A CASE-CONTROL BASED GENOMIC ANALYSIS OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

ANJANA RAMNATH

David Moxley, Thesis Supervisor

MAY 2023

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

A CASE-CONTROL BASED GENOMIC ANALYSIS OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE

presented by Anjana Ramnath,

a candidate for the degree of Master of Science,

and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Professor David Moxley

_____

Professor Sue Boren

_____

Professor Iris Zachary

DEDICATION

I would like to thank my father, Mr. Ramnath Narasimhan, my mother, Mrs. Usha Ramnath and my sister, Ms. Neerja Ramnath for all the sacrifices they made and their endless love, support and encouragement without which this work would not have been possible.

ACKNOWLEDGMENTS

I would like to thank the faculty at the University of Missouri who have helped and guided me in my studies.

I would like to also thank my mentors for their continued support and help in my research. I would also like to thank my committee members. I am also grateful to my fellow friends and colleagues at MU for their friendship and support.

Finally, thanks to all my extended family members and friends for all their support, encouragement and guidance.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

LIST OF ABBREVIATIONS

PM	Precision Medicine
COPD	Chronic Obstructive Pulmonary Disease
MeSH	Medical Subject Headings
HT-NGS	High Throughput-Next Generation Sequencing
DE	Differential Expression
PPI	Protein-Protein Interaction Network
GO	Gene Ontology
EHR	Electronic Health Record
AI	Artificial Intelligence
ML	Machine Learning

ABSTRACT

Chronic Obstructive Pulmonary Disease is a respiratory illness that affects a large number of people all over the world. It is a major cause of chronic morbidity and mortality and a serious global public health problem. COPD is the fourth leading cause of death worldwide. Although the environmental causes of COPD which predominantly include cigarette smoking are well-documented, to this date the genetic underpinnings of COPD remain largely unknown. Furthermore, in the current landscape of a respiratory pandemic, COPD patients are at a much higher risk for developing other respiratory illnesses and co-morbidities. Treatment methods for this disease have remained the same over the years. In this study we use genomic data from case-control based cohorts to study the genetic patterns and profiles of patients with this illness in order to identify key genetic factors and thereby gain a deeper understanding of the disease. This understanding would lead to greater insight on how to better diagnose, manage and treat this disease. A clearer insight at the genomic level would assist in actionable outcomes than could be leveraged to adopt a more Precision Medicine based modality to manage this disease thereby leading to more effective and better treatment options which would improve overall patient health outcomes

CHAPTER 1: Introduction to Chronic Obstructive Pulmonary Disease

Chronic Obstructive Pulmonary Disease (COPD) is a common, preventable disease that is a leading cause of mortality worldwide (Duffy & Criner, 2019).

1.1: Symptoms of COPD

It is characterized by persistent respiratory symptoms and airflow limitation due to airway or alveolar abnormalities usually caused by significant exposure to noxious particles or gases. Emphysema and chronic bronchitis are the most common symptoms of COPD (Silverman, 1998).

1.2: Risk Factors of COPD

Tobacco use remains the main risk factor for COPD (Diaz-Guzman & Mannino, 2014). The sources for COPD remain different in developed vs developing countries. In developed nations, cigarette smoke continues to remain the number one risk factor whereas in developing countries environmental pollution remains a major risk factor.

1.3: Diagnosis and Detection

Diagnosis of COPD is done via Lung (Spirometry) tests, Chest X-ray, CT Scan, Arterial blood gas analysis, and specific lab tests (Labaki & Rosenberg, 2020). People with COPD are at increased risk of developing heart disease, lung cancer and a variety of other conditions.

1.4: Treatment of COPD

Treatment for these symptoms of COPD include inhalers to stimulate bronchodilation as well as pharmacotherapy (Labaki & Rosenberg, 2020). Given the heterogeneous nature of this disease, a pharmacogenomics and precision medicine based approach taking into consideration sample size as well as COPD subtypes and genetics would prove to be beneficial to obtain more efficient treatment outcomes.

Clinical implementation of these modalities of treatment nevertheless has not yet been achieved (Wood et al., 2009; Hersh, 2019). In this paper, we outline a genomics based network analysis approach to gain a more thorough understanding of genetics underpinnings of this disease. This would lead to more targeted treatment based on genetics and better overall clinical outcomes.

CHAPTER 2: COPD in the context of Precision Medicine

2.1: Introduction to Precision Medicine

Precision medicine (PM) is a new and emerging field that looks to treat diseases by studying genetic, socio-environmental, and clinical factors that influence disease. It thereby attempts to provide a holistic view of a person's health. It usually deals with a lot of details about the individual especially their genetics. Precision medicine essentially has the potential to personalize or tailor treatment to the individual patient by using their genetic information to guide treatment choices (Velmovitsky et al., 2021; Wynn et al., 2018) .

Precision Medicine shows the most potential in treating oncology related diseases, where it is possible to tailor the specific drug and treatment based on what is known as the patient's genomic profile (Gameiro et al., 2018). The rationale underlying the use of genetics is based on certain important factors. One of the most important factors is that each individuals genetic make-up varies and is unique to that particular individual. As such, no two individuals have the exact same genetic make-up, thereby leading to differences that could change the way they react not only to the same medication but also the same illness. For example, some people react differently to the same dosage of acetaminophen than others. No two people also respond the same way when they are infected with the same virus.

Furthermore, in certain highly specialized disciplines such as oncology, the PM approach allows the provider to give drugs that are highly specific to the particular type of tumor a person has. Physicians can use genetic testing to assess genetic variants for example BRCA1 and tailor their patient's treatment based on that. In oncology, genomics and bioinformatics play key roles in analyzing and helping tailor treatment

modalities (Canzoneri et al., 2019; McGowan et al., 2014) . Several workflows and methodologies have in fact been developed to better analyze and study how treatment can be tailored based on genomics of the patient (Jäger, 2022).

A driving factor in increasing use of PM is reduction in cost of genome sequencing. Initially it cost thousands just to sequence one genome. Now, however, with the advent of new library prep methods and other technological advancements, genomes can be sequenced for much less (McCombie et al., 2019; Pareek et al., 2011). This has in turn made sequencing of several patient's genomes possible thereby leading to the advent of precision medicine.

One of the main reasons, however was the mapping of the entire human genome via the Human Genome Project (Green et al., 2015). This project set out the ambitious goal of mapping all the regions in the entire human genome. This was done under the auspices of the National Institutes of Health under Dr. Francis Collins. This project successfully mapped out the entire Human Genome and in fact reached its goal earlier than anticipated (Collins et al., 2003).

This in turn has given rise to several new technologies and advancements in the field beginning with genomics and ranging from HT-NGS technologies such as analysis of chromatin immunoprecipitation coupled to DNA microarray or bisulfite sequencing (ChIP-seq), RNA sequencing (RNAseq), whole genome genotyping, genome wide structural variation to de novo assembling and re-assembling of the genome (Pareek et al., 2011).

Further initiatives were taken by the government including *All of Us* research project which used PM based methods(Llanto et al., 2020; Sankar & Parker, 2017). This project is an initiative that helps to better highlight precision medicine. It seeks to create

a database with precision medicine based information that is reflective of the diverse US population. Participants information about genetic data, biological samples, and other information about their health is collected. Participants can access their health information, as well as research that uses their data, during the study. Researchers including Clinicians and Physicians can use these data to study a large range of diseases, with the goals of better predicting disease risk, understanding how diseases occur, and finding improved diagnosis and treatment strategies. Other initiatives by the government also include the Obama administration's Precision Medicine Initiative (Matthew, 2019).

2.2: Addressing COPD through Precision Medicine

Precision medicine approaches offer promising avenues to improve not only diagnosis, but also treatment and management of COPD.

Precision medicine's objective is to tailor medical care to an individual's specific genetic, environmental, and lifestyle factors. In the context of COPD, a precision medicine approach could involve identifying specific genetic or biomarker-based subtypes of the disease and developing targeted therapies or interventions that are more effective and have fewer side effects (Leung et al., 2019).

The use of bronchodilators that target specific molecular pathways is an important example of how precision medicine can be leveraged in COPD treatment and long-term management. An example would be patients with COPD who have a specific genetic mutation could potentially benefit from bronchodilators that target that particular mutation (Wood et al., 2009).

Another approach is using biomarkers to identify patients who are at higher risk of exacerbations or disease progression. These patients may be given more aggressive

treatments or more specific, targeted interventions either to prevent exacerbations or slow disease progression (Stockley et al., 2019).

Further, precision medicine approaches can help identify specific factors that contribute to the development of COPD. These factors could range from environmental exposures, lifestyle factors and various comorbidities. This information can be used to develop prevention and management strategies that are tailored to suit a specific patient. Combining a Precision Medicine approach within the broader framework of Population health brought about the concept of P4 medicine whereby genomics is taken into account within the broader framework of improving overall public health with an aim of getting patients also more involved (Khoury et al., 2012).

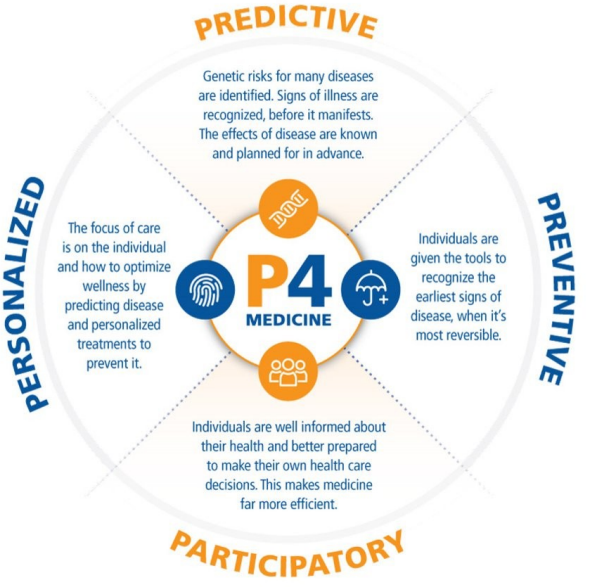


Figure 1.1: P4 Medicine

Overall, precision medicine has the potential to revolutionize the diagnosis, treatment, and management of COPD by providing tailored interventions that are more effective and with fewer side effects. This would lead to better outcomes for patients. Although prior work on studying impact of precision medicine in healthcare has been

performed, there does not exist a specific systematic review looking at provider approaches and experiences of precision medicine clinical practice.

