Integrative genetic and genomic networks identify microRNA associated with COPD and ILD

Ana B. Pavel[1,2*], Carly Garrison[1], Lingqi Luo[1], Gang Liu[1], Daniel Taub[1], Ji Xiao[1], Brenda Juan-Guardela[3], John Tedrow[3**], Yuriy O. Alekseyev[4], Ivana V. Yang[5], Mark W. Geraci[5***], Frank Sciurba[3], David A. Schwartz[5], Naftali Kaminski[3***], Jennifer Beane[1,2], Avrum Spira[1,2], Marc E. Lenburg[1,2,4], Joshua D. Campbell[1,2]

1. Department of Medicine, Boston University School of Medicine, Boston, MA, USA.
2. Bioinformatics Graduate Program, Boston University, Boston, MA, USA.
3. Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA, USA.
4. Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA, USA.
5. Department of Medicine, University of Colorado, Aurora, CO, USA.

* Now at Department of Biomedical Engineering, University of Mississippi, University, MS, USA.

** Now at St. Elizabeth's Medical Center-Brighton, Brighton, MA, USA.

*** Now at Department of Medicine, University of Pittsburgh Medical Center, PA, USA.

**** Now at Department of Medicine, Yale School of Medicine, New Haven, CT, USA.

Corresponding authors:

Joshua D. Campbell

Email: camp@bu.edu

72 East Concord St

Boston, MA 02118 USA

Ana B. Pavel

Email: apavel@olemiss.edu

303 Brevard Hall

University, MS 38655 USA

ABSTRACT

Chronic obstructive pulmonary disease (COPD) and interstitial lung disease (ILD) are clinically and molecularly heterogeneous diseases. We utilized clustering and integrative network analyses to elucidate roles for microRNAs (miRNAs) and miRNA isoforms (isomiRs) in COPD and ILD pathogenesis. Short RNA sequencing was performed on 351 lung tissue samples of COPD (n=145), ILD (n=144) and controls (n=64). Five distinct subclusters of samples were identified including 1 COPD-predominant cluster and 2 ILD-predominant clusters which associated with different clinical measurements of disease severity. Utilizing 262 samples with gene expression and SNP microarrays, we built disease-specific genetic and expression networks to predict key miRNA regulators of gene expression. Members of miR-449/34 family, known to promote airway differentiation by repressing the Notch pathway, were among the top connected miRNAs in both COPD and ILD networks. Genes associated with miR-449/34 members in the disease networks were enriched among genes that increase in expression with airway differentiation at an air-liquid interface. A highly expressed isomiR containing a novel seed sequence was identified at the miR-34c-5p locus. 47% of the anticorrelated predicted targets for this isomiR were distinct from the canonical seed sequence for miR-34c-5p. Overexpression of the canonical miR-34c-5p and the miR-34c-5p isomiR with an alternative seed sequence down-regulated NOTCH1 and NOTCH4. However, only overexpression of the isomiR down-regulated genes involved in Ras signaling such as CRKL and GRB2. Overall, these findings elucidate molecular heterogeneity inherent across COPD and ILD patients and further suggest roles for miR-34c in regulating disease-associated gene-expression.

INTRODUCTION

Complex chronic lung diseases arise from heterogeneous molecular processes and are influenced by multiple factors including exposure to toxins and genetic susceptibility. Chronic obstructive pulmonary disease (COPD) is a progressive lung disease and the fourth leading cause of death worldwide,[1] with an incidence of 2.8 cases per 1,000 persons per year[2]. Patients with COPD suffer from breathing difficulty, wheezing, excess mucus production, and chronic cough. Although biological processes, such as chronic inflammation, apoptosis, and oxidative stress, have been implicated in COPD pathogenesis, knowledge of the key molecular drivers of this disease remains limited[3]. Interstitial lung disease (ILD) is a collection of chronic lung diseases characterized by fibrosis or inflammation of the alveolar tissue in the lung parenchyma[4]. One of the most common subtypes of ILD, idiopathic pulmonary fibrosis (IPF), has an incidence of 6.8-8.8 per 100,000 persons per year[5] with a median survival from diagnosis of 3–5 years[6–8].

MicroRNAs (miRNAs) are short RNA transcripts about 20-23 nucleotides long that can modulate expression levels or translation rates of specific mRNA targets via sequence-specific binding to their 3' UTR[9]. MicroRNAs are involved in a wide variety of developmental processes and aberrant activity of miRNAs can also contribute to disease pathogenesis[10]. Previous studies have performed transcriptomic profiling of affected lung tissue to understand the molecular processes associated with complex lung diseases such as COPD and ILD[11–13]. Additional studies sought to identify microRNA (miRNA) expression profiles associated with the presence of disease to gain insights into the regulation of aberrant gene expression[14–16]. Despite the information gained from these initial studies, larger sample sets are needed to identify novel molecular subtypes of disease and more data modalities need to be measured on each sample to perform integrative network analyses.

Network approaches that integrate multiple data types have been used extensively to study complex diseases[17]. Integrative genetic and genomic network approaches have been used to identify molecular drivers of late-onset Alzheimer's disease and breast cancer risk[18,19]. Integrative analysis of DNA methylation and gene expression has identified key regulators in the setting of COPD[20]. Several computational approaches have been applied to infer causality from biological data, including Bayesian networks,[21–23] factor graphs[18,19] and ridge and least absolute shrinkage and selection operator[24]. Statistical framework such as the Causlity

Inference Test (CIT), can be used to infer mediators of genetic or epigenetic factors associated with quantitative traits[25]. The CIT has also been used to characterize the role of microRNAs (miRNAs) in gene regulatory networks[26] and can be applied in settings where profiling of miRNA expression, mRNA expression, and genetic or epigenetic variation have been captured on the same samples.

Previous miRNA studies in COPD and ILD have relied primarily on microarray technology for quantifying expression. Microarrays only allow for the profiling of a canonical miRNA sequences. With the advent of small RNA sequencing, additional variation in miRNA sequences have been observed including variation on the 5' end of mature miRNAs[27]. Variation on the 5' end of a miRNA produces a different seed sequence. The seed sequence is the primary feature for determining the binding specificity of the miRNA to the 3' UTR of mRNA transcripts. These noncanonical miRNAs, often called isomiRs, can have alternative functional roles compared to the canonical miRNA sequence at that locus due to the targeting of distinct mRNAs[28]. Despite the potentially important role of isomiRs in regulating gene expression, the expression patterns of isomiRs have not been well-characterized in tissues from subjects with chronic lung disease.

In order to characterize molecular heterogeneity in chronic lung disease and predict key regulators of gene expression, we profiled miRNA expression via small-RNA sequencing from a large number of samples from the Lung Genome Research Consortium (LGRC). Unsupervised clustering revealed subgroups of subjects with distinct clinical and molecular characteristics. Using the CIT, we developed integrative networks and found increased connectivity in the disease cohorts for miRNAs involved in airway differentiation and ciliogenesis. Finally, we identified and characterized a 5' isomiR of miR-34c-5p with putative roles in the regulation of Ras pathway members. Overall, these analyses provide a comprehensive view of miRNA expression patterns COPD and ILD and elucidate the potential roles of these miRNA in regulating biological pathways within the lung.

