IDENTIFICATION OF NOVEL CODING SINGLE NUCLEOTIDE POLYMORPHISMS ASSOCIATED WITH ACUTE RESPIRATORY DISTRESS SYNDROME

A THESIS IN
Bioinformatics

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
The requirements for the degree

MASTER OF SCIENCE

By
KATHERINE ANNE SHORTT

B.S, Indiana University, 2007

Kansas City, Missouri

2014

IDENTIFICATION OF NOVEL CODING SINGLE NUCLEOTIDE POLYMORPHISMS ASSOCIATED WITH ACUTE RESPIRATORY DISTRESS SYNDROME

Katherine Anne Shortt, Candidate for the Master of Science Degree

University of Missouri-Kansas City, 2014

ABSTRACT

ARDS is a lung condition characterized by impaired gas exchange with systemic release of inflammatory mediators, causing inflammation, hypoxemia and multiple organ failure. Disease susceptibility and progression are poorly understood and there are few effective therapeutic options. Existing biomarkers have limited effectiveness as diagnostic and therapeutic targets. Whole-exome sequencing is an effective tool in detection of disease-causing genetic variants in complex genetic conditions such as acute respiratory distress syndrome (ARDS).

To identify disease-causing variants in ARDS patients, whole-exome sequencing was performed on 96 patient DNA samples from the National Heart, Lung and Blood Institute's ARDS Network (ARDSnet). By comparing these exome data with 625 participants of the 1000 Genomes Project, we have tentatively identified a number of single nucleotide polymorphisms (SNP) which are potentially associated with ARDS. In this study, we validated three SNPs (rs78142040, rs9605146, and rs3848719) in an

additional 117 ARDS patients using TaqMan SNP genotyping assays (Life Technologies) to substantiate their associations with the susceptibility, severity and outcome of ARDS.

rs78142040 (C>T) occurs within a histone mark in intron 6 of the Arylsulfatase D gene. rs9605146 (G>A)(also known as rs114989947) causes a coding change (proline to leucine) with a deleterious effect in the XK, Kell blood group complex subunit-related family, member 3 gene. rs3848719 (G>A) is a synonymous SNP in exon 5 of gene Zinc-Finger/Leucine-Zipper Co-Transducer NIF1. rs78142040 and rs9605146 are significantly associated with susceptibility to ARDS[minor allele frequency (MAF): 0.219 versus 0.003 (control), $p<2.95\times10^{-7}$ and 0.386 versus 0.046 (control), $p<2.95\times10^{-7}$; respectively]. rs3848719 (MAF 0.394 in cases, 0.287 in controls) is associated with APACHE II score when the score quartiles are compared (p=0.032 OR=0.549, 95%CI=0.313-0.96) and Rs78142040 approaches significant association with APACHE II score (p=0.061). rs78142040 is associated with the 60-day mortality in the 213 ARDS patient population (p=0.017, OR=2.039, 95%CI=1.130-3.681). The same trends hold by stratification of patient population and comorbidity variables. All SNPs are in Hardy-Weinberg Equilibrium (HWE $p>1\times10^{-4}$) in our cases. These 3 SNPs have not been previously associated with ARDS and represent potential new genetic biomarkers for ARDS. More validations in larger patient populations and further exploration of underlying molecular mechanisms are warranted.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Medicine, have examined a thesis titled "Identification of Novel Coding Single Nucleotide Polymorphisms Associated With Acute Respiratory Distress Syndrome", presented by Katherine Anne Shortt, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Shui Qing Ye, M.D, Ph.D.
Department of Biomedical and Health Informatics and Department of Pediatrics

Dmitry Grigoryev, M.D, Ph.D.
Department of Biomedical and Health Informatics and Department of Pediatrics

Lakshmi Venkitachalam, Ph.D., MPH,
Department of Biomedical and Health Informatics

CONTENTS

ABSTRACT iii

LIST OF ILLUSTRATIONS vii

LIST OF TABLES viii

GLOSSARY ix

ACKNOWLEDGMENTS xii

Chapter

1. INTRODUCTION 1

2. REVIEW OF THE LITERATURE 4

3. RESEARCH QUESTION 6

4. METHODS 8

    Sampling 8

    Research Design 10

    Analysis 11

    Genotyping of Selected Candidates 18

5. RESULTS 20

6. DISCUSSION 39

7. CONCLUSIONS 46

Appendix

A. SUPPLEMENTARY FIGURES AND TABLES 47

REFERENCE LIST 59

VITA 62

ILLUSTRATIONS

Figure Page

1. The Steps of Next-Generation DNA Sequence Analysis 11
2. Pipeline of the exome-seq data analysis workflow.....13
3. Manhattan Plot of ARDS Patients and 1000 Genomes Project Controls 21
4. Quantile-Quantile Plots of Genotypic Trend Test χ^2 Values for the ARDS and 1000 Genomes Project Population.....35

TABLES

Table Page

1. Participant Demographics and Comorbidities for the ARDS Cases....9
2. The Comparison Groups for Genetic Association Analysis....12
3. The Populations and Subpopulations Used for Analyses....14
4. Summary of the Filtering Applied to Candidate SNPs....22
5. Top Canonical Pathways of Genes Containing SNPs Associated with ARDS....23
6. Overall Association Summary....24
7. Rs78142040 Statistics....25
8. Chi-Square Test P-Values for Rs78142040 in Patients and Controls....26
9. Logistic Regression of Genotype and APACHE II Score by Quartile....27
10. Logistic Regression of Genotype and 60-Day Mortality....28
11. Predicted Effects of the 4 SNPs on Amino Acid Coding....29
12. Rs9605146 Statistics....30
13. Chi-Square Test P-Values for Rs9605146 in Patients and Controls....31
14. Rs3848719 Statistics....32
15. Chi-Square Test P-Values for Rs3848719 in Patients and Controls....33
16. A Summary of the 4 Genotyped SNPs and the Effect of the PCA and Outlier Removal on Their Genotypic Trend Test P-Values....38

GLOSSARY

Acute respiratory distress syndrome: (ARDS) A severe condition resulting from illness or injury and is characterized by inflammation and fluid buildup in the lungs. Oxygen is incapable of crossing from the lungs into the blood.

Allele: One of a number of alternate nucleotides of a specific locus on a strand of DNA.

Amino acid: Organic acids that form the building blocks of proteins. Each of the 20 amino acids are coded by a 3 nucleotide DNA sequence.

Deoxyribonucleic Acid (DNA): Nucleic acids that encode genetic instructions of all living organisms.

Eigenvalue (principal component): magnitude of the variation along the eigenvector which is identified using principal component analysis. The product of a non-zero eigenvector and a square matrix.

Eigenvector: Axis of variation that is identified using principal component analysis. Eigenvectors are defined as non-zero vectors associated with matrix equations.

Exon: DNA sequence that remains in the mature RNA of a gene after introns are removed.

Exome: Genome sequence that is transcribed to the mature RNA, comprised of all of the exons.

Exome-seq: Exome sequencing is a method of DNA sequencing that selectively sequences protein coding regions of the genome (exons).

Genome: The total genetic material of an organism.

Genome-wide association study: case-control study examining the association of common genetic variants with other traits.

Genotype: The set of alleles that an organism possesses for a given genetic locus. Normally a genotype contains 2 alleles due to the presence of 2 chromosomes.

Haplotype: A combination of nearby alleles that are inherited together more frequently than expected.

Hardy-Weinberg Equilibrium (HWE): The expected distribution of genotypes giving the existing allele counts according to the equation $(p+q)^2=p^2+2pq+q^2$, where the alleles are p and q. In HWE, allele and genotype frequencies remain consistent across generations.

Intron: DNA sequence in a gene but is not transcribed into the mature RNA.

Linkage disequilibrium: When alleles for two genetic loci are not distributed randomly the loci are in linkage disequilibrium.

Major allele: For a genomic position the allele that occurs most frequently in a population is the major allele.

Manhattan plot: A plot of the genetic variant location vs. the –log10 (χ^2 p-value). It is used to visualize associations across the genome in a population.

Mendelian disease: Diseases that have simple genetic inheritance patterns and are caused by a small number of genes.

Minor allele: For a genomic position an allele that occurs in a population less frequently than a different allele at the same position is called a minor allele.

Next generation sequencing (NGS): Current high throughput genetic sequencing technology, which produces the order of nucleotides within a DNA molecule.

Nonsynonymous mutation: A genetic variant which causes a change in amino acid coding.

Principal component: See eigenvalue.

Principal component analysis (PCA): A mathematical method for identifying and adjusting for axes of variation within a dataset. PCA uses an orthogonal transformation to convert variables into linearly uncorrelated variables we call eigenvalues.

Q-Q plot (quantile-quantile plot): Compares the expected and actual values for a probability distribution (commonly the χ^2, χ^2 p-value, or the Hardy-Weinberg p-value).

Scree plot: A plot of the magnitude of the eigenvalues. This type of plot is used to visualize eigenvalues produced by principal components analysis that may be retained in association tests.

Single nucleotide polymorphism (SNP): Genetic variant that occurs at one nucleotide (base pair). Different alleles are observed at one position across a population.

Synonymous mutation: A genetic variant in a protein coding region that causes no change in amino acid coding.

Whole-exome sequencing: (WES) see Exome-Seq

ACKNOWLEDGMENTS

Thanks to Dr. Ye, Dr. Grigoryev, and Dr. Venkitachalam as well as the faculty and staff in Division of Experimental and Translational Genetics, Children's Mercy Hospital and the faculty and staff of the Department of Biomedical and Health Informatics for their time and guidance in the completion of this project. Thanks to my friends and family for constantly encouraging me to pursue a field that I love and listening to my musings along the way. We would additionally like to acknowledge the sponsorship of Roy G Brower, MD, Johns Hopkins University School of Medicine and the ARDSnet team (www.ardsnet.org) for providing 213 ARDS patient DNA samples in this study.

FUNDING

Funding was provided by the NHLBI/NIH Grant (HL080042 & HL080042-S1, Ye, SQ), start-up fund of Children's Mercy Hospitals and Clinics, UMKC (Ye, SQ), a Sarah Morrison Student Research Award of UMKC (Shortt, K), and a UMKC GAF Award of UMKC (Shortt,K).

CHAPTER 1

INTRODUCTION

Acute respiratory distress syndrome (ARDS), a severe form of acute lung injury, is characterized by the inflammation and fluid build-up of the alveoli in the lungs, which reduces the ability of oxygen to cross over into the blood stream[1,2]. ARDS has an extremely high mortality rate where over a third of sufferers die, and many of the survivors experience complications such as brain damage due to prolonged oxygen deprivation[3,4]. The etiology and pathology of the disease are not fully understood and there is a paucity of unique and effective therapies. ARDS is estimated to have an age-adjusted incidence of 86.2 new cases per 100,000 person-years in adults age 15 and older and the total number of cases estimated to occur yearly in the US is about 190,000[3].

It is recognized that ARDS is caused by a complex interplay between multiple genetic and environmental factors, yet few advances in prevention and reduction of mortality and morbidity have been made[3,5-7]. It has recently been shown that complex diseases can be between 50 and 90% genetically determined[8]. Sepsis is the leading cause of ARDS, followed by elective surgery, pulmonary infection, trauma, and other causes[9]. The disease remains in need of specific and efficient preventative measures and treatments that can be developed using genetic investigation to identify novel biomarkers and treatment targets.

With the development of next-generation sequencing technologies and improvements in data analysis capabilities, it is now feasible to sequence and analyze whole genomes in a couple of days[10]. However, the cost for whole genome sequencing is

still a prohibitive factor for sequencing many samples. Whole exome sequencing (WES) is faster and less expensive, making it ideal for the study of variants that cause structural changes as candidate gene mutants to human diseases[11]. WES is used to study both mendelian and complex diseases. Complex diseases, such as ARDS, can arise in populations that are genetically very heterogeneous. This makes complex diseases difficult to predict and treat, necessitating the use of more complex genetic studies, such as association studies and linkage analyses[12,13]. In the study of ARDS it becomes important to take into account the make-up of the study populations. In our study, ethnicity, presence of pneumonia and presence of sepsis are all considered in the association analyses.

The advancements in sequencing technology have created a need for increased bioinformatics and data analysis. Genome-wide association studies (GWAS) are used to quantify the association of genetic variations and phenotypes in populations. This data analysis technique was developed in response to the increasing volume of genomic data in 2005[6,14]. Genetic association studies compare the association of genotype frequencies and specific traits between the case and control populations[12,15,16]. Whole exome sequence analysis can be applied to perform whole exome wide association studies to focus on the identification of coding SNPs associated with human diseases. By pursuing SNPs that are above a threshold of significance and are suspected to have structural effects, it becomes possible to pare down the list of candidate variants. By comparing our patient population to the freely available, healthy 1000 Genomes Project control genomes and to the different stages of diseases, we will be able to identify SNPs

associated with ARDS susceptibility, severity and outcome, which will be determined using linear and logistic regressions of patient data, respectively[17].

CHAPTER 2

REVIEW OF THE LITERATURE

A primary goal of SNP target identification is to identify biomarkers for diagnosis and treatment. ARDS Biomarkers that have been previously studied that are present in blood serum include surfactant-associated proteins (SP-A, B, and C), Mucin-associated antigens (KL-6 and MUC1), Cytokines (IL-1, 2, 6, 8, 10, and 15, TNFα)endothelium activation markers (E,L selectim, I-CAM-1, V-CAM-1, and VWF), and neutrophil activation markers (MMP-9, LTB4, and Ferritin). Cytokine levels have been identified as a moderately effective measure of severity[17]. Other biomarkers of ARDS severity that can be obtained from breath include hydrogen peroxide, acidity of the breath[18]. Despite the identification of these biomarkers, their diagnostic capabilities are inconsistent and limited The mortality rate of ARDS remain high, especially the mortality of cases with common comorbidities sepsis with suspected pulmonary source (40.6%) and witnessed aspiration(43.6%) [19]. This warrants that further research be conducted to identify novel genetic targets for ARDS prevention and management.

One of the recently characterized biomarkers of ARDS is Pre-B cell Colony Enhancing Factor (PBEF). Dr. Shui Qing Ye's group identified several SNPs in the human PBEF gene promoter that affect the function of the PBEF gene. They reported PBEF was a novel biomarker for ARDS[20,21]. Their research surrounding the effect of the two PBEF promoter polymorphisms suggest that the -1535T (originally labeled as -1543T) variant allele is associated with a decreased susceptibility to ALI/ARDS and a better outcome in septic patients in a Caucasian population when compared with patients

without the variation. The -1001G variant allele was associated with increased susceptibility to acute lung injury and ARDS in African American and Caucasian populations. The -1001G variant is also associated with a higher ICU mortality rate in septic patients in a Caucasian population[22-25].

However, it is increasingly recognized that ARDS is a complex disease with multiple genetic components that warrant further study. The lack of currently identified biomarkers that have consistently effective application in predicting ARDS susceptibility and progression is the fundamental basis for our study, where the ultimate goal is to connect the dots between genetics and ARDS outcome. With current NGS technology, it is possible to conduct whole exome sequencing of a sample population and validate the population in a secondary or separate population or data. This data can be used to identify new diagnostic and therapeutic targets. Genotyping of selected SNPs in an additional population can confirm the findings of candidate SNPs.

CHAPTER 3

RESEARCH QUESTION

The purpose of this study is to discover new candidate genes for ARDS risk assessment, diagnosis, and treatment. The identification and validation of new biomarkers hold promise for mechanistic insights and novel therapeutic targets. To accomplish this, whole-exome sequencing had been performed with the intent of identifying coding single nucleotide polymorphisms (SNPs) whose minor allele frequencies are significantly different in ARDS than those of healthy controls.

Aim 1: Whole-Exome Association

Using SNP analysis of whole exome sequence data from an ARDS patient population and normal healthy controls from Coriell and the 1000 Genomes Project, it was possible to identify coding SNPs associated with ARDS susceptibility.

Aim 2: Further Measures of Association and SNP Effects

Regressions of selected SNPs within the patient population were used to assess association with severity and outcome. Functional effect predictions of the SNPs were obtained to substantiate potential causality.

Aim 3: Validate Selected SNPs in a Larger Patient Population

These data were used to choose candidate SNPs for further validation in a larger patient population using genotyping. Genetic association analysis was performed to determine associations between the SNPs and ARDS susceptibility, severity and outcome. Exome sequencing of 96 patient DNA samples from the ARDSnet was performed using Illumina's HiScanSQ system. Using these data and the 1000 Genomes Project we identified a number of SNPs which are potentially associated with ARDS. This study validates three SNPs (rs78142040, rs9605146 and rs3848719) in an additional 117 ARDS patients for a total of 2013 cases using TaqMan genotyping assays (Life Technologies) to substantiate their associations with the susceptibility, severity and outcome of ARDS.

