A computational approach to the identification of target genes regulated by Systemic Sclerosis-associated miRNAs

Kalliopi Nikitopoulou

Supervisors: Julia Drylewicz, Marzia Rossato, Ad Feelders

Utrecht University

UMC Utrecht

Table of contents

Abstract

In the light of previous research (Chouri et al, unpublished results) performed on Systemic sclerosis (SSc), arises the importance of determining the molecular pathways affected by SSc-associated miRNAs. To accomplish this, it is crucial to identify putative target genes affected by miRNAs that are dysregulated in SSc. The goal of our research is to implement a computational approach to identify putative target genes of miRNAs associated with SSc. We propose here an *in-silico* pipeline based on Pearson correlation of miRNAs and RNAseq data. We applied this method to a dataset from a cohort of SSc patients.

Our approach also considers the multiple testing problem as well as different significance thresholds and various target resources. To evaluate the resulting interactions we determine the strength (significance) and type (negative correlation) of association and compare our approach with the method proposed by van Iterson et al. In the latter study (van Iterson et al, 2013), a multiple linear regression approach called the global test was proposed and was proved to successfully identify target genes by using an integrated analysis of miRNA and mRNA expression. It was also compared to correlation and LASSO methods in terms of predictive performance. We show that our method successfully identifies miRNA-mRNA interactions found in reliable target databases (published validated and experimentally supported) but can also identify potential novel interactions by investigating sequence-based predicted interactions. Finally, the most relevant findings of our analysis might be used for further validation through wet-lab experiments as a next step of investigation.

Acknowledgements

I would like to thank Dr Jose Borghans, Dr. Julia Drylewicz and Dr Marzia Rossato for offering me the opportunity to conduct my research project in the Laboratory of Translational Immunology. I wish to express my sincere gratitude to Dr. Julia Drylewicz, for the daily supervision, the precious feedback and her help in order to fulfill my project. I would like to thank Dr. Marzia Rossato and Dr Ad Feelders for their useful guidance and knowledge shared. Finally, I would like to take this opportunity to express my gratefulness towards my family and my friends for their encouragement and support.

I. Introduction

1. Biological background

Systemic Sclerosis (SSc) is an autoimmune disease characterized by extensive fibrosis of the skin and internal organs, associated with high mortality (Gabrielli et al, 2009). Currently, there is no successful therapy for SSc and its diagnosis is very difficult in the early phase of the disease. Consequently, patient treatment is delayed until skin and/or internal organ involvement is evident and already irreversible. There is thus an extreme need of both effective therapies and early diagnosis markers in order to treat SSc patients at early stage before the development of fibrosis and severe organ complications.

Immune system dysfunction has been demonstrated to underlie SSc pathogenesis (=mechanisms that lead to the diseased state). In particular, plasmacytoid dendritic cells (pDCs) constitute an important subset of leukocytes that is increased in the fibrotic skin of SSc patients (van Bon et al, 2014) and is dysregulated in patients presenting early-symptoms of SSc (van Bon et al, 2013; van den Hoogen et al, 2013). In order to investigate why pDCs are dysregulated early in SSc, a previous study (Chouri et al, unpublished results) has investigated the expression profiling of messenger-RNAs (mRNA = RNA molecules transferring genetic information from gene to protein) and microRNAs (miRNA = short RNA molecules that inhibit gene expression) of pDCs in patients at different stages of SSc and in healthy individuals. This study demonstrated that numerous miRNAs and mRNAs are altered in pDCs of SSc patients. However the molecular mechanisms connecting miRNAs and mRNAs and leading to pDC deregulation in SSc remain unclear.

2. Target Prediction of miRNAs

miRNAs are short RNA molecules that interact with specific mRNAs and inhibit their translation, or promote their degradation. The miRNA-mRNA interaction is driven by means of sequence complementarity and causes reduction of mRNA and/or protein levels. Given that miRNAs are involved in important biological processes and their deregulation is implicated in various diseases, unraveling miRNA targets is of vital importance for both diagnostic and therapeutic aims. However, the lack of high-throughput and low-cost experimental methods for the identification of miRNA-target genes, combined with the complexity of these molecular mechanisms, makes target identification a challenging field.

Based on the sequence complementarity of the interaction between miRNAs and mRNAs, various computational approaches have been developed to predict miRNA-targets (see Alexiou et al, 2011, for a review of sequence-based target prediction algorithms). The most commonly used databases collecting miRNA-target predictions are miRanda, TargetScan, DIANAmicroT, and PicTar. These databases contain a large number of predicted targets for each miRNA. However, the false positive rate for these predictions varies from 24% to 70% (Setupathy et al, 2006; Bentwich, 2005) and the inconsistency among these methods has highlighted the need of developing new strategies for target prediction.

Recently, it has been suggested (Miniategui et al, 2012) that integration of miRNA and mRNA expression data with putative interactions predicted by sequence-based methods can refine target prediction in a given biological context. The main steps of this approach are as follows:

i. Retrieve the existing predicted interactions of the miRNA with potential target genes from prediction databases (e.g. miRanda, TargetScan, DIANAmicroT, and PicTar)
ii. Model the relationship between the expression of miRNA-target gene pairs
iii. Select the miRNA-mRNA pairs significantly inversely correlated

Various mathematical approaches have been implemented to model the relationship between miRNA and targets varying from simple correlation to Bayesian inference methods (Muniategui et al, 2012). These methods have successfully identified valid and biologically relevant interactions among the ones present in the databases of putative targets.

3. Objective of our study

The main objective of our research is to implement a computational approach for the identification of putative target genes of SSc associated miRNAs. We propose here an *in-silico* pipeline that integrates the sequence-based predicted targets and other experimentally validated targets with the miRNA and mRNA data from a cohort of SSc patients.

Our computational approach is based on Pearson correlation between miRNA and mRNA expression data (Figure 1). Since no target prediction analysis has yet been performed on SSc associated miRNAs, we have chosen Pearson correlation as this approach has been successfully used in several studies (Liu et al, 2010; van Der Auwera et al, 2012; Fulci et al, 2009). There are also numerous web-based tools and databases using this approach for target prediction (Gennarino et al, 2011; Sales et al, 2010; Ritchie et al, 2010; Cho et al, 2011).

When integrating expression data and sequence based target predictions we aimed to tackle the inherent problem of false positives in this type of predictions (Witkos et al, 2001). The common procedure to integrate the results of different pure prediction software is to use the "union" or the "intersection" (i.e. common predictions) of putative-targets lists. However, this process does not improve the sensitivity/specificity ratio of true-target identification (Witkos et al, 2011). By using the union list, the problem of false positive targets will tend to deteriorate because false positives of many different software will be taken along. The disadvantage of the intersection approach is that it prevents from exploiting the potential complementarity of these algorithms; i.e. possible true positives found using one algorithm might be excluded in others. We propose a strategy specifically for these target resources in order to avoid these issues.

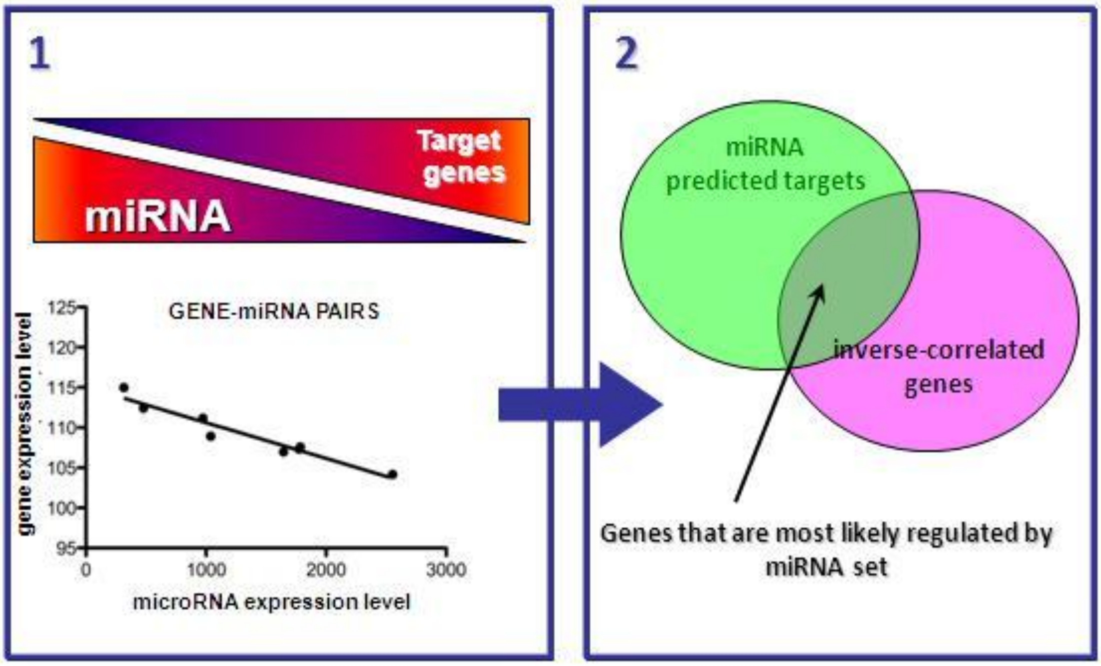

Figure 1. Schematic representation of integrated analysis of gene and miRNA expression with sequence-based prediction data.

