# DEVELOPING MULTI-OMICS ANALYSIS PIPELINE TO IDENTIFY NOVEL DRUG REPURPOSING TARGETS FOR COPD

---

A Dissertation
Submitted to
The Temple University Graduate Board

---

In Partial Fulfillment
Of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Fang Wang
December 2020

Examining Committee Members:

Dr. Carlos A. Barrero, Advisory Chair, Department of Pharmaceutical Sciences
Dr. Daniel J. Canney, Department of Pharmaceutical Sciences
Dr. Ken Korzekwa, Department of Pharmaceutical Sciences
Dr. Jozef Madzo, External member, Coriell Institute for Medical Research

# ABSTRACT

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease characterized by breathlessness due to airflow obstruction. COPD is the third leading cause of death worldwide. So far, none of the existing pharmacological treatments for COPD can stop the progressive decline in lung function. Drug repurposing is the application of existing approved therapeutic compounds for new disease indications, which may reduce the cost and time of new drug development. So far, there is not any systematic multi-omics data integration for drug repurposing in COPD. The goal of this project is to develop a systems biology pipeline for the identification of biological-relevant gene targets with drug repurposing potential for COPD treatment using multi-omics integration.

Here we implemented a computational methodology to identify drug repurposing targets for COPD. We integrated multi-omics COPD data including, genome, transcriptome, proteome, metabolome, interactome data, and drug-target information. A distance-based network model was created to rank the potential candidate genes. Fifty genes were prioritized as COPD signature genes for their overall proximity to signature genes identified at all omics levels. Forty of them may be considered as “druggable” targets. Literature search reported CRCX4 – Plerixafor as one prioritized targets-gene pair for drug repurposing. The bone marrow stimulant Plerixafor is currently being evaluated for COPD treatment in clinical trials, suggesting that our pipeline is finding promising drug repurposing targets. Our work, for the first time, applied a systematic approach integrating multiple omics data to find drug repurposing targets for COPD.

# DEDICATION

This work is dedicated to my family,

for their unconditioned love and support.

# ACKNOWLEDGEMENTS

The PhD journey can be challenging in many ways. I sincerely appreciate the support and encouragement I have received from many people.

I am deeply grateful to my PhD advisor Dr. Carlos Barrero. He is an exceptional advisor. He always appreciated my work and gave constructive suggestions. He continuously encouraged me to grow my skills as a scientist, like my critical thinking and communication. He is a kind and sincere mentor who cared about my well-being. He is also a great friend. His passion for science and life inspired me every day.

I am very thankful to my dissertation committee members. Dr. Daniel Canney has been extremely supportive through various stages of my PhD program. The insightful feedback and scientific curiosity from Dr. Ken Korzekwa always motivated me to think more deeply and broadly about science. Dr. Jozef Madzo from Coriell Institute for Medical Research gave valuable inputs in the aspect of bioinformatics, as well as helpful suggestions on setting up the server for our lab.

I would like to thank my fellow graduate students, Di and Kim. We started the PhD program in the same year. Their peer support cheered me all the way toward my graduation. I would like to thank other graduate students, postdocs, and scientists at Temple University School of Pharmacy, including Eva, Casey, John, Connor, Erickson, Dr. Mario Rico, and others. I appreciated their constant encouragement and enjoyed the friendly conversation with them. Even when we are social distancing during the COVID time, I still feel supported by them.

I would like to thank Dr. Krynetskiy and Dr. Suraj Peri, who guided me through the research of short tandem repeats and cancer bioinformatics in my early PhD years.

I would also like to thank the School of Pharmacy and Temple University for providing TA/RA assistantship and various resources for graduate study. I received enormous help from the student success center, especially from their program of academic coach, writing tutoring, and conversation partners.

Lastly, I would like to give special thanks to my family, especially my Mom and my brother, for their love and support, even though we were physically apart in three different countries. I would like to thank my boyfriend Meng for being by my side and providing continuous encouragement and support.

**TABLE OF CONTENTS**

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF EQUATIONS

# ABBREVIATIONS

| | |
|---|---|
| AAT | Alpha-1 antitrypsin |
| BA | Beta2-agonists |
| BALF | Bronchoalveolar lavage fluid |
| cAMP | Cyclic adenosine monophosphate |
| COPD | Chronic obstructive pulmonary disease |
| DEGs | Differentially expressed genes |
| DR | Drug repurposing |
| EBC | Exhaled breath condensate |
| ELISA | Enzyme-linked immunosorbent assay |
| eQTL | Expression quantitative trait loci |
| $FEV_1$ | Forced expiratory volume in 1 second |
| FVC | Forced vital capacity |
| GEO | Gene Expression Omnibus |
| GOLD | Global Initiative for Chronic Obstructive Lung Disease |
| GSEA | Gene Set Enrichment Analysis |
| GWAS | Genome-wide association studies |
| InChI | International Chemical Identifier |
| LA | Long-acting |
| MA | Muscarinic antagonists |
| MAGI | Metabolome and genome integration |
| NGS | Next-generation sequencing |

| | |
|---|---|
| SA | Short-acting |
| SMR | Summary Mendelian Randomization |
| WGCNA | Weighted gene co-expression network analysis |

# CHAPTER 1

# INTRODUCTION

## 1.1 COPD

### 1.1.1 Disease definition

Chronic obstructive pulmonary disease (COPD) is the most common severe chronic disease and is currently the third leading cause of death worldwide (Lozano et al., 2012). It is characterized by increasing breathlessness due to airflow obstruction and an accelerated lung function decline. These phenotypes are associated with long-term dysregulated inflammatory responses in the airways and the lungs (Barnes, 2014; Brusselle, Joos, & Bracke, 2011).

In the clinic, spirometry is used to measure airway obstruction for the assessment of exhalation ability in patients (Hogg, 2004; Rabe et al., 2007). One key readout is forced vital capacity (FVC), which measures the largest possible amount of air one can forcefully exhale after a deep inhalation. Another key readout is forced expiratory volume in 1 second ($FEV_1$). $FEV_1$ measures the amount of air one can exhale in the first second. Their ratio, $FEV_1/FVC$, is essential for the diagnosis of COPD along with other clinical factors. Generally, a ratio of less than 70% after bronchodilator usage is used to identify COPD patients, as suggested by Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines (Global Initiative for Chronic Obstructive Lung Disease, 2011).

One major attribute of COPD is the exacerbation, which is the sudden worsening of COPD symptoms (Viniol & Vogelmeier, 2018). This acute symptom is what makes the chronic lung disease devastating, and can be seen as analogous to a "stroke of the lungs" (Hillas, Perlikos, & Tzanakis, 2016). Indeed, the severe exacerbations of COPD were found to significantly worsen the survival outcome of patients (Soler-Cataluña et al., 2005). Hence, exacerbation is a critical aspect of COPD and imposes a substantial burden for COPD treatment.

### 1.1.2 Disease epidemiology

COPD is associated with both high morbidity and high mortality (Quaderi & Hurst, 2018). According to the Global Burden of Disease Study in 2016, 251 million people were affected by COPD around the world (Vos et al., 2016). In a meta-analysis consisting of 194 studies from 58 different countries, the prevalence of COPD was estimated to be around 6-9% (Ntritsos et al., 2018). The prevalence will probably continue to increase due to the aging population and continued exposure to risk factors like smoking (Chapman et al., 2006). In one study, the number of COPD patients was predicted to increase by 155% from 2010 to 2030 (Khakban et al., 2016). As a consequence, COPD will inflict a substantial economic and social burden on both individual patients and health care systems worldwide (Anees ur Rehman et al., 2019; James et al., 2018).

Globally, the total number of deaths caused by COPD was 2.8 million in 2010 (Burney, Patel, Newson, Minelli, & Naghavi, 2015). In the United States, the total number of deaths from COPD was averaged around 146 thousand for the years 2014-2015. The mortality rate per 100,000 persons was reported to be 45.7 (Sullivan et al., 2018).  The mortality rate is especially high for patients with very severe COPD. In a population-based study, 26% of patients with very severe COPD died after a 1-year follow-up, while the number was only 2.8% for non-COPD subjects (Afonso, Verhamme, Sturkenboom, & Brusselle, 2011).

### 1.1.3 Disease heterogeneity

Traditionally, COPD has been classified as chronic bronchitis and emphysema. Chronic bronchitis is characterized by increased mucus and inflammation in the bronchi, whereas emphysema is the destruction and enlargement of the alveoli, the air spaces in the lungs. However, there is significant heterogeneity among COPD patients, which cannot be simply classified into only these two categories. COPD patients exhibit a great variety of disease manifestations, severity, clinical symptoms, spirometry result, pathophysiology, exacerbation risk, and presence of comorbidities (Castaldi et al., 2017; Roca et al., 2014). Thus, disease severity assessment is very critical for the selection of treatments and patient prognosis. Table 1 shows different symptoms, spirometry, and therapies based on different severities of COPD (Global Initiative for Chronic Obstructive Lung Disease, 2011). There are also other guidelines for COPD assessment, like LNN (lower limit of normal as the $FEV_1/FVC$ threshold for COPD diagnose), $FEV_1/FEV_6$ (Bhatt et al., 2014; Jing, Huang, Cui, Xu, & Shen, 2009), and the new ABCD classification based on symptoms and exacerbation history (Calverley, 2013; Global Initiative for Chronic Obstructive Lung Disease, 2017). In general, they are different from the GOLD guidelines in the test readout calculation, threshold to identify decreased lung function, or inclusion of additional symptoms for diagnosis. However, these alternative guidelines are beyond the discussion of this thesis.

Table 1. Classification of airflow limitation severity in COPD

| Severity | Symptoms | Spirometry | Treatment |
|---|---|---|---|
| GOLD I: Mild | Chronic cough and sputum (maybe) | $FEV_1 \geq 80\%$ predicted | Avoid risk factors<br>Obtain flu vaccine<br>Short-acting bronchodilator |
| GOLD II: Moderate | Progression of above, plus dyspnea on exertion | $50\% \leq FEV_1 < 80\%$ predicted | Above,<br>Plus regular with one or more long-acting bronchodilators and rehabilitation. |
| GOLD III: Severe | Increased shortness of breath and repeated exacerbations | $30\% \leq FEV_1 < 50\%$ predicted | Above,<br>Plus inhaled glucocorticosteroids for repeated exacerbations |
| GOLD IV: Very Severe | Impaired quality of life and exacerbations may be life-threatening | $FEV_1 < 30\%$ predicted | Above,<br>Plus long-term oxygen therapy for chronic respiratory failure.<br>Lung surgery. |

*$FEV_1$: The forced expiratory volume in one second; Post-bronchodilator $FEV_1$ in patients with $FEV_1/FVC < 70\%$.

### 1.1.4 Risk factors for COPD

#### *1.1.4.1 Cigarette smoke*

Cigarette smoke is a major risk for COPD. Studies found that about 10-30% of smokers develop COPD (D. S. Kim et al., 2005; YOO et al., 2011). Especially in the elderly population, up to 50% of smokers may develop COPD (Lundbäck et al., 2003). In addition, 40-70% of patients with COPD are previous smokers or current smokers (Bridevaux et al., 2010; Celli, Halbert, Nordyke, & Schau, 2005; Lamprecht et al., 2011; Wheaton, Cunningham, Ford, & Croft, 2015; Zhou et al., 2009). Cigarette smoke causes oxidative stress, triggers inflammatory-immune responses (Stämpfli & Anderson, 2009), and affects the ability of hosts to activate appropriate immune and inflammatory responses (Yao & Rahman, 2011). There is still a gap in the full understanding of the molecular mechanisms of smoke effects on cellular systems and the underlying pathogenesis of COPD. This has impeded the progress in developing effective treatments for COPD.

Aside from cigarette smoking, long-term exposure to other lung irritants can also contribute to COPD, like occupational exposure or indoor air pollution (Blanc et al., 2008; Kurmi, Semple, Simkhada, Smith, & Ayres, 2010; United States Department of Health and Human Services, 2014). A recent study on vaping has found that e-cigarettes are similar to cigarette smoke in their ability to stimulate inflammation and lung damage (Bozier et al., 2019).

#### *1.1.4.2 Genetic factors*

Genetic factors may also contribute to the risk of COPD. Unlike Mendelian diseases, which are caused by mutations in a single gene, complex diseases like COPD may involve multiple genetic susceptibility variants. These genetic determinants themselves may only have a modest effect size, but when they team together and interact with the environment, they are likely to give rise to COPD phenotypes (Hardin & Silverman, 2014).

The earliest finding on the genetic components in COPD was related to alpha-1 antitrypsin (AAT) deficiency, a rare hereditary disorder (Eriksson, 1964). Patients with AAT deficiency tend to develop emphysema at an early age, and AAT deficiency is associated with an increased risk of COPD in smokers (Black & Kueppers, 1978).

Table 2. Candidate genes associated with COPD

| Gene | Potential Function of Variants | Encoded protein and description | Reference |
|---|---|---|---|
| ***AAT*** | Protease inhibition | Alpha-1 antitrypsin: a protein belonging to the serpin superfamily, which is serine protease inhibitors. | (Eriksson, 1964) |
| ***IL6R*** | Immune mechanisms | Interleukin-6 (IL-6) receptor subunit alpha: IL-6 is involved in both pro-inflammatory and anti-inflammatory activity. | (Wilk, Walter, Laramie, Gottlieb, & O'Connor, 2007) |
| ***GSTO2*** | Arsenic biotransformation | Glutathione S-transferase omega 2: an enzyme involved in metabolizing xenobiotics and carcinogens. | (Wilk et al., 2007) |
| ***HHIP*** | Lung development by hedgehog signaling | Hedgehog-interacting protein: regulates the signaling pathway of developmental processes by interacting with proteins in the hedgehog family. | (Wilk et al., 2009) |
| ***IREB2*** | Pulmonary iron homeostasis | Iron responsive element binding protein 2: an RNA-binding protein that regulates iron hemostasis by affecting ferritin mRNA translation and transferrin receptor mRNA stability. | (Pillai et al., 2009) |
| ***MMP12*** | Elastase activity | Matrix metallopeptidase 12: a protease that breaks down the extracellular matrix | (Hunninghake et al., 2009) |
| ***FAM13A*** | Oxidative stress and impaired apoptosis | Family with sequence similarity 13 member A: may inhibit the signal transduction molecule Rho A. | (Michael H Cho et al., 2010) |
| ***AGER*** | Immune mechanisms | Advanced glycosylation endproduct receptor: a pattern recognition receptor that binds to multiple ligands | (Repapi et al., 2010) |

Following this early identification of genetic risk factors, many studies and initiatives have been carried out to reveal other genetic determinants of COPD and understand the pathogenesis of COPD (M. H. Cho et al., 2012; Michael H Cho et al., 2010; W. J. Kim et al., 2014; W. J. Kim, Oh, Kim, et al., 2013; W. J. Kim, Oh, Lee, et al., 2013; J. H. Lee et al., 2015; Regan et al., 2010; Vestbo et al., 2008). These large-scale studies have suggested multiple genetic loci for COPD susceptibility. Table 2 lists several candidate genes associated with COPD that have been replicated in numerous populations.

### 1.1.5 Current Treatment

Due to the heterogeneity of COPD symptoms, treatment may vary based on different disease conditions, as shown in Table 1. Current therapies of COPD include medication, smoking cessation, rehabilitation, oxygen therapy, surgery, or a combination of these. These treatments aim to improve patient quality of life by lessening the current symptoms, slowing the progression of the disease, and preventing future exacerbation.

#### *1.1.5.1 Bronchodilators*

Bronchodilators are administered regularly for COPD patients to help relieve shortness of breath and coughing, by relaxing the muscles around airways. They are used in the form of inhalers to help deliver medications to airways and lungs. There are three types of bronchodilators, including beta2-agonists (BA), anticholinergics/muscarinic antagonists (MA), and methylxanthines. Each type is different in their pharmacological mechanisms, effect onset time, and duration of actions.

Beta2-agonists stimulate beta2 adrenoceptors in airway smooth muscles, which increases messenger cyclic adenosine monophosphate (cAMP), inducing bronchodilation by airway smooth muscle relaxation (Soneson & Delorenzi, 2013). Anticholinergics, also known as muscarinic antagonists, exert the bronchodilation effect by blocking the bronchoconstriction effects of acetylcholine on M3 muscarinic receptors in airways smooth muscles (Melani, 2015). Methylxanthine drugs, like theophylline, target adenosine receptors and block adenosine mediated bronchoconstriction (Daly, Jacobson, & Ukena, 1987). They also work as phosphodiesterase inhibitors which can increase cAMP levels for the effect of airway smooth muscle relaxation and bronchodilation (Essayan, 2001).

The duration of actions is an essential characteristic of the bronchodilators. Short-acting (SA) bronchodilators can be used before activities for immediate effect. In contrast, long-acting (LA) bronchodilators can be used regularly to control COPD symptoms. Usually, a combination of drugs will be recommended for patients, as combination therapies have been shown to increase bronchodilation with a decreased risk of side-effect (Cazzola & Molimard, 2010). Table 3 shows different bronchodilators used in the market for COPD.

Table 3. Current COPD medications in the market

| name | type | form | duration |
|---|---|---|---|
| Ipratropium bromide | SAMA | Inhaler | 4-6h |
| Salbutamol sulfate | SABA | Inhaler | 4-6h |
| Terbutaline sulfate | SABA | Inhaler | 4-6h |
| Salbutamol sulfate | SABA | Inhaler | 4-6h |
| Umeclidinium bromide | LAMA | Inhaler | 24h |
| Glycopyrronium bromide | LAMA | Inhaler | 24h |
| Tiotropium bromide monohydrate | LAMA | Inhaler | 24h |
| Aclidinium bromide | LAMA | Inhaler | 12h |
| Formoterol fumarate | LABA | Inhaler | 12h |
| Indacaterol maleate | LABA | Inhaler | 24h |
| Salmeterol xinafoate | LABA | Inhaler | 12h |
| Olodaterol hydrochloride | LABA | Inhaler | 24h |
| Fluticasone propionate/ salmeterol xinafoate | ICS/LABA | Inhaler | 12h |
| Fluticasone furoate/ vilanterol trifenatate | ICS/LABA | Inhaler | 24h |
| Budesonide/formoterol fumarate | ICS/LABA | Inhaler | 12h |
| Fluticasone furoate/ umeclidinium bromide/ vilanterol trifenatate | ICS/LAMA/LABA | Inhaler | 24h |
| Ipratropium bromide/salbutamol sulfate | SAMA and SABA | Inhaler | 4-6h |
| Umeclidinium bromide/vilanterol trifenatate | LAMA and LABA | Inhaler | 24h |
| Aclidinium bromide/formoterol fumarate dehydrate | LAMA and LABA | Inhaler | 12h |
| Tiotropium bromide monohydrate/olodaterol hydrochloride | LAMA and LABA | Inhaler | 24h |
| Glycopyrronium bromide/ indacaterol maleate | LAMA and LABA | Inhaler | 24h |
| Theophylline | Methylxanthines | Oral | 8/12h |
| Roflumilast | Phosphodiesterase-4 inhibitor | Oral | 24h |

SAMA - Short-Acting Muscarinic Antagonist. SABA - Short-Acting Beta2-Agonist. LAMA - Long-Acting Muscarinic Antagonist. LABA - Long-Acting Beta2-Agonist. ICS - Inhaled Corticosteroid.

#### *1.1.5.2 Corticosteroids*

Corticosteroids are anti-inflammatory drugs that are widely prescribed for a variety of diseases. In COPD, the inhaled corticosteroid is usually administrated to prevent frequent exacerbations by reducing airway inflammation (Falk, Minai, & Mosenifar, 2008). Long-term use of the corticosteroids is not recommended, as it may lead to systemic complications of weight gain, hyperglycemia, osteoporosis, or pneumonia (Wood-Baker, Gibson, Hannay, Walters, & Walters, 2005; Yang, Clarke, Sim, & Fong, 2012). Combinational therapy with both inhaled corticosteroids and long-acting bronchodilators is often suggested as it has shown to be more effective than each individual component alone (Nannini, Lasserson, & Poole, 2012; Nannini, Poole, Milan, Holmes, & Normansell, 2013).

#### *1.1.5.3 Oxygen therapy*

Oxygen therapy is a treatment that provides supplemental oxygen to patients. Long-term use of supplemental oxygen was found to improve survival in patients with COPD (Kvale, Conway, & Coates, 1980; Stoller et al., 2010; Stuart et al., 1981). Oxygen supplement devices are common in clinical settings, and there are also devices that patients can use at home (Branson, 2018; Furlanetto & Pitta, 2017). Oxygen is usually delivered through nasal prongs or a face mask.

#### *1.1.5.4 Surgery*

The chronic airflow limitation in COPD is an outcome caused by malfunctions in airways and/or lung tissues, which are termed as bronchitis and emphysema, respectively. Although β2 agonists and anticholinergic agents are widely administrated for the treatment of COPD, they are less effective in emphysema dominant COPD (J.-H. Lee et al., 2010). Thus patients with advanced emphysema may need to opt for more invasive treatment, like lung volume reduction surgery, bullectomy, or lung transplantation (Marchetti & Criner, 2015).

Lung volume reduction surgery is a procedure to remove about 30 percent of the most diseased lung tissues. It can facilitate the ventilation of the remaining healthier portions of the lung and reduce air trapping (Brantigan & Mueller, 1957; Shah, Herth, van Geffen, Deslee, & Slebos, 2017). There is also another type of surgery, bullectomy, that focus on removing only the most destroyed bullae (Dartevelle, Macchiarini, & Chapelier, 1995). The bullae are large air sacs that consist of hundreds of alveoli. Bullectomy helps improving lung function and reducing dyspnea in COPD (Travaline, Addonizio, & Criner, 1995).

Lung transplant is an option for very severe or advanced COPD. It is a procedure to replace one or both lungs in patients with lungs from donors (Trulock, 1998). However, the waiting time for available donor lungs can be very long, and there is also a risk for organ rejection and other complications. Thus patients need to be carefully selected for the recommendation of lung transplants (Vazquez Guillamet, 2019).

## 1.2 Drug repurposing

### 1.2.1 Drug repurposing overview

New drug research and development is becoming increasingly costly both in terms of time and financial investment. On average, pharmaceutical companies need to spend 6-15 years and 2.6 billion dollars before a drug gets FDA approval (DiMasi, Grabowski, & Hansen, 2016; Prasad & Mailankody, 2017); in fact, the cost has increased 10-fold in the past decades. This steep increase in the cost has been referred to as “Eroom’s Law” (Scannell, Blanckley, Boldon, & Warrington, 2012), as it is the exact opposite trend of the Moore’s Law in the information technology industry. Moore’s Law predicts that the capability of our computers will increase exponentially over the years, and the cost will drop exponentially at the same time. Thus, new strategies are urgently needed to facilitate new drug research and development processes. One strategy called drug repurposing has the potential to reduce the enormous cost and time that are traditionally required for new drug development.

Drug repurposing (DR) is the application of existing approved therapeutic compounds for new disease indications. One famous example is sildenafil. It was originally developed for cardiovascular disease treatment, but it was repurposed by Pfizer and marketed as Viagra for the treatment of erectile dysfunction. This successful example of drug repurposing is based on clinical retrospective analysis. Another more recent example is the repurposing of thalidomide. It was notoriously known to cause birth defects in newborns around 1960. It led to the biggest medical disaster in US history and was pulled off the market. However, in 2016, it was approved by the FDA for the treatment of multiple myeloma. Thalidomide is now used to treat other diseases, including Crohn's disease, leprosy, and some cancers (Vargesson, 2017). With these two examples, we can see that existing drugs can be successfully repurposed for new disease indications, and the knowledge from those existing drugs can help diminish the time and cost of getting a new treatment approval.

A variety of approaches has been proposed to find potential drug repurposing opportunities. They were summarized nicely as different experimental and computational approaches by a recent review. These approaches help generate drug repurposing hypotheses based on three components, population phenotype, disease biology, and drug property. Population phenotype, or, more specifically, pharmacoepidemiology, relies on retrospective clinical analysis to reveal unintended drug effects that may be beneficial for the treatment of other diseases. Such analysis can be performed on real-world evidence or electronic health records (Breckenridge, Breckenridge, & Peck, 2019; Xu et al., 2015). Recently, biobanks with a large number of human samples and rich subject phenotypes have been proposed to be a novel source to explore for drug repurposing clues (Pulley et al., 2017). On the other hand, studies on disease biology may lead to the identification of potential drug repurposing targets on the molecular levels. Genes or proteins associated with the disease can serve as potential drug targets. These molecules may be novel targets or genes that are already the target of marketed drugs. On the functional level, pathways involved in the disease may also provide insight into the potential genes that can be targeted. Drug repurposing opportunities may be identified if there is an overlap between underlying pathways of different diseases. Last but not the least, directly screening on the desired drug properties is another strategy that has been used to identify drug repurposing opportunities. Drugs can be screened on in vitro cell lines or in vivo animal disease models for the evaluation of their potential effect on the disease. The screening can also be performed in silico. Databases with transcriptomic profiles of cell lines treated with compounds can be used as the resources to compare to the disease profile to identify potential drug repurposing opportunities (Subramanian et al., 2017). In addition to the evaluation of the potential effect of drugs, drug-target binding can be evaluated to identify novel targets of current drugs. Specifically, ligand binding assay or computational prediction of molecular docking can be used to reveal drug-target affinities.

In summary, various computational or experimental approaches can be used to make predictions or generate hypotheses on the potential for drug repurposing from various sources of input, including population phenotype, disease biology, and drug property.

### 1.2.2 Drug repurposing in COPD

Some efforts have been made in terms of identifying drug repurposing opportunities in COPD with different strategies. Insights from experts have helped generate testable drug repurposing hypotheses. For example, Bisphosphonate alendronate is widely used in the treatment of bone disease such as osteoporosis. It has also been shown that it can induce apoptosis in macrophage cell lines. Since alveolar macrophages play a critical role in inducing inflammation and destruction of the alveolar, it is hypothesized that Bisphosphonate alendronate may also induce apoptosis in alveolar macrophage and thus prevent emphysema. Alendronate inhalation induces apoptosis in alveolar macrophages and prevented airspace enlargement in mice emphysema model

(Ueno et al., 2015). In another example, the repurposing of cardiovascular statin drugs is reviewed for improved treatment of COPD. Statins use has been associated with improvement in COPD symptoms and decreased mortality in clinical settings. In vitro experiment showed that statins affect neutrophil function, which is a crucial cellular component of innate immune systems and correlates with inflammation and COPD severity (Walton et al., 2016).

Drug screening is another strategy to identify drug repurposing opportunities. Based on previously known pathways involved in COPD, assays are designed to measure the activation or inhibition of the pathway that may reverse the underlying COPD molecular mechanism. For example, WNT/β-catenin signaling is known to decrease in patients with COPD. High-throughput drug screening identified several compounds that activate WNT/β-catenin signaling *in vitro* and may induce lung repair (Costa et al., 2018). In another drug screening study, niclosamide and nitazoxanide were suggested as promising repurposing drug for COPD. They were identified as inhibitors of the CaCC TMEM16A that can induce airways bronchodilation by the screening of 580,000 clinical compounds (Miner et al., 2019).

Lastly, omics data have also helped propose potential drug repurposing targets for COPD. In a COPD genome-wide association study with as many as 150,000 individuals, 243 genes were identified with a possible causal effect on lung function. Seven of them are targets of clinical drugs, including CHRM3, SLC6A4, and CRHR1 (Wain et al., 2017).

Although these drug repurposing strategies helped nominated several drug repurposing targets, there has not been any effort that takes advantage of multiple omics data and integrates them to identify potential drug repurposing targets for COPD treatment. We proposed to use a systems biology approach to expand the pool of potential drug targets for COPD.

## 1.3 Systems biology in drug target identification

### 1.3.1 Target identification through systems biology

Target identification is commonly the first step in the drug discovery pipeline, followed by lead identification, lead optimization, preclinical assessment, and clinical trials. Historically, scientists would search for potential drug targets through literature, basic biological research, or patients with extreme cases of the disease. Yet these traditional approaches of target identification may be limited by prior scientific knowledge, experiment throughput, and availability of rare patient cases. If fact, a 2006 study counted only 324 distinct molecular drug targets based on US FDA pharmacopeia, which is far less than the 21,000 drug products or 1,357 unique drugs at that time (Overington, Al-Lazikani, & Hopkins, 2006). In 2016, the number of drug targets in the human genome was updated to 667 (Santos et al., 2016). Although this number only makes up less than 5% of the total 20,000 genes in the human genome, this number will continue to grow as our knowledge of diseases increase.

Systems biology is one field that can help us gain more knowledge about the disease on the molecular level to increase the pool of drug targets. It is defined as “the comprehensive quantitative analysis of the manner in which all the components of a biological system interact functionally over time” (Aderem, 2005). This field emerges as we are now entering an exciting era that more biological data than ever is available to the scientific community due to the advance in omics technology. Omics technologies have enabled the high-throughput detection of molecules across the same level, as opposed to investigating molecular components (e.g., gene or protein) one at a time (Horgan & Kenny, 2011). The massive amount of omics data poses new challenges to process, analyze, and interpret the data. The progressing field of systems biology offers a more quantitative and computational way to address these challenges. There is some progress made on each omics level that contributes to our understanding of the COPD. However, we are still far from the ideal scenario where we fully understand the functional role of each molecule in the systems, how they cross-talk between different omics levels and their contribution to the disease.

### 1.3.2 Omics analysis in COPD

#### *1.3.2.1 Genome*

The first widely used genomic technology is the DNA chip or microarray (Stekel, 2003). They are microscope slides attached with thousands or millions of probes of known DNA sequences. The fragmented DNA from a sample can bind specifically to these probes and generate fluorescence signals to distinguish different genotypes across numerous genetic loci simultaneously. This technology gave rise to a significant number of genome-wide association studies (GWAS) that compare the genomic profile between COPD patients and healthy subjects. Hundreds of loci were found to be associated with COPD or lung functions by GWAS research, and several candidate genes were proposed to contribute to COPD susceptibility, including CHRNA3/5, FAM13A, IREB2, AGER, and HHIP (W. J. Kim & Lee, 2015; Lamontagne et al., 2018).

Another widely used genomic technology is the next-generation sequencing (NGS) (Park & Kim, 2016). Its capability of genotyping is not limited to the user-defined probes as in the microarray. As it sequences the genome to the resolution of each single nucleotides, this technology can also be used to detect unknown genetic mutations.

Several new genetic variants were recently identified by a study in lung cancer patients with COPD (Lunardi et al., 2017).

#### *1.3.2.2 Transcriptome*

Transcriptomic analysis studies all the RNA molecules in a biological sample. As RNAs are also composed of nucleotides, they can be reverse-transcribed into cDNA fragments and then measured by similar genomic technologies like microarray or NGS. It is widely used to study the gene expression (mRNA) profiling in tissue or cells undergoing specific treatment. The up- or down-regulated genes can be identified in these samples and offers more insight into the dynamic regulation of the underlying molecular mechanisms. It is proposed that the airway and lung tissue transcriptomics can be utilized to identify molecular subtypes of COPD and personalized the treatment despite the same clinical manifestation of COPD patients (Steiling, Lenburg, & Spira, 2013). Through transcriptomics studies on activated blood CD4+ T cells *in-vivo*, it was found that the mRNA transcripts of CD4+ T cells can help differentiate patients with COPD and control smokers, as well as subtyping of COPD patients (Roberts et al., 2015).

#### *1.3.2.3 Proteome*

Proteomics is the next omics level after genomics and transcriptomics, based on the molecular flow of the central dogma of biology. However, it poses a more significant challenge for the large scale analysis of the entire set of proteins (Graves & Haystead, 2002). Unlike genomics or transcriptomics, which are based only on four nucleotide components of adenine (A), thymine (T), guanine (G), and cytosine (C), the human proteins are composed of 20 amino acids. This dramatically limits the throughput and coverage of proteomics analysis. A conventional genomic analysis can capture nearly all 20,000 genes in the human genome, while a proteomic analysis measured thousands of proteins would be considered as high coverage. However, proteomics still offers an indispensable insight into biological systems in the sense that proteins are closer to the functional endpoint than genomics or transcriptomics. Western blotting and enzyme-linked immunosorbent assay (ELISA) are widely used as the traditional way to study individual proteins (Lequin, 2005). Higher throughput proteomics assays include protein microarray, two-dimensional polyacrylamide gel electrophoreses (2D-PAGE), and mass spectrometry-based assays (Issaq & Veenstra, 2008; Macbeath, 2002; Yates, 2011). Generally, hundreds of proteins can be detected in a proteomics experiment (Mulvey, Thur, Crawford, & Godovac-Zimmermann, 2010). This helps the measurement of protein abundance changes under various conditions and contributed to many publications related to COPD proteomics change. For example, proteomics methodologies have identified several biomarkers related to COPD, such as apolipoprotein E, inter-α-trypsin-inhibitor heavy chain H4, histone 3.3, and matrix metalloproteinase 13 (Bandow et al., 2008; Barrero et al., 2013; E. J. Lee et al., 2009).

#### *1.3.2.4 Epigenome and metabolome*

Aside from the three main components in the central dogma of biology, there are also other important omics levels that play a vital role in the physiology and pathology of biological systems.

The epigenome is the study of the chemical changes to the DNA and nuclear histone proteins that help regulate the gene expression by changing the chromatin structure (Bernstein, Meissner, & Lander, 2007). The underlying epigenetic changes in COPD can bring insights for targeting epigenetic modulation treatments (Schamberger, Mise, Meiners, & Eickelberg, 2014). Corticosteroids are used in COPD as anti-inflammatory drugs. However, resistance to corticosteroids may arise due to epigenetic mechanisms, including the reduced activity and expression of histone deacetylase-2 (HDAC2) (Barnes & Adcock, 2009). Theophylline is used to account for the corticosteroid resistance, as it targets this epigenetic mechanism of drug resistance by reversing the reduced HDAC2 expression (Barnes, 2006).

Metabolites in biological systems are composed of a wide range of chemicals, including lipids, carbohydrates, amino acids, and so on. Metabolomics is the comprehensive measurement of all these chemical metabolites within any biological sample, such as cell, tissue, or organism. These metabolites include small molecules like carbohydrates, lipids, amino acids, fatty acids, alkaloids and etc. Metabolomics represents the functional end products formed by biochemical reactions in our body to regulate the physiological state (Hollywood, Brison, & Goodacre, 2006). Therefore, many metabolomics studies have identified several dysregulated metabolites as the biomarkers for COPD (McClay et al., 2010; Naz et al., 2017; Ubhi et al., 2012).

### 1.3.3 Network analysis

While omics technologies allow unbiased measurement of molecules at high-throughput, they also bring new challenges of detecting significant signals out of the plethora of results and understanding their biological significance. New types of analysis methods are developed continuously to process the data at the omics scale.

#### *1.3.3.1 Pathway analysis*

It is not uncommon to get the result of dozens or hundreds of significant genes from an omics experiment. The long list of genes makes it impossible to perform an in-depth literature search for the function of all the genes and their relevance to the disease. New methods such as pathway analysis are developed to organize the genes in terms of their functions. Pathway analysis looks at the set of genes that work together to fulfill a specific function, to see if they are over-represented in disease. One standard pathway analysis method is enrichment analysis (Reimand et al., 2019). After a list of differentially expressed genes/proteins in disease is obtained, we can compare it to all known pathways. If there is a greater abundance of differential expressed genes in a specific pathway than it would appear randomly in the whole genome, this particular pathway will be considered as enriched in the disease. There are also other types of pathway analysis like Gene Set Enrichment Analysis and signaling pathway impact analysis (Subramanian et al., 2005; Tarca et al., 2008). All these help us to understand the genes at the level of functional pathways.

