Blood-Based Transcriptomic and Proteomic Biomarkers of 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 A. Pratte[5], Russell P. Bowler[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 (smlutz@hsph.harvard.edu), 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 A. Pratte (prattek@njhealth.org), Russell 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 K08 HL141601, K08 HL146972, K08 HL136928, K01 HL157613, R01 HL124233, U01 HL089897, R01 HL147326, R01 HL133135, P01 HL114501, U01 HL089897, and U01 HL089856. The COPDGene study

(NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprised of 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: 3,479

AT A GLANCE COMMENTARY (162/200 words)

Scientific Knowledge on the Subject:

Differential gene expression and protein analyses have uncovered some of the molecular underpinnings of emphysema. However, no studies have assessed alternative splicing mechanisms and analyzed proteomic data from 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 the large and well-phenotyped COPDGene study. In addition to confirming earlier findings, our differential gene expression, alternative splicing, and protein analyses identified novel biomarkers and pathways of chest CT-quantified emphysema. Our mediation analysis detected varying degrees of transcriptomic and proteomic mediation due

to BMI. Our supervised machine learning modeling demonstrated the utility of incorporating multi-omics data in enhancing the prediction of emphysema.

Keywords: Emphysema; Biomarkers; Transcriptomics; Proteomics; Prediction

This article has an online data supplement, which is accessible from this issue's table of content online at www.atjsjournals.org

ABSTRACT

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

Objectives: Discover blood omics biomarkers for chest CT-quantified emphysema 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 low 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. Elastic net models were also developed in the training sample sequentially using clinical, complete blood count (CBC) cell proportions, RNA sequencing, and proteomic biomarkers to predict quantitative emphysema. Model accuracy was assessed in the testing sample by the 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%)* and yielded 109 biological pathways. 75% of the genes and 77% of the proteins associated with emphysema showed evidence of mediation by BMI. The highest-performing predictive model used

clinical, CBC, and protein biomarkers, distinguishing the top from the bottom tertile of emphysema with an AUROC of 0.92.

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

Abstract word count: 250/250

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality (1). Emphysema, the anatomic destruction of lung parenchyma frequently observed in COPD subjects, has been independently associated with an increased risk for cardiovascular disease, lung cancer, and mortality (2-4). 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. Emphysema blood biomarkers would also overcome the issues of radiation exposure and false positive findings associated with CT scans (5). In addition, early disease biomarkers and a stronger understanding of the molecular basis of emphysema are needed to develop novel personalized therapies to improve the prognosis of affected individuals (2, 6, 7).

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) primarily from gene expression analyses using lung tissue samples (8-12). However, fewer studies have evaluated the associations of emphysema with blood transcriptomics, alternative splicing, or proteomics. 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 (13-19). Protein levels have also been studied for potential emphysema biomarker identification, and sRAGE, ICAM1, CCL20, and adiponectin levels in blood and eotaxin levels in bronchoalveolar lavage fluid were found to be associated with emphysema (20-23), though the protein panels used for these studies included fewer proteins than the more recently developed panels. Finally, previous research that used blood-based emphysema predictive models had small sample sizes and only tested one ‘omic modality at a time (20, 24-27).

We hypothesized that 1) transcriptomic and proteomic characterization of smokers would elucidate emphysema pathobiology and yield novel disease biomarkers, 2) many emphysema associations with transcripts and proteins are influenced by BMI, and 3) multi-omic modeling would provide improved prediction of emphysema relative to readily available clinical variables. To test these hypotheses, we analyzed whole-blood genome-wide RNA sequencing (RNA-seq) and plasma SomaScan proteomic data from the large and well-phenotyped COPDGene study. Given the high clinical correlation between emphysema and BMI (28), we performed mediation analysis to understand the influence of BMI on emphysema-associated genes and proteins. We also developed machine learning predictive models for emphysema using transcriptomic and proteomic biomarkers. Some of these results have been previously reported as an abstract (29) and a preprint (30).

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 subjects, 44-90 years old with an average of 44 pack-years of lifetime cigarette smoking history (31). 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. 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 3,386 subjects with available clinical, RNA-seq, and SomaScan proteomic data.

Emphysema quantification

Using the Thirona software (www.thirona.eu), emphysema was quantified as the 15th percentile of the attenuation histogram + 1,000 Hounsfield units (HU), corrected for the inspiratory depth variations using Multi-Ethnic Study of Atherosclerosis normative equations (predicted lung volume using baseline age, time-varying height, and BMI) (adjusted Perc15 density) (32-34). This correction was made since it had been demonstrated to provide a more robust measure of longitudinal changes in emphysema (33).

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 using the training sample. The validation of the identified biomarkers and constructed predictive model 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 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 (37). Samples with low volume, failed hybridization control, or failed dilution scale were removed. Proteomic data for 5,670 participants passed quality control. The protein counts were transformed to log2 RFU (relative fluorescent units) 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 (38, 39). 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), complete blood count (CBC) cell proportions, CT scanner model, and library preparation batch for RNA-seq or 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. A sensitivity analysis was performed in which the list of covariates from the primary model was expanded to include BMI. To select biomarkers for inclusion in the prediction model, we ran additional models only adjusted for the technical factors (CT scanner model and library preparation batch for RNA-seq or 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% (40).

