Polygenic and transcriptional risk scores identify chronic obstructive pulmonary disease subtypes

Matthew Moll1,2,3,24, Julian Hecker1,24, John Platig4, Jingzhou Zhang5, Auyon J. Ghosh6, Katherine A. Pratte7, Rui-Sheng Wang1, Davin Hill8, Iain R. Konigsberg9, Joe W. Chiles III10, Craig P. Hersh1,2,24, Peter J. Castaldi1,10,24, Kimberly Glass1,24, Jennifer G. Dy8, Don D. Sin12, Ruth Tal-Singer13, Majd Mouded14, Stephen I. Rennard15, Gary P. Anderson16, Gregory L. Kinney17, Russell P. Bowler18, Jeffrey L. Curtis19,20, Merry-Lynn McDonald9,21,22, Edwin K. Silverman1,2,24, Brian D. Hobbs23†, Michael H. Cho1,2,24†

Affiliations:
1 Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital, Boston, MA, 02115.

2 Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women’s Hospital, Boston, MA, 02115.

3 Division of Pulmonary, Critical Care, Sleep and Allergy. Veterans Affairs Boston Healthcare System, West Roxbury, MA, 02123

4 Center for Public Health Genomics, University of Virginia, Charlottesville, VA, 22903

5 The Pulmonary Center, Boston University Medical Center, Boston, MA 02118

6 Division of Pulmonary, Critical Care, and Sleep Medicine, SUNY Upstate Medical University, Syracuse, NY, 13210.

7 Department of Biostatistics, National Jewish Health, Denver, CO, 80206.

8 Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, 02115.

9 Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045.

10 Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, 35233.

11 Division of General Internal Medicine and Primary Care, Department of Medicine, Brigham and Women’s Hospital, Boston, MA, 02115.

12 Centre for Heart Lung Innovation, St. Paul's Hospital, and Department of Medicine (Respiratory Division), University of British Columbia, Vancouver, BC, Canada.

13 Global Allergy and Airways Patient Platform, Vienna, AU.
14 Novartis Institute for Biomedical Research, Cambridge, MA.

15 Division of Pulmonary, Critical Care, and Sleep Medicine, University of Nebraska, Omaha, NE 68198

16 Lung Health Research Centre, Department of Biochemistry and Pharmacology, University of Melbourne, Melbourne, Victoria, Australia.

17 Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, 80045

18 Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Health, Denver, CO, 80206.

19 Division of Pulmonary and Critical Care Medicine, University of Michigan School of Medicine, Ann Arbor, MI, 48109.

20 Medical Service, Veterans Affairs Ann Arbor Healthcare System, Ann Arbor, MI, 48109

21 Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, 701, 19th Street S., LHRB 440, Birmingham, AL, 35233.

22 Department of Genetics, University of Alabama at Birmingham, Birmingham, AL, 35233.

23 Regeneron Pharmaceutical, Tarrytown, NY

24 Harvard Medical School, Boston, MA, 02115.

† = Jointly supervised and co-senior authors

Correspondence:
Michael H. Cho, MD, MPH
181 Longwood Avenue
Boston, MA 02115
remhc@channing.harvard.edu

This article has an online data supplement.

Funding:
MM is supported by K08HL159318.
BDH is supported by NIH K08HL136928, U01 HL089856, and an Alpha-1 Foundation Research Grant.
JH is supported by P01 HL132825.
JHP is supported by NIH K25HL140186.
CPH is supported by NIH R01HL157879 and P01HL114501.
MHC is supported by NIH R01HL137927, R01HL135142, HL147148, and HL089856.
PJC is supported by NIH R01HL124233 and R01HL147326. RPB is supported by NIH R01 HL137995 and R01 HL152735. EKS is supported by NIH R01 HL147148, U01 HL089856, R01 HL133135, R01 HL152728, and P01 HL114501. Proteomic data generated for this proposal was supported by R01 HL137995. The COPDGene project was supported by NHLBI grants U01 HL089897 and U01 HL089856 and by NIH contract 75N92023D00011. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. COPDGene 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. The ECLIPSE study (NCT00292552; GSK code SCO104960) was funded by GlaxoSmithKline.

**Disclosures:**
EKS received grant support from Bayer and Northpond Labs. BDH received grant support from Bayer. MHC has received grant support from Bayer. MM received grant support from Bayer and consulting fees from Sitka, TheaHealth, 2ndMD, and TriNetX. CPH reports grant support from Boehringer-Ingelheim, Novartis, Bayer and Vertex, outside of this study. PJC has received grant support from GlaxoSmithKline and Bayer and consulting fees from GlaxoSmithKline and Novartis. RTS received consulting fees from GSK, AstraZeneca, Roche, Itai and Beyond, Samay Health, Immunomet, ENA Respiratory, Teva, COPD Foundation and Vocalis Health. She is a retiree and shareholder of GSK and holds share options at ENA Respiratory. JLC received consulting fees from AstraZeneca PLC, CSL Behring, LLC, and Novartis Corporation. SIR received consulting fees from Verona Pharma, Sanofi, BeyondAir and the Alpha 1 Foundation. He is a founder and president of Great Plains Biometrix. He was an employee of AstraZeneca from 2015-2019 during which he received shares as part of his compensation that he owns.

