ANALYSIS OF INTERSTITIAL LUNG DISEASES AND INTEGRATED ANALYSIS BETWEEN MICRO RNA AND MRNA

ENROLLMENT NUMBER : 131565

STUDENT'S NAME : NAMIT BHAGWANANI

SUPERVISOR'S NAME : DR. CHITTARANJAN ROUT

Submitted in partial fulfillment of the Degree Bachelor of Technology in Bioinformatics

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT

CERTIFICATE

This is to certify that the work entitled "**Analysis of interstitial lung diseases and integrated analysis between microRNA and mRNA "** pursued by **Namit Bhagwanani** (**131565**) in partial fulfillment for the award of degree Bachelor of Technology in Biotechnology from Jaypee University of Information Technology, Wakhnaghat has been carried out under my supervision. This part of work has not been submitted partially or wholly to any other University or Institute for the award of any degree or appreciation.

Dr. Chittranjan Rout

Assistant Professor

Department of Biotechnology and BIoinformatics

Jaypee University of Information Technology

Wakhnaghat, Solan

ACKNOWLEDGEMENT

All praise belongs to the almighty lord to whom I thank for the strength, courage and perseverance bestowed upon to me to undertake the course of the study.I hereby acknowledge with deep gratitude the cooperation and help given by all members of Jaypee University in helping with my project.

With proud privilege and profound sense of gratitude, I acknowledge my indebtedness to my guide Dr. Chittranjan Rout, Assistant professor, Jaypee University of Information and technology his valuable guidance, suggestions, constant encouragement and cooperation.I express my thanks to Prof. Rajinder Singh Chauhan, Dean, Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology.

I would also like to extend my gratitude towards Ms.Smriti Mishra, and other staff members for their constant help and motivation for successfully carrying my research work.

DATE :

PLACE :

Namit Bhagwanani

LIST OF CONTENTS

INTRODUCTION

1.1 ABSTRACT

Interstitial lung diseases are a unique group of lung disorders that damage the tissues. The alveolar epithelium, base membrane, perivascular and pulmonary capillary endothelium are all affected by these lung disorders. These disorders happen due to degeneration or damage to the lungs which results in abnormal immune response in the organism. Generally, the body produces enough amounts of fiber bundle tissues to compensate the damage caused to various parts of lungs however in interstitial lung diseases, the repair of lung tissues does not happen properly causing the tissue around the air sacs to become swollen and scarred .

Scarring of lung tissues makes the transport of oxygen in blood very difficult.

Interstitial lung disease identifies a different group of diseases which show a lot of variation than obstructive pathway diseases.

2013 showed cases of 600,231 interstitial lung diseases globally resulting in 523,112 deaths.

When we consider the fact that microRNAs in more amounts tends to change or affect the expression of a considerable part of genome and the complex biological pathway in which they function, we can understand that their deregulation has a significant repercussion on the proper functions and metabolism in an organism. An increasing number of groups of miRNAs are known to be involved in different interstitial lung diseases that have hugely enabled us to better understand their metabolic pathways and stages of maturation.

1.2 OBJECTIVE

MicroRNAs “miRNA” are very small RNA molecules that do not have coding regions and they also are responsible for negatively regulating the gene expression. Not only are they very responsive molecules but they also participate in performing significant cell functions.

They play a big role in causing the disorders such as pulmonary fibrosis, chronic obstructive pulmonary diseases and severe asthma.

If we can understand the specific roles that these microRNAs have in these disorders, it can help us in developing better diagnostics and medical therapies. Our study sheds new light on the role of some microRNAs in different lung diseases and the most probable techniques that can be developed in upcoming years which will help us in clinical diagnostics.

We will use two expression data sets to find out the correlation between miRNA and mRNa and their impact on interstitial lung diseases.

BACKGROUND

2.1 chILD

Significant findings have culminated in the last 10 years which have better helped us to pin point the most significant causes for interstitial lung disease. It has been observed that disorders in the lung tissue most probable starts in stages of infancy and kept growing despite maximum medical approaches and diagnostics. Interstitial Lung Diseases were often known to show that the genetic mechanisms could be a significant factor in causing cancers and other forms of diseases in lung tissues in children.

These highly varied class of lung disorders was first studied and then grouped using a model schema. This was done by considering lung disorders in adult groups and the nature of the biological networks in the lungs. Some varied forms of surfactant molecules were annoyed showing that interstitial lung diseases are not spontaneous in nature .We acknowledge the fact that very unique set of genetic changes and metabolism cause forms of chILD and should be sent for specific non-invasive diagnostics, apprehending the Relatives and friends about the concerns of recurrence in the tissues, predicting evolutionary information, and performing a schema analysis that relies on internal functioning and mechanisms of disease. Diseased Samples help us in recognizing the normal metabolisms in the lungs and identifying the major causes which may mature into any type of lung disorders or pulmonary fibrosis.

Gene expression in single gene diseases generate specific protein sequences that play an important role in the proper functioning of surfactant but there is a very high chance that the genes that directly cause lung diseases continue to mature.

2.2 SURFACTANT METABOLIC DYSFUNCTION DISORDERS

Mutation caused in some genes encode 3 different proteins with important significant mechanism in surfactant function, SP-B protein, SP-C protein, and a functional protein A3 of ATP-binding family of transporters. They are shown to have resulted in diseases with very identical clinical and metabolic pathways in lungs. SP-B (79 base pair) amino acid is a special class of water repelling protein that is secreted by only one gene on chromosome 2 of SFTPB.

This protein is further known to product of much bigger class of protein, the expressed SP-B peptide chain which is available to us and is produced by post-translational photolytic processing at N and C Terminals respectively.

A 35 base pair amino acid termed as SP_C is also a very hydrophobic protein produced by a minute gene. This is present on chromosome 8 in gene SFTPC.

Expressed SP-C is produced by post-translational proteolytic processing at both the N and C terminal of evolutionary older protein proSP-C much like that of SP-B.

The two surfactants SP-B and SP-C are found in surfactants which have been deducted from mammals. These surfactants are used for treatments of infant children with ILDs such as replacement therapy etc., and are significant for better functioning of these therapies.

A member of family of transporters called ABCA3 is known to hydrolyze the ATP molecules to transport molecules inside biological membranes. An amino acid protein of one thousand and seventy four contains two membrane spanning and two nucleotide binding domains.

A huge varieties of tissues shows the expression of ABCA3 but it is mostly expressed in lungs in which it functions only inside the lamellar bodies in the alveolar cells of the lungs.

A lot of members of this family of proteins is known to transport the lipids inside the body

so there is a very high chance that ‘ABCA3’ also performs the movement of lipids who are very important for biological functions in the surfactant molecules.

Newborns who are deficient in ABCA3 showed variable signs of reduced surface tension and very specific amounts of surfactant lipids such as DSPC and PG in lung fluid.

2.3 Pulmonary Surfactant Components and their Metabolism

Pulmonary surfactant can be termed as combination of proteins and phospholipids which are required for the reduction in surface tension in the alveoli and also prevent apoptosis. Slow or inadequate generation of surfactant molecules is the known to be the most significant cause of the RDS in premature babies. These lung disorders are caused in infants when the surfactant generation comes to a halt due to specific genetic mechanisms that are known to resemble respiratory distress syndrome in premature babies.

The pulmonary surfactant is expressed in alveolar cells. It is then compressed into organelles produced from lysosomes which are called as lamellar bodies. Disaturated phosphatidylcholine is the most significant lipid present on the surfactant as it is known to reduce surface tension due to its nature. This class of lipid is known to slowly adsorb to the interface present. Surfactant proteins SP-B and SP-C who are known to be highly hydrophobic in nature, pertains the significant properties upon surfactant lipids that reduces the alveolar surface tension. It constitutes 2 different water repelling and closely related proteins (SP-A and SP-D) structurally who are a part of the collectin family and have important function in immunity. It is then reverted into type II cells. This alveolar macrophage matures with time.

The maturity corresponds to the GM-CSF whose function is to bind to the surface of the macrophage.

Surfactant dysfunction happens when the lungs are unable to produce the SP-B protein inaccurately.

The affected newborns have symptoms that increase dramatically and lung disease progresses as soon as the child is born. They also depict the lung disease that matches RDS in premature babies. The disease is grows so fast that the rate of mortality in the infants is reduced to a few months even after appropriate medical treatment is available including change of surfactant. Diagnosis and treatment is done after correctly identifying the factors that cause the disease. This can be done by prediction of changes in the gene which causes the disorder.