#### *1.3.3.2 WGCNA*

A transcriptomic experiment can provide around 20,000 gene expression profiles for each sample. For any study that has over fifty samples, it can easily generate over a million data points of gene expression. A common way to downsize the overwhelming data for further focused analysis is to identify the differentially expressed genes, i.e. the individual genes that have significantly higher or lower expression levels in the disease condition. Yet this approach treats each gene independently when analyzing the expression level, and thus lose potentially valuable information on the relationship between genes. Weighted gene co-expression network analysis (WGCNA) offers an alternative approach to maintain the relationship among genes by identifying their correlation patterns on gene expression levels. Instead of solely getting a list of differentially expressed genes from microarray data, WGCNA takes into consideration the relationships between measured transcripts. It accomplishes this by means of pair-wise correlations between gene expression profiles, and output modules based on highly correlated genes. The relationship between modules can be further explored and then be related to external information like clinical data or biological assay readings for their association with the disease. WGCNA also has the advantage in that it can decrease the false-negative rate as a result of multiple correction adjustments on tens of thousands of genes. When genes are grouped in modules, it reduces the number of statistical tests and lowers the penalty from the multiple correction adjustment. WGCNA was successfully applied to the blood transcriptome of COPD patients in a recent study (Obeidat et al., 2017). The study identified three modules that significantly correlated with the lung function reading $FEV_1$. Compared to pathway enrichment analysis, which relies on previous knowledge of genes in the pathway, WGCNA may reveal novel functional pathways by leveraging the co-expression pattern of genes.

### 1.3.4 Multi-omics integration

There have been some efforts to integrate different COPD-related omics data for the identification of disease biomarkers, signature genes, and associated pathways.

The earliest and more well-established integration is genomic and transcriptomic integration. While genomics analysis is able to identify genetic variants at specific sites in the genome, it lacks the capability to reveal the association between the genetic variants and the genes whose expression level might be affected by them. The genetic variants within a gene region can be readily linked to the expression of the gene, but the genetic variants in noncoding RNAs intergenic regions require other more advanced approaches to identify the associated genes. The most common approach is expression quantitative trait loci (eQTL) analysis, along with many more methods being developed, like TWAS, Summary Mendelian Randomization (SMR) (Gusev et al., 2016; He et al., 2013; Nica et al., 2010; Zhu et al., 2016). One study on COPD integrated transcriptomic and GWAS with S-PrediXcan, which estimates the effects of gene expression on phenotypes through association analysis. Several genes were suggested to be significantly associated with severe COPD by this analysis (Barbeira et al., 2018).

With the rise of proteomics, metabolomics, there are also more challenges in terms of the integration of the data. Several pathways associated with COPD were identified in a study by combining data from metabolomics and transcriptomics (Cruickshank-Quinn et al., 2018). Another study also used the pathway approach to integrate transcriptomics, proteomics, and metabolomics to reveal long-term effects and mechanisms of a traditional Chinese formula on COPD (J. Li et al., 2016). In addition to the pathway level integration, similarity network-fusion-based integration was applied to transcriptomics, proteomics, and metabolomics to enable molecular classification of COPD (C. X. Li, Wheelock, Magnus Sköld, & Wheelock, 2018).

There is still great room for improvement in multi-omics integration. First, the data collected from different omics levels are usually not from matching samples. Different research groups generated the omics data related to COPD on a wide range of biological systems, including *in vitro* cell lines, primary cells, and tissues from model organisms, or human subjects. Secondly, although there are several studies with large cohorts of COPD patients, omics data from lung tissue are still limited. This can be explained by the fact that less invasive approaches are more frequently used to collect samples from human subjects, such as peripheral blood, sputum, bronchoalveolar lavage fluid (BALF). Lastly, more quantitative integration approaches are needed to align with the quantitative nature of the omics data. The pathway-level integration approach helps reveal the underlying dysregulated pathways involved in COPD. However, a more quantitative approach may reveal the candidate genes associated with COPD and explore their roles as a drug target.

# CHAPTER 2

# DEVELOPING A MULTI-OMICS ANALYSIS PIPELINE TO IDENTIFY NOVEL DRUG REPURPOSING TARGETS FOR COPD

## 2.1 Background and project goal

COPD has become the third leading cause of death worldwide (Lozano et al., 2012). However, none of the existing pharmacological treatments for COPD have been shown to modify the long-term decline in lung function successfully. Therefore, there is an unmet need to discover new treatments for COPD. In order to bring potential new therapies to COPD patients at a faster pace and lower cost, recent efforts have been made to explore the drug repurposing strategy. This strategy aims to find approved drugs with different disease indications that can also improve COPD symptoms.

As introduced in chapter 1, several potential compounds have been proposed to be repurposed for COPD. They are the findings of different drug repurposing approaches, including expert insight, high-throughput drug screening, and omics study (Table 4). However, there are some limitations to these approaches. First, they may be limited or biased by our previous knowledge about COPD. The evaluation of both *in vitro* and *in vivo* experiments based on expert hypothesis and drug screening strategy relies on a specific pathway or target known to be affected in COPD. These approaches would prevent drug discovery on novel mechanisms underlying COPD. Second, current drug repurposing omics studies for COPD only utilize a single omics layer in the biological systems. They are limiting the identification of a more comprehensive list of potential drug targets because they may not capture the full molecular changes in COPD. To date, there is not any published report that systematically integrates multiple omics data to identify potential drug repurposing targets for COPD treatment.

Table 4. Current findings for drug repurposing in COPD

| Drug | Original indication | Drug repurposing approach | Reference |
|---|---|---|---|
| **-Bisphosphonate alendronate** | bone disease: osteoporosis, Paget's disease | expert insight: alendronate can induce apoptosis in macrophages, which play a critical role in including inflammation and destruction of the alveolar in COPD | (Ueno et al., 2015) |
| **-Statins** | cardiovascular disease | expert insight: statins have been associated with improvement in COPD symptoms and decreased mortality in clinical settings | (Walton et al., 2016) |
| **-Riluzole**<br>**-and 5 others** | amyotrophic lateral sclerosis | drug screening: WNT/β-catenin signaling assay as in-vitro COPD model | (Costa et al., 2018) |
| **-Niclosamide**<br>**-Nitazoxanide** | tapeworm infestations<br>diarrhea caused by a parasite infection | drug screening: CaCC TMEM16A Antagonists assay as in-vitro COPD model | (Miner et al., 2019) |
| **-Carbachol**<br>**-Pilocarpine**<br>**-Oxybutynin chloride**<br>**-Acetylcholine**<br>**-and 3 others in clinical trials** | Glaucoma<br>Glaucoma<br>Hyperhidrosis, Polyuria<br>Glaucoma | omics study: genome-wide association analysis | (Wain et al., 2017) |
| **-Tripeptide GHK** | A natural wound healing agent | omics study: Transcriptomic analysis | (McDonough et al., 2012) |

Thus, the goal of this project was to develop a systems biology pipeline for the identification of biological-relevant and feasible drug targets for potential COPD treatment. Through systems biology approaches, multi-level omics data outcomes from COPD were used to find the genes relevant for the physiopathology of the disease. Their drug repurposing potential evaluated the feasibility of these genes as potential pharmacological targets. This project is be the first that has applied a systematic approach integrating multiple omics data to find drug repurposing targets for COPD.

## 2.2 Specific aims rationale

This project was divided into three specific aims:

- Aim 1. Obtain initial signature genes in COPD by transcriptomic analysis.
- Aim 2. Expand and prioritize the signature gene list by integrating additional omics data, including genomics, proteomics, metabolomics, and interactomics.
- Aim 3. Identify drug repurposing targets within the COPD signature genes

The human genome consists of more than 20,000 protein-coding genes (Salzberg, 2018). To search for potential gene targets for COPD treatment among them, we need first to identify genes that are relevant for COPD. Transcriptomic data, which are the expression levels for all genes in biological samples, are selected as the first omics level to be screened for several reasons. Firstly, the transcriptome is differentially expressed on each tissue or cell type, as opposed to the same genome sequence across all tissues in the body. It has the advantage of capturing underlying molecular modifications to the main organs (airways and lungs) involved in COPD. Secondly, as transcriptome is able to reveal more dynamic regulations than the stable genome, it can indicate how genes are dis-regulated in specific disease status compared to healthy subjects. Thirdly, transcriptome has a higher coverage of almost all the 20,000 genes in the human genome, while proteomics studies are limited to hundreds or thousands of proteins at most. Lastly, transcriptomic data are one of the most abundant omics data. This large sample size offers the advantage of greater statistical power, as well as the possibility of further subdividing the cohort with less penalty from small subgroup size. Thus, transcriptomics is a good data source for initial screening to find signature genes for COPD.

However, some gene signals may not be captured at the transcriptomic level. Additional data at different omics levels are able to offer complementary molecular profiles that underline COPD. Gene transcripts are translated to proteins to exert biological functions. Thus proteome provides an additional omics level to reveal proteins that are dysregulated in COPD patients. However, only thousands or hundreds of proteins can be detected in a proteomics experiment. Thus, we also included metabolomics data to compensate for the limitation on protein coverage. If a metabolite is dysregulated in COPD, the protein involved in the biomedical reaction to generate the specific metabolite might also be dysregulated. Based on this hypothesis, proteins associated with the metabolites dysregulated in COPD were also added to our pipeline.

Genomics data from GWAS were also included in the pipeline. With the decrease in the cost of genomic microarray and next-generation sequencing, tens of thousands of COPD patients were recruited in GWAS research to search for the genetic association to COPD. Genetic variants with a smaller effect size are more likely to be identified with such a large cohort.

Interactome, which is the comprehensive knowledge of protein-protein interaction, can also help the identification of additional proteins that may involve in COPD. Protein-protein interaction information can be utilized to identify the genes that closely interact with the signature genes we found in the earlier step. The hypothesis is that those neighbor genes may work in the same pathways to contribute to COPD. With the additional knowledge on the protein-protein interaction network, we can then prioritize the genes that are in close proximity to signature genes on all omics levels.

It needs to be noted that the sample sources of different omics data should be consistent. COPD is a disease of the respiratory systems. Airways and lungs are the two main tissues involved in COPD. At the same time, COPD omics studies also rely on additional sample sources like blood, sputum, and urine due to their easier accessibility. In addition, bronchoalveolar lavage fluid (BALF), and exhaled breath condensate (EBC) have also been collected from COPD patients for omics studies. The omics profiling will be greatly affected by the tissue of choice, except for the genomics. Therefore, we focused on the collection of omics data from the lungs in this project to directly represent the underlying molecular disturbance in COPD. Omics data from BALF were used as an alternative when there was a lack of lung omics data. BALF has been considered as a reliable representation of lung alveoli and lower respiratory airways (Goldstein et al., 1990).

Finally, we should be able to look for potential drug repurposing opportunities among this prioritized gene list. Some genes in the list may be the target for existing drugs for the treatment of other diseases. Some genes might be similar to the known gene target for existing drugs. These genes can lead us to the drugs that may have the potential to be repurposed for COPD treatment. The effectiveness of the pipeline in identifying potential drug repurposing targets can be supported by a follow-up literature search on these genes or drugs. We expect that most of the potential candidates will be identified de novo and will require further confirmation and validation in future projects.

## 2.3 Aim 1: Obtain initial signature genes in COPD by transcriptomic analysis

### 2.3.1 Material and method

#### *2.3.1.1 Transcriptomic data*

We obtained COPD lung transcriptomics datasets for the identification of initial COPD signature genes from publicly available sources. We searched for transcriptomics data generated by microarray and RNA-Sequencing experiments, as both are high-throughput and can measure the gene expression profile at the whole genome level. The search was performed on the Gene Expression Omnibus (GEO) database (www.ncbi.nlm.nih.gov/geo/), one of the biggest data repositories for gene expression data. GEO is a public repository for transcriptomics data generated and submitted by the research community, mostly from experiments on microarray and next-generation sequencing platforms. It covers a wide range of organisms, tissues, and diseases, which makes it a valuable resource for transcriptomics studies (Barrett et al., 2012).

Specifically, we queried the GEO database with the keyword “COPD” and “Homo sapiens” filter option. Then we checked the data description and manually screened for gene expression datasets generated from the lungs of COPD patients and matching controls. Finally, the dataset with the most number of human samples and the most detailed COPD patients status was used as the discovery dataset for the identification of initial signature genes in COPD. Additional datasets with subjects of similar COPD stages (based on the COPD GOLD classification) were included as the validation datasets.

All datasets were downloaded within R by the GEOquery package for a streamlined analysis process (Davis & Meltzer, 2007). The patient clinical characteristics of spirometry measurement, $FEV_1$, were extracted from the datasets. The GOLD stages of the COPD patients were then determined by the $FEV_1$ value of the subjects according to the GOLD guideline, as shown in Table 1. Additional clinical characteristics including gender, age, and smoking status, were obtained for further statistical analysis in R. The comparison of gender and smoking status between case and control groups was made using Fisher's exact test. The difference in the age between case and control groups was analyzed with the Mann-Whitney U test.

#### *2.3.1.2 Differential expression analysis*

Before running differential expression analysis, we need to make sure that the expression matrices in the dataset have been properly pre-processed. The raw microarray intensity values need to be background corrected, normalized to transform into reliable expression levels (Page et al., 2007). The expression data in the discovery dataset has already been background corrected and normalized using a pairwise cyclic loess approach, which is a commonly used normalization method for microarray platforms like Agilent (Ballman, Grill, Oberg, & Therneau, 2004). The expression levels for probes have already been summarized to one probe per gene by selecting the highest intensity probes.

In addition, we need to select a subset of patient samples to focus on very severe COPD for the initial analysis. Due to the heterogeneity of COPD, patients at different stages may manifest a variety of symptoms and be receive different treatment options.

For this analysis, we started with the patients in the COPD GOLD IV group, which represents the most severe type of patients. We may expand the analysis to other subsets of the patients in the future after we have tested and validated our pipeline.

Next, for the differential expression analysis, we used the limma package (Ritchie et al., 2015) as it is a widely used method for microarray data. Basically, it fits the gene expression to a linear model and used empirical Bayes to borrow information between genes in a dynamic way. Gender and age of the subjects were used in the linear model as covariates. Then t-statistics was applied to identify the up-regulated and down-regulated probes in GOLD IV patients vs. healthy controls. In order to adjust for the random appearance of significant results during multiple statistics analysis, adjusted p-values were obtained by applying Benjamini and Hochberg multiple comparison adjustment to the raw p-values (Benjamini & Hochberg, 1995). The probe IDs were mapped to gene symbols with the Agilent chip annotation package "hgug4112a.db" in R (Carlson, 2016). Finally, COPD signature genes were selected based on both adjusted p-values and fold change of the gene expression levels. The differential gene analysis results were visualized as a volcano plot.

#### *2.3.1.3 Functional enrichment*

Genes work in a team rather than individually. Together they perform specific tasks, which are called pathways. Although differential gene expression analysis is able to detect the individual up/down-regulated genes in disease status, we can also look at the pathway levels to see if any pathways are over-represented in disease. One way to do this is through pathway enrichment analysis after the identification of differentially expressed genes. With the knowledge of the overall relative abundance of differentially expressed genes in the dataset, we can compare to it the percent of significant genes in a specific pathway. The higher abundance of differential expressed genes within a particular pathway than in the overall level, the greater that pathway will be considered to be enriched in the disease status.

We performed over-representation analysis on the DEGs from differential expression analysis. Enriched biological processes in COPD were identified with the pathway annotations in the gene ontology database (http://geneontology.org). This analysis was carried out by DOSE package in R and visualized with clusterProfiler package (Yu, Wang, Yan, & He, 2015; Yu, Wang, Han, & He, 2012).

#### *2.3.1.4 Gene set enrichment analysis*

Gene set enrichment analysis (GSEA) was performed to evaluate whether the discovery set DEGs were enriched in the validation COPD cohort. GSEA algorithm compares gene sets to a ranked gene list and calculates enrichment score and p-values by permutation test (Subramanian et al., 2005). First, a ranked gene list was constructed by sorting genes in the validation set by the signed p-values from differential expression analysis (equation 1) in decreasing order. Next, differentially expressed genes (DEGs) in the discovery set were split into two gene sets of up-regulated and down-regulated DEGs. A negative control gene set of genes up-regulated in Alzheimer's disease was obtained from GSEA Molecular Signatures Database (Blalock et al., 2004; Liberzon et al., 2011). The selection of the background genome may affect the performance of GSEA by sample source bias(Simillion, Liechti, Lischer, Ioannidis, & Bruggmann, 2017). Therefore, we selected the genes that were detected in both discovery and validation sets as background genome. The ranked gene list and gene sets were updated to only include genes from the background genome. Finally, the GSEA analysis and visualization were carried out by the fgsea package in R (Korotkevich, Sukhov, & Sergushichev, 2019).

$$\text{Ranking metrics} = sign(log_2\text{fold\_change}) \times -log_{10}(\text{p\_value})$$

Equation 1. Determination of ranking metrics

#### *2.3.1.5 Over-representation test of the differentially expressed genes*

The over-representation test was performed to assess whether the differentially expressed genes (DEGs) identified in the discovery set were associated with the DEGs identified in the validation set. In principle, the ratio of the overlapping DEGs will be compared to the random chance that such overlap may happen to another gene set in the genome (DEGs in the validation set in this case) with hypergeometric distribution by fisher exact test (Sprent, 2011). In order to properly run the over-representation test, we first defined a uniform background genome to address the inconsistent numbers of probe genes in the microarray platforms in the discovery and validation set. The set of genes screened by both microarray platforms was used as the background genome. Next, we identified DEGs in GOLD IV compared to healthy controls in the validation set with the same age-and-gender-adjusted linear model as in the discovery set. Then we updated both discovery and validation DEGs by only including the ones in the previously defined background genome. Finally, the Fisher exact test was performed on the 2 x 2 contingency table with the number of overlapping DEGs, the number of remaining non-overlapping DEGs in the discovery set, the number of DEGs in the validation datasets, the number of genes not identified as validation DEGs in the genome (Table 5).

Table 5. Contingency table for over-representation test

| | DEGs in discovery set (D) | Genome (G) |
|---|---|---|
| DEGs in validation set (V) | $D \cap V$ | $G \cap V$ |
| DEGs not in validation set | $D - D \cap V$ | $G - G \cap V$ |

DEGs: Differentially expressed genes. D: DEGs in discovery set. V: DEGs in validation set. G: Common genes screened by both discovery and validation set

### 2.3.2 Results and discussion

#### *2.3.2.1 COPD lung transcriptomics datasets and patient stratification*

The search on the GEO database with the keyword "COPD" returned 5,091 results (searched on 08/29/2019). After filtering on the Homo Sepian organism and manually checking the dataset descriptions, we found seven lung transcriptomics datasets related to COPD (Table 6).

The dataset with the highest number of human samples was dataset GEO23546 (Bossé et al., 2012), but there was no corresponding COPD disease status. The lack of patient disease information prevented proper re-analysis of the transcriptomics data. Therefore, the dataset GSE47460 was selected as the source for the discovery dataset in the transcriptomics analysis due to its 2$^{nd}$ largest sample size and the most complete COPD disease status. This dataset provided microarray transcriptomics data generated from total RNA extracted from the whole lung homogenate of 582 total subjects. Among all, 254 were subjects with interstitial lung disease, 220 were subjects with COPD, and 108 were healthy controls. The expression levels of 15,261 probe genes were reported in the microarray dataset. The original study for this dataset focused on the expression profile in idiopathic pulmonary fibrosis (Tan et al., 2016). Here, we re-analyzed the data focusing on COPD patients. The COPD cases in this dataset consisted of GOLD I to GOLD IV patients. The controls were a mix of healthy smokers, non-smokers, or ex-smokers. Since COPD is a disease of great heterogeneity, we stratified the patient samples based on their COPD disease stages and chose the microarray of GOLD IV patients as the discovery dataset for transcriptomics analysis.

Table 6. COPD Lung Transcriptomics Datasets

| Dataset | Tissue | No. of COPD | No. of Control | Case | Control | Platform |
|---|---|---|---|---|---|---|
| GSE23546* | Lung | 446 | 389 | COPD | healthy subjects | Microarray: Affymetrix 2.0-HuRSTA |
| GSE47460 | Lung | 220 | 108 | GOLD I - IV | healthy subjects (smoker/non-smoker/ex-smoker) | Microarray: Agilent-014850 Agilent-028004 |
| GSE76925 | Lung | 110 | 40 | GOLD III - IV | healthy ex-smoker | Microarray: Illumina HumanHT-12 V4.0 expression beadchip |
| GSE38974 | Lung | 23 | 9 | COPD | healthy smokers | Microarray: Agilent-G4112F |
| GSE1650 | Lung | 18 | 12 | severe emphysema | no or mild emphysema | Microarray: Affymetrix HG-U133A |
| GSE37768 | Lung | 18 | 9 | GOLD II | healthy smoker | Microarray: Affymetrix HG-U133_Plus_2 |
| GSE57148 | Lung | 98 | 91 | GOLD I-II | healthy subjects | RNA-Seq: Illumina-HiSeq 2000 |

* there is no corresponding individual COPD disease status for this dataset

The dataset GSE76925 was chosen as the source for the validation dataset in the transcriptomics analysis due to its large sample size and the similar subject disease status as the discovery dataset. This dataset provided microarray data generated from the RNA extracted from the homogenized lung tissues. There were a total of 111 COPD patients and 40 healthy controls. The COPD patients belonged to GOLD III or GOLD IV stage, and the controls were healthy ex-smokers who quitted smoking at least one month before the study (Morrow et al., 2017). Similarly, we stratified the patient samples to include only GOLD IV patients as cases in the validation dataset.

In order to account for the effect of smoking, we only included the ex-smokers as the controls for both discovery and validation set. In total, there were 54 cases and 68 controls in the discovery set and 71 cases and 40 controls in the validation set. The other datasets found in the GEO database were not used in the transcriptomics analysis because of their smaller sample sizes or different case/control grouping strategies (Ezzie et al., 2012; Jin Kim et al., 2015; Spira et al., 2004). Namely, they were not comprised of GOLD IV samples as cases and healthy ex-smokers as controls. Nevertheless, these datasets will be useful if we decide to focus on the other subgroups of COPD patients in the future.

#### *2.3.2.2 Sample demographics*

The discovery set and validation set were extracted from GSE47460 and GSE76925 datasets, respectively. In both discovery and validation, the cases were defined as the most severe COPD subjects, namely, the GOLD IV patients. The controls were defined as healthy ex-smokers. The subjects with missing gender, age, or smoking status were excluded. There were 121 subjects (54 cases, 68 controls) in the discovery set and 111 subjects (71 cases, 40 controls) in the validation set. The sample demographics of the discovery and validation sets are shown in Table 7, with information about subject age, gender, and smoking status.

Table 7. Sample demographics of discovery and validation transcriptomics sets

| | **Discovery set** | | **Validation set** | |
|---|---|---|---|---|
| | **GOLD IV** | **Control** | **GOLD IV** | **Control** |
| **Number of subjects** | 54 | 68 | 71 | 40 |
| **Age** (years) | 57.2 ± 8.4 | 65.8 ± 10.2 | 61.9 ± 5.6 | 65.7 ± 9 |
| **Gender** (male/female) | 24/30 | 38/30 | 37/34 | 15/25 |
| **Smoking status** (non/ex-smokers) | 1/53 | 0/68 | 0/71 | 0/40 |

In both the discovery set and the validation set, the GOLD IV group was slightly younger compared to the control group. The statistical test showed that the age distributions were significantly different between cases and controls (Mann-Whitney-Wilcoxon Test, p-values<0.01). This suggested that we need to account for the age difference when performing differential gene expression analysis. One of the potential reasons for the age difference could be that patients died of COPD before they reached an older age.

The gender differences between cases and controls in both discovery and validation datasets are not statistically significant (Fisher exact test, p-values > 0.1). Nevertheless, we still chose to adjust for the gender in the differential gene expression analysis as a study showed that lung transcriptomics was affected by sex (Dugo et al., 2016).

It also needs to be stated that the subject characteristics in the original source datasets were more evenly distributed across different groups (e.g., age, gender) by study design, as the original datasets consisted of GOLD IV patients as well as patients of other lung diseases. However, since we only focused on the GOLD IV cohort, the proportion of age and gender within this cohort might be slightly skewed. Most GOLD IV cases in the datasets were ex-smokers. This was in accordance with the fact that smoking is the major risk factor for COPD (U.S. Department of Health and Human Services, 2004), and COPD patients are strongly advised to quit smoking to help control the progression of the disease. As smoking may affect the lung transcriptomics, we only included healthy ex-smokers as the controls to match the smoking status of the cases.

##### *2.3.2.3 Differentially expressed genes in GOLD IV vs. control subjects*

The expression levels for 15,261 probe genes were compared between the GOLD IV very severe COPD patients and the healthy controls. 157 differentially expressed genes were identified between GOLD IV patients and healthy controls. The differentially expressed genes (DEGs) were selected based on the p-values, and the fold changes reported by the statistics test in the differential gene expression analysis. The p-values were first adjusted to address the issue of random significant results caused by multiple testing of all the 15,261 genes. The adjusted p-value of DEGs needed to be less than 0.05. Further, there needed to be at least two-fold change of gene expression levels for DEGs in either up-regulated or down-regulated directions. In other words, the absolute value of log2 fold change needed to be greater or equal to 1. Overall, there were 37 down-regulated genes and 120 up-regulated genes, as illustrated in the volcano plot in Figure 1. The top 10 up-regulated genes and down-regulated genes are summarized in Table 8. For a full list of the DEGs, please refer to APPENDIX A.

As shown in the volcano plot, there are more up-regulated genes in the list of 157 DEGs in GOLD IV patients vs. the healthy controls. This is in line with the previous findings that smoking dis-regulates the gene expression pattern by opening up chromosomes for more active transcription (Kopa & Pawliczak, 2018).

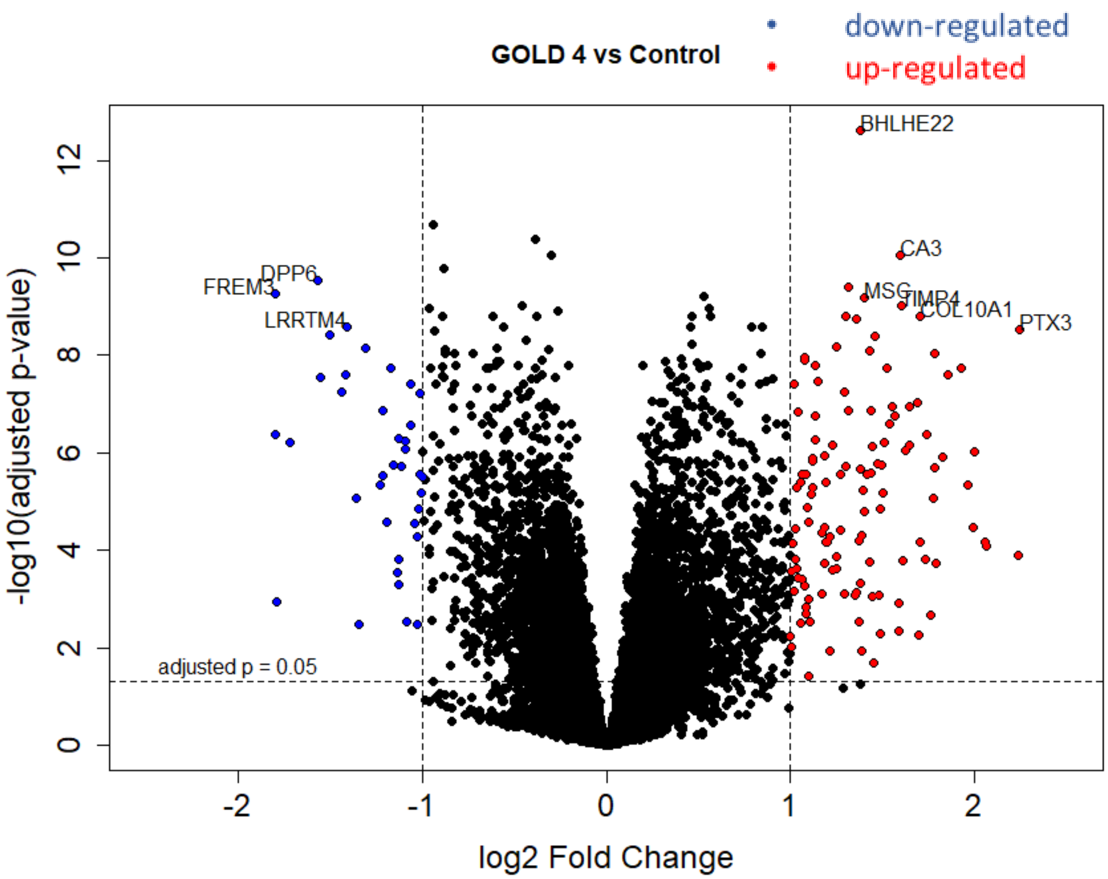



Figure 1. A total of 157 differentially expressed genes were identified between healthy subjects and GOLD IV COPD patients

Table 8. The top 10 up-regulated and down-regulated genes in GOLD IV patients vs. controls

| Probe ID | Gene | P-Value | Adjusted P-value | Fold Change | Average Expression | t-statistics |
|---|---|---|---|---|---|---|
| **up-regulated** | | | | | | |
| A_32_P75581 | BHLHE22 | 1.56E-17 | 2.38E-13 | 2.61 | 8.98 | 9.07 |
| A_23_P20316 | CA3 | 2.43E-14 | 9.01E-11 | 3.02 | 8.81 | 8.01 |
| A_23_P209625 | CYP1B1 | 2.05E-13 | 3.91E-10 | 2.48 | 11.47 | 7.69 |
| A_23_P256948 | MSC | 4.64E-13 | 6.44E-10 | 2.64 | 4.87 | 7.57 |
| A_32_P70315 | TIMP4 | 7.59E-13 | 9.60E-10 | 3.03 | 5.74 | 7.49 |
| A_23_P324754 | CEMIP | 2.33E-12 | 1.62E-09 | 2.46 | 8.05 | 7.31 |
| A_23_P214144 | COL10A1 | 2.38E-12 | 1.62E-09 | 3.26 | 8.00 | 7.31 |
| A_23_P162918 | SERPINA3 | 2.86E-12 | 1.82E-09 | 2.56 | 11.21 | 7.28 |
| A_23_P121064 | PTX3 | 5.90E-12 | 3.00E-09 | 4.74 | 9.17 | 7.17 |
| A_23_P131024 | ZBTB32 | 8.87E-12 | 4.10E-09 | 2.74 | 6.86 | 7.10 |
| A_23_P127584 | NNMT | 1.56E-11 | 6.60E-09 | 2.38 | 14.35 | 7.01 |
| **down-regulated** | | | | | | |
| A_23_P170888 | DPP6 | 1.36E-13 | 2.96E-10 | 0.34 | 6.78 | -7.76 |
| A_24_P738168 | FREM3 | 3.20E-13 | 5.42E-10 | 0.29 | 5.15 | -7.62 |
| A_24_P174294 | LRRTM4 | 4.72E-12 | 2.56E-09 | 0.38 | 3.75 | -7.20 |
| A_23_P217277 | SLITRK2 | 8.06E-12 | 3.84E-09 | 0.35 | 4.40 | -7.12 |
| A_32_P227605 | RGS9BP | 1.80E-11 | 7.24E-09 | 0.40 | 5.79 | -6.99 |
| A_23_P382240 | TMEM26 | 8.01E-11 | 1.77E-08 | 0.44 | 5.86 | -6.74 |
| A_23_P41789 | SLC27A6 | 1.31E-10 | 2.56E-08 | 0.37 | 4.36 | -6.66 |
| A_24_P403150 | FLJ34503 | 1.55E-10 | 2.84E-08 | 0.34 | 5.81 | -6.63 |
| A_23_P354855 | RS1 | 2.48E-10 | 3.87E-08 | 0.48 | 4.78 | -6.55 |
| A_23_P2674 | KRT4 | 4.19E-10 | 5.82E-08 | 0.37 | 9.44 | -6.46 |

#### *2.3.2.4 Functional enrichment*

Gene Ontology enrichment analysis revealed 160 statistically significant over-represented biological processes based on the differentially expressed genes in GOLD IV patients. The top 20 most significant biological processes are shown in Figure 2, together with the number of genes involved in the specific pathways. Most of these pathways are immune response pathways, including granulocyte chemotaxis, leukocyte chemotaxis, and neutrophil migration. This is in accordance with the findings that the disrupted immune response in the lung is the hallmark of COPD (Bhat, Panzica, Kalathil, & Thanavala, 2015). Other over-represented biological processes suggested an involvement of the lung extracellular matrix in COPD, e.g., extracellular structure organization. This is in line with a previous study that the major extracellular composition in COPD patients was altered compared with healthy controls (Annoni et al., 2012).

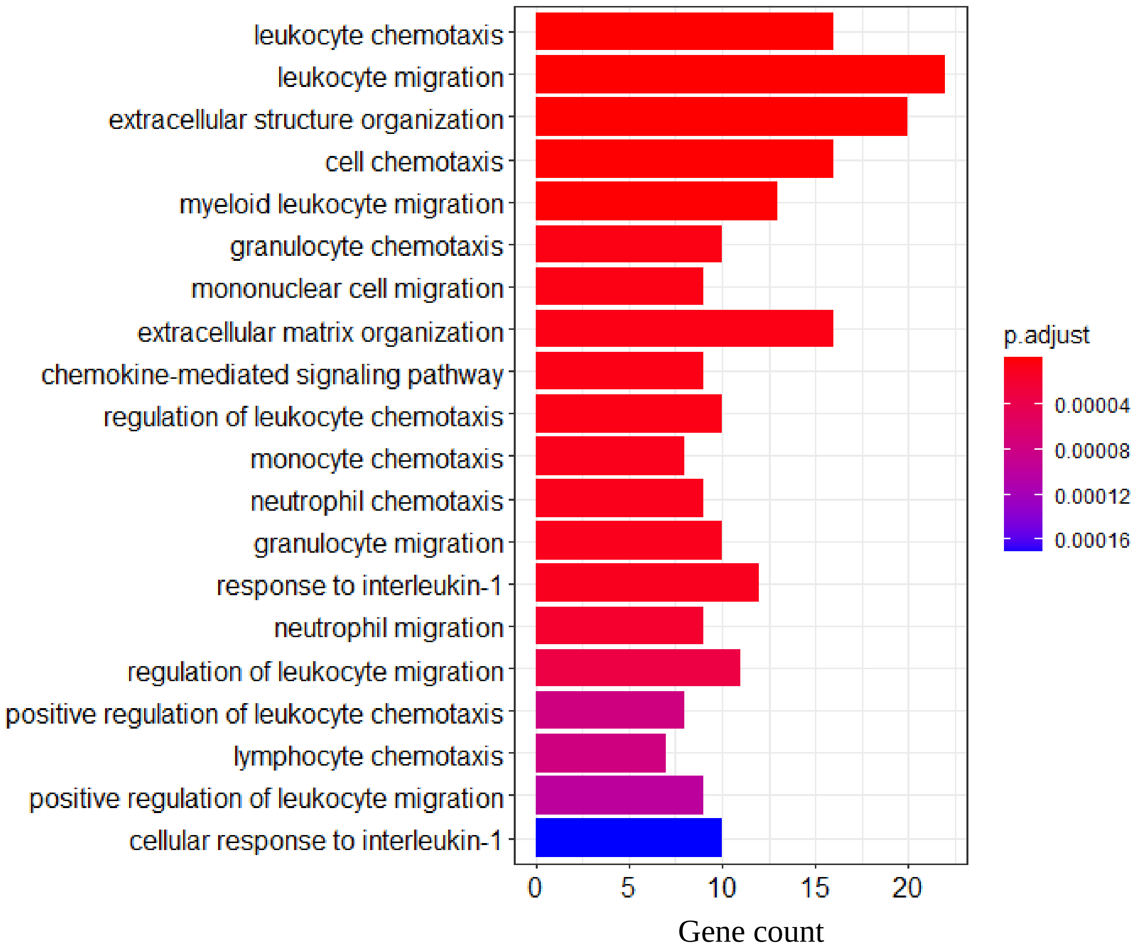
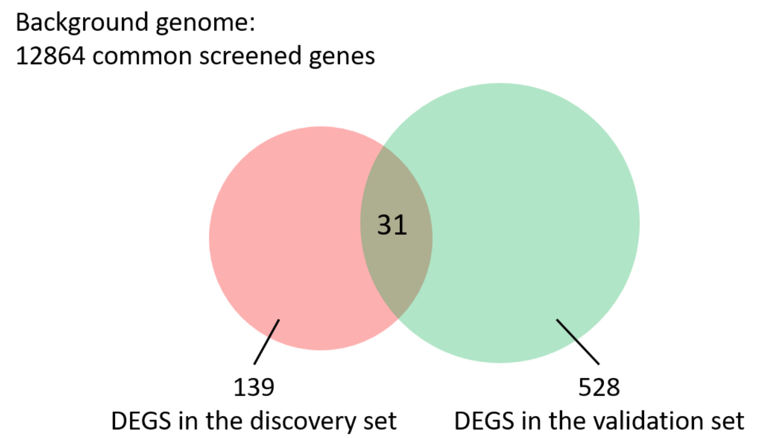



Figure 2. Top 20 statistically significant over-represented biological processes in GOLD IV patients.

