Blood-Based Transcriptomic and Proteomic Biomarkers of Radiologic Emphysema

Rahul Suryadevara[1], Andrew Gregory[1], Robin Lu[1], Zhonghui Xu[1], Aria Masoomi[2], Sharon M. Lutz[3], Seth Berman[1], Jeong H. Yun[1,4], Aabida Saferali[1], Craig P. Hersh[1,4], Edwin K. Silverman[1,4], Jennifer Dy[2], Katherine Pratte[5], Russel P. Bowler[1,5], Peter J. Castaldi[1,6]*, Adel Boueiz[1,4]* for the COPDGene investigators.

* Equal contribution

[1]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; [2]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA; [3]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA; [4]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; [5]Department of Biostatistics, National Jewish Health, Denver, CO; [5]Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Health, Denver, CO; [6]Division of General Medicine and Primary Care, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

Corresponding Author: Adel Boueiz, Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA, 02115, Email: adel.boueiz@channing.harvard.edu

Authors' email addresses: Rahul Suryadevara (rahul.suryadevara@channing.harvard.edu), Andrew Gregory (andrew.gregory@channing.harvard.edu), Robin Lu (robin.lu@channing.harvard.edu), Zhonghui Xu (zhonghui.xu@channing.harvard.edu), Aria Masoomi (masoomi.a@northeastern.edu), Sharon M. Lutz (sharon.m.lutz@gmail.com), Seth

Berman (seth.berman@channing.harvard.edu), Jeong H. Yun (jeong.yun@channing.harvard.edu), Aabida Saferali (aabida.saferali@channing.harvard.edu), Craig P. Hersh (craig.hersh@channing.harvard.edu), Edwin K. Silverman (ed.silverman@channing.harvard.edu), Jennifer Dy (jdy@ece.neu.edu), Katherine Pratte (prattek@njhealth.org), Russel P. Bowler (bowlerr@njhealth.org), Peter J. Castaldi (peter.castaldi@channing.harvard.edu), Adel Boueiz (adel.boueiz@channing.harvard.edu)

Author Contributions:

Drs. Boueiz and Castaldi had full access to all the data in the study, take responsibility for the integrity of the data and the accuracy of the data analysis, had authority over manuscript preparation and the decision to submit the manuscript for publication.

Study concept and design: Boueiz, Castaldi, Xu

Acquisition, analysis, or interpretation of data: All authors

Drafting of the manuscript: Suryadevara, Boueiz, Castaldi

Critical revision of the manuscript for important intellectual content: All authors

Statistical analysis: Xu, Lutz, Castaldi, Boueiz

Obtained funding: Boueiz, Castaldi, Silverman

Study supervision: All authors

All authors gave final approval of the version to be published.

Funding Sources:

This work was supported by NHLBI K08HL141601, K08HL146972, K08HL136928, K01HL157613, R01HL124233, U01 HL089897, R01 HL147326, and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through

contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion.

Running Head: Emphysema blood biomarkers

Descriptor: 9.03 COPD: Clinical Phenotypes

Manuscript Word Count: 4,404/3,500

AT A GLANCE COMMENTARY (200/200 words)

Scientific Knowledge on the Subject:

Emphysema may have unidentified treatment targets due to a distinct pathophysiology from other forms of COPD. Blood-based biomarkers may facilitate the identification of emphysema in smokers and reveal key therapeutic targets. Differential gene expression and protein analyses have uncovered some of the molecular underpinnings of emphysema. However, no studies have assessed the alternative splicing mechanisms and analyzed data from the recently developed high throughput panels. In addition, although emphysema has been associated with low body mass index (BMI), it is still unclear how BMI affects the transcriptome and proteome of the disease. Finally, the effectiveness of multi-omic biomarkers in determining the severity of emphysema has not yet been investigated.

What This Study Adds to the Field:

We performed whole-blood genome-wide RNA sequencing and plasma SomaScan proteomic analyses in a large, well-phenotyped cohort of smokers. In addition to confirming earlier findings, our differential gene expression, alternative splicing, and protein analyses

identified novel emphysema biomarkers. Our mediation analysis detected varying degrees of transcriptomic and proteomic mediation due to BMI and assisted in differentiating whether pathways are primarily affected by emphysema or BMI. Finally, our supervised machine learning emphysema prediction modeling demonstrated the utility of incorporating multi-omic data.

Keywords: Emphysema; Biomarkers; Transcriptomics; Proteomics; Prediction

ABSTRACT

Rationale: Emphysema is a key component of COPD with important prognostic implications. Identifying blood-based biomarkers of emphysema will facilitate early diagnosis and possible development of targeted therapies.

Objectives: Discover blood transcriptomic and proteomic biomarkers for chest computed tomography-quantified emphysema in smokers and develop predictive biomarker panels.

Methods: Emphysema blood biomarker discovery was performed using differential gene expression, alternative splicing, and protein association analyses in a training set of 2,370 COPDGene participants with available whole blood RNA sequencing, plasma SomaScan proteomics, and clinical data. Validation was conducted in a testing set of 1,016 COPDGene subjects. Since body mass index (BMI) and emphysema often co-occur, we performed a mediation analysis to quantify the effect of BMI on gene and protein associations with emphysema. Predictive models were also developed using elastic net to predict quantitative emphysema from cell blood count, RNA sequencing, and proteomic biomarkers. Model accuracy was assessed by area under the receiver-operator-characteristic-curves (AUROC) for subjects stratified into tertiles of emphysema severity.

Measurements and Main Results: 4,913 genes, 1,478 isoforms, 386 exons, and 881 proteins were significantly associated with emphysema *(FDR 10%)*. 75% and 77% of genes and proteins, respectively, were mediated by BMI. The significantly enriched biological pathways were involved in inflammation and cell differentiation, differing between the most

and least BMI-mediated genes. The cell blood count plus protein model achieved the highest performance with an AUROC of 0.89.

Conclusions: Blood transcriptome and proteome-wide analyses reveal key biological pathways of emphysema and enhance the prediction of emphysema.

Abstract word count: 246/250

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality (1). Emphysema, a major COPD phenotype, has been independently associated with an increased risk for cardiovascular disease, lung cancer, and mortality (2-4). While some progress has been made in treating COPD, much work remains in identifying therapeutic targets specifically for emphysema (5). Furthermore, timely diagnosis calls for a blood-based predictive model as it may identify emphysema in subjects where computed tomography (CT) scans are not clinically indicated. An emphysema blood biomarker would also overcome the issues of radiation exposure and false positive findings associated with CT scans (6, 7). In addition, early disease biomarkers and a stronger understanding of the molecular bases of emphysema are needed to develop novel personalized therapies to improve the prognosis of affected individuals (2, 8, 9).

Previous transcriptomic studies have identified emphysema-associated genes (such as *COL6A1*, *CD19*, *PTX3*, and *RAGE*) and biological processes (such as innate and adaptive immunity, inflammation, and tissue remodeling) (7, 10-14). However, most studies to date have not evaluated emphysema-associated alternative splicing mechanisms. Alternative splicing, the regulatory process in which multi-exon human genes are expressed in multiple transcript isoforms, has been implicated in the pathophysiology of several lung diseases such as asthma, pulmonary fibrosis, pulmonary arterial hypertension, and COPD (15-21). Protein levels have also been studied for potential emphysema biomarker identification and it was found that sRAGE, ICAM1, CCL20, and adiponectin levels in blood and eotaxin levels in bronchoalveolar lavage fluid are associated with emphysema severity (22-25), though the protein panels used for these studies included fewer proteins than more recently developed panels (26, 27). Finally, previous research that used blood-based emphysema prognostic models had small sample sizes and only tested one -omic modality at a time (22, 26-29).

We used whole-blood genome-wide RNA sequencing (RNA-seq) and plasma SomaScan proteomic data from a well-phenotyped cohort of current and former smokers of the COPDGene study to determine the associations of genes, alternative splicing, and proteins with CT-quantified emphysema. Given the high clinical correlation between emphysema and BMI (30), we also performed a mediation analysis to understand the influence of BMI on emphysema-associated genes and proteins. Finally, we developed machine learning predictive models for emphysema using transcriptomic and proteomic biomarkers. We hypothesized that transcriptomic and proteomic characterization of smokers would elucidate emphysema pathobiology and yield novel disease biomarkers. We also hypothesized that most differentially expressed genes and proteins would be mediated by BMI and enriched for a distinct set of biological processes relative to genes not mediated by BMI. Lastly, a multi-omic prediction model might distinguish between smokers with low and high emphysema severity. Some of these results have been previously reported as an abstract (31).

METHODS

Study description

Participants were recruited from the COPDGene study (NCT00608764, www.copdgene.org), a longitudinal study investigating the genetic basis of COPD. The COPDGene population consists of 10,371 non-Hispanic white and African-American smokers 45-80 years old, with at least ten pack-years of lifetime cigarette smoking history (32). Subjects had varying degrees of COPD severity, as measured by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) grading system. COPDGene obtained 5-year follow-up data and is currently obtaining 10-year follow-up data of available subjects.

Questionnaires, chest CT scans, and spirometry have been gathered at 21 clinical facilities in the United States. In addition, whole blood genome-wide RNA-seq and plasma proteomic measurements were obtained from a subset of subjects at their 5-year follow-up visit (Visit 2). Each center acquired institutional review board approval and written informed consents. In our analyses, we used the COPDGene Visit 2 data, which included RNA-seq and SomaScan plasma proteomic data.

Emphysema quantification

Using the Thirona software (www.thirona.eu), radiologic emphysema was quantified as the Hounsfield units (HU) at the 15th percentile of CT density histogram at total lung capacity, corrected for the inspiratory depth variations (adjusted Perc15 density) (33, 34). Adjusted Perc15 density values are reported as the HU + 1,000. The lower the adjusted Perc15 values are, i.e., the closer to -1,000 HU, the more CT-quantified emphysema is present.

Training and testing samples

We randomly partitioned our studied cohort into training and testing samples comprising 70% and 30% of the subjects, respectively. All association and mediation analyses, as well as prediction model training, were conducted in the training data. Validation was carried out in the testing sample.

RNA isolation, library preparation, filtering, and normalization

Illumina sequencers were utilized to obtain gene, isoform, and exon counts from total blood RNA isolated from Visit 2 participants. Genomic features with very low expression (average counts per million (CPM) < 0.2 or number of subjects with CPM < 0.5 less than 50)

or extremely highly expressed genes (number of subjects with CPM > 50,000 less than 50) were filtered out prior to applying trimmed mean of M values normalization from edgeR (v3.24.3), which accounts for differences in sequencing depth (35). Counts were transformed to log2 CPM values and quantile-normalized to further remove systematic noise from the data.

Protein measurements and filtering

At Visit 2, plasma samples were assayed for 4,979 proteins in 6,018 COPDGene participants using the SomaScan Human Plasma 5.0K assay, a multiplex aptamer-based assay (SomaLogic, Boulder, Colorado) (36). The SomaScan data was standardized per the SomaLogic protocol to control for inter-assay variation between analytes and batch differences between plates. Samples with low volume, failed hybridization control, or failed dilution scale were removed. Protein counts were also transformed to log2 CPM values.

RNA-seq differential expression, usage, and protein association analyses

We used the limma-voom linear modeling approach (as implemented in limma v3.38.3) to test for the associations between emphysema and whole blood RNA transcripts (37, 38). The diffSplice function from limma was used to test for differential usage of isoforms and exons. While differential expression refers to the change in the *absolute* expression levels of a feature, differential usage captures alternative splicing and refers to the change in the *relative* expression levels of the isoforms/exons within a given gene. The associations of the SomaScan proteins with emphysema were tested using multivariable linear modeling. In the emphysema "primary" model, we adjusted for age, race, sex, pack-years of smoking, current smoking status, forced expiratory volume in one second (FEV_1), cell blood count (CBC) proportions, CT scanner model, and library preparation batch for

RNA-seq / clinical center for proteins. The validation rate in the testing sample was determined based on a threshold P-value < 0.1 and a consistent direction of effect in the training and testing datasets. In the emphysema plus BMI model performed on just genes and proteins ("sensitivity analysis"), we added BMI to the list of covariates. To select biomarkers for inclusion in the prediction model, we ran additional models only adjusted for the CT scanner model and library preparation batch for RNA-seq / clinical center for proteins. Multiple comparisons were corrected with the Benjamini-Hochberg method using a threshold of significance of a false discovery rate (FDR) of 10% (39).

