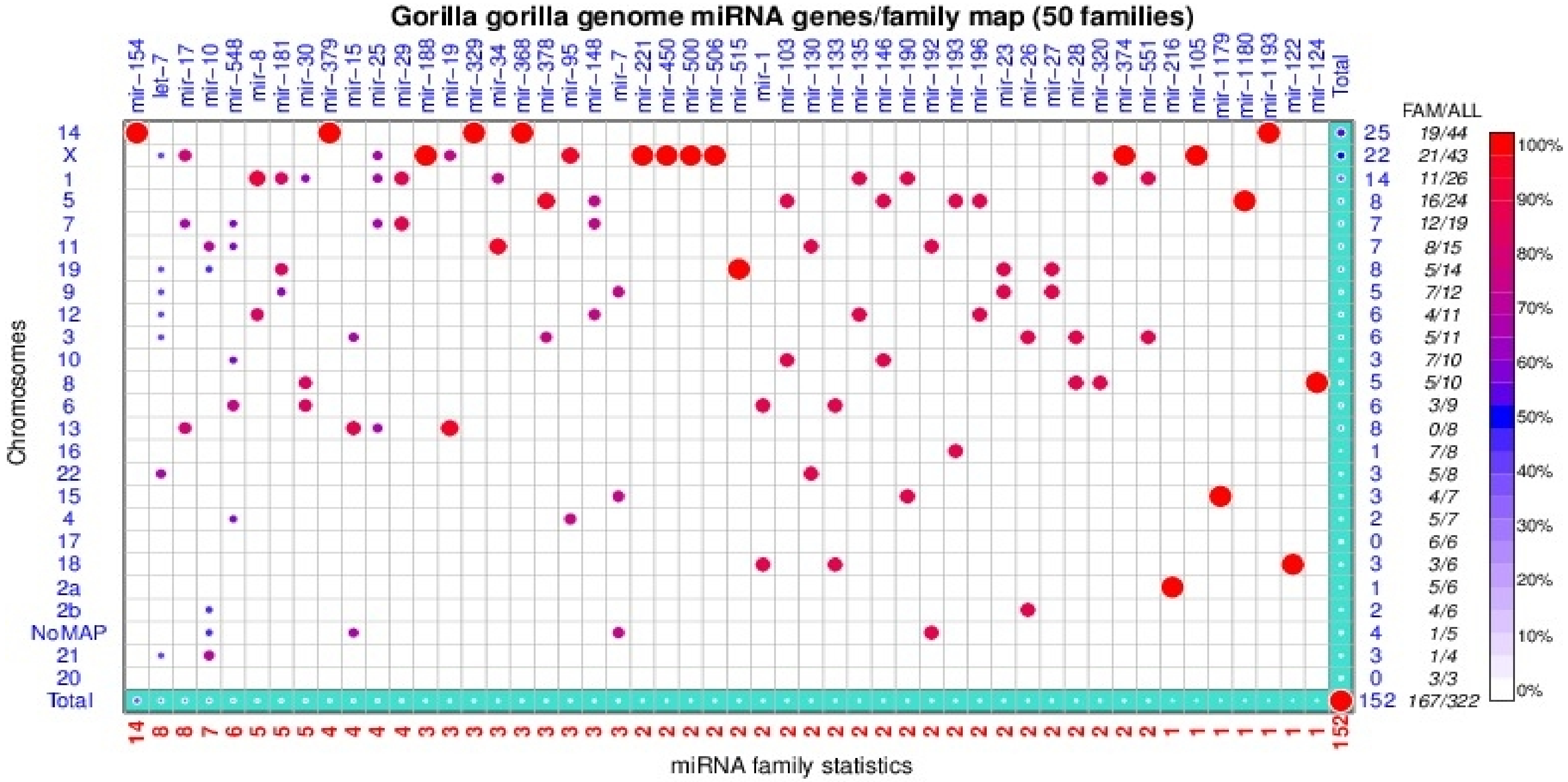
Supplementary Figure S2: miRNAVISA (v1.0) web interface. Three forms are provided for inter-species comparisons, intra-species comparison, and specific keyword queries based on miRBase nomenclature for miRNA genes and their families.



