MicroRNA Markers for Acute Respiratory Distress Syndrome and Shared Genetic Architecture of Asthma With Allergic Diseases: A Genome-Wide Cross Trait Analysis of 112,000 Individuals From UK Biobank

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42066872">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42066872</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
MICRORNA MARKERS FOR ACUTE RESPIRATORY DISTRESS SYNDROME AND SHARED GENETIC ARCHITECTURE OF ASTHMA WITH ALLERGIC SYMPTOMS: A GENOME-WIDE CROSS TRAIT ANALYSIS OF 112,000 INDIVIDUALS FROM UK BIOBANK

ZHAOZHONG ZHU

A Dissertation Submitted to the Faculty of
The Harvard T.H. Chan School of Public Health
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science
in the Department of Environmental Health

Harvard University

Boston, Massachusetts.

May 2017
MicroRNA Markers for Acute Respiratory Distress Syndrome and Shared Genetic Architecture of Asthma with Allergic Diseases: A Genome-wide Cross Trait Analysis of 112,000 Individuals from UK Biobank

Abstract
MicroRNAs (miRNAs) mediate inflammation and infection, common manifestations of acute respiratory distress syndrome (ARDS). I proposed to examine whether miRNAs from whole blood serve as potential diagnostic or prognostic biomarkers for ARDS. This nested case-control study (N=530) included 2 cohorts of ARDS patients and critically ill at-risk controls. Whole blood miRNAs were profiled and logistic regression analyses or survival analysis were performed to identify miRNA correlations with ARDS. Receiver operating characteristic (ROC) analysis was used for evaluating miRNA diagnostic performance along with the Lung Injury Prediction Score (LIPS). ARDS patient whole blood miRNA profiling revealed that miR-181a and miR-92a were risk biomarkers of ARDS, whereas miR-424 was a protective biomarker. Addition of these miRNAs increased the risk diagnosis of ARDS from LIPS. We also built a miRNA classifier for predicting ARDS 28-day mortality. In addition, I also did a genome-wide cross trait analysis of 112,000 individuals from UK Biobank to investigate shared genetic architecture of asthma with allergic diseases. Clinical and epidemiological data suggest that asthma and allergic diseases, such as allergic rhinitis and eczema, are associated. One hypothesis to account for the relationship is that these diseases share a common genetic etiology. Heritability has been estimated at high level for asthma and allergic diseases, which suggests that the genetic contribution to them can be significant. We analyzed genome-wide single-nucleotide polymorphism (SNP) data for the asthma and allergic diseases in 35,783 cases and 76,768
controls of European ancestry. We used a logistic regression, linkage disequilibrium score regression, cross trait meta-analysis and GTEx tissue enrichment analysis in this study. We have found a strong genetic correlation between asthma and allergic diseases ($R_g=0.75$, $P=6.84 \times 10^{-62}$). Cross trait analysis and identified 90 loci. GTEx tissue enrichment analysis showed shared genetic loci between asthma and allergic diseases were enriched in skin and esophageal tissue. Our GWAS study has highlighted shared genetic pathway in immune and inflammatory systems and skin/esophageal tissue by asthma and allergic diseases. This work should support the idea that common patterns of association between asthma and allergy implicate shared biological processes and advance understanding of the molecular mechanisms underlying co-morbid asthma and allergic diseases.
# Table of Contents

Abstract ................................................................................................................................. II
List of Tables ............................................................................................................................ VI
List of Figures ........................................................................................................................ VI
Acknowledgments .................................................................................................................. VIII

Chapter 1 : Introduction ........................................................................................................ 1

Chapter 2 : Whole Blood MicroRNA Markers are Associated with Acute Respiratory Distress Syndrome .................................................................................................................. 10

Abstract ................................................................................................................................. 11
Introduction ............................................................................................................................. 12
Methods ................................................................................................................................... 13
  Study design .......................................................................................................................... 13
  RNA isolation ...................................................................................................................... 15
  miRNA profiling ................................................................................................................ 15
  Statistical analysis ............................................................................................................. 15
Results ..................................................................................................................................... 16
  miRNA screening and validation ......................................................................................... 16
  Sepsis and pneumonia stratification analysis ..................................................................... 19
  miRNA diagnostic performance ......................................................................................... 20
Discussion ............................................................................................................................... 22
Conclusions ............................................................................................................................. 26
References .............................................................................................................................. 27

Chapter 3 : Whole Blood MicroRNAs as a Prognostic Classifier for Acute Respiratory Distress Syndrome 28-day Mortality ............................................................................................................................. 30

Introduction ............................................................................................................................. 31
Method ..................................................................................................................................... 31
  RNA isolation and complementary DNA (cDNA) synthesis ............................................ 32
  miRNA profiling ............................................................................................................... 32
  miRNA data quality control ............................................................................................ 32
  Imputation ......................................................................................................................... 33
Results ..................................................................................................................................... 33
Conclusion ............................................................................................................................... 33
List of Tables
Table 2.1 miRNA associations with ARDS in discovery cohort, validation cohorts, combined cohorts, and meta-analysis................................................................. 18
Table 2.2 Gene set enrichment analysis of 22 candidate miRNAs in whole miRNA set......... 19
Table 2.3 Stratification analysis of miRNA associations with ARDS in validation cohorts.... 20
Table 2.4 Diagnostic performance of sepsis/pneumonia model, and miRNA combined model for ARDS........................................................................................................... 21
Table 4.1 Genetic correlation between each trait............................................................. 50
Table 4.2 Cross trait analysis result of asthma and eczema in replication cohort ............... 54
List of Figures
Figure 2.1 Study design of discovery cohort, validation cohort, and gene set enrichment analysis (GSEA) ................................................................. 14
Figure 3.1 Kaplan–Meier survival analysis according to the miRNA classifier grouped by expression level ........................................................................ 34
Figure 4.1 Study population and design ...................................................................... 42
Figure 4.2 Manhattan plot of cross trait analysis of asthma and allergic diseases .......... 47
Figure 4.3 GTEx tissue expression enrichment analysis ............................................... 52
Figure 4.4 GO enrichment analysis identified pathway related to shared genes between asthma and allergy ........................................................................ 55
Acknowledgments

I am thankful to Dr. David Christiani with Harvard T.H. Chan School of Public Health to understand the acute respiratory distress syndrome (ARDS), also thankful to Dr David Christiani providing all resources related to ARDS project. In addition, we were able to tease out the specific autonomic components (accelerations and decelerations) that are affected by work exposures on different time scales. It is my hope that this research would inform further research on mechanisms of these effects to provide a platform for targeted interventions.

This research would not have been possible without the help of many people. First, I would like to thank Drs. David Christiani, Liming Liang, Quan Lu and Andrea Baccarelli for serving on my research committee. Their feedback and guidance during this work was tremendous. I would also like to express my sincere appreciation to my advisor, Dr. David Christiani, for giving me the opportunity to work on this project for my doctoral degree, and for his continued support throughout this work. I also want to appreciate Dr. Phil Lee, Dr Wonil Chung, Dr Po-Ru Loh for their guidance in statistical analysis. My third project has been conducted using the UK Biobank resource under application number 16549. We would like to thank the participants and researchers from the UK Biobank who significantly contributed or collected data. We thank to Dr Vernerri Anttila and Dr Steven Gazal for their statistical advice.

Finally, I am forever grateful to my wife Conglin for her love and support, my son Charles, my parents Shan and Xi, and my friends for all their encouragements.

Zhaozhong Zhu

May 2017
Chapter 1: Introduction
Respiratory diseases are prevalent worldwide. The diagnosis, prognosis and treatment of them are insufficiently provided. In this thesis, I aim to use both genetics and epigenetics methods to investigate the markers for two respiratory diseases. Acute respiratory distress syndrome (ARDS) is a complex syndrome of acute inflammation and infection caused by direct and indirect injury to the lung. This syndrome affects approximately 200,000 people annually in the United States, carries a mortality rate of 40% and is a major cause of intensive care unit (ICU) morbidity and mortality worldwide (Bellani et al. 2016). Emerging viral infections, such as severe acute respiratory syndrome coronavirus and H1N1 swine-origin influenza virus have become important causes of ARDS in humans and possess the potential for pandemic spread (Ramsey and Kumar 2011). Importantly, its complex etiology and a lack of reliable biomarkers have complicated its diagnosis and treatment. A number of protein-based biomarkers have been identified from plasma (Ware et al. 2013), but none has been translated to clinical diagnostic routines and more comprehensive study design is needed to identify new mediators for ARDS pathogenic mechanisms (Rocco and Nieman 2016). Further, no previous studies have evaluated the potential or performance of whole blood microRNAs (miRNAs) to diagnose ARDS or predict the disease outcome.

miRNAs, a group of small non-coding RNAs, regulate gene expression by binding to specific target sites on messenger RNA to either repress or degrade the targets. Previous studies have discovered important roles for miRNAs in many disorders, including inflammation and infection (Lu et al. 2005, Necela et al. 2011). Thus, miRNA expression patterns may be used to construct a diagnostic classifier for better disease detection (Schultz et al. 2014, Lin et al. 2015). Studies suggested an involvement of miRNAs in the development of ARDS. In a rat model of ARDS,
miRNA profiling from lung tissue demonstrated the changes of multiple miRNAs over control tissues (Huang et al. 2014). Also we recently demonstrated whole blood miRNAs can be potentially valuable for predicting ARDS 28-day mortality since they are comparable to APACHE III prediction ability (Zhu et al. 2016). Yet it has not been tested whether whole blood miRNAs may serve as a biomarker for the risk of human ARDS or in experimental animals. Whole blood offers several advantages for miRNA profiling compared to other tissue types. It contains rich immune cell- and tissue-specific miRNAs with low risk of miRNA expression noise from additional serum or plasma isolation steps or sample contamination (Patnaik et al. 2012).

In first two studies we compared miRNA expression in whole blood preparations from large populations of ARDS patients and critically ill at-risk controls or ARDS survival patients and non-survival patient.

In the third study, we aim to find the shared genetic architecture of asthma with allergic diseases. Asthma is a chronic respiratory syndrome that is characterized by abnormal and inflamed mucosa of the airways, wheezing, and shortness of breath. Allergic diseases are immune responses for allergies, such as allergic rhinitis and atopic dermatitis (eczema). Asthma, allergic rhinitis and eczema all belong to type I hypersensitivity, which is an immune response to foreign antigen and associated with immunoglobulin E (IgE)-mediated inflammation (Lenz 2007, Wallace et al. 2008). Genetic studies offer a structured means of understanding the causes of asthma and allergic diseases, as well as identifying targets that can be used to treat the syndrome (Torgerson et al. 2011, Moffatt et al. 2010, Bunyavanich et al. 2014, Ramasamy et al. 2011, Pillai et al. 2009).
Clinical and epidemiological data suggest that asthma and allergy are associated (Leynaert et al. 2000, Brauer et al. 2002). Several studies have identified allergic diseases, such as allergic rhinitis and eczema, as a risk factor for asthma, with the prevalence of allergic rhinitis in asthmatic patients being 80% to 90% (Leynaert et al. 2000, Pariente et al. 1997). These studies and others demonstrate that the coexistence of asthma and allergy is frequent, that allergy usually precedes asthma. Also our previous epigenetic study has identified methylation loci linked to asthma and allergy via IgE pathway (Liang et al. 2015).