Mediation analysis

We conducted a mediation analysis to distinguish how much of the effect of emphysema on gene expression acted through BMI (referred to as the indirect effect) and how much of the effect of emphysema directly influenced gene expression (referred to as the direct effect). The medflex R package v0.6-7 was employed (40). The analysis was performed both on the genes and proteins with statistically significant total effects (the sum of the indirect and direct effects) from the emphysema model without BMI adjustment. A mediated proportion representing the ratio of the indirect effect over the total effect was computed for each gene.

Gene set enrichment analyses

The biological enrichment of the gene sets derived from the gene expression, transcript usage, and protein association analyses was evaluated using the topGO (v2.33.1) weight01 algorithm, which accounts for the dependency in the Gene Ontology (GO) topology (41). GO enrichment analysis was also conducted on the top 250 most mediated (i.e., with the lowest significant indirect effect FDR) and least mediated (i.e., with the lowest significant

direct effect FDR and mediated proportions between -1.2 and 0.2) genes and proteins. We only reported GO pathways with at least three significant genes and an adjusted P-value < 0.005.

Development of predictive models

To predict CT-quantified emphysema, we constructed supervised elastic net models. Elastic net offers several well-known benefits, including the ability to account for multi-collinear features and avoid overfitting (42). The outcome variable was the adjusted Perc15 density. The predictors were the RNA-seq and proteins that reached statistical significance in the transcriptomic and proteomic association analyses (adjusted only for the scanner model and library preparation batch or clinical center). We first used CBC with either genes, isoforms, or exons as predictors. The highest-performing RNA-seq data type was utilized to build a second set of prediction models (CBC only, CBC + RNA-seq, CBC + proteins, and CBC + RNA-seq + proteins) with or without readily available clinical variables (age, BMI, sex, and race). The outcome and the predictors were centered and scaled. The models were trained using 10-fold cross-validation, minimizing the mean squared error (MSE) (43) on the left-out fold. After model training on the continuous emphysema variable, we classified subjects into tertiles of adjusted Perc15 density. We evaluated the predictive performances of the models using R^2 for the continuous emphysema and the area-under-receiver-operator-characteristic curve (AUROC) for the model accuracy to distinguish those in the highest and lowest tertiles of emphysema severity. We compared AUROCs with the DeLong test using the pROC R package (44). Finally, predictors were ranked by the absolute values of their coefficients from the regression model.

Statistical analysis

Data were reported as mean with standard deviations or counts with percentages. Continuous variables were tested with Kruskal-Wallis and categorical variables with chi-square. Upregulated versus downregulated genes as well as positive versus negative signs of the protein coefficients are provided with respect to their relationships with adjusted Perc15 density (i.e., they have opposite directions for their associations with emphysema).

Additional methods are available in the Supplement.

RESULTS

Subject characteristics

3,386 subjects from COPDGene Visit 2 with complete RNA-seq, protein, and clinical data necessary were included in our analyses (Figure 1). As shown in Table 1, the included subjects were mostly non-Hispanic whites with a balanced representation by sex, a mean age of 65, a mean BMI of 29, and a mean of 41 pack-years of smoking. The subjects' characteristics did not significantly differ between the training and testing data, which consisted of 2,370 and 1,016 subjects, respectively. A comparison of subjects with and without missing data showed that the two groups were largely similar in characteristics (Table E1). A schematic overview of the analyses performed is illustrated in Figure 2.

Differential gene expression analysis

We performed differential gene expression (DGE) analysis on the gene level RNA-seq counts obtained from the 2,370 subjects of the training dataset. 4,913 out of 19,177 genes reached significance at 10% FDR for CT-quantified emphysema (Table E2). 2,339 genes

were up-regulated, and 2,574 were down-regulated (Figure 3A). The GO enrichment analysis performed on the differentially expressed genes identified 44 significantly enriched biological processes, including neutrophil degranulation, regulation of NF-kappaB (NF-κB) signaling, viral transcription, T cell proliferation, and regulation of tumor necrosis factor (TNF)-mediated signaling pathway (Table 3, E3).

Differential isoform and exon usage analyses

We next performed differential isoform usage (DIU) and differential exon usage (DEU) analyses on the training dataset to investigate the changes in relative isoform and exon levels within single parent genes. Out of 78,837 isoforms and 209,707 exons tested, 1,478 isoforms and 368 exons reached significance (*FDR 10%*), respectively (Table 2). The differentially used isoforms (DUIs) mapped to 1,209 individual genes; 45% of these genes (542/1,209) were also identified in the DGE analysis (Table E4). The differentially used exons (DUEs) mapped to 251 genes (Table E6); 68% of these genes (171/251) were also differentially expressed. 788 isoforms and 142 exons were up-regulated. 690 isoforms and 244 exons were down-regulated (Figure 3B, 3C). The GO enrichment analyses performed on the DUIs and DUEs yielded 35 and 13 significantly enriched biological processes, respectively. Top processes included autophagy of the mitochondrion, regulation of NF-κB signaling, negative regulation of wingless-related integration site (WNT) signaling, and viral transcription (Table 3, E5, E7).

Protein association analysis

We tested 4,979 proteins measured with the SomaScan v2 panel in the training dataset using multivariate linear regression modeling. 18% (881/4,979) of the evaluated proteins were associated with emphysema (*FDR 10%*) (Table E8, Figure E1). From the GO

enrichment analyses performed on the proteins, we found 17 significantly enriched biological processes (Table E9). The top enriched pathways were related to complement activation, classical pathway, and WNT-signaling. Figure 4 summarizes the overlap of the biomarkers and GO terms between DGE, DIU, DEU, and protein analyses, showing that over 90% of the total reported biomarkers and GO terms are unique to each individual analysis.

Validation analyses

We analyzed 1,016 subjects with RNA-seq and proteomic data in the testing samples to provide independent validation of the emphysema biomarkers identified in the training sample. We observed that the effect sizes were highly correlated between training and testing DGE, DEU, and protein analyses (Pearson's r = 0.80, 0.86, and 0.88, respectively). A lower correlation (r = 0.29) was observed in the DIU analysis. We further determined whether biomarkers were validated by using a threshold of (testing) P-value < 0.1 coupled with whether the training and testing data had a consistent direction of effect. 46% (2,252/4,913), 30% (449/1,478), 60% (233/368), and 47% (416/881) of the DGE, DIU, DEU, and protein biomarkers, respectively, were validated (Tables E2, E4, E6, and E8).

Mediation analysis

Since severe emphysema is often associated with low BMI, we performed sensitivity analyses that also adjusted for BMI in our transcriptomic and proteomic analyses. We observed that 96% (4,728/4,913) of the differentially expressed genes and 80% (703/881) of the proteins (Figure E2) associated with emphysema from the primary analysis were no longer significant after adjustment for BMI, suggesting that BMI mediates many of the emphysema-associated transcriptomic and proteomic changes. BMI may therefore be involved in the causal pathway linking genes and proteins to emphysema. To investigate this,

we performed mediation analysis to compare the direct effect of emphysema on genes and proteins and the indirect effect of emphysema on genes and proteins that are mediated by BMI, as visualized by the directed acyclic graph (DAG) in Figure 2. The analyses were performed on the 4,913 differentially expressed genes and proteins from the association analyses without BMI adjustment. We found that 70% of genes (3,456/4,913) and 61% of proteins (537/881) showed evidence of mediation with a significant indirect effect and no significant direct effect. 229 genes and 138 proteins had significant direct and indirect effects, 234 genes and 103 proteins had significant direct effects only, and 994 genes and 103 proteins had no significant effect in the mediation analysis (Tables E12 and E13).

The top 250 most and least mediated genes and proteins were analyzed for shared and unique biological pathways. Most of the pathways significantly enriched for the least mediated genes were related to the immune response, such as interferon-gamma production and chemokine-mediated signaling pathway. The pathways significantly enriched for the most mediated genes included immune and iron-related pathways such as iron ion homeostasis and the protoporphyrinogen IX (PPIX) metabolic process (Table E14). No pathways were enriched for the top 250 and bottom 250 mediated proteins at the P-value < 0.005 threshold.

Prediction

To develop predictive models for emphysema using blood biomarkers, we performed association analyses in the training dataset, adjusting only for technical factors (CT scanner model and library preparation batch for transcriptomic or clinical center for proteomic). 13,066 genes, 4,254 isoforms, 2,263 exons, and 1,719 proteins reached significance (*FDR 10%)*. To evaluate whether gene expression data was more informative at the gene, isoform, or exon level, we trained three models in the training sample (CBC + gene, CBC + isoform,

and CBC + exon). The AUROC were 0.80, 0.70, and 0.76, respectively (Table E17 and Figure E4). Accordingly, we focused on gene-level quantifications exclusively for the subsequent models. CBC, CBC + gene, CBC + protein, and CBC + gene + protein elastic net models were run along with a set of models also using clinical predictors. The adjusted Perc15 density was then classified into tertiles (Figure E3), and the ability of the predictive models to distinguish subjects in the highest and lowest tertiles was assessed in the testing sample.

The model using only CBC achieved an AUROC of 0.64. Adding genes to this model improved the performance to an AUROC of 0.80 *(DeLong P-value 0.05)*. However, the performance was even better when the protein data was added to the CBC-only model (AUROC 0.89, *DeLong P-value 0.05*). Adding both gene and protein data to CBC gave an AUROC of only 0.87, suggesting that gene data do not provide additional predictive information to the protein data. Figure 5 summarizes the model results, and Table E15 summarizes each model's AUROC, alpha, and L1 parameters. Each elastic net model repeated with clinical predictors had a higher AUROC than the corresponding models without. However, they did not impact the ranking of the model performances (i.e., CBC + protein remained the highest-performing model) (Figure E5).

Ranked by absolute beta coefficients, the top-10 predictors of the all-inclusive model included sRAGE (soluble receptor for advanced glycation end products) and biomarkers that have not been previously connected to emphysema: *MIR124-1HG* (MIR124-1 Host Gene) and PSMP (PC3-secreted microprotein) (Figure 6).

DISCUSSION

In this study, we performed the largest blood transcriptomic and proteomic profiling of CT-quantified emphysema to date, including investigations into the alternative splicing

mechanisms of emphysema. We uncovered thousands of biomarker associations. The biological relevance of these findings was assessed through GO pathway analyses, which demonstrated enrichment for inflammatory pathways such as neutrophil degranulation as well as those involved in cell differentiation such as NF-κB and WNT signaling. The mediation analysis revealed that 70% of differentially expressed genes and 61% of associated proteins in our emphysema cohort are mediated through BMI, shedding light on distinct biological pathways associated with the mediated or non-mediated genes and proteins. We also showed that prediction models using blood biomarkers achieve high accuracy in discriminating between smokers with substantial versus mild emphysema.

A growing but limited number of studies have examined emphysema biomarkers and biological pathways. The extracellular matrix (ECM), NF-κB, transforming growth factor beta (TGF-ß), B cell antigen receptor (BCR), and oxidative phosphorylation pathways are among the most reported in these studies (10, 45, 46). However, although researched from various sources, including peripheral blood, lung tissue, and sputum, most identified pathways and biomarkers originate from studying a single 'omics modality at a time (22, 27, 47). Furthermore, the alternative splicing mechanisms of emphysema have not been examined extensively. Our investigations examined the blood genes, isoforms, exons, and proteins, confirming many of the known emphysema-associated pathways and revealing additional ones that may also be at play.

Systemic inflammation and immunological dysfunction due to noxious particle exposure potentiate alveolar damage in emphysema and increase susceptibility to viral infections (47-52). Our pathway analysis identified neutrophil degranulation and the TNF pathway, both emphysema-related inflammatory signals previously implicated in COPD and murine models of emphysema (53-56). Additionally, our pathway analysis revealed T cell proliferation, which aligns with studies that have correlated the number of T cells with the

level of alveolar damage in COPD (57). Due to its enhancement of degranulation and cytokine release, the C5 complement factor has been linked to emphysematous changes (58, 59). Although we did not look into particular complement components, our pathway analysis revealed that the classical complement pathway is enriched for emphysema. C5 can be activated through the classical pathway, supporting its putative role in emphysema.

