Whole-Genome Sequencing in Severe Chronic Obstructive Pulmonary Disease

Dmitry Prokopenko¹, Phuwanat Sakornsakolpat¹, Heide Loehlein Fier^{2,3}, Dandi Qiao¹, Margaret M. Parker¹, Merry-Lynn N. McDonald⁴, Ani Manichaikul⁵, Stephen S. Rich⁵, R. Graham Barr^{6,7}, Christopher J. Williams⁸, Mark L. Brantly⁸, Christoph Lange³, Terri H. Beaty⁹, James D. Crapo¹⁰, Edwin K. Silverman¹, and Michael H. Cho¹; on behalf of COPDGene Study Investigators* and NHLBI TOPMed Investigators[†]

¹Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts; ²Working Group of Genomic Mathematics, University of Bonn, Bonn, Germany; ³Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; ⁴Division of Pulmonary, Allergy and Critical Care Medicine, University of Alabama at Birmingham, Birmingham, Alabama; ⁵Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia; ⁶Department of Medicine, College of Physicians and Surgeons and ⁷Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York; ⁸Division of Pulmonary Critical Care and Sleep Medicine, College of Medicine, University of Florida, Gainesville, Florida; ⁹Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland; and ¹⁰National Jewish Health, Denver, Colorado

ORCID IDs: 0000-0002-1844-5652 (D.P.); 0000-0002-4308-0974 (P.S.); 0000-0002-4907-1657 (M.H.C.).

Abstract

Genome-wide association studies have identified common variants associated with chronic obstructive pulmonary disease (COPD). Whole-genome sequencing (WGS) offers comprehensive coverage of the entire genome, as compared with genotyping arrays or exome sequencing. We hypothesized that WGS in subjects with severe COPD and smoking control subjects with normal pulmonary function would allow us to identify novel genetic determinants of COPD. We sequenced 821 patients with severe COPD and 973 control subjects from the COPDGene and Boston Early-Onset COPD studies, including both non-Hispanic white and African American individuals. We performed single-variant and grouped-variant analyses, and in addition, we assessed the overlap of variants between sequencing- and array-based imputation. Our most significantly associated variant was in a known region near *HHIP* (combined $P = 1.6 \times 10^{-9}$); additional variants

approaching genome-wide significance included previously described regions in *CHRNA5*, *TNSI*, and *SERPINA6/SERPINA1* (the latter in African American individuals). None of our associations were clearly driven by rare variants, and we found minimal evidence of replication of genes identified by previously reported smaller sequencing studies. With WGS, we identified more than 20 million new variants, not seen with imputation, including more than 10,000 of potential importance in previously identified COPD genome-wide association study regions. WGS in severe COPD identifies a large number of potentially important functional variants, with the strongest associations being in known COPD risk loci, including *HHIP* and *SERPINA1*. Larger sample sizes will be needed to identify associated variants in novel regions of the genome.

Keywords: whole-genome sequencing; chronic obstructive pulmonary disease; association studies

(Received in original form March 12, 2018; accepted in final form June 27, 2018)

*A full listing of COPDGene investigators can be found at: <http://www.copdgene.org/directory/>.

†A full listing of TOPMed investigators can be found in the data supplement for this article.

Supported by National Institutes of Health (NIH) grants R01 HL113264 (E.K.S.), R01 HL137927 (M.H.C. and E.K.S.), R01 HL135142 (M.H.C.), P01 HL089856 (E.K.S.), R01 HL113264 (M.H.C. and E.K.S.), R01 HL089897 (J.D.C.), P01 HL132825, R00 HL121087 (M.-L.N.M.), and R01 HL131565 and the Alpha-1 Foundation (A.M.). The Genetic Epidemiology of Chronic Obstructive Pulmonary Disease (COPDGene) Study was supported by National Heart, Lung, and Blood Institute (NHLBI) awards R01 HL089897 and R01 HL089856. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHLBI or the NIH. The COPDGene Study (NCT00608764) is also supported by the Chronic Obstructive Pulmonary Disease (COPD) Foundation through contributions made to an industry advisory board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens, and Sunovion. The Multi-Ethnic Study of Atherosclerosis (MESA) and the MESA SNP Health Association Resource (SHARe) project are conducted and supported by the NHLBI in collaboration with MESA investigators. Support for MESA is provided by NIH contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. The MESA Family Study is conducted and supported by the NHLBI in collaboration with MESA investigators. Support is provided by NIH grants and contracts R01 HL071051, R01 HL071205, R01 HL071250, R01 HL071251, R01 HL071258, and R01 HL071259; by National Center for Research Resources grant UL1RR033176; and by National Center for Advancing Translational Sciences grant UL1TR001881. Funding for SHARe genotyping was provided by NHLBI contract N02-HL-64278. Genotyping was performed at Affymetrix and the Broad Institute of Harvard and MIT using the Affymetrix Genome-Wide Human SNP Array 6.0. The MESA Lung and MESA COPD studies are funded by NIH grants R01 HL077612 and R01 HL093081. Whole-genome sequencing for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the NHLBI. Whole-genome sequencing for TOPMed phs000951 and phs000946 was performed at the University of Washington Northwest Genomics Center (3R01 HL089856-08S1). Centralized read mapping and genotype calling, together with variant quality metrics and filtering, were provided by the TOPMed Informatics Research Center (3R01 HL-117626-02S1). Phenotype harmonization, data management, sample identity quality control, and general study coordination were provided by the TOPMed Data Coordinating Center (3R01 HL-120393-02S1).