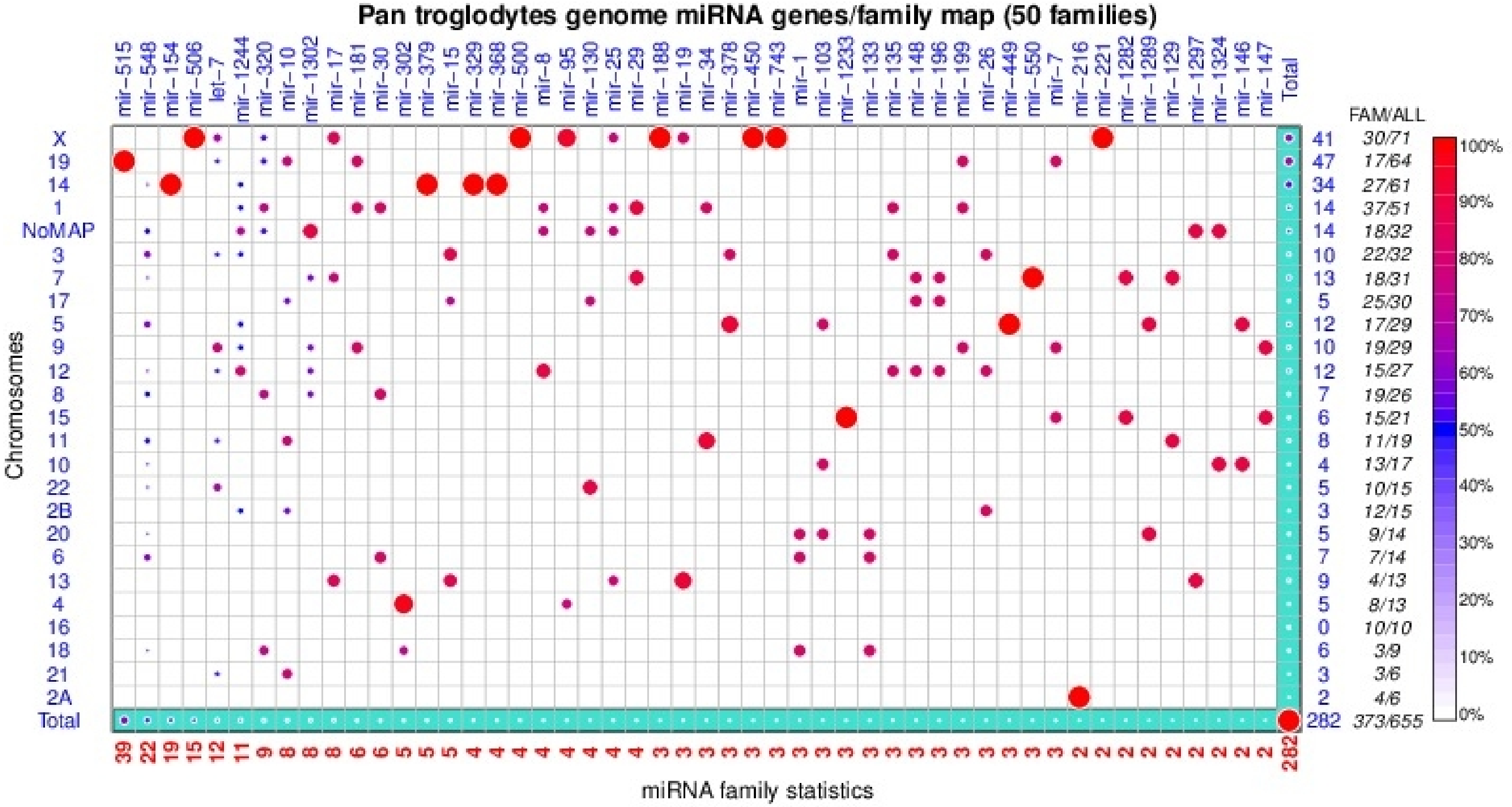
Supplementary Figure S3: Genomic distribution of family-annotated miRNA genes in the monkey (*Macaca mulatta*, MML) genome. The row-wise totals in blue show the number of chromosome-specific miRNA genes (hairpins) in the miRNA families analyzed in the figure. Totals under the “FAM” and “ALL” column labels are the number of miRNA genes in other miRNA families that were not selected for analysis and the total (family-annotated and non-family-annotated) number of registered miRNA genes on a given chromosome, respectively. Genes in miRNA families mir-515, mir-154, mir-663, mir-650, mir-302, mir-329, mir-368, mir-379, mir-743, mir-188, mir-450, mir-500, and say mir-216 are each exclusively co-located on a specific chromosome. A total of 535 miRNA genes are annotated in the MML genome in miRBase R19 and about 98.1% (237 plus 288) of these genes have been classified into different miRNA families. Only 237 family-annotated miRNA genes constituting the largest 50 families are analyzed and summarized in this figure. Other customized miRNA family queries can be specified using the intra-species comparison option given at the miRNAVISA web server.