NF-κB promotes innate immune and T cell differentiation, suppresses apoptosis, and enhances pro-inflammatory genes (60). Higher amounts of the NF-κB p65 subunit protein have been found in sputum samples and bronchial biopsies of COPD patients compared to controls (61, 62) and have been implicated in emphysema pathogenesis (63, 64). This is supported by our data. Our pathway analysis also corroborates prior literature on the possible role of mitochondrial autophagy (mitophagy) in emphysema (65, 66). The canonical WNT signaling pathway, which maintains tissue homeostasis and regulates cell differentiation and apoptosis (67), is a known activator of mitochondrial biogenesis (68). According to prior research, WNT signaling is inactive in emphysematous lung tissue (69, 70), as supported by its downregulation in our analysis.

Severe emphysema is known to be associated with low BMI, muscle wasting, and cachexia (71-73). Most of the transcriptomic and proteomic associations with emphysema in our study were sensitive to the adjustment for BMI, suggesting that BMI may be involved in the causal pathways underlying these associations. According to a recent study, COPD patients' cachexia may be influenced by impaired heme biosynthesis, which leads to excess iron accumulation and oxidative tissue damage (74). Interestingly, the PPIX metabolic pathway, which is directly involved in heme biosynthesis (75), and iron ion homeostasis were enriched in our top 250 most mediated genes. On the other hand, the top 250 least mediated genes (i.e., those less influenced by BMI) were more involved with immune response-related than iron-related pathways.

CT scan is the best currently available non-invasive method for detecting emphysema. However, CT has several drawbacks, including increased costs, radiation exposure, and high rates of unrelated false-positive findings (76). Accurate risk prediction tools that use the best available data sources to stratify patients based on their specific risk profiles could help with more efficient early and targeted interventions. Until recently, such prediction models were only created using data from a single 'omics type with or without standard clinical features (77-80). As the first study to utilize gene, alternative splicing, and protein predictors combined with CBC, we developed models that could classify upper and lower tertiles of emphysema severity with reasonable accuracy. While alternative splicing predictors were worth exploring, gene data had a higher AUROC and the highest number of features selected. While genes outperformed clinical and CBC features, protein predictors yielded the best AUROC across all models.

From the top 10 predictors of the CBC + gene + protein model, sRAGE, which minimizes tissue injury and inflammation, has consistently been recognized as a candidate emphysema biomarker (7, 12, 81, 82). Even though its function is not fully known, PSMP has been implicated in inflammation and cancer development (77). The putative role and function of PSMP in emphysema require further investigation. Also not previously connected to emphysema, *MIR124-1HG* is involved in the sensory perception of sound by modulating inner ear stem cell growth (83). Bayat et al. concluded that COPD patients had more hearing loss than control patients (84). However, the speculated mechanisms leading to hearing impairment, such as brain hypoxia (85) and increased inflammatory cytokines (86), require further investigation. Another study found that miR-124 (microRNA of *MIR124-1HG* host gene) regulates the sensory region of the cochlea by targeting inhibitors of the WNT signaling pathway (83).

This study has several strengths. Our findings come from a large, well-phenotyped cohort of smokers. This is the first study that, to our knowledge, has looked at alternative splicing mechanisms in emphysema in addition to differential gene expression and protein association analyses. As a result, we were able to contrast the various biological pathway enrichments, discovering new emphysema mechanisms and support existing ones. Additionally, we performed validation analyses of our reported biomarkers. In order to understand how BMI impacts emphysema, we also conducted a mediation analysis that allowed us to assess the contribution of the most and least mediated genes and proteins to the enriched biological pathways. Finally, our multi-omic approach model enabled us to construct prediction models that accurately discriminate between more severe and less severe emphysema.

This study also has several limitations. CBC quantifications do not capture the variability of the immune cell subpopulations, which limits the ability to localize these effects to specific cell types. Our results could, therefore, partially be due to cell subpopulations not represented in the CBC quantifications. Future studies may address this by using single-cell data. Next, the mediation analysis is based on the following assumptions: no unmeasured confounding of the emphysema-BMI-gene expression/protein level relationship, no measurement error for the exposure or mediator, and the arrows in the DAG are correctly specified. However, the specification of the DAG is reasonable based on prior medical knowledge. In addition, while mediation and subsequent pathway analyses were able to detect interesting biological pathways that may be involved in disease pathogenesis, such analyses are only hypothesis-generating and require functional confirmation. Lastly, the large sample size for our primary analysis reduces the risk of false positive associations, but further validation of these results in comparable cohorts will provide greater confidence in these associations.

CONCLUSION

Collectively, our transcriptomic and proteomic analyses illustrated the inflammatory and cell differentiation pathways leading to emphysematous changes in addition to identifying novel biomarkers predictive of emphysema. While not ready to be used for clinical practice, our prediction model opens the possibility of assessing emphysema severity in the clinical setting using a minimally invasive blood sample. This could inform patient enrollment in clinical trials and minimize radiation exposure. Future work is required to compare blood and lung tissue biomarkers, understand how they change as emphysema progresses, and evaluate the impact of implementing the developed predictive models to personalize and improve patient care.

Table 1. Characteristics of subjects in the training and testing datasets in COPDGene Visit 2.

	Training (N = 2,370)	Testing (N = 1,016)	P-value
Age	65.07 (8.78)	65.42 (8.85)	0.28
Sex, % male	51.35%	49.80%	0.41
Race, % NHW	72.87%	76.18%	0.04
BMI	28.94 (6.33)	28.70 (6.01)	0.31
Smoking pack-years	41.65 (25.72)	41.23 (25.90)	0.66
Current Smoker	858 (36.20%)	346 (34.06%)	0.23
FEV_1 (mL)	2.22 (0.84)	2.21 (0.85)	0.75
FEV_1, % predicted	80.57 (24.35)	80.52 (24.33)	0.95
FVC (mL)	3.20 (0.95)	3.19 (0.96)	0.88
Bronchodilator responsiveness (FVC, % predicted)	2.99 (9.67)	3.25 (10.65)	0.50
Adjusted Perc15 density	85.96 (24.8)	85.72 (24.91)	0.79
% Segmental airway wall thickness	49.70 (8.37)	49.54 (8.37)	0.62
Gas trapping	19.81 (18.37)	19.78 (18.59)	0.97
Pi10	2.24 (0.57)	2.23 (0.56)	0.54
GOLD grade			
PRISm	299 (12.62%)	120 (11.81%)	
0	996 (42.03%)	415 (40.85%)	
1	232 (9.79%)	100 (9.84%)	0.89
2	425 (17.93%)	187 (18.41%)	
3	209 (8.82%)	95 (9.35%)	
4	84 (3.54%)	35 (3.44%)	
Exacerbation frequency	152 (14.96%)	356 (15.03%)	0.73
Severe exacerbation frequency	185 (7.81%)	78 (7.68%)	0.90
SGRQ score	24.94 (24.35)	24.40 (23.11)	0.55
MMRC dyspnea score			
0	1294 (54.60%)	562 (55.31%)	
1	284 (11.98%)	133 (13.09%)	
2	276 (11.65%)	117 (11.52%)	0.78
3	361 (15.23%)	144 (14.17%)	
4	155 (6.54%)	60 (5.91%)	
CAD	199 (8.40%)	74 (7.28%)	0.28
Diabetes	383 (16.16%)	143 (14.07%)	0.12
Hypertension	1136 (47.93%)	508 (50.00%)	0.27

Continuous variables are expressed as means and standard deviations. Categorical variables are expressed as absolute values and/or percentages. Participant characteristics reported here are from Visit 2, when 'omics data were obtained.

Adjusted Perc15 density: Hounsfield units at the 15th percentile of CT density histogram at total lung capacity, corrected for the inspiratory depth (per convention, adjusted Perc15 density values are reported as the HU + 1000); Bronchodilator responsiveness (FVC, % predicted): Percentage of subjects with post-

bronchodilator increase in FVC of at least 12% from baseline; CAD: Coronary Artery disease; Exacerbation history: At least one COPD exacerbation (acute worsening of respiratory symptoms that required systemic steroids and/or antibiotics) in the previous year; FEV_1: Forced expiratory volume in one second; GOLD: Global Initiative for Chronic Obstructive Lung Disease; GOLD 0: Normal spirometry (defined as post-bronchodilator $FEV_1/FVC \geq 0.7$ and $FEV_1 \geq 80\%$ predicted); GOLD 1: $FEV_1/FVC < 0.70$ and post-bronchodilator $FEV_1 \geq 80\%$ predicted; GOLD 2: $FEV_1/FVC < 0.70$ and post-bronchodilator FEV_1 50-79% predicted; GOLD 3: $FEV_1/FVC < 0.70$ and post-bronchodilator FEV_1 30-49% predicted; GOLD 4: FEV1/FVC < 0.70 and post-bronchodilator $FEV_1 < 30\%$ predicted; MMRC: Modified medical research council dyspnea scoring system; Pi10: Square root of the wall area of a hypothetical airway of a 10-mm internal perimeter; PRISm: Preserved Ratio Impaired Spirometry (defined as $FEV_1/FVC \geq 0.70$ but with $FEV_1 < 80\%$ predicted); Race: Self-reports as either non-Hispanic white (NHW) or African American; Severe exacerbation history: COPD exacerbation requiring an emergency department visit or hospital admission; SGRQ: St. George's Respiratory Questionnaire.

Table 2. Top 5 differentially expressed genes (DGE), differentially used isoforms (DIU), and differentially used exons (DEU) associated to adjusted Perc15 density.

	ID	HUGO Gene Name	Log Fold Change	Average Log Expression	FDR
DGE	ENSG00000160179	*ABCG1*	-0.007	4.907	$4x10^{-19}$
	ENSG00000138772	*ANXA3*	0.006	4.825	$8x10^{-17}$
	ENSG00000164674	*SYTL3*	-0.004	5.212	$8x10^{-17}$
	ENSG00000253981	*ALG1L13P*	-0.006	2.721	$3x10^{-15}$
	ENSG00000169877	*AHSP*	0.012	4.573	$6x10^{-15}$
DIU	ENST00000432854	*DBNL*	0.017	-1.701	$1x10^{-20}$
	ENST00000483180	*NFKBIZ*	-0.015	-1.759	$2x10^{-13}$
	ENST00000357428	*USP33*	0.013	-2.868	$7x10^{-13}$
	ENST00000315939	*WNK1*	0.012	2.770	$5x10^{-12}$
	ENST00000339486	*RIOK3*	0.008	8.065	$5x10^{-12}$
DEU	360147	*PSMA1*	-0.004	1.511	$1x10^{-7}$
	413338	*FRY*	0.004	1.744	$2x10^{-7}$
	450397	*CCNDBP1*	-0.004	2.388	$3x10^{-7}$
	514701	*VMP1*	0.002	4.936	$1x10^{-6}$
	510631	*ATP6V0A1*	0.003	2.087	$4x10^{-6}$

Adjusted Perc15 density: Hounsfield units at the 15th percentile of CT density histogram at total lung capacity, corrected for the inspiratory depth (per convention, adjusted Perc15 density values are reported as the HU + 1000). The lower the Perc15 values are, i.e., the closer to -1,000 HU, the more CT-quantified emphysema is present.

For the DGE, DIU, and DEU analyses performed in the training and testing samples, the covariates used were age, race, sex, pack-years of smoking, current smoking status, forced expiratory volume in one second (FEV1), CBC cell count proportions, library preparation batch, and CT scanner model. False discovery rate (FDR) was used for multiple testing corrections.

Genes and isoforms are represented by their Ensembl Gene ID and Ensembl Transcript ID, respectively. Exonic part IDs with genomic positions are available in Supplemental Table E2. HUGO Gene Name corresponds to the unique gene identified by the Ensembl Gene ID (DGE), and the gene associated with the isoform or exon (DIU and DEU). Log fold change values indicate change per unit increase in adjusted Perc15. Positive log fold change values represent upregulated genes, while negative ones correspond to downregulated ones with respect to adjusted Perc15 density (i.e., they have opposite signs for their associations with emphysema). Average log expression is the average of the log-transformed counts of the gene in analyzed subjects. A threshold of FDR 10% was applied.

Table 3. Selected top 5 gene ontology (GO) biological processes enriched in differentially expressed genes (DGE), differentially used isoforms (DIU), and differentially used exons (DEU) associated to adjusted Perc15 density. GO terms were selected based on potential biological relevance to emphysema.

