Cell Genomics Transparent Peer Review Record

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

Sinan Uğur Umu[1], Vanessa M. Paynter[2], Håvard Trondsen[1], Tilo Buschmann[3], Trine B. Rounge[4,5], Kevin J. Peterson[6], Bastian Fromm[2*]

Summary

Initial submission:	Received : 12/6/2022
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 2 Revision invited : 2/27/2023 Revision received : 3/15/2023
Second round of review:	Number of reviewers: 2 Accepted : 5/26/2023
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer #1: The authors develop MirMachine, a tool to find orthologs of known miRNA genes in sequenced genomes. They use MirMachine to efficiently annotate miRNAs in many diverse newly sequenced genomes. Overall, this appears to be an accurate tool from the leaders in miRNA genomics and evolution.

Major concerns:
1) The method is very poorly described—two sentences in the results (lines 132-136), and five sentences in the methods (lines 150-159). It would help if the main text would show and describe a couple representative covariance models (CMs) for miRNA genes. (The CM shown in the graphical abstract does not appear to correspond to a miRNA gene or to the alignment shown in the graphical abstract.) I assume that CMs for miRNA genes would typically have relatively little canonical covariation, since there are typically many different sites in the hairpin at which a compensatory substitution could stabilize a hairpin weakened by a destabilizing substitution. Is this signature of relatively rare covariation in miRNA CMs recognized by the machine learning?

2) The authors do not seem to compare the accuracy of their method to the established method of simply annotating orthologous positions in whole-genome alignments. This comparison should be added. They also emphasize the utility of their method for extinct species—I wonder if this is warranted. Don't all of the extinct species have close paralogs for which miRNA genes could be annotated by using whole-genome alignments? The authors also tout their method as of practical use for assessing genome completeness. Is this true—does their method have overall advantages that would argue strongly for using it over standard metrics (such as the N50 value) that are already in use?

Minor concerns:
1. When first describing the miRNA families, the authors should clarify that their families can differ from the conventional seed families that some miRNA researchers might be assuming.

2. When first mentioning the MCC, the authors should mention (for non-statisticians) that this metric is sensitive to both false negatives and false positives.

3. Line 355. Spell out "FPs"

Reviewer #2: The manuscript by Uğur Umu et al. describes MirMachine, a new software for annotating conserved microRNA complements from genomes. This new tool is shown to be useful to annotate miRNAs in a diverse set of animal species, including extinct ones. Altogether, I find the results in this manuscript convincing and I believe this tool can be useful for researchers interested in annotating miRNAs based on genomes. I have mainly very minor comments:
1. The authors write about animal genomes in general, but miss that non-bilaterian animals might not be annotated efficiently by miRMachine. For cnidarians, miRNA turnover rate seem to be high, meaning that most of the miRNA complement is not conserved between species (see Praher et al. 2021 Proceeding of the Royal Society B 288: 20203169 for Anthozoa and Nong et al. 2020 Nature Communications11: 3051 for Medusozoa) and for poriferan miRNA precursors seem to be quite unusual (see Grimson et al. 2008 Nautre 1193-1197) and not so well conserved between species (see Liew et al. 2016 PLOS One 11: e0153731). This is worth mentioning in 1-2 sentences in order to be more accurate about the generality of the approach for animals.
2. The reason for choosing Capitella teleta as the "protostome model" for testing the program should be mentioned (even if its just a random choice).
3. It should be emphasized in the discussion that this approach is not suitable for predicting de novo miRNAs that lack sequence homology to previously described miRNAs.

4. The argument that such a tool is extremely valuable for annotating miRNAs in extinct animals is convincing. Another argument worth mentioning is that some miRNAs are temporally regulated quite tightly and for some invertebrates obtaining specific life stages might be practically impossible.
5. Figure 4A, let-7 should not appear in capital letters and the names/origin of the sequences on the left side of the multiple sequence alignment could be bigger so one could read them more easily.
6. Figure 5, in panels A and B the titles seem to be in too large fonts as they appear on top of the figures.