**Author contributions:**


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


Obtained funding: Edwin K. Silverman, Michael H. Cho, Matthew Moll

**Word counts:**
Abstract: 249
Manuscript: 3404
ABSTRACT

Rationale: Genetic variants and gene expression predict risk of chronic obstructive pulmonary disease (COPD), but their effect on COPD heterogeneity is unclear.

Objectives: Define high-risk COPD subtypes using both genetics (polygenic risk score, PRS) and blood gene expression (transcriptional risk score, TRS) and assess differences in clinical and molecular characteristics.

Methods: We defined high-risk groups based on PRS and TRS quantiles by maximizing differences in protein biomarkers in a COPDGene training set and identified these groups in COPDGene and ECLIPSE test sets. We tested multivariable associations of subgroups with clinical outcomes and compared protein-protein interaction networks and drug repurposing analyses between high-risk groups.

Measurements and Main Results: We examined two high-risk omics-defined groups in non-overlapping test sets (n=1,133 NHW COPDGene, n=299 African American (AA) COPDGene, n=468 ECLIPSE). We defined “High activity” (low PRS/high TRS) and “severe risk” (high PRS/high TRS) subgroups. Participants in both subgroups had lower body-mass index (BMI), lower lung function, and alterations in metabolic, growth, and immune signaling processes compared to a low-risk (low PRS, low TRS) reference subgroup. “High activity” but not “severe risk” participants had greater prospective FEV$_1$ decline (COPDGene: -51 mL/year; ECLIPSE: -40 mL/year) and their proteomic profiles were enriched in gene sets perturbed by treatment with 5-lipoxygenase inhibitors and angiotensin-converting enzyme (ACE) inhibitors.

Conclusions: Concomitant use of polygenic and transcriptional risk scores identified clinical and molecular heterogeneity amongst high-risk individuals. Proteomic and drug repurposing analysis
identified subtype-specific enrichment for therapies and suggest prior drug repurposing failures may be explained by patient selection.
INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality worldwide. Although COPD is characterized by irreversible airflow obstruction, there is marked heterogeneity amongst individuals in emphysema and airway pathology, exacerbation incidence, and lung function decline. Identifying individuals at high risk for rapid COPD progression or eventual severe disease is critically important to implement personalized therapeutic approaches.

Large-scale omics data offer the potential to identify via a simple blood test high risk groups that share distinct, targetable pathobiology. Genetics, quantified with polygenic risk scores (PRSs), can identify individuals at high risk for coronary artery disease and guide consideration of statin therapy earlier than advised by current guidelines. In cancer, integrating genetic and transcriptomic profiling can improve therapy recommendations and outcomes. We demonstrated that both a PRS and a transcriptional risk score (TRS) independently predict COPD. Although the PRS and TRS were both based on spirometry measures, the scores are not correlated and likely capture different aspects of lung pathobiology. Specifically, COPD genetic risk loci are enriched for aspects of lung development and have a greater effect in early COPD; in contrast, the COPD TRS is associated with markers of inflammation and lung function decline and may reflect disease activity and propensity toward disease progression. Thus, it may be possible to leverage the different features of the PRS and TRS to identify clinically and biologically distinct COPD subtypes.

Despite advances in omics-based risk prediction, important clinical translation questions remain. Omics risk scores are usually standardized for statistical analyses, leaving the question of how to use them to risk-stratify individuals. Risk scores are also continuous measures, often
normally distributed; the issue of attempting to identify subtypes along a continuum has been
previously recognized\textsuperscript{10}. Despite this limitation, there is a need to classify individuals to link
omics-defined high-risk groups, which might benefit from specific therapies, with specific
pathobiological processes and treatment decisions. COPD drug and drug repurposing candidates
have high failure rates in clinical trials\textsuperscript{11}, but it remains unknown if these therapies have failed
because of patient selection, which currently does not utilize omics or other biomarkers.

The PRS was effective for predicting COPD severity and incident COPD, while the TRS
was better at predicting FEV\textsubscript{1} decline; combining both risk scores may identify subgroups at risk
for multiple important COPD outcomes. Therefore, we hypothesized that our published PRS and
TRS, both based on spirometry, could identify COPD subtypes (i.e., heterogeneity) within high-risk
groups with clinical and biological differences in two cohorts of ever-smokers. We aimed to
develop a novel approach for how omics risk scores can be applied to populations and leveraged
for precision medicine. We used proteomics to obtain an additional biological view of omics-defined
subgroups and performed in silico drug repurposing analyses to identify potential
subgroup-specific drug repurposing candidates.

