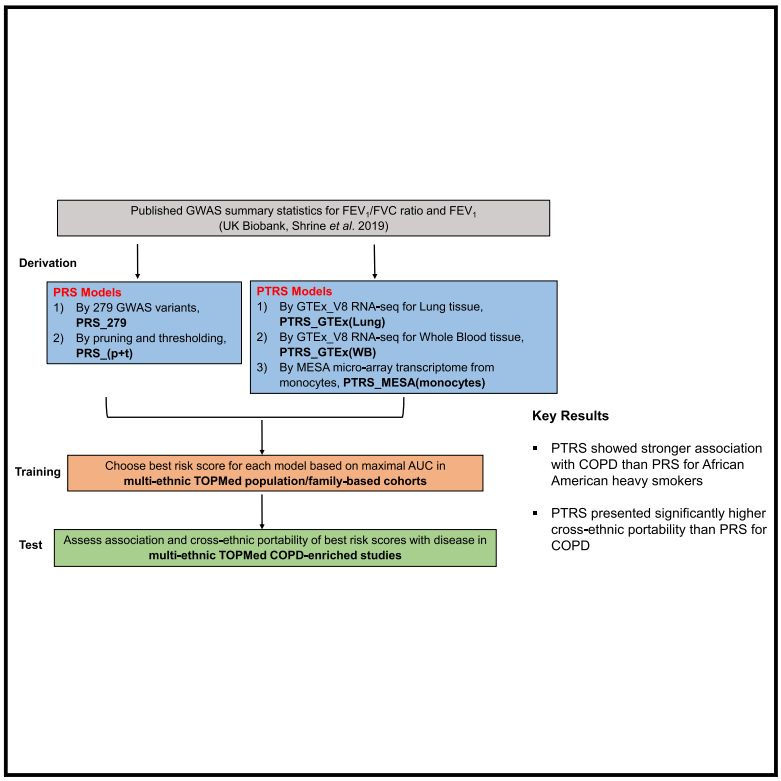
Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program

Graphical abstract



Authors

Xiaowei Hu, Dandi Qiao, Wonji Kim, ..., Michael H. Cho, Hae Kyung Im, Ani Manichaikul

Correspondence

am3xa@virginia.edu

Hu et al., 2022, The American Journal of Human Genetics *109*, 857–870
May 5, 2022
https://doi.org/10.1016/j.ajhg.2022.03.007

CellPress

Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program

Xiaowei Hu,[1] Dandi Qiao,[2] Wonji Kim,[2] Matthew Moll,[2,3] Pallavi P. Balte,[4] Leslie A. Lange,[5] Traci M. Bartz,[6,7] Rajesh Kumar,[8,9] Xingnan Li,[10] Bing Yu,[11] Brian E. Cade,[12,13] Cecelia A. Laurie,[6] Tamar Sofer,[12,13] Ingo Ruczinski,[14] Deborah A. Nickerson,[15,38] Donna M. Muzny,[16] Ginger A. Metcalf,[16] Harshavardhan Doddapaneni,[16] Stacy Gabriel,[17] Namrata Gupta,[17] Shannon Dugan-Perez,[16] L. Adrienne Cupples,[18,39] Laura R. Loehr,[19] Deepti Jain,[6] Jerome I. Rotter,[20] James G. Wilson,[21] Bruce M. Psaty,[22] Myriam Fornage,[11,23] Alanna C. Morrison,[11] Ramachandran S. Vasan,[24,25] George Washko,[26] Stephen S. Rich,[1] George T. O'Connor,[27] Eugene Bleecker,[10] Robert C. Kaplan,[28,29] Ravi Kalhan,[30] Susan Redline,[12,13] Sina A. Gharib,[31] Deborah Meyers,[10] Victor Ortega,[32] Josée Dupuis,[18] Stephanie J. London,[33] Tuuli Lappalainen,[34,35] Elizabeth C. Oelsner,[4] Edwin K. Silverman,[2,3] R. Graham Barr,[4] Timothy A. Thornton,[6] Heather E. Wheeler,[36] TOPMed Lung Working Group, Michael H. Cho,[2,3] Hae Kyung Im,[37] and Ani Manichaikul[1,*]

Summary

While polygenic risk scores (PRSs) enable early identification of genetic risk for chronic obstructive pulmonary disease (COPD), predictive performance is limited when the discovery and target populations are not well matched. Hypothesizing that the biological mechanisms of disease are shared across ancestry groups, we introduce a PrediXcan-derived polygenic transcriptome risk score (PTRS) to improve cross-ethnic portability of risk prediction. We constructed the PTRS using summary statistics from application of PrediXcan on large-scale GWASs of lung function (forced expiratory volume in 1 s [FEV_1] and its ratio to forced vital capacity [FEV_1/FVC]) in the UK Biobank. We examined prediction performance and cross-ethnic portability of PTRS through smoking-stratified analyses both on 29,381 multi-ethnic participants from TOPMed population/family-based cohorts and on 11,771 multi-ethnic participants from TOPMed COPD-enriched studies. Analyses were carried out for two dichotomous COPD traits (moderate-to-severe and severe COPD) and two quantitative lung function traits (FEV_1 and FEV_1/FVC). While the proposed PTRS showed weaker associations with disease than PRS for European ancestry, the PTRS showed stronger association with COPD than PRS for African Americans (e.g., odds ratio

[1]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA; [2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [3]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; [4]Departments of Medicine and Epidemiology, Columbia University Medical Center, New York, NY 10032, USA; [5]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado School of Medicine Anschutz Medical Campus, Aurora, CO 80045, USA; [6]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [7]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98101, USA; [8]Division of Allergy and Clinical Immunology, Ann and Robert H. Lurie Children's Hospital, Chicago, IL 60611, USA; [9]Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; [10]Department of Medicine, University of Arizona, Tucson, AZ 85724, USA; [11]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [12]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; [13]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA 02115, USA; [14]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA; [15]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [16]The Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [17]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [18]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA; [19]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA; [20]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [21]Division of Cardiovascular Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02115, USA; [22]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Systems and Population Health, University of Washington, Seattle, WA 98101, USA; [23]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [24]Boston University and the National Heart Lung and Blood Institute's Framingham Heart Study, Framingham, MA 01702, USA; [25]Department of Preventive Medicine and Epidemiology, School of Medicine and Public Health, Boston University, Boston, MA 02118, USA; [26]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; [27]Pulmonary Center, Boston University, School of Medicine, Boston, MA 02118, USA; [28]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA; [29]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [30]Department of Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; [31]Division of Pulmonary, Critical Care and Sleep Medicine, University of Washington, Seattle, WA 98109, USA; [32]Pulmonary and Critical Care, School of Medicine, Wake Forest University, Winston-Salem, NC 27157, USA; [33]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Durham, NC 27709, USA; [34]New York Genome Center, New York, NY 10013, USA; [35]Department of Systems Biology, Columbia University, New York, NY 10032, USA; [36]Department of Biology, Loyola University Chicago, Chicago, IL 60660, USA; [37]Section of Genetic Medicine, The University of Chicago, Chicago, IL 60637, USA
[38]Deceased December 24, 2021
[39]Deceased January 14, 2022
*Correspondence: am3xa@virginia.edu
https://doi.org/10.1016/j.ajhg.2022.03.007.

[OR] = 1.24 [95% confidence interval [CI]: 1.08–1.43] for PTRS versus 1.10 [0.96–1.26] for PRS among heavy smokers with ≥ 40 pack-years of smoking) for moderate-to-severe COPD. Cross-ethnic portability of the PTRS was significantly higher than the PRS (paired t test $p < 2.2 \times 10^{-16}$ with portability gains ranging from 5% to 28%) for both dichotomous COPD traits and across all smoking strata. Our study demonstrates the value of PTRS for improved cross-ethnic portability compared to PRS in predicting COPD risk.

Introduction

Chronic obstructive pulmonary disease (COPD), characterized by irreversible airflow obstruction, is currently a leading cause of death in the United States[1,2] and worldwide.[3] COPD is diagnosed using two spirometric measures of lung function, namely forced expiratory volume in one second (FEV_1) and its ratio to forced vital capacity (FEV_1/FVC). While the main risk factor for COPD is cigarette smoking, non-smokers can also develop COPD,[4,5] which suggests genetic variation in susceptibility to the disease. Furthermore, COPD is a highly heterogeneous disease with heritability estimates ranging from 35% to 60% even after accounting for smoking behavior.[6–8] Currently, there is no convincing therapy that prevents the development and progression of COPD, which reflects limited understanding of its biological mechanisms. Thus, early diagnosis can provide a crucial path toward prevention of more severe disease.

Large-scale genome-wide association studies (GWASs) of COPD and lung function have identified numerous genetic variants associated with COPD risk.[9–13] However, the individual contribution of the identified disease-associated variants to complex disease is generally very small.[14,15] The polygenic risk score (PRS) framework, aggregating the cumulative effects of genetic variants, tends to capture a reasonable proportion of variation in COPD risk and exhibits generally stronger association with disease when more genetic variants are included in the risk score.[9,11,16–18]

Additionally, PRS has the benefit that it can translate the results of GWASs into clinical application for the early identification of genetic risk of complex diseases. With recent increases in the scale of GWASs, the PRS approach has become more powerful. Many studies have demonstrated the predictive power of PRS on a wide range of complex diseases or traits.[19–23] However, populations with varying genetic ancestry may possess different allelic frequencies and linkage structures, and as a result, the predictive power of PRS is limited when the discovery and target populations are from different genetic ancestry groups, which is referred to as limited cross-ethnic portability. For example, a study of seventeen anthropometric and blood-panel quantitative traits in the UK Biobank has shown that prediction accuracy was far lower for non-European-ancestry populations when the PRS was derived from summary statistics for studies of individuals with European ancestry.[24]

For COPD, Shrine and colleagues derived a 279-variant weighted PRS from large-scale GWASs of lung function carried out in individuals with European ancestry from the UK Biobank.[11] Their results show that the derived PRS performed significantly better for individuals with European ancestry than for individuals with African ancestry in the external validation cohorts. In addition, Moll and colleagues derived an expanded PRS for COPD using Shrine's GWAS[11] results and showed that the gap in odds ratio of COPD between individuals with European and non-European ancestry increased with the decile of PRS.[18] As the majority of GWASs have been performed in individuals with European ancestry, disparity in prediction accuracy across non-European-ancestry individuals from GWAS-derived PRS is a major concern in consideration of potential clinical applications.[24,25,26]

While heterogeneity in genetic architectures limits cross-ethnic portability of PRS,[27–30] the results from several cross-ethnic GWAS replication studies provide evidence for some causal variants shared across populations.[31–34] Given that a substantial proportion of GWAS variants demonstrate gene regulation effects,[35] constructing risk scores built on expression quantitative-trait locus (eQTL) variants presents a promising path toward incorporating biological information in genetic prediction. Motivated by the hypothesis that the underlying biological mechanisms of trait or disease are shared across ancestry groups, Liang and colleagues proposed to improve cross-ethnic risk prediction using a polygenic transcriptome risk score (PTRS) constructed using multi-SNP predictors of gene expression.[36] The proposed PTRS builds on the widely used PrediXcan approach, an integrative method that leverages gene expression information to identify trait-associated genes from GWAS.[37] Compared with more conventional PRS approaches, the PTRS uses the cumulative effect of genes to construct genetic predictors of traits. Their results, which focused on seventeen anthropomorphic and blood phenotypes, showed substantial benefits in cross-ethnic portability from PTRS prediction compared with standard PRS approaches.[36]. Another benefit of the PTRS as a gene-based risk score is that it provides an additional layer of biological interpretability, as the PrediXcan-based predictors can be tied directly to gene expression traits corresponding to specific genes.

In this paper, we explored the benefits of PTRS for predicting COPD risk across self-reported race/ethnic groups by adapting the PTRS approach proposed by Liang et al.,[36] with the main difference being that the current work leverages summary statistics from published GWASs rather than individual-level data. We constructed PTRSs for prediction of two quantitative lung function traits (FEV_1 and FEV_1/FVC ratio) and two definitions of COPD (moderate-to-severe COPD and severe COPD) using summary statistics from the recent large-scale GWAS of pulmonary function traits (FEV_1 and FEV_1/FVC ratio) conducted in individuals with European ancestry from the UK

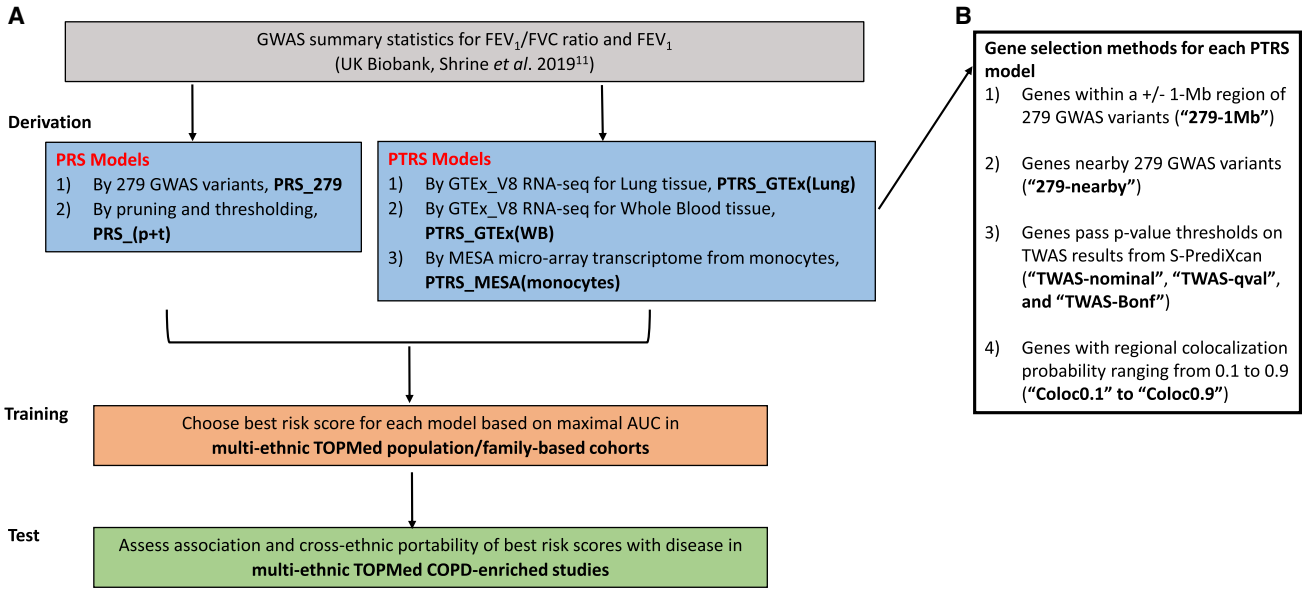


Figure 1. Study design
(A) Study workflow.
(B) Four methods of selecting genes included in each PTRS model.
PRS, polygenic risk score; PTRS, polygenic transcriptome risk score; 279 variants are the SNPs used to derive weighted genetic risk score for COPD in Shrine's work; TWAS, transcriptome-wide association study; S-PrediXcan, Summary-PrediXcan.

Biobank.[11]. We focused on these specific quantitative traits and definitions of COPD (1) for consistency with prior genetic studies[11–13] and risk scores for COPD[11,18] and (2) for their clinical relevance to diagnosis of COPD. We further proposed multiple approaches for the construction of PTRS, leveraging gene expression prediction functions from GTEx[38] and the Multi-Ethnic Study of Atherosclerosis (MESA).[39] To assess the prediction performance and the cross-ethnic portability of our proposed PTRS candidates, we leveraged multi-ethnic participants in the NHLBI Trans-Omics for Precision Medicine (TOPMed) program to select the best-performing candidates from population/family-based cohorts and then tested their performance on COPD-enriched studies.

Material and methods

Overview of approach

An overview of the study design is shown in Figure 1. For the derivation of risk scores (both PRS and PTRS), we leveraged summary statistics from the recent large-scale GWASs of pulmonary function traits (FEV_1 and FEV_1/FVC ratio) conducted in individuals with European ancestry from the UK Biobank.[11] For the two COPD definitions (moderate-to-severe COPD and severe COPD), the risk-score candidates were constructed using GWAS results for FEV_1/FVC ratio. We assessed these candidate scores on the TOPMed population/family-based cohorts (training data) to select the best-performing candidate by maximal area under the curve (AUC) from the set of risk scores derived for each model and trait. Both association and cross-ethnic portability (prediction accuracy for non-European ancestry relative to European ancestry) of the best risk scores with disease were then tested on TOPMed COPD-enriched studies.

PTRS derivation

Gene expression prediction models for integrative analysis

We selected two existing types of gene expression prediction models for investigation in construction of the PTRS:

- PTRS_GTEx models: based on European ancestry-dominant GTEx_V8 RNA-seq data. We used MASHR-based[40] prediction models that are recommended by the PrediXcan team for GTEx_V8 RNA-seq data.[38] In the analysis, both lung (n = 444, PTRS_GTEx[Lung]) and whole blood (n = 573, PTRS_GTEx[WB]) tissues of MASHR-based GTEx models were applied.
- PTRS_MESA model: based on multi-ethnic microarray transcriptome data collected from monocytes (total, n = 1,163; non-Hispanic Whites [NHW], n = 578; African Americans [AA], n = 233; and Hispanics/Latinos [HIS], n = 352).[39] We used the multi-ethnic elastic net prediction model (trained with mixing parameter $\alpha = 0.5$) as presented in Mogil et al.[39] for our analysis, and we refer this model as PTRS_MESA(monocytes).

Construction of risk scores

Building on the concept of genetically regulated gene expression (GReX) introduced as part of the widely used PrediXcan framework,[37] we calculated the PTRS for the i^{th} individual using the following formula:

$$PTRS_i = \sum_{j=1}^{m} T_{ij}\widehat{\gamma}_j,$$

where T_{ij} is the GReX of gene j for individual i, the calculation of T_{ij} is detailed in PrediXcan framework,[37] $\widehat{\gamma}_j$ is the estimated effect size of gene j, the calculation of $\widehat{\gamma}_j$ is detailed in Summary-PrediXcan (S-PrediXcan),[41] and m is the total number of genes. The PTRS approach used in the current manuscript was adapted from

previous work,[36] with the main difference being that the current work leverages summary statistics from published GWASs rather than individual-level data. We applied S-PrediXcan[41] using the selected gene expression prediction models with the published lung function GWASs to obtain the estimated effects corresponding to each gene ($\widehat{\gamma}_j$). Finally, the PTRS was inverse-normal transformed in the analysis.

