**Integrative analysis to identify microRNA drivers of COPD and ILD**

**Abstract**

The molecular events leading to the development of Chronic Obstructive Pulmonary Disease (COPD) and Interstitial Lung Disease (ILD), are poorly understood. We hypothesized that integrative genetic and expression networks can help identify drivers of these diseases and elucidate mechanisms of genetic susceptibility.

We utilized 262 lung tissue specimens from patients with COPD or ILD that were profiled with microRNA (miRNA) sequencing, microarray gene expression and SNP chip genotyping. We first identified associations between SNP genotypes and either gene or miRNA expression. Next, we built condition specific integrative networks using a Causality Inference Test for predicting SNP-miRNA-mRNA triplets, where the miRNA is a predicted mediator of the SNP’s effect on gene expression. Lastly, we identified the miRNAs predicted to affect the most genes within each network and their associated SNPs.

Members of miR-34/449 family, known to promote airway differentiation by repressing the Notch pathway, were among the top ranked miRNAs in COPD and ILD networks, but not in the non-disease network. Furthermore, the set of genes that positively correlated with the miR-34/449 members were enriched among genes that increase in expression over time when airway basal cells are differentiated at an air-liquid interface. This set of genes was also associated with genes that are positively correlated with airway wall thickening in patients with emphysema. In addition, we found SNPs that are associated with COPD and ILD status in addition to miR-34/449 expression.

We used integrative genetic and genomic networks to reveal novel miRNA drivers of COPD and ILD. In addition, miRNAs may be important mediators of some genetic risk factors for these lung diseases.

**Introduction**

Previous studies have used models of regulatory networks to explore the differences between disease and normal states. For example, an integrative network based approach to identify genetic nodes important to Late-Onset Alzheimer’s Disease has been proposed, highlighting an immune- and microglia-specific module1. Also, the role of microRNAs (miRNAs) within gene regulatory networks has been previously characterized2,3.

**Methods**

**High-throughput sequencing of small RNA**

45 samples were prepared with Small RNA Sample Prep Kit v1.5 (Illumina) and sequenced on the Genome Analyzer IIx (Illumina) according to the manufacturer’s protocol.  Multiplexed small RNA sequencing was conducted on the Illumina HiSeq 2000 for 319 lung tissue samples.  Briefly, one microgram of total RNA from each sample was used for library preparation with a TruSeq Small RNA Sample Prep Kit (Illumina).  RNA adapters were ligated to 3’ and 5’ end of the RNA molecule and the adapter-ligated RNA was reverse transcribed into single-stranded cDNA. The RNA 3’ adapter was specifically designed to target miRNAs and other small RNAs that have a 3’ hydroxyl group resulting from enzymatic cleavage by Dicer or other RNA processing enzymes.  The cDNA was then PCR amplified using a common primer and a primer containing one of 10 index sequences. The introduction of the six-base index tag at the PCR step allowed multiplexed sequencing of different samples in a single lane of a flowcell.  Ten individual PCR-enriched cDNA libraries with unique indices in equal amount were pooled and gel purified together.  A 0.5% PhiX spike-in was also added in all lanes for quality control.  Each library was hybridized to one lane of the 8-lane single-read flowcell on a cBot Cluster Generation System (Illumina) using TruSeq Single-Read Cluster Kit (Illumina).  The clustered flowcell was loaded onto HiSeq 2000 sequencer for a multiplexed sequencing run that consists of a standard 36-cycle sequencing read with the addition of a 7-cycle index read.

**miRNA alignment and quality control**

To estimate miRNA expression we used a small RNA sequencing pipeline previously described4. Briefly, the 3′ adapter sequence was trimmed using the FASTX toolkit. Reads longer than 15 nt were aligned to hg19 using Bowtie v0.12.75 allowing up to one mismatch and up to 10 genomic locations. miRNA expression was quantified by counting the number of reads aligning to mature miRNA loci (miRBase v20) using Bedtools v2.9.06,7. miRNA counts within each sample were normalized to RPM values by adding a pseudocount of one to each miRNA, dividing by the total number of reads that aligned to all miRNA loci within that sample, multiplying by 1 × 106, and then applying a log2 transformation4. The batch effects of the two sequencing protocols were removed by Combat8 and 13 outliers were removed by PCA; 351 patients were included in the downstream analysis.