	GO.ID	GO Term	Total number of genes in category	Number of adjusted Perc15 density-associated genes in category	Adjusted P-value
DGE	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	99	72	$2x10^{-21}$
	GO:0006413	Translational initiation	185	97	$2x10^{-19}$
	GO:0000184	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	120	77	$5x10^{-14}$
	GO:0019083	Viral transcription	174	87	$3x10^{-13}$
	GO:0043312	Neutrophil degranulation	466	212	$9x10^{-12}$
	GO:0002181	Cytoplasmic translation	98	44	$7x10^{-7}$
	GO:0051092	Positive regulation of NF-kappaB transcription factor activity	144	70	$3x10^{-6}$
	GO:0046718	Viral entry into host cell	111	58	$6x10^{-6}$
	GO:0042102	Positive regulation of T cell proliferation	84	47	$1x10^{-5}$
	GO:0010803	Regulation of tumor necrosis factor-mediated signaling pathway	51	25	$2x10^{-5}$
DIU	GO:0006413	Translational initiation	176	35	$2x10^{-5}$
	GO:0045070	Positive regulation of viral genome replication	32	13	$3x10^{-5}$
	GO:0043044	ATP-dependent chromatin remodeling	63	13	$7x10^{-5}$
	GO:0090263	Positive regulation of canonical WNT signaling pathway	107	26	$7x10^{-5}$

	GO:0018105	Peptidyl-serine phosphorylation	209	45	$1x10^{-4}$
	GO:0032092	Positive regulation of protein binding	59	17	$1x10^{-4}$
	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	94	23	$2x10^{-4}$
	GO:0000184	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	118	28	$2x10^{-4}$
	GO:0090263	Positive regulation of transcription by RNA polymerase II	690	106	$3x10^{-4}$
	GO:0019083	Viral transcription	172	31	$5x10^{-4}$
	GO:0019079	Viral genome replication	103	23	$5x10^{-4}$
DEU	GO:0006413	Translational initiation	181	13	0
	GO:0006995	Cellular response to nitrogen starvation	9	3	0.001
	GO:1904667	Negative regulation of ubiquitin protein ligase activity	9	3	0.001
	GO:1901991	Negative regulation of mitotic cell cycle phase transition	182	8	0.002
	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	94	8	0.002
	GO:0000422	Autophagy of mitochondrion	72	8	0.002
	GO:0071560	Cellular response to transforming growth factor beta stimulus	133	9	0.002
	GO:0045722	Positive regulation of gluconeogenesis	11	3	0.002
	GO:0043124	Negative regulation of I-kappaB kinase/NF-kappaB signaling	38	5	0.002
	GO:0050821	Protein stabilization	142	10	0.002

Adjusted Perc15 density: Hounsfield units at the 15th percentile of CT density histogram at total lung capacity, corrected for the inspiratory depth (per convention, adjusted Perc15 density values are reported as the HU + 1000). The lower the Perc15 values are, i.e., the closer to -1,000 HU, the more CT-quantified emphysema is present.

For the DGE, DIU, and DEU analyses, covariates used were age, race, sex, pack-years of smoking, current smoking status, forced expiratory volume in one second (FEV1), CBC cell count proportions, library preparation batch, and CT scanner model. False discovery rate (FDR) was used for multiple testing corrections.

For the GO analyses, we only reported the GO pathways with at least 3 significant genes. Enriched GO terms were identified using the weighted Fisher's test P-values < 0.005. Selected GO terms with the lowest P-values in the DGE, DIU, and DEU analyses are listed. Total number of genes in category refers to all genes studied that fall under the GO term. The number of adjusted Perc15-associated genes in category refers to the genes that reached significance (*FDR 10%*) in the DGE, DIU, and DEU analyses.

Table 4. Mediated proportions and direct, indirect, and total effects of the top 5 most and least mediated differentially expressed genes significantly associated to adjusted Perc15 density.

	Ensembl Gene ID	HUGO Gene Name	Mediated Proportion	Direct Effect		Indirect Effect		Total Effect	
				Beta Coefficient	FDR	Beta Coefficient	FDR	Beta Coefficient	FDR
Most mediated genes (genes with significant indirect effect)	ENSG00000160179	*ABCG1*	0.822	-0.001	0.324	-0.005	$1x10^{-31}$	-0.006	$2x10^{-18}$
	ENSG00000169877	*AHSP*	0.882	0.001	0.559	0.009	$4x10^{-31}$	0.011	$7x10^{-15}$
	ENSG00000118113	*MMP8*	1.054	-0.001	0.859	0.010	$4x10^{-26}$	0.009	$1x10^{-9}$
	ENSG00000158578	*ALAS2*	0.912	0.001	0.724	0.008	$1x10^{-24}$	0.009	$3x10^{-11}$
	ENSG00000119326	*CTNNAL1*	0.928	0.000	0.795	0.006	$3x10^{-24}$	0.007	$1x10^{-10}$
Least mediated genes (genes with significant direct effect)	ENSG00000189430	*NCR1*	-0.124	-0.004	$1x10^{-4}$	$4x10^{-4}$	0.318	-0.003	$2x10^{-6}$
	ENSG00000179841	*AKAP5*	0.129	-0.004	0.002	$-7x10^{-4}$	0.199	-0.005	$8x10^{-9}$
	ENSG00000165071	*TMEM71*	0.094	-0.001	0.002	$-1x10^{-4}$	0.374	-0.001	$5x10^{-8}$
	ENSG00000170298	*LGALS9B*	-0.065	-0.005	0.002	$3x10^{-4}$	0.642	-0.005	$1x10^{-5}$
	ENSG00000162909	*CAPN2*	-0.217	0.001	0.002	$-2x10^{-4}$	0.128	0.001	$5x10^{-5}$

Mediation analysis was performed to distinguish how much of the effect of emphysema on gene expression acted through BMI (referred to as the indirect effect) and how much of the effect of emphysema directly influenced gene expression (referred to as the direct effect). Covariates: BMI, sex, age, race, pack-years of smoking, current smoking status, and forced expiratory volume in one second (FEV_1).

Mediated proportions of top 5 genes are listed along with the coefficients and false discovery rates (FDR) of their direct, indirect, and total effects. Mediated proportion is defined as the ratio of indirect effect to the sum of the indirect and direct effects. Genes are sorted in order of decreasing FDR for the total effect.

FIGURE LEGENDS

Figure 1. COPDGene Visit 2 participant flow diagram. Abbreviations: NHW = Non-Hispanic White, AA = African-American, FEV_1 = Forced expiratory volume during the first second, CBC = Cell blood count.

Figure 2. Study overview figure. Abbreviations: CT = Computed tomography, BMI = Body mass index, CBC = Cell blood count.

Figure 3. (A) Volcano plot of differentially expressed genes without BMI adjustment. (B) Volcano plot of differentially used isoforms without BMI adjustment. (C) Volcano plot of differentially used exons without BMI adjustment. Genes significantly associated with adjusted Perc15 density, therefore differentially used, appear above the red line marked at a threshold of 10% false discovery rate (FDR). Genes that are up-regulated are in blue and those that are down-regulated are in red. Isoforms/exons that are not differential used are gray and appear below the threshold line. Adjusted Perc15 density: Hounsfield units at the 15th percentile of CT density histogram at total lung capacity, corrected for the inspiratory depth (per convention, adjusted Perc15 density values are reported as the HU + 1000). The lower the Perc15 values are, i.e., the closer to -1,000 HU, the more CT-quantified emphysema is present. Upregulated versus downregulated genes are reported with respect to adjusted Perc15 density (i.e., they have opposite directions for their associations with emphysema).

Figure 4. (A) Number of significant genes associated with adjusted Perc15 density from the differential gene expression (DGE), differential isoform usage (DIU), differential exon usage (DEU), and protein association analyses. (B) Number of significant enriched gene ontology (GO) terms from the DGE, DIU, DEU, and protein association analyses. Adjusted Perc15 density: Hounsfield units at the 15th percentile of CT density histogram at total lung capacity,

corrected for the inspiratory depth (per convention, adjusted Perc15 density values are reported as the HU + 1000). The lower the Perc15 values are, i.e., the closer to -1,000 HU, the more CT-quantified emphysema is present. Upregulated versus downregulated are reported with respect to adjusted Perc15 density (i.e., they have opposite directions for their associations with emphysema).

Figure 5. The receiver operating characteristic curves for all four models from the elastic net prediction. The models compared are CBC (cell blood count) only, CBC plus gene, CBC plus protein, and CBC plus gene plus protein. The table summarizes the pairwise DeLong P-values of the models, in which significant differences in model performance (P-values < 0.05) are marked.

Figure 6. Top 10 predictors sorted in descending order by the absolute magnitude of their beta-coefficients from the elastic net model using CBC (cell blood count), gene, and protein data. The horizontal lines represent the magnitude of the coefficient for each feature. All predictors were centered and scaled.

REFERENCES

1. Lindberg A, Lindberg L, Sawalha S, Nilsson U, Stridsman C, Lundbäck B, Backman H. Large underreporting of COPD as cause of death-results from a population-based cohort study. *Respir Med* 2021; 186: 106518.

2. Li Y, Swensen SJ, Karabekmez LG, Marks RS, Stoddard SM, Jiang R, Worra JB, Zhang F, Midthun DE, de Andrade M, Song Y, Yang P. Effect of emphysema on lung cancer risk in smokers: a computed tomography-based assessment. *Cancer Prev Res (Phila)* 2011; 4: 43-50.

3. Rahman HH, Niemann D, Munson-McGee SH. Association between asthma, chronic bronchitis, emphysema, chronic obstructive pulmonary disease, and lung cancer in the US population. *Environ Sci Pollut Res Int* 2022.

4. Morgan AD, Zakeri R, Quint JK. Defining the relationship between COPD and CVD: what are the implications for clinical practice? *Ther Adv Respir Dis* 2018; 12: 1753465817750524.

5. Barnes PJ, Stockley RA. COPD: current therapeutic interventions and future approaches. *Eur Respir J* 2005; 25: 1084-1106.

6. Shaw JG, Vaughan A, Dent AG, O'Hare PE, Goh F, Bowman RV, Fong KM, Yang IA. Biomarkers of progression of chronic obstructive pulmonary disease (COPD). *J Thorac Dis* 2014; 6: 1532-1547.

7. Carolan BJ, Hughes G, Morrow J, Hersh CP, O'Neal WK, Rennard S, Pillai SG, Belloni P, Cockayne DA, Comellas AP, Han M, Zemans RL, Kechris K, Bowler RP. The association of plasma biomarkers with computed tomography-assessed emphysema phenotypes. *Respir Res* 2014; 15: 127.

8. Lopez-Campos JL, Alcazar B. Evaluation of symptomatic patients without airflow obstruction: back to the future. *J Thorac Dis* 2016; 8: E1657-e1660.

9. Guo NL, Wan YW. Network-based identification of biomarkers coexpressed with multiple pathways. *Cancer Inform* 2014; 13: 37-47.

10. Zuo Q, Wang Y, Yang D, Guo S, Li X, Dong J, Wan C, Shen Y, Wen F. Identification of hub genes and key pathways in the emphysema phenotype of COPD. *Aging (Albany NY)* 2021; 13: 5120-5135.

11. Zhang Y, Tedrow J, Nouraie M, Li X, Chandra D, Bon J, Kass DJ, Fuhrman CR, Leader JK, Duncan SR, Kaminski N, Sciurba FC. Elevated plasma level of Pentraxin 3 is associated with emphysema and mortality in smokers. *Thorax* 2021; 76: 335-342.

12. Paci P, Fiscon G, Conte F, Licursi V, Morrow J, Hersh C, Cho M, Castaldi P, Glass K, Silverman EK, Farina L. Integrated transcriptomic correlation network analysis identifies COPD molecular determinants. *Sci Rep* 2020; 10: 3361.

13. Lamontagne M, Timens W, Hao K, Bossé Y, Laviolette M, Steiling K, Campbell JD, Couture C, Conti M, Sherwood K, Hogg JC, Brandsma CA, van den Berge M, Sandford A, Lam S, Lenburg ME, Spira A, Paré PD, Nickle D, Sin DD, Postma DS. Genetic regulation of gene expression in the lung identifies CST3 and CD22 as potential causal genes for airflow obstruction. *Thorax* 2014; 69: 997-1004.