\textbf{METHODS}

\textbf{Study populations}

All study participants provided written informed consent and studies were approved by
local Institutional Review Boards. Details regarding genotyping, RNA-sequencing, gene
expression microarray, and proteomics data acquisition and processing are available in the
Supplement.
We included Genetic Epidemiology of COPD (COPDGene) study (ClinicalTrials.gov Identifier: NCT00608764) participants with single nucleotide polymorphism (SNP) genotyping, RNA-sequencing, and SomaScan proteomic data to calculate the PRS, TRS, and performed differential protein expression analyses, respectively.

As previously described, we included Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE) study (ClinicalTrials.gov Identifier: NCT00292552) participants with SNP genotyping data, whole blood microarray data, and at least two FEV₁ measurements.

Additional details for both cohorts are in the supplement.

Polygenic and transcriptional risk scores

The COPD PRS and TRS were both based on spirometry and previously described, more details can be found in the supplement.

Proteomic data
Blood proteomic data were measured using SomaScan v4.0, which uses aptamers (i.e. SOMAmers) to quantify 4,776 unique human proteins. Further details regarding SomaScan data and preparation can be found in the Supplement and here\textsuperscript{12}.

**Statistical analysis**

**Overview of study design**

To identify high-risk subgroups based on continuous scores, we used the same previously defined COPDGene training set\textsuperscript{7}, and tested among PRS and TRS quantiles to maximize the number of associated differentially expressed proteins across the resulting subtype partitions (Fig. 1). We then determined the raw (non-standardized) score cut offs associated with the corresponding percentile values in the COPDGene training set. This approach facilitated classifying each participant in the COPDGene testing set and the external ECLIPSE validation set into an omics-defined subtype. We characterized the newly-defined subtypes using proteomic network and drug repurposing analyses, and applied multivariable linear regressions to test the association of subtypes with COPD-related outcomes.

**Determining risk score divisions and identifying omics-defined subtypes**

Oomics-based risk scores are typically standardized prior to statistical analysis, and therefore, have a normal distribution and by design do not lend themselves to clustering analyses (Figure S1). Yet, for clinical application, patients need to be categorized into groups. First, to determine whether one or more clusters are optimal, we calculated the gap statistic based on the PRS and TRS using the clusGap function (cluster R package) with a maximum of 8 kmeans clusters and 500 bootstrap iterations. The gap statistic provides a measure of dispersion for each
cluster and compares this dispersion metric to the expected dispersion under the null
distribution; thus, the difference (or “gap”) between the observed and expected within cluster
dispersion is used to calculate the gap statistic and the maximum gap statistic over a range of
cluster numbers indicates the optimal number of clusters.

As an alternative to clustering, individuals are commonly placed into omics risk-score
quantiles, and participants in each quantile are compared to those in the lowest risk quantile. To
extend this approach to two separate omics risk scores is more complex, as fewer
subdivisions lead to larger group sizes and more statistical power, but more subdivisions allow a
more extreme comparison group. As genetic and transcriptomic data were used to define the
subtypes, proteomics would provide a third “view” of the data. Thus, to determine the optimal
quantiles we split the PRS and TRS into 2 to 3 quantiles (minimum of 4 and maximum of 9
groups) and tested to see what group divisions maximized the number of associated differentially
expressed proteins using limma, comparing each quantile category to the lowest quantile. To
test the sensitivity of the groups to partitioning, we also examined clinical characteristics for
each combination of these partitions. Benjamini-Hochberg false discovery rate (FDR)-adjusted
p-values less than 0.05 were considered significant. The number of significantly differentially
expressed proteins associated with each omics risk score category was summed.

Clinical comparisons of omics-defined subtypes

We compared clinical characteristics across omics-defined subtypes using the tableone R
package. In COPDGene, transcriptomic and proteomic data were collected at the 5-year follow
up visit (i.e., “Phase 2”), so we examined differences in anthropometry (including change in BMI
per year (Kg/m²/year) from enrollment to the 5-year follow up visit), spirometry (including prospective FEV₁ change (from 5- to 10-year follow up visits)), and CT measures of emphysema (quantitative emphysema on inspiratory CT scans (% LAA < -950 HU)¹⁷, 15th percentile of lung density histogram on inspiratory CT scans (Perc15)¹⁸) and of airway thickening (wall area percent (WA%)¹⁷ and square root of wall area of a hypothetical internal perimeter of 10 mm (Pi10)¹⁹) at the 5-year follow up visit. In ECLIPSE, we examined the same outcomes but longitudinal follow up was from the time of study enrollment to the 3-year follow up visit.

Further details regarding outcomes and regression model specifications are in the Supplementary Methods.