One hypothesis to account for the similar symptoms and conditions is that these diseases share a common genetic etiology. Cotsapas et al discovered nearly half of loci in genome-wide association studies (GWAS) of an individual disease influence risk to at least two diseases, indicating the shared genetic architecture of similar diseases (Cotsapas et al. 2011). As each of the shared or similar risk factors has strong genetic influences on disease risk, the observed clustering of multiple risk factors could be due to an overlap in the causal genes and pathways (Criswell et al. 2005, Cross-Disorder Group of the Psychiatric Genomics 2013, Global Lipids Genetics et al. 2013, Pickrell et al. 2016, Shi, Kichaev, and Pasaniuc 2016, Lane et al. 2016, Hobbs et al. 2017, Emdin et al. 2017). In addition, grouping variants by the traits they influence should provide insight into the specific biological processes underlying co-morbidity and disease risk. Clinical and epidemiological studies have found asthma and allergic diseases can occur either in the same individual or in closely related family members (Ober and Yao 2011, Cookson 2004, Belsky et al. 2013, Holgate 1999), suggesting potential pleiotropic effect. The heritability has been estimated at varying between 35% and 95% for asthma (Ober and Yao 2011, Duffy et al. 1990) and 33% and 91% for allergic rhinitis, 71% and 84% for atopic dermatitis (eczema), 34% and 84% for serum IgE levels, which suggests that the genetic contribution to them can be
significant (Ober and Yao 2011). Hinds and colleagues found the shared genetic etiology for 38 allergic diseases (Hinds et al. 2013). However, understanding has been limited only 16 genome-wide shared susceptibility loci which were based on self-reported phenotype.

Thus, in order to increase our knowledge of shared genetic determinants influencing physician diagnosed allergic diseases and asthma, and potentially discover novel loci, we proposed to investigate the genetic commonality in two diseases, asthma and allergic diseases (hay fever/allergic rhinitis or eczema). We conducted a large scale GWAS analysis based on these traits to explore genetic correlations and shared genetic components among these traits using UK Biobank data, which is the largest and most complete European biobank available at present. Also we applied two independent public available GWAS studies, the GABRIEL asthma GWAS study (Moffatt et al. 2010) and the EArly Genetics & Lifecourse Epidemiology (EAGLE) eczema consortium study (Paternoster et al. 2015) to replicate our findings.
References


Chapter 2 : Whole Blood MicroRNA Markers are Associated with Acute Respiratory Distress Syndrome

Zhaozhong Zhu1,ScD, Liming Liang2,3, PhD, Ruyang Zhang1,4, PhD, Yongyue Wei1,4, PhD, Li Su1, MS, Paula Tejera1, PhD, Yichen Guo1, Zhaoxi Wang1, PhD, Quan Lu1, PhD, Andrea A. Baccarelli1, PhD, Xi Zhu5, MD, Ednan K. Bajwa6, MD/MPH, B. Taylor Thompson6, MD, Guo-Ping Shi7, ScD, David C. Christiani1,6, MD/MPH/MS

1 Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA
2 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
3 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA
4 Department of Environmental Health, Department of Epidemiology and Biostatistics, Ministry of Education Key Laboratory for Modern Toxicology, School of Public Health, Nanjing Medical University, Nanjing, China
5 Department of Critical Care Medicine, Peking University Third Hospital, Beijing, China
6 Pulmonary and Critical Care Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA
7 Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

Corresponding author: David C. Christiani, Department of Environmental Health, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue, Boston, MA, USA. (Email: dchris@hsph.harvard.edu; Tel: 617-432-1641; Fax: 617-726-9274)

Running head: ARDS miRNA biomarkers

Keywords: ARDS, microRNA, LIPS, diagnosis, whole blood

Main text word counts: 2724

Abstract word counts: 248
Abstract

Background: MicroRNAs (miRNAs) mediate inflammation and infection, common manifestations of acute respiratory distress syndrome (ARDS). We proposed to examine whether miRNAs from whole blood serve as potential diagnostic biomarkers for human ARDS.

Methods: This nested case-control study (N=530) included 2 cohorts of ARDS patients and critically ill at-risk controls. Whole blood miRNAs were profiled and logistic regression analyses were performed to identify miRNA correlations with ARDS. Selected miRNA markers were also assessed for their role in sepsis and pneumonia associated with ARDS by stratification analysis. Receiver operating characteristic (ROC) analysis was used for evaluating miRNA diagnostic performance along with the Lung Injury Prediction Score (LIPS).

Results: Statistical analysis was performed on 294 miRNAs selected from 754 miRNAs after quality control screening. Logistic regression identified 22 miRNAs from the 156-patient discovery cohort as potential risk or protective markers of ARDS. Three miRNAs (miR-181a, miR-92a, and miR-424) remained significantly associated with ARDS from 156-patient discovery cohort, 373-patient independent validation cohort (FDR q<0.05), and meta-analysis (p<0.001). ROC analysis demonstrated a LIPS baseline area-under-the-curve (AUC) value of ARDS at 0.708 (95% CI: 0.651-0.766). Addition of miR-181a, miR-92a, and miR-424 increased the baseline AUC to 0.723 (95% CI: 0.667-0.778) with a relative integrated discrimination improvement of 2.40 (p=0.005) and the category-free net reclassification index at 27.21% (p=0.01).

Conclusions: ARDS patient whole blood miRNA profiling revealed that miR-181a and miR-92a were risk biomarkers of ARDS, whereas miR-424 was a protective biomarker. Addition of these miRNAs increased the risk diagnosis of ARDS from LIPS.
Introduction

Acute respiratory distress syndrome (ARDS) is a complex syndrome of acute inflammation and infection caused by direct and indirect injury to the lung. This syndrome affects approximately 200,000 people annually in the United States, carries a mortality rate of 40% and is a major cause of intensive care unit (ICU) morbidity and mortality worldwide (Bellani et al. 2016). Emerging viral infections, such as severe acute respiratory syndrome coronavirus and H1N1 swine-origin influenza virus have become important causes of ARDS in humans and possess the potential for pandemic spread (Ramsey and Kumar 2011). Importantly, its complex etiology and a lack of reliable biomarkers have complicated its diagnosis and treatment. A number of protein-based biomarkers have been identified from plasma (Ware et al. 2013), but none has been translated to clinical diagnostic routines and more comprehensive study design is needed to identify new mediators for ARDS pathogenic mechanisms (Rocco and Nieman 2016). Further, no previous studies have evaluated the potential or performance of whole blood microRNAs (miRNAs) to diagnose ARDS.

miRNAs, a group of small non-coding RNAs, regulate gene expression by binding to specific target sites on messenger RNA to either repress or degrade the targets. Previous studies have discovered important roles for miRNAs in many disorders, including inflammation and infection (Lu et al. 2005, Necela et al. 2011). Thus, miRNA expression patterns may be used to construct a diagnostic classifier for better disease detection (Schultz et al. 2014, Lin et al. 2015). Studies suggested an involvement of miRNAs in the development of ARDS. In a rat model of ARDS, miRNA profiling from lung tissue demonstrated the changes of multiple miRNAs over control tissues (Huang et al. 2014). Also we recently demonstrated whole blood miRNAs can be potentially valuable for predicting ARDS 28-day mortality since they are comparable to
APACHE III prediction ability (Zhu et al. 2016). Yet it has not been tested whether whole blood miRNAs may serve as a biomarker for the risk of human ARDS or in experimental animals. Whole blood offers several advantages for miRNA profiling compared to other tissue types. It contains rich immune cell- and tissue-specific miRNAs with low risk of miRNA expression noise from additional serum or plasma isolation steps or sample contamination (Patnaik et al. 2012).

In this study we compared miRNA expression in whole blood preparations from large populations of ARDS patients and critically ill at-risk controls. We identified three miRNAs as novel ARDS markers that survived from 2 cohorts. Combination of these miRNAs with standard clinical test enhanced ARDS risk diagnostic power.

**Methods**

**Study design**
This nested case-control study was part of a molecular epidemiology study of ARDS (MEARDS) that has over 4000 patients and includes both ARDS patients and at-risk controls who were critically ill patients admitted to the ICU of the Massachusetts General Hospital (MGH, Boston, MA) and Beth Israel Deaconess Medical Center (BIDMC, Boston, MA), Harvard Medical School since 2000 (Gong et al. 2005). Common known ARDS risk factors are listed in Table S1. Inclusion and exclusion criteria were described in the supplementary methods and illustrated in Figure S1. All patients enrolled in this cohort immediately after meeting all inclusion criteria. All ARDS patients met with Berlin definition (Force et al. 2012). The institutional review boards of the MGH, BIDMC, and Harvard T.H. Chan School of Public Health approved this study.

We applied a two-phase study with a total of 530 participants, including 199 ARDS and 330 at-risk controls. One patient without information was excluded (Figure 1). The discovery
population included 78 ARDS patients (cases) and 78 at-risk patients (controls), matched by age (±5 years) and sex. Twenty two miRNAs were selected and used for building an ARDS risk factor panel that included one independent validation cohort containing 121 ARDS and 252 at-risk controls.

Figure 2.1 Study design of discovery cohort, validation cohort, and gene set enrichment analysis (GSEA).

*1 sample (at-risk control) excluded due to few detectable miRNAs. ** Discovery cohort ARDS and at-risk controls were matched on age (±5 years) and genders.
RNA isolation
Peripheral whole blood from 530 participants was collected in Tri Reagent solution (Molecular Research Center, Cincinnati, OH) within 24 hours of participant enrollment and stored at −80°C. Tri Reagent is a robust miRNA stabilization method for long-term storage and can generate reproducible results without degradation (Mraz et al. 2009). Total RNA containing small RNA was extracted from whole blood. RNA quality was assessed by Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA), and the RNA integrity numbers (RIN) were reported. Total RNA with RIN between 6.5 and 10 was processed for complementary DNA synthesis using TaqMan Megaplex RT primer pools A or B then amplified with Megaplex PreAmp Primers. One sample in validation cohort (as-risk control) was excluded due to few detectable miRNAs.

miRNA profiling
The discovery phase detected the expression of 754 human miRNA transcripts from TaqMan OpenArray Human MicroRNA Panel (Applied Biosystems, Foster City, CA), according to manufacturer’s instructions. After quality control screening, we selected 294 miRNAs for this discovery cohort data analysis. Twenty two miRNAs were selected from the discovery cohort for further validation in an independent validation cohort. In the validation cohort, miRNA expression arrays were customized using OpenArray QuantStudio system. Technical consistency of assays within and across different cohorts was also tested using samples from three donors.

Statistical analysis
We performed both univariate and multivariate logistic regression to identify miRNA candidates that are associated with ARDS status. The odds ratios (OR) and 95% confidence intervals (CI) were calculated. Discovery miRNA candidates’ selection was based on the following criteria: fold change > 1.5 or < 0.67 from the logistic regression model in at least one of the four
normalization methods (Lin et al. 2015). Gene set enrichment analysis (GSEA) was used to investigate if the candidate miRNAs were significantly enriched in the whole miRNA gene set (Subramanian et al. 2005).