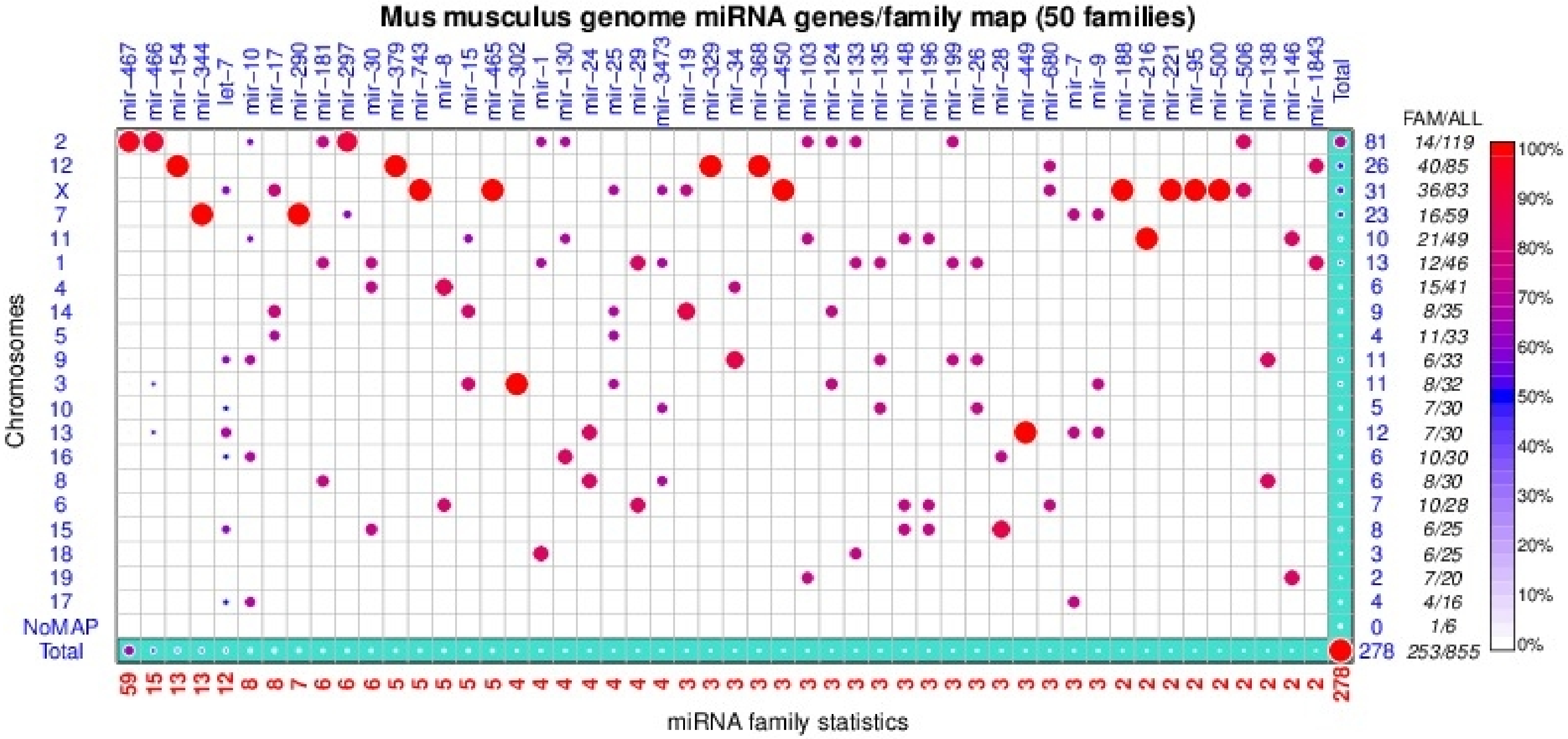
Supplementary Figure S4: Genomic distribution of some family-annotated miRNA genes in the orangutan (*Pongo pygmaeus*; PPY) genome. The row-wise totals in blue show the number of chromosome-specific miRNA genes (hairpins) in the miRNA families analyzed in the figure. Totals under the "FAM" and "ALL" column labels are the number of miRNA genes in other miRNA families that were not selected for analysis and the total (family-annotated and non-family-annotated) number of registered miRNA genes on a given chromosome, respectively. Genes in miRNA families mir-515 (excluding the unmapped miRNA genes), mir-154, mir-379, mir-329, mir-368, mir-743, mir-188, mir-302, and say mir-450 are each co-located on a specific chromosome. All but one of all the 64 miRNA genes found on chromosome X chromosome have been classified into different miRNA families.