Mediation analysis

We conducted mediation analysis using the medflex R package (41) to distinguish how much of the effect of emphysema on gene expression or protein levels acted through BMI (referred to as the indirect effect) and how much of the effect of emphysema directly influenced gene expression or protein levels (referred to as the direct effect). The analysis was performed on the significant genes and proteins identified in the primary association analysis. A mediated proportion representing the ratio of the indirect effect over the total effect was computed for each gene with significant total effect. The P-values of the direct, indirect, and total effects for each biomarker were subject to a threshold significance of 10% FDR.

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 (42). We only reported GO pathways with at least three significant genes and an adjusted P-value < 0.005.

176

177 ***Predictive modeling***

178 We constructed elastic net models to predict cross-sectional emphysema (43). The
179 outcome variable was the adjusted Perc15 density. We utilized clinical variables that are
180 readily available in the primary care setting (age, race, sex, BMI, pack-years of smoking, and
181 current smoking status), CBC (proportions of neutrophils, eosinophils, monocytes,
182 lymphocytes, and platelets), and the RNA-seq and proteins that reached statistical
183 significance in the association analyses performed in the training data (adjusted only for the
184 scanner model and library preparation batch or clinical center). To determine the top
185 performing RNA data type to be used in the main models, we first constructed models using
186 clinical + gene, clinical + isoform, and clinical + exon counts. We then constructed models in
187 the following order: clinical only, clinical + CBC, clinical + CBC + RNA-seq, clinical +
188 CBC + proteins, and clinical + CBC + RNA-seq + proteins. The outcome and the predictors
189 were centered and scaled. The models were trained using 10-fold cross-validation,
190 minimizing the mean squared error (44) on the left-out fold. After model training on the
191 continuous emphysema variable, we classified subjects into tertiles of adjusted Perc15
192 density. We evaluated the predictive performances of the models in the testing sample using
193 R^2 for the continuous emphysema and the area-under-receiver-operator-characteristic curve
194 (AUROC) for the model accuracy to distinguish subjects in the highest and lowest tertiles of
195 emphysema severity. We compared the AUROCs with the DeLong test using the pROC R
196 package (45). Finally, predictors were ranked by the absolute values of their coefficients from
197 the regression model.

198

199 ***Statistical analysis***

Data were reported as means with standard deviations or counts with percentages. Continuous variables were tested with Kruskal-Wallis and categorical variables were tested 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., negative coefficients indicate a greater extent of emphysema).

Additional methods are available in the Supplement.

RESULTS

Subject characteristics

3,386 subjects from COPDGene Visit 2 with complete clinical, RNA-seq, and protein data 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 (Table E1). A schematic overview of the analyses performed is illustrated in Figure E1.

Differential gene expression analysis

We performed differential gene expression (DGE) analysis in the 2,370 subjects of the training dataset. 4,913 out of 19,177 genes reached significance at 10% FDR (Table 2 and E2). 2,339 genes were up-regulated and 2,574 were down-regulated with respect to adjusted Perc15 density (i.e., they have opposite directions for their associations with

emphysema) (Figure 2A). The GO enrichment analysis identified 44 significantly enriched biological processes, including neutrophil degranulation, regulation of NF-κB signaling, viral transcription, T cell proliferation, and regulation of TNF-mediated signaling (Tables 3 and 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%)* (Table 2). The differentially used isoforms mapped to 1,209 individual genes (Table E4), 45% of which (542/1,209) were also identified in the DGE analysis. The differentially used exons mapped to 251 genes (Table E5), 68% of which (171/251) were also differentially expressed. 53% (788/1478) of the significant isoforms and 66% (244/368) of the significant exons were up-regulated with respect to adjusted Perc15 density (Figure 2B, 2C). The GO enrichment analyses performed on the differentially used isoforms and differentially used exons yielded 35 and 13 significantly enriched biological processes, respectively. Top processes included mitophagy, regulation of NF-κB signaling, negative regulation of WNT signaling, and viral transcription (Table 3, E6 E7).

Protein association analysis

We tested 4,979 SomaScan proteins measured in the training dataset using multivariable linear regression. 18% (881/4,979) were significantly associated with emphysema *(FDR 10%)* (Table E8, Figure E2). Seventeen significantly enriched biological processes were identified, including pathways related to classical complement pathway and

WNT signaling (Table E9). Figure 3 summarizes the overlap of the biomarkers and GO terms between DGE, DIU, DEU, and protein analyses, showing that most of the significant biomarkers and enriched pathways are unique to each analysis.

Validation

We analyzed 1,016 subjects with RNA-seq and proteomic data in the testing sample to provide independent validation of the emphysema biomarkers identified in the training sample. We observed that the effect sizes were highly correlated between the analyses performed in the training and testing data for the 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 and checking if the training and testing data had a consistent direction of effect. Respectively, 46% (2,252/4,913), 30% (449/1,478), 60% (233/368), and 47% (416/881) of the DGE, DIU, DEU, and protein biomarkers were validated (Tables E2, E4, E5, and E8).

Mediation analysis

Since severe emphysema is often associated with low BMI, we performed sensitivity analyses that also adjusted for BMI. We observed that 96% (4,728/4,913) of the genes and 80% (703/881) of the proteins (Figure E3) associated with emphysema from the primary analysis were no longer significant after adjustment for BMI (Tables E10 and E11). These observations suggest that BMI mediates many of the emphysema-associated transcriptomic and proteomic changes. To investigate this, we performed mediation analysis based on the directed acyclic graph (DAG) (Figure E4) to divide each observed biomarker association into a direct pathway (emphysema directly affects gene/protein expression) and an indirect

pathway (emphysema affects gene/protein expression via its effects on BMI). Of the 4,913 differentially expressed genes and 881 proteins that reached significance in the primary model (i.e., no BMI adjustment), 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 *(FDR 10%)* (Tables E12 and E13).