We then applied four methods to select genes for inclusion in the PTRS:

(1) The first method was to take genes within a ± 1-Mb region of 279 variants from previous GWASs[11] and then overlap with genes from each of the three PTRS models ("279-1Mb").
(2) The second method is a variation of the first method (above). We restricted to 261 genes identified in previous GWAS (Table S9 in Shrine's work[11]) harboring the 279 variants and then selected genes overlapping with each of the three PTRS models ("279-nearby").
(3) The third method was to apply different p value thresholds on transcriptome-wide association study (TWAS) results from S-PrediXcan. Nominal, qval, and Bonf represented the p value, q value, and Bonferroni corrected cut-offs at 0.05, respectively ("TWAS-nominal," "TWAS-qval," and "TWAS-Bonf").
(4) The last method was to select genes by regional colocalization probability (RCP). We applied FastEnloc[42] on eQTLs and GWASs[11] to compute RCP for genes. For eQTLs, we adopted GTEx_V8 eQTL[43] for PTRS_GTEx models and MESA eQTLs[39] for PTRS_MESA models. The genes were selected first by the range of RCP (from 0.1 to 0.9) and then overlapped with genes from each of the three PTRS models ("Coloc0.1" to "Coloc0.9").

A summary of the number of genes selected by each method is reported in Table S1.

PRS calculation

To provide a comparison with our proposed PTRS, we incorporated in our study two models that reflect a more standard PRS framework. We further applied inverse normal transformation to the PRS values to standardize the scores.

- PRS_279 model: denotes the previously published genetic risk score that leverages weights for 279 selected variants from previous GWASs[11] ("SNPs-279").
- PRS_(p+t) model: applied pruning and thresholding by PLINK 1.90b –clump and –score,[44] a p value and linkage disequilibrium (LD)-driven procedure, to build additional PRS candidates where 1,864 European ancestry samples from MESA who had-whole genome sequence data through TOPMed were used to construct a LD reference panel. For each trait, a range of p values (5×10^{-4} and 5×10^{-8}) and pairwise correlation r^2 (0.2, 0.4, 0.6, and 0.8) thresholds were used to create an additional eight PRS candidates ("5e-4_0.2" to "5e-8_0.8").

Study samples

The training data comprised the participants from eight population/family-based cohorts (the Atherosclerosis Risk in Communities [ARIC] study, the Coronary Artery Risk Development in Young Adults [CARDIA] study, the Cleveland Family Study [CFS], the Cardiovascular Health Study [CHS], the Framingham Heart Study [FHS], the Hispanic Community Health Study/Study of Latinos [HCHS/SOL], the Jackson Heart Study [JHS], and the Multi-Ethnic Study of Atherosclerosis [MESA]). The test data consisted of the participants from two COPD-enriched studies (the Genetic Epidemiology of COPD [COPDGene] study and the SubPopulations and Intermediate Outcome Measures in COPD Study [SPIROMICS]). For all of the included studies, Institutional Review Boards at each field center approved study protocols, and written informed consent was obtained from all participants. Detailed cohort descriptions are provided in the supplemental methods.

Whole-genome sequence data

Whole-genome sequencing (WGS) in TOPMed had, on average, deep (~30×) coverage with joint-sample variant calling and variant level quality control in ~140,000 TOPMed samples for Freeze 8 and ~159,000 samples for Freeze 9b.[45] Analyses in the population/family-based cohorts, as well as COPDGene, used WGS from TOPMed Freeze 8. Study-specific analyses in SPIROMICS used the newer Freeze 9b WGS genotypes as (1) these analyses were carried out at a later stage in our research, and (2) the SPIROMICS WGS data were only available starting from the newer Freeze 9b release. Additional details regarding quality control of genotype data for the present analyses are included in the supplemental methods.

Phenotype definition

The phenotype harmonization of pulmonary function traits (pre-bronchodilator FEV_1 and FEV_1/FVC ratio) was conducted following the protocol of the NHLBI Pooled Cohorts Study (supplemental methods, Oelsner et al.[17]). We followed Zhao et al.[13] to proceed with phenotype QC and calculate the race/ethnic-specific predicted values of FEV_1 using the equations of Hankinson[46] that were determined for NHW, AA, and HIS, respectively, COPD cases, and controls were then defined as follows:

- Moderate-to-severe COPD: pre-bronchodilator $FEV_1 < 80\%$ predicted and $FEV_1/FVC < 0.7$
- Severe COPD: pre-bronchodilator $FEV_1 < 50\%$ predicted and $FEV_1/FVC < 0.7$
- Controls: pre-bronchodilator $FEV_1 \geq 80\%$ predicted and $FEV_1/FVC \geq 0.7$

Statistical analysis to examine prediction performance

We carried out pooled analyses across self-reported race/ethnic groups for the training data (population/family-based cohorts). For the test data (COPD-enriched studies), analyses were stratified by self-reported race/ethnic group (NHW versus AA) and then meta-analyzed using an inverse-variance weighted fixed effect model. Statistical analyses were conducted using R/GENESIS v.2.21.3,[47] and meta-analyses were implemented in R/meta v4.13-0.[48]

For dichotomous traits, the score with the best prediction accuracy for each set of risk scores corresponding to each model was determined by the maximal AUC. The AUC was calculated using a generalized linear mixed model for association of the dichotomous trait with thescore candidate and including additional covariate adjustment for age, sex, race, study, sequence center,

pack-years of smoking, ever versus never smoking, and principal components (PCs) of ancestry. The genetic relationship matrix (GRM) was used to specify the covariance structures of the random effects term in the model. The AUC was calculated by risk score only, and the confidence intervals of AUC were calculated using R/pROC v.1.17.0.1.[49]

For quantitative traits, we followed Zhao et al.[13] and Sofer et al.[50] to obtain study-specific variance adjusted residuals as phenotypes for the analyses. More specifically, we applied linear mixed models to obtain study-specific residuals, along with study-specific standard deviations of the residuals. The inverse-normal transformed residuals were scaled by their study-specific standard deviations. The resulting values were used to assess the association with each proposed risk score using a linear mixed model. The linear mixed models included covariate adjustment for age, age-squared, sex, height, height-squared, race, study, sequence center, pack-years of smoking, current smoking, former smoking, PCs of ancestry, and the GRM. Prediction performance of each score was quantified as the proportion of variance explained (%), estimated as 100 × the squared correlation (R^2) between the observed phenotypes and the predicted phenotypes by score only.

The GRM of samples and PCs of ancestry for all studies except SPIROMICS were generated on TOPMed Freeze 8 and obtained directly from the TOPMed Data Coordinating Center. For SPIROMICS, the GRM and PCs were based on TOPMed Freeze 9b and were computed by following TOPMed Freeze 8 procedures using R/GENESIS v.2.21.3.[47] Analyses in TOPMed Freeze 8 (population/family-based and COPDGene) included adjustment for the first 11 PCs of ancestry, whereas analyses in SPIROMICS included adjustment for the first 4 PCs of ancestry after checking pairwise PC plots.

Smoking interaction

For the best-performing risk-score candidate identified for each model, the smoking × score interaction effects were assessed by adding an interaction term for pack-years of smoking × score in the (generalized) linear mixed models for each of the four traits (moderate-to-severe COPD, severe COPD, FEV_1, and FEV_1/FVC ratio) on population/family-based cohorts.

Portability analysis

Cross-ethnic portability was defined as the prediction accuracy (AUC) ratio for non-European- versus European-ancestry populations. We applied bootstrapped sampling (i.e., random sampling with replacement) on two COPD-enriched studies to generate 95% confidence intervals of cross-ethnic portability. For each bootstrapped sample of COPD cases and controls, we calculated the cross-ethnic portability. The 95% CIs for portability estimates were then obtained using the percentile method on 10,000 bootstrapped samples, separately for each of the two COPD-enriched studies.

Examination of a combined risk score

To explore the performance of a single risk score that combines PRS and PTRS, we selected one candidate to represent each approach (PRS_279: SNPs-279 and PTRS_GTEx(Lung): 279-nearby) for further investigation. These two risk scores are relevant but provide different levels of genetic risk information. We first examined the interaction between these two scores in the training data (population/family-based cohorts). The interaction was assessed by adding a score interaction term in the same prediction model as for risk-score prediction evaluation. We then explored two schemes to combine two individual risk scores, unweighted sum and weighted sum. The unweighted sum is obtained as the direct sum of the two individual risk scores. The weights in the weighted sum were obtained as the regression coefficients of two individual risk scores in the score interaction model using the population/family-based cohorts. The weights were also applied to COPD-enriched studies to calculate the combined risk score. Finally, we evaluated the predictive performance of the risk scores for each COPD trait (1) using the same prediction model used for our primary analyses as described above, (2) using a baseline set of clinical risk factors alone (age, sex, race, pack-years of smoking), (3) using the risk score alone, and (4) using the combination of clinical risk factors and risk score.

Results

Participant characteristics

Demographic and clinical characteristics of our study samples are summarized in Table 1, which includes 29,381 participants from the eight population/family-based cohorts and 11,771 participants from the COPD-enriched studies. Based on participant self-reported race/ethnicity, 50% and 74% of the participants were categorized as NHW in the population/family-based cohorts and COPD-enriched studies, respectively. The remaining participants represented AA (24% and 26% in the population/family-based and COPD-enriched studies, respectively) and HIS (26% in the population/family-based cohorts).

Selection of best-performing risk-score candidate for each model

The overview of study design is shown in Figure 1. We used large-scale GWASs of individuals with European ancestry from the UK Biobank reported by Shrine et al.[11] (n = 321,047) to derive both PRS and PTRS candidates for FEV_1/FVC ratio and FEV_1, respectively. For the two COPD definitions (moderate-to-severe COPD and severe COPD), the risk-score candidates were constructed using GWAS results for FEV_1/FVC ratio. Specifically, we derived 42 candidates for PTRS by three different transcriptome reference models (i.e., PTRS_GTEx[Lung], PTRS_GTEx[WB], and PTRS_MESA[monocytes]) and by four different gene selection methods for each transcriptome reference model (i.e., "279-1Mb," "279-nearby," "TWAS-nominal, qval, and Bonf," and "Coloc0.1 to Coloc0.9"), and 9 candidates for PRS by two models (i.e., PRS_279 and PRS_[p+t]) for each of the complex traits (material and methods). We assessed these candidate scores on the TOPMed population/family-based cohorts (training data) to select the best-performing risk score by maximum AUC for each model and for each trait. Both association strength and cross-ethnic portability of the best risk scores with diseases were then tested in the TOPMed COPD-enriched studies (Figure 1).

Figure 2 shows that the two definitions of COPD (moderate-to-severe COPD and severe COPD) shared the same

Table 1. Characteristics of the study-participants included in the analyses

Stratum	Study	NHW	AA	HIS	Age, years	Female (%)	Smoking pack-years	$FEV_{1\%}$ predicted	FEV_1/FVC ratio	Moderate-to-severe COPD	Severe COPD
Population- and family-based	ARIC	5,717	1,458	–	62.81 ± 9.69	3,998 (56)	17.28 ± 23.41	92 ± 19	0.73 ± 0.09	1,115	199
	CARDIA	1,386	1,373	–	42.17 ± 6.62	1,525 (55)	0.71 ± 1.58	95 ± 14	0.79 ± 0.06	–	–
	CFS	402	388	–	47.21 ± 16.07	432 (55)	10.29 ± 16.26	91 ± 20	0.79 ± 0.07	64	16
	CHS	2,321	312	–	78.74 ± 6.09	1,574 (60)	16.08 ± 24.85	91 ± 25	0.72 ± 0.11	500	155
	FHS	3,321	–	–	48.86 ± 11.85	1,771 (53)	6.72 ± 15.76	96 ± 15	0.76 ± 0.07	239	26
	HCHS/SOL	–	–	6,750	46.65 ± 13.64	3,969 (59)	7.43 ± 15.97	92 ± 15	0.80 ± 0.07	394	69
	JHS	–	2,511	–	54.54 ± 12.54	1,621 (65)	–	93 ± 18	0.81 ± 0.08	123	26
	MESA	1,580	983	879	66.40 ± 9.84	2,088 (52)	10.69 ± 20.90	94 ± 18	0.75 ± 0.08	408	52
	Total	14,727	7,025	7,629	–	–	–	–	–	2,843	543
COPD-enriched	COPDGene	6,609	3,258	–	59.55 ± 9.04	4,602 (47)	44.27 ± 24.87	73 ± 26	0.65 ± 0.16	3,981	2,022
	SPIROMICS	1,535	369	–	63.50 ± 9.06	886 (47)	47.60 ± 27.91	67 ± 27	0.59 ± 0.16	1,115	547
	Total	8,144	3,627	–	–	–	–	–	–	5,096	2,569

Mean ± standard deviation. ARIC, Atherosclerosis Risk in Communities; CARDIA, Coronary Artery Risk Development in Young Adults; CFS, Cleveland Family Study; CHS, Cardiovascular Health Study; FHS, Framingham Heart Study; HCHS/SOL, Hispanic Community Health Study/Study of Latinos; JSH, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis; COPDGene, Genetic Epidemiology of COPD; SPIROMICS, Sub-Populations and Intermediate Outcome Measures in COPD Study; NHW, non-Hispanic White; AA, African American; HIS, Hispanic; FEV_1, forced expiratory volume in 1 s; FEV_1/FVC ratio, FEV_1 ratio to forced vital capacity.

pattern of the best-performing risk-score candidates for each model. Taking moderate-to-severe COPD for example, PRS_279 had the best prediction accuracy overall (AUC = 0.579 [95% CI: 0.567–0.590]). The best-performing candidate for the PRS pruning and thresholding model was from the accumulation of independently genome-wide significant variants (i.e., PRS_[p+t]: 5e-8_0.2 with AUC = 0.566 [95% CI: 0.555–0.578]). Among our proposed PTRSs, the PTRS with genes near 279 variants had the best AUC for PTRS_GTEx(Lung) model (i.e., PTRS_GTEx[Lung]: 279-nearby with AUC = 0.549 [95% CI: 0.537–0.560]). Among the PTRS_GTEx(WB) model risk scores, the candidate with genes selected using the q value method was the best performing (i.e., PTRS_GTEx[WB]: TWAS-qval with AUC = 0.525 [95% CI: 0.513–0.536]). The PTRS with the second-largest gene size was the best performing for the MESA model (i.e., PTRS_MESA[monocytes]: TWAS-nominal with AUC = 0.537 [95% CI: 0.525–0.548]). The best PRS candidate (PRS_279) has significantly higher AUC than the best PTRS candidate (PTRS_GTEx[Lung]: 279-nearby) for moderate-to-severe COPD (Delong p value = 7.46×10^{-6}). The detailed prediction performance of all proposed risk scores for the two COPD traits are shown in Tables S2 and S3.

The genes included in the best-performing PTRS for two COPD traits may provide additional information to prioritize genes for further investigation of the underlying biological mechanism of COPD (Tables S4–S6). Taking the best-performing PTRS of the GTEx(Lung) model (i.e., PTRS_GTEx[Lung]: 279-nearby) for example, there are 126 genes included in this PTRS that represent a subset of the 261 genes identified based on GWASs of lung function in the UK Biobank.[11] Among these, 43 of the genes included in the risk score achieved Bonferroni significance (i.e., TWAS p value < 0.05/126, Table S4). Furthermore, the small number of overlapped genes among the three best-performing PTRS candidates suggested that the best candidate from each PTRS model provided information on a relatively distinct set of genes (Table S1 and Figures S2–S4).

For the two quantitative lung function traits, PRS_279 has overall better prediction accuracy than the other risk scores examined for both traits, and the best-performing PTRS candidate differed by trait (Figure S1). Overall, the risk scores from the model PTRS_GTEx(Lung) presented higher R^2 among all PTRS candidates for both traits. The detailed prediction results for all proposed risk scores are shown in Tables S7 and S8, and the genes included for the best-performing score from each PTRS model for two lung function quantitative traits are shown in Tables S9–S13.

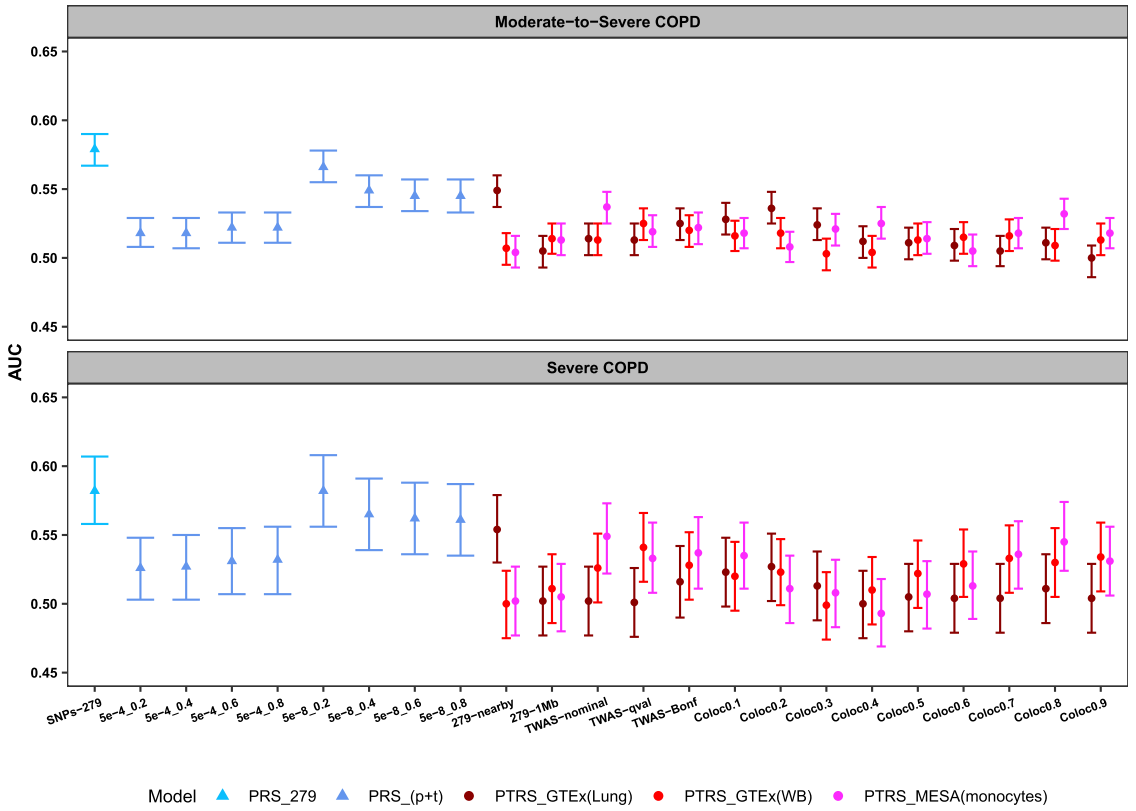


Figure 2. Prediction accuracy of all risk-score candidates on multi-ethnic population/family-based cohorts for two COPD traits
AUC, area under the curve, was evaluated by the risk score only. Data are shown as AUC with 95% CI. PRS_279, PRS derived by previously published 279 variants for FEV_1/FVC ratio; PRS_(p+t), PRS derived by pruning and thresholding (a range of p value and pairwise correlation thresholds were used to create eight candidates, 5e-4_0.2 to 5e-8_0.8); 279-nearby and 279-1Mb, PTRS derived by genes nearby and within ± 1 Mb region of 279 variants, respectively; TWAS-nominal, TWAS-qval, and TWAS-Bonf, PTRS derived by genes passing TWAS p value threshold of 0.05, q value, and Bonferroni, respectively; Coloc0.1 to Coloc0.9, PTRS derived by genes with regional colocalization probability ranging from 0.1 to 0.9.

