**Whole Genome Sequence Analysis of Pulmonary Function and COPD in >19,000 Multi-ethnic Participants of the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program**

Xutong Zhao1, Dandi Qiao2, Chaojie Yang3, Nick Shrine4, Chiara Batini5, Tamar Sofer6,7, Sarah A. Gagliano1, Yanlin Ma3, Phuwanat Sakornsakolpat2, Pallavi P. Balte8, Dmitry Prokopenko2, Bing Yu9, Leslie A. Lange10, Josée Dupuis11, Brian E. Cade6,7, Jiwon Lee7, Sina A. Gharib12, Michelle Daya10, Cecelia A. Laurie13, Jennifer N. Nguyen3, Ingo Ruczinski14, Laura R. Loehr15, Traci M. Bartz13, Alanna C. Morrison16, Bruce M. Psaty17, L. Adrienne Cupples11,18, Ramachandran Vasan18,19, James G. Wilson20, Jerome I. Rotter21, Kathleen C. Barnes10, Stephen S. Rich3, George T. O’Connor22, Susan Redline6,7,23, Robert M. Reed24, Cathy C. Laurie13, Kristin M. Burkhart8, Robert C. Kaplan25,26, Elizabeth C. Oelsner8, Stephanie J. London27, Goncalo R Abecasis1, Edwin K Silverman2, R Graham Barr8, Michael H. Cho2,28, Ani Manichaikul3,28, On behalf of the TOPMed PFT, COPD and Lung Working Groups

1. Center for Statistical Genetics, and Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

2. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

3. Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA

4. Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, United Kingdom

5. Department of Health Sciences, University of Leicester, Leicester, United Kingdom

6. Department of Medicine, Harvard Medical School, Boston, Massachusetts. USA

7. Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, Massachusetts, USA

8. Department of Medicine, Columbia University College of Physicians and Surgeons

, New York, New York, USA

9. Department of Epidemiology, Human Genetics & Environmental Sciences, UTHealth School of Public Health, Houston, Texas, USA

10. Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado School of Medicine Anschutz Medical Campus, Aurora, Colorado, USA

11. Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA

12. Division of Pulmonary, Critical Care and Sleep Medicine, University of Washington, Seattle, Washington, USA

13. Department of Biostatistics, University of Washington, Seattle, Washington, USA

14. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

15. Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina , Chapel Hill, North Carolina, USA

16. 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, Texas, USA

17. Departments of Epidemiology and Medicine, University of Washington, Seattle, Washington, USA

18. Boston University and the National Heart Lung and Blood Institute’s Framingham Heart Study, Framingham, Massachusetts, USA

19. Department of Preventive Medicine and Epidemiology, Boston University  
School of Medicine and Public Health, Boston, Massachusetts, USA

20. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, USA

21. Institute for Translational Genomics and Population Sciences and Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, California, USA

22. Boston University School Of Medicine, Pulmonary Center, Boston, Massachusetts,

23. Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

24. Division of Pulmonary and Critical Care Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA

25. Albert Einstein College of Medicine, Department of Epidemiology and Population Health, Bronx, New York, USA

26. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

27. Epidemiology Branch, National Institute of Environmental Health Sciences, Durham, North Carolina, USA

28. These authors jointly supervised this work

**Correspondence:**

Michael H. Cho, MD, MPH

Assistant Professor of Medicine

Channing Division of Network Medicine,

Brigham and Women's Hospital,

Boston, MA USA

e-mail:

Phone:

Fax:

Ani Manichaikul, PhD

Assistant Professor of Public Health Sciences

Center for Public Health Genomics

University of Virginia School of Medicine

West Complex Room 6115

Charlottesville, VA USA

e-mail: [amanicha@virginia.edu](mailto:amanicha@virginia.edu)