**miRNA expression in lung tissue**

Differentially expression analysis was performed on the 351 patients. A generalized linear model (*glm.nb(),* MASS R package) was applied, with miRNA counts as response variable and protocol, smoking status, age, gender and disease status as regression terms. The significance of the associations were assessed using ANOVA9 and p-values were adjusted for false discovery rate10 (FDR). Next, groups of miRNAs or samples were identified using consensus clustering11 (*ConsensusClusterPlus* R package). Sample clusters were then associated with clinical variables of disease severity including DLCO (diffusing capacity of the lungs for carbon monoxide), FEV1/FVC ratio (forced expiratory volume 1/ forced vital capacity), FEV1 percent predicted, percent emphysema, BODE (body-mass index), by Kruskal-Wallis test12 (*kruskal.test,* R *stats* package).

**Building causal disease specific networks**

By “anchoring” expression data with genetic information, we can identify key miRNA regulators of gene expression associated with COPD. We utilized a subset of 262 lung tissue samples that had miRNA sequencing data as well as an Agilent gene expression array and Affymetrix SNP chip. We first identified all genes and miRNAs associated with a SNP (i.e. eQTL) by ANOVA while correcting for age, gender, smoking status, and population structure (p<0.0005). Next, we built integrative networks within the COPD, ILD, and control patients using the causality inference test (CIT)2. This test is a previously established method for predicting SNP-miRNA-mRNA triplets where the SNP is regulating the expression of the miRNA and the miRNA is regulating the expression of the gene2. CIT assesses the hypothesis that a potential mediator between an initial randomized variable and an outcome variable is causal for that outcome. Causal and reactive models are defined as series of conditions of associations between the three variables, corresponding to SNP, microRNA and mRNA nodes (figure 2A). The significance of the test is computed for both the causal and reactive models. If the causal p-value is lower than 0.05 and the reactive higher than 0.05 then the call is considered causal. If both p-values are greater than 0.05 then the call is independent, and if both p-values are lower than 0.05, then the causality cannot be inferred.

Next we compared the disease networks with the control network and evaluated those miRNAs that were differentially connected with their targets between the two states. The genes found to be regulated by the top differentially connected miRNAs were significantly enriched by both GSVA13 and GSEA14, in two independent gene expression datasets15.

**Results**

**miRNA Sequencing Data**

We profiled small-RNA sequencing from 351 lung tissue samples (table 1).

|  |  |  |  |
| --- | --- | --- | --- |
| Covariates | Control (n=62) | ILD (n=144) | COPD (n=145) |
| Smoking Status ˦ ǂ | 2 current,  38 former, 19 never,  3 NA | 5 current,  85 former,  50 never,  4 NA | 8 current,  129 former,  6 never,  2 NA |
| Age ǂ | 63.1 (12.0) | 61.2 (10.2) | 64.4 (9.9) |
| Pack Years \* ˦ ǂ | 41.1 (36.6) | 26.3 (19.9) | 55.9 (39.0) |
| Gender | 31 males,  31 females | 78 males,  66 females | 86 males,  59 females |
| FEV1/FVC \* ˦ ǂ | 0.77 (0.1) | 0.83 (0.1) | 0.5 (0.2) |
| Percent Emphysema ˦ ǂ | 0.7 (1.0) | 0.8 (1.7) | 16.6 (18.0) |

**Table 1.** Demographics table of the small RNA sequencing samples. \*Significantly different between ILD and Control (p<0.05); ˦Significantly different between COPD and Control (p<0.05); ǂSignificantly different between ILD and COPD (p<0.05); p-values for gender and smoking status were calculated by using *Fisher’s exact test*; p-values for age, pack years, FEV1/FVC and Percent Emphysema were calculated by using *Student’s t-test*.

**miRNA expression in lung tissue**

The expression profiles of 506 miRNA isoforms were found to be significantly associated with the presence of disease (FDR<0.05), as shown in figure 1A. Next, five main clusters were determined by consensus clustering (figure 1B). The majority of control samples are clustered in sample cluster 1. Sample clusters 2 and 4 are associated with ILD (p<0.001), while cluster 5 is associated with both ILD and COPD (p<0.001). Figure 1C presents the association of the five clusters with relevant clinical variables, such as DLCO (p=0.01) for ILD, and FEV1 percent predicted (p=0.01), FEV1/FVC (p=0.07), percent emphysema (p=0.07), and BODE (p=0.05) for COPD.