Prediction

To select blood biomarkers for inclusion in cross-sectional predictive models for emphysema, we performed association analyses in the training dataset, adjusting only for technical factors, which yielded 13,066 genes, 4,254 isoforms, 2,263 exons, and 1,719 proteins that were used as candidate predictors. To evaluate whether RNA-seq is more informative at the gene, isoform, or exon level, we trained three separate models (clinical + gene, clinical + isoform, and clinical + exon). The AUROCs were 0.86, 0.74, and 0.86, respectively (Table E14 and Figure E5). Although no statistical significance was attained, a slightly higher AUROC was achieved with the genes compared to exons. Accordingly, we focused on gene-level quantifications for the subsequent models. We next evaluated the relative contribution of CBC proportions, genes, and proteins compared to a baseline model using clinical variables alone. The model using only clinical variables explained 35% of the variance of emphysema. Substantial improvement was seen from adding gene and protein data respectively (R^2 = 0.38 for clinical + CBC + gene and 0.50 for clinical + CBC + protein, Table E14). The model with clinical + CBC + gene + protein data did not perform as well as the model with clinical + CBC + protein data. The same pattern was seen when we evaluated model performance for distinguishing subjects in the top versus bottom emphysema tertile; the highest-performing model was the clinical + CBC + protein model with an AUROC of

0.92. Figure 4 summarizes the model results, and Table E14 summarizes the AUROCs, alphas, and L1 ratios.

Ranked by absolute beta coefficients, the top-10 predictors of the all-inclusive (clinical + CBC + gene + protein) model included BMI, sRAGE, and two biomarkers that have not been previously linked to emphysema (the *MIR124-1HG* gene and the PSMP protein) (Figure 5).

DISCUSSION

In this study, we performed the largest blood transcriptomic and proteomic profiling of CT-quantified emphysema to date, including investigations into alternative splicing mechanisms, identifying thousands of validated blood biomarker associations. The biological relevance of these findings was assessed through GO pathway analyses, which mostly demonstrated enrichment for inflammatory pathways and cell differentiation. Mediation analysis revealed that 70% of the differentially expressed genes and 61% of the associated proteins are mediated through BMI, implying that blood biomarker associations to emphysema largely reflect shared biological processes with BMI. We also demonstrated the utility of incorporating multi-omics data in enhancing the prediction of emphysema.

In previous biomarker studies of emphysema, the extracellular matrix (ECM), NF-κB, transforming growth factor beta (TGF-ß), B cell antigen receptor (BCR), and oxidative phosphorylation pathways were among the most frequently reported emphysema-associated pathways (8, 46, 47). However, most of these studies focused on a single 'omics modality (20, 24, 48). Our investigation of blood-based transcriptomic and proteomic biomarkers supported numerous established emphysema-associated pathways and discovered new ones. In addition, our alternative splicing analysis for the first time revealed widespread evidence of alternative splicing associated with emphysema.

Most blood biomarker associations with emphysema occur through BMI, as indicated by the significant mediation of the tested genes and proteins. This suggests that some of the molecular processes identified in this analysis may be causally related to both emphysema and BMI. We must keep in mind, though, that our 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. While our specified DAG is reasonable based on prior knowledge, there are other plausible alternatives DAGs, but no currently available methods to simultaneously test these possibilities.

CT scan is the best non-invasive method for detecting emphysema. However, CT has several drawbacks, including increased costs, radiation exposure, and high rates of unrelated false-positive findings (5). 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 (49-52). As the first study to utilize genes, alternative splicing, and proteins combined with clinical and CBC predictors, we developed models that could classify upper and lower tertiles of emphysema severity with good accuracy. While alternative splicing predictors were worth exploring, gene data had a higher AUROC. While genes outperformed clinical and CBC features, protein predictors yielded the best AUROC across all models.

From the top 10 predictors of the clinical + CBC + gene + protein model, sRAGE, which minimizes tissue injury and inflammation, has consistently been recognized as a candidate emphysema biomarker (5, 10, 53, 54). Not previously connected to emphysema, PSMP has been implicated in inflammation and cancer development (55, 56) and *MIR124*-

1HG has been shown to affect Wnt-signaling and inflammation (57-60).Their putative roles and functions in emphysema require further investigation.

This study has several strengths. The large sample size allowed us to identify many more significant associations than any previous study, and we could split our sample to allow for the validation of our findings. This is the first study that has examined alternative splicing mechanisms in emphysema in addition to differential gene expression and protein association analyses. Because emphysema often co-occurs with low BMI, we performed mediation analyses to better understand the relationship between molecular markers, emphysema, and BMI, providing suggestive evidence of shared biology between emphysema and BMI. Finally, we were able to improve emphysema prediction models with the use of multi-omic data.

This study also has several limitations. Complete blood count quantifications do not capture the variability of the immune cell subpopulations, which limits the ability to localize these effects to specific cell types. Future studies may address this by using single-cell transcriptomics data. Limitations to the SomaScan proteomics include the lack of SOMAmers for small molecules such as desmosine, fibrinogen degradation product (Aa-Val360, a specific product generated by elastase cleavage of fibrinogen), and sphingomyelin, which have been suggested to be emphysema biomarkers in other studies (37, 60-62). Lastly, the mediation analysis needs to be viewed as hypothesis-generating since it is based on a number of assumptions.