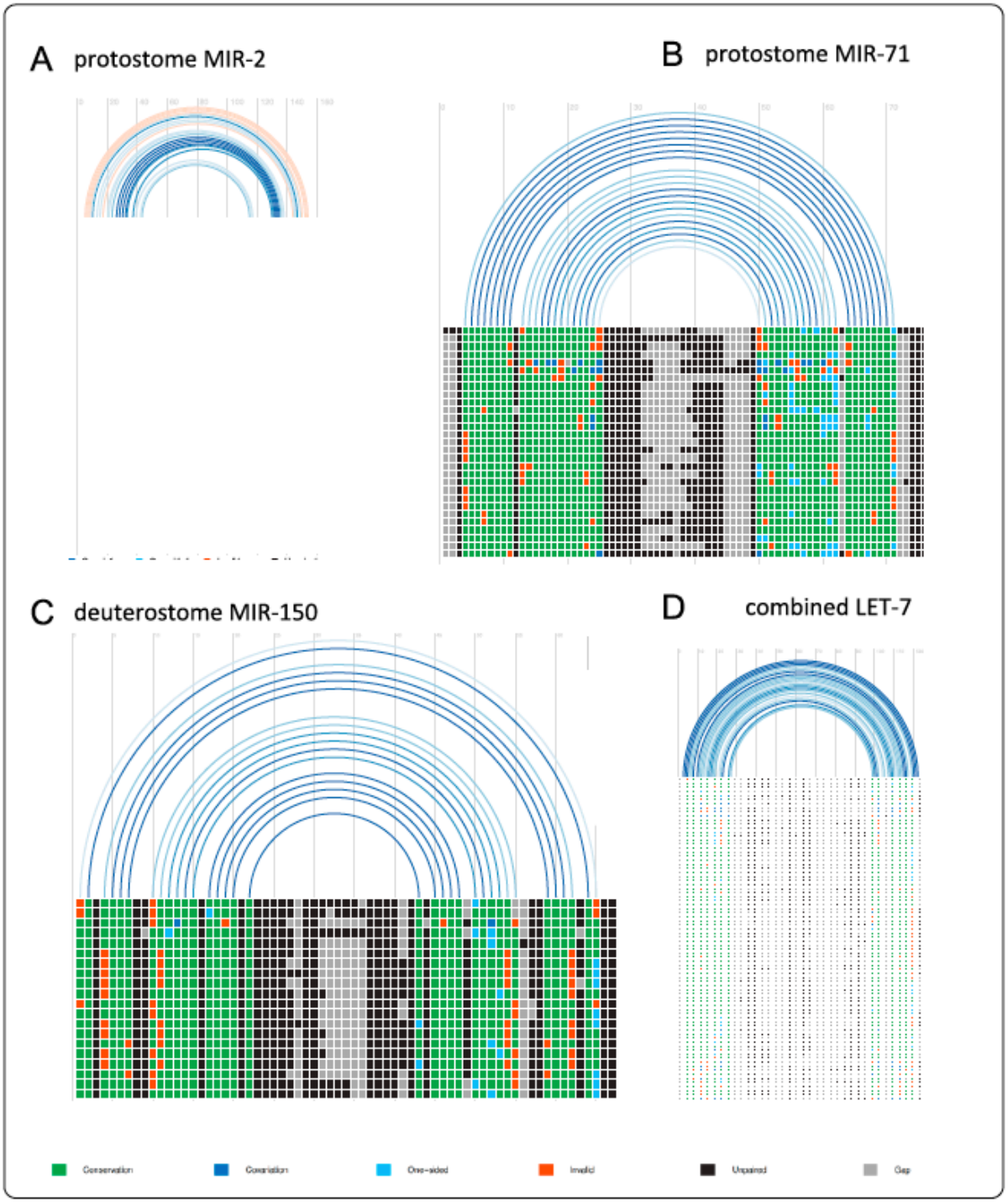
Authors' response to the first round of review

Reviewer #1: The authors develop MirMachine, a tool to find orthologs of known miRNA genes in sequenced genomes. They use MirMachine to efficiently annotate miRNAs in many diverse newly sequenced genomes. Overall, this appears to be an accurate tool from the leaders in miRNA genomics and evolution.
Answer: Thank you very much for this constructive and overall very positive review. We addressed all points below.
Major concerns:
1) The method is very poorly described—two sentences in the results (lines 132-136), and five sentences in the methods (lines 150-159).