**miR-34/449 family is differentially connected in disease compared to control**

To identify disease specific drivers, we performed an integrative analysis on a subset of patients that had available SNP, miRNA and mRNA data (table 2).

|  |  |  |  |
| --- | --- | --- | --- |
| Covariates | Control (n=38) | ILD (n=113) | COPD (n=111) |
| Smoking Status ˦ ǂ | 2 current,  24 former, 12 never | 4 current,  71 former,  38 never | 7 current,  99 former,  5 never |
| Age | 65.5 (11.5) | 62.2 (9.2) | 63.8 (9.2) |
| Pack Years \* ǂ | 49.9 (40.8) | 26.3 (20.4) | 55.1 (37.8) |
| Gender | 22 males,  16 females | 61 males,  52 females | 65 males,  46 females |
| FEV1/FVC \* ˦ ǂ | 0.76 (0.06) | 0.83 (0.07) | 0.49 (0.24) |
| Percent Emphysema ˦ ǂ | 0.7 (1.0) | 0.74 (1.7) | 17.0 (18.3) |

**Table 2.** Demographics table of samples with available miRNA and mRNA data. \*Significantly different between ILD and Control (p<0.05); ˦Significantly different between COPD and Control (p<0.05); ǂSignificantly different between ILD and COPD (p<0.05); p-values for gender and smoking status were calculated by using *Fisher’s exact test*; p-values for age, pack years, FEV1/FVC and Percent Emphysema were calculated by using *Student’s t-test*.

We first identified all genes and miRNAs associated with a SNP (i.e. eQTL) by ANOVA while correcting for age, gender, smoking status, and population structure (p<0.0005). Then, we built integrative networks within the COPD, ILD, and control patients using the causality inference test (CIT)2. The number of significant interactions obtained at each step of the network construction are presented in supplementary figure 1. Next, we explored the *scale-free* property of the networks by computing the frequency of node degree in log scale. As expected, the networks present a negative linear relationship between the node degree and the frequency of node degree in log scale (figure 2B). Furthermore, we identified the miRNAs predicted to interact with the most genes in each network (supplementary table 1). Members of the miR-449 and miR-34 families were found to be among the top ranked in COPD and ILD networks (figure 2C), indicating that miR-449/34 family has a greater impact on gene expression regulation in disease compared to control group. Members of miR-34/449 family can promote airway differentiation by repressing the Notch pathway16. These miRNAs were found to share an increased number of associated genes, as illustrated in figure 3C. We observed that the combined set of genes that positively correlated with these miRNAs were enriched among genes that increase in expression over time when airway basal cells are differentiated at an air-liquid interface (ALI)15. Gene enrichment results were significant by both GSVA, p<10-3 (figure 3B) and GSEA,q<<0.001 (supplementary figure 2A). This set of genes was also associated with genes that are positively correlated with airway wall thickening in patients with emphysema, by both GSVA, p<10-4 (figure 3C) and GSEA, q<<0.001 (supplementary figure 2B). These results suggest that the miR-449/34 family is playing a role in differentiation associated with the airway wall thickening phenotypes.

**SNPs associated with disease that regulate miR-34/449**

Using the causality inference test2, we found 75 SNPs in COPD (supplementary file COPD\_SNPs.xlsx) and 60 SNPs in ILD (supplementary file ILD\_SNPs.xlsx) and that may regulate miR-449/34 family. Some of these SNPs have been previously associated with asthma, inflammation, cancer and other degenerative diseases. Top significantly associated SNPs with COPD or ILD by a Fisher’s exact test (q<0.25) are shown in figure 4.

**Discussion**

Besides miR449/34 family, miR-4423 was also differentially connected in COPD. Expression of this miRNA has been previously associated with airway differentiation in smokers with lung cancer17.

**Acknowledgements**

**References**

1. Zhang B, Gaiteri C, Bodea L-G, et al. Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer’s Disease. *Cell*. 2013;153(3):707-720. doi:10.1016/j.cell.2013.03.030.

2. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. *BMC Genet*. 2009;10(1):23. doi:10.1186/1471-2156-10-23.

3. Su W-L, Kleinhanz RR, Schadt EE. Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques. *Mol Syst Biol*. 2011;7:490. doi:10.1038/msb.2011.23.

4. Campbell JD, Liu G, Luo L, et al. Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data. *RNA*. 2015;21(2):164-171. doi:10.1261/rna.046060.114.

5. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25.

6. Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res*. 2004;32(Database issue):D109-11. doi:10.1093/nar/gkh023.

7. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033.

8. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037.

9. Chambers J. *Linear Models*. (Chambers J, Hastie T, eds.). Pacific Grove: Wadsworth & Brooks/Cole; 1992.

10. Yoav Benjamini YH. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995;57(1):289-300. http://www.jstor.org/stable/2346101.

11. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26(12):1572-1573. doi:10.1093/bioinformatics/btq170.

12. Myles Hollander and Douglas A. Wolfe. *Nonparametric Statistical Methods*. New York: John Wiley & Sons; 1973.

13. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7. doi:10.1186/1471-2105-14-7.

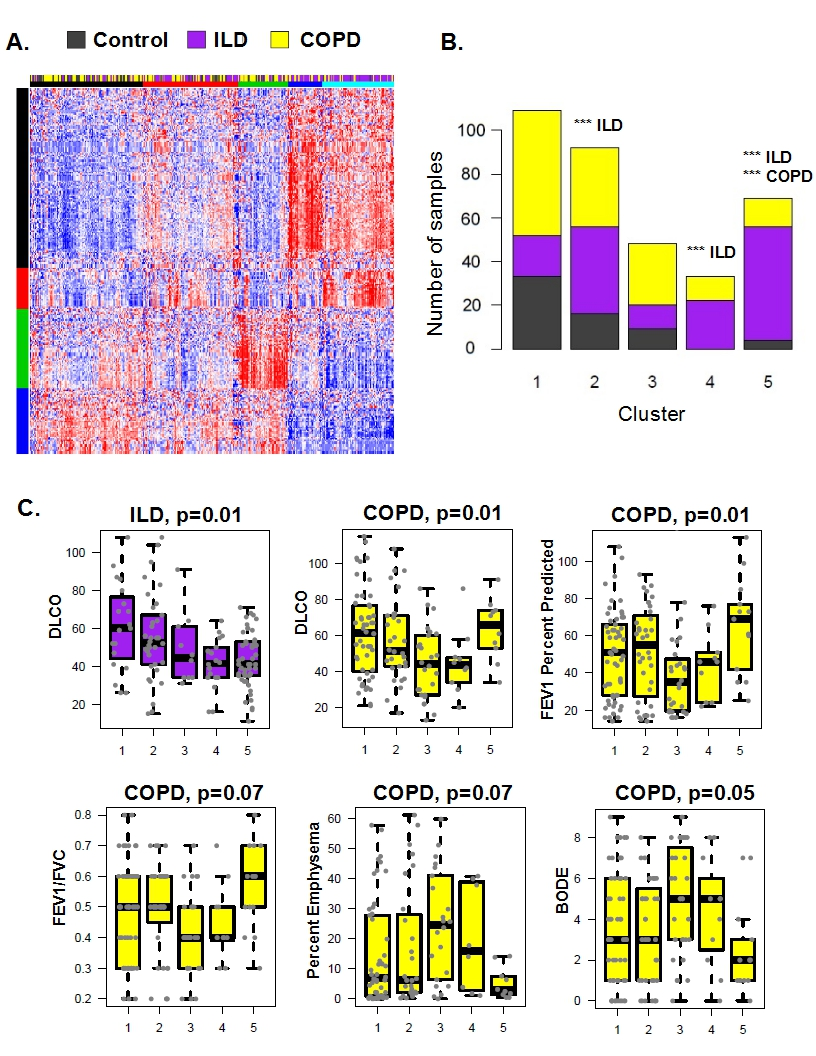
14. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102.

15. Ross AJ, Dailey LA, Brighton LE, Devlin RB. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am J Respir Cell Mol Biol*. 2007;37(2):169-185. doi:10.1165/rcmb.2006-0466OC.