To build the miRNA risk factor model, we used miRNAs that were validated in all cohorts with the same effect directions, and computed sensitivity, specificity, accuracy, and area-under-the-receiver operating characteristic-curve (AUC) to assess performance of the risk factors. Further, to compare the miRNA risk factors, we used a lung injury prediction score (LIPS) (Gajic et al. 2011) for ARDS risk factor assessment as the base model. Finally, based on miRNA results, we selected miR-181a, miR-92a, and miR-424 to build a miRNA classifier for ARDS risk factor evaluation with combination with LIPS model.

A value of p<0.05 or false discovery rate (FDR) q<0.05 was considered significant. All analyses were performed with R software (v.3.3.0) and Statistical Analysis System software (v.9.3, SAS Institute).

**Results**

Demographic features and clinical variables of each cohort are presented in Table S2. In all experiments, we distributed samples such that age, sex, case-control status, and RNA quality were balanced with respect to day of purification and day of analysis or plate number and randomized within each day and plate. This aspect is important to reduce confounding effect from technical variation such as plate-to-plate variation and variation due to purification (Rieu and Powers 2009).

**miRNA screening and validation**

In total, 754 miRNA transcripts were determined from the discovery cohort, among which 294 miRNA transcripts passed quality controls and were included in statistical analysis. Twenty-two
miRNAs were selected based on fold change as candidate risk factors of ARDS from the logistic regression model (Table 1.1). Of the 22 miRNAs, 3 miRNAs (miR-181a, miR-92a and miR-424) remained significant risk factors (OR>1.0) or protecting factors (OR<1.0) of ARDS from validation cohort after adjusting multiple testing by FDR (Benjamini–Hochberg) (Table 1). From the fixed effect meta-analysis, 14 miRNAs appeared significant risk/protecting factors of ARDS. Of the 14 miRNAs, miR-181a, miR-92a and miR-424 are associated with the ARDS most (p<0.001) (Table 1.1).

miR-181a, miR-92a, and miR-424 are three most significant miRNAs among all 22 miRNAs from the discovery cohort and remained significant in validation cohort and meta-analysis. Indeed, under null hypothesis GSEA (Subramanian et al. 2005), miR-181a, miR-92a, miR-424 were found to be significantly overrepresented and enriched among top 6 genes in a global miRNA scale (Table 1.2/Figure S4).
Table 2.1 miRNA associations with ARDS in discovery cohort, validation cohorts, combined cohorts, and meta-analysis.

Discovery screening based on OR>1.5 or OR<0.67. Meta-analysis was conducted based on fixed-effect model.
Abbreviations: OR, odds ratio; CI, confidence interval; FDR, false discovery rate
### Table 2.2 Gene set enrichment analysis of 22 candidate miRNAs in whole miRNA set.

17 of them found to be significantly overrepresented (FDR q < 0.001) in ARDS vs at-risk control. miR-181a, miR-92a, and miR-424 are among the top enrich score miRNAs.

<table>
<thead>
<tr>
<th>MicroRNA</th>
<th>Rank in gene list</th>
<th>Enrichment score</th>
<th>Core Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-424</td>
<td>0</td>
<td>0.654</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-181a</td>
<td>1</td>
<td>0.560</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-1291</td>
<td>2</td>
<td>0.545</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-744*</td>
<td>3</td>
<td>0.545</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-331</td>
<td>5</td>
<td>0.513</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-92a</td>
<td>6</td>
<td>0.480</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-1244</td>
<td>7</td>
<td>0.446</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-486-3p</td>
<td>8</td>
<td>0.383</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-493</td>
<td>11</td>
<td>0.328</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-204</td>
<td>13</td>
<td>0.323</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-34a</td>
<td>18</td>
<td>0.298</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-642</td>
<td>21</td>
<td>0.282</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-29b</td>
<td>22</td>
<td>0.281</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-483-5p</td>
<td>24</td>
<td>0.278</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-340</td>
<td>28</td>
<td>0.259</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-148a</td>
<td>29</td>
<td>0.257</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-590-3P</td>
<td>33</td>
<td>0.253</td>
<td>Yes</td>
</tr>
<tr>
<td>miR-1290</td>
<td>47</td>
<td>0.211</td>
<td>No</td>
</tr>
<tr>
<td>miR-21</td>
<td>52</td>
<td>0.203</td>
<td>No</td>
</tr>
<tr>
<td>miR-579</td>
<td>76</td>
<td>0.151</td>
<td>No</td>
</tr>
<tr>
<td>miR-20a</td>
<td>86</td>
<td>0.138</td>
<td>No</td>
</tr>
<tr>
<td>miR-155</td>
<td>211</td>
<td>0.051</td>
<td>No</td>
</tr>
</tbody>
</table>

Sepsis and pneumonia stratification analysis

Sepsis and pneumonia are the two most common ARDS-predisposing clinical risks and the highest percentage risks in our study cohort, thus stratification of them can help to validate biomarkers in different risk aspects (Ware and Calfee 2016). According to the risk factor assessment from the validation cohort in Table 3, we selected the 6 miRNAs that showed significant associations with ARDS and performed stratification analysis according to those with sepsis or pneumonia. Among patients with sepsis or those with pneumonia, miR-424, miR-1290, and miR-29a remained significant ARDS protecting factors, and miR-92a, miR-181a, and miR-331 remained significant risk factors of ARDS (Table 1.3).
<table>
<thead>
<tr>
<th>Sepsis</th>
<th>miR-424</th>
<th>miR-92a</th>
<th>miR-181a</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>OR (95% CI)</td>
<td>p</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>At-risk control</td>
<td>197</td>
<td>Ref.</td>
<td>Ref.</td>
</tr>
<tr>
<td>ARDS</td>
<td>112</td>
<td>0.78 (0.66-0.94)</td>
<td>0.007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>miR-1290</th>
<th>miR-92b</th>
<th>miR-331</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>OR (95% CI)</td>
<td>p</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>At-risk control</td>
<td>197</td>
<td>Ref.</td>
<td>Ref.</td>
</tr>
<tr>
<td>ARDS</td>
<td>112</td>
<td>0.84 (0.72-0.98)</td>
<td>0.026</td>
</tr>
</tbody>
</table>

**Table 2.3 Stratification analysis of miRNA associations with ARDS in validation cohorts**

Stratify on sepsis or pneumonia only. All models were adjusted for age and gender.
Abbreviations: OR, odds ratio; CI, confidence interval.

**miRNA diagnostic performance**

Receiver operating characteristic (ROC) analysis demonstrated that LIPS model (AUC: 0.708, 95% CI: 0.651-0.766), sepsis (AUC: 0.572, 95% CI: 0.573-0.607) and pneumonia (AUC: 0.695, 95% CI: 0.651-0.740) diagnosed ARDS when patients from validation cohort were considered.

AUC values of most of the six miRNAs (miR-181a, miR-92a and miR-424) were larger than that of sepsis, but smaller than LIPS and pneumonia. Importantly, the specificity accuracy of the three miRNAs were all larger than those from LIPS, sepsis or pneumonia (Table S4), suggesting that these miRNAs have better performance in correctly classifying at-risk controls. Addition of any one of the three miRNAs increased significantly the baseline LIPS AUC, sensitivity, specificity, and accuracy (Table 4). When three risk and protective factors miRNAs (miR-181a, miR-92a, and miR-424) (Table 4) were computed together with the baseline, AUC increased significantly to 0.723 (95% CI: 0.667-0.778, p=0.005). Computation of all six miRNA (miR-
181a, miR-92a, miR-424, miR-1290, miR-29b, and miR-331) together with the baseline increased further the AUC to 0.728 (95% CI: 0.674-0.783, p=0.001) (Table 1.4). These observations suggest that determination of three miRNAs miR-181a, miR-92a, and miR-424 from the whole blood greatly increased the risk evaluation of ARDS of this population, including their AUC, sensitivity, specificity and accuracy.

<table>
<thead>
<tr>
<th>Combined cohort (N=373)</th>
<th>ARDS vs. at-risk controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC (95% CI)</td>
</tr>
<tr>
<td>LIPS model*</td>
<td>0.708 (0.651-0.766)</td>
</tr>
<tr>
<td>miR-181a + LIPS</td>
<td>0.719 (0.661-0.776)</td>
</tr>
<tr>
<td>miR-92a + LIPS</td>
<td>0.716 (0.659-0.773)</td>
</tr>
<tr>
<td>miR-424 + LIPS</td>
<td>0.715 (0.659-0.771)</td>
</tr>
<tr>
<td>Extended model 1**</td>
<td>0.723 (0.667-0.778)</td>
</tr>
<tr>
<td>Extended model 2***</td>
<td>0.728 (0.674-0.783)</td>
</tr>
</tbody>
</table>

Table 2.4 Diagnostic performance of sepsis/pneumonia model, and miRNA combined model for ARDS.

Abbreviations: AUC, area under the curve; IDI, integrated discrimination improvement; NRI, net reclassification index.
*LIPS: lung injury prediction score; **Extend model 1: LIPS + miR-181a + miR-92a + miR-424, ***Extend model 2: LIPS + miR-181a + miR-92a + miR-424 + miR-1290 + miR-29b + miR-331.
To assess further the incremental diagnostic power that three miRNAs have when added to baseline LIPS risk model, we computed integrated discrimination improvement (IDI) and net reclassification index (NRI) (Supplement), which offer an intuitive way of quantifying improvement offered by new biomarkers (Pencina et al. 2008). The IDI values for miR-181a, miR92a, and miR-424 were 1.36 (95% CI: 0.14-2.58, p=0.029), 1.22 (95% CI: 0.09-2.35, p=0.034), and 1.43 (95% CI: 0.12-2.74, p=0.033), respectively. Of note, combination of miR-181a, miR-92a, and miR-424 increased IDI and NRI to 2.40 (95% CI: 0.72-4.08, p=0.005) and 27.21% (95% CI: 5.72-48.70, p=0.014) respectively (Table 1.4).

Discussion
ARDS is a life threatening inflammatory disease of the alveoli. There is currently no proven pharmacological treatment, although a mechanical ventilation strategy may improve the symptom and there are over 30 finished and on-going clinical trials targeting ARDS. Its mortality rate remains high (Gong and Thompson 2016, Beitler, Schoenfeld, and Thompson 2014). Therefore, early diagnosis and prevention become essential. The LIPS model has been used to detect potential risk factors of ARDS with appropriate sensitivity using clinical predisposing conditions based on clinical observations, which may not accurately reflect the pathophysiological process (Gajic et al. 2011).