II. Materials and Methods & preparatory work

1. Data-sets

Our datasets consisted of a miRNA profiling and a gene transcriptome analysis generated in parallel from pDC of 36 subjects, including 9 healthy individuals and 27 patients at early stages of SSc (patients with SSc-related symptoms i.e. Rhaynaud's phenomenon, early and definite SSc patients). miRNA profiling was generated by RT-qPCR (a technique used to quantify miRNA expression), while the transcriptome (= the set of all RNA molecules in one cell or a population of cells) analysis by RNA-sequencing (a technology that uses the capabilities of next-generation sequencing to reveal a snapshot of RNA presence and quantity from a genome at a given moment in time.). These datasets contain respectively the expression profiles of 45 miRNAs and ~20,000 mRNAs across the 36 subjects.

1.1 miRNA profile

miRNA expression was given as relative fold change (FC) compared to a reference sample of healthy controls. Patients were classified into 4 groups according to the symptoms and the score of the classification criteria based on the van den Hoogen et al (2013) study. In particular, patients were divided into healthy controls (HC), patients showing Raynaud's phenomenon (RP), early scleroderma patients with score 8 (eaSSc) and patients that scored higher than 8, which are definite scleroderma patients (defSSc). Our dataset included miRNAs that were selected as being significantly differentially expressed in at least one group of SSc patients compared to healthy controls. A miRNA was selected if FC > 2 or FC < 0.5 with a p-value lower than 0.05, in at least one patient group. A two-sided independent samples Student t-test was used since the original values were close-to-normally distributed. We obtained *t*-statistic for the null hypothesis H_0: $\mu_X = \mu_Y$ versus the alternate hypothesis H_1: $\mu_X \neq \mu_{Y,}$ where μ_X denotes the average expression of each miRNA at patient group X and μ_Y denotes the average expression of each miRNA at healthy group Y. All differences were considered significant different at the level of 0.05.

A total of 45 miRNAs were selected following these criteria. Before integrating our miRNA dataset with the transcriptome dataset, the miRNA expression data were $\log_2$-transformated. This normalization was performed in order to reduce skewness and variability between samples (Sylvain et al, 2009).

Among the 45 SSc-associated miRNAs, 3 were highly differentially expressed (Chouri E, unpublished results) in patients at early stages of the disease (Figure 2) making them possible candidate prognostic markers or potential therapeutic targets. Therefore, they are examined in more detail later on in the analysis.

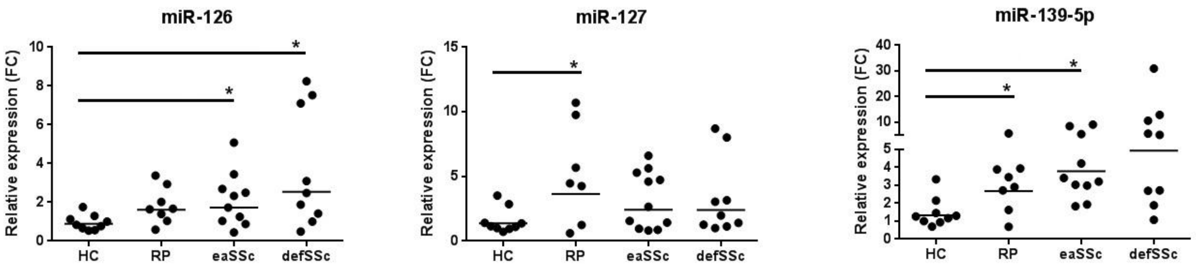

Figure 2. Expression of the 3 miRNAs of interest (miR-126, miR-127, miR139-5p) in healthy controls and patient groups: Healthy controls (HC); patients with Rhayneaud's Phenomenon (RP); early SSc patients (eaSSc); definite SSc patients (defSSc). Expression values of each miRNA is shown as relative fold change across 36 samples and the median FC of each subgroup is indicated by an horizontal bar. Denoted above with * are the statistically significant differentially expressed (FC>2) miRNAs in each group of patients (p<0.05): miR-126 is differentially expressed in eaSSc group and defSSc group, miR-127 is differentially expressed in RP and miR-139-5p in RP and eaSSc.

1.2. Transcriptome profile

The transcriptome dataset included the unbiased measurements of all mRNAs present in pDC of patients and healthy controls, expressed as RPKM (Reads Per Kilobase per Million mapped reads). The initial dataset included 20,365 gene expression measurements. It was reduced to a list of 11,890 after a quality control step as follows:

(i) We replaced missing values by 0 given that usually missing values in RNAseq data correspond to really low measurements (Wang et al, 2009). The missing value in RNAseq can be either due to absence of expression or lack of detection because of limits in the sensitivity of the technique. However, considering the type of sequencing that has been used (20 million reads/sample) this allows the detection of more than 90% of the transcriptome. The genes not detected are therefore either not expressed or expressed at lower than 1 transcript/cell. Consequently we can safely approximate missing values to 0. In order to avoid the log0 issue later on, we added to all measurements the pseudocount 0.1.

(ii) We performed quantile normalization in order to make the sample distributions for each gene identical in statistical properties. Quantile normalization was originally used for gene expression microarrays but it is now also applied in several data types including RNA-Seq (Hicks et al, 2014). In this scheme, the following algorithm (Bolstad et al, 2003) is applied for normalizing a set of data vectors by giving them the same distribution: 1. given n arrays (genes) of length p (sample size), form X of dimension p × n where each array is a column; 2. sort each column of X to give X_{sort}; 3. take the means across rows of X_{sort} and assign this mean to each element in the row to get X'_{sort}; 4. get $X_{normalized}$ by rearranging each column of X'_{sort} to have the same ordering as original X. In other words, in the quantile normalization process, a reference sample distribution is first computed by taking the average across all ordered (ascending order) gene expression observations in each sample distribution. The original expressions are then replaced by the entries of the reference sample distribution with the same rank; the highest value from each array is replaced by the average of all of the highest values, the second highest by the average of all of the second highest, and so on. In a formula, the transform is: $x_{norm} = F_i^{-1}(F_{ref}(x))$, where F_i is the distribution function of array *i*, and F_{ref} is the distribution function of the reference array.

(iii) Pathway enrichment analysis was performed, using DAVID© software, to define a cutoff of gene expression: we removed from the mRNA dataset genes with low expression level in the range of measurement noise. We considered two options for the aforementioned filtering: exclude genes for which the median gene expression value is lower than 0.5 or lower than 1 RPKM. The list of genes we would lose with the higher gene expression cutoff threshold was checked in order to identify their functionality. The pathway enrichment analysis revealed that many genes were indeed relevant, so we decided to use a cutoff at 0.5

(iv) Normalization to a normal distribution with $\log_2$ transformation was performed, as in the miRNA profile data. Initially the RNAseq data are assumed to have a Poisson distribution or a negative binomial distribution (Srivastava et al, 2010). This is explained when thinking that in RNA-Sequencing you choose a location at random from the transcriptome to produce a read; this is a Poisson process.

2. Collection of predicted miRNA-targets from available databases

As previously mentioned, there are some existing target resources predicting target genes for miRNAs. They can be classified in three categories according to the type of predicted-target:

a. *validated targets*: genes predicted to interact with a certain miRNA and also experimentally proven in publications to be actively regulated by it. This database is a result of meticulous literature curation.
b. *experimentally supported targets*: predicted miRNA-target pairs for which some experimental evidences are present. They integrate data from high-throughput techniques as well as individual miRNA studies providing either direct or indirect evidence for interaction.
c. *purely predicted targets*: novel miRNA-targets identified by computational analyses based on sequence complementarity between the miRNA and the target mRNA.

It was of most importance to consider all these type of resources, as the databases of validated and experimentally-supported targets include the most reliable miRNA-mRNA pairs, while the predicted targets can reveal novel interactions specifically present in pDCs, which were not previously identified in other experimental models.

We retrieved the targets present in the 3 different databases for the 45 miRNAs included in our dataset. We used the list of targets existing in each database for our miRNA set as input for the correlation computation analysis.

2.1. Validated targets

In the case of the published validated targets, the information reported is manually curated (according to PubMed).

We used three different databases of validated targets:

- miRecords: http://mirecords.biolead.org
- mirTarBase: http://mirtarbase.mbc.nctu.edu.tw

- TarBase v6: http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index.

As this class of targets is the most reliable one, we wanted to include as many targets as possible from these resources. Therefore we used the union of the lists retrieved from these databases. In the same aim, we included the targets identified for both human and mouse, as miRNA-target interactions are usually conserved among species. Finally to maximize the gain of information, we enlarged the selection of miRNA-target pairs to all miRNA-families as miRNAs belonging to the same family differ only in a single nucleotide and usually share their targets.