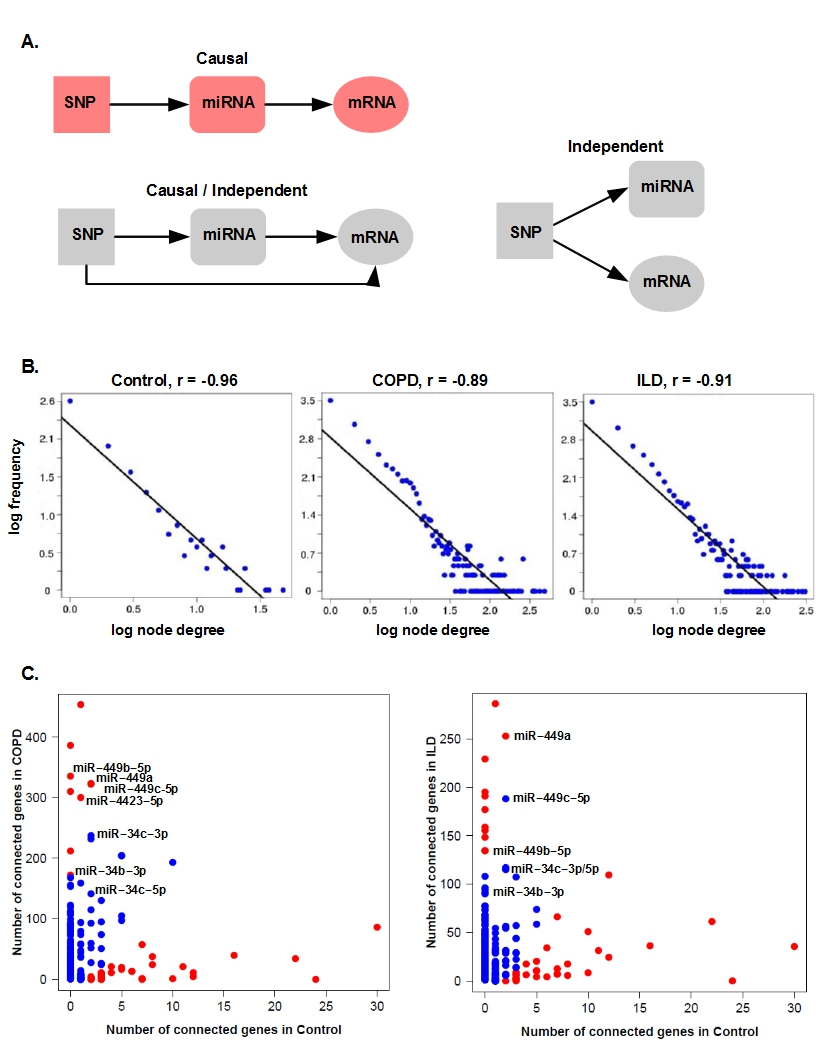
16. Chevalier B, Adamiok A, Mercey O, et al. miR-34/449 control apical actin network formation during multiciliogenesis through small GTPase pathways. *Nat Commun*. 2015;6:8386. doi:10.1038/ncomms9386.

17. Perdomo C, Campbell JD, Gerrein J, et al. MicroRNA 4423 is a primate-specific regulator of airway epithelial cell differentiation and lung carcinogenesis. *Proc Natl Acad Sci U S A*. 2013;110(47):18946-18951. doi:10.1073/pnas.1220319110.