CHAPTER 4

METHODOLGY

Sampling

To perform this case-control study, we used information from DNA samples obtained from various study groups (table 1) and from the 1000 Genomes Project. Data on ARDS cases was derived from a 213-participant population obtained from the NHLBI ARDS network 05: Fluid and Catheter Treatment Trial[26,27]. Phenotype information for this population was collected from the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC, http://biolincc.nhlbi.nih.gov.home/). Exome sequencing was performed on 96 cases using HiScanSQ (Illumina, CA, USA) in the Core of Genetic Research directed by Dr. Shui Qing Ye. TaqMan genotyping of selected candidate SNPs were performed on the remaining 117 patients.

These data were compared with 625 European Ancestry (EUR) and African Ancestry (AFR) samples from the 1000 Genomes Project. 1000 Genome data includes genomic context, population genetics, flanking sequence, and any synonyms used in publication. The 1000 Genomes Project AFR population (n=246, 46.8% male) was used as a control population for the African American ARDS samples and the EUR population (n=379, 47.0% male) will be used as a control for the Caucasian ARDS population. The data identifies about 15 million SNPs[28]. Descriptors of the ARDS case data include race, gender, ARDS etiology (sepsis or pneumonia), ventilator-free days per 28 days after admission, APACHE II scores and 60 day mortality (table 1).

Table 1: Participant Demographics and comorbidities for the ARDS Cases

Whole-exome Sequencing Sample group	Participants*	Gender (%male)	Age± SD	APACHE II score ±SD	Ventilator-free days per 28 days ±SD	60-day mortality (% dead at 60 days post-onset)
ARDS total	96	46.9	50±14.6	97.2±30.4	12.6±10.2	29.4
Caucasian ARDs	70	44.3	50.7±13.8	94.5±30.2	13.6±10.0	23.9
African American ARDS	26	53.8	47.8±16.6	104.6±30.2	9.9±10.5	44
ARDS with Sepsis total	48	50	52.0±15.1	105.2±32.2	10.2±10.1	38.3
Caucasian ARDS with Sepsis	37	43.2	51.8±15.0	102.6±31.1	10.5±9.8	33.3
African American ARDS with Sepsis	11	72.7	52.9±16.1	113.6±35.8	9.3±11.7	54.5
ARDS with Pneumonia total	48	43.8	47.9±13.9	88.9±26.2	14.9±9.87	20
Caucasian ARDS with Pneumonia	33	45.5	49.5±12.4	85.0±26.6	17.0±9.2	12.9
African American ARDS with Pneumonia	15	40	44.1±16.6	97.6±24.0	10.3±10.0	35.7
TaqMan Genotyping Sample group	Participants	Gender (%male)	Age± SD	APACHE II score ±SD	Ventilator-free days per 28 days ±SD	60-day mortality (% dead at 60 days post-onset)
ARDS Total	117	46.2	50.8±16.6	105.1±32.7	12.0±10.5	31
Caucasian ARDs	75	40	52.1±15.4	104.3±31.4	11.8±10.5	29.2
African American ARDS	17	64.7	51.6±22.8	113.3±33.8	12.2±10.6	29.4
Other ancestry ARDS	25	52	46.4±15.3	101.8±36.2	12.6±10.7	37.5
ARDS with Sepsis total	59	52.5	51.5±17.3	112.9±31.5	11.3±11.0	37.9
Caucasian ARDS with Sepsis	34	50	51.1±15.6	113.6±29.7	10.9±11.1	33.3
African American ARDS with Sepsis	9	66.7	56.6±27.1	121.8±33.0	12.9±11.5	33.3
Other ancestry ARDS with Sepsis	16	50	49.6±14.8	106.4±34.7	11.4±11.0	50
ARDS with Pneumonia total	58	39.7	50.1±16.0	96.9±32.3	12.7±9.9	23.6
Caucasian ARDS with Pneumonia	41	31.7	52.9±15.3	96.4±31.0	12.6±10.0	25.6
African American ARDS with Pneumonia	8	62.5	46±16.9	103.8±34.3	11.4±10.2	25
Other ancestry ARDS with Pneumonia	9	55.6	46.8±15.2	92.4±39.7	14.9±10.3	12.5
Total ARDS	Participants	Gender (%male)	Age± SD	APACHE II score ±SD	Ventilator-free days per 28 days ±SD	60-day mortality (% dead at 60 days post-onset)
All ARDS patients	213	46.5	50.4± 15.7	101.6± 31.9	12.3± 10.3	30.2

*, 8 samples, 4 exome sequenced and 4 TaqMan genotyped ARDS patients did not have these phenotypes available. An additional 2 exome sequenced ARDS patients did not have ventilator-free days data. In addition to the 8 patients missing severity and mortality phenotype data, 2 patients were excluded from the regression

Ultimately, SNPs that had the largest impact on the susceptibility, severity and outcome were picked up for validation. Measures of ARDS severity include the number of ventilator-free days per 28 days (an indication of a patient's ability to breathe on their own) and the APACHE II score (a measure of the severity of a disease in adult patients). ARDS outcome is measured by 60 day mortality. These measures are compared with the SNP genotype frequency and participant population to determine the presence of correlations.

Research Design

The exome sequence libraries of the NHLBI ARDS samples and the Coriell normal controls were created using an Illumina HiScanSQ platform. Prior to sequencing, the DNA was prepared using the TrueSeq Exome Enrichment Kit (http://www.illumina.com). Paired-end sequencing with 101 base pair read lengths was preformed, providing a minimum average coverage depth of 50 ×.

Consensus Assessment of Sequence And variations (CASAVA) software was used for the conversion of .bcl files to .fastq format and for demultiplexing (http://www.illumina.com). The sequences were aligned to the hg19 human reference genome and variants alleles were called using the Genome Analysis Toolkit (http://www.broadinstitute.org/gatk/). Both the lab-processed ARDS patient samples and the 1000 Genomes Project controls were processed using the GATK methodology (figure 1)[22].

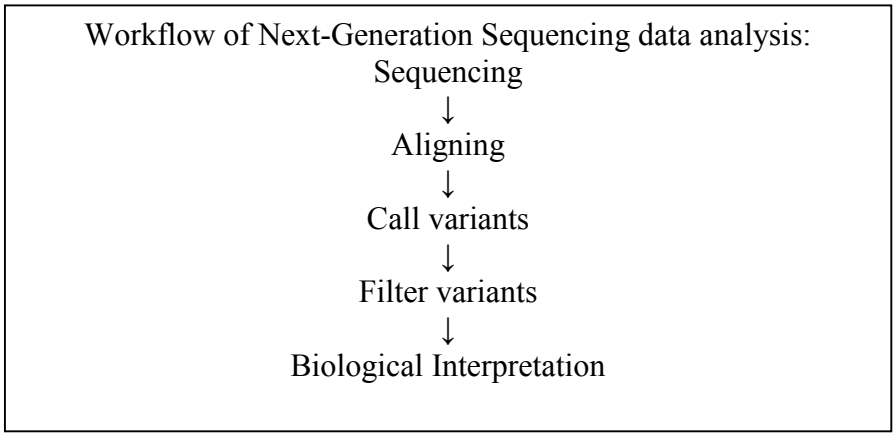


Figure 1: The Steps of Next-Generation DNA Sequence Analysis.

GATK was also used to generate the list of SNPs from the sequence data and match the SNPs in the sample population to their identifying rs number if one was available. Genotypes of the patient samples were considered to be high-confidence if the Phred-like quality was a minimum of 20 and there were at least 4× coverage depths. The exome sequence was chosen rather than the whole genome because it is more likely that variants causing structural changes will be observed in exon sequence and the cost and time are greatly reduced from whole genome sequencing[29].

Analysis

SNP Filtering and Susceptibility Association Test

The SNP analysis was done in several steps using SNP & Variation Suite v8.1.0 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). The SNP associations were measured by comparing the following groups as listed in Table 3. By comparing the SNP frequencies of smaller subsets of the data it is possible to identify relationships that could be obscured in the admixed data set (table 2).

Table 2: The Comparison Groups for Genetic Association Analysis.

Exome Seq Cases	**Secondary Controls**
96 Cases ARDS: whole exome sequenced group	625 1000 Genomes Controls
Cases stratified by race	1000 Genomes Controls stratified by race
Cases stratified by ARDS etiology (Sepsis or Pneumonia)	625 1000Genomes Controls
Cases stratified by population and ARDS etiology (Sepsis or Pneumonia)	1000 Genomes Controls stratified by Race

To conduct the statistical analyses the ARDS population VCF files were imported to SNP & Variation Suite 8 (SVS V8.1.0). The 1000 genomes exome sequence data was provided by Golden Helix in a *.dsf file. These files and their phenotype data was joined and used to calculate the χ^2 values and χ^2 p-values for the SNPs that were matched by location in SVS V8.1.0 using a basic allelic statistical model. Calculations were also stratified by ancestry and ARDS comorbidity (sepsis and pneumonia). 1,382,399 SNPs were identified in the ARDS patients and 714,071 SNPs were available from the 1000 Genomes Project. The Bonferroni corrected p-value was determined to be 2.95×10^{-7} because 169,376 SNPs were matched between the 1000 Genomes data and the ARDS patient data (figure 2).

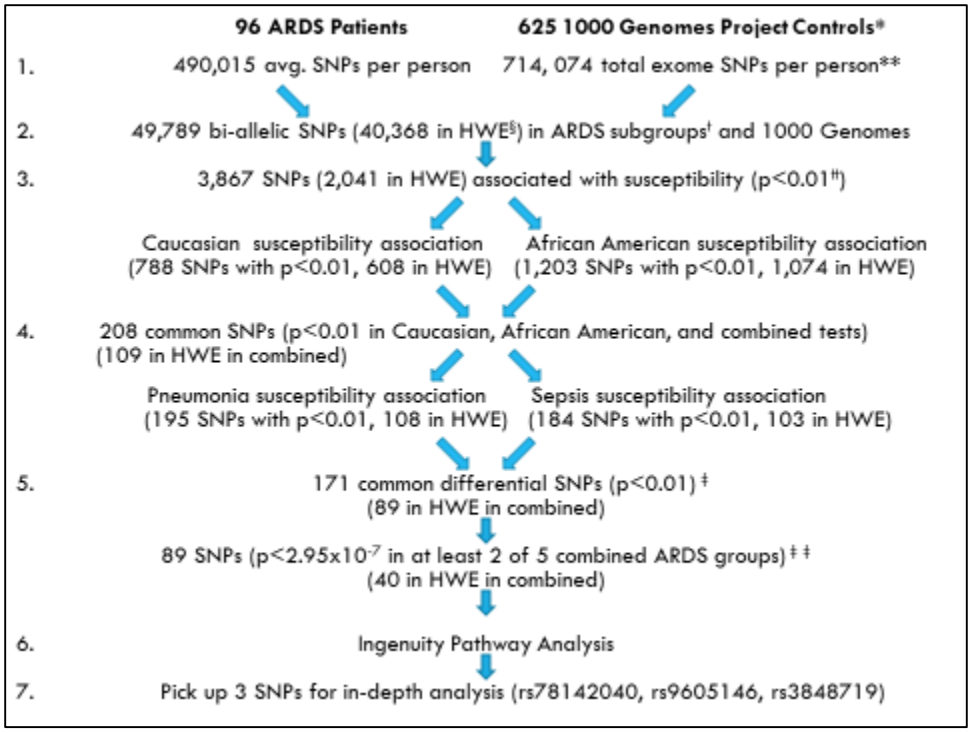


Figure 2: Pipeline of the exome-seq data analysis workflow. After processing the data using the GATK pipeline, this filtering workflow was derived to identify SNPs which were associated with measures of susceptibility across the racial and etiology groups of cases. SNPs were filtered based on strength of association, coding effect, and functional prediction prior to testing for association with other ARDS phenotypes. *, The sample contains African American and Caucasian patients, so the EUR and AFR healthy controls from 1000 Genomes were used for comparison; **, In the 1000 Genomes Project exome sequence, the same 714, 074 SNPs are present for all 625 EUR and AFR; §, HWE=Hardy Weinberg Equilibrium, p>0.0001; Ɨ, African American with pneumonia, African American with Sepsis, Caucasian with pneumonia, Caucasian with sepsis; Ɨ Ɨ, $\chi 2$ test of ARDS vs. respective 1000 Genomes Project control groups; ǂ, SNPs with P-value <0.01 in the overall comparison, Caucasian ARDS comparison, and African American comparison with 1000 Genomes were filtered further by p<0.01 in the Sepsis comparison and Pneumonia comparison; ǂ ǂ, All ARDS cases, all pneumonia cases, all sepsis cases, all African American cases, all Caucasian cases.

SNPs that were present in all 4 of the ARDS sub populations Caucasians with sepsis, Caucasians with pneumonia, African Americans with sepsis, and African Americans with pneumonia were selected for the statistical association tests, totaling

49,789 (table 3). SNPs with a chi square p-value smaller than the Bonferroni-corrected p-value in at least 2 of the 5 main populations (All ARDS vs. 1000 Genomes, All Sepsis vs. 1000 Genomes, All Pneumonia vs. 1000 Genomes, African American ARDS vs. AFR 1000 Genomes, and Caucasian ARDS vs. EUR 1000 Genomes) were considered further[23,24]. The χ^2 test of the categorical data was used to identify novel targets for study. This was represented graphically by Manhattan plots generated using SVS v7.8.

Hardy-Weinberg equilibrium and SNP call rate were calculated using the Marker Statistics feature of SVS V8.1.0. SNPs were considered to be in HWE if the p-value was $>1\times10^{-4}$. The call rate of the 1000 genomes project controls was 100%, meaning there were no missing genotypes. The call rate of the ARDS SNPs was considered to be good if it was >95% in the total exome sequenced population[17,30-37].

Table 3: The Populations and Subpopulations Used for Analyses

Population	SNPs Called (SVS)	X2 test compared with:	SNPs matched by location in SVS exome
AA pneumonia	629,987	AA 1000 Genomes	108,180
AA sepsis	533,174	AA 1000 Genomes	95,917
EA pneumonia	748,096	EA 1000 Genomes	94,039
EA sepsis	769,702	EA 1000 Genomes	89,380
All ARDS*	272,963 (1,382,399)	1000 Genomes	50,190
All sepsis*	324,514 (978,362)	1000 Genomes	57,247
All pneumonia*	365,324 (1,012,760)	1000 Genomes	63,755
All EA ARDS*	473,138 (1,044,661)	EA 1000 Genomes	74,359
All AA ARDS*	383,188 (779,973)	AA 1000 Genomes	75,120
1000 Genomes Project	714,071	NA	NA
1000 Genomes Project AA	714,071	NA	NA
1000 Genomes Project EA	714,071	NA	NA

*= SNPs common across included subgrous, followed by (all detected). Only SNPs that were in all subgroups were included in the X2 tests for association with case-control status. **The same ARDS populations were used for the regressions against ARDS phenotypes. Chi-square tests were run on SNPs that were in both the controls and the cases.

Regressions with Comorbidities Among ARDS Cases

The next phase of the data analysis was to investigate the association of candidate SNPs in the exome sequence with ARDS severity and outcome. The analyses were conducted using numerically coded genotypes in an additive model. The frequencies of the genotypes of the SNPs were compared with the APACHE II score using a linear regression and the APACHE II score quartiles 1 and 4 using a logistic regression. The genotypes were also compared using a linear regression of ventilator free days. The logistic regression of genotype vs. 60 day mortality will provide a measure of relationship between the SNP and the outcome. The significance level for these test were set at $p<0.05$ for the logistic regressions and $R^2 >0.8$ for the linear regressions. All regressions were run on the overall population as well as the subpopulations with the exception of the APACHE II 1st quartile versus 4th quartile regression, which was only run on the All ARDS group. In addition to the 8 patients missing severity and mortality phenotype data, 2 patients were excluded from the regression because their phenotypes were missing at the moment of the analysis. These same statistics were repeated for the TaqMan genotyped patients (117 patients) and the total population (213 patients) when the candidate SNPs were chosen and genotyped further down the workflow.