14. Sakornsakolpat P, Morrow JD, Castaldi PJ, Hersh CP, Bossé Y, Silverman EK, Manichaikul A, Cho MH. Integrative genomics identifies new genes associated with severe COPD and emphysema. *Respir Res* 2018; 19: 46.

15. Kowalski ML, Borowiec M, Kurowski M, Pawliczak R. Alternative splicing of cyclooxygenase-1 gene: altered expression in leucocytes from patients with bronchial asthma and association with aspirin-induced 15-HETE release. *Allergy* 2007; 62: 628-634.

16. Deng N, Sanchez CG, Lasky JA, Zhu D. Detecting splicing variants in idiopathic pulmonary fibrosis from non-differentially expressed genes. *PLoS One* 2013; 8: e68352.

17. Cogan J, Austin E, Hedges L, Womack B, West J, Loyd J, Hamid R. Role of BMPR2 alternative splicing in heritable pulmonary arterial hypertension penetrance. *Circulation* 2012; 126: 1907-1916.

18. Saferali A, Yun JH, Parker MM, Sakornsakolpat P, Chase RP, Lamb A, Hobbs BD, Boezen MH, Dai X, de Jong K, Beaty TH, Wei W, Zhou X, Silverman EK, Cho MH, Castaldi PJ, Hersh CP. Analysis of genetically driven alternative splicing identifies FBXO38 as a novel COPD susceptibility gene. *PLoS Genet* 2019; 15: e1008229.

19. Saferali A, Xu Z, Sheynkman GM, Hersh CP, Cho MH, Silverman EK, Laederach A, Vollmers C, Castaldi PJ. Characterization of a COPD-Associated NPNT Functional Splicing Genetic Variant in Human Lung Tissue via Long-Read Sequencing. *medRxiv* 2020.

20. Faiz A, van den Berge M, Vermeulen CJ, Ten Hacken NHT, Guryev V, Pouwels SD. AGER expression and alternative splicing in bronchial biopsies of smokers and never smokers. *Respir Res* 2019; 20: 70.

21. Kim WJ, Lim JH, Lee JS, Lee SD, Kim JH, Oh YM. Comprehensive Analysis of Transcriptome Sequencing Data in the Lung Tissues of COPD Subjects. *Int J Genomics* 2015; 2015: 206937.

22. Zhang YH, Hoopmann MR, Castaldi PJ, Simonsen K, Midha M, Cho MH, Criner GJ, Bueno R, Liu J, Moritz R, Silverman EK. Lung proteomic biomarkers associated with chronic obstructive pulmonary disease. *Am J Physiol Lung Cell Mol Physiol* 2021.

23. Faner R, Tal-Singer R, Riley JH, Celli B, Vestbo J, MacNee W, Bakke P, Calverley PM, Coxson H, Crim C, Edwards LD, Locantore N, Lomas DA, Miller BE, Rennard SI,

Wouters EF, Yates JC, Silverman EK, Agusti A. Lessons from ECLIPSE: a review of COPD biomarkers. *Thorax* 2014; 69: 666-672.

24. Miller M, Ramsdell J, Friedman PJ, Cho JY, Renvall M, Broide DH. Computed tomographic scan-diagnosed chronic obstructive pulmonary disease-emphysema: eotaxin-1 is associated with bronchodilator response and extent of emphysema. *J Allergy Clin Immunol* 2007; 120: 1118-1125.
25. Bracke KR, D'Hulst A I, Maes T, Moerloose KB, Demedts IK, Lebecque S, Joos GF, Brusselle GG. Cigarette smoke-induced pulmonary inflammation and emphysema are attenuated in CCR6-deficient mice. *J Immunol* 2006; 177: 4350-4359.
26. Zemans RL, Jacobson S, Keene J, Kechris K, Miller BE, Tal-Singer R, Bowler RP. Multiple biomarkers predict disease severity, progression and mortality in COPD. *Respir Res* 2017; 18: 117.
27. Keene JD, Jacobson S, Kechris K, Kinney GL, Foreman MG, Doerschuk CM, Make BJ, Curtis JL, Rennard SI, Barr RG, Bleecker ER, Kanner RE, Kleerup EC, Hansel NN, Woodruff PG, Han MK, Paine R, 3rd, Martinez FJ, Bowler RP, O'Neal WK. Biomarkers Predictive of Exacerbations in the SPIROMICS and COPDGene Cohorts. *Am J Respir Crit Care Med* 2017; 195: 473-481.
28. Celli BR, Cote CG, Marin JM, Casanova C, Montes de Oca M, Mendez RA, Pinto Plata V, Cabral HJ. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med* 2004; 350: 1005-1012.
29. Thomsen M, Dahl M, Lange P, Vestbo J, Nordestgaard BG. Inflammatory biomarkers and comorbidities in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2012; 186: 982-988.

30. McNicholas WT. COPD-OSA Overlap Syndrome: Evolving Evidence Regarding Epidemiology, Clinical Consequences, and Management. *Chest* 2017; 152: 1318-1326.

31. Suryadevara R, Gregory A, Masoomi A, Xu Z, Berman S, Yun JH, Saferali A, Hersh CP, Silverman EK, Dy J, Castaldi PJ, El Boueiz A. Blood Transcriptomics-Based Machine Learning Prediction of Emphysema in Smokers. *CHEST Journal* 2021; Volume 160, Issue 4, Supplement, A1841-A1842, October 01, 2021.

32. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. *Copd* 2010; 7: 32-43.

33. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MS, Zheng J, Stocks J. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012; 40: 1324-1343.

34. Parr DG, Sevenoaks M, Deng C, Stoel BC, Stockley RA. Detection of emphysema progression in alpha 1-antitrypsin deficiency using CT densitometry; methodological advances. *Respir Res* 2008; 9: 21.

35. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26: 139-140.

36. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, Carter J, Dalby AB, Eaton BE, Fitzwater T, Flather D, Forbes A, Foreman T, Fowler C, Gawande B, Goss M, Gunn M, Gupta S, Halladay D, Heil J, Heilig J, Hicke B, Husar G, Janjic N, Jarvis T, Jennings S, Katilius E, Keeney TR, Kim N, Koch TH, Kraemer S, Kroiss L, Le N, Levine D, Lindsey W, Lollo B, Mayfield W, Mehan M, Mehler R, Nelson SK, Nelson M, Nieuwlandt D, Nikrad M, Ochsner U, Ostroff RM, Otis M, Parker T, Pietrasiewicz

S, Resnicow DI, Rohloff J, Sanders G, Sattin S, Schneider D, Singer B, Stanton M, Sterkel A, Stewart A, Stratford S, Vaught JD, Vrkljan M, Walker JJ, Watrobka M, Waugh S, Weiss A, Wilcox SK, Wolfson A, Wolk SK, Zhang C, Zichi D. Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE* 2010; 5: e15004.

37. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; 43: e47.

38. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014; 15: R29.

39. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995; 57: 289-300.

40. Steen J, Loeys T, Moerkerke B, Vansteelandt S. medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models. *Journal of Statistical Software* 2017; 76: 1 - 46.

41. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2006; 22: 1600-1607.

42. Quan D, Ren J, Ren H, Linghu L, Wang X, Li M, Qiao Y, Ren Z, Qiu L. Exploring influencing factors of chronic obstructive pulmonary disease based on elastic net and Bayesian network. *Sci Rep* 2022; 12: 7563.

43. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436-444.

44. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12: 77.

45. Faner R, Cruz T, Casserras T, López-Giraldo A, Noell G, Coca I, Tal-Singer R, Miller B, Rodriguez-Roisin R, Spira A, Kalko SG, Agustí A. Network Analysis of Lung Transcriptomics Reveals a Distinct B-Cell Signature in Emphysema. *Am J Respir Crit Care Med* 2016; 193: 1242-1253.

46. Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, Bowler R, Reisdorph N. Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. *Sci Rep* 2018; 8: 17132.

47. Qiu W, Cho MH, Riley JH, Anderson WH, Singh D, Bakke P, Gulsvik A, Litonjua AA, Lomas DA, Crapo JD, Beaty TH, Celli BR, Rennard S, Tal-Singer R, Fox SM, Silverman EK, Hersh CP. Genetics of sputum gene expression in chronic obstructive pulmonary disease. *PLoS One* 2011; 6: e24395.

48. Wiegman CH, Li F, Ryffel B, Togbe D, Chung KF. Oxidative Stress in Ozone-Induced Chronic Lung Inflammation and Emphysema: A Facet of Chronic Obstructive Pulmonary Disease. *Front Immunol* 2020; 11: 1957.

49. Lugg ST, Scott A, Parekh D, Naidu B, Thickett DR. Cigarette smoke exposure and alveolar macrophages: mechanisms for lung disease. *Thorax* 2022; 77: 94-101.

50. Yun JH, Morrow J, Owen CA, Qiu W, Glass K, Lao T, Jiang Z, Perrella MA, Silverman EK, Zhou X, Hersh CP. Transcriptomic Analysis of Lung Tissue from Cigarette Smoke-Induced Emphysema Murine Models and Human Chronic Obstructive Pulmonary Disease Show Shared and Distinct Pathways. *Am J Respir Cell Mol Biol* 2017; 57: 47-58.

51. Serban KA, Pratte KA, Strange C, Sandhaus RA, Turner AM, Beiko T, Spittle DA, Maier L, Hamzeh N, Silverman EK, Hobbs BD, Hersh CP, DeMeo DL, Cho MH, Bowler RP. Unique and shared systemic biomarkers for emphysema in Alpha-1 Antitrypsin deficiency and chronic obstructive pulmonary disease. *EBioMedicine* 2022; 84: 104262.

52. Meyer M, Jaspers I. Respiratory protease/antiprotease balance determines susceptibility to viral infection and can be modified by nutritional antioxidants. *Am J Physiol Lung Cell Mol Physiol* 2015; 308: L1189-1201.

53. Koethe SM, Kuhnmuench JR, Becker CG. Neutrophil priming by cigarette smoke condensate and a tobacco anti-idiotypic antibody. *Am J Pathol* 2000; 157: 1735-1743.

54. Jasper AE, McIver WJ, Sapey E, Walton GM. Understanding the role of neutrophils in chronic inflammatory airway disease. *F1000Res* 2019; 8.

55. Hoenderdos K, Condliffe A. The neutrophil in chronic obstructive pulmonary disease. *Am J Respir Cell Mol Biol* 2013; 48: 531-539.

56. Ernst M, Inglese M, Scholz GM, Harder KW, Clay FJ, Bozinovski S, Waring P, Darwiche R, Kay T, Sly P, Collins R, Turner D, Hibbs ML, Anderson GP, Dunn AR. Constitutive activation of the SRC family kinase Hck results in spontaneous pulmonary inflammation and an enhanced innate immune response. *J Exp Med* 2002; 196: 589-604.

57. Bagdonas E, Raudoniute J, Bruzauskaite I, Aldonyte R. Novel aspects of pathogenesis and regeneration mechanisms in COPD. *Int J Chron Obstruct Pulmon Dis* 2015; 10: 995-1013.

58. Marc MM, Korosec P, Kosnik M, Kern I, Flezar M, Suskovic S, Sorli J. Complement factors c3a, c4a, and c5a in chronic obstructive pulmonary disease and asthma. *Am J Respir Cell Mol Biol* 2004; 31: 216-219.

59. Westwood JP, Mackay AJ, Donaldson G, Machin SJ, Wedzicha JA, Scully M. The role of complement activation in COPD exacerbation recovery. *ERJ Open Res* 2016; 2.

60. Liu T, Zhang L, Joo D, Sun SC. NF-κB signaling in inflammation. *Signal Transduct Target Ther* 2017; 2: 17023-.

61. Brown V, Elborn JS, Bradley J, Ennis M. Dysregulated apoptosis and NFkappaB expression in COPD subjects. *Respir Res* 2009; 10: 24.

62. Di Stefano A, Caramori G, Oates T, Capelli A, Lusuardi M, Gnemmi I, Ioli F, Chung KF, Donner CF, Barnes PJ, Adcock IM. Increased expression of nuclear factor-kappaB in bronchial biopsies from smokers and patients with COPD. *Eur Respir J* 2002; 20: 556-563.

63. Guo-Parke H, Linden D, Weldon S, Kidney JC, Taggart CC. Mechanisms of Virus-Induced Airway Immunity Dysfunction in the Pathogenesis of COPD Disease, Progression, and Exacerbation. *Front Immunol* 2020; 11: 1205.

