STATISTICAL ANALYSIS

In supervised multivariate statistics, the goal is to model the relationship between input data (samples derived from the affected subjects) and their corresponding labels (WHO MDS classes) by looking for a function depending on several variables (miRNAs) at a time. Such a function should be able to: (1) have good prediction performance on the given data, (2) generalize on previously unseen data and (3) effectively describe the interplay between the measured variables.

The final outcome of a classifier is the prediction of labels associated to a set of input examples. To assess a prediction performance, one should define appropriate metric, such as the accuracy score. Accuracy is defined as the ratio of correctly predicted labels over the total number of analyzed samples.

In this context, regularization methods are a popular class of machine learning techniques that can be expressed as the minimization problem of one *loss* function V that measures the adherence of the objective function to the data and one or more *regularization penalites* R that introduce additional information used to solve the problem: $min_f V(f(X), y) + \lambda R(f)$.
Different choices of V and R lead to different algorithms. The parameter λ controls the trade-off between the adherence of the model to the data and the regularity of the function f. Choosing the regularization parameter appropriately leads to unbiased statistical models (Ambroise and MacLachlan 2002). In the remainder of this section we will illustrate the regularization method used in the present work.

Variable selection with sparse regularization

For multivariate variable selection, we chose l1l2, an embedded regularization method that combines two penalty terms, one enhancing sparsity (l1 norm) and the other retaining correlated variables (l2 norm). The algorithm can be tuned to give a minimal set of discriminative variables or larger sets including correlated ones. A detailed description of the functional and the optimization method is beyond the scope of this paper, we refer to (De Mol et al. 2009, Hastie et al. 2015) for a detailed description of the method.
Assume we are given a collection of n samples, each represented by a d-dimensional vector x of measurements (e.g. the miRNAs expression vector). Each sample is also associated with a binary label y, assigning it to a class (*e.g.,* treated vs. not treated).

The dataset is therefore represented by a $n \times d$ matrix X and a n-dimensional labels vector y. Using only a subset of the given data (learning set), l1l2 looks for a linear model $f(x) = \beta^* x$, whose sign gives the classification rule that can be used to associate a sample to one of the two classes. The classification performance of $f(x)$ is then assessed on the remaining samples (test set) that were not used to build the model.

Note that the vector of weights β^* is forced to be a sparse vector, that is some of its entries are zero, then some variables will not contribute in building the estimator $f(x)$. The weight vector β^* is found in the so-called model selection phase, which consists in selecting the optimal values for the regularization parameters.

Model selection and classification accuracy assessment are performed within two nested cross-validation loops, similarly to (Barla et al. 2008) in order to guarantee an unbiased result. As a consequence of the external loop of cross validation, l1l2 provides a set of B=3 lists of discriminant variables (see Figure 3), therefore it is necessary to choose an appropriate criterion (Barla et al. 2008) in order to assess a common list of relevant variables. We based ours on the absolute frequency, i.e., we decided to promote as relevant variables the most stable genes across the lists. The threshold we used to select the final lists was chosen according to the slope variation of the number of selected genes vs. frequency, its value being 75%. In this way we managed to cut out those variables that were not stable across the cross-validation lists.

PALLADIO

We chose l1l2 method within PALLADIO, a machine learning framework, implemented in Python, based on regularization methods (Barbieri et al. 2016). Specifically PALLADIO aims at providing, along with the relevant variables, an estimate of the predictive performance of a model that takes into account only those variables. The reliability of the result is evaluated by performing two sets of experiments, which we refer to as *regular batch* and *permutation batch*. For each experiment, PALLADIO resamples different learning and test sets a large number of times in order to estimate the distribution of performance scores for both batches. In the analyses performed and described in this work, we chose to perform 50 regular and 50 permutations experiments. The *regular batch* performs experiments on the given dataset, while the *permutation batch* performs experiments where the relationship between input and output is destroyed by shuffling the labels in the learning set. The two distributions are then compared testing the null hypothesis H0, which assumes no difference between them, by means of a non-

parametric two samples Kolmogorov-Smirnov test. This is a principled way to measure the statistical robustness of the obtained result. It is possible to reject the null hypothesis when the computed p- value is smaller than the confidence level of choice. Rejecting H0 implies that there is a clear difference between the two distributions and that the sample size is large enough to describe the relationship between input and output.

