medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Polygenic and transcriptional risk scores identify chronic obstructive pulmonary disease subtypes

Matthew Moll[1,2,3,24], Julian Hecker[1,24], John Platig[4], Jingzhou Zhang[5], Auyon J. Ghosh[6], Katherine A. Pratte[7], Rui-Sheng Wang[1], Davin Hill[8], Iain R. Konigsberg[9], Joe W. Chiles III[10], Craig P. Hersh[1,2,24], Peter J. Castaldi[1,10,24], Kimberly Glass[1,24], Jennifer G. Dy[8], Don D. Sin[12], Ruth Tal-Singer[13], Majd Mouded[14], Stephen I. Rennard[15], Gary P. Anderson[16], Gregory L. Kinney[17], Russell P. Bowler[18], Jeffrey L. Curtis[19,20], Merry-Lynn McDonald[9,21,22], Edwin K. Silverman[1,2,24], Brian D. Hobbs[23†], Michael H. Cho[1,2,24†]

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.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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:
Study Design: Matthew Moll, Julian Hecker, Michael H. Cho, Brian D. Hobbs.

Acquisition, analysis, or interpretation of the data: Matthew Moll, Julian Hecker, Edwin K. Silverman, Auyon Ghosh, Don D. Sin, Peter J. Castaldi, Katherine Pratte, Russell Bowler, Brian D. Hobbs, Michael H. Cho.

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

Statistical analysis: Matthew Moll, Julian Hecker, John Platig, Kimberly Glass, Brian D. Hobbs, Michael H. Cho.

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

Word counts:
Abstract: 249
Manuscript: 3404

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

identified subtype-specific enrichment for therapies and suggest prior drug repurposing failures may be explained by patient selection.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality worldwide[1]. 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[2,3]. 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[4]. In cancer, integrating genetic and transcriptomic profiling can improve therapy recommendations and outcomes[5]. We demonstrated that both a PRS and a transcriptional risk score (TRS) independently predict COPD[6,7]. Although the PRS and TRS were both based on spirometry measures, the scores are not correlated[7] 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[6,8]; 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[7]. 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[9]. Risk scores are also continuous measures, often

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

normally distributed; the issue of attempting to identify subtypes along a continuum has been previously recognized[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[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_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.

METHODS

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

COPDGene

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.

ECLIPSE

As previously described[7], 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 FEV_1 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[6,7]; more details can be found in the supplement.

Proteomic data

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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[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[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

Omics-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

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

cluster and compares this dispersion metric to the expected dispersion under the null distribution[13]; 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[4,6,14]. 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[15], 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[16] 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

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

per year (Kg/m^2/year) from enrollment to the 5-year follow up visit), spirometry (including prospective FEV_1 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 interactome[20] and performed Reactome[21] pathway enrichment, STRING[24] and Enrichr[22–24] analyses; additional details are in the supplement.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

PRS/high TRS" and "high PRS/high PRS") and lowest risk ("low PRS/low TRS") groups (**Table S2**).

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_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_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_1 measurements while ECLIPSE participants had up to six FEV_1 measurements. Given the clinical relevance of accelerated FEV_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_1 decline across cohorts, this group had the lowest lung function and most emphysema; therefore,

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 (β = -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[20] 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[26]. Both subtypes demonstrated enrichment proteomic profiles suggesting potential treatment with ACE inhibitors, thyroid

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 FEV_1 decline across both replication cohorts, albeit with a non-significant (though directionally consistent) association in multivariable models adjusted for baseline FEV_1. "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[27]. Other approaches have incorporated gene expression into PRSs to improve prediction and imply

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

biological mechanisms[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_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[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

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

clinical trial showed failure of losartan to decrease emphysema progression[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[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 [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[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[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[40] and as a proinflammatory cytokine essential for host defense[41,42].

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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[41,43]. In COPD, elevated leptin concentrations have been reported in the plasma and airways in some[44–48], but not all studies[49], and has been identified as a potential biomarker of emphysema progression[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

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 FEV_1 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[52]. 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