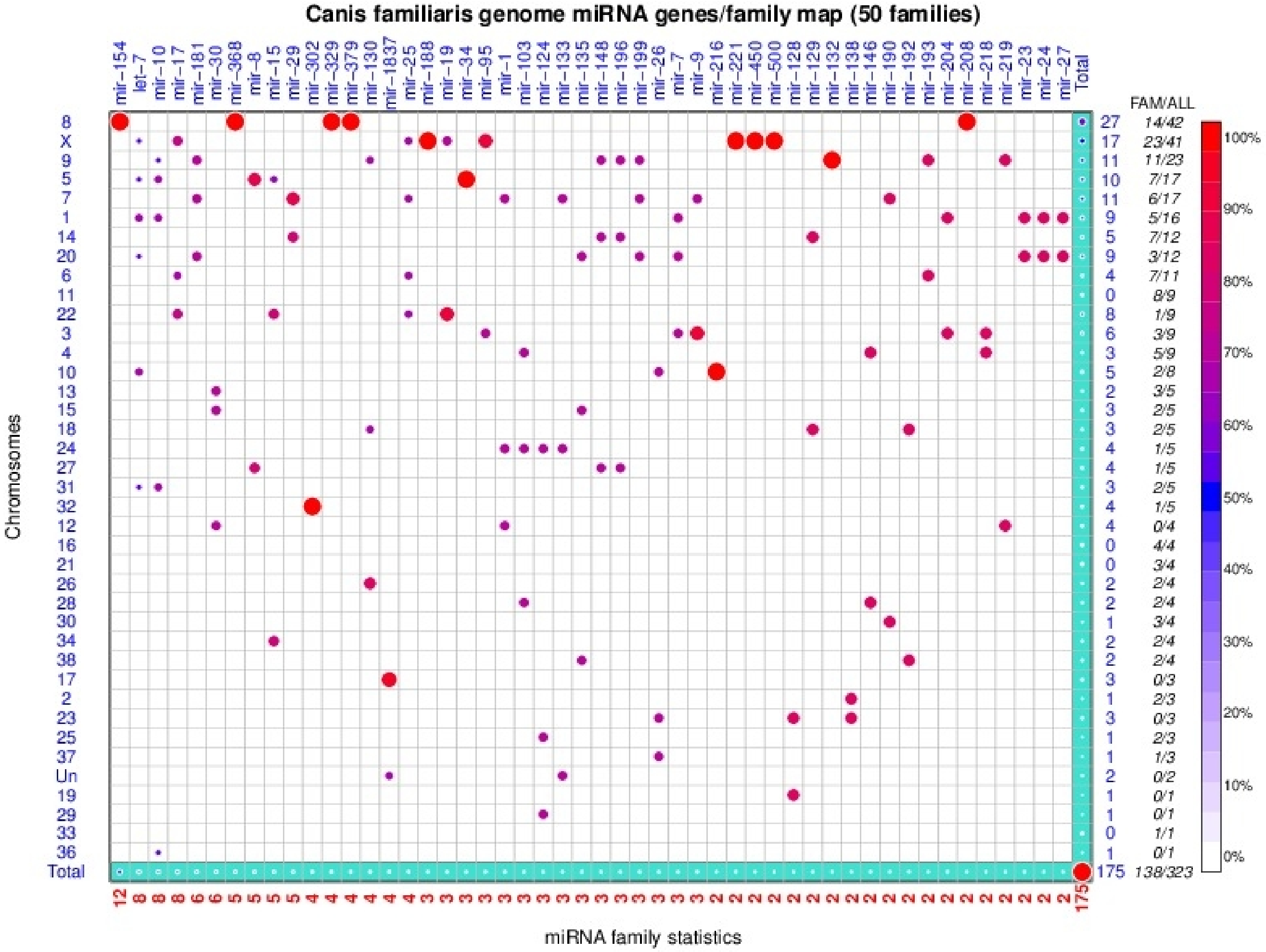
Supplementary Figure S5: Genome-wide distribution of some family-annotated *Gorilla gorilla* (gorilla; GGO) miRNA genes. The row-wise totals in blue show the number of chromosome-specific miRNA genes (hairpins) in the miRNA families analyzed in the figure. Totals under the “FAM” and “ALL” column labels are the number of miRNA genes in other miRNA families that were not selected for analysis and the total (family-annotated and non-family-annotated) number of registered miRNA genes on a given chromosome, respectively. Genes in miRNA families mir-154, mir-379, mir-188, mir-329, mir-221, mir-368, mir-450, mir-500, and say mir-515 are each co-located on a specific chromosome. All the eight known miRNA genes on chromosome 13 of the genome belong to the four miRNA families marked at the respective row.