Biological characterization of omics-defined subtypes

We performed differential gene and protein expression analyses (accepting FDR-adjusted p-values <0.05), comparing high risk subtypes to the reference group, defined as the group with the lowest PRS and TRS quantiles. We mapped differentially expressed proteins to the human protein-protein interactome²⁰ and performed Reactome²¹ pathway enrichment, STRING²⁴ and Enrichr²²–²⁴ analyses; additional details are in the supplement.
RESULTS

Characteristics of study populations

We included 3,274 participants across two cohorts of individuals who smoked. The COPDGene training and testing sets are similar in demographic and spirometry characteristics (Table 1). Compared to COPDGene, ECLIPSE participants were more likely to be younger, male, to have a greater number of smoking pack-years, a lower FEV$_1$ % predicted, and lower FEV$_1$/FVC, and were less likely to be current smokers.

Defining polygenic and transcriptional risk score divisions

Participants plotted on the axes of PRS and TRS exist along a continuum (Figure S1), as is the case for spirometric measures of COPD severity (i.e., FEV$_1$ and FEV$_1$/FVC$^{10,25}$). We calculated the gap statistic over a range of kmeans cluster numbers in the COPDGene training set, and consistent with visual inspection of Figure S1, we observed that one cluster yields the highest gap statistic, indicating that there are no clusters (Figure S2). As an alternative approach to clustering, we applied the common practice of dividing risk scores into quantiles, though the optimal quantiles balancing sufficiently high risk yet adequate sample size are not clear. Thus, we tested four combinations of partitioned omics risk scores, using protein expression differences (not used in the PRS and TRS) between groups. We observed that dichotomizing PRS and dividing TRS into tertiles yielded the greatest number of differentially expressed proteins (Table S1). We then applied these same quantiles to the COPDGene testing set and ECLIPSE participants (Fig. 2). To test the robustness of subgroups to specific partitions, we also examined 4 to 9 subdivisions, and noted stable clinical characteristics of the highest (“low
Polygenic and transcriptional risk scores identify “high disease activity” and “severe disease risk” subtypes

We observed, as expected, heterogeneity amongst the two high-risk (i.e., high TRS) groups (Table 2). We compared these high-risk subtypes to a reference group, which was defined as the lowest omics risk group (i.e., “low PRS/low TRS” subtype). Compared to the reference group, the two high-TRS risk groups (i.e., “low PRS/high TRS” and “high PRS/high TRS”) demonstrated decreased BMI, lower spirometry measures, more emphysema, and thicker airways across testing cohorts (Table 2). Both groups had similar mean adjusted prospective FEV\textsubscript{1} decline in the COPDGene testing set compared to the reference group, but this finding was only consistent for the “low PRS/high TRS” group in ECLIPSE (-40 mL/year).

We then performed linear regression analyses on selected COPD-related outcomes. We compared anthropometric, spirometry, CT, and other COPD-related outcomes across COPDGene and ECLIPSE (Table 3, Table S3). Compared to the reference group, high-risk groups exhibited lower spirometry, more emphysema, and thicker airways (Table 3). While none of the adjusted FEV\textsubscript{1} decline measures were statistically significant, the “low PRS, high TRS” means were consistent between COPDGene and ECLIPSE (-30ml/yr and -24ml/year, P = 0.15 and P = 0.11, respectively), despite that COPDGene participants had only two FEV\textsubscript{1} measurements while ECLIPSE participants had up to six FEV\textsubscript{1} measurements. Given the clinical relevance of accelerated FEV\textsubscript{1} decline in the “low PRS/high TRS” group, we renamed this group the “high disease activity” subtype. While the “high PRS/high TRS” group did not have replicable FEV\textsubscript{1} decline across cohorts, this group had the lowest lung function and most emphysema; therefore,
we renamed this group the “severe disease risk” subtype. In COPDGene NHW participants only, the “high disease activity” subtype also exhibited a trend toward greater decline in BMI compared to the reference group ($\beta = -0.154$ [95% CI: -0.333 to 0.0253], p=0.094).

**Biological characterization and drug repurposing analyses of subtypes**

Having identified clinical differences between the two high-risk groups, we sought to characterize biological differences between these subtypes. We performed differential gene and protein expression analyses between the “high disease activity” and “severe disease risk” subtypes and the reference group in the COPDGene NHW testing set (Tables S4-S6). The “high disease activity” subtype had 14 and the “severe disease risk” subtype had 2 differentially expressed proteins (Table S6). We did not observe differentially expressed genes or proteins when directly comparing high risk groups. We examined how the PRS affects differential gene expression associated with COPD case-control status as detailed in the supplement (Supplementary Results and Table S7).