PTRS presents stronger association than PRS with COPD in African Americans

Defining subgroups for stratified analysis

As shown in Table 1, the participants in COPD-enriched studies had heavier smoking history than those in TOPMed population/family-based cohorts. Hence, to examine the association strength of best risk scores on COPD-enriched studies, we first examined the impact of pack-years of smoking on the relationship between the proposed risk scores and COPD risk via smoking interaction analysis (material and methods). For each of the four traits (two definitions of COPD and two quantitative traits: FEV_1 and FEV_1/FVC ratio), at least one selected candidate score showed nominally significant interaction with smoking (i.e., interaction p value < 0.05, Table S14). We then conducted smoking-stratified analyses on COPD-enriched studies to examine the association performance of best risk scores on different smoking strata.

Within smoking strata, we undertook separate analyses for NHW and AA. Due to the limited samples with pack-years of smoking < 20 in SPIROMICS (Table S15), we only applied analyses in COPDGene for the light smokers (i.e., pack-years of smoking < 20) for all four traits. For each of the five risk-score models (i.e., PRS_279, PRS_[p+t], PTRS_GTEx[Lung], PTRS_GTEx[WB], and PTRS_MESA [monocytes]), we selected the best risk scores based on their AUCs in each smoking stratum on population/family-based cohorts and then applied them in analysis of COPD-enriched studies (Tables S16–S19). The meta-analyzed odds ratios for the association of best risk scores with COPD traits on COPD-enriched studies are shown in Figure 3A. Overall, PRS_279 still showed stronger association with both COPD traits in NHW participants from COPD-enriched studies and for each smoking strata (e.g., odds ratio [OR] = 1.57 [95% CI: 1.48–1.67] for moderate-to-severe COPD and OR = 1.66 [95% CI: 1.55–1.79] for severe COPD for smoking pack years ≥ 0). For light smokers (i.e., pack-years of smoking < 20) in AA, PTRS showed either a similar or stronger association than PRS with two COPD traits (OR = 1.33 [95% CI: 1.06–1.68] from PTRS_GTEx[WB] versus 1.33 [95% CI: 1.06–1.68] from PRS_279 for moderate-to-severe COPD, and OR = 1.51 [95% CI: 1.04–2.19] from PTRS_GTEx[WB] versus 1.31 [95% CI: 0.87–1.96] from PRS_279 for severe COPD). For heavy smokers (i.e., pack-years of smoking ≥ 40) in AA, the PTRS presented a stronger association than the PRS for moderate-to-severe COPD (OR = 1.24 [95% CI: 1.08–1.43] from PTRS_GTEx[Lung] versus OR = 1.10 [95%

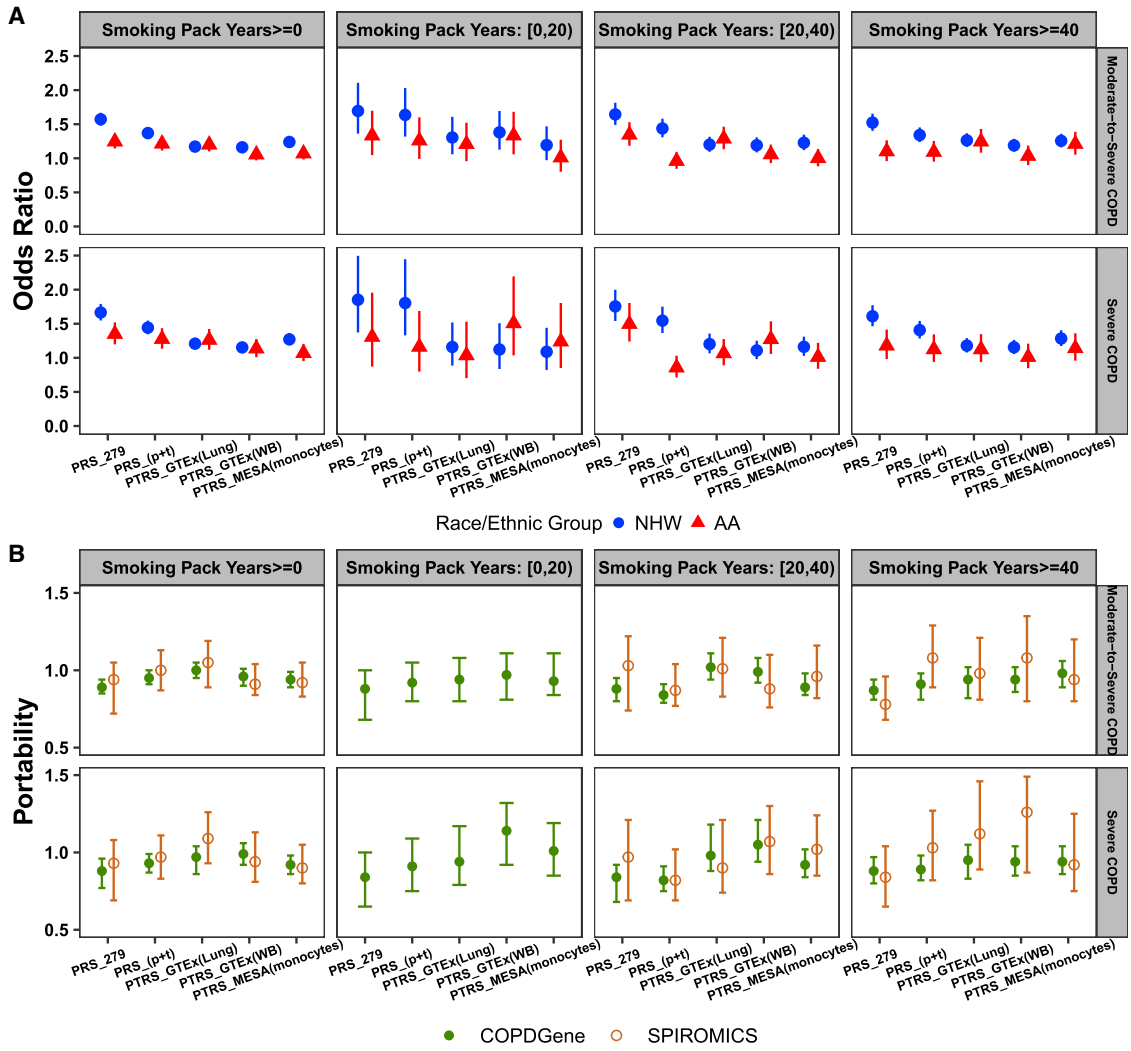


Figure 3. Association and cross-ethnic portability of best risk scores with two COPD traits in COPD-enriched studies
The risk-score candidates used in the analyses were based on the best AUC of each smoking stratum on multi-ethnic population/family-based cohorts for both PRS and PTRS.
(A) Association of the best risk scores with two COPD traits. NHW, non-Hispanic Whites; AA, African Americans. Data are shown as meta-analyzed odds ratio with 95% CI.
(B) Cross-ethnic portability of the best risk scores; portability was calculated as the ratio of AUC of AA over NHW. Data are shown as raw portability, and error bars are 95% CIs from 10,000 bootstrapped samples.

CI: 0.96–1.26] from PRS_279). For the other two lung function traits, FEV_1/FVC ratio and FEV_1, PRS_279 outperformed in both NHW and AA (Figures S5 and S6).

PTRS improves cross-ethnic portability of prediction

The noticeably decreased performance of PRS from NHW to AA is shown in Figure 3A. For example, considering the performance of all participants (i.e., pack-years of smoking ≥ 0) for moderate-to-severe COPD, the OR was 1.57 [95% CI: 1.48–1.67] by PRS_279 for NHW, but it dropped to 1.24 [95% CI: 1.14–1.36] for AA (Table S16). To test the cross-ethnic portability of prediction for both PRS and PTRS, we generated 10,000 bootstrapped samples for two COPD-enriched studies (material and methods). By definition of cross-ethnic portability, the reference portability is 1. As shown in Figure 3B, the PTRS models retained overall better portability than that from PRS models for both definitions of COPD. More specifically, we compared the performance of the best portability between PTRS and PRS. For example, in pooled analysis across all smoking strata (i.e., pack-years of smoking ≥ 0) for moderate-to-severe COPD, the PTRS with the best portability was PTRS_GTEx(Lung) model (portability = 1 [95% CI: 0.95–1.05] for COPDGene and portability = 1.05 [95% CI: 0.89–1.19] for SPIROMICS), whereas the PRS with the best portability was the PRS_(p+t) model (portability = 0.95 [95% CI: 0.91–1] for COPDGene and portability = 1 [95% CI: 0.87–1.13] for SPIROMICS). Hence, the gain of portability from PTRS was 5% from both cohorts in this smoking strata for moderate-to-severe COPD, and based on a paired t test comparing the bootstrapped distributions, the PTRS_GTEx(Lung) model has significantly higher portability

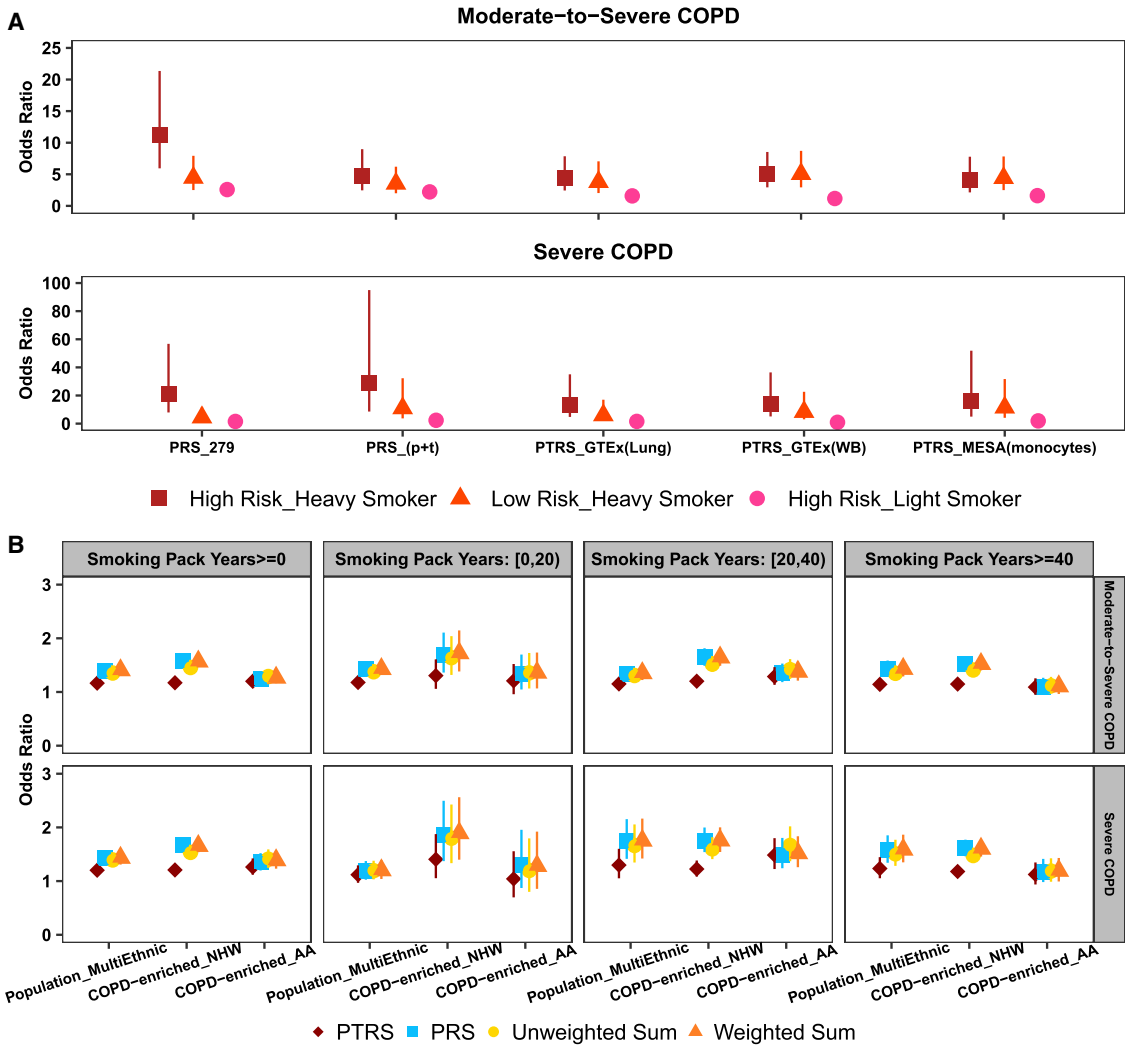


Figure 4. Risk for two COPD traits by different risk groups and by the combined risk scores

(A) Odds ratios of two COPD traits for different risk groups on multi-ethnic population/family-based cohorts. The reference group was defined as low risk and light smoker; low and high risk are referring to the 1st and the 5th quintile of genetic risk score, respectively; light and heavy smokers are referring to the participants with pack-years of smoking < 20 and ≥ 40, respectively; the risk-score candidates used in the analyses were based on the prediction performance of non-smoking stratum on population/family-based cohorts. Data are shown as odds ratio with 95% CI.

(B) Association of the combined risk scores with two COPD traits. The risk-score candidates, PTRS_GTEx(Lung): 279-nearby and PRS_279: SNPs-279 were used for PTRS and PRS, respectively, in the analyses. Unweighted sum and weighted sum refer to the direct summation and the weighted summation of PTRS and PRS, respectively. Data are shown as odds ratio with 95% CI. For COPD-enriched studies, the odds ratios were meta-analyzed. NHW, non-Hispanic Whites; AA, African Americans; Population_MultiEthnic, multi-ethnic samples in population/family-based cohorts.

than that from the PRS_(p+t) model for both cohorts ($p < 2.2 \times 10^{-16}$). We also observed the significantly improved portability from PTRS (gain ranges from 5% to 28%, Tables S16 and S17) for the other three smoking strata for both definitions of COPD.

Comparison of genetic versus smoking-related risk of disease

To aid with practical interpretation, we examined the impact of the combination of genetic risk scores and smoking history on the risk of COPD. These exploratory analyses were conducted on population/family-based cohorts only, as these studies allowed us to examine the population-level risk of disease. Participants were divided into risk categories by the values of both risk scores (quintiles) and pack-years of smoking (< 20 and ≥ 40). The reference group was defined as the participants with low genetic risk (i.e., 1st quintile of genetic risk score) and light smoking (i.e., pack-years of smoking < 20). Comparing the individuals with high genetic risk (i.e., 5th quintile of genetic risk score) and heavy smoking history (i.e., pack-years of smoking ≥ 40) to those in the reference group, the OR by the PRS_279 model was 11.26 (95% CI: 5.94–21.35) for moderate-to-severe COPD and 21.21 (95% CI: 7.93–56.74) for severe COPD (Figure 4A). The impact of smoking history was observed to be greater than genetic risk, as quantified by

either the PRS or the PTRS. Taking the results from PRS_279 for example, if an individual has low genetic risk but heavy smoking, then the OR for moderate-to-severe COPD was 4.45 (95% CI: 2.50–7.93). However, for an individual in the light smoking group, even with high genetic risk, the OR was comparably lower at 2.58 (95% CI: 2.14–3.11) for moderate-to-severe COPD (Figure 4A). The same pattern was also observed for severe COPD.

Combined PRS and PTRS improves association strength

To explore the performance of a single risk score that borrows information from both PRS and PTRS for COPD, we selected one risk score representing each approach, PRS_279: SNPs-279 and PTRS_GTEx(Lung): 279-nearby, for further examination. These two risk scores are correlated (Pearson's correlation = 0.27, $p < 2.2 \times 10^{-16}$) but provide different levels of genetic risk information. We first examined the interaction between the two scores in the population/family-based cohorts (material and methods). The significant main effects (i.e., PRS and PTRS effects) and non-significant interactions indicated that two individual risk scores provided independent effects for all four traits (Table S20). Then we applied unweighted- and weighted-sum schemes to combine two individual risk scores into a single score. Figure 4B presents the association results of combined risk score with two COPD traits on both training and test data. In general, the weighted-sum score and PRS showed similar strength of association, and both presented stronger association than unweighted-sum score and PTRS on both NHW dominant training data (population/family-based cohorts) and NHW participants in test data (COPD-enriched studies). The similar performance between the weighted-sum score and the PRS for NHW can be explained by the major contribution from PRS to this combined score (i.e., the weights mainly come from PRS, Table S20). For AA, the unweighted-sum score that equally borrows information from both PRS and PTRS produced noticeable improvement. Taking the AA participants with pack-years of smoking between 20 and 40 for example, the OR was increased from 1.50 (95% CI: 1.24–1.80) by PRS to 1.66 (95% CI: 1.37–2.02) by the unweighted-sum score for severe COPD, which produced 10% increment for OR. The same pattern of improvement was also observed for the two lung traits FEV_1/FVC and FEV_1 (Figures S7 and S8). For the prediction accuracy evaluated by AUC, the weighted-sum score presented overall outperformance for both COPD traits among risk scores in population/family-based cohorts (Tables S21 and S22). Although the AUC achieved by a baseline model (i.e., AUC based on clinical risk factors) was higher than the AUC for the risk score alone, we observed a trend of lower baseline AUC and high risk-score AUC among heavy smokers compared to those with reduced smoking exposures. Taking moderate-to-severe COPD for example, the baseline AUC dropped from 0.742 (95% CI: 0.729–0.754) for light smokers (i.e., pack-years of smoking < 20) to 0.63 (95% CI: 0.604–0.657) for heavy smokers (i.e., pack-years of smoking ≥ 40), whereas the weighted-sum risk-score AUC increased from 0.592 (95% CI: 0.576–0.608) to 0.599 (95% CI: 0.572–0.625) (Table S21).