#### *2.3.2.5 Validation in independent transcriptomics dataset*

In order to validate the 157 differentially expressed genes (DEGs) identified in the discovery dataset, we compared this gene list to the DEGs identified in a validation dataset with similar COPD cohort. Although different microarray platforms were used in the two datasets, 12,864 common genes were screened by both platforms in the discovery and validation dataset. In the discovery dataset, 139 out of the 157 DEGs belong to the commonly screened genes. In the validation dataset, 528 genes were identified as DEGs in the GOLD IV patients by the same analysis approach in the discovery dataset. There are 31 overlapping DEGs between the discovery and validation sets (Figure 3). Moreover, the fold change directions of all the overlapping DEGs are consistent (APPENDIX C). Next, we statistically evaluated whether the ratio of 31/139 overlapping genes would appear due to random chance by comparing with the ratio of 528 validation DEGs in the 12,864 background genome. The over-representation test reported a p-value $< 10^{-12}$ based on the Fisher exact test on the contingency table in (Figure 3), indicating that our DEGs identified in the discovery set was also associated with the DEGs in the validation set.




| | DEGs in discovery set | Genome |
|---|---|---|
| DEGs in validation set | 31 | 528 |
| DEGs not in validation set | 139 – 31 = 108 | 12864 – 528 = 12336 |

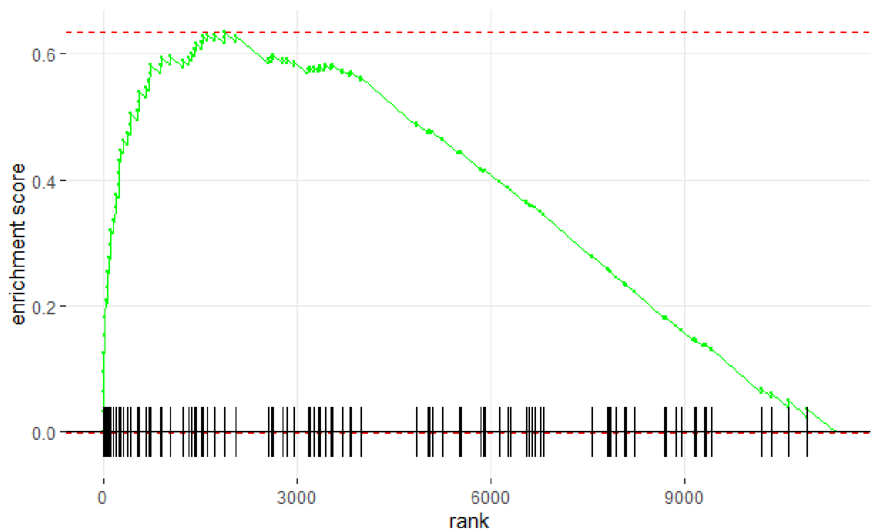
Figure 3. Differentially expressed genes identified by discovery and validation sets

In addition, we further validate our DEGs with a second approach of gene set enrichment analysis. Both the up-regulated and down-regulated DEGs in the discovery set were found to be significantly enriched in the validation COPD cohort when they were compared to the genes of the validation set ranked by differential expression analysis p-values (Table 9). As shown in the enrichment plot, genes in the validation set were ranked by the signed log-transformed p-values. Hence, genes most significantly up-regulated will be at the top of the ranked list, and genes most significantly down-regulated will be at the bottom of the ranked list (Figure 4). GSEA found that up-regulated genes in the discovery set showed up more frequently at the top of the ranked validation gene list. Similarly, the down-regulated genes in the discovery set located more frequently at the bottom of the ranked validation gene list. This suggested an enrichment of both the up-regulated and down-regulated genes in the validation set. On the other hand, the negative control of genes up-regulated in Alzheimer's disease was not shown to be enriched in the validation set.

Table 9. Gene set enrichment analysis result table for DEGs in the discovery set

| Gene set | P-value | Adjusted p-value | Enrichment Score | Normalized Enrichment Score | No. of overlapping genes |
|---|---|---|---|---|---|
| DEG_Discovery_up | 0.00116 | 0.00349 | 0.63554 | 2.13579 | 106 |
| DEG_Discovery_down | 0.00437 | 0.00655 | -0.70377 | -2.2467 | 31 |
| Control_Alzheimer_up | 0.94206 | 0.94206 | 0.23381 | 0.87969 | 1228 |

Enrichment Plot: up-regulated DEGs in discovery set




Enrichment Plot: down-regulated DEGs in discovery set

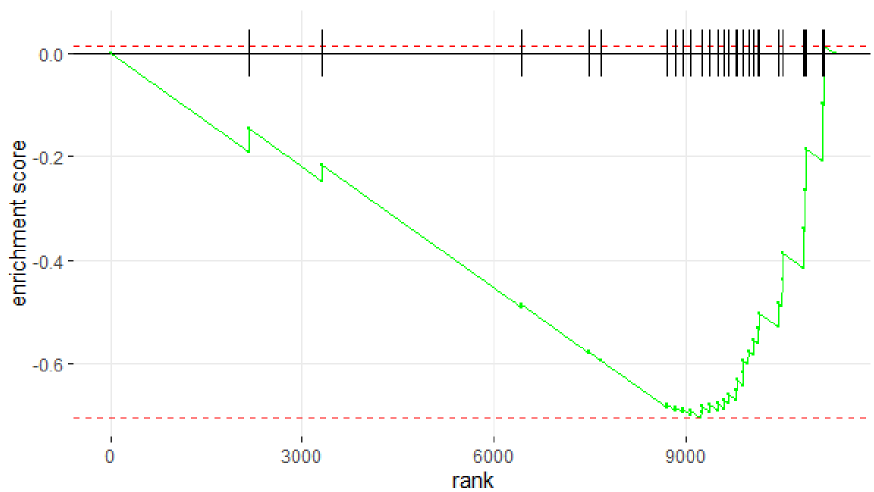



Figure 4. Gene set enrichment analysis plot for up-regulated and down-regulated DEGs in the discovery set.

**Rank**: after differential gene expression analysis, genes in the validation set are ranked in decreasing order by –log10(p-values) * sign of the fold change. **Black vertical line**: gene set hits - DEGs in the discovery set are found in the validation set. **Green line**: Enrichment score profile - when there is a hit, the enrichment score will increase and the green line will climb up. Otherwise, the enrichment score will decrease and the green line will go down. **Red dotted line**: max and min enrichment score

Besides the gene level comparison, we also evaluate the similarity of the DEGs between the discovery and the validation dataset on the pathway level. Gene Ontology enrichment analysis on the 528 DEGs from the validation dataset revealed 33 same over-represented biological processes in GOLD IV patients, including leukocyte activation (GO:0045321), cellular response to cytokine stimulus (GO:0071345) and positive regulation of cell-cell adhesion (GO:0022409).

Overall, compared with a similar COPD transcriptomics dataset, our DEGs were well supported by two different validation approaches. The over-representation test and GSEA are widely used in the bioinformatics fields as approaches for pathway enrichment analysis. Here, we employed these two approaches for the validation of DEGs identified in the discovery set, as we can readily extend the use of the two methods when we treated our DEGs as a pathway or a gene set related to COPD. The same approach has been used by another published study to compare their DEGs to other transcriptomic datasets (Campbell, et al., 2012). However, there are still some limitations to such validation approaches. The percentage of DEGs detected by both discovery and validation set was not high. A similar inconsistency was also observed in another COPD transcriptomic study. Several factors may contribute to this. The first reason was the biological difference among subjects. Even though we have limited our analysis to GOLD IV patients to control the disease heterogeneity, the patients may still represent different subgroups of the disease in terms of other pathophysiological status related to COPD. The second reason is the sample size. Despite that this is already the largest COPD transcriptomics dataset with the necessary disease clinical information for re-analysis, thousands of samples may be needed to obtain a more consistent list of DEGs due to the heterogeneous nature of the disease and small variations on the expression level changes. The third reason is the different platforms. The discovery set was based on the Agilent platform and the validation set was based on the Illumina platform, and cross-platform variation was shown to exist for the expression pattern of some genes (Chen, Hsueh, Delongchamp, Lin, & Tsai, 2007). Nevertheless, it has been demonstrated in a microarray reproducibility study that most of the DEGs identified in a microarray study were still likely true positives even the percent of overlapping genes with other studies appeared to be low (Zhang et al., 2008). Therefore, we moved forward with the DEGs in the discovery set for further analysis in the pipeline.

#### *2.3.2.6 Differentially expressed genes across different stages of COPD*

In order to explore the DEGs across different stages of COPD, differential gene expression analysis was carried out between the control group vs. mild (GOLD I), moderate (GOLD II), and severe COPD (GOLD III) patients. There are 23, 94, 68 patients in GOLD I, II III groups respectively. 1, 6, 157 DEGs were identified between the control group vs. GOLD I, GOLD II, and GOLD III group with the same age and gender adjusted linear model (adjusted p-values $< 0.05$ and fold change $> 2$ or $< 0.5$). Figure 5 shows the overlap between the DEGs identified across different stages of COPD.

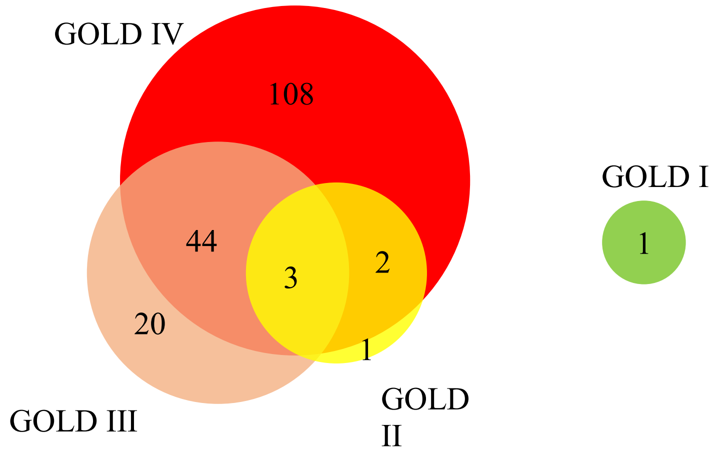



Figure 5. The overlap between the DEGs identified across different stages of COPD

Each circle represents the number of DEGs identified in individual GOLD levels.
DEG: Differentially expressed genes

The number of DEGs increased as the COPD progressed into later stages, suggesting the late-stage COPD of GOLD III and GOLD IV presented a more distinct transcriptomics profile than the early-stage COPD compared to the healthy subjects. A limited number of DEGs were identified in the early stage COPD of GOLD I and II. One potential reason is that the detection of DEGs may be inhibited by the smaller fold change of lung transcriptomics in the early-stage COPD. To increase the detection power, we expanded the sample size by combining both GOLD I and II subjects into one group to represent early-stage COPD. Only three DEGs were identified between early-stage COPD subjects and healthy controls, potentially affected by the fold change criteria in the selection of DEGs. When we excluded the fold change criteria and determined the DEGs solely by an adjusted p-value smaller than 0.05, more DEGs will be identified across all four stages. However, the trend still remained that more DEGs were identified in the later stage COPD. Figure 6 shows the change of DEGs across four stages. 62.5% of the genes were not identified as DEGs across all four GOLD stages and were excluded from the figure. There were some genes constantly up-regulated or down-regulated across stages. On the other hand, there were genes only dysregulated on specific stages. Future work may focus on genes that are constantly dysregulated across stages and genes specific to the individual stage to help the discovery of general or personalized drug targets.

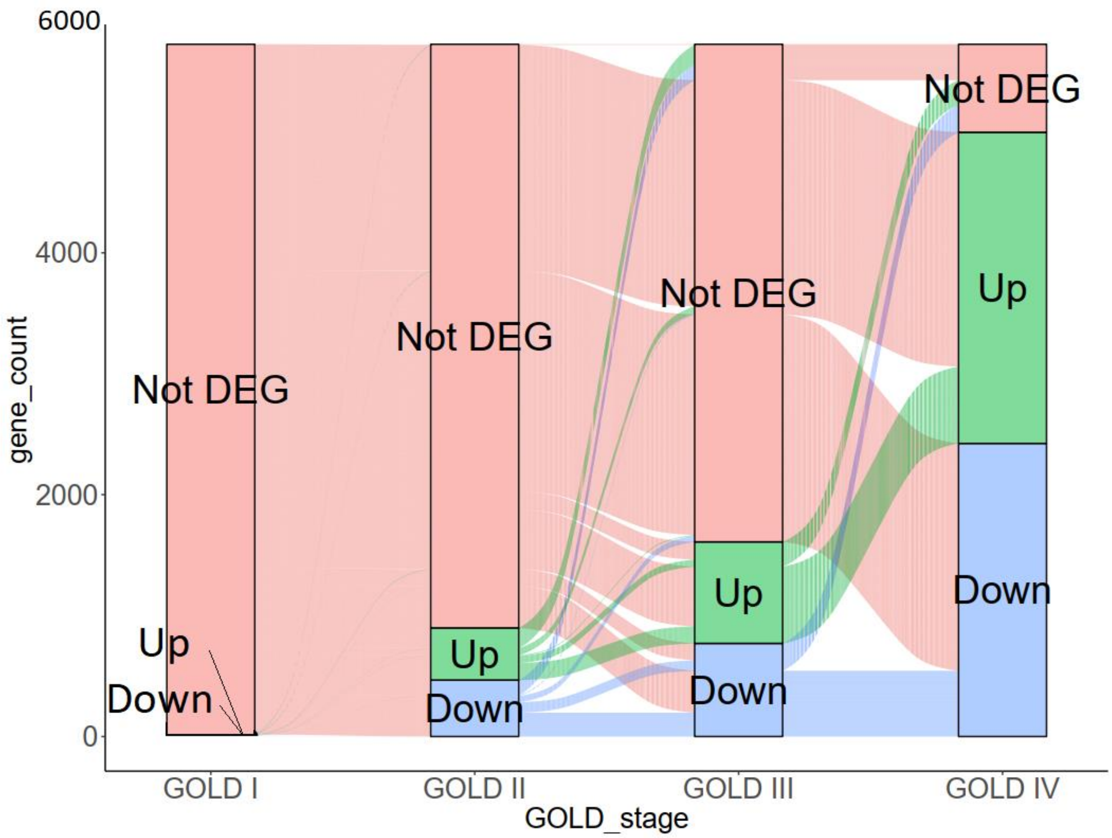



Figure 6. The trend of DEGs across GOLD I to GOLD IV

Not DEG: genes that were not differentially expressed genes in GOLD I-IV patients vs healthy controls; Up: up-regulated differentially expressed genes in GOLD I-IV patients; Down: down-regulated differentially expressed genes in GOLD I-IV patients; gene count: the number of genes belonging to Not DEG, Up, or Down categories. Red flows between each GOLD column represent the status change of genes in “Not DEG” category, i.e. remain “Not DEG”, change to “Up”, or change to “Down”; Blue and green flows represent status changes in “Down” and “Up” genes in a similar way as red flows.

### 2.3.3 Conclusions

157 genes were identified as the initial COPD signature genes through differential gene expression analysis on the transcriptomics data from the largest COPD cohort with complete disease information. Functional enrichment analysis on the 157 genes revealed 160 statistically significant over-represented biological processes, including immune response pathways and extracellular structure organization pathways. Compared with a similar COPD transcriptomics dataset, our DEGs were well supported by two different validation approaches of the over-representation test and GSEA.

## 2.4 Aim 2: Expand and prioritize COPD signature genes through multi-omics integration

### 2.4.1 Material and method

#### *2.4.1.1 GWAS data*

GWAS related to COPD was identified through PubMed search with the keyword "COPD AND ('genome-wide association study')". The abstracts and material sections were reviewed to filter out unrelated publications. In addition, publications mentioned in the COPD GWAS meta-analysis were reviewed to complement the GWAS literature that may be missed during the PubMed search. Information on the subject disease status, genetic loci associated with COPD, and susceptible genes were obtained from GWAS literature.

#### *2.4.1.2 Proteomics data*

A list of differentially expressed proteins in lung tissues of COPD patients was obtained from COPD proteomics literature. The literature search was done by searching the keyword "(COPD lung) AND (proteome OR proteomics)" on the PubMed website (pubmed.ncbi.nlm.nih.gov). Then the abstracts and the material sections of the papers in the search results were checked for the information on the disease status of subjects, the tissue source, and the experimental approaches on sample analysis. Finally, a list of differentially expressed proteins was compiled based on the results of the papers containing proteomics analysis on lung tissues from COPD patients. The protein names were matched to corresponding gene symbols by searching on the UniProt website (www.uniprot.com) for the purpose of multi-omics integration.

#### 2.4.1.3 Metabolomics data

Metabolomics study related to COPD was identified through PubMed search with the keyword "(COPD lung) AND (metabolomics OR metabolome)". Information on the subject disease status, sample source, and experimental approaches were obtained from the abstracts and the material sections of the papers in the search results. Metabolomics studies based on bronchoalveolar lavage fluid from COPD patients were included as the source for the identification of metabolites associated with COPD. Metabolomics class information was obtained from the Metabolomics Workbench (www.metabolomicsworkbench.org).

#### 2.4.1.4 Integration of metabolomics data with MAGI

The MAGI (metabolome and genome integration) software was used to identify proteins and reactions linked to the metabolites associated with COPD (Erbilgin et al., 2019). The metabolites compound names were mapped to the IUPAC International Chemical Identifier (InChI) keys as MAGI inputs (Heller, McNaught, Pletnev, Stein, & Tchekhovskoi, 2015). MAGI score of greater than 4 was used as criteria to filter proteins with a high confidence level of the metabolites and genes association. Protein names were mapped to gene symbols for the uniform integration with other omics levels.

#### 2.4.1.5 Protein-protein interaction data

We obtained the physical and functional protein-protein interaction data from the STRING database v11 (Szklarczyk et al., 2019). The interactions were based on several types of sources, including experimental data, computational prediction, and public literature text mining. We filtered for interactions with high confidence level interactions (interaction score >=0.7) to be included in the analysis.

### *2.4.1.6 Calculation of distances between candidate genes and omics signature genes*

The distances between genes were calculated from the protein-protein interaction network. Figure 7 is an illustration of the distances between a candidate gene and two other genes in the network, as well as the distance of the candidate gene to itself.

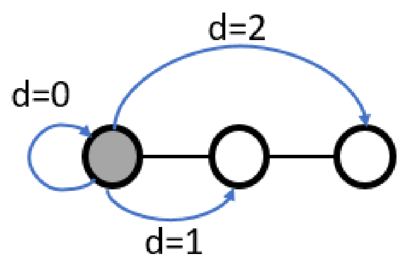



Figure 7. Illustration of distances between genes in the network

Gray circle: candidate gene; white circle: additional genes in the network; black line: physical or functional interaction between two genes; blue line: distances of candidate genes to genes in the network

First, three distances matrices were created from the paired protein interaction table retrieved from STRING database: 1) distance between COPD genomics signature genes and all candidate genes in the genome, 2) distance between COPD transcriptomics signature genes and all candidate genes in the genome, and 3) distance from all candidate genes in the genome to COPD signature genes identified from proteomics study or proteins linked to COPD metabolomics. The distance matrices were created with the igraph package and visualized with the pheatmap package in R.

Next, the shortest distances of an individual candidate gene to three sets of omics signature genes were calculated as $d^{Transcriptomics}$, $d^{GWAS}$, and $d^{Protein}$, as shown in the example in Figure 8. For each candidate gene, it had multiple distance values representing distances to all genes in the signature gene set. The shortest distance was used to represent the proximity of the candidate gene to the individual omics signature gene set.

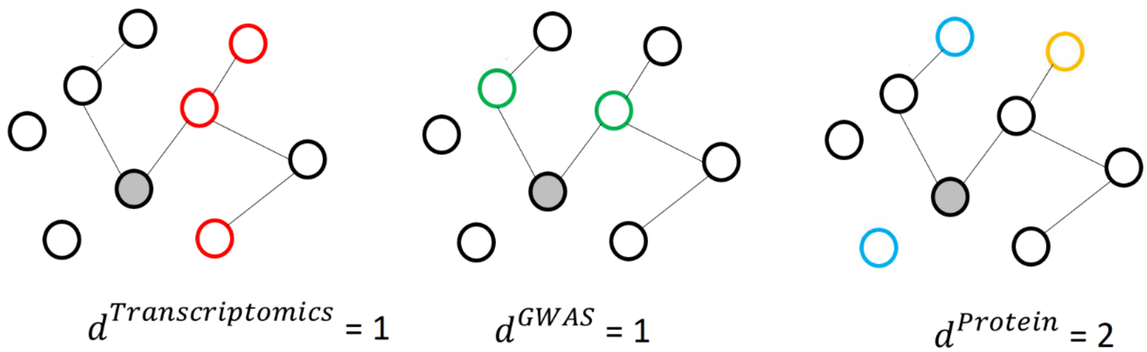



Figure 8. Shortest distances of a candidate gene to three sets of omics signature genes

Circle: genes; grey circle: candidate gene; red circle: COPD genomics signature genes; green circle: COPD transcriptomics signature genes; blue circle: COPD proteomics signature genes; yellow circle: proteins linked to the COPD-associated metabolomics

Finally, $d^{Transcriptomics}$, $d^{GWAS}$, and $d^{Protein}$ were summed up to a sum distance to represent the overall proximity of a candidate gene to all omics signature genes (Equation 2). A smaller sum distance indicates a closer association to signature genes from all omics levels.

$$\hat{d}_i^{sum} = \min_j d_{ij}^{Transcriptomics} + \min_j d_{ij}^{GWAS} + \min_j d_{ij}^{Protein}$$

Equation 2. Determination of sum distance as overall proximity to all omics signature genes

*i is the candidate gene and j is the gene to calculate distance from in the network

#### *2.4.1.7 Heat diffusion network analysis with Cytoscape*

Cytoscape 3.7.0 (Shannon et al., 2003), a network analysis software, was used for the implementation of the heat diffusion network analysis. First, the whole human interactome was retrieved from the STRING database through the EnrichmentMap app 3.1.0 in Cytoscape. Only interactions with greater than 0.7 STRING interaction scores were loaded into the software. Next, the signature genes identified from genomics, transcriptomics, proteomics, metabolomics levels were imported into Cytoscape as gene node annotation table. Then, heat diffusion network analysis was carried out with the diffusion plugin 1.6.0 in Cytoscape (Carlin, Demchak, Pratt, Sage, & Ideker, 2017). The nodes representing signature genes from all omics levels were selected as diffusion network input, and the analysis was run with the default settings.

### 2.4.2 Result and discussion

#### *2.4.2.1 Susceptible genes identified by COPD GWAS*

20 GWAS publications investigating the genetic association with COPD, emphysema, or lung function in COPD were discovered through PubMed search and manual screening. This list of publications was then compared to the GWAS included in a recent COPD and lung function GWAS meta-analysis to ensure the comprehensive inclusion of all related GWAS (Lamontagne et al., 2018). In comparison, our publication list included one more recent GWAS publication in 2018 (Burkart et al., 2018), and excluded GWAS related to lung function in other conditions, or biomarkers association studies. APPENDIX E shows the genetic loci and corresponding genes associated with

COPD. In total, there were 213 genetic loci and 223 suspected genes associated with COPD based on the 20 GWAS studies.

GWAS provides valuable insight to reveal the potential genetic factors in complex diseases like COPD. There are usually thousands or even tens of thousands of subjects in a GWAS. This large sample size offers the potential to identify genetic variants with a smaller effect size, which is very common in complex disease. With the development of array-based assay and next-generation sequencing technology, GWAS has become one of the most abundant types of omics data. We can take advantage of this genetic information and integrate it with our multi-omics pipeline. However, genomic data are also limited on the functional annotation. Some susceptible genes reported by GWAS are based on the location of the variants to the nearby genes. eQTL analysis may offer stronger evidence in mapping genetic variants to causal genes (Nica et al., 2010).

#### *2.4.2.2 Differentially expressed proteins in COPD patients*

Proteomics data from COPD patients were included in the pipeline to complement the other omics levels. PubMed search and manual screening identified four publications related to proteomics studies performed on lung tissues from COPD patients and healthy controls (Table 10). Collectively, 58 proteins were found to be differentially expressed in COPD patients compared to healthy controls based on the results from the four publications.

Table 10. Proteomics studies on lung tissues from COPD patients

| Year | Authors | Tissue | Groups | Differentially expressed proteins (shown as the corresponding gene symbols) |
|---|---|---|---|---|
| 2013 | Barrero, Carlos A., et al | Lung | Healthy smokers vs COPD GOLD IV patients | H3F3B/H3F3A, H2BFS, H4C1, ICAM1, MES, MYH9, MYH10 , VIM , FIBA , FIBG , EHD2 , TLN1 , PERM1 , DHE3, ATP1A1 , TGM2 , COL6A3 |
| 2016 | Ohlmeier, Steffen, et al | Lung | Healthy smokers vs GOLD I-II patients | CTSD, TGM2, TPP1, FBP1, COL6A1, EEF1G, MYL6, ASAH1, FTL, PRDX3, SERPINB1, TPM1, UQCRC1, AMBP, APOA1, HBB, ITIH4, TF |
| | | | Healthy smokers vs COPD III-IV patients | TGM2, S100A4, ACTN1, ACTA2, CCT5, COL6A1, EEF1G, HSPB1, HSP90AB1, MFAP4, MYL6, NAPA, TPM1, TPM2, USP5, ANXA7, PTRF, TPM4, UQCRC1, AMBP, APOA1, TF, CLU |
| 2009 | Lee, Eun Joo, et al | Lung | Healthy smokers vs COPD patients | MMP13, TFAP4, PSME3IP1, HDGF, NAPA, CETN3, CAV1, GLRX3, TPT1, CALML3 |
| 2008 | Ohlmeier, Steffen, et al | Lung | Healthy controls (never smokers and ex-smokers) vs COPD IV patients | SFTPA2/SFTPA1 |

Several proteins in this list of 58 proteins have been previously linked to biological processes in the lung. Matrix metalloproteinase 13 (MMP13) is found to be upregulated in COPD patients in the lung proteomics study from Eun Joo Lee and colleagues (E. J. Lee et al., 2009). MMP13 belongs to the matrix metalloproteinases family. Several members in this family are found to be involved in COPD (Ilumets et al., 2007; Omachi, Eisner, Rames, Markovtsova, & Blanc, 2011), potentially by the breakdown of lung matrix and as inflammatory modulators (Hautamaki, Kobayashi, Senior, & Shapiro, 1997; Nénan et al., 2005). In a recent paper, MMP13 was reported to slow the rate of epithelial repair in vitro, and MMP13 inhibitor was able to promote the epithelial repair(Howell, Smith, & Shute, 2018). This suggested that MMP13 may be a promising drug target for COPD, since COPD is heavily involved in epithelial injury and repair (Crosby & Waters, 2010).

Another example is the N-ethylmaleimide-sensitive factor attachment protein α (αSNAP). This protein is encoded by NAPA gene, and the protein expression levels were found to be dysregulated in COPD lungs in two proteomics studies (E. J. Lee et al., 2009; Ohlmeier et al., 2016). It plays a major role in the intracellular membrane fusion of vesicles (Clary, Griff, & Rothman, 1990). It is involved in the surfactant secretion in the lung (Abonyo et al., 2003), as well as the secretory processes in other tissues (De Paola, Bello, & Michaut, 2015; Südhof & Rizo, 2011; Tomes et al., 2005).

Overall, 58 differentially expressed proteins were identified from literature mining. However, the differentially expressed proteins identified by the four COPD lung proteomics studies do not overlap with the differentially expressed genes identified by COPD lung transcriptomics analysis. Although proteins are translated from mRNA transcripts, protein abundance does not always correlate well with mRNA expression (Ghazalpour et al., 2011; J. Wang, Wu, Chen, & Zhang, 2016). Therefore, the inclusion of proteomic data is necessary to gain a comprehensive view of the biology systems on the molecular level.

#### 2.4.2.3 COPD-associated metabolites identified in metabolomics studies

The literature search identified several COPD metabolomics studies. There is currently no metabolomics study on COPD lung tissues. Current COPD metabolomics studies were conducted on various human sample sources, including blood, urine, bronchoalveolar lavage fluid (BALF), and exhaled breath condensate (EBC) (Adamko et al., 2015; Halper-Stromberg et al., 2019; Kilk et al., 2018; Naz et al., 2017). BALF is obtained by flushing saline into the lung and collecting the fluid back through a bronchoscopy tube that passes through the lower respiratory airway (Baughman, 2007). During this process, lung epithelial lining fluid is sampled with the saline and can be used for diagnostic purposes or biological studies. On the other hand, EBC is a less invasive approach to sample the airway lining. The air exhaled by patients is collected and cooled into condensate, which may contain volatile as well as some non-volatile metabolites produced by the lung and airway lining (Kubáň & Foret, 2013; Liang, Yeligar, & Brown, 2012). We chose to include the BALF metabolomics study in our pipeline as BALF is more closely linked to lung tissue metabolites compared to blood and urine. Moreover, BALF is considered more reliable as a representation of lung alveoli and lower respiratory airways, while EBC may be affected by additional parts in the respiratory systems like upper respiratory airways, throat, and mouth.

Table 11 lists the BALF metabolites most significantly associated with the lung function in COPD patients based on the results from the metabolomics study by Halper-Stonmberg et al. These metabolites belong to several classes, including glycerophospholipid, sphingolipid, organoheterocyclic compounds and others. Several of the metabolites in this list were previously reported as blood biomarkers in COPD patients, including ceramides and phosphatidylcholines (Cruickshank-Quinn et al., 2018). Ceramides are a type of lipids that belong to the sphingolipid family. The increased level of ceramides found by the COPD metabolomics study was also observed in other studies (Petrache et al., 2006; Scarpa et al., 2013). Therefore, ceramides have been proposed by the previous study as a potential drug target (Tibboel, Reiss, de Jongste, & Post, 2013). Phosphatidylcholines are a class of phospholipids that is the main component of the cell membrane and lung surfactant (Holm, Wang, Egan, & Notter, 1996; Vance & Vance, 2008). The level of phosphatidylcholines has been found to correlate with lung function in COPD (Agudelo et al., 2020).

In summary, a list of 25 metabolites was identified as lung metabolites associated with COPD through the literature search. Metabolomics data provided an additional layer of biological information into our target identification pipeline. While the omics data from genomics to proteomics levels can suggest the key gene products in specific cellular functions and processes, metabolomics represents the end products catalyzed by those gene products and reflects the physiological state of biological systems (Hollywood et al., 2006). However, the aims of many COPD metabolomics studies are to identify the dysregulated metabolites that may serve as the biomarkers for the disease (Celli et al., 2018; Deja et al., 2014; Maniscalco et al., 2019). Therefore, these studies tend to employ a less invasive approach to collect samples from the patients which does not involve the samples from the respiratory system. Here, we took the use of a recent metabolomics study on BALF from COPD patients and obtained a list of lung metabolites associated with COPD.

Table 11. BALF metabolites associated with the lung function in COPD

| Compound | Compound Class | Compound super class |
|---|---|---|
| Ceramide (d18:1/16:0) | Sphingolipid | Sphingolipid |
| Decaprenyl phosphate | Polyprenols | Prenol Lipids |
| Glycocholic acid | Bile acids | Sterol Lipids |
| Homocysteine | Amino acids and peptides | Organic acids |
| Lophocerine | Tetrahydroisoquinolines | Organoheterocyclic compounds |
| L-Threonylcarbamoyladenylate | Purines | Nucleic acids |
| MGDG (36:5) | Glycosyldiradylglycerols | Glycerolipids |
| Mycalamide B | / | / |
| N-palmitoyl glycine | Fatty amides | Fatty Acyls |
| p-cresol | Phenols | Benzenoids |
| PC (32:1) | Glycerophosphocholines | Glycerophospholipid |
| PC (34:2) | Glycerophosphocholines | Glycerophospholipid |
| PC (36:4) | Glycerophosphocholines | Glycerophospholipid |
| PC (40:6) | Glycerophosphocholines | Glycerophospholipid |
| PC (40:6) (isomer) | Glycerophosphocholines | Glycerophospholipid |
| CL (70:0) | Cardiolipins | Glycerophospholipid |
| CL (72:7) | Cardiolipins | Glycerophospholipid |
| PE (34:2) | Glycerophosphoethanolamines | Glycerophospholipid |
| PE (35:1) | Glycerophosphoethanolamines | Glycerophospholipid |
| PE (36:3) | Glycerophosphoethanolamines | Glycerophospholipid |
| PE (38:3) | Glycerophosphoethanolamines | Glycerophospholipid |
| PS (37:3) | Glycerophosphoethanolamines | Glycerophospholipid |
| S-(Phenylacetothiohydroximoyl)-L-cysteine | / | / |
| SM (d18:1/24:1) | Sphingomyelins | Sphingolipids |
| SQMG (16:1) | Sulfolipid | Glycolipids |

MGDG - Monogalactosyldiacylglycerol. PC - Phosphatidylcholine. CL - Cardiolipin.
PE - Phosphatidylethanolamine. PS - Phosphatidylserine. SM - Sphingomyelin.
SQMG - Sulfoquinovosyl monoacylglycerol.

*2.4.2.4 Proteins associated with the metabolites from metabolomics data*

The biomedical reactions and protein enzymes linked with the COPD-associated metabolites were identified by the MAGI software. 16 out of 25 metabolites were searched by MAGI, as the remaining compounds did not present in the MAGI database. The full list of inputs can be found in APPENDIX D. In total, 3,761 biomedical reactions and 533 proteins were reported to be associated with the 16 input metabolites. The associations were with a MAGI score ranging from 0.02 to 6.33 (Figure 9). 33 unique proteins were identified with high confidence (MAGI score >4), corresponding to 13 genes. Multiple isoform proteins were mapped to one single gene. The 13 proteins were included in the pipeline for further analysis.

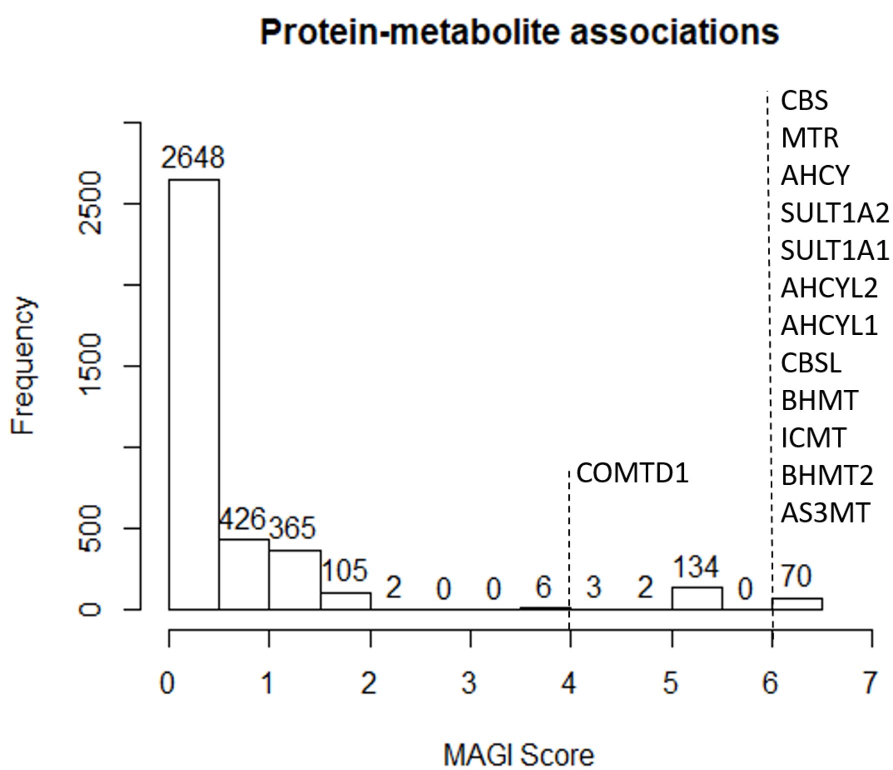



Figure 9. Protein-metabolite associations identified by MAGI

Frequency: the number of protein-metabolite associations within specific MAGI score range; 12 and 1 proteins with a MAGI score higher than 6, and between 4 to 6 is marked on the plot, respectively

When integrated with the other omics data, metabolomics can compensate for the knowledge gained from the gene-centric omics levels. In a living organism, proteins work as enzymes that catalyze the biological reactions to produce metabolites. Therefore, if a metabolite is associated with a certain phenotype or disease, the protein/gene involved in the reaction to generate the specific metabolite might also be dysregulated. Thus, it will offer another resource and supportive evidence to identify COPD signature genes as potential targets, by finding the genes or proteins associated with the dysregulated metabolites. Based on this rationale, we added 13 proteins associated with the dysregulated metabolites to our COPD signature genes. For example, one of the metabolites associated with COPD was homocysteine. MAGI identified the hydrolysis reaction involved with homocysteine, and linked it to the adenosylhomocysteinase protein (Figure 10).