-“A frameshift mutation resulting in insertion of 2 bases into codon 121 and coined 121ins2 is the most widely found SFTPB mutation, has accounted for about 2/3rd of the mutant alleles identified to date, and its occurrence in distant subjects is due to common evolutionary ancestral origin.”

At the moment the only effective and recognizable way to treat the worst cases of ILDs in infants is lung transplantation.

These newborns are unable to encode the SP-B class of protein. There have been a few occurrences in which the infants expressed very light symptoms of ILDs as they are somehow able to encode for a small amount of SP-B protein. They have been shown to thrive for years without proper medical treatment.

At the onset of twentieth century some cases of ILDs were reported which may have been caused by some *specific mutations. Individuals who were affected by these mutations had variable signs of maturity and the extent of the. The newborns also faced similar problems including difficulty to inhale in stable conditional environment and low rate of survival. One such example of this mutation is *mutation.

This specific change is present in too many distant relatives causes 30 percent of the total mutations in genes in recent memory. There are major reasons that cause the onset of the disorder is the development of a mechanism which causes abrupt functioning in the protein which results in aggregation and improper folds in the structure . All of the above mentioned activities lead to general instability in the protein causing tension in the reticulum.

Efficient therapies for individuals with SFTPC mutations are still not known. Therapeutic lung lavage in infancy, corticosteroids, and hydroxychloroquine have been used to improve the clinical condition in case patients, but the highly variable history of the disease makes deductions from these uncontrolled observations problematic. No randomized studies of these treatments have been reported. Lung transplant has been performed in individuals with progressive degradation in lung function.

Several mutations have been studied and a schema classification has been proposed based on mutations that either preclude ABCA3 production or intracellular transport or stop the ability of protein to bind or hydrolyze ATP or transport phospholipids through membranes .However 1

specific mutation-p.E292V or c.875A>T-has been expressed in multiple unrelated children with much milder disease and the phenotype of genetic disorder of chILD. Some studies show that this mutation results in less functional restriction in ABCA3 function than other type II mutations. These findings show that retained function may progress disease severity and that the genotype may predict phenotype to some extent and even some boost in ABCA3 production should improve the clinical condition of such patients.

*for reference

1.” SFTPC mutations”

2.” c.218T>C, p.I73T”

2.4 GM-CSF RECEPTOR DEFICIENCY

One important or significant feature of inefficient surfactant function is the aggregation of granular, eosinophilic substance in distant airspaces. This study shows the resemblance of pulmonary alveolar proteinases' with that of adults. Pulmonary alveolar proteinases' shows signs of disease which mature at a very slow rate, this is due to the fluids and organelles that occupy a major part of the spaces in lungs. Most of the metamorphic order of lungs is extremely non conserved and the surfactant disease corresponds with onset of pulmonary fibrosis.

Several measures are taken in medical advances to prevent the protein rich fluid from invading the alveolar spaces of infants.

There is a unique receptor molecule which is made up of two small units called the alpha and beta chains. These chains also pair with receptors in MK 8-9. The GM-CSF is activated when it interacts with these receptor molecules. There have been several difficulties in the identification of the mutation in gene that encodes the beta strand. This is because the physical dataset of the infants could not relate to the pulmonary alveolar proteinosis. Strong evidence suggested that the gene that encode the alpha strand showed mutational problems and this could lead to alveolar proteinosis in babies.These events helped us to conclude that defects in genes could lead to onset of alveolar proteinosis in infancy. It is stressful to find the course of action for the disease as its features of maturity and biological metabolism are not available. Lung pathology therefore is more specific as when used as treatment for various types of Interstitial lung diseases.

2.5 ALVEOLAR CAPILLARY DYSPLASIA

It has been reported that the prematured pulmonary veins and capillaries are present in similar regions as the arteries instead of being present in its designated system of lymph nodes. This causes alveolar capillary dysplasia. Affected babies who are down and show high symptoms of pulmonary stress are not recognized to show proper response to medical therapy and naturally succumb to the disease. Sometimes rarely milder cases whose onset comes later have been presented and a higher normality is observed in infants. Treatment procedure is carried out mostly through thorough analysis and diagnostic of tissues present in lungs, cardiac characterization can also be carried out. Most cases of progressive lung disorders that have been calculated show that fifteen out of thirty four extreme cases of ILDs have been seen in Great Britain. This has been directly related to ACD. Extra-pulmonary disorderss were reported in 55–78% of the cases hinting at the most probable reason of genetic defects.

*for reference

-“Very recently microdeletions were found in a group of people with lung pathological findings of ACD along with other disorders which include cardiac and genitourinary deformities in fox (forkhead) family of transcription factors and sequence analysis revealed heterozygous-loss-of-function(FoxF1) mutation in 20 of 84 patients with ACD examined which signified the role transcription factor in the pathogenesis of the pulmonary phenotype but the mechanism still remains a mystery.”

Table 1.

Known Genetic Causes of ChILD Syndrome

	Alveolar capillary dysplasia	*SP-B deficiency*	*ABCA3 deficiency*	*SP-C dysfunction*	*Brain–thyroid–lung syndrome*	*GM-CSF receptor deficiency, α chain*	*Lysinuric protein intolerance*
OMIM	#265380	#265120 SMDP1	#610921 SMDP3	#610913 SMDP2	#610978	#300770 SMDP4	#222700
Locus	FoxF1	SFTPB	ABCA3	SFTPC	TTF1 (Nkx2.1)	CSF2RA	SLC7A7
Chromosomal location	16q24.1	2p12-p11.2	16p13.3	8p21	14q13.3	Xp22.32, Yp11.3	14q11.2
Inheritance	AD, sporadic	AR	AR	AD, sporadic	AD, sporadic	AR	AR
Mechanism	Haplo insufficiency	Loss of function	Loss of function	Toxic gain of function	Haplo insufficiency	Loss of function	Loss of function
Onset of pulmonary symptoms	Neonate	Neonate	Neonate, infancy, childhood	Neonate < infancy to adult	Newborn, infancy	Childhood	Infancy to childhood
Principal histology	ACD/MPV	SDM	SDM	SDM	SDM	PAP	PAP
Other findings	Cardiac, GI, or GU malformations				Hypothyroidism Neurological		Hyperammonemia Vomiting Failure to thrive
Course	Severe, fatal	Severe, fatal	Variable	Variable	Variable		Progressive

Table 1 : Genetic causes for ILDs

GENETIC APPROACH TO DIAGNOSTICS

The table above depicts information of genes corresponding to its relative interstitial lung disorder in infants. It is important to understand the specific functions of genes that cause ILDs as it helps us to perform better diagnostics. Following up on the above mentioned action promotes efficiency in choosing from varied range of therapies so as to minimize cost and risk and also to suggest optimal methods for treatment such as lung transplant in extreme cases. There is a very high possibility that the disease is not caused by a defect in the genetic mechanism,and considering the expensive treatment that is available, therefore there is not sufficient data on the genes related to ILDs as current approaches are not very efficient. . Rapidly progressive disorders show complete maturation by the time the diagnosis is complete and genetic tests come forwards therefore the most efficient and probable procedure for diagnosis till date is lung biopsy.

3.0 CORRELATION BETWEEN MICRO RNA AND MRNA

MicroRNA or miRNA refers to a form of RNA with non codin regions with a length of twenty two base pairs.