Am J Respir Cell Mol Biol Vol 59, Iss 5, pp 614–622, Nov 2018

Copyright © 2018 by the American Thoracic Society

Originally Published in Press as DOI: 10.1165/rcmb.2018-0088OC on June 27, 2018

Internet address: www.atsjournals.org

Chronic obstructive pulmonary disease (COPD) is among the top four leading causes of death in the United States and worldwide (1). Although COPD is highly influenced by smoking behavior, genetic factors do play a significant role and can explain up to 60% of disease susceptibility (2–4). Several genome-wide association studies (GWAS) (with up to ~15,000 cases and ~50,000 control subjects) have been successful at identifying and replicating common variants associated with moderate effects on risk for COPD (4–10). Still, a large proportion of heritability remains unexplained (11). Although the major contributors to this “missing” heritability may be additional common variants, rare variants may explain some portion of this missing heritability (12) and, more importantly, may identify new biologic insights and be important for individual risk prediction. Indeed, in addition to *SERPINA1* (9, 13), studies have identified rare variants in cutis laxa genes (including *ELN* [14] and *FBLN5* [15]) and telomerase pathway genes (including *TERT* [16]) that increase risk of COPD.

Whole-genome sequencing (WGS) allows a comprehensive analysis of the entire genome. In contrast to array-based genotyping and imputation, the standard approach in GWAS, WGS allows investigation of rare variants that cannot be imputed well using standard reference panels. In contrast to exome sequencing, the most common current method of identifying rare variants, WGS includes important noncoding regions but also improves coverage of coding regions (17). Recently, the development and remarkable drop in cost of sequencing techniques has allowed several consortia to provide population-based WGS data (18, 19). Through the Trans-Omics in Precision Medicine (TOPMed) program, we sought to identify novel genetic determinants for COPD and identify additional variants in previously associated genomic regions that contribute to risk for this complex and heterogeneous disease.

In this article, we report a WGS study of severe COPD in two ethnic populations. We

performed an association analysis of 1,794 unrelated individuals from the Boston Early-Onset COPD Study and COPDGene cohorts, assessed the impact of these common and rare sequenced variants on risk of severe COPD, and compared the variants in the same subjects from two sources of data: sequenced and imputed.

Methods

Subject Selection

The COPDGene and Boston Early-Onset Study have been described previously (3, 20). Individuals were selected for sequencing from two populations: non-Hispanic white (NHW) and African American (AA) individuals. Case individuals had severe and very severe COPD (as defined by the Global Initiative for Chronic Obstructive Lung Disease [GOLD]: GOLD 3 and GOLD 4) with a mean (SD) FEV₁ percent predicted of 33% ($\pm 11\%$). Control subjects had normal lung function with an FEV₁ percent predicted of 101% ($\pm 11\%$) and a history of past or present smoking. The cases and control subjects were ascertained by age (young cases vs. old control subjects); therefore, an age covariate was not included in the models used in this association analysis (Table 1).

Population Structure

We used a modified principal component analysis method (21) to correct for population structure. This approach uses the Jaccard similarity on variants with low minor allele frequencies (MAFs) and hence has the advantage of recovering more recent admixture based on WGS data.

Single-Variant Analysis

Quality control procedures for genomic data are described in the data supplement. EPACTS v3.2.6 (22) was used to perform a Firth logistic regression for case/control status, correcting for pack-years of smoking and Jaccard-based principal components (JPCs) of genetic ancestry within racial group. Analysis was performed separately in NHW and AA individuals, for whom we used two JPCs. Furthermore, when the two

populations were analyzed together, we used 10 JPCs. The genome-wide significance level was set to 5×10^{-8} , owing to block correlation (linkage disequilibrium [LD]) structure of the variants. The suggestive significance level was set to $5 \times 10^{-8} < P < 5 \times 10^{-6}$. Only variants on autosomes were analyzed in this study.