METHODS

High-throughput sequencing of small RNA. RNA was obtained from the National Heart, Lung, and Blood Institute–sponsored Lung Tissue Research Consortium (LTRC)[11,13]. Tissue samples from the LTRC are labeled with IDs that cannot be used to identify to the study subjects. 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 320 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 described[29]. Briefly, the 3' adapter sequence was trimmed using the FASTX toolkit. Reads longer than 15 nt were aligned to hg19 using Bowtie v0.12.7[30] 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 v2.9.0.[31,32] For quality control, we examined the distribution of read lengths for each sample after trimming to ensure that the sequences we observed were of the proper length for miRNA. 13 of 365 samples clustered differently than the rest of the samples based on the read length

distribution and were excluded from subsequent analyses (**Supplementary Figure 1**). One additional sample was excluded as a duplicate leaving 351 samples for expression analysis (**Table 1**).

Differential expression. To identify miRNAs associated with disease a generalized negative binomial model (*glm.nb,* MASS R package) was applied to each miRNA. The count of each miRNA was used as the response variable and sequencing read depth, sequencing protocol, smoking status, age, gender, as well as COPD status and ILD status were used as predictor variables. The significance of the associations were assessed by performing an ANOVA[33] between a model with both the COPD and ILD terms and a second model without the either disease term. P-values were adjusted with the Benjamini-Hochberg false discovery rate[34] (FDR). MiRNAs were considered differentially expressed if they had an FDR q-value < 0.1 and the absolute value of the coefficient for either the COPD or ILD term was greater than 0.22, corresponding to a Fold Change (FC) of at least 1.25.

Consensus Clustering. For clustering and display in heatmaps, 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 1×10^6, and then applying a $\log_2$ transformation.[29] The batch effects of the two sequencing protocols were removed by Combat.[35] Groups of miRNAs or samples were identified using consensus clustering[36] (*ConsensusClusterPlus* R package) on the normalized and batch-corrected miRNA expression data. Sample clusters were assessed for enrichment of disease samples by using two logistic regression models where disease status (either COPD or ILD) was the response and sequencing read depth, sequencing protocol, smoking status, age, gender, and cluster status were dependent variables. Sample clusters were also 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, and BODE score (i.e., a measure of the degree of obstruction, dyspnea, and exercise capacity). Linear models were fit where each clinical phenotype was the response variable and sequencing read depth, sequencing protocol, smoking status, age,

gender, and cluster status were dependent variables. Two separate models were fit for ILD and COPD patients.

Building disease specific networks. We utilized a subset of 262 lung tissue samples with miRNA expression profiled by sequencing, as well as an Agilent gene expression microarray 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 ($p<0.0005$) using the *MatrixEQTL* package.[37] Next, we built integrative networks within the COPD, ILD, and control patients using the causality inference test (CIT).[25] This test is a previously established method for predicting SNP-miRNA-mRNA triplets where the SNP is modulating the expression of the miRNA and the miRNA is modulating the expression of the gene.[25] CIT assesses the hypothesis that a potential mediator between a genetic variable and an outcome variable is potentially 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. 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 causal relationship is indicated. 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 causality cannot be inferred. We select those SNP-miRNA-mRNA triplets where the SNP-mRNA relationship is defined by a miRNA mediator and did not examine triplets where the SNP is not associated with the miRNA. The number of mRNA predicted to be regulated by each miRNA was compared between control and disease networks. The genes found to be regulated by the top differentially connected miRNAs were examined by GSVA[38] and GSEA[39] in an independent dataset examining gene expression patterns associated with differentiation of airway epithelium at an air-liquid interface (ALI)[40].

Validation by qRT-PCR. To measure the expression of miR-34a-5p, miR-34b-5p and miR-34c-5p, 10 ng of total RNA was used in a Taqman MiRNA Assay (Life Technologies, Catalog #4427975, ID #000426, 000427, 000428, Carlsbad, CA) as per manufacturer's protocol and the results were normalized to RNU44 expression (Life Technologies, Catalog #4427975, ID #001094, Carlsbad, CA). To measure the expression of *RALA*,

GRB2, *CRK*, *CRKL*, *GRAP*, *RHOA*, *RHOC*, *EGF*, *ARAP2, NOTCH4* and *NOTCH1*, 500 ng of total RNA was reverse transcribed using RT2 First Strand Kits (Qiagen, Catalog #330401, Valencia, CA) according to the manufacturer's protocol. cDNA product was added to SYBR Green qPCR Mastermix (Qiagen, Catalog #330523, Valencia, CA) and the appropriate primer (Qiagen, Catalog #PPH07458A, PPH00714C, PPH00731A, PPH01982A, PPH13173A, PPH00305G, PPH01089E, PPH00137B, PPH20012A, PPH06021F, PPH00526C, Valencia, CA). Data was normalized to the expression of UBC (Qiagen, Catalog #PPH00223F, Valencia, CA) and analyzed using the comparative CT method.

Quantification and transfection of isomiRs. IsomiRs were identified within each canonical miRNA locus by grouping reads with the same 5' start position. Targetscan v6.0[41] was used to predict mRNA targets for each canonical and isomiR seed. IMR90 cells and HBEpCs were transiently transfected with hsa-miR-34c-5p miRIDIAN miRNA mimic (Dharmacon, Catalog #C-300655-03-0020, Lafayette, CO), a custom miR-34c 5' isomiR miRIDIAN miRNA mimic (Dharmacon, Lafayette, CO) or miRIDIAN miRNA mimic Negative Control #1 (Dharmacon, Catalog #C-001000-01, Lafayette, CO). IMR90 cell transfection was completed using Lipofectamine RNAiMAX transfection reagent (Life Technologies, Catalog #13778150, Carlsbad, CA) according to the manufactures protocol. Transfection of HBEpCs was done using Cytofect Epithelial Cell Transfection Kit (Cell Applications, Catalog #TF102K, San Diego, CA).

Data availability

SNP data was provided by the Lung Genomics Research Consortium (LGRC; http://lung-gemomics.org; 1RC2HL101715) using tissue samples and clinical data collected through the Lung Tissue Research Consortium (LTRC; http://www.ltrcpublic.com/). This data is available from dbGaP under the accession phs000624.v1.p1. The microRNA expression datasets generated and analyzed during the current study are available in the Raw and normalized data is available at the Gene Expression Omnibus (GEO) under the accession number GSE201121.

)9

RESULTS

Subject cohort

MicroRNA expression was profiled with small-RNA sequencing for 364 lung tissue samples collected by the Lung Tissue Research Consortium. Thirteen samples with low quality were removed (**Methods; Supplementary Table 1**) resulting in 351 samples for downstream analyses including 145 subjects with COPD, 144 subjects with ILD, 62 Controls (**Table 1, Supplementary Table 2**). Controls were mostly derived from tissue adjacent normal to cancer as previously described[11]. Subjects with COPD had a significantly higher proportion of former smokers, higher Pack Years, lower FEV1/FVC ratios, and higher Percent Emphysema compared to Control subjects. Subjects with ILD had significantly lower Pack Years and higher FEV1/FVC ratios compared to Control subjects. Compared to individuals with ILD, COPD subjects had a significantly higher proportion of former smokers, higher Age, higher FEV1/FVC ratios, and higher Percent Emphysema.

Table 1. Sample demographics.

	Control n=62	ILD n=144	COPD n=145
Smoking Status	2 current, 38 former, 19 never, 3 N/A	5 current, 85 former, 50 never, 4 N/A	8 current, 129 former, 6 never, 2 N/A
Age	63.1 +/- 12.0	61.2 +/- 10.2	64.4 +/- 9.9
Pack Years *	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 *	0.77 +/- 0.1	0.83 +/- 0.1	0.5 +/- 0.2
Percent Emphysema	0.7 +/- 1.0	0.8 +/- 1.7	16.6 +/- 18.0

*Significantly different between ILD and Control ($p<0.05$)
Significantly different between COPD and Control ($p<0.05$)
Significantly 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*.