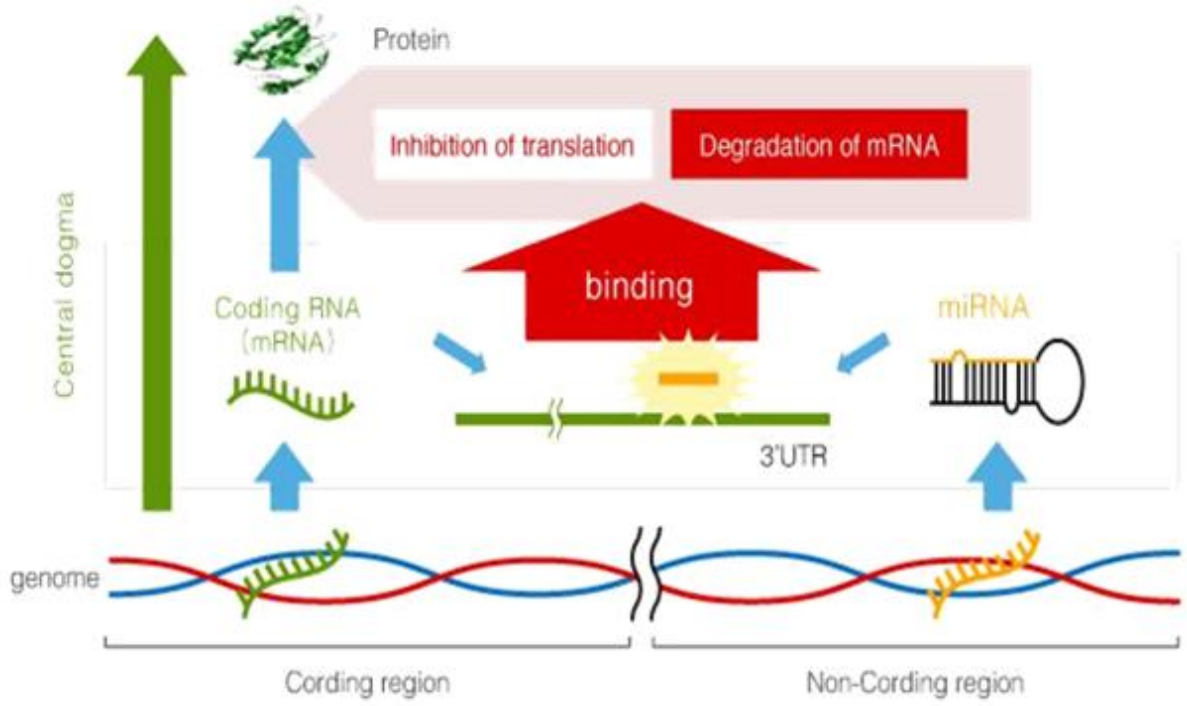
It specifically interacts and then binds with a sequence which is complimentary in nature. It binds in the 3' UTR prime end of mRNA that corresponds to the coding region of RNA. It therefore is known to regulate the gene expression.

It has been deducted that 30-50% of all genes are regulated by the correlation between these two types of RNA.

The microRNA has two forms of regulatory mechanisms:

(i).Degradation in the transcript

(ii).Slowing the process rate of translation.



The regulation mechanism at times is known to show high flexibility, this is when its complementarity with the 3'end sequence is more than 70 percent. This causes gene regulation by the microRNA. MicroRNA and mRNA can regulate each other in high amounts. Correlation between mRNAs and miRNAs for regulation isn't simple but actually very comples corresponding to many-to-many relationship.

3.1 IN SILICO ANALYSIS OF EXPRESSION DATA SETS

TALASSO: Two different datasets were used showing expression ranges in human mmicroRNAs and mRNAs and talasso algorithm was used.

Best or better ranking interactions resulting from TaLasso are very specifically enriched in cross validated targets.

Other algorithms don't provide the same insight. TaLasso is available as R code. TaLasso is being validated because it has special features such as predicting the cross validated putative targets accurately and calculating how relevant the predicted interactions can be. Also it is better to consider the functional aspect of genes in high ranked features rather than referring to experimentally derived ones. It has also been seen that incorporating non +ve restrictions mostly advances the specificity while using Lasso regression analysis for prediction.

When we are using datasets of interactions in microrna and mRNA , their expression scores, the TaLasso algorithm starts calculating the nature of down regulation of every miRNA corresponding to its mrna targets..

Due to the complex biological and metabolical functionality it has to be made less complex by simplification and normalization.

DATASET AND METHODOLGY

ASSUMPTIONS MADE:

Our assumptions are made in accordance to making our computations much easier. We start by assuming thar mRNa expression is solely regulated by microRNAs. This leads us to the conclusion that only the most significant interactions will be depicted by the talasso algorithm.

We also have to assume that no microRNAs can be present as *TFs.
This is because we can only allow miRNAs to cause down regulation in accordance with its putative RNAs to ensure the stability of the model that is being developed.

Then TaLasso will only quantify the down-regulation effect on those miRNA-mRNA interactions from an initial set of putative miRNA-mRNA pairs (i.e. predicted from sequence based algorithms). Consequently, TaLasso will not be able to recover those interactions not included in this initial set of putative targets.

4.1 EXPRESSION DATASETS FOR ANALYSIS

TaLasso AND MAGIA were analysed on two specific data sets that were downloaded from GEO database.

1.GPL16770 -031181 Unrestricted human affymetrix Microarray

2.GPL20907 -021827 Human miRNA Microarray from MirBase

To effectively perform the encrichment analysis ,ranking of the interations between mircrornas and mrnas was done based on the scores. Then we took the highest ranked features in the and counted the experimentally croass validated putative targets . After performing the above computations we use specific distribution in order to compute p-values. The hypergeometric distribution is a discrete probability distribution that describes the probability of obtaining *p* successes (experimentally-validated interactions) when *n* elements (selected interactions) from a finite population without replacement (the union of all the putative interactions) are drawn.

The values that we comuted can be used for the comparison between algorithms.
For example: the lower the value of p value in our interactive system ,more enrichment will be observed in validated targets.

We compared the algorithms using two scores: the number of experimentally validated interactions in the top-500 predicted interactions and the minimum p-value on the enrichment curves. We then took the experimentally analysed targets and the number of predicted interactions found and included them in the test as well.

4.2 IMPLEMENTATION

To get the aforementioned results we used an online development tool called Talasso which uses algorithms such as GenMiR++ and correlation.

We choose two specific expression data files of microRNA ans mRNA respectively. These files are tab separated, one with mRNA expression data and the other with miRNA expression data. The user can then perform various types of analysis be specifically choosing the database, the units of miRNA data (ΔΔCts or expression values) and the algorithm to score the interactions.

The resulting targets are sorted by computed score and it is also shown if they are included in the given experimental validated databases.

4.3 METHODOLOGY:

Correlation between microRNA and mRNA

Download expression matrix files from GEO

Analyse in G2R for Fold changes

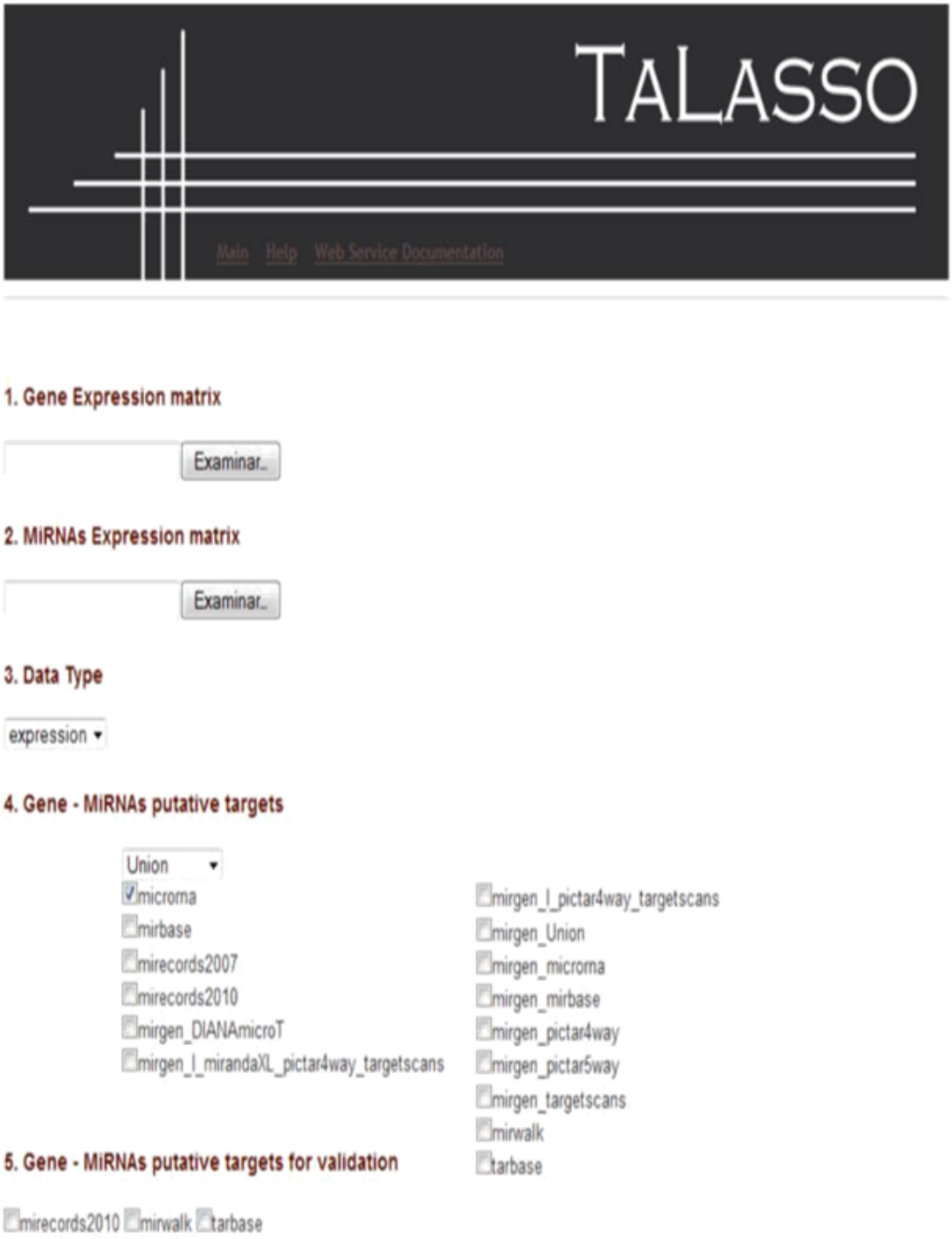
Upload matrices in Analyzation tools :