64. Schuliga M. NF-kappaB Signaling in Chronic Inflammatory Airway Disease. *Biomolecules* 2015; 5: 1266-1283.

65. Lin Q, Zhang CF, Guo JL, Su JL, Guo ZK, Li HY. Involvement of NEAT1/PINK1-mediated mitophagy in chronic obstructive pulmonary disease induced by cigarette smoke or PM(2.5). *Ann Transl Med* 2022; 10: 277.

66. Mizumura K, Cloonan SM, Nakahira K, Bhashyam AR, Cervo M, Kitada T, Glass K, Owen CA, Mahmood A, Washko GR, Hashimoto S, Ryter SW, Choi AM. Mitophagy-dependent necroptosis contributes to the pathogenesis of COPD. *J Clin Invest* 2014; 124: 3987-4003.

67. Aros CJ, Pantoja CJ, Gomperts BN. Wnt signaling in lung development, regeneration, and disease progression. *Commun Biol* 2021; 4: 601.

68. Qu J, Yue L, Gao J, Yao H. Perspectives on Wnt Signal Pathway in the Pathogenesis and Therapeutics of Chronic Obstructive Pulmonary Disease. *J Pharmacol Exp Ther* 2019; 369: 473-480.

69. Carlier FM, Dupasquier S, Ambroise J, Detry B, Lecocq M, Biétry-Claudet C, Boukala Y, Gala JL, Bouzin C, Verleden SE, Hoton D, Gohy S, Bearzatto B, Pilette C. Canonical WNT pathway is activated in the airway epithelium in chronic obstructive pulmonary disease. *EBioMedicine* 2020; 61: 103034.

70. Shi J, Li F, Luo M, Wei J, Liu X. Distinct Roles of Wnt/β-Catenin Signaling in the Pathogenesis of Chronic Obstructive Pulmonary Disease and Idiopathic Pulmonary Fibrosis. *Mediators Inflamm* 2017; 2017: 3520581.

71. Stratelis G, Fransson SG, Schmekel B, Jakobsson P, Mölstad S. High prevalence of emphysema and its association with BMI: a study of smokers with normal spirometry. *Scand J Prim Health Care* 2008; 26: 241-247.

72. Divo MJ, Cabrera C, Casanova C, Marin JM, Pinto-Plata VM, de-Torres JP, Zulueta J, Zagaceta J, Sanchez-Salcedo P, Berto J, Cote C, Celli BR. Comorbidity Distribution, Clinical Expression and Survival in COPD Patients with Different Body Mass Index. *Chronic Obstr Pulm Dis* 2014; 1: 229-238.

73. McDonald MN, Wouters EFM, Rutten E, Casaburi R, Rennard SI, Lomas DA, Bamman M, Celli B, Agusti A, Tal-Singer R, Hersh CP, Dransfield M, Silverman EK. It's more than low BMI: prevalence of cachexia and associated mortality in COPD. *Respir Res* 2019; 20: 100.

74. Wilson AC, Kumar PL, Lee S, Parker MM, Arora I, Morrow JD, Wouters EFM, Casaburi R, Rennard SI, Lomas DA, Agusti A, Tal-Singer R, Dransfield MT, Wells JM, Bhatt SP, Washko G, Thannickal VJ, Tiwari HK, Hersh CP, Castaldi PJ, Silverman EK,

McDonald MN. Heme metabolism genes Downregulated in COPD Cachexia. *Respir Res* 2020; 21: 100.

75. Sachar M, Anderson KE, Ma X. Protoporphyrin IX: the Good, the Bad, and the Ugly. *J Pharmacol Exp Ther* 2016; 356: 267-275.
76. Boulesteix AL, Sauerbrei W. Added predictive value of high-throughput molecular data to clinical data and its validation. *Brief Bioinform* 2011; 12: 215-229.
77. Liu Q, Sun D, Wang Y, Li P, Jiang T, Dai L, Duo M, Wu R, Cheng Z. Use of machine learning models to predict prognosis of combined pulmonary fibrosis and emphysema in a Chinese population. *BMC Pulm Med* 2022; 22: 327.
78. Humphries SM, Notary AM, Centeno JP, Strand MJ, Crapo JD, Silverman EK, Lynch DA. Deep Learning Enables Automatic Classification of Emphysema Pattern at CT. *Radiology* 2020; 294: 434-444.
79. Castaldi PJ, Boueiz A, Yun J, Estepar RSJ, Ross JC, Washko G, Cho MH, Hersh CP, Kinney GL, Young KA, Regan EA, Lynch DA, Criner GJ, Dy JG, Rennard SI, Casaburi R, Make BJ, Crapo J, Silverman EK, Hokanson JE. Machine Learning Characterization of COPD Subtypes: Insights From the COPDGene Study. *Chest* 2020; 157: 1147-1157.
80. Castaldi PJ, Dy J, Ross J, Chang Y, Washko GR, Curran-Everett D, Williams A, Lynch DA, Make BJ, Crapo JD, Bowler RP, Regan EA, Hokanson JE, Kinney GL, Han MK, Soler X, Ramsdell JW, Barr RG, Foreman M, van Beek E, Casaburi R, Criner GJ, Lutz SM, Rennard SI, Santorico S, Sciurba FC, DeMeo DL, Hersh CP, Silverman EK, Cho MH. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* 2014; 69: 415-422.
81. Castaldi PJ, Cho MH, San Jose Estepar R, McDonald ML, Laird N, Beaty TH, Washko G, Crapo JD, Silverman EK, Investigators CO. Genome-wide association identifies

regulatory Loci associated with distinct local histogram emphysema patterns. *Am J Respir Crit Care Med* 2014; 190: 399-409.

82. Pratte KA, Curtis JL, Kechris K, Couper D, Cho MH, Silverman EK, DeMeo DL, Sciurba FC, Zhang Y, Ortega VE, O'Neal WK, Gillenwater LA, Lynch DA, Hoffman EA, Newell JD, Jr., Comellas AP, Castaldi PJ, Miller BE, Pouwels SD, Hacken N, Bischoff R, Klont F, Woodruff PG, Paine R, Barr RG, Hoidal J, Doerschuk CM, Charbonnier JP, Sung R, Locantore N, Yonchuk JG, Jacobson S, Tal-Singer R, Merrill D, Bowler RP. Soluble receptor for advanced glycation end products (sRAGE) as a biomarker of COPD. *Respir Res* 2021; 22: 127.

83. Huyghe A, Van den Ackerveken P, Sacheli R, Prévot PP, Thelen N, Renauld J, Thiry M, Delacroix L, Nguyen L, Malgrange B. MicroRNA-124 Regulates Cell Specification in the Cochlea through Modulation of Sfrp4/5. *Cell Rep* 2015; 13: 31-42.

84. Bayat A, Saki N, Nikakhlagh S, Mirmomeni G, Raji H, Soleimani H, Rahim F. Is COPD associated with alterations in hearing? A systematic review and meta-analysis. *Int J Chron Obstruct Pulmon Dis* 2019; 14: 149-162.

85. Ortapamuk H, Naldoken S. Brain perfusion abnormalities in chronic obstructive pulmonary disease: comparison with cognitive impairment. *Ann Nucl Med* 2006; 20: 99-106.

86. Barnes PJ. The cytokine network in chronic obstructive pulmonary disease. *Am J Respir Cell Mol Biol* 2009; 41: 631-638.

ACKNOWLEDGEMENTS

COPDGene Investigators - Core Units:

Administrative Center: James D. Crapo, MD (PI); Edwin K. Silverman, MD, PhD (PI); Barry J. Make, MD; Elizabeth A. Regan, MD, PhD

Genetic Analysis Center: Terri H. Beaty, PhD; Peter J. Castaldi, MD, MSc; Michael H. Cho, MD, MPH; Dawn L. DeMeo, MD, MPH; Adel Boueiz, MD, MMSc; Marilyn G. Foreman, MD, MS; Auyon Ghosh, MD; Lystra P. Hayden, MD, MMSc; Craig P. Hersh, MD, MPH; Jacqueline Hetmanski, MS; Brian D. Hobbs, MD, MMSc; John E. Hokanson, MPH, PhD; Wonji Kim, PhD; Nan Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD; Dmitry Prokopenko, PhD; Matthew Moll, MD, MPH; Jarrett Morrow, PhD; Dandi Qiao, PhD; Elizabeth A. Regan, MD, PhD; Aabida Saferali, PhD; Phuwanat Sakornsakolpat, MD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD; Jeong Yun, MD, MPH

Imaging Center: Juan Pablo Centeno; Jean-Paul Charbonnier, PhD; Harvey O. Coxson, PhD; Craig J. Galban, PhD; MeiLan K. Han, MD, MS; Eric A. Hoffman, Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex

Kluiber; David A. Lynch, MB; Pietro Nardelli, PhD; John D. Newell, Jr., MD; Aleena Notary; Andrea Oh, MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; Gonzalo Vegas SanchezFerrero, PhD; Lucas Veitel; George R. Washko, MD; Carla G. Wilson, MS

PFT QA Center, Salt Lake City, UT: Robert Jensen, PhD

Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO: Douglas Everett, PhD; Jim Crooks, PhD; Katherine Pratte, PhD; Matt Strand, PhD; Carla G. Wilson, MS

Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO: John E. Hokanson, MPH, PhD; Erin Austin, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Kendra A. Young, PhDVersion Date: March 26, 2021

Mortality Adjudication Core: Surya P. Bhatt, MD; Jessica Bon, MD; Alejandro A. Diaz, MD, MPH; MeiLan K. Han, MD, MS; Barry Make, MD; Susan Murray, ScD; Elizabeth Regan, MD; Xavier Soler, MD; Carla G. Wilson, MS

Biomarker Core: Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; Farnoush BanaeiKashani, PhD

COPDGene Investigators - Clinical Centers:

Ann Arbor VA: Jeffrey L. Curtis, MD; Perry G. Pernicano, MD

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS; Mustafa Atik, MD; Aladin Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Amit Parulekar, MD

Brigham and Women's Hospital, Boston, MA: Dawn L. DeMeo, MD, MPH; Craig Hersh, MD, MPH; Francine L. Jacobson, MD, MPH; George Washko, MD

Columbia University, New York, NY: R. Graham Barr, MD, DrPH; John Austin, MD; Belinda D'Souza, MD; Byron Thomashow, MD

Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD; H. Page McAdams, MD; Lacey Washington, MD

HealthPartners Research Institute, Minneapolis, MN: Charlene McEvoy, MD, MPH; Joseph Tashjian, MD

Johns Hopkins University, Baltimore, MD: Robert Wise, MD; Robert Brown, MD; Nadia N. Hansel, MD, MPH; Karen Horton, MD; Allison Lambert, MD, MHS; Nirupama Putcha, MD, MHS

Lundquist Institute for Biomedical Innovation at Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD; Alessandra Adami, PhD; Matthew Budoff, MD; Hans Fischer, MD; Janos Porszasz, MD, PhD; Harry Rossiter, PhD; William Stringer, MD

Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, PhD; Charlie Lan, DO

Minneapolis VA: Christine Wendt, MD; Brian Bell, MD; Ken M. Kunisaki, MD, MS

Morehouse School of Medicine, Atlanta, GA: Eric L. Flenaugh, MD; Hirut Gebrekristos, PhD; Mario Ponce, MD; Silanath Terpenning, MD; Gloria Westney, MD, MS

National Jewish Health, Denver, CO: Russell Bowler, MD, PhD; David A. Lynch, MB

Reliant Medical Group, Worcester, MA: Richard Rosiello, MD; David Pace, MD

Temple University, Philadelphia, PA: Gerard Criner, MD; David Ciccolella, MD; Francis Cordova, MD; Chandra Dass, MD; Gilbert D'Alonzo, DO; Parag Desai, MD; Michael Jacobs, PharmD; Steven Kelsen, MD, PhD; Victor Kim, MD; A. James Mamary, MD; Nathaniel Marchetti, DO; Aditi Satti, MD; Kartik Shenoy, MD; Robert M. Steiner, MD; Alex Swift, MD; Irene Swift, MD; Maria Elena Vega-Sanchez, MD

University of Alabama, Birmingham, AL: Mark Dransfield, MD; William Bailey, MD; Surya P. Bhatt, MD; Anand Iyer, MD; Hrudaya Nath, MD; J. Michael Wells, MD

