MMRA MicroRNA Master Regulator Analysis

Laura Cantini (✉ laura.cantini@unito.it)

  Enzo Medico Lab (University of Torino, Candiolo Cancer Institute)

Claudio Isella

Consalvo Petti

Gabriele Picco

Simone Chiola

Elisa Ficarra

Michele Caselle

Enzo Medico

Abstract

MMRA is an analysis pipeline that can be used for the identification of microRNAs driver of cancer subtypes. It starts with \(i) a paired microRNA/mRNA expression dataset \(ii) classification of the samples into subtypes \(iii) Subtype-specific mRNA expression signatures, and outputs lists of microRNAs with subtype-specific expression significantly contributing to the expression of subtype signature genes. MMRA consists of four sequential steps, each aimed at progressively reducing the number of candidate microRNAs: \(i) differential expression analysis to highlight microRNAs with subtype-specific expression; \(ii) target transcript enrichment analysis, to further select those microRNAs whose predicted targets are enriched in the associated subtype mRNA signature; \(iii) network analysis, in which an mRNA network is constructed around each microRNA and tested for enrichment in signature genes; \(iv) identification of microRNAs whose expression "explains" the expression of subtype signature genes. The pipeline is available at http://eda.polito.it/MMRA/.

Introduction

In many cancer types, gene expression signatures able to discriminate subgroups of cases with different prognosis or drug response have been proposed. However, the biological mechanisms and regulatory networks underlying these subtypes are mostly unexplained. Here we considered as potential subtype drivers the microRNAs, small non-coding RNAs of $20-22$ nucleotides that bind complementary sequences in target mRNAs and thus reduce their stability and translation rate[1]. Identification of microRNAs driving cancer suptypes requires integrative analysis of paired microRNA-mRNA expression profiles. Recently, integrative computational methods have been proposed to discover microRNA-mRNA interactions possibly involved in tumour development[2,3]. However, these methods have been typically applied to distinguish tumor from normal tissue, a comparison characterized by much wider variation than between two tumor subtypes. Moreover, the methods only take into account microRNA-mRNA interactions supported by anticorrelation, while it has been recently observed that microRNAs can act also indirectly through e.g. regulation of silencing complexes[4]. Finally, the above methods do not prioritize the identified microRNA-mRNA interactions. To overcome all these limitations, we propose the MMRA analysis pipeline, aimed at discovering which microRNAs potentially regulate which cancer subtype, and applied it to colorectal cancer \(CRC). MMRA is subdivided in four sequential steps, each aimed at progressively reducing the number of candidate microRNAs: I. Differential expression analysis to highlight microRNAs with subtype-specific expression. II. Target transcript enrichment analysis, to further select those microRNAs whose predicted targets are enriched in the associated subtype mRNA signature. III. Network analysis, in which an mRNA network is constructed around each microRNA using ARACNe [5], and tested for enrichment in signature genes. IV. Identification of microRNAs whose expression "explains" the expression of subtype signature genes, using Stepwise Linear Regression \(SLR) analysis[6]. An overview of the workflow and of the algorithmic steps is provided in Figure 1.

Equipment

The source code provided with this protocol implements the pipeline introduced in7 and available at http://eda.polito.it/MMRA/. It requires the following freely-available software/libraries: • R \(http://www.r-project.org/) • R packages: preprocessCore, plyr, Matching • Perl • ARACNE: o From http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE download the file aracne.zip o save the file in MMRA_pipeline/codes o unzip aracne.zip