1.MAGIA

2.TaLasso

Devise correlation and stringency through scores

Optimization of results



The first step is uploading the expression data set file for microrna and mrna. To avoid any form of errors it should be unsured that both files are tab delimited both should contain names followed by gene IDs of miRNA names in two columns. The uploading is made by selecting the text files in points 1 and 2 of figure 4.

1) microRNA files generally consist of its expression values.

2) We choose and select from different interacted targets in the databses that has been made available to us. It must be also determined whether the putative interactions will be the union or the intersection of the selected databases.

3) We also have to ensure that the selected targets go through a statistical analysis .(only in case union option is selected and if the corresponding experimentally validated database has been selected on point 4).

4) Finally, users have to select the method for interaction scoring: Pearson Correlation, GenMiR++ or TaLasso (step 6). For this last method, the tuning parameter has to be determined: global (a general one for all the mRNAs) or local (one for each mRNA) and its value (step 7).

After running the talasso algorithm the results will be showed in a html web . The hierarchy of the model is relatively easy to understand:

1) The calculated scores are arranged from highest to lowest in a decreasing manner. These are the scores of the interactions

2) The results are depicted in a tabular format with specific images shown next to the genes present in our dataset. Interaction with these images helps us to annotate all micrornas who down regulate the target rna. It also shows the regulated genes by the microRNA. Clicking on the image next to the gene name shows all its functional and structural aspects by referring to ensemble database.

3) There are three types of databses that contain all the information about the interactions. IT can be checked which databse they belong to i.e. TaRBase or GEnMIR, MiRBase etc.

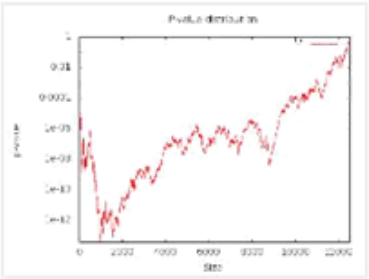
4) The p-values of the multiple linear regression of the results are placed in the column pValue.

The figure on the top of the results page corresponds to the hypergeometric test on the experimentally validated targets selected and a hit distribution map of the interactions. The images become bigger by clicking on them.

Results for geneExpression-134

Download results files: [Targets] [Genes] [MiRNAs]

Validation of results (tarbase)



Hits distribution map (tarbase)

Table of results

First [1] >> Last

Gene	miRNA	Score	pValue	Mirecords	Mirwalk	Tarbase
SLC7A7 [ENSG00000155465]	hsa-miR-205	0.076312	0.0024005			
APOE [ENSG00000130203]	hsa-miR-1	0.056631	0.00094027		✓	
FXYD2 [ENSG00000137731]	hsa-miR-205	0.025325	0.0011771			
ZCCHC7 [ENSG00000147905]	hsa-miR-130a	0.02436	0.07315			
FAM83D [ENSG00000101447]	hsa-miR-99a	0.021542	0.087336			
GPX2 [ENSG00000176153]	hsa-miR-205	0.019674	0.0024048			
ACPP [ENSG00000014257]	hsa-miR-10a	0.019504	0.00027084			
ALDH3B1 [ENSG00000006534]	hsa-miR-205	0.015956	0.0018284			
CTSA [ENSG00000064601]	hsa-miR-205	0.013761	0.0003944			
CARHSP1 [ENSG00000153048]	hsa-miR-1	0.013555	7.5497e-05		✓	

From the tarbase analysis we found that microRNAs from family of hsa-mir-1 are targeting the genes who are known for de regulation of mRNAs and therefore causing interstitial lung diseases.

4.4 RUNNING TALASSO IN MATLAB AND R

TaLasso can be run using Matlab or R. Both R and Matlab share the folder structure.

These folders are stored in TaLasso and TaLasso_R folders for Matlab and R softwares respectively.

- The experimental data must be included within a folder in the "*data*" folder. The name of the folder is used to describe the particular experimental data.

- The folder *library* contains the putative targets databases, and gene and miRNA names.

- The folder *code* contains all the code functions (*.m or *.R files). „main.m" or „main.R" are the only code to be executed to run TaLasso algorithm.

- Once TaLasso is run, the algorithm will save the results on the folder *result*. TaLasso assign to each solution a particular name indicating: 1) the value of *Gamma*, 2) data normalization method (*M* = median), 3) the type of tuning parameter used (*global* or *local*), 4) the name of the experimental data (i.e. *MCC* or *LDS*) and 5) the tuning parameters" factor used.

THE STRUCTURE FOR EXPRESSION DATA SETS

MRNA and miRNA expressions data for each experiment must be a tab delimited text file named *geneExpression.txt* and *mirnaExpression.txt* respectively. The first row on each file must contain sample names (with no free spaces) and the first column must contain miRNA (miRBase ID) or gene (Ensembl ID) names.

RUNNING TALASSO

The only code to be executed to run TaLasso in both softwares is the file „main.m‟ or „main.R‟. However, „Rcplex‟ package must be installed before R code is run. This installation is a little bit involved. First of all there must be a running copy of Cplex. Afterwards, the user must also download and install the „cplex‟ software. Once „cplex‟ is installed, the „Rcplex‟ package must be compiled.

The code included on „main.m‟ is shown in the following paragraph (is almost identical for main.R). Each of the input parameters are:

1) *Directory* = is the directory where the folder TaLasso has been placed.

2) *Data* = is the folder name where the files „geneExpression.txt‟and „mirnaExpression.txt‟ for the experiment to analyze are stored.

3) *Data_Type* = refers to miRNA data: „expression‟ in case miRNA data corresponds to expression levels and „DDCt‟ in case miRNA data refers to ΔΔCt data (qPCR experiment).

4) *databases_num* = is a vector of numbers referring to the databases chosen and from the list on„main.m‟.

5) *option* = a variable indicating whether the putative targets matrix must be the union („union‟) or the intersection („intersect‟) of those chosen on databases_num.

6) *TuningFactor_Type* = indicates if TaLasso must use „global‟ or „local‟ tuning parameters

7) TuningFactor = is the value of the tuning factor to use.

ANALYSIS AND PROTOCOLS

5.1 PROTOCOLS

After getting putative microRNA and mRNA interactions we decided to find out the roles of the microRNAs in the development of interstitial lung diseases.
Since there is very scarce data available for Lung tissues at molecular levels,we will use integrated bioinformatics tools to identify important microRNas that cause lung diseases. We used a set protocol for functional analysis of regulated microRNAs in interstitial lung diseases.

The protocols were as follows:

1. Using five specific target prediction tools and then combining those results. The tools used are PicTar, MiRDB, TargetScan, Diana-microT and miRanda.

2. Information derived from expression analysis performed before.

3. Biological pathway and enrichment analysis of interactions in datasets.

4. Using the results to link with interstitial lung diseases.

5.2 APPROACHES FOR INTEGRATIVE ANALYSIS

To get the required microRNAs we used the different analysis tools to find out the interactions or RNAs at sequential level.