Phone: 434-982-1612

Fax: 434-982-1815

**Abstract**

Chronic obstructive pulmonary disease (COPD), the fourth leading cause of death in the United States, is diagnosed by a decrease in lung function, namely forced expiratory volume in one second (FEV1) and its ratio to forced vital capacity (FEV1/FVC). Genome-wide association studies (GWAS) of lung function and COPD have already identified over 200 loci associated with one or more of these measures and conditions. We performed whole genome sequence (WGS) analysis of lung function traits (FEV1, FEV1/FVC and FVC) and COPD in the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program to examine evidence for rare and/or common trait-associated variants that were not identified in previous GWAS. WGS had deep (~30X) coverage with joint-sample variant calling and comprehensive variant level quality control in >50,000 TOPMed samples (freeze 5b). Lung function measures and COPD were analyzed in a subset of 10,686 subjects from four population-based and two family-based studies, as well as 8,499 individuals from the COPD-ascertained studies. These samples included 4,378 moderate-to-severe COPD cases out of which 1,722 had severe COPD. We conducted pooled analyses combining White (n=11,673), African American (n=6,288), Hispanic (n=709) and Asian (n=515), in addition to race/ethnic stratified analyses, with covariate adjustment for age, sex, height, weight, current and former smoking, pack-years of smoking, PCs of ancestry and sequencing center. After removing variants with Heterozygosity Count < 30, we recapitulated signals of association for 11 known GWAS loci and further identified 21 distinct novel association signals associated with one or more measures of pulmonary function or COPD. The most strongly associated SNPs for eight of the novel associated regions were common variants with minor allele frequency (MAF) greater than 0.1, and four of these novel common variant associations were identified only in stratified analysis of African Americans. Two novel SNPs from our TOPMed analyses showed evidence of replication in White British samples (n~320,000) from the UK Biobank (*CMIP* region SNP rs74469188, FVC replication *P*=1.4 x 10-5; and *FTO* region SNP rs7188378, FEV1/FVC ratio replication *P*=2.0 x 10-4). Our study thus identified multiple novel signals for lung function in African Americans and multi-ethnic samples, demonstrating the benefits of dense coverage across the genome by WGS, particularly for genome regions and ancestry groups not well-represented by previous GWAS efforts. [Need to update abstract text based on any new results. Also, numbers highlighted in yellow need to be checked and/or updated]

**Introduction**

Lung function, as measured by spirometry, is an important measure of health and a predictor of morbidity and mortality in the general population.1,2 Chronic obstructive pulmonary disease (COPD), characterized by chronic airflow limitation and an a abnormal response to noxious stimuli, is the fourth leading cause of death in the United States.3 COPD is defined by a decrease in lung function, namely forced expiratory volume in one second (FEV1) and its ratio to forced vital capacity (FEV1/FVC). While the main risk factor for COPD is cigarette smoking, the risk of COPD also increases with age, and can progress despite smoking cessation.4 Despite the enormous burden of COPD, there are no current pharmacologic therapies that convincingly slow progression of disease or affect mortality, and the genetic risk factors for disease are poorly understood.

COPD is a major cause of death in the United States5 and worldwide6, and has shown continued increases in prevalence in recent years.7 This rise has been higher among minorities and women, with an increase of 74% among African Americans and 88% among ‘other’ race/ethnicities (including Asian) from 1980 to 2000, compared a 65% increase among Whites8; during the same time, the number of cases in women doubled. Studies of COPD in Hispanics indicate large variations in morbidity and mortality across Latin America as well as in the United States9,10, reflecting a wide variety in exposure to risk factors including tobacco smoke, air pollution, socioeconomic status and access to care.11

COPD has substantial heritability, even accounting for differences in cigarette smoking behavior, with estimates from 35-60%.12–14 Lung function in the population is also similarly heritable: over 40% of variation in FEV1, FVC and FEV1/FVC is attributed to genetic factors.15 Genome-wide association studies (GWAS) have identified a number of loci for both COPD and pulmonary function. Studies led by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE)/SpiroMeta16–18 and the UK Biobank19, including European ancestry subjects, identified over nearly 100 gene regions that contain common single nucleotide polymorphisms (SNPs) significantly associated with FEV1 or FEV1/FVC. A GWAS of pulmonary function from the CHARGE consortium, combining participants of European, African, African and Hispanic/Latino ethnicities identified an additional 50 loci in ancestry specific and multi-ethnic analyses.16 More recently, a large scale GWAS including >400,000 European ancestry participants from the UK Biobank identified an additional >100 loci for pulmonary function traits.20 Thus, increased ethnic diversity as well as order of magnitude increases in sample size have proven to be effective ways of identifying novel genetic loci related pulmonary function.

The largest published GWAS of COPD to date, from the International COPD Genetics Consortium, included 35,735 cases 222,076 controls from the UK Biobank and additional studies from the International COPD Genetics Consortium.21 Of the 35 novel COPD-associated loci identified in that study, 13 were associated with lung function in independent samples from the SpiroMeta consortium after Bonferroni correction for multiple testing, 14 showed nominal association with lung function (*P*<0.05), while the remaining 8 loci did not demonstrate evidence of association with lung function traits. Thus, though there is a high degree of overlap between COPD and lung function, there are relative advantages to studying quantitative versus dichotomous outcomes, and both approaches can be fruitful.