We mapped differentially expressed proteins associated with each high-risk subtype in the COPDGene testing set to the human protein-protein interactome and used the mapped proteins as seed proteins to construct STRING PPI networks (Figures 3 and 4) with associated MCL clusters (Table S8) and perform pathway enrichment analyses (Table S9). To identify subtype-specific drug repurposing candidates, we used these same seed proteins to perform enrichment analyses on the MAGMA Drugs and Disease database. Both subtypes demonstrated enrichment proteomic profiles suggesting potential treatment with ACE inhibitors, thyroid
medications, carvedilol, bromocriptine, and lovastatin; the “high disease activity” subtype also had significant findings for 5-lipoxygenase inhibitors, fomepizole, and galantamine, while the “severe disease risk” (Table S10) subtype had significant findings for atypical antipsychotics.

DISCUSSION

In this study of 3,274 ever-smokers from two cohorts, we used blood-based polygenic (PRS) and transcriptional (TRS) risk scores to identify heterogeneity within high-risk individuals, defining “high disease activity” and “severe disease risk” COPD subtypes. Compared to a reference group, both subtypes had lower mean BMI values and alterations in metabolic, growth, and immune signaling processes. “High disease activity” participants exhibited prospective FEV₁ decline across both replication cohorts, albeit with a non-significant (though directionally consistent) association in multivariable models adjusted for baseline FEV₁. “Severe disease risk” participants had low spirometry measures with high quantitative emphysema and thick airways. We identified biological processes and drug repurposing candidates associated with each subtype, including therapies previously tested in COPD clinical trials. Our study demonstrates how omics risk scores can identify COPD subtypes with associated clinical and biological characteristics that can be leveraged for therapeutic interventions.

Linking omics-defined high-risk groups to specific pathobiology is an active area of research that we now extend to lung disease. In schizophrenia, PRS-defined high risk groups were biologically characterized using weighted gene co-expression network analyses. Other approaches have incorporated gene expression into PRSs to improve prediction and imply
biological mechanisms\textsuperscript{28–30}. Here, we used genetic and transcriptomic data to define subtypes, and then leveraged proteomic differences between groups to understand subtype biology. Importantly, our results suggest that different omics risk scores are not interchangeable, i.e., a higher omics risk score will not always have the same association with specific outcomes. Individuals with the highest transcriptomic quantile exhibited different clinical and biological features depending on their underlying polygenic risk. We also demonstrate that individuals may exist along a continuum of COPD risk – i.e., there are no clusters – yet omics-defined subgroups may have important clinical and biological differences. While we did not observe clusters, we observed that omics-defined subgroups were robust to varying PRS and TRS subdivisions, with the two high-risk groups ("low PRS/high TRS" and "high PRS/high TRS") demonstrating stable clinical characteristics across risk score subdivisions. Thus, genetics and transcriptomics may provide alternative yet complementary views of lung function biology.

Of direct clinical relevance, the association of lower spirometry, greater emphysema and observed FEV\textsubscript{1} decline across two cohorts suggests that the "high disease activity" subtype is a targetable trait, the risk of which, might be modified by approved medications (5-lipoxygenase inhibitors, angiotensin-converting enzyme (ACE) inhibitors, fomepizole, galantamine). Fomepizole has not previously been implicated as a COPD drug repurposing candidate to our knowledge. Galantamine is known to cause bronchospasm, and enrichment for proteins targeted by galantamine suggests this drug is most likely to cause harm in this subtype of patients. Although a randomized trial of the 5-lipoxygenase inhibitor, Zileuton, did not reduce length of stay or treatment failure in patients hospitalized for COPD exacerbations, it was likely underpowered\textsuperscript{31}, and did not examine longer term outcomes. ACE inhibitors and angiotensin receptor blockers (ARB) have been identified as COPD drug repurposing candidates. A recent
clinical trial showed failure of losartan to decrease emphysema progression\textsuperscript{32}. Our drug repurposing analyses implicated utility of captopril – not all ACE inhibitors and ARBs - in the “high disease activity” but not the “severe disease risk” subtype. Conversely, our analysis suggests potential benefit of atypical antipsychotics in the “severe disease risk” subtype. The coincidence of schizophrenia and COPD is largely attributed to smoking, though a phenome-wide association and polygenic risk analysis suggests that schizophrenia and obstructive lung disease may have shared genetic mechanisms\textsuperscript{33}. While we adjusted for self-reported cigarette smoking status, these measures are imperfect, and thus we cannot address whether shared mechanisms are due to smoking behavior. A broader implication of the drug repurposing analyses that merits validation is that previous failure of drugs in clinical trials was due to heterogeneity in patient selection that could be overcome by omics-based subtyping.