To our knowledge, this current study is the first study using whole blood samples from a large population of ARDS patients and critically ill at-risk control patients from discovery and independent validation design with rigorous statistical analyses of a high-throughput miRNA set. Primary selected 22 miRNAs from the discovery cohort were further validated in an independent validation cohort and subsequent meta-analysis. Such diligent analyses may offer advantage for
miRNA profiling and face a much reduced risk of misrepresentation from miRNA expression noise that typically results from additional serum or plasma isolation steps and sample contamination. Although patient collection spanned 10 years, we carefully selected high quality samples and controlled all experiments by grouping patients into a randomly selected discovery cohort and a validation cohort according to the dates of sample collection. We used identical amounts of RNA input in all experiments to control biased variation due to different RNA input amount. We also distributed samples such that age, sex, case-control status, and RNA quality were balanced with respect to day of purification and day of analysis or plate number and randomized within each day and plate to reduce confounding from technical variation such as plate-to-plate variation and variation due to purification.

miRNAs have been used successfully as biomarkers for chronic diseases such as pancreatic and gastric cancers (Lin et al. 2015, Zhu et al. 2014). This study identified three most promising miRNAs, miR-181a, miR-92a, or miR-424, which associated with human ARDS; and GSEA confirmed that they were significantly overrepresented in the ARDS patients vs at-risk controls, although there is no direct evidence linking these miRNAs to ARDS. Yet multiple recent studies provided indirect evidence for their involvement in the dysregulated signaling pathways of ARDS (Li et al. 2007, Loyer et al. 2014, Kim et al. 2013). Here we reported that miR-181a and miR-92a are associated with the risk of ARDS in all tested cohorts and meta-analysis. These findings are consistent with prior studies in inflammation and endothelial cell injury, which are common in ARDS (Moussa et al. 2015). miR-181a is a key regulator of T cell development and T cell receptor signaling threshold (Li et al. 2007). Increased miR-181a expression in mature T cells augments cell sensitivity to peptide antigens. T cell responses decline with age due to an age-associated defect in T cell receptor signaling, which is caused by the increase expression
phosphatase 6 and miR-181a. miR-92a inhibits endothelial cell (EC) angiogenesis and impairs EC function (Bonauer et al. 2009, Loyer et al. 2014, Wu et al. 2012, Fang and Davies 2012). miR-92a targets Krüppel-like Factor 2 (KLF2), KLF4, and Sirtuin 1, thereby promoting inflammatory responses (Loyer et al. 2014, Wu et al. 2011, Fang and Davies 2012). Lung microvascular endothelium injury-associated pulmonary edema is a hallmark of ARDS (Maniatis and Orfanos 2008, Ware 2006). When miR-92a is overexpressed, the blood vessel growth and functional recovery of damaged tissue is restricted (Bonauer et al. 2009), which may enhance the incidence of pulmonary edema and ARDS.

In contrast, this study reported miR-424 as a protective factor of ARDS (Table1). miR-424 was found down-regulated in pulmonary artery hypertension (PAH) via apelin and fibroblast growth factor 2 signaling in pulmonary artery EC (Kim et al. 2013). PAH is commonly observed in ARDS patients, in whom hypoxemia can promote the pulmonary vasoconstriction that gives rise to PAH. Hypoxia-induced miR-424 played an important role in vascular remodeling and angiogenesis in EC (Ghosh et al. 2010). Low oxygen level affects cells and tissues during wound healing as well as during pathological conditions such as stroke. As a consequence, miR-424 signaling system is activated in EC to stabilize hypoxia-inducible factors, thus overcome hypoxia to restore oxygen (Ghosh et al. 2010). These prior studies support our finding that miR-424 expression may exert a protective effect against ARDS.

Of note, LIPS is currently considered as a standard clinical diagnostic model and associate with the risk and complications of ARDS (Gajic et al. 2011). All three miRNAs selected from our two cohorts had the similar specificity and accuracy in diagnosing ARDS to that of LIPS (Table S4). Incorporation of these miRNA or the three most significant miRNAs, miR-181a, miR-92a, and miR-424, to LIPS further increased the potency and accuracy in diagnosis of ARDS (Table 4).
Therefore, miRNAs identified from this study may have values to that of LIPS for future ARDS risk evaluation and clinical diagnosis.

This study focused mainly on 3 miRNAs selected from the validation cohort. It does not mean that the remaining 19 miRNAs selected from the discovery cohort are irrelevant (Table 1). Some of these miRNAs have also been implicated in inflammatory signaling pathways and may also be ARDS candidate risk factors. For example, miR-155 and miR-21 are functionally related and contributed to NF-κB signaling (Ma et al. 2011), an important pathway for innate and adaptive immunity and inflammation. Further investigation is needed to confirm the involvement of these miRNAs in ARDS, which might provide a better understanding of the mechanisms underlying ARDS.

Yet, we also acknowledge limitations in our study. First of all, the diagnostic power of our miRNAs might not intense. However, unlike other similar studies (Schultz et al. 2014, Zhu et al. 2014), our control subjects were at-risk patients, providing less deviation from cases. It means that use of at-risk patients as controls in our study may have reduced confounding noise compared to that from the use of completely healthy controls. ARDS is considered as a complicated syndrome with multiple etiologies, single or few miRNAs might not show strong signals for all ARDS patients. This concept was recently confirmed in ARDS randomized clinical trial study, which concluded no beneficial effect of aspirin for ARDS prevention (Kor et al. 2016). Also, our study was based on a single geographic region; a geographically different external cohort in a similar study setting will be helpful to further validate our findings.
Conclusions
This study links the expression of miR-181a, miR-92a, and miR-424 in whole blood to ARDS. Inflammatory response markers miR-181a and miR-92a were significantly elevated in ARDS patients, while the pulmonary artery EC anti-inflammation marker miR-424 was significantly reduced in ARDS patients. Further, the expression patterns of our miRNA biomarkers may provide an in depth molecular understanding between ARDS and at-risk patients on top of sepsis and pneumonia. By combining our miRNA biomarkers with LIPS, diagnosis of ARDS may be improved. Circulation levels of these miRNAs may have the potential to guide ARDS treatments by targeting these miRNAs.

Authors’ contributions: Z.Z., D.C.C, L.L., A.B., Q.L. designed the study. D.C.C, E.B., B.T., L.S. established the MEARDS cohort and collected the samples and clinical information. Z.Z. and L.S. performed the experiment and data collection. Z.Z., L.L., R.Z., W.Y., X.Z. and Y.G. performed data analysis. Z.Z., G.S, X.Z. drafted the manuscript. All authors reviewed and edited the final paper. D.C.C had full access to all of the data in the study and take responsibility for the integrity of the data.

Funding: This study was supported by grants R01 HL060710 (DCC), R56HL134356 (DCC), P30 ES000002 (DCC) and HL123568 (GPS) from National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health.

Acknowledgements: We thank Dr. Xinghui Sun from Brigham and Women's Hospital and Harvard Medical School for his critical reading of this manuscript, Andrea Shafer and Sean Levy from Massachusetts General Hospital and Beth Israel Deaconess Medical Center for their help in searching missing clinical data and Dr. Boyu Ren from Harvard T.H. Chan School of Public Health for his assistance in data analysis. We also thank Ms. Sheila Cherry from Fresh Eyes Editing for her editorial assistance.

Ethical standards statement: The institutional review boards of the MGH, BIDMC, and Harvard T.H. Chan School of Public Health approved this study.

Competing interests: The authors declare no competing interests.

Consent to publish: We confirm that we have obtained consent to publish from the participant (or legal parent or guardian for children) to report individual patient data.
References


Chapter 3 : Whole Blood MicroRNAs as a Prognostic Classifier for Acute Respiratory Distress Syndrome 28-day Mortality

Zhaozhong Zhu¹, Ruyang Zhang¹,², Liming Liang³,⁴, Li Su¹, Quan Lu¹, Andrea A. Baccarelli¹, Ednan K. Bajwa⁵, B. Taylor Thompson⁵, David C. Christiani¹,⁵

¹ Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA
² Department of Environmental Health, Department of Epidemiology and Biostatistics, Ministry of Education Key Laboratory for Modern Toxicology, School of Public Health, Nanjing Medical University, Nanjing, China
³ Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
⁴ Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA
⁵ Pulmonary and Critical Care Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA

Corresponding author: David C. Christiani, Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA. (Email: dchris@hsph.harvard.edu; Tel: 617-432-1641; Fax: 617-726-9274)

Authors’ contributions: Z.Z., D.C.C, L.L., A.B., Q.L. designed the study. D.C.C, E.B., B.T., L.S. established the MEARDS cohort and collected the sample. Z.Z. and L.S. performed the experiment and data collection. Z.Z. and R.Z. did the data analysis and interpretation. Z.Z. drafted the manuscript.

Funding: This study was supported by grants R01 HL060710 and P30 ES000002 (DCC) from National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health.

Acknowledgements: The authors thank Andrea Shafer and Sean Levy from Massachusetts General Hospital and Beth Israel Deaconess Medical Center for assistance in clinical data retrieval missing clinical data.

Competing interests: The authors declare no competing interests.

Ethical standards statement: The institutional review boards of the MGH, BIDMC, and Harvard T.H. Chan School of Public Health approved this study.

Word counts: 497
Introduction
The acute respiratory distress syndrome (ARDS) is the leading cause of respiratory-related death disease in both intensive care unit (ICU) and hospital-wide, with a mortality rate of up to 40% (Bellani et al. 2016). Despite a decreasing mortality rate of ARDS over time due to improved management (Villar et al. 2011), this syndrome is under-diagnosed and insufficiently treated, and as a result remains highly deadly (Bellani et al. 2016).

MicroRNAs (miRNAs) are small non-coding RNAs, usually ~22 nucleotides in length. They regulate gene expression by binding to specific target sites on messenger RNAs to either repress the translation of or degrade the transcript. MiRNAs play important role in inflammation and infection (O'Connell, Rao, and Baltimore 2012), both of which are common manifestations in ARDS (Sheu et al. 2010). In addition, miRNAs are used to construct prognostic classifier for early prediction of disease outcomes, including cancer (Volinia and Croce 2013).

Method
Here we report a survival analysis as part of the Molecular Epidemiology Study of ARDS (MEARDS) from the ICU at Massachusetts General Hospital and Beth Israel Deaconess Medical Center. We collected 78 whole blood RNA samples from MEARDS. Expression of 754 human miRNAs identified by TaqMan OpenArray Human MicroRNA Panel was measured. After quality control screening (supplement), we selected 294 miRNAs for data analysis. Imputation was used to handle missing miRNA data (supplement). We used multi-variate Cox proportional regression analysis to estimate the hazards ratio (HR) of miRNA for ARDS 28-day mortality. The Kaplan-Meier and log-rank method was performed to test the equality for survival distributions in different groups. All analyses were performed with R software (version 3.2.3) and Statistical Analysis System software (v.9.4, SAS Institute).
RNA isolation and complementary DNA (cDNA) synthesis
We collected peripheral blood from 78 participants on day 1 of participant enrollment in Tri Reagent solution (Molecular Research Center, Cincinnati, OH) and stored at −80°C. Tri Reagent is a robust miRNA stabilization method for long-term storage and can generate reproducible results without degradation (Mraz et al. 2009). Total RNA containing small RNA was extracted from whole blood. Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA) was used to assess extracted total RNA and the RNA integrity numbers (RIN) were reported. Total RNA with RIN between 6.5 and 10 was processed for cDNA synthesis using TaqMan Megaplex RT primer pools A or B then amplified with Megaplex PreAmp Primers (pools A or B). For each sample, we used 200 ng total RNA containing the small RNA fraction for reverse transcription with Megaplex™. The small RNA fraction was not enriched to avoid loss of longer control transcripts (snoRNAs).

miRNA profiling
TaqMan OpenArray Human MiRNA Panel (Applied Biosystem,s, Foster City, CA) was used to profile 754 miRNAs. Each microRNA plate contained 3 TaqMan® MicroRNA assay endogenous controls (RNU44, RNU48, U6) and 1 exogenous control (ath-miR-159a) to aid in data normalization. We used average of RNU44 and RNU48 expression as endogenous normalization since it provided the most stable expression level. The cDNA was run through 40 cycles of 16 °C for 2 minutes, 42 °C for 1 minute, and 50 °C for 1 second. Reactions were then held at 85 °C for 5 minutes and cooled to 4 °C for storage.

miRNA data quality control
For all quality control procedures, we used the following criteria to identify reliable miRNAs: amplification score >1.1, Cq confidence >0.8, high expression (Ct<30), and missing percent less than half of samples (Thermo Fisher Co. 2015). Quality control screening of 754 miRNAs left
294 miRNAs for statistical analysis, due to exclusion of poor-quality and low-expression microRNAs.