Unsupervised clustering identifies novel subgroups associated with clinical phenotypes

693 of 2104 mature miRNAs were detected with 2 counts in at least 50% of samples. The expression profiles of 255 miRNAs were significantly associated with the presence of disease (ANOVA FDR q-value < 0.10 and fold change > 1.25 in either disease group compared to controls; **Figure 1A; Supplementary Table 2**). Five clusters of samples (S1-5) and 4 clusters of miRNA (M1-4) were determined by Consensus Clustering (**Methods; Supplementary Figure 2**). The majority of control samples (52%) were found in S1 (**Figure 1B**). Clusters S2, S4, and S5 had significantly more ILD samples compared to S1 ($p<0.05$). While the fractions of COPD samples in clusters S2-S5 were not significantly higher compared to cluster S1 ($p>0.05$), sample cluster S3 contained the highest proportion of COPD cases (53%).

The ILD patients in clusters S4 and S5 had significantly lower DLCO (diffusing capacity of the lung for carbon monoxide) compared to ILD patients in cluster 1 ($p<0.05$; **Figure 1C**). Clusters S4 and S5 had an up-regulation in the expression of miRNAs from M1 and a down-regulation of miRNAs in M4. Sixty of the 125 miRNAs in M1 were regionally co-located on chromosome 14q32 and were previously reported to be up-regulated in IPF[16]. Additional miRNAs in M1, including miR-21-5p/3p, miR-199a-3p, and miR-155-5p, have also been implicated in the ILD subtype IPF[14,42,43]. MiR-199a-5p was also associated with ILD status ($p = 0.002$) but did not pass our fold change cutoff[42]. The S4 and S5 clusters showed down-regulation in the expression of miRNAs from cluster M4. Several of the miRNAs in the M4 cluster were a part of in the miR-30 family, including miR-30a-5p/3p, miR-30d-5p/3p, miR-30b-5p, and miR-30c-2-3p. The two ILD-associated sample clusters S4 and S5 with more severe disease could be distinguished by higher levels of miRNA cluster M2. M2 contained many miRNAs that are major regulators of airway differentiation and ciliogenesis including miR-34b-5p/3p, miR-34c-5p/3p miR-449a, miR-449b-5p, miR-449c-5p, and miR-4423-5p[44,45]. Other studies have identified two subclasses of IPF that are characterized by differences in expression of ciliary genes[46]. The different patterns of expression of ciliary-related miRNA between clusters S4 and S5 may also be indicative of this subtype. Cluster S2, which was also enriched for ILD samples, had intermediate levels of cluster M1/M2 up-regulation and M4 down-

regulation compared to sample clusters S4 and S5, potentially indicating intermediate levels of disease severity.

COPD samples in cluster S3 had significantly lower FEV1 percent predicted ($p=0.01$), DLCO ($p<0.001$), and FEV1/FVC ($p=0.03$), as well as significantly higher BODE score (body-mass index, airflow obstruction, dyspnea, and exercise, $p=0.01$) and percent emphysema ($p=0.01$) compared to the COPD samples in cluster S1. Cluster S3 was largely defined by the up-regulation of miRNAs in M3. M3 miRNA included proximal miR-144 and miR-451a cluster on chromosome 17, miR-222-5p and miR-223-5p/-3p on chromosome X, as well as miR-18a and miR-92a-3p from the miR-17-92 polycistronic cluster on chromosome 13. Although not included in our clustering analysis due to the fold change cutoff, other miRNAs in the miR-17-92 polycistronic cluster were also associated with disease status, including miR-17-5p/3p, miR-19b-3p, and miR-20a-5p/3p (FDR q-value < 0.05). Overall, we identified expression patterns of miRNAs that can distinguish unique subsets of patients with COPD and ILD, including patients with more severe clinical phenotypes.

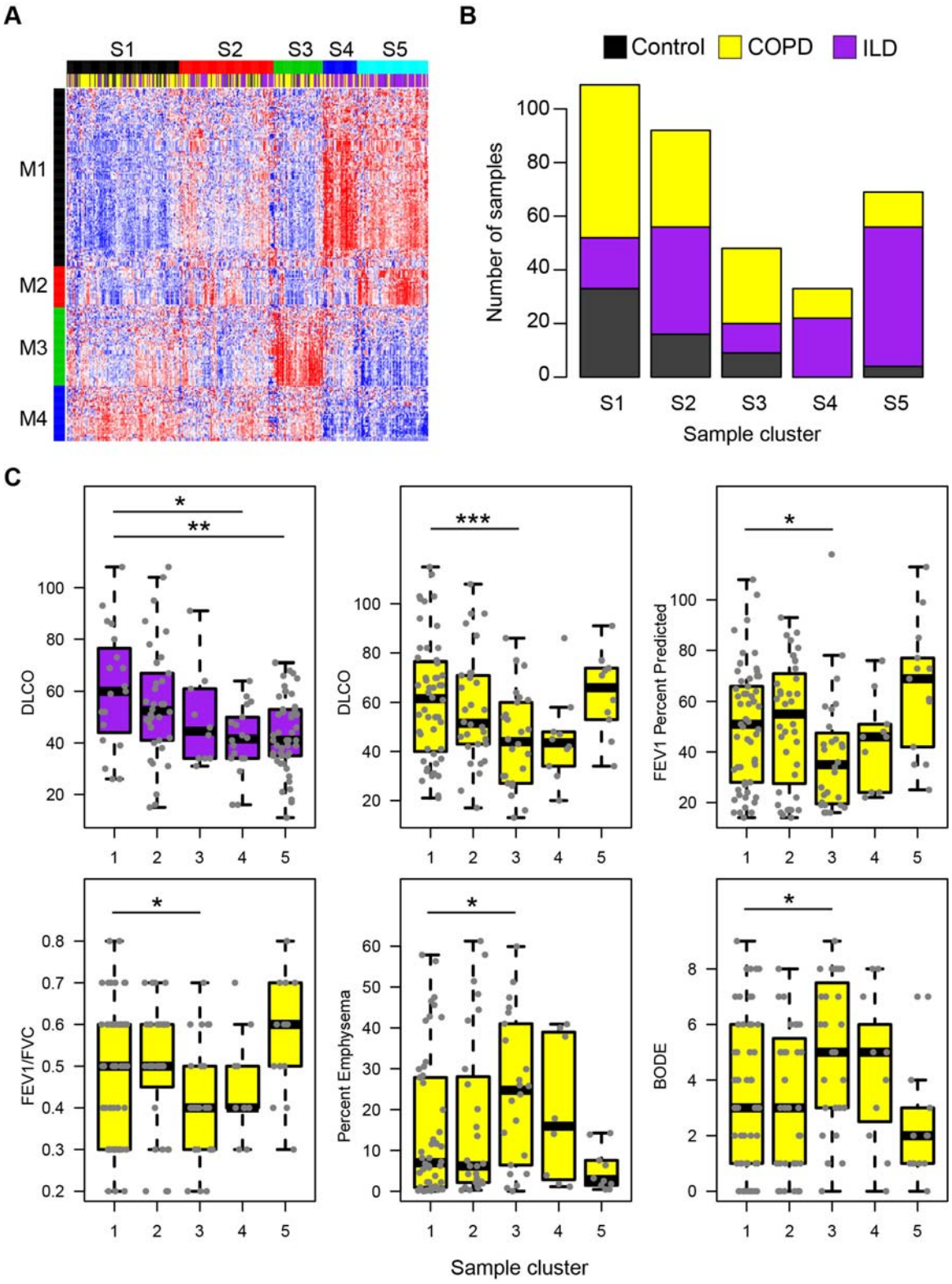

Figure 1. Heterogeneity of miRNA expression profiles associated with COPD or ILD. A. The expression profiles of 255 miRNAs were significantly associated with the presence of disease (ANOVA FDR q-value < 0.10 and fold change > 1.25 in either disease group compared to controls). Consensus clustering was used to identify 5 distinct samples clusters and 4 distinct miRNA clusters. **B.** Stacked barplots display the proportion of disease and control samples within each sample cluster. The majority of control samples (53%) fell into cluster S1. Clusters S2, S4, and S5 were enriched with ILD patients compared to cluster S1. **C.** ILD samples in clusters S4 and S5 has significantly lower DLCO compared to ILD samples in cluster S1 ($p < 0.05$). COPD samples in cluster S3 had significantly lower DLCO, FEV1 percent predicted, FEV1/FVC ratios and significantly higher percent emphysema and BODE scores compared to COPD samples in cluster S1. Asterisks indicate significance: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Comparison of the number of gene and miRNA eQTLs in COPD and ILD