University of California, San Diego, CA: Douglas Conrad, MD; Xavier Soler, MD, PhD; Andrew Yen, MD

University of Iowa, Iowa City, IA: Alejandro P. Comellas, MD; Karin F. Hoth, PhD; John Newell, Jr., MD; Brad Thompson, MD

University of Michigan, Ann Arbor, MI: MeiLan K. Han, MD MS; Ella Kazerooni, MD MS; Wassim Labaki, MD MS; Craig Galban, PhD; Dharshan Vummidi, MD

University of Minnesota, Minneapolis, MN: Joanne Billings, MD; Abbie Begnaud, MD; Tadashi Allen, MD

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD; Jessica Bon, MD; Divay Chandra, MD, MSc; Joel Weissfeld, MD, MPH

University of Texas Health, San Antonio, San Antonio, TX: Antonio Anzueto, MD; Sandra Adams, MD; Diego Maselli-Caceres, MD; Mario E. Ruiz, MD; Harjinder Singh

Figure 1

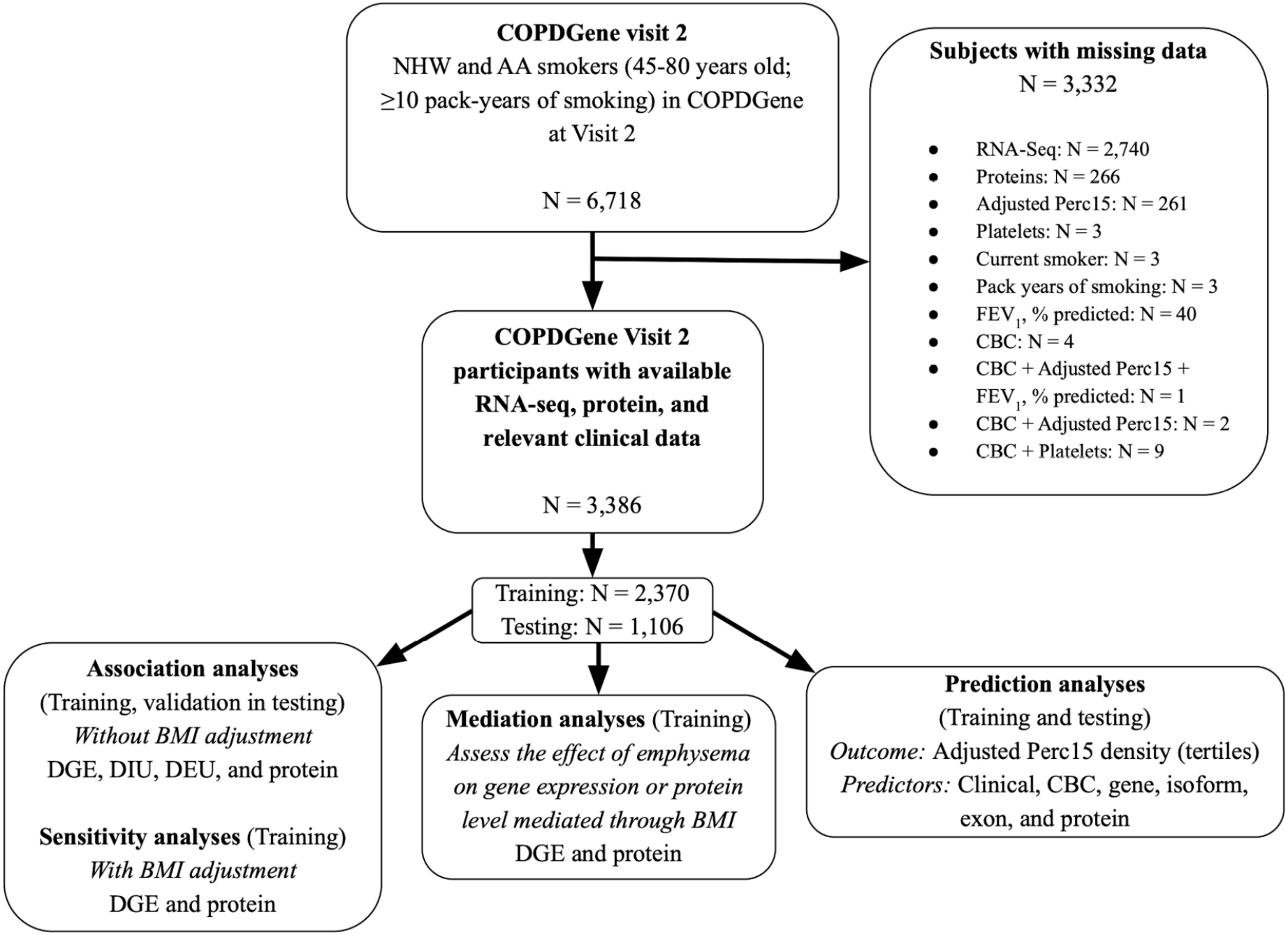
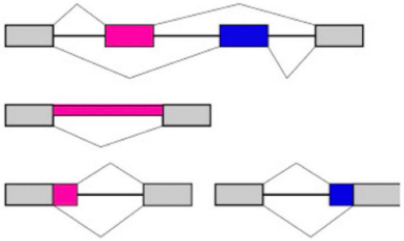
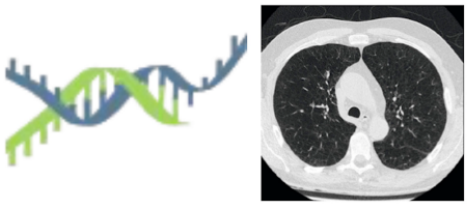

COPDGene visit 2
NHW and AA smokers (45-80 years old; ≥10 pack-years of smoking) in COPDGene at Visit 2
N = 6,718
Subjects with missing data
N = 3,332
RNA-Seq: N = 2,740
Proteins: N = 266
Adjusted Perc15: N = 261
Platelets: N = 3
Current smoker: N = 3
Pack years of smoking: N = 3
FEV1, % predicted: N = 40
CBC: N = 4
CBC + Adjusted Perc15 + FEV1, % predicted: N = 1
CBC + Adjusted Perc15: N = 2
CBC + Platelets: N = 9
COPDGene Visit 2 participants with available RNA-seq, protein, and relevant clinical data
N = 3,386
Training: N = 2,370
Testing: N = 1,106
Association analyses
(Training, validation in testing)
Without BMI adjustment
DGE, DIU, DEU, and protein
Sensitivity analyses (Training)
With BMI adjustment
DGE and protein
Mediation analyses (Training)
Assess the effect of emphysema on gene expression or protein level mediated through BMI
DGE and protein
Prediction analyses
(Training and testing)
Outcome: Adjusted Perc15 density (tertiles)
Predictors: Clinical, CBC, gene, isoform, exon, and protein

Figure 2



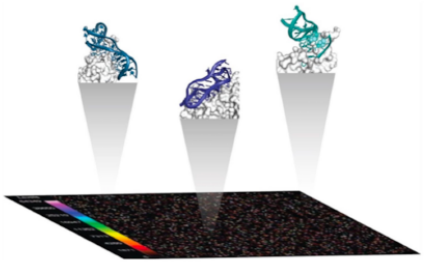
Differential gene expression, isoform/exon usage, and gene ontology enrichment analyses (Training)

- *Primary analysis*: Emphysema model without BMI adjustment
- *Sensitivity analysis*: Emphysema model with BMI adjustment (DGE only)



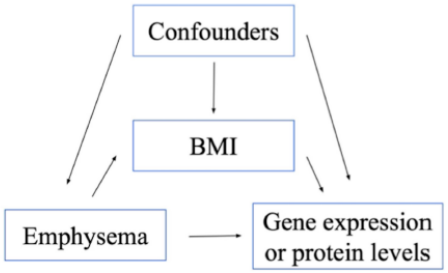
Clinical, transcriptomic, and proteomic data collection in COPDGene visit 2

- CT-quantified emphysema
- Whole blood RNA sequencing
- Plasma SomaScan proteomic assay

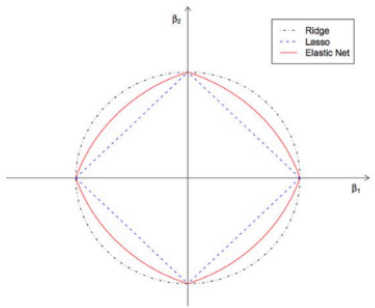


Protein association analyses (Training)

- *Primary analysis*: Emphysema model without BMI adjustment
- *Sensitivity analysis*: Emphysema model with BMI adjustment



Mediation analysis

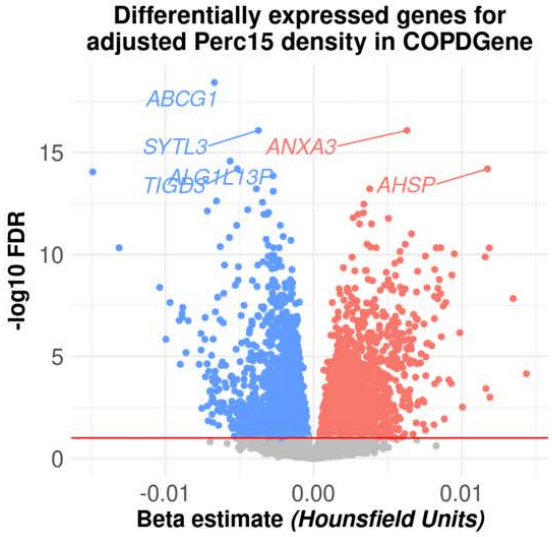


Elastic net prediction of emphysema
Clinical, CBC, gene, isoform, exon, and protein predictors

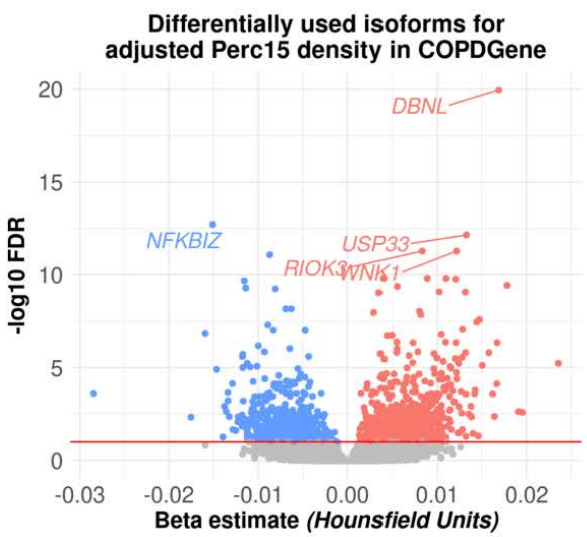
Validation analyses (Testing)
Differential gene expression, isoform/exon usage, and protein association analysis

- Emphysema model without BMI adjustment

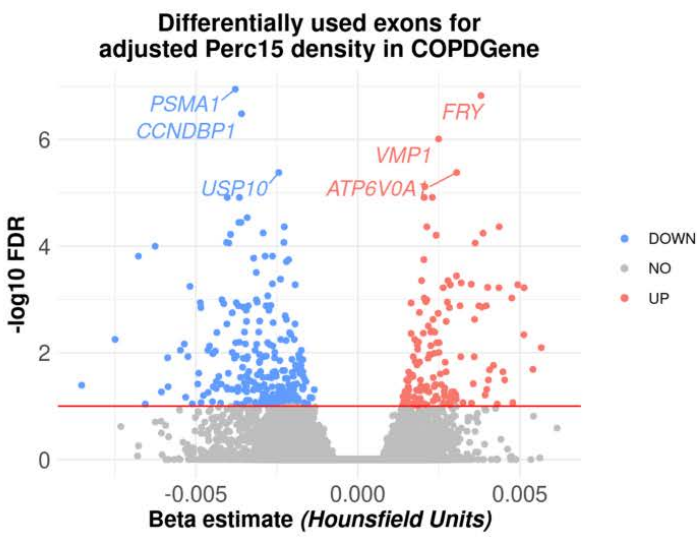
Figure 3


Differentially expressed genes for
adjusted Perc15 density in COPDGene
ABCG1
SYTL3
ANXA3
ALG1L13P
TIGD3
AHSP
-log10 FDR
15
10
5
0
-0.01
0.00
0.01
Beta estimate (Hounsfield Units)