Discussion

In the current manuscript, we proposed and applied an integrative framework to quantify genetic risk of COPD and predict quantitative lung function traits. Our proposed polygenic transcriptomic risk score (PTRS) framework, built on the widely used PrediXcan approach used for systematic integration of GWASs with reference eQTL data,[39,43] bears a more direct connection to underlying disease biology than standard PRS approaches. Hypothesizing that the underlying biology of complex disease traits is shared across diverse race/ethnic groups, we further anticipated that risk scores constructed under our PTRS framework would have greater portability than the standard PRS. Our application of PTRS to prediction of COPD in African American individuals from COPD-enriched studies demonstrated that our proposed PTRS had better portability for prediction of both moderate-to-severe COPD and severe COPD than the PRS approaches that we examined. Further, examining correlation of our PTRS with a standard PRS, we showed that the two classes of scores are not strongly correlated and thus present independent and complementary information that can be combined.

As the PTRS approaches are restricted primarily to eQTL variants, the number of possible predictors available for construction of these risk scores is relatively constrained in relation to more standard PRS approaches. Thus, we did not hypothesize that the PTRS would show overall stronger predictive performance than comparable PRS approaches. As expected, the PRS approaches showed performance advantages in prediction of COPD risk in individuals with European ancestry. In examining performance specifically among African Americans, we did note a stronger association with COPD for the PTRS compared to PRS, particularly in heavy smokers with 40 or more pack-years of smoking for moderate-to-severe COPD and in light smokers with pack-years of smoking less than 20 for severe COPD. Although our present work did not include a direct comparison to the more recently published COPD PRS[18] that is a weighted sum of two individual PRSs for FEV_1 and FEV_1/FVC and includes more variants not reaching genome-wide significance by lassosum,[51] we observed the same pattern as the Moll paper[18] that the AUC based on clinical risk factors was higher than the risk-score AUC, but the combined AUC (including both clinical and genetic factors) improved upon each of the separate models. In addition, we observed that the AUC obtained by clinical risk factors alone decreased with increasing smoking history, whereas the AUC achieved by the risk score alone was higher in strata with greater smoking exposures. This result reflects the likely larger effects of the

underlying SNPs in the presence of smoking and warrants further investigation. While the Moll PRS[18] demonstrated improvements in predictive performance over the Shrine et al.[11] risk score for both individuals with European ancestry and individuals with African ancestry, the performance gap between these two ancestry groups was increased (e.g., based on comparison of odds ratios for the respective risk scores observed for COPDGene NHW and AA from the Moll versus Shrine PRS). In our study, the improved association performance of the PTRS in African Americans is concordant with the portability advantages that we also observed for the PTRS. Combined, these results underscore the value of using PTRS approaches to leverage large-scale genomic resources of primarily European ancestry to construct risk scores that can be extended to non-European ancestry populations.

The specific reasons contributing to the particular value of PTRS in improving portability across ancestry groups are not entirely straightforward. While we hypothesized initially that the PTRS may show advantages due to its use of eQTL variants that tie it to biological mechanisms that may be shared across ancestry groups, part of the portability of the PTRS may also stem in part from the methods used to construct the gene expression prediction models. The GTEx prediction models[38] incorporated statistical fine-mapped variants in selection of SNPs for gene expression prediction, which may have helped in enriching the resulting predictors for causal variants. While the MESA prediction models were built using the elastic net model without initial selection based on fine mapping, these MESA prediction models were constructed leveraging the diverse and multi-ethnic MESA participants,[39] such that the variants ultimately included in the predictors were also enriched for eQTL variants exhibiting shared effects across ancestry groups.

Comparing performance of the PTRS across the different gene expression prediction models used to construct these risk scores, we did not observe a clear trend in terms of which model resulted in better predictive performance overall. Direct comparison of PTRS performance across different gene expression prediction models is not straightforward, in part because the properties of the underlying models from GTEx and MESA differ on multiple levels. Besides the differences in statistical approaches used to develop these prediction models noted earlier, other differences between the GTEx and MESA models include (1) source tissues represented, (2) race/ethnic composition of the underlying studies, with GTEx including roughly 15% non-European-ancestry individuals[52] compared to 50% non-European-ancestry individuals in MESA,[39] and (3) sample sizes of the underlying models in GTEx lung ($n = 444$) and whole blood ($n = 573$) versus MESA monocytes ($n = 1{,}163$).[38,39] Further, the PTRS constructed using different gene expression prediction models differed markedly in the specific genes included in the final risk scores. While these differences may reflect distinct biology captured by each of the corresponding sources of tissue, we should keep in mind that the sample sizes used to construct the underlying gene expression prediction models were limited. In addition, the quality and disease relevance of the GTEx lung data for application to chronic lung diseases are limited.[53] As resources used for eQTL mapping expand in sample size and improve in tissue quality in the future, we expect to gain additional resolution in examining the finer difference in performance among the various PTRSs derived from distinct gene expression prediction models.

In summary, we applied the PTRS framework toward genetic risk prediction of COPD and demonstrated its value in providing biologically interpretable disease risk prediction that is portable across ancestry groups and provides information complementary to the standard polygenic risk scores. A particular strength of our approach is that the PTRS can be constructed using summary statistics from large-scale GWASs and/or TWASs alone, allowing us to leverage the abundant genome-wide summary data from prior GWASs to construct these risk scores. Limitations of our study include our use, for construction of the PTRS, of existing gene expression prediction functions, which themselves rely on limited sample sizes and provide limited ability to compare performance of PTRS approaches across underlying tissue models for gene expression prediction. In future work, we will work to build gene expression prediction models in expanded RNA-seq resources from TOPMed and other sources, allowing us to leverage larger sample sizes for gene expression prediction, while also applying more consistent methods for gene expression prediction to allow more direct comparison across tissues. In addition, our proposed PTRS framework extends naturally to other molecular omics types, and we intend to extend our PTRS approach to leverage proteomics or additional omics as they become more widely available.

While we have demonstrated the value of integrative approaches leveraging eQTLs toward improvement of cross-ancestry portability of risk scores, we further emphasize that constructing more portable risk scores represents just one line of investigation toward achieving equity in risk prediction and personalized medicine. Ultimately, a crucial step toward improving performance of genomic risk prediction for non-European-ancestry groups will be to increase the diversity of participants included and analyzed in genetic studies. Our long-term hope is that our field can make sufficient progress on expanding diverse ancestry resources for genomics, for example from TOPMed, the Population Architecture using Genomics and Epidemiology (PAGE) Consortium,[30] and All of Us Research Program,[54] such that we can leverage these diverse ancestry resources more directly toward improved prediction in diverse ancestry populations. As there remains a long road ahead toward recruitment, phenotyping, and analysis of diverse ancestry samples for large-scale diverse ancestry genetic studies, we suggest that construction of more portable risk scores will contribute toward improvements in equity in the near future.

Data and code availability

Individual whole-genome sequence data for TOPMed whole genomes are available through dbGaP. The dbGaP accession numbers are: Atherosclerosis Risk in Communities (ARIC) phs001211, Coronary Artery Risk Development in Young Adults (CARDIA) phs001612, Cardiovascular Health Study (CHS) phs001368, Cleveland Family Study (CFS) phs000954, Framingham Heart Study (FHS) phs000974, Hispanic Community Health Study/Study of Latinos (HCHS/SOL) phs001395, Jackson Heart Study (JHS) phs000964, Multi-Ethnic Study of Atherosclerosis (MESA) phs001416, Genetic Epidemiology of COPD (COPDGene) phs000951, and SubPopulations and InteRmediate Outcome Measures in COPD Study (SPIROMICS) phs001927. Data in dbGaP can be downloaded by controlled access with an approved application submitted through dbGaP website. All PrediXcan code used is available in the GitHub repository.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.03.007.

Acknowledgments

We dedicate this manuscript to our co-authors, long-time collaborators, and dear friends Dr. Debbie Ann Nickerson and Dr. L. Adrienne Cupples. Debbie was a pioneer in human genomics and advocate for diversity and inclusion in science and research. Adrienne was committed to excellence in statistical methodology and collaborative consortium-based genomic research, with major contributions to teaching and service in biostatistics. This research was supported by NIH/NHLBI R01 HL131565 (X.H. and A.M.); R01 HL153248 (A.M. and M.H.C.); R01 HL135142, R01 HL137927, R01 HL089856, and R01 HL147148 (M.H.C.); and K01-HL129039 (D.Q.). Study-specific acknowledgments are given in the supplemental information. We gratefully acknowledge the cohorts and participants who provided biological samples and data for TOPMed. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services. A full list of authors for the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium is provided at https://www.nhlbiwgs.org/topmed-banner-authorship.

Declaration of interests

In the past three years, E.K.S. and M.H.C. have received institutional grant support from GlaxoSmithKline and Bayer. M.H.C. has received consulting and speaking fees from Illumina and AstraZeneca. B.M.P. serves on the Steering Committee of the Yale Open Data Access project funded by Johnson & Johnson. T.L. is an advisor for Variant Bio, Goldfinch Bio, and GSK. T.L. also has stock in Variant Bio. All other authors have declared no competing interests.

Received: November 5, 2021
Accepted: March 4, 2022
Published: April 5, 2022

Web resources

dbGaP, https://www.ncbi.nlm.nih.gov/gap
PredictDB, https://predictdb.org/
PrediXcan GitHub repository, https://github.com/hakyimlab/PrediXcan

References

1. Heron, M. (2018). Deaths: Leading Causes for 2016. Natl. Vital Stat. Rep. *67*, 1–77.
2. Murphy, S.L., Xu, J., Kochanek, K.D., and Arias, E. (2018). Mortality in the United States, 2017. NCHS Data Brief (328), 1–8.
3. Global Health Estimates Life expectancy and leading causes of death and disability. Accessed Jan 18, 2022. URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death
4. Tan, W.C., Sin, D.D., Bourbeau, J., Hernandez, P., Chapman, K.R., Cowie, R., FitzGerald, J.M., Marciniuk, D.D., Maltais, F., Buist, A.S., et al.; CanCOLD Collaborative Research Group (2015). Characteristics of COPD in never-smokers and ever-smokers in the general population: results from the CanCOLD study. Thorax *70*, 822–829.
5. Smith, B.M., Kirby, M., Hoffman, E.A., Kronmal, R.A., Aaron, S.D., Allen, N.B., Bertoni, A., Coxson, H.O., Cooper, C., Couper, D.J., et al.; MESA Lung, CanCOLD, and SPIROMICS Investigators (2020). Association of Dysanapsis With Chronic Obstructive Pulmonary Disease Among Older Adults. JAMA *323*, 2268–2280.
6. Silverman, E.K., Chapman, H.A., Drazen, J.M., Weiss, S.T., Rosner, B., Campbell, E.J., O'Donnell, W.J., Reilly, J.J., Ginns, L., Mentzer, S., et al. (1998). Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. Am. J. Respir. Crit. Care Med. *157*, 1770–1778.
7. Ingebrigtsen, T., Thomsen, S.F., Vestbo, J., van der Sluis, S., Kyvik, K.O., Silverman, E.K., Svartengren, M., and Backer, V. (2010). Genetic influences on Chronic Obstructive Pulmonary Disease - a twin study. Respir. Med. *104*, 1890–1895.
8. Zhou, J.J., Cho, M.H., Castaldi, P.J., Hersh, C.P., Silverman, E.K., and Laird, N.M. (2013). Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. Am. J. Respir. Crit. Care Med. *188*, 941–947.
9. Wain, L.V., Shrine, N., Artigas, M.S., Erzurumluoglu, A.M., Noyvert, B., Bossini-Castillo, L., Obeidat, M., Henry, A.P., Portelli, M.A., Hall, R.J., et al.; Understanding Society Scientific Group; and Geisinger-Regeneron DiscovEHR Collaboration (2017). Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nat. Genet. *49*, 416–425.
10. Wyss, A.B., Sofer, T., Lee, M.K., Terzikhan, N., Nguyen, J.N., Lahousse, L., Latourelle, J.C., Smith, A.V., Bartz, T.M., Feitosa, M.F., et al. (2018). Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. Nat. Commun. *9*, 2976.
11. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M., Jackson, V.E., Hobbs, B.D., Melbourne, C.A., Batini, C., Fawcett, K.A., Song, K., Sakornsakolpat, P., et al.; Understanding Society Scientific Group (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. *51*, 481–493.

12. Sakornsakolpat, P., Prokopenko, D., Lamontagne, M., Reeve, N.F., Guyatt, A.L., Jackson, V.E., Shrine, N., Qiao, D., Bartz, T.M., Kim, D.K., et al.; SpiroMeta Consortium; and International COPD Genetics Consortium (2019). Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. Nat. Genet. *51*, 494–505.
13. Zhao, X., Qiao, D., Yang, C., Kasela, S., Kim, W., Ma, Y., Shrine, N., Batini, C., Sofer, T., Taliun, S.A.G., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Lung Working Group (2020). Whole genome sequence analysis of pulmonary function and COPD in 19,996 multi-ethnic participants. Nat. Commun. *11*, 5182.
14. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
15. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
16. Busch, R., Hobbs, B.D., Zhou, J., Castaldi, P.J., McGeachie, M.J., Hardin, M.E., Hawrylkiewicz, I., Sliwinski, P., Yim, J.-J., Kim, W.J., et al.; National Emphysema Treatment Trial Genetics; Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-Points; International COPD Genetics Network; and COPDGene Investigators (2017). Genetic Association and Risk Scores in a Chronic Obstructive Pulmonary Disease Meta-analysis of 16,707 Subjects. Am. J. Respir. Cell Mol. Biol. *57*, 35–46.
17. Oelsner, E.C., Ortega, V.E., Smith, B.M., Nguyen, J.N., Manichaikul, A.W., Hoffman, E.A., Guo, X., Taylor, K.D., Woodruff, P.G., Couper, D.J., et al. (2019). A Genetic Risk Score Associated with Chronic Obstructive Pulmonary Disease Susceptibility and Lung Structure on Computed Tomography. Am. J. Respir. Crit. Care Med. *200*, 721–731.
18. Moll, M., Sakornsakolpat, P., Shrine, N., Hobbs, B.D., DeMeo, D.L., John, C., Guyatt, A.L., McGeachie, M.J., Gharib, S.A., Obeidat, M., et al.; International COPD Genetics Consortium; and SpiroMeta Consortium (2020). Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. Lancet Respir. Med. *8*, 696–708.
19. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
20. Knowles, J.W., and Ashley, E.A. (2018). Cardiovascular disease: The rise of the genetic risk score. PLoS Med. *15*, e1002546.
21. Sharp, S.A., Rich, S.S., Wood, A.R., Jones, S.E., Beaumont, R.N., Harrison, J.W., Schneider, D.A., Locke, J.M., Tyrrell, J., Weedon, M.N., et al. (2019). Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. Diabetes Care *42*, 200–207.
22. Restuadi, R., Garton, F.C., Benyamin, B., Lin, T., Williams, K.L., Vinkhuyzen, A., van Rheenen, W., Zhu, Z., Laing, N.G., Mather, K.A., et al. (2021). Polygenic risk score analysis for amyotrophic lateral sclerosis leveraging cognitive performance, educational attainment and schizophrenia. Eur. J. Hum. Genet. https://doi.org/10.1038/s41431-021-00885-y.
23. Forrest, I.S., Chaudhary, K., Paranjpe, I., Vy, H.M.T., Marquez-Luna, C., Rocheleau, G., Saha, A., Chan, L., Van Vleck, T., Loos, R.J.F., et al. (2021). Genome-wide polygenic risk score for retinopathy of type 2 diabetes. Hum. Mol. Genet. *30*, 952–960.
24. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
25. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human Genetic Studies. Cell *177*, 26–31.
26. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun. *10*, 3328.
27. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet. *97*, 576–592.
28. Chen, C.-Y., Han, J., Hunter, D.J., Kraft, P., and Price, A.L. (2015). Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. Genet. Epidemiol. *39*, 427–438.
29. Márquez-Luna, C., Loh, P.-R., Price, A.L.; South Asian Type 2 Diabetes (SAT2D) Consortium; and SIGMA Type 2 Diabetes Consortium (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823.
30. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
31. Marigorta, U.M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genet. *9*, e1003566.
32. Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. *6*, 91.
33. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.
34. Shi, H., Burch, K.S., Johnson, R., Freund, M.K., Kichaev, G., Mancuso, N., Manuel, A.M., Dong, N., and Pasaniuc, B. (2020). Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. Am. J. Hum. Genet. *106*, 805–817.
35. Porcu, E., Rüeger, S., Lepik, K., Santoni, F.A., Reymond, A., Kutalik, Z.; eQTLGen Consortium; and BIOS Consortium (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat. Commun. *10*, 3300.
36. Liang, Y., Pividori, M., Manichaikul, A., Palmer, A.A., Cox, N.J., Wheeler, H.E., and Im, H.K. (2022). Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. Genome Biol. *23*, 23.
37. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015).

A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
38. Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., et al.; GTEx GWAS Working Group; and GTEx Consortium (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol. *22*, 49.
39. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene expression traits across diverse populations. PLoS Genet. *14*, e1007586.
40. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. *51*, 187–195.
41. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., et al.; GTEx Consortium (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. *9*, 1825.
42. Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet. *13*, e1006646.
43. Pividori, M., Rajagopal, P.S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., Park, Y., Consortium, G., Wen, X., Im, H.K.; and GTEx Consortium (2020). PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci. Adv. *6*, eaba2083.
44. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
45. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.
46. Hankinson, J.L., Odencrantz, J.R., and Fedan, K.B. (1999). Spirometric reference values from a sample of the general U.S. population. Am. J. Respir. Crit. Care Med. *159*, 179–187.
47. Gogarten, S.M., Sofer, T., Chen, H., Yu, C., Brody, J.A., Thornton, T.A., Rice, K.M., and Conomos, M.P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics *35*, 5346–5348.
48. Balduzzi, S., Rücker, G., and Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. Evid. Based Ment. Health *22*, 153–160.
49. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.
50. Sofer, T., Zheng, X., Gogarten, S.M., Laurie, C.A., Grinde, K., Shaffer, J.R., Shungin, D., O'Connell, J.R., Durazo-Arvizo, R.A., Raffield, L., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. Genet. Epidemiol. *43*, 263–275.
51. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480.
52. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
53. McCall, M.N., Illei, P.B., and Halushka, M.K. (2016). Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. Am. J. Hum. Genet. *99*, 624–635.
54. Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., Dishman, E.; and All of Us Research Program Investigators (2019). The "All of Us" Research Program. N. Engl. J. Med. *381*, 668–676.