For mirTarBase, we removed records of the database that were identified as weak evidences. From TarBase (version 6) we selected records identified as "positive" indicating that the effect of the miRNA on the target gene expression is a classical inhibition as expected.

The final table of validated miRNA-target pairs contained ~3,900 miRNA-target gene pairs for the 45 miRNAs that we studied. This list of validated targets is very restrictive, although very well-founded and reliable.

2.2. Experimentally supported targets

We retrieved experimentally-supported targets from three databases:

- StarBase: http://starbase.sysu.edu.cn/. We retrieved miRNAs-target pairs predicted by at least 3 prediction software and supported by 2 experimental evidences (medium threshold). The database allows also other filtering options that we ignored.
- mirTarBase: http://mirtarbase.mbc.nctu.edu.tw/. We kept all data without applying any filters, except from removing records with the indication "non-functional MTI" which indicates a false-positive prediction.
- TarBase: http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=site/index. We made the selection of "positive" miRNA-target interactions, as described above, without any further filtering of the database.

Similarly to the validated target databases, we also used the union of the experimentally supported databases to maximize the information. But the selection of miRNA-target pairs was here restricted to human and to the specific miRNA-isoforms included in the list of 45 miRNAs of interest.

The final table of experimentally supported targets included ~10,000 miRNA-target gene pairs for the studied 45 miRNAs.

2.3. Pure predictions

We used six software predicting miRNA targets according to miRNA-mRNA binding rules: DIANAmT (Kiriakidou et al, 2004), miRanda (John et al, 2004), PICTAR5 (Krek et al, 2005; Lall et al, 2006), PITA (Kertesz et al, 2007), RNA22 (Miranda et al, 2006) and TargetScan

(Lewis et al, 2005). We used the web tool miRWalk (http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/) to retrieve the targets predicted by at least 3 out of the 6 mentioned software by considering the longest 3′UTR of the target (3′UTR is the possible miRNA binding region on a mRNA). The final table of pure predictions included ~90,000 miRNA-mRNA pairs for the studied 45 miRNAs.

Although the common process to integrate the targets of different prediction software is to use the "union" or the "intersection" of the putative-targets lists, other integration strategies can be implemented. Such strategies are: (i) to use only one/two selected prediction software e.g. TargetScan (Lewis et al, 2005), or PITA (Kertesz et al, 2007), (ii) to apply the intersection approach by selecting the best performing algorithms i.e. TargetScan, Pictar (Krek et al, 2005; Lall et al, 2006) and Miranda (John et al, 2004), (iii) to include targets predicted by at least some of these software by selecting the desired number to include. In our case, to avoid the limitations of the intersection approach and the restricted results of one specific algorithm, we used the 3rd option.

3. Target prediction based on computational models

Several computational approaches have been proposed to predict miRNA-mRNA interactions from experimental data. These methods use information of mRNA expression, miRNA expression and putative interactions (target predictions).

Let define the matrices $X_{J\times T} = [x_{jt}]$ and $Z_{K\times T} = [z_{kt}]$ of the expression values of mRNAs (j=1,…,J and J=11,890) and miRNAs (k=1,…,K and K=45) in sample t (t=1,…,T and T=36), respectively and the binary matrix $C_{J\times K} = [c_{jk}]$ of putative interactions, where $c_{jk} = 1$ if the mRNA j is predicted as target of the miRNA k from a sequence-based method and 0 otherwise.

The most frequently applied statistical approaches are the following:

3.1. Pearson correlation

The Pearson correlation approach is a simple method to evaluate the relationship between miRNAs and mRNAs. It is a measure of linear dependency and its mathematical representation is given in equation (1) where r_{jk} represents the Pearson correlation coefficient of the mRNA j and the miRNA k.

$$r_{jk} = \frac{\sum_{t=1}^{T}(x_{jt}-\bar{x}_j)(z_{kt}-\bar{z}_k)}{\sqrt{\sum_{t=1}^{T}(x_{jt}-\bar{x}_j)^2}\sqrt{\sum_{t=1}^{T}(z_{kt}-\bar{z}_k)^2}} \qquad (1)$$

In this formula, $\bar{x}_j = \frac{1}{T}\sum_{t=1}^{T} x_{jt}$ is the mean expression of the mRNA j and $\bar{z}_k = \frac{1}{T}\sum_{t=1}^{T} z_{kt}$ is the mean expression of the miRNA k.

We applied the Pearson correlation method because our dataset was transformed to a normal distribution.

3.2. Multiple linear regression

It is important to point out that each miRNA can target (bind to) various different mRNAs according to a matching sequence complementarity of nucleotides. Therefore there is a many to many relationship between miRNAs and target mRNAs; one mRNA can have several different miRNAs targeting it and one miRNA can target several different mRNAs. However, correlation does not imply causation in this case; there is no causal relationship between miRNAs and target mRNAs but a simple linear relationship of their expression levels.

Multiple linear regressions describe the relation of a given miRNA k and the complete set of targets at the same time, rather than evaluating separately each interaction as done by computing Pearson correlations.

The multiple linear regression model can be formulated as follows:

$$z_{kt} = \sum_{j=1}^{J^k} b_{jk} \cdot x_{jt} + a_k + e_t \qquad (2)$$

where J^k represents the total number of putative targets of the miRNA k and $b_{jk} = \{b_{0k}, b_{1k}, \ldots, b_{Jk}\}$ is the vector of regression coefficients, a_k the intercept and e_t the error term.

Note that in this equation, the regression coefficients (b_{jk}) represent the independent contributions of each independent variable x_j to the dependent variable z_k. In other words, the expression of target-mRNA j is correlated with the expression of miRNA k after controlling for all other predicted targets' expression values.

3.3 Global Test as proposed by van Iterson et al. (2013)

In the approach of van Iterson et al. (2013), an integrated analysis of miRNA and mRNA expression based on a model called "Global Test" was developed. The global test is a multiple linear regression (MLR) model where the expression of a miRNA is modeled as a function of the expression of the complete set of their predicted targets.

This approach was implemented in an R package called miRNAmRNA, (available from www.humgen.nl/MicroarrayAnalysisGroup.html).

As input the algorithm uses the matrix of miRNA expressions, the matrix of mRNAs expressions and the incidence matrix relating miRNAs to mRNAs.

The algorithm gives as output the miRNA-mRNA associations ordered, according to their p-value (p-value of the coefficient in the regression model), from the strongest to the weakest association. They use a statistical significance level of 0.05 and do not perform False Discovery Rate Correction.

In each regression model computation the two-sided t-test is used for the null hypothesis H_0: $b_j=0$ versus the alternative hypothesis H_1: $b_j \neq 0$. We obtain the *t*-statistic $t = \frac{b_j}{\sigma_j}$, where b_j are the coefficient estimates and σ_j their standard error. The test statistic follows a t-distribution under H_0 with: df=n-2=36-2=34 (n = sample size) degrees of freedom, since the samples follow independent normal distributions.

In their paper, van Iterson et al. (2013) compared the performance of the proposed global test with Pearson correlation and LASSO approaches. They found that the global test performed similarly to Pearson correlation in terms of sensitivity in the identification of experimentally validated miRNA-mRNA pairs. They also found that global test outperformed LASSO. We considered the global test in addition to our Pearson correlation approach in order to compare results of the two methods.

4. Multiple testing correction

When many statistical tests are performed simultaneously, arises the problem of multiple testing. To correct for the multiple testing problem, we endeavor to control the probability of committing a Type I error. The common approach to address the multiplicity problem involves controlling the family-wise error rate (FWER), which is the probability of getting at least one false positive (committing a Type I error). This is achieved by setting a level of significance for an entire family of related hypotheses. A different and less stringent approach presented by Benjamini and Hochberg, suggests controlling the fraction of false significant results among the significant results found (Benjamini et al, 1995). This approach is called false discovery rate correction (FDR = expected proportion of erroneous rejections among all rejections of the null hypothesis).

The table below summarizes the notation used by Benjamini and Hochberg (1995).

		Decision Made		
		Not reject	**Reject**	**Total**
True State Of Nature	**Null is true**	U	V	m_0
	Null is false	T	S	$m-m_0$
	Total	m-R	R	m

Where,

- m is the total number hypotheses tested
- m_0 is the number of true null hypotheses
- $m-m_0$ is the number of true alternative hypotheses
- V is the number of false positives (Type I error), also called "false discoveries"
- S is the number of true positives, also called "true discoveries"
- T is the number of false negatives (Type II error)
- U is the number of true negatives
- R is the number of rejected null hypotheses, also called "discoveries"

R is an observable random variable, and S, T, U, and V are unobservable random variables. According to the above notation, the False Discovery Rate is: $R = \frac{V}{V+S} = \frac{V}{R}$. While in the FWER approach we try to control $P(V \geq 1)$, in the FDR approach we try to control $E(\frac{V}{R})$ for $R > 0$.