Answer: Thank you for this careful observation. Following your comment, we have substantially expanded the description of our method in both results (L 121-129) and STAR method section (L449-461). We would also like to point out to our Supplementary Figure 3 which contains a graphical representation of the MirMachine workflow and, more importantly, the CM creation pipeline.

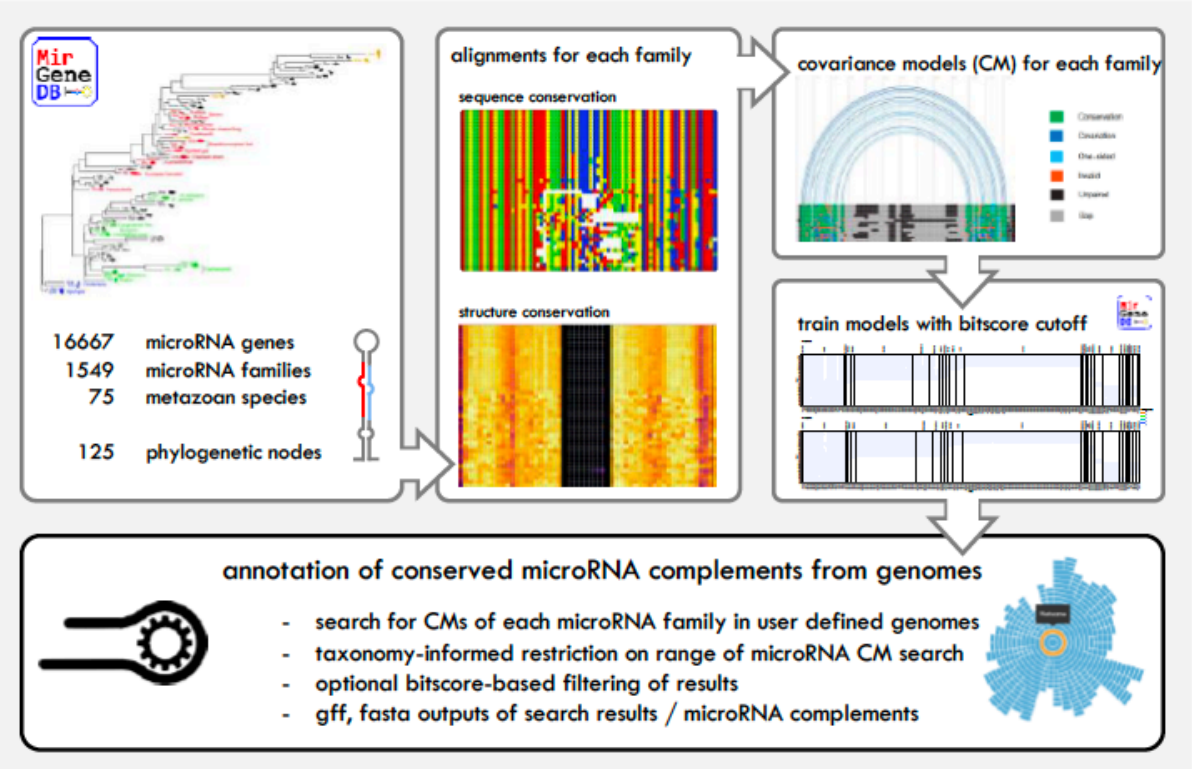
It would help if the main text would show and describe a couple representative covariance models (CMs) for miRNA genes. (The CM shown in the graphical abstract does not appear to correspond to a miRNA gene or to the alignment shown in the graphical abstract.)