**Imputation**
We assumed that the Ct values were not missing at random but missing due to low expression of microRNA in the sample. Therefore, we imputed missing data using the 95th percentile of the same microRNA for cases or controls (Schultz et al. 2014).

**Results**
Demographic characteristics can be found in the supplement (Table s1). We identified nineteen miRNAs potentially associated with ARDS survival in patients with moderate to severe ARDS (Table s2). Among them, five miRNAs were most differentially expressed, miR-628.3p (HR=1.70, \(p<0.01\)), miR-922 (HR=1.05, \(p<0.01\)), miR-505* (HR=1.65, \(p<0.01\)), miR-130b* (HR=1.44, \(p<0.01\)), miR-624 (HR=1.38, \(p<0.01\)). In addition, based on all statistical significant miRNAs, we used backwards elimination methods with Akaike information criterion to select miRNAs that have potential to predict ARDS 28-day mortality (miR-628.3p, miR-922, miR-766, miR-194 and miR-7). The final miRNA classifier was obtained by both most differential expression and backwards elimination. Expression of miRNA classifier larger than median was assigned as high expression, lower than median was assigned as low expression. The Kaplan–Meier curves for 28-day mortality groups, using the eight-miRNA classifier, are shown in Figure1. Time to death is shorter in patients with higher eight-miRNA classifier expression (\(p=0.04\)), which is comparable to APACHE III (supplement).

**Conclusion**
To our knowledge, this is the first study of miRNA as a prognostic classifier from whole blood for ARDS 28-day mortality. Whole blood contains both immune cell- and tissue-specific miRNAs and thus offers a major advantage for miRNA profiling compared with other tissue
types. While our study confidence is limited by sample size and the mortality rate in this small cohort is high and may not be representative of general ARDS cohort, the classifier containing miRNAs discovered in this study offers a potentially valuable, novel biomarker signature in the clinical practice to better ARDS 28-day mortality prognosis.

Figure 3.1 Kaplan–Meier survival analysis according to the miRNA classifier grouped by expression level.
References

Bellani, G., J. G. Laffey, T. Pham, E. Fan, L. Brochard, A. Esteban, L. Gattinoni, F. van Haren, A.
Larsson, D. F. McAuley, M. Ranieri, G. Rubenfeld, B. T. Thompson, H. Wrigge, A. S. Slutsky, A.
Pesenti, Lung Safe Investigators, and Esicm Trials Group. 2016. "Epidemiology, Patterns of Care,
and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in


Schultz, N. A., C. Dehlendorff, B. V. Jensen, J. K. Bjerregaard, K. R. Nielsen, S. E. Bojesen, D.

"The influence of infection sites on development and mortality of ARDS." Intensive Care Med


Villar, J., J. Blanco, J. M. Anon, A. Santos-Bouza, L. Blanch, A. Ambros, F. Gandia, D. Carriedo, F.
ALIEN study: incidence and outcome of acute respiratory distress syndrome in the era of lung

of patients with invasive breast cancer." Proc Natl Acad Sci U S A 110 (18):7413-7. doi:
10.1073/pnas.1304977110.
Chapter 4 : Shared Genetic Architecture of Asthma with Allergic diseases: A Genome-wide Cross Trait Analysis of 112,000 Individuals from UK Biobank

Zhaozhong Zhu\textsuperscript{1,2,3}, Phil H. Lee\textsuperscript{1,3,4,5}, Wonil Chung\textsuperscript{6}, Mark Chaffin\textsuperscript{7}, Po-Ru Loh\textsuperscript{6,7}, Alkes Price\textsuperscript{6,7,8}, Quan Lu\textsuperscript{2}, David C. Christiani\textsuperscript{2}, Liming Liang\textsuperscript{7,8}

\textsuperscript{1} Center for Human Genetics Research, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.
\textsuperscript{2} Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
\textsuperscript{3} Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.
\textsuperscript{4} Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
\textsuperscript{5} Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
\textsuperscript{6} Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
\textsuperscript{7} Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
\textsuperscript{8} Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.

Corresponding author: Liming Liang, Department of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. (Email: liang@hsph.harvard.edu; Tel: 617-432-5896)

Authors’ contributions: ZZ, LL, PHL, QL, AP designed the study.
ZZ, MC, WC, LL, PRL performed the statistical analysis.
ZZ, LL, MC, QL wrote the manuscript.
All authors helped interpret the data, reviewed and edited the final paper, and approved the submission.
ZZ and LL had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Competing interests: The authors declare no competing interests.
Abstract

Introduction: Clinical and epidemiological data suggest that asthma and allergic diseases, such as allergic rhinitis and eczema, are associated. Several studies have demonstrated that the coexistence of asthma and allergic diseases is frequent. One hypothesis to account for the relationship is that these chronic diseases share a common genetic etiology. Heritability has been estimated at high level for asthma and allergic diseases, which suggests that the genetic contribution to them can be significant.

Methods: We analyzed genome-wide single-nucleotide polymorphism (SNP) data for the asthma and allergic diseases in 35,783 cases and 76,768 controls of European ancestry from UKBiobank. We used a multinomial logistic regression to find the association SNPs with each disease. And we used the linkage disequilibrium score regression to estimate heritability and genetic correlation between each trait from their summary statistics. Cross trait meta-analysis models were applied to find the shared genetic components between correlated traits. And finally we used GETx tissue enrichment analysis to find the tissue enrichment of shared GWAS loci. We have further included two public available independent genome wide association study (GWAS) data, asthma GWAS and eczema GWAS as replication phase to confirm our finding.

Results: We identified 78 genome-wide significant ($P < 5 \times 10^{-8}$) loci for asthma and 54 for allergic diseases. We have found a strong genetic correlation between asthma and allergic diseases ($R_g=0.75$, $P=6.84 \times 10^{-62}$). Eight SNPs were found to be exactly overlap between asthma and allergic diseases (rs34290285 on D2HGDH and GAL3ST2, rs4833097, rs7705653 on SLC25A46 and TMEM232, rs1837253, rs12413578, rs7936070 on C11orf30, rs56062135 on SMAD3 and rs10414065), 24 loci contains same protein-coding genes between asthma and allergic diseases even though not at the same exact SNP. Cross trait analysis and identified 90
loci, which could be taken to suggest that it is this shared genetic loci between two diseases driving the overall significant positive correlation. GTEx tissue enrichment analysis showed shared genetic loci between asthma and allergic diseases were enriched in skin tissue. Seventeen loci demonstrated genome-wide significance. Of these, genes IL13, C11orf30, IL33, IL4, IL1RL1, IL18R1, IL18RAP, MIR4772, SLC9A4 and GSDMA were replicated in two independent GWAS data.

Conclusion: Our GWAS study has highlighted shared genetic pathway in immune and inflammatory systems and skin tissue by asthma and allergic diseases. This work should support the idea that common patterns of association between asthma and allergy implicate shared biological processes and advance understanding of the molecular mechanisms underlying co-morbid asthma and allergic diseases.
Introduction
Asthma is a chronic respiratory syndrome that is characterized by abnormal and inflamed mucosa of the airways, wheezing, and shortness of breath. Allergic diseases are immune responses for allergies, such as allergic rhinitis and atopic dermatitis (eczema). Asthma, allergic rhinitis and eczema all belong to type I hypersensitivity, which is an immune response to foreign antigen and associated with immunoglobulin E (IgE)-mediated inflammation (Lenz 2007, Wallace et al. 2008). Genetic studies offer a structured means of understanding the causes of asthma and allergic diseases, as well as identifying targets that can be used to treat the syndrome (Torgerson et al. 2011, Moffatt et al. 2010, Bunyavanich et al. 2014, Ramasamy et al. 2011, Pillai et al. 2009).

Clinical and epidemiological data suggest that asthma and allergy are associated (Leynaert et al. 2000, Brauer et al. 2002). Several studies have identified allergic diseases, such as allergic rhinitis and eczema, as a risk factor for asthma, with the prevalence of allergic rhinitis in asthmatic patients being 80% to 90% (Leynaert et al. 2000, Pariente et al. 1997). These studies and others demonstrate that the coexistence of asthma and allergy is frequent, that allergy usually precedes asthma. Also our previous epigenetic study has identified methylation loci linked to asthma and allergy via IgE pathway (Liang et al. 2015).

One hypothesis to account for the similar symptoms and conditions is that these diseases share a common genetic etiology. Cotsapas et al discovered nearly half of loci in genome-wide association studies (GWAS) of an individual disease influence risk to at least two diseases, indicating the shared genetic architecture of similar diseases (Cotsapas et al. 2011). As each of the shared or similar risk factors has strong genetic influences on disease risk, the observed
clustering of multiple risk factors could be due to an overlap in the causal genes and pathways (Criswell et al. 2005, Cross-Disorder Group of the Psychiatric Genomics 2013, Global Lipids Genetics et al. 2013, Pickrell et al. 2016, Shi, Kichaev, and Pasaniuc 2016, Lane et al. 2016, Hobbs et al. 2017, Emdin et al. 2017). In addition, grouping variants by the traits they influence should provide insight into the specific biological processes underlying co-morbidity and disease risk. Clinical and epidemiological studies have found asthma and allergic diseases can occur either in the same individual or in closely related family members(Ober and Yao 2011, Cookson 2004, Belsky et al. 2013, Holgate 1999), suggesting potential pleiotropic effect. The heritability has been estimated at varying between 35% and 95% for asthma(Ober and Yao 2011, Duffy et al. 1990) and 33% and 91% for allergic rhinitis, 71% and 84% for atopic dermatitis (eczema), 34% and 84% for serum IgE levels, which suggests that the genetic contribution to them can be significant(Ober and Yao 2011). Hinds and colleagues found the shared genetic etiology for 38 allergic diseases(Hinds et al. 2013). However, understanding has been limited only 16 genome-wide shared susceptibility loci which were based on self-reported phenotype.