REFERENCES

1. Safiri, S. *et al.* Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the Global Burden of Disease Study 2019. *BMJ* **378**, e069679 (2022).
2. Hurst, J. R. *et al.* Susceptibility to Exacerbation in Chronic Obstructive Pulmonary Disease. *N. Engl. J. Med.* **363**, 1128–1138 (2010).
3. Wedzicha, J. A. The heterogeneity of chronic obstructive pulmonary disease. *Thorax* **55**, 631–632 (2000).
4. Natarajan, P. *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101 (2017).
5. Rodon, J. *et al.* Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nat. Med.* **25**, 751–758 (2019).
6. Moll, M. *et al.* Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. *Lancet Respir. Med.* **8**, 696–708 (2020).
7. Moll, M. *et al.* Development of a Blood-based Transcriptional Risk Score for Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **205**, 161–170 (2022).
8. Zhang, J. *et al.* A Polygenic Risk Score and Age of Diagnosis of Chronic Obstructive Pulmonary Disease. *Eur. Respir. J.* 2101954 (2022) doi:10.1183/13993003.01954-2021.
9. Sun, J. *et al.* Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat. Commun.* **12**, 5276 (2021).
10. Castaldi, P. J. *et al.* Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* **69**, 415–422 (2014).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

11. Martinez, F. J. *et al.* Treatment Trials in Young Patients with Chronic Obstructive Pulmonary Disease and Pre-Chronic Obstructive Pulmonary Disease Patients: Time to Move Forward. *Am. J. Respir. Crit. Care Med.* **205**, 275–287 (2022).
12. Serban, K. A. *et al.* Unique and shared systemic biomarkers for emphysema in Alpha-1 Antitrypsin deficiency and chronic obstructive pulmonary disease. *EBioMedicine* **84**, 104262 (2022).
13. Tibshirani, R., Walther, G. & Hastie, T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 411–423 (2001).
14. Khera, A. V. *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596.e9 (2019).
15. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
16. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
17. Han, M. K. *et al.* Chronic Obstructive Pulmonary Disease Exacerbations in the COPDGene Study: Associated Radiologic Phenotypes. *Radiology* **261**, 274–282 (2011).
18. Parr, D. G., Sevenoaks, M., Deng, C. Q., Stoel, B. C. & Stockley, R. A. Detection of emphysema progression in alpha 1-antitrypsin deficiency using CT densitometry; Methodological advances. *Respir. Res.* **9**, 1–8 (2008).
19. Van Tho, N. *et al.* A mixed phenotype of airway wall thickening and emphysema is associated with dyspnea and hospitalization for chronic obstructive pulmonary disease. *Ann. Am. Thorac. Soc.* **12**, 988–996 (2015).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

20. Wang, R.-S. & Loscalzo, J. Network module-based drug repositioning for pulmonary arterial hypertension. *CPT Pharmacomet. Syst. Pharmacol.* **10**, 994–1005 (2021).
21. Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142 (2017).
22. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
23. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-97 (2016).
24. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).
25. Lowe, K. E. *et al.* COPDGene® 2019: Redefining the Diagnosis of Chronic Obstructive Pulmonary Disease. *Chronic Obstr. Pulm. Dis. Miami Fla* **6**, 384–399 (2019).
26. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
27. Radulescu, E. *et al.* Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol. Psychiatry* **25**, 791–804 (2020).
28. Hu, X. *et al.* Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program. *Am. J. Hum. Genet.* **109**, 857–870 (2022).
29. Mareckova, K. *et al.* Novel polygenic risk score as a translational tool linking depression-related changes in the corticolimbic transcriptome with neural face processing and anhedonic symptoms. *Transl. Psychiatry* **10**, 1–10 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

30. Bandres-Ciga, S. *et al.* Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease. *Acta Neuropathol. (Berl.)* **140**, 341–358 (2020).
31. Woodruff, P. G. *et al.* Randomized trial of zileuton for treatment of COPD exacerbations requiring hospitalization. *COPD* **8**, 21–29 (2011).
32. Wise, R. A. *et al.* Clinical Trial of Losartan for Pulmonary Emphysema: Pulmonary Trials Cooperative Losartan Effects on Emphysema Progression Clinical Trial. *Am. J. Respir. Crit. Care Med.* **206**, 838–845 (2022).
33. Zhang, R. *et al.* Novel disease associations with schizophrenia genetic risk revealed in ~400 000 UK Biobank participants. *Mol. Psychiatry* **27**, 1448–1454 (2022).
34. McDonald, M.-L. N. *et al.* It's more than low BMI: prevalence of cachexia and associated mortality in COPD. *Respir. Res.* **20**, 100 (2019).
35. Grigsby, M. R. *et al.* Low Body Mass Index Is Associated with Higher Odds of COPD and Lower Lung Function in Low- and Middle-Income Countries. *COPD* **16**, 58–65 (2019).
36. Wada, H. *et al.* Low BMI and weight loss aggravate COPD mortality in men, findings from a large prospective cohort: the JACC study. *Sci. Rep.* **11**, 1531 (2021).
37. Mason, S. E. *et al.* Respiratory exacerbations are associated with muscle loss in current and former smokers. *Thorax* **76**, 554–560 (2021).
38. Eriksson, B. *et al.* Only severe COPD is associated with being underweight: results from a population survey. *ERJ Open Res.* **2**, 00051–02015 (2016).
39. Vanfleteren, L. E. G. W. *et al.* Clusters of comorbidities based on validated objective measurements and systemic inflammation in patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **187**, 728–735 (2013).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