We matched our target search with the databases available such as MiRDB.

The overlapping results were then used to infer the most important microRNAs that cause the disease.The protocol is described as follows:

(i) We collected datasets of microRNA expression files for different lung diseases.

(ii) Then we performed target analysis using the diferent tools.

(iii) Intersection of results and identification of new targets with experimentally validated ones.

(iv) Optimization by calculating binding energies and pathway enrichment Analysis.

RESULTS AND CALCULATIONS

GPL16770 Mrna DATASET RESULT

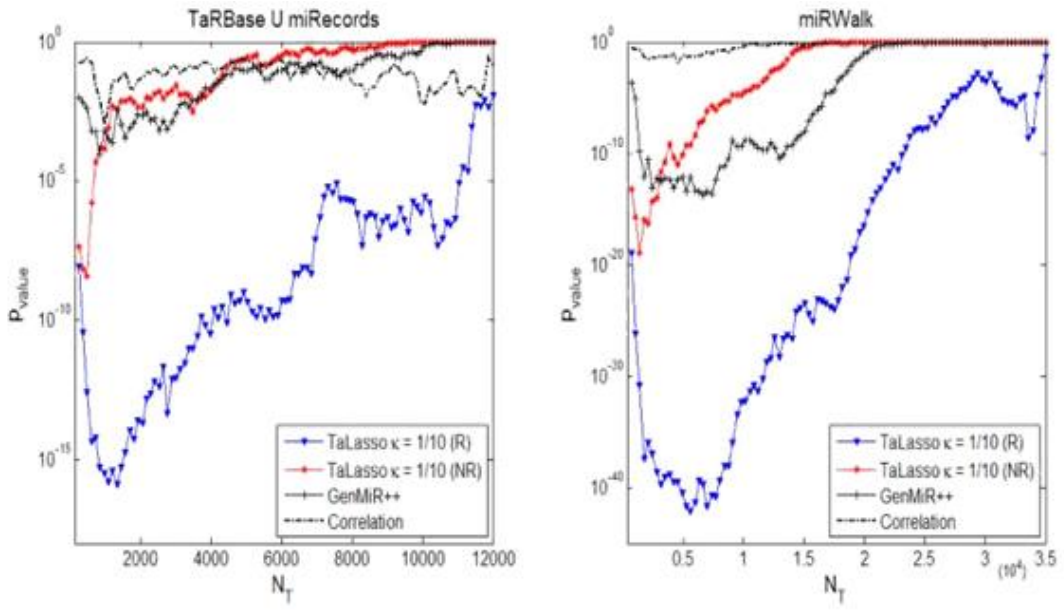


Figure 1: Enrichment analysis and down regulation effect on GPL16770

GPL20907 MICROrna DATASET RESULT

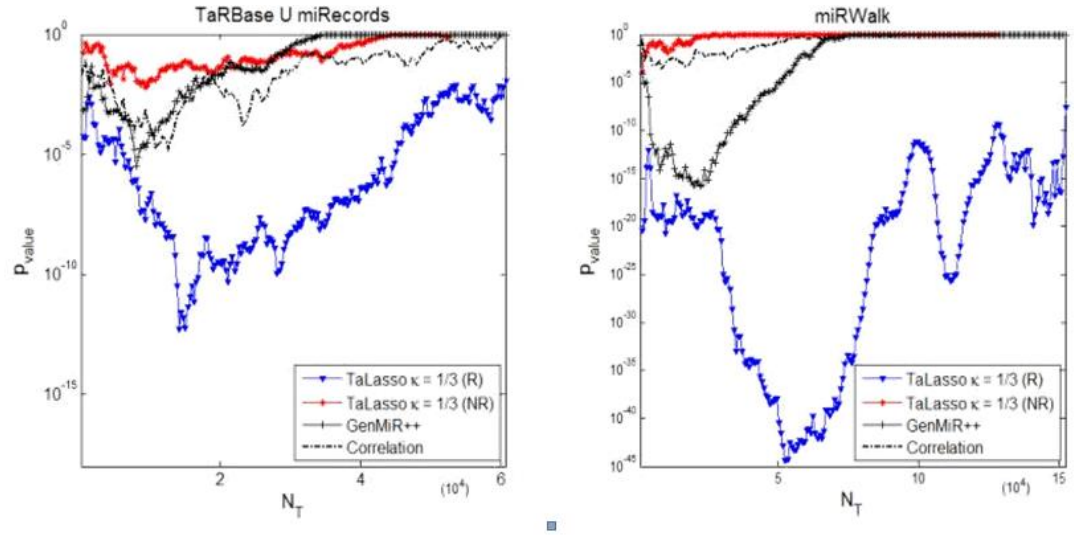


Figure 2: Enrichment analysis and down regulation effect on GPL20907

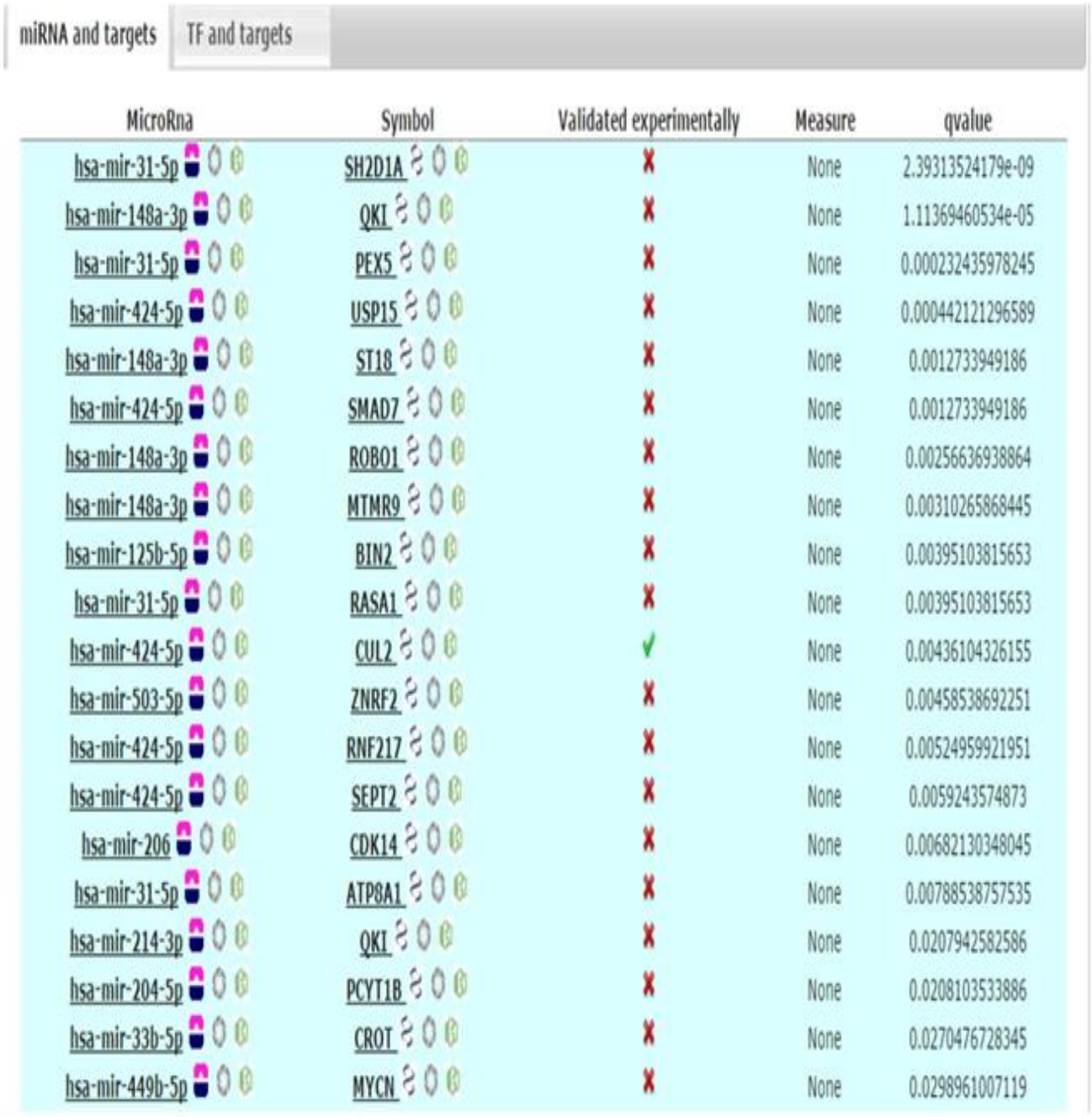
GPL20907 DATASET RESULT

Both these graphs depict the mechanism of down regulation by microRNA in its corresponding mRNA targets. They also show how enriched the pathways are.