Thus, in order to increase our knowledge of shared genetic determinants influencing physician diagnosed allergic diseases and asthma, and potentially discover novel loci, we proposed to investigate the genetic commonality in two diseases, asthma and allergic diseases (hay fever/allergic rhinitis or eczema). We conducted a large scale GWAS analysis based on these traits to explore genetic correlations and shared genetic components among these traits using UK Biobank data, which is the largest and most complete European biobank available at present. Also we applied two independent public available GWAS studies, the GABRIEL asthma GWAS study(Moffatt et al. 2010) and the EArly Genetics & Lifecourse Epidemiology (EAGLE) eczema consortium study(Paternoster et al. 2015) to replicate our findings.
Methods

Study population and design.
Study participants were from the UK Biobank study, described in detail elsewhere (Collins 2012, Sudlow et al. 2015, Allen et al. 2014). In brief, the UK Biobank is a prospective study of >500,000 people living in the UK. All people in the National Health Service registry who were aged 40–69 years and living <25 miles from a study center were invited to participate from 2006–2010. In total, 503,325 participants were recruited from over 9.2 million mailed invitations. Self-reported baseline data were collected by questionnaire, and anthropometric assessments were performed. For the current analysis, individuals of non-white ethnicity were excluded to avoid confounding effects. All participants provided informed consent to the UK Biobank.

To identify genetic variants that contribute to doctor-diagnosed asthma and allergic diseases and link them with other conditions, we performed GWAS using phenotype measures in UK Biobank participants of European ancestry (n=152,249). We have removed 39,698 samples that are due to non-European, related, withdraw from UK Biobank or missing phenotype information. Thus, total of 112,551 European descents with high-quality genotyping and complete phenotype/covariate data were used for these analyses, including 25,685 allergic diseases (hay fever, allergic rhinitis or eczema), 14,085 asthma and 76,768 controls for the analysis. There were 53,094 males and 59,457 females in the population.

Furthermore, we have included an independent asthma GWAS (GABRIEL study, 10,365 cases 16,110 controls) (Moffatt et al. 2010) and an independent eczema GWAS (EAGLE eczema consortium, 13,287 cases and 41,345 controls) (Paternoster et al. 2015) summary statistics for replication analysis (Figure 3.1).
Figure 4.1 Study population and design
Data summary, quality control (QC) and imputation
Detailed genotyping and quality control procedures of UK Biobank were described previously (http://biobank.ctsu.ox.ac.uk/) and in supplementary material. In summary, after QC procedures were applied, the current released UK Biobank data contains 806,466 SNPs that passed SNP QC in at least one batch. Post-imputation QC was performed as previously outlined (http://biobank.ctsu.ox.ac.uk/); and genotype imputation procedure provided a total of 73,355,677 SNPs. In addition, we performed stringent quality-control standards by PLINK 1.90. (Purcell et al. 2007) We selected variants that did not deviate from Hardy-Weinberg Equilibrium (HWE) (P>1×10^{-12}), per variant missing call rates<5%, per-sample missing rate<40%, minor allele frequency (MAF) >1% and an imputation quality score (INFO) >0.8. Quantile-Quantile (QQ) plots were produced and checked for each phenotype. R package GenABEL was used to compute genomic inflation values. A total of 7,489,529 SNPs passed QC on whole genome and with complete information in all three phenotypes, which were eligible for statistical association analyses.

GWAS analysis
We performed multinomial additive logistic regression genetic association analysis adjusting for age, sex, genotyping array and ten principal components ancestry using PLINK 1.90(Purcell et al. 2007) to assess association between phenotype and genotype on each individual study. After association analysis, we have applied conditional analyses to assess evidence for multirisk loci in a region using PLINK clumping function (parameters: --clump-p1 5e-8 --clump-p2 1e-5 --clump-r2 0.2 --clump-kb 500) to find top loci region which has r^2 less than 0.2 and within 500 kb away from a peak variant, and are assigned to that peak variant's clump with association p-value smaller than 1×10^{-5}. We have used GWAS catalog (https://www.ebi.ac.uk/gwas) to identify previously reported genes by other GWAS studies.
**LD score regression analysis.**
Post-GWAS genome-wide genetic correlation analysis by LD score regression (LDSC) (Bulik-Sullivan et al. 2015) was conducted using all UK Biobank SNPs after QC also found in 1000 genome (Genomes Project et al. 2015). LDSC estimates genetic correlation between two traits (ranging from −1 to 1) from summary statistics using the facts that the GWAS effect size estimate for each SNP incorporates the effects of all SNPs in linkage disequilibrium with that SNP, SNPs with high linkage disequilibrium have higher χ² statistics than SNPs with low linkage disequilibrium, and a similar relationship is observed when single-study test statistics are replaced with the product of the z scores from two studies of traits with some correlation (Bulik-Sullivan et al. 2015). LDSC applied the self-estimated intercept during the analysis to take care of the shared subjects between studies (Bulik-Sullivan et al. 2015).

**Cross trait meta-analysis**
After assessing genetic correlations among all traits, we have applied cross-trait GWAS meta-analysis by using a R package association analysis based on subsets (ASSET) to combine the association evidence for asthma and allergic diseases (Bhattacharjee et al. 2012). This method combines effect estimate and standard error of the GWAS summary statistics. It can also account the correlation among studies/subjects that might arise due to shared subjects across distinct studies or due to correlation among related traits in the same study by using case-control overlap matrices.

**GTEx tissue expression enrichment analysis**
Gene sets were generated from lists obtained from analyses for asthma, allergic diseases and cross-trait, and by including those which have a matching HGNC name. Genes identified in this way were analyzed for tissue enrichment using publicly available expression data from pilot phase of the Genotype-Tissue Expression project (GTEx) (Consortium 2013), version 4
(www.gtexportal.org). In the project, postmortem samples from a wide variety of tissues and donors have been used for bulk RNA sequencing according to a unified protocol. All samples were sequenced using Illumina 76 base-pair paired-end reads. The pilot phase of the GTEx project was completed in December of 2014, and includes data from 50 human tissues and three cell line preparations.

Collapsed reads per kilobase per million mapped reads (RPKM) values for each of the 52,577 included transcripts, filtered for unique HGNC IDs (n=20,932) were organized by tissue and individual (N\text{tissues} = 53, N\text{samples} = 2,921). All transcripts were ranked by mean RPKM across all samples, and 500,000 permutations of each credible set gene list were generated by selecting a random transcript for each entry in the credible set within +/-100 ranks of the transcript for that gene. For each sample, the RPKM values were converted into ranks for that transcript, and sums of ranks for each within each tissue were computed for each gene. P-values for each tissue were calculated by taking the total number of cases where the gene list of interest had a lower sum of ranks than the permuted sum of ranks, and dividing by the total number of permutations. To assess the significance of the enrichment after testing multiple tissues, we used a Bonferroni correction adjusted for the number of independent tissues, estimated via the matSpD(Nyholt 2004) (www.neurogenetics.qimrberghofer.edu.au/matSpD) tool to arrive at 27 independent tests and a significance threshold of P=1.90×10^{-3}.

**Replication analysis**  
In order to confirm our finding, we have further included two public available independent GWAS data, asthma GWAS, the GABRIEL consortium study (Moffatt et al. 2010) and eczema GWAS, the EArly Genetics & Lifecourse Epidemiology (EAGLE) eczema consortium study (Paternoster et al. 2015), for LD score regression analysis, cross trait analysis and GTEx tissue
expression enrichment analysis. We have used LiftOver tool
(http://genome.sph.umich.edu/wiki/LiftOver) to convert asthma GWAS reference genome from
hg18 to hg19. ImpG-Summary was used to impute asthma GWAS summary statistics to 1000
genome(Pasaniuc et al. 2014). SNPs with imputation quality r2pred<0.8 were removed. Cross
trait analysis between two replication GWAS datasets was conducted using R package Cross-
Phenotype Meta-Analysis (CPMA), which combines the empirical evidence based on their P-
values(Cotsapas et al. 2011).

**Overrepresentation Enrichment Analysis**
In order to understand the biological insights of the shared genes, we have used WebGestalt tool
(Wang et al. 2013) to assess overrepresentation enrichment of the identified shared gene set
between asthma and allergic diseases in KEGG(Kanehisa and Goto 2000) pathways, Gene

**Results**
The phenotype-genotype association test was carried out on all 112,551 samples from UK
Biobank. We identified 78 genome-wide significant (P < 5×10^{-8}) loci for asthma and 54 for
allergic diseases (Fig. 3.2). The λ_{1000} (λ rescaled to a sample of 1000 cases and 1000 controls)
was 1.005 for asthma, 1.003 for allergic diseases and 1.006 for cross trait meta-analysis(Devlin
Figure 4.2 Manhattan plot of cross trait analysis of asthma and allergic diseases

For asthma, we confirmed most of the previously reported association loci, including IL1R1, IL1RL1, IL18R1, IL13, SLC25A46, HLA regions, SMAD3, GSDBM and many others (Supplementary Table 1). Out of the 78 independent loci reported here, 11 represent new associations with asthma, 6 of these new loci were mapped to protein-coding genes (Supplementary Table 1). For allergic diseases, genes such as C11orf30, FLG, SLC25A46, TMEM232, CAMK4, TSLP, WDR36, CLEC16A, DEXI, HLA and many others were confirmed in our analysis (Supplementary Table 2). Out of the 54 independent loci identified for allergic
diseases, 12 represent the new associations, of which 8 were mapped to protein-coding genes (Supplementary Table 2). Eight SNPs were found to be exactly overlap between asthma and allergic diseases (rs34290285 on D2HGDH and GAL3ST2, rs4833097, rs7705653 on SLC25A46 and TMEM232, rs1837253, rs12413578, rs7936070 on C11orf30, rs56062135 on SMAD3 and rs10414065), 24 loci contains same protein-coding genes between them even though not at the same exact SNP (Supplementary Table 3).

Among 8 exactly overlap SNPs, 4 of them were mapped to protein-coding genes. TMEM232 and SLC25A46 (rs7705653) encode transmembrane protein 232 that belongs to tetraspanin family and a solute carrier protein involved in transport of metabolites respectively. C11orf30 (rs7936070) is associated with total serum IgE levels in asthma (Li et al. 2012); also the locus C11orf30 increases susceptibility to poly-sensitization (Amaral et al. 2015). rs56062135 on SMAD3 involves in development of asthmatic inflammation that are due to responses of immune cells, such as cytokins and T-helper 2 (Th2) cells, which are known to be important in generating an inflammation that characterizes asthma and allergic disease. (Anthoni et al. 2007) The fourth SNP is rs34290285 on two novel genes for asthma and allergic diseases, D2HGDH and GAL3ST2, which are responsible for encoding metabolism related D-2hydroxyglutarate dehydrogenase enzyme (Struys et al. 2005) and tumor metastasis related Galactose-3-O-sulfotransferase 2 enzyme (Seko et al. 2002). Among all the protein coding genes that each peak locus mapped on within 500kb, 101 genes are shared by asthma and allergic diseases, 158 genes are specific to asthma and 38 genes are specific to allergic diseases (Supplementary Table 4).

The genetic correlation was calculated using common SNPs. There was strong positive genetic correlation between asthma and allergic diseases ($R_g=0.75$, $P=6.84\times10^{-62}$). We have also extended our LDSC analysis to other 1 type of immune complex mediated (Type III),
rheumatoid arthritis (Okada et al. 2014), and 2 types of delayed cell mediated (Type IV) immune
diseases, Crohn's disease and ulcerative colitis (Jostins et al. 2012), however we didn’t find
significant genetic correlation between asthma and other immune traits, but we confirmed the
high genetic correlation between Crohn's disease and ulcerative colitis (Table 1). Thus this
empirical evidence of shared genetic etiology for asthma and allergic diseases can encourage the
investigation of common pathophysiology for them. Estimates of the SNP-based heritabilities by
using GWAS summary statistics are presented in Supplementary Table 5. The proportion of
variance in the traits explained by common genetic variants is 7.2% (s.e. 0.7%) for asthma and
7.5% (s.e. 0.7%) for allergic diseases.