CHAPTER 3: Literature Review

3.1: Methods

A search of PubMed (2017-2022) and Ovid Medline (2017-2022) for research articles was conducted in November 2022 using the following search terms: precision medicine (Medical Subject Headings [MeSH]) and physician (MeSH). The term genomics (MeSH) was also used. These search terms and filters were used to find the most contemporarily relevant articles that embraced all methods of physician's experiences with precision medicine.

3. 2: Inclusion and Exclusion Criteria

Inclusion criteria were any research article evaluating physician or provider's experiences with precision medicine in a clinical setting. Only articles published in peer-reviewed journals from 2017-2022 were considered. Conference papers were not included. Furthermore, only studies that were in English and on Humans were included. Studies were excluded if they did not specifically address physician perspectives or focused on settings outside the US. Articles that were reviews themselves or were Randomized Control Trials were also excluded.

3.3: Study Selection and Data Extraction

The titles, abstracts and results were reviewed and screened based on the inclusion and exclusion criteria described earlier. Data that was relevant was individually collected from each screened article including type of study, type of intervention/method used and results. The provider and result data were sorted and then organized into a table, including the intervention type. Data were studied for patterns based on differences in the intervention type used.

The PRIMSA guidelines were used to construct the four phase flow diagram which included Identification, Screening, Eligibility and articles Included (*PRISMA*

Guidelines, n.d.). The aim of the PRISMA statement is to help authors improve the reporting of systematic reviews and meta-analyses. PRISMA guidelines can also be used for appraisal of published systematic reviews. Zotero version 6.0.18 was used to manage the articles. It is important keep in mind that since this is such a new field, the number of articles overall was less and this influenced to a great extent the size of the recall set.

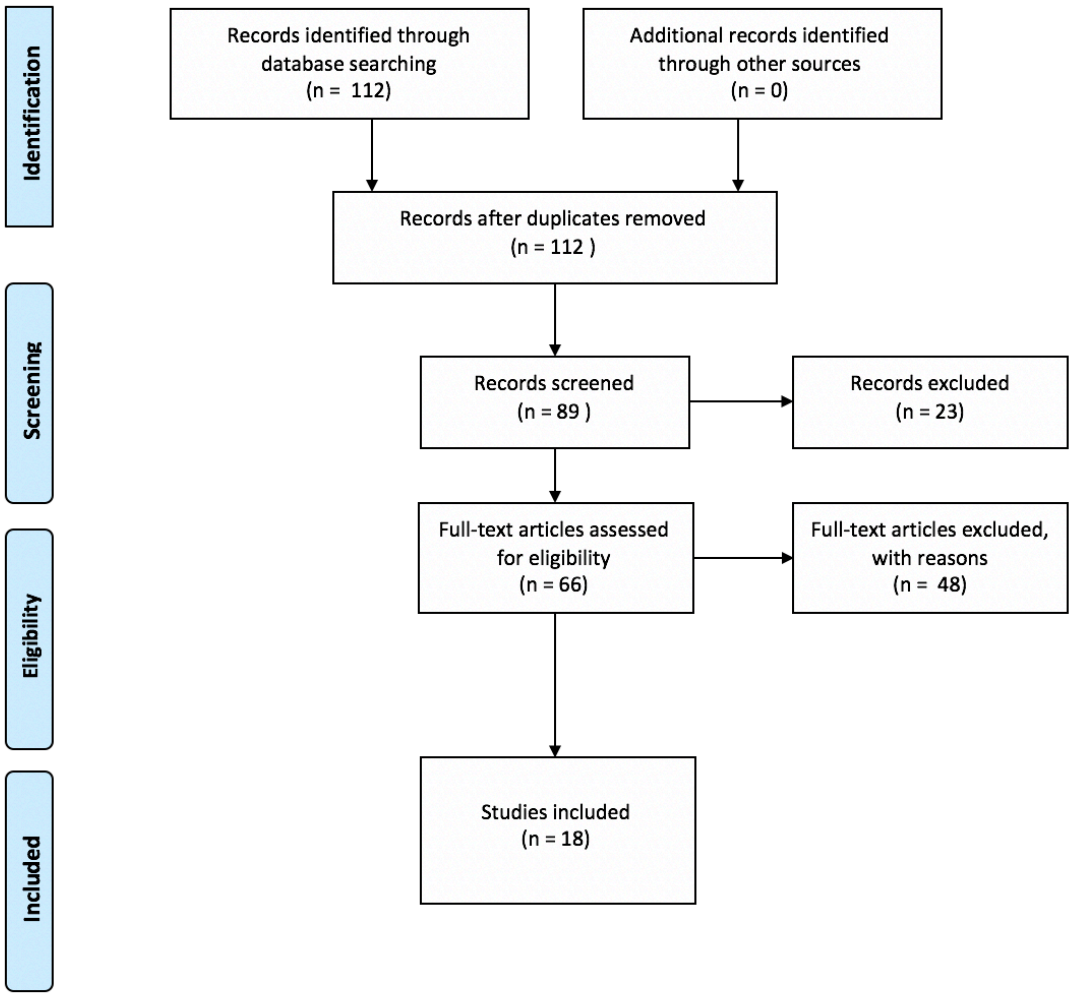


Figure 3.1: Flow chart of article selection

3.4: Literature Review Results

Comprehensive literature searches identified 112 articles (Figure 1). The titles and abstracts of these articles were studied and 14 articles were ultimately determined to be of relevance. The distribution of articles included surveys and interviews as well as discussions in conferences and clinical settings and one case study (Table 1). Four studies focused on approaches of specific disciplines and five on primary care setting and the rest focused on general provider experiences. Most of the studies were on perceived challenges and provider focused efforts to overcome. Studies were mostly focused on provider initiated PM and one study focused on provider experience with patient initiated direct-to-consumer testing. Studies were broadly classified into five categories: Surveys done on provider experiences, Opinion pieces, Cohort studies done in a Clinical Genomics setting, Studies done after Educational Interventions, and studies done based on discussions in Conferences and Summits. The author of each study also put forth their own assessment of provider perceptions and approaches. Nine of the eighteen studies (13%) were surveys that were conducted, 3 were done as cohort studies in clinical setting, 2 were opinion pieces, 1 was an educational intervention and 1 was a case-study.

These studies stated that various perceptions of precision medicine existed between various disciplines and in certain cases the provider approach towards precision medicine depended on the disciples. Four of the studies stated that demographics also played a role in assessing provider experiences and perceptions towards precision medicine. One of the studies also looked at the legal ramifications in

provider experience towards PM.

3.5.1: Surveys and Interviews of Physician experience and attitude towards Precision Medicine

Among the studies reporting survey or interview data, authors conducted a thorough assessment of provider as well as patient experience and the challenges they faced thereof with implementation of precision medicine. In one study the authors conducted a study assessing implementation outcomes to inform spread and scale, using mixed methods of semi-structured interviews of physicians. Fifty of 69 patients (72%) invited by primary care providers participated in the Humanwide pilot. Reviews were performed for the 50 participating patients. Participants were diverse overall (50% non-white, 66% female). Over half of the participants were obese and 58% had one or more major cardiovascular risk factor: dyslipidemia, hypertension, diabetes. Reach/penetration of the four components varied: pharmacogenomics testing 94%, health coaching 80%, genetic testing 72%, and digital health 64%. Interview participants (n=27) included patients (n=16), providers (n=9) (Brown-Johnson et al., 2021). The authors focused on four main components, namely health coaching, genetic screening, pharmacogenomics and digital health. This was a report of implementation of a multi-component precision health model embedded in team-based primary care. The authors found acceptance from both patients and providers. The authors found that barriers to implementation of precision health in a team-based primary care clinic were straightforward, though not necessarily easy to overcome. The authors suggested based on their survey results that future implementation endeavors should invest in basics such as education, workflow, and reflection/evaluation. Other surveys showed that less than half of physicians had a clear understanding of pharmacogenomics, which is a mainstay of PM (J. DeLuca et al.,

2020; McGrath et al., 2019).

Further survey based study such as the one done with pathologist residents showed that There was significantly lower perceived understanding of genetics compared with non-genetics topics (Haspel et al., 2021; McCauley et al., 2017).

Primary care providers also felt unprepared to work with patients at high risk for genetic conditions and were not confident about interpreting test results (Hauser et al., 2018; McCauley et al., 2017; Sethi et al., 2022; Vashistha et al., 2020; Yabroff et al., 2020) . The surveys showed that, the overall consensus seemed to be that providers need better training and education on PM based concepts in order to successfully implement them.

3.5.2: Case Study of Provider experience with Direct-to-Consumer (DTC) testing

There was one case study which looked at provider experience with patient ordered DTC. This was a case where a couple presented their 23andMe test to provider. The article explores how the physician should present and discuss the results of the such genetics test to consumers and the challenges and scenarios therein (Artin et al., 2019).

3.5.3: Studies in Clinical setting assessing Physician experience with Clinical Genomics testing

There were three studies done which looked at physician experience with clinical genomics tests. Patient's were subjected to genomics tests and physicians experience with assessing and interpreting these tests was studied. The sample size here ranged from 10 to 814. Studies were done mostly in the discipline of oncology and one was done in primary case setting. In overall primary care setting, around 61 were for genetic cancer risk assessment, 29 for pharmacogenomic testing, and 9 for validation and

interpretation of direct-to-consumer testing. The largest genomic testing sample was one from a retrospective cohort study of women with incident breast cancer which consisted of 156,229 patients (Zipkin et al., 2021) .