(A)


Differentially used isoforms for
adjusted Perc15 density in COPDGene
DBNL
NFKBIZ
USP33
RIOK3
WNK1
-log10 FDR
20
15
10
5
0
-0.03
-0.02
-0.01
0.00
0.01
0.02
Beta estimate (Hounsfield Units)

(B)

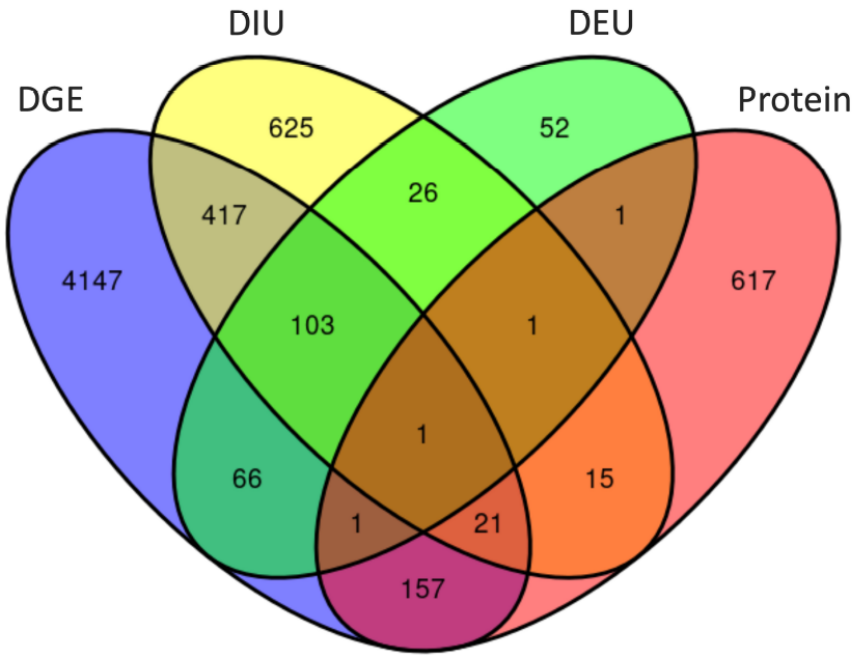

Differentially used exons for
adjusted Perc15 density in COPDGene
PSMA1
CCNDBP1
USP10
FRY
VMP1
ATP6V0A1
-log10 FDR
6
4
2
0
-0.005
0.000
0.005
Beta estimate (Hounsfield Units)
DOWN
NO
UP

(C)

Figure 4

Genes and GO terms significant for adjusted Perc15 density in DGE, DIU, DEU, and protein analyses

(A) GENES



(B) GO TERMS

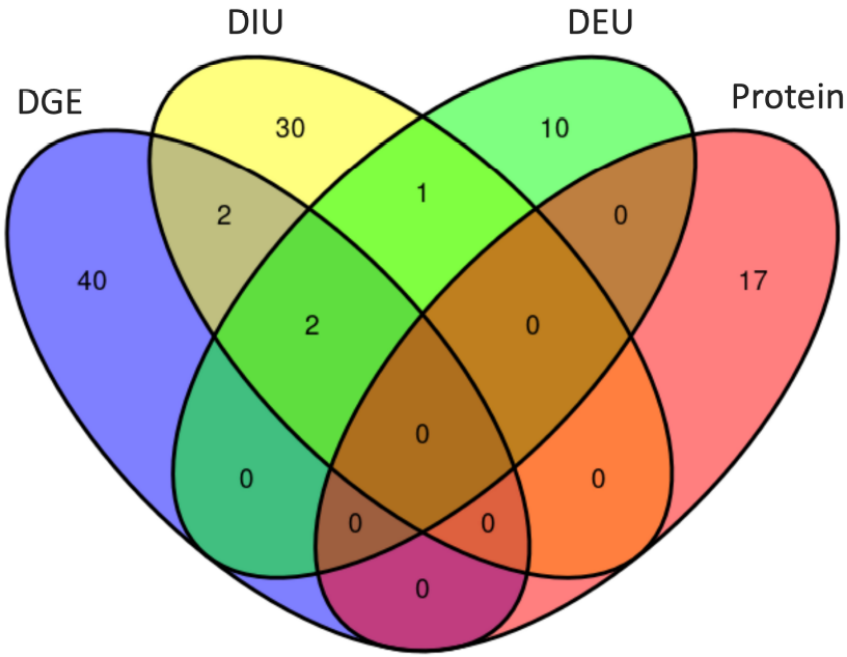
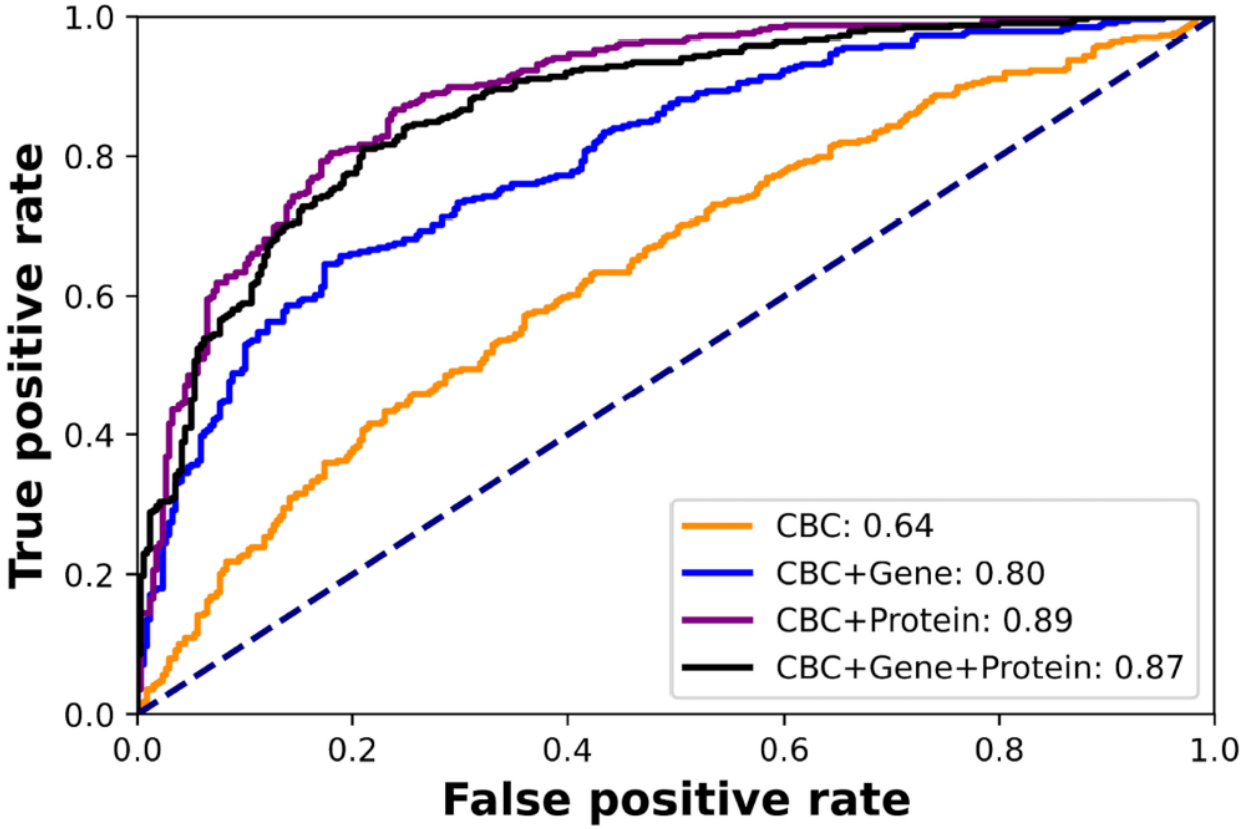


Figure 5

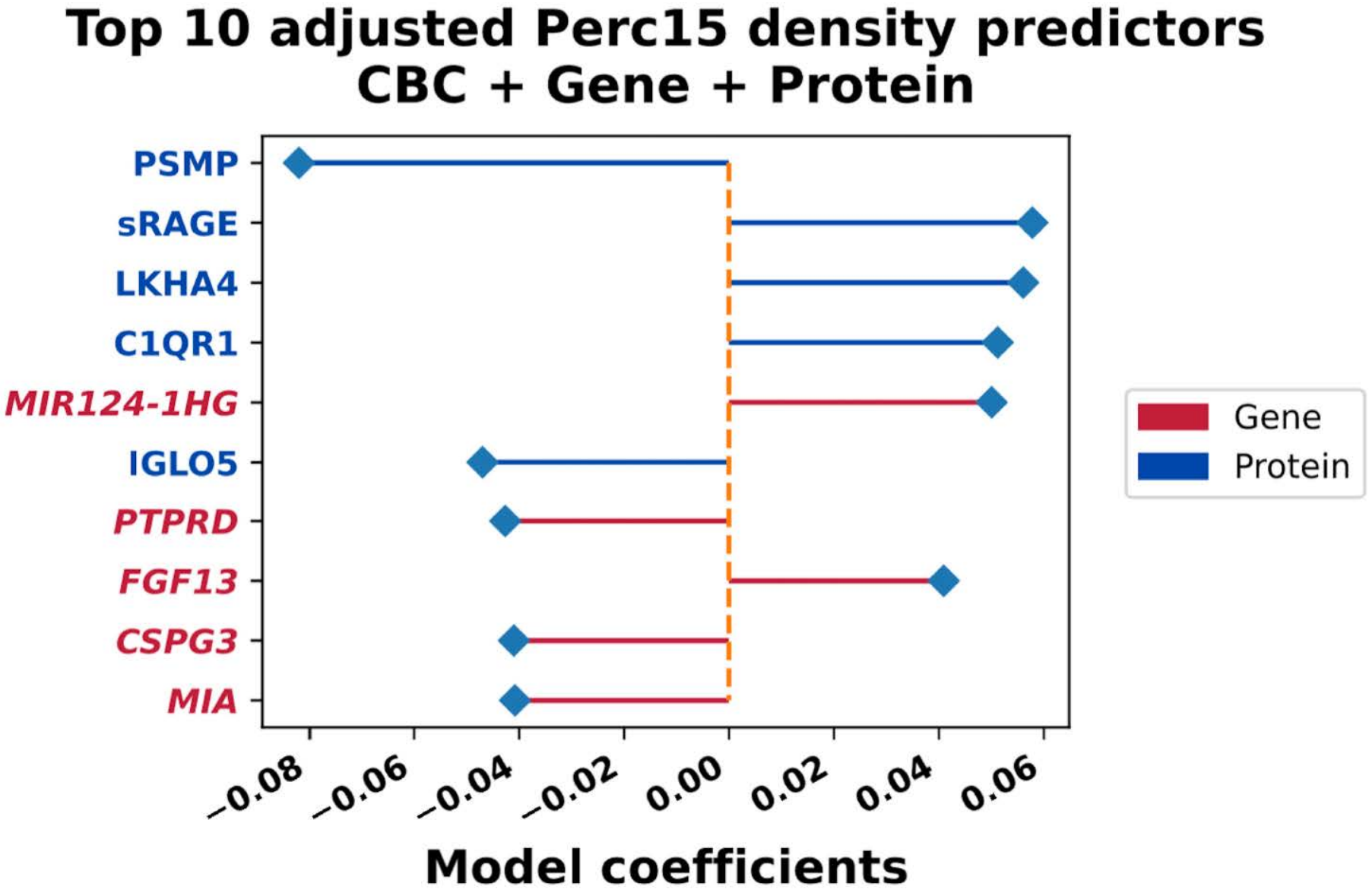
Model AUROC comparison

Upper versus Lower Tertiles of Adjusted Perc15



	CBC only	CBC+Gene	CBC+Protein
CBC+Gene	**< 0.05**		
CBC+Protein	**< 0.05**	**< 0.05**	
CBC+Gene+Protein	**< 0.05**	**< 0.05**	**< 0.05**

Figure 6


Top 10 adjusted Perc15 density predictors
CBC + Gene + Protein
PSMP
sRAGE
LKHA4
C1QR1
MIR124-1HG
IGLO5
PTPRD
FGF13
CSPG3
MIA
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
Model coefficients
Gene
Protein