The cachectic COPD patient, who may be prone to more exacerbations, is a well-described clinical phenotype\textsuperscript{34–37}, and we observed that the “high disease activity” and “severe disease risk” subtypes have lower mean BMI than other subtypes. Although pulmonary cachexia has been proposed to occur only in severe COPD\textsuperscript{38}, we demonstrate additional heterogeneity within lower BMI ever-smokers and identify a “high disease activity” subtype with less spirometric severity and distinct biology. The “high disease activity” group may also overlap with a previously identified comorbidity cachexia subgroup\textsuperscript{39}, and in the current analysis, we identified molecular profiles that might guide therapy. Relevant to body composition, the adipocyte product leptin exhibited altered expression in both subtypes, which has several potential, not mutually exclusive interpretations. Leptin acts both as a hormone negatively regulating hunger and adipocyte fat storage\textsuperscript{40} and as a proinflammatory cytokine essential for host defense\textsuperscript{41,42}. 
The role of leptin in regulating hunger suggests it could be relevant to the “severe disease risk” subtype, which had the largest effect size in leptin expression (-0.71 log-fold change); the observed decrease in leptin could be a compensatory response to or share a causal relationship with pulmonary cachexia. In addition, the lower leptin levels observed in malnourished populations are associated with dysfunctional cell-mediated immunity and increased susceptibility to infections\textsuperscript{41,43}. In COPD, elevated leptin concentrations have been reported in the plasma and airways in some\textsuperscript{44–48}, but not all studies\textsuperscript{49}, and has been identified as a potential biomarker of emphysema progression\textsuperscript{50,51}. Our results suggest that “high disease activity” or “severe disease risk” subtype individuals might exhibit humoral and cytokine profiles similar to those seen in malnourished individuals with increased susceptibility to infections. It remains unclear whether the observed peripheral blood proteomic alterations are merely a consequence of disease activity, or a causal component of a positive feedback loop involved in driving disease progression. The biomarkers used in this study are blood-based, making it difficult to identify the relevant mechanisms in lung tissue. However, blood-based biomarkers are practical in a clinical setting and follow up studies linking changes in these biomarkers to pathophysiologic mechanism in lung tissue can help to bridge the gap between prediction and precision therapeutics.

This study leveraged multiple omics data in the form of validated, replicated, risk scores, to identify COPD subtypes in two cohorts of ever-smokers. One major challenge of this analysis was determining how to subset subjects from a standardized, continuous distribution. To overcome this limitation, we varied the number of quantiles and assessed which combination of divisions yielded the greatest number of differentially expressed proteins, then identified corresponding raw score cut-off values that allowed each participant in the COPDGene and
ECLIPSE testing sets to be categorized into an omics-defined subtype. We acknowledge that there are other reasonable approaches for applying omics risk scores to identify COPD subtypes. Individuals near the cut-off value for PRS might exist on a continuum between “high disease activity” and “severe disease risk” subtypes, and these participants might be able to transition between subtypes; regardless, defining thresholds and categories are an important step toward clinical translation. We observed similar results in COPDGene AA participants, including the “high disease activity” subtype association with FEV₁ decline in multivariable analyses; however, additional work to optimize multi-ancestry genetic and omics-based prediction is crucial to prevent omics technologies from becoming a vehicle for worsening existing healthcare disparities. We based our drug repurposing candidates on enrichment analyses of proteomic profiles, but directionality of biological processes was not accounted for in enrichment analyses; additional mechanistic and clinical trial validations are needed. Finally, we identified two high-risk subtypes using this approach, but these do not explain all of COPD heterogeneity, and there are almost certainly other important subtypes.

In conclusion, polygenic and transcriptional risk scores, both based on spirometry, identified “high disease activity” and “severe disease risk” subtypes with distinct clinical and biological characteristics. Proteomic and drug repurposing analysis identified subtype-specific enrichment for therapies, some of which were previously hypothesized in COPD.
REFERENCES


Table legends

Table 1: Characteristics of study populations. COPD = chronic obstructive pulmonary disease. COPDGene = Genetic Epidemiology of COPD. ECLIPSE = Evaluation of COPD to Longitudinally Identify Predictive Surrogate Endpoints study. FEV₁ = forced expiratory volume in 1 second. FEV₁/FVC = FEV₁/forced vital capacity. NHW = non-Hispanic white. AA=African American.

Table 2: Omics-defined subtypes in the COPDGene testing set and ECLIPSE. The was defined as the "Low PRS, Low TRS" group. BMI = body-mass index. LAA = low attenuation area. HU = Hounsfield units. Perc15 = 15th percentile of lung density histogram on inspiratory CT scans. WA % = wall area percent. Pi10 = square root of wall area of a hypothetical internal perimeter of 10 mm. ACO = asthma-COPD overlap. See Table 1 legend for other abbreviations. NHW = non-Hispanic white. AA=African American.

Table 3: Multivariable linear regressions in the COPDGene testing sets and ECLIPSE. Models were adjusted or age, sex, current smoking status, pack-years of smoking, and principal components of genetic ancestry. Computed tomography imaging outcomes were additionally adjusted for CT scanner. Abbreviations are detailed in the Table 1 and 2 legends.