Variant analysis and Functional Effect Predictions

Following the results of the genetic association study the data was further analyzed by Variant Analysis component of SVS V8.1.0. This software ranked the SNPs in order of likely importance based on location as well as make amino acid change predictions. This information was joined with predictions of protein functional effect

changes made by Sift and Provean (http://provean.jcvi.org/index.php) as well as Polyphen2 (http://genetics.bwh.harvard.edu/pph2/index.shtml).

Pathway Analysis

Ingenuity® Pathway Analysis™ Software (www.ingenuity.com/) from Ingenuity Systems was used to screen for SNPs which are likely to alter the function of relevant biologic pathways. To accomplish this I submitted a list of the genes that contain SNPs with χ^2 p-value of <0.01 in the 5 main comparisons of exome sequence data (All ARDS vs. 1000 Genomes, All sepsis vs. 1000 Genomes, All pneumonia vs. 1000 Genomes, African American ARDS vs. AFR 1000 Genomes, and Caucasian ARDS vs. EUR 1000 Genomes).

Assessment of Population Structure

Quantile-quantile plots (QQ plots) are used to visualize skew in data distributions between the transformed expected p-values or χ^2 values of a test and the actual p-values or χ^2 values of the test. In a QQ plot, the expected and actual values are expected to approximate a line of X=Y when plotted against each other[37]. Population structure differences between the ARDS exome sequenced patients and the 1000 Genomes healthy controls are visualized using QQ plots of the trend test χ^2 values and by computing the principal components using the Eigensoft PCA method in SVS for the All ARDS and All 1000 Genomes group, Caucasian ARDS and EUR 1000 Genomes subgroup, and African American ARDS and AFR 1000 Genomes subgroup[37].

Principal component analysis (PCA) is used to identify the components that contribute to genetic structure in a dataset. Principal components (PCs) that have a large contribution to the population structure can be corrected for in association tests to adjust

for genetic variability that isn't related to the exposure of interest. The data for the ARDS and the ancestry subgroups was filtered to include the SNPs that were considered to be in Hardy-Weinberg Equilibrium (HWE) in the 1000 Genomes project controls (HWE p-value >1x10^04), in linkage equilibrium in the ARDS+1000 Genomes populations (pairwise R^2>0.2 in a 200 SNP window moving in 25 SNP increments), and have a good call rate (>0.95 in the ARDS+1000 Genomes populations) per SNP. Sample call rate was not used in filtering due to the small population size. SNPs in chromosomes 1-X were used to determine principal components (eigenvalues). The maximum number of principal components is equal to the number of study participants, so N-1 principal components were calculated for the total population and the 2 racial subpopulations. The largest eigenvalues calculated using PCA were corrected for in a genotype trend test. The cutoff for correction was determined using Scree plots and the regression with cases-control status, where the PCs with the largest eigenvalues down to the point where the p-values were no longer significant ($p<0.001$) were corrected in the trend tests. Outlying samples were identified as being more than 6 SD away from the number of PCs chosen for correction in a genotype trend test and the PCs were recomputed 5 times[37]. PC corrected trend tests and PC and outlier corrected trend tests were run with correction for 9 components for the All ARDS and 1000 Genomes group, 6 components for the Caucasian ARDS and EA 1000 Genomes subgroup, and both 2 and 20 components for the AA ARDS and AA 1000 Genomes subgroup).

The genomic inflation factors (λ) were calculated for the trend test for the unfiltered exome cases vs. controls, filtered but pre-PCA data, and PCA corrected data to

quantify the amount of error that could be attributed to population stratification or sequencing error[37].

Genotyping of Selected Candidates

These data were compiled and used to determine the top SNP candidates using the criteria that they were present in all 4 individual ARDS populations (African American Sepsis, African American Pneumonia, Caucasian Sepsis, and Caucasian Pneumonia), had a χ^2 p-value < 0.01 in the 5 main ARDS populations, had a $\chi^2 < 2.95 \times 10^{-7}$ in at least 2 of the 5 main populations, and had some association with severity or outcome. Functional prediction information was also taken into account, and nonsynonymous SNPs were preferred because of their likelihood of causing direct functional effects.

Ultimately, three SNPs (rs78142040; rs9605146, also known as rs114989947; and rs3848719) in the additional 117 ARDS patient DNA samples from the NHLBI ARDSNet were genotyped using TaqMan human SNP genotyping assays on the ViiA 7 machine (http://www.invitrogen.com/) according to the supplier's instruction (Life Technologies).Using the sequence we provided, the first allele is detected using a VIC dye probe and the second allele is detected using a FAM dye probe. The sample DNA undergoes PCR amplification, using an output of the allele calls for each sample, the genotypes can be identified and analyzed. The genotype data from ARDS patients were compared with the genotype frequency calculated from the population genetics available for the SNPs from the 1000 Genomes Project[37]. Genotyping by the TaqMan assay is fast and targeted, making it the ideal tool for confirming SNP frequencies. Genotyping

accuracy was confirmed using the reference samples in-lab. Genotyping data from the additional samples was used to validate the associations observed in the ARDS vs. 1000 Genomes comparisons by calculating the association statistics described above.

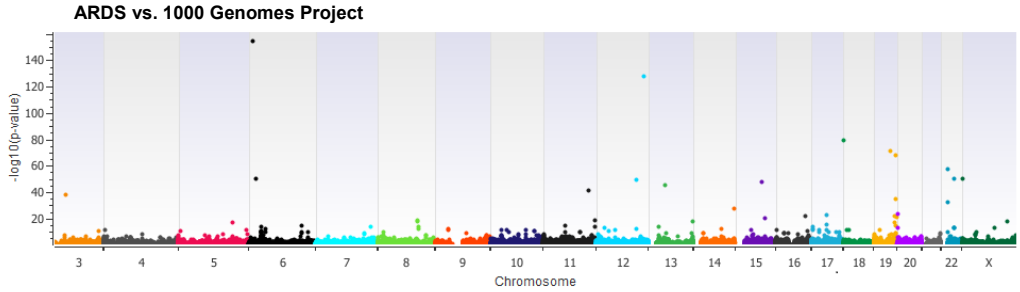
CHAPTER 5

RESULTS

Whole-Exome Sequencing and Identification of SNPs Associated With Susceptibility

The overall study population included 625 genome-sequenced controls from the 1000 Genomes Project and 213 ARDS patients (96 exome-sequenced patients and 117 patients genotyped for the 3 specific SNPs). These patients consist of 70 Caucasian and 26 African-Americans (table 1). In Caucasian patients, 37 cases were due to the initiating etiology of sepsis and 33 were due to pneumonia. In African American patients, 11 cases were due to the initiating etiology of sepsis and 15 were due to pneumonia. We detected 1,382,399 SNPs in 96 ARDS patients by exome-seq (table 4) and 490,015 SNPs per person on average. Among them, 169,376 SNPs matched records from the 625 healthy control subjects in the 1000 Genomes Project. From the 169,376 SNPs, there are 49,789 bi-allelic SNPs(50,190 total) in all ARDS patient subgroups based on race and initiating etiologies: Caucasian sepsis, Caucasian pneumonia, African-American sepsis and African American pneumonia (figure 3).

3.A



3.B

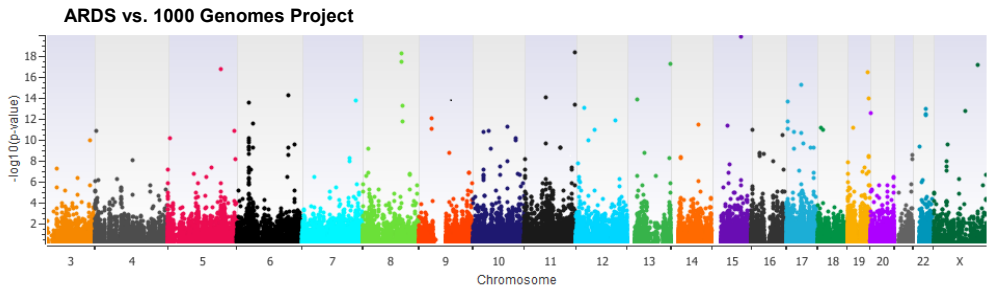


Figure 3: Manhattan Plot of ARDS Patients and 1000 Genomes Project Controls.

(A)A Manhattan plot of the whole exome sequence all ARDS cases vs. the 1000 Genomes Project controls created using SVS v8.1.0. This is a graphic representation of the chromosome location (x axis) vs. the –log10 (χ^2 p-value) of the allele frequencies. SNPs whose chi-square tests yield a smaller p-value fall higher on the log scale are more significant[37]. (B) The same Manhattan plot with a zoomed Y-axis.

Of our 1,382,399 ARDS SNPs, 608,723 were common between the sepsis and pneumonia cases (369,639 and 404,037 non-overlapping SNPs, respectively) and 442,235 SNPs were common between our African American cases and Caucasian cases (337,738 and 602,426 non-overlapping SNPs, respectively). 85.4% of the 1,213,023 ARDS SNPs not found in the 1000 Genomes Project Exome were assigned RS numbers, suggesting our data collection and processing is reliable.

By comparing the frequencies of the minor alleles in those SNPs in our 96 ARDS patients with 625 Caucasian and African-American healthy control subjects of the 1000 human genome project, we found that there are 3,867 differential SNPs (chi-square test

$p<0.01^{H}$). In Caucasians between ARDS patients and healthy controls, there are 788 differential SNPs ($p<0.01^{H}$). In African-Americans between ARDS patients and healthy controls, there are 1,203 differential SNPs ($p<0.01^{H}$). There are 208 common differential SNPs ($p<0.01$) in both Caucasian and African American patients or healthy controls. When we separately examined sepsis- or pneumonia-initiated ARDS, we found that 195 and 184 differential SNPs ($p<0.01$), respectively. Between them, there are 171 common differential SNPs ($p<0.01$). When the Bonferroni correction ($p<2.95x10^{-7}$) was applied, 89 SNPs remains significantly different. These SNPs are potentially novel coding SNPs associated with the ARDS susceptibility. Functional prediction information was also taken into account, and nonsynonymous SNPs were focused because of their likelihood of causing direct functional effects (table 4).

Table 4: Summary of the Filtering Applied to Candidate SNPs

Criteria for filtering	Number of remaining variants
SNPs detected in ARDS SNPs	1,382,399
matched in 1000 Genomes Project	169,376
common across 4 ARDS race and etiology groups*	50,190
χ^2 p-value <0.01 in top 5 ARDS groups**	171
χ^2 p-value <$2.95x10^{-7}$ in at least 2/5 top 5 ARDS groups	89 (in 76 genes)
coding variants	49 (in 38 genes)
nonsynonymous and splice variants	23
synonymous variants	26

*, African American with pneumonia, African American with Sepsis, Caucasian with pneumonia, Caucasian with sepsis subgroups. Out of the 50, 190 SNPs, 49,789 are bi-allelic. **, All ARDS, all pneumonia, all sepsis, all African American, all Caucasian groups

Of these SNPs 89 have a χ^2 p-value of <2.95x10^07 in at least 2 of the 5 main populations. 49 of these are expected to be in coding regions of the DNA. Of these, 22

cause nonsynonymous amino acid coding changes, 26 cause synonymous coding changes, 1is a splicing variant, and 1 was indeterminate (presumed to be intronic). Regressions with ARDS severity and mortality phenotypes were performed on 25 SNPs including the 22 nonsynonymous SNPs, 1 splice variant, 1 of the synonymous SNPs, and 1 unknown (presumed intronic) SNP.

Pathway Analysis

Ingenuity pathway analysis was conducted on the 76 genes that contain the top 89 SNPs to identify biologic pathways in which these genes function. The top canonical pathway is Nur77 signaling in T Lymphocytes and included 5 genes that contained associated SNPs, comprising 8% of the genes involved in the pathway ($p=1.47x10^{-6}$) (table 5).

Table 5: Top Canonical Pathways of Genes Containing SNPs Associated with ARDS.

Top Canonical Pathways	p-value*	Ratio**	Molecules
Nur77 Signaling in T Lymphocytes	1.47x10-6	0.08	CASP9,HLA-DQA1,HLA-DRB1,HLA-DQB1,SIN3A
Calcium-induced T Lymphocyte Apoptosis	6.93x10-5	0.06	HLA-DQA1,HLA-DRB1,HLA-DQB1,ITPR1
B Cell Development	1.97x10-4	0.08	HLA-DQA1,HLA-DRB1,HLA-DQB1
iCOS-iCOSL Signaling in T Helper Cells	5.69x10-4	0.03	HLA-DQA1,HLA-DRB1,HLA-DQB1,ITPR1
Graft-versus-Host Disease Signaling	5.82x10-4	0.03	CASP9,HLA-DQA1,HLA-DRB1,HLA-DQB1

Top canonical pathways as predicted from the 76 genes containing the 89 SNPs that were identified using 2 tests. Pathway predictions were done using the Core Analysis function of Ingenuity Pathway Analysis.
*, P-Value of <0.05 indicates a non-random association between the genes and pathway

**, Ratio of the number of genes in the dataset involved in the pathway to the total number of genes in the pathway

The top 5 canonical pathways additionally include calcium-induced T-lymphocyte signaling, B-cell development, iCOS-iCOSL signaling in T helper cells, and graft-versus-host disease signaling.

Validation of 3 SNPs

3 SNPs (rs78142040, rs9605146, and rs3848719) were genotyped in a larger population and statistics were re-computed (table 6).

Table 6: Overall association summary

	rs3848719	rs9605146	rs78142040
Susceptibility (cases vs. controls)			
Chi-squared p-value	9.39E-5	8.64E-71	6.68E-61
Odds Ratio (95% CI)[1]	1.61(1.27-2.05)	12.90(9.29-17.93)	87.76(32.00-240.65)
Severity (ventilator-free days/28 days)			
p-value	NS	NS	NS
Odds Ratio (95% CI)[2]	NS	NS	NS
Severity (APACHE II score)			
p-value	0.032	NS	0.061**
Odds Ratio (95% CI)[2]	0.55 (0.31-0.96)	NS	2.60(0.93-7.26)
Outcome (60-day mortality)			
p-value	NS*	NS	0.017
Odds Ratio (95% CI)[2]	NS	NS	2.04 (1.13-3.68)

NS, Not significant; * significantly associated in genotyped Caucasian, pneumonia, and Caucasian pneumonia subgroups; **, genotyped samples only; 1, Odds ratio for alternate allele (allelic test); 2, additive genotypic model

Rs78142040 has a major allele C and a minor allele T and is found on the X chromosome position X:2832771 in the Arylsulfatase D gene (*ARSD*). The SNP we have identified is significantly associated with susceptibility ($p<2.95x10^{-7}$) in the total 213 patient population (MAF=0.22) and the subgroups when compared with controls from the 1000 genomes project (MAF=0.003) (tables 7, 8).

Table 7: rs78142040 Statistics.

SNP		rs78142040	
position		X:2832771	
Gene (s)		ARSD	
	96 Exome	117 TaqMan	Total 213
χ^2 P-value*	1.14E-50	2.26E-62	6.68E-61
χ^2	224.13	277.51	271.06
Odds Ratio (Alternate Allele)	77.88	95.71	87.76
OR Lower Confidence Bound (Alt.)	27.32	34.27	32.00
OR Upper Confidence Bound (Alt.)	221.94	267.30	240.65
Call Rate	0.99	1.00	0.99
Call Rate (Cases)	0.94	1.00	0.97
HWE P-value (Cases)	1.77E-2	2.17E-2	1.18E-3
HWE P-value (Controls)	6.11E-138	6.11E-138	6.11E-138
HWE P-value	4.75E-2	6.56E-3	4.07E-1
Number of Distinct Alleles	2	2	2
Alternate Allele	T	T	T
Alternate Allele Frequency	0.03	0.04	0.06
Alt. Allele Freq. (Cases)	0.20	0.24	0.22
Alt. Allele Freq. (Controls)	0.003	0.003	0.003
Reference Allele	C	C	C
Reference Allele Frequency	0.97	0.96	0.94
Ref. Allele Freq. (Cases)	0.80	0.76	0.78
Ref. Allele Freq. (Controls)	0.997	0.997	0.997
Genotype AA Count			
AA (Cases)	0	2	2
AA (Controls)	2	2	2
Genotype Ar Count	36	51	87
Ar (Cases)	36	51	87
Ar (Controls)	0	0	0
Genotype rr	677	687	741
rr (Cases)	54	64	118
rr (Controls)	623	623	623
Alternate Allele A Count	40	59	95
A (Cases)	36	55	91
A (Controls)	4	4	4
Reference Allele r	1390	1425	1569
r (Cases)	144	179	323
r (Controls)	1246	1246	1246

A summary of the SNP rs78142040 in the exome sequenced ARDS, TaqMan genotyped ARDS patients, and total ARDS patients, where the controls are

1000 Genomes Project participants. "A" represents the alternate allele and "r" represents the reference allele.
*, Chi-square tests were run on SNPs that were in both the controls and the cases.