Answer: You are right, and we have now created graphical representations of all CMs MirMachine uses. Those are over one thousand pdfs and are available on github (https://github.com/sinanugur/MirMachine-supplementary/tree/main/CM_figures). In addition, we have added Supplementary Figure 1(see below), which includes four representative CM representations that are discussed in the text. It is noteworthy that these models all show covarying sites (Supplementary Figure 1, blue).



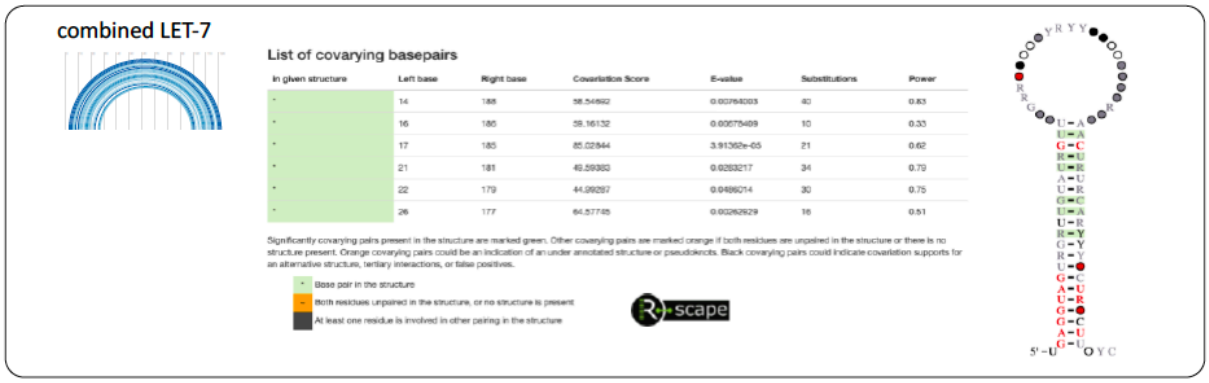
Supplementary Figure 1: Graphical representation of CMs of representative microRNA families. Graphical representations of all CMs used by MirMachine can be found on github (https://github.com/sinanugur/MirMachine-supplementary/tree/main/CM_figures).

We also corrected the graphical abstract that incorrectly contained a CM of a different RNA and now includes a microRNA.



I assume that CMs for miRNA genes would typically have relatively little canonical covariation, since there are typically many different sites in the hairpin at which a compensatory substitution could stabilize a hairpin weakened by a destabilizing substitution. Is this signature of relatively rare covariation in miRNA CMs recognized by the machine learning?

Answer: Thank you very much for the highly appreciated set of points raised. For non-coding RNA search, CMs outperform sequence-based methods and other homology-based methods1. The Infernal tool by default uses structure information which was captured by CMs. States within the CMs capture paired and unpaired regions while allowing insertions and deletions. When studying the actual number of covarying sites in microRNA CMs we find numerous covarying sites. Specifically, using R-scape2, we investigated this issue further on their webserver (eddylab.org/R-scape). As an example, we used the CM of LET-7 and found 6 paired, covarying sites (see Figure below).



Additionally, we have run R-scape analyses on all models and found, on average one covarying site in each CM. All plots can be found on github (https://github.com/sinanugur/MirMachine-supplementary/tree/main/R-scape).

2) The authors do not seem to compare the accuracy of their method to the established method of simply annotating orthologous positions in whole-genome alignments. This comparison should be added. They also emphasize the utility of their method for extinct species—I wonder if this is warranted. Don't all of the extinct species have close paralogs for which miRNA genes could be annotated by using whole-genome alignments?

Answer: Thank you for raising these points and identifying an area where we should have informed in more detail. In the microRNA annotation field, it is very uncommon to base the annotation of microRNAs on genomes only and then, commonly, people use blast or relatively crude RNA family models, and, in fact, no reliable in silico method currently exists to annotated microRNA complements from genomes only. Having said this, whole-genome alignments are usually not available for non-model organisms and even when they exist, will some non-align-able regions contain microRNAs. We have introduced these issues - as an actual motivation for the project - in L 82 -98. In the case of extinct species whole genome alignments are more common, but in the specific case of the Mammoth, the alignment extant genome of the elephant does not contain microRNA annotations.

