BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Bioinformatics AI, Models and Tools
Title	miR-RF: a machine learning based tool for the in-silico pre-miRNAs annotation/classification
All Authors	Canevotti Alessia(1), Traversa Daniele(1), Chiara Matteo(1)
Affiliation	
(1) Department of Biosciences, University of Milan, Milano	

Motivation

MicroRNAs (miRNAs) are short non-coding RNAs that mediate translational repression through degradation of their target mRNAs. Currently, miRBase and mirGene are the main resources for miRNAs annotation; however, large discrepancies/inconsistencies in the number of annotated miRNAs and their function are observed between the 2 databases. For example mirBase reports 1917 miRNAs genes for homo sapiens, while the equivalent figure for mirGene is 504. Accurate annotation of miRNA genes is crucial for their study and implications in human disorders.
Here, we developed miR-RF, an upgraded ML based method for the in-silico annotation of miRNAs with the following aims: provide an independent assessment of miRNAs annotated in the human genome; study the effect of genetic variants on these miRNAs, in the healthy population and in disease (also identify relevant miRNAs) and study the conservation of human miRNAs/miRNAs families at a broad taxonomic level.

Methods

miR-RF was developed in R v4.1.2, using a Random Forest classifier. The Caret (Classification And REgression Training) package was used to provide machine learning algorithms and training methods. Training was performed by “train” function, using a 10-fold Cross-Validation. miR-RF computes and evaluates a collection of 124 descriptive features (representing RNA hairpin structures, thermodynamic stability and sequence composition) and outputs a binary label for each input sequence: “YES”=miRNA; “NO” otherwise. The method was trained on the reference dataset introduced by Xue et al., and tested on two independent test sets.
miRNAs sequences were retrieved by MirGene (v2.1) and mirBase (v22) databases in fasta format; secondary structures were predicted by RNAfold and processed by custom Python scripts.
To evaluate the potential impact of single nucleotide variants (SNVs) on miRNAs structural stability, each position in a miRNA sequence was mutated in silico. SNVs were scored according to the following categorization: Deactivating, Neutral, Activating and No impact. Identification of candidate orthologous human miRNAs in 10 selected species was performed by NcOrtho. Somatic variants in cancer were assessed by the COSMIC v3.4 db, while gnomAD v4 was used to interrogate standing genomic variation at miRNA loci.

Results

miR-RF showed performances on par or superion compared to other computational methods for the in-silico annotation of miRNAs: accuracy of 96%, sensitivity of 93% and specificity of 98% on a reference benchmark test set.
When publicly available annotations of miRNAs in the human genome were evaluated, miRF classified 93% of miRGene entries as “YES” (total n. 504), while for miRBase this proportion dropped to 63% (1917); these results are consistent with previous reports in the literature- suggesting a large number of inaccurate annotations in miRBase- but could also indicate that a large number of actual miRNAs are not currently included in the mirGene db. Similar patterns were observed when a broader taxonomic sampling was considered, although the difference between the number of miRNAs reported in mirGene and mirBase was less marked. Notably, in miRGene, the conservation Human-Rhesus of “YES” miRNAs was 93%, compared to miRBase, where this percentage decreased to 82%.
Human miRNAs were mutated in silico to evaluate the potential effect of SNVs on their structural stability. The analysis revealed that a significant portion of miRNAs remains structurally stable, and consistently classified as miRNA irrespective of the mutation inserted. SNVs predicted to deactivate miRNAs (miR-RF classification switched from YES to NO) were significantly underrepresented in databases of standing human genomic variation, potentially indicating negative selection in the human population. Interestingly, “deactivating” SNVs showed a positional bias toward the central region of the miRNA hairpin.
According to our analyses, susceptibility of pre-miRNAs to SNV disruption did not correlate with large miRNA families (Fisher test: p-val~0.1), with a slight overrepresentation of vulnerable miRNAs within miRGene families.
miR-RF was then applied to investigate somatic SNVs in miRNAs in cancer. Several known oncosuppressor miRNAs (as miR-433 and miR-1911) were associated with potential disruptive mutations in distinct samples, suggesting a mutation-specific mechanism, which miR-RF can qualitatively assess. Similar patterns of “inactivation” were observed for miR-3655,miR-1182 and miR-664b, not previously linked with carcinogenesis. Remarkably, they tended to be downregulated in cancer, indicating a selective pressure for their downregulation and a potential tumor suppressor activity.
In conclusion, miR-RF is a robust and reliable method for miRNAs annotation and can be used to refine/augment publicly available databases such as mirGene and miRBase. Analyses of somatic SNPs in miRNAs in cancer, demonstrate that our method can be successfully applied to score clinically relevant mutations.

Info	
-	
filename	-
Figure	
-	
Availability	-

Dissemination Material

Social	
-	
Summary	
-	

Corresponding Author

Name, Surname	Alessia, Canevotti
Email	alessia.canevotti@unimi.it
Submitted on	29.04.2024