Table 8: Chi-Square Test P-Values for rs78142040 in Patients and Controls.

SNP	rs78142040		
Gene		ARSD	
	96 Exome	117 TaqMan	Total 213
ARDS χ^2 P-value			
Caucasian χ^2 P-value	1.14E-50	2.16E-62	6.68E-61
African American χ^2 P-value	8.59E-32	8.65E-35	4.30E-34
Pneumonia χ^2 P-value	2.14E-22	1.18E-26	7.97E-28
Sepsis χ^2 P-value	4.10E-33	1.55E-42	1.71E-42
African American Sepsis χ^2 P-value	4.20E-59	1.81E-72	3.15E-72
African American Pneumonia χ^2 P-value	3.30E-27	1.46E-22	1.31E-30
Caucasian Sepsis χ^2 P-value	1.73E-11	1.06E-19	2.50E-18
Caucasian Pneumonia χ^2 P-value	1.88E-37	1.60E-46	9.85E-43

The allelic chi-square test p-values for the exome sequenced ARDS patients ARDS patients and subgroups compared with the 1000 Genomes Project participants and subgroups, TaqMan genotyped ARDS patients and subgroups compared with the 1000 Genomes Project participants and subgroups , and the total ARDS patient population and subgroups compared with the 1000 Genomes Project participants and subgroups. P-values were considered to be significant if they were smaller than the Bonferroni corrected p-value of 2.95x10^-7.

Rs78142040 approaches association with APACHE II score when the score quartiles are compared for the genotyped ARDS patients (p=0.061, OR=2.603, 95% CI=0.933-7.260) (table 9).

Table 9: Logistic Regression of Genotype and APACHE II Score by Quartile.

SNP	rs3848719	rs9605146	rs78142040
Gene (s)	ZNF335	XKR3	ARSD
Exome samples			
p-value	0.10	0.15	na
Odds Ratio	0.50	2.00	na
OR Lower Conf. Bound	0.22	0.76	na
OR Upper Conf. Bound	1.16	5.26	na
TaqMan samples			
p-value	0.30	0.36	0.06
Odds Ratio	0.68	0.70	2.60
OR Lower Conf. Bound	0.33	0.33	0.93
OR Upper Conf. Bound	1.42	1.50	7.26
All samples			
p-value	0.03	0.78	0.29
Odds Ratio	0.55	1.08	1.51
OR Lower Conf. Bound	0.31	0.62	0.70
OR Upper Conf. Bound	0.96	1.89	3.25

The APACHE II scores are split into quartiles and the 1st and 4th quartiles are used in a logistic regression against genotype using an additive model in the ARDS exome samples, TaqMan genotyped samples, and total ARDS samples. Regressions were also run on the stratified sub-populations of the ARDS patients. Associations were considered to be significant if $P<0.05$.

Rs78142040 is associated with the 60-day mortality in the genotyped ARDS group (p=0.034, OR=2.245, 95%CI 1.053-4.786), the total ARDS population (p=0.017, OR=2.039, 95%CI 1.130-3.681) and the all-sepsis subgroups of the genotyped patients and total ARDS population (p=0.029, OR=2.981, 95%CI 1.072-8.289 and p=0.035, OR=2.276, 95%CI 1.044-4.961, respectively) (table 10).

Table 10: Logistic Regression of Genotype and 60-Day Mortality.

SNP		rs78142040	
Gene (s)		ARSD	
	96 Exome	117 TaqMan	Total 213
ARDS p-value	0.2630728	0.0339629	0.0173745
Pneumonia p-value	0.6602961	0.8575796	0.6581823
Sepsis p-value	0.5246962	0.028991	0.0348283
Caucasian p-value	0.2222798	0.9495742	0.408666
African American p-value	0.8547592	na	0.3606608
African American Pneumonia p-value	0.9282386	na	0.5363197
African American Sepsis p-value	0.4870023	na	0.596097
Caucasian Sepsis p-value	0.373584	0.8254455	0.4565353
Caucasian Pneumonia p-value	0.8171329	0.6294941	0.7551327
SNP		rs9605146	
Gene (s)		XKR3	
	96 Exome	117 TaqMan	Total 213
ARDS p-value	0.199116	0.1613938	0.8628899
Pneumonia p-value	0.0799541	0.1352877	0.9207455
Sepsis p-value	0.6639858	0.4980102	0.8681921
Caucasian p-value	0.676358	0.0857127	0.3488488
African American p-value	0.2706209	0.4411772	0.6577734
African American Pneumonia p-value	0.3251523	0.4184522	0.6835628
African American Sepsis p-value	0.553167	0.7626348	0.7754479
Caucasian Sepsis p-value	0.9381507	0.258087	0.527152
Caucasian Pneumonia p-value	0.4336663	0.1939841	0.4733411
SNP		rs3848719	
Gene (s)		ZNF335	
	96 Exome	117 TaqMan	Total 213
ARDS p-value	0.2983921	0.3583787	0.9136127
Pneumonia p-value	0.1851452	0.0323794	0.380281
Sepsis p-value	0.5992432	0.455294	0.3463909
Caucasian p-value	0.6700216	0.0122873	0.2039698
African American p-value	0.9736738	na	0.7412062
African American Pneumonia p-value	0.9149283	na	0.5810341
African American Sepsis p-value	0.8580141	na	0.8671714
Caucasian Sepsis p-value	0.919974	0.306832	0.5617649
Caucasian Pneumonia p-value	0.2672277	0.011951	0.2015335

The p-values of the logistic regression of 60-day mortality against genotype using an additive model in the ARDS exome samples, TaqMan genotyped samples, and total ARDS samples. Regressions were also run on the stratified sub-populations of the ARDS patients.

The SNP is in Hardy-Weinberg Equilibrium (HWE $p>1x10^{-4}$) in the EUR 1000genomes controls and in the ARDS population and subgroups. The SNP was

determined to lie within a histone mark of intron 6 using the UCSC Genome Browser and could potentially play a role in regulation of expression (http://genome.ucsc.edu/).

rs9605146 has a major allele G and minor allele A. It is a nonsynonymous SNP found within exon 4 of chromosome 22 (22: 17265194) in the "XK, Kell blood group complex subunit-related family, member 3" gene and causes a predicted amino acid change from proline to leucine (*XKR3*). This amino acid change has a deleterious effect predicted by a Provean score of -5.494, where a score of maximum -2.5 is considered to be deleterious (table 11).

Table 11: Predicted Effects of the 4 SNPs on Amino Acid Coding.

dbSNP_ID	rs3848719	rs9605146	rs78142040
Gene (s)	ZNF335	XKR3	ARSD
SNP Classification	Coding	Coding	Coding
Coding Classification	Synonymous	Nonsyn SNV	Intronic
Reference amino acid	S	P	NK
Alternate amino acid	S	L	NK**
TYPE	Synonymous	Single AA Change	NK
Provean prediction (cutoff=-25)	Neutral	Deleterious	NK
Sift prediction (cutoff=0.05)	Tolerated	Tolerated	NK

Sift, Provean, and Polyphen2 were used to predict the functional changes expected to be caused by the SNPs. *, PolyPhen2; **, Not Known

The SNP is in HWE ($p>1x10^{-4}$) in the EUR 1000 Genomes Project as well as in the ARDS population and subgroups. Rs9605146 is associated with disease susceptibility (p<2.95x10^-7) in the total ARDS population (MAF=0.39) and subgroups when compared with the 1000 genomes controls (MAF=0.05) (tables 12, 13).

Table 12: rs9605146 Statistics.

SNP	rs9605146		
position	22:17265194		
Gene (s)	XKR3		
	96 Exome	117 TaqMan	Total 213
χ^2 P-value*	8.37E-58	2.38E-51	8.64E-71
χ^2	256.84	227.25	316.44
Odds Ratio (Alternate Allele)	14.52	11.72	12.91
OR Lower Confidence Bound (Alt.)	9.80	8.06	9.29
OR Upper Confidence Bound (Alt.)	21.51	17.05	17.93
Call Rate	1.00	1.00	1.00
Call Rate (Cases)	0.97	1.00	0.99
HWE P-value (Cases)	3.78E-1	8.22E-1	4.22E-1
HWE P-value (Controls)	5.13E-15	5.13E-15	5.13E-15
HWE P-value	2.35E-21	7.83E-16	4.34E-19
Number of Distinct Alleles	2	2	2
Alternate Allele	A	A	A
Alternate Allele Frequency	0.09	0.10	0.13
Alt. Allele Freq. (Cases)	0.41	0.36	0.39
Alt. Allele Freq. (Controls)	0.05	0.05	0.05
Reference Allele	G	G	G
Reference Allele Frequency	0.91	0.90	0.87
Ref. Allele Freq. (Cases)	0.59	0.64	0.61
Ref. Allele Freq. (Controls)	0.95	0.95	0.95
Genotype AA Count	28	26	44
AA (Cases)	18	16	34
AA (Controls)	10	10	10
Genotype Ar Count	79	91	132
Ar (Cases)	41	53	94
Ar (Controls)	38	38	38
Genotype rr	611	625	659
rr (Cases)	34	48	82
rr (Controls)	577	577	577
Alternate Allele A Count	135	143	220
A (Cases)	77	85	162
A (Controls)	58	58	58
Reference Allele r	1301	1341	1450
r (Cases)	109	149	258
r (Controls)	1192	1192	1192

A summary of the SNP rs9605146 in the exome sequenced ARDS, TaqMan genotyped ARDS patients, and total ARDS patients, where the controls are 1000 Genomes Project participants. “A” represents the alternate allele and “r” represents the reference allele.

*, Chi-square tests were run on SNPs that were in both the controls and the cases.

Associations with severity and outcome are not statistically significant. The SNP also approaches significant association with 60-day mortality in the exome-sequenced patients with pneumonia (p=0.080).

Table 13: Chi-Square Test P-Values for rs9605146 in Patients and Controls.

SNP	rs9605146		
Gene		XKR3	
	96 Exome	117 TaqMan	Total 213
ARDS χ^2 P-value	8.37E-58	2.37E-51	8.64E-71
Caucasian χ^2 P-value	5.26E-34	7.31E-37	1.19E-44
African American χ^2 P-value	1.53E-28	3.81E-13	3.05E-30
Pneumonia χ^2 P-value	2.50E-42	9.95E-33	4.55E-53
Sepsis χ^2 P-value	9.64E-38	1.12E-39	2.46E-55
African American Sepsis χ^2 P-value	1.72E-16	7.96E-8	1.02E-18
African American Pneumonia χ^2 P-value	1.66E-21	6.34E-9	7.32E-24
Caucasian Sepsis χ^2 P-value	4.71E-25	6.52E-26	4.94E-35
Caucasian Pneumonia χ^2 P-value	3.91E-24	2.82E-28	6.46E-36

The allelic chi-square test p-values for the exome sequenced ARDS patients ARDS patients and subgroups compared with the 1000 Genomes Project participants and subgroups, TaqMan genotyped ARDS patients and subgroups compared with the 1000 Genomes Project participants and subgroups , and the total ARDS patient population and subgroups compared with the 1000 Genomes Project participants and subgroups. P-values were considered to be significant if they were smaller than the Bonferroni corrected p-value of 2.95x10^-7.

rs3848719 is a synonymous SNP in the 5th exon of the Zinc-Finger/Leucine-Zipper Co-Transducer NIF1 gene in chromosome 20 (location 20: 44596545) (*ZNF335*). The SNP rs3848719 has a major allele of G and a minor allele A. The SNP approaches association with susceptibility ($p=9.39x10^{-5}$) in the total ARDS population (MAF=0.39) and in the exome ARDS subgroup when compared with the 1000 genomes controls (MAF=0.29) (table 14, table 15).

Table 14: rs3848719 Statistics.

SNP	Rs3848719		
position	20:44596545		
Gene (s)	ZNF335		
	96 Exome	117 TaqMan	Total 213
χ^2 P-value*	2.68E-6	9.26E-2	9.39E-5
χ^2	22.03	2.83	15.25
Odds Ratio (Alternate Allele)	2.28	1.29	1.61
OR Lower Confidence Bound (Alt.)	1.61	0.96	1.27
OR Upper Confidence Bound (Alt.)	3.24	1.73	2.05
Call Rate	0.97	1.00	0.97
Call Rate (Cases)	0.74	1.00	0.88
HWE P-value (Cases)	3.81E-6	5.86E-1	2.86E-4
HWE P-value (Controls)	4.76E-3	4.76E-3	4.76E-3
HWE P-value	2.30E-6	4.63E-3	4.90E-6
Number of Distinct Alleles	2	2	2
Alternate Allele	A	A	A
Alternate Allele Frequency	0.31	0.30	0.31
Alt. Allele Freq. (Cases)	0.48	0.34	0.39
Alt. Allele Freq. (Controls)	0.29	0.29	0.29
Reference Allele	G	G	G
Reference Allele Frequency	0.69	0.70	0.69
Ref. Allele Freq. (Cases)	0.52	0.66	0.61
Ref. Allele Freq. (Controls)	0.71	0.71	0.71
Genotype AA Count	92	81	107
AA (Cases)	26	15	41
AA (Controls)	66	66	66
Genotype Ar Count	243	277	293
Ar (Cases)	16	50	66
Ar (Controls)	277	277	277
Genotype rr	361	384	413
rr (Cases)	29	52	81
rr (Controls)	332	332	332
Alternate Allele A Count	427	439	507
A (Cases)	68	80	148
A (Controls)	359	359	359
Reference Allele r	965	1045	1119
r (Cases)	74	154	228
r (Controls)	891	891	891

A summary of the SNP rs3848719 in the exome sequenced ARDS, TaqMan genotyped ARDS patients, and total ARDS patients, where the controls are 1000 Genomes Project participants. “A” represents the alternate allele and “r” represents the reference allele.
*, Chi-square tests were run on SNPs that were in both the controls and the cases.

rs3848719 is also associated with is associated with APACHE II score when the score quartiles are compared for total ARDS patients (p=0.032, OR=0.549,

95%CI=0.313-0.96). The SNP is associated with 60-day mortality in the TaqMan genotyped Caucasian ARDS (p=0.012, OR=2.753, 95% CI=1.196-6.336), TaqMan genotyped pneumonia (p=0.032, OR=2.511, 95% CI=1.053-5.984), and TaqMan genotyped Caucasians with pneumonia (p=0.012, OR=4.045, 95% CI=1.219-13.433). The SNP is in HWE ($p>1x10^{-4}$) in the Overall, EUR and AFR 1000 Genomes Project groups as well as in the overall ARDS population and subgroups with the exception of the combined African American ARDS cases.

Table 15: Chi-Square Test P-Values for Rs3848719 in Patients and Controls.

SNP	rs3848719		
Gene		ZNF335	
	96 Exome	117 TaqMan	Total 213
ARDS χ^2 P-value	2.68E-6	9.26E-2	9.39E-5
Caucasian χ^2 P-value	4.24E-3	4.22E-1	2.78E-1
African American χ^2 P-value	1.32E-4	6.73E-1	2.68E-3
Pneumonia χ^2 P-value	1.26E-3	1.92E-1	4.03E-3
Sepsis χ^2 P-value	2.72E-4	2.37E-1	2.26E-3
African American Sepsis χ^2 P-value	2.79E-3	2.48E-1	1.55E-1
African American Pneumonia χ^2 P-value	6.12E-3	7.29E-2	1.87E-3
Caucasian Sepsis χ^2 P-value	3.76E-2	4.58E-1	4.26E-1
Caucasian Pneumonia χ^2 P-value	3.57E-2	6.47E-1	3.99E-1

The allelic chi-square test p-values for the exome sequenced ARDS patients ARDS patients and subgroups compared with the 1000 Genomes Project participants and subgroups, TaqMan genotyped ARDS patients and subgroups compared with the 1000 Genomes Project participants and subgroups , and the total ARDS patient population and subgroups compared with the 1000 Genomes Project participants and subgroups. P-values were considered to be significant if they were smaller than the Bonferroni corrected p-value of 2.95x10^-7.

Two of the 3 SNPs selected for further study (rs78142040 and rs3848719) are significantly associated with outcome as measured by 60-day mortality in at least one subgroup of the ARDS patients (table 10).