The authors also tout their method as of practical use for assessing genome completeness. Is this true—does their method have overall advantages that would argue strongly for using it over standard metrics (such as the N50 value) that are already in use?

Answer: Thank you for highlighting this point. Indeed, we had already previously thought that, given their conservation, microRNAs could be a good indicator of completeness of assemblies. As we lay out in the manuscript, our microRNA score does indeed correlate very well with N50 indicating a strong connection of microRNA presence and genome contiguity and possibly completeness (Figure 5B & 5C). Our data clearly suggested that the microRNA score predicts N50, which is a standard measure easy to get from assemblies. However, currently, we did not systematically test the microRNA score against actual completeness measures such as BUSCO scores, or OMark. We will follow up on this in the near future with careful comparisons and will also formulate an actual tool to derive the microRNA scores from assemblies directly, which, however, is outside the scope of this study.

Minor concerns:
When first describing the miR blast or relatively crude RNA family models, the authors should clarify that their families can differ from the conventional seed families that some miRNA researchers might be assuming.

Answer: Thank you for this comment, we now highlight this in L70-71.

2. When first mentioning the MCC, the authors should mention (for non-statisticians) that this metric is sensitive to both false negatives and false positives.

Answer: Thank you for the comment, we added a sentence in L 471-472.

3. Line 355. Spell out "FPs"

Answer: Done.

Reviewer #2: The manuscript by Uğur Umu et al. describes MirMachine, a new software forannotating conserved to annotate miRNAs in a diverse set of animal species, including extinct ones. Altogether, I find the results in this manuscript convincing and I believe this tool can be useful for researchers interested in annotating miRNAs based on genomes. I have mainly very minor comments:

Answer: We are delighted by your review and highly appreciate your kind words. We have addressed all the points you raised below.

1. The authors write about animal genomes in general, but miss that non-bilaterian animals might not be annotated efficiently by miRMachine. For cnidarians, miRNA turnover rate seem to be high, meaning that most of the miRNA complement is not conserved between species (see Praher et al. 2021 Proceeding of the Royal Society B 288: 20203169 for Anthozoa and Nong et al. 2020 Nature Communications11: 3051 for Medusozoa) and for poriferan

miRNA precursors seem to be quite unusual (see Grimson et al. 2008 Nautre 1193-1197) and not so well conserved between species (see Liew et al. 2016 PLOS One 11: e0153731). This is worth mentioning in 1-2 sentences in order to be more accurate about the generality of the approach for animals.

Answer: Despite the fact that we have CMs for both groups, it was indeed worthwhile mentioning how little conservation is observed and how aberrant some of these microRNAs are. We have now added a short section in the discussion L.428-431.

2. The reason for choosing Capitella teleta as the "protostome model" for testing the program should be mentioned (even if its just a random choice).

Answer: Excellent point. We added a section highlighting what we chose C. teleta because of the relatively complete microRNA complement and the lack of species from the same genus in our database L. 156-159

3. It should be emphasized in the discussion that this approach is not suitable for predicting de novo miRNAs that lack sequence homology to previously described miRNAs.

Answer: We highlight this in the discussion L. 418-424

4. The argument that such a tool is extremely valuable for annotating miRNAs in extinct animals is convincing. Another argument worth mentioning is that some miRNAs are temporally regulated quite tightly and for some invertebrates obtaining specific life stages might be practically impossible.

Answer: Really a great point we were surprised to have missed ourselves. We added a corresponding section at the beginning of our discussion L. 369-371.

5. Figure 4A, let-7 should not appear in capital letters and the names/origin of the sequences on the left side of the multiple sequence alignment could be bigger so one could read them more easily.

Answer: In Figure 4 A, we present all microRNAs family members of the LET-7 family, which according to nomenclature rules is spelled in all caps. We have increased the font size of the other microRNAs as requested.

6. Figure 5, in panels A and B the titles seem to be in too large fonts as they appear on top of the figures.

Answer: fixed.

