Benchmarking procedure: MicroRNA discovery by similarity search to a database of RNA-seq profiles

Sachin Pundhir[1,2], Jan Gorodkin[1,2]

1 Center for non-coding RNA in Technology and Health

2 IKVH, University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark.

∗ **E-mail: Corresponding gorodkin@rth.dk**

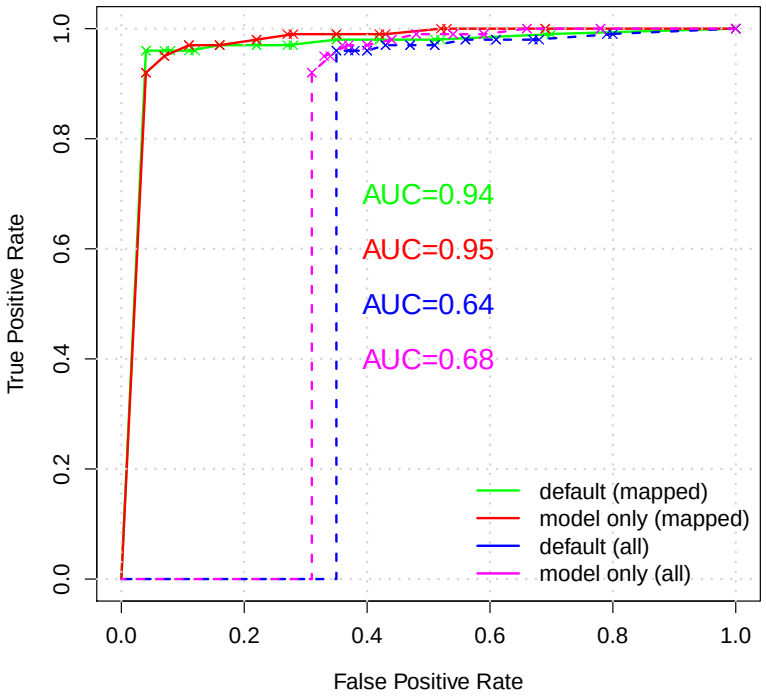
1 Performance of miRanalyzer on the dataset of short reads corresponding to 1,361 ncRNA read profiles

As part of benchmark, we evaluated the performance of an already published tool, miRanalyzer [1] for miRNA prediction. We used the reads corresponding to 1,361 ncRNAs that were used to evaluate the performance of our method as an input dataset to miRanalyzer. We evaluated the performance of miRanalyzer at 11 different posterior probability thresholds (0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00) using two modes

- **Default (miRBase and model)**: In this mode, miRanalyzer made predictions by first mapping reads to known miRNAs from miRBase followed by using random forest model for the remaining set of reads.
- **Model**: In this mode, all the predictions are exclusively based on random forest model.

We first determine all the coordinates corresponding to 1,361 ncRNAs that have a prediction from miRanalyzer at the threshold of 0.0. Next, we consider these set of coordinates as the 'mapped coordinates' corresponding to which miRanalyzer have successfully mapped the reads and have made predictions. In total, we obtained 989 and 1,054 'mapped coordinates' when using default and model search modes of miRanalyzer, respectively. Using the 'mapped coordinates', we then evaluated the performance of miRanalyzer for rest of the thresholds. We determine the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) as follows

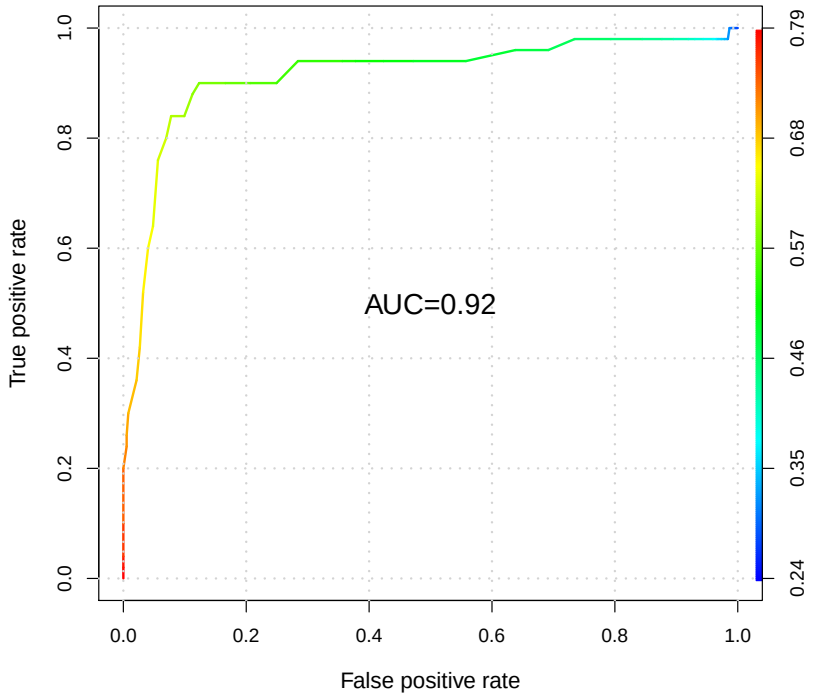
- TP: the total number of mapped coordinates (miRNA) that have a corresponding prediction.
- TN: the total number of mapped coordinates (non-miRNA) that do not have a corresponding prediction.
- FN: the total number of mapped coordinates (miRNA) that do not have a corresponding prediction.
- FP: the total number of mapped coordinates (non-miRNA) that have a corresponding prediction.



Supplementary Figure S1. Performance of miRanalyzer on the dataset comprised of short reads corresponding to 1,361 ncRNAs that were used to evaluate the performance of our method. A high AUC of 0.94 and 0.95 is observed for the set of coordinates (mapped) corresponding to which miRanalyzer made a prediction using default and model search modes, respectively. Also shown is the performance of miRanalyzer evaluated for all the coordinates corresponding to 1,361 ncRNAs irrespective of their mapping status. All the coordinates for which no prediction is made are considered false positives. Here, a low AUC of 0.68 and 0.64 is observed for default and model modes of search, respectively. In this context, our method showed an AUC of 0.93 on this dataset. Please note that for all the four ROC curves, lines have been extrapolated to 0,0 coordinate for completeness. However, the AUC has been computed prior to extrapolation of the curves using the 'Trapezoidal rule' that is commonly used to compute the AUC, such as in pROC [2]

2 Performance of our method based on alignment of miRNA read profiles on the short RNA-seq dataset (GSE10829) used in the original paper on miRanalyzer

We retrieved short Reads from GEO [3] and mapped it against human genome assembly (hg19) using `segemehl` [4] with default parameters. The mapped reads are then analyzed for miRNA prediction based on their alignment to a database of miRNA read profiles (miRRPdb).



Supplementary Figure S2. ROC curve analysis for the prediction performance of our method on an independent dataset that was used to benchmark miRanalyzer in their original study. A high AUC of 0.92 was observed that is consistent with the AUC of 0.93 that we observed on the benchmark dataset comprised of reads corresponding to 1,361 ncRNAs. An AUC of 0.98 has been reported for miRanalyzer on the same dataset.

References

1. Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic acids research 39: W132–W138.

2. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, et al. (2011) proc: an open-source package for R and S+ to analyze and compare roc curves. BMC bioinformatics 12: 77.

3. Barrett T, Troup D, Wilhite S, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets - 10 years on. Nucleic acids research 39: D1005.

4. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, et al. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comp Biol 5: e1000502.