Completed GWAS have been limited by imperfect matching and sample size of reference panels (*e.g.* HapMap22 or 1000 Genomes23), resulting in missing information on both common and rare variants, particularly in African Americans and Hispanics. Rare variants affect COPD susceptibility: severe alpha-1 antitrypsin deficiency has been recognized for decades as a genetic cause of COPD14; however, additional rare variants have been difficult to identify. To address these limitations, we leveraged deep sequencing in the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program to perform the first large scale, multi-ethnic whole genome sequence (WGS) analysis of pulmonary function and COPD.

**Methods**

**Study Samples**

Participants were included from five population- and family-based cohorts (the Atherosclerosis Risk in Communities [ARIC] Study, the Cleveland Family Study [CFS], the Cardiovascular Health Study [CHS], the Framingham Heart Study [Framingham], the Jackson Heart Study [JHS], and the Multi-Ethnic Study of Atherosclerosis [MESA]) and two COPD-ascertained studies (the Genetic Epidemiology of COPD [COPDGene] Study and the Boston Early Onset COPD [EOCOPD] Study). Detailed cohort descriptions are provided in the **Supplementary Text**.

**Phenotype definition**

Phenotype harmonization of Pulmonary Function Test (PFT) measures, including pre-bronchodilator FEV1, FVC, and FEV1/FVC ratio, was conducted following the protocol of the NHLBI Pooled Cohorts Study (**Supplementary Text**).24 Based on the quantitative measures of PFT and self-reported categories of race/ethnicity, we calculated race/ethnic-specific predicted values of FEV1 for White, African American, Hispanic and Asian participants using the equations of Hankinson25 that were determined for White, African American, Mexican American, and White reference populations, respectively. COPD cases and controls were then defined as follows:

* **Moderate-to-Severe COPD:** pre-bronchodilator FEV1 < 80% predicted and FEV1/FVC < 0.7,
* **Severe COPD:** pre-bronchodilator FEV1 < 50% predicted and FEV1/FVC < 0.7, and
* **Controls:** pre-bronchodilator FEV1 ≥ 80% predicted and FEV1/FVC ≥ 0.7.

**Whole genome sequence data**

Whole Genome Sequencing (WGS) in TOPMed had deep (~30X) coverage with joint-sample variant calling and variant level quality control in >50,000 TOPMed samples (freeze 5b).26

**Single Variant analyses**

Analyses were conducted using SAIGE-LMM27 and stratified by study design (population- and family-based studies vs. COPD-ascertained studies), as well as combined. Within strata, separate analyses in White vs. African Americans, as well as pooled across race/ethnic groups.

*Quantitative trait analysis of FEV1. FVC and FEV1/FVC:* We applied inverse normal transform of phenotypes and implemented a heterogeneous variance model28 to account for different phenotype distributions across studies. We incorporated covariate adjustment for age, sex, height, weight (FVC only), current / former smoking, pack-years of smoking, PCs of ancestry and sequencing center

*Dichotomous trait analysis of COPD:*Case-control analyses incorporated covariate adjustment for age, sex, pack-years, and ever / never smoking, PCs of ancestry and sequencing center.

*Variant-level filter:* In addition to standard quality control filters applied to the TOPMed Freeze 5b data set26, whole genome sequence analysis results were filtered on heterozygosity count > 30. We applied a genome-wide significance threshold of *P*<5x10-8 for reporting novel and known variants in this manuscript.

*Annotation of novel variants:* Novel variants identified by WGS analyses were annotated using the WGS Annotator (WGSA).29

**Concordance of TOPMed WGS calls with genotypes imputed by GWAS**

We examined the concordance of genotypes for the novel associated variants using variant calls from TOPMed Freeze 5b compared to genotypes obtained using imputation of genome-wide genotypes to various reference panels including the 1000 Genomes Phase 123, 1000 Genomes Phase 330, and the Haplotype Reference Consortium. Need to add text regarding genome-wide imputation in MESA and COPDGene.

**Conditional analysis based on GWAS**

We assessed independence of signals using GCTA-COJO (REF). GCTA-COJO is a summary statistics based approach for conditional analysis. The analysis was focused on variants within 1-Mb region of the most strongly associated signals. Since the MHC region has more complex structure, we conducted the analysis on a 9-Mb region. Conditioning on the most strongly associated variants, any variant that had conditional p-value < (?) was assigned as an independent signal.