The Benjamini and Hochberg procedure is a stepwise process involving the following steps:

1. Sort the p-values from the entire set of m tests from smallest to largest: p_i refers to the ith smallest p-value.

2. Define *k* as the largest value of *i* for which $p_i \leq \left(\frac{i}{m}\right) \cdot \alpha$, where α is the predefined significance threshold. $\left(\frac{i}{m}\right) \cdot \alpha$ is an “adjusted threshold value” for each individual p-value (Lemma proved in Benjamini et al, 1995), so the procedure compares p_m with α, p_{m-1} with $\left(\frac{m-1}{m}\right) \cdot \alpha, \ldots, p_1$ with $\frac{\alpha}{m}$.

3. If at least one value of *i* satisfies this relationship, then hypotheses 1 to k are rejected, otherwise no hypotheses are rejected.

Keselman et al. (1999) have shown that when the number of comparisons in the set of tests increases, the Benjamini and Hochberg (BH) approach loses less power compared to other approaches based on family-wise control. Therefore, we decided to use the BH approach in our study.

5. Our approach based on Pearson correlation

In our approach, we correlated mRNA expression levels with those of putative targeting miRNAs using Pearson correlations. Our approach was implemented in R (http://www.r-project.org/) and the general pipeline is presented in figure 3.The algorithm uses as input the database target list, the gene expression data and their matched miRNA expression data as well as a statistical significance threshold and a boolean variable stating whether FDR correction will be applied or not.

miRNA and mRNA data were first normalized and filtered as described previously (quality control step). The mRNA data were then merged with the list of *in silico* predicted targets lists generated by using the existing databases (pure predictions, validated or experimentally supported). Pearson correlation coefficients between a particular miRNA and its predicted target mRNAs were computed. We only considered negative correlations reaching the chosen significance threshold. The p-values were obtained, and Benjamini-Hochberg adjustment was used as a multiple testing correction. Finally, the output included each miRNA-mRNA pair that was significantly negatively correlated according the chosen criteria, ranked in ascending order according to the correlation coefficient and p-value.

In each correlation computation in our approach the two-sided t-test was used. We obtained *t*-statistic $t = r\sqrt{\frac{n-2}{1-r^2}}$ (n = sample size) for the null hypothesis H_0: r=0 versus the alternative hypothesis H_1: r≠0. The test statistic follows a t-distribution under H_0 with: df=n-2=36-2=34 degrees of freedom, since the samples follow independent normal distributions.

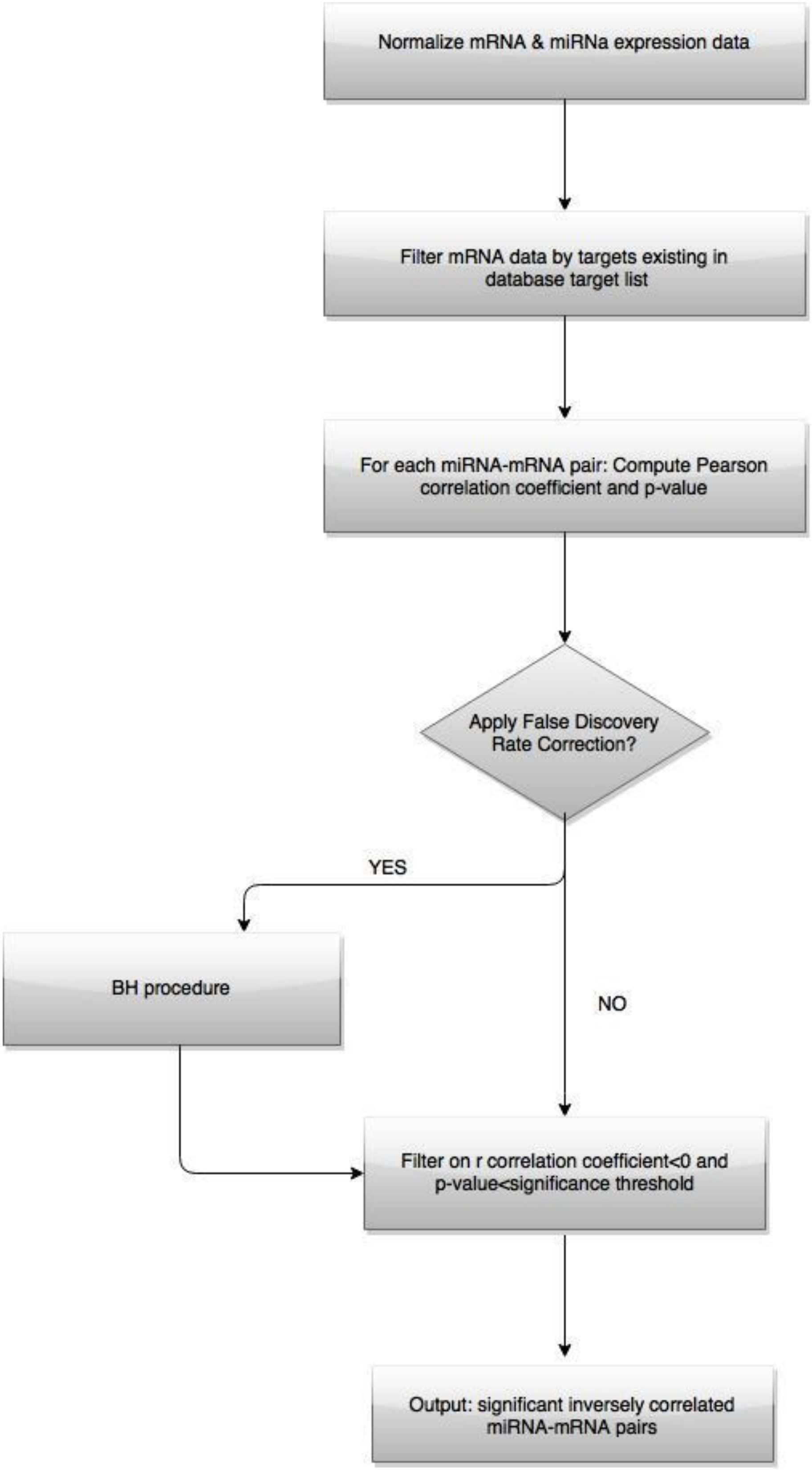
Normalize mRNA & miRNa expression data
Filter mRNA data by targets existing in database target list
For each miRNA-mRNA pair: Compute Pearson correlation coefficient and p-value
Apply False Discovery Rate Correction?
YES
NO
BH procedure
Filter on r correlation coefficient<0 and p-value<significance threshold
Output: significant inversely correlated miRNA-mRNA pairs

Figure 3. Pipeline of our approach to identify miRNAs target genes.

III. Results

We applied our approach on our datasets of miRNA and mRNA expression using different statistical parameter settings: significance threshold at 0.05 and 0.1, and FDR correction or not. These settings gave different results for the 3 different lists of predicted targets, thus emerged the need to discriminate among them and determine an optimal setting for each type of target list. In this section, we present the results obtained using our pipeline based on Pearson correlation and compare them with the results obtained when applying the global test proposed by van Iterson et al. (2013). We compared these approaches in terms of the number of significant interactions found in the lists of pure predictions, experimentally supported and published validated interactions.

1. Results of Pearson correlation analysis

We first applied our correlation approach without FDR correction and with significance threshold of 0.05. For the total of 45 miRNAs under study, in this setting we found 169 validated targets, 484 experimentally supported and 5557 pure predictions (Figure 4.A). For the 3 miRNAs of interest, only 1 validated target was significantly inversely correlated with a miRNA, while in the database of experimentally supported targets 6 pairs were found and for the pure predictions target list, 174 miRNA-mRNA pairs passed the significance threshold (Figure 5.A).

When applying the FDR correction while keeping the significance threshold at 0.05, for the total of 45 miRNAs, 6 gene-miRNA pairs were found in the database of validated targets, 21 in the experimentally supported and 435 in the pure predictions list (Figure 4.B). In this setting, for the 3 miRNAs of higher interest, no significant inversely correlated interactions were obtained from the databases of validated and experimentally supported targets. However, 10 negatively correlated pairs present in the pure predictions target list passed the significance threshold (Figure 5.B).

When changing the significance threshold of 0.1 and still applying FDR correction, 17 pairs were found in the validated target list, 69 in the experimentally supported and 1072 in the pure predictions, for the total of 45 miRNAs (Figure 4.C). In this setting, for the 3 miRNAs of higher interest, our approach still gave 0 interactions included in the validated database, only 1 in the experimentally supported list of targets and 24 in the pure predictions (Figure 5.C).

2. Comparison with the global test method

We compared our results with those obtained when applying the global test proposed by van Iterson et al on our dataset. In this approach, no FDR correction was applied and the significance threshold was 0.05. Compared to the correlation approach, the global test resulted in less targets for the pure predictions and for the experimentally supported target list (553 miRNA-mRNA pairs less) and the experimentally supported target list (1 pair less), but more on the validated target list (1 pair more) (Figure 4.D). Interestingly, for the case of the 3 miRNAs of interest, this method produced the same resulting interactions as our approach of Pearson correlation without FDR correction and significance threshold at 0.05 (Figure 5.D).