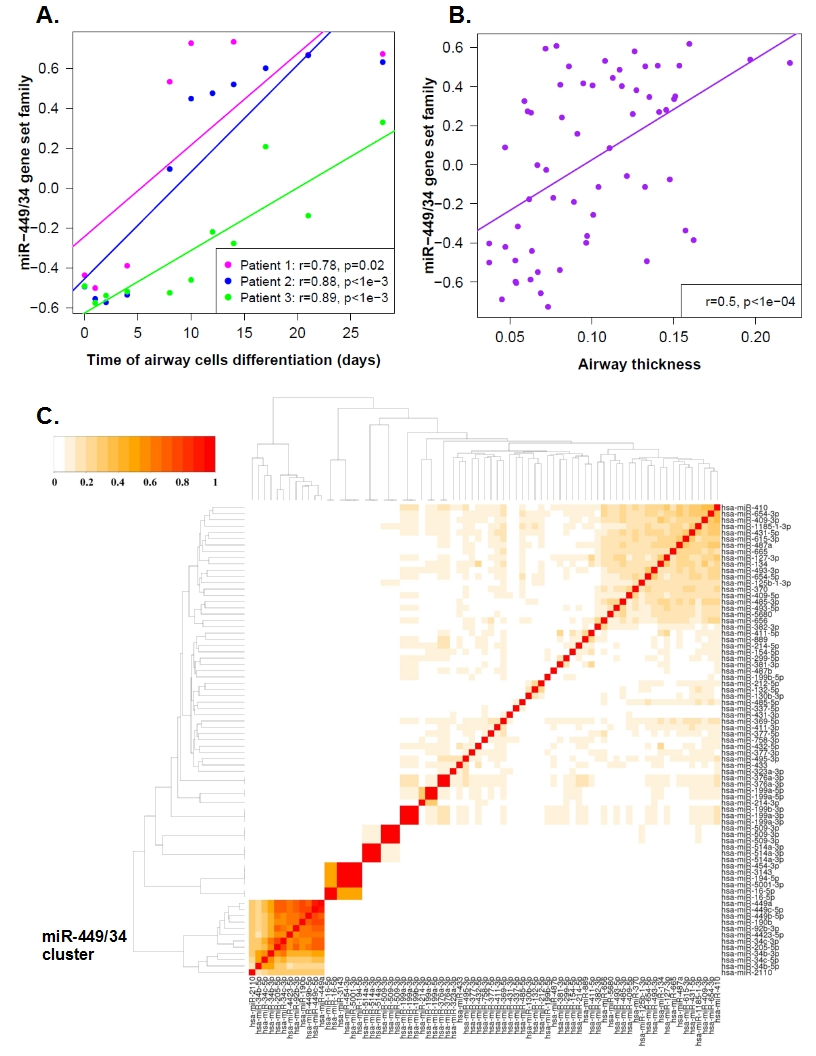
Figures

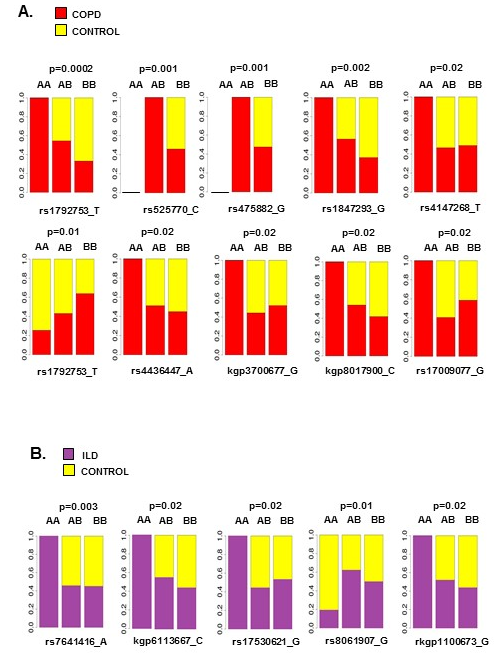


**Figure 1. miRNA expression profiles.** **A.** The expression profiles of 506 miRNA isoforms were found to be significantly associated with the presence of disease (FDR<0.05). **B.** The majority of control samples are in sample cluster 1. Sample clusters 2 and 4 are associated with ILD (p<0.001), while cluster 5 is associated with both ILD and COPD (p<0.001); \*\*\* p<0.001; **C.** Association of the 5 clusters with relevant clinical variables by Kruskal-Wallis test.



**Figure 2.** **Network construction and evaluation.** **A.** Causality inference test.The significance of the test is computed for both the causal and reactive models. If the causal p-value is lower than 0.05 and the reactive higher than 0.05 then the call is considered causal. If both p-values are greater than 0.05 then the call is independent, and if both p-values are lower than 0.05, then the causality cannot be inferred. **B.** The CIT networks follow a power law. The negative correlation between the frequency of node degree and the node degree indicates that the networks are scale free. **C.** Number of genes regulated by each miRNA. miR-449/34 family members were found to be among the top 20 differentially connected in COPD and ILD compared to control group. The red dots indicate the significantly differentially connected miRNAs by a Fisher’s exact test (FDR<0.2).