CONCLUSION

Our transcriptomic and proteomic analyses yielded numerous inflammatory and cell differentiation pathways connected to emphysema as well as novel potential blood biomarkers of the disease. While not yet ready to be used in clinical practice, with further

validation, our prediction model might be helpful as a less invasive indicator of emphysema severity that could guide patient enrollment in clinical trials. Future research is necessary to compare blood and lung tissue biomarkers, understand how they change as emphysema progresses, and evaluate the impact of implementing these 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 (L)	2.22 (0.84)	2.21 (0.85)	0.75
FEV_1, % predicted	80.57 (24.35)	80.52 (24.33)	0.95
FVC (L)	3.20 (0.95)	3.19 (0.96)	0.88
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
Pi10	2.24 (0.57)	2.23 (0.56)	0.54
GOLD grade			
PRISm	299 (12.62%)	120 (11.81%)	
Normal spirometry	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 history (%)	14.96%	15.03%	0.73
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%)	0.78
2	276 (11.65%)	117 (11.52%)	
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

Participant characteristics reported here are from Visit 2, when 'omics data were obtained.

Continuous variables are expressed as means and standard deviations. Categorical variables are expressed as absolute values and/or percentages.

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 HU + 1000); CAD: Self-reported history of 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 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; 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 HU + 1000). The lower the Perc15 values are, the more CT-quantified emphysema is present.

For the DGE, DIU, and DEU analyses, the covariates used were age, race, sex, pack-years of smoking, current smoking status, forced expiratory volume in one second (FEV_1), CBC cell count proportions, library preparation batch, and CT scanner model. A threshold of FDR 10% was applied.

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.

Table 3. Selected top 10 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	$1x10^{-4}$
	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 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 (FEV_1), CBC cell count proportions, library preparation batch, and CT scanner model.

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: AA = African American, CBC = Complete blood count, DGE = Differential gene expression, DIU = Differential isoform usage, DEU = Differential exon usage, FEV_1 = Forced expiratory volume, NHW = Non-Hispanic White.

Figure 2. Volcano plots of the primary model representing (A) differentially expressed genes (B) differentially used isoforms, and (C) differentially used exons. Genes significantly associated with adjusted Perc15 density appear above the red line marked at FDR 10%. Up-regulated genes are in blue and down-regulated genes 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 HU + 1000). The lower the Perc15 values are, 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 3. (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. HUGO gene symbols were used to find the intersection of biomarkers between the DGE, DIU, DEU, and protein analyses. Multiple proteins may map to a single gene. Therefore, the diagram does not reflect the total number of proteins significantly associated with adjusted Perc15 density. (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 HU + 1000). The lower the Perc15 values are, 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 4. The receiver operating characteristic curves for the elastic net prediction models: clinical (age, race, sex, BMI, pack-years of smoking, and current smoking status) only, clinical + complete blood count (CBC) proportions of neutrophils, eosinophils, monocytes, lymphocytes, and platelets, clinical + CBC + genes, clinical + CBC + proteins, and clinical + CBC + genes + proteins. The table summarizes the pairwise DeLong P-values of the model comparisons. P-values < 0.05 are bolded.

Figure 5. Top 10 predictors sorted in descending order by the absolute values of their beta-coefficients from the elastic net model using clinical (age, race, sex, BMI, pack-years of smoking, and current smoking status), complete blood count (CBC) proportions of neutrophils, eosinophils, monocytes, lymphocytes, and platelets, 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. 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.

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

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

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

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

29. 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.

30. Suryadevara R, Gregory A, Lu R, Xu Z, Masoomi A, Lutz SM, Berman S, Yun JH, Saferali A, Hersh CP, Silverman EK, Dy J, Pratte K, Bowler RP, Castaldi PJ, Boueiz A. Blood-Based Transcriptomic and Proteomic Biomarkers of Radiologic Emphysema. *medRxiv* 2022: 2022.2010.2025.22281458.

31. 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.

32. 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.

33. 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.

34. Pompe E, Strand M, van Rikxoort EM, Hoffman EA, Barr RG, Charbonnier JP, Humphries S, Han MK, Hokanson JE, Make BJ, Regan EA, Silverman EK, Crapo JD, Lynch DA. Five-year Progression of Emphysema and Air Trapping at CT in Smokers with and Those without Chronic Obstructive Pulmonary Disease: Results from the COPDGene Study. *Radiology* 2020; 295: 218-226.

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. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

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

45. 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.

46. 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.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. She S, Wu X, Zheng D, Pei X, Ma J, Sun Y, Zhou J, Nong L, Guo C, Lv P, Song Q, Zheng C, Liang W, Huang S, Li Q, Liu Z, Song Z, Li Y, Zhang Y, Kong W, You H, Xi J, Wang Y. PSMP/MSMP promotes hepatic fibrosis through CCR2 and represents a novel therapeutic target. *J Hepatol* 2020; 72: 506-518.
56. Pei X, Sun Q, Zhang Y, Wang P, Peng X, Guo C, Xu E, Zheng Y, Mo X, Ma J, Chen D, Zhang Y, Zhang Y, Song Q, Guo S, Shi T, Zhang Z, Ma D, Wang Y. PC3-secreted microprotein is a novel chemoattractant protein and functions as a high-affinity ligand for CC chemokine receptor 2. *J Immunol* 2014; 192: 1878-1886.
57. 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.
58. 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.
59. Ortapamuk H, Naldoken S. Brain perfusion abnormalities in chronic obstructive pulmonary disease: comparison with cognitive impairment. *Ann Nucl Med* 2006; 20: 99-106.
60. Cantor J, Ochoa A, Ma S, Liu X, Turino G. Free Desmosine is a Sensitive Marker of Smoke-Induced Emphysema. *Lung* 2018; 196: 659-663.
61. Manon-Jensen T, Langholm LL, Rønnow SR, Karsdal MA, Tal-Singer R, Vestbo J, Leeming DJ, Miller BE, Bülow Sand JM. End-product of fibrinogen is elevated in

emphysematous chronic obstructive pulmonary disease and is predictive of mortality in the ECLIPSE cohort. *Respir Med* 2019; 160: 105814.