From this comparison, we observed that although the two methods perform similarly for the case of the 3 miRNAs of interest (Table 2), our approach produces more results for the total of 45 miRNAs (Table 1).

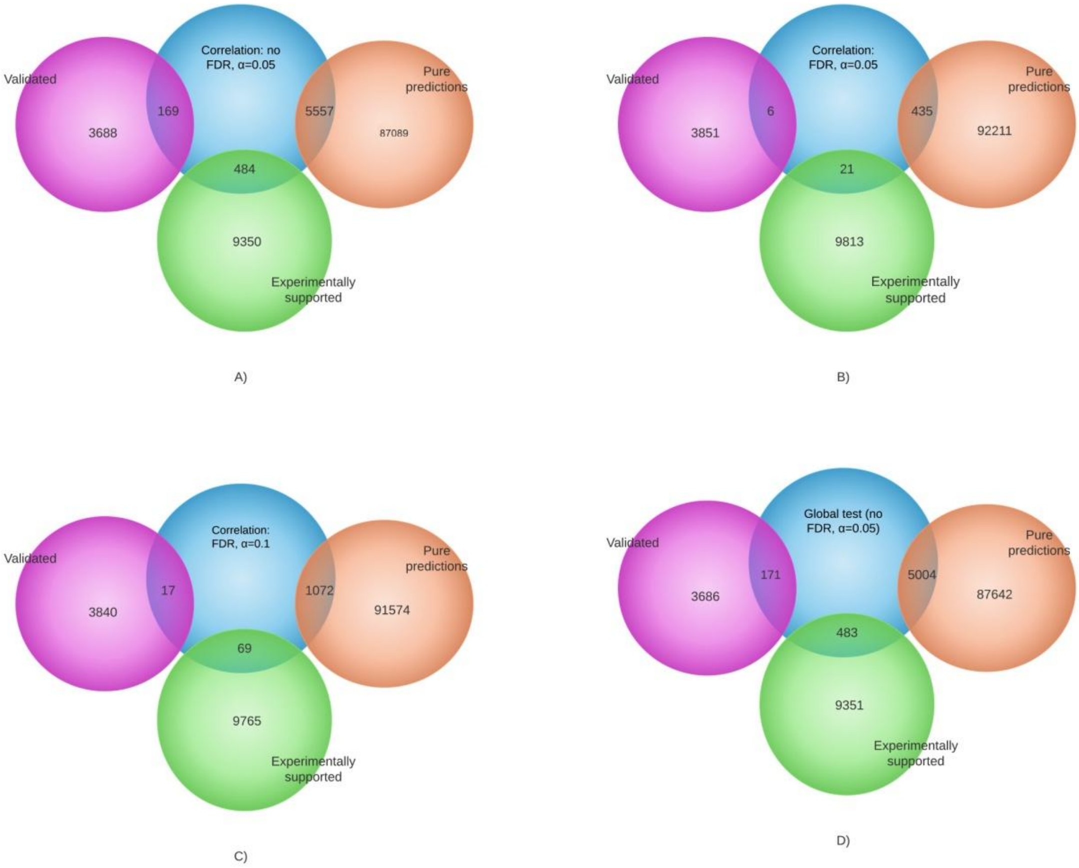

Figure 4. Venn diagrams showing, for the 45 miRNAs under study, the number of significantly inversely correlated miRNA-mRNA pairs found when merging the results of our Pearson correlation approach (A,B,C) and the global test (D) with each list of predicted targets: validated (purple), experimentally supported (green) and pure predictions (orange). In A the Pearson correlation approach (blue) was applied with significance threshold 0.05 and no FDR correction was performed. In B the correlation approach (blue) was applied with significance threshold 0.05 and with FDR correction. In C our correlation approach (blue) was applied with significance threshold 0.1 and FDR correction. In D the global test (blue) was applied. Possible common targets between the three target lists are not shown here.

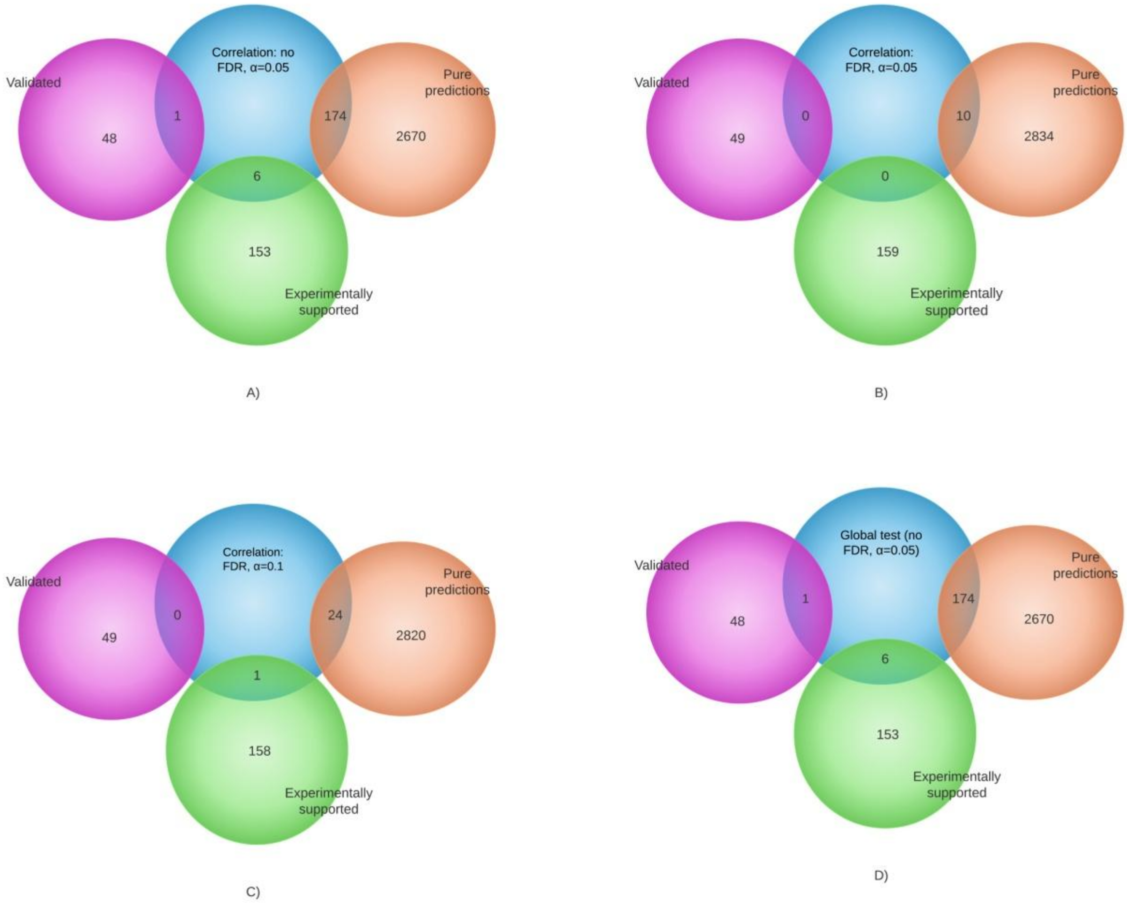

Figure 5. Venn diagrams showing, for the 3 miRNAs of higher interest, the number of significantly inversely correlated miRNA-mRNA pairs found when merging the results of our Pearson correlation approach (A,B,C) and the global test (D) with each list of predicted targets: validated (purple), experimentally supported (green) and pure predictions (orange). In A the Pearson correlation approach (blue) was applied with significance threshold 0.05 and no FDR correction was performed. In B the correlation approach (blue) was applied with significance threshold 0.05 and with FDR correction. In C our correlation approach (blue) was applied with significance threshold 0.1 and FDR correction. In D the global test (blue) was applied. Possible common targets between the three target lists are not shown here.

Table 1. Comparison of results obtained for the total of 45 miRNAs with different parameter settings for the correlation approach and the Global test of van Iterson et al. We demonstrate the percentage of significant interactions found within the validated, experimentally supported and pure predictions databases. The correlation approach with no FDR correction and the Global test give a much higher findings rate (4.5 - 6% approximately) compared to the correlation approaches with FDR correction (0.2 – 1.2% approximately).

	Correlation: no FDR, α=0.05	Correlation: FDR, α=0.05	Correlation: FDR, α=0.1	Global test (no FDR, α=0.05)
Validated	4.48%	0.15%	0.44%	4.43%
Experimentally supported	4.92%	0.21%	0.70%	4.91%
Pure predictions	5.99%	0.46%	1.15%	5.40%

Table 2. Comparison of results obtained for the 3 most important miRNAs with different parameter settings for the correlation approach and the Global test of van Iterson et al. We demonstrate the percentage of significant interactions found within the validated, experimentally supported and pure predictions databases. Similarly to the case of the 45 miRNAs, the correlation approach with no FDR correction and the Global test give a higher findings rate compared to the correlation approaches with FDR correction.