Oncologists who adopted specific genomic test (ODX) were associated with a 1.38-fold increase in the odds of the medical oncologist adopting ODX in 2010-2011 (95% CI = 1.04-1.83), as was co-location with early-adopting surgeons (odds ratio [OR] = 1.25, 95% CI = 1.00-1.58). Patients whose primary medical oncologist was linked to an early-adopting surgeon through co-location (OR = 1.17, 95% CI = 1.04-1.32) or both patient-sharing and co-location (OR = 1.17, 95% CI = 1.03-1.34) were more likely to receive ODX. It is worth noting that the larger the sample size, the greater the log-odds ratio of physicians adopting use of genomic testing (Gutierrez et al., 2017; Massart et al., 2022) .

3.5.4: Opinion pieces on provider experience with precision medicine

There was a total of 1 study where the authors put forth their opinions on clinical genomics testing and provider approaches to their same. The authors in these pieces talked in general about challenges in integrating precision medicine into the clinical setting and strategies to overcome these (McGrath et al., 2021; Ta et al., 2019) .

3.5.5: Study of Provider experience with Educational Interventions to integrate precision medicine

One study focused on an educational intervention that was used specifically to improve provider understanding of genomics and precision medicine. In this study, the authors assessed pharmacogenomics clinical implementation education initiatives for providers (Rohrer Vitek et al., 2017).

The summary of the findings is included in Table 1.

3.6: Discussion of Literature Review

The findings from the reviewed articles provided evidence that providers as a whole had a positive approach towards integration of precision medicine. Overall, providers expressed challenges in certain key areas including a lack of standardization and reimbursement concerns. Providers believed that the integration of precision medicine into healthcare, while showing great promise, faces many barriers. There was a consensus that pharmacogenomic education and system-wide implementation is necessary to overcome some of these challenges. A large-scale expansion of pharmacogenomics education is a step toward producing knowledgeable providers who are in a much better position to apply its methodology and promote its benefits in individualized care.

Training and educating providers in genomic medicine will become more important in the years to come as physicians increasingly interact with genomic and other precision medicine technologies. Currently, as per the articles, physicians report little to no interaction with genetic specialists such as genetic counselors. Demographics does appear to play a crucial role given the fact that in one articles the younger residents report greater exposure to genetics and genetic testing. It is further evident that cost and legal ramifications are further issues that providers are concerned about.

However, it is heartening to note that they are also willing to discuss these hitherto unexplored topics such as legalities behind genomic testing. It is of singular importance that based on the feedback from some of the surveys, providers express one of the biggest issues around how to interpret these results and the ambiguities therein. Providers express explicit concern over the how insurance will attempt to use this genetic information and the consequences around that. Many were concerned that

genetic testing might lead to insurance discrimination and seemed to lack trust in companies that offer genetic tests, such as 23andMe. These findings point to some of the attitudes and knowledge gaps among the providers that should be considered in the clinical implementation of genomic medicine for chronic conditions. A more targeted and improved training, guidelines, clinical tools, and awareness of patient protections might support the effective adoption of precision medicine by primary care providers.

Chapter 4: Methods of Genomic Analysis

4.1: Differential Expression Analysis

The study included 2611 participants of the COPDGene study as per the protocol outlined for COPDGene (EA Regan, 2011). In this study, self-identified Non-Hispanic Whites and African Americans between the ages of 45 and 80 years with a minimum of 10 pack-years lifetime smoking history (1 pack-year = 1 pack of cigarettes smoked daily for 1 year) were enrolled at 21 centers across the United States. Subjects returned for a second study visit approximately 5 years after initial enrollment. At this point they completed detailed questionnaires, pre- and post- bronchodilator spirometry, volumetric computed tomography of the chest, and provided blood for complete blood counts (CBCs) with differentials and RNA sequencing. All subjects enrolled in the trial were cancer-free at time of initial study enrollment.

Smoking history was ascertained by filling out self-reporting questionnaires. Participants were classified as current and former smokers based on their self-report. Sequenced subjects included COPD cases classified based on GOLD staging. The

GOLD staging was done according to available computed tomography and spirometry analysis which classified the subjects into GOLD stage 2, 3 or 4 (Vogelmeier et al., 2017). Smokers with normal lung function were given classification of GOLD stage 0 or 1. Institutional review board approval and written informed consent was obtained for all subjects.

4.1.1: RNA extraction

RNA-seq data was taken from peripheral blood samples. Total RNA was extracted from PAXgene™. Blood RNA tubes using the Qiagen PreAnalytiX PAXgene Blood miRNA Kit (Qiagen, Valencia, CA). The extraction protocol was performed either manually or with the Qiagen QIAcube extraction robot according to the company's standard operating procedure. Extracted RNA samples with RIN > 7 and concentration >=25 µg/ul were sequenced.

4.1.2: cDNA library preparation and sequencing

Globin reduction and cDNA library preparation for total RNA was performed with the Illumina TruSeq Stranded Total RNA with Ribo-Zero Globin kit (Illumina, Inc., San Diego, CA). Libraries were QCed by quantification with Picogreen, size analysis on an Agilent Bioanalyzer or Tapestation 2200 (Agilent, Santa Clara, CA) and qPCR quantitation against a standard curve. 75 bp paired end reads were generated on a HiSeq 2500 flow cell. Libraries are loaded at an empirically determined concentration in order to generate the optimal number of clusters per lane of the flow cell. Samples were sequenced to an average depth of 20 million reads.

4.1.3: Read alignment, expression quantification, and sequencing quality control

Reads were trimmed of TruSeq adapters using Skewer with default parameters (Jiang et al., 2014). Trimmed reads were aligned to the GRCH38 genome using the

STAR aligner version 2.4.0 (Dobin et al., 2013). Gene and exon level counts were generated using RSubreads (Liao et al., 2013) with the Ensembl version 81 annotation . Quality control was performed using the FastQC (*Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*, n.d.; Kersey et al., 2016) and RNA-SeQC programs (D. S. DeLuca et al., 2012). Samples were included for subsequent analysis if they had >10 million total reads, >80% of reads mapped to the reference genome, XIST and Y chromosome expression was consistent with reported gender, <10% of R1 reads in the sense orientation, Pearson correlation >= 0.9 with samples in the same library construction batch, and concordant genotype calls between variants called from RNA sequencing reads and DNA genotyping. The gene count data used for this analysis are available in GEO (Barrett et al., 2013) (accession number GSE9753).

4.1.4: Technical covariates

In order to remove unwanted batch effects and confounders, we applied SVAseq (Leek, 2014) to the gene or exon count matrices. Surrogate variables (SVs) were estimated while specifying the following covariates: age, gender, race, pack-years of smoking history, library construction batch and cell count percentages.

4.1.5: Gene-level differential expression analyses

Differential gene expression analysis was performed using the voom (Law et al., 2014) /limma (Smyth, 2005; Ritchie et al., 2015) R package. Transcripts that were expressed at >= 1 count per million mapped reads in >= 10 subjects were used for analyses. The variables controlled for were age, gender, race, pack-years of smoking history, monocyte percentage, lymphocyte percentage, eosinophil percentage, neutrophil percentage, library construction batch, and significant SVs (n = 27).

Differentially expressed genes were defined with as those with an empirical Bayes corrected p-value <0.05. Genes were considered significant if their Bonferroni corrected p-value was <0.05 (corrected for the number of differentially expressed genes).

4.1.6: Gene ontology (GO) enrichment analyses

To identify gene sets over or under-represented in differentially expressed genes, we performed GO gene ontology enrichment analyses (Ashburner et al., 2000; Mi et al., 2013). Analysis input included all significant differentially expressed genes, and queries included gene sets in the “biological processes” ontology (database version released 2017–01-26). Significant gene sets were defined as those with a Bonferroni corrected p value <0.05.

Subjects with a COPD GOLD stage of 2, 3, and 4 (Moderate, Severe and Very Severe disease respectively) were considered as “Case”. Those with a GOLD stage of 0 were controls. Subjects with a GOLD stage of 1 (Mild disease) were not used for the purpose of this analysis.

Covariates used were sex, race, age, number of pack-years smoked and cell count percentage. A surrogate variable based analysis was performed.

A differential pathway analysis was performed between the Severe vs Control and Case vs Control group (elaborate). A Protein-Protein interaction network was constructed using the STRING database using default parameters. The STRING database is a curated database that consists of known and predicted protein-protein interactions (Szklarczyk et al., 2017; von Mering et al., 2003). The interactions include physical and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases .

4.2: Weighted Gene Correlation Networks

Out of the original gene set used for the differential expression analysis, 3000 genes were selected at random for ease of computation. All 2611 samples were considered for the analysis. The WGCNA package for co-expression analysis was used via Bioconductor (http://bioconductor.org/biocLite.R)(Langfelder & Horvath, 2008; Dai et al., 2018) . The soft threshold method for Pearson correlation analysis of the expression profiles was used to determine the connection strengths between among the genes in order to construct a weighted correlation network using default parameters. Average linkage hierarchical clustering was carried out to group genes based on topological overlap dissimilarity which in turn was used to determine network connection strengths. Genes were grouped into different "modules" based on their correlation profile. To ensure we got an accurate determination of number of modules and clarify gene interaction, we set the restricted minimum gene number to 30 for each module and used the default threshold to merge the similar modules.

4.2.1: Correlation of Modules with clinically significant traits

Modules were then correlated with phenotype or clinical traits. Module eigengenes (MEs) are the first principal component of genes in a given module which have the same expression profile. The relationship between MEs and clinical traits was hence analyzed clinically relevant modules were identified.

A log10 transform of the p-value from the linear regression between gene expression and clinical stage was then performed, which was defined as gene significance. Average gene significance in a module was defined as module significance.