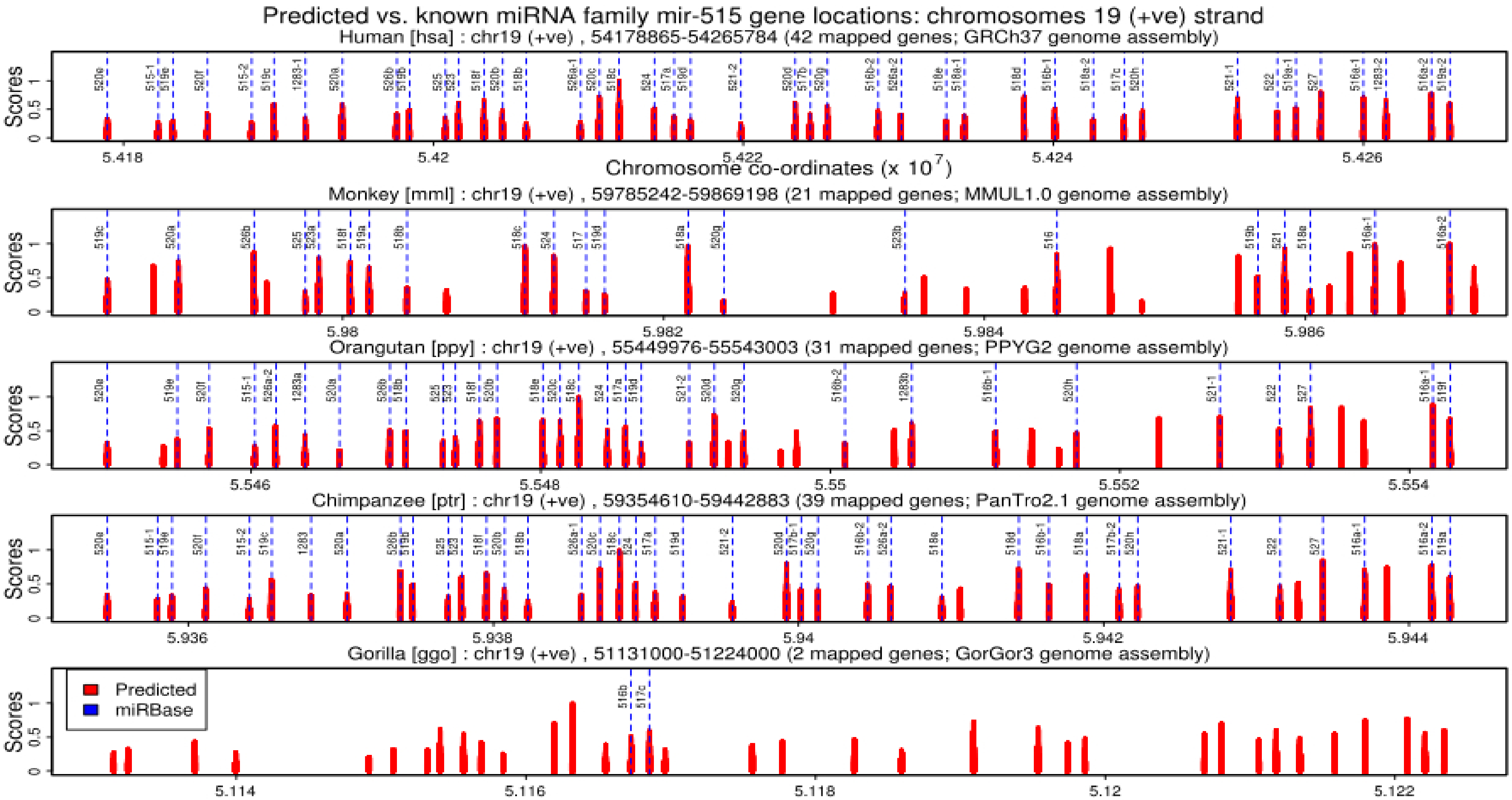
Supplementary Figure S6: Genome-wide distribution of some family-annotated miRNA genes in the *Pan troglodytes* (Chimpanzee; PTR) genome. The row-wise totals in blue show the number of chromosome-specific miRNA genes (hairpins) in the miRNA families analyzed in the figure. Totals under the "FAM" and "ALL" column labels are the number of miRNA genes in other miRNA families that were not selected for analysis and the total (family-annotated and non-family-annotated) number of registered miRNA genes on a given chromosome, respectively. Genes in miRNA families mir-515, mir-154, mir-379, mir-329, mir-368, mir-743, mir-188, mir-302 and say mir-450 each co-located on a specific chromosome.