eQTL analysis can reveal insights into specific biological effects that genetic variants have across different tissues or to disease phenotypes.[47] To compare the numbers of eQTLs between ILD and COPD and the number of eQTL effecting mRNA or miRNA expression in the setting of ILD and COPD, we utilized a subset of 262 lung tissue samples that had data from miRNA sequencing, SNP chips, and mRNA microarrays including 111 COPD, 113 ILD, and 38 Controls (**Supplementary Table 4**). Protein-coding genes and miRNAs associated with a SNP were identified by ANOVA while correcting for age, gender, smoking status, and population structure within ILD, COPD and Control groups (FDR < 0.05; **Table 2**; **Supplementary Figure 3**). The COPD cohort had larger numbers of *trans* eQTLs for both genes and miRNAs compared to the ILD cohort (**Gene:** 110,424 in COPD, 68,050 in ILD; **miRNA:** 557 in COPD, 362 in ILD). In contrast, the ILD cohort had larger numbers of *cis* gene eQTLs and nearly the same number of miRNA eQTLs as the COPD cohort (**Gene:** 8195 in COPD, 10,941 in ILD; **miRNA:** 53 in COPD, 52 in ILD). The proportion of unique genes and miRNAs with at least one *cis* eQTL was similar between COPD and ILD cohorts (***cis* gene:** 6% in COPD, 7% in ILD; ***cis* miRNA**: 2% in COPD, 2% in ILD). However, there was 1.56-fold more *trans* gene eQTLs and 1.53-fold more *trans* miRNA eQTLS in the COPD cohort compared to the ILD cohort. Lastly, the proportion of miRNAs with any eQTL was significantly lower than the proportion of protein-coding genes with any eQTL in both the COPD and ILD cohorts ($p < 0.001$; Fisher's exact test; **Table 2**). Overall, these results suggest that there are more *trans* associations contributing to variability in expression in COPD compared to ILD and that miRNAs have fewer proportions of *cis* and *trans* eQTLs than protein-coding genes in both diseases.

Table 2. Number of gene and miRNA eQTLs in different diseases.

	Interaction Type	COPD (n=111)	ILD (n=113)
No. of Gene-SNP eQTLs	*cis*	8195	10941
	trans	110424	68050
No. of miRNA-SNP eQTLs	*cis*	53	52
	trans	557	362
No. of unique genes with eQTL	*cis*	938 (6%)	1142 (7%)
	trans	6629 (43%)	4249 (28%)
No. of unique miRNAs with eQTL	*cis*	17 (2%)	14 (2%)
	trans	221 (32%)	144 (21%)

Integrated network analysis implicates the miR-34/449 family in COPD and ILD.

Key miRNA regulators of gene expression can more readily be identified in disease states can be more readily identified by "anchoring" expression data with genetic information and building integrative networks.[26] Genes and miRNAs associated with at least one SNP were included in the network analysis ($p < 0.0005$). To build integrative networks within each cohort, we leveraged the causality inference test (CIT).[25] CIT assesses the hypothesis that a potential mediator between an initial random variable and an outcome variable is causal for that outcome. Causal and independent relationships are defined as series of conditions of associations between the three variables, corresponding to SNP, microRNA and mRNA nodes (**Figure 2A**). The number of significant associations obtained at each step of the network construction are presented in **Supplementary Figure 4**. For our study, we focused on relationships where the miRNA is predicted to be the modulator of mRNA expression. A common property of biological networks is that they often display a *scale-free* topology where a few nodes contain the majority of interactions in the network (i.e. the power law).[48,49] We confirmed that our three networks followed a scale-free topology by observing a strong negative linear relationship between the number of predicted interactions for each microRNA and the frequency of microRNA with a certain number of interactions in log scale (**Figure 2B**). We further examined the miRNAs predicted to interact with the most genes in each network (**Figure 2C**, **Supplementary Table 5**).