Procedure

1. Perform MicroRNAs differential expression analysis To perform differential microRNA expression analysis, we used a combination of Kolmogorov-Smirnov \(KS) test and fold-change \(FC). For the KS test we used the function ks.boot implemented in the R package 'Matching'[8]. The thresholds for this step are chosen by a permutation-based estimate of the false discovery rate \(FDR), i.e. the estimated percentage of microRNAs identified by chance. For each pair of chosen KS P-value and FC thresholds, the FDR was computed reshuffling 1000 times the samples constituting the microRNA dataset. The mean value of microRNAs significantly differentially expressed in these 1000 experiments was computed and then compared with the number of microRNAs differentially expressed in our step of the pipeline. **2. Perform Target transcripts enrichment analysis** In the second MMRA step, for each microRNA differentially expressed in a given CRC subtype, we performed a target enrichment analysis in the gene signature corresponding to the subtype in which the microRNA was differentially expressed. MicroRNA's target transcripts were predicted following the procedure discussed in Riba and colleagues[9]. To evaluate an enrichment of predicted targets in the signature of the miRNA associated subtype, we calculated a Bonferroni-adjusted Hypergeometric test P-value and the observed/expected \(O/E) ratio. To choose optimal P-value and O/E ratio thresholds, we implemented a FDR computation as follows. **3. Perform network analysis** Network analysis was performed using the ARACNe information-theoretic algorithm for inferring transcriptional interactions[5]. The software was downloaded \ (http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE) and included in the pipeline to infer interactions between each microRNA selected by the previous steps and any mRNA from the paired dataset. For each microRNA selected at the previous steps, data preparation for ARACNEinvolved the setting up of an expression matrix \(X) row-wise combining the entire mRNA expressionTCGA dataset with the expression values of the single microRNA under analysis. To generate a matrix compatible with the standard ARACNE pre-processing steps, we inverted log2 transformation of the expression dataset: naming Xij the elements of the expression matrix previously described, we obtained the called "linear expression matrix" Y through the following operation $Y_{ij}=2^{X_{ij}}$. Then, standardARACNE pre-processing involves quantile normalization of the dataset Y, log2 transformation and filtering of those genes with a standard deviation lower than 1.2. For MMRA the only edges of interest are those connecting the microRNA to mRNAs, therefore the algorithm is run imposing the microRNA as the only hub of the network. The chosen MI P-value significance threshold \(10^{-7}) and bootstrapping P-value threshold \(10^{-12} after 100 bootstrapped networks) are the originally recommended ones[10]. Subsequently, each of the consensus networks constructed around the selected microRNAs \(the "regulons"), is tested for significant enrichment in subtype signature genes respect to a random null model. To this end the Master

Regulator Analysis \(MRA) algorithm is used as previously described[11,12], evaluating the statistical significance \(P-values computed by Fisher's exact test, FET) of the overlap between the "regulon" of each microRNA, and the gene signature of the subtype in which the microRNA was identified as differentially expressed at the previous steps. To assess the sensitivity and specificity of our approach, we built a null model selecting the microRNAs that were expressed \(detected in more than 45 of the 450 samples) but not differential in any subtype of any classifier \(signal to noise ratio – i.e. fold-change over standard deviation – < 0.05). The regulons of the microRNAs constituting the null model were also required to have an intersection with any regulon of the previously selected candidate microRNAs lower than 70%. Then, we chose the MRA Pvalue threshold of the MRApvalues obtained in the null model. **4. Perform stepwise linear regression analysis** In this step, to filter out weak microRNA-mRNA relations within the regulons, MMRA employs stepwise linear regression \(SLR), a procedure previously adopted for transcription factor / target analysis[11,12]. The SLR procedure involved the construction of a linear model for each signature gene, as follows: the log2-expression level of the gene was considered the response variable, and the log2-expression levels of microRNAs linked by ARACNE to the gene were considered as the explanatory variables. Then, a stepwise algorithm is used to select the best minimal set of explanatory variables within the model. Akaike information criterion \(AIC) was used as the stop criterion. The output of SLR was reorganized at the microRNA level, to include, for each microRNA, a list of response variables \(subtype signature genes associated by ARACNE) to which it was associated by SLR. The extent of modulation of a given subtype by a given microRNA can then be estimated as the fraction of signature genes for that subtype whose expression is approximated by the microRNA according to SLR analysis \(positive or negative coefficient). To estimate a significance threshold for this step we considered the distribution of the results for all the selected microRNAs in all the colorectal cancer subtypes. These results are expected to include a small subset of true associations, also selected across the previous steps, and a larger set of random associations. We therefore selected the 90th percentile of the fraction values. To generate the final output of the MMRA pipeline, significant fractions of associated subtype genes are provided only for the microRNA-subtype associations also selected in the previous steps.

Timing

Due to the computational complexity of the network analysis, the pipeline needs to run on a computational cluster. The scripts provided with the ARACNE distribution, are designed for the Rocks 3.2 distribution of Linux and the Sun Grid Engine \(SGE) job scheduler. Thus these scripts use the 'qsub' command to submit jobs to the cluster, and this command must be modified based on the user's job-scheduling software.

Anticipated Results

List of microRNAs with subtype-specific expression providing a significant contribution to the expression of subtype signature genes. An example of MMRA output is shown in7 \(Table 1).