The datasets were scanned in different available databses of RNA expressions.

miRNA and targets | TF and targets

MicroRna	Symbol	Validated experimentally	Measure	qvalue
hsa-mir-31-5p	SH2D1A	✗	None	2.39313524179e-09
hsa-mir-148a-3p	QKI	✗	None	1.11369460534e-05
hsa-mir-31-5p	PEX5	✗	None	0.000232435978245
hsa-mir-424-5p	USP15	✗	None	0.000442121296589
hsa-mir-148a-3p	ST18	✗	None	0.0012733949186
hsa-mir-424-5p	SMAD7	✗	None	0.0012733949186
hsa-mir-148a-3p	ROBO1	✗	None	0.00256636938864
hsa-mir-148a-3p	MTMR9	✗	None	0.00310265868445
hsa-mir-125b-5p	BIN2	✗	None	0.00395103815653
hsa-mir-31-5p	RASA1	✗	None	0.00395103815653
hsa-mir-424-5p	CUL2	✓	None	0.00436104326155
hsa-mir-503-5p	ZNRF2	✗	None	0.00458538692251
hsa-mir-424-5p	RNF217	✗	None	0.00524959921951
hsa-mir-424-5p	SEPT2	✗	None	0.0059243574873
hsa-mir-206	CDK14	✗	None	0.00682130348045
hsa-mir-31-5p	ATP8A1	✗	None	0.00788538757535
hsa-mir-214-3p	QKI	✗	None	0.0207942582586
hsa-mir-204-5p	PCYT1B	✗	None	0.0208103533886
hsa-mir-33b-5p	CROT	✗	None	0.0270476728345
hsa-mir-449b-5p	MYCN	✗	None	0.0298961007119



miRNA and targets | TF and targets

MicroRna	Symbol	Validated experimentally	Measure	qvalue
hsa-mir-31-5p	SH2D1A	✗	None	2.39313524179e-09
hsa-mir-148a-3p	QKI	✗	None	1.11369460534e-05
hsa-mir-31-5p	PEX5	✗	None	0.000232435978245
hsa-mir-424-5p	USP15	✗	None	0.000442121296589
hsa-mir-148a-3p	ST18	✗	None	0.0012733949186
hsa-mir-424-5p	SMAD7	✗	None	0.0012733949186
hsa-mir-148a-3p	ROBO1	✗	None	0.00256636938864
hsa-mir-148a-3p	MTMR9	✗	None	0.00310265868445
hsa-mir-125b-5p	BIN2	✗	None	0.00395103815653
hsa-mir-31-5p	RASA1	✗	None	0.00395103815653
hsa-mir-424-5p	CUL2	✓	None	0.00436104326155
hsa-mir-503-5p	ZNRF2	✗	None	0.00458538692251
hsa-mir-424-5p	RNF217	✗	None	0.00524959921951
hsa-mir-424-5p	SEPT2	✗	None	0.0059243574873
hsa-mir-206	CDK14	✗	None	0.00682130348045
hsa-mir-31-5p	ATP8A1	✗	None	0.00788538757535
hsa-mir-214-3p	QKI	✗	None	0.0207942582586
hsa-mir-204-5p	PCYT1B	✗	None	0.0208103533886
hsa-mir-33b-5p	CROT	✗	None	0.0270476728345
hsa-mir-449b-5p	MYCN	✗	None	0.0298961007119

Figure : 3 has-mir-424-5p binds to CUL2 gene.

The above two results show that has-mir-424-5p and has-mir-34-5p are two microRNAs that down regulate the Human RNA datasets. Their symbol is also given in the result table.We will now use the annotated families of these microRNAs to conclude our analysis by searching the functions of these specific class of microRNAs.

The mir-34 and mier-424 belong to a family of micro RNAs who have been known to cause various forms of cancers mainly the lungs.

The de-regulation of these microRNAs cause tumor as they are involved in the tumor suppresent network. These microRNAs will now be further analysed and correlated o other families who are responsible for disorders in the interstitial lung diseases.

6.1 INTEGRATED ANALYSIS OF MICRORNAs

We used diiferent bioinformatics tools to overlap our results for comparing with new validated sets and their annotation so that significant disease causing microRNAs can be identified.

6.2 TargetSCAN

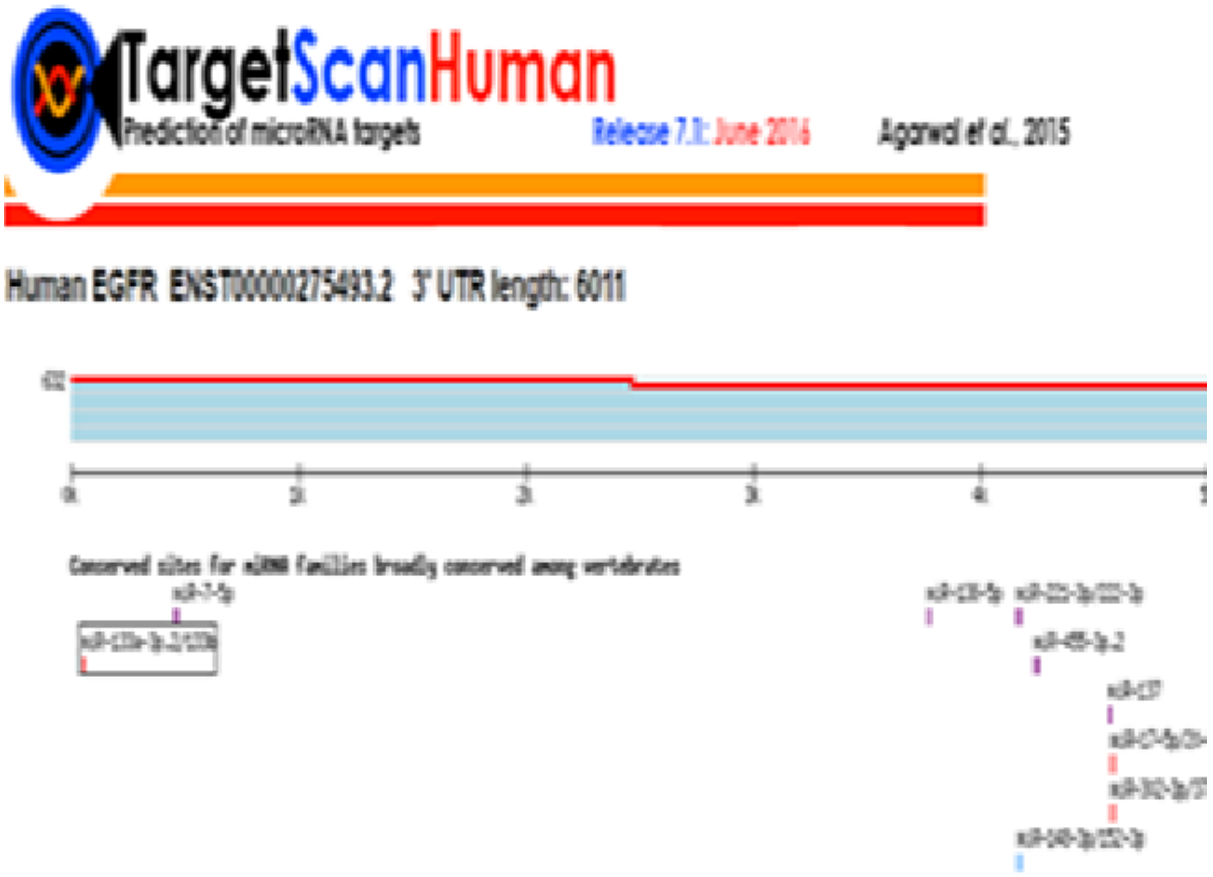


Figure: 4

TargetScan predicts biological targets of miRNAs by searching for the presence of conserved 8mer, 7mer, and 6mer sites that match the seed region of each miRNA. TargetScanHuman considers matches to human 3' UTRs and their orthologs, as defined by UCSC whole-genome alignments.