4.2.2: Hub Genes

Out of the genes studied, there were certain genes in each specific module that had a high degree of connectivity. A high degree of connectivity meant that a number of other genes connected or interacted with this gene in some way. These genes were termed as “Hub” genes and were further analyzed for functionality. A measure of intramodular connectivity was used to classify the genes as “Hub” genes. Intramodular connectivity is a measure of the connectivity of nodes to other nodes within the same module. A gene that has several such connections is termed as a “Hub” gene. We analyzed each module for a Hub gene and thus got 14 Hub genes. A GO Ontology analysis was also performed on the set of Hub genes.

4.2.3: Verification of Co-expression network

In order to corroborate the findings on co-expression of genes in the co-expression network created through WGCNA, we ran the genes of the two most populous clinically relevant modules through GeneMania which searches various databases checking for different types of gene interactions including co-expression (Warde-Farley et al., 2010; Franz et al., 2018).

Chapter 5: Results and Discussion

5.1: Results

There were 44 DE genes in Case vs Control and 45 DE genes in Severe vs Control. Out of these, there were 14 common differentially expressed genes.

Out of these, the gene C1QB and microRNA MIR4433B had highest fold-change in Case. In the Severe group, gene C1QB, RAP1GAP and microRNA MIR4477B had highest fold-change. The GPR15 gene was found to be significant only in Case vs Control group and not in the Severe vs Control group.

Case-Control	genes_0.1	genes_0.05	genes_0.01	top_genes
No cell count, NO SVA	5414	3774	1892	LMO2,UBE2J1,ZNF831,LINC01410,ZNF117,MIR4477B,ST3GAL6,METTL9,TLR2,TNRC6C
finalGold_DE_casecontrol_5_SVA	123	74	28	MIR4477B,C9orf156,CDC42EP3,UBE2J1,LMO2,LINC01410,LYPLAL1,IL10RA,SLC38A1,SMARCD1
finalGold_DE_casecontrol_9	629	259	10	GPR15,UBE2J1,ZNF117,LMO2,SON,MIR4477B,ZNF831,SMARCD1,URB2,ADGRG3
finalGold_DE_casecontrol_9_SVA	83	44	11	C9orf156,MIR4477B,UBE2J1,CDC42EP3,EXTL3,SMARCD1,ADGRG3,LINC01410,LYPLAL1,AFF4

Analysis	genes_0.1	genes_0.05	genes_0.01	top_genes
finalGold_DE_casecontrol_5	7318	5527	3171	ST3GAL6,TLR2,RNF175,ZNF831,PDLIM1P4,ZNF117,IKZF3,CLEC4E,HGF,MDN1
finalGold_DE_casecontrol_5_SVA	210	122	33	ABHD2,IL10RA,UBE2J1,ATG2A,MIR4477B,ST3GAL6,RNF175,HIF1A,ZNF831,LMO2
finalGold_DE_casecontrol_9	83	25	0	ST3GAL6,ZNF117,ZNF831,ITGB8,PTGER3,HIVEP2,CPA3,LYPLAL1,TLR2,ADAM17
finalGold_DE_casecontrol_9_SVA	86	45	7	ABHD2,ASXL2,UBE2J1,IL10RA,MIR4477B,C1QB,HIF1A,ATG2A,SLC38A1,MGAT5

Table 5.1.1: Differentially Expressed Genes

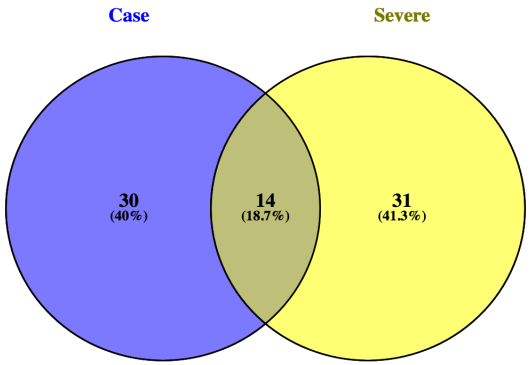


Figure 5.1.1: Number of genes distributed in Case vs Control

Common DE Genes between the two groups
MIR4477B
UBE2J1
EXTL3
ADGRG3
LYPLAL1
IL10RA
SMAP1
C1QB

FAM20A
ASXL2
SLC38A1
LMO2
ATG2A
PAPOLG

Table 5.1.2: Differentially expressed genes common to Case and Control groups

The Significance was based on adjusted P-value of < 0.05. We found 44 DE genes in overall Case group and 45 DE genes in Severe group. There were 14 common DE genes across both groups. C1QB and MIR4433B had highest fold-change in Case group whereas C1QB, MIR4477B and RAP1GAP had highest fold-change in Severe group. GPR15 which is a gene associated with current smokers, was found to be significant only in Case-Control and not in the Severe group.

fitIndices.Power	fitIndices.SFT.R.sq	fitIndices.slope	fitIndices.truncated.R.sq	fitIndices.mean.k.	fitIndices.median.k.	fitIndices.max.k.
1	0.09692486	-0.3109533	0.78514066	407.304236	400.076521	849.019998
2	0.73674297	-1.0393783	0.90732516	108.702444	88.1522211	375.490393
3	0.76128852	-1.2771187	0.91157141	41.074552	25.0127815	207.231717
4	0.82797942	-1.384142	0.95544991	18.9805342	8.7285353	128.817038
5	0.82175249	-1.5094822	0.94923403	9.97072418	3.48399991	85.5059327
6	0.83984541	-1.5519339	0.96354237	5.72761877	1.4644304	59.2876435
7	0.84608123	-1.5805503	0.9693178	3.51945456	0.96089618	42.4189911
8	0.8704053	-1.6024104	0.97769157	2.28277365	0.47970425	31.0830368
9	0.88505795	-1.6249594	0.98249484	1.54990781	0.24128949	23.2107806
10	0.90294211	-1.6414386	0.98720472	1.09556916	0.13048439	17.6008674
11	0.9196701	-1.6483681	0.98757737	0.80327694	0.06847941	13.5185738
12	0.94368142	-1.6284375	0.99162385	0.6092979	0.03735617	10.495913
13	0.89083134	-1.4099619	0.9713999	0.47709895	0.02083579	8.22484367
14	0.84714596	-1.4188625	0.92557064	0.38490505	0.01156865	6.58521849
15	0.84308542	-1.4085884	0.91713012	0.3192978	0.00641215	5.57668678

Table 5.1.3: Computing the beta threshold

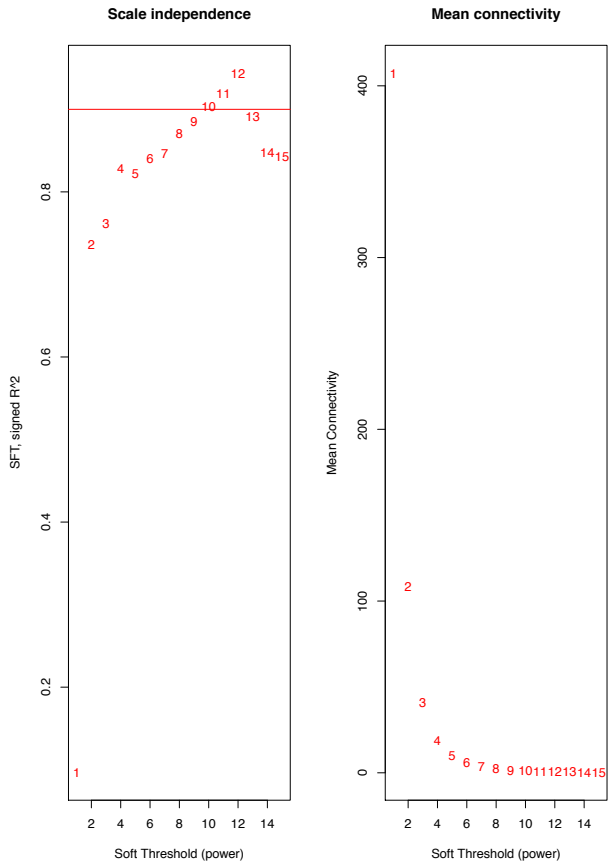


Figure 5.1.2: Computing the beta threshold

Given the above figure and table we looked at both the scale independence as well as mean connectivity, we selected the beta power where it plateaus off at the lowest possible threshold. In this case, it was found to be at around β=6 and hence this is the power that was chosen for the soft-threshold.

According to Horvath et al., selecting the beta soft threshold this way would give the most accurate results.

The network was then scaled and raised to this power in order to approach a scale-free topology model for the gene network. Out of the 3000 genes, 228 genes were removed due to presence of too many missing values or zero variance. The final sample gene network was based on squared Euclidean distance. Samples were

designated as outliers if their value was below the threshold. Initially, a Cluster tree of COPD subjects was formed. The leaves of the tree correspond to the subjects. The first color band underneath the tree indicates which subjects appear to be outlying.

The Module eigengene is defined as the first principal component of the expression matrix of the corresponding module. However, the calculation may fail if the expression data has too many missing entries. Since the module eigengene is an optimal summary of the gene expression profiles of a given module, the eigengene can be correlated with traits and to look for the most significant associations.