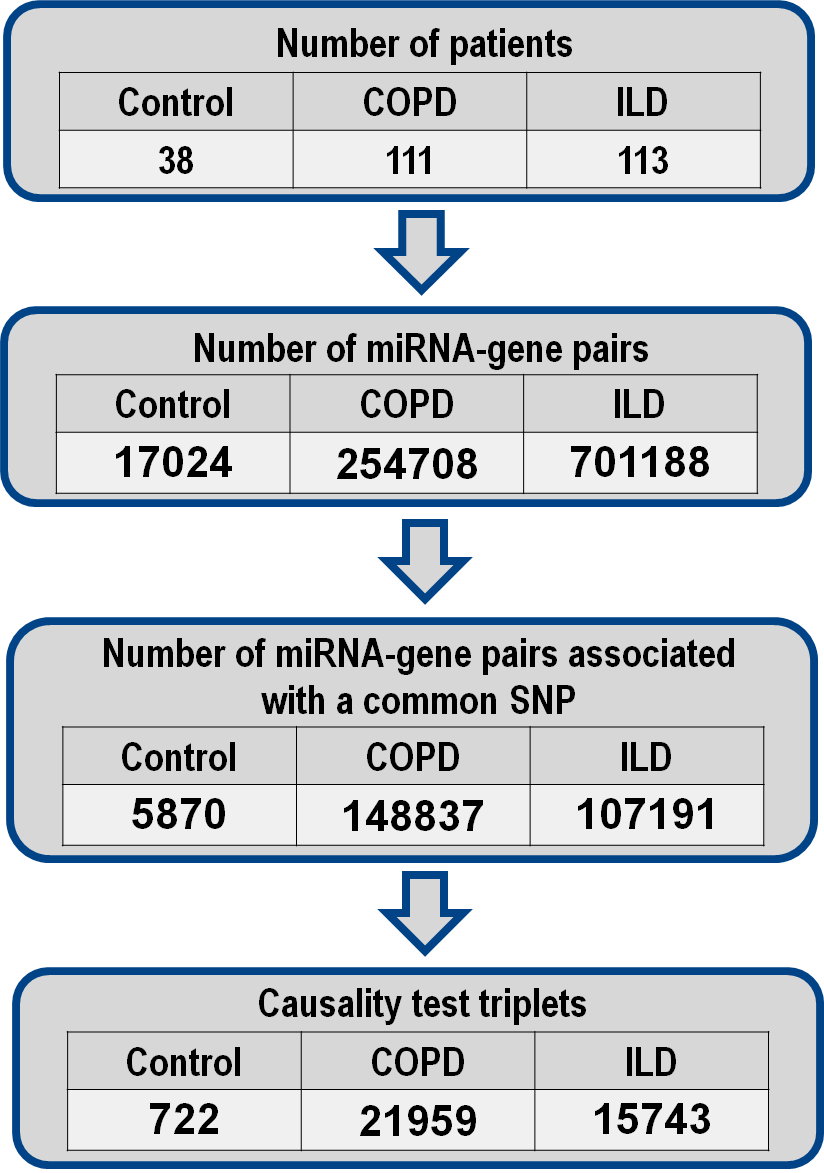
 **Figure 3. Enrichment of miR-449/34 modules. A.** Enrichment of miR-449/34 gene set family with airway cells differentiation by GSVA. The set of genes that positively correlated with miR-449/34 family were enriched among genes that increase in expression with the airway epithelial cells differentiation in COPD **C.** Enrichment of miR-449/34 gene set family with increasing airway wall thickness in patients with emphysema by GSVA.The set of genes that positively correlated with miR-449/34 family were enriched among genes that increase in expression with airway wall thickening of patients with emphysema in COPD. **C.** miR-449/34 modules present an increased number of shared genes by Jaccard index.



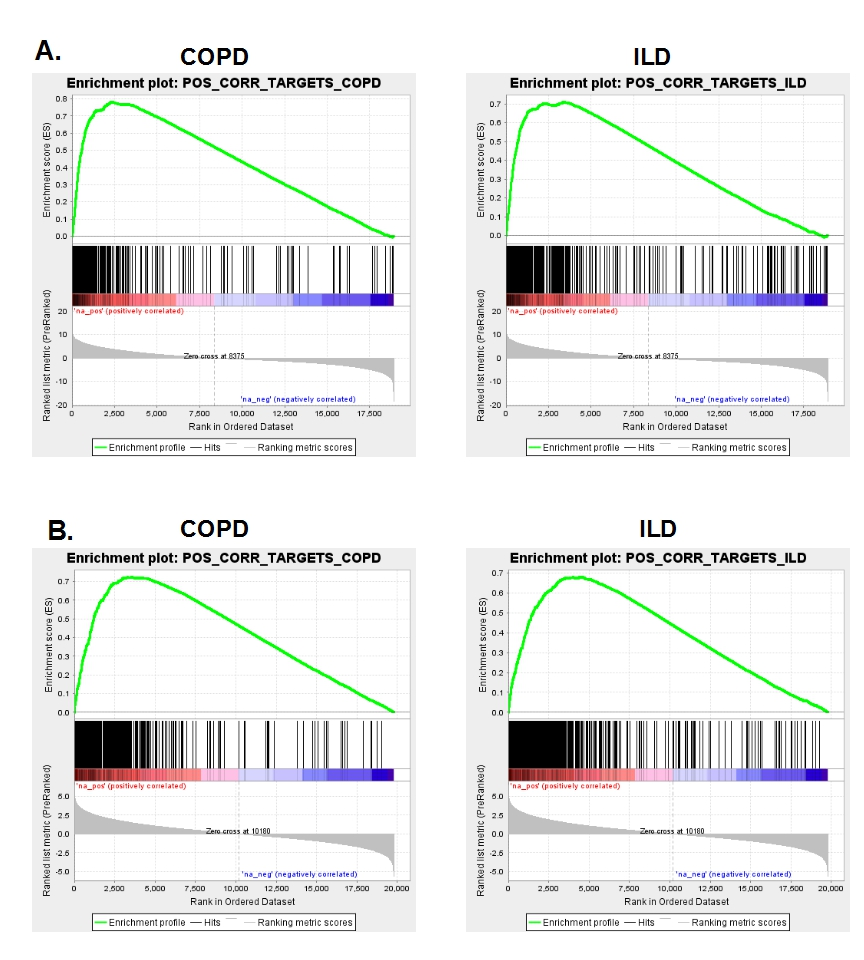
**Figure 4.** **A.** Top SNPs that are significantly associated with COPD by a Fisher’s exact test (q<0.25). **B.** Top SNPs that are significantly associated with ILD by a Fisher’s exact test (q<0.25).