62. Bowler RP, Jacobson S, Cruickshank C, Hughes GJ, Siska C, Ory DS, Petrache I, Schaffer JE, Reisdorph N, Kechris K. Plasma sphingolipids associated with chronic obstructive pulmonary disease phenotypes. *Am J Respir Crit Care Med* 2015; 191: 275-284.

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, PhD

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, MD

Figure 1

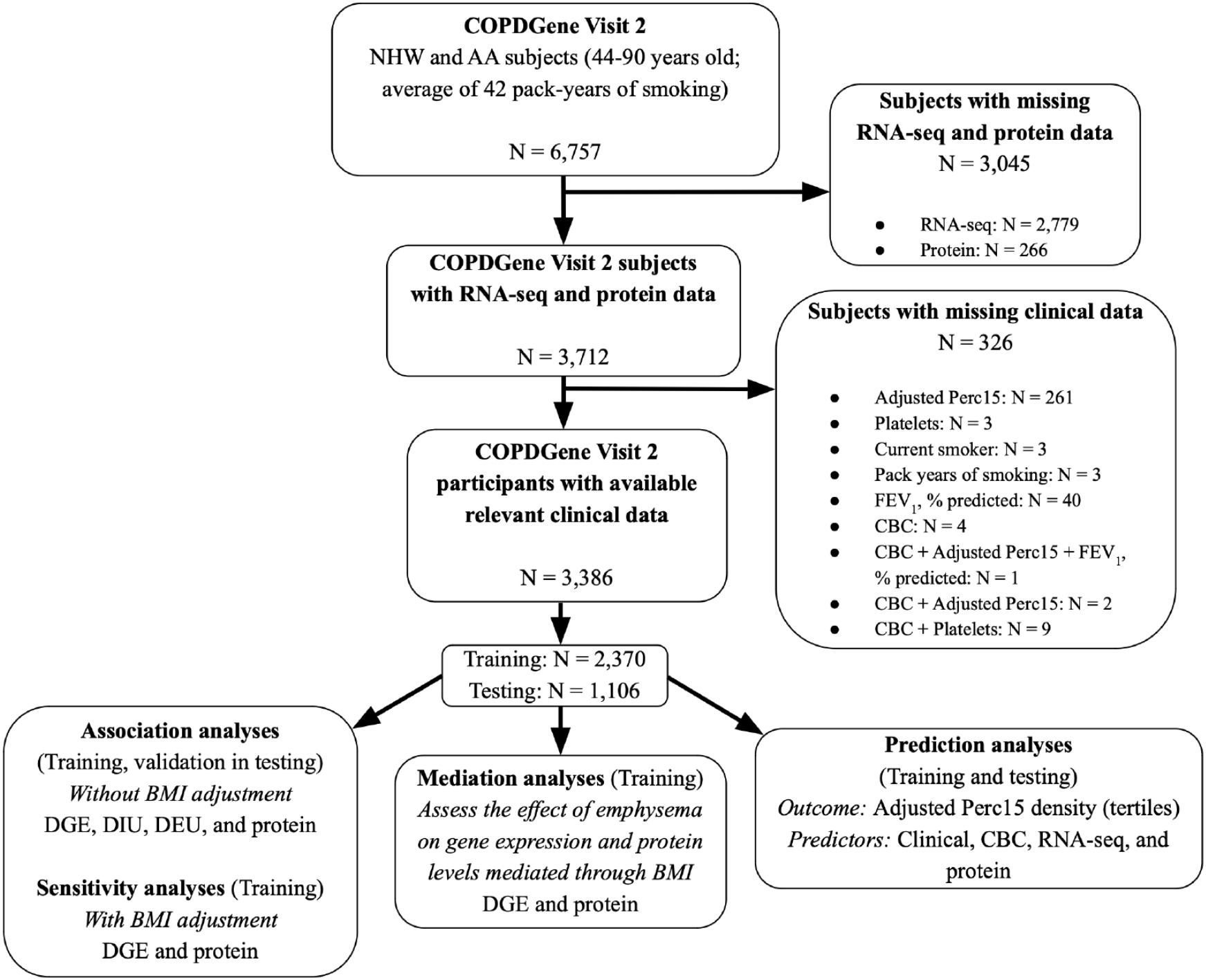
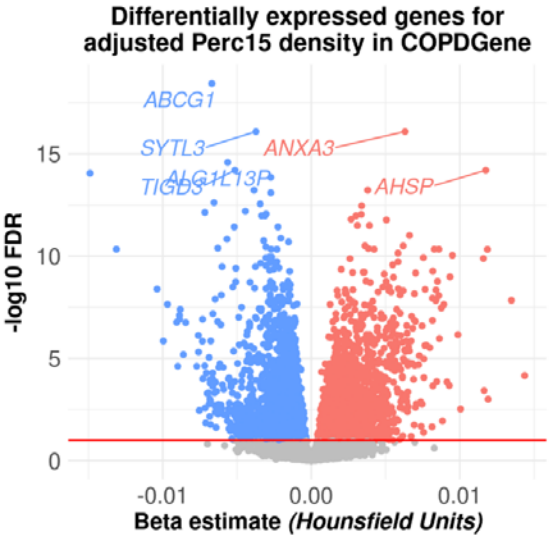
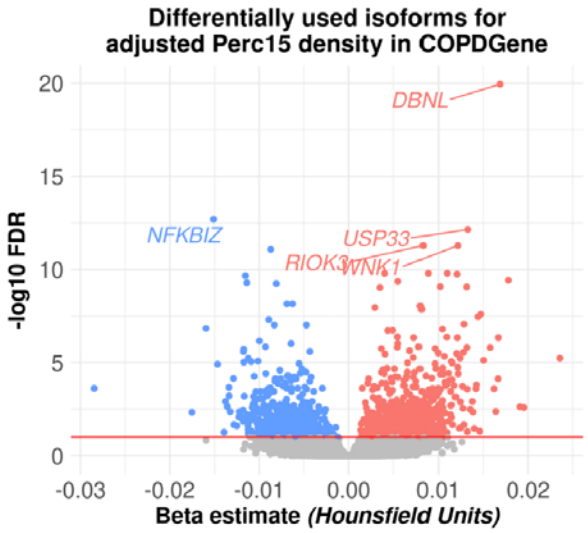