Of the 3 SNPs selected for further study, one SNP (rs3848719) is significantly associated with a reduction in severity as measured by a comparison of the highest and

lowest APACHE II score quartiles ($p<0.05$) and one SNP (rs8142040) approaches significant association with APACHE II score quartile (table 9).

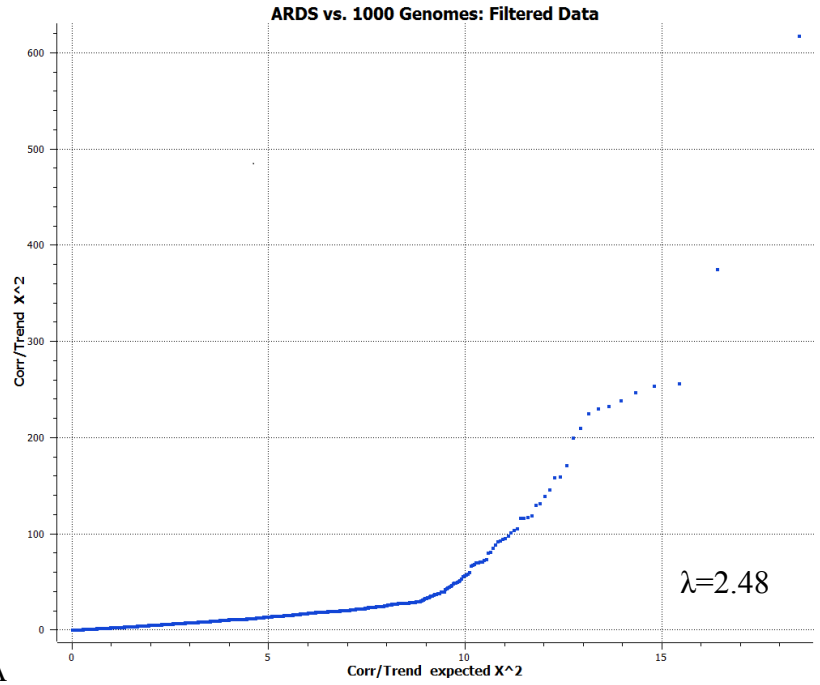
SNPs rs9605146 is predicted by Provean to cause deleterious changes in amino acid coding and predicted to cause neutral changes by Sift and Polyphen2. SNP rs3848719 is predicted to have a neutral effect by Sift and Provean, and is not found using Polyphen2. SNP rs78142040 is assumed to be in a non-protein coding region of DNA.

Assessment of Population Structure

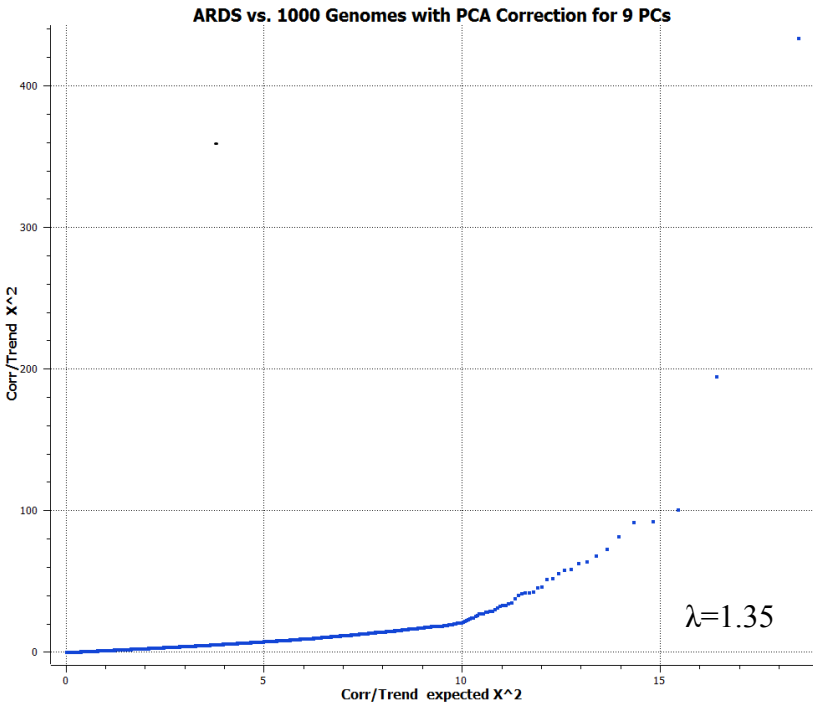
Genotype trend tests were conducted before filtering, after filtering, after PCA, and after PCA with outlier sample removal to visualize the skew in the expected vs. actual χ^2 values of the genotype trend test with case/control status across the populations. Genomic inflation factors (lambda) were calculated for the trend tests as a metric of the amount of genomic inflation observed in the populations. Lambda values >1 indicates the presence of population structure in the combined cases and controls possibly due to divergent ancestry or genotyping error.

For the All ARDS and 1000 Genomes group, 168,309 SNPs with 2 alleles were matched by location between the cases and controls (of 169,376 total SNPs). The trend test lambda=5.25, indicating the need for data filtering and correction for population structure. When the data is filtered using the earlier described method (29,617 SNPs remaining) the lambda decreases to 2.48. Of the principal components, PCs 1-6 and 9 were significantly associated ($p<1x10^{-3}$) with case-control status. After correction for 9 principal components, the lambda decreases to 1.35. When outlier samples are removed

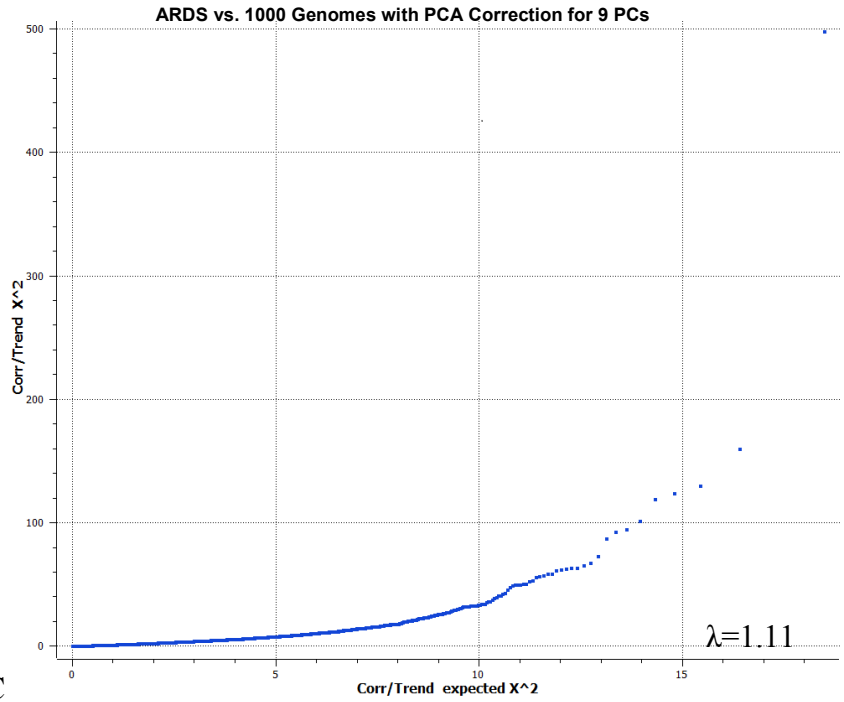
from the same 9 components, lambda decreases to 1.11. When outlier removal is included in the PCA corrected trend test 45 samples are removed (all Caucasian cases) (figure 4).



4.A



4.B



4.C

Figure 4: Quantile-Quantile Plots of Genotypic Trend Test χ^2 Values for the ARDS and 1000 Genomes Project Population.
The straight line on each plot represents y=x. A) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data is filtered on HWE, LD, and SNP call rate but not PCA corrected. B) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data have been filtered and corrected for 9 PCs. C) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data have been filtered and corrected for 9 PCs and undergone sample outlier removal.

The Caucasian ARDS patients and EUR have 110,862 SNPs with 2 alleles in the combined population. The trend test of the unfiltered data has a lambda value of 2.07. When the data is filtered (28, 681 SNPs remaining) the lambda decreases to 1.3. Of the principal components calculated, 1, 3, 5, and 6 were significantly associated with case-control status ($p<1x10^{-3}$). 6 PCs were identified for correction and the lambda of the

PCA corrected trend test is 1.20. After PCA correction and outlier removal of 21 samples (all cases) the lambda is 1.12.

In the unfiltered data there are 129, 612 SNPs that are present in the African American ARDS patients and the AFR were used to assess the population structure. The trend test of the unfiltered data had a lambda of 2.11 and after filtering (48,852 SNPs remaining) 1.70 (48, 852 SNPs remaining). Principal components 1, 2, 19, and 20 were significantly associated with case control status ($p<1x10^{-3}$), so two versions of the PCA correction were done- the first, for 2 PCs and the 2nd for 20 PCs. A trend test was first corrected for 2 PCs, with a lambda of 1.54. No outliers were identified for removal from the first 2 PCs. When a trend test is PCA corrected for 20 PCs the lambda value is 1.06. When a trend test is PCA corrected with outlier removal on 20 PCs 51 samples were removed including 16 cases and 35 controls, and the lambda is 1.00.

To provide reference for PC adjusted genotypic trend tests, the trend test p-values for the 3 genotyped SNPs are provided for the uncorrected tests as well as the p-values from the PCA corrected tests that passed the pre-PCA data filtering. Rs9605146 and rs78142040 passed filtering in the Caucasian ARDS and EA 1000 Genomes sub population. By looking in the changes in p-values between uncorrected, PCA corrected, and PCA corrected with outlier removal, it is seen that the significance in genotype association with case-control status decreases when the association test is corrected for PCs, however the trend remains (table 16).

Table 16: A Summary of the 4 Genotyped SNPs and the Effect of the PCA and Outlier Removal on Their Genotypic Trend Test P-Values.

		rs3848719	rs9605146	rs78142040
ARDS and all 1000 Genomes	Passed filtering?	no	no	no
no PCA	trend p-value	1.57E-5	4.27E-43	3.11E-47
corrected for 10 PCs	trend p-value	NA	NA	NA
corrected for 10 PCs with outlier removal on 10 PCs	trend p	NA	NA	NA
AA ARDS and AA 1000 Genomes	passed filtering?	no	no	no
no PCA	trend p-value	2.89E-04	5.38E-20	1.03E-18
corrected for 2 PCs	trend p-value	NA	NA	NA
corrected for 2 PCs with outlier removal on 2 PCs	trend p-value	NA	NA	NA
corrected for 20 PCs	trend p-value	NA	NA	NA
corrected for 20 PCs with outlier removal on 20 PCs	trend p-value	NA	NA	NA
EA ARDS and EA 1000 Genomes	passed filtering?	no	yes	yes
no PCA	trend p-value	4.76E-03	5.67E-27	1.52E-32
corrected for 26PCs	trend p-value	NA	2.62E-13	1.19E-11
corrected for 6 PCs with outlier removal on 6 PCs	trend p-value	NA	2.17E-19	2.62E-04

CHAPTER 6

DISCUSSION

The ARDS population has similar ages across the subgroups. The 60-day mortality is higher in the sepsis cases than pneumonia and higher in the African American cases than Caucasians, which correlates with previously reported literature[37]. The same is observed for the APACHE II scores. Ventilator-free days were lower in sepsis cases than pneumonia cases as well as lower in African American cases than Caucasian.

Whole-Exome Sequencing and Identification of SNPs Associated With Susceptibility

Whole-exome sequencing was previously performed in 96 ARDS patients from the ARDSnet with the intent of identifying coding SNPs whose minor allele frequencies are significantly different in ARDS than those of healthy controls as well as identifying those novel SNPs who may be predictors of ARDS severity and outcome. In the overall ARDS population 1,382,399 SNPs were detected by exome-seq (Table 2) and 490,015 SNPs per person on average compared to 714, 074 SNPs per person from 625 healthy control subjects in the 1000 human genome project. Among them, 169,376 SNPs overlapped between cases and controls. The majority of non-overlapping SNPs in ARDS patients may represent ARDS specific SNPs barring the individual variability, sequencing error and data analysis discrepancy. From 169,376 SNPs, there are 49,789 bi-allelic SNPs in all ARDS patient subgroups based on race and initiating etiologies: Caucasian sepsis, Caucasian pneumonia,

African-American sepsis and African American pneumonia. These SNPs may represent sepsis or pneumonia etiology specific SNPs of ARDS. The reason why we initially focused on the identification of novel coding SNPs associated with ARDS in sepsis and pneumonia origins was that in the original ARDS patient population, sepsis and pneumonia etiologies accounted for most cases.

Validation of selected three SNPs

Among three selectively validated SNPs(rs78142040, rs9605146 and rs3848719) in additional 117 ARDS patients, rs78142040 in *ARSD* is associated with increased ARDS susceptibility in the overall ARDS population (213 patients) as well as all racial and comorbidity subpopulations. It approaches significant association with an increase in APACHEII score ($p=0.061$) when samples in the highest and lowest score quartiles are compared in ARDS patients. rs78142040 is significantly associated ($p<0.05$) with an increase in 60-day mortality in the total ARDS population ($p=0.017$, OR=2.039, 95%CI 1.130-3.681). The lack of HWE in the 1000 Genomes Project population is contributed to by a lack of heterozygous samples (and very few minor alleles). Associations with ventilator-free days and APACHE II (not in quartiles) score were not observed. The molecular mechanism underlying these associations are presently unknown. The *ARSD* gene encodes a sulfatase that is associated with bone and cartilage development and has been previously identified as having involvement in sphingolipid metabolism (involved in signal transmission and cellular recognition) and as a potential biomarker for chronic lymphocytic leukemia[21]. ARSD protein isoforms have a highly conserved catalytic

peptide domain when compared with other arylsulfatases[26,27]. ARSD is widely expressed and is suspected to play a role in housekeeping or multiple other processes, however specific substrates have not been identified[28]. It was reported that there were changes in activities of lung lysosomal enzymes including sulfatase during ARDS[38]. The SNP lies within a histone mark in intron 6 and possibly affects transcription thought this theory requires validation. It may be interesting to explore whether rs78142040 causes the differential expression of *ARSD*, thus sulfatase activity, which may link its role in the pathogenesis of ARDS. While the SNP of interest in this gene is intronic, the strength of the association within the cases as well as in comparison with the healthy controls indicates its potential application as a biomarker.

rs9605146 in *XKR3* gene is associated with increased susceptibility in the ARDS population and all subgroups. Associations with severity and mortality were not observed. XKR3 is a member of the XK/Kell complex in the Kell blood group system. XKR3 is a homolog of XK, which is a putative membrane transporter. XK is associated with McLeod syndrome (characterized by late-onset abnormalities in the central nervous system and neuromuscular system) and red cell acanthocytosis[22]. *XKR3* has previously been indicated as a potential biomarker for blood transfusion compatibility[29]. While it is currently unknown what underlies the association of rs9605146 with susceptibility in the ARDS, it causes a deleterious amino acid coding change from proline to leucine in *XKR3* as predicted by Provean (score=-5.494), make rs9605146 a legitimate candidate for further study of its role in the pathogenesis of ARDS.

rs3848719 in *ZNF335* approaches association with increased susceptibility in the exome-sequenced and overall ARDS population. The SNP is associated with a reduction

in APACHE II score when the highest and lowest score quartiles are compared in the total ARDS population (p=0.032, OR=0.55, 95%CI 1.27-2.05). The SNP is also associated with increased 60-day mortality in Caucasian and pneumonia groups in the genotyped samples, suggesting that this SNP could be a good predictor of outcome in Caucasian populations and in patients with pneumonia, however theses associations were not observed in the overall population. . It is a synonymous SNP in the 5th exon of the Zinc-Finger/Leucine-Zipper Co-Transducer NIF1 gene (*ZNF335*). ZNF335 is involved in neural progenitor cell proliferation and self-renewal as a component of the vertebrate-specific, trithorax H3K4-methylation complex. *ZNF335* is associated with microcephaly (a neurodevelopmental disorder), small somatic size and neonatal death. The gene is essential as homozygous knockout mouse models have a lethal effect [23,24]. The role of the gene in cellular differentiation and gene expression could implicate an effect on the fundamental physiology and neural signaling in the lungs, contributing to the pathogenesis of ARDS.