	Correlation: no FDR, α=0.05	Correlation: FDR, α=0.05	Correlation: FDR, α=0.1	Global test (no FDR, α=0.05)
Validated	2.04%	0%	0%	2.04 %
Experimentally supported	3.77%	0%	0.62%	3.77%
Pure predictions	6.11%	0.35%	0.84%	6.11%

3. Correlation approach: using different statistical settings for different target lists

Looking at our results described previously for different parameter settings for the 3 different databases of targets, we observed that the threshold of 0.05 limited a lot the results we got from the most reliable target resources -validated and experimentally supported- for the total of 45 miRNAs and for the 3 most interesting miRNAs. This was mostly the case when FDR correction was applied, where no correlation passed the significance threshold of 0.05 in the validated target list, and only 1 experimentally supported interaction passed the threshold of 0.1, for the 3 miRNAs of higher interest. Consequently, we needed to apply our pipeline by using a less strict threshold, specifically for the two reliable databases (validated and experimentally supported).

Since these databases contain proven and verified targets, it is tempting to conclude that no FDR correction is needed when a negatively correlated miRNA-mRNA pair included in these two target lists is found to be significant. As the aim of our approach is to identify putative target genes that will be further investigated by new experiments, the impact of including false positives is in fact less severe than the risk of excluding a possible true positive since the number of pairs found is so limited. Therefore, we decided to not apply a FDR correction for the validated and experimentally supported targets and we identified 7 significantly inverse correlated pairs (Table 3) when considering the 3 miRNAs of higher interest.

Table 3. miRNA-mRNA interactions for the 3 miRNAs of interest predicted by our analysis and included in the validated and experimentally supported databases without FDR correction at a significance level of 0.05. For each pair the value of the correlation coefficient and the p-value are given. For a given miRNA the interactions are shown in ascending order according to their p-value.

miRNA name	Gene symbol	Correlation coefficient r	P-value
miR.126	PTPN7	-0.341446528	0.041541
miR.139.5p	UHMK1	-0.489253018	0.00246
miR.139.5p	PTPRS	-0.436076564	0.007845
miR.139.5p	MEF2A	-0.43266476	0.0084
miR.139.5p	AP3M1	-0.401533736	0.015211
miR.139.5p	IGF1R	-0.398113108	0.016185
miR.139.5p	TCF12	-0.344125275	0.039867

However, for the pure predictions target list, the problem of false positives already exists, so the necessity of applying FDR correction in the correlated pairs found in this database should be considered as substantial. Additionally, from a practical point of view, the large number of miRNA-mRNA pairs found in this database made it crucial to reduce the number of pairs considered for experimental validation later on.

For the 45 miRNAs under study, when no FDR correction was applied and the significance threshold was set at 0.05, the number of miRNA-mRNA pairs included in the results was very large (5557 interactions found) which would make any further experimental investigation very difficult. On the contrary, when applying FDR correction the number of results on the pure prediction list was 435 and included 10 targets for miR-135-5p (1 out of the 3 most interesting miRNAs). When the significance level was set at 0.1 we obtained 1072 pairs for the total of 45 miRNAs including 24 pairs (Table 4) for 2 out of the 3 most interesting miRNAs (miR126 and miR-139-5p) thus improving our findings on the 3 miRNAs.

Table 4. miRNA-mRNA interactions found in the pure predictions list, for the 3 miRNAs of interest, using FDR correction and a significance level of 0.1. For each pair the value of the correlation coefficient and the p-value are given. For a given miRNA the interactions are shown in ascending order according to their p-value.

miRNA name	Gene symbol	Correlation coefficient r	P-value
miR.126	ZC3H12B	-0.517002576	0.001243925
miR.139.5p	UST	-0.590135345	0.00015171
miR.139.5p	TMEM67	-0.576971794	0.000229896
miR.139.5p	RAB3IP	-0.570598083	0.000279389

miR.139.5p	INCENP	-0.562424786	0.000356695
miR.139.5p	ZNF260	-0.555682197	0.000434275
miR.139.5p	ZNF793	-0.552807559	0.000471687
miR.139.5p	SLC26A2	-0.551803626	0.000485413
miR.139.5p	CYP20A1	-0.550501319	0.000503747
miR.139.5p	ICA1L	-0.544749284	0.000592312
miR.139.5p	SPATA5	-0.537840665	0.000716772
miR.139.5p	ASB1	-0.537575401	0.000721982
miR.139.5p	HERC2	-0.524643754	0.001020408
miR.139.5p	ANKH	-0.520830962	0.001127059
miR.139.5p	RANBP10	-0.520415918	0.001139242
miR.139.5p	NFAT5	-0.515710761	0.001285701
miR.139.5p	EML5	-0.507546149	0.001579501
miR.139.5p	MAP9	-0.493329015	0.002233578
miR.139.5p	HTT	-0.490493828	0.002389189
miR.139.5p	UHMK1	-0.489253018	0.002460219
miR.139.5p	MKL2	-0.488635232	0.002496266
miR.139.5p	AP4E1	-0.484953081	0.002720837
miR.139.5p	RPS6KA3	-0.484147352	0.002772262
miR.139.5p	C7orf42	-0.622306789	5.0723E-05

Another advantage of FDR correction on the abundant list of targets obtained when merging our correlation results with the pure predictions was that FDR correction reduced the number of weak interactions found within the pure predictions. This was revealed when examining the distribution of the r correlation coefficient.

For the setting of no FDR correction and significance threshold 0.05 the total number of significant inverse correlated miRNA-mRNA pairs found in the database of pure predictions was 5557. Checking the distribution of the values of the correlation coefficient r for these interactions (Figure 6), we observed that the large majority (83%) lies in the interval [-0.5, -0.3]. Comparing these results with the 435 FDR corrected correlations significant at 0.05, we saw (Figure 6) that after applying the correction for multiple testing the interactions excluded from the not FDR corrected results contained the weak linear relationships rather than the strong ones. Specifically, for 97% of the FDR corrected correlation results, the r correlation coefficient lies in the interval [-0.7, -0.5].

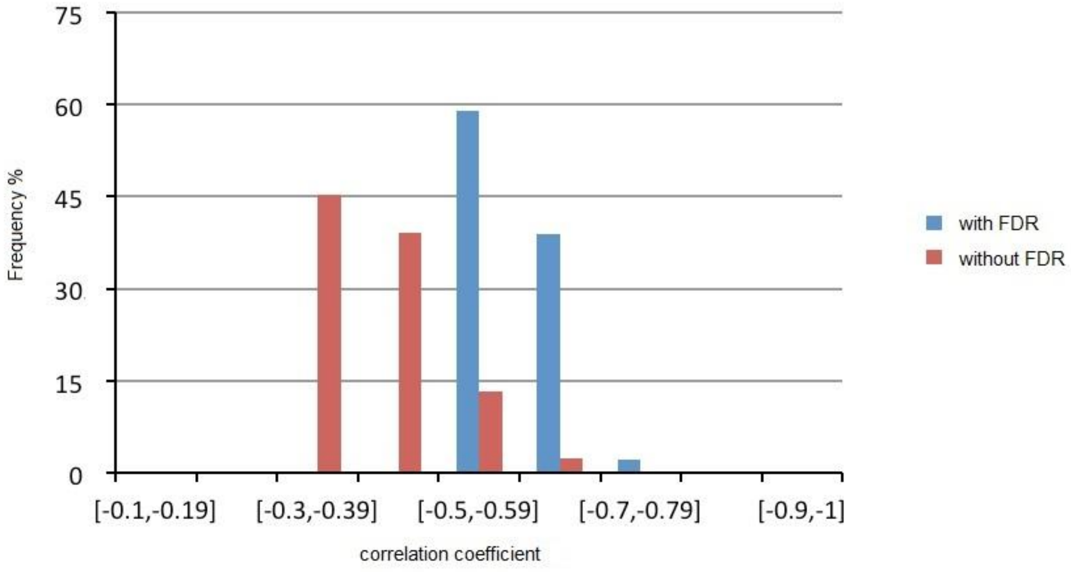

Figure 6. Distribution of correlation coefficient r for the significantly inversely correlated miRNA-mRNA pairs found in the pure predictions list with (blue bars) and without FDR correction (red bars) at a significance level of 0.05.

Similarly, when setting the significance threshold at 0.1, applying FDR correction did not exclude the most strongly related miRNA-mRNA pairs: the correlation coefficient for the majority (approximately 90%) of the inversely correlated pairs in the FDR corrected results was between -0.7 and -0.5, while for the not FDR it was lower than -0.5 (Figure 7).

For the pure predictions list, we can conclude that applying a FDR correction excludes the weakest correlations and keeps only the strongest ones.