Region-based Analysis

The whole genome was partitioned into nonoverlapping regions using a spatial clustering approach (23). Briefly, it is assumed that the occurrence of variants along the genome varies spatially, and these occurrences are modeled as a Poisson process. Nonoverlapping regions are formed around variants in close proximity to each other. Only groups with at least five variants were kept for this analysis. First, regions were built on the basis of all variants and tested with SKAT-RC (24). Next, regions were built by considering only rare variants (MAF ≤ 0.01). Those regions were tested using SKAT-O (25). The significance level was set to 5×10^{-8} for this analysis on the basis of Bonferroni correction.

Gene-based Analysis

Gene borders were extracted from the National Center for Biotechnology Information Reference Sequence Database, and an additional window of 50 kb was taken on either side of each gene. We used a faster implementation of the SKAT (SNP-set [Sequence] Kernel Association Test) (26) to incorporate the large number of sequenced variants within genes. Similarly to region-based analysis, we first used all variants in the defined genes and then considered only rare variants. The significance level was set to 2×10^{-6} on the basis of Bonferroni correction.

Results

Baseline characteristics of sequenced samples are shown in Table 1 and the METHODS section. Briefly, 821 cases included only GOLD 3 (severe COPD) and

Author Contributions: Study design: E.K.S. and M.H.C.; data collection: S.S.R., R.G.B., C.J.W., M.L.B., T.H.B., J.D.C., E.K.S., and M.H.C.; data quality control and analysis: D.P., P.S., M.M.P., M.-L.N.M., A.M., and M.H.C.; statistical support: D.P., H.L.F., D.Q., and C.L.; and manuscript writing: D.P., E.K.S., and M.H.C. All authors revised the manuscript.

Correspondence and requests for reprints should be addressed to Dmitry Prokopenko, Ph.D., Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115. E-mail: redpr@channing.harvard.edu.

This article has a data supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Table 1. Subject Characteristics for Whole-Genome Sequencing in the Trans-Omics in Precision Medicine Study

Population	Affection Status	Number of Individuals	Median Age (Yr) (IQR)	Median Pack-Years (IQR)	Median FEV ₁ % Predicted (IQR)
Non-Hispanic white individuals	Case	512	56.8 (9.0)	43.5 (30.5)	32.8 (18.7)
	Control	438	69.7 (5.7)	36.0 (27.0)	98.1 (14.7)
African American individuals	Case	309	60.1 (12.5)	38.3 (28.4)	36.4 (16.8)
	Control	535	55.7 (5.6)	38.5 (20.9)	99.0 (15.9)

Definition of abbreviations: IQR = interquartile range; FEV₁ = forced expiratory volume in 1 second.

GOLD 4 (very severe COPD), and 973 control subjects had normal function and a history of smoking. After variant calling and quality filtering, there were 63 million single-nucleotide variants (SNVs) and 3.7 million indels. Among these SNVs, there were 30 million singletons (i.e., variants that occurred exactly once in the entire dataset).

Single-Variant Analysis

Our primary SNV analysis combined all subjects and adjusted for ancestry using JPCs. This analysis confirmed previously reported genome-wide significant associations near the *HHIP* gene and suggestive associations in the 15q25 region (Figure 1, Table 2; see also Figures E1–E3 and Table E1 in the data supplement).

Next, we performed single-variant analysis within each population (NHW and AA). In the NHW population, we found suggestive ($5.0 \times 10^{-8} < P < 5.0 \times 10^{-6}$) associations near *HHIP*, *TNSI* (rs191759874; MAF = 0.01). In the AA population, we found suggestive

associations near *IRX5* (rs142208844; MAF = 0.01) and *SERPINA6/SERPINA1* (rs2402444; MAF = 0.088) (Figures E4 and E5 and Table 2). We have previously described associations in *SERPINA1* with severe COPD in NHW individuals due to LD with the Z allele (9). In the present study, we sought evidence that the association in AA individuals was due to well-recognized variants in *SERPINA1*. However, although rs2402444 is in strong LD with the Z allele ($D' = 1$; $R^2 = 0.06$), the association was only minimally attenuated after conditioning on the Z and S alleles in *SERPINA1* (conditional $P = 9 \times 10^{-6}$). This same variant was not significantly associated with COPD in NHW individuals (MAF = 0.206; $P = 0.52$); furthermore, this variant was imputed ($R^2 = 0.81$) in AA individuals, allowing us to test all of the COPD Gene AA population, including those with less severe (GOLD 2–4) disease. This association was less significant ($P = 0.03$), indicating that the association in our study may be limited to those with severe disease. We hypothesized that such an association

could be due to rare coding variants in strong LD with this variant among AA but not NHW individuals (27). However, we were unable to find any such rare coding variants in our WGS data. Next, we searched for noncoding variants in strong LD ($D' > 0.7$) with this variant, a higher MAF among AA individuals, and $P < 0.05$ or absent among control subjects. Of the 219 variants within 250 kb, none attenuated the signal (P value after adjusting for our best variant, rs138899312, 2.6×10^{-3}). Finally, we searched for correlation with alpha-1 concentrations using imputed genotyping data in 1,570 unrelated AA individuals from the Multi-Ethnic Study of Atherosclerosis (MESA) (28, 29) (see data supplement). Neither rs2402444 nor rs138899312 was associated with alpha-1 antitrypsin concentrations ($P > 0.8$); in addition, of the 219 variants mentioned above, only 145 were present in MESA, and of those, two had a nominally significant P value (rs191351588, $P = 0.005$; rs546321637, $P = 0.03$). Thus, despite a fairly comprehensive evaluation,