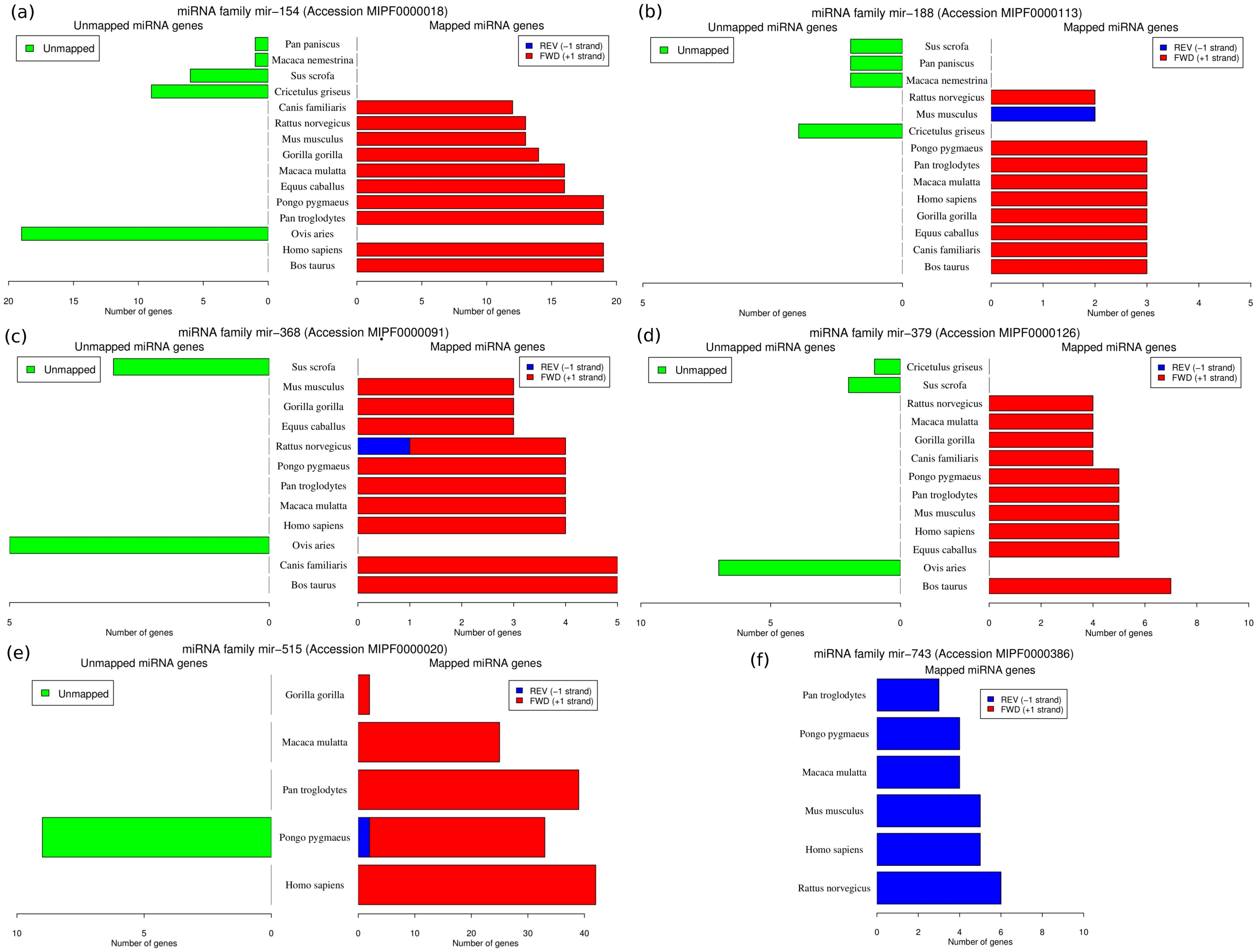
Supplementary Figure S7: Genome-wide distribution of some family-annotated mouse (*Mus musculus)* miRNA genes. The row-wise totals in blue show the number of chromosome-specific miRNA genes (hairpins) in the miRNA families analyzed in the figure. Totals under the “FAM” and “ALL” column labels are the number of miRNA genes in other miRNA families that were not selected for analysis and the total (family-annotated and non-family-annotated) number of registered miRNA genes on a given chromosome, respectively. Genes in say miRNA families mir-154, mir-379, mir-743, mir-302, mir-329, mir-368, mir-450, mir-188 and mir-216 are each co-located on a specific chromosome.