**Rare putative loss of function (pLOF) variant burden test**

Gene-based burden test was conducted on 228,966 pLoF variants. These variants were previously identified using Loss-Of-Function Transcript Effect Estimator (LOFTEE) v0.3-beta and Variant Effect Predictor (VEP) v94 (REF). Variants used in our analysis included stop-gained, frameshift, and splice site disturbing variants annotated as high-confidence (HC) pLoF. The genes were defined based on GENCODE v29. Only SNPs with MAF of <0.5% were included. Gene-level burden was generated by aggregating low frequency pLoF variants together. To be specific, for each gene if an individual carries the alternate allele in at least one of the pLoF variants, we assigned 1 for that individual, 0 if not. Then we tested the burden variables for association with traits using SAIGE-LMM. **(1)** **[**We applied a genome-wide significance threshold of for reporting significant signals.**] (2) [**We focused on genes around loci identified in the current or previous studies. The candidate genes include 24 Mendelian gene and genes that overlap with 100kb flank region of the GWAS top associated variants. In total, we checked 129 genes(regions) for FEV1, 124 genes for FVC, 161 genes for FEV1/FVC, 183 genes for COPD (Supplementary table). The Bonferroni-corrected p-values were derived as 0.05 divided by the number of genes to test, which are for FEV1, for FVC, for FEV1/FVC and for COPD. **] (3) [**We focused on 24 candidate Mendelian genes. The Bonferroni-corrected p-values was derived as 0.05 divided by the number of genes to test, which is .**]**

**Replication cohorts and analysis**

For those variants demonstrating novel associations with one or more measures of pulmonary function or COPD, we examined evidence of replication in the UK Biobank and the Hispanic Community Health Study / Study of Latinos (HCHS/SoL). Details provided in the **Supplementary Text**.

**Results**

**Participant characteristics**

Our study sample included a total of >19,177 participants, including 10,678 participants from population- and family-based studies, as well as 8,499 participants from COPD-ascertained studies (**Table 1**). Using participant self-reported race/ethnicitiy, 11,673 and 6,280 participants were categorized as non-Hispanic White or African American, respectively, while the remaining 1,224 participants represented Hispanic, Asian and other race/ethnicities. The combined samples included 4,232 moderate-to-severe COPD cases and 1722 severe COPD cases. Among these, 1191 moderate-to-severe and 203 severe COPD cases were contributed by population- and family-based cohorts, and the remaining COPD cases were from the COPD-ascertained studies (**Supplementary Table 1**).

**Single variant analysis**

*Overlap with known GWAS loci:* In single variant analysis, we recapitulated at genome-wide significance the signals of association for 13 known GWAS loci. In particular, we report association with *HTR4* in population- and family-based analysis, associations with *CHRNA3/5*; *FAM13A*; *EEFSEC*; *RIN3*; and *HHIP* in analysis of COPD-ascertained samples, and associations with *CHRNA3/5*; *GSTCD*; *AGER*; *DSP*; *RIN3*; *HHIP*; *FAM13A*; *THSD4*; *C1GALT1*; and *EEFSEC* in combined analysis incorporating both population-based and COPD-ascertained strata (**Supplementary Table 2**).

*Novel associated variants:* Among our genome-wide significant findings, we report 26 association signals across all strata included in analyses, covering 21 distinct loci (depending on conditional analysis) (**Table 2, Supplementary Figure 1**). These variants include four common variant signals identified in stratified analysis of African Americans, 5 distinct signals on the X chromosome, and 4 rare and infrequent variants with minor allele frequency less than 5%. Among these, we identified one novel association with FEV1/FVC at rs7188378, located near *FTO* (T>C, and ). For this variant, we performed sensitivity analysis with additional covariate adjustment for weight, and noted only a modest attenuation in the association signal under this model (T>C, and ).

Our novel variants demonstrated largely consistent directions of effect across cohorts. However, for those variants identified in combined analysis including both population- and family-based cohorts and COPD-ascertained studies, we observed several cases in which the effect sizes were substantially larger in the COPD-ascertained studies compared to population-based cohorts. For example, rs7188378, which is significantly associated with FEV1/FVC in the combined analysis, had larger effect sizes in COPDGene and EOCOPD studies than most of the other studies (COPDGene: ; EOCOPD: vs. ARIC: ; FHS: ; JHS: ; MESA: ). We did not compare with CFS and CHS here due to their small sample sizes. (We should provide some specific examples. Xutong is also working on a **Supplementary Figure 2** to demonstrate cohort-specific results.)

**Characterizing novel associated variants**

*Concordance of whole genome sequence with imputed genotypes:* We observed generally very good concordance between TOPMed Freeze 5b WGS variant calls and those obtained by imputation of genome-wide genotypes. There were a few exceptions with some of the variants demonstrating poorer imputation quality, in particular for variants with minor allele frequencies less than 1% (**Supplementary Table 3**).