A total of fourteen modules were discovered corresponding to various clinical traits. The Euclidian distance between the modules was computed as well as those modules falling below the minimum threshold for Euclidian distance were merged into one module. The figure below gives the various modules:

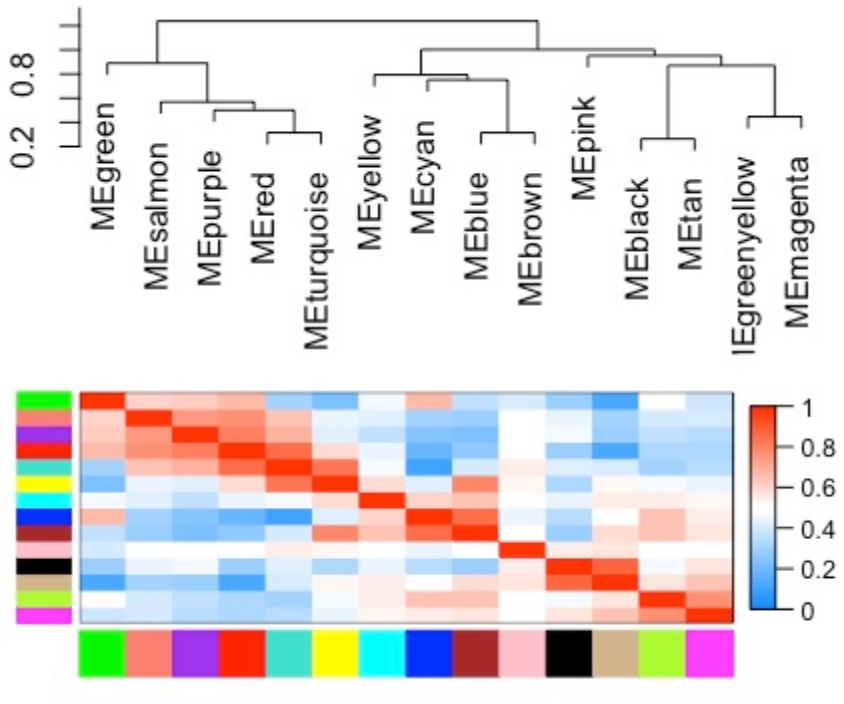


Figure 5.1.3: Modules

The table below gives the percent gene table among the modules which is an indication of gene-module membership and the number of genes belonging to each module.

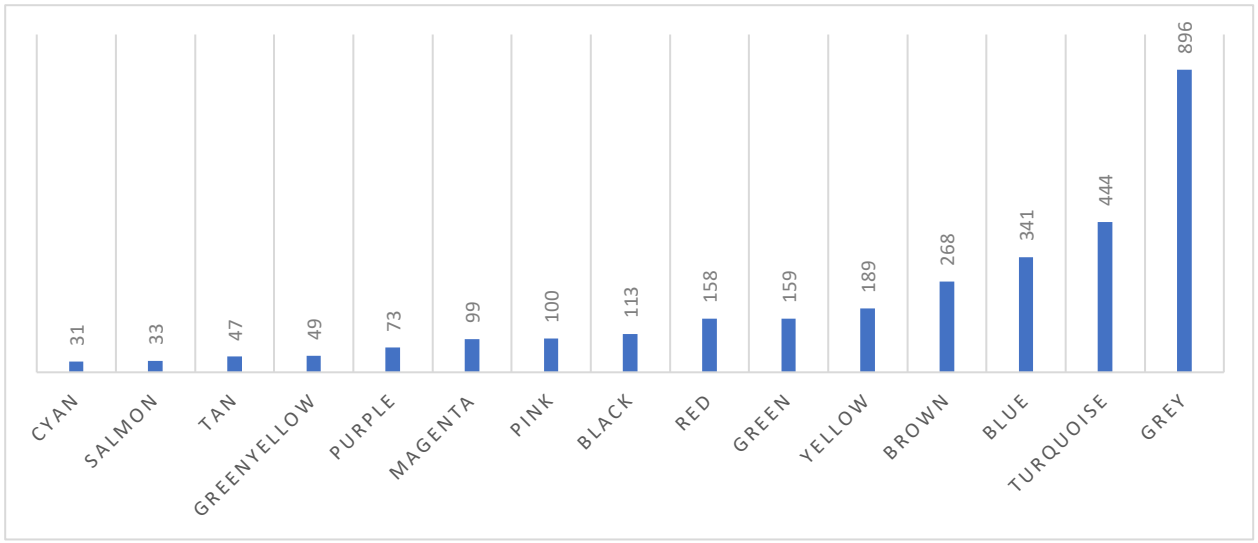


Figure 5.1.4: Gene distribution across modules

From the above table, we see that the most number of genes are in the Blue and Brown model since both the Turquoise and Grey modules are considered as ‘junk modules’. The Consensus gene dendrogram gives us a visual idea of which genes are belong to which module.

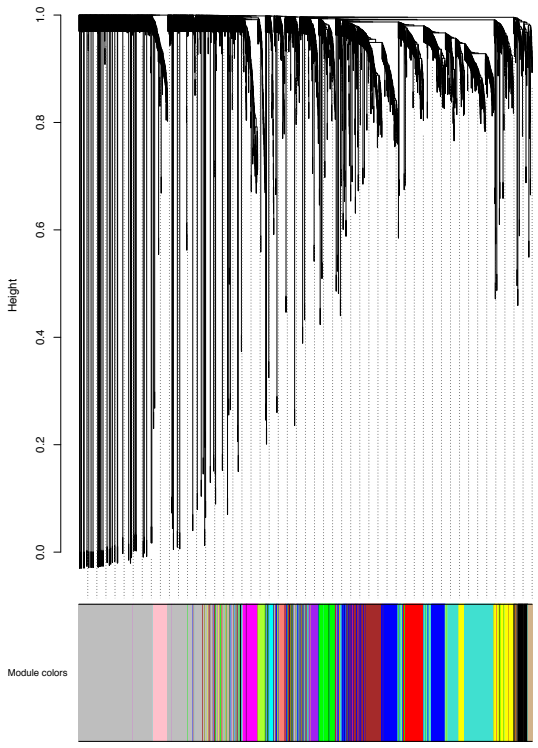


Figure 5.1.5: Consensus Dendrogram

The module-trait correlation was then performed, and we found that there were a significant number of DE genes corresponding to different traits.

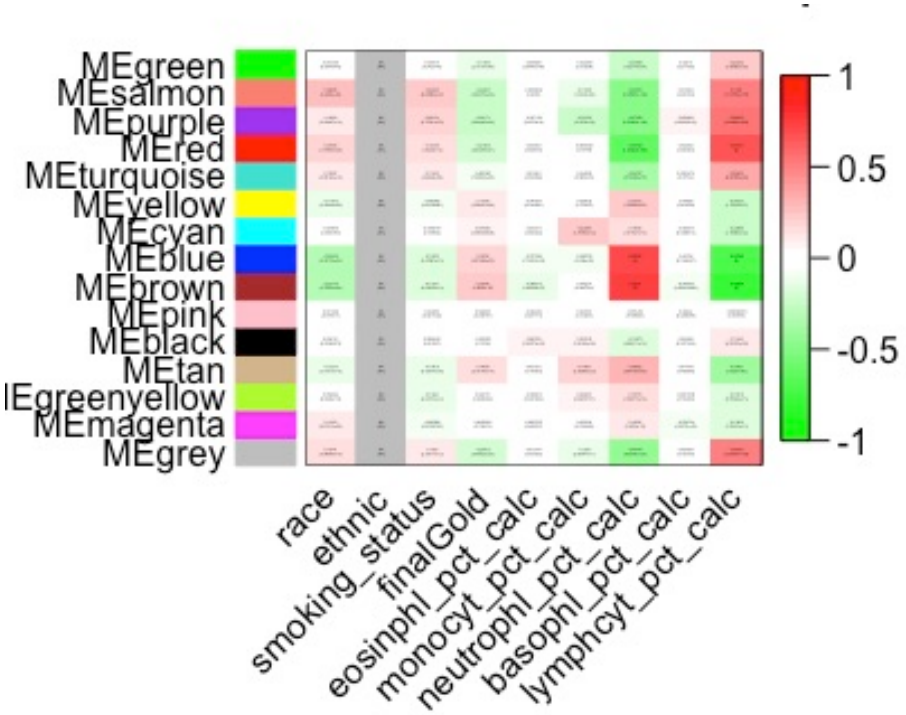


Figure 5.1.5: Module-Trait relationship

We then constructed a PPI to see if the gene correlation was paralleled at the protein level.

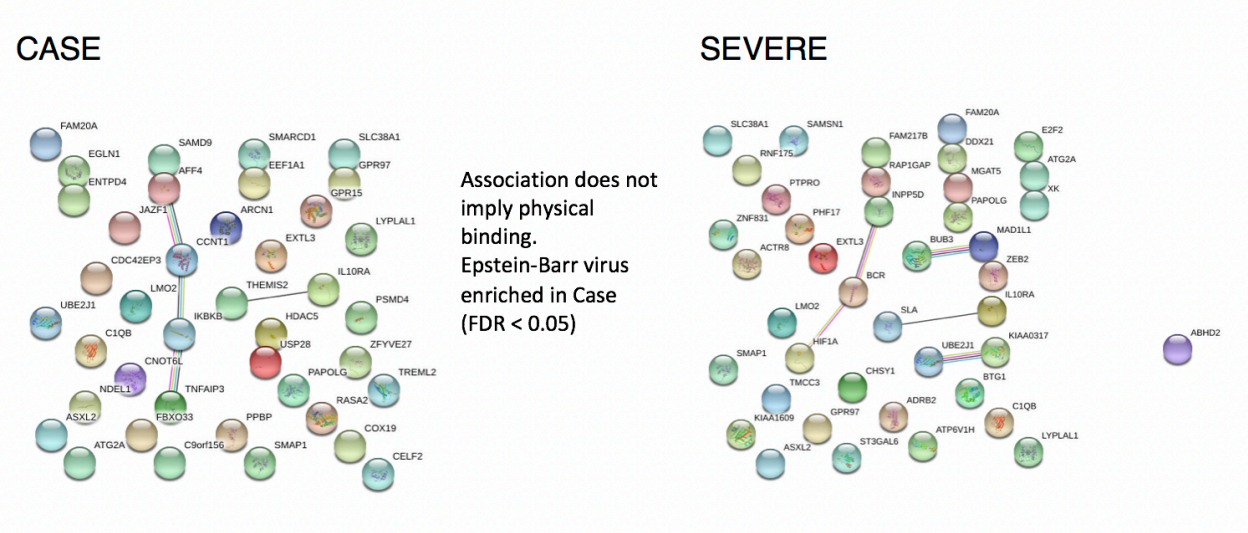


Figure 5.1.6: Protein-Protein Interaction Network between Case and Control groups

In order to validate the co-expression network, we mined the genes against various databases. From our results, we concluded that the network did indeed have a high degree of co-expression thereby validating our network.