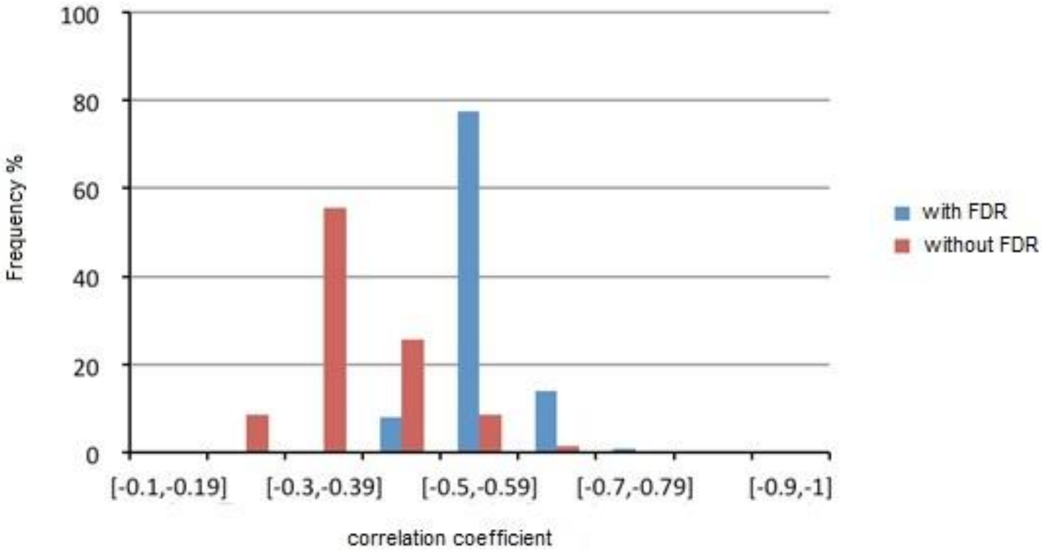

Figure 7. Distribution of correlation coefficient r for the significantly inversely correlated miRNA-mRNA pairs found in the pure predictions list with (blue bars) and without FDR correction (red bars) at a significance level of 0.1.

4. Correcting for the purity of samples

During our experiment we encountered a problem with the purity of the pDC samples. (purity=the proportion of cells object of the study -pDC- over the total amount of cells used in the analysis, composed by pDC and other contaminating cell types) The purity was found to be quite variable (70-90%). Impure samples could possibly affect the results of our analysis, consequently we decided to correct for this problem.

Our purity data consist of: the 1st cohort where we have the expression of 11 genes of different cell type (pDCs, Monocytes-lineage, Monocytes, Tcells, Bcells) that are used as markers for the purity, and the sample purity given as percentage for 30 samples; the 2nd cohort where we have the expression of 12 gene-markers of the aforementioned cell types and the sample purity as percentage for 31 samples.

We aimed at identifying the markers that might explain the variation of purity within the different samples. The contamination of the samples by other cell types might influence the expression of the genes included in the RPKM table and therefore be confounders. In order to analyze the correlation between miRNA and genes cohort we would like to correct for these confounders by taking them into account together with the target gene expression in a multiple linear regression model thus identifying the correlation of miRNA and mRNA expression while simultaneously controlling/adjusting for the purity markers’ expression.

We performed a multiple linear regression analysis using the purity of pDCs in the sample (given as percentage) and the lists of markers in each cohort defining different contaminating cell types. The form of the regression model was: purity ~ $\text{marker}_1 + \text{marker}_2 + \ldots + \text{marker}_m$, where m is the number of given markers. The best model was selected stepwise using the Akaike Information Criterion.

Following this process, for the 1st cohort, we concluded in 4 markers explaining the purity: CD14, FCGR2A, FCGR2C, CD19 (Table 5). For the 2nd we concluded in: FCGR2B, CD14, CD19 and CD3D.

We decided to use all markers found for the 1st cohort (CD14, FCGR2A, FCGR2C, CD19) and add one extra marker found for the 2nd cohort (FCGR2B) to correct for the purity in our miRNA-mRNA correlation analysis. We decided to exclude CD3D given its low R^2 score.

Table 5. Resulting purity markers obtained by performing multivariate linear regression analysis for the 1st cohort. Presented in the table is the following information, for the statistically significant variables (p-value<0.05) of the best model: R^2 for each variable and the p-value. The specific p-value refers to the test for the hypothesis that the coefficient of the gene is zero. The R^2 reported shows how much variation of the purity is explained by the variation of the marker. Model ’s R^2 = 0.5445 (selected with stepwise AIC).

Purity marker	R^2	p-value
CD14	-0.4625821	0.02784
FCGR2A	0.7150392	0.00987
FCGR2C	0.2396062	0.02704

CD19	0.09153679	0.00347

Table 6. Resulting purity markers obtained by performing multivariate linear regression for the 2nd cohort. Presented in the table is the following information, for the statistically significant variables (p-value<0.05) of the best model: R^2 for each variable and the p-value. The specific p-value refers to the test for the hypothesis that the coefficient of the gene is zero. The R^2 reported shows how much variation of the purity is explained by the variation of the marker the effect of the gene expression on the purity controlling/adjusting for the purity markers expression. Model ’s R^2 = 0. 0.8365 (selected with stepwise AIC).

Purity marker	R^2	p-value
FCGR2B	0.2412635	0.0160
CD14	0.2191132	0.0109
CD19	0.3650827	1.79e-05
CD3D	-0.07045658	0.0014

Next, in order to address the “purity problem”, we modified our approach to predict interaction between gene expression and miRNA expression accounting for purity. We attempted to correct for this problem by including the 5 markers found to be related to the purity as covariates in a linear model where miRNA expression is the dependent variable and the independent variables are each mRNA examined for correlation together with the “purity markers”. We performed multivariate linear regressions for each miRNA-mRNA pair. The form of the model was defined as follows:

miRNA ~ mRNA + CD14 + FCGR2A + FCGR2C + CD19 + FCGR2B

Unfortunately this approach did not produce results for our standards: when applying FDR correction at significance level α=0.1 there were found 0 inversely correlated pairs in the pure prediction target list for the 3 miRNAs of interest. When examining the experimentally supported database only 6 miRNA-mRNA pairs (for the 3 most interesting miRNAS) were found significant and no validated interactions were obtained (Table 7).

Table 7. Comparison of results obtained for the 3 most important miRNAs (red) and the total of 45 miRNAs (blue) with different parameter settings for the purity correction model using 5 purity markers. We demonstrate the number of significant interactions found within the validated, experimentally supported and pure predictions databases.

	Purity model: no FDR, α=0.05		Purity model: no FDR, α=0.1		Purity model: FDR, α=0.1	
Validated	132	0	226	0	0	0
Experimentally supported	258	6	469	10	0	0
Pure predictions	2254	94	4110	190	1	0

In an effort to obtain more interactions for the 3 most important miRNAs, we performed this analysis with less stringent “purity correction”. As we suspected some correlation between “purity” markers, we decided to use a model that included a subset of the markers accounting for the purity. Using FCGR2A, FCGR2C and FCGR2B may not be necessary since they all account for the same type of cells as CD14 (Monocytes). Thus, we performed the analysis using as explanatory variables accounting for the purity only CD14 and CD19 (we selected the 2 out of the 5 markers best explaining the purity, based on R^2 statistics as seen in tables 5 & 6). However, this alternative approach did not bring about the expected improvement. On the contrary, less significant inversely correlated pairs were found for the 3 miRNAS under study.
Comparing the miRNA-mRNA interactions found using 2 markers with the ones obtained using 5 markers for the pure predictions database, we observed the following:

- When running the model without FDR correction at $\alpha=0.1$, looking at the 45 miRNAs we found more inversely correlated pairs using 2 markers than using 5 markers (~6000 vs. ~4100) ; on the contrary we found less when we looked at the 3 miRNAs (122 vs. 190)
- When applying FDR correction, only a few number of pairs were still significantly inversely correlated: only 1 at $\alpha=0.05$, and 56 at $\alpha=0.1$ in which none of the 3 miRNAs were included (Table 8)

Table 8. Comparison of results obtained for the 3 most important miRNAs (red) and the total of 45 miRNAs (blue) with different parameter settings for the purity correction model using 2 purity markers. We demonstrate the number of significant interactions found within the validated, experimentally supported and pure predictions databases.

	Purity model: no FDR, $\alpha=0.05$		Purity model: no FDR, $\alpha=0.1$		Purity model: FDR, $\alpha=0.1$	
Validated	123	0	198	0	0	0
Experimentally supported	200	2	328	6	0	0
Pure predictions	3841	58	5981	121	56	0

On the initial “purity” analysis, we found that FCGR2A had the largest R squared among the variables, therefore we decided to add FCGR2A to the “purity correction” now using 3 markers: CD19, CD14 and FCGR2A, in another attempt to achieve a balance between adequate purity correction and number of significant results obtained for the 3 miRNAs under study. Yet this effort produced even less results that the previous 2 (Table 9).

Table 9. Comparison of results obtained for the 3 most important miRNAs (red) and the total of 45 miRNAs (blue) with different parameter settings for the purity correction model using 3 purity markers. We demonstrate the number of significant interactions found within the validated, experimentally supported and pure predictions databases.