Figure 1. Manhattan plot for combined single-variant analysis of 1,794 individuals done with Firth logistic regression for case/control status, as produced by Efficient and Parallelizable Association Container Toolbox (EPACTS). Red line indicates genome-wide significance level.

Table 2. Selected Top Signals from Each Type of Analysis

Type of Analysis	Non-Hispanic White Individuals			African American Individuals			Combined		
	Top Variant, Region, or Gene	Nearest Gene	P Value	Top Variant, Region, or Gene	Nearest Gene	P Value	Top Variant, Region, or Gene	Nearest Gene	P Value
Single variant	rs1032295 (MAF = 0.371)	HHIP	6.4×10^{-8}	rs142208844 (MAF = 0.011)	IRX5	2.5×10^{-7}	rs1512281 (MAF = 0.257)	HHIP	1.6×10^{-9}
Region based, all variants	rs191759874 (MAF = 0.013)	TNS1	1.5×10^{-6}	rs2402444 (MAF = 0.088)	SERPINA6	7.4×10^{-7}	rs12914385 (MAF = 0.331)	CHRNA3	1.0×10^{-7}
	4:145432542-145434901	HHIP	1.1×10^{-7}	1:89316701-89317931	GTF2B	1.1×10^{-7}	16: 77738639-77741680	NUDT7	2.7×10^{-7}
	14:42228523-42230762	LRFN5	3.6×10^{-6}	7: 9879855-9881080	PER4	7.4×10^{-7}	5:163473111-163479934	LOC101927835	6×10^{-7}
Gene based, all variants	OPCML	OPCML	5×10^{-4}	LOC645967, NUP43	LOC645967, NUP43	1.2×10^{-4}	NUDT18	NUDT18	5.5×10^{-5}
Gene based, rare variants (MAF $\geq 1\%$)	FBXL4	FBXL4	3.7×10^{-4}	LOC645513, GTF2IP12	LOC645513, GTF2IP12	1.7×10^{-4}	KRT16P2	KRT16P2	2.5×10^{-4}

Definition of abbreviation: MAF = minor allele frequency. For additional single-variant information, see Table E1 in the data supplement. Positions are based on GRCh37/Hg19.

Figure 2. Bar plots of variant loadings in WGS and HRC-imputed data for (A) non-Hispanic white individuals and (B) African American individuals. The quality of the overlap was measured on the basis of an NRD cutoff of 5%. Common variants correspond to variants with MAF greater than 5%, low-frequency variants with MAF between 1% and 5%, and rare variants with MAF less than or equal to 1%. HRC = Haplotype Reference Consortium; MAF = minor allele frequency; NRD = nonreference discordance; WGS = whole-genome sequencing.

we were unable to find strong supportive evidence for a specific causal variant(s) for COPD in the *SERPINA1* locus.

Grouped-Variant Analysis

Aggregating variants may improve statistical power in studies of rare variants. We first performed analysis using known genes as the functional unit. We used a fast implementation of a variance component test (26). Neither combining rare and common variants together nor using only rare variants alone identified any significant associations (see the METHODS section, Table 2, and Figures E6 and E7). To aggregate signals from rare variants outside known genes in a systematic way, we partitioned the whole genome into nonoverlapping regions based on the spatial distribution of the variants; this approach has been demonstrated to result in higher power (23) (see the METHODS section and Table E2).

We used SKAT-RC (24) for all variants (rare and common) and SKAT-O (25) for rare variants only. Again, no region achieved a Bonferroni-corrected genome-wide

significance level (Table 2 and Figures E8 and E9). Our most significant association in NHW individuals was near *HHIP* ($P = 1.1 \times 10^{-7}$), which contained the top associated variant from our single variant analysis. The identified functional variant previously reported (30) is located nearby, and this was ranked as the fourth most significant regional hit ($P = 4.2 \times 10^{-7}$).