In our replication analysis, we have identified similar level of genetic correlation between asthma
and eczema by using two independent GWAS data: the GABRIEL GWAS study (Moffatt et al.
2010) and EAGLE eczema study (Paternoster et al. 2015). LD score regression results showed
there is strong positive genetic correlation between asthma and eczema (Rg=0.46, P=0.0044)
(Table 3.1 and Supplementary Table 5).
Furthermore, we have strengthened this finding in cross trait analysis and identified 90 loci contain SNPs with $P<5\times10^{-8}$ (Supplementary Table 6). Figure 1 shows the Manhattan plot of the cross trait analysis between asthma and allergic diseases. 51 loci were found to overlap with asthma single trait analysis, 35 loci were found to overlap with allergic diseases. Eight exactly overlap SNPs between asthma and allergic diseases were also identified through cross trait analysis. The strongest association signal was observed on chromosome 6 at HLA-DQ region (rs9273374, $P_{\text{meta}}=7.87\times10^{-35}$)(Supplementary Table 6), which was the first asthma-susceptibility locus to be identified (Marsh, Meyers, and Bias 1981), and extended haplotypes encompassing HLA-DQ and HLA-DR have been studied for their effects on specific allergen sensitization (Moffatt et al. 2001). The second strongest signal was an overlap SNP (rs7936070, $P_{\text{meta}}=2.81\times10^{-28}$) between asthma and allergic diseases mapped on C11orf30 on chromosome 11q13 (Supplementary Table 6), where it is an overlap gene associated with total serum IgE levels and increases susceptibility to poly-sensitization. The third strongest signal was observed

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Asthma</th>
<th>Allergic diseases</th>
<th>Rheumatoid arthritis</th>
<th>Crohn’s disease</th>
<th>Ulcerative colitis</th>
<th>Asthma (Moffat et al)</th>
<th>Ezcema (Paternoster et al)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>1</td>
<td>0.7455</td>
<td>0.0942</td>
<td>0.0428</td>
<td>-0.0514</td>
<td>0.6307</td>
<td>0.4138</td>
</tr>
<tr>
<td>Allergic diseases</td>
<td>6.84E-62</td>
<td>1</td>
<td>0.0042</td>
<td>0.1415</td>
<td>0.0611</td>
<td>0.4738</td>
<td>0.5308</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>0.0793</td>
<td>0.9259</td>
<td>1</td>
<td>0.067</td>
<td>0.1003</td>
<td>0.0369</td>
<td>-0.0463</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>0.5925</td>
<td>0.0371</td>
<td>0.318</td>
<td>1</td>
<td>0.5711</td>
<td>-0.007</td>
<td>0.119</td>
</tr>
<tr>
<td>ulcerative colitis</td>
<td>0.4942</td>
<td>0.3721</td>
<td>0.1032</td>
<td>7.98E-23</td>
<td>1</td>
<td>-0.0814</td>
<td>-0.0121</td>
</tr>
<tr>
<td>Asthma (Moffat et al)</td>
<td>1.09E-11</td>
<td>1.99E-08</td>
<td>0.6478</td>
<td>0.9477</td>
<td>0.4753</td>
<td>1</td>
<td>0.4611</td>
</tr>
<tr>
<td>Ezcema (Paternoster et al)</td>
<td>1.29E-09</td>
<td>4.04E-15</td>
<td>0.5846</td>
<td>0.1982</td>
<td>0.9134</td>
<td>0.0044</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 4.1 Genetic correlation between each trait**

Furthermore, we have strengthened this finding in cross trait analysis and identified 90 loci contain SNPs with $P<5\times10^{-8}$ (Supplementary Table 6). Figure 1 shows the Manhattan plot of the cross trait analysis between asthma and allergic diseases. 51 loci were found to overlap with asthma single trait analysis, 35 loci were found to overlap with allergic diseases. Eight exactly overlap SNPs between asthma and allergic diseases were also identified through cross trait analysis. The strongest association signal was observed on chromosome 6 at HLA-DQ region (rs9273374, $P_{\text{meta}}=7.87\times10^{-35}$)(Supplementary Table 6), which was the first asthma-susceptibility locus to be identified (Marsh, Meyers, and Bias 1981), and extended haplotypes encompassing HLA-DQ and HLA-DR have been studied for their effects on specific allergen sensitization (Moffatt et al. 2001). The second strongest signal was an overlap SNP (rs7936070, $P_{\text{meta}}=2.81\times10^{-28}$) between asthma and allergic diseases mapped on C11orf30 on chromosome 11q13 (Supplementary Table 6), where it is an overlap gene associated with total serum IgE levels and increases susceptibility to poly-sensitization. The third strongest signal was observed
on IL1R1 genes (rs72823641, $P_{\text{meta}}=1.58\times 10^{-27}$) [ref: Interleukin-1 Receptor Cluster: Gene Organization of IL1R2, IL1R1, IL1RL2(IL-1Rrp2), IL1RL1(T1/ST2), and IL18R1(IL-1Rrp) on Human Chromosome 2q] (Fig. 3.1 and Supplementary Table 6), the cytokine receptors have impact on many cytokine induced immune and inflammatory responses, such as asthma and allergy (Moffatt et al. 2010, Gudbjartsson et al. 2009, Paternoster et al. 2015). The fourth strongest signal observed on SMAD3 (rs56062135, $P_{\text{meta}}=1.56\times 10^{-22}$) is another loci found overlap in both asthma and allergic diseases, which is known for have a role in responses of immune cells (Anthoni et al. 2007). We have also found FLG (Filaggrin) gene (rs61816766, $P_{\text{meta}}=4.63\times 10^{-12}$) that are associated with both asthma and allergic diseases, which is crucial for maintaining normal skin layer function (McLean 2011). Our cross trait results showed most of the loci are significant for two diseases, could be taken to suggest that it is this shared genetic loci between two diseases driving the overall significant positive correlation.

To understand whether the 90 loci as a group are enriched for expression in certain tissue types, we used the GTEx pilot data (Consortium 2013). Our GTEx enrichment analysis identified three tissues that were significantly enriched (after BH correction) for expression of the cross trait-associated genes (Fig 3.3). The two most strongly enriched tissues were part of the integumentary system (skin sun exposed and not sun exposed). The one other significantly enriched tissue was from the digestive system (esophageal mucosa). Whole blood and vagina also showed moderate enrichment although it’s not significant. Skin, esophageal mucosa and vagina tissues are all characterized by stratified squamous epithelium (Ganesan, Comstock, and Sajjan 2013). This suggests that the enrichment of asthma and allergy risk variants in genes expressed in tissues with stratified squamous epithelium component that is not specific to certain organ or tissue; rather, it seems to be generalizable across stratified squamous epithelium types.
Furthermore, in order to understand which tissues are more responsible for each disease, we applied the GTEx analysis again in the result of single trait analysis.

In our replication analysis, we have CPMA to investigate the shared genes between asthma and one allergic disease, eczema. Seventeen loci demonstrated genome-wide significance (P<5×10\(^{-8}\)). Of these, genes IL13, C11orf30, IL33, IL4, IL1RL1, IL18R1, IL18RAP, MIR4772, SLC9A4 and GSDMA were replicated comparing with UK Biobank cohort’s results (Table 3.2). SMAD3
was found as a suggestive gene (rs744910, $P_{\text{meta}}=1.46 \times 10^{-7}$) even though it does not reach to the genome-wide significance (Supplementary Table S6). While the phenotypes in our replication phase are not exactly same as UK Biobank’s, these results underscore the emerging overlap in genetic bases of asthma and allergy. Although additional replication in other different types of allergy is recommended to confirm the associations, many likely represent true risk alleles shared by asthma and allergy.
WebGester highlighted several biological processes significantly enriched by the shared gene set between asthma and allergic diseases in GO terms, which mainly includes epithelium development, keratinization, skin development and T helper 1 type immune response (Figure 3.4 and Supplementary Table 7). Also KEGG pathway analysis found the shared genes were significantly enriched in immune related pathway (Supplementary Table 8).