The American Journal of Human Genetics, Volume *109*

Supplemental information

Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program

Xiaowei Hu, Dandi Qiao, Wonji Kim, Matthew Moll, Pallavi P. Balte, Leslie A. Lange, Traci M. Bartz, Rajesh Kumar, Xingnan Li, Bing Yu, Brian E. Cade, Cecelia A. Laurie, Tamar Sofer, Ingo Ruczinski, Deborah A. Nickerson, Donna M. Muzny, Ginger A. Metcalf, Harshavardhan Doddapaneni, Stacy Gabriel, Namrata Gupta, Shannon Dugan-Perez, L. Adrienne Cupples, Laura R. Loehr, Deepti Jain, Jerome I. Rotter, James G. Wilson, Bruce M. Psaty, Myriam Fornage, Alanna C. Morrison, Ramachandran S. Vasan, George Washko, Stephen S. Rich, George T. O'Connor, Eugene Bleecker, Robert C. Kaplan, Ravi Kalhan, Susan Redline, Sina A. Gharib, Deborah Meyers, Victor Ortega, Josée Dupuis, Stephanie J. London, Tuuli Lappalainen, Elizabeth C. Oelsner, Edwin K. Silverman, R. Graham Barr, Timothy A. Thornton, Heather E. Wheeler, TOPMed Lung Working Group, Michael H. Cho, Hae Kyung Im, and Ani Manichaikul

Supplemental Figures

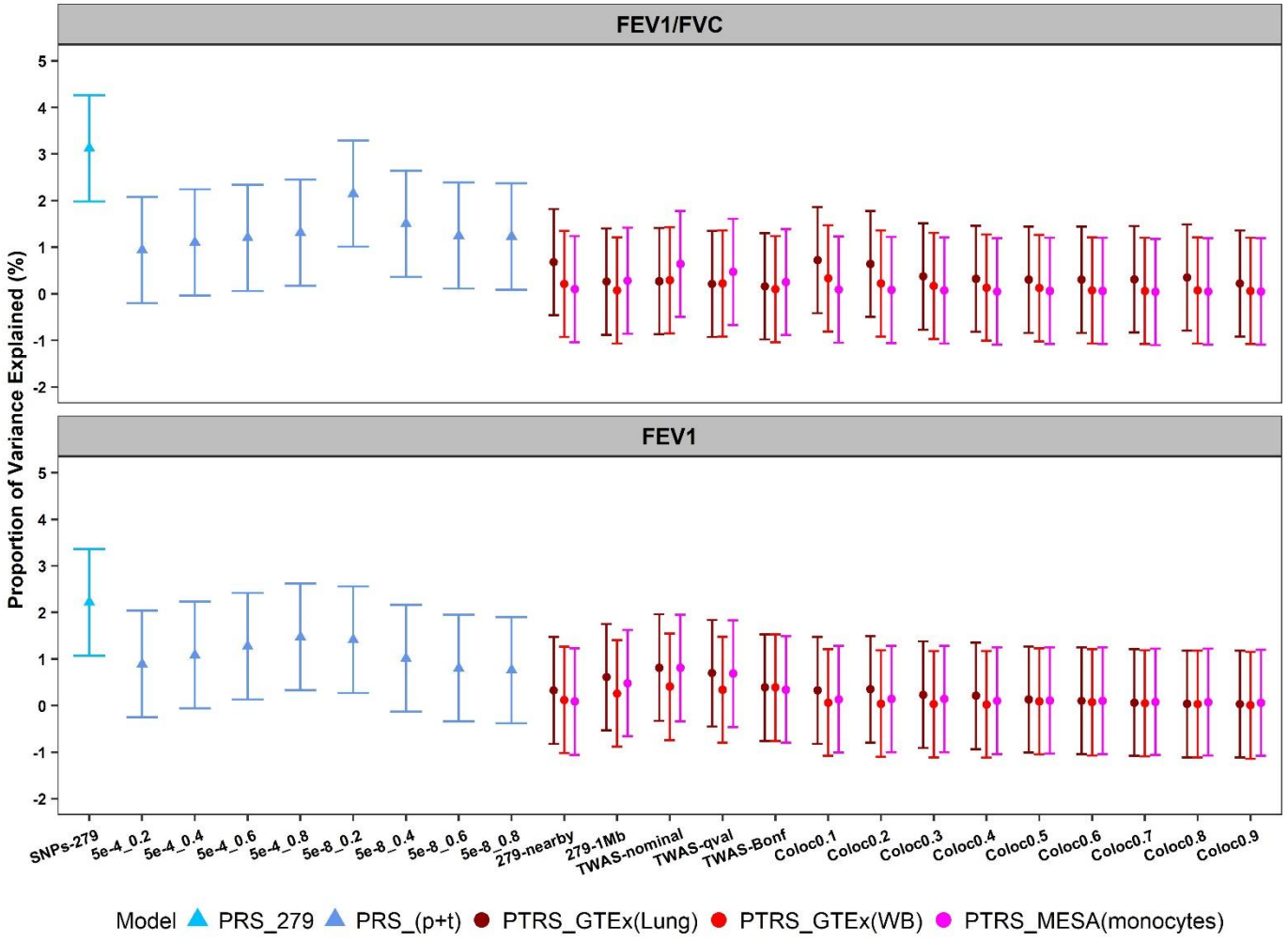


Figure S1: Prediction performance of all risk score candidates in multi-ethnic population/family-based cohorts for FEV1/FVC ratio and FEV1. Proportion of variance explained (%), estimated by 100*squared correlation between the observed phenotypes and the predicted phenotypes by risk score only; Data are shown as proportion of variance explained with 95% CI; PRS_279, PRS derived by previously published 279 variants for FEV1/FVC ratio or FEV1; PRS_(p+t), PRS derived by pruning and thresholding, a range of p-value and pairwise correlation thresholds were used to create eight candidates (5e-4_0.2 to 5e-8_0.8); 279-nearby and 279-1Mb, PTRS derived by genes nearby and within +/- 1Mb region of previously published 279 variants (for FEV1/FVC ratio or FEV1) respectively; TWAS-nominal, TWAS-qval, and TWAS-Bonf, PTRS derived by genes passing TWAS p-value threshold of 0.05, q-value, and Bonferroni respectively; Coloc0.1 to Coloc0.9, PTRS derived by genes with regional colocalization probability ranging from 0.1 to 0.9.

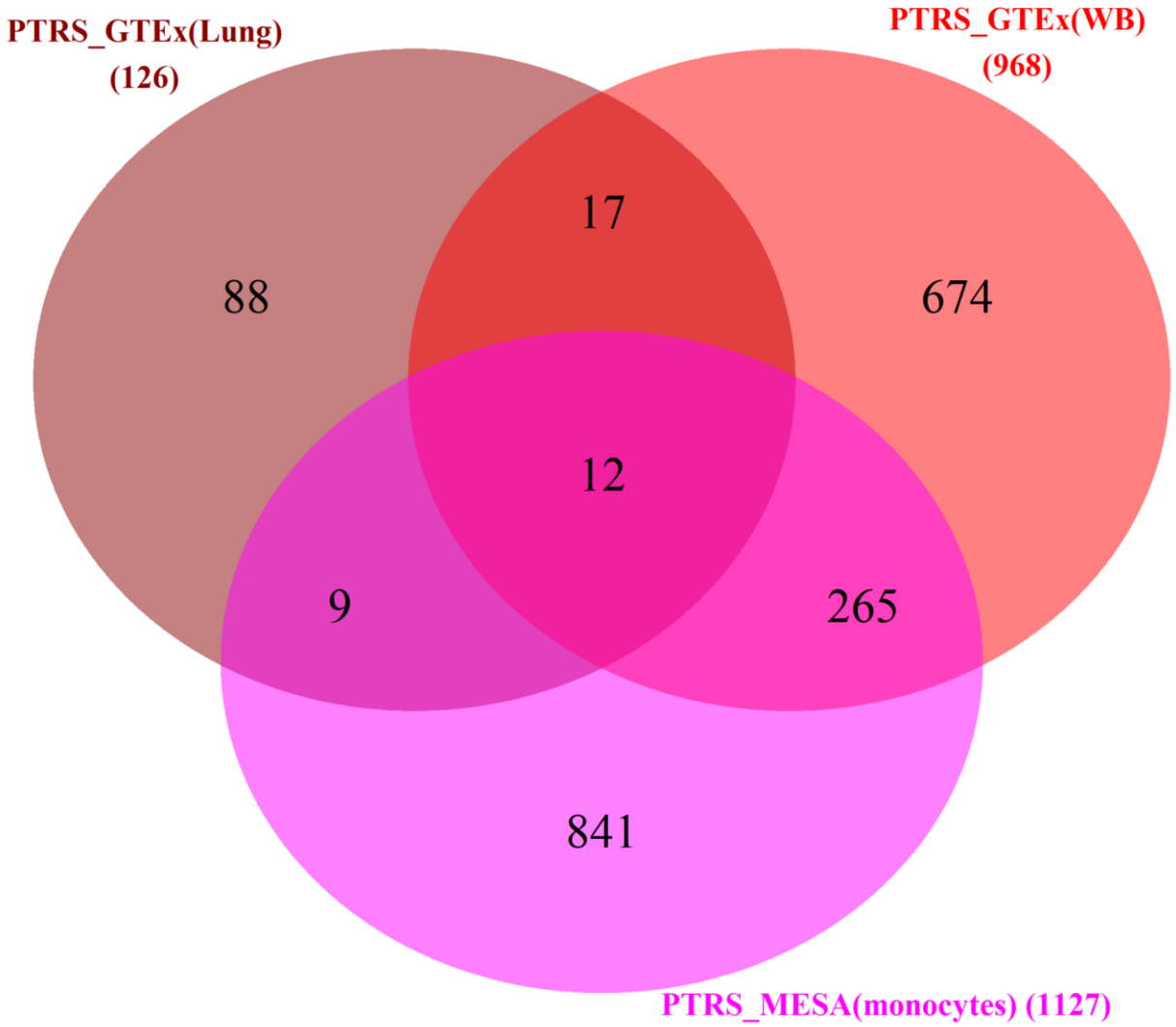


Figure S2: Venn diagram of the number of genes included in the best risk score candidate of each model for two COPD traits. The best candidates of three models are PTRS_GTEx(Lung): 279-nearby that was derived by genes nearby previously published 279 variants for FEV1/FVC ratio; PTRS_GTEx(WB): TWAS-qval that was derived by genes passing TWAS q-value threshold; PTRS_MESA(monocytes): TWAS-nominal that was derived by genes passing TWAS p-value threshold of 0.05. The number in the parenthesis shows the gene size of the best candidate.

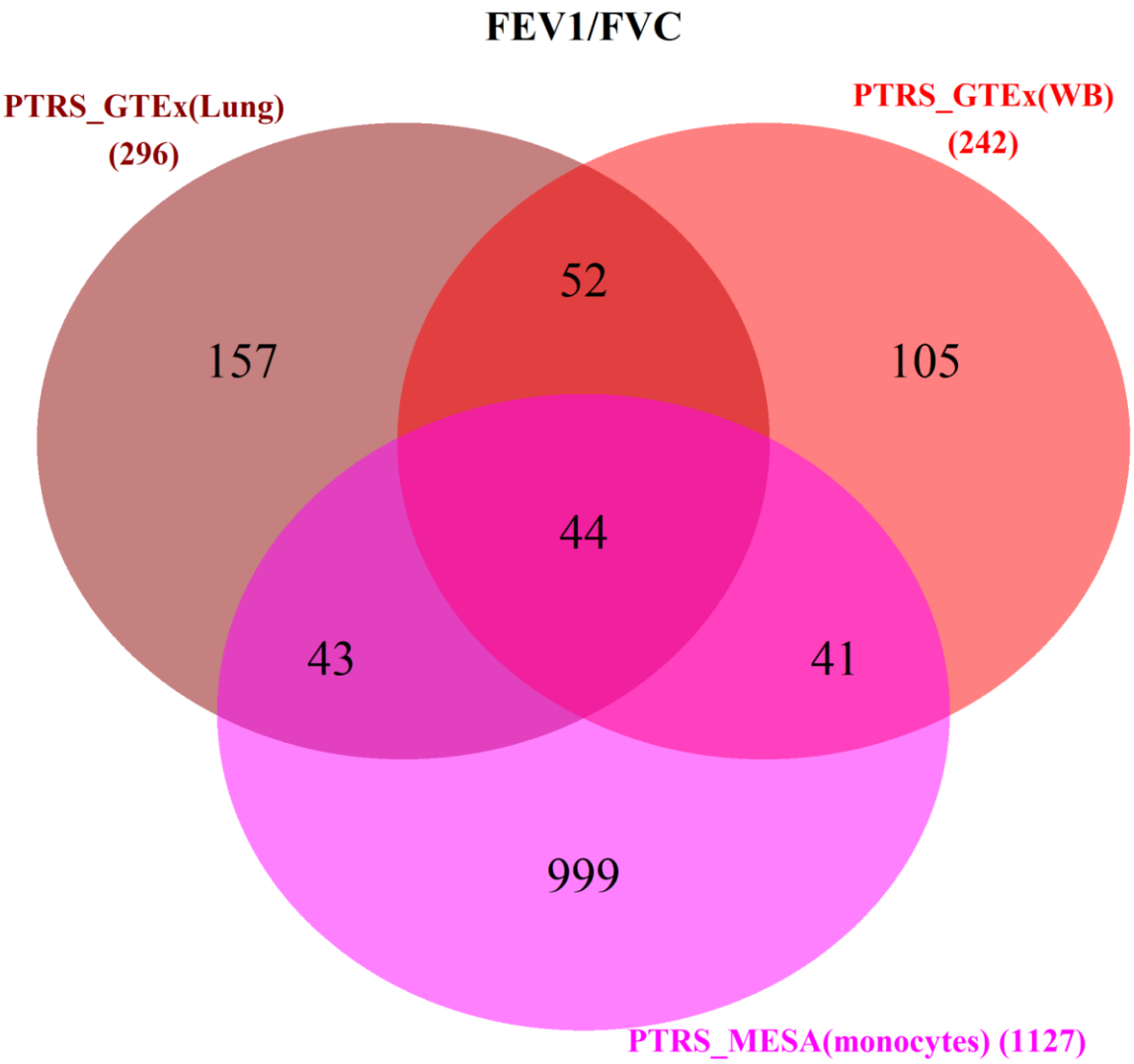


Figure S3: Venn diagram of the number of genes included in the best risk score candidate of each model for FEV1/FVC ratio. The best candidates of three models are PTRS_GTEx(Lung): Coloc0.1; PTRS_GTEx(WB): Coloc0.1; Coloc0.1 was derived by genes with regional colocalization probability greater than 0.1; PTRS_MESA(monocytes): TWAS-nominal that was derived by genes passing TWAS p-value threshold of 0.05. The number in the parenthesis shows the gene size of the best candidate.

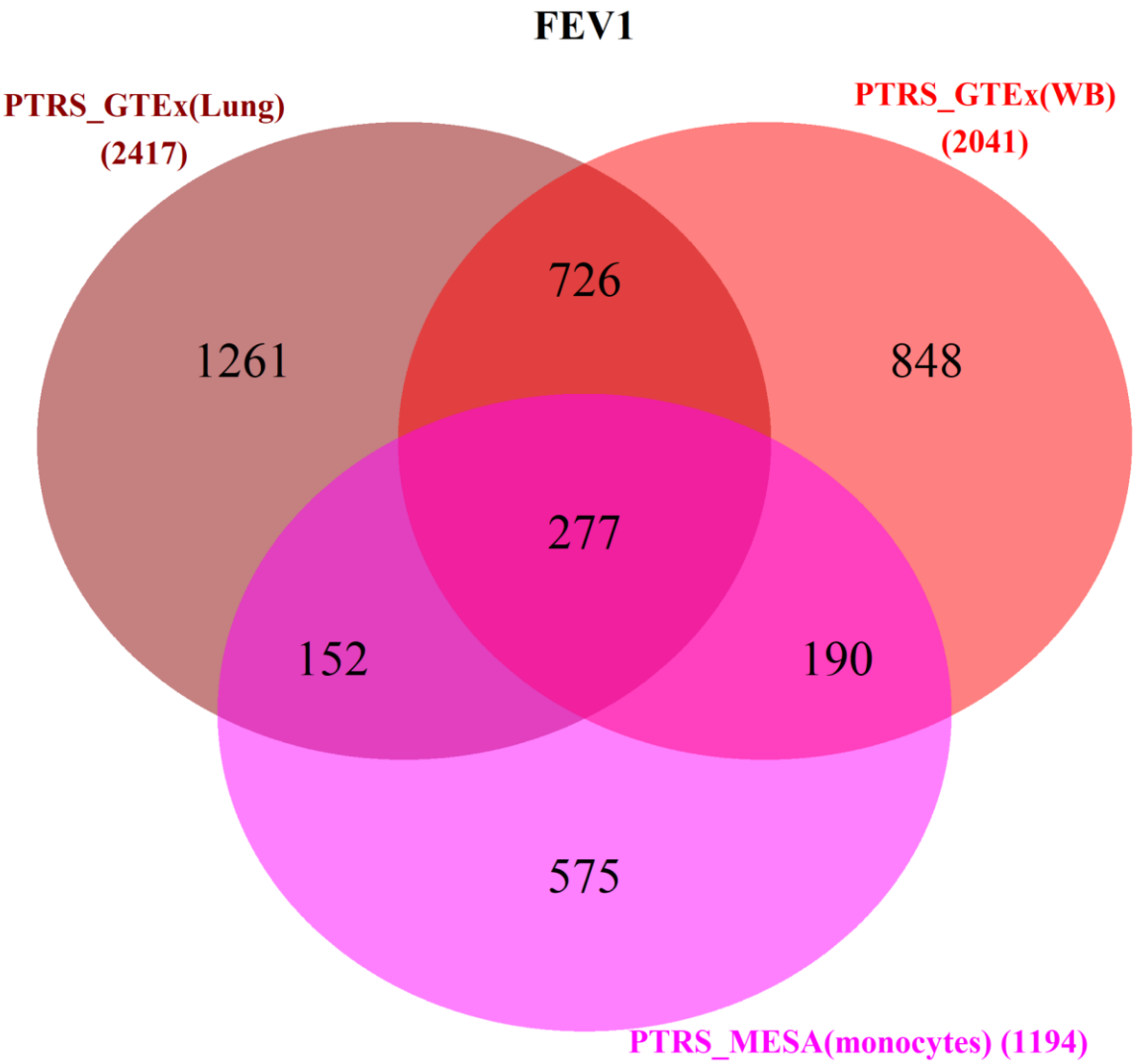


Figure S4: Venn diagram of the number of genes included in the best risk score candidate of each model for FEV1. The best candidates of three models are PTRS_GTEx(Lung): TWAS-nominal; PTRS_GTEx(WB): TWAS-nominal; PTRS_MESA(monocytes): TWAS-nominal. TWAS-nominal was derived by genes passing TWAS p-value threshold of 0.05. The number in the parenthesis shows the gene size of the best candidate.