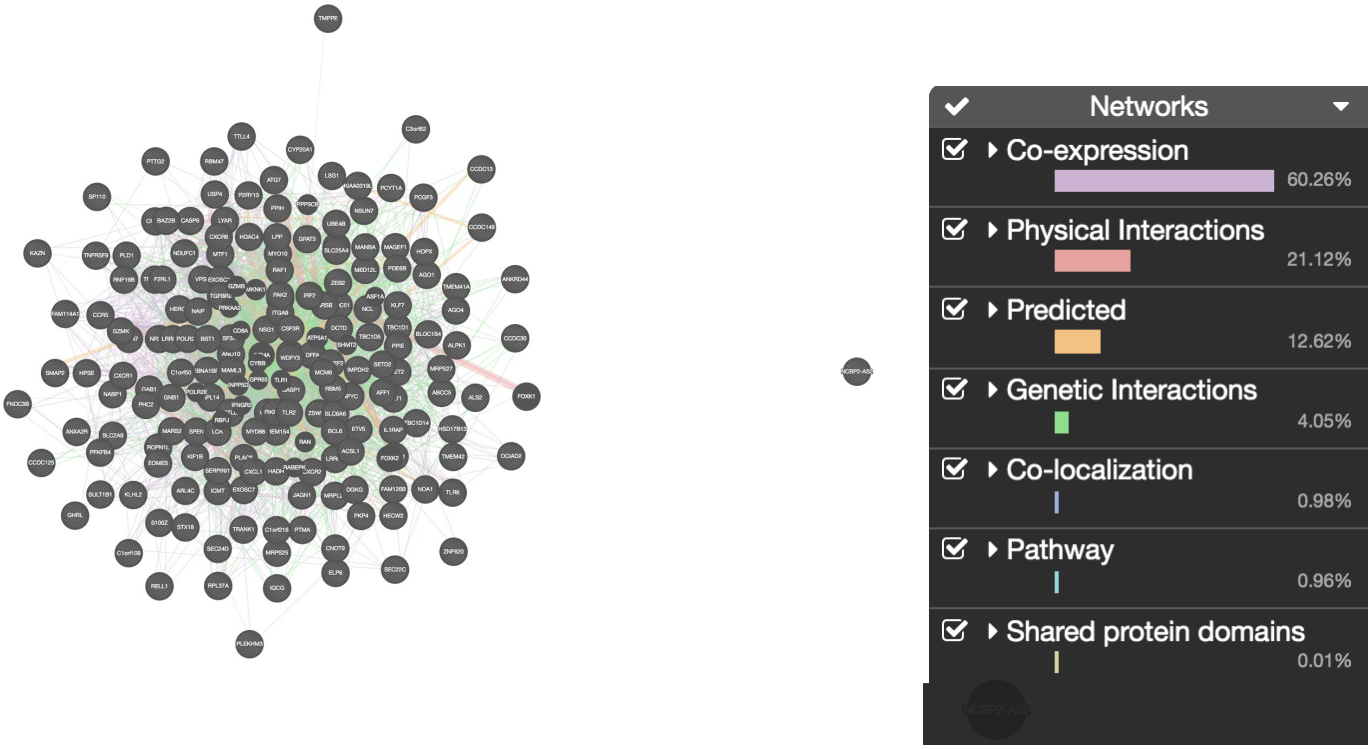


Figure 5.1.7: Validation of Co-expression Network using GeneMania

Module Color	Hub genes-Ensembl Gene Symbol	Gene Symbol	Name	Function
black	ENSG00000164062	APEH	acylaminoacyl-peptide hydrolase	This gene encodes the enzyme acylpeptide hydrolase, which catalyzes the hydrolysis of the terminal acetylated amino acid preferentially from small acetylated peptides. Associations with lung and renal carcinomas.
blue	***ENSG00000163625***	***WDFY3***	***WD repeat and FYVE domain containing 3-Master TF***	***Master Transcription Factor. This gene encodes a phosphatidylinositol 3-phosphate-binding protein that functions as a master conductor for aggregate clearance by autophagy.***
brown	ENSG00000091317	CKLF	CKLF like MARVEL transmembrane domain containing 6.	The product of this gene is a cytokine. This protein may play important roles in inflammation and in the regeneration of skeletal muscle.
cyan	ENSG00000115267	IFIH1	interferon induced with helicase C domain 1.	Encoding protein MDA5. MDA5 has been implicated in autoimmune and autoinflammatory diseases such as type 1 diabetes, systemic lupus erythematosus, and Aicardi-Goutieres syndrome. Note: This gene has been reviewed for its involvement in coronavirus biology, and is involved in immune response or antiviral activity. This gene has been reviewed for its involvement in coronavirus biology, and is involved in immune response or antiviral activity.
green	ENSG00000048707	VPS13D	vacuolar protein sorting 13 homolog D	This gene encodes a protein belonging to the vacuolar-protein-sorting-13 gene family.
greenyellow	ENSG00000168497	CAVIN2	caveolae associated protein 2	Cancer implications-metastasis supressor
magenta	ENSG00000125037	EMC3	ER membrane protein complex subunit 3	transmembrane protein
pink	ENSG00000187653	TMSB4XP8	TMSB4X pseudogene 8	implicated in ubiquitin biosynthesis
purple	ENSG00000138795	LEF1	lymphoid enhancer binding factor 1-TF	This gene encodes a transcription factor belonging to a family of proteins that share homology with the high mobility group protein-1.
red	ENSG00000114353	GNAI2	G protein subunit alpha i2	The encoded protein contains the guanine nucleotide binding site and is involved in the hormonal regulation of adenylate cyclase.
salmon	ENSG00000153064	BANK1	B cell scaffold protein with ankyrin repeats 1	The protein encoded by this gene is a B-cell-specific scaffold protein that functions in B-cell receptor-induced calcium

				mobilization from intracellular stores. This protein can also promote Lyn-mediated tyrosine phosphorylation of inositol 1,4,5-trisphosphate receptors. Polymorphisms in this gene are associated with susceptibility to systemic lupus erythematosus.
tan	ENSG00000010256	UQCRC1	ubiquinol-cytochrome c reductase core protein 1	Enables ubiquitin protein ligase binding activity. Predicted to be involved in oxidative phosphorylation and mitochondrial electron transport, ubiquinol to cytochrome c. Located in mitochondrion. Implicated in Alzheimer's disease. Biomarker of Alzheimer's disease.
turquoise	ENSG00000164163	ABCE1	ATP binding cassette subfamily E member 1	The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intra-cellular membranes.
yellow	ENSG00000196504	PRPF40A	pre-mRNA processing factor 40 homolog A	Enables RNA binding activity

Table 5.1.4: Hub Genes distribution across modules

5.2: Discussion

There were 30 DE genes in Case vs Control and 31 in Severe vs Control and 14 common genes between the two groups. However, the gene associated with former smokers, namely GRP15 was not found to be significant in either groups. This was interesting to note because it was initially expected that this gene would be significantly expressed in at least severe group. The fact that it was not showed that perhaps population that fell into the “former smokers” category and have not quit smoking may not be fully accorded the protection that comes with their current smoking status i.e the fact that they have now quit smoking.

Among the top percentage of upregulated genes, it was interesting to note that the terpenoid biosysthesis pathway and the NOTCH signaling pathway were highlighted.

5.2.1: Limitations

WGCNA was written originally for microarray data which is normally distributed. However, the Rna-seq data we used is based on actual counts. Therefore, instead of using FPKM values, Limma requires raw counts as input data. Raw counts are the number of reads overlapping a given gene. In the Limma pipeline used, the VOOM package was first used to normalize the raw counts. When these raw counts which were normalized were used as input for WGCNA, it could potentially could cause problems. This is due to the statistical distribution on which Voom is based and using which it does its transformation of the raw expression matrix.

Hence in order to make it model agnostic, the raw expression matrix was used.

5.2.2: Hub Genes

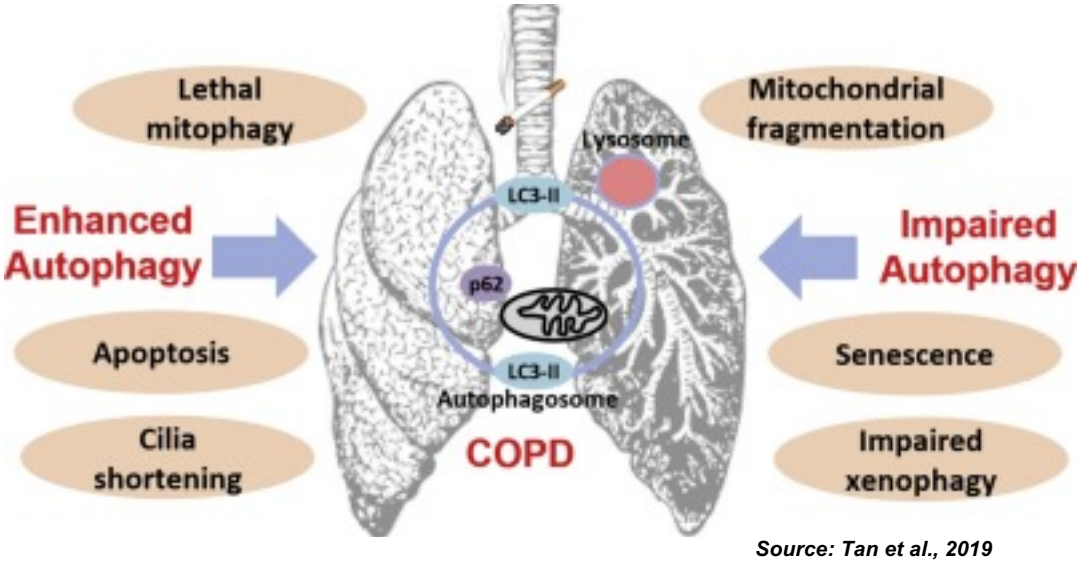
A hub gene or hub node is a gene that is highly connected to other genes and is also centrally located in the module. Hub nodes have been found to play important roles in networks. Similarly, highly connected hub genes are also expected to play an important role as far as biological networks are concerned. Knock-out studies have also shown that hub proteins are found to play an important role in microorganisms.