### Table 4.2 Cross trait analysis result of asthma and eczema in replication cohort

<table>
<thead>
<tr>
<th>SNP</th>
<th>P</th>
<th>N</th>
<th>CHR</th>
<th>POS</th>
<th>KB</th>
<th>RANGES</th>
<th>Nearest genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1218891</td>
<td>1.98E-19</td>
<td>105</td>
<td>5</td>
<td>chr5:131881605..132002125</td>
<td>120.521</td>
<td>[IL13, RAD50]</td>
<td></td>
</tr>
<tr>
<td>rs2212434</td>
<td>2.13E-19</td>
<td>68</td>
<td>11</td>
<td>chr11:76136482..76302073</td>
<td>165.992</td>
<td>[C11orf30]</td>
<td></td>
</tr>
<tr>
<td>rs479844</td>
<td>3.87E-17</td>
<td>80</td>
<td>11</td>
<td>chr11:65401336..65562257</td>
<td>160.922</td>
<td>[AP5B1, KATS5, MIR4489, MIR4690, OVOL1, OVOL1-AS1, PCNXL3, RELA, RNASEH2C, SIPA1]</td>
<td></td>
</tr>
<tr>
<td>rs7216389</td>
<td>7.53E-17</td>
<td>276</td>
<td>17</td>
<td>chr17:37810218..38215117</td>
<td>404.9</td>
<td>[CSF3, ERBB2, GRB7, GSDMA, GSDMB, IKZF3, LRRC3C, MED24, MIEN1, MIR4728, MIR6884, ORMD3, PGAP3, PNMT, PSMD3, SNORD124, STARD3, TCAP, ZPB2]</td>
<td></td>
</tr>
<tr>
<td>rs992969</td>
<td>1.22E-12</td>
<td>46</td>
<td>9</td>
<td>chr9:6146441..6237547</td>
<td>91.107</td>
<td>[IL33]</td>
<td></td>
</tr>
<tr>
<td>rs1174781</td>
<td>2.33E-12</td>
<td>82</td>
<td>5</td>
<td>chr5:132009154..132071212</td>
<td>62.059</td>
<td>[IL4, KIF3A]</td>
<td></td>
</tr>
<tr>
<td>rs1020829</td>
<td>4.82E-12</td>
<td>212</td>
<td>2</td>
<td>chr2:102858490..103091618</td>
<td>233.129</td>
<td>[IL18RL1, IL18R1, IL18RAP, MIR4772, SLC9A4]</td>
<td></td>
</tr>
<tr>
<td>rs1257676</td>
<td>5.13E-10</td>
<td>3</td>
<td>11</td>
<td>chr11:65354926..65488556</td>
<td>133.631</td>
<td>[ACSL6, CSorf56, CSF2, IL3, IRF1, LOC553103, MIR3976, MIR6830, P4HA2, P4HA2-AS1, PDLIM4, SLC22A4, SLC22A5]</td>
<td></td>
</tr>
<tr>
<td>rs2073643</td>
<td>1.20E-09</td>
<td>171</td>
<td>5</td>
<td>chr5:131336287..131835999</td>
<td>497.317</td>
<td>[ACSL6, CSorf56, CSF2, IL3, IRF1, LOC553103, MIR3976, MIR6830, P4HA2, P4HA2-AS1, PDLIM4, SLC22A4, SLC22A5]</td>
<td></td>
</tr>
<tr>
<td>rs6419573</td>
<td>1.43E-09</td>
<td>96</td>
<td>2</td>
<td>chr2:102927726..103188785</td>
<td>261.06</td>
<td>[IL18RL1, IL18R1, IL18RAP, MIR4772, SLC9A4]</td>
<td></td>
</tr>
<tr>
<td>rs2918299</td>
<td>8.61E-09</td>
<td>24</td>
<td>19</td>
<td>chr19:8785744..8796160</td>
<td>10.417</td>
<td>ADAMTS10, ZNF558</td>
<td></td>
</tr>
<tr>
<td>rs1295301</td>
<td>1.63E-08</td>
<td>60</td>
<td>17</td>
<td>chr17:47341580..47461433</td>
<td>119.854</td>
<td>[LOC102724596, MIR6129, ZNF652]</td>
<td></td>
</tr>
<tr>
<td>rs1837253</td>
<td>2.09E-08</td>
<td>1</td>
<td>5</td>
<td>chr5:110401872..110401872</td>
<td>0.001</td>
<td>TSLB</td>
<td></td>
</tr>
<tr>
<td>rs1041973</td>
<td>2.13E-08</td>
<td>4</td>
<td>2</td>
<td>chr2:102955468..103092584</td>
<td>137.117</td>
<td>[IL18RL1, IL18R1, IL18RAP, MIR4772, SLC9A4]</td>
<td></td>
</tr>
<tr>
<td>rs4129267</td>
<td>2.70E-08</td>
<td>27</td>
<td>1</td>
<td>chr1:154395212..154428283</td>
<td>33.072</td>
<td>[IL6R]</td>
<td></td>
</tr>
<tr>
<td>rs2284033</td>
<td>3.10E-08</td>
<td>8</td>
<td>22</td>
<td>chr22:37531436..37537514</td>
<td>6.079</td>
<td>[IL2RB]</td>
<td></td>
</tr>
<tr>
<td>rs3859189</td>
<td>4.62E-08</td>
<td>89</td>
<td>17</td>
<td>chr17:38112608..38216933</td>
<td>104.326</td>
<td>[CSF3, GSDMA, MED24, MIR6884, PSMD3, SNORD124]</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4. GO enrichment analysis identified pathway related to shared genes between asthma and allergy.
Discussion
To our knowledge, this is the first large scale cross trait GWAS analysis of asthma and allergic
diseases based on doctor diagnosis. We have identified strong genetic correlation between
asthma and allergic diseases (hayfever, allergic rhinitis or eczema) and shared genetic loci
between them. Further we have replicated our finding in two independent GWAS
studies.(Moffatt et al. 2010, Paternoster et al. 2015) We have also found shared genetic loci
between asthma and allergic diseases were significantly enriched in skin and esophagus mucosa
tissue expression, whole blood and vagina also showed enrichment although it’s not significant.

Our predominant finding in whole blood tissue enrichment indicated that these diseases were
caused by some malfunction of the shared immune system. An important gene that we found to
be associated with both allergic diseases and asthma is CD247, which is a Protein Coding gene
that plays important role in T-Helper Cell ICos-ICosL Pathway. Indeed, patients who suffer from
both allergic diseases such as eczema and asthma have an overactive T helper 2 (Th2) immune
system, the arm of the adaptive immune response that spurs IgE antibody production against
allergens. While Th2 cytokines are indeed present, it’s not a primary malfunction in the immune
system, but a genetic malfunction in the skin that initiates eczema and, in some cases,
asthma.(McLean 2011) RUNX3 [rs6600246, OR=0.89, \(P_{\text{meta}}=2.81\times10^{-8}\)] is another gene that
found to be associated with both asthma and allergic diseases. It was previously reported as a
risk variant for many immune-related inflammatory diseases, such as asthma, Crohn’s disease,
ulcerative colitis and eczema(Laprise 2014, Lotem et al. 2015, Guo et al. 2010, Esparza-Gordillo
et al. 2013). The lack of RNUX3 expression can lead to series of inflammatory responses related
to leukocytes, such as CD8\(^+\) T cells, NK and DCs(Lotem et al. 2015), which is consistent with
our finding that RUNX3 showed protective effect in both asthma and allergic diseases.
We found genes that are associated with IgE function can consist of another crucial pathway for atopic diseases, such as asthma and eczema. We have found C11orf30 on chromosome 11 (Supplementary table 1,2,3) as another shared gene between asthma and allergic diseases. C11orf30 was reported to be associated with total serum IgE levels and increases susceptibility to poly-sensitization (Li et al. 2012). When individuals are exposed to allergen, IgE are released from immune system and travel to local organ or tissues to release type 2 cytokines such as IL-4 and IL-13 and other inflammatory mediators. (Novak, Kraft, and Bieber 2001) These cytokines can further cause the allergic diseases and asthma, such as reddish/ dry skin and coughing, wheezing or shortness of breath (Wynn 2015, Barnes 2001).

Also what at first appeared to be a surprising association of respiratory disease with skin tissue began to make sense on a molecular and genetics level as we explore of the shared genes function. We have found FLG gene on chromosome 1 (Supplementary table 1,2,3) was associated with both allergic diseases and asthma. FLG is important for the formation of the stratum corneum and also for the hydration of this barrier layer (McLean 2011). People who have mutations in the FLG gene can be sensitive to external allergens and produce dry and flaky skin (McLean 2011). Meanwhile it can activate a strong allergic immune response in through peripheral blood to many organs, such as lung, causing the inflammation in the lungs that results in shortness of breath (McLean 2011).

Thanks to the benefit of large sample size from UK Biobank, we have been able to identify novel loci for asthma and allergic diseases. Total of 12 new loci identified for asthma and 12 for allergic diseases. Of these, D2HGDH and GAL3ST2, which are responsible for encoding
metabolism related D-2hydroxyglutarate dehydrogenase enzyme(Struys et al. 2005) and tumor metastasis related Galactose-3-O-sulfotransferase 2 enzyme(Seko et al. 2002).

Our findings support for significant skin and esophageal mucosa tissue enrichment and moderate vagina and whole blood tissue enrichment of the loci for shared genes between asthma and allergic diseases, with the replication of the independent data sets. Skin, esophageal mucosa and vagina tissues are all characterized by stratified squamous epithelium, which is a squamous epithelial cells arranged in layers upon a cutaneous membrane(Presland and Dale 2000, Ganesan, Comstock, and Sajjan 2013). This type of epithelium is often permeable and occurs where small molecules pass quickly through membranes via filtration or diffusion. The airway, containing stratified squamous epithelium, increased their permeability in asthma patients for vascular remodeling. Also whole blood is responsible for production of anti-inflammatory cytokines and circulating them in the whole body (Meijer et al. 2003).

Clinical and epidemiologic observations have shown that immune-mediated inflammatory and autoimmune diseases can be associated due to overlap of genes(Cotsapas et al. 2011). Indeed, our LDSC analysis showed high genome-wide genetic correlation between asthma and allergic diseases and identified shared genes between them. However we didn’t find significant genetic correlation between asthma and other 3 immune related diseases, rheumatoid arthritis, Crohn's disease and ulcerative colitis, and any overlapping genes, suggesting that the patterns of shared genes among immune related traits can still be distinct. Asthma and allergic diseases both belong to IgE mediated hypersensitivity (Type I), while the other three traits are immune related, rheumatoid arthritis belongs to immune complex mediated (Type III) hypersensitivity, Crohn's disease and ulcerative colitis belong delayed cell mediated (Type IV) hypersensitivity. Thus we
found strong genetic correlation for the traits within each hypersensitivity category. Different types of hypersensitivity are characterized by different types of gene sets.

We also acknowledge the limitations of our study. First of all, more allergic diseases will be needed to investigate if genes those are specific to certain types of allergy sharing with asthma, such as food allergy. Also, our allergic symptom phenotype contained both allergic rhinitis/hay fever and eczema. The allergy genes we found may not be associated with specific type of allergy. However, since allergic rhinitis and eczema are both type I hypersensitivity and characterized by epithelial cell, they should share highly similar physiology. (Alonso and Fuchs 2003, Pawankar et al. 2011) Finally, our UK Biobank cohort traits are not independent, there are shared cases and complete shared controls between asthma and allergic diseases. However, for all of our analysis, we have applied methods that can adjust potential effect due to overlapping samples; and we have successfully replicated our finding in two independent GWAS data.

These shared genetic components can provide opportunity to investigate asthma and allergy co-morbidity on molecular level. Drugs that targets to shared genetic loci can be useful to have doubled effect of treating diseases. And our findings can also inspire patients and doctors to be more cautious about co-morbidity of asthma and allergic diseases in the clinical practice.

**Conclusion**
Understanding the genetic overlap between asthma and allergic diseases can be beneficial to treat disease outcome more efficiently. Our GWAS study has highlighted shared genetic pathway in immune and inflammatory systems and skin tissue by asthma and allergic diseases. This work should advance understanding of the molecular processes underlying asthma and allergy and open new avenues of treatment for them.

**Acknowledgments**
This research has been conducted using the UK Biobank Resource under application number 16549. We would like to thank the participants and researchers from the UK Biobank who significantly contributed or collected data. We thank to Dr Vernerri Anttila and Dr Steven Gazal for their statistical advice.
References


Summary and Conclusion

My first study links the expression of miR-181a, miR-92a, and miR-424 in whole blood to ARDS. Inflammatory response markers miR-181a and miR-92a were significantly elevated in ARDS patients, while the pulmonary artery EC anti-inflammation marker miR-424 was significantly reduced in ARDS patients. Further, the expression patterns of our miRNA biomarkers may provide an in depth molecular understanding between ARDS and at-risk patients on top of sepsis and pneumonia. By combining our miRNA biomarkers with LIPS, diagnosis of ARDS may be improved. Circulation levels of these miRNAs may have the potential to guide ARDS treatments by targeting these miRNAs.

My second study, to our knowledge, is the first study of miRNA as a prognostic classifier from whole blood for ARDS 28-day mortality. Whole blood contains both immune cell- and tissue-specific miRNAs and thus offers a major advantage for miRNA profiling compared with other tissue types. While our study confidence is limited by sample size and the mortality rate in this small cohort is high and may not be representative of general ARDS cohort, the classifier containing miRNAs discovered in this study offers a potentially valuable, novel biomarker signature in the clinical practice to better ARDS 28-day mortality prognosis.

My third study provides a comprehensive understanding the genetic overlap between asthma and allergic diseases can be beneficial to treat disease outcome more efficiently. Our GWAS analysis has highlighted shared genetic pathway in immune and inflammatory systems and skin tissue by asthma and allergic diseases. This work should advance understanding of the molecular processes underlying asthma and allergy and open new avenues of treatment for them.