References

1. Freyhult, E.K., Bollback, J.P., and Gardner, P.P. (2007). Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res. 17, 117–125.
2. Rivas, E., Clements, J., and Eddy, S.R. (2020). Estimating the power of sequence covariation for detecting conserved RNA structure. Bioinformatics 36, 3072–3076.

Referees' report, second round of review

Reviewer #1: The authors have satisfactorily addressed some of my concerns. The remaining concerns are listed below.

1) The method is now described much better, with appropriate illustrations of covariation models (CMs). However, there is still a question of whether it is appropriate to call them covariation models, since covariation of paired residues is so infrequently observed. The authors highlight an example with 6 covarying pairs, and observe an average of one covarying site per CM. What is the FDR of these covarying sites? And doesn't an average of one covarying site per CM imply that a large fraction of CM's have no covarying sites? Would it be more appropriate to describe these as "pairing models," rather than covariation models. If the authors want to continue to use the CM terminology, they should state in the main text that

they are using this terminology even though many (or most) of the CMs have no statistically significant covariation.

2, part 1) The authors have not performed the requested analysis (comparing their method to using whole-genome alignments). Perhaps in the "miRNA annotation field" most people do not use whole-genome alignments to find orthologs, but biologists typically do use whole-genome alignments to find orthologs of an miRNA in other species. Moreover, even though whole-genome alignments are not available for most non-model organisms, they are available for hundreds of non-model organisms. Thus, it should be straightforward to compare the results of using a whole-genome alignment to the result of using MirMachine on the same species as in the whole-genome alignment.

2, part 2) In response to my concern the authors say that learning whether the miRNA score is in fact useful for measuring completeness is outside the scope of the current study. Given this uncertainty they should tone down the claims in the paper. For example, they should delete lines 37-38 in the abstract, delete lines 105-106 in the introduction, reword lines 220-221 of the heading to replace the "reveal" and instead simply mention a correlation between miRNA score and genome contiguity, replace "predicts" in line 252 with "correlates with," delete line 379, and delete 388-391.

Minor concerns
1) Most readers will not understand what the authors added in lines 70-71. Perhaps it would be better to wait until the results to add this clarification. For example, at line 121, they could insert a sentence saying something like, "Note that the miRNA families of this analysis sometimes include related miRNAs that have divergent seed sequences and thus are predicted to target different mRNAs."

New minor concerns:
1) The legend to Supplementary Figure 1 does not describe the color key. (What is meant by One-sided or Invalid?)

2) The inserted text at lines 155-158 seems to be talking about two species, but it is unclear what the second species is.

3) Figure 3 panels A,B,D have no key for the colors of the heatmaps.

4) Line 151, don't the results of this section show that the models can be dependent on an individual species?

Reviewer #2: The authors have addressed all my comments adequately and I have no further concerns.

Authors' response to the second round of review

Reviewer #1
The authors have satisfactorily addressed some of my concerns. The remaining concerns are listed below.

1) The method is now described much better, with appropriate illustrations of covariation models (CMs). However, there is still a question of whether it is appropriate to call them covariation models, since covariation of paired residues is so infrequently observed. The authors highlight an example with 6 covarying pairs, and observe an average of one covarying site per CM. What is the FDR of these covarying sites? And doesn't an average of one covarying site per CM imply that a large fraction of CM's have no covarying sites? Would it be more appropriate to describe these as "pairing models," rather than covariation models. If the authors want to continue to use the CM terminology, they should state in the main text that

they are using this terminology even though many (or most) of the CMs have no statistically significant covariation.

Response: We realize now that there is a simple misunderstanding between us and the reviewer concerning our methodology. We use covariance models (CMs), not a model solely based on covariation of nucleotides. Specifically, we used 'Infernal' (ref 1)1 that creates multi-sequence alignment-based 'covariance models'. This term was coined by Eddy and Durbin 1994 (ref 2)2, which 'describe probabilistic models that flexibly describe the secondary structure and primary sequence consensus of an RNA sequence family'. Indeed, this is the standard approach used by Rfam (ref 3)3 to group all RNAfamily entries. Hence, our use of 'CM terminology' is consistent with established protocols in the field and in no way implies any covariation of nucleotides (or the lack thereof). We have added several new parts to the manuscript to clarify this even more.