40. Collins, S. *et al.* Role of leptin in fat regulation. *Nature* **380**, 677 (1996).

41. La Cava, A. Leptin in inflammation and autoimmunity. *Cytokine* **98**, 51–58 (2017).

42. La Cava, A. & Matarese, G. The weight of leptin in immunity. *Nat. Rev. Immunol.* **4**, 371–379 (2004).

43. Matarese, G. *et al.* Balancing susceptibility to infection and autoimmunity: a role for leptin? *Trends Immunol.* **23**, 182–187 (2002).

44. Suzuki, M. *et al.* Lower Leptin/Adiponectin Ratio and Risk of Rapid Lung Function Decline in Chronic Obstructive Pulmonary Disease. *Ann. Am. Thorac. Soc.* **11**, 1511–1519 (2014).

45. Takabatake, N. *et al.* Circulating leptin in patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **159**, 1215–1219 (1999).

46. Hansel, N. N. *et al.* Leptin receptor polymorphisms and lung function decline in COPD. *Eur. Respir. J.* **34**, 103–110 (2009).

47. Bruno, A. *et al.* Does leptin play a cytokine-like role within the airways of COPD patients? *Eur. Respir. J.* **26**, 398–405 (2005).

48. Vernooy, J. H. J. *et al.* Enhanced pulmonary leptin expression in patients with severe COPD and asymptomatic smokers. *Thorax* **64**, 26–32 (2009).

49. Sueblinvong, V. & Liangpunsakul, S. Relationship between serum leptin and chronic obstructive pulmonary disease in US adults: Results from NHANESIII. *J. Investig. Med. Off. Publ. Am. Fed. Clin. Res.* **62**, 934–937 (2014).

50. Curtis, J. L. Queens beat one-eyed jacks, but nobody's played the ace yet. Adipokines as chronic obstructive pulmonary disease biomarkers. *Ann. Am. Thorac. Soc.* **12**, 971–973 (2015).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

51. Oh, Y.-M. *et al.* Association of plasma adipokines with chronic obstructive pulmonary disease severity and progression. *Ann. Am. Thorac. Soc.* **12**, 1005–1012 (2015).

52. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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. FEV_1 = forced expiratory volume in 1 second. FEV_1/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 = Hounsfeld 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 = FEV1/forced vital capacity. NHW = non-Hispanic white. AA=African American.

Characteristic	*COPDGene training set*	*COPDGene testing set*		*ECLIPSE*
	NHW	*NHW*	*AA*	*NHW*
n	1374	1133	299	468
Age in years (mean (SD))	67.38 (8.25)	68.13 (8.30)	61.01 (6.95)	64.43 (6.09)
Sex (No. % female)	677 (49.3)	560 (49.4)	153 (51.2)	156 (33.3)
Pack-years of smoking (mean (SD))	45.47 (24.61)	45.47 (23.65)	39.55 (20.74)	49.33 (26.87)
Current smoking (No %)	343 (25.0)	289 (25.5)	182 (60.9)	70 (15.0)
FEV1 % predicted (mean (SD))	78.37 (24.21)	77.85 (24.81)	81.79 (23.54)	44.22 (14.64)
FEV1/FVC (mean (SD))	0.67 (0.15)	0.66 (0.15)	0.71 (0.14)	0.44 (0.11)

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 = Hounsfeld 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.