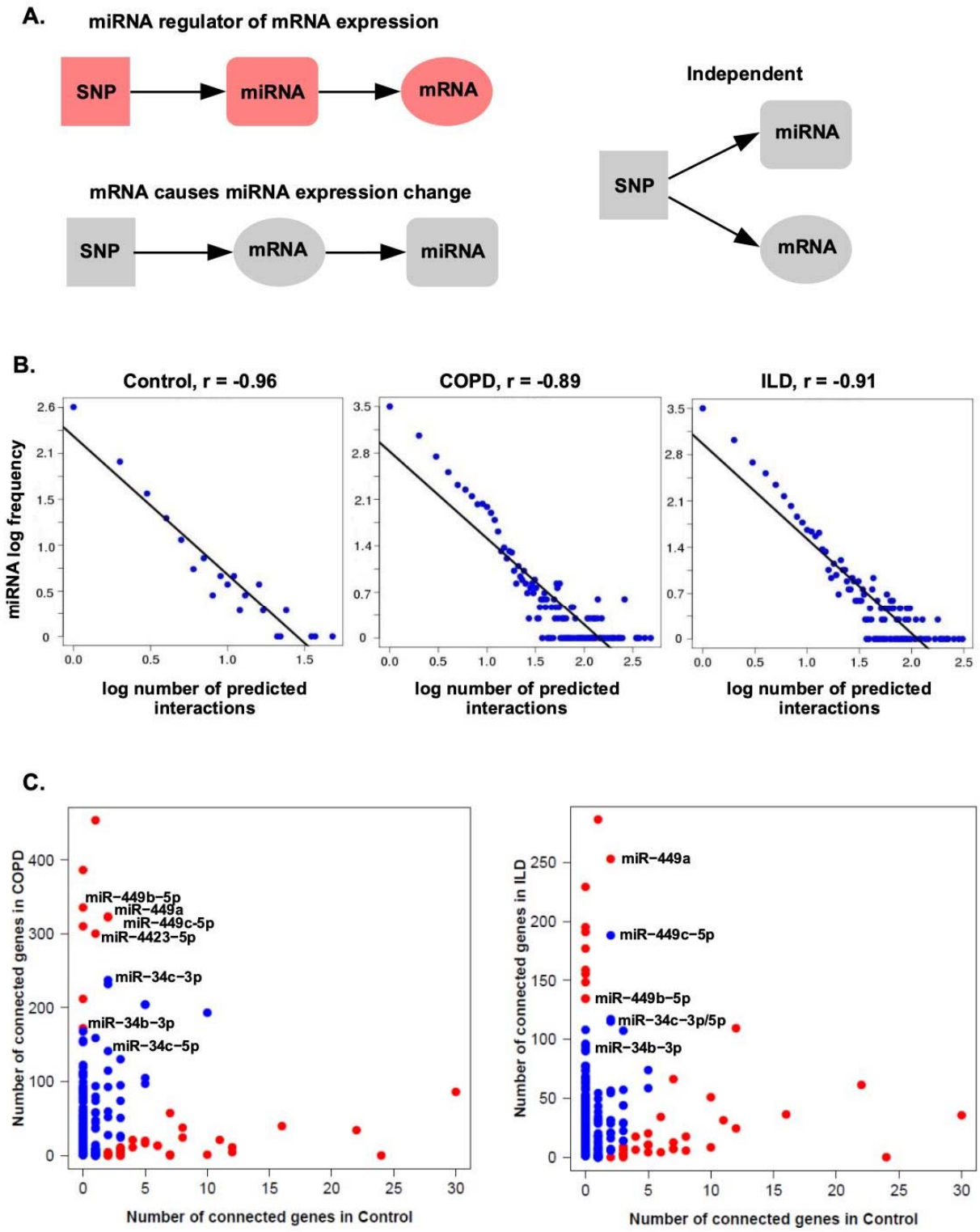

Figure 2. Examining top miRNA in disease-specific integrative networks. A. We select those SNP-miRNA-mRNA triplets where the SNP-mRNA relationship is defined by a miRNA mediator; we filter out independent relationships and those triplets where the SNP is not associated with the miRNA. **B.** The CIT 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).

Members of the miR-449 and miR-34 families were found to be among the most connected to genes in COPD and ILD networks (**Figure 2C**), indicating that the miR-449/34 family has a greater impact on gene expression regulation in the disease groups compared to the control group. The miRNAs in the miR449/34 family had larger numbers of associated genes compared to the network in control samples (**Figure 3A**). Members of miR-449/34 family can promote airway differentiation by repressing the Notch pathway[50]. We observed that the union set of genes (n=406) that positively correlated with any of the miRNAs in this family in COPD or ILD was enriched among genes that increase in expression over time when airway basal cells are differentiated at an air-liquid interface (ALI)[40]. Gene enrichment results were significant by both GSVA ($p < 0.05$; **Figure 3B**) and GSEA ($q < 0.001$; **Figure 3C**). 75 SNPs in COPD (**Supplementary Table 6**) and 60 SNPs in ILD (**Supplementary Table 6**) were associated with members of the miR-449/34 family using the CIT. Some of these SNPs have been previously found to be associated with asthma, inflammation, cancer and other degenerative diseases in the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP)[51]. Allele frequencies for 10 of these SNPs were also significantly associated with COPD and 4 of them with ILD by a Fisher's exact test ($q<0.25$; **Supplementary Figure 5)**.

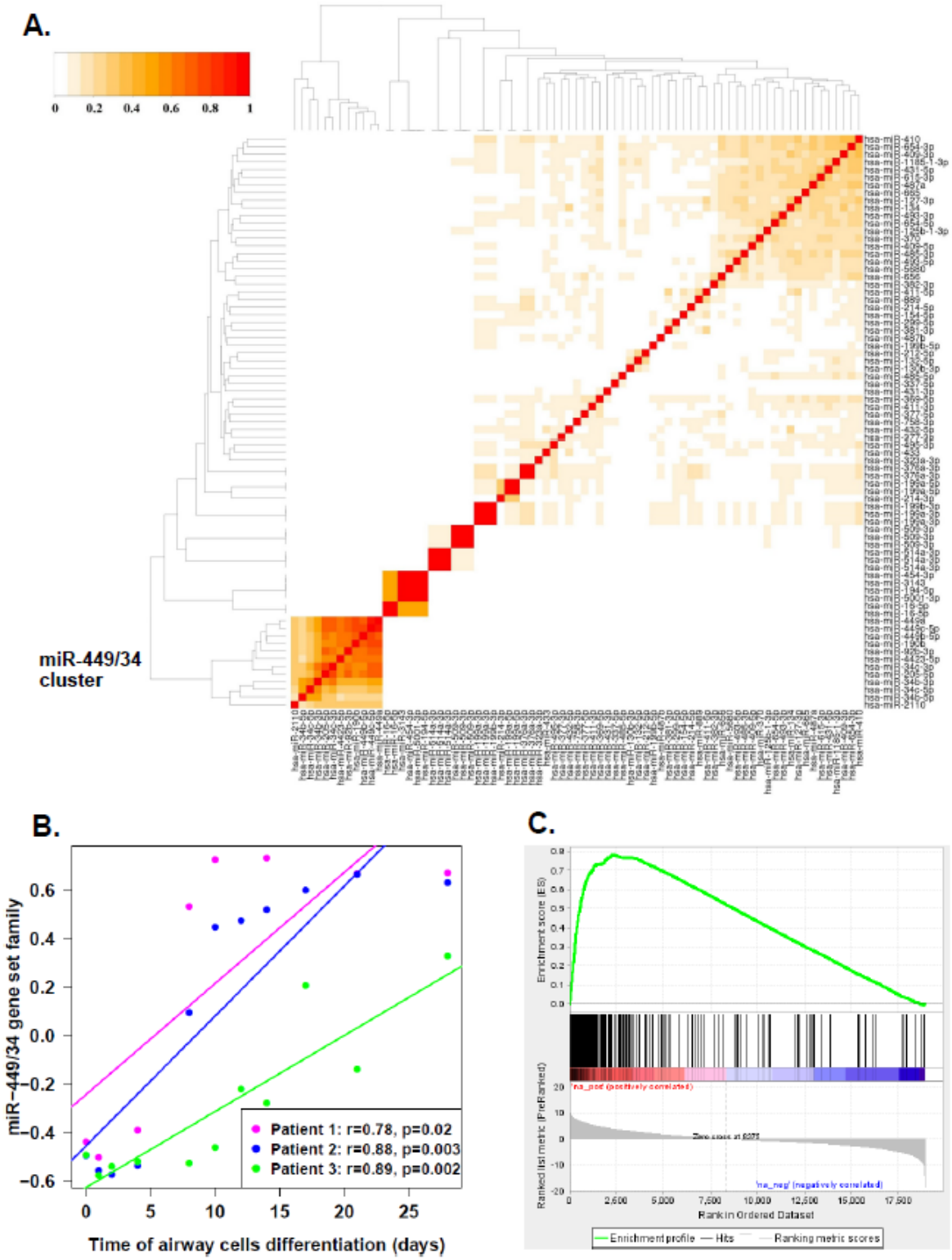

Figure 3. Enrichment of miR-449/34 modules. A. Clustering of miRNA modules based on the Jaccard index revealed a group of strongly overlapping miRNA in the miR-449/34 family. **B.** GSVA was used to predict the activity of the miR-449/34 family in a gene-expression dataset of airway epithelial differentiation. The set of genes that positively correlated with miR-449/34 family (406 genes) were enriched among genes that increase in expression with the airway epithelial cells differentiation ($p < 0.05$; Linear mixed-effects model). **C.** Similarly, enrichment of miR-449/34 gene set family with the airway cells differentiation is shown by GSEA (FDR q-value <0.001).