**Supplementary figures and tables**

**Supplementary Figures and Tables**



**Supplementary figure 1.** Number of significant interactions at each step of network construction in Control, COPD and ILD groups.



**Supplementary figure 2. Enrichment of miR-449/34 targets by GSEA. A.** Enrichment of miR-449/34 associated genes with airway differentiation. The set of genes that positively correlated with miR-449/34 family were enriched among genes that increase in expression with the airway epithelial cells differentiation (q≈0), both in COPD and ILD. Genes were ranked from those that increased in expression with differentiation over time (red) to those that decreased in expression (blue). **B.** Enrichment of miR-449/34 associated genes with increasing airway wall thickness in patients with emphysema.The set of genes that positively correlated with miR-449/34 family were enriched among genes that increase in expression with airway wall thickening of patients with emphysema (q≈0), both in **A.** COPD and **B.** ILD. Genes were ranked from those that increased in expression with thicker airway walls (red) to those that decreased in expression in thicker airway walls (blue), in an independent dataset of 60 airway samples.

|  |  |  |
| --- | --- | --- |
| Top mostly connected miRNAs in COPD | Top mostly connected miRNAs in ILD | Top mostly connected miRNAs in Control |
| hsa-miR-27a-5p \*  hsa-miR-190b \*  hsa-miR-449b-5p \*  hsa-miR-449a \*  hsa-miR-449c-5p \*  hsa-miR-4423-5p \*  hsa-miR-92b-3p \*  hsa-miR-34c-3p  hsa-miR-205-5p  hsa-miR-23a-5p \*  hsa-miR-509-3p-2  hsa-miR-509-3p-3  hsa-miR-509-3p-1  hsa-miR-30a-3p  hsa-miR-34b-3p \*  hsa-miR-1185-1-3p  hsa-miR-125b-1-3p  hsa-miR-654-5p  hsa-miR-485-5p  hsa-miR-34c-5p | hsa-miR-92b-3p \*  hsa-miR-449a \*  hsa-miR-200a-5p \*  hsa-miR-31-5p \*  hsa-miR-92b-5p \*  hsa-miR-449c-5p  hsa-miR-200b-3p \*  hsa-miR-31-3p \*  hsa-miR-190b \*  hsa-miR-449b-5p \*  hsa-miR-511-1 \*  hsa-miR-511-2 \*  hsa-miR-34c-5p  hsa-miR-34c-3p  hsa-miR-146b-5p \*  hsa-miR-2110 \*  hsa-miR-34b-5p  hsa-miR-450b-5p  hsa-miR-200a-3p  hsa-miR-34b-3p | hsa-miR-21-5p \*  hsa-miR-4802-3p \*  hsa-miR-146a-5p \*  hsa-miR-378c \*  hsa-miR-142-3p \*  hsa-miR-146b-5p \*  hsa-miR-421 \*  hsa-miR-30a-3p  hsa-miR-378a-5p \*  hsa-miR-378a-3p \*  hsa-miR-330-5p \*  hsa-miR-425-5p \*  hsa-miR-378i \*  hsa-miR-26a-5p-1 \*  hsa-miR-26a-5p-2 \*  hsa-miR-223-5p \*  hsa-miR-191-5p \*  hsa-miR-30a-5p  hsa-miR-509-3p-2  hsa-miR-509-3p-3  hsa-miR-509-3p-1  hsa-miR-5571-3p  hsa-miR-301b \*  hsa-miR-766-3p \*  hsa-miR-199b-5p \*  hsa-miR-34a-5p \* |

**Supplementary table 1.** Top 20 mostly connected miRNAs in each phenotype; \* indicates the significant FDR-adjusted p-values (q<0.2) by a Fisher’s exact test that determines the difference in the connectivity frequencies between the two phenotypes for each miRNA.