Characteristic	**COPDGene NHW testing set**			**COPDGene AA testing set**			**ECLIPSE**		
	Reference group *(Low PRS, Low TRS)*	*Low PRS, High TRS*	*High PRS, High TRS*	**Reference group** *(Low PRS, Low TRS)*	*Low PRS, High TRS*	*High PRS, High TRS*	**Reference group** *(Low PRS, Low TRS)*	*Low PRS, High TRS*	*High PRS, High TRS*
n	196	65	68	61	16	11	57	63	56
Age in years (mean (SD))	67.85 (8.52)	67.16 (6.46)	69.04 (8.11)	61.26 (7.55)	60.58 (7.24)	65.95 (8.05)	63.00 (5.10)	65.43 (6.65)	66.59 (5.95)
Sex (No % female)	114 (58.2)	19 (29.2)	24 (35.3)	36 (59.0)	5 (31.2)	8 (72.7)	30 (52.6)	10 (15.9)	11 (19.6)
Pack-years of smoking (mean (SD))	36.51 (22.27)	60.07 (26.49)	57.28 (23.35)	37.01 (20.85)	43.42 (16.14)	61.36 (38.87)	45.49 (26.24)	55.29 (35.25)	54.29 (21.10)
Current smoking (No. %)	20 (10.2)	36 (55.4)	36 (52.9)	31 (50.8)	15 (93.8)	8 (72.7)	6 (10.5)	7 (11.1)	12 (21.4)
BMI (Kg/m^2) (mean (SD))	30.00 (6.84)	26.51 (5.81)	25.70 (5.58)	30.61 (6.36)	29.14 (8.52)	25.26 (6.57)	27.54 (5.06)	26.70 (6.22)	25.49 (4.76)
FEV1 % predicted (mean (SD))	90.15 (18.13)	66.50 (27.07)	60.55 (24.11)	90.83 (17.41)	80.61 (25.43)	71.08 (10.67)	50.25 (13.42)	43.41 (15.65)	37.44 (13.44)
FEV1/FVC (mean (SD))	0.75 (0.09)	0.57 (0.15)	0.53 (0.15)	0.77 (0.10)	0.73 (0.14)	0.61 (0.12)	45.84 (11.13)	43.16 (12.79)	39.71 (10.29)
% LAA < -950 HU (mean (SD))	2.78 (4.71)	8.50 (10.39)	11.00 (11.98)	1.71 (4.00)	3.51 (8.08)	5.49 (7.30)	16.01 (11.46)	20.88 (13.50)	22.44 (13.22)
Perc15 (mean (SD))	-913.62 (20.43)	-927.23 (27.21)	-931.05 (31.55)	-895.53 (26.89)	-901.19 (29.94)	-913.62 (34.95)	-943.75 (50.13)	-960.00 (49.44)	-971.11 (45.84)
Pi10 (mean (SD))	2.01 (0.50)	2.55 (0.65)	2.60 (0.51)	2.11 (0.48)	2.56 (0.59)	2.42 (0.44)	4.38 (0.20)	4.42 (0.23)	4.43 (0.19)
WA % (mean (SD))	46.97 (8.11)	54.46 (8.35)	54.61 (7.41)	47.48 (7.76)	54.86 (10.52)	51.62 (4.06)	64.35 (2.79)	65.05 (3.58)	67.22 (3.44)
Change in FEV1 mL/year (prospective) (mean (SD))	-32.11 (48.07)	-51.21 (74.41)	-66.45 (61.58)	-42.28 (36.59)	-87.86 (156.98)	-56.26 (76.62)	-32.21 (59.54)	-40.13 (75.31)	-15.22 (67.46)