**adenosylhomocysteinase**

H2O + S-adenosyl-L-homocysteine <===========> adenosine + L-**homocysteine**

Figure 10. Hydrolysis reaction of homocysteine

### *2.4.2.5 Comparison of signature genes from each omics level*

4 separate lists of signature genes related to COPD were identified through the re-analysis of transcriptomics data and literature mining on genomics, proteomics, and metabolomics studies related to COPD (Figure 11). However, there was only one gene identified by more than one omics levels. TMEM26 gene was identified as signature genes based on genomics and transcriptomics data. This minimum overlap between results from different omics levels was also observed in another study of genomics and transcriptomics integration in COPD (Morrow et al., 2017).

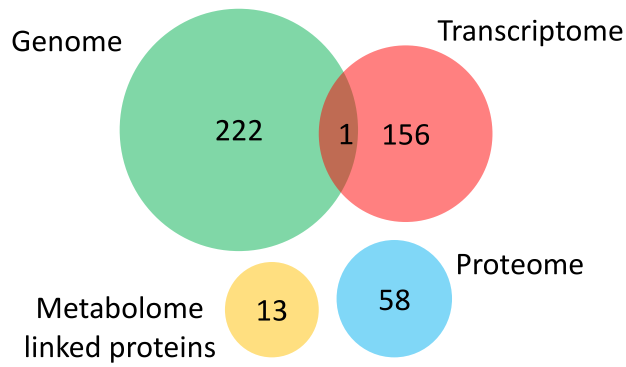



Figure 11. Comparison of signature genes from each omics level

Each circle indicates the number of genes identified from each omics level

There are potentially several reasons behind this limited overlap. Firstly, GWAS relies on very stringent p-value cutoff (usually $5 \times 10^{-8}$) to find the genetic variants associated with the disease and avoid false positives (Pe'er, Yelensky, Altshuler, & Daly, 2008; The International HapMap Consortium, 2005). This framework helped in the identification of the genes that contribute significantly to the disease, like the tumor-suppressing gene TP53 in prostate cancer and glioma risk (Enciso-Mora et al., 2013; Stacey et al., 2011), or genes in Mendelian genetic disorders (Fadista, Manning, Florez, & Groop, 2016). However, for a complex disease like COPD, the genetic disposition to the development of the disease might lie in multiple genes. In addition, different combinations of genes might lead to the same or similar COPD symptoms. Furthermore, there are also environmental contributions to diseases and the interplay between environment and genes. Therefore, the contribution of individual genes might not produce enough significant signals for the GWAS studies to pick up by its stringent p-value cutoff.

Secondly, the transcriptomic data has its own limitations too. Gene expression is a very dynamic and highly regulated process. Our current technology can only capture one transcriptomic snapshot at one time point and one location (cell/tissue) for each experiment. So it might not be a good overall representation of the disease physiological status. There will be noise coming from the intrinsic changes in mRNA levels during the circadian rhythm, other co-occurrent diseases in people, the intake of the drugs, respond to different environmental stimuli and etc. Therefore, additional supporting evidence from the genomics level may help identify more constant regulators of the disease. Genes affected by genetic variants associated with COPD can be viewed as genes that are constantly dysregulated compared to the wildtype gene in healthy controls. And if this constant dysregulation is found more frequently in the disease population, they might be a potential contributing factor to the disease.

Thirdly, the coverage of protein detection in the published proteomics studies is relatively low. Only hundreds of proteins were detected in the current COPD proteomics studies, while a genomic or transcriptomic study may screen more than 20,000 genes. Although more advanced proteomics platforms are catching up with the protein coverage, results from such state-of-art platforms are not currently available for COPD yet (M. Wang et al., 2018). We managed to compensate for the proteomics coverage by identifying additional proteins linked to COPD-associated metabolites. However, the increase in the number of proteins is still limited as metabolomics also suffers from the low detection coverage, as well as a more difficult challenge in assigning metabolite identity or structure to the metabolomics feature (Vuckovic, 2018).

Ideally, COPD signatures genes may be prioritized over genes supported by data from multiple omics levels. The ranking of the genes may be further refined with a quantitative scoring system based on the raw p-values from GWAS, fold-change from transcriptomic analysis, protein-protein interaction confidence score, and metabolite-gene association MAGI score. However, applying such a quantitative model to the current case will only help with the prioritization within individual omics levels. Therefore, a model that integrates the results from the four omics levels is necessary to facilitate the target gene prioritization.

##### *2.4.2.6 Identification of genes close to COPD signature genes from all omics levels*

The selection of candidate genes was expanded to genes interacting with the signature genes from each omics level. 11,759,455 protein-protein interactions of physical and functional associations were retrieved from the STRING database for 19,344 genes in the genome. 841,069 (7.15%) of the association were with a score greater than 0.7, which indicated a high confidence level of interactions. Based on the high confidence level interactions, three distance matrices were created to represent the shortest distance of candidate genes to individual omics level:

- 19,344 candidate gene * 217 COPD signature genes from genomic levels
- 19,344 candidate gene * 157 COPD signature genes from transcriptomic levels
- 19,344 candidate gene * 58+13 COPD signature genes from proteomics and proteins linked to metabolomics levels

The distances of gene pairs in the three matrices ranged from 0 to 7. The median of the distance in all three matrices is 2. The heatmap visualizations of distances matrices between signatures genes from all omics levels can be viewed in Figure 12, Figure 13, and Figure 14. As we can see in the figures, some of the signature genes do not have interacting neighbors. This may be due to a lack of high confidence level interactions. Many signature genes are linked to other signature genes with a short distance.

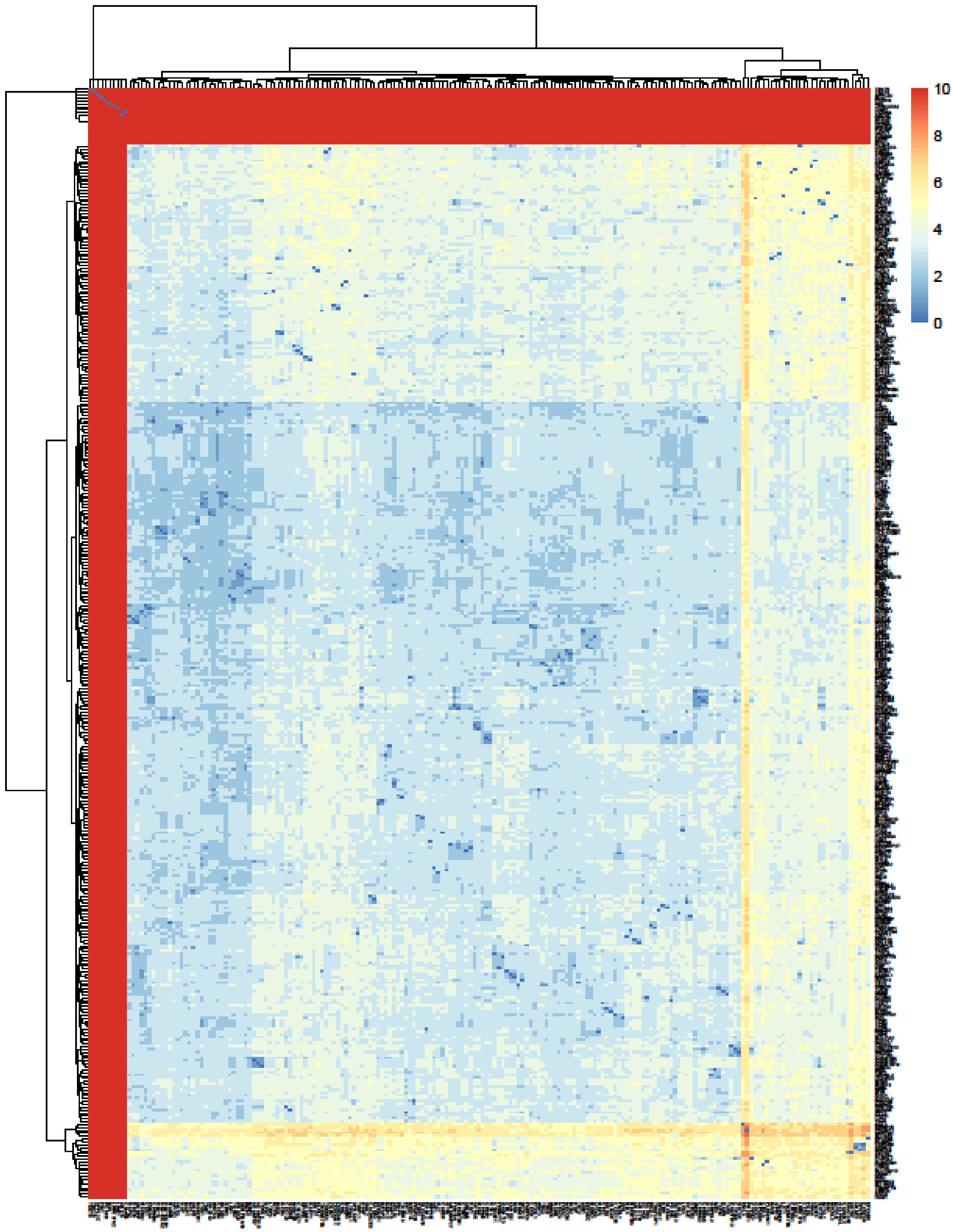



Figure 12. Heatmap visualizations of distances matrices between 217 genomics signatures genes and signature genes from all omics levels

Columns are 217 genomic signature genes; rows are signature genes from all omics levels. Colors indicate the distance between gene pairs. Blue indicates a smaller distance, and red indicates greater distance. Genes with no interacting pairs are assigned a distance value of 10.

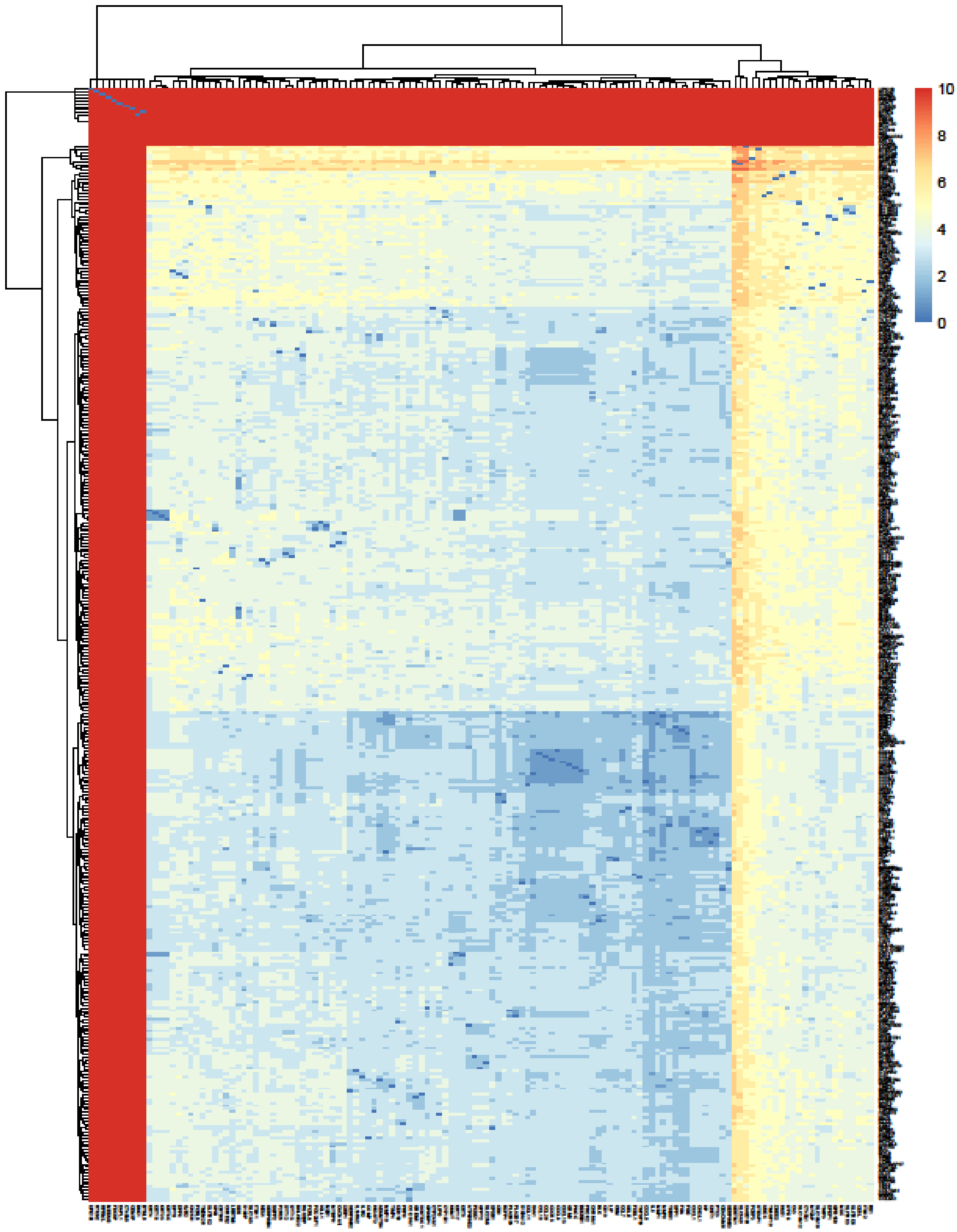



Figure 13. Heatmap visualizations of distances matrices between 157 transcriptomics signatures genes and signature genes from all omics levels

Columns are 157 transcriptomics signature genes; rows are signature genes from all omics levels. Blue indicates a smaller distance, and red indicates a greater distance between genes. Genes with no interacting pairs are assigned a distance value of 10.

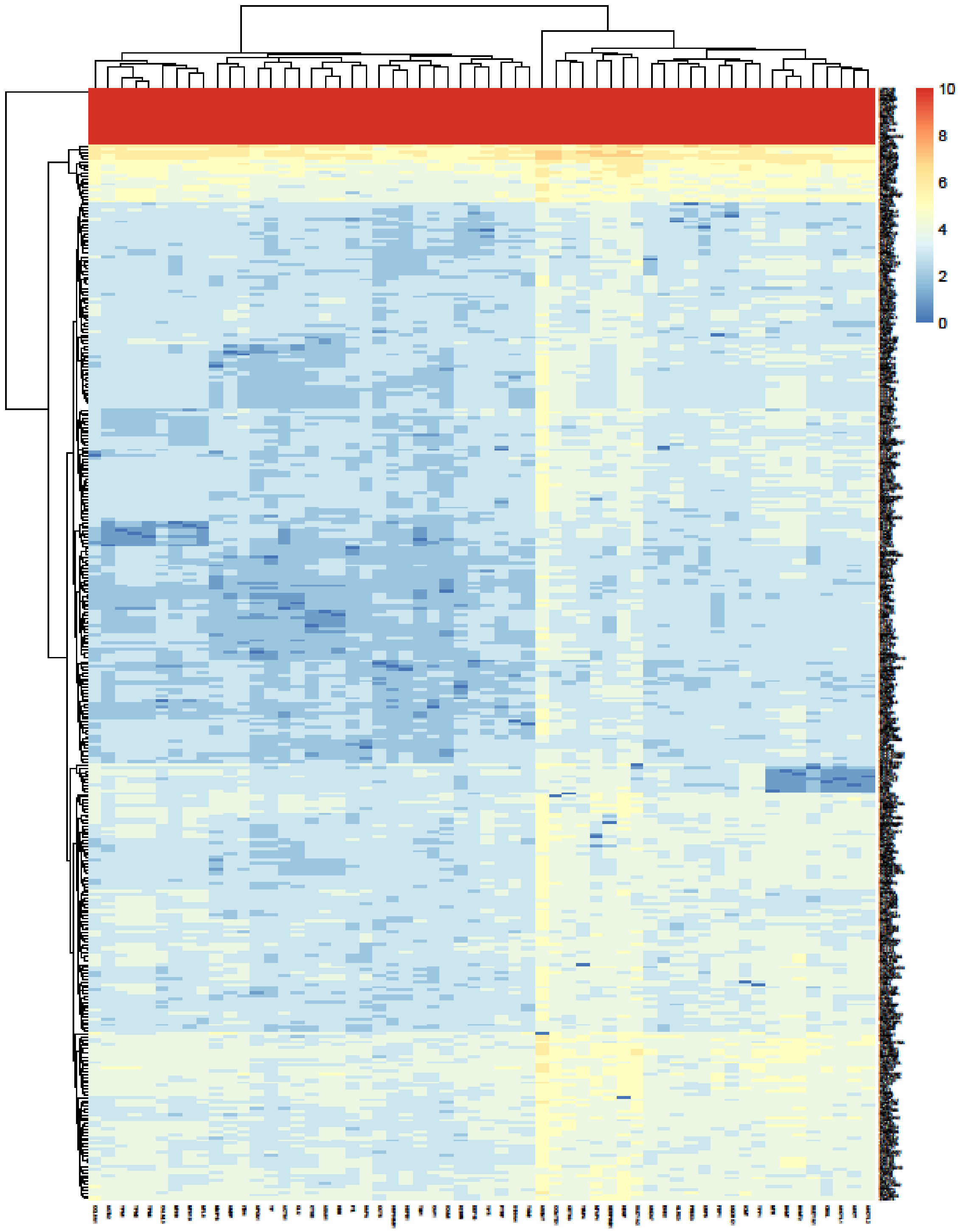



Figure 14. Heatmap visualizations of distances matrices between 71 protein signatures and signature genes from all omics levels

Columns are 71 protein signatures genes from proteomics and metabolomics level; rows are signature genes from all omics levels. Blue indicates a smaller distance, and red indicates a greater distance between genes. Genes with no interacting pairs are assigned a distance value of 10.

For each gene in the genome, the sum of its shortest distance to three sets of signature genes was calculated to represent the overall proximity of a candidate gene to the signature genes identified from different omics levels. The sum distance ranged from 2 to 21 (Figure 15). 50 of the candidate genes were with a sum distance of 2, all of which were themselves identified on one omics level and were direct neighbor to signature genes on other omics levels (Table 12). Therefore, these 50 genes will be our prioritized candidate genes. More candidate COPD genes will be proposed if the distance sum criteria is changed to 3. At the same time, it will require further prioritization to narrow down the number of potential gene targets.

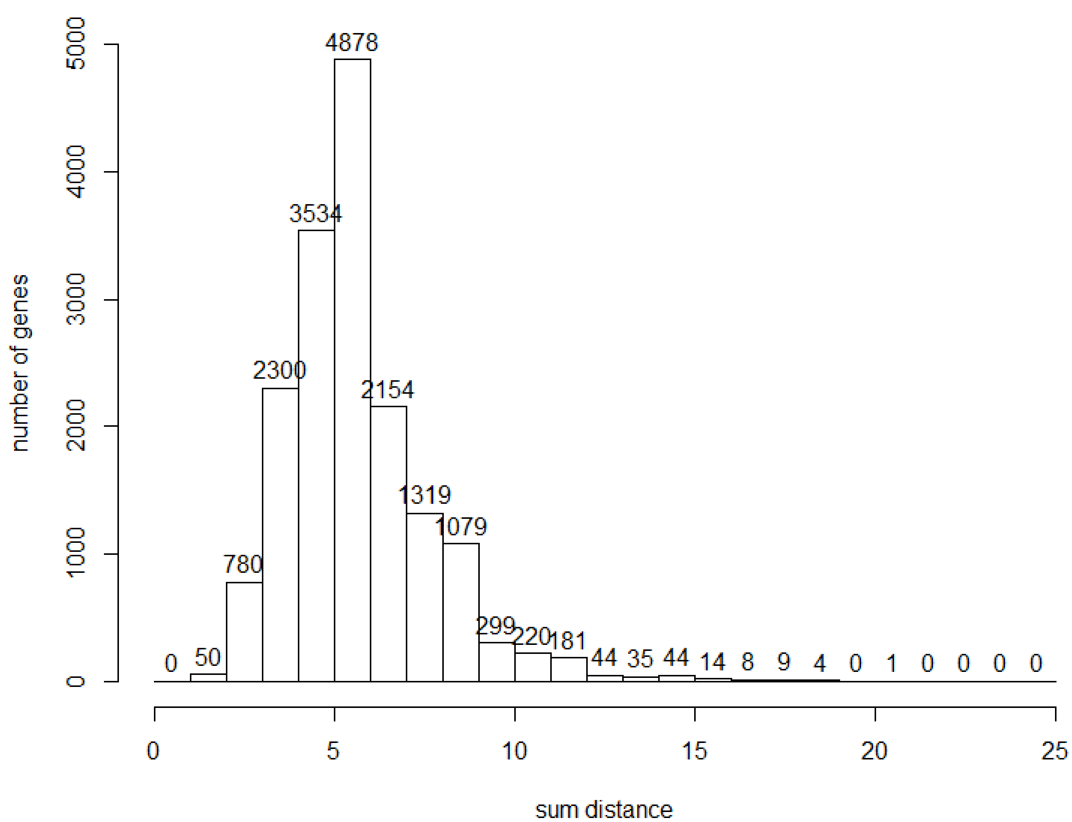



Figure 15. Distribution of sum distances for the candidate genes

We compared our distance model to a heat diffusion network model and found there was minimal overlap between the results. The key assumptions behind the two models were different. The heat diffusion network model normalizes the impact of a gene by the number of neighboring genes. Thus, the top results in this model are mostly genes with only one neighbor. Since the current interactome is still far from complete, the results from heat diffusion may be biased toward genes with less known interactions. On the other hand, the prioritized genes in our model were mostly high degree genes, i.e. genes connected with many neighbors. Although the high degree may increase the chance of linking a candidate gene to omics signatures genes, the links identified are supported by a high confidence level to suggest a real close association of the candidate gene to the omics signature genes and we can rely on this known interaction to prioritize the target.

Table 12. 50 prioritized candidate genes whose distance sum equal 2

| Gene | Distance to Protein signature | Distance to transcriptomics signature gene | Distance to genomics signature gene | Distance sum |
|---|---|---|---|---|
| AGER | 1 | 1 | 0 | 2 |
| APOM | 1 | 1 | 0 | 2 |
| BCAR1 | 1 | 1 | 0 | 2 |
| BMP2 | 1 | 1 | 0 | 2 |
| CXCR4 | 1 | 1 | 0 | 2 |
| GRB2 | 1 | 1 | 0 | 2 |
| ITGA1 | 1 | 1 | 0 | 2 |
| LRP1 | 1 | 1 | 0 | 2 |
| MAN2B1 | 1 | 1 | 0 | 2 |
| MCL1 | 1 | 1 | 0 | 2 |
| MMP3 | 1 | 1 | 0 | 2 |
| PPAP2B | 1 | 1 | 0 | 2 |
| PRSS23 | 1 | 1 | 0 | 2 |

| | | | | |
|---|---|---|---|---|
| SERPINA10 | 1 | 1 | 0 | 2 |
| SNRPF | 1 | 1 | 0 | 2 |
| TGFB2 | 1 | 1 | 0 | 2 |
| WNT5A | 1 | 1 | 0 | 2 |
| ADAMTS4 | 1 | 0 | 1 | 2 |
| COL10A1 | 1 | 0 | 1 | 2 |
| CXCL1 | 1 | 0 | 1 | 2 |
| CXCL8 | 1 | 0 | 1 | 2 |
| FGA | 1 | 0 | 1 | 2 |
| FGG | 1 | 0 | 1 | 2 |
| HP | 1 | 0 | 1 | 2 |
| IL6 | 1 | 0 | 1 | 2 |
| MMP1 | 1 | 0 | 1 | 2 |
| MMP10 | 1 | 0 | 1 | 2 |
| MMP9 | 1 | 0 | 1 | 2 |
| PLA2G7 | 1 | 0 | 1 | 2 |
| SAA1 | 1 | 0 | 1 | 2 |
| SERPINB3 | 1 | 0 | 1 | 2 |
| SGPP2 | 1 | 0 | 1 | 2 |
| SPP1 | 1 | 0 | 1 | 2 |
| THBS1 | 1 | 0 | 1 | 2 |
| TIMP4 | 1 | 0 | 1 | 2 |
| ACTN1 | 0 | 1 | 1 | 2 |
| AMBP | 0 | 1 | 1 | 2 |
| APOA1 | 0 | 1 | 1 | 2 |
| ASAH1 | 0 | 1 | 1 | 2 |
| CLU | 0 | 1 | 1 | 2 |
| COL6A1 | 0 | 1 | 1 | 2 |
| CTSD | 0 | 1 | 1 | 2 |
| FTL | 0 | 1 | 1 | 2 |
| H2BFS | 0 | 1 | 1 | 2 |
| HBB | 0 | 1 | 1 | 2 |
| ICAM1 | 0 | 1 | 1 | 2 |
| ITIH4 | 0 | 1 | 1 | 2 |
| MMP13 | 0 | 1 | 1 | 2 |
| TF | 0 | 1 | 1 | 2 |
| TLN1 | 0 | 1 | 1 | 2 |

### 2.4.3 Conclusion

Additional sets of COPD signatures genes based on genomics, proteomics, and metabolomics data were identified through literature mining. With the integration of human interactome data from the STRING database, a distance-based network algorithm was created to rank the candidate genes close to signature genes identified at all omics levels. 50 genes were prioritized as the final set of COPD candidate genes. They were not only identified as COPD signature genes at one omics level, and also were the direct interacting proteins with the signature genes identified at all other omics levels.

## 2.5 Aim 3: Identify drug repurposing targets within COPD signature genes

### 2.5.1 Material and method

#### *2.5.1.1 Druggability evaluation*

The readiness of a candidate gene being able to targeted by drugs, or the druggability of a candidate gene, was evaluated according to the druggable genome study (Finan et al., 2017). Following the definition of the druggable genome study, the druggability was further categorized into three tires based on their approval status in drug clinical trials or their stages in the drug development processes.

#### *2.5.1.2 The retrieval of drug-gene target information*

The drugs targeting the druggable genes were found on the DrugBank database (version 5.1.4, released 2019-07-02). This database is a well-organized resource that holds both comprehensive drug information and drug-target data (Wishart et al., 2018). Specifically, drug-gene pairs were obtained by parsing the xml database files with the XML package in R.

#### *2.5.1.3 The automated literature search for COPD studies with repurposing drugs*

The literature search was carried out to find supportive information on drug repurposing opportunities. The PubMed search was performed on batch with an R script and NCBI API. For each drug targeting the candidate genes, the number of literature found in PubMed search by keyword “COPD + drug” was obtained. This number was used for the preliminarily screening of drug relevance to COPD. A manual literature search was carried out for the follow-up with the drugs that have been involved in COPD studies.

#### *2.5.1.4 Connectivity map analysis*

The COPD transcriptomic profiles obtained from the previous differential gene expression analysis were analyzed on the Connectivity Map (CMap) database (Subramanian et al., 2017). This database offers tools to evaluate the gene-disease connectivity based on transcriptomic profiles of several human cell lines that underwent knockout or overexpression genetic modifications. First, COPD differentially expressed genes identified in aim 1 were organized as up-regulated and down-regulated gene lists. Next, the two lists were uploaded to CMap query website (https://clue.io/query) as COPD disease signature. The L1000 gene expression Touchstone dataset, which provides the transcriptomic profiles of genetically modified cell lines, was chosen as the reference in the analysis. From the initial analysis result on disease connectivity, we filtered out the results from the genetically modified cell lines undergone knockout or overexpression on the 50 candidate genes we identified in aim 2. Specifically, we examined the genes with the connectivity scores greater than 90% or smaller than -90% in the lung cancer cell line A549 and HCC515. The other seven cell lines in the reference Touchstone dataset were not included in the analysis as the cell lines are not derived from the lung tissue.

#### *2.5.1.5 R for data analysis*

All the data analysis, including data cleaning, pre-processing, integration, statistical test, and visualization, was implemented with R version 3.5.1 (2018-07-02). The coding was saved as R scripts for reproducibility. The detailed R version information and the running environment are listed in APPENDIX F: R VERSION AND RUNNING ENVIRONMENT.

### 2.5.2 Result and discussion

#### *2.5.2.1 Drug repurposing targets within the candidate genes*

The druggable genome study suggested 4,479 protein-coding genes to be druggable (Finan et al., 2017). The druggable genes are mainly the targets of existing drugs, or genes with similar protein structure or sequence to existing drugs, or genes in well-studied drug target families like kinases. Their druggability is further stratified into three tiers, as defined by the druggable genome study (Table 13).

Table 13. Druggable gene targets from the druggable genome study

| Tier | Gene count | Description |
|---|---|---|
| Tier 1 | 1,427 | - targets of small molecule and biologics drugs<br>- genes targeted by clinical trial drug candidates |
| Tier 2 | 682 | - genes similar to tier 1 genes in terms of<br>    o drug binding domains<br>    o protein peptide sequence similarity |
| Tier 3 | 2,370 | - extracellular proteins<br>- genes with slightly lower similarity to tier 1 genes than the tier 2 genes<br>- genes in well-studied drug target families<br>    o G protein-coupled receptors<br>    o nuclear receptors<br>    o ion channels<br>    o kinases<br>    o phosphodiesterases |

Out of the 50 prioritized candidate genes we identified in Aim 2, 40 genes were considered as “druggable” targets based on the definition from the druggable genome (Figure 16). Interestingly, the 40 druggable genes accounted for 80% of the 50 candidate genes, in contrast to the findings that only 22% of the genes in the human genome has been considered as “druggable” (Finan et al., 2017). The greater ratio suggested that our omics integration pipeline identified more candidate genes that have been considered as or similar to the known drug targets. Since COPD is a complex disease and involves multiple pathways and genes, it is not surprising to find COPD related genes targeted by drugs of other diseases. The remaining 10 candidate genes were not considered as druggable based on the current pool of known drugs. More resources may be needed to develop novel therapeutics targeting these genes.

Among these 40 “druggable” genes, 14 genes showed a tier 1 druggability, indicating that they were targets of approved drugs or drug candidates in clinical trials (Figure 16). Therefore, they might be considered as prioritizing targets to look for potential COPD treatment.

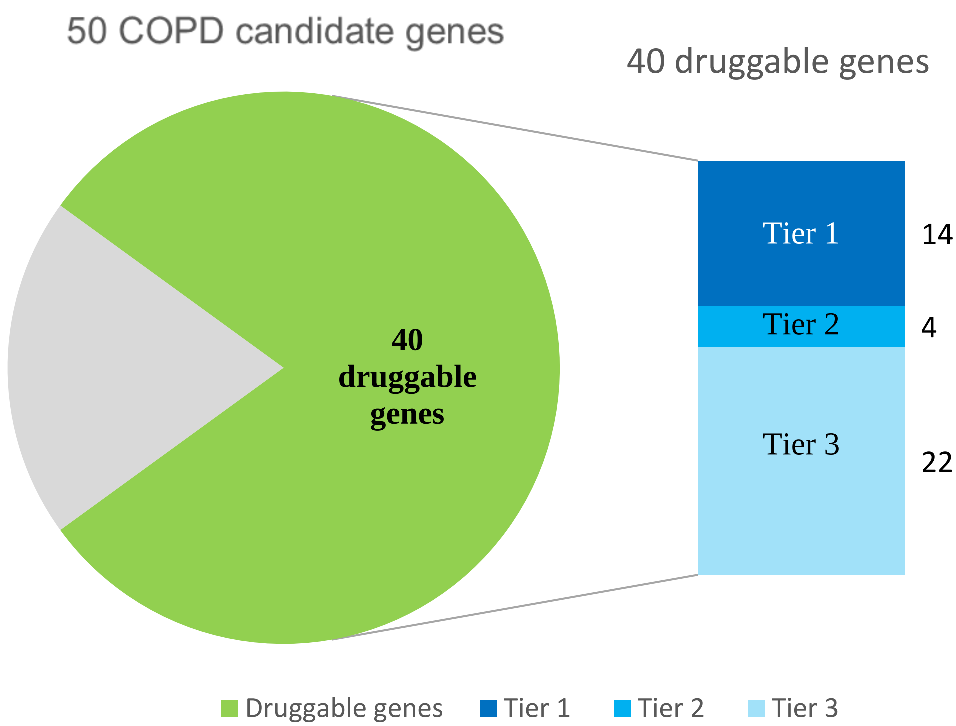



Figure 16. Composition of druggable targets in COPD

The circle represents the 50 COPD candidate genes; The bar plot represents the different tiers of the druggable genes.

*2.5.2.2 Drugs targeting the tier 1 candidate genes*

By extracting drug-gene target pair information from the DrugBank database, we found 55 drugs or biomedical entities targeting the 14 tier 1 druggable genes. Some of these drugs like Natalizumab were indicated for immune systems disease, which was similar to COPD in terms of the dysregulated immune responses. Other drugs like Dilmapimod or Captopril were used for the treatment of nervous systems disease or cardiovascular disease. Both diseases belong to comorbidities in COPD (Cavaillès et al., 2013; Spitzer et al., 2011).

We searched for COPD studies involved with the 55 drugs targeting our 14 druggable genes, to see whether the findings of our pipeline is supported by independent studies. Table 14 shows some example of drugs which target the tier 1 candidate genes. Drug-gene target information is not available for 3 of the tier 1 genes on the Drugbank database, thus they were not included. The number of COPD studies involved with the example drug was listed in Table 14. A table with all the 55 drugs targeting the 14 tier 1 druggable genes can be found in APPENDIX G. Hyaluronic acid is used widely for skincare purposes (Papakonstantinou, Roth, & Karakiulakis, 2012). However, a recent study also found that hyaluronic acid showed a protective role in the emphysema animal model (Turino, Ma, Lin, & Cantor, 2018). Another example is the Plerixafor, which is originally indicated for non-Hodgkin's lymphoma or multiple myeloma. Recently, Plerixafor completed phase 1 trials as a stem cell therapy for COPD treatment (Balkissoon, 2018). These two examples support the concept that a drug repurposing strategy may offer new treatment opportunities for a complex disease like COPD.

Table 14. Evaluation of drug relevance to COPD

| gene | druggability tier | Example Drug | No. of COPD Papers on Example Drug |
|---|---|---|---|
| **CTSD** | 1 | Insulin pork | 0 |
| **CXCR4** | 1 | Plerixafor | 4 |
| **FGA** | 1 | Prothrombin | 39 |
| **FGG** | 1 | Thrombin | 52 |
| **ICAM1** | 1 | Hyaluronic acid | 50 |
| **IL6** | 1 | Dilmapimod | 2 |
| **MMP1** | 1 | Marimastat | 3 |
| **MMP13** | 1 | Endostatin | 1 |
| **MMP3** | 1 | Marimastat | 3 |
| **MMP9** | 1 | Glutathione | 465 |
| **PLA2G7** | 1 | Rilapladib | 0 |

It needs to be noted that the number of studies was only intended for initial screening and is not necessarily supportive of the drug repurposing related to COPD. For example, many COPD research investigated the glutathione to investigate the oxidative stress in COPD, not the potential therapeutical effect.

### *2.5.2.3 CXCR4 as drug repurposing target for COPD*

CXCR4 is one of the prioritized potential drug targets identified by our pipeline for COPD treatment. This gene was not only reported to be genetically associated with COPD (Wain et al., 2015), but also was found to be directly linked to COPD signature genes identified on transcriptomics and proteomics levels (Figure 17).