2, part 1) The authors have not performed the requested analysis (comparing their method to using whole-genome alignments). Perhaps in the "miRNA annotation field" most people do not use whole-genome alignments to find orthologs, but biologists typically do use whole-genome alignments to find orthologs of an miRNA in other species. Moreover, even though wholegenome alignments are not available for most non-model organisms, they are available for hundreds of non-model organisms. Thus, it should be straight-forward to compare the results of using a whole-genome alignment to the result of using MirMachine on the same species as in the whole-genome alignment.

Response: We have now performed the requested analyses and incorporated them as supplementary data in the manuscript. Specifically, we have used the 470 MULTIZ whole genome alignment of 470 mammalian genomes and intersected the human microRNA complement (567 genes; MirGeneDB.org) with the 1.2 TB sized alignment file. We find that the majority of loci indeed produced alignments in most species, but that there was a high number of 1) missing families and genes and 2) a very high number of false positives calls in these microRNA alignments. We show that the latter are either false-alignments or alignments to homologues loci which are clearly not giving rise to microRNA genes. This identification of loci that do show sequence similarity but have no microRNA function could be an interesting avenue for future research on the evolution and pseudogenization of microRNAs.

Furthermore, WGA based approaches aiming at microRNA complement wide analyses require substantial computational resources (500 000 CPU hours for the 470 mammal alignment vs 4000 CPU hours for 90 mammals with MirMachine) and skills (one cannot search for the full microRNA complement on UCSC, but must rather use commandline) and, hence, are not suited - or sustainable - for the standardized annotation of full microRNA complements. We have added a small section describing all these new analyses and 3 new figures (Supplementary Figures 3-5) to the manuscript.

2, part 2) In response to my concern the authors say that learning whether the miRNA score is in fact useful for measuring completeness is outside the scope of the current study. Given this uncertainty they should tone down the claims in the paper. For example, they should delete lines 37-38 in the abstract, delete lines 105-106 in the introduction, reword lines 220-221 of the heading to replace the "reveal" and instead simply mention a correlation between miRNA score and genome contiguity, replace "predicts" in line 252 with "correlates with," delete line 379, and delete 388-391.

Response: We agree with the reviewer that we, at this stage, cannot use "predict" as we have indeed not conducted a more in-depth regression analysis for the microRNA score. Accordingly, we have toned down these statements in the revised version of the manuscript.

Minor concerns

1) Most readers will not understand what the authors added in lines 70-71. Perhaps it would be better to wait until the results to add this clarification. For example, at line 121, they could insert a sentence saying something like, "Note that the miRNA families of this analysis sometimes include related miRNAs that have divergent seed sequences and thus are predicted to target different mRNAs."

Response: We agree and have removed addition in L.70-71 and added in the proposed part this sentence now: "Given the evolutionary microRNA family definition used by MirGeneDB, microRNA families can include nucleotide differences in mature and seed that are captured and summarized in the models." .

New minor concerns:
1) The legend to Supplementary Figure 1 does not describe the color key. (What is meant by One-sided or Invalid?)
Response: We have added the requested text to the legend. One-sided means a mutation that does not necessarily disturb structure, but is therefore, not covarying. Invalid is a proposed non-canonical base pairing (i.e., G-U).
2) The inserted text at lines 155-158 seems to be talking about two species, but it is unclear what the second species is.
Response: Indeed, although we discuss human and the polychaete annelid Capitella, we did not specify the species (C. teleta) but will rectify that in a resubmission.
3) Figure 3 panels A,B,D have no key for the colors of the heatmaps.
Response: Thank you for this comment. Those are added.
4) Line 151, don't the results of this section show that the models can be dependent on an individual species?
Response: This is true as we discuss in this section: Our approach requires microRNAs to be conserved in at least two species. Hence, when we remove one of two species that represent an entire phylum, the microRNA families specific to the phylum can no longer be used for algorithmic training. Further, because our bit score cut-offs are determined based on all analyzed species, the removal of any one species, especially one with a particularly derived sequence composition, despite these microRNAs still being predicted, are now below this newly defined bit score cutoff. In a resubmission, we have changed the current section heading from “MirMachine CMs models are not dependent on individual species” to “MirMachine CMs models are largely independent of any single species” to help clarify this section, and we thank the reviewer for helping us clarify this important point.