WGS versus Genotyping and Imputation

To determine the impact of WGS versus the standard approach of genotyping and imputation, we compared sequenced variants in the TOPMed WGS data of COPDGene NHW and AA individuals with sites that were genotyped and imputed using the Haplotype Resource Consortium (HRC) (Figure 2). We excluded all indels because these are not present in the HRC dataset. Overall, WGS identified 30,119,417 variants in NHW individuals and 47,280,725 in AA individuals. As expected, there were large numbers of variants unique to WGS: 14,570,118 (NHW) and 26,625,214 (AA). In the imputed dataset, there were 16,326,605 (NHW) and 21,571,470 (AA)

variants, of which a substantial number were not present in the quality-controlled WGS dataset: 777,306 NHW and 915,959 in AA (see corresponding section of the data supplement). For variants present in both datasets, we calculated the discordance between imputed data and sequencing data. The nonreference discordance (NRD) was higher for SNVs with low MAF, even when they appeared to be of high imputation quality (Figure E10). For common variants, the discordance was substantially higher in AA than in NHW individuals.

Surprisingly, we found a greater discrepancy between imputation quality and NRD for common variants located within long-range LD regions: unusually long blocks of variants in high LD with each other. In these regions, we saw a very good concordance between WGS data and imputed data, even for variants with a low imputation quality (Figure E11). In NHW individuals, of 196 variants with 0.5 less than or equal to R^2 less than or equal to 0.7 and NRD less than or equal to 0.05, 180 variants were found to be part of long-range LD regions, as defined previously (31). In AA individuals, of 105 variants with the same

Figure 3. Treemap for a list of variants in 22 previously described genome-wide association study regions in some measure of linkage disequilibrium with the lead SNP using $D' \geq 0.8$ in the same subset of individuals with sequenced and imputed genotypes for (A) non-Hispanic white individuals and (B) African American individuals. D = a subset of variants in category PF AND being within one of the six most “damaging” consequence categories as per ANNOVAR (see data supplement for definition); HRC = variants found with imputed genotypes; PF = a subset of variants having an annotation suggesting some possible functional role (see data supplement for definition); WGS = variants found with whole-genome sequencing.

parameters, there were 68 variants found in long-range LD regions.

Candidate Gene and Loci Analysis

We further investigated 22 regions found to be associated with COPD in our most recent GWAS on more than 63,000 individuals (10). We first compared overall results from genotyping and sequencing data. We compared WGS association analysis with

imputed genotype association analysis using a subset of these same individuals (1,725 subjects) (see data supplement). In general, the association statistics based on WGS were highly correlated in NHW (Pearson’s correlation chi-square = 0.91). The correlation in AA individuals was lower, 0.79, consistent with poorer coverage of non-European subjects by existing genotyping arrays and imputation methods.

In addition to discovery of novel regions, sequencing may identify rare variants in previously described regions (32) or may identify variants of interest that might be causal in a fine-mapping context. Using WGS data, we found an additional 297 nominally ($P < 0.05$) significant variants in NHW individuals and an additional 832 nominally significant variants in AA individuals, which were not previously seen in analysis of imputed variants on this same set of individuals for these 22 regions. Of those variants, 53 and 620 in NHW and AA individuals, respectively, had an MAF less than or equal to 1%. However, none of these variants was significant after Bonferroni correction for the total number of variants in the 22 regions.

We also searched for variants seen only in WGS in LD (using $D' \leq 0.8$) with the top previously reported variants or proxy ($R^2 > 0.95$) (see data supplement). Of 51,788 novel variants (among NHW individuals) and 100,588 (among AA individuals) variants, we found 9,085 (including 9,013 rare, with $MAF \leq 1\%$) in NHW individuals and 16,850 (16,744 rare, with $MAF \leq 1\%$) in AA individuals present only in the WGS data to be in LD with the top variant and to have an annotation suggesting some possible functional role (see data supplement for definition of *functional*). Among these, 495 variants in NHW individuals and 702 variants in AA individuals were annotated by ANNOVAR (33) as being within one of the six most “damaging” consequence categories (see data supplement and Tables E3 and E4). Together, these data illustrate the large number of potential COPD functional variants found by sequencing (Figure 3).

Previously, Radder and colleagues performed a WGS study on emphysema with severe airflow obstruction in 129 individuals (34) and identified suggestive associations for variants in the *ZNF816* and *PTPRO* genes. We evaluated whether these findings were replicated in our gene- and region-based association tests, but neither *ZNF816* nor *PTPRO* was significant (all $P > 0.05$) in our populations. We further looked for the four reported nonsynonymous variants in *PTPRO* by Radder and colleagues. rs61754411 gave $P = 0.013$ in our combined single-variant analysis, but this variant had an effect in the opposite (protective) direction from the

previous work. We further assessed variants in *CCDC38* and *MECOM*, which were previously shown to be associated with lung function in exome and targeted sequencing (35, 36). *CCDC38* and rs10859974 had nonsignificant *P* values ($P > 0.05$) in our aggregate and single-variant analyses, although rs10859974 in the NHW population had a higher MAF seen among control subjects, directionally consistent with the previous reports (37). Rare variants in *MECOM* ($P = 0.014$) were found to be nominally significantly associated with severe COPD in our rare variant region-based analysis.