References

1. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281-97 \(2004). 2. Fu, J. et al. Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis. BMC Syst Biol 6, 68 \(2012). 3. Pizzini, S. et al. Impact of microRNAs on regulatory networks and pathways in human colorectal carcinogenesis and development of metastasis. BMC Genomics 14, 589 \(2013). 4. Fabbri, M., Calore, F., Paone, A., Galli, R. & Calin, G.A. Epigenetic regulation of miRNAs in cancer. Adv Exp Med Biol 754, 137-48 \(2013). 5. Margolin, A.A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1, S7 \(2006). 6. Carro, M.S. et al. The transcriptional network for mesenchymal transformation of brain tumours. 7. Cantini, L., et al. \(2015). MicroRNA-mRNA interactions underlying colorectal cancer molecular subtypes. Nature Communications, 6. 8. Sekhon, J.S. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. Vol. 42 \(Journal of Statistical Software, 2011). 9. Riba, A., Bosia, C., El Baroudi, M., Ollino, L. & Caselle, M. A combination of transcriptional and microRNA regulation improves the stability of the relative concentrations of target genes. PLoS Comput Biol 10, e1003490 \(2014). 10. Margolin, A.A. et al. Reverse engineering cellular networks. Nat Protoc 1, 662-71 \(2006). 11. Carro, Maria Stella, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature 463.7279 \(2010): 318-325. 12. Bae, T. et al. Identification of upstream regulators for prognostic expression signature genes in colorectal cancer. BMC Syst Biol 7, 86 \(2013).

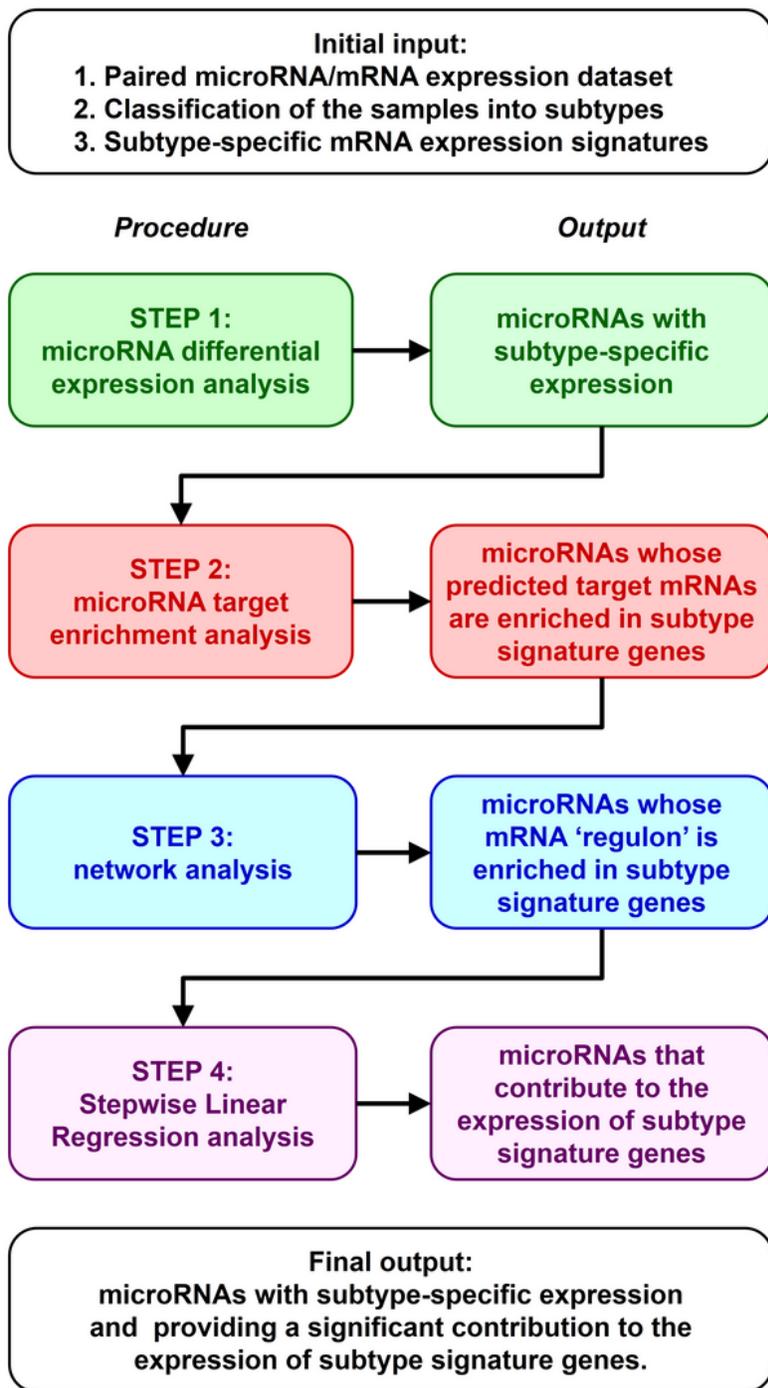# Acknowledgements

Figures

Figure 1

Schematic representation of the microRNA Master Regulator Analysis (MMRA) workflow. The schema reports the data required as initial input, the four analytic steps with the respective outputs, and the final output of the pipeline.