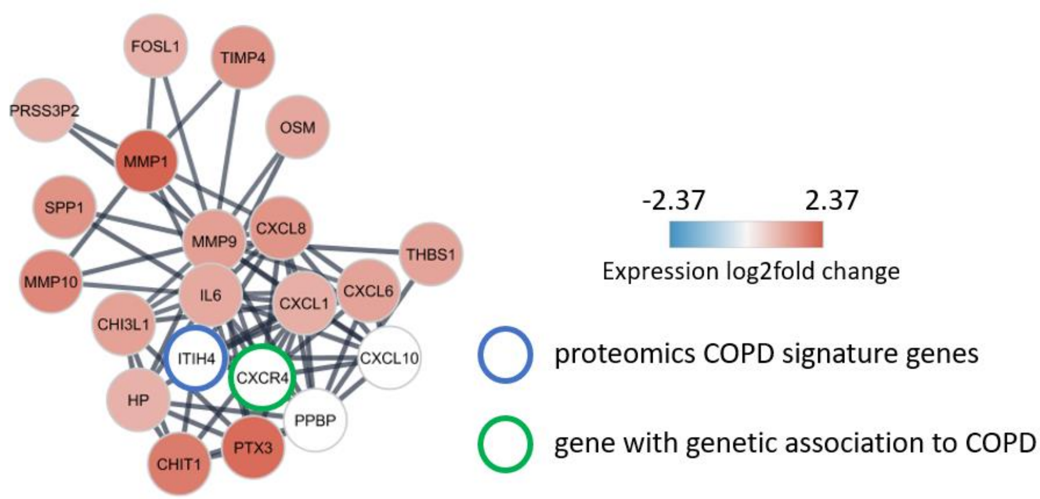



Figure 17. CXCR4 is close to signature genes identified at all omics levels

Circles filled with red color indicate up-regulated genes on transcriptomics levels. Circles filled with white color indicate genes not differentially expressed on transcriptomics levels. Blue and green outlined circles indicate COPD signature identified from proteomics and genomics levels, respectively.

This gene encodes a C-X-C motif chemokine receptor 4, which is known to interact with stromal cell-derived factor-1 (SDF-1/CXCR4) (Caruz et al., 1998; Moriuchi, Moriuchi, Turner, & Fauci, 1997). They participate in mobilizing bone marrow-derived stem cells (Hopman & DiPersio, 2014). Therefore, it is mostly expressed in bone marrow, and as well as other tissues, including liver, lung, kidney, gastrointestinal tract, and endocrine tissues (Figure 18).

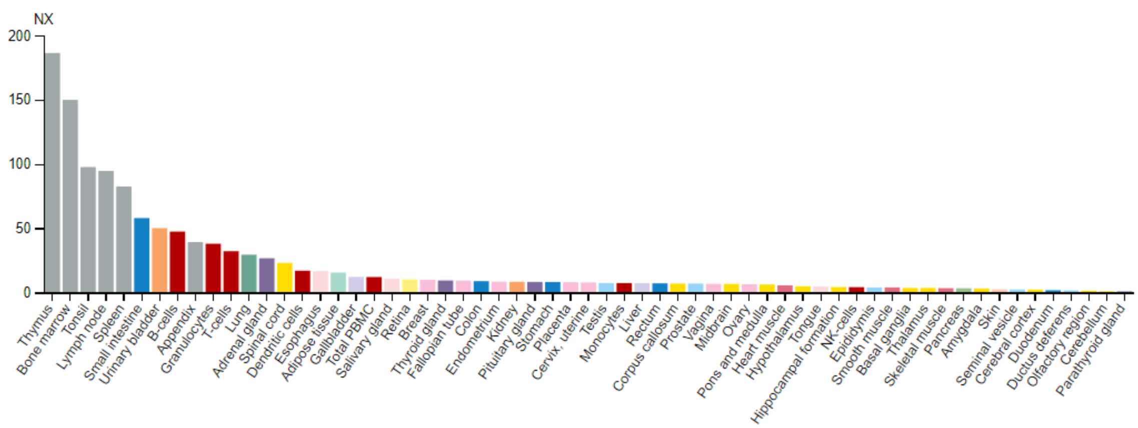



Figure 18. Gene expression levels of CXCR4 across different human tissues

Expression levels were obtained by search on the Human Protein Atlas database (Uhlen et al., 2015). Y axis indicates the gene expression level.

There are several CXCR4-targeted therapeutic agents approved by the FDA or currently in clinical trials (

Table 15). Two of the drugs that target CXCR4, Ibalizumab, and Mogamulizumab, are monoclonal antibodies for HIV treatment and mycosis fungoides and Sézary disease, respectively (Beccari et al., 2019; Duvic et al., 2015). Another drug targeting CXCR4 is Framycetin, which is a topical antibiotic for infection treatment or prevention. Framycetin can also be used to treat hepatic coma when administered orally. Among all these drugs targeting CXCR4, the most promising drug to our COPD target identification pipeline is Plerixafor. Plerixafor is a bone marrow stimulant drug that treats non-

Hodgkin's lymphoma or multiple myeloma. Of importance, Plerixafor has recently completed phase 1 trials for COPD treatment (Balkissoon, 2018). The research team proposed the clinical trial based on in vitro experiment observation and expert knowledge, which is different from our systematic approach. Our pipeline independently identified CXCR4 as a potential COPD drug target, suggesting that our pipeline is finding promising drug repurposing hypothesis.

Table 15. Drugs targeting CXCR4 and their indications

| Drugs Targeting CXCR4 | Indications for the drug |
|---|---|
| Framycetin | Antibiotics. |
| Plerixafor | Bone marrow stimulant: It treats non-Hodgkin's lymphoma or multiple myeloma. |
| Ibalizumab | A monoclonal antibody for HIV treatment. |
| Mogamulizumab | A monoclonal antibody for the treatment of mycosis fungoides and Sézary disease. |

#### *2.5.2.4 In silico gene knockout or overexpression evaluation*

Conventionally, in vitro or in vivo experiments are performed to validate the involvement of a candidate gene in the disease. Candidate genes are knocked-out or over-expressed in cell lines or animal models to observe whether the genetic modifications will lead to phenotypic changes associated with the disease. Here, we carried out an in silico evaluation of the potential effect of the genetic modifications of the 50 candidate genes with the Connectivity Map database. If the knockout or overexpression of the candidate genes induce a similar or reversed transcriptomic profile in COPD, the involvement of the gene in dysregulating gene expression profiles in COPD may be validated.

The most recent version of the database (July 10, 2020) provides transcriptomic profiles of human cell lines that underwent knockout or overexpression experiments for 3,848 unique genes. This covers 27 of the 50 COPD candidate genes we identified in aim 2.

The differentially expressed genes in GOLD IV COPD patients were uploaded to Connectivity Map as the molecular profiling associated with COPD. Through connectivity map analysis, knockout or overexpression of 6 COPD candidate genes were found to result in transcriptomic profiles positively or negatively connected with COPD (Table 16). The connections were identified with a connectivity score of greater than 90% or smaller than -90%. The scores were generated based on the comparison to the transcriptomics patterns resulted from all the other genetic modifications and compound treatments to the cell lines conducted by the Connectivity Map project. If a specific cell line treatment reports a connectivity score of 90% for a set of disease signature genes, it means this treatment results in a transcriptomic pattern that is more similar to the disease compared with the remaining 90% of the treatments. Reversely, a score of -90% indicates a more reversed transcriptomic profile compared with the other 90% treatments to the cell line.

Table 16. COPD signature genes with high connectivity scores

| gene symbol | gene name | experiment | cell line | connectivity score |
|---|---|---|---|---|
| APOA1 | Apolipoprotein A-I | over-expression | A549 | 96.29% |
| SPP1 | Secreted phosphoprotein 1 | over-expression | A549 | 99.89% |
| PRSS23 | Serine Protease 23 | over-expression | A549 | 97.44% |
| GRB2 | Growth Factor Receptor Bound Protein 2 | over-expression | HCC515 | -99.54% |
| ASAH1 | N-Acylsphingosine Amidohydrolase 1 | knockdown | HCC515 | -96.38% |
| CTSD | Cathepsin D | knockdown | HCC515 | 98.6% |

A549 is a cell line of human adenocarcinoma alveolar basal epithelial cells
HCC515 is a cell line of human non-small cell lung adenocarcinoma

The knockout of ASAH1 or GRB2 genes was found to be negatively connected with the COPD transcriptomics profile. It suggested that the inhibition of ASAH1 or GRB2 may result in a reversed transcriptomics pattern compared with COPD. Therefore, we can speculate that if the inhibitors or antagonists targeting the ASAH1 or GRB2 gene present the same knockout effect, they may similarly lead to a reversed transcriptomic change pattern compared with COPD. The reversed transcriptomic pattern may be a potentially desirable outcome that may reverse the underlying molecular changes in COPD.

On the other hand, the overexpression of SPP1, APOA1, PRSS23 genes, and the knockout of CTSD gene in the lung carcinoma cell lines were found to be positively connected with COPD transcriptomics. In other words, the over-expression or knockout experiment led to transcriptomic patterns that were very similar to COPD. This helped validate the involvement of SPP1, APOA1, PRSS23, and CTSD in COPD. Moreover, it suggested that these genes are more likely to be causal genes or upstream genes that may drive the cell transcriptomics to a status similar to COPD.

It needs to be noted that the cell lines selected for connectivity map evaluation here are two lung adenocarcinoma cell lines. Currently, connectivity map data are generated from eight cancerous cell lines and one non-cancer cell line from the immortalized human kidney epithelial cell. Although we have selected the cell lines that are more representative of the lung tissue, they belong to cancer cell lines and may behave differently compared to the non-cancer lung cell lines.

In summary, six COPD candidate genes were found to be connected with the COPD transcriptomic profile through in silico screening of the genetically modified cell lines in the Connectivity Map database. These genes are likely the upstream genes contributing to the dysregulated gene expression profiles in COPD patients. Future in vitro experiments targeting these genes in the non-cancer cell lines can help with the further evaluation of their role in COPD.

### 2.5.3 Conclusion

Among the 50 candidate genes, 40 were considered as druggable targets. Literature search and in silico screening further prioritized several genes with high potential for drug repurposing. One of the prioritized drug targets, CXCR4, is targeted by the drug Plerixafor, which is currently being evaluated for COPD treatment in clinical trials. This suggests that our pipeline is finding promising drug repurposing targets for COPD.

# CHAPTER 3

# CONCLUSION AND FUTURE DIRECTION

We identified 157 COPD signature genes through an initial lung transcriptomics analysis from COPD patients and healthy subjects. Additional sets of COPD signatures genes based on genomics, proteomics, and metabolomics data were identified through literature mining. With the integration of human interactome data from the STRING database, a distance-based network algorithm was created to rank the candidate genes close to signature genes previously identified from all omics levels. This pipeline helped prioritize 50 genes as the final set of COPD candidate genes. They not only belong to the previously identified COPD signature genes, and also are the directly interacting proteins with the signature genes identified from all other omics levels. Among these candidate genes, 40 genes may be considered as “druggable” targets and 14 genes are targeted by approved or clinical phase drugs. One of the prioritized drug targets, CXCR4, is targeted by a drug that is currently being investigated for repurposing in COPD treatment, suggesting that our pipeline is finding promising drug repurposing targets for COPD.

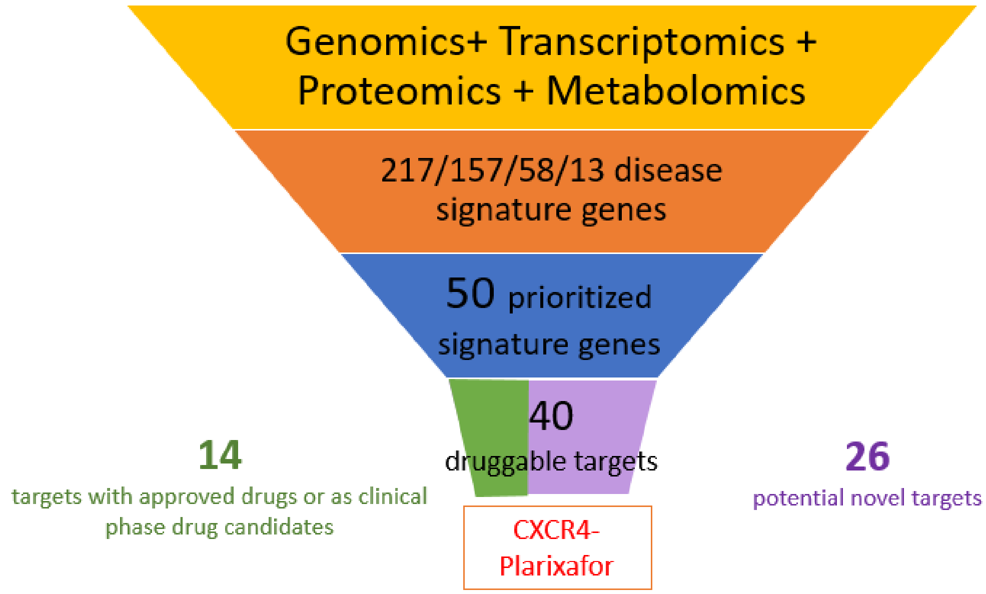



Figure 19. Integration with drug-target information and genomic association data helped identified potential drug repurposing targets

This is the first project that uses a systematic approach integrating multiple types of omics data to find drug repurposing targets for COPD. The pipeline being developed is supported by rich information from omics data, not limited by our prior knowledge, balanced on disease relevance and development feasibility of targets, and can be readily applied to other diseases.

Future directions will include: 1) expanding the pipeline to the earlier stage of COPD, 2) delineating patient heterogeneity by stratification on molecular subtypes, 3) capturing network relationships across omics levels with advanced computation models, and 4) in vitro and in in vivo validation of the prioritized candidate genes with lung cell lines and lung tissues.

# REFERENCE


Abonyo, B. O., Wang, P., Narasaraju, T. A., Rowan, W. H., McMillan, D. H., Zimmerman, U. J., & Liu, L. (2003). Characterization of α-soluble N-ethylmaleimide-sensitive fusion attachment protein in alveolar type II cells: Implications in lung surfactant secretion. *American Journal of Respiratory Cell and Molecular Biology*, *29*(3 I), 273–282. https://doi.org/10.1165/rcmb.2002-0189OC

Adamko, D. J., Nair, P., Mayers, I., Tsuyuki, R. T., Regush, S., & Rowe, B. H. (2015). Metabolomic profiling of asthma and chronic obstructive pulmonary disease: A pilot study differentiating diseases. *Journal of Allergy and Clinical Immunology*, *136*(3), 571-580.e3. https://doi.org/10.1016/j.jaci.2015.05.022

Aderem, A. (2005). Systems biology: its practice and challenges. *Cell*, *121*(4), 511–513. https://doi.org/10.1016/j.cell.2005.04.020

Afonso, A. S. M., Verhamme, K. M. C., Sturkenboom, M. C. J. M., & Brusselle, G. G. O. (2011). COPD in the general population: Prevalence, incidence and survival. *Respiratory Medicine*, *105*(12), 1872–1884. https://doi.org/10.1016/j.rmed.2011.06.012

Agudelo, C. W., Kumley, B. K., Area-Gomez, E., Xu, Y., Dabo, A. J., Geraghty, P., … Garcia-Arcos, I. (2020). Decreased surfactant lipids correlate with lung function in chronic obstructive pulmonary disease (COPD). *PLOS ONE*, *15*(2), e0228279. https://doi.org/10.1371/journal.pone.0228279

Anees ur Rehman, Ahmad Hassali, M. A., Muhammad, S. A., Shah, S., Abbas, S., Hyder Ali, I. A. Bin, & Salman, A. (2019). The economic burden of Chronic Obstructive Pulmonary Disease (COPD) in USA, Europe and Asia: results from a systematic review of the literature. *Expert Review of Pharmacoeconomics & Outcomes Research*, 14737167.2020.1678385. https://doi.org/10.1080/14737167.2020.1678385

Annoni, R., Lançąs, T., Tanigawa, R. Y., Matsushita, M. D. M., Fernezlian, S. D. M., Bruno, A., … Mauad, T. (2012). Extracellular matrix composition in COPD. *European Respiratory Journal*, *40*(6), 1362–1373. https://doi.org/10.1183/09031936.00192611

Balkissoon, R. (2018). Stem Cell Therapy for COPD: Where are we? *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *5*(2), 148–153. https://doi.org/10.15326/jcopdf.5.2.2018.0140

Ballman, K. V., Grill, D. E., Oberg, A. L., & Therneau, T. M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, *20*(16), 2778–2786. https://doi.org/10.1093/bioinformatics/bth327

Bandow, J. E., Baker, J. D., Berth, M., Painter, C., Sepulveda, O. J., Clark, K. A., … VanBogelen, R. A. (2008). Improved image analysis workflow for 2-D gels enables large-scale 2-D gel-based proteomics studies - COPD biomarker discovery study. *Proteomics*, *8*(15), 3030–3041. https://doi.org/10.1002/pmic.200701184

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., … Im, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03621-1

Barnes, P. J. (2006, September). Theophylline for COPD. *Thorax*, *61*(9), 742–744.

https://doi.org/10.1136/thx.2006.061002

Barnes, P. J. (2014). Cellular and Molecular Mechanisms of Chronic Obstructive Pulmonary Disease. *Clinics in Chest Medicine*, *35*(1), 71–86. https://doi.org/10.1016/j.ccm.2013.10.004

Barnes, P. J., & Adcock, I. M. (2009). Glucocorticoid resistance in inflammatory diseases. *The Lancet*, *373*(9678), 1905–1917. https://doi.org/10.1016/S0140-6736(09)60326-3

Barrero, C. A., Perez-Leal, O., Aksoy, M., Moncada, C., Ji, R., Lopez, Y., … Merali, S. (2013). Histone 3.3 Participates in a Self-Sustaining Cascade of Apoptosis That Contributes to the Progression of Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, *188*(6), 673–683. https://doi.org/10.1164/rccm.201302-0342OC

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995. https://doi.org/10.1093/nar/gks1193

Baughman, R. P. (2007). Technical aspects of bronchoalveolar lavage: Recommendations for a standard procedure. *Seminars in Respiratory and Critical Care Medicine*, *28*(5), 475–485. https://doi.org/10.1055/s-2007-991520

Beccari, M. V, Mogle, B. T., Sidman, E. F., Mastro, K. A., Asiago-Reddy, E., & Kufel, W. D. (2019). Ibalizumab, a Novel Monoclonal Antibody for the Management of Multidrug-Resistant HIV-1 Infection. *Antimicrobial Agents and Chemotherapy*, *63*(6), e00110-19. https://doi.org/10.1128/AAC.00110-19

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bernstein, B. E., Meissner, A., & Lander, E. S. (2007, February 23). The Mammalian Epigenome. *Cell*. https://doi.org/10.1016/j.cell.2007.01.033

Bhat, T. A., Panzica, L., Kalathil, S. G., & Thanavala, Y. (2015). Immune dysfunction in patients with chronic obstructive pulmonary disease. In *Annals of the American Thoracic Society* (Vol. 12, pp. S169–S175). American Thoracic Society. https://doi.org/10.1513/AnnalsATS.201503-126AW

Bhatt, S. P., Kim, Y., Wells, J. M., Bailey, W. C., Ramsdell, J. W., Foreman, M. G., … Investigators, for the Copdg. (2014). FEV $_1$ /FEV $_6$ to Diagnose Airflow Obstruction. Comparisons with Computed Tomography and Morbidity Indices. *Annals of the American Thoracic Society*, *11*(3), 335–341. https://doi.org/10.1513/AnnalsATS.201308-251OC

Black, L. F., & Kueppers, F. (1978). alpha1-Antitrypsin deficiency in nonsmokers. *The American Review of Respiratory Disease*, *117*(3), 421–428. https://doi.org/10.1164/arrd.1978.117.3.421

Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R., & Landfield, P. W. (2004). Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(7), 2173–2178. https://doi.org/10.1073/pnas.0308512100

Blanc, P. D., Menezes, A. M. B., Plana, E., Mannino, D. M., Hallal, P. C., Toren, K., … Zock, J.-P. (2008). Occupational exposures and COPD: an ecological analysis of international data.

*European Respiratory Journal*, *33*(2), 298–304. https://doi.org/10.1183/09031936.00118808

Bossé, Y., Postma, D. S., Sin, D. D., Lamontagne, M., Couture, C., Gaudreault, N., … Laviolette, M. (2012). Molecular signature of smoking in human lung tissues. *Cancer Research*, *72*(15), 3753–3763. https://doi.org/10.1158/0008-5472.CAN-12-1160

Bozier, J., Rutting, S., Xenaki, D., Peters, M., Adcock, I., & Oliver, B. G. (2019). Heightened response to e-cigarettes in COPD. *ERJ Open Research*, *5*(1), 00192–02018. https://doi.org/10.1183/23120541.00192-2018

Branson, R. D. (2018). Oxygen Therapy in COPD. *Respiratory Care*, *63*(6), 734–748. https://doi.org/10.4187/respcare.06312

Brantigan, O. C., & Mueller, E. (1957). Surgical treatment of pulmonary emphysema. *The American Surgeon*, *23*(9), 789–804. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13470332

Breckenridge, A. M., Breckenridge, R. A., & Peck, C. C. (2019). Report on the current status of the use of real-world data (RWD) and real-world evidence (RWE) in drug development and regulation. *British Journal of Clinical Pharmacology*, *85*(9), 1874–1877. https://doi.org/10.1111/bcp.14026

Bridevaux, P.-O., Probst-Hensch, N. M., Schindler, C., Curjuric, I., Felber Dietrich, D., Braendli, O., … Russi, E. W. (2010). Prevalence of airflow obstruction in smokers and never-smokers in Switzerland. *European Respiratory Journal*, *36*(6), 1259–1269. https://doi.org/10.1183/09031936.00004110

Brusselle, G. G., Joos, G. F., & Bracke, K. R. (2011). New insights into the immunology of chronic obstructive pulmonary disease. *The Lancet*, *378*(9795), 1015–1026. https://doi.org/10.1016/S0140-6736(11)60988-4

Burkart, K. M., Sofer, T., London, S. J., Manichaikul, A., Hartwig, F. P., Yan, Q., … Graham Barr, R. (2018). A genome-wide association study in hispanics/latinos identifies novel signals for lung function the hispanic community health study/study of latinos. *American Journal of Respiratory and Critical Care Medicine*, *198*(2), 208–219. https://doi.org/10.1164/rccm.201707-1493OC

Burney, P. G. J., Patel, J., Newson, R., Minelli, C., & Naghavi, M. (2015). Global and regional trends in COPD mortality, 1990–2010. *European Respiratory Journal*, *45*(5), 1239–1247. https://doi.org/10.1183/09031936.00142414

Calverley, P. M. A. (2013). The ABCD of GOLD made clear. *European Respiratory Journal*, *42*(5), 1163–1165. https://doi.org/10.1183/09031936.00117413

Campbell, J. D., McDonough, J. E., Zeskind, J. E., Hackett, T. L., Pechkovsky, D. V., Brandsma, C. A., … Spira, A. (2012). A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK. *Genome Medicine*, *4*(12), 67. https://doi.org/10.1186/gm367

Campbell, J. D., McDonough, J. E., Zeskind, J. E., Hackett, T. L., Pechkovsky, D. V, Brandsma, C.-A., … Avrum, S. (2012). A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK. *Genome Medicine*, *4*(8), 67. https://doi.org/10.1186/gm367

Carlin, D. E., Demchak, B., Pratt, D., Sage, E., & Ideker, T. (2017). Network propagation in the cytoscape cyberinfrastructure. *PLOS Computational Biology*, *13*(10), e1005598. https://doi.org/10.1371/journal.pcbi.1005598

Carlson, M. (2016). Agilent "Human Genome, Whole" annotation data (chip hgug4112a). R package version 3.2.3.

Caruz, A., Samsom, M., Alonso, J. M., Alcami, J., Baleux, F., Virelizier, J. L., … Arenzana-Seisdedos, F. (1998). Genomic organization and promoter characterization of human CXCR4 gene. *FEBS Letters*, *426*(2), 271–278. https://doi.org/10.1016/s0014-5793(98)00359-7

Castaldi, P. J., Benet, M., Petersen, H., Rafaels, N., Finigan, J., Paoletti, M., … Garcia-Aymerich, J. (2017). Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax*, *72*(11), 998–1006. https://doi.org/10.1136/thoraxjnl-2016-209846

Cavaillès, A., Brinchault-Rabin, G., Dixmier, A., Goupil, F., Gut-Gobert, C., Marchand-Adam, S., … Diot, P. (2013). Comorbidities of COPD. *European Respiratory Review : An Official Journal of the European Respiratory Society*, *22*(130), 454–475. https://doi.org/10.1183/09059180.00008612

Cazzola, M., & Molimard, M. (2010). The scientific rationale for combining long-acting beta2-agonists and muscarinic antagonists in COPD. *Pulmonary Pharmacology & Therapeutics*, *23*(4), 257–267. https://doi.org/10.1016/j.pupt.2010.03.003

Celli, B. R., Halbert, R. J., Nordyke, R. J., & Schau, B. (2005). Airway obstruction in never smokers: Results from the Third National Health and Nutrition Examination Survey. *The American Journal of Medicine*, *118*(12), 1364–1372. https://doi.org/10.1016/j.amjmed.2005.06.041

Celli, B. R., Locantore, N., Tal-Singer, R., Riley, J., Miller, B., Vestbo, J., … Agusti, A. (2018). Emphysema and extrapulmonary tissue loss in COPD: A multi-organ loss of tissue phenotype. *European Respiratory Journal*, *51*(2). https://doi.org/10.1183/13993003.02146-2017

Chapman, K. R., Mannino, D. M., Soriano, J. B., Vermeire, P. A., Buist, A. S., Thun, M. J., … Beasley, R. (2006). Epidemiology and costs of chronic obstructive pulmonary disease. *European Respiratory Journal*, *27*(1), 188–207. https://doi.org/10.1183/09031936.06.00024505

Chen, J. J., Hsueh, H. M., Delongchamp, R. R., Lin, C. J., & Tsai, C. A. (2007). Reproducibility of microarray data: A further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics*, *8*(1), 412. https://doi.org/10.1186/1471-2105-8-412

Cho, M. H., Castaldi, P. J., Wan, E. S., Siedlinski, M., Hersh, C. P., Demeo, D. L., … COPDGene Investigators. (2012). A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Human Molecular Genetics*, *21*(4), 947–957. https://doi.org/10.1093/hmg/ddr524

Cho, Michael H, Boutaoui, N., Klanderman, B. J., Sylvia, J. S., Ziniti, J. P., Hersh, C. P., … Silverman, E. K. (2010). Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature Genetics*, *42*(3), 200–202. https://doi.org/10.1038/ng.535

Clary, D. O., Griff, I. C., & Rothman, J. E. (1990). SNAPs, a family of NSF attachment proteins involved in intracellular membrane fusion in animals and yeast. *Cell*, *61*(4), 709–721.

https://doi.org/10.1016/0092-8674(90)90482-T

Costa, R., Wagner, D. E., De Santis, M. M., Bailey, K. E., Doryab, A., Schorpp, K., … Königshoff, M. (2018). Novel compounds for WNT/ß-catenin induced lung repair in chronic obstructive pulmonary disease (p. PA584). European Respiratory Society (ERS). https://doi.org/10.1183/13993003.congress-2018.pa584

Crosby, L. M., & Waters, C. M. (2010, June). Epithelial repair mechanisms in the lung. *American Journal of Physiology - Lung Cellular and Molecular Physiology*. American Physiological Society Bethesda, MD. https://doi.org/10.1152/ajplung.00361.2009

Cruickshank-Quinn, C. I., Jacobson, S., Hughes, G., Powell, R. L., Petrache, I., Kechris, K., … Reisdorph, N. (2018). Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-35372-w

Daly, J. W., Jacobson, K. A., & Ukena, D. (1987). Adenosine receptors: development of selective agonists and antagonists. *Progress in Clinical and Biological Research*, *230*, 41–63.

Dartevelle, P., Macchiarini, P., & Chapelier, A. (1995). Operative technique of bullectomy. *Chest Surgery Clinics of North America*, *5*(4), 735–749. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8574560

Davis, S., & Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor, *23*(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

De Paola, M., Bello, O. D., & Michaut, M. A. (2015). Cortical granule exocytosis is mediated by alpha-SNAP and N-ethilmaleimide sensitive factor in mouse oocytes. *PLoS ONE*, *10*(8). https://doi.org/10.1371/journal.pone.0135679

Deja, S., Porebska, I., Kowal, A., Zabek, A., Barg, W., Pawelczyk, K., … Mlynarz, P. (2014). Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease. *Journal of Pharmaceutical and Biomedical Analysis*, *100*, 369–380. https://doi.org/10.1016/j.jpba.2014.08.020

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

Dugo, M., Cotroneo, C. E., Lavoie-Charland, E., Incarbone, M., Santambrogio, L., Rosso, L., … Colombo, F. (2016). Human Lung Tissue Transcriptome: Influence of Sex and Age. *PLOS ONE*, *11*(11), e0167460. https://doi.org/10.1371/journal.pone.0167460

Duvic, M., Pinter-Brown, L. C., Foss, F. M., Sokol, L., Jorgensen, J. L., Challagundla, P., … Kim, Y. H. (2015). Phase 1/2 study of mogamulizumab, a defucosylated anti-CCR4 antibody, in previously treated patients with cutaneous T-cell lymphoma. *Blood*, *125*(12), 1883–1889. https://doi.org/10.1182/blood-2014-09-600924

Enciso-Mora, V., Hosking, F. J., Di Stefano, A. L., Zelenika, D., Shete, S., Broderick, P., … Houlston, R. S. (2013). Low penetrance susceptibility to glioma is caused by the TP53 variant rs78378222. *British Journal of Cancer*, *108*(10), 2178–2185. https://doi.org/10.1038/bjc.2013.155

Erbilgin, O., Rübel, O., Louie, K. B., Trinh, M., Raad, M. de, Wildish, T., … Bowen, B. P. (2019). MAGI: A Method for Metabolite Annotation and Gene Integration. *ACS Chemical*

*Biology*, *14*(4), 704–714. https://doi.org/10.1021/acschembio.8b01107

Eriksson, S. (1964). Pulmonary Emphysema and Alpha1-Antitrypsin Deficiency. *Acta Medica Scandinavica*, *175*(2), 197–205. https://doi.org/10.1111/j.0954-6820.1964.tb00567.x

Essayan, D. M. (2001). Cyclic nucleotide phosphodiesterases. *The Journal of Allergy and Clinical Immunology*, *108*(5), 671–680. https://doi.org/10.1067/mai.2001.119555

Ezzie, M. E., Crawford, M., Cho, J. H., Orellana, R., Zhang, S., Gelinas, R., … Nana-Sinkam, S. P. (2012). Gene expression networks in COPD: microRNA and mRNA regulation. *Thorax*, *67*(2), 122–131. https://doi.org/10.1136/thoraxjnl-2011-200089

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, *24*(8), 1202–1205. https://doi.org/10.1038/ejhg.2015.269

Falk, J. A., Minai, O. A., & Mosenifar, Z. (2008, May). Inhaled and systemic corticosteroids in chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society*. https://doi.org/10.1513/pats.200707-096ET

Finan, C., Gaulton, A., Kruger, F. A., Lumbers, R. T., Shah, T., Engmann, J., … Casas, J. P. (2017). The druggable genome and support for target identification and validation in drug development. *Science Translational Medicine*, *9*(383), eaag1166. https://doi.org/10.1126/scitranslmed.aag1166

Furlanetto, K. C., & Pitta, F. (2017). Oxygen therapy devices and portable ventilators for improved physical activity in daily life in patients with chronic respiratory disease. *Expert Review of Medical Devices*, *14*(2), 103–115. https://doi.org/10.1080/17434440.2017.1283981

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., … Lusis, A. J. (2011). Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genetics*, *7*(6). https://doi.org/10.1371/journal.pgen.1001393

Global Initiative for Chronic Obstructive Lung Disease. (2011). *Global Strategy for the Diagnosis, Management and Prevention of COPD, GOLD 2011*. Retrieved from https://goldcopd.org

Global Initiative for Chronic Obstructive Lung Disease. (2017). *Global Strategy for the Diagnosis, Management and Prevention of COPD, GOLD 2017*. Retrieved from https://goldcopd.org

Goldstein, R. A., Rohatgi, P. K., Bergofsky, E. H., Block, E. R., Daniele, R. P., Dantzker, D. R., … Metzger, W. J. (1990). Clinical role of bronchoalveolar lavage in adults with pulmonary disease. *The American Review of Respiratory Disease*, *142*(2), 481–486. https://doi.org/10.1164/ajrccm/142.2.481

Graves, P. R., & Haystead, T. A. J. (2002). Molecular Biologist's Guide to Proteomics. *Microbiology and Molecular Biology Reviews*, *66*(1), 39–63. https://doi.org/10.1128/mmbr.66.1.39-63.2002

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), 245–252. https://doi.org/10.1038/ng.3506

Halper-Stromberg, E., Gillenwater, L., Cruickshank-Quinn, C., O'neal, W. K., Reisdorph, N., Petrache, I., … Bowler, R. P. (2019). Bronchoalveolar lavage fluid from COPD patients reveals more compounds associated with disease than matched plasma. *Metabolites*, *9*(8). https://doi.org/10.3390/metabo9080157

Hardin, M., & Silverman, E. K. (2014). Chronic Obstructive Pulmonary Disease Genetics: A Review of the Past and a Look Into the Future. *Chronic Obstructive Pulmonary Diseases (Miami, Fla.)*, *1*(1), 33–46. https://doi.org/10.15326/jcopdf.1.1.2014.0120

Hautamaki, R. D., Kobayashi, D. K., Senior, R. M., & Shapiro, S. D. (1997). Requirement for macrophage elastase for cigarette smoke-induced emphysema in mice. *Science*, *277*(5334), 2002–2004. https://doi.org/10.1126/science.277.5334.2002

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., & Li, H. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics*, *92*(5), 667–680. https://doi.org/10.1016/j.ajhg.2013.03.022

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, *7*(1), 23. https://doi.org/10.1186/s13321-015-0068-4

Hillas, G., Perlikos, F., & Tzanakis, N. (2016). Acute exacerbation of COPD: is it the "stroke of the lungs"? *International Journal of Chronic Obstructive Pulmonary Disease*, *11*, 1579–1586. https://doi.org/10.2147/COPD.S106160

Hogg, J. C. (2004). Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. *The Lancet*, *364*(9435), 709–721. https://doi.org/10.1016/S0140-6736(04)16900-6

Hollywood, K., Brison, D. R., & Goodacre, R. (2006). Metabolomics: Current technologies and future trends. *PROTEOMICS*, *6*(17), 4716–4723. https://doi.org/10.1002/pmic.200600106

Holm, B. A., Wang, Z., Egan, E. A., & Notter, R. H. (1996). Content of dipalmitoyl phosphatidylcholine in lung surfactant: Ramifications for surface activity. *Pediatric Research*, *39*(5), 805–811. https://doi.org/10.1203/00006450-199605000-00010

Hopman, R. K., & DiPersio, J. F. (2014). Advances in stem cell mobilization. *Blood Reviews*, *28*(1), 31–40. https://doi.org/10.1016/j.blre.2014.01.001

Horgan, R. P., & Kenny, L. C. (2011). ‘Omic’ technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, *13*(3), 189–195. https://doi.org/10.1576/toag.13.3.189.27672

Howell, C., Smith, J. R., & Shute, J. K. (2018). Targeting matrix metalloproteinase-13 in bronchial epithelial repair. *Clinical and Experimental Allergy*, *48*(9), 1214–1221. https://doi.org/10.1111/cea.13215

Hunninghake, G. M., Cho, M. H., Tesfaigzi, Y., Soto-Quiros, M. E., Avila, L., Lasky-Su, J., … Celedón, J. C. (2009). MMP12, lung function, and COPD in high-risk populations. *New England Journal of Medicine*, *361*(27), 2599–2608. https://doi.org/10.1056/NEJMoa0904006

Ilumets, H., Rytilä, P., Demedts, I., Brusselle, G. G., Sovijärvi, A., Myllärniemi, M., … Kinnula, V. L. (2007). Matrix metalloproteinases -8, -9 and -12 in smokers and patients with Stage 0 COPD. *International Journal of COPD*, *2*(3), 369–379.