COPDGene Visit 2
NHW and AA subjects (44-90 years old; average of 42 pack-years of smoking)
N = 6,757
Subjects with missing RNA-seq and protein data
N = 3,045
RNA-seq: N = 2,779
Protein: N = 266
COPDGene Visit 2 subjects with RNA-seq and protein data
N = 3,712
Subjects with missing clinical data
N = 326
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 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 and protein levels mediated through BMI
DGE and protein
Prediction analyses
(Training and testing)
Outcome: Adjusted Perc15 density (tertiles)
Predictors: Clinical, CBC, RNA-seq, and protein

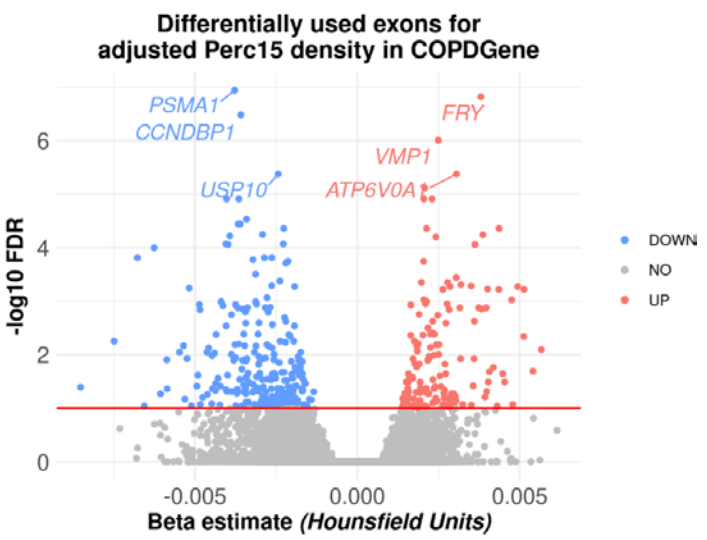
Figure 2


Differentially expressed genes for adjusted Perc15 density in COPDGene
ABCG1
SYTL3
ANXA3
ALG1L13P
TIGD5
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)

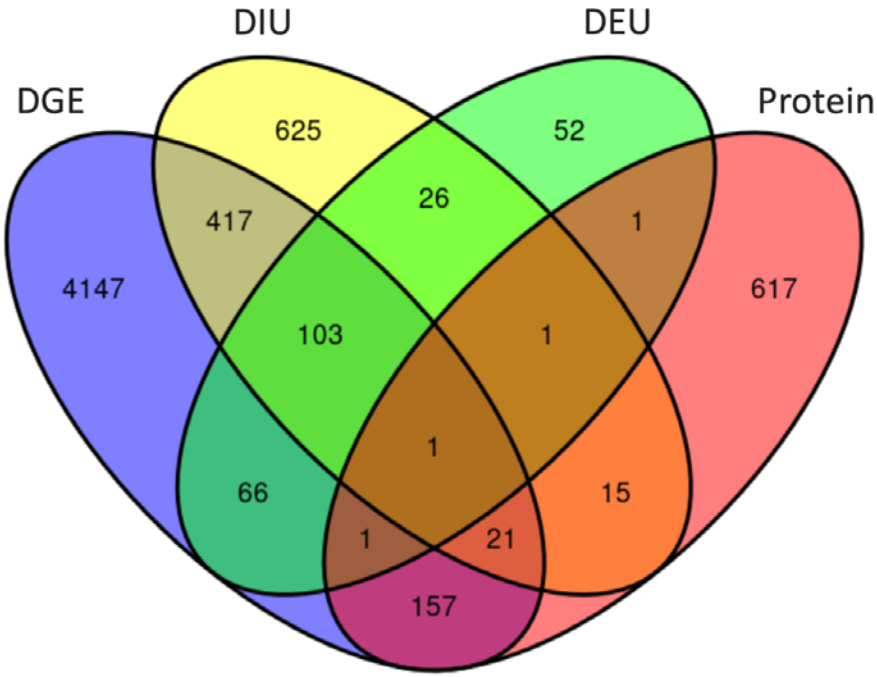

Differentially used exons for adjusted Perc15 density in COPDGene
PSMA1
CCNDBP1
USP10
FRY
VMP1
ATP6V0A
-log10 FDR
6
4
2
0
-0.005
0.000
0.005
Beta estimate (Hounsfield Units)
DOWN
NO
UP