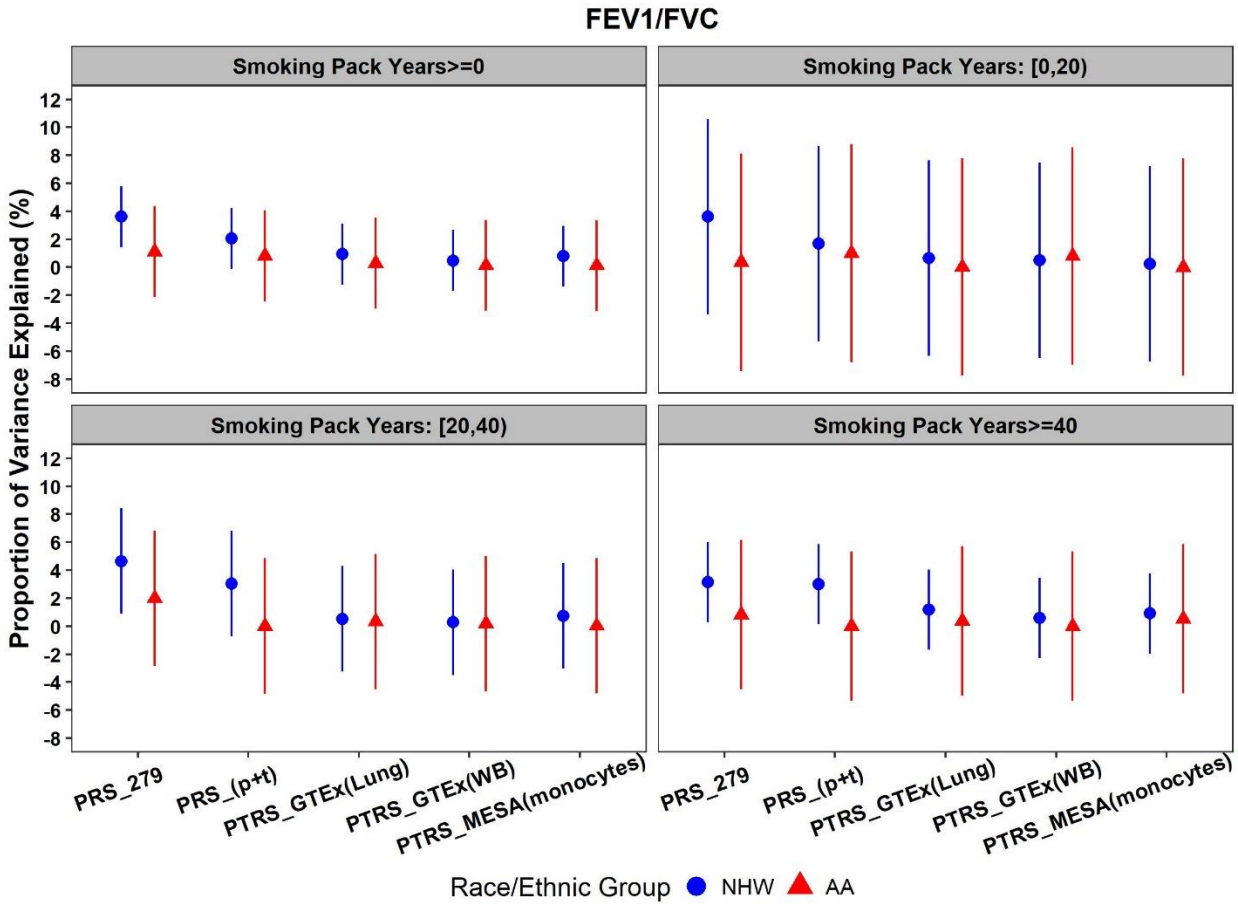


Figure S5: Prediction performance of the best risk scores with FEV1/FVC ratio in COPD-enriched studies. NHW, Non-Hispanic Whites; AA, African Americans; the risk scores used in the analyses were based on the prediction performance of each smoking stratum on population/family-based cohorts for FEV1/FVC ratio; proportion of variance explained (%) was estimated by 100*squared correlation between the observed phenotypes and the predicted phenotypes by risk score only. Data are shown as meta-analyzed results with error bars as 95% CIs of proportion of variance explained.

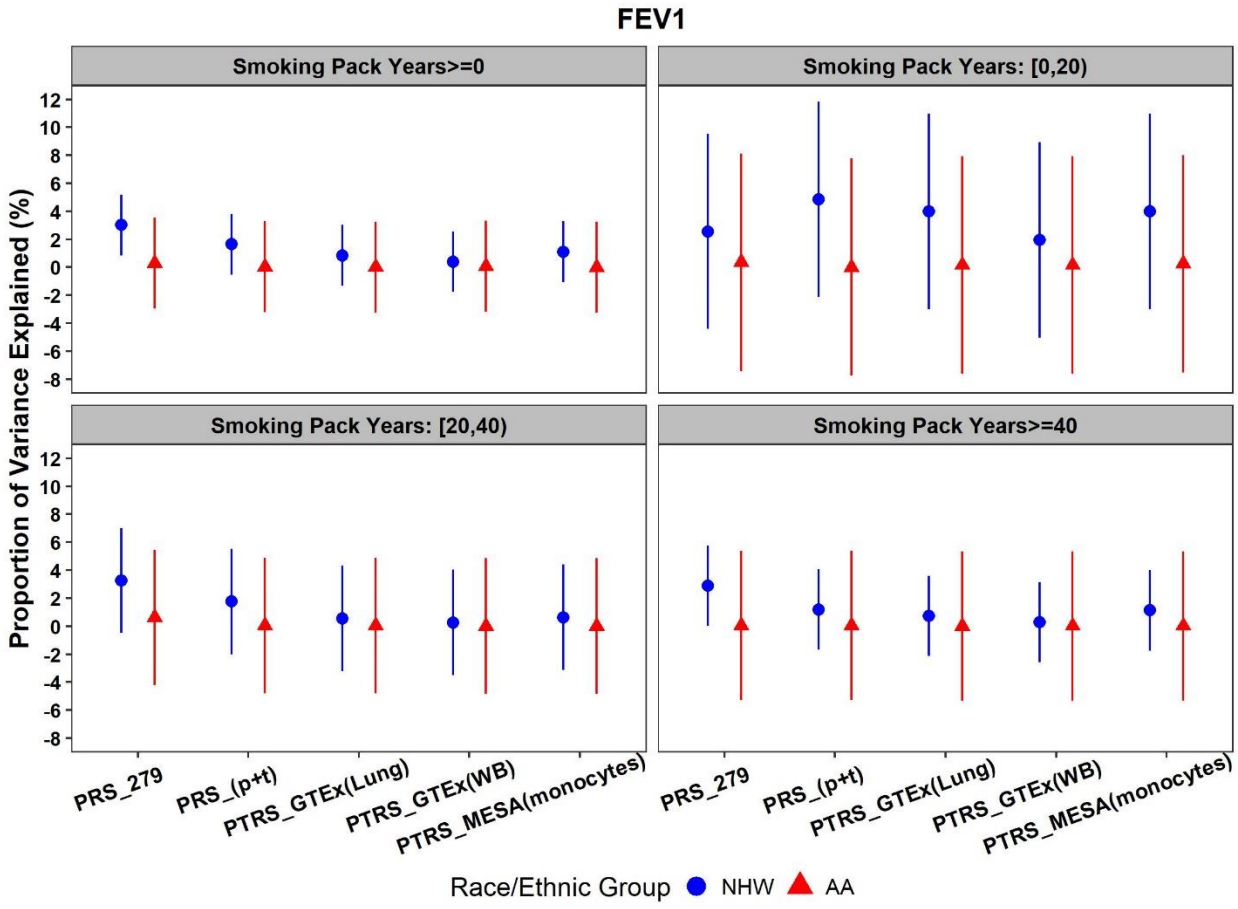


Figure S6: Prediction performance of the best risk scores with FEV1 in COPD-enriched studies. NHW, Non-Hispanic Whites; AA, African Americans; the risk scores used in the analyses were based on the prediction performance of each smoking stratum on population/family-based cohorts for FEV1; proportion of variance explained (%) was estimated by 100*squared correlation between the observed phenotypes and the predicted phenotypes by risk score only. Data are shown as meta-analyzed results with error bars as 95% CIs of proportion of variance explained.

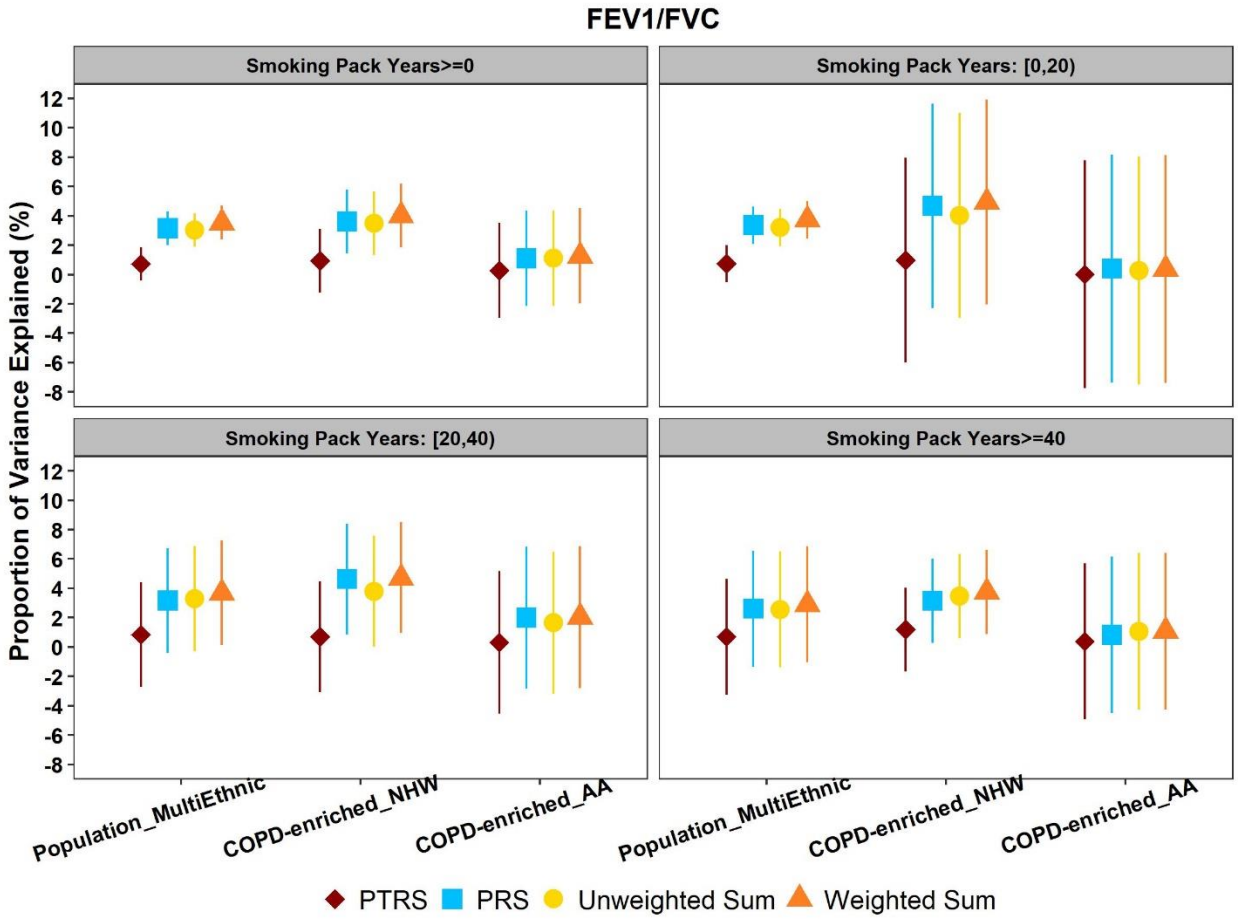


Figure S7: Prediction performance of the combined risk scores for FEV1/FVC ratio. The risk score candidates, PTRS_GTEx(Lung): 279-nearby and PRS_279: SNPs-279, were used for PTRS and PRS respectively in the analyses; Unweighted Sum and Weighted Sum refer to the direct summation and the weighted summation of PTRS and PRS respectively; Data are shown as proportion of variance explained (%) which was estimated by 100*squared correlation between the observed phenotypes and the predicted phenotypes by combined risk score only, error bars are 95% CIs; For COPD-enriched studies, the results were meta-analyzed. NHW, Non-Hispanic Whites; AA, African Americans; Population_MultiEthnic, multi-ethnic samples in population/family-based cohorts.

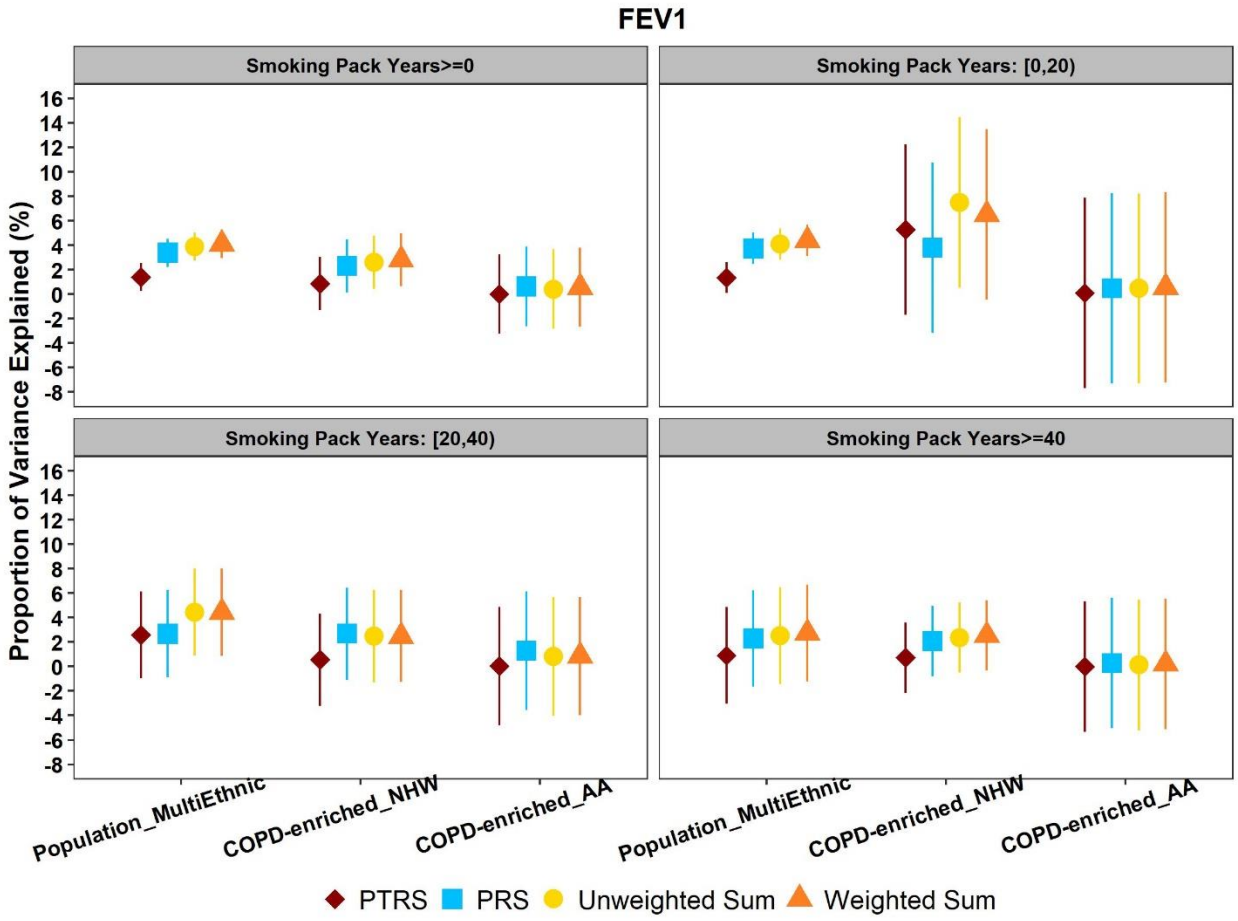


Figure S8: Prediction performance of the combined risk scores for FEV1. The risk score candidates, PTRS_GTEx(Lung): 279-nearby and PRS_279: SNPs-279, were used for PTRS and PRS respectively in the analyses; Unweighted Sum and Weighted Sum refer to the direct summation and the weighted summation of PTRS and PRS respectively; Data are shown as proportion of variance explained (%) which was estimated by 100*squared correlation between the observed phenotypes and the predicted phenotypes by combined risk score only, error bars are 95% CIs; For COPD-enriched studies, the results were meta-analyzed. NHW, Non-Hispanic Whites; AA, African Americans; Population_MultiEthnic, multi-ethnic samples in population/family-based cohorts.

Supplemental Tables

In this section we will include Table S1, Tables S14-15, and Tables S20-22. The Excel file will contain Tables S2-13 and Tables S16-19.

