Integrated computational strategies for annotating microRNA target genes over genome-wide dataset

Corrada D[1,2], Battaglia C[3], Milanesi L[1]

Motivation

MicroRNAs (miRNAs) are small non-coding RNAs that suppress expression and translation of the target genes by binding through a complex interaction to the 3'-UTR of a cognate target mRNA in animals. The prediction of target genes may represent an approach to study the function of the related miRNAs. Due to the structural variability of the RNA duplex, there are no univocal rules for describing the binding modes till now; the lack of these rules constitutes a big issue for a selective identification of the miRNA targets. The large number of predicted targets represents a problem for global analysis and the biological interpretation of the regulatory impact of the miRNAs. The list of putative target genes needs to be characterized according to existing functional annotation systems, such as ontologies or databases of known signaling pathways. Statistical procedures known as enrichment analysis can establish wich annotations are significantly overrepresented. The purpose of this work consists of building a workflow which can generate a pattern of target genes associated with specific miRNAs. Our protocol is intended to assign functional annotation to predicted target list in order to unravel the network of biological functions that are mainly affected.

Methods

The input datasets on which our protocol is based are: a) the mature sequences of miRNAs of interest; b) a whole genome non redundant dataset of 3'-UTR sequences. The procedure proposed here starts by submitting each miRNA to three target prediction algorithms which rely on different assumption: miRanda, TargetScan and RNAhybrid. The management about launching jobs and monitoring them is addressed by Perl script developed for running under dedicated bioinformatic cluster. The output flow will be stored in a relational database, where each entry is accompained by external annotation tables about the genes wich 3'-UTR target belongs. The miRNA binding sites predicted on each 3'-UTR target are further characterized by a scoring function which takes into account several criteria about the position of the site along the sequence and the closeness of other binding sites. Each 3'-UTR target is so ranked by joining together the score of sites which it hosts. Our workflow provides a further refinement of the predicted target

[1] Institute for Biomedical Technologies, National Research Council (ITB-CNR), Segrate [2] Fellowship of the Doctorate School of Molecular Medicine, Università di Milano, Milano [3] Department of Science and Biomedical Technologies, Università di Milano, Milano

list by looking for regions which are really accessible for miRNA by applying PITA target prediction algorithm. We have developed an R package which is composed of functions that are able to query the target rank tables hosted on database and perform statistical tests for the detection of significant associations between target genes and GO terms. These tests are based on the hypergeometric distribution and the Fisher's exact test. In order to carry on a pathway analysis of the predicted targets the package is also suitable for performing association tests with KEGG pathways.

Results

We have taken into account three mature sequences of miRNA which they are supposed to be involved in the processes of cardiac remodelling related to pathological forms of hypertrophy during heart failure. Most of the evidence for the deregulation of miRNAs in cardiac hypertrophy are available for mouse models, so we began by considering the murine variants (mmu-miR-199-3p; mmu-miR-199-5p; mmu-miR-214). The 3'-UTR dataset was obtained by all sequences whose transcripts are annotated on RefSeq (genome assembly NCBI m37 [mm9]). The number of 3'-UTR proposed as target is widely different (2,077 for mir-199-3p; 3,360 for mir-199-5p; 4,944 for mir-214). Only a small portion of these results is commonly predicted by all the algorithms adopted (8.70, 23.5 and 39.7 percent respectively): this observation confirms the benchmarks evidence on prediction methods showing how small variations in the ranking criteria could lead to very different sets of the predicted targets. Filtering targets by the evaluation of binding sites accessibility effectively reduces the number of predictions (408 for mir-199-3p; 692 for mir-199-5p; 2,945 for mir-214). It should to keep in mind that the subset of accessible target doesn't represents the complete panel of real targets, but it almost resembles. All the predicted targets were evaluated for assessing if 3'-UTR/miR interactions are also conserved in other mammals. In particular we have classified the murine predictions as orthologues or non-orthologues with reference to the human ones: this feature may help to highlight which portion of regulatory mechanisms underlying miRNA are conserved among different species. From the analisys of overrepresented annotations related to predicted targets we have found several annotations that are shared among the different miRNA target subsets (GObp: 16 out of 88 found; Gocc: 10/45; GOmf: 12/54; KEGG pathway: 9/44). The evaluation of ontologies suggests that the miRNAs shares common regulation features; an overview of shared pathways from KEGG confirms the literature evidence about the involvement of these miRNA in cancer.

Contact e-mail

dario.corrada@itb.cnr.it