Focusing on intramodular hubs leads to key findings in trait analysis. (Langfelder & Horvath, 2008)(Langfelder et al., 2013). Hubs that are located within a module i.e intramodular hubs in disease related modules are often of clinical importance (Carlson et al., 2006) (Ivliev et al., 2010). Hubs can also be used to find genes of importance associated with the various stages of COPD (Di et al., 2019).

In this analysis, it was found that the among the hub genes in the various modules, there was one that was a master transcription factor (TF). A master TF is also known as a “master regulator” or “master regulatory gene.” This term was first proposed by Susumu Ohno for a “gene that occupies the very top of a regulatory hierarchy,” and is not controlled by any other gene (Chan & Kyba, 2013). These “master TFs” have the potential to be excellent drug targets and are thus of great interest from a pharmacological and therapeutic standpoint.

Among the fourteen modules with a hub gene corresponding to each module, it was found that there was one hub gene that was a master TF. It was interesting to note that this gene was involved in regulation of autophagy. Autophagy is a process conserved evolutionarily which is responsible for maintaining cellular homeostasis. It achieves this via degradation of damaged protein, lipid and various cellular organelles (Tan et al., 2019). Autophagy is disrupted by cigarette smoking, which is an oft

implicated cause in many cases of COPD. Autophagy plays an important role in COPD and this mechanism is of therapeutic interest in management of COPD.



Source: Tan et al., 2019

Figure 5.2.1: COPD and Autophagy

The NOTCH signaling pathway was also found to be differently expressed. This is yet another area of interest as NOTCH has been implicated in various mechanisms of cancer. There are studies that link disrupted NOTCH signaling with cancer (Allenspach et al., 2002). Further analysis would be required to see if these pathways are getting triggered in patients with COPD to study if COPD exacerbations could have implications in cancer and tumorigenesis. Given that NOTCH is being increasingly implicated in lungs diseases, this avenue would be of interest for further study.

Chapter 6: Conclusion and Future Direction

COPD like many other lung diseases thus has a varied etymology and treatment options and management would vary based on an individual's situation. This genomic analysis provides a deeper and more comprehensive understanding of the genetic underpinnings of this disease. A better understanding at the genetic level including understanding how genes and their corresponding pathways are differentially regulated would help to explore new avenues for treatment as well as diagnosis of COPD.

Further analysis could potentially incorporate machine learning (ML) and artificial intelligence (AI) methods to draw out patterns of interest. A wide range of applications from using images to differentially regulated gene data for the ML models is of great interest. Studies have shown that a combination of genomic analysis data along with AI/ML technology can be useful in adopting a precision medicine based approach for the benefit of patients (Feng et al., 2021)(Pirooznia et al., 2021).

Adopting more precise individualized treatments can reduce over or undertreatment caused by errors that can occur due to human error in a clinical setting. Use of ML models in studying patterns in COPD and other lung disease while having a lot of potential, is hampered by issues such as missing or messy data sets. Adopting approaches to surmount these problems would be of immense benefit paving the way for new ways to not only diagnose but also treat and manage COPD.

Bibliography

1. Allenspach, E. J., Maillard, I., Aster, J. C., & Pear, W. S. (2002). Notch signaling in cancer. *Cancer Biology & Therapy*, *1*(5), 466–476. https://doi.org/10.4161/cbt.1.5.159
2. Artin, M. G., Stiles, D., Kiryluk, K., & Chung, W. K. (2019). Cases in Precision Medicine: When Patients Present With Direct-to-Consumer Genetic Test Results. *Annals of Internal Medicine*, *170*(9), 643–650. https://doi.org/10.7326/M18-2356
3. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556
4. *Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data*. (n.d.). Retrieved April 16, 2023, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
5. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991-995. https://doi.org/10.1093/nar/gks1193
6. Brown-Johnson, C. G., Safaeinili, N., Baratta, J., Palaniappan, L., Mahoney, M., Rosas, L. G., & Winget, M. (2021). Implementation outcomes of Humanwide: Integrated precision health in team-based family practice primary care. *BMC Family Practice*, *22*(1), 28. https://doi.org/10.1186/s12875-021-01373-4

7. Canzoneri, R., Lacunza, E., & Abba, M. C. (2019). Genomics and bioinformatics as pillars of precision medicine in oncology. *Medicina*, *79*(Spec 6/1), 587–592.
8. Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., & Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: Predictions from modular yeast co-expression networks. *BMC Genomics*, *7*(1), 40. https://doi.org/10.1186/1471-2164-7-40
9. Chan, S. S.-K., & Kyba, M. (2013). What is a Master Regulator? *Journal of Stem Cell Research & Therapy*, *3*, 114. https://doi.org/10.4172/2157-7633.1000e114
10. Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from large-scale biology. *Science (New York, N.Y.)*, *300*(5617), 286–290. https://doi.org/10.1126/science.1084564
11. Dai, R., Xia, Y., Liu, C., & Chen, C. (2018). CsuWGCNA: a combination of signed and unsigned WGCNA to capture negative correlations. *BioRxiv*. https://doi.org/10.1101/288225
12. DeLuca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., & Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics (Oxford, England)*, *28*(11), 1530–1532. https://doi.org/10.1093/bioinformatics/bts196
13. DeLuca, J., Selig, D., Poon, L., Livezey, J., Oliver, T., Barrett, J., Turner, C., & Hellwig, L. (2020). Toward Personalized Medicine Implementation: Survey of Military Medicine Providers in the Area of Pharmacogenomics. *Military Medicine*, *185*(3–4), 336–340. https://doi.org/10.1093/milmed/usz419

14. Di, Y., Chen, D., Yu, W., & Yan, L. (2019). Bladder cancer stage-associated hub genes revealed by WGCNA co-expression network analysis. *Hereditas 2019 156:1*, *156*(1), 1–11. https://doi.org/10.1186/S41065-019-0083-Y

15. Diaz-Guzman, E., & Mannino, D. M. (2014). Epidemiology and Prevalence of Chronic Obstructive Pulmonary Disease. *Clinics in Chest Medicine*, *35*(1), 7–16. https://doi.org/10.1016/j.ccm.2013.10.002

16. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

17. Duffy, S. P., & Criner, G. J. (2019). Chronic Obstructive Pulmonary Disease: Evaluation and Management. *The Medical Clinics of North America*, *103*(3), 453–461. https://doi.org/10.1016/j.mcna.2018.12.005

18. EA Regan, J. H., JR Murphy. (2011). Genetic epidemiology of COPD (COPDgene) study design. *Epidemiology*, *7*(1), 1–10. https://doi.org/10.3109/15412550903499522.genetic

19. Feng, Y., Wang, Y., Zeng, C., & Mao, H. (2021). Artificial Intelligence and Machine Learning in Chronic Airway Diseases: Focus on Asthma and Chronic Obstructive Pulmonary Disease. *International Journal of Medical Sciences*, *18*(13), 2871–2889. https://doi.org/10.7150/ijms.58191

20. Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., & Morris, Q. (2018). GeneMANIA update 2018. *Nucleic Acids Research*, *46*(W1), W60–W64. https://doi.org/10.1093/nar/gky311

21. Gameiro, G., Sinkunas, V., Liguori, G., & Auler-Júnior, J. (2018). Precision Medicine: Changing the way we think about healthcare. *Clinics*, *73*. https://doi.org/10.6061/clinics/2017/e723

22. Green, E. D., Watson, J. D., & Collins, F. S. (2015). Human Genome Project: Twenty-five years of big biology. *Nature*, *526*(7571), 29–31. https://doi.org/10.1038/526029a

23. Gutierrez, M. E., Choi, K., Lanman, R. B., Licitra, E. J., Skrzypczak, S. M., Pe Benito, R., Wu, T., Arunajadai, S., Kaur, S., Harper, H., Pecora, A. L., Schultz, E. V., & Goldberg, S. L. (2017). Genomic Profiling of Advanced Non-Small Cell Lung Cancer in Community Settings: Gaps and Opportunities. *Clinical Lung Cancer*, *18*(6), 651–659. https://doi.org/10.1016/j.cllc.2017.04.004

24. Haspel, R. L., Genzen, J. R., Wagner, J., Fong, K., Group, U. T. in G. (UTRIG) W., & Wilcox, R. L. (2021). Call for improvement in medical school training in genetics: Results of a national survey. *Genetics in Medicine*, *23*(6), 1151–1157. https://doi.org/10.1038/s41436-021-01100-5

25. Hauser, D., Obeng, A. O., Fei, K., Ramos, M. A., & Horowitz, C. R. (2018). Views Of Primary Care Providers On Testing Patients For Genetic Risks For Common Chronic Diseases. *Health Affairs (Project Hope)*, *37*(5), 793–800. https://doi.org/10.1377/hlthaff.2017.1548

26. Hersh, C. P. (2019). Pharmacogenomics of chronic obstructive pulmonary disease. *Expert Review of Respiratory Medicine*, *13*(5), 459–470. https://doi.org/10.1080/17476348.2019.1601559

27. Ivliev, A. E., ’t Hoen, P. A. C., & Sergeeva, M. G. (2010). Coexpression Network Analysis Identifies Transcriptional Modules Related to Proastrocytic Differentiation and

Sprouty Signaling in Glioma. *Cancer Research*, *70*(24), 10060–10070. https://doi.org/10.1158/0008-5472.CAN-10-2465