*Annotation of the identified variants:*Among our novel associated variants (**Table 2**), we identified one missense mutation (rs5376; p.Ser334Asn in *GALR1*). We further identified three non-coding exon variants (rs4076943, rs74469188, rs7046490), suggesting these variants may play a role in regulation of other genes and/or transcripts (**Supplementary Table 4**).

**Conditional analysis**

We used GCTA-COJO to identify variants that were independent of lead variants in

pulmonary function and COPD loci identified in the current study. Conditioning on the lead signals, we did not identify independent variants within 1-Mb of the lead variants achieving genome-wide significance.

* Should we also do some analysis of fine mapping (like we did for Dmitry’s paper?)

**Rare putative loss of function (pLOF) variant burden test**

(1) In the gene-based burden test using rare pLOF variants, a burden of rare pLoF variants in ACKR1 (comprised of 9 rare pLoF variants; cumulative allele frequency cases vs. controls= 0 vs. ) was significantly associated with moderate-to-severe COPD in the pooled-ethnic population- and family-based studies. The result were driven by frameshift variants in this gene (chr1:159206292:CT:C, chr1:159206156:AG:A and chr1:159205445:TG:T).

(2) Among 24 Mendelian candidate genes we investigated in the rare pLOF variants burden test, a burden of rare pLoF variants in ATP6V1E1 (comprised of 2 rare pLoF variants) was significantly associated with moderate-to-severe COPD in the pooled-ethnic population- and family-based studies (cumulative allele frequency cases vs. controls= vs. 0) as well as the combined studies (cumulative allele frequency cases vs. controls= vs. 0). The result were driven by one splice acceptor variant in this gene (chr22:17619129:T:C).

(3) Among the Mendelian candidate genes and genes near GWAS loci we investigated in the rare pLOF variants burden test, a burden of rare pLoF variants in ARHGEF17 (comprised of 26 rare pLoF variants) was significantly associated with FEV1/FVC in the combined study (). The result were driven by stop gained variants (chr11:73308663:C:T, chr11:73308682:C:A, chr11:73308729:G:T, chr11:73309239:C:T, chr11:73347126:G:T, chr11:73347129:G:T), frameshift variants (chr11:73310488:CT:C, chr11:73310815:AG:A, chr11:73346973:G:GC, chr11:73356179:T:TTC, chr11:73362161:TG:T) and splice donor (chr11:73357135:G:A). DENND2D (comprised of 9 rare pLoF variants) was significantly associated with moderate-to-severe COPD in the African Americans in the combined study (cumulative allele frequency cases vs. controls= vs. ). The result were driven by stop gained variants (chr1:111200399:G:A, chr1:111200414:G:A), frameshift variant (chr1:111188329:GCACAAAGGGGCCTGAAA:G) and splice donor (chr1:111189211:C:T). MTCL1 (comprised of 28 rare pLoF variants) was significantly associated with severe COPD in the Whites of the COPD-ascertained studies (cumulative allele frequency cases vs. controls= vs. ). The result were driven by stop gained variant (chr18:8807058:C:T) and frameshift variants (chr18:8706455:G:GC, chr18:8784799:TCCCC:T).

**Examination of replication in independent cohorts**

* Look-up in published GWAS of PFT (Nick Shrine), smoking intensity (Chiara Batini), and COPD (Michael / Dandi) from UK Biobank / ICGC: The *FTO* locus identified in our TOPMed analysis of FEV1/FVC replicates with consistent direction of effect in UK Biobank White British samples. The locus near CMIP also shows replication in UK Biobank, but the direction of effect is not consistent with that observed in TOPMed. It does not appear that either of these replication signals can be attributed to smoking intensity.
* Look-up in HCHS/SOL GWAS results (Tamar Sofer): No significant replication, perhaps due to sample size.
* Waiting for African ancestry results from UK Biobank
* Currently can’t include UK Biobank results from Sarah and Xutong due to rules about publication

**Need to add:**

**eQTL/pQTL or other look-ups for novel SNPs (Table 2)**

* MESA (Francois Aguet): re-contact Francois to see if he can contribute (Ani)
* COPDGene (Dandi, can we summarize these results in a Table?)

**Phewas:** Michael, what are your suggestions for this?

**Other:**

* Finalize what Tables / Figures belong in manuscript

**Discussion**

Need to update / expand this section (Ani will draft)

* In this first pooled cohorts WGS analysis of pulmonary function and COPD from the NHLBI TOPMed Program, we recapitulated known GWAS loci and further identified multiple novel loci.
* Our study demonstrates the value of WGS approaches, particularly for identification of associations with variants that may be difficult to impute, including common variants in African Americans. In addition, our study’s inclusion of COPD cases give us a unique ability to identify variants that have stronger effects in that subset.
* Future work will include:
  + - Replication efforts, possibly leveraging expanded sample sizes through genome-wide imputation using TOPMed as a reference panel.
    - Expanded pooled cohorts WGS analyses with additional samples recently added to TOPMed.