	FEV1/FVC			**FEV1**		
Candidate	PTRS_GTEx (Lung)	PTRS_GTEx (WB)	PTRS_MESA (monocytes)	PTRS_GTEx (Lung)	PTRS_GTEx (WB)	PTRS_MESA (monocytes)
279-nearby	126	102	75	126	102	75
279-1Mb	2797	2447	1400	2797	2447	1400
TWAS-nominal	2407	2065	1127	2417	2041	1194
TWAS-qval	1175	968	526	1164	928	513
TWAS-Bonf	271	219	143	206	200	142
Coloc0.1	296	242	289	301	260	352
Coloc0.2	185	147	207	182	161	230
Coloc0.3	126	107	147	128	102	177
Coloc0.4	96	72	119	97	85	149
Coloc0.5	72	50	98	56	50	131
Coloc0.6	55	32	80	43	38	111
Coloc0.7	40	26	61	22	20	82
Coloc0.8	35	16	42	8	14	67
Coloc0.9	19	8	25	2	2	46

Table S1: The number of genes included in each risk score candidate corresponding to three PTRS models for both FEV1/FVC ratio and FEV1. 279-nearyby and 279-1Mb, PTRS derived by genes nearby and within +/- 1Mb region of previously published 279 variants (for FEV_1/FVC ratio or FEV1) respectively; TWAS-nominal, TWAS-qval, and TWAS-Bonf, PTRS derived by genes passing TWAS p-value threshold of 0.05, q-value, and Bonferroni respectively; Coloc0.1 to Coloc0.9, PTRS derived by genes with regional colocalization probability ranging from 0.1 to 0.9.

Moderate-to-Severe COPD					
Model	Candidate	Score beta	Interaction beta	Interaction se	Interaction P-value
PRS_279	SNPs-279	0.3143	0.0012	0.0009	0.2121
PRS_(p+t)	5e-8_0.2	0.2455	0.0015	0.0009	0.0767
PTRS_GTEx (Lung)	279-nearby	0.1501	0.0001	0.0009	0.8872
PTRS_GTEx (WB)	TWAS-qval	0.0225	0.0019	0.0009	0.0355
PTRS_MESA (monocytes)	TWAS-nominal	0.1309	0.0014	0.0009	0.1119
Severe COPD					
PRS_279	SNPs-279	0.2002	0.0052	0.0015	0.0007
PRS_(p+t)	5e-8_0.2	0.2458	0.004	0.0014	0.0042
PTRS_GTEx (Lung)	279-nearby	0.1151	0.0025	0.0016	0.1143
PTRS_GTEx (WB)	TWAS-qval	0.0066	-0.004	0.0015	0.0074
PTRS_MESA (monocytes)	TWAS-nominal	0.1441	0.0026	0.0015	0.0789
FEV1/FVC					
PRS_279	SNPs-279	-1.2297	-0.0075	0.0021	0.0004
PRS_(p+t)	5e-8_0.2	-0.9944	-0.0078	0.0020	0.0001
PTRS_GTEx (Lung)	Coloc0.1	-0.5938	-0.0032	0.0021	0.1384

PTRS_GTEx (WB)	Coloc0.1	-0.4156	-0.0027	0.0022	0.2189
PTRS_MESA (monocytes)	TWAS-nominal	-0.5215	-0.0069	0.0022	0.0017
FEV1					
PRS_279	SNPs-279	-0.0703	-0.0003	0.0001	0.0288
PRS_(p+t)	5e-4_0.8	-0.0571	-0.0004	0.0001	0.0024
PTRS_GTEx (Lung)	TWAS-nominal	-0.0491	-0.0001	0.0001	0.4462
PTRS_GTEx (WB)	TWAS-nominal	-0.0272	-0.0002	0.0001	0.1013
PTRS_MESA (monocytes)	TWAS-nominal	-0.0397	-0.0002	0.0001	0.1249

Table S14: Risk score by smoking pack-years interaction analysis for the best performing risk score candidates in population/family-based cohorts. PRS_279, PRS derived by previously published 279 variants for FEV1/FVC ratio or FEV1; PRS_(p+t), PRS derived by pruning and thresholding, a range of p-value and pairwise correlation thresholds were used to create candidates (5e-4_0.8 and 5e-8_0.2) ; 279-nearyby, PTRS derived by genes nearby previously published 279 variants for FEV1/FVC ratio or FEV1; TWAS-nominal and TWAS-qval, PTRS derived by genes passing TWAS p-value threshold of 0.05 and q-value respectively; Coloc0.1, PTRS derived by genes with regional colocalization probability greater than 0.1.

		Moderate-to-Severe COPD			Severe COPD			FEV1/FVC and FEV1
Study	Smoking Pack-Years	N	Control	Case	N	Control	Case	N
COPDGene_NHW	>=0	5183	2122	3061	3729	2122	1607	6609
	[0,20)	559	400	159	467	400	67	715
	[20,40)	1721	939	782	1289	939	350	2245
	>=40	2903	783	2120	1973	783	1190	3649
COPDGene_AA	>=0	2504	1584	920	1999	1584	415	3258
	[0,20)	434	307	127	362	307	55	612
	[20,40)	1137	769	368	924	769	155	1450
	>=40	933	508	425	713	508	205	1196
SPIROMICS_NHW	>=0	1284	346	938	803	346	457	1535
	[0,20)	66	63	3	63	63	0	74
	[20,40)	368	143	225	253	143	110	457
	>=40	850	140	710	487	140	347	1004
SPIROMICS_AA	>=0	316	139	177	229	139	90	369
	[0,20)	24	23	1	23	23	0	25
	[20,40)	162	74	88	118	74	44	188
	>=40	130	42	88	88	42	46	156

Table S15: Sample size by smoker group for four traits in COPD-enriched studies. NHW, Non-Hispanic Whites; AA, African Americans.

Trait	Effect	Beta	se	P-value
Moderate-to-Severe COPD	PRS	0.3181	0.0238	8.50E-41
	PTRS	0.0571	0.0238	0.0163
	Interaction	0.0075	0.0212	0.7249
Severe COPD	PRS	0.3248	0.0502	9.90E-11
	PTRS	0.085	0.0511	0.0961
	Interaction	0.0049	0.0446	0.9124
FEV1/FVC	PRS	-1.2434	0.0426	9.43E-185
	PTRS	-0.4556	0.0425	9.71E-27
	Interaction	0.0187	0.0415	0.6523
FEV1	PRS	-0.0666	0.0028	3.68E-127
	PTRS	-0.0369	0.003	9.17E-36
	Interaction	0.0013	0.0026	0.6190

Table S20: Risk score interaction analysis in population/family-based cohorts. PRS, PRS_279: SNPs-279 that was derived by previously published 279 variants for FEV1/FVC ratio or FEV1; PTRS, PTRS_GTEx(Lung): 279-nearby that was derived by genes nearby previously published 279 variants for FEV1/FVC ratio or FEV1.

	Smoking Pack-Years≥0			Smoking Pack-Years: [0,20)		
	AUC	L95_AUC	U95_AUC	AUC	L95_AUC	U95_AUC
Clinical factors	0.807	0.798	0.815	0.742	0.729	0.754
PRS	0.579	0.567	0.59	0.589	0.574	0.605
PTRS	0.549	0.537	0.56	0.548	0.532	0.564
Unweighted Sum	0.58	0.568	0.591	0.586	0.571	0.602
Weighted Sum	0.582	0.571	0.593	0.592	0.576	0.608
Clinical+PRS	0.813	0.805	0.821	0.753	0.741	0.766
Clinical+PTRS	0.808	0.8	0.816	0.744	0.732	0.756
Clinical+Unweighted Sum	0.812	0.804	0.82	0.751	0.739	0.764
Clinical+Weighted Sum	0.813	0.805	0.822	0.754	0.741	0.766
	Smoking Pack-Years: [20,40)			Smoking Pack-Years≥40		
	AUC	L95_AUC	U95_AUC	AUC	L95_AUC	U95_AUC
Clinical factors	0.647	0.623	0.672	0.63	0.604	0.657
PRS	0.585	0.558	0.611	0.598	0.571	0.625
PTRS	0.547	0.52	0.574	0.542	0.515	0.569
Unweighted Sum	0.581	0.554	0.608	0.585	0.559	0.612
Weighted Sum	0.587	0.56	0.613	0.599	0.572	0.625
Clinical+PRS	0.667	0.642	0.691	0.659	0.634	0.684
Clinical+PTRS	0.651	0.626	0.676	0.635	0.61	0.661
Clinical+Unweighted Sum	0.662	0.637	0.686	0.651	0.625	0.676
Clinical+Weighted Sum	0.667	0.643	0.691	0.659	0.634	0.684

Table S21: Prediction accuracy for Moderate-to-Severe COPD in population/family-based cohorts from models using clinical risk factors alone (including age, sex, race, and smoking pack-years), risk score alone, and the combination of clinical risk factors and risk score. PRS, PRS_279: SNPs-279 that was derived by previously published 279 variants for FEV1/FVC ratio; PTRS, PTRS_GTEx(Lung): 279-nearby that was derived by genes nearby previously published 279 variants for FEV1/FVC ratio; Unweighted Sum and Weighted Sum refer to the direct summation and the weighted summation of PTRS and PRS respectively; L95_AUC and U95_AUC are lower and upper 95% CIs of AUC respectively.

	Smoking Pack-Years≥0			Smoking Pack-Years: [0,20)		
	AUC	L95_AUC	U95_AUC	AUC	L95_AUC	U95_AUC
Clinical factors	0.854	0.837	0.87	0.762	0.731	0.792
PRS	0.582	0.558	0.607	0.545	0.502	0.587
PTRS	0.554	0.53	0.579	0.527	0.486	0.568
Unweighted Sum	0.587	0.562	0.612	0.545	0.503	0.587
Weighted Sum	0.588	0.563	0.613	0.547	0.505	0.589
Clinical+PRS	0.853	0.836	0.87	0.761	0.731	0.792
Clinical+PTRS	0.853	0.836	0.87	0.761	0.731	0.792
Clinical+Unweighted Sum	0.853	0.836	0.87	0.761	0.731	0.791
Clinical+Weighted Sum	0.853	0.836	0.87	0.761	0.73	0.791
	Smoking Pack-Years: [20,40)			**Smoking Pack-Years≥40**		
	AUC	L95_AUC	U95_AUC	AUC	L95_AUC	U95_AUC
Clinical factors	0.703	0.657	0.748	0.693	0.658	0.727
PRS	0.659	0.607	0.711	0.617	0.577	0.657
PTRS	0.574	0.515	0.633	0.556	0.517	0.596
Unweighted Sum	0.639	0.583	0.695	0.612	0.573	0.651
Weighted Sum	0.66	0.608	0.712	0.621	0.581	0.66
Clinical+PRS	0.743	0.698	0.789	0.723	0.688	0.758
Clinical+PTRS	0.71	0.663	0.757	0.699	0.664	0.734
Clinical+Unweighted Sum	0.733	0.685	0.781	0.717	0.682	0.752
Clinical+Weighted Sum	0.743	0.698	0.789	0.723	0.688	0.758

Table S22: Prediction accuracy for Severe COPD in population/family-based cohorts from models using clinical risk factors alone (including age, sex, race, and smoking pack-years), risk score alone, and the combination of clinical risk factors and risk score. PRS, PRS_279: SNPs-279 that was derived by previously published 279 variants for FEV1/FVC ratio; PTRS, PTRS_GTEx(Lung): 279-nearby that was derived by genes nearby previously published 279 variants for FEV1/FVC ratio; Unweighted Sum and Weighted Sum refer to the direct summation and the weighted summation of PTRS and PRS respectively; L95_AUC and U95_AUC are lower and upper 95% CIs of AUC respectively.

Supplemental Methods

Phenotype harmonization

Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). Phenotype harmonization for pulmonary traits was contributed by the NHLBI Pooled Cohorts Study with funding from NIH/NHLBI R21 HL121457, R21 HL129924, K23 HL130627, R01 HL077612.

The NHLBI Pooled Cohorts Study (PCS)[1] harmonized and pooled data from nine large US epidemiologic cohorts that conducted lung function assessments over the last four decades. Additionally, data on self-administered questionnaires, with detailed questions regarding tobacco consumption, past medical history, medications, respiratory symptoms, lipids, renal biomarkers, etc., were also harmonized. Data on CLRD events was harmonized using either adjudicated CLRD hospitalizations or ICD data for all hospitalizations occurring over follow-up.

Among the cohorts included in the NHLBI PCS, the follow cohorts were also included in our TOPMed WGS analysis, for which we utilized harmonized data sets from NHLBI PCS: Atherosclerosis Risk in Communities (ARIC) study; Cardiovascular Health Study (CHS); Coronary Artery Risk Development in Young Adults (CARDIA); Framingham Heart Study (FHS); Hispanic Community Health Study/Study of Latinos (HCHS/SOL); Jackson Heart Study (JHS); and Multi-Ethnic Study of Atherosclerosis (MESA). For the purpose of phenotype harmonization in TOPMed, harmonized data sets for each of the participating cohorts were provided by the NHLBI Pooled Cohorts Study for analyses in the current WGS analysis. We note that a subset of the ARIC participants was later recruited into JHS. For TOPMed purposes, participants in the ARIC-JHS overlap group were not included as part of Exam 4 in ARIC. For JHS, ARIC participants were not excluded. For TOPMed studies not included in the NHLBI Pooled Cohorts Study (Cleveland Family Study (CFS); Genetic Epidemiology of COPD (COPDGene) and Sub-Populations and Intermediate Outcome Measures in COPD Study (SPIROMICS)), phenotype harmonization was conducted separately by each of the participating cohorts, following as closely as possible with the NHLBI Pooled Cohorts Study variable definitions and procedures.

Study descriptions: population/family-based cohorts

The Atherosclerosis Risk in Communities Study (ARIC)

The ARIC study is a population-based prospective cohort study of cardiovascular disease sponsored by the National Heart, Lung, and Blood Institute (NHLBI). ARIC included 15,792 individuals, predominantly European American and African American, aged 45-64 years at baseline (1987-89), chosen by probability sampling from four US

communities. Cohort members completed three additional triennial follow-up examinations, a fifth exam in 2011-2013, a sixth exam in 2016-2017, and a seventh exam in 2018-2019. The ARIC study has been described in detail previously (The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: Design and objectives. American Journal of Epidemiology 1989; 129:687-702).

At each visit, spirometry testing protocols were standardized across the four ARIC field centers, calibration checks were performed daily, and the standardization of data collection and management was coordinated across field centers by a single pulmonary function reading center. Each participant's best FEV_1 and FVC of three acceptable maneuvers, based on the centralized expert review, was used for analysis[3]. For the current cross-sectional WGS analysis, data from the most recent spirometry exam were utilized for participants having multiple longitudinal measures.

The Coronary Artery Risk Development in Young Adults (CARDIA)

During 1985 -1986, CARDIA recruited 5,115 black and white men and women, aged 18 to 30 years, from the general population at Birmingham, Alabama; Chicago, Illinois; and Minneapolis, Minnesota; and from the membership of the Oakland Kaiser-Permanente Health Plan in Oakland, California. The participants were selected so that there would be approximately the same number of people in subgroups of race, gender, education (high school or less and more than high school) and age (18-24 and 25-30) in each of 4 centers. Detailed methods, instruments, and quality control procedures are described at the CARDIA website (http://www.cardia.dopm.uab.edu/ex_mt.htm) and in other published reports[4,5]. Spirometric pulmonary function testing was performed using the Collins survey 8-liter water-sealed spirometer and the Eagle II microprocessor (Warren E. Collins, Inc., Braintree, MA) in a sitting position with noseclips, as per the 1979 American Thoracic Society criteria[6]. Specifically, each subject performed a minimum of three trials with expirations recorded to the FVC plateau, which occurs after six seconds of expiration in adult males and was maintained for at least one second before terminating the forced expiratory maneuver. If, at the end of the three trials, there were at least three acceptable tracings, and with the maximum FVC and FEV1 reproduced to within 5% or 100 mL, whichever is greater, no more trials were performed. For the current cross-sectional WGS analysis, data from the most recent spirometry exam were utilized for participants having multiple longitudinal measures.

The Cardiovascular Health Study (CHS)

CHS is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers7. The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional

predominantly African-American cohort of 687 persons was enrolled in 1992-1993 for a total sample of 5,888. Blood samples were drawn from all participants at their baseline examination and DNA was subsequently extracted from available samples. Pulmonary function testing was conducted at the 1989-1990 visit and follow-up visits four and seven years later. The spirometry procedures for pulmonary function testing have been previously described[8,9]. Briefly, spirometry technicians were centrally trained and certified prior to recruitment of participants. A standard spirometry system, including a Collins Survey I water-seal spirometer (Collins Medical, Inc., Braintree, Massachusetts) and software from S&M Instruments (Doylestown, Pennsylvania), was used by technicians at all four recruitment centers. Stringent quality assurance procedures for spirometry testing exceeded ATS recommendations[8]. For the current cross-sectional WGS analysis, data from the most recent spirometry exam were utilized for participants having multiple longitudinal measures.

European ancestry and African American ancestry CHS participants that had been selected for inclusion in the second phase of the TOPMed sequencing program were included in our discovery analyses. CHS was approved by institutional review committees at each field center and individuals in the present analysis had available DNA and gave informed consent including consent to use of genetic information for the study of cardiovascular disease.

The Cleveland Family Study (CFS)

CFS is a family-based longitudinal study that includes participants with laboratory diagnosed sleep apnea, their family members and neighborhood control families followed between 1990 and 2006. Four examinations over 16 years provided measurements of sleep apnea with overnight polysomnography, anthropometry, and other related phenotypes, as detailed previously[10]. At each exam, forced vital capacity (FVC) and forced expiratory flow (FEV1) was obtained using a calibrated spirometer (Multi-Spiro). While seated, participants were encouraged to perform between 5-8 maneuvers to obtain 3 curves that met ATS standards for acceptability and reproducibility. For the current cross-sectional WGS analysis, data from the most recent spirometry exam were utilized for participants having multiple longitudinal measures.

The Framingham Heart Study (FHS)

The Original Cohort of the Framingham Study was established between 1948 and 1952 as a random sample of 5,209 adult residents of the town of Framingham, Massachusetts. Between 1971 and 1975, the Framingham Study was expanded to include a second generation, the Offspring Cohort, comprising 5,124 adults who were the offspring, or spouses of the offspring, of Original Cohort participants[11]. The Offspring

Cohort has returned for examinations approximately every 4 years since enrollment, and spirometry data are available for the 3rd, 5th, 6th, 7th, ,8th, and 9th examinations.

Spirometry for the Offspring Cohort 3rd examination (1983-87) was performed with a Collins Survey II spirometer interfaced with an Eagle II microprocessor (Warren E. Collins, Inc., Braintree, MA). Spirometry for the 5th (1991-95), 6th (1995-98), and 7th (1998-2001) examinations were performed with a Collins Survey II spirometer interfaced with a personal computer equipped with software developed by S & M Instruments (Doylestown, PA) and adapted for use in epidemiologic studies. Spirometry for the 8th (2005-08) and 9th (2011-14) examinations was performed with the Collins Comprehensive Pulmonary Laboratory (CPL) system with Collins 2000 Plus/SQL Software (Nspire Health, Inc., Longmont, CO). Spirometry was performed in accordance with contemporaneous guidelines of the American Thoracic Society. For the current cross-sectional WGS analysis, data from the first available spirometry exam were utilized for participants having multiple longitudinal measures.