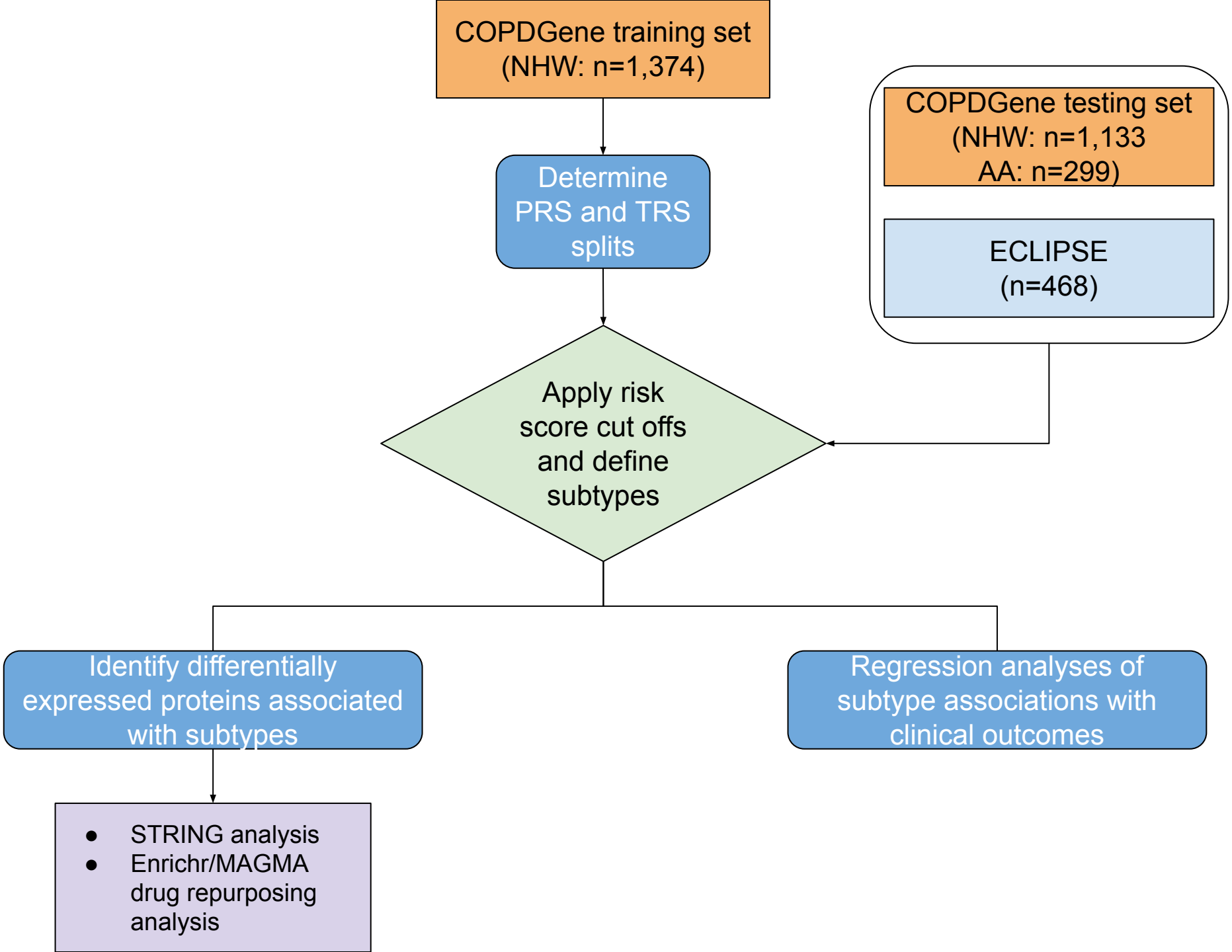
medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

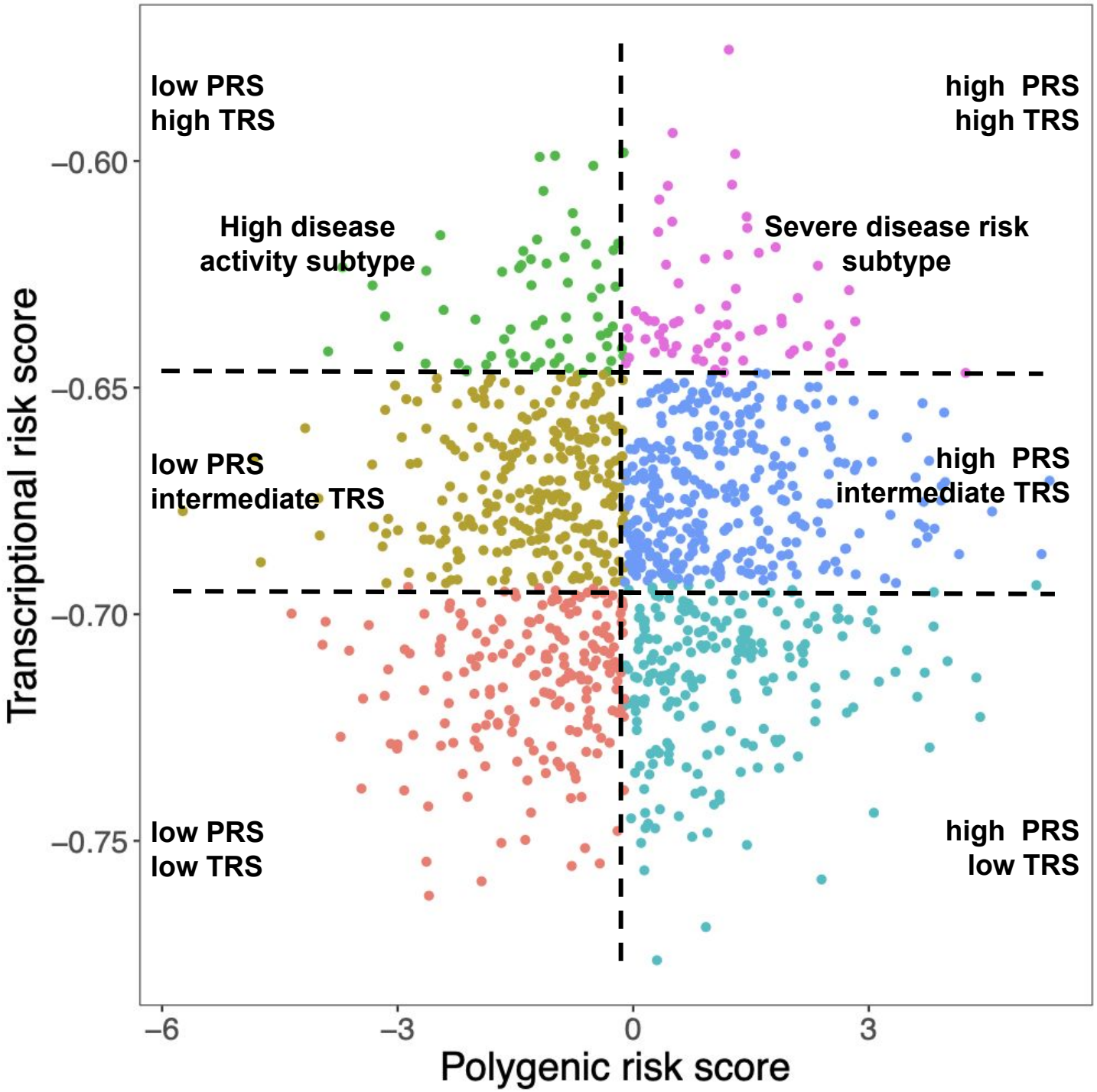
Table 3. Multivariable linear regressions in the COPDGene testing sets and ECLIPSE. Models were adjusted or age, sex, current smoking status, pack-years of smoking, and principal components of genetic ancestry. Computed tomography imaging outcomes were additionally adjusted for CT scanner. Abbreviations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