Issaq, H., & Veenstra, T. (2008). Two-dimensional polyacrylamide gel electrophoresis (2D-

PAGE): advances and perspectives. *BioTechniques*, *44*(5), 697–698, 700. https://doi.org/10.2144/000112823

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., … Murray, C. J. L. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Jin Kim, W., Hyun Lim, J., Seung Lee, J., Lee, S.-D., Han Kim, J., & Oh, Y.-M. (2015). Comprehensive Analysis of Transcriptome Sequencing Data in the Lung Tissues of COPD Subjects. https://doi.org/10.1155/2015/206937

Jing, J., Huang, T., Cui, W., Xu, F., & Shen, H. (2009). Should FEV 1 /FEV 6 Replace FEV 1 /FVC Ratio To Detect Airway Obstruction? *Chest*, *135*(4), 991–998. https://doi.org/10.1378/chest.08-0723

Khakban, A., Sin, D. D., FitzGerald, J. M., McManus, B., Ng, R., Hollander, Z., & Sadatsafavi, M. (2016). The Projected Epidemic of COPD Hospitalizations Over the Next 15 Years: A Population Based Perspective. *American Journal of Respiratory and Critical Care Medicine*, *195*(3), rccm.201606-1162PP. https://doi.org/10.1164/rccm.201606-1162PP

Kilk, K., Aug, A., Ottas, A., Soomets, U., Altraja, S., & Altraja, A. (2018). Phenotyping of chronic obstructive pulmonary disease based on the integration of metabolomes and clinical characteristics. *International Journal of Molecular Sciences*, *19*(3). https://doi.org/10.3390/ijms19030666

Kim, D. S., Kim, Y. S., Jung, K.-S., Chang, J. H., Lim, C.-M., Lee, J. H., … Korean Academy of Tuberculosis and Respiratory Diseases. (2005). Prevalence of Chronic Obstructive Pulmonary Disease in Korea. *American Journal of Respiratory and Critical Care Medicine*, *172*(7), 842–847. https://doi.org/10.1164/rccm.200502-259OC

Kim, W. J., & Lee, S. Do. (2015). Candidate genes for COPD: Current evidence and research. *International Journal of COPD*. Dove Medical Press Ltd. https://doi.org/10.2147/COPD.S80227

Kim, W. J., Lee, M. K., Shin, C., Cho, N. H., Lee, S. Do, Oh, Y.-M., & Sung, J. (2014). Genome-wide association studies identify locus on 6p21 influencing lung function in the Korean population. *Respirology*, *19*(3), 360–368. https://doi.org/10.1111/resp.12230

Kim, W. J., Oh, Y.-M., Kim, T.-H., Lee, J.-H., Kim, E.-K., Lee, J. H., … Lee, S. Do. (2013). ***CHRNA3*** Variant for Lung Cancer Is Associated with Chronic Obstructive Pulmonary Disease in Korea. *Respiration*, *86*(2), 117–122. https://doi.org/10.1159/000342976

Kim, W. J., Oh, Y.-M., Lee, J. H., Park, C.-S., Park, S. W., Park, J. S., & Lee, S. Do. (2013). Genetic variants in *HHIP* are associated with $FEV_1$ in subjects with chronic obstructive pulmonary disease. *Respirology*, *18*(8), 1202–1209. https://doi.org/10.1111/resp.12139

Kopa, P. N., & Pawliczak, R. (2018). Effect of smoking on gene expression profile – overall mechanism, impact on respiratory system function, and reference to electronic cigarettes. *Toxicology Mechanisms and Methods*, *28*(6), 397–409. https://doi.org/10.1080/15376516.2018.1461289

Korotkevich, G., Sukhov, V., & Sergushichev, A. (2019). Fast gene set enrichment analysis. *BioRxiv*.

Kubáň, P., & Foret, F. (2013, December 17). Exhaled breath condensate: Determination of non-volatile compounds and their potential for clinical diagnosis and monitoring. A review. *Analytica Chimica Acta*. Anal Chim Acta. https://doi.org/10.1016/j.aca.2013.07.049

Kurmi, O. P., Semple, S., Simkhada, P., Smith, W. C. S., & Ayres, J. G. (2010). COPD and chronic bronchitis risk of indoor air pollution from solid fuel: a systematic review and meta-analysis. *Thorax*, *65*(3), 221–228. https://doi.org/10.1136/thx.2009.124644

Kvale, P. A., Conway, W. A., & Coates, E. O. (1980). Continuous or nocturnal oxygen therapy in hypoxemic chronic obstructive lung disease. A clinical trial. *Annals of Internal Medicine*, *93*(3), 391–398. https://doi.org/10.7326/0003-4819-93-3-391

Lamontagne, M., Bérubé, J.-C., Obeidat, M., Cho, M. H., Hobbs, B. D., Sakornsakolpat, P., … Bossé, Y. (2018). Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. *Human Molecular Genetics*, *27*(10), 1819–1829. https://doi.org/10.1093/hmg/ddy091

Lamprecht, B., McBurnie, M. A., Vollmer, W. M., Gudmundsson, G., Welte, T., Nizankowska-Mogilnicka, E., … BOLD Collaborative Research Group. (2011). COPD in Never Smokers. *Chest*, *139*(4), 752–763. https://doi.org/10.1378/chest.10-1253

Lee, E. J., In, K. H., Kim, J. H., Lee, S. Y., Shin, C., Shim, J. J., … Uhm, C. S. (2009). Proteomic analysis in lung tissue of smokers and COPD patients. *Chest*, *135*(2), 344–352. https://doi.org/10.1378/chest.08-1583

Lee, J.-H., Lee, Y. K., Kim, E.-K., Kim, T.-H., Huh, J. W., Kim, W. J., … Lee, S. Do. (2010). Responses to inhaled long-acting beta-agonist and corticosteroid according to COPD subtype. *Respiratory Medicine*, *104*(4), 542–549. https://doi.org/10.1016/j.rmed.2009.10.024

Lee, J. H., Cho, M. H., Hersh, C. P., McDonald, M.-L. N., Wells, J. M., Dransfield, M. T., … Silverman, E. K. (2015). *IREB2* and *GALC* Are Associated with Pulmonary Artery Enlargement in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory Cell and Molecular Biology*, *52*(3), 365–376. https://doi.org/10.1165/rcmb.2014-0210OC

Lequin, R. M. (2005). Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clinical Chemistry*, *51*(12), 2415–2418. https://doi.org/10.1373/clinchem.2005.051532

Li, C. X., Wheelock, C. E., Magnus Sköld, C., & Wheelock, Å. M. (2018). Integration of multi-omics datasets enables molecular classification of COPD. *European Respiratory Journal*, *51*(5). https://doi.org/10.1183/13993003.01930-2017

Li, J., Zhao, P., Yang, L., Li, Y., Tian, Y., & Li, S. (2016). System biology analysis of long-term effect and mechanism of Bufei Yishen on COPD revealed by system pharmacology and 3-omics profiling. *Scientific Reports*, *6*. https://doi.org/10.1038/srep25492

Liang, Y., Yeligar, S. M., & Brown, L. A. S. (2012). Exhaled Breath Condensate: A Promising Source for Biomarkers of Lung Disease. *The Scientific World Journal*, *2012*. https://doi.org/10.1100/2012/217518

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J. P., & Bateman, A. (2011). Molecular signatures database (MSigDB) 3.0. *BIOINFORMATICS APPLICATIONS NOTE*, *27*(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., … Murray, C. J.

(2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, *380*(9859), 2095–2128. https://doi.org/10.1016/S0140-6736(12)61728-0

Lunardi, F., Pasqualini, L., Nardo, G., Nannini, N., Vuljan, S., Schiavon, M., … Calabrese, F. (2017). Next generation sequencing in lung adenocarcinoma of smokers with and without chronic obstructive pulmonary disease (COPD) (p. PA2037). European Respiratory Society (ERS). https://doi.org/10.1183/1393003.congress-2017.pa2037

Lundbäck, B., Lindberg, A., Lindström, M., Rönmark, E., Jonsson, A. ., Jönsson, E., … Studies, O. L. D. in N. S. (2003). Not 15 But 50% of smokers develop COPD?—Report from the Obstructive Lung Disease in Northern Sweden Studies. *Respiratory Medicine*, *97*(2), 115–122. https://doi.org/10.1053/rmed.2003.1446

Macbeath, G. (2002). Protein microarrays and proteomics. *Nature Genetics*. https://doi.org/10.1038/ng1037

Maniscalco, M., Fuschillo, S., Paris, D., Cutignano, A., Sanduzzi, A., & Motta, A. (2019). Clinical metabolomics of exhaled breath condensate in chronic respiratory diseases. In *Advances in Clinical Chemistry* (Vol. 88, pp. 121–149). Academic Press Inc. https://doi.org/10.1016/bs.acc.2018.10.002

Marchetti, N., & Criner, G. J. (2015). Surgical Approaches to Treating Emphysema: Lung Volume Reduction Surgery, Bullectomy, and Lung Transplantation. *Seminars in Respiratory and Critical Care Medicine*, *36*(4), 592–608. https://doi.org/10.1055/s-0035-1556064

McClay, J. L., Adkins, D. E., Isern, N. G., O'Connell, T. M., Wooten, J. B., Zedler, B. K., … Van Den Oord, E. J. C. G. (2010). 1 H nuclear magnetic resonance metabolomics analysis identifies novel urinary biomarkers for lung function. *Journal of Proteome Research*, *9*(6), 3083–3090. https://doi.org/10.1021/pr1000048

Melani, A. S. (2015). Long-acting muscarinic antagonists. *Expert Review of Clinical Pharmacology*, *8*(4), 479–501. https://doi.org/10.1586/17512433.2015.1058154

Miner, K., Labitzke, K., Liu, B., Wang, P., Henckels, K., Gaida, K., … Sullivan, J. K. (2019). Drug repurposing: The anthelmintics niclosamide and nitazoxanide are potent TMEM16A antagonists that fully bronchodilate airways. *Frontiers in Pharmacology*, *10*(FEB). https://doi.org/10.3389/fphar.2019.00051

Moriuchi, M., Moriuchi, H., Turner, W., & Fauci, A. S. (1997). Cloning and analysis of the promoter region of CXCR4, a coreceptor for HIV-1 entry. *Journal of Immunology (Baltimore, Md. : 1950)*, *159*(9), 4322–4329. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9379028

Morrow, J. D., Zhou, X., Lao, T., Jiang, Z., Demeo, D. L., Cho, M. H., … Hersh, C. P. (2017). Functional interactors of three genome-wide association study genes are differentially expressed in severe chronic obstructive pulmonary disease lung tissue. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/srep44232

Mulvey, C., Thur, B., Crawford, M., & Godovac-Zimmermann, J. (2010). How Many Proteins are Missed in Quantitative Proteomics Based on MS/MS Sequencing Methods? *Proteomics Insights*, *3*(1), 61–66. https://doi.org/10.4137/PRI.S5882

Nannini, L. J., Lasserson, T. J., & Poole, P. (2012). Combined corticosteroid and long-acting beta(2)-agonist in one inhaler versus long-acting beta(2)-agonists for chronic obstructive pulmonary disease. *The Cochrane Database of Systematic Reviews*, (9), CD006829. https://doi.org/10.1002/14651858.CD006829.pub2

Nannini, L. J., Poole, P., Milan, S. J., Holmes, R., & Normansell, R. (2013, November 10). Combined corticosteroid and long-acting beta2-agonist in one inhaler versus placebo for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews*. John Wiley and Sons Ltd. https://doi.org/10.1002/14651858.CD003794.pub4

Naz, S., Kolmert, J., Yang, M., Reinke, S. N., Kamleh, M. A., Snowden, S., … Wheelock, C. E. (2017). Metabolomics analysis identifies sex-associated metabotypes of oxidative stress and the autotaxin-lysoPA axis in COPD. *The European Respiratory Journal*, *49*(6). https://doi.org/10.1183/13993003.02322-2016

Nénan, S., Planquois, J. M., Berna, P., De Mendez, I., Hitier, S., Shapiro, S. D., … Bertrand, C. P. (2005). Analysis of the inflammatory response induced by rhMMP-12 catalytic domain instilled in mouse airways. *International Immunopharmacology*, *5*(3), 511–524. https://doi.org/10.1016/j.intimp.2004.10.011

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, *6*(4). https://doi.org/10.1371/journal.pgen.1000895

Ntritsos, G., Franek, J., Belbasis, L., Christou, M. A., Markozannes, G., Altman, P., … Evangelou, E. (2018). Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis. *International Journal of Chronic Obstructive Pulmonary Disease*, *13*, 1507–1514. https://doi.org/10.2147/COPD.S146390

Obeidat, M., Nie, Y., Chen, V., Shannon, C. P., Andiappan, A. K., Lee, B., … Sin, D. D. (2017). Network-based analysis reveals novel gene signatures in peripheral blood of patients with chronic obstructive pulmonary disease. *Respiratory Research*, *18*(1). https://doi.org/10.1186/s12931-017-0558-1

Ohlmeier, S., Nieminen, P., Gao, J., Kanerva, T., Rönty, M., Toljamo, T., … Pulkkinen, V. (2016). Lung tissue proteomics identifies elevated transglutaminase 2 levels in stable chronic obstructive pulmonary disease. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, *310*(11), L1155–L1165. https://doi.org/10.1152/ajplung.00021.2016

Omachi, T. A., Eisner, M. D., Rames, A., Markovtsova, L., & Blanc, P. D. (2011). Matrix metalloproteinase-9 predicts pulmonary status declines in α1-antitrypsin deficiency. *Respiratory Research*, *12*(1), 35. https://doi.org/10.1186/1465-9921-12-35

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, *5*(12), 993–996. https://doi.org/10.1038/nrd2199

Page, G. P., Zakharkin, S. O., Kim, K., Mehta, T., Chen, L., & Zhang, K. (2007). Microarray analysis. *Methods in Molecular Biology (Clifton, N.J.)*. Humana Press. https://doi.org/10.1007/978-1-59745-530-5_20

Papakonstantinou, E., Roth, M., & Karakiulakis, G. (2012). Hyaluronic acid: A key molecule in skin aging. *Dermato-Endocrinology*. Landes Bioscience. https://doi.org/10.4161/derm.21923

Park, S. T., & Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. *International Neurourology Journal*. Korean Continence Society. https://doi.org/10.5213/inj.1632742.371

Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, *32*(4), 381–385. https://doi.org/10.1002/gepi.20303

Petrache, I., Natarajan, V., Zhen, L., Medler, T. R., Richter, A., Berdyshev, E. V., & Tuder, R. M. (2006). Ceramide causes pulmonary cell apoptosis and emphysema: A role for sphingolipid homeostasis in the maintenance of alveolar cells. In *Proceedings of the American Thoracic Society* (Vol. 3, p. 510). Proc Am Thorac Soc. https://doi.org/10.1513/pats.200603-071MS

Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., … Goldstein, D. B. (2009). A Genome-Wide association study in chronic obstructive pulmonary disease (COPD): Identification of two major susceptibility loci. *PLoS Genetics*, *5*(3). https://doi.org/10.1371/journal.pgen.1000421

Prasad, V., & Mailankody, S. (2017). Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. *JAMA Internal Medicine*, *177*(11), 1569. https://doi.org/10.1001/jamainternmed.2017.3601

Pulley, J. M., Shirey-Rice, J. K., Lavieri, R. R., Jerome, R. N., Zaleski, N. M., Aronoff, D. M., … Denny, J. C. (2017). Accelerating Precision Drug Development and Drug Repurposing by Leveraging Human Genetics. *Assay and Drug Development Technologies*, *15*(3), 113–119. https://doi.org/10.1089/adt.2016.772

Quaderi, S. A., & Hurst, J. R. (2018). The unmet global burden of COPD. *Global Health, Epidemiology and Genomics, 3,* e4. https://doi.org/10.1017/gheg.2018.1

Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., … Zielinski, J. (2007). Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, *176*(6), 532–555. https://doi.org/10.1164/rccm.200703-456SO

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., … Crapo, J. D. (2010). Genetic epidemiology of COPD (COPDGene) study design. *COPD*, *7*(1), 32–43. https://doi.org/10.3109/15412550903499522

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., … Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, *14*(2), 482–517. https://doi.org/10.1038/s41596-018-0103-9

Repapi, E., Sayers, I., Wain, L. V., Burton, P. R., Johnson, T., Obeidat, M., … Tobin, M. D. (2010). Genome-wide association study identifies five loci associated with lung function. *Nature Genetics*, *42*(1), 36–44. https://doi.org/10.1038/ng.501

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–e47. https://doi.org/10.1093/nar/gkv007

Roberts, M. E. P., Higgs, B. W., Brohawn, P., Pilataxi, F., Guo, X., Kuziora, M., … White, W. I. (2015). CD4+ T-Cell Profiles and Peripheral Blood Ex-Vivo Responses to T-Cell Directed Stimulation Delineate COPD Phenotypes. *Chronic Obstructive Pulmonary Diseases:*

*Journal of the COPD Foundation*, *2*(4), 268–280. https://doi.org/10.15326/jcopdf.2.4.2015.0131

Roca, J., Vargas, C., Cano, I., Selivanov, V., Barreiro, E., Maier, D., … Synergy-COPD consortium, the S.-C. (2014). Chronic Obstructive Pulmonary Disease heterogeneity: challenges for health risk assessment, stratification and management. *Journal of Translational Medicine*, *12 Suppl 2*(Suppl 2), S3. https://doi.org/10.1186/1479-5876-12-S2-S3

Salzberg, S. L. (2018). Open questions: How many genes do we have? *BMC Biology*, *16*(1), 94. https://doi.org/10.1186/s12915-018-0564-x

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., … Overington, J. P. (2016). A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, *16*(1), 19–34. https://doi.org/10.1038/nrd.2016.230

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, *11*(3), 191–200. https://doi.org/10.1038/nrd3681

Scarpa, M. C., Baraldo, S., Marian, E., Turato, G., Calabrese, F., Saetta, M., & Maestrelli, P. (2013). Ceramide expression and cell homeostasis in chronic obstructive pulmonary disease. *Respiration*, *85*(4), 342–349. https://doi.org/10.1159/000341185

Schamberger, A. C., Mise, N., Meiners, S., & Eickelberg, O. (2014). Epigenetic mechanisms in COPD: Implications for pathogenesis and drug discovery. *Expert Opinion on Drug Discovery*. Informa Healthcare. https://doi.org/10.1517/17460441.2014.913020

Shah, P. L., Herth, F. J., van Geffen, W. H., Deslee, G., & Slebos, D. J. (2017, February 1). Lung volume reduction for emphysema. *The Lancet Respiratory Medicine*. Lancet Publishing Group. https://doi.org/10.1016/S2213-2600(16)30221-1

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Simillion, C., Liechti, R., Lischer, H. E. L., Ioannidis, V., & Bruggmann, R. (2017). Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, *18*(1), 151. https://doi.org/10.1186/s12859-017-1571-6

Soler-Cataluña, J. J., Martínez-García, M. Á., Sánchez, P. R., Salcedo, E., Navarro, M., & Ochando, R. (2005). Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. *Thorax*, *60*(11), 925–931. https://doi.org/10.1136/THX.2005.040527

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*, 91. https://doi.org/10.1186/1471-2105-14-91

Spira, A., Beane, J., Pinto-Plata, V., Kadar, A., Liu, G., Shah, V., … Brody, J. S. (2004). Gene expression profiling of human lung tissue from smokers with severe emphysema. *American Journal of Respiratory Cell and Molecular Biology*, *31*(6), 601–610. https://doi.org/10.1165/rcmb.2004-0273OC

Spitzer, C., Gläser, S., Grabe, H. J., Ewert, R., Barnow, S., Felix, S. B., … Schäper, C. (2011). Mental health problems, obstructive lung disease and lung function: Findings from the general population. *Journal of Psychosomatic Research*, *71*(3), 174–179. https://doi.org/10.1016/j.jpsychores.2011.03.005

Sprent, P. (2011). Fisher Exact Test. In *International Encyclopedia of Statistical Science* (pp. 524–525). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_253

Stacey, S. N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D. F., … Stefansson, K. (2011). A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nature Genetics*, *43*(11), 1098–1103. https://doi.org/10.1038/ng.926

Stämpfli, M. R., & Anderson, G. P. (2009). How cigarette smoke skews immune responses to promote infection, lung disease and cancer. *Nature Reviews Immunology*, *9*(5), 377–384. https://doi.org/10.1038/nri2530

Steiling, K., Lenburg, M. E., & Spira, A. (2013, December). Personalized management of chronic obstructive pulmonary disease via transcriptomic profiling of the airway and lung. *Annals of the American Thoracic Society*. https://doi.org/10.1513/AnnalsATS.201306-190AW

Stekel, D. (2003). *Microarray bioinformatics*. Cambridge University Press.

Stoller, J. K., Panos, R. J., Krachman, S., Doherty, D. E., Make, B., & Long-term Oxygen Treatment Trial Research Group,  the L. O. T. T. R. (2010). Oxygen therapy for patients with COPD: current evidence and the long-term oxygen treatment trial. *Chest*, *138*(1), 179–187. https://doi.org/10.1378/chest.09-2555

Stuart, H., C, S. H., J, B., TJH, C., AC, D., JE, C., … PD, O. (1981). LONG TERM DOMICILIARY OXYGEN THERAPY IN CHRONIC HYPOXIC COR PULMONALE COMPLICATING CHRONIC BRONCHITIS AND EMPHYSEMA: REPORT OF THE MEDICAL RESEARCH COUNCIL WORKING PARTY. *LONG TERM DOMICILIARY OXYGEN THERAPY IN CHRONIC HYPOXIC COR PULMONALE COMPLICATING CHRONIC BRONCHITIS AND EMPHYSEMA: REPORT OF THE MEDICAL RESEARCH COUNCIL WORKING PARTY*.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., … Golub, T. R. (2017). A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell*, *171*(6), 1437. https://doi.org/10.1016/J.CELL.2017.10.049

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Südhof, T. C., & Rizo, J. (2011). Synaptic vesicle exocytosis. *Cold Spring Harbor Perspectives in Biology*, *3*(12), a005637. https://doi.org/10.1101/cshperspect.a005637

Sullivan, J., Pravosud, V., Mannino, D. M., Siegel, K., Choate, R., & Sullivan, T. (2018). National and State Estimates of COPD Morbidity and Mortality — United States, 2014-2015. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *5*(4), 324–333. https://doi.org/10.15326/jcopdf.5.4.2018.0157

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., … Mering, C. von. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Tan, J., Tedrow, J. R., Dutta, J. A., Juan-Guardela, B., Nouraie, M., Chu, Y., … Kass, D. J. (2016). Expression of RXFP1 is decreased in idiopathic pulmonary fibrosis implications for relaxin-based therapies. *American Journal of Respiratory and Critical Care Medicine*, *194*(11), 1392–1402. https://doi.org/10.1164/rccm.201509-1865OC

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J., … Spia, I. A. (2008). SPIA Supplementary Material. *Statistics*, *00*(00), 1–2. https://doi.org/10.1093/bioinformatics/btn577

The International HapMap Consortium, T. I. H. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Tibboel, J., Reiss, I., de Jongste, J. C., & Post, M. (2013). Ceramides: A potential therapeutic target in pulmonary emphysema. *Respiratory Research*, *14*(1), 96. https://doi.org/10.1186/1465-9921-14-96

Tomes, C. N., De Blas, G. A., Michaut, M. A., Farré, E. V., Cherhitin, O., Visconti, P. E., & Mayorga, L. S. (2005). α-SNAP and NSF are required in a priming step during the human sperm acrosome reaction. *Molecular Human Reproduction*, *11*(1), 43–51. https://doi.org/10.1093/MOLEHR/GAH126

Travaline, J. M., Addonizio, V. P., & Criner, G. J. (1995). Effect of bullectomy on diaphragm strength. *American Journal of Respiratory and Critical Care Medicine*, *152*(5 Pt 1), 1697–1701. https://doi.org/10.1164/ajrccm.152.5.7582315

Trulock, E. P. (1998). Lung Transplantation for COPD. *Chest*, *113*(4), 269S-276S. https://doi.org/10.1378/chest.113.4_supplement.269s

Turino, G. M., Ma, S., Lin, Y. Y., & Cantor, J. O. (2018). The Therapeutic Potential of Hyaluronan in COPD. *Chest*, *153*(4), 792–798. https://doi.org/10.1016/j.chest.2017.12.016

U.S. Department of Health and Human Services. (2004). The Health Consequences of Smoking: A Report of the Surgeon General. *National Library of Medicine*, *2012*, 51576–51576. https://doi.org/10.1002/yd.20075

Ubhi, B. K., Riley, J. H., Shaw, P. A., Lomas, D. A., Tal-Singer, R., MacNee, W., … Connor, S. C. (2012). Metabolic profiling detects biomarkers of protein degradation in COPD patients. *The European Respiratory Journal*, *40*(2), 345–355. https://doi.org/10.1183/09031936.00112411

Ueno, M., Maeno, T., Nishimura, S., Ogata, F., Masubuchi, H., Hara, K., … Kurabayashi, M. (2015). Alendronate inhalation ameliorates elastase-induced pulmonary emphysema in mice by induction of apoptosis of alveolar macrophages. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms7332

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419–1260419. https://doi.org/10.1126/science.1260419

United States Department of Health and Human Services. (2014). The Health Consequences of Smoking—50 Years of Progress A Report of the Surgeon General. *A Report of the Surgeon General*.

Vance, D. E., & Vance, J. E. (2008). Phospholipid biosynthesis in eukaryotes. In *Biochemistry of Lipids, Lipoproteins and Membranes* (pp. 213–244). Elsevier. https://doi.org/10.1016/B978-044453219-0.50010-6

Vargesson, N. (2017). Thalidomide. *Reproductive and Developmental Toxicology*, 467–479. https://doi.org/10.1016/B978-0-12-804239-7.00026-3

Vazquez Guillamet, R. (2019). Chronic Obstructive Pulmonary Disease and the Optimal Timing of Lung Transplantation. *Medicina (Kaunas, Lithuania)*, *55*(10). https://doi.org/10.3390/medicina55100646

Vestbo, J., Anderson, W., Coxson, H. O., Crim, C., Dawber, F., Edwards, L., … investigators, on behalf of the E. (2008). Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *European Respiratory Journal*, *31*(4), 869–873. https://doi.org/10.1183/09031936.00111707

Viniol, C., & Vogelmeier, C. F. (2018). Exacerbations of COPD. *European Respiratory Review : An Official Journal of the European Respiratory Society*, *27*(147), 170103. https://doi.org/10.1183/16000617.0103-2017

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., … Murray, C. J. L. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6

Vuckovic, D. (2018). Improving metabolome coverage and data quality: Advancing metabolomics and lipidomics for biomarker discovery. *Chemical Communications*, *54*(50), 6728–6749. https://doi.org/10.1039/c8cc02592d

Wain, L. V., Shrine, N., Artigas, M. S., Erzurumluoglu, A. M., Noyvert, B., Bossini-Castillo, L., … Tobin, M. D. (2017). Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nature Genetics*, *49*(3), 416–425. https://doi.org/10.1038/ng.3787

Wain, L. V, Shrine, N., Miller, S., Jackson, V. E., Ntalla, I., Soler Artigas, M., … Hall, I. P. (2015). Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet. Respiratory Medicine*, *3*(10), 769–781. https://doi.org/10.1016/S2213-2600(15)00283-0

Walton, G., Stockley, J., Griffiths, D., Sadhra, C., Purvis, T., & Sapey, E. (2016). Repurposing Treatments to Enhance Innate Immunity. Can Statins Improve Neutrophil Functions and Clinical Outcomes in COPD? *Journal of Clinical Medicine*, *5*(10), 89. https://doi.org/10.3390/jcm5100089

Wang, J., Wu, G., Chen, L., & Zhang, W. (2016). Integrated analysis of transcriptomic and proteomic datasets reveals information on protein expressivity and factors affecting translational efficiency. *Methods in Molecular Biology*, *1375*, 123–136. https://doi.org/10.1007/7651_2015_242

Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., & Bandeira, N. (2018). Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems*, *7*(4), 412-421.e5. https://doi.org/10.1016/j.cels.2018.08.004

Wheaton, A. G., Cunningham, T. J., Ford, E. S., & Croft, J. B. (2015). Employment and Activity Limitations Among Adults with Chronic Obstructive Pulmonary Disease — United States, 2013. *Morbidity and Mortality Weekly Report*, *64*(11).

Wilk, J. B., Chen, T. H., Gottlieb, D. J., Walter, R. E., Nagle, M. W., Brandler, B. J., … O'Connor, G. T. (2009). A genome-wide association study of pulmonary function measures in the framingham heart study. *PLoS Genetics*, *5*(3). https://doi.org/10.1371/journal.pgen.1000429

Wilk, J. B., Walter, R. E., Laramie, J. M., Gottlieb, D. J., & O'Connor, G. T. (2007). Framingham Heart Study genome-wide association: Results for pulmonary function measures. *BMC Medical Genetics*, *8*(SUPPL. 1). https://doi.org/10.1186/1471-2350-8-S1-S8

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., … Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037

Wood-Baker, R. R., Gibson, P. G., Hannay, M., Walters, E. H., & Walters, J. A. E. (2005). Systemic corticosteroids for acute exacerbations of chronic obstructive pulmonary disease. *The Cochrane Database of Systematic Reviews*, (1), CD001288. https://doi.org/10.1002/14651858.CD001288.pub2

Xu, H., Aldrich, M. C., Chen, Q., Liu, H., Peterson, N. B., Dai, Q., … Denny, J. C. (2015). Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association*, *22*(1), 179–191. https://doi.org/10.1136/amiajnl-2014-002649

Yang, I. A., Clarke, M. S., Sim, E. H. A., & Fong, K. M. (2012). Inhaled corticosteroids for stable chronic obstructive pulmonary disease.[Update of Cochrane Database Syst Rev. 2007;(2):CD002991; PMID: 17443520]. *Cochrane Database of Systematic Reviews*, *7*, CD002991. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=prem&AN=22786484

Yao, H., & Rahman, I. (2011). Current concepts on oxidative/carbonyl stress, inflammation and epigenetics in pathogenesis of chronic obstructive pulmonary disease. *Toxicology and Applied Pharmacology*, *254*(2), 72–85. https://doi.org/10.1016/j.taap.2009.10.022

Yates, J. R. (2011, August). A century of mass spectrometry: From atoms to proteomes. *Nature Methods*. https://doi.org/10.1038/nmeth.1659

YOO, K. H., KIM, Y. S., SHEEN, S. S., PARK, J. H., HWANG, Y. I., KIM, S.-H., … HAN, S. K. (2011). Prevalence of chronic obstructive pulmonary disease in Korea: The fourth Korean National Health and Nutrition Examination Survey, 2008. *Respirology*, *16*(4), 659–665. https://doi.org/10.1111/j.1440-1843.2011.01951.x

Yu, G., Wang, L.-G., Yan, G.-R., & He, Q.-Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis, *31*(4), 608–609. https://doi.org/10.1093/bioinformatics/btu684

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang, M., Yao, C., Guo, Z., Zou, J., Zhang, L., Xiao, H., … Troyanskaya, O. (2008). Apparently low reproducibility of true differential expression discoveries in microarray studies. *BIOINFORMATICS ORIGINAL PAPER*, *24*(18), 2057–2063. https://doi.org/10.1093/bioinformatics/btn365

Zhou, Y., Wang, C., Yao, W., Chen, P., Kang, J., Huang, S., … Ran, P. (2009). COPD in Chinese nonsmokers. *European Respiratory Journal*, *33*(3), 509–518. https://doi.org/10.1183/09031936.00084408

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., … Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. https://doi.org/10.1038/ng.3538