(C)

Figure 3

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

(A) GENES



(B) GO TERMS

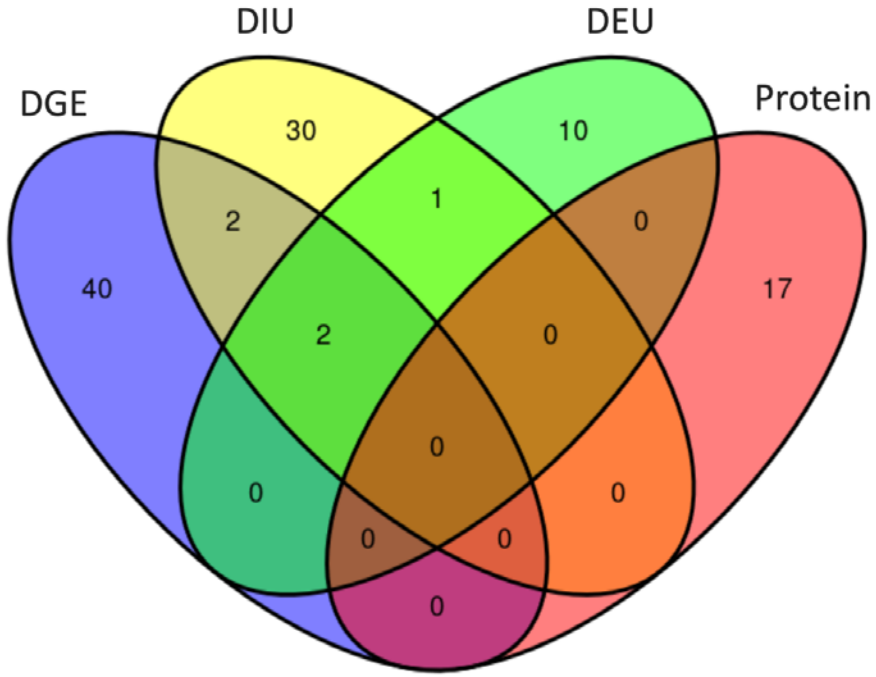
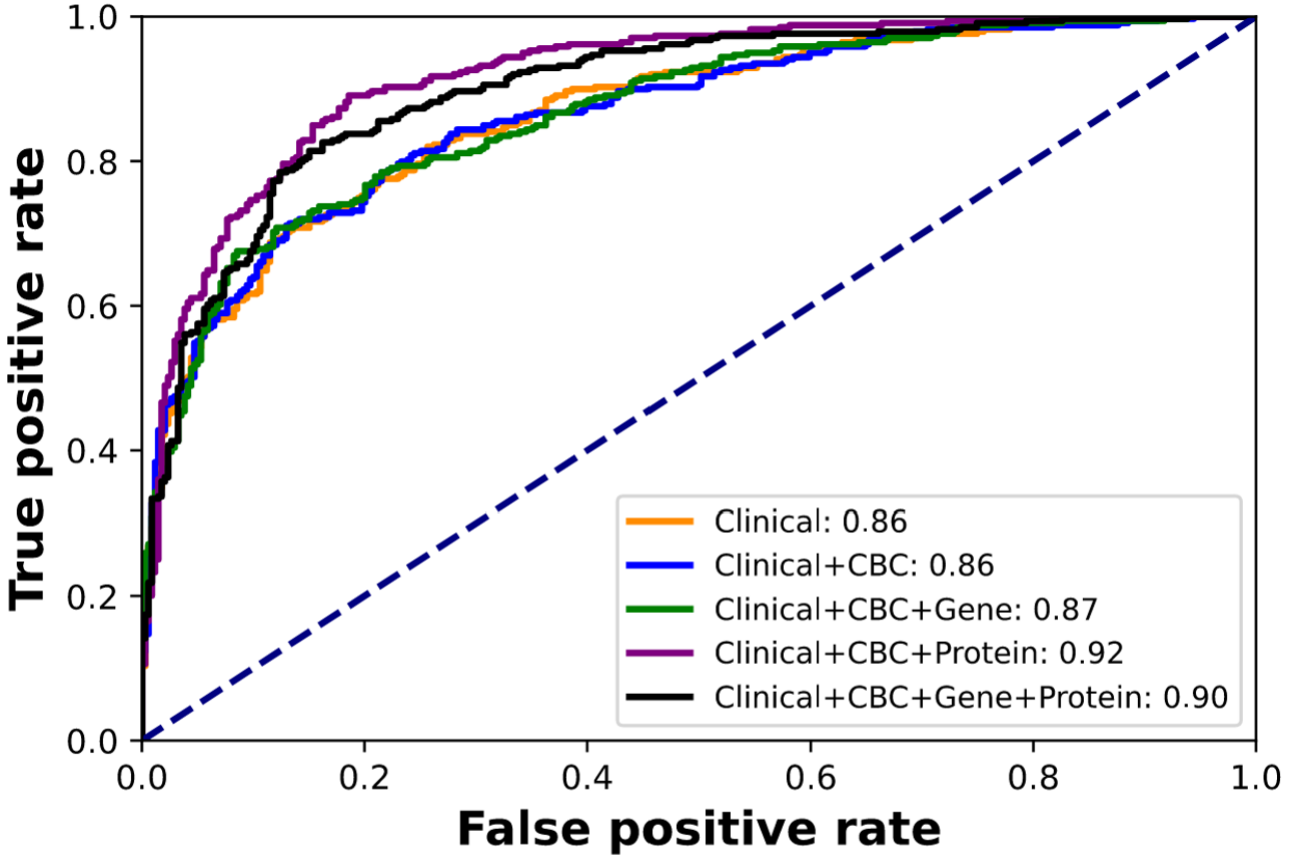