28. Jäger, N. (2022). Bioinformatics workflows for clinical applications in precision oncology. *Seminars in Cancer Biology*, *84*, 103–112. https://doi.org/10.1016/j.semcancer.2020.12.020
29. Jiang, H., Lei, R., Ding, S.-W., & Zhu, S. (2014). Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, *15*, 182. https://doi.org/10.1186/1471-2105-15-182
30. Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Humphrey, J., Kerhornou, A., Khobova, J., Aranganathan, N. K., Langridge, N., Lowy, E., McDowall, M. D., Maheswari, U., Nuhn, M., … Staines, D. M. (2016). Ensembl Genomes 2016: More genomes, more complexity. *Nucleic Acids Research*, *44*(D1), D574-580. https://doi.org/10.1093/nar/gkv1209
31. Khoury, M. J., Gwinn, M. L., Glasgow, R. E., & Kramer, B. S. (2012). A population approach to precision medicine. *American Journal of Preventive Medicine*, *42*(6), 639–645. https://doi.org/10.1016/j.amepre.2012.02.012
32. Labaki, W. W., & Rosenberg, S. R. (2020). Chronic Obstructive Pulmonary Disease. *Annals of Internal Medicine*, *173*(3), ITC17–ITC32. https://doi.org/10.7326/AITC202008040
33. Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-559

34. Langfelder, P., Mischel, P. S., & Horvath, S. (2013). When Is Hub Gene Selection Better than Standard Meta-Analysis? *PLoS ONE*, *8*(4), e61505. https://doi.org/10.1371/journal.pone.0061505

35. Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

36. Leek, J. T. (2014). svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, *42*(21), e161. https://doi.org/10.1093/nar/gku864

37. Leung, J. M., Obeidat, M., Sadatsafavi, M., & Sin, D. D. (2019). Introduction to precision medicine in COPD. *European Respiratory Journal*, *53*(4), 1802460. https://doi.org/10.1183/13993003.02460-2018

38. Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, *41*(10), e108. https://doi.org/10.1093/nar/gkt214

39. Llanto, K., Lim, F., & Ea, E. (2020). Impact of the All of Us research program. *Nursing*, *50*(3), 67–68. https://doi.org/10.1097/01.NURSE.0000654172.46117.18

40. Massart, M., Berenbrok, L. A., Munro, C., Berman, N. R., & Empey, P. E. (2022). A Multidisciplinary Precision Medicine Service in Primary Care. *Annals of Family Medicine*, *20*(1), 88. https://doi.org/10.1370/afm.2764

41. Matthew, D. B. (2019). Two Threats to Precision Medicine Equity. *Ethnicity & Disease*, *29*(Suppl 3), 629–640. https://doi.org/10.18865/ed.29.S3.629

42. McCauley, M. P., Marcus, R. K., Strong, K. A., Visotcky, A. M., Shimoyama, M. E., & Derse, A. R. (2017). Genetics and Genomics in Clinical Practice: The Views of Wisconsin Physicians. *WMJ*, *116*(2), 69–74.

43. McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019). Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine*, *9*(11), a036798. https://doi.org/10.1101/cshperspect.a036798

44. McGowan, M. L., Settersten, R. A., Juengst, E. T., & Fishman, J. R. (2014). Integrating genomics into clinical oncology: Ethical and social challenges from proponents of personalized medicine. *Urologic Oncology*, *32*(2), 187–192. https://doi.org/10.1016/j.urolonc.2013.10.009

45. McGrath, S. P., Peabody, A. E., Walton, D., & Walton, N. (2021). Legal Challenges in Precision Medicine: What Duties Arising From Genetic and Genomic Testing Does a Physician Owe to Patients? *Frontiers in Medicine*, *8*, 663014. https://doi.org/10.3389/fmed.2021.663014

46. McGrath, S. P., Walton, N., Williams, M. S., Kim, K. K., & Bastola, K. (2019). Are providers prepared for genomic medicine: Interpretation of Direct-to-Consumer genetic testing (DTC-GT) results and genetic self-efficacy by medical professionals. *BMC Health Services Research*, *19*(1), 844. https://doi.org/10.1186/s12913-019-4679-8

47. Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092

48. Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413–435. https://doi.org/10.1007/s13353-011-0057-x

49. Pirooznia, M., Han, S., & Lee, R. S. (2021). Editorial: Machine Learning and Network-Driven Integrative Genomics. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/FGENE.2021.660201/FULL

50. *PRISMA guidelines*. (n.d.).

51. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

52. Rohrer Vitek, C. R., Abul-Husn, N. S., Connolly, J. J., Hartzler, A. L., Kitchner, T., Peterson, J. F., Rasmussen, L. V., Smith, M. E., Stallings, S., Williams, M. S., Wolf, W. A., & Prows, C. A. (2017). Healthcare provider education to support integration of pharmacogenomics in practice: The eMERGE Network experience. *Pharmacogenomics*, *18*(10), 1013–1025. https://doi.org/10.2217/pgs-2017-0038

53. Sankar, P. L., & Parker, L. S. (2017). The Precision Medicine Initiative's All of Us Research Program: An agenda for research on its ethical, legal, and social issues. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *19*(7), 743–750. https://doi.org/10.1038/gim.2016.183

54. Sethi, S., Oh, S., Chen, A., Bellinger, C., Lofaro, L., Johnson, M., Huang, J., Bhorade, S. M., Bulman, W., & Kennedy, G. C. (2022). Percepta Genomic Sequencing Classifier and decision-making in patients with high-risk lung nodules: A decision impact study. *BMC Pulmonary Medicine*, *22*(1), 26. https://doi.org/10.1186/s12890-021-01772-4

55. Silverman, E. (1998). Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am. J. Respir. Crit. Care Med.*, *157*, 1770–1778.

56. Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397–420). Springer-Verlag. https://doi.org/10.1007/0-387-29362-0_23

57. Stockley, R. A., Halpin, D. M. G., Celli, B. R., & Singh, D. (2019). Chronic Obstructive Pulmonary Disease Biomarkers and Their Interpretation. *American Journal of Respiratory and Critical Care Medicine*, *199*(10), 1195–1204. https://doi.org/10.1164/rccm.201810-1860SO

58. Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., & Von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, *45*(D1), D362–D368. https://doi.org/10.1093/NAR/GKW937

59. Tan, W. S. D., Shen, H.-M., & Wong, W. S. F. (2019). Dysregulated autophagy in COPD: A pathogenic process to be deciphered. *Pharmacological Research*, *144*, 1–7. https://doi.org/10.1016/j.phrs.2019.04.005

60. Vashistha, V., Poonnen, P. J., Snowdon, J. L., Skinner, H. G., McCaffrey, V., Spector, N. L., Hintze, B., Duffy, J. E., Weeraratne, D., Jackson, G. P., Kelley, M. J., & Patel, V. L. (2020). Medical oncologists' perspectives of the Veterans Affairs National Precision Oncology Program. *PLoS ONE [Electronic Resource]*, *15*(7), e0235861. https://doi.org/10.1371/journal.pone.0235861

61. Velmovitsky, P. E., Bevilacqua, T., Alencar, P., Cowan, D., & Morita, P. P. (2021). Convergence of Precision Medicine and Public Health Into Precision Public Health: Toward a Big Data Perspective. *Frontiers in Public Health*, *9*. https://www.frontiersin.org/articles/10.3389/fpubh.2021.561873

62. Vogelmeier, C. F., Criner, G. J., Martinez, F. J., Anzueto, A., Barnes, P. J., Bourbeau, J., Celli, B. R., Chen, R., Decramer, M., Fabbri, L. M., Frith, P., Halpin, D. M. G., Varela, M. V. L., Nishimura, M., Roche, N., Rodriguez-Roisin, R., Sin, D. D., Singh, D., Stockley, R., … Agustí, A. (2017). Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. *American Journal of Respiratory and Critical Care Medicine*, *195*(5), 557–582. https://doi.org/10.1164/rccm.201701-0218PP

63. von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). STRING: A database of predicted functional associations between proteins. *Nucleic Acids Research*, *31*(1), 258–261. https://doi.org/10.1093/nar/gkg034

64. Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, *38*(Web Server issue), W214-220. https://doi.org/10.1093/nar/gkq537

65. Wood, A. M., Tan, S. L., & Stockley, R. A. (2009). Chronic obstructive pulmonary disease: Towards pharmacogenetics. *Genome Medicine*, *1*(11), 112. https://doi.org/10.1186/gm112

66. Wynn, R. M., Adams, K. T., Kowalski, R. L., Shivega, W. G., Ratwani, R. M., & Miller, K. E. (2018). The Patient in Precision Medicine: A Systematic Review Examining Evaluations of Patient-Facing Materials. *Journal of Healthcare Engineering*, *2018*, e9541621. https://doi.org/10.1155/2018/9541621

67. Yabroff, K. R., Zhao, J., de Moor, J. S., Sineshaw, H. M., Freedman, A. N., Zheng, Z., Han, X., Rai, A., & Klabunde, C. N. (2020). Factors Associated With Oncologist Discussions of the Costs of Genomic Testing and Related Treatments. *Journal of the National Cancer Institute*, *112*(5), 498–506. https://doi.org/10.1093/jnci/djz173

68. Zipkin, R., Schaefer, A., Chamberlin, M., Onega, T., O'Malley, A. J., & Moen, E. L. (2021). Surgeon and medical oncologist peer network effects on the uptake of the 21-gene breast cancer recurrence score assay. *Cancer Medicine*, *10*(4), 1253–1263. https://doi.org/10.1002/cam4.3720

VITA

Anjana Ramnath is a master's graduate in Health Informatics from the MU Department of Health Management and Informatics. She has gained a bachelor's degree in Biochemistry from Randolph College and the University of Madras followed by a master's degree in Bioinformatics from the University of Madras in India. Her research interest lies in the fields of Bioinformatics as well as Health Informatics with domain knowledge spanning across a broad spectrum from Genomics to Precision Medicine to Predictive modeling and Epidemiology, with a specific focus on modeling and exploratory analysis of genomics with statistical, data mining, and machine learning methods. She has co-authored journal articles and was one of the National finalists at AMIA's Design Challenge to integrate Patient Reported Outcomes with Precision Medicine. In her master's thesis project, she applied various Genomic modeling and Network analysis methods to gain a more comprehensive understanding of Chronic Obstructive Pulmonary Disease.