Reviewer #2
The authors have addressed all my comments adequately and I have no further concerns.

Response: Thank you.

References
1. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–2935.
2. Eddy, S.R., and Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic Acids Res. 22, 2079–2088.
3. Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 49, D192–D200.
4. Mohammed, J., Flynt, A.S., Siepel, A., and Lai, E.C. (2013). The impact of age, biogenesis, and genomic clustering on Drosophila microRNA evolution. RNA 19, 1295–1308.
5. Mohammed, J., Flynt, A.S., Panzarino, A.M., Mondal, M.M.H., DeCruz, M., Siepel, A., and Lai, E.C. (2018). Deep experimental profiling of microRNA diversity, deployment, and evolution across the Drosophila genus. Genome Res. 28, 52–65.
6. Castellano, L., and Stebbing, J. (2013). Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. Nucleic Acids Res. 41, 3339–3351.
7. Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev. 24, 992–1009.
8. Jones-Rhoades, M.W. (2012). Conservation and divergence in plant microRNAs. Plant Mol. Biol. 80, 3–16.
9. Ludwig, N., Becker, M., Schumann, T., Speer, T., Fehlmann, T., Keller, A., and Meese, E. (2017). Bias in recent miRBase annotations potentially associated with RNA quality issues. Sci. Rep. 7, 5162.
10. Langenberger, D., Bartschat, S., Hertel, J., Hoffmann, S., Tafer, H., and Stadler, P.F. (2011). MicroRNA or Not MicroRNA? In Advances in Bioinformatics and Computational Biology (Springer Berlin Heidelberg), pp. 1–9.

11. Meng, Y., Shao, C., Wang, H., and Chen, M. (2012). Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants. RNA Biol. 9, 249–253.
12. Tarver, J.E., Donoghue, P.C., and Peterson, K.J. (2012). Do miRNAs have a deep evolutionary history? Bioessays 34, 857–866.
13. Taylor, R.S., Tarver, J.E., Hiscock, S.J., and Donoghue, P.C. (2014). Evolutionary history of Plant microRNAs. Trends Plant Sci. 10.1016/j.tplants.2013.11.008.
14. Wang, X., and Liu, X.S. (2011). Systematic Curation of miRBase Annotation Using Integrated Small RNA High-Throughput Sequencing Data for C. elegans and Drosophila. Front. Genet. 2, 25.
15. Fromm, B., Billipp, T., Peck, L.E., and Johansen, M. (2015). A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. Annual review of.
16. Axtell, M.J., and Meyers, B.C. (2018). Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. Plant Cell 30, 272–284.
17. Guo, Z., Kuang, Z., Wang, Y., Zhao, Y., Tao, Y., Cheng, C., Yang, J., Lu, X., Hao, C., Wang, T., et al. (2020). PmiREN: a comprehensive encyclopedia of plant miRNAs. Nucleic Acids Res. 48, D1114–D1121.
18. Fromm, B., Domanska, D., Høye, E., Ovchinnikov, V., Kang, W., Aparicio-Puerta, E., Johansen, M., Flatmark, K., Mathelier, A., Hovig, E., et al. (2019). MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Res., 258749.
19. Fromm, B., Keller, A., Yang, X., Friedlander, M.R., Peterson, K.J., and Griffiths-Jones, S. (2020). Quo vadis microRNAs? Trends Genet. 36, 461–463.
20. Fromm, B., Høye, E., Domanska, D., Zhong, X., Aparicio-Puerta, E., Ovchinnikov, V., Umu, S.U., Chabot, P.J., Kang, W., Aslanzadeh, M., et al. (2022). MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res. 50, D204–D210.