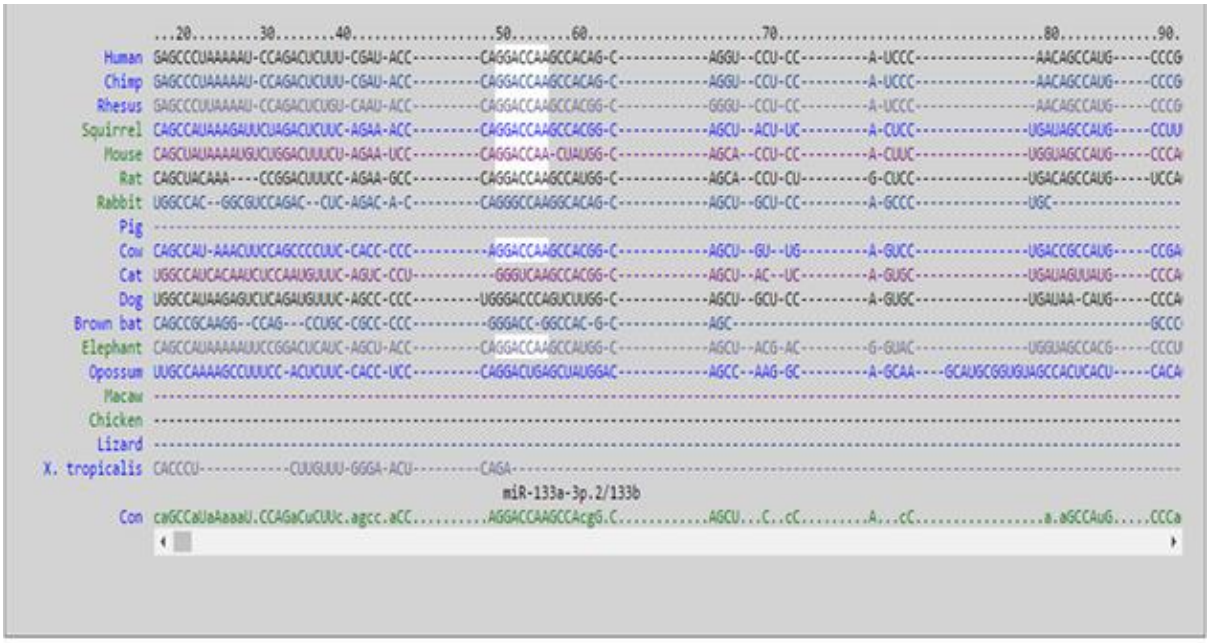


Figure 5: the conserved regions of microRNAs between mammals.

TargetScan is used to predict the targets for miRNAs by identifying different l-mer sites that match the conserved regions. It has a search option with databases including vertebrates and non vertebrates.

Mammalian analysis is also provide with designated ranks according to similarity I the conserved regions.

6.3 PicTar

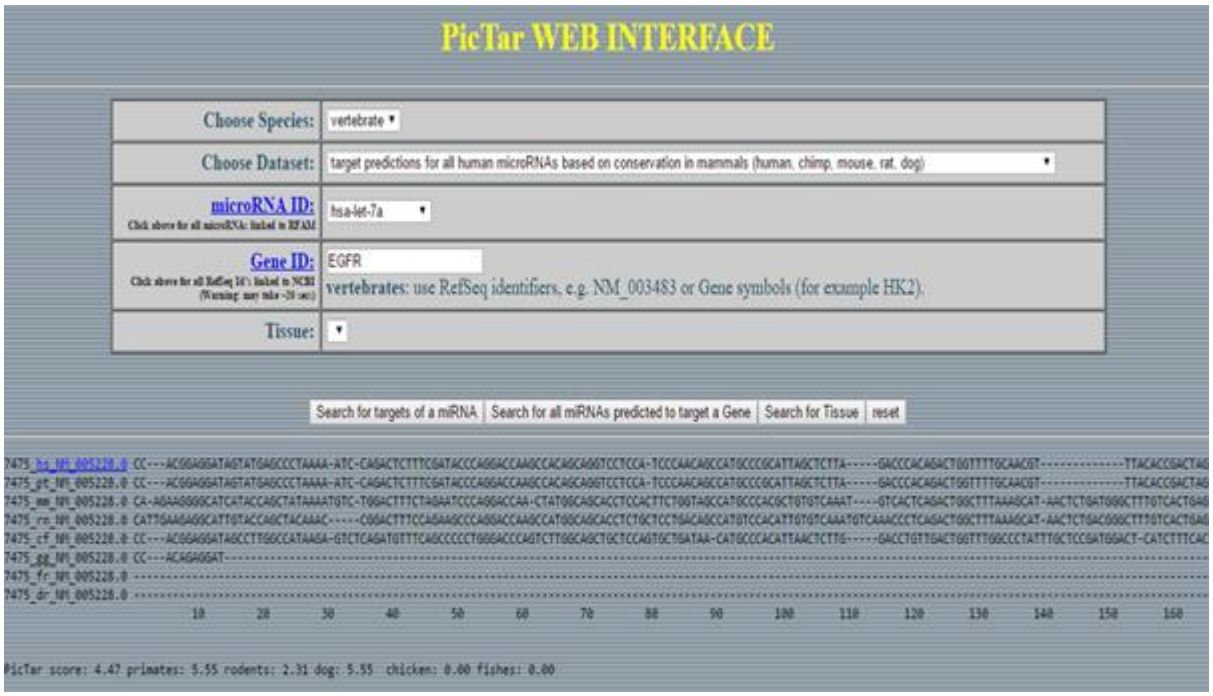


Figure: 6

PicTar is also an algorithm that has been designed to identify microRNA targets.

Here we performed the analysis by using Epidermal Growth Factor gene on the microRNA databse to see the conserved regions to predict the target regions and specific micriRNAs.

6.4 MiRDB

This is a microRNA prediction Database which accurately identifies targets for mRNa datasets. It also provides functional annotations of all microRNAs such as their sequence and other parameters

Gene 1956 is predicted to be targeted by 127 miRNAs in miRDB.

Target Detail	Target Rank	Target Score	miRNA Name	Gene Symbol	Gene Description
Details	1	100	hsa-miR-141-5p	EGFR	epidermal growth factor receptor
Details	2	98	hsa-miR-514a-3p	EGFR	epidermal growth factor receptor
Details	3	98	hsa-miR-514b-3p	EGFR	epidermal growth factor receptor
Details	4	96	hsa-miR-4328	EGFR	epidermal growth factor receptor
Details	5	96	hsa-miR-5093	EGFR	epidermal growth factor receptor
Details	6	94	hsa-miR-551b-5p	EGFR	epidermal growth factor receptor
Details	7	91	hsa-miR-302e	EGFR	epidermal growth factor receptor
Details	8	91	hsa-miR-6876-5p	EGFR	epidermal growth factor receptor
Details	9	91	hsa-miR-4476	EGFR	epidermal growth factor receptor
Details	10	90	hsa-miR-133b	EGFR	epidermal growth factor receptor
Details	11	90	hsa-miR-4504	EGFR	epidermal growth factor receptor
Details	12	90	hsa-miR-133a-3p	EGFR	epidermal growth factor receptor
Details	13	87	hsa-miR-6875-3p	EGFR	epidermal growth factor receptor
Details	14	87	hsa-miR-27b-3p	EGFR	epidermal growth factor receptor
Details	15	87	hsa-miR-27a-3p	EGFR	epidermal growth factor receptor
Details	16	86	hsa-miR-6778-5p	EGFR	epidermal growth factor receptor
Details	17	86	hsa-miR-1233-5p	EGFR	epidermal growth factor receptor
Details	18	86	hsa-miR-8066	EGFR	epidermal growth factor receptor

Figure:7

We used EGFR gene which was targeted by 127 miRNAs in the databse.

The highest ranked microRNAs were then selected among all the databases and

Their optimization was done such as calculating the binding energy and enrichment analysis of their pathways for annotation of those microRNAs.

6.5 DIANA micro-T

Diana performs target based prediction methods and databses of micro-t which include validated microRNA putative targets of coding and non coding regions.

We performed the analysis using EGFR gene and the available databse to search for microRNA which down regulates the mRNAs.

The results showed a member has-mir-579-3p which was binding to the target and affected regulation.