# APPENDICES

## APPENDIX A: DIFFERENTIALLY EXPRESSED GENES IN GOLD IV COPD PATIENTS VS HEALTHY CONTROLS

| Probe ID | gene | FoldChange | t-statistics | P.Value | adj.P.Val |
|---|---|---|---|---|---|
| A_32_P75581 | BHLHE22 | 2.61 | 9.07 | 1.56E-17 | 2.38E-13 |
| A_23_P20316 | CA3 | 3.02 | 8.01 | 2.43E-14 | 9.01E-11 |
| A_23_P170888 | DPP6 | 0.34 | -7.76 | 1.36E-13 | 2.96E-10 |
| A_23_P209625 | CYP1B1 | 2.48 | 7.69 | 2.05E-13 | 3.91E-10 |
| A_24_P738168 | FREM3 | 0.29 | -7.62 | 3.20E-13 | 5.42E-10 |
| A_23_P256948 | MSC | 2.64 | 7.57 | 4.64E-13 | 6.44E-10 |
| A_32_P70315 | TIMP4 | 3.03 | 7.49 | 7.59E-13 | 9.60E-10 |
| A_23_P324754 | CEMIP | 2.46 | 7.31 | 2.33E-12 | 1.62E-09 |
| A_23_P214144 | COL10A1 | 3.26 | 7.31 | 2.38E-12 | 1.62E-09 |
| A_23_P162918 | SERPINA3 | 2.56 | 7.28 | 2.86E-12 | 1.82E-09 |
| A_24_P174294 | LRRTM4 | 0.38 | -7.20 | 4.72E-12 | 2.56E-09 |
| A_23_P121064 | PTX3 | 4.74 | 7.17 | 5.90E-12 | 3.00E-09 |
| A_23_P217277 | SLITRK2 | 0.35 | -7.12 | 8.06E-12 | 3.84E-09 |
| A_23_P131024 | ZBTB32 | 2.74 | 7.10 | 8.87E-12 | 4.10E-09 |
| A_23_P127584 | NNMT | 2.38 | 7.01 | 1.56E-11 | 6.60E-09 |
| A_32_P227605 | RGS9BP | 0.40 | -6.99 | 1.80E-11 | 7.24E-09 |
| A_23_P137665 | CHI3L1 | 2.69 | 6.95 | 2.19E-11 | 8.15E-09 |
| A_23_P123853 | CCL19 | 3.44 | 6.93 | 2.59E-11 | 8.99E-09 |
| A_24_P286114 | SLC1A3 | 2.11 | 6.88 | 3.48E-11 | 1.13E-08 |
| A_23_P62115 | TIMP1 | 2.11 | 6.85 | 4.17E-11 | 1.27E-08 |
| A_24_P945096 | CACNA1I | 2.19 | 6.77 | 6.62E-11 | 1.63E-08 |
| A_23_P165598 | DAPL1 | 3.80 | 6.74 | 7.96E-11 | 1.77E-08 |
| A_23_P382240 | TMEM26 | 0.44 | -6.74 | 8.01E-11 | 1.77E-08 |
| A_23_P40174 | MMP9 | 2.87 | 6.73 | 8.26E-11 | 1.77E-08 |
| A_23_P13548 | CHRDL2 | 3.62 | 6.66 | 1.27E-10 | 2.54E-08 |
| A_23_P41789 | SLC27A6 | 0.37 | -6.66 | 1.31E-10 | 2.56E-08 |
| A_24_P403150 | FLJ34503 | 0.34 | -6.63 | 1.55E-10 | 2.84E-08 |
| A_23_P147255 | PCBP3 | 2.22 | 6.58 | 2.10E-10 | 3.52E-08 |
| A_23_P16834 | FNDC4 | 2.02 | 6.56 | 2.38E-10 | 3.80E-08 |
| A_23_P354855 | RS1 | 0.48 | -6.55 | 2.48E-10 | 3.87E-08 |
| A_23_P20311 | DPYS | 2.45 | 6.47 | 4.00E-10 | 5.70E-08 |
| A_23_P2674 | KRT4 | 0.37 | -6.46 | 4.19E-10 | 5.82E-08 |
| A_23_P121253 | TNFSF10 | 0.50 | -6.45 | 4.54E-10 | 6.19E-08 |
| A_24_P63019 | IL1R2 | 3.23 | 6.35 | 7.67E-10 | 9.43E-08 |
| A_24_P252945 | CXCR5 | 2.94 | 6.31 | 9.80E-10 | 1.12E-07 |

| | | | | | |
|---|---|---|---|---|---|
| A_23_P2645 | SDS | 3.13 | 6.30 | 1.02E-09 | 1.16E-07 |
| A_23_P118158 | HS3ST2 | 2.48 | 6.27 | 1.25E-09 | 1.35E-07 |
| A_23_P145096 | PLA2G7 | 2.70 | 6.27 | 1.26E-09 | 1.35E-07 |
| A_24_P37253 | LYPD6 | 0.43 | -6.27 | 1.26E-09 | 1.35E-07 |
| A_23_P108751 | FHL2 | 2.06 | 6.25 | 1.36E-09 | 1.41E-07 |
| A_23_P113572 | CD19 | 2.97 | 6.22 | 1.68E-09 | 1.72E-07 |
| A_23_P413760 | P2RX5 | 2.19 | 6.20 | 1.81E-09 | 1.77E-07 |
| A_23_P31725 | BLK | 2.90 | 6.12 | 2.97E-09 | 2.58E-07 |
| A_24_P10226 | SEMA6D | 0.48 | -6.10 | 3.24E-09 | 2.71E-07 |
| A_23_P90888 | CHRNA1 | 3.34 | 6.01 | 5.34E-09 | 4.12E-07 |
| A_23_P145841 | SOSTDC1 | 0.29 | -6.01 | 5.47E-09 | 4.18E-07 |
| A_23_P72697 | GPIHBP1 | 0.46 | -5.97 | 6.79E-09 | 4.91E-07 |
| A_23_P140760 | ADGRG3 | 2.20 | 5.95 | 7.46E-09 | 5.25E-07 |
| A_23_P433050 | RXFP1 | 0.47 | -5.93 | 8.18E-09 | 5.65E-07 |
| A_23_P85209 | IL13RA2 | 2.85 | 5.92 | 8.75E-09 | 5.96E-07 |
| A_32_P8221 | GRM8 | 0.30 | -5.92 | 8.85E-09 | 6.00E-07 |
| A_24_P264943 | COMP | 3.13 | 5.89 | 1.05E-08 | 6.84E-07 |
| A_23_P253321 | PNOC | 2.35 | 5.88 | 1.07E-08 | 6.92E-07 |
| A_23_P312920 | POU2AF1 | 2.71 | 5.87 | 1.16E-08 | 7.41E-07 |
| A_23_P113793 | ZBED2 | 0.47 | -5.84 | 1.36E-08 | 8.61E-07 |
| A_23_P77493 | TUBB3 | 3.09 | 5.83 | 1.43E-08 | 8.93E-07 |
| A_23_P31755 | CRH | 4.00 | 5.82 | 1.50E-08 | 9.33E-07 |
| A_23_P310274 | PRSS2 | 2.28 | 5.77 | 2.00E-08 | 1.18E-06 |
| A_23_P46039 | FCRLA | 3.55 | 5.75 | 2.19E-08 | 1.26E-06 |
| A_23_P114983 | TRIM63 | 2.17 | 5.75 | 2.24E-08 | 1.27E-06 |
| A_23_P115161 | ACKR1 | 2.17 | 5.71 | 2.70E-08 | 1.49E-06 |
| A_23_P74001 | S100A12 | 2.78 | 5.68 | 3.19E-08 | 1.68E-06 |
| A_24_P122137 | LIF | 2.82 | 5.67 | 3.36E-08 | 1.74E-06 |
| A_24_P366607 | SERTM1 | 0.45 | -5.66 | 3.43E-08 | 1.77E-06 |
| A_23_P347777 | SYCP2L | 0.46 | -5.65 | 3.80E-08 | 1.90E-06 |
| A_24_P142118 | THBS1 | 2.46 | 5.64 | 3.92E-08 | 1.95E-06 |
| A_23_P432947 | GREM1 | 3.44 | 5.63 | 4.08E-08 | 2.02E-06 |
| A_24_P917819 | ANKRD30BP2 | 2.61 | 5.62 | 4.30E-08 | 2.10E-06 |
| A_23_P201211 | FCRL5 | 2.71 | 5.57 | 5.65E-08 | 2.66E-06 |
| A_23_P372946 | TM4SF19 | 2.67 | 5.57 | 5.76E-08 | 2.70E-06 |
| A_32_P345659 | XIRP1 | 2.42 | 5.56 | 5.82E-08 | 2.71E-06 |
| A_23_P200096 | SPSB1 | 2.11 | 5.56 | 5.88E-08 | 2.73E-06 |
| A_24_P285055 | PCDH11Y | 0.50 | -5.55 | 6.13E-08 | 2.80E-06 |
| A_23_P31453 | STEAP1 | 2.08 | 5.55 | 6.24E-08 | 2.84E-06 |
| A_24_P13041 | RTKN2 | 0.43 | -5.54 | 6.60E-08 | 2.94E-06 |
| A_23_P368779 | ZNF114 | 0.50 | -5.52 | 7.26E-08 | 3.16E-06 |

| | | | | | |
|---|---|---|---|---|---|
| A_23_P144916 | GFPT2 | 2.29 | 5.46 | 9.85E-08 | 3.99E-06 |
| A_23_P104471 | DUSP13 | 2.08 | 5.46 | 1.02E-07 | 4.09E-06 |
| A_23_P65307 | SLITRK6 | 0.43 | -5.43 | 1.17E-07 | 4.58E-06 |
| A_23_P1691 | MMP1 | 3.90 | 5.43 | 1.17E-07 | 4.59E-06 |
| A_23_P39067 | SPIB | 2.17 | 5.39 | 1.40E-07 | 5.27E-06 |
| A_23_P334870 | TMEM217 | 2.04 | 5.39 | 1.43E-07 | 5.32E-06 |
| A_23_P152262 | DPEP1 | 2.63 | 5.37 | 1.57E-07 | 5.73E-06 |
| A_23_P148737 | MYBPH | 2.83 | 5.33 | 1.91E-07 | 6.67E-06 |
| A_23_P129144 | MYZAP | 0.50 | -5.33 | 1.94E-07 | 6.73E-06 |
| A_23_P127288 | IL2RA | 2.16 | 5.32 | 2.05E-07 | 7.02E-06 |
| A_24_P349117 | GPR158 | 0.39 | -5.28 | 2.53E-07 | 8.35E-06 |
| A_23_P7313 | SPP1 | 3.43 | 5.27 | 2.66E-07 | 8.68E-06 |
| A_24_P335620 | SLC7A5 | 2.13 | 5.15 | 4.59E-07 | 1.36E-05 |
| A_24_P71904 | HPGD | 0.49 | -5.14 | 4.87E-07 | 1.43E-05 |
| A_23_P366936 | KRT6C | 2.80 | 5.14 | 4.96E-07 | 1.45E-05 |
| A_23_P204947 | GJB2 | 2.63 | 5.10 | 5.91E-07 | 1.65E-05 |
| A_23_P81280 | BTNL9 | 0.44 | -4.99 | 1.04E-06 | 2.58E-05 |
| A_23_P208747 | PGLYRP1 | 2.15 | 4.98 | 1.09E-06 | 2.67E-05 |
| A_23_P18362 | SLITRK3 | 0.48 | -4.97 | 1.14E-06 | 2.76E-05 |
| A_24_P245379 | SERPINB2 | 2.27 | 4.91 | 1.47E-06 | 3.35E-05 |
| A_23_P126278 | CHIT1 | 3.97 | 4.91 | 1.51E-06 | 3.42E-05 |
| A_24_P37409 | DUSP2 | 2.03 | 4.89 | 1.66E-06 | 3.71E-05 |
| A_32_P3476 | RPRML | 2.42 | 4.88 | 1.72E-06 | 3.83E-05 |
| A_23_P84596 | MZB1 | 2.26 | 4.84 | 2.05E-06 | 4.38E-05 |
| A_23_P78248 | KRT23 | 2.61 | 4.81 | 2.37E-06 | 4.94E-05 |
| A_23_P421306 | SYT12 | 2.31 | 4.80 | 2.52E-06 | 5.16E-05 |
| A_24_P244706 | SSTR1 | 0.49 | -4.79 | 2.63E-06 | 5.33E-05 |
| A_23_P166408 | OSM | 2.58 | 4.74 | 3.22E-06 | 6.19E-05 |
| A_23_P58266 | S100P | 2.29 | 4.73 | 3.48E-06 | 6.60E-05 |
| A_23_P148088 | FGG | 4.16 | 4.72 | 3.61E-06 | 6.74E-05 |
| A_24_P335092 | SAA1 | 3.25 | 4.71 | 3.72E-06 | 6.87E-05 |
| A_23_P66635 | CCL11 | 2.29 | 4.71 | 3.73E-06 | 6.88E-05 |
| A_24_P57993 | SHISAL1 | 2.02 | 4.70 | 3.97E-06 | 7.20E-05 |
| A_23_P121695 | CXCL13 | 4.19 | 4.66 | 4.76E-06 | 8.29E-05 |
| A_23_P360754 | ADAMTS4 | 4.71 | 4.54 | 8.05E-06 | 1.26E-04 |
| A_23_P86470 | CH25H | 2.38 | 4.53 | 8.56E-06 | 1.32E-04 |
| A_23_P42897 | MGAM | 2.03 | 4.49 | 1.03E-05 | 1.53E-04 |
| A_23_P96158 | KRT17 | 3.33 | 4.48 | 1.05E-05 | 1.55E-04 |
| A_24_P53778 | ITLN2 | 0.46 | -4.48 | 1.07E-05 | 1.57E-04 |
| A_23_P380240 | CEACAM8 | 3.05 | 4.47 | 1.11E-05 | 1.62E-04 |
| A_23_P38537 | KRT16 | 2.69 | 4.45 | 1.21E-05 | 1.72E-04 |

| | | | | | |
|---|---|---|---|---|---|
| A_23_P146274 | STMN2 | 2.27 | 4.44 | 1.28E-05 | 1.80E-04 |
| A_23_P375372 | FGA | 3.46 | 4.42 | 1.37E-05 | 1.90E-04 |
| A_24_P97342 | PROK2 | 2.37 | 4.36 | 1.79E-05 | 2.33E-04 |
| A_24_P125469 | LIPG | 2.05 | 4.36 | 1.79E-05 | 2.33E-04 |
| A_23_P161624 | FOSL1 | 2.35 | 4.33 | 2.04E-05 | 2.58E-04 |
| A_23_P79518 | IL1B | 2.01 | 4.31 | 2.18E-05 | 2.72E-04 |
| A_23_P144656 | CDH10 | 0.45 | -4.31 | 2.25E-05 | 2.78E-04 |
| A_24_P941217 | SGPP2 | 2.05 | 4.21 | 3.32E-05 | 3.77E-04 |
| A_23_P206760 | HP | 2.09 | 4.20 | 3.53E-05 | 3.97E-04 |
| A_23_P22134 | BNC1 | 2.61 | 4.15 | 4.40E-05 | 4.67E-04 |
| A_24_P242581 | SLC5A9 | 0.46 | -4.11 | 5.01E-05 | 5.15E-04 |
| A_23_P386478 | TNIP3 | 2.11 | 4.11 | 5.08E-05 | 5.21E-04 |
| A_23_P79769 | BIRC7 | 2.02 | 4.03 | 7.02E-05 | 6.73E-04 |
| A_23_P78037 | CCL7 | 2.56 | 4.00 | 7.81E-05 | 7.37E-04 |
| A_24_P200219 | UPK1B | 2.25 | 3.99 | 8.18E-05 | 7.59E-04 |
| A_24_P174793 | PCSK1 | 2.45 | 3.98 | 8.52E-05 | 7.84E-04 |
| A_23_P128744 | BDKRB1 | 2.79 | 3.96 | 9.18E-05 | 8.29E-04 |
| A_23_P99515 | MEDAG | 2.55 | 3.96 | 9.26E-05 | 8.33E-04 |
| A_23_P13094 | MMP10 | 2.73 | 3.93 | 1.03E-04 | 9.12E-04 |
| A_23_P218626 | NEU4 | 2.14 | 3.90 | 1.21E-04 | 1.03E-03 |
| A_32_P780817 | CT45A1 | 0.29 | -3.86 | 1.38E-04 | 1.15E-03 |
| A_32_P87013 | CXCL8 | 3.00 | 3.84 | 1.52E-04 | 1.24E-03 |
| A_23_P89431 | CCL2 | 2.12 | 3.77 | 1.95E-04 | 1.50E-03 |
| A_32_P385587 | ALAS2 | 2.12 | 3.69 | 2.71E-04 | 1.95E-03 |
| A_23_P362694 | FDCSP | 3.38 | 3.66 | 3.03E-04 | 2.14E-03 |
| A_23_P321949 | PLA2G2A | 2.59 | 3.55 | 4.46E-04 | 2.90E-03 |
| A_23_P7727 | HAPLN1 | 0.47 | -3.55 | 4.51E-04 | 2.93E-03 |
| A_23_P146146 | ATP6V0D2 | 2.16 | 3.54 | 4.61E-04 | 2.97E-03 |
| A_23_P208768 | FCAR | 2.07 | 3.52 | 4.98E-04 | 3.16E-03 |
| A_32_P170547 | CT45A5 | 0.39 | -3.51 | 5.17E-04 | 3.25E-03 |
| A_23_P161659 | SYT13 | 0.49 | -3.50 | 5.32E-04 | 3.32E-03 |
| A_23_P97112 | SELE | 3.00 | 3.38 | 8.10E-04 | 4.66E-03 |
| A_23_P71037 | IL6 | 2.81 | 3.34 | 9.42E-04 | 5.23E-03 |
| A_23_P27400 | HAS1 | 3.25 | 3.31 | 1.03E-03 | 5.63E-03 |
| A_24_P788878 | C2CD4B | 2.00 | 3.30 | 1.09E-03 | 5.88E-03 |
| A_23_P7144 | CXCL1 | 2.01 | 3.11 | 2.05E-03 | 9.73E-03 |
| A_23_P218047 | KRT5 | 2.61 | 3.05 | 2.49E-03 | 1.14E-02 |
| A_23_P337689 | WT1-AS | 2.32 | 3.03 | 2.62E-03 | 1.18E-02 |
| A_23_P95790 | ITLN1 | 2.74 | 2.81 | 5.21E-03 | 2.02E-02 |
| A_23_P251031 | IL13 | 2.14 | 2.54 | 1.15E-02 | 3.79E-02 |

## APPENDIX B: STATISTICALLY SIGNIFICANT OVER-REPRESENTED BIOLOGICAL PROCESSES IN GOLD IV PATIENTS

| ID | Description | Gene Ratio | Bg Ratio | p.adjust | geneID |
|---|---|---|---|---|---|
| GO:0030595 | leukocyte chemotaxis | 16/145 | 202/17653 | 2.42E-08 | CCL19/CXCR5/PLA2G7/S100A12/THBS1/GREM1/SAA1/CCL11/CXCL13/CH25H/IL1B/CCL7/CXCL8/CCL2/IL6/CXCL1 |
| GO:0050900 | leukocyte migration | 22/145 | 470/17653 | 4.79E-08 | CCL19/MMP9/CXCR5/PLA2G7/S100A12/THBS1/GREM1/MMP1/SLC7A5/SAA1/CCL11/CXCL13/CH25H/CEACAM8/IL1B/CCL7/BDKRB1/CXCL8/CCL2/SELE/IL6/CXCL1 |
| GO:0043062 | extracellular structure organization | 20/145 | 395/17653 | 6.68E-08 | CYP1B1/COL10A1/TIMP1/MMP9/PLA2G7/GPIHBP1/COMP/PRSS2/THBS1/GREM1/MMP1/SPP1/FGG/ADAMTS4/FGA/LIPG/MMP10/PLA2G2A/HAPLN1/HAS1 |
| GO:0060326 | cell chemotaxis | 16/145 | 273/17653 | 5.38E-07 | CCL19/CXCR5/PLA2G7/S100A12/THBS1/GREM1/SAA1/CCL11/CXCL13/CH25H/IL1B/CCL7/CXCL8/CCL2/IL6/CXCL1 |
| GO:0097529 | myeloid leukocyte migration | 13/145 | 177/17653 | 1.17E-06 | CCL19/PLA2G7/S100A12/THBS1/GREM1/SAA1/CCL11/IL1B/CCL7/CXCL8/CCL2/IL6/CXCL1 |
| GO:0071621 | granulocyte chemotaxis | 10/145 | 105/17653 | 5.75E-06 | CCL19/S100A12/THBS1/SAA1/CCL11/IL1B/CCL7/CXCL8/CCL2/CXCL1 |
| GO:0071674 | mononuclear cell migration | 9/145 | 79/17653 | 5.75E-06 | CCL19/PLA2G7/S100A12/THBS1/GREM1/CCL11/CCL7/CCL2/IL6 |
| GO:0030198 | extracellular matrix organization | 16/145 | 341/17653 | 6.23E-06 | CYP1B1/COL10A1/TIMP1/MMP9/COMP/PRSS2/THBS1/GREM1/MMP1/SPP1/FGG/ADAMTS4/FGA/MMP10/HAPLN1/HAS1 |
| GO:0070098 | chemokine-mediated signaling pathway | 9/145 | 82/17653 | 6.23E-06 | CCL19/CXCR5/ACKR1/CCL11/CXCL13/CCL7/CXCL8/CCL2/CXCL1 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0002688 | regulation of leukocyte chemotaxis | 10/145 | 111/17653 | 6.24E-06 | CCL19/PLA2G7/THBS1/GREM1/CXCL13/CCL7/CXCL8/CCL2/IL6/CXCL1 |
| GO:0002548 | monocyte chemotaxis | 8/145 | 62/17653 | 8.66E-06 | CCL19/PLA2G7/S100A12/GREM1/CCL11/CCL7/CCL2/IL6 |
| GO:0030593 | neutrophil chemotaxis | 9/145 | 88/17653 | 8.74E-06 | CCL19/S100A12/SAA1/CCL11/IL1B/CCL7/CXCL8/CCL2/CXCL1 |
| GO:0097530 | granulocyte migration | 10/145 | 119/17653 | 9.38E-06 | CCL19/S100A12/THBS1/SAA1/CCL11/IL1B/CCL7/CXCL8/CCL2/CXCL1 |
| GO:0070555 | response to interleukin-1 | 12/145 | 192/17653 | 1.07E-05 | CHI3L1/CCL19/IL1R2/TRIM63/FGG/CCL11/IL1B/CCL7/CXCL8/CCL2/SELE/IL6 |
| GO:1990266 | neutrophil migration | 9/145 | 99/17653 | 1.96E-05 | CCL19/S100A12/SAA1/CCL11/IL1B/CCL7/CXCL8/CCL2/CXCL1 |
| GO:0002685 | regulation of leukocyte migration | 11/145 | 176/17653 | 3.38E-05 | CCL19/PLA2G7/THBS1/GREM1/CXCL13/CCL7/BDKRB1/CXCL8/CCL2/IL6/CXCL1 |
| GO:0002690 | positive regulation of leukocyte chemotaxis | 8/145 | 87/17653 | 7.81E-05 | CCL19/PLA2G7/THBS1/CXCL13/CCL7/CXCL8/IL6/CXCL1 |
| GO:0048247 | lymphocyte chemotaxis | 7/145 | 60/17653 | 7.81E-05 | CCL19/SAA1/CCL11/CXCL13/CH25H/CCL7/CCL2 |
| GO:0002687 | positive regulation of leukocyte migration | 9/145 | 123/17653 | 9.97E-05 | CCL19/PLA2G7/THBS1/CXCL13/CCL7/BDKRB1/CXCL8/IL6/CXCL1 |
| GO:0071347 | cellular response to interleukin-1 | 10/145 | 170/17653 | 0.00017 | CHI3L1/CCL19/IL1R2/FGG/CCL11/IL1B/CCL7/CXCL8/CCL2/IL6 |
| GO:0032102 | negative regulation of response to external stimulus | 13/145 | 310/17653 | 0.000195 | FNDC4/IL1R2/THBS1/GREM1/IL2RA/SPP1/PGLYRP1/SERPINB2/FGG/SAA1/CXCL13/FGA/CCL2 |
| GO:0032496 | response to lipopolysaccharide | 13/145 | 327/17653 | 0.000335 | TIMP4/HPGD/GJB2/CXCL13/IL1B/TNIP3/BDKRB1/CXCL8/CCL2/SELE/IL6/CXCL1/IL13 |
| GO:0007565 | female pregnancy | 10/145 | 187/17653 | 0.000345 | TIMP1/MMP9/PNOC/CRH/LIF/SPP1/HPGD/GJB2/FOSL1/IL1B |

| | | | | | |
|---|---|---|---|---|---|
| GO:0042119 | neutrophil activation | 16/145 | 498/17653 | 0.000345 | SERPINA3/PTX3/CHI3L1/MMP9/ADGRG3/PRSS2/S100A12/PGLYRP1/CHIT1/S100P/MGAM/CEACAM8/HP/CXCL8/FCAR/CXCL1 |
| GO:0002446 | neutrophil mediated immunity | 16/145 | 500/17653 | 0.000348 | SERPINA3/PTX3/CHI3L1/MMP9/ADGRG3/PRSS2/S100A12/PGLYRP1/CHIT1/S100P/MGAM/CEACAM8/HP/FCAR/IL6/CXCL1 |
| GO:0032103 | positive regulation of response to external stimulus | 12/145 | 289/17653 | 0.000444 | CCL19/PLA2G7/S100A12/THBS1/OSM/CXCL13/IL1B/CCL7/CXCL8/PLA2G2A/IL6/CXCL1 |
| GO:0002237 | response to molecule of bacterial origin | 13/145 | 345/17653 | 0.00049 | TIMP4/HPGD/GJB2/CXCL13/IL1B/TNIP3/BDKRB1/CXCL8/CCL2/SELE/IL6/CXCL1/IL13 |
| GO:0050920 | regulation of chemotaxis | 10/145 | 202/17653 | 0.00057 | CCL19/PLA2G7/THBS1/GREM1/CXCL13/CCL7/CXCL8/CCL2/IL6/CXCL1 |
| GO:0043312 | neutrophil degranulation | 15/145 | 485/17653 | 0.000895 | SERPINA3/PTX3/CHI3L1/MMP9/ADGRG3/PRSS2/S100A12/PGLYRP1/CHIT1/S100P/MGAM/CEACAM8/HP/FCAR/CXCL1 |
| GO:0050921 | positive regulation of chemotaxis | 8/145 | 130/17653 | 0.000895 | CCL19/PLA2G7/THBS1/CXCL13/CCL7/CXCL8/IL6/CXCL1 |
| GO:0002283 | neutrophil activation involved in immune response | 15/145 | 488/17653 | 0.000895 | SERPINA3/PTX3/CHI3L1/MMP9/ADGRG3/PRSS2/S100A12/PGLYRP1/CHIT1/S100P/MGAM/CEACAM8/HP/FCAR/CXCL1 |
| GO:0022617 | extracellular matrix disassembly | 7/145 | 94/17653 | 0.000895 | TIMP1/MMP9/PRSS2/MMP1/SPP1/ADAMTS4/MMP10 |
| GO:0072676 | lymphocyte migration | 7/145 | 94/17653 | 0.000895 | CCL19/SAA1/CCL11/CXCL13/CH25H/CCL7/CCL2 |
| GO:0044706 | multi-multicellular organism process | 10/145 | 218/17653 | 0.000916 | TIMP1/MMP9/PNOC/CRH/LIF/SPP1/HPGD/GJB2/FOSL1/IL1B |
| GO:0051928 | positive regulation of calcium | 7/145 | 108/17653 | 0.00209 | CEMIP/CD19/P2RX5/CRH/BDKRB1/CCL2/IL13 |

| | | | | | |
|---|---|---|---|---|---|
| | ion transport | | | | |
| GO:0050727 | regulation of inflammatory response | 13/145 | 411/17653 | 0.002339 | FNDC4/IL1R2/PLA2G7/CD19/S100A12/IL2RA/PGLYRP1/OSM/SAA1/IL1B/PLA2G2A/SELE/IL6 |
| GO:0006953 | acute-phase response | 5/145 | 48/17653 | 0.002919 | SERPINA3/SAA1/IL1B/HP/IL6 |
| GO:0030335 | positive regulation of cell migration | 14/145 | 486/17653 | 0.003049 | CEMIP/CCL19/MMP9/PLA2G7/SEMA6D/THBS1/CCL11/CXCL13/IL1B/CCL7/BDKRB1/CXCL8/IL6/CXCL1 |
| GO:0060627 | regulation of vesicle-mediated transport | 14/145 | 490/17653 | 0.003245 | PTX3/CCL19/CACNA1I/IL13RA2/GREM1/SYT12/FGG/FGA/IL1B/CCL2/SYT13/SELE/C2CD4B/IL13 |
| GO:0007566 | embryo implantation | 5/145 | 50/17653 | 0.003297 | TIMP1/MMP9/LIF/SPP1/IL1B |
| GO:0042730 | fibrinolysis | 4/145 | 27/17653 | 0.003801 | THBS1/SERPINB2/FGG/FGA |
| GO:0010524 | positive regulation of calcium ion transport into cytosol | 5/145 | 52/17653 | 0.003801 | CEMIP/CD19/P2RX5/BDKRB1/IL13 |
| GO:0034368 | protein-lipid complex remodeling | 4/145 | 28/17653 | 0.004187 | PLA2G7/GPIHBP1/LIPG/PLA2G2A |
| GO:0034369 | plasma lipoprotein particle remodeling | 4/145 | 28/17653 | 0.004187 | PLA2G7/GPIHBP1/LIPG/PLA2G2A |
| GO:0046677 | response to antibiotic | 11/145 | 326/17653 | 0.004312 | CA3/CYP1B1/SLC1A3/CRH/HPGD/GJB2/FOSL1/HP/CCL7/IL6/IL13 |
| GO:0006959 | humoral immune response | 11/145 | 329/17653 | 0.004518 | CD19/POU2AF1/PRSS2/S100A12/PGLYRP1/CXCL13/FGA/CCL2/PLA2G2A/IL6/ITLN1 |
| GO:0034367 | protein-containing complex remodeling | 4/145 | 29/17653 | 0.004518 | PLA2G7/GPIHBP1/LIPG/PLA2G2A |
| GO:0034116 | positive regulation of | 3/145 | 12/17653 | 0.00566 | FGG/FGA/IL1B |

| | | | | | |
|---|---|---|---|---|---|
| | heterotypic cell-cell adhesion | | | | |
| GO:0030213 | hyaluronan biosynthetic process | 3/14 5 | 13/1765 3 | 0.00716 5 | CEMIP/IL1B/HAS1 |
| GO:0017157 | regulation of exocytosis | 8/14 5 | 186/176 53 | 0.00731 7 | CACNA1I/IL13RA2/SYT12/FGG/F GA/SYT13/C2CD4B/IL13 |
| GO:0050714 | positive regulation of protein secretion | 9/14 5 | 238/176 53 | 0.00737 1 | CCL19/BLK/CRH/FGG/SAA1/FGA/ IL1B/IL6/IL13 |
| GO:0030336 | negative regulation of cell migration | 10/1 45 | 296/176 53 | 0.00784 2 | CYP1B1/TIMP1/SEMA6D/THBS1/ GREM1/DPEP1/CXCL13/KRT16/C CL2/HAS1 |
| GO:1903532 | positive regulation of secretion by cell | 11/1 45 | 361/176 53 | 0.00895 6 | CCL19/CACNA1I/BLK/CRH/SPP1/F GG/SAA1/FGA/IL1B/IL6/IL13 |
| GO:0030574 | collagen catabolic process | 5/14 5 | 68/1765 3 | 0.01067 3 | COL10A1/MMP9/PRSS2/MMP1/ MMP10 |
| GO:0070372 | regulation of ERK1 and ERK2 cascade | 10/1 45 | 311/176 53 | 0.01068 7 | CHI3L1/CCL19/LIF/FGG/CCL11/FG A/IL1B/CCL7/CCL2/PLA2G2A |
| GO:2000146 | negative regulation of cell motility | 10/1 45 | 311/176 53 | 0.01068 7 | CYP1B1/TIMP1/SEMA6D/THBS1/ GREM1/DPEP1/CXCL13/KRT16/C CL2/HAS1 |
| GO:0023061 | signal release | 12/1 45 | 435/176 53 | 0.01068 7 | BLK/CRH/LIF/SPP1/SYT12/OSM/F GG/FGA/IL1B/SYT13/IL6/C2CD4B |
| GO:0051281 | positive regulation of release of sequestered calcium ion into cytosol | 4/14 5 | 38/1765 3 | 0.01074 5 | CEMIP/CD19/BDKRB1/IL13 |
| GO:0019730 | antimicrobia l humoral response | 6/14 5 | 108/176 53 | 0.01084 3 | PRSS2/S100A12/PGLYRP1/FGA/P LA2G2A/ITLN1 |
| GO:0002793 | positive regulation of peptide secretion | 9/14 5 | 259/176 53 | 0.01174 6 | CCL19/BLK/CRH/FGG/SAA1/FGA/ IL1B/IL6/IL13 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0009612 | response to mechanical stimulus | 8/145 | 209/17653 | 0.013175 | CHI3L1/SLC1A3/THBS1/SLITRK6/FOSL1/IL1B/BDKRB1/IL13 |
| GO:0070268 | cornification | 6/145 | 113/17653 | 0.013175 | KRT4/KRT6C/KRT23/KRT17/KRT16/KRT5 |
| GO:0051047 | positive regulation of secretion | 11/145 | 388/17653 | 0.013952 | CCL19/CACNA1I/BLK/CRH/SPP1/FGG/SAA1/FGA/IL1B/IL6/IL13 |
| GO:0070371 | ERK1 and ERK2 cascade | 10/145 | 328/17653 | 0.01446 | CHI3L1/CCL19/LIF/FGG/CCL11/FGA/IL1B/CCL7/CCL2/PLA2G2A |
| GO:1904035 | regulation of epithelial cell apoptotic process | 5/145 | 76/17653 | 0.014911 | THBS1/FGG/FGA/IL6/IL13 |
| GO:0002544 | chronic inflammatory response | 3/145 | 19/17653 | 0.017382 | THBS1/CCL11/CXCL13 |
| GO:0071827 | plasma lipoprotein particle organization | 4/145 | 45/17653 | 0.017577 | PLA2G7/GPIHBP1/LIPG/PLA2G2A |
| GO:0051271 | negative regulation of cellular component movement | 10/145 | 340/17653 | 0.017577 | CYP1B1/TIMP1/SEMA6D/THBS1/GREM1/DPEP1/CXCL13/KRT16/CCL2/HAS1 |
| GO:0046427 | positive regulation of JAK-STAT cascade | 5/145 | 80/17653 | 0.017577 | CYP1B1/LIF/OSM/IL6/IL13 |
| GO:1903305 | regulation of regulated secretory pathway | 6/145 | 122/17653 | 0.017577 | CACNA1I/IL13RA2/SYT12/SYT13/C2CD4B/IL13 |
| GO:0006022 | aminoglycan metabolic process | 7/145 | 172/17653 | 0.018129 | CEMIP/CHI3L1/HS3ST2/PGLYRP1/CHIT1/IL1B/HAS1 |
| GO:0032963 | collagen metabolic process | 6/145 | 124/17653 | 0.018129 | COL10A1/MMP9/PRSS2/MMP1/MMP10/IL6 |
| GO:0050729 | positive regulation of inflammatory response | 6/145 | 124/17653 | 0.018129 | PLA2G7/S100A12/OSM/IL1B/PLA2G2A/IL6 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0046697 | decidualizati on | 3/14 5 | 20/1765 3 | 0.01812 9 | LIF/SPP1/GJB2 |
| GO:1904894 | positive regulation of STAT cascade | 5/14 5 | 82/1765 3 | 0.01823 3 | CYP1B1/LIF/OSM/IL6/IL13 |
| GO:0007259 | JAK-STAT cascade | 7/14 5 | 174/176 53 | 0.01823 3 | CYP1B1/LRRTM4/LIF/OSM/CCL2/I L6/IL13 |
| GO:0071622 | regulation of granulocyte chemotaxis | 4/14 5 | 47/1765 3 | 0.01823 3 | CCL19/THBS1/CXCL8/CXCL1 |
| GO:0071825 | protein-lipid complex subunit organization | 4/14 5 | 47/1765 3 | 0.01823 3 | PLA2G7/GPIHBP1/LIPG/PLA2G2A |
| GO:0097696 | STAT cascade | 7/14 5 | 176/176 53 | 0.01923 4 | CYP1B1/LRRTM4/LIF/OSM/CCL2/I L6/IL13 |
| GO:0034114 | regulation of heterotypic cell-cell adhesion | 3/14 5 | 21/1765 3 | 0.01944 7 | FGG/FGA/IL1B |
| GO:1903035 | negative regulation of response to wounding | 5/14 5 | 85/1765 3 | 0.01997 8 | THBS1/SPP1/SERPINB2/FGG/FGA |
| GO:0040013 | negative regulation of locomotion | 10/1 45 | 353/176 53 | 0.01997 8 | CYP1B1/TIMP1/SEMA6D/THBS1/ GREM1/DPEP1/CXCL13/KRT16/C CL2/HAS1 |
| GO:0006874 | cellular calcium ion homeostasis | 11/1 45 | 420/176 53 | 0.02037 1 | CEMIP/CCL19/CD19/P2RX5/SAA1 /CCL11/CXCL13/PROK2/CCL7/BD KRB1/IL13 |
| GO:0048661 | positive regulation of smooth muscle cell proliferation | 5/14 5 | 86/1765 3 | 0.02037 1 | MMP9/THBS1/HPGD/IL6/IL13 |
| GO:1901623 | regulation of lymphocyte chemotaxis | 3/14 5 | 22/1765 3 | 0.02050 6 | CXCL13/CCL7/CCL2 |
| GO:0045807 | positive regulation | 6/14 5 | 131/176 53 | 0.02050 6 | PTX3/CCL19/GREM1/IL1B/CCL2/S ELE |

| | | | | | |
|---|---|---|---|---|---|
| | of endocytosis | | | | |
| GO:2000351 | regulation of endothelial cell apoptotic process | 4/145 | 50/17653 | 0.020506 | THBS1/FGG/FGA/IL13 |
| GO:0007204 | positive regulation of cytosolic calcium ion concentration | 9/145 | 295/17653 | 0.020506 | CEMIP/CCL19/CD19/P2RX5/SAA1/CXCL13/PROK2/BDKRB1/IL13 |
| GO:0030195 | negative regulation of blood coagulation | 4/145 | 51/17653 | 0.021497 | THBS1/SERPINB2/FGG/FGA |
| GO:1900047 | negative regulation of hemostasis | 4/145 | 51/17653 | 0.021497 | THBS1/SERPINB2/FGG/FGA |
| GO:0070374 | positive regulation of ERK1 and ERK2 cascade | 8/145 | 239/17653 | 0.021497 | CHI3L1/CCL19/FGG/CCL11/FGA/CCL7/CCL2/PLA2G2A |
| GO:0043270 | positive regulation of ion transport | 8/145 | 240/17653 | 0.021847 | CEMIP/CD19/P2RX5/CRH/IL1B/BDKRB1/CCL2/IL13 |
| GO:0051043 | regulation of membrane protein ectodomain proteolysis | 3/145 | 23/17653 | 0.022009 | TIMP4/TIMP1/IL1B |
| GO:0055074 | calcium ion homeostasis | 11/145 | 431/17653 | 0.022347 | CEMIP/CCL19/CD19/P2RX5/SAA1/CCL11/CXCL13/PROK2/CCL7/BDKRB1/IL13 |
| GO:0045765 | regulation of angiogenesis | 10/145 | 366/17653 | 0.022729 | CYP1B1/CHI3L1/THBS1/GREM1/CCL11/CXCL13/PROK2/IL1B/CXCL8/IL6 |
| GO:0046879 | hormone secretion | 9/145 | 306/17653 | 0.024227 | BLK/CRH/LIF/SPP1/OSM/FGG/FGA/IL1B/IL6 |