Finally, we assessed a list of candidate genes based on Mendelian syndromes, including cutis laxa and telomerase pathway genes (Table E5). Although none of these genes survived correction for multiple hypothesis testing, we found three genes in NHW individuals (*PYCR1*, $P = 0.007$; *RTEL1*, $P = 0.05$; *RIN2*, $P = 0.03$) and two genes in AA individuals (*ALDH18A1*, $P = 0.05$; *FBLN5*, $P = 0.05$) to have a nominally significant *P* value in the gene-based analysis. We next aimed to find unique nominally significant sequenced variants within candidate Mendelian genes, adding a 50-kb window to both sides, that were missed in analysis of imputed genotypes. We found 37 nominally significant variants in NHW individuals and 228 nominally significant SNPs in AA individuals. Of these variants, 43 had an annotation suggesting possible functional evidence (see data supplement section for definition, as well as Table E6). Finally, within candidate genes, we searched for SNVs with a known relationship to existing phenotypes as reported in the ClinVar database (38). Among the Mendelian genes listed in Table E5, we found 49 variants reported in ClinVar. For three of these variants associated with diseases under an autosomal recessive inheritance pattern, no individual in our dataset was homozygous for the alternate allele. Seven variants were associated with diseases with autosomal dominant inheritance patterns, though most were of unclear clinical significance (Table E7).

Discussion

We performed a WGS study of severe COPD in NHW and AA individuals drawn

from the COPDGene and Boston Early-Onset COPD studies. We searched for novel genetic signals for severe COPD and examined previously identified genes and regions using these sequencing data. We performed single-variant and grouped-variant analyses (both genome-wide and in known candidate genes and regions). We also performed a thorough comparison of sequenced and imputed variants. This work is the largest WGS study in COPD and among the largest in chronic lung disease to date.

Our single-variant and region-based analyses confirmed associations between variants near *HHIP* and *CHRNA3* genes as influencing risk for severe COPD. Although no additional variants or regions were significant, our top-ranked associations included *TNSI*, which was previously described to be associated with lung function but not previously reported in COPD (39), and a suggestive association (for rs2402444) in AA individuals upstream of *SERPINA6* and near *SERPINA1*. Many studies have shown that variants in *SERPINA1* cause alpha-1 antitrypsin deficiency. We have previously described associations in this region with emphysema and severe COPD (9, 40, 41) driven by heterozygous associations with the Z allele. In contrast, this association was not attenuated by conditioning on the well-recognized PI*Z allele. Despite the use of sequencing data, we were unable to find conclusive evidence that this signal was driven by rare variants in or near the *SERPINA1* gene. Additional investigation in AA individuals with WGS data may be needed to detect possible rare variant contributions in controlling risk for severe COPD in this region; in general, our study suggests that the present sample size does not yield enough power to identify novel rare variants in risk for severe COPD.

Our study highlights the additional contributions and challenges of analyzing WGS data for lung disease in relation to the standard GWAS, in which the investigator must rely on reference panels to impute variants not genotyped. Although most of our association signals in single-variant analysis were common, the top hits in NHW individuals near *TNSI* and in AA individuals near *IRX5* were rare (allele frequency around 1%). We discovered 1,129 new nominally significant variants in 22 loci previously associated with COPD;

of those variants, 673 were rare. We showed that a small but not insignificant number of common variants and most rare variants (66%) could not be identified when relying on existing imputation panels, and this problem is dramatically worse in AA samples, which not only harbor more rare genetic variation but also have less representation on current imputation panels. Furthermore, imputed rare variants in the same set of individuals show poor concordance with sequenced variants found by WGS, even when filtering for imputed variants of highest quality. Imputation works much better on variants with higher MAF, as expected. One interesting observation was related to common variants in regions of long-range LD. We found that poorly imputed common variants tend to show a very good concordance with the sequenced variants in those regions. This implies that imputation quality statistics might be misleading for these regions of long-range LD. We also found a small percentage of variants present in the imputed data but completely absent in the WGS dataset. This could indicate that some proportion of variants is being missed by current sequencing technologies or being excluded by current quality control approaches.

WGS also identified rare variants of potential pathogenic significance not identified by GWAS and imputation. We could not find strong evidence of potential association for candidate genes in the cutis laxa and telomerase pathways; however, lack of functional knowledge of individual variants could contribute to a lack of association in these gene-based tests. We also were not able to replicate recently described associations based on sequencing data; for example, we found no association between markers in the *PTPRO* and *ZNF816* genes, identified on the basis of WGS data. The top SNP in *PTPRO* reported by Radder and colleagues (34) was nominally significant in our combined single-variant analysis, but it displayed a protective effect on risk of severe COPD, opposite to that previously reported. Potential reasons for these null findings might include lack of power to replicate other studies with our modest sample size in this WGS study, differences in study populations, or that the reported findings were false-positive results.