The canonical and 5' isomiR seeds for miR-34c-5p regulate members of distinct signaling pathways

As noted previously, expression for members of the miR-449/34 family were a part of a miRNA module (M2) that could distinguish two distinct groups of ILD patients and were found among the most connected microRNAs in COPD and ILD specific networks. The miR-449/34 family has been shown to promote ciliogenesis by down-regulating anti-differentiation genes such as NOTCH1[52]. The majority of miRNAs in this family share the same seed sequence, GGCAGTG, which is a conserved heptametrical sequence on the 5' end. Previous studies have revealed that variation in the 5' end of miRNAs can create a novel seed sequence, which allows the 5' isomiR to target a distinct set of genes from the canonical seed[27]. Within the miR-34/449 family, we found relatively high expression of isomiRs within the miR-34c-5p locus that contained an alternative seed sequence representing a 1-base shift to the left from the canonical seed, AGGCAGT (**Figure 4A**). 24 of the top 25 sequences from the miR-34c-5p locus were up-regulated in ILD sample compared to the Control samples (**Supplementary Table 8**), demonstrating that the majority of sequences follow the same expression pattern across samples with respect to disease status.

To explore similarities and differences in putative mRNA targets between canonical and isomiR seeds of miR-34c-5p, Targetscan v6.0[41] was used to predict mRNA targets for each seed. mRNAs were grouped into four categories: predicted targets of the canonical and 5' isomiR seeds, predicted targets of the canonical seed only, predicted targets of the the 5' isomiR seed only, or those not predicted to be a target of either seed. Additionally, the expression of each gene was correlated to the overall expression levels of miR-34c-5p using Spearman correlation within the ILD samples. The distribution of correlation coefficients for groups of genes that were predicted targets of the canonical and/or 5' isomiR seeds were more negative compared to non-predicted targets (Kolmogorov-Smirnov test; $p < 1e\text{-}7$; **Figure 4B**), suggesting that both the canonical seed and the 5' isomiR seed may be negatively regulating target gene expression. We also explored the degree of overlap between genes that were significantly negatively correlated to miR-34c-5p expression (Spearman correlation; FDR q-value < 0.25) and were also a predicted target of either the canonical or 5' isomiR (**Figure 4C**). Interestingly, 47% of the miR-34c-5p 5' isomiR targets were distinct from the miR-34c-5p canonical targets. Using Enrichr,[53] we found that the anti-correlated predicted targets specific to the 5' isomiR were

enriched for genes in the "Ras protein signal transduction" pathway including GRAP, GRB2, YWHAB, RHOA, RAPGEF6, MAPKAPK3, RALA, and SHC3 (p = 0.0001; **Supplementary Table 9**). Anti-correlated predicted targets specific to the miR-34c-5p canonical seed were enriched for the "notch signaling pathway" which contained other Notch-related genes beyond NOTCH1 including ADAM10, PSEN1, HEY1, DLL4, and NOTCH4 (p = 0.003 **Supplementary Table 10**). The predicted targets specific to the canonical seed were also enriched in other Ras-related pathways such as "Ras GTPase binding" and "small GTPase binding" suggesting that the canonical seed and the isomiR seed may be regulating different members of the same signaling pathway.

Validation of expression and miR-34c-5p isomiR activity

The expression of miR-34c-5p was measured in a subset of ILD and Control samples (n = 10 per group) via qRT-PCR and was significantly upregulated with disease (p<0.05, **Supplementary Table 11**, **Supplementary Figure 6**). Similarly, NOTCH1 which is a known target of the canonical miR-34c-5p[54,55] and a predicted target of the 5' isomiR, was validated to be down-regulated in ILD samples compared to Controls when measured by qRT-PCR (p < 0.05; **Supplementary Figure 7**). Finally, qRT-PCR was used to measure the expression of genes in Ras signaling pathway that were anti-correlated predicted targets of the canonical seed (CRK) or the 5' isomiR seed (RALA, GRB2, GRAP, RHOA, ARAP2, CRKL). Five of the seven predicted targets were significantly down-regulated in the subset of ILD samples compared to Controls (p < 0.05; **Supplementary Figure 7**). Additionally, two genes not predicted to be targets of miR-34c-5p seed were examined. An association between RHOC expression and ILD was observed with the mRNA microarrays (p = 0.002) and confirmed with qRT-PCR (p < 0.01) while a lack of association between EGF expression and ILD was observed with the mRNA microarrays (p = 0.230) and confirmed with qRT-PCR (p > 0.05; **Supplementary Figure 7**). These results confirm that associations with ILD determined by mRNA microarrays are largely recapitulated by qRT-PCR.