**References**

1. Schünemann, H. J., Dorn, J., Grant, B. J., Winkelstein, W. & Trevisan, M. Pulmonary function is a long-term predictor of mortality in the general population: 29-year follow-up of the Buffalo Health Study. *Chest* (2000). doi:10.1378/chest.118.3.656

2. Young, R. P., Hopkins, R. & Eaton, T. E. Forced expiratory volume in one second: Not just a lung function test but a marker of premature death from all causes. *European Respiratory Journal* (2007). doi:10.1183/09031936.00021707

3. Heron, M. Deaths: Leading Causes for 2016. *Natl. Vital Stat. Rep.* (2018). doi:10.1016/S0140-6736(15)00057-4

4. Disease, G. I. for chronic obstructive lung. *Chronic Obstructive Pulmonary Disease Updated 2010 Global Initiative for Chronic Obstructive Lung Disease*. *Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease* (2016). doi:10.1097/00008483-200207000-00004

5. Murphy, S. L., Xu, J., Kochanek, K. D. & Arias, E. Mortality in the United States, 2017. *NCHS Data Brief* (2018).

6. Mathers C., Stevens, G. A., Mahanani W.R., Fat, D. M. & Hogan, D. WHO | Disease burden and mortality estimates. *Who* 1–65 (2018). doi:2013

7. May, S. M. & Li, J. T. C. Burden of chronic obstructive pulmonary disease: Healthcare costs and beyond. *Allergy Asthma Proc.* (2015). doi:10.2500/aap.2015.36.3812

8. Mannino, D. M., Homa, D. M., Akinbami, L. J., Ford, E. S. & Redd, S. C. Chronic obstructive pulmonary disease surveillance--United States, 1971-2000. *Respir Care* **47,** 1184–1199 (2002).

9. Brehm, J. M. & Celedón, J. C. Chronic obstructive pulmonary disease in hispanics. *American Journal of Respiratory and Critical Care Medicine* (2008). doi:10.1164/rccm.200708-1274PP

10. Barr, R. G. *et al.* Pulmonary disease and age at immigration among hispanics: Results from the hispanic community health study/study of latinos. *Am. J. Respir. Crit. Care Med.* (2016). doi:10.1164/rccm.201506-1211OC

11. Eisner, M. D. *et al.* Socioeconomic status, race and COPD health outcomes. *J. Epidemiol. Community Health* (2011). doi:10.1136/jech.2009.089722

12. Zhou, J. J. *et al.* Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am. J. Respir. Crit. Care Med.* (2013). doi:10.1164/rccm.201302-0263OC

13. Ingebrigtsen, T. *et al.* Genetic influences on chronic obstructive pulmonary disease - A twin study. *Respir. Med.* (2010). doi:10.1016/j.rmed.2010.05.004

14. Silverman, E. K. *et al.* 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.* (1998). doi:10.1164/ajrccm.157.6.9706014

15. Klimentidis, Y. C. *et al.* Heritability of pulmonary function estimated from pedigree and whole-genome markers. *Front. Genet.* (2013). doi:10.3389/fgene.2013.00174

16. Wyss, A. B. *et al.* Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat. Commun.* (2018). doi:10.1038/s41467-018-05369-0

17. Artigas, M. S. *et al.* Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat. Genet.* **43,** 1082–1090 (2011).

18. Loth, D. W. *et al.* Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat. Genet.* **46,** (2014).

19. Wain, L. V *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet* **49,** 416–425 (2017).

20. Shrine, N. *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* (2019). doi:10.1038/s41588-018-0321-7

21. Sakornsakolpat, P. *et al.* Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* (2019). doi:10.1038/s41588-018-0342-2

22. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* (2007). doi:10.1038/nature06258

23. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* (2012). doi:10.1038/nature11632.An

24. Oelsner, E. C. *et al.* Harmonization of Respiratory Data From 9 US Population-Based Cohorts. *Am. J. Epidemiol.* (2018). doi:10.1093/aje/kwy139

25. Hankinson, J. L., Odencrantz, J. R. & Fedan, K. B. Spirometric reference values from a sample of the general U.S. Population. *Am. J. Respir. Crit. Care Med.* (1999). doi:10.1164/ajrccm.159.1.9712108

26. Taliun, D., Harris, D. N., Kessler, M. D. & Al., E. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nat. (under Rev.*

27. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0184-y

28. Sofer, T. *et al.* A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet. Epidemiol.* **Epub ahead,** (2019).

29. Liu, X. *et al.* WGSA: an annotation pipeline for human genome sequencing studies. *J. Med. Genet.* (2015). doi:10.1136/jmedgenet-2015-103423

30. Auton, A. *et al.* A global reference for human genetic variation. *Nature* (2015). doi:10.1038/nature15393

**Acknowledgments**

We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. WGS was supported by the National Heart, Lung and Blood Institute (NHLBI). 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 TOPMed cohorts was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C) and at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). TOPMed phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). 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 and R01-HL142028. This research was supported by NIH/NHLBI R01 HL131565 and U01 HL117626, Study-specific acknowledgments are given in the Supplementary Text. We are grateful to all members of the TOPMed consortium who recruited study participants, collected data, and contributed to this project in many other ways.

**Table 1:** Summary of study-participants included in analyses.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Population- and family-based** | | **COPD-ascertained** | | **All combined** | |
| **Race/ethnic group** | **Study** | **N** | **Study** | **N** | **Study** | **N** |
|  |  |  |  |  | ARIC | 2603 |
|  | ARIC | 2603 |  |  | CFS | 202 |
|  | CFS | 202 |  |  | CHS | 41 |
|  | CHS | 41 | COPDGene | 5713 | COPDGene | 5713 |
| **White** | Framingham | 1835 | EOCOPD | 55 | EOCOPD | 55 |
|  | MESA | 1224 | **Total** | **5768** | Framingham | 1835 |
|  | **Total** | **5905** |  |  | MESA | 1224 |
|  |  |  |  |  | **Total** | **11673** |
|  | ARIC | 154 |  |  | ARIC | 154 |
|  | CFS | 203 |  |  | CFS | 203 |
| **African** | JHS | 2388 | COPDGene | 2731 | COPDGene | 2731 |
| **American** | MESA | 804 | **Total** | **2731** | JHS | 2388 |
|  | **Total** | **3549** |  |  | MESA | 804 |
|  |  |  |  |  | **Total** | **6280** |
| **Other** | MESA  **Total** | 1224  **1224** | - | - | MESA  **Total** | 1224  **1224** |
|  |  |  |  |  | ARIC | 2757 |
|  | ARIC | 2757 |  |  | CFS | 405 |
|  | CFS | 405 |  |  | CHS | 41 |
|  | CHS | 41 | COPDGene | 8444 | COPDGene | 8444 |
| **All** | Framingham | 1835 | EOCOPD | 55 | EOCOPD | 55 |
| **combined** | JHS | 2388 | **Total** | **8499** | Framingham | 1835 |
|  | MESA | 3252 |  |  | JHS | 2388 |
|  | **Total** | **10678** |  |  | MESA | 3252 |
|  |  |  |  |  | **Total** | **19177** |

