Does microRNA seed match conservation guarantee miRNA targeting?

Blaise Li and Hervé Seitz

IGH (UPR 1142 CNRS), 141, rue de la Cardonille, 34396 Montpellier Cedex 5, France

INSTITUT DE GENETIQUE HUMAINE

1) miRNA target prediction

A microRNA (miRNA) is a small non-coding RNA that represses gene expression via complementarity between its *seed* (positions 2–8) and (generally) 3'UTRs of mature mRNAs (its *targets*).
Example: Target region of miR-134 in the 3'UTR of USP9X.

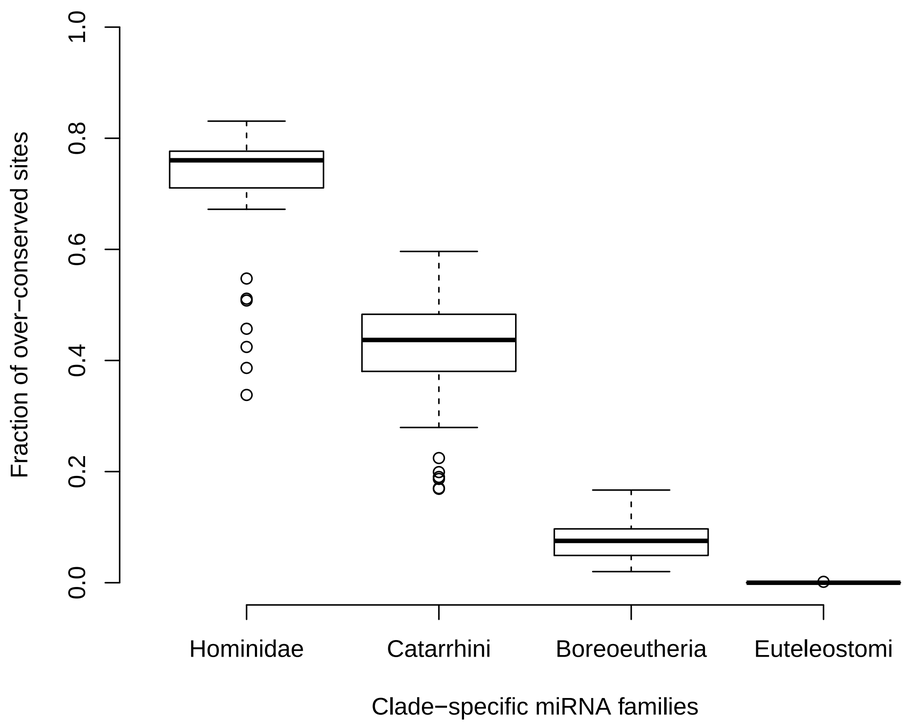
```
5'  ...GACACUGAUUCAGGAUUU CAGUCAC AUCAAG...  3'        USP9X mRNA
                         |||||||
        3' GGGGAGACCAGUUG GUCAGUG U  5'                miR-134
```

Most miRNA **target site prediction** programs (like TargetScan) search for **seed matches that were conserved during evolution**. They typically predict hundreds of targets for a given miRNA, and 60% of the human coding genes are predicted targets for at least one miRNA.