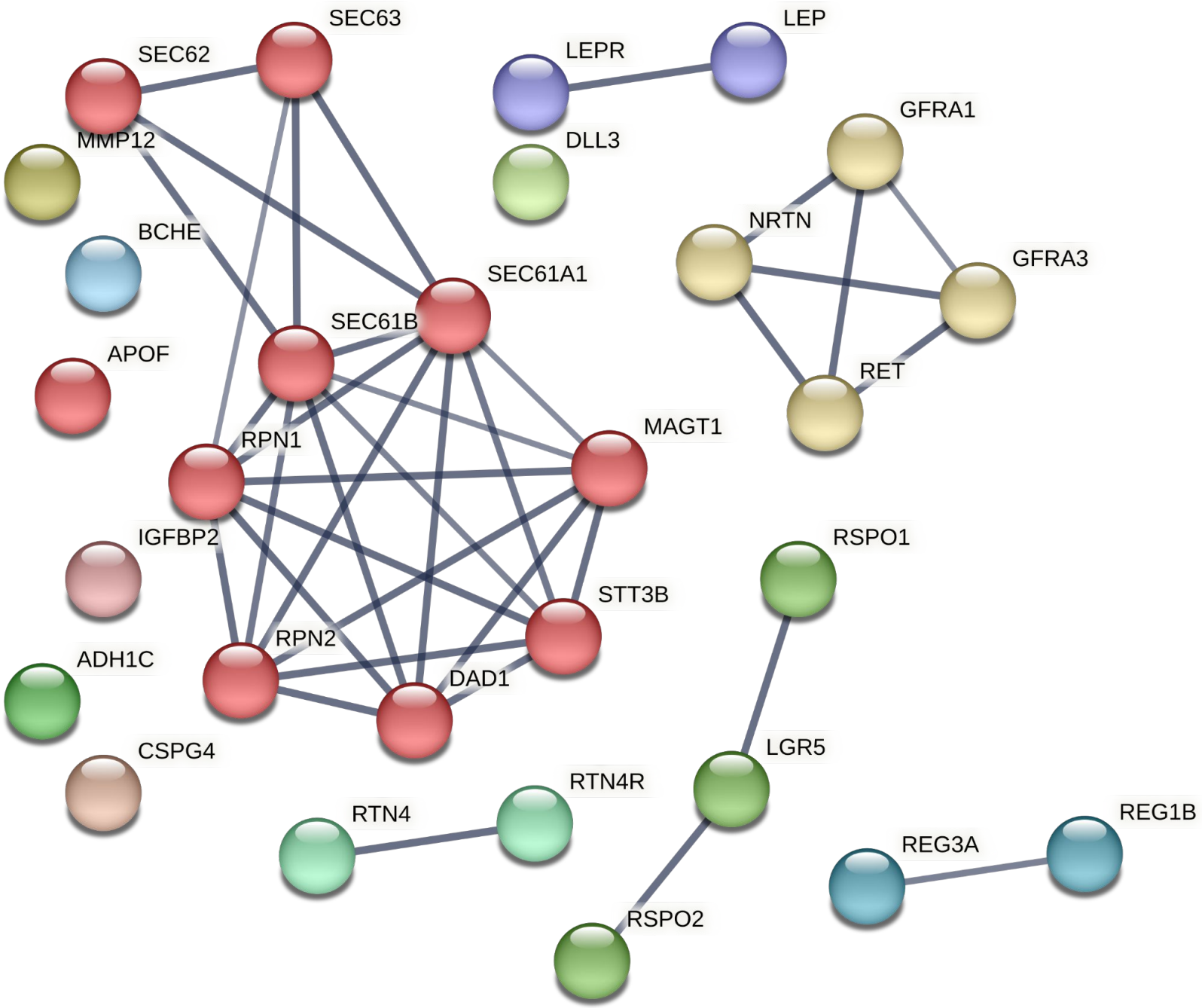
	COPDGene NHW testing set				**COPDGene AA testing set**				**ECLIPSE**			
COPD-related outcome	*Low PRS, High TRS*		*High PRS, High TRS*		*Low PRS, High TRS*		*High PRS, High TRS*		*Low PRS, High TRS*		*High PRS, High TRS*	
	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p
FEV1 % predicted	-22.3 (-29.5 to -15.2)	4.3E-09	-31.5 (-38.2 to -24.8)	1E-17	-4.11 (-16.8 to 8.59)	0.53	-19.6 (-32.7 to -6.51)	0.0048	-7.8 (-13.6 to -1.95)	0.01	-15.5 (-21.4 to -9.61)	1.3E-06
FEV1/FVC	-0.147 (-0.184 to -0.11)	3E-13	-0.197 (-0.233 to -0.161)	4.8E-22	-0.00368 (-0.0731 to 0.0657)	0.92	-0.118 (-0.193 to -0.043)	0.0032	-4.5 (-9.07 to 0.0772)	0.057	-8.83 (-13.5 to -4.16)	0.0003
% LAA < -950 HU	3.1 (1.8 to 5.2)	3.30E-05	3.4 (2.1 to 5.6)	2.30E-06	1.3 (0.44 to 3.7)	0.66	0.94 (0.23 to 3.9)	0.93	2.33 (-3.34 to 8)	0.42	6.97 (-0.0748 to 14)	0.056
Perc15	-11.2 (-18.6 to -3.72)	0.0037	-15.3 (-22.8 to -7.89)	7.6E-05	-0.305 (-17.1 to 16.5)	0.97	-4.66 (-26.2 to 16.9)	0.67	-0.994 (-24.1 to 22.1)	0.93	-19.2 (-49.3 to 10.8)	0.21
Pi10	0.498 (0.298 to 0.698)	2.10E-06	0.605 (0.429 to 0.78)	1.40E-10	0.346 (-0.032 to 0.724)	0.079	0.475 (0.00489 to 0.946)	0.054	-0.0786 (-0.147 to -0.00988)	0.029	-0.0753 (-0.163 to 0.0124)	0.095
WA %	5.88 (3.19 to 8.58)	2.90E-05	7.17 (4.63 to 9.71)	8.90E-08	4.55 (-1.6 to 10.7)	0.15	7.48 (-0.0534 to 15)	0.058	0.516 (-0.972 to 2)	0.5	2.18 (0.376 to 3.98)	0.02
FEV1 change (mL/year)	-30 (-70.6 to 10.7)	0.15	-43.9 (-91.7 to 3.94)	0.077	-103 (-294 to 87.5)	0.32	-60.5 (-105 to -15.6)	0.033	-23.7 (-52.9 to 5.47)	0.11	6.97 (-25.8 to 39.8)	0.68