**Table 2:** Summary of novel 22 genome-wide significant results not previously identified by whole genome sequence analysis of the TOPMed Participants.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Trait** | **Stratum** | **Race / ethnic group** | **Chr** | **Pos (b38)** | **rsid** | **Effect / other allele** | **Effect allele FREQ** | **Het count** | **Beta** | **SE** | ***P*-value** | **Nearest gene** |
| **FEV1** | Population- and family-based | African American | 18 | 77,268,853 | rs5376 | A/G | 0.71 | 1450 | 0.06 | 0.01 | 4.97E-08 | *GALR1* |
| All | 3 | 37,598,577 | rs73064792 | A/G | 0.15 | 2693 | 0.05 | 0.01 | 4.98E-08 | *ITGA9* |
| All | X | 46,670,331 | rs142712254 | T/C | 0.01 | 141 | -0.13 | 0.02 | 4.41E-08 | *SLC9A7* |
| COPD-ascertained | White | X | 80,958,253 | rs142755000 | A/G | 0.03 | 141 | -0.18 | 0.03 | 3.58E-08 | *BRWD3* |
| African American | 11 | 11,239,853 | rs4076943 | T/A | 0.28 | 1128 | 0.12 | 0.02 | 2.37E-08 | *GALNT18* |
| Combined | White | X | 143,199,689 | rs145318069 | G/A | 0.01 | 152 | 0.15 | 0.03 | 4.48E-08 | *SPANXN4* |
| Combined | All | X | 143,199,689 | rs145318069 | G/A | 0.01 | 168 | 0.13 | 0.02 | 4.31E-08 | *SPANXN4* |
| **FVC** | Population- and family-based | White | X | 46,687,945 | rs182915372 | C/T | 0.01 | 79 | -0.20 | 0.04 | 3.26E-08 | *SLC9A7* |
| All | X | 46,687,945 | rs182915372 | C/T | 0.01 | 90 | -0.19 | 0.03 | 1.22E-08 | *SLC9A7* |
| COPD-ascertained | White | 1 | 196,989,333 | rs371740347 | C/T | 0.01 | 93 | 0.43 | 0.07 | 2.67E-09 | *CFHR5* |
| African American | 16 | 81,611,365 | rs74469188 | C/T | 0.15 | 714 | 0.14 | 0.03 | 2.30E-08 | *CMIP* |
| All | 1 | 196,989,333 | rs371740347 | C/T | 0.01 | 104 | 0.38 | 0.07 | 1.07E-08 | *CFHR5* |
| All | 9 | 673,533 | rs7046490 | A/G | 0.39 | 3840 | 0.06 | 0.01 | 2.47E-08 | *KANK1* |
| All | X | 47,087,005 | rs12556310 | G/C | 0.56 | 1901 | 0.05 | 0.01 | 3.30E-08 | *RGN* |
| Combined | White | 8 | 68,126,989 | rs78101151 | G/A | 0.003 | 71 | -0.41 | 0.08 | 4.90E-08 | *PREX2* |
| White | X | 47,317,317 | rs5953026 | G/A | 0.60 | 2743 | 0.04 | 0.01 | 3.26E-08 | *ZNF157* |
| All | X | 47,084,756 | rs6611328 | C/A | 0.57 | 4697 | 0.03 | 0.01 | 1.46E-08 | *RGN* |
| **FEV1/FVC** | Population- and family-based | African American | 13 | 84,871,276 | rs149004949 | A/G | 0.01 | 66 | -5.69 | 0.96 | 3.58E-09 | *SLITRK6* |
| All | 14 | 36,176,982 | rs145538273 | G/A | 0.004 | 89 | -4.40 | 0.80 | 4.37E-08 | *MBIP* |
| COPD-ascertained | African American | 8 | 133,793,876 | rs144870669 | C/T | 0.01 | 36 | 0.83 | 0.15 | 1.67E-08 | *ST3GAL1* |
| Combined | White | 6 | 102,490,266 | rs572153283 | A/AT | 0.01 | 329 | -3.49 | 0.63 | 3.08E-08 | *GRIK2* |
| African American | 6 | 107,949,366 | rs1032155362 | T/G | 0.003 | 34 | -10.44 | 1.69 | 6.84E-10 | *SEC63* |
| All | 16 | 53,872,940 | rs7188378 | C/T | 0.49 | 9325 | -0.61 | 0.11 | 3.44E-08 | *FTO* |
| **Moderate-to-Severe COPD** | Population- and family-based | African American | 5 | 60,908,891 | rs114353081 | T/A | 0.03 | 215 | 1.81 | 0.33 | 3.19E-08 | *ERCC8* |
| African American | 9 | 24,295,928 | rs17197726 | G/T | 0.17 | 1004 | 0.76 | 0.14 | 3.79E-08 | *IZUMO3* |
| COPD-ascertained | All | 15 | 53,390,153 | rs72740913 | T/G | 0.01 | 147 | 1.08 | 0.20 | 3.27E-08 | *WDR72* |

**Supplementary Tables**

**Supplementary Table 1:** Detailed cohort-specific descriptive characteristics

**Supplementary Table 2:** Detailed list of genome-wide significant variants, including those identified in previously published studies

**Supplementary Table 3:** Concordance of whole genome sequence calls with imputed genotypes

**Supplementary Table 4:** Annotation of novel variants

**Supplementary Table 5:** Examination of association of novel variants with measures of pulmonary function and COPD in White British samples from the UK Biobank

**Supplementary Table 6:** Examination of association of novel variants with measures of smoking intensity in White British samples from the UK Biobank

**Supplementary Table 7:** Examination of association of novel variants with measures of pulmonary function and COPD in African samples from the UK Biobank

**Supplementary Table 8:** Examination of association of novel variants with measures of smoking intensity in African samples from the UK Biobank

**Supplementary Table 9:** Examination of association of novel variants with measures of pulmonary function and COPD in HCHS/SoL

**Supplementary Table 10:** Genomic inflation factors from Q-Q plots of observed and expected p-values for all stratified analysis.

**Supplementary Table 11:** Candidate gene for pLOF analysis.

**Supplementary Figures**

**Supplementary Figure 1:** Local association plots for novel variants

**Supplementary Figure 2:** Forest plots demonstrating cohort-specific effects for novel variants