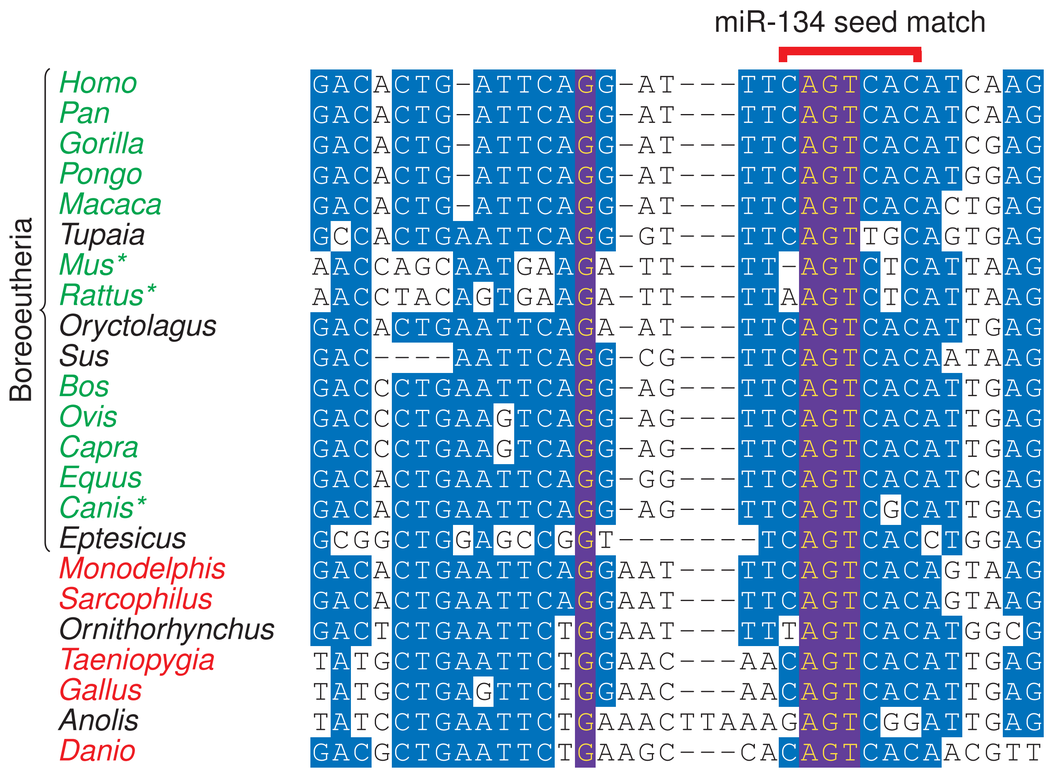
Our goal is to test the assumption that seed match conservation is an appropriate criterion for predicting miRNA targets. →

3) Results

Many miRNA families have seed matches outside their phylogenetic distribution.



Example: Alignment extract from the USP9X miR-134 target region, restricted to species present in miRBase (genus names only, for brevity).



In green, the species having a miRNA with seed GUGACUG in miRBase. The miRNA family is present in 75% of miRBase Boreoeutheria (species with * may have seed matches, but not in this alignment extract).
In red, the non-Boreoeutheria species for which the seed match is conserved. The seed is present in marsupials and birds, and even in *Danio*. miR-134 family is not recorded for these species in miRBase.
Such seed match conservation is not imputable to miRNA targeting.

2) The test procedure

1. Extract seeds from miRBase, together with the species having miRNAs with a given seed (miRNA family↔ species association).
   Example: a miRBase record for a member of the miR-134 family.
   `>ggo-miR-134 MIMAT0002288 Gorilla gorilla miR-134`
   `UGUGACUGGUUGACCAGAGGG`
   Seed and species identifier are in **bold**.
2. Scan the human genome to determine seed match coordinates in 3'UTRs.
3. Use the available UCSC genome alignment (used by TargetScan, 100 vertebrate species) to determine conservation.
   - At each seed match in the human sequence, list the species having the same sequence in the alignment.
4. For each miRNA family, compare miRNA phylogenetic distribution with seed match distribution.
   - Determine the most recent common ancestor (MRCA) of the species having miRNAs in the family.
   - If at least 75% of the descendants of MRCA have such a miRNA, look for seed matches outside that clade (*outgroup* seed matches).

←The outgroup seed matches indicate **over-conserved sites**.

4) Evaluating tests result robustness (in progress)

To see if our results are robust, we can propose some ways to **make the test more stringent**.

1. **Seed matches are expected by chance**. We could require more outgroups with seed matches to count a miRNA family as having over-conserved seed matches.
2. **miRBase is probably not exhaustive**. We could look for missing miRNA orthologs in outgroup genomes.

We are currently implementing the second point. For a given miRNA family, missing members are searched using the following procedure.

1. Partition known miRNA genes using similarity.
2. Build a HMMER profile for each part of the partition.
3. Scan genomes of the species without known miRNA using the HMMER profiles.
4. Check secondary structure predictions of HMMER hits for miRNA-like features (stable unbranched hairpin).
5. Check seed conservation in candidates (seed conservation at positions 2–8 of the mature predicted miRNA).

5) Applications and perspectives

Over-conserved seed matches identified using our protocol could be black-listed from the predictions of TargetScan and similar programs.
The causes for over-conservation of seed matches are unknown to us. It could be interesting to mutate such sites to identify possible functions in outgroups.

6) Tools

Analyses were performed using python, perl, R, bedtools, blast, clustalw, HMMER and various UNIX command-line general usage tools.
A biopython unofficial module to parse UCSC MAF alignment format was corrected as part of this project and is available at:
`https://github.com/blaiseli/biopython/tree/master`
This poster is composed using the baposter LaTeX class:
`http://www.brian-amberg.de/uni/poster/`

7) Funding

cnrs SANOFI HFSP FONDATION RECHERCHE MÉDICALE