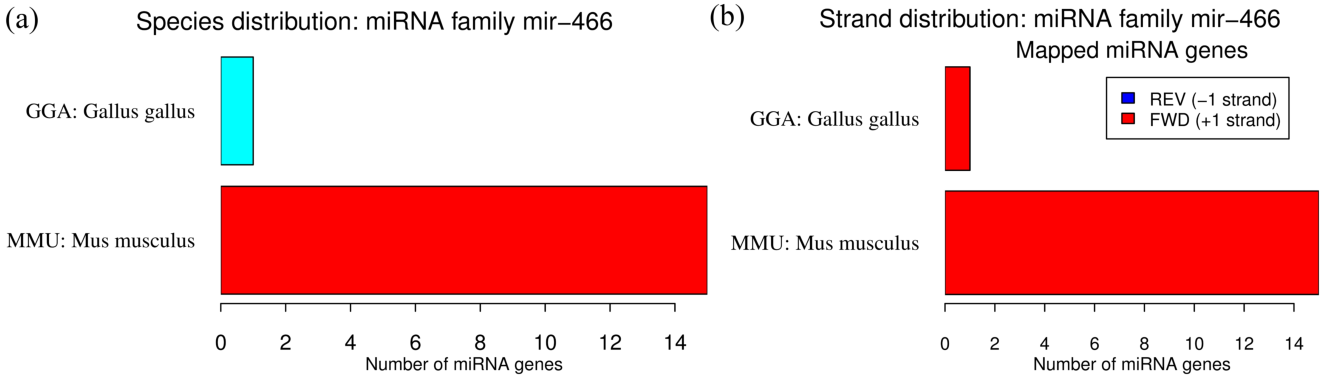
Supplementary Figure S8: Genome-wide distribution of some family-annotated *Canis familiaris* miRNA genes. The row-wise totals in blue show the number of chromosome-specific miRNA genes (hairpins) in the miRNA families analyzed in the figure. Totals under the “FAM” and “ALL” column labels are the number of miRNA genes in other miRNA families that were not selected for analysis and the total (family-annotated and non-family-annotated) number of registered miRNA genes on a given chromosome, respectively.Genes in miRNA families mir-154, mir-368, mir-302, mir-329, mir-379, mir-188, mir-216, and say mir-221 are each co-located on a specific chromosome.



Supplementary Figure S9: Totality of known (recovered) and novel (unregistered) miRNA family mir-515 genes in primate species. The superimposed blue dashed lines show the miRBase gene start co-ordinate of the corresponding genes (truncated gene names) whereas the red bars span the predicted gene sequence. Predicted novel miRNA genes do no have corresponding dashed lines. All miRNA genes were exclusively detected on the FWD (+1) strand with no significant signal in the opposite orientation on the strand and similarly on the negative strand. A majority of the mir-515 family genes are known in human, chimpanzee, orangutan and monkey. Only two gorilla miRNA family mir-515 genes are registered at miRBase. All (42 genes) the human miRNA family mir-515 genes comprise 50% of the family-annotated miRNA genes on chromosome 19 (Figure 3). The miRNA family mir-515 genes have an average intergenic distance of 2113 (42 genes), 2315 (29 genes), 2991 (32 genes) and 4183 (21 genes) nucleotides in the human, chimpanzee, orangutan, monkey genomes, respectively. The minimum and maximum intergenic distances vary due to the differences in number of known miRNA genes in each species at within these clusters are 820-6124, 814-6115, and 849-9912 and 846-12517 nucleotides, respectively. Notably, this distance is less than the average pairwise intergenic distance for clustered miRNAs of 3,000bp suggested in Altuvia *et al.* (2005)[34].



Supplementary Figure S10: Strand distribution of miRNA genes in different miRNA families. (a) mir-154, (b) mir-188, (c) mir-368, (d) mir-379, (e) mir-515, and (f) mir-743. The red and blue bars indicate number of miRNA genes annotated to the positive strand [FWD (+1) strand] and the reverse-complementary strand [REV; (-1) strand], respectively. The number of unmapped family-annotated miRNA genes is shown by the green bars. In (e) i.e., the mir-515 family, the blue bars represent two miRNA genes (ppy-mir-518d and ppy-mir-515-2 with accession numbers MI0014998 and MI0015010, respectively) that are mapped to non-chromosomal DNA as elaborated in Supplementary Table S5.



Supplementary Figure S11: Species and strand distribution of miRNA family mir-466 (Accession MIPF0000208) genes determined using "miRNAVISA family query" at www.cbrc.kaust.edu.sa/mirnavisa/family.php. (a) miRNA genes belonging to the mir-466 family are specific to *Gallus gallus* and *Mus musculus*.(Hypothesis H1) (b) miRNA genes annotated to the mir-466 family are located on the positive strand [FWD (+1) strand].