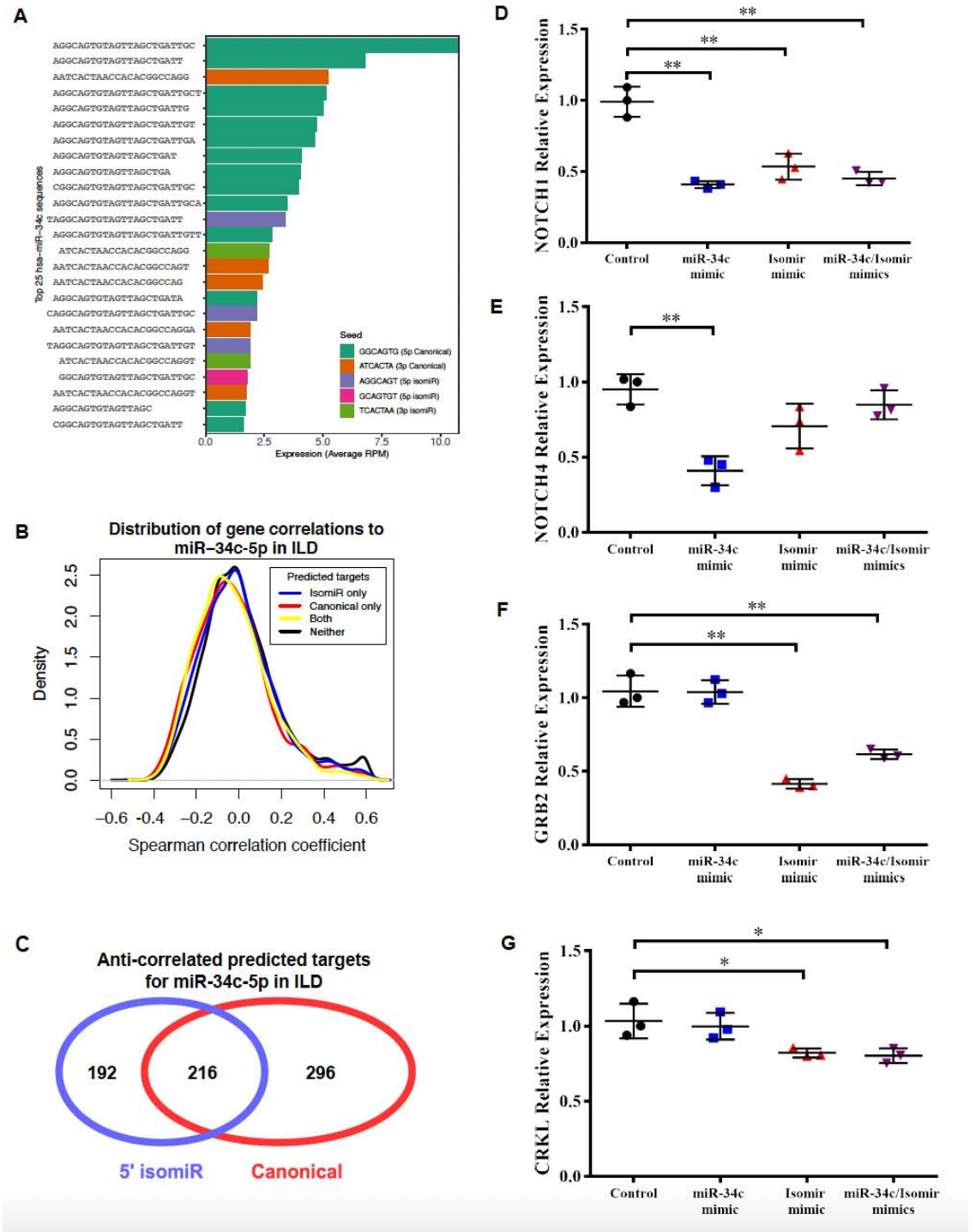

Figure 4. Functional roles of an isomiR for hsa-miR-34c. A. The top 25 highest expressed sequences are shown for the hsa-miR-34c locus**.** Three of these sequences represented an isomiR on the 5' end which contained a non-canonical miRNA seed (purple). **B.** 5' UTR-site predicted mRNA targets by Targetscan are more negatively correlated with the miR-34c isomiR seed than non-predicted targets (Kolmogorov-Smirnov test, p<1e-7). **C.** The overlap of negatively correlated (FDR<0.25) and 5' UTR-site predicted targets of miR-34c and 5' isomiR. miR-34c targets are significantly enriched for Notch signaling pathway by Enrichr (p<0.02); 5' isomiR targets are significantly enriched for Ras signaling pathway by Enrichr (p<0.002). **D.** IMR90 fibroblast cells show significant repression of NOTCH1 with all experimental transfections (p<0.005, p<0.005, p<0.005). **E.** NOTCH4 expression is only downregulated by miR-34c mimic transfection (p<0.005) and not by transfection of the isomiR. **F.** GRB2 is significantly downregulated with the miR-34c 5' isomiR mimic transfection (p<0.005, p<0.005) but not the miR-34c canonical transfection. **G.** CRKL is also significantly downregulated with the miR-34c 5' isomiR mimic transfection (p<0.05, p<0.05) but not the miR-34c canonical transfection.

Given the differences in the sets of predicted target genes for each miR-34c-5p seed sequence, we next sought to validate activity of the miR-34c 5' isomiR. Human lung fibroblasts (IMR90) were transfected with mimics of miR-34c-5p, the miR-34c-5p 5' isomiR, or both sequences. In the ILD cohort, NOTCH1, a known target of miR-34c-5p, was significantly anti-correlated with miR-34c-5p expression (FDR q-value = 0.016) and was a predicted target of the 5' isomiR seed as well. Fibroblasts transfected with any mimics had significant down-regulation of NOTCH1 expression ($p<0.005$, **Figure 4D**). NOTCH4 is a validated target of the canonical form of miR-34c-5p,[55] but was not a predicted targeted by the miR-34c-5p 5' isomiR. Expression of NOTCH4 was only significantly decreased with canonical miR-34c-5p overexpression ($p<0.005$, **Figure 4E**). GRB2 and CRKL, genes involved in the Ras pathway and predicted targets of only the isomiR, were significantly downregulated only with the miR-34c-5p 5' isomiR mimic transfections and not the canonical miR-34c-5p transfection ($p<0.05$, **Figure 4F,G**). Overall, these results demonstrate the ability of the miR-34c-5p isomiR to modulate the expression of predicted targets distinct from the canonical seed sequence.

Discussion

We applied unsupervised and integrative analyses to multi-omic data to characterize the role of miRNAs in the setting of COPD and ILD. Novel subgroups of patients were identified with miRNAs that were differentially expressed in either disease. Several sample subgroups were enriched for disease patients and/or had significantly worse lung function phenotypes compared the subgroup with the most Control subjects (sample cluster S1). Interestingly, several disease patients were also in cluster S1. We have previously shown that both gene and miRNA expression can vary with regional emphysema severity across different sections within the lungs of patients with COPD[15]. Therefore, COPD or ILD patients clustering with Control patients could be due to variable sampling of diseased regions within the lung. Conversely, a smaller number of control patients clustered in one of the disease-associated subgroups. This could be due to the fact that in diseases such as COPD, aberrant processes like emphysema can begin to occur before the onset of overall lung function decline[56]. Overall, these molecular subgroups defined by miRNA expression represent previously unappreciated patient subclasses that may require distinct therapeutic modalities.

In order to identify potential miRNA regulators of gene expression, we leveraged genetic and gene expression data available on a subset of samples. We first identified eQTLs for both miRNAs and genes and found that, in contrast to gene expression, relatively fewer miRNAs were associated a *cis* SNP compared to protein-coding genes, suggesting that their regulation is more dependent on other factors within this disease setting. However, we were able to detect a large number of *trans* interactions, suggesting the lack of *cis* interactions is not simply due to lack of power. To identify potential miRNA-gene interactions within each disease, we applied the CIT and observed a significantly higher proportion of interactions for specific miRNA between disease and normal networks, including interactions for the miR-34 and miR-449 families. These miRNA families regulates mucociliary differentiation by directly targeting the NOTCH pathway[50,52,54,55,57]. Gene modules for these miRNAs in the COPD and ILD integrative networks were associated with airway epithelial cell differentiation in an independent dataset. Interestingly, miR-34b and miR-34c have been associated with emphysema severity[58]. We also found that the primate-specific miR-4423 was differentially connected in the COPD network. Expression of this miRNA is highly connected with the miR-449/34 family and has been previously associated with airway differentiation in smokers with lung cancer[45]. As the SNPs associated with the miR-449/34 family were on different chromosomes than the miRNAs, future studies will be needed to elucidate the mechanisms by which the trans-genetic variants can modulate the expression of these miRNAs.

Finally, we leveraged the ability of small-RNA sequencing to characterize sequence variation beyond expression levels and identified an isomiR with a novel seed sequence at the miR-34c-5p locus. This seed sequence was predicted to target a distinct set of genes from the canonical seed sequence and was enriched for genes involved in Ras signaling. This pathway has been previously implicated in tight junction formation in normal airway epithelial barrier formation.[59] Down-regulation of this pathway may be necessary for normal differentiation of ciliary cells in the airway as well as the aberrant differentiation observed in a subset of ILD patients. Additional experiments in animal models will be necessary to determine if inhibition of these miRNAs can ameliorate disease phenotypes. The aggregation of these findings suggests a role for aberrant miRNA and isomiR regulation of airway differentiation in a subset of COPD or ILD patients and the inhibition of this process may represent a novel therapeutic approach for disease treatment.

Acknowledgements

This work was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute with funding from the Lung Genomics Research Consortium (RC2-HL101715) and R01HL118542 (M.E.L., A.S.).

.

Author contributions

A.B.P., J.D.C. and J.B. contributed to data analysis. L.L., G.L., J.X., and Y.O.A. contributed to data generation. C.G. and D.T. performed experiments. B.J.G., J.T., I.V.Y., and F.S. contributed to sample collection and processing. A.B.P. and J.D.C. wrote the manuscript. M.W.G., D.A.S., N.K., A.S., and M.E.L. contributed to the overall study design. All authors reviewed the manuscript.