Figure legends

Figure 1: Schematic of study design. COPD = chronic obstructive pulmonary disease. COPDGene = Genetic Epidemiology of COPD study. ECLIPSE = Evaluation of COPD to Longitudinally Identify Predictive Surrogate Endpoints study. PRS = polygenic risk score. TRS = transcriptional risk score. STRING = Search Tool for the Retrieval of Interacting Genes/Proteins. MAGMA = Multi-marker Analysis of GenoMic Annotation.

Figure 2. Omics-defined groups or subtypes overlaid on a plot of the polygenic risk score (PRS; $x$-axis) and transcriptional risk score (TRS; $y$-axis) in the COPDGene testing set.
Figure 3. High disease activity ("low PRS, high TRS") subtype STRING protein-protein interaction networks using differentially expressed proteins in Omics-defined groups (subtypes) in the COPDGene testing set as seed proteins, permitting up to 10 interactors in the first shell and 5 interactors in the second shell. Only high-confidence interactions were included and greater line thickness indicates greater confidence. Differentially expressed proteins were identified by comparing group assignments to the reference group. Colors represent MCL (Markov) clusters.

Figure 4: Severe disease risk ("high PRS, high TRS") subtype STRING protein-protein interaction networks using differentially expressed proteins in Omics-defined groups (subtypes) in the COPDGene testing set as seed proteins, permitting up to 10 interactors in the first shell and 5 interactors in the second shell. Only high-confidence interactions were included and greater line thickness indicates greater confidence. Differentially expressed proteins were identified by comparing group assignments to the reference group. Colors represent MCL (Markov) clusters.
Table 1: Characteristics of study populations. COPD = chronic obstructive pulmonary disease. COPDGene = Genetic Epidemiology of COPD. ECLIPSE = Evaluation of COPD to Longitudinally Identify Predictive Surrogate Endpoints study. FEV$_1$ = forced expiratory volume in 1 second. FEV$_1$/FVC = FEV$_1$/forced vital capacity. NHW = non-Hispanic white. AA=African American.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>COPDGene training set</th>
<th>COPDGene testing set</th>
<th>ECLIPSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NHW</td>
<td>NHW</td>
<td>AA</td>
</tr>
<tr>
<td>n</td>
<td>1374</td>
<td>1133</td>
<td>299</td>
</tr>
<tr>
<td>Age in years (mean (SD))</td>
<td>67.38 (8.25)</td>
<td>68.13 (8.30)</td>
<td>61.01 (6.95)</td>
</tr>
<tr>
<td>Sex (No. % female)</td>
<td>677 (49.3)</td>
<td>560 (49.4)</td>
<td>153 (51.2)</td>
</tr>
<tr>
<td>Pack-years of smoking (mean (SD))</td>
<td>45.47 (24.61)</td>
<td>45.47 (23.65)</td>
<td>39.55 (20.74)</td>
</tr>
<tr>
<td>Current smoking (No %)</td>
<td>343 (25.0)</td>
<td>289 (25.5)</td>
<td>182 (60.9)</td>
</tr>
<tr>
<td>FEV$_1$ % predicted (mean (SD))</td>
<td>78.37 (24.21)</td>
<td>77.85 (24.81)</td>
<td>81.79 (23.54)</td>
</tr>
<tr>
<td>FEV$_1$/FVC (mean (SD))</td>
<td>0.67 (0.15)</td>
<td>0.66 (0.15)</td>
<td>0.71 (0.14)</td>
</tr>
</tbody>
</table>
Table 2: Omics-defined subtypes in the COPDGene testing set and ECLIPSE. The was defined as the "Low PRS, Low TRS" group. BMI = body-mass index. LAA = low attenuation area. HU = Hounsfield units. Perc15 = 15th percentile of lung density histogram on inspiratory CT scans. WA % = wall area percent. Pi10 = square root of wall area of a hypothetical internal perimeter of 10 mm. ACO = asthma-COPD overlap. See Table 1 legend for other abbreviations. NHW = non-Hispanic white. AA=African American.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>COPDGene NHW testing set</th>
<th>COPDGene AA testing set</th>
<th>Reference group (Low PRS, Low TRS)</th>
<th>ECLIPSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>196</td>
<td>65</td>
<td>61</td>
<td>57</td>
</tr>
<tr>
<td>Age in years (mean (SD))</td>
<td>67.85 (8.52)</td>
<td>67.16 (6.46)</td>
<td>61.26 (7.55)</td>
<td>63.00 (5.10)</td>
</tr>
<tr>
<td>Sex (No % female)</td>
<td>114 (58.2)</td>
<td>19 (29.2)</td>
<td>36 (59.0)</td>
<td>30 (52.6)</td>
</tr>
<tr>
<td>Pack-years of smoking (mean (SD))</td>
<td>36.51 (22.27)</td>
<td>60.07 (26.49)</td>
<td>37.01 (20.85)</td>
<td>30.61 (6.36)</td>
</tr>
<tr>
<td>Current smoking (No. %)</td>
<td>20 (10.2)</td>
<td>36 (55.4)</td>
<td>31 (50.8)</td>
<td>6 (10.5)</td>
</tr>
<tr>
<td>BMI (Kg/m^2) (mean (SD))</td>
<td>30.00 (6.84)</td>
<td>26.51 (5.81)</td>
<td>30.61 (6.36)</td>
<td>27.54 (5.06)</td>
</tr>
<tr>
<td>FEV1 % predicted (mean (SD))</td>
<td>90.15 (18.13)</td>
<td>66.50 (27.07)</td>
<td>90.83 (17.41)</td>
<td>50.25 (13.42)</td>
</tr>
<tr>
<td>FEV1/FVC (mean (SD))</td>
<td>0.75 (0.09)</td>
<td>0.57 (0.15)</td>
<td>0.77 (0.10)</td>
<td>16.01 (11.13)</td>
</tr>
<tr>
<td>% LAA &lt;-950 HU (mean (SD))</td>
<td>2.78 (4.71)</td>
<td>8.50 (10.39)</td>
<td>1.71 (4.00)</td>
<td>20.88 (13.50)</td>
</tr>
<tr>
<td>Perc15 (mean (SD))</td>
<td>-913.62 (20.43)</td>
<td>-927.23 (27.21)</td>
<td>-895.53 (26.89)</td>
<td>-943.75 (50.13)</td>
</tr>
<tr>
<td>Pi10 (mean (SD))</td>
<td>2.01 (0.50)</td>
<td>2.55 (0.65)</td>
<td>2.11 (0.48)</td>
<td>4.38 (0.20)</td>
</tr>
<tr>
<td>WA % (mean (SD))</td>
<td>46.97 (8.11)</td>
<td>54.46 (8.35)</td>
<td>47.48 (7.76)</td>
<td>64.35 (2.79)</td>
</tr>
<tr>
<td>Change in FEV1 mL/year (prospective)</td>
<td>-32.11 (48.07)</td>
<td>-51.21 (74.41)</td>
<td>-42.28 (36.59)</td>
<td>-32.21 (59.54)</td>
</tr>
</tbody>
</table>