All SNPs are in Hardy-Weinberg Equilibrium (HWE $p>1x10^{-4}$) in the total ARDS cases. Rs3848719 is in HWE in the EUR + AFR 1000 Genomes and AFR 1000 Genomes subgroup and all 3 are in HWE in the EUR 1000 Genomes subgroup (supplementary table 1).

We selectively genotyped and validated three SNPs(rs78142040, rs9605146 and rs3848719) in additional 117 ARDS patients using the TaqMan genotype assay and performed in depth association analyses of these SNPs with the susceptibility, severity and outcome to ARDS in a combined 213 ARDS patients (96 by exome-seq +117 by TaqMan=213). These validations lend a solid support to the validity and prowess of

novel ARDS associated SNP identifications by exome-seq. This study provides a rich resource for further experimentation and replication to develop and establish new genetic biomarkers and therapeutic targets to ARDS. The further replicates of all these SNPs in different and larger patient populations and mechanistic explorations of these SNPs in cell culture and animal models may lead to the development of these novel SNPs as new diagnostic biomarkers and therapeutic targets to ARDS in the near future.

Limitations

Although we applied the Bonferroni correction ($p<2.95x10^{-7}$) and several SNP filtering steps during our data analysis as well as validations of three selected candidate SNPs, our data come with some potential limitations. Firstly, we only performed exome-sequencing of 96 ARDS patient samples. Although we would argue that this is a very reasonable sample size considering the high cost of exome-sequencing (even the exome-seq cost per sample is cheaper than whole genome-seq per sample), the sample size is not large. Our 89 SNPs which are strongly associated with susceptibility are all present in an age and race matched 48 sample control set (coriell.org), which will be used to validate our findings in further studies (117,35 out of the 169,376 SNPs which are in the ARDS cases and 1000 Genomes Project are found in this control set). Confirmation of our findings in additional larger patient populations is warranted. Second, during analysis of SNP associations with ARDS susceptibility, we used the healthy control subjects from the 1000 genome project. Both ARDS patients and healthy control subjects aren't derived from the same population. We have applied HWE and LD filtering, PCA analysis and Q-Q plot visualization as well as race specific comparison to filter the identified SNPs to

better quantify the amount of structure, though these steps may not totally correct the population admixtures.

It is expected that population admixture may be responsible for some of the observed population structure in our dataset, however we can't determine if genomic inflation is uniform across the genome. Lambda values improved with the pre-PCA filtering, which suggests that this method is only effective for quantifying error in SNPs that are inherited in LE and HWE. It is possible that SNPs that are genuinely associated with a disease may not follow the expected linkage equilibrium (LE) or Hardy-Weinberg Equilibrium (HWE) distributions, and these SNPs are typically removed from the dataset prior to principal components analysis (PCA) because they can bias corrections made for population stratification SNPs that don't meet the criteria for PCA are still interesting and viable therapeutic targets. Adjusting the trend tests for the largest principal components caused a decrease in Lambda confirming a reduction in population structure with correction, and again when samples that were outliers from the top principal components were removed from analysis. In all principal component adjustments the resulting qq plots better approximate the expected x=y line for correlation $\chi 2$ vs. expected trend $\chi 2$ after adjustment for the PCs with the exception of the African American ARDS and AFR 1000 Genomes Project population where correcting for 20 PCs with outlier removal overcorrected the data. Since the number of cases is small the outlier removal step of the PCA analysis is not practical for drawing for subsequent statistical analyses and the identification of candidate variants in this population.

Replication of our findings in larger and different populations may strengthen and develop here identified candidate SNPs as true genetic biomarkers of ARDS. Mechanistic

explorations of these SNPs in cell culture and animal models may lead to the clinical application of these novel SNPs as new diagnostic biomarkers and therapeutic targets to ARDS in the near future. With a better understanding of the genetic components of ARDS it will be possible to create novel gene therapies that could significantly reduce the mortality and morbidity of ARDS.

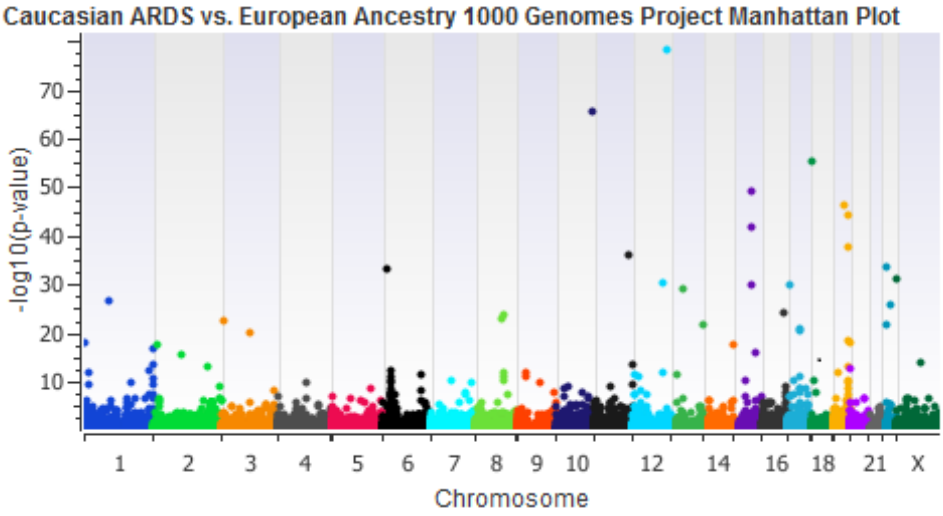
CHAPTER 7

CONCLUSIONS

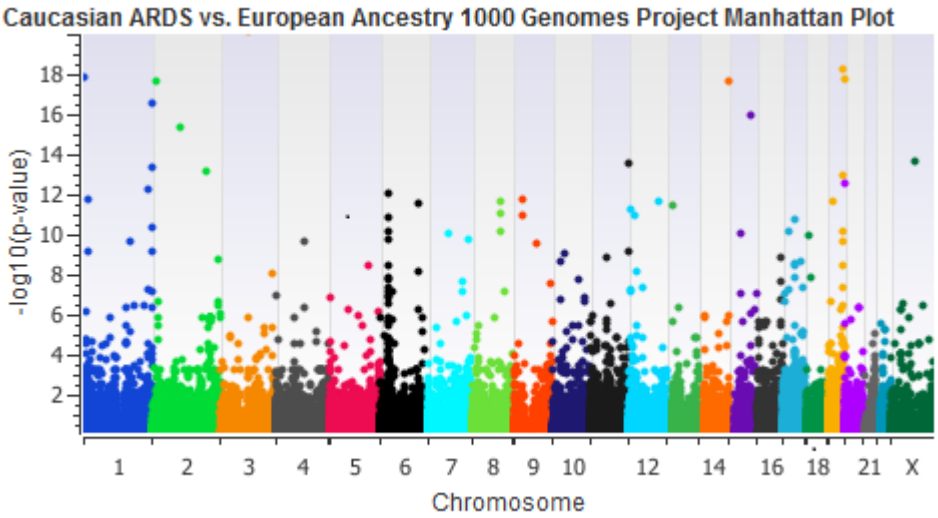
The primary focus of this study is to identify genetic components of ARDS susceptibility, severity and outcome. Whole exome-sequencing has identified a number of potential ARDS association SNPs. This study selectively validated 3 SNPs that are associated with the susceptibility (rs78142040 and rs9605146), severity (rs3848719) and outcome (rs78142040 and rs3848719) of ARDS. More validations in a larger patient population with matched controls as well as further investigation of the underlying molecular mechanisms are needed. Further research on the topic can include biological validation in cellular and mouse models as well as computational modeling validation. Possible implications of the study could be to provide novel diagnostic and therapeutic targets for ARDS. With a better understanding of the genetic components of ARDS it could be possible to develop new or better therapeutic modalities that could significantly reduce mortality and morbidity of ARDS in the long run.

APPENDIX A

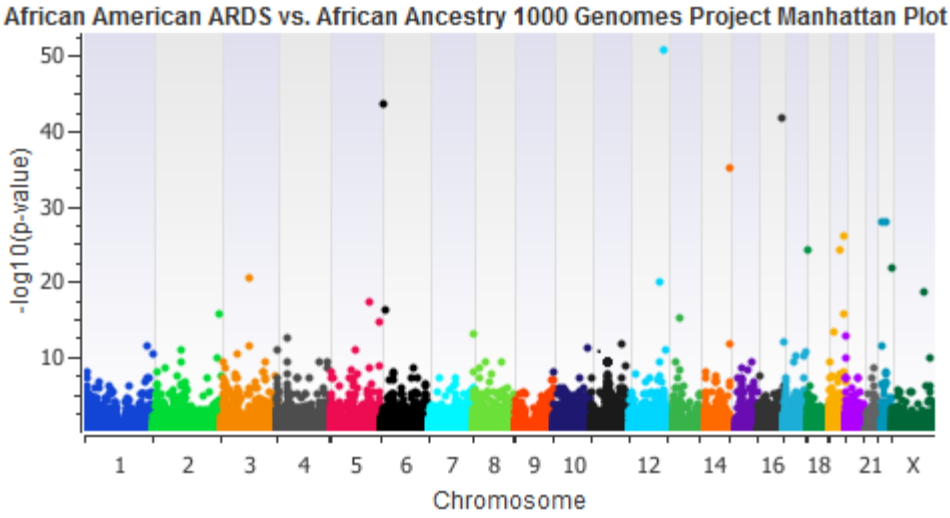
SUPPLEMENTARY TABLES AND FIGURES



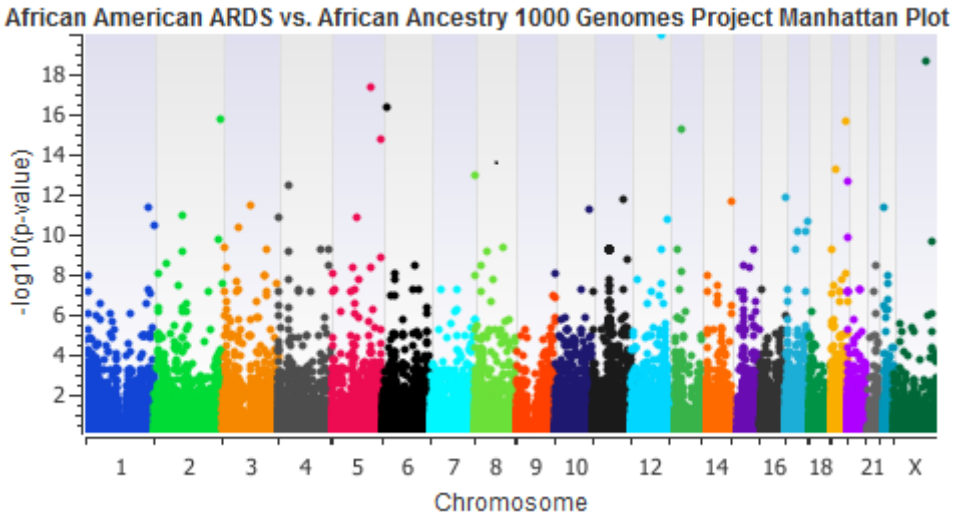
S1.A



S1.B



S1.C



S1.D

Supplementary figure 1. Manhattan Plots of the EUR Cases and Controls and the AFR Cases and Controls, Respectively.
(A)A Manhattan plot of –log10 (χ^2 p-value) by the chromosomal location of the Caucasian ARDS compared with the EUR 1000 Genomes participants. (B) A Manhattan plot of the same data as A with a zoomed in Y-axis. (C)A Manhattan plot of –log10 (χ^2 p-value) by the chromosomal location of the African American ARDS compared with the AFR 1000 Genomes participants. (D) A Manhattan plot of the same data as C with a zoomed in Y-axis.

Supplementary table 1: The Call Rate and Hardy-Weinberg Equilibrium P-values of the 4 SNPs in the 96 Exome Sequenced Patients and 1000 Genomes Project Population and Subgroups.

SNP	rs78142040		
Gene	ARSD		
	Cases	Controls	Cases + Controls
ARDS P-value	1.77E-2	6.11E-138	4.75E-2
Caucasian P-value	8.31E-2	1	5.75E-1
African American P-value	7.90E-2	1.93E-55	5.29E-4
Pneumonia P-value	2.23E-1	6.11E-138	4.19E-8
Sepsis P-value	2.70E-2	6.11E-138	3.92E-4
African American Sepsis P-value	5.81E-2	1.93E-55	3.61E-2
African American Pneumonia P-value	4.16E-1	1.93E-55	2.41E-12
Caucasian Sepsis P-value	1.22E-1	1	7.26E-1
Caucasian Pneumonia P-value	3.55E-1	1	8.22E-1
SNP	rs9605146		
Gene	XKR3		
	Cases	Controls	Cases + Controls
ARDS P-value	3.78E-1	5.13E-15	2.35E-21
Caucasian P-value	2.02E-1	4.36E-2	6.27E-9
African American P-value	3.88E-1	2.16E-16	3.45E-14
Pneumonia P-value	7.61E-1	5.13E-15	1.83E-16
Sepsis P-value	1.16E-1	5.13E-15	1.60E-22
African American Sepsis P-value	1.22E-1	2.16E-16	1.17E-12
African American Pneumonia P-value	8.47E-1	2.16E-16	2.85E-17
Caucasian Sepsis P-value	9.24E-3	4.36E-2	3.09E-11
Caucasian Pneumonia P-value	3.84E-1	4.36E-2	5.22E-3
SNP	rs3848719		
Gene	ZNF335		
	Cases	Controls	Cases + Controls
ARDS P-value	3.81E-6	4.76E-3	2.30E-6
Caucasian P-value	2.64E-3	4.41E-1	6.23E-1
African American P-value	7.26E-4	2.44E-1	7.67E-2
Pneumonia P-value	2.09E-3	4.76E-3	1.88E-4
Sepsis P-value	5.60E-4	4.76E-3	9.10E-5
African American Sepsis P-value	2.04E-2	2.44E-1	6.48E-1
African American Pneumonia P-value	1.40E-2	2.44E-1	6.43E-1
Caucasian Sepsis P-value	1.46E-2	4.41E-1	9.94E-1
Caucasian Pneumonia P-value	7.42E-2	4.41E-1	8.31E-1

These values were taken into account in the selection of the SNPs for genotyping in the additional population.

Supplementary table 2: Linear Regression with APACHE II Score.

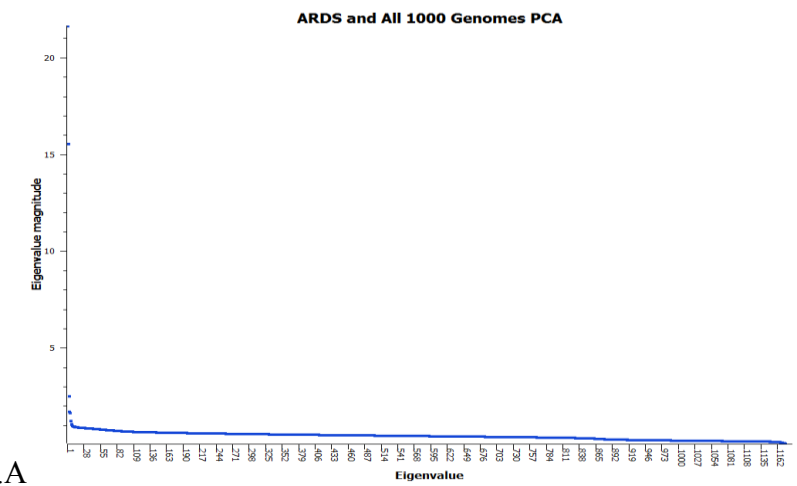
SNP	rs78142040		
Gene	ARSD		
	96 Exome	117 TaqMan	Total 213
ARDS R^2	0.005521	0.01224	0.010852
EA ARDS R^2	0.018964	0.002823	0.000723
AA ARDS R^2	0.022106	0.215676	0.016704
Pneumonia R^2	0.000523	0.000903	0.000365
Sepsis R^2	0.003156	0.021868	0.014571
AA Sepsis R^2	0.084892	0.118034	0.000143
AA Pneumonia R^2	0.049196	0.367781	0.028425
EA Sepsis R^2	0.0198	0.003489	0.001723
EA Pneumonia R^2	0.000702	0.043586	0.014079
SNP	rs9605146		
Gene	XKR3		
	96 Exome	117 TaqMan	Total 213
ARDS R^2	0.048143	0.012997	0.000497
EA ARDS R^2	0.049544	0.020126	0.001013
AA ARDS R^2	0.01239	0.101041	0.013936
Pneumonia R^2	0.039067	0.085527	0.013693
Sepsis R^2	0.071783	0.001684	0.02047
AA Sepsis R^2	0.13201	0.004314	0.009831
AA Pneumonia R^2	0.002553	0.343422	0.092252
EA Sepsis R^2	0.053535	0.002677	0.022779
EA Pneumonia R^2	0.056131	0.085977	0.009475
SNP	rs3848719		
Gene	ZNF335		
	96 Exome	117 TaqMan	Total 213
ARDS R^2	0.062995	0.005162	0.023751
EA ARDS R^2	0.001195	0.000516	0.000861
AA ARDS R^2	0.215301	0.001016	0.070074
Pneumonia R^2	0.166849	0.001253	0.014435
Sepsis R^2	0.029758	0.035016	0.041493
AA Sepsis R^2	0.340003	0.191067	0.177648
AA Pneumonia R^2	0.155128	0.044786	0.00434
EA Sepsis R^2	0.064841	0.001605	4.24E-05
EA Pneumonia R^2	0.058004	0.008149	0.002669

This table shows the results of the linear regression (additive model). Associations were considered to be significant if R^2>0.8.