References

1. Osei ET, Florez-Sampedro L, Timens W, Postma DS, Heijink IH, Brandsma C-A. Unravelling the complexity of COPD by microRNAs: it's a small world after all. *Eur Respir J*. 2015;46(3).
2. Raherison C, Girodet P-O. Epidemiology of COPD. *Eur Respir Rev*. 2009;18(114).
3. Steiling K, van den Berge M, Hijazi K, et al. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. *Am J Respir Crit Care Med*. 2013;187(9):933-942. doi:10.1164/rccm.201208-1449OC
4. Skolnik K, Ryerson CJ. Unclassifiable interstitial lung disease: A review. *Respirology*. 2016;21(1):51-56. doi:10.1111/resp.12568
5. Nalysnyk L, Cid-Ruzafa J, Rotella P, Esser D. Incidence and prevalence of idiopathic pulmonary fibrosis: review of the literature. *Eur Respir Rev*. 2012;21(126):355-361. doi:10.1183/09059180.00002512
6. Gribbin J, Hubbard RB, Le Jeune I, Smith CJP, West J, Tata LJ. Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK. *Thorax*. 2006;61(11):980-985. doi:10.1136/thx.2006.062836
7. Raghu G, Weycker D, Edelsberg J, Bradford WZ, Oster G. Incidence and prevalence of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*. 2006;174(7):810-816. doi:10.1164/rccm.200602-163OC
8. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med*. 2011;183(6):788-824. doi:10.1164/rccm.2009-040GL
9. Winter J, Jung S, Keller S, Gregory RI, Diederichs S. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol*. 2009;11(3):228-234. doi:10.1038/ncb0309-228
10. Sayed D, Abdellatif M. MicroRNAs in development and disease. *Physiol Rev*. 2011;91(3):827-887. doi:10.1152/physrev.00006.2010
11. Kusko RL, Brothers JF, Tedrow J, et al. Integrated Genomics Reveals Convergent Transcriptomic Networks Underlying Chronic Obstructive Pulmonary Disease and Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 2016;194(8):948-960. doi:10.1164/rccm.201510-2026OC

12. Campbell JD, McDonough JE, Zeskind JE, et al. A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK. *Genome Med*. 2012;4(8):67. doi:10.1186/gm367

13. Yang I V, Pedersen BS, Rabinovich E, et al. Relationship of DNA methylation and gene expression in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*. 2014;190(11):1263-1272. doi:10.1164/rccm.201408-1452OC

14. Liu G, Friggeri A, Yang Y, et al. miR-21 mediates fibrogenic activation of pulmonary fibroblasts and lung fibrosis. *J Exp Med*. 2010;207(8):1589-1597. doi:10.1084/jem.20100035

15. Christenson SA, Brandsma C-A, Campbell JD, et al. miR-638 regulates gene expression networks associated with emphysematous lung destruction. *Genome Med*. 2013;5(12):114. doi:10.1186/gm519

16. Milosevic J, Pandit K, Magister M, et al. Profibrotic role of miR-154 in pulmonary fibrosis. *Am J Respir Cell Mol Biol*. 2012;47(6):879-887. doi:10.1165/rcmb.2011-0377OC

17. Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat Genet*. 2012;44(8):841-847. doi:10.1038/ng.2355

18. Ng S, Collisson EA, Sokolov A, et al. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*. 2012;28(18):i640-i646. doi:10.1093/bioinformatics/bts402

19. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26(12):i237-45. doi:10.1093/bioinformatics/btq182

20. Yoo S, Takikawa S, Geraghty P, et al. Integrative Analysis of DNA Methylation and Gene Expression Data Identifies EPAS1 as a Key Regulator of COPD. 2015;(January). doi:10.1371/journal.pgen.1004898

21. Vignes M, Vandel J, Allouche D, et al. Gene Regulatory Network Reconstruction Using Bayesian Networks, the Dantzig Selector, the Lasso and Their Meta-Analysis. Rattray M, ed. *PLoS One*. 2011;6(12):e29165. doi:10.1371/journal.pone.0029165

22. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical

Evaluation. *J Mach Learn Res*. 2010;11:171-234.

23. Dondelinger F, Husmeier D, Lèbre S. Dynamic Bayesian networks in molecular plant science: inferring gene regulatory networks from multiple gene expression time series. *Euphytica*. 2012;183(3):361-377. doi:10.1007/s10681-011-0538-3

24. Omranian N, Eloundou-Mbebi JMO, Mueller-Roeber B, et al. Gene regulatory network inference using fused LASSO on multiple data sets. *Sci Rep*. 2016;6:20533. doi:10.1038/srep20533

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

26. 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):490. doi:10.1038/msb.2011.23

27. Tan GC, Dibb N. IsomiRs have functional importance. *Malays J Pathol*. 2015;37(2):73-81. http://www.ncbi.nlm.nih.gov/pubmed/26277662.

28. Tan GC, Chan E, Molnar A, et al. 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res*. 2014;42(14):9424-9435. doi:10.1093/nar/gku656

29. 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

30. 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

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

32. 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

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

34. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc*. 1995;57(1):289-300. doi:10.2307/2346101

35. 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

36. 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

37. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012;28(10):1353-1358. doi:10.1093/bioinformatics/bts163

38. 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

39. 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

40. 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

41. 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. doi:10.1016/j.cell.2004.12.035

42. Lino Cardenas CL, Henaoui IS, Courcot E, et al. miR-199a-5p Is upregulated during fibrogenic response to tissue injury and mediates TGFbeta-induced lung fibroblast activation by targeting caveolin-1. *PLoS Genet*. 2013;9(2):e1003291. doi:10.1371/journal.pgen.1003291

43. Pottier N, Maurin T, Chevalier B, et al. Identification of keratinocyte growth factor as a target of microRNA-155 in lung fibroblasts: implication in epithelial-mesenchymal interactions. *PLoS One*. 2009;4(8):e6718. doi:10.1371/journal.pone.0006718

44. Lizé M, Herr C, Klimke A, Bals R, Dobbelstein M. MicroRNA-449a levels increase by several orders of magnitude during mucociliary differentiation of airway epithelia. *Cell Cycle*. 2010;9(22):4579-4583. doi:10.4161/cc.9.22.13870

45. 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

46. Yang I V, Coldren CD, Leach SM, et al. Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis. *Thorax*. June 2013:1-8. doi:10.1136/thoraxjnl-2012-202943

47. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1620):20120362. doi:10.1098/rstb.2012.0362

48. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 2004;5(2):101-113. doi:10.1038/nrg1272

49. Barabasi AL, Bonabeau E. Scale-free networks. *Sci Am*. 2003;288(5):60-69.

50. 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

51. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*. 2014;30(12):i185-i194. doi:10.1093/bioinformatics/btu273

52. Marcet B, Chevalier B, Luxardi G, et al. Control of vertebrate multiciliogenesis by miR-449 through direct repression of the Delta/Notch pathway. *Nat Cell Biol*. 2011;13(6):693-699. doi:10.1038/ncb2241

53. Kuleshov M V, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-7. doi:10.1093/nar/gkw377

54. Liu X-D, Zhang L-Y, Zhu T-C, Zhang R-F, Wang S-L, Bao Y. Overexpression of miR-34c inhibits high glucose-induced apoptosis in podocytes by targeting Notch signaling pathways. *Int J Clin Exp Pathol*. 2015;8(5):4525-4534.

55. Bae Y, Yang T, Zeng H-C, et al. miRNA-34c regulates Notch signaling during bone development. *Hum Mol Genet*. 2012;21(13):2991-3000. doi:10.1093/hmg/dds129

56. McDonough JE, Yuan R, Suzuki M, et al. Small-airway obstruction and emphysema in chronic obstructive pulmonary disease. *N Engl J Med*. 2011;365(17):1567-1575. doi:10.1056/NEJMoa1106955

57. Lizé M, Klimke A, Dobbelstein M. MicroRNA-449 in cell fate determination. *Cell Cycle*. 2011. doi:10.4161/cc.10.17.17181

12 58. Savarimuthu Francis SM, Davidson MR, Tan ME, et al. MicroRNA-34c is associated with emphysema
13 severity and modulates SERPINE1 expression. *BMC Genomics*. 2014;15:88. doi:10.1186/1471-2164-
14 15-88
15 59. Durgan J, Tao G, Walters MS, et al. SOS1 and Ras regulate epithelial tight junction formation in the
16 human airway through EMP1. *EMBO Rep*. 2015;16(1):87-96. doi:10.15252/embr.201439218
17
18