SUPPLEMENTARY TABLES

Supplementary Table S1: miRNA families and their gene distribution within the ~44kb C14MC in the human genome (miRBase R19).

miRNA Family	mir-154	mir-379	mir-329	mir-368	mir-1193	mir-134	mir-299	mir-412	mir-485	mir-541	mir-544	mir-654	mir-668	mir-889
Number of genes	19	5	4	4	1	1	1	1	1	1	1	1	1	1

Supplementary Table S2: Predicted known (recovered) miRNA family mir-515 genes in the *Homo sapiens* genome (GRCh37 assembly) that are used in this study. The prediction method that was used is reported separately as mentioned in the text. Details of the predicted gene sequences, the computational evaluation of support of atleast one experimentally validated (miRBase) family-mir-515-specific mature miRNA, secondary structure information [estimated minimum folding energy (MFE) and the number of base pairs on the stem regions] and inferred registered homologous genes. The estimated secondary structures of the predicted gene sequences are provided at the miRNAVISA web server. All the 42 (100%) known family genes in the genome were recovered but their length was on the average greater than that of their respective sequences deposited at miRBase (see Fig. 3 and Supplementary Fig. S9).

Supplementary Table S3: Predicted known (recovered) and novel (unregistered) miRNA family mir-515 genes in the *Gorilla gorilla* genome (GorGor3 assembly). The prediction method that was used is reported separately as mentioned in the text. Details of the predicted gene sequences, the computational evaluation of support of atleast one experimentally validated (miRBase) family-mir-515-specific mature miRNA, secondary structure information [estimated minimum folding energy (MFE) and the number of base pairs on the stem regions] and inferred registered homologous genes. The estimated secondary structures of the predicted gene sequences are provided at the miRNAVISA web server. All the mapped family genes in the genome were recovered. The length of the predicted sequences was on the average greater than that of the respective sequences deposited at miRBase. A total of 32 novel miRNA genes are reported in this study and their sequences support the presence of one or more experimentally validated mature miRNA gene deposited at miRBase [see Supplementary Figs. S5, S9 and S10].

Supplementary Table S4: Predicted known (recovered) and novel (unregistered) miRNA family mir-515 genes in the *Macaca mulatta* genome (MMUL1.0 assembly). The prediction method that was used is reported separately as mentioned in the text. Details of the predicted gene sequences, the computational evaluation of support of atleast one experimentally validated (miRBase) family-mir-515-specific mature miRNA, secondary structure information [estimated minimum folding energy (MFE) and the number of base pairs on the stem regions] and inferred registered homologous genes. The estimated secondary structures of the predicted gene sequences are provided at the miRNAVISA web server. All the mapped family genes in the genome were recovered. The length of the predicted sequences was on the average greater than that of the respective sequences deposited at miRBase. A total of 14 unknown miRNA gene sequences are reported in this study [see Supplementary Figs. S9 and S10]. Six of latter predictions are offered as high confidence homologous miRNA genes because they support the presence of one or more family-specific experimentally validated mature miRNA deposited at miRBase. However, the remaining eight sequences are offered here as candidate miRNA genes that await experimental validation of the expressed mature miRNA.

Supplementary Table S5: Predicted known (recovered) and novel (unregistered) miRNA family mir-515 genes in the *Pongo pygmaeus* genome (PPYG2 assembly). The prediction method that was used is reported separately as mentioned in the text. Details of the predicted gene sequences, the computational evaluation of support of atleast one experimentally validated (miRBase) family-mir-515-specific mature miRNA, secondary structure information [estimated minimum folding energy (MFE) and the number of base pairs on the stem regions] and inferred registered homologous genes. The estimated secondary structures of the predicted gene sequences are provided at the miRNAVISA web server. All the mapped family genes in the genome were recovered. The length of the predicted sequences was on the average greater than that of the respective sequences deposited at miRBase. A total of 25 new miRNA genes are reported in this study. Nine of latter predicted genes are offered as high confidence homologs since they support the presence of one or more family-specific experimentally validated mature miRNA deposited at miRBase [see Supplementary Figs. S9 and S10]. However, the remaining 16 sequences were predicted from non-chromosomal DNA and this include most of the known but unmapped family genes that are shown in Supplementary Figure S10.

Supplementary Table S6: Predicted known (recovered) and novel (unregistered) miRNA family mir-515 genes in the *Pan troglodytes* genome (PanTro2.1.4 assembly). The prediction method that was used is reported separately as mentioned in the text. Details of the predicted gene sequences, the computational evaluation of support of atleast one experimentally validated (miRBase) family-mir-515-specific mature miRNA, secondary structure information [estimated minimum folding energy (MFE) and the number of base pairs on the stem regions] and inferred registered homologous genes. The estimated secondary structures of the predicted gene sequences are provided at the miRNAVISA web server. All the mapped family genes in the genome were recovered. The length of the predicted sequences was on the average greater than that of the respective sequences deposited at miRBase. Three novel homologous miRNA genes are reported in this study and their sequences support the presence of one or more experimentally validated mature miRNA gene deposited at miRBase.