**Halictid miRNA analysis**

**Karen Kapheim**

**10 August 2018**

This document briefly describes the results of an analysis of microRNAs expressed in halictid bee brains.

***Associated files***

I have uploaded the files I think will be most useful to you in the google drive `Kocher\_halictid\_miRs`. Other files are stored on my lab-owned server (backed up to tape on a quarterly basis), and can be transferred to you upon request.

1. `halictid\_miRs\_27jun2018.md` This is a complete description of my workflow, including code and supplementary results.

2. `final\_miRsets` This directory contains two subdirectories: ‘fasta’ and ‘miR\_info’.

* `fasta` contains fasta files for the final set of mature miRs in each species. These are in RNA format. If you would like the fasta files for the pre-miR sequences or the star sequences, let me know and I can send those along too. I reran the my MGEN and NMEL miR sets through this step to identify homologs in the rest of the halictids, so those are included too. Let me know if you need any of the other information for mgen and nmel. I have just included them in this section, but I am happy to share previous data too.
* `miR\_info` contains two types of files with information about the final set of miRs in each species.
  + `species\_homologs\_r2.txt` is a tab-delimited file with information about the final set of miRs in each species.
  + `result...species\_homologs\_r2.html` contains similar information, but with important differences. This is an unfiltered set of results output by the mirDeep2 program. It therefore has more miRs than are in the final dataset. The benefit of this file, and the reason I included it, is that it also contains the information about the command that was run and live links to blast each miR. You might find these handy. It also has information about the column names, which you can see if you hover your mouse over them one at a time. You might also find this useful.
* In both sets of files, the ‘*example miRBase miRNA with the same seed*’ field is what I would use to identify homologous miRs across species.

3. `pre-miR\_overlap\_gene\_models` is a directory containing tab-delimited text files that are output from the bedtools intersect function I used to determine where miRs are found within the genome. A summary of these results is below, but you may wish to do further filtering with these files (e.g., proportion of overlapping nts).

4. `target\_predictions` is a directory containing the predicted targets for each of the final miRs in each species. I used two programs for this (miranda, RNAhybrid), and results of each are included. Some people choose to report only targets predicted by >1 programs, but others do not. I included both outputs here so that you can decide for yourself. There are two files for the RNAhybrid results: one is unfiltered and the other is filtered based on p < 0.001 (field 9, “:” delimited file). I included the unfiltered output incase you want to adjust the filtering. I chose p < 0.001, because this produced the most reasonable number of hits (i.e., comparable to those of miranda). For both programs, I specified the minimum free energy < -20.

***Results Summary***

Characterization of microRNAs expressed in the brains of each species/population.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Genomic Location | | | | | Target Prediction | | | |
| Species | Final set of detected miRs | intergenic | exon | intron | 5’ UTR | 3’ UTR | Miranda hits | miRs with miranda predicted targets | RNAhybrid hits | miRs with RNAhybrid predicted targets |
| AAUR | 69 | 41 | 6 | 18(4) | 0 | 0 | 2730 | 69 | 2195 | 69 |
| APAU | 69 | 41 | 1(1) | 21(5) | 0 | 0 | 2398 | 67 | 1684 | 68 |
| AVIR | 75 | 54 | 1 | 16(4) | 0 | 0 | 3418 | 75 | 2749 | 73 |
| HLIG | 102 | 77 | 0 | 19(6) | 0 | 0 | 4466 | 102 | 3001 | 102 |
| HRUB | 84 | 55 | 2 | 23(4) | 0 | 0 | 3027 | 84 | 2762 | 84 |
| LALB\_SOC | 87 | 58 | 1(1) | 24(3) | 0 | 0 | 3493 | 86 | 2643 | 86 |
| LALB\_SOL | 107 | 68 | 2(2) | 30(5) | 0 | 0 | 4157 | 106 | 2549 | 106 |
| LCAL | 77 | 49 | 2(3) | 22(1) | 0 | 0 | 2813 | 77 | 1875 | 77 |
| LFIG | 74 | 53 | 0(1) | 18(2) | 0 | 0 | 2552 | 73 | 2064 | 74 |
| LLEU | 87 | 57 | 1 | 25(4) | 0 | 0 | 3208 | 87 | 1570 | 87 |
| LMAL | 83 | 53 | 1(2) | 21(6) | 0 | 0 | 3338 | 82 | 2001 | 83 |
| LMAR | 69 | 39 | 2(1) | 24(3) | 0 | 0 | 2901 | 69 | 1792 | 69 |
| LOEN | 94 | 57 | 2(1) | 28(6) | 0 | 0 | 3206 | 93 | 1919 | 94 |
| LPAU | 78 | 55 | 0(2) | 17(4) | 0 | 0 | 5155 | 78 | 3148 | 78 |
| LVIE | 66 | 46 | 1 | 19 | 0 | 0 | 2536 | 66 | 1724 | 66 |
| LZEP | 48 | 32 | 2 | 12(1) | 1 | (1) | 1947 | 48 | 1218 | 48 |

\* For genomic location, numbers not in parentheses represent features on the same strand as the pre-miR. Numbers in parentheses indicate strand mismatch.

\* For LZEP, one pre-miR overlapped with a genes on both the same and opposite strands, and is thus counted twice.