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

HCHS/SOL is a community-based cohort study of 16,415 self-identified Hispanic/Latino persons aged 18 to 74 years at baseline recruited from four U.S. communities (Bronx NY, Chicago IL, San Diego CA, Miami FL). The baseline clinic visit took place in 2008-2011 and included spirometry data collection. The study design, cohort recruitment[12,13] and baseline clinical examination[14] have been previously described. Institutional Review Boards at each field center approved study protocols, and written informed consent was obtained from all participants. For the current cross-sectional WGS analysis, only baseline spirometry data were utilized.

The Jackson Heart Study (JHS)

JHS is a large, population-based observational study evaluating the etiology of cardiovascular diseases and related disorders among African Americans residing in the three counties (Hinds, Madison, and Rankin) that make up the Jackson, Mississippi metropolitan area[15,16]. Data and biologic materials have been collected from 5,301 participants, including a nested family cohort of 1,498 members of 264 families. The age at enrollment for the unrelated cohort was 35-84 years; the family cohort included related individuals >21 years old. During a baseline examination (2000-2004) and two follow-up examinations (2005-2008 and 2009-2012), participants provided extensive medical and social history, had an array of physical and biochemical measurements and diagnostic procedures, and provided blood for genomic DNA[17]. The study population is characterized by a high prevalence of diabetes, hypertension, obesity, and related disorders. Annual follow-up interviews and cohort surveillance are ongoing. For the current cross-sectional WGS analysis, only baseline spirometry data were utilized.

The Multi-Ethnic Study of Atherosclerosis (MESA)

MESA is a longitudinal study of subclinical cardiovascular disease and risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease[18]. Between 2000 and 2002, MESA recruited 6,814 men and women 45 to 84 years of age from Forsyth County, North Carolina; New York City; Baltimore; St. Paul, Minnesota; Chicago; and Los Angeles. Exclusion criteria were clinical cardiovascular disease, weight exceeding 136 kg (300 lb.), pregnancy, and impediment to long-term participation. The MESA Lung Study performed spirometry following the 2005 ATS/ERS guidelines in a subset of the MESA Study, as previously described[19]. All participants provided informed consent and the protocols of MESA were approved by the IRBs of collaborating institutions and the National Heart, Lung and Blood Institute. For the current cross-sectional WGS analysis, data from the earliest spirometry exam were utilized for participants having multiple longitudinal measures.

Study descriptions: COPD-enriched studies

Genetic Epidemiology of COPD (COPDGene)

COPDGene[20] is a multi-center observational cohort for epidemiologic and genetic study of over 10,000 subjects (2/3 non-Hispanic White and 1/3 African Americans) with at least 10 pack-years of cigarette smoking with and without COPD. All subjects underwent extensive phenotyping, including lung function, chest CT phenotypes (including emphysema and expiratory gas trapping). Pre- and post-bronchodilator spirometry measures were obtained using a standardized protocol and spirometer (ndd EasyOne Spirometer, Zurich, Switzerland). All study sites obtained local IRB approval to enroll participants and all subjects provided written informed consent. For the current cross-sectional WGS analysis, only baseline spirometry data were utilized.

Sub-Populations and Intermediate Outcome Measures in COPD Study (SPIROMICS)

SPIROMICS is a prospective cohort study that enrolled 3,200 participants into four strata (non-smokers, smokers without airflow obstruction, mild/moderate COPD, and severe COPD). Participants may be enrolled in concurrent observational studies, excluding the COPDGene Study[20], which facilitates combined analyses between SPIROMICS and COPDGene. The Institutional Review Boards/Ethics Committees of all the cooperating institutions have approved the study protocols. SPIROMICS participants were 40 to 80 years of age with a smoking history ≥20 pack-years with COPD (GOLD spirometric grades 1-4) and without COPD21,22. Participants were comprehensively characterized with annual pre- and post-bronchodilator lung function measures for up to 3 years, computed tomography scans, and standardized

questionnaires[21]. For the current cross-sectional WGS analysis, only baseline spirometry data were utilized.

Quality control of samples and variants included in TOPMed cohorts for analyses

For the pooled cohort, we first removed subjects who failed sample-level quality control. The filters included checking for pedigree errors, discrepancies between self-reported and genetic sex. Details regarding the quality control are described on the website (https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8). In addition, only one subject from each pair of duplicates was kept. There were 29,381 subjects from population/family-based cohorts and 11,771 subjects from COPD-enriched studies that passed the filtering.

For site-level quality control, variants were removed based on Mendelian discordance, a support vector machine (SVM) quality filter and excess heterozygosity filter. Details are described on the online document, https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8.

Acknowledgments for TOPMed whole genome sequencing

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Atherosclerosis Risk in Communities (ARIC)" (phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute for MIT and Harvard (3R01HL092577-06S1). WGS for "NHLBI TOPMed: Coronary Artery Risk Development in Young Adults (CARDIA)" (phs001612) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I). WGS for "NHLBI TOPMed: The Cleveland Family Study (CFS)" (phs000954) was performed at the University of Washington Northwest Genomics Center (3R01HL098433-05S1 and HHSN268201600032I). WGS for "NHLBI TOPMed: Cardiovascular Health Study (CHS)" (phs001368) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I) and Broad Institute Genomics Platform (HHSN268201600034I). WGS for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study (FHS)" (phs000974) was performed at the Broad Institute Genomics Platform (3U54HG003067-12S2and 3R01HL092577-06S1). WGS for "NHLBI TOPMed: Hispanic Community Health Study/Study of Latinos (HCHS/SOL)" (phs001395) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I). WGS for "NHLBI TOPMed: The Jackson Heart Study (JHS)" (phs000964) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C). WGS for "NHLBI TOPMed: Multi-

Ethnic Study of Atherosclerosis (MESA)" (phs001416) was performed at the Broad Institute Genomics Platform (3U54HG003067-13S1 and HHSN268201600034I). WGS for "NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene) in the TOPMed Program" (phs000951) was performed at the University of Washington Northwest Genomics Center (3R01 HL089856-08S1) and the Broad Institute Genomics Platform (HHSN268201500014C). WGS for "NHLBI TOPMed: SubPopulations and InteRmediate Outcome Measures in COPD Study (SPIROMICS)" (phs001927) was performed at Broad Institute Genomics Platform (HHSN268201600034I). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I).

Study specific acknowledgments: population/family-based cohorts

The Atherosclerosis Risk in Communities Study

The Genome Sequencing Program (GSP) was funded by the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The Centers for Common Disease Genomics (CCDG) program was supported by NHGRI and NHLBI, and whole genome sequencing was performed at the Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898 and R01HL059367). The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. SJL is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ZO1 ES043012).

The Coronary Artery Risk Development in Young Adults Study

The Coronary Artery Risk Development in Young Adults study is conducted at the University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I). CARDIA was also partially supported by the Intramural Research Program of the National Institute on Aging (NIA) and an intra-agency agreement between NIA and NHLBI (AG0005).

The Cleveland Family Study

The Cleveland Family Study and SR were supported by NIH grants HL 046389, HL113338, and 1R35HL135818. BC is supported by the NIH grant K01 HL135405 and an American Thoracic Society Foundation Unrestricted Grant (Sleep) (http://foundation.thoracic.org).

The Cardiovascular Health Study

This Cardiovascular Health Study (CHS) research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268200960009C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006; and NHLBI grants U01HL080295, U01HL130114, R01HL087652, R01HL105756, R01HL103612, R01HL085251, and R01HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The Framingham Heart Study

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I, and 75N92019D00031 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. Dr. Vasan is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine.

The Hispanic Community Health Study/Study of Latinos

The authors thank the staff and participants of HCHS/SOL for their important contributions. The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine

(HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03).

The Jackson Heart Study

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

The Multi-Ethnic Study of Atherosclerosis

MESA and the MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. The MESA Lung Study was funded by R01-HL077612 and R01-HL093081. Funding for genotyping was provided by NHLBI Contract N02-HL-64278. Also supported in part by NHLBI CHARGE Consortium Contract HL105756. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. Infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute (NHLBI) grant R01HL105756.

Study specific acknowledgments: COPD-enriched studies

Genetic Epidemiology of COPD (COPDGene)

The project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. 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.

COPD Foundation Funding

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.

COPDGene® Investigators – Core Units

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

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

Imaging Center: Juan Pablo Centeno; Jean-Paul Charbonnier, PhD; Harvey O. Coxson, PhD; Craig J. Galban, PhD; MeiLan K. Han, MD, MS; Eric A. Hoffman, Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber; David A. Lynch, MB; Pietro Nardelli, PhD; John D. Newell, Jr., MD; Aleena Notary; Andrea Oh, MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; Gonzalo Vegas Sanchez-Ferrero, PhD; Lucas Veitel; George R. Washko, MD; Carla G. Wilson, MS

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

Data Coordinating Center and Biostatistics, *National Jewish Health, Denver, CO*: Douglas Everett, PhD; Jim Crooks, PhD; Katherine Pratte, PhD; Matt Strand, PhD; Carla G. Wilson, MS
Epidemiology Core, *University of Colorado Anschutz Medical Campus, Aurora, CO*: John E.Hokanson, MPH, PhD; Erin Austin, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Kendra A. Young, PhD

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

Biomarker Core: Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; Farnoush Banaei-Kashani, Ph.D

COPDGene® Investigators – Clinical Centers

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

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS; Mustafa Atik, MD; Aladin Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Amit Parulekar, MD
Brigham and Women's Hospital, Boston, MA: Dawn L. DeMeo, MD, MPH; Craig Hersh, MD, MPH; Francine L. Jacobson, MD, MPH; George Washko, MD

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

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

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

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

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

Michael E. DeBakey VAMC, Houston, *TX*: Amir Sharafkhaneh, MD, PhD; Charlie Lan, DO
Minneapolis VA: Christine Wendt, MD; Brian Bell, MD; Ken M. Kunisaki, MD, MS

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

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

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

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

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

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

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

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

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

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

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

Sub-Populations and Intermediate Outcome Measures in COPD Study (SPIROMICS)

The authors thank the SPIROMICS participants and participating physicians, investigators and staff for making this research possible. More information about the study and how to access SPIROMICS data is available at www.spiromics.org. The authors would like to acknowledge the University of North Carolina at Chapel Hill BioSpecimen Processing Facility for sample processing, storage, and sample disbursements (http://bsp.web.unc.edu/).

We would like to acknowledge the following current and former investigators of the SPIROMICS sites and reading centers: Neil E Alexis, MD; Wayne H Anderson, PhD; Mehrdad Arjomandi, MD; Igor Barjaktarevic, MD, PhD; R Graham Barr, MD, DrPH; Patricia Basta, PhD; Lori A Bateman, MSc; Surya P Bhatt, MD; Eugene R Bleecker, MD; Richard C Boucher, MD; Russell P Bowler, MD, PhD; Stephanie A Christenson, MD; Alejandro P Comellas, MD; Christopher B Cooper, MD, PhD; David J Couper, PhD; Gerard J Criner, MD; Ronald G Crystal, MD; Jeffrey L Curtis, MD; Claire M Doerschuk, MD; Mark T Dransfield, MD; Brad Drummond, MD; Christine M Freeman, PhD; Craig Galban, PhD; MeiLan K Han, MD, MS; Nadia N Hansel, MD, MPH; Annette T Hastie, PhD; Eric A Hoffman, PhD; Yvonne Huang, MD; Robert J Kaner, MD; Richard E Kanner, MD; Eric C Kleerup, MD; Jerry A Krishnan, MD, PhD; Lisa M LaVange, PhD; Stephen C Lazarus, MD; Fernando J Martinez, MD, MS; Deborah A Meyers, PhD; Wendy C Moore, MD; John D Newell Jr, MD; Robert Paine, III, MD; Laura Paulin, MD, MHS; Stephen P Peters, MD, PhD; Cheryl Pirozzi, MD; Nirupama Putcha, MD, MHS; Elizabeth C Oelsner, MD, MPH; Wanda K O'Neal, PhD; Victor E Ortega, MD, PhD; Sanjeev Raman, MBBS, MD; Stephen I. Rennard, MD; Donald P Tashkin, MD; J Michael Wells, MD; Robert A Wise, MD; and Prescott G Woodruff, MD, MPH. The project officers from the Lung Division of the National Heart, Lung, and Blood Institute were Lisa Postow, PhD, and Lisa Viviano, BSN; SPIROMICS was supported by contracts from the NIH/NHLBI (HHSN268200900013C, HHSN268200900014C, HHSN268200900015C, HHSN268200900016C, HHSN268200900017C, HHSN268200900018C, HHSN268200900019C, HHSN268200900020C), grants from the NIH/NHLBI (U01 HL137880 and U24 HL141762), and supplemented by contributions made through the Foundation for the NIH and the COPD Foundation from AstraZeneca/MedImmune; Bayer; Bellerophon Therapeutics; Boehringer-Ingelheim Pharmaceuticals, Inc.; Chiesi Farmaceutici S.p.A.; Forest Research Institute, Inc.; GlaxoSmithKline; Grifols Therapeutics, Inc.; Ikaria, Inc.; Novartis Pharmaceuticals Corporation; Nycomed GmbH; ProterixBio; Regeneron Pharmaceuticals, Inc.; Sanofi; Sunovion; Takeda Pharmaceutical Company; and Theravance Biopharma and Mylan.

References

1. Oelsner, E.C., Balte, P.P., Cassano, P.A., Couper, D., Enright, P.L., Folsom, A.R., Hankinson, J., Jacobs, D.R., Kalhan, R., Kaplan, R., et al. (2018). Harmonization of Respiratory Data From 9 US Population-Based Cohorts: The NHLBI Pooled Cohorts Study. Am. J. Epidemiol. *187*, 2265–2278.

2. (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am. J. Epidemiol. *129*, 687–702.

3. Mirabelli, M.C., Preisser, J.S., Loehr, L.R., Agarwal, S.K., Barr, R.G., Couper, D.J., Hankinson, J.L., Hyun, N., Folsom, A.R., and London, S.J. (2016). Lung function decline over 25 years of follow-up among black and white adults in the ARIC study cohort. Respir. Med. *113*, 57–64.

4. Hughes, G.H., Cutter, G., Donahue, R., Friedman, G.D., Hulley, S., Hunkeler, E., Jacobs, D.R., Liu, K., Orden, S., and Pirie, P. (1987). Recruitment in the Coronary Artery Disease Risk Development in Young Adults (Cardia) Study. Control. Clin. Trials *8*, 68S-73S.

5. Friedman, G.D., Cutter, G.R., Donahue, R.P., Hughes, G.H., Hulley, S.B., Jacobs, D.R., Liu, K., and Savage, P.J. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. J. Clin. Epidemiol. *41*, 1105–1116.

6. (1979). ATS statement--Snowbird workshop on standardization of spirometry. Am. Rev. Respir. Dis. *119*, 831–838.

7. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., and Newman, A. (1991). The Cardiovascular Health Study: design and rationale. Ann. Epidemiol. *1*, 263–276.

8. Enright, P.L., Kronmal, R.A., Higgins, M., Schenker, M., and Haponik, E.F. (1993). Spirometry reference values for women and men 65 to 85 years of age. Cardiovascular health study. Am. Rev. Respir. Dis. *147*, 125–133.

9. Enright, P.L., Kronmal, R.A., Higgins, M.W., Schenker, M.B., and Haponik, E.F. (1994). Prevalence and correlates of respiratory symptoms and disease in the elderly. Cardiovascular Health Study. Chest *106*, 827–834.

10. Larkin, E.K., Patel, S.R., Goodloe, R.J., Li, Y., Zhu, X., Gray-McGuire, C., Adams, M.D., and Redline, S. (2010). A Candidate Gene Study of Obstructive Sleep Apnea in European Americans and African Americans. Am. J. Respir. Crit. Care Med. *182*, 947–953.

11. (2017). An Investigation of Coronary Heart Disease in Families: The Framingham Offspring Study. Am. J. Epidemiol. *185*, 1093–1102.

12. Lavange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and

cohort selection in the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 642–649.

13. Barr, R.G., Avilés-Santa, L., Davis, S.M., Aldrich, T.K., Gonzalez, F., Henderson, A.G., Kaplan, R.C., LaVange, L., Liu, K., Loredo, J.S., et al. (2016). Pulmonary Disease and Age at Immigration among Hispanics. Results from the Hispanic Community Health Study/Study of Latinos. Am. J. Respir. Crit. Care Med. *193*, 386–395.

14. Sorlie, P.D., Avilés-Santa, L.M., Wassertheil-Smoller, S., Kaplan, R.C., Daviglus, M.L., Giachello, A.L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. (2010). Design and implementation of the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 629–641.

15. Taylor, H.A. (2005). The Jackson Heart Study: an overview. Ethn. Dis. *15*, S6-1–3.

16. Taylor, H.A., Wilson, J.G., Jones, D.W., Sarpong, D.F., Srinivasan, A., Garrison, R.J., Nelson, C., and Wyatt, S.B. (2005). Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn. Dis. *15*, S6-4–17.

17. Wilson, J.G., Rotimi, C.N., Ekunwe, L., Royal, C.D.M., Crump, M.E., Wyatt, S.B., Steffes, M.W., Adeyemo, A., Zhou, J., Taylor, H.A., et al. (2005). Study design for genetic analysis in the Jackson Heart Study. Ethn. Dis. *15*, S6-30–37.

18. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881.

19. Hankinson, J.L., Kawut, S.M., Shahar, E., Smith, L.J., Stukovsky, K.H., and Barr, R.G. (2010). Performance of American Thoracic Society-recommended spirometry reference values in a multiethnic sample of adults: the multi-ethnic study of atherosclerosis (MESA) lung study. Chest *137*, 138–145.

20. Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., and Crapo, J.D. (2010). Genetic epidemiology of COPD (COPDGene) study design. COPD *7*, 32–43.

21. Couper, D., LaVange, L.M., Han, M., Barr, R.G., Bleecker, E., Hoffman, E.A., Kanner, R., Kleerup, E., Martinez, F.J., Woodruff, P.G., et al. (2014). Design of the Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS). Thorax *69*, 491–494.

22. Vogelmeier, C.F., Criner, G.J., Martinez, F.J., Anzueto, A., Barnes, P.J., Bourbeau, J., Celli, B.R., Chen, R., Decramer, M., Fabbri, L.M., et al. (2017). Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. Am. J. Respir. Crit. Care Med. *195*, 557–582.