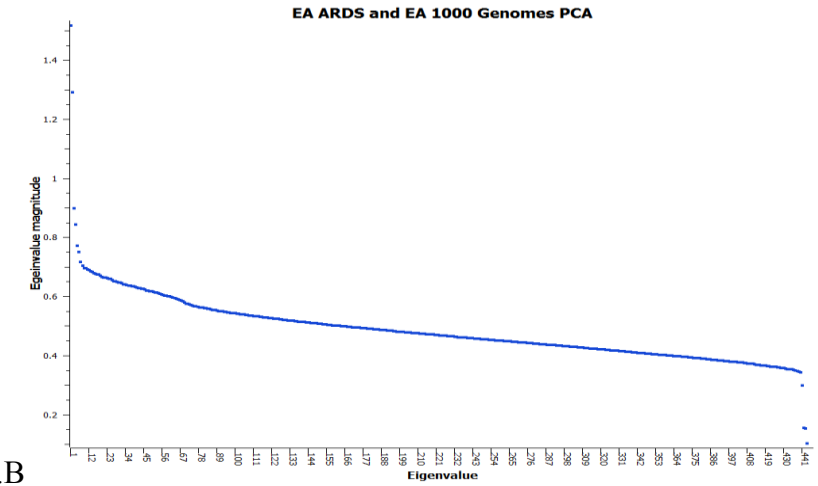
Supplementary table 3: Linear Regression with Ventilator-Free Days per 28 Days.

SNP	rs78142040		
Gene	ARSD		
	96 Exome	117 TaqMan	Total 213
ARDS R^2	0.015259	0.033802	0.026054
EA ARDS R^2	0.025028	0.000407	0.006519
AA ARDS R^2	0.001615	0.184596	0.019431
Pneumonia R^2	0.013639	5.37E-05	0.003439
Sepsis R^2	0.003739	0.088682	0.042303
AA Sepsis R^2	0.022886	0.406296	0.053339
AA Pneumonia R^2	0.005096	0.03435	0.007386
EA Sepsis R^2	0.017498	0.002952	0.007018
EA Pneumonia R^2	0.007463	0.004356	9.83E-05
SNP	rs9605146		
Gene	XKR3		
	96 Exome	117 TaqMan	Total 213
ARDS R^2	0.018928	0.007354	0.000112
EA ARDS R^2	0.004794	0.029358	0.003427
AA ARDS R R^2	0.048353	1.17E-05	0.01857
Pneumonia R^2	0.083303	0.022224	0.00073
Sepsis R^2	0.00051	0.001202	4.27E-05
AA Sepsis R^2	0.030121	0.02358	0.000308
AA Pneumonia R^2	0.065061	0.037774	0.056542
EA Sepsis R^2	9.04E-05	0.000829	0.000312
EA Pneumonia R^2	0.035209	0.088483	0.012099
SNP	rs3848719		
Gene	ZNF335		
	96 Exome	117 TaqMan	Total 213
ARDS R^2	7.29E-05	7.24E-05	0.000112
EA ARDS R^2	0.022345	0.017442	0.014008
AA ARDS R^2	0.024184	0.006926	0.009665
Pneumonia R^2	0.031582	0.033899	0.001198
Sepsis R^2	0.012915	0.037174	0.0031
AA Sepsis R^2	0.024291	0.114525	0.000487
AA Pneumonia R^2	0.032883	0.034291	0.034885
EA Sepsis R^2	0.071066	0.008406	0.005712
EA Pneumonia R^2	0.000707	0.10277	0.028252

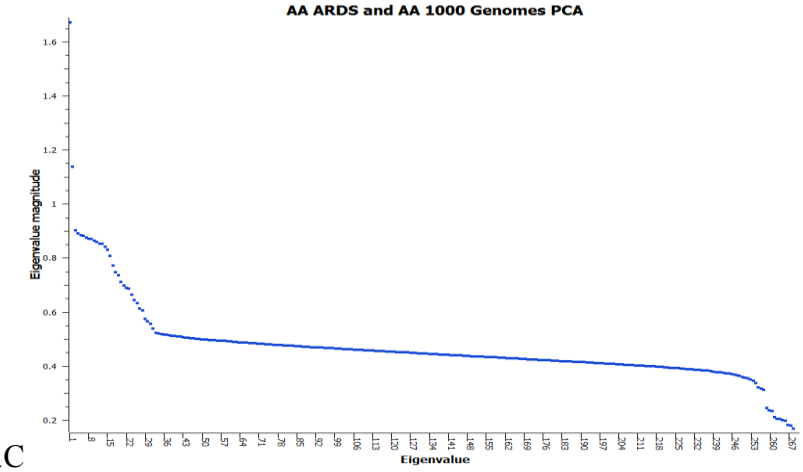
This table shows the R^2 results of the linear regression (additive model). Associations were considered to be significant if $R^2 > 0.8$.



S2.A

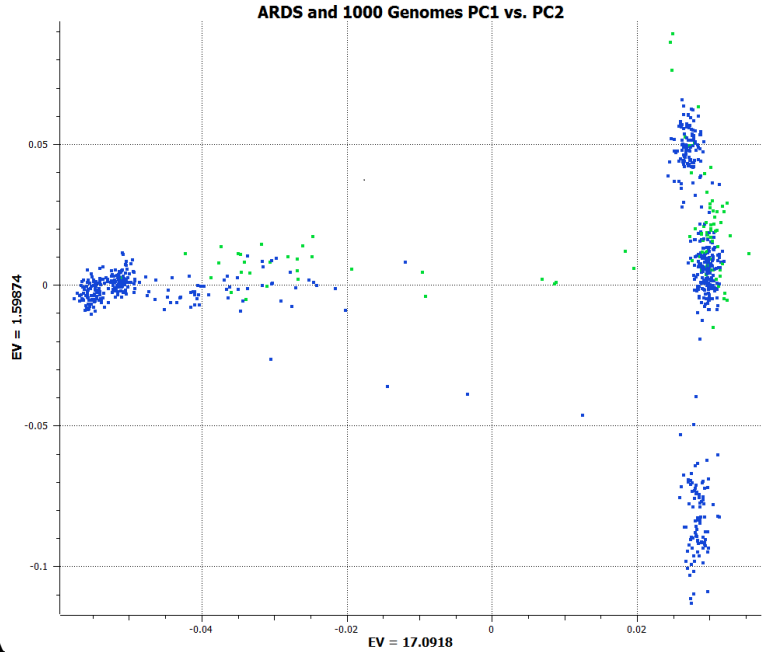


S2.B

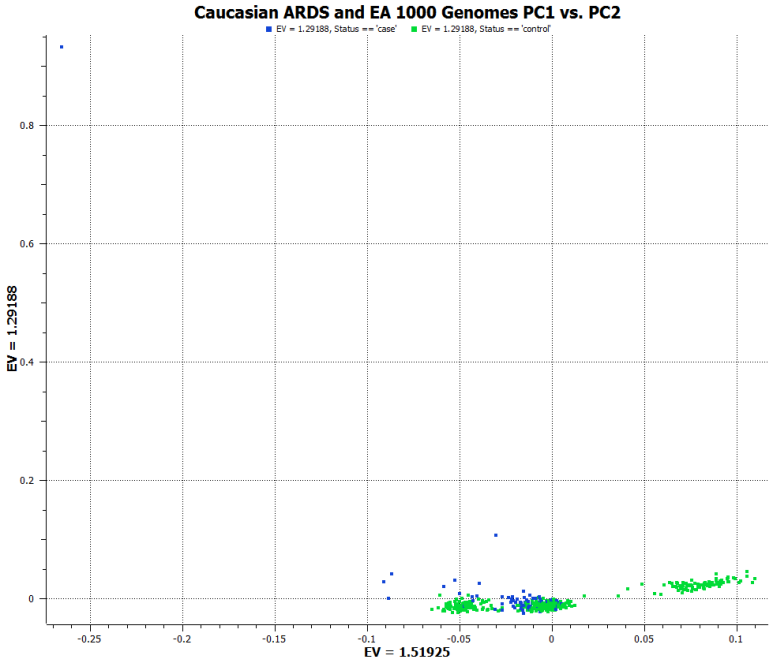


S2.C

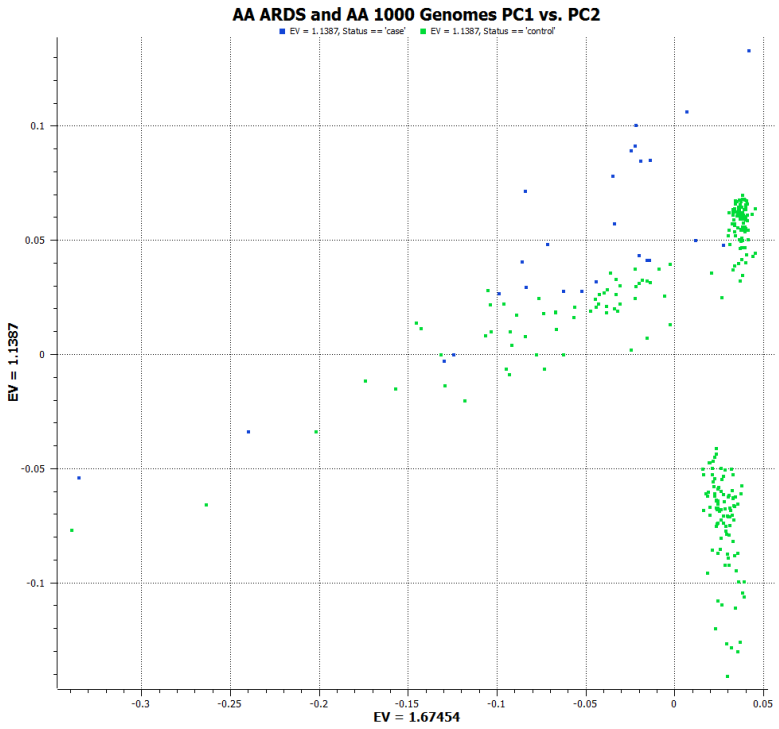
Supplementary Figure 2: Scree Plots of the Eigenvalues Generated by Principal Component Analysis. The largest eigenvalues are used in corrections for population structure. A) The All ARDS and AFR+EUR 1000 Genomes population eigenvalues. 720 principal components were measured and the largest eigenvalue is 17.09. B) The Caucasian ARDS and EUR ARDS population eigenvalues. 449 principal components were measured and the largest eigenvalue is 1.52. C) The African American ARDS and AFR 1000 Genomes population eigenvalues. 272 principal components were measured and the largest eigenvalue is 1.67.


ARDS and 1000 Genomes PC1 vs. PC2
EV = 1.59874
EV = 17.0918
0.05
0
-0.05
-0.1
-0.04
-0.02
0
0.02

S3.A

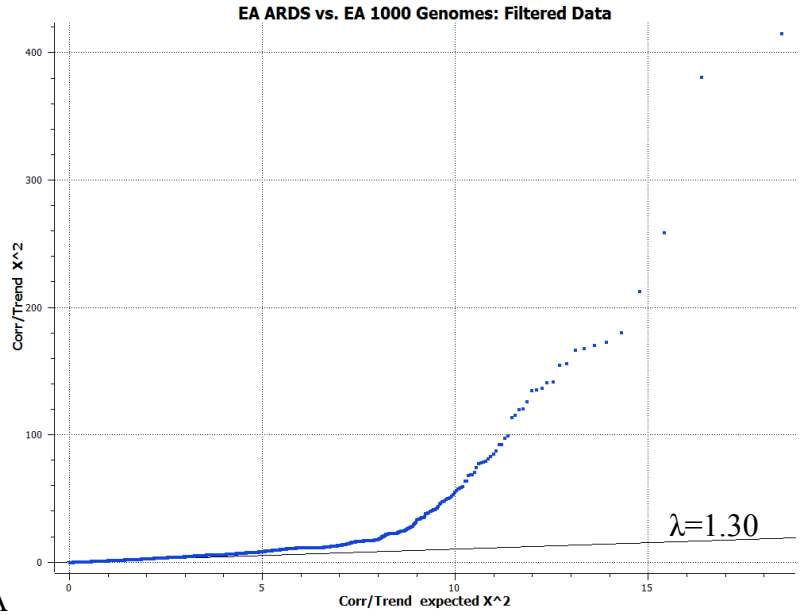

Caucasian ARDS and EA 1000 Genomes PC1 vs. PC2
EV = 1.29188, Status == 'case'
EV = 1.29188, Status == 'control'
EV = 1.29188
EV = 1.51925
0.8
0.6
0.4
0.2
0
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1

S3.B

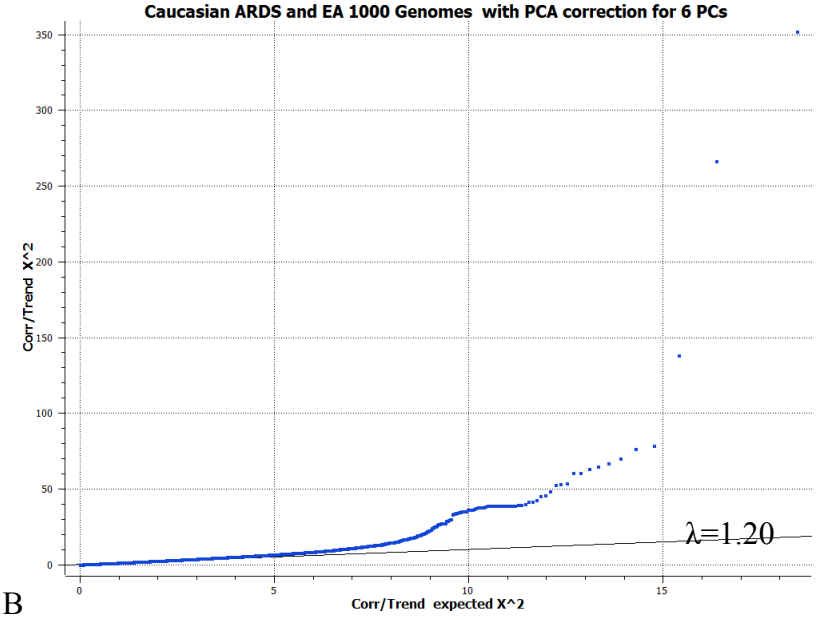


S3.C

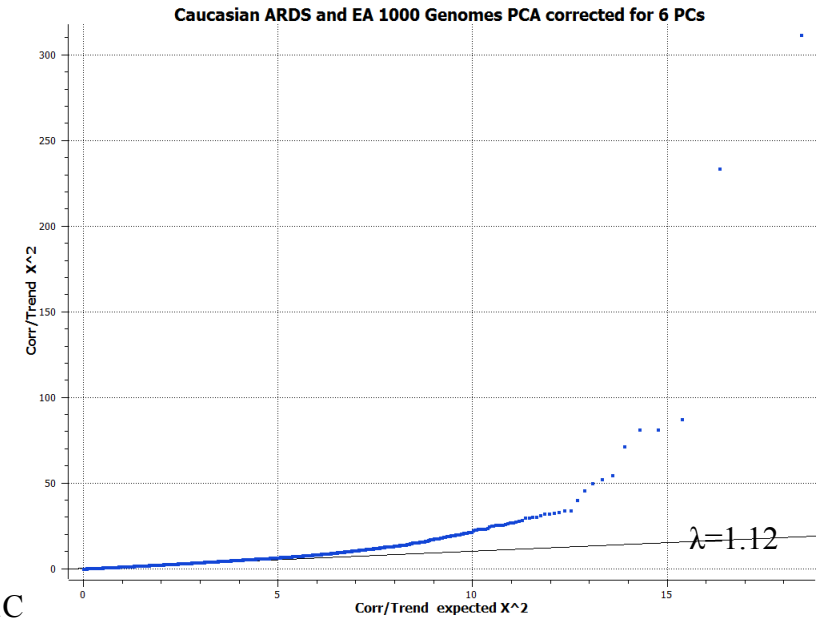
Supplementary figure 2: Scatter Plots of the Principal Component Values. The principal component values for the first 2 eigenvalues are plotted against each other where the cases are in blue and the controls in green. A) The All ARDS and 1000 Genomes population principal components. B) The Caucasian ARDS and EUR ARDS population principal components. C) The African American ARDS and AFR 1000 Genomes population principal components.