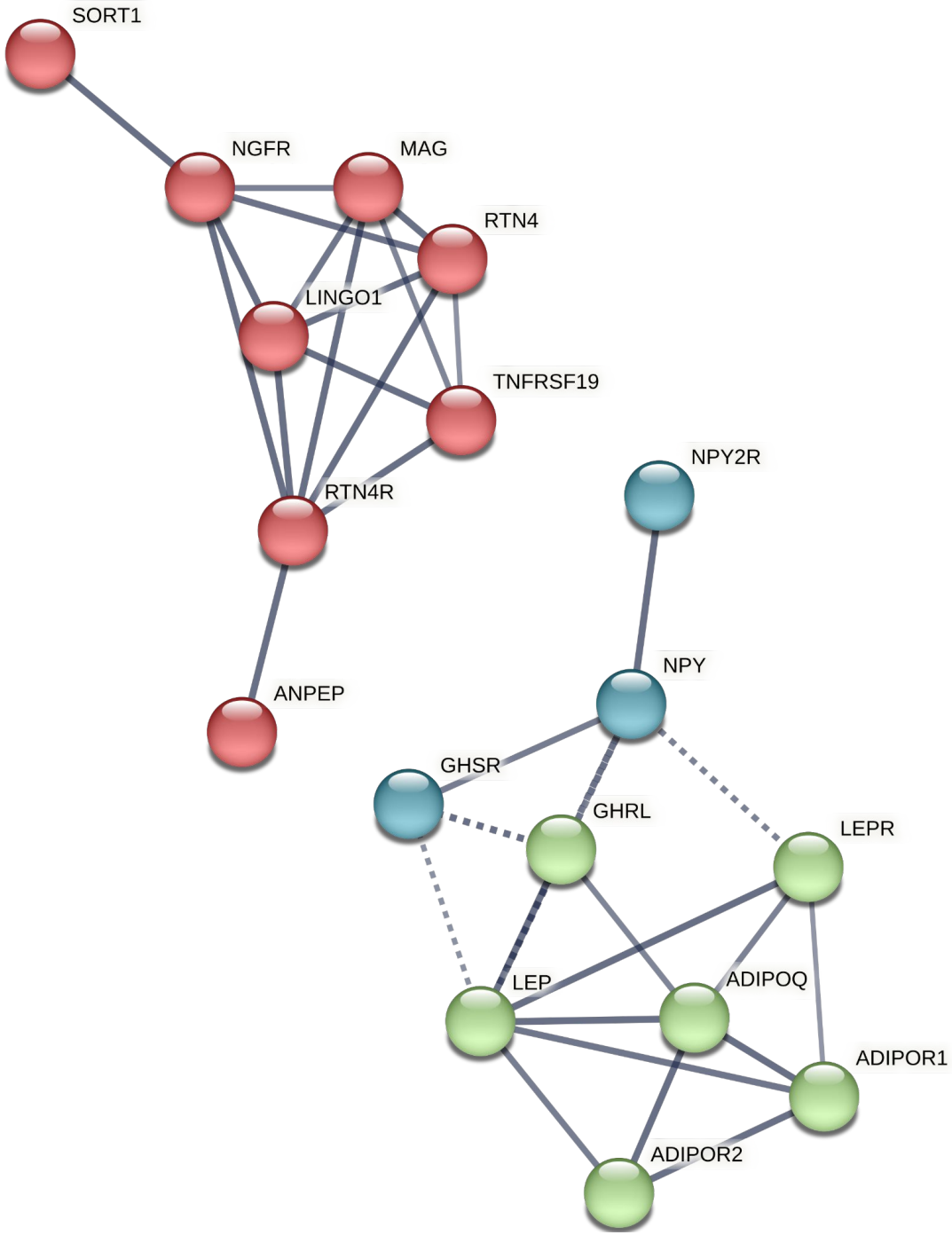
medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.20.24307621; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .


COPDGene training set
(NHW: n=1,374)
COPDGene testing set
(NHW: n=1,133
AA: n=299)
ECLIPSE
(n=468)
Determine PRS and TRS splits
Apply risk score cut offs and define subtypes
Identify differentially expressed proteins associated with subtypes
Regression analyses of subtype associations with clinical outcomes
STRING analysis
Enrichr/MAGMA drug repurposing analysis


low PRS
high TRS
high PRS
high TRS
High disease
activity subtype
Severe disease risk
subtype
low PRS
intermediate TRS
high PRS
intermediate TRS
low PRS
low TRS
high PRS
low TRS
–0.60
–0.65
–0.70
–0.75
Transcriptional risk score
–6
–3
0
3
Polygenic risk score


SEC63
SEC62
MMP12
BCHE
APOF
SEC61A1
SEC61B
RPN1
MAGT1
IGFBP2
STT3B
RPN2
ADH1C
DAD1
CSPG4
LEP
LEPR
DLL3
GFRA1
NRTN
GFRA3
RET
RSPO1
LGR5
RTN4R
RTN4
REG1B
REG3A
RSPO2


SORT1
NGFR
MAG
RTN4
LINGO1
TNFRSF19
RTN4R
ANPEP
NPY2R
NPY
GHSR
GHRL
LEPR
LEP
ADIPOQ
ADIPOR1
ADIPOR2