	Purity model: no FDR, α=0.05		Purity model: no FDR, α=0.1		Purity model: FDR, α=0.1	
Validated	126	0	200	0	0	0
Experimentally supported	311	2	500	3	0	0
Pure predictions	3700	46	5700	100	4	0

In conclusion, miRNA-mRNA interaction results of this approach are not shown here since this analysis was not completed and relevant results for our objective were not achieved so far. Further investigation is needed to determine a model for purity correction analysis that is more suitable for our standards.

IV. Discussion

To identify miRNA-mRNA interactions in a specific biological context, we integrated miRNA-target predictions with miRNA and mRNA expression data produced from the same samples. We have developed an approach where each miRNA was tested for association, using Pearson correlation, with expression levels of known and putative mRNA-targets. We applied this approach to miRNAs and mRNAs expression in a cohort of SSc patients and focused on miRNAs previously shown to be associated with this disease.

We compared our results with the ones obtained using the global test proposed by van Iterson et al (2013). We found that the efficiency of our approach in identifying targets for the list of 3 miRNAs of higher interest was similar to the one of van Iterson. However, our results differed from the results obtained using the global test when investigating the whole list of 45 miRNAs. In particular, for the pure predictions list, many more significant interactions were included in our results. Despite the larger number of predicted targets obtained using our method within the pure predictions list, when considering the issue of false positives for this target database (Witkos et al, 2011), we believed that not all interactions identified could be trusted in this case. Consequently, we have recognized the need of using a more stringent significance threshold for this list of potential targets, i.e.by controlling the False Discovery Rate (Benjamini & Hochberg, 1995). Additionally, Pearson correlation compared to multiple linear regression may yield more interactions but also involves an important multiple testing problem: a pair-wise approach is more likely to be influenced by individual pairs with large associations, which may occur by chance. While in our correlation approach we had to control for the False Discovery Rate, van Iterson’s approach does not perform any correction for multiple testing. This is due to the fact that, in the case of van Iterson′s study, the number of tests performed is greatly reduced compared to ours: their model considers the complete set of predicted targets for each miRNA instead of testing every single miRNA-mRNA pair individually. Hence in their approach the multiplicity problem is not as severe.

Furthermore, we tried to tackle an implication of our study as regard the possibility to correct for possible biases in the dataset, e.g. the purity of the samples. As the purity of initial pDC samples was found to be relatively variable and even low for some samples, our results might include some associations that are not specific for pDCs but due to the impurity of the samples. Starting from the purity of pDCs in the sample and the expression of genes defining different cell types that could have contaminated our sample, we performed a multiple linear regression analysis to identify the best gene-set explaining the purity. On the basis of this analysis, we corrected our pipeline for the purity-bias, by using a linear model where miRNA expression was the dependent variable and the independent variables were each gene examined for correlation together with the set of “purity markers”. Unfortunately this approach gave us very limited number of significantly negatively correlated miRNA-mRNA pairs and did not bring the expected improvement: this could be due to the fact that the “purity markers” were correlated to each other, therefore the model was over-corrected. Future work could attempt to tackle the purity problem using a different approach such as a penalized regression model (Tibshirani, 1996), e.g. LASSO. The advantage of LASSO lies in the fact that this approach gives a relatively small set of resulting associations, since the penalty used tends to produce some coefficients that are exactly equal to 0 hence it is inherently performing feature selection (Lu et al, 2011). This is can be effective in performing purity correction since within a large number of predictors we would like to determine a smaller subset that exhibits the strongest effects. The elastic net (Zou et al, 2005) is another penalized regression model that could be used to address the purity correction problem. The elastic net performs a “grouping” of the predictors, where strongly correlated predictors tend to be in or out of the model together. Therefore this method could address the problem of purity markers correlated to each other.

The advantage of our approach entails the ability for the user to define different significance thresholds and the choice to apply or not a FDR correction when investigating targets predicted from databases of different degree of reliability. The necessity to adjust the significance settings is supported from our results: the large number of identified miRNA-mRNA pairs in the pure predictions database is greatly reduced when applying FDR correction thus producing results that are more probable to be true interactions. Another of our findings in support of the need of FDR correction in this case, is the fact that the large majority of the interactions included in the not-FDR corrected results were weak correlations. Contrary to the pure predictions target list, for the reliable databases of proven targets (validated and experimentally supported), we recommend a less stringent significance setting: no FDR correction is needed given that these databases are very restrictive but also very reliable. The number of significant results produced in this way is very limited -especially for the differentially expressed miRNAs- and the cost of including a possible false positive would be trivial in the prospect of further experimental validation. Future investigations could be also aimed at a further refinement of our results that would also facilitate later validation experiments. One way to reduce the large number of predicted targets from the pure predictions target list could be a strategy based on pathway enrichment analysis, as it would identify target genes that are most relevant for the specific biological condition of interest.

In conclusion, we propose a method able to add biological relevance to existing miRNA-target prediction using Pearson correlation between expression levels of miRNAs and mRNAs. We applied our approach in the context of SSc and we show that depending of the target list source different statistical settings should be applied. Our correlation pipeline is available as a function in R, making it possible to apply to other studies of miRNA target prediction.

References

1. Alexiou P, Maragkakis M, Hatzigeorgiou A. Online resources for microRNA analysis. Journal of Nucleic Acids Investigation 2011; 2:e4.

2. Baek D, Villen J, Shin C, et al. The impact of microRNAs on protein output. Nature 2008;455(7209): 64–71.

3. Benjamini Y, Hochberg Y, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society 1995;57:289-300.

4. Bentwich I. Prediction and validation of microRNAs and their targets. FEBS Lett 2005;579:5904-5910.

5. Bolstad BM, Irizarry RA, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-193.

6. Cho S, Jun Y, Lee S, et al. miRGator v2.0: an integrated system for functional investigation of microRNAs. Nucleic Acids Res 2011;39:D158–D162.

7. Fulci V, Colombo T, Chiaretti S, et al. Characterization of B- and T-lineage acute lymphoblastic leukemia by integrated analysis of MicroRNA and mRNA expression profiles. Genes Chromosomes Cancer 2009;48:1069–1082.

8. Gabrielli A, Avvedimento EV, Krieg T. Scleroderma. N Engl J Med 2009;360:1989-2003.

9. Gennarino VA, Sardiello M, Mutarelli M, et al. HOCTAR database: a unique resource for microRNA target prediction. Gene 2011;480:51–8.

10. Hicks S, Irizarry R. When to use Quantile normalization? Biorxiv 2014.

11. John B, Enright AJ, Aravin A, et al. Human MicroRNA targets. PLoS Biol. 2004;2(11):e363.

12. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39(10):1278–84.

13. Keselman HJ, Cribbie R, Holland B, The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparison wise Type I error control, Psychological Methods 1999;4:58-69.

14. Kiriakidou M, Nelson PT, Kouranov A, et al. A combined computational-experimental approach predicts human microRNA targets. Genes Dev. 2004;18(10):1165–78.

15. Krek A, Grun D, Poy MN, et al. Combinatorial microRNA target predictions. Nat Genet.2005;37(5):495–500.

16. Lall S, Grun D, Krek A, et al. A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol. 2006;16(5):460–71.

17. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120(1):15–20.

18. Liu H, Brannon AR, Reddy AR, et al. Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell Renal Cell Carcinoma. BMC Syst Biol 2010;4:51.

19. Lu Y, Zhou Y, et al. A Lasso regression model for the construction of microRNA-target regulatory networks. Bioinformatics 2011;27(17):2406-13.

20. Miniategui A, Pey J, Planes F, et al. Joint analysis of miRNA and mRNA expression data. Briefings in Bioinformatics 2013;14(3):263-78.

21. Miranda KC, Huynh T, Tay Y, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell. 2006;126(6):1203–17.

22. Ritchie W, Flamant S, Rasko JEJ. mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. Bioinformatics 2010;26:223–227.

23. Sales G, Coppe A, Bisognin A, et al. MAGIA, a web-based tool for miRNA and genes integrated analysis. Nucleic Acids Res 2010;38:W352–W359.

24. Setupathy P et al. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods 2006;3:881-886.

25. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res 2010; 38(17):e170.

26. Sylvain P, Johann W, Jérôme T, et al., Impact of normalization on miRNA microarray expression profiling. RNA 2009, 15: 493-501.

27. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the royal statistical society;58:267-288.

28. Van Bon L, Affandi A, Broen J, et al. Proteome-wide Analysis and CXCL4 as a Biomarker in Systemic Sclerosis. N Engl J Mead 2014;370:433-443.

29. van den Hoogen F, Khanna D, Fransen J, et al. Classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. Ann Rheum Dis 2013;72(11):1747-55.

30. Van Der Auwera I, Limame R, Van Dam P, et al. Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype. Br J Cancer 2010;103:532–541.

31. Van Iterson M, Bervoets S, et al. Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions, Nucleic Acids Res 2013;41:e146.

32. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics 2009;10(1):57-63.

33. Witkos T, Koscianska E, et al. Practical Aspects of microRNA Target Prediction. Curr Mol Med. 2011;11(2): 93–109.

34. Zou H, Hastie T, Regularization and variable selection via the elastic net. J. R. Statist. Soc. B 2005;67(2):301–320 .