Functional characterization of the signature

For the functional analysis of the miRNA signature we used the on-line gene set analysis toolkit WebGestalt (Wang et al. 2017). The toolkit performs the functional characterization by a gene set enrichment analysis in several databases including the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2000) (KEGG). Given a KEGG pathway and a reference set (such as the entire human genome) the enrichment is based on the comparison between the fraction of signature genes in the pathway and the fraction of pathway genes in the reference set. The signature is enriched in the KEGG pathway if the former is larger than the latter fraction. To perform the enrichment analysis in KEGG, we selected the WebGestalt human genome as reference set, p–value ≤ 0.05 as level of significance, the Benjamini-Hochberg correction to correct for multiple hypothesis.

Target gene prediction tools

Several published target prediction algorithms have been used for the prediction of miRNA targets. Generally, such software mainly use sequence complementarity, evolutionary conservation among different species, and thermodynamic criteria to estimate the probability of a miRNA:mRNA duplex formation. For further reading, the review by Bartel published in 2009 is recommended (Bartel et al. 2009).

For gene target prediction, we considered six prediction tools: microRNA (http://www.microrna.org, Betel et al. 2008), mirDB (http://mirdb.org/miRDB/index.html. Wong et al. 2014), Pictar (http:// http://pictar.mdc-berlin.de/, Krek et al. 2005), PITA (http://genie.weizmann.ac.il/pubs/mir07. Kertesz et al. 2007)), TargetScan (http://targetscan.org, Friedman et al. 2008), ElMMo (http://www.mirz.unibas.ch/ElMMo2, Gaidatzis et al. 2007).

Concerning TargetScan we used the Predicted Targets (default predictions) file belonging to the 7.0 release. Concerning PITA we considered the PITA_sites_hg18_0_0_ALL file of the last release (number 6). Concerning PicTar we downloaded and used all the targets from doRiNA database (Blin K. et al. 2014), as suggested in the home page of this

prediction tool. Concerning microRNA we considered the predictions from the August 2010 release. Specifically we used the most comprehensive file that includes the “Non-good mirSVR Score and Non-conserved miRNAs” that includes the most restricted file (that includes only those miRNAs target predictions with good mirSVR scores and conserved miRNAs target sites). Concerning mirDB we considered the last version (5.0). Concerning EImmo we used the last release, number 3, updated in January 2009.

References:

- Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA. (2002) 99:6562–6. 10.1073/pnas.102102699.

- De Mol C, Mosci S, Traskine M, Verri A. A regularized method for selecting nested groups of relevant genes from microarray data. J Comput Biol. (2009) 16:677–90. 10.1089/cmb.2008.0171.

- Tibshirani R, Wainwright M, Hastie T. Statistical Learning With Sparsity: The Lasso and Generalizations. New York, NY: Chapman and Hall/CRC; (2015).

- Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. Brief Bioinform. (2008) 9:119–28. 10.1093/bib/bbn008.

- Barbieri M, Fiorini S, Tomasi F, Barla A. (eds.). PALLADIO: a parallel framework for robust variable selection in high-dimensional data. In: Workshop on Python for High-Performance and Scientific Computing (PyHPC). Salt Lake City, UT: (2016).

- Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 2017 Jul 3;45(W1):W130-W137. doi: 10.1093/nar/gkx356.

- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. (2000) 28:27–30. 10.1093/nar/28.1.27.

- Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009 Jan 23;136(2):215-33. doi: 10.1016/j.cell.2009.01.002. PMID: 19167326; PMCID: PMC3794896.

- Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. (2010) 11:R90. 10.1186/gb-2010-11-8-r90

- Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res. 2015 Jan;43(Database issue):D146-52. doi: 10.1093/nar/gku1104. Epub 2014 Nov 5. PMID: 25378301; PMCID: PMC4383922.

- Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. . Combinatorial microRNA target predictions. Nat Genet. (2005) 37:495–500. 10.1038/ng1536

- Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. (2007) 39:1278–84. 10.1038/ng2135

- Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. (2009) 19:92–105. 10.1101/gr.082701.108

- Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics (2007) 8:69. 10.1186/1471-2105-8-69

- Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin A. DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res. (2015)