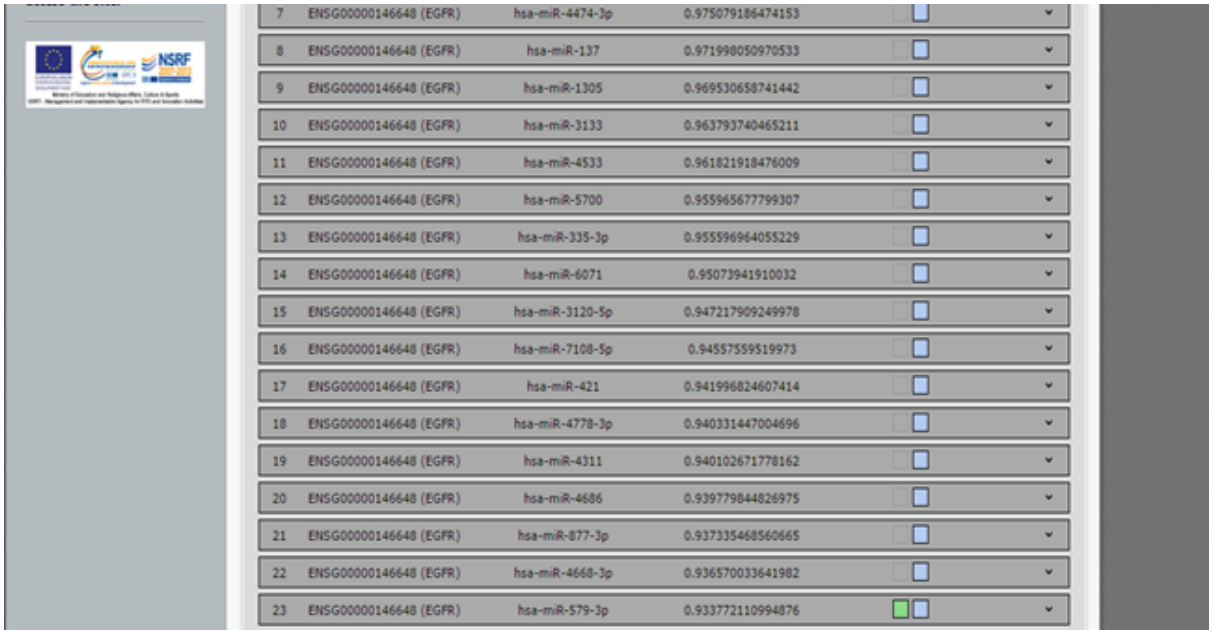


Figure 8: microRNA predicts target site

After performing the integrated analysis we use the microRNAs for targeting the major interstitial lung disease causing genes in the body.

The microRNAs that we got from our analysis were as follows :

(i) Cel-let-7
(ii) Has-mir-7-1
(iii) Has-mir-3934

These were then targeted to genes causing ILDs and it was shown that these families of microRNAs were responsible for down regulating the mRNA by binding to these genes and therefore causing Interstitial Lung Diseases. The genes that they targeted were as follows:

(i) EGFR
(ii) MUC5B
(iii) INTERLEUKIN

```
Version: RNAhybrid 2.2
Command line:/vol/bioapps/bin/RNAhybrid.bin -n 22 -q /var/bi
s 3utr_human -t /var/bibiserv2/anonymous/rnahybrid/25/09/44/
searching
dataset: 1
mde of cel-let-7: -43.000000
Individual hits
--------------------------------------------------------------

dataset: 1
target: AAH70123.1
length: 153
miRNA : cel-let-7
length: 22

mfe: -6.0 kcal/mol
p-value: 1.000000e+00

position  15
target 5'       N   G              3'
                 ACA
                 UGU
miRNA  3' UUGAUA   UGGAUGAUGGAGU 5'

--------------------------------------------------------------
```

Figure 9: cel-let-7 complements interleukin gene

```
Version: RNAhybrid 2.2
Command line:/vol/bioapps/bin/RNAhybrid.bin -n 110 -q /var/bibiserv2/anonym
-s 3utr_human -t /var/bibiserv2/anonymous/rnahybrid/25/09/30/bibiserv2_2017
searching
dataset: 1
mde of hsa-mir-7-1: -216.099976
Individual hits
-------------------------------------------------------------

dataset: 1
target: AAI28420.1
length: 388
miRNA : hsa-mir-7-1
length: 110

mfe: -14.7 kcal/mol
p-value: 1.000000e+00

position  351
target 5'                N   N     GN NN   NUNGNNNNNNA   NN    NN C
                          CU   GCU   G   GC             UG   GUU   A
                          GA   CGG   U   CG             AC   CAG   U
miRNA   3' GACAUCUCCGUACCG   CA    UA AC   UCUGACACUAA   AA    C   AAAUCAAUAGAUUUU

-------------------------------------------------------------
```

Figure 10: has-mir-7-1 binds to EGFR gene

```
Version: RNAhybrid 2.2
Command line:/vol/bioapps/bin/RNAhybrid.bin -n 107 -q /var/bibiserv2/anonymous/rnahybrid/25/09/52/bibiser
-s 3utr_human -t /var/bibiserv2/anonymous/rnahybrid/25/09/52/bibiserv2_2017-04-25_095205_cXcDJ/rnahybrid_
searching
dataset: 1
mde of hsa-mir-3934: -236.899963
Individual hits
-------------------------------------------------------------

dataset: 1
target: CAA96577.1
length: 3570
miRNA : hsa-mir-3934
length: 107

mfe: -15.3 kcal/mol
p-value: 1.000000e+00

position  1998
target 5'              N   N   NN         NA  NN  NNNNUN N   NNNNU     N   GN  NNNN N   NNNNNN     U
                        GUU UAU  UUUU  UU  UG  AU      U  GU     UUAUU UAU  UU    U GUU      UUUAU
                        CGA GUG  AGGG  GA  AC  UG      A  CG     AAUGA GUG  GG    G CGG      AGGUG
miRNA  3' CCUAAGUUUAGGGA   G   A     UC  C   GU  G      CU  UUC       A   AC  AG   A   AGUCAA     UGGACUU

-------------------------------------------------------------
```

Figure 11: has-mir-3934 complements MUC5B gene

In the above figures we can clearly see that the microRNAs bind to the complimentary gene sequence that was used.

The tool that I used was BIBISERV. We used this tool for the optimization and cross-validation of predicted microRNA and RNA interactions.

Here the microRNA binds to the best possible location on the gene sequence.

Also hybridization energy is also calculated.

This tool helped us to verify that the correlation between these two types of RNAs can cause Interstitial lung diseases in organisms when the mRNa is down regulated.

REFERENCES

1. Wienholds E, Plasterk RH. MicroRNA function in animal development. FEBS Lett. 2005; 579(26):5911–5922.

2. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010; 466(7308):835–840.

3. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: Tools for microRNA genomics. Nucleic Acids Res. 2008; 36(Database issue):D154–8.

4.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.

5. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, et al. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. Mol Cell. 2007;27(1):91–105.

6. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved Targets of microRNAs. Genome Res. 2009;19(1):92–105.

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

8. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: Targets and expression. Nucleic Acids Res. 2008; 36(Database issue):D149–53.

9. Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG. miRGen: A database for the study of animal microRNA genomic organization and function. Nucleic Acids Res. 2007;35(Database issue):D149–55.

10. Dinwiddie R. Sharief N. Crawford O. Idiopathic interstitial pneumonitis in children: a national survey in the United Kingdom and Ireland. Pediatr Pulmonol. 2002; 34:23–29.

11. Fan LL. Langston C. Chronic interstitial lung disease in children. Pediatr Pulmonol. 1993; 16:184–196.

12. Goerke J. Pulmonary surfactant: functions and molecular composition. Biochim Biophys Acta. 1998;1408:79–89.

13. Farrell PM. Avery ME. Hyaline membrane disease. Am Rev Respir Dis. 1975;111:657– 688.

14. Weaver TE. Na CL. Stahlman M. Biogenesis of lamellar bodies, lysosome-related organelles involved in storage and secretion of pulmonary
surfactant. Semin Cell Dev Biol. 2002;13:263– 270.

15. Weaver TE. Conkright JJ. Function of surfactant proteins B and C. Annu Rev Physiol. 2001;63:555–578.

16. Crouch E. Wright JR. Surfactant proteins a and d and pulmonary host defense. Annu Rev Physiol. 2001;63:521–554.