Moll et al.
Table 3. Multivariable linear regressions in the COPDGene testing sets and ECLIPSE. Models were adjusted or age, sex, current smoking status, pack-years of smoking, and principal components of genetic ancestry. Computed tomography imaging outcomes were additionally adjusted for CT scanner. Abbreviations are detailed in the Table 1 and 2 legends.

<table>
<thead>
<tr>
<th>COPD-related outcome</th>
<th>COPDGene NHW testing set</th>
<th>COPDGene AA testing set</th>
<th>ECLIPSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low PRS, High TRS</td>
<td>High PRS, High TRS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>beta (95% CI)</td>
<td>p</td>
<td>beta (95% CI)</td>
</tr>
<tr>
<td>FEV1 % predicted</td>
<td>-22.3 (-29.5 to -15.2)</td>
<td>4.3E-09</td>
<td>-31.5 (-38.2 to -24.8)</td>
</tr>
<tr>
<td>FEV1/FVC</td>
<td>-0.147 (-0.184 to -0.11)</td>
<td>3E-13</td>
<td>3.1 (1.8 to 5.2)</td>
</tr>
<tr>
<td>% LAA &lt; -950 HU</td>
<td>-11.2 (-18.6 to -3.72)</td>
<td>0.0037</td>
<td>-15.3 (-22.8 to -7.89)</td>
</tr>
<tr>
<td>Perc15</td>
<td>-11.2 (-18.6 to -3.72)</td>
<td>0.0037</td>
<td>0.498 (0.298 to 0.698)</td>
</tr>
<tr>
<td>Pi10</td>
<td>-1.003 (0.252 to 1.936)</td>
<td>0.031</td>
<td>0.588 (3.19 to 8.58)</td>
</tr>
<tr>
<td>WA %</td>
<td>-30 (-70.6 to 10.7)</td>
<td>0.15</td>
<td>-43.9 (-91.7 to 3.94)</td>
</tr>
<tr>
<td>FEV1 change (mL/year)</td>
<td>-30 (-70.6 to 10.7)</td>
<td>0.15</td>
<td>-43.9 (-91.7 to 3.94)</td>
</tr>
</tbody>
</table>
COPDGene training set (NHW: n=1,374) → Determine PRS and TRS splits → Apply risk score cut offs and define subtypes → ECLIPSE (n=468)

COPDGene testing set (NHW: n=1,133 AA: n=299)

Identify differentially expressed proteins associated with subtypes
- STRING analysis
- Enrichr/MAGMA drug repurposing analysis

Regression analyses of subtype associations with clinical outcomes
High disease activity subtype

Severe disease risk subtype

low PRS
high TRS

low PRS
intermediate TRS

low PRS
low TRS

high PRS
high TRS

high PRS
intermediate TRS

high PRS
low TRS