| | | | | | |
|---|---|---|---|---|---|
| GO:1905952 | regulation of lipid localization | 6/145 | 138/17653 | 0.024227 | CRH/THBS1/SPP1/LIPG/IL1B/IL6 |
| GO:0050922 | negative regulation of chemotaxis | 4/145 | 54/17653 | 0.024613 | THBS1/GREM1/CXCL13/CCL2 |
| GO:0032370 | positive regulation of lipid transport | 4/145 | 55/17653 | 0.026101 | CRH/SPP1/LIPG/IL1B |
| GO:0045766 | positive regulation of angiogenesis | 7/145 | 193/17653 | 0.026101 | CYP1B1/CHI3L1/THBS1/GREM1/CCL11/IL1B/CXCL8 |
| GO:0010951 | negative regulation of endopeptidase activity | 8/145 | 251/17653 | 0.026554 | TIMP4/SERPINA3/TIMP1/MMP9/THBS1/DPEP1/SERPINB2/BIRC7 |
| GO:0002637 | regulation of immunoglobulin production | 4/145 | 56/17653 | 0.026855 | IL13RA2/MZB1/IL6/IL13 |
| GO:0050819 | negative regulation of coagulation | 4/145 | 56/17653 | 0.026855 | THBS1/SERPINB2/FGG/FGA |
| GO:0010522 | regulation of calcium ion transport into cytosol | 5/145 | 97/17653 | 0.027626 | CEMIP/CD19/P2RX5/BDKRB1/IL13 |
| GO:0031639 | plasminogen activation | 3/145 | 26/17653 | 0.027626 | THBS1/FGG/FGA |
| GO:0048245 | eosinophil chemotaxis | 3/145 | 26/17653 | 0.027626 | CCL11/CCL7/CCL2 |
| GO:0072577 | endothelial cell apoptotic process | 4/145 | 57/17653 | 0.027626 | THBS1/FGG/FGA/IL13 |
| GO:0060402 | calcium ion transport into cytosol | 6/145 | 145/17653 | 0.02802 | CEMIP/CCL19/CD19/P2RX5/BDKRB1/IL13 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0009914 | hormone transport | 9/145 | 318/17653 | 0.02802 | BLK/CRH/LIF/SPP1/OSM/FGG/FGA/IL1B/IL6 |
| GO:0022409 | positive regulation of cell-cell adhesion | 8/145 | 257/17653 | 0.028345 | CCL19/IL2RA/FGG/CXCL13/FGA/IL1B/CCL2/IL6 |
| GO:1904019 | epithelial cell apoptotic process | 5/145 | 99/17653 | 0.029064 | THBS1/FGG/FGA/IL6/IL13 |
| GO:0000187 | activation of MAPK activity | 6/145 | 147/17653 | 0.029064 | CCL19/THBS1/SAA1/PROK2/IL1B/BIRC7 |
| GO:1902041 | regulation of extrinsic apoptotic signaling pathway via death domain receptors | 4/145 | 59/17653 | 0.02951 | TNFSF10/THBS1/FGG/FGA |
| GO:0072503 | cellular divalent inorganic cation homeostasis | 11/145 | 457/17653 | 0.02951 | CEMIP/CCL19/CD19/P2RX5/SAA1/CCL11/CXCL13/PROK2/CCL7/BDKRB1/IL13 |
| GO:0046883 | regulation of hormone secretion | 8/145 | 261/17653 | 0.029902 | BLK/CRH/LIF/SPP1/OSM/FGG/FGA/IL1B |
| GO:0051480 | regulation of cytosolic calcium ion concentration | 9/145 | 325/17653 | 0.030598 | CEMIP/CCL19/CD19/P2RX5/SAA1/CXCL13/PROK2/BDKRB1/IL13 |
| GO:0045785 | positive regulation of cell adhesion | 10/145 | 392/17653 | 0.030872 | CCL19/PRSS2/IL2RA/FGG/SAA1/CXCL13/FGA/IL1B/CCL2/IL6 |
| GO:0002675 | positive regulation of acute inflammatory response | 3/145 | 28/17653 | 0.031059 | OSM/IL1B/IL6 |
| GO:0010466 | negative regulation of peptidase activity | 8/145 | 264/17653 | 0.031059 | TIMP4/SERPINA3/TIMP1/MMP9/THBS1/DPEP1/SERPINB2/BIRC7 |

| | | | | | |
|---|---|---|---|---|---|
| GO:2001237 | negative regulation of extrinsic apoptotic signaling pathway | 5/145 | 103/17653 | 0.032071 | TNFSF10/THBS1/FGG/FGA/IL1B |
| GO:1901342 | regulation of vasculature development | 10/145 | 396/17653 | 0.032177 | CYP1B1/CHI3L1/THBS1/GREM1/CCL11/CXCL13/PROK2/IL1B/CXCL8/IL6 |
| GO:0031349 | positive regulation of defense response | 11/145 | 466/17653 | 0.032177 | PLA2G7/CRH/S100A12/PGLYRP1/OSM/FGG/FGA/IL1B/TNIP3/PLA2G2A/IL6 |
| GO:0046425 | regulation of JAK-STAT cascade | 6/145 | 154/17653 | 0.033549 | CYP1B1/LRRTM4/LIF/OSM/IL6/IL13 |
| GO:0046888 | negative regulation of hormone secretion | 4/145 | 63/17653 | 0.034609 | CRH/LIF/OSM/IL1B |
| GO:1904892 | regulation of STAT cascade | 6/145 | 156/17653 | 0.035089 | CYP1B1/LRRTM4/LIF/OSM/IL6/IL13 |
| GO:0070663 | regulation of leukocyte proliferation | 7/145 | 212/17653 | 0.035089 | CCL19/GREM1/IL2RA/MZB1/IL1B/IL6/IL13 |
| GO:0072677 | eosinophil migration | 3/145 | 30/17653 | 0.035089 | CCL11/CCL7/CCL2 |
| GO:0090023 | positive regulation of neutrophil chemotaxis | 3/145 | 30/17653 | 0.035089 | CCL19/CXCL8/CXCL1 |
| GO:0060401 | cytosolic calcium ion transport | 6/145 | 157/17653 | 0.035089 | CEMIP/CCL19/CD19/P2RX5/BDKRB1/IL13 |
| GO:0072507 | divalent inorganic cation homeostasis | 11/145 | 475/17653 | 0.035089 | CEMIP/CCL19/CD19/P2RX5/SAA1/CCL11/CXCL13/PROK2/CCL7/BDKRB1/IL13 |
| GO:1904427 | positive regulation of calcium ion transmembr | 4/145 | 65/17653 | 0.03641 | CEMIP/CD19/BDKRB1/IL13 |

| | | | | | |
|---|---|---|---|---|---|
| | ane transport | | | | |
| GO:0002526 | acute inflammatory response | 7/145 | 215/17653 | 0.03641 | SERPINA3/CD19/OSM/SAA1/IL1B/HP/IL6 |
| GO:1904018 | positive regulation of vasculature development | 7/145 | 215/17653 | 0.03641 | CYP1B1/CHI3L1/THBS1/GREM1/CCL11/IL1B/CXCL8 |
| GO:0032368 | regulation of lipid transport | 5/145 | 109/17653 | 0.03641 | CRH/THBS1/SPP1/LIPG/IL1B |
| GO:0001893 | maternal placenta development | 3/145 | 31/17653 | 0.03641 | LIF/SPP1/GJB2 |
| GO:2000352 | negative regulation of endothelial cell apoptotic process | 3/145 | 31/17653 | 0.03641 | FGG/FGA/IL13 |
| GO:0042531 | positive regulation of tyrosine phosphorylation of STAT protein | 4/145 | 67/17653 | 0.039074 | LIF/OSM/IL6/IL13 |
| GO:0010574 | regulation of vascular endothelial growth factor production | 3/145 | 32/17653 | 0.039074 | CYP1B1/IL1B/IL6 |
| GO:0071624 | positive regulation of granulocyte chemotaxis | 3/145 | 32/17653 | 0.039074 | CCL19/CXCL8/CXCL1 |
| GO:0051209 | release of sequestered calcium ion into cytosol | 5/145 | 112/17653 | 0.039442 | CEMIP/CCL19/CD19/BDKRB1/IL13 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0051283 | negative regulation of sequestering of calcium ion | 5/145 | 112/17653 | 0.039442 | CEMIP/CCL19/CD19/BDKRB1/IL13 |
| GO:0051282 | regulation of sequestering of calcium ion | 5/145 | 113/17653 | 0.040711 | CEMIP/CCL19/CD19/BDKRB1/IL13 |
| GO:1902624 | positive regulation of neutrophil migration | 3/145 | 33/17653 | 0.04153 | CCL19/CXCL8/CXCL1 |
| GO:0006026 | aminoglycan catabolic process | 4/145 | 69/17653 | 0.041643 | CEMIP/CHI3L1/PGLYRP1/CHIT1 |
| GO:0045861 | negative regulation of proteolysis | 9/145 | 352/17653 | 0.042054 | TIMP4/SERPINA3/TIMP1/MMP9/IL1R2/THBS1/DPEP1/SERPINB2/BIRC7 |
| GO:0071902 | positive regulation of protein serine/threonine kinase activity | 9/145 | 355/17653 | 0.043476 | CEMIP/CHI3L1/CCL19/S100A12/THBS1/SAA1/PROK2/IL1B/BIRC7 |
| GO:0010573 | vascular endothelial growth factor production | 3/145 | 34/17653 | 0.043476 | CYP1B1/IL1B/IL6 |
| GO:0048246 | macrophage chemotaxis | 3/145 | 34/17653 | 0.043476 | THBS1/SAA1/CCL2 |
| GO:0051208 | sequestering of calcium ion | 5/145 | 116/17653 | 0.043476 | CEMIP/CCL19/CD19/BDKRB1/IL13 |
| GO:0000302 | response to reactive oxygen species | 7/145 | 227/17653 | 0.043777 | CYP1B1/CCL19/MMP9/DPEP1/FOSL1/HP/IL6 |
| GO:0050708 | regulation of protein secretion | 10/145 | 426/17653 | 0.043846 | CCL19/IL1R2/BLK/CRH/FGG/SAA1/FGA/IL1B/IL6/IL13 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0051222 | positive regulation of protein transport | 10/145 | 427/17653 | 0.044295 | CEMIP/CCL19/BLK/CRH/FGG/SAA1/FGA/IL1B/IL6/IL13 |
| GO:0048262 | determination of dorsal/ventral asymmetry | 2/145 | 10/17653 | 0.044988 | SOSTDC1/GREM1 |
| GO:1901072 | glucosamine-containing compound catabolic process | 2/145 | 10/17653 | 0.044988 | CHI3L1/CHIT1 |
| GO:0061045 | negative regulation of wound healing | 4/145 | 72/17653 | 0.044988 | THBS1/SERPINB2/FGG/FGA |
| GO:0090022 | regulation of neutrophil chemotaxis | 3/145 | 35/17653 | 0.044988 | CCL19/CXCL8/CXCL1 |
| GO:0051924 | regulation of calcium ion transport | 7/145 | 230/17653 | 0.044988 | CEMIP/CD19/P2RX5/CRH/BDKRB1/CCL2/IL13 |
| GO:0030212 | hyaluronan metabolic process | 3/145 | 36/17653 | 0.04801 | CEMIP/IL1B/HAS1 |
| GO:0070741 | response to interleukin-6 | 3/145 | 36/17653 | 0.04801 | CHI3L1/FGG/IL6 |
| GO:0002699 | positive regulation of immune effector process | 6/145 | 175/17653 | 0.048872 | CCL19/IL13RA2/MZB1/IL1B/IL6/IL13 |

## APPENDIX C: DIFFERENTIALLY EXPRESSED GENES IN GOLD IV PATIENTS IDENTIFIED BY BOTH DISCOVERY AND VALIDATION SET

| | discovery | | validation | | comparison |
|---|---|---|---|---|---|
| gene | FoldChange | adj.P.Val | FoldChange | adj.P.Val | foldChangeDirection |
| TIMP4 | 3.030766 | 9.60E-10 | 2.44282 | 0.000306 | Same |
| COL10A1 | 3.261507 | 1.62E-09 | 2.938952 | 0.001922 | Same |
| SLC1A3 | 2.111202 | 1.13E-08 | 2.148357 | 0.000142 | Same |
| DAPL1 | 3.796642 | 1.77E-08 | 2.662397 | 0.003054 | Same |
| MMP9 | 2.869956 | 1.77E-08 | 2.739713 | 0.000811 | Same |
| DPYS | 2.446332 | 5.70E-08 | 3.38565 | 1.33E-05 | Same |
| CD19 | 2.968273 | 1.72E-07 | 3.967283 | 5.39E-05 | Same |
| CHRNA1 | 3.343385 | 4.12E-07 | 2.58651 | 0.013068 | Same |
| COMP | 3.125228 | 6.84E-07 | 2.997249 | 0.001287 | Same |
| FCRLA | 3.553258 | 1.26E-06 | 3.500201 | 0.000162 | Same |
| GREM1 | 3.441354 | 2.02E-06 | 2.602995 | 0.001739 | Same |
| SLITRK6 | 0.426058 | 4.58E-06 | 0.410249 | 0.000719 | Same |
| SPP1 | 3.43214 | 8.68E-06 | 2.887195 | 0.000322 | Same |
| CHIT1 | 3.973664 | 3.42E-05 | 2.577022 | 0.040434 | Same |
| RPRML | 2.416566 | 3.83E-05 | 2.230537 | 0.008482 | Same |
| SSTR1 | 0.489822 | 5.33E-05 | 0.408891 | 0.01397 | Same |
| FGG | 4.160678 | 6.74E-05 | 3.232993 | 0.000839 | Same |
| SAA1 | 3.250574 | 6.87E-05 | 2.238221 | 0.007314 | Same |
| CXCL13 | 4.186925 | 8.29E-05 | 3.549575 | 0.002005 | Same |
| KRT17 | 3.330074 | 0.000155 | 2.358718 | 0.018116 | Same |
| FGA | 3.461159 | 0.00019 | 2.091014 | 0.038842 | Same |
| BIRC7 | 2.023296 | 0.000673 | 2.118124 | 0.003321 | Same |
| BHLHE22 | 2.606016 | 2.38E-13 | 2.256409 | 1.65E-05 | Same |
| CA3 | 3.0223 | 9.01E-11 | 3.4648 | 2.35E-05 | Same |
| CCL19 | 3.44047 | 8.99E-09 | 3.052061 | 0.000205 | Same |
| CXCR5 | 2.937334 | 1.12E-07 | 2.97256 | 3.2E-05 | Same |
| HS3ST2 | 2.481425 | 1.35E-07 | 2.589938 | 0.000535 | Same |
| PLA2G7 | 2.704196 | 1.35E-07 | 2.380378 | 0.000452 | Same |
| BLK | 2.900955 | 2.58E-07 | 2.946401 | 7.93E-05 | Same |
| POU2AF1 | 2.714459 | 7.41E-07 | 2.86574 | 4.4E-05 | Same |
| TM4SF19 | 2.671553 | 2.70E-06 | 2.470754 | 0.000561 | Same |

## APPENDIX D: METABOLITES INPUT FOR MAGI

| Compound | inchi key (MAGI input) | searched by MAGI* |
|---|---|---|
| Mycalamide B | AAABMNXUOFPYQK-GRMLSQNXSA-N | No |
| S-(Phenylacetothiohydroximoyl)-L-cysteine | DPHQZNQYOOCWSR-CVZWZAMJSA-N | No |
| Homocysteine | FFFHZYDWPBMWHY-VKHMYHEASA-N | Yes |
| Lophocerine | GEHUGSUAESFIIV-UHFFFAOYSA-N | Yes |
| L-Threonylcarbamoyladenylate | GHLUPQUHEIJRCU-DWVDDHQFSA-N | Yes |
| p-cresol | IWDCLRJOBJJRNH-UHFFFAOYSA-N | Yes |
| N-palmitoyl glycine | KVTFEOAKFFQCCX-UHFFFAOYSA-N | Yes |
| Glycocholic acid | RFDAIACWWDREDC-FRVQLJSFSA-N | Yes |
| SM (d18:1/24:1) | WKZHECFHXLTOLJ-QYKFWSDSSA-N | Yes |
| Decaprenyl phosphate | XBEJBEIXLWRYBT-DJNGBRKISA-N | Yes |
| Ceramide (d18:1/16:0) | YDNKGFDKKRUKPY-TURZORIXSA-N | Yes |
| PS (37:3) | GTQLLUSBBJKGBX-WWYFVLCKSA-N | No |
| PE (38:3) | MSYFXMJDQLPRGO-JTBMWNAQSA-N | No |
| PC(40:6) | TXHZYNSTTCIWMJ-SQKXFFESSA-N | No |
| PC (40:6) (isomer) | KZSXWBSDQDXESU-PPEVQBTRSA-N | Yes |
| PC (32:1) | QIBZFHLFHCIUOT-ZPHPHTNESA-N | No |
| MGDG (36:5) | DRLQFBRXASRGDP-XJYZHZQCSA-N | Yes |
| PE (35:1) | NDZIIJNXYISRGX-YWTUKGCKSA-N | Yes |
| PC (36:4) | NKQPOVROGSWLTO-NVPMBMBWSA-N | Yes |
| PE (36:3) | LMWFNZUKABEGHS-CISNCOODSA-N | Yes |
| PC (34:2) | JLPULHDHAOZNQI-UHFFFAOYSA-O | No |
| SQMG (16:1) | OGARSMGVMPSJRA-FEASXFNBSA-N | Yes |
| PE (34:2) | HBZNVZIRJWODIB-NHCUFCNUSA-N | Yes |
| CL (70:0) | HIPHRGUIEMLVAI-UHFFFAOYSA-N | Yes |
| CL (72:7) | ZGKKGDOIMJSRJH-YHMFKNNJSA-N | No |

*Compounds not in MAGI database were not searched

## APPENDIX E: GENETIC LOCI AND SUSPECTED GENES ASSOCIATED WITH COPD

| Author | Year | Phenotype | Suspected genes from GWAS | SNP |
|---|---|---|---|---|
| Burkart | 2018 | Lung function | KIF25 | rs76656601 |
| | | | ZSWIM7 | rs4791658 |
| | | | HAL | rs145174011 |
| | | | KCNE2 | rs28593428 |
| | | | GPR126 | rs262113 |
| | | | LOC105375250 | rs74444778 |
| | | | GEMIN5 | rs115745680 |
| | | | KIF4B | |
| Hobbs | 2017 | COPD | HHIP | rs13141641 |
| | | | CHRNA5 | rs17486278 |
| | | | HTR4 | rs7733088 |
| | | | ADGRG6 | rs9399401 |
| | | | THSD4 | rs1441358 |
| | | | FAM13A | rs6837671 |
| | | | GSTCD | rs11727735 |
| | | | RIN3 | rs754388 |
| | | | ADAM19 | rs113897301 |
| | | | TET2 | rs2047409 |
| | | | EEFSEC | rs2955083 |
| | | | CFDP1 | rs7186831 |
| | | | TGFB2 | rs10429950 |
| | | | AGER | rs2070600 |
| | | | CCDC101 | rs17707300 |
| | | | ARMC2 | rs2806356 |
| | | | PID1 | rs16825267 |
| | | | DSP | rs2076295 |
| | | | MTCL1 | rs647097 |
| | | | RARB | rs1529672 |
| | | | SFTPD | rs721917 |
| | | | CYP2A6 | rs12459249 |
| Wain | 2017 | Lung function | MFAP2 | rs2284746 |
| | | | LOC101929516<br>PABPC4 | rs17513135 |
| | | | CDC7<br>TGFBR3 | rs1192404 |
| | | | TGFBR3<br>BRDT | rs12140637 |

| | | |
|---|---|---|
| | TGFB2 | rs993925 |
| | RNU5F-1<br>LYPLAL1<br>SLC30A10 | rs4328080 |
| | CHRM3 | rs6688537 |
| | KCNS3 | rs62126408 |
| | PID1<br>SPHKAP | rs10498230 |
| | TRAF3IP1<br>ASB1 | rs61332075 |
| | HDAC4 | rs12477314 |
| | RARB | rs1529672 |
| | CACNA2D3<br>WNT5A | rs1458979 |
| | EEFSEC<br>RUVBL1 | rs2811415 |
| | LOC100507661<br>MECOM | rs56341938 |
| | FAM13A | rs13110699 |
| | FAM13A | rs2045517 |
| | NPNT | rs34712979 |
| | HHIP<br>GYPA<br>AS1 | rs138641402 |
| | ITGA1 | rs1551943 |
| | SPATA9<br>RHOBTB3 | rs153916 |
| | C5orf56 | rs7713065 |
| | CYFIP2<br>ADAM19 | rs10515750 |
| | ADAM19 | rs1990950 |
| | LST1<br>MICB<br>MICA | rs28986170 |
| | NCR3<br>AIF1<br>MICB | rs2857595 |
| | AGER | rs2070600 |
| | KCNQ5 | rs141651520 |
| | ARMC2 | rs2768551 |
| | LOC153910<br>GPR126 | rs7753012 |
| | GPR126 | rs148274477 |

| | | |
|---|---|---|
| | C1GALT1 | rs10246303 |
| | ZKSCAN1 | rs72615157 |
| | PTCH1 | rs16909859 |
| | ASTN2 | rs803923 |
| | CDC123 | rs7090277 |
| | SVIL<br>KIAA1462 | rs3847402 |
| | LRP1 | rs11172113 |
| | FGD6 | rs113745635 |
| | CCDC38 | rs12820313 |
| | LINC01467<br>LINC00911 | rs1698268 |
| | MGA | rs72724130 |
| | THSD4 | rs10851839 |
| | THSD4 | rs12591467 |
| | SH3GL3 | rs66650179 |
| | TEKT5 | rs12149828 |
| | MMP15 | rs12447804 |
| | CFDP1<br>TMEM170A<br>BCAR1 | rs3743609 |
| | EFCAB5<br>CRYBA1<br>SSH2<br>SLC6A4 | rs62070270 |
| | CISD3 | rs11658500 |
| | LTBP4 | rs113473882 |
| | KCNE2 | rs2834440 |
| | AP1S2 | rs7050036 |
| | SPAG17<br>TBX15 | rs200154334 |
| | EFEMP1 | rs1430193 |
| | SUCLG2 | rs1490265 |
| | RSRC1 | rs1595029 |
| | LOC340113<br>TARS | rs91731 |
| | ARL15 | rs2441026 |
| | BMP6 | rs6924424 |
| | LHX3<br>QSOX2 | rs10858246 |
| | DNLZ<br>INPP5E<br>CARD9 | rs10870202 |

| | | |
|---|---|---|
| | MYPN | rs7095607 |
| | HSD17B12 | rs4237643 |
| | PRDM11 | rs2863171 |
| | CCDC91<br>FLJ35252 | rs2348418 |
| | TBX3<br>MED13L | rs35506 |
| | WWOX | rs1079572 |
| | KCNJ2 | rs6501431 |
| | CASC20<br>BMP2 | rs6140050 |
| | ENSA<br>MCL1<br>GOLPH3L | rs6681426 |
| | TNS1 | rs2571445 |
| | MECOM | rs1344555 |
| | TET2 | rs34480284 |
| | GSTCD<br>INTS12<br>NPNT | rs10516526 |
| | HTR4 | rs7715901 |
| | ABLIM3<br>GRPEL2 | rs3839234 |
| | ZKSCAN3<br>LINC01012<br>OR2B2 | rs34864796 |
| | HLA-DQB1<br>HLA-DQA2<br>APOM<br>RNF5 | rs114544105 |
| | HLA-DQB1<br>HLA-DQA2 | rs114229351<br>(rs9275068) |
| | LOC389602<br>LOC285889 | rs12698403 |
| | GLIS3 | rs7872188 |
| | C10orf11 | rs2637254 |
| | AHNAK<br>ROM1<br>EML3<br>MTA2<br>GANAB<br>C11orf83 | rs2509961 |

| | | | | |
|---|---|---|---|---|
| | | | ME3<br>PRSS23 | rs145729347 |
| | | | CDON<br>RPUSD4 | rs567508 |
| | | | MSRB3 | rs1494502 |
| | | | RBM19 | 12:114743533 |
| | | | TBX3 | rs10850377 |
| | | | TRIP11<br>ATXN3 | rs7155279 |
| | | | RIN3 | rs117068593 |
| | | | ARHGAP27<br>ARL17A<br>ARL17B<br>CRHR1<br>FMNL1<br>KANSL1<br>LRRC37A<br>LRRC37A2<br>LRRC37A4<br>MAPT<br>NSF<br>NUDT1<br>PLEKHM1<br>SPPL2C<br>WNT3 | rs35524223 |
| | | | TSEN54<br>CASKIN2 | rs7218675 |
| | | | ZGPAT<br>LIME1 | rs72448466 |
| | | | MICAL3 | rs11704827 |
| | | | MN1 | rs134041 |
| | | | MN1 | rs2283847 |
| Hobbs | 2016 | COPD | CHRNA5 | rs16969968 |
| | | | IL27<br>TUFM | rs181206 |
| | | | CHRNA5 | rs16969968 |
| | | | ADAMTS7 | rs3825807 |
| | | | MMP3 | rs679620 |
| Chen | 2015 | COPD | KLHL7<br>NUPL2 | rs858249 |
| | | | DLG2 | rs286499 |
| de Jong | 2015 | Lung function | PCDH9 | rs17490056 |
| | | | ZMAT4 | rs13278529 |

| | | | | |
|---|---|---|---|---|
| | | | OLIG3 | rs473892 |
| | | | GALNT13 | rs6751439 |
| | | | PDE4D | rs159497 |
| | | | ODZ2 | rs516732 |
| | | | TMEM176A | rs2888674 |
| Wain | 2015 | Lung function | RBM19<br>TBX5 | Chr12:114743533 |
| | | | TET2 | rs2047409 |
| | | | INTS12<br>GSTCD<br>NPNT | rs34712979 |
| | | | HLA-DQB1<br>HLA-DQA2 | rs9274600 |
| | | | KANSL1 | rs2532349 |
| | | | TSEN54<br>GRB2 | rs7218675 |
| | | | MICAL3 | rs11704827 |
| | | | SPAG17<br>TBX15 | rs200154334 |
| | | | DARS<br>CXCR4 | rs35337335 |
| | | | CCDC91 | rs200840970 |
| | | | SLMAP | rs7652294 |
| | | | FERMT1<br>BMP2 | rs979012 |
| Lutz | 2015 | Lung function | CHRNA3<br>AGPHD1<br>CHRNA5<br>PSMA4<br>CHRNB4<br>IREB2<br>ADAMTS7 | rs56077333<br>rs8042849<br>rs17486278<br>rs58365910<br>rs17487223<br>rs17484524<br>rs12913260<br>rs8031948<br>rs2568494 |
| | | | LOC646576<br>HHIP | rs13141641 |
| | | | TGFB2 | rs72738834 |
| | | | FAM13A | rs6837671 |
| | | | MMP3<br>MMP12 | rs72981684<br>rs72981675 |
| | | | RIN3 | rs754388 |
| | | | CYP2A6 | rs56113850 |
| | | | TGFB2 | rs10863398 |

| | | | | |
|---|---|---|---|---|
| | | | LOC646576<br>HHIP | rs138641402 |
| | | | FAM13A | rs6837671 |
| | | | DBH | rs1108581 |
| | | | CYP2A6<br>CYP2A7 | rs56113850 |
| Cho | 2015 | Emphysema | HHIP | rs13141641 |
| | | | CHRNA3<br>CHRNA5<br>IREB2 | rs55676755 |
| | | | AGER | rs2070600 |
| | | | DLC1 | rs75200691 |
| | | | SERPINA10 | rs45505795 |
| | | | DLC1 | rs74834049 |
| | | | HHIP | rs13141641 |
| | | | MIR2054 | rs142200419 |
| | | | AGER | rs2070600 |
| | | | LINC00310<br>KCNE2 | rs55706246 |
| Cho | 2014 | COPD | CHRNA3<br>CHRNA5<br>CHRNB4<br>IREB2<br>PSMA4<br>HYKK | rs12914385<br>rs13180<br>rs3743073 |
| | | | FAM13A | rs4416442 |
| | | | HHIP | rs13141641 |
| | | | RIN3 | rs754388<br>rs11849228 |
| | | | MMP12 | rs626750 |
| | | | TGFB2 | rs4846480 |
| Manichaikul | 2014 | Emphysema | SNRPF<br>CCDC38 | rs7957346 |
| | | | PPT2<br>AGER | rs10947233<br>rs9391855 |
| | | | MAN2B1 | rs10411619 |
| | | | DHX15 | rs7698250 |
| | | | MGAT5B | rs7221059 |
| Castaldi | 2014 | Emphysema | CHRNA5 | rs17486278 |
| | | | HHIP | rs138641402 |
| | | | TGFB2 | rs1690789<br>rs75011710 |
| | | | MMP12 | rs17368659 |

| | | | | |
|---|---|---|---|---|
| | | | CHRNA3 | rs114205691 |
| | | | CYP2A6 | rs56113850<br>rs4560023 |
| | | | MMP12 | rs17368582 |
| | | | TGFB2 | rs1690789 |
| | | | AGPHD1 | rs9788721 |
| | | | MYO1D | rs379123 |
| | | | AGPHD1 | rs11852372 |
| | | | VWA8 | rs9590614 |
| Hansel | 2013 | Lung function decline | ANK3<br>CDK1<br>RHOBTB1<br>TMEM26 | rs7911302<br>rs10761570 |
| | | | FOXA1 | rs177852 |
| Cho | 2012 | COPD | RAB4B<br>EGLN2<br>MIA<br>CYP2A6 | rs7937<br>rs2604894 |
| | | | FAM13A | rs1964516<br>rs7671167 |
| | | | HHIP | rs13141641<br>rs13118928 |
| | | | CHRNA3<br>CHRNA5<br>IREB2 | rs11858836<br>rs13180 |
| Wilk | 2012 | Lung function | AGPHD1<br>IREB2<br>CHRNA5<br>CHRNA3<br>CHRNB4<br>PSMA4 | rs8031948<br>rs16969968 |
| | | | HCP5<br>HLA-A | rs3094013 |
| | | | HTR4 | rs7733088 |
| | | | ADAM19 | rs11744671 |
| | | | RARB | rs1997352 |
| | | | PPAP2B | rs12744110 |
| | | | ADAMTS19 | rs9632471 |
| Kong | 2011 | Emphysema | BICD1 | rs10844154<br>rs161976 |
| Cho | 2010 | COPD | FAM13A | rs7671167<br>rs1903003 |

| | | | | |
|---|---|---|---|---|
| | | | CHRNA3<br>CHRNA5<br>IREB2 | rs1062980 |
| | | | HHIP | rs1828591 |
| Repapi | 2010 | Lung function | HHIP | rs12504628 |
| | | | TNS1 | rs2571445 |
| | | | GSTCD | rs10516526 |
| | | | HTR4 | rs3995090 |
| | | | AGER | rs2070600 |
| | | | THSD4 | rs12899618 |
| Pillai | 2009 | COPD | CHRNA3<br>CHRNA5 | rs8034191<br>rs1051730 |
| | | | HHIP | rs1828591<br>rs13118928 |
| Wilk | 2009 | Lung function | HHIP | rs13147758 |
| Wilk | 2007 | Lung function | GSTO2 | rs156697 |

## APPENDIX F: R VERSION AND RUNNING ENVIRONMENT

```
> sessionInfo()
R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
 [1] parallel  stats4    grid      stats     graphics  grDevices
 [7] utils     datasets  methods   base

other attached packages:
 [1] clusterProfiler_3.10.1 ggalluvial_0.11.3
 [3] limma_3.36.5           annotate_1.58.0
 [5] XML_3.98-1.19          hgug4112a.db_3.2.3
 [7] org.Hs.eg.db_3.6.0     AnnotationDbi_1.44.0
 [9] IRanges_2.14.12        S4Vectors_0.18.3
[11] Biobase_2.40.0         BiocGenerics_0.26.0
[13] tidyr_1.0.2            magrittr_1.5
[15] ggplot2_3.3.0          plyr_1.8.4
[17] VennDiagram_1.6.20     futile.logger_1.4.3
[19] data.table_1.11.8

loaded via a namespace (and not attached):
 [1] bitops_1.0-6        enrichplot_1.2.0     bit64_0.9-7
 [4] progress_1.2.0      RColorBrewer_1.1-2   httr_1.4.1
 [7] UpSetR_1.4.0        tools_3.5.1          R6_2.3.0
[10] DBI_1.0.0           colorspace_1.4-0     withr_2.1.2
[13] prettyunits_1.0.2   tidyselect_1.0.0     gridExtra_2.3
[16] bit_1.1-14          compiler_3.5.1       formatR_1.5
[19] xml2_1.3.2          triebeard_0.3.0      labeling_0.3
[22] scales_1.0.0        ggridges_0.5.1       stringr_1.4.0
[25] digest_0.6.18       DOSE_3.8.2           pkgconfig_2.0.2
[28] rlang_0.4.5         rstudioapi_0.11      RSQLite_2.1.1
[31] gridGraphics_0.4-1  farver_1.1.0         jsonlite_1.6
[34] BiocParallel_1.14.2 GOSemSim_2.8.0       dplyr_0.8.3
[37] RCurl_1.95-4.12     ggplotify_0.0.4      GO.db_3.7.0
[40] Matrix_1.2-14       Rcpp_1.0.2           munsell_0.5.0
[43] viridis_0.5.1       lifecycle_0.2.0      stringi_1.2.4
[46] yaml_2.2.0          ggraph_2.0.0         MASS_7.3-50
[49] qvalue_2.14.1       blob_1.1.1           ggrepel_0.8.0
[52] DO.db_2.9           crayon_1.3.4         lattice_0.20-35
[55] graphlayouts_0.5.0  cowplot_0.9.4        splines_3.5.1
[58] hms_0.5.3           pillar_1.4.4         fgsea_1.8.0
[61] igraph_1.2.4.2      reshape2_1.4.3       futile.options_1.0.1
[64] fastmatch_1.1-0     glue_1.3.0           lambda.r_1.2.3
[67] urltools_1.7.3      vctrs_0.2.4          tweenr_1.0.1
```

```
[70] gtable_0.2.0        purrr_0.3.4         polyclip_1.10-0
[73] assertthat_0.2.0    ggforce_0.3.1       europepmc_0.3
[76] xtable_1.8-3        tidygraph_1.1.2     viridisLite_0.3.0
[79] tibble_3.0.1        rvcheck_0.1.3       memoise_1.1.0
[82] ellipsis_0.3.0
```

## APPENDIX G: NUMBER OF COPD STUDIES INVOLVED WITH DRUGS TARGETING THE 14 TIER 1 DRUGGABLE GENES

| gene | druggability tier | example Drug | No. of COPD Papers on Example Drug | No. of drugs targeting the gene | No. of COPD Papers on All Drugs |
|---|---|---|---|---|---|
| **CTSD** | 1 | Insulin pork | 0 | 3 | S-Methylcysteine:1, Insulin pork:0, N-Aminoethylmorpholine:0 |
| **CXCR4** | 1 | Plerixafor | 4 | 5 | Plerixafor:4, Framycetin:0, AMD-070:0, Mogamulizumab:0, Ibalizumab:0 |
| **FGA** | 1 | Prothrombin | 39 | 17 | Zinc:108, Thrombin:52, Prothrombin:39, Zinc chloride:3, Tenecteplase:2, Alteplase:1, Sucralfate:1, Reteplase:0, Anistreplase:0, Alfimeprase:0, Ancrod:0, EP-2104R:0, Lanoteplase:0, Human Thrombin:0, Thrombin alfa:0, Anti-inhibitor coagulant complex:0, Zinc acetate:0 |
| **FGG** | 1 | Thrombin | 52 | 4 | Thrombin:52, Sucralfate:1, Human Thrombin:0, Thrombin alfa:0 |
| **ICAM1** | 1 | Hyaluronic acid | 50 | 3 | Hyaluronic acid:50, Natalizumab:1, Nafamostat:0 |
| **IL6** | 1 | Dilmapimod | 2 | 13 | Ginseng:18, Andrographolide:6, Dilmapimod:2, Siltuximab:1, Polaprezinc:1, YSIL6:0, VX-702:0, Atiprimod:0, CRx-139:0, Tapinarof:0, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Foreskin fibroblast (neonatal):0, Foreskin keratinocyte (neonatal):0, Binimetinib:0 |
| **MMP1** | 1 | Marimastat | 3 | 1 | Marimastat:3 |
| **MMP13** | 1 | Marimastat | 3 | 7 | Marimastat:3, 3-Methylpyridine:1, Endostatin:1, WAY-151693:0, Hydroxyaminovaline:0, 4-methoxybenzenesulfinate:0, CTS-1027:0 |
| **MMP3** | 1 | Marimastat | 3 | 4 | Marimastat:3, 3-Methylpyridine:1, Hydroxyaminovaline:0, 4-methoxybenzenesulfinate:0 |
| **MMP9** | 1 | Glutathione | 465 | 11 | Glutathione:465, Zinc:108, Captopril:24, Minocycline:11, Glucosamine:10, Marimastat:3, Zinc chloride:3, Endostatin:1, AE-941:0, PG-530742:0, Zinc acetate:0 |
| **PLA2G7** | 1 | Rilapladib | 0 | 1 | Rilapladib:0 |