EA ARDS vs. EA 1000 Genomes: Filtered Data
Corr/Trend X^2
Corr/Trend expected X^2
λ=1.30

S4.A

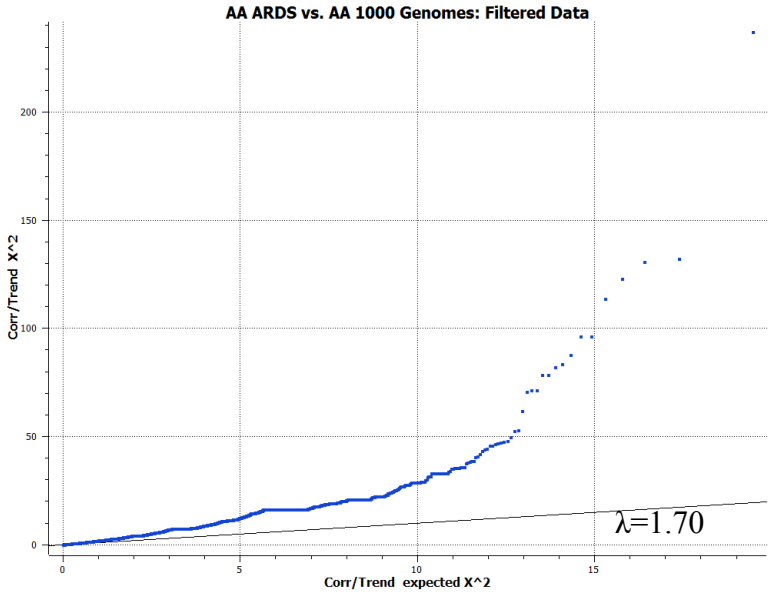

Caucasian ARDS and EA 1000 Genomes with PCA correction for 6 PCs
Corr/Trend X^2
Corr/Trend expected X^2
λ=1.20

S4.B

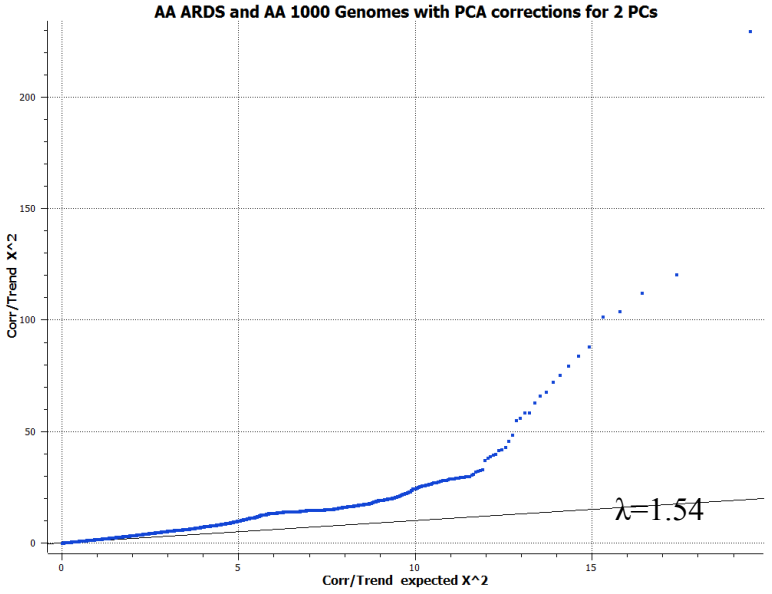


S4.C

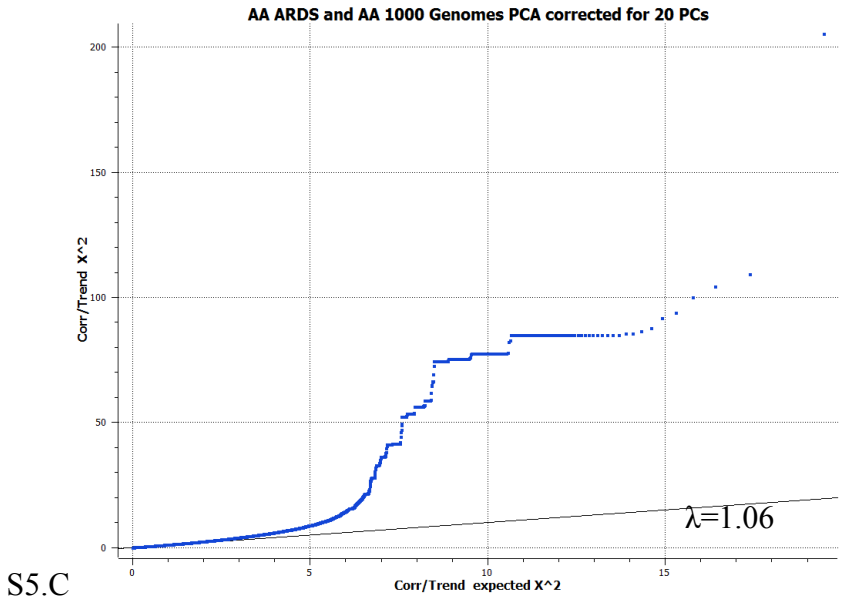
Figure 5: Quantile-Quantile Plots of Genotypic Trend Test χ^2 Values for the Caucasian ARDS and EUR 1000 Genomes Comparison.
The straight line on each plot represents y=x. A) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data is filtered on HWE, LD, and SNP call rate but not PCA corrected. B) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data have been filtered and corrected for 6 PCs. C) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data have been filtered and corrected for 6 PCs and undergone sample outlier removal.


AA ARDS vs. AA 1000 Genomes: Filtered Data
Corr/Trend X^2
Corr/Trend expected X^2
λ=1.70

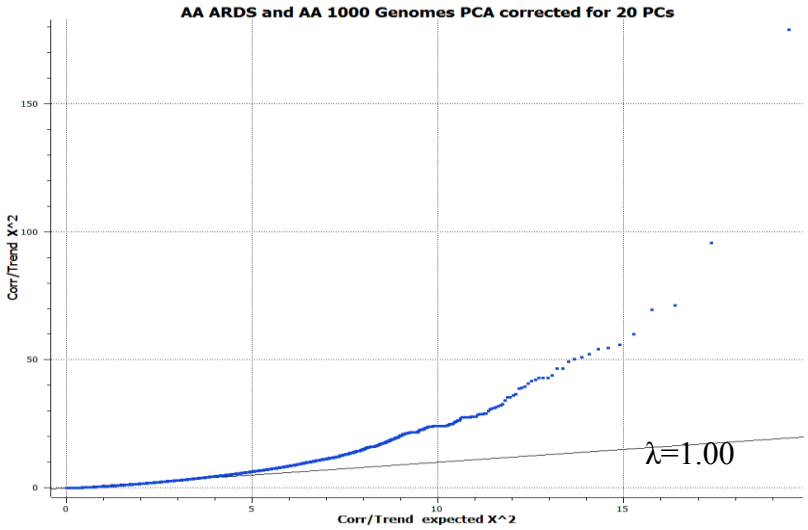
S5.A


AA ARDS and AA 1000 Genomes with PCA corrections for 2 PCs
Corr/Trend X^2
Corr/Trend expected X^2
λ=1.54

S5.B



S5.C



S5.D

Figure 6: Quantile-Quantile plots of Genotypic Trend test χ^2 Values for the African American ARDS and AFR 1000 Genomes Comparison.
The straight line on each plot represents y=x. A) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data is filtered on HWE, LD, and SNP call rate but not PCA corrected. B) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. No samples were identified for outlier removal. C) QQ plot of expected χ^2 values versus the actual χ^2 for the genotypic trend test of case-control status. The data have been filtered and corrected for 20 PCs. D) QQ plot of expected χ^2 values versus the actual χ^2 values for the genotypic trend test of case-control status. The data have been filtered and corrected for 20 PCs and undergone sample outlier removal.

REFERENCES

1. Ashbaugh DG, Bigelow DB, Petty TL, Levine BE. Acute respiratory distress in adults. *Lancet.* Aug 12 1967;2(7511):319-323.
2. Bernard GR, Artigas A, Brigham KL, et al. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *American journal of respiratory and critical care medicine.* Mar 1994;149(3 Pt 1):818-824.
3. Rubenfeld GD, Caldwell E, Peabody E, et al. Incidence and outcomes of acute lung injury. *The New England journal of medicine.* Oct 20 2005;353(16):1685-1693.
4. Blank R, Napolitano LM. Epidemiology of ARDS and ALI. *Critical care clinics.* Jul 2011;27(3):439-458.
5. Flores C, Pino-Yanes MM, Casula M, Villar J. Genetics of acute lung injury: past, present and future. *Minerva anestesiologica.* Oct 2010;76(10):860-864.
6. Garcia JG. Searching for candidate genes in acute lung injury: SNPs, Chips and PBEF. *Transactions of the American Clinical and Climatological Association.* 2005;116:205-219; discussion 220.
7. Gong MN. Genetic epidemiology of acute respiratory distress syndrome: implications for future prevention and treatment. *Clinics in chest medicine.* Dec 2006;27(4):705-724; abstract x.
8. McGlothlin JR, Gao L, Lavoie T, et al. Molecular cloning and characterization of canine pre-B-cell colony-enhancing factor. *Biochemical genetics.* Apr 2005;43(3-4):127-141.
9. Parks BW, Nam E, Org E, et al. Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell metabolism.* Jan 8 2013;17(1):141-152.
10. Tzouvelekis A, Pneumatikos I, Bouros D. Serum biomarkers in acute respiratory distress syndrome an ailing prognosticator. *Respiratory research.* 2005;6:62.
11. Crader KM RJ, Repine JE. Breath Biomarkers and the Acute Respiratory Distress Syndrome. *J Pulmonar Respirat Med.* 01/01/2012 2012;2(111).
12. Liu Y, Shao Y, Yu B, Sun L, Lv F. Association of PBEF gene polymorphisms with acute lung injury, sepsis, and pneumonia in a northeastern Chinese population. *Clinical chemistry and laboratory medicine : CCLM / FESCC.* Nov 2012;50(11):1917-1922.
13. Ye SQ, Simon BA, Maloney JP, et al. Pre-B-cell colony-enhancing factor as a potential novel biomarker in acute lung injury. *American journal of respiratory and critical care medicine.* Feb 15 2005;171(4):361-370.
14. Bajwa EK, Yu CL, Gong MN, Thompson BT, Christiani DC. Pre-B-cell colony-enhancing factor gene polymorphisms and risk of acute respiratory distress syndrome. *Critical care medicine.* May 2007;35(5):1290-1295.
15. Lee KA, Gong MN. Pre-B-cell colony-enhancing factor and its clinical correlates with acute lung injury and sepsis. *Chest.* Aug 2011;140(2):382-390.

16. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *Journal of pathology informatics.* 2012;3:40.
17. Goh G, Choi M. Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics & informatics.* Dec 2012;10(4):214-219.
18. Takata A, Kato M, Nakamura M, et al. Exome sequencing identifies a novel missense variant in RRM2B associated with autosomal recessive progressive external ophthalmoplegia. *Genome biology.* 2011;12(9):R92.
19. Fang X, Bai C, Wang X. Bioinformatics insights into acute lung injury/acute respiratory distress syndrome. *Clinical and translational medicine.* 2012;1(1):9.
20. Genomes Project C, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* Nov 1 2012;491(7422):56-65.
21. Trojani A, Di Camillo B, Tedeschi A, et al. Gene expression profiling identifies ARSD as a new marker of disease progression and the sphingolipid metabolism as a potential novel metabolism in chronic lymphocytic leukemia. *Cancer biomarkers : section A of Disease markers.* 2011;11(1):15-28.
22. Calenda G, Peng J, Redman CM, Sha Q, Wu X, Lee S. Identification of two new members, XPLAC and XTES, of the XK family. *Gene.* Mar 29 2006;370:6-16.
23. Mahajan MA, Murray A, Samuels HH. NRC-interacting factor 1 is a novel cotransducer that interacts with and regulates the activity of the nuclear hormone receptor coactivator NRC. *Molecular and cellular biology.* Oct 2002;22(19):6883-6894.
24. Yang YJ, Baltus AE, Mathew RS, et al. Microcephaly gene links trithorax and REST/NRSF to control neural stem cell proliferation and differentiation. *Cell.* Nov 21 2012;151(5):1097-1112.
25. Rubenfeld GD, Herridge MS. Epidemiology and outcomes of acute lung injury. *Chest.* Feb 2007;131(2):554-562.
26. Franco B, Meroni G, Parenti G, et al. A cluster of sulfatase genes on Xp22.3: mutations in chondrodysplasia punctata (CDPX) and implications for warfarin embryopathy. *Cell.* Apr 7 1995;81(1):15-25.
27. Urbitsch P, Salzer MJ, Hirschmann P, Vogt PH. Arylsulfatase D gene in Xp22.3 encodes two protein isoforms. *DNA and cell biology.* Dec 2000;19(12):765-773.
28. Dooley TP, Haldeman-Cahill R, Joiner J, Wilborn TW. Expression profiling of human sulfotransferase and sulfatase gene superfamilies in epithelial tissues and cultured cells. *Biochemical and biophysical research communications.* Oct 14 2000;277(1):236-245.
29. Le Goff GC, Bres JC, Rigal D, Blum LJ, Marquette CA. Robust, high-throughput solution for blood group genotyping. *Analytical chemistry.* Jul 15 2010;82(14):6185-6192.
30. Lunetta KL. Genetic association studies. *Circulation.* Jul 1 2008;118(1):96-101.
31. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA : the journal of the American Medical Association.* Mar 19 2008;299(11):1335-1344.

32. Wheeler AP, Bernard GR, Thompson BT, et al. Pulmonary-artery versus central venous catheter to guide treatment of acute lung injury. *The New England journal of medicine.* May 25 2006;354(21):2213-2224.
33. Wiedemann HP, Wheeler AP, Bernard GR, et al. Comparison of two fluid-management strategies in acute lung injury. *The New England journal of medicine.* Jun 15 2006;354(24):2564-2575.
34. Abecasis GR, Altshuler D, Auton A, et al. A map of human genome variation from population-scale sequencing. *Nature.* Oct 28 2010;467(7319):1061-1073.
35. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010 Sep 20 2010;20:1297-1303.
36. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.).* Apr 15 2005;308(5720):385-389.
37. Melum E, Franke A, Schramm C, et al. Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. *Nature genetics.* Jan 2011;43(1):17-19.
38. Anasiewicz A, Maciejewski R, Juskiewicz W, Lakowska H, Madej B, Szkodziak P. [Changes of lysosomal enzyme activity in the lungs during the course of acute pancreatitis]. *Wiad Lek.* 1997;50 Suppl 1 Pt 2:96-100.

VITA

Katherine is a Master degree candidate for Bioinformatics in the Department of Biomedical and Health Informatics, University of Missouri School of Medicine in Kansas City, Missouri. She was born on April 28, 1989 in St. Charles, Missouri. She began assisting with biotechnology research at age 15 and graduated from Indiana University, Bloomington in 2011 with a bachelor's of science in Biology with a minor in French and a minor equivalent in chemistry, and has been a member of the Omicron Delta Kappa National Leadership Honor Society since 2013. She published an abstract on the effect of polychlorinated biphenyls on hepatic oxidative stress enzymes in 2010 (Society of Toxicology) and an abstract on novel single nucleotide polymorphisms associated with acute respiratory distress syndrome in 2014 (American Thoracic Society). After graduation, she plans to further pursue a PhD degree in the same field with a career goal as a translational researcher in medical science.