Figure 4

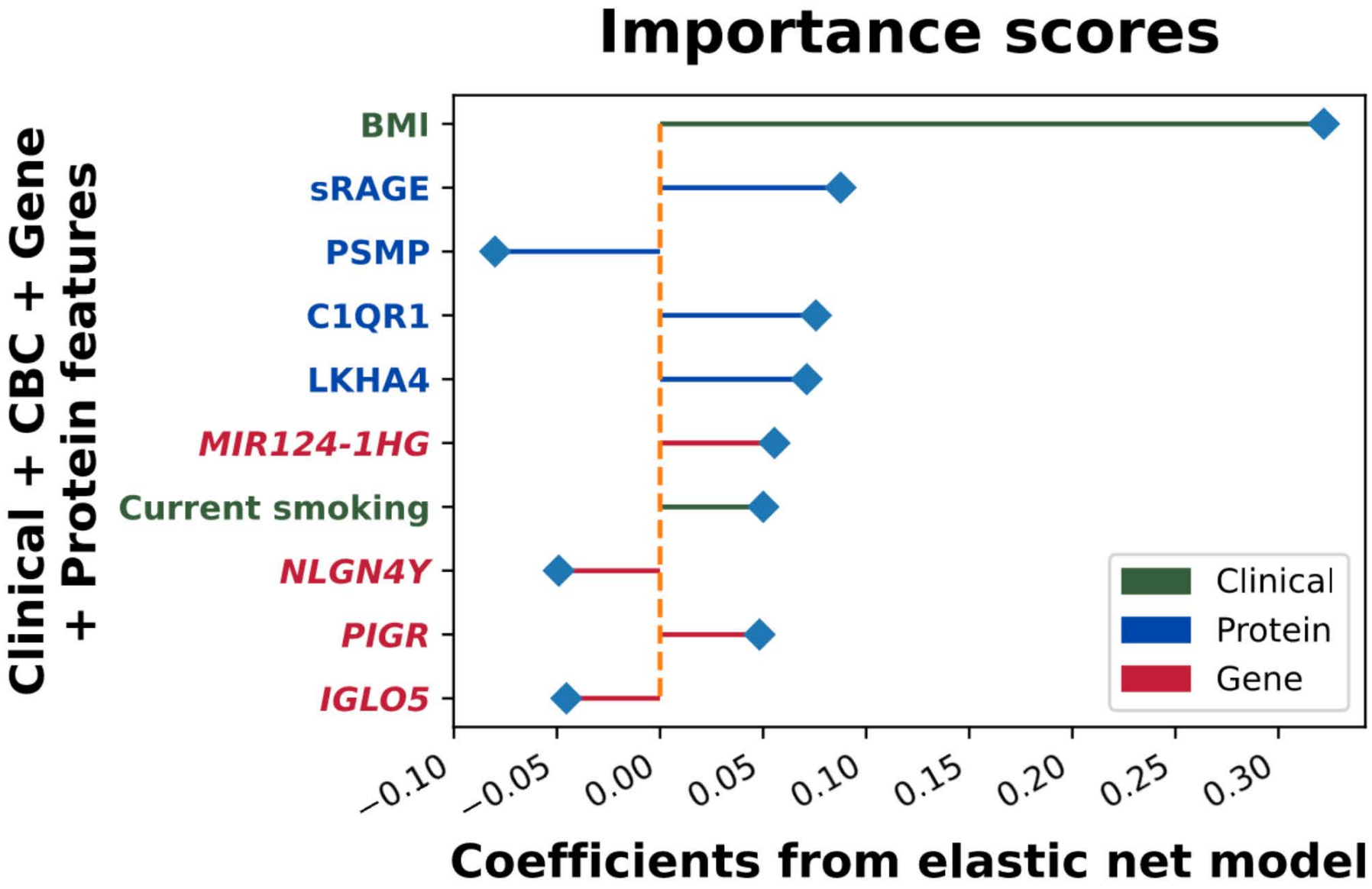
Model AUROC comparison

Upper versus lower tertiles of adjusted perc15 density



	Clinical	Clinical+CBC	Clinical+CBC+Gene	Clinical+CBC+Protein
Clinical+CBC	0.90			
Clinical+CBC+Gene	0.85	0.78		
Clinical+CBC+Protein	**< 0.001**	**< 0.001**	**< 0.001**	
Clinical+CBC+Gene+Protein	**0.0019**	**0.0012**	**< 0.001**	**0.0088**

Figure 5


Importance scores
Clinical + CBC + Gene + Protein features
BMI
sRAGE
PSMP
C1QR1
LKHA4
MIR124-1HG
Current smoking
NLGN4Y
PIGR
IGLO5
Clinical
Protein
Gene
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Coefficients from elastic net model