Our study has several limitations. First, although the largest WGS study of COPD to date to our knowledge, our sample size was limited compared with studies using genotyping, such as those in recent meta-analysis of multiple GWASs (8, 10). We had two different populations, which further reduced our sample size to detect population-specific genes and variants controlling risk for COPD. Our AA population was not only smaller but also more diverse; recently, others have argued that owing to increased genetic diversity, a more stringent significance-level cutoff should be used in genomes with African ancestry (42). Thus, an even greater effort to expand AA samples is likely necessary to achieve comparable power. We used a spatial clustering algorithm to group our variants for region-based testing, which allowed us to reduce the multiple testing burden; however, the optimal method to test for rare variants is still not clear, and one issue for nearly all rare variant testing

methods is the inclusion of too many variants exerting no effect on the phenotype, which reduces their power.

Although GWAS using genotyping arrays have been the most common tool for identifying genetic associations for complex diseases such as COPD, WGS provides a comprehensive assessment of the entire genome. Our findings demonstrate the utility of WGS for variant identification, particularly for rare variants and in non-European samples. The best strategy for genetic association studies will depend on the nature of the variants (rare vs. common) and the population studied. The advent of TOPMed and other larger WGS efforts, which include large numbers of non-European populations, will allow the construction of better imputation reference panels. Thus, we expect that genotyping arrays will continue to be used for studies of common variants and leveraging these panels for increasingly accurate imputation for rarer variants.

However, studies of rare or poorly imputed variants and in unique populations, as well as studies that require comprehensive variant identification such as fine mapping, will benefit from sequencing. In addition, with the drop in sequencing price, technological improvements, and increases in sequencing quality, WGS may eventually obviate the need for genotyping and imputation. Our study also highlights the importance of larger sample sizes in sequencing studies and the need to improve tools for functional annotation of the noncoding genomic regions to allow maximal use of WGS data and to discover new associations with COPD. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors gratefully acknowledge the study researchers and participants who provided biological samples and data for TOPMed.

References

- Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med* 2013;187:347–365.
- Ingebrigtsen T, Thomsen SF, Vestbo J, van der Sluis S, Kyvik KO, Silverman EK, *et al.* Genetic influences on chronic obstructive pulmonary disease – a twin study. *Respir Med* 2010;104:1890–1895.
- Silverman EK, Chapman HA, Drazen JM, Weiss ST, Rosner B, Campbell EJ, *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;157:1770–1778.
- Cho MH, McDonald MLN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, *et al.*; NETT Genetics, ICGN, ECLIPSE and COPDGene Investigators. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med* 2014;2:214–225.
- Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, *et al.* Variants in *FAM73A* are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;42:200–202.
- Cho MH, Castaldi PJ, Wan ES, Siedlinski M, Hersh CP, Demeo DL, *et al.*; ICGN Investigators; ECLIPSE Investigators; COPDGene Investigators. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet* 2012;21:947–957.
- Kim WJ, Lee SD. Candidate genes for COPD: current evidence and research. *Int J Chron Obstruct Pulmon Dis* 2015;10:2249–2255.
- Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, *et al.*; Understanding Society Scientific Group; Geisinger-Regeneron DiscovEHR Collaboration. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet* 2017;49:416–425.
- Busch R, Hobbs BD, Zhou J, Castaldi PJ, McGeachie MJ, Hardin ME, *et al.*; National Emphysema Treatment Trial Genetics; Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-Points; International COPD Genetics Network; COPDGene Investigators. Genetic association and risk scores in a chronic obstructive pulmonary disease meta-analysis of 16,707 subjects. *Am J Respir Cell Mol Biol* 2017;57:35–46.
- Hobbs BD, de Jong K, Lamontagne M, Bossé Y, Shrine N, Artigas MS, *et al.*; COPDGene Investigators; ECLIPSE Investigators; LifeLines Investigators; SPIROMICS Research Group; International COPD Genetics Network Investigators; UK BiLEVE Investigators; International COPD Genetics Consortium. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat Genet* 2017;49:426–432.
- Zhou JJ, Cho MH, Castaldi PJ, Hersh CP, Silverman EK, Laird NM. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am J Respir Crit Care Med* 2013;188:941–947.
- Mandolia TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, *et al.* Finding the missing heritability of complex diseases. *Nature* 2009;461:747–753.
- Silverman EK, Sandhaus RA. Alpha₁-antitrypsin deficiency. *N Engl J Med* 2009;360:2749–2757.
- Kelleher CM, Silverman EK, Broekelmann T, Litonjua AA, Hernandez M, Sylvia JS, *et al.* A functional mutation in the terminal exon of elastin in severe, early-onset chronic obstructive pulmonary disease. *Am J Respir Cell Mol Biol* 2005;33:355–362.
- Brandsma CA, van den Berge M, Postma DS, Jonker MR, Brouwer S, Paré PD, *et al.* A large lung gene expression study identifying fibulin-5 as a novel player in tissue repair in COPD. *Thorax* 2015;70:21–32.
- Stanley SE, Chen JJL, Podlevsky JD, Alder JK, Hansel NN, Mathias RA, *et al.* Telomerase mutations in smokers with severe emphysema. *J Clin Invest* 2015;125:563–570.
- Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 2015;112:5473–5478.
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;526:68–74.
- McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, *et al.*; Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–1283.

20. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, *et al.*; COPDGene Investigators. Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2010;7:32–43.
21. Prokopenko D, Hecker J, Silverman EK, Pagano M, Nöthen MM, Dina C, *et al.* Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics* 2016;32:1366–1372.
22. EPACTS. Available from: <https://genome.sph.umich.edu/wiki/EPACTS>.
23. Loehlein Fier H, Prokopenko D, Hecker J, Cho MH, Silverman EK, Weiss ST, *et al.* On the association analysis of genome-sequencing data: a spatial clustering approach for partitioning the entire genome into nonoverlapping windows. *Genet Epidemiol* 2017;41:332–340.
24. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* 2013;92:841–853.
25. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, *et al.*; NHLBI GO Exome Sequencing Project—ESP Lung Project Team. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;91:224–237.
26. Lumley T, Brody J, Peloso G, Morrison A, Rice K. FastSKAT: sequence kernel association tests for very large sets of markers. *Genet Epidemiol* [online ahead of print] 22 Jun 2018; DOI: 10.1002/gepi.22136.
27. Thun GA, Imboden M, Ferrarotti I, Kumar A, Obeidat M, Zorzetto M, *et al.* Causal and synthetic associations of variants in the *SERPINA* gene cluster with alpha1-antitrypsin serum levels. *PLoS Genet* 2013;9:e1003585.
28. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol* 2002;156:871–881.
29. Rodriguez J, Jiang R, Johnson WC, MacKenzie BA, Smith LJ, Barr RG. The association of pipe and cigar use with cotinine levels, lung function, and airflow obstruction: a cross-sectional study. *Ann Intern Med* 2010;152:201–210.
30. Zhou X, Baron RM, Hardin M, Cho MH, Zielinski J, Hawrylkiewicz I, *et al.* Identification of a chronic obstructive pulmonary disease genetic determinant that regulates *HHIP*. *Hum Mol Genet* 2012;21:1325–1335.
31. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, *et al.* Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 2008;83:132–135; author reply 135–139.
32. Mousas A, Ntritsos G, Chen MH, Song C, Huffman JE, Tzoulaki I, *et al.* Rare coding variants pinpoint genes that control human hematological traits. *PLoS Genet* 2017;13:e1006925.
33. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
34. Radder JE, Zhang Y, Gregory AD, Yu S, Kelly NJ, Leader JK, *et al.* Extreme trait whole-genome sequencing identifies *PTPRO* as a novel candidate gene in emphysema with severe airflow obstruction. *Am J Respir Crit Care Med* 2017;196:159–171.
35. Wain LV, Sayers I, Soler Artigas M, Portelli MA, Zeggini E, Obeidat M, *et al.* Whole exome re-sequencing implicates *CCDC38* and cilia structure and function in resistance to smoking related airflow obstruction. *PLoS Genet* 2014;10:e1004314.
36. Artigas MS, Wain LV, Shrine N, McKeever TM, UK BiLEVE, Sayers I, *et al.*; Lung GO; NHLBI Exome Sequencing Project; COPDGene Investigators. Exome sequencing analysis in severe, early-onset chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2016;193:1353–1363.
37. Qiao D, Lange C, Beaty TH, Crapo JD, Barnes KC, Bamshad M, *et al.*; Lung GO; NHLBI Exome Sequencing Project; COPDGene Investigators. Exome sequencing analysis in severe, early-onset chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2016;193:1353–1363.
38. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;44:D862–D868.
39. Soler Artigas M, Wain LV, Repapi E, Obeidat M, Sayers I, Burton PR, *et al.*; SpiroMeta Consortium. Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. *Am J Respir Crit Care Med* 2011;184:786–795.
40. Cho MH, Castaldi PJ, Hersh CP, Hobbs BD, Barr RG, Tal-Singer R, *et al.*; NETT Genetics, ECLIPSE, and COPDGene Investigators. A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *Am J Respir Crit Care Med* 2015;192:559–569.
41. Foreman MG, Wilson C, DeMeo DL, Hersh CP, Beaty TH, Cho MH, *et al.*; Genetic Epidemiology of COPD (COPDGene) Investigators. Alpha-1 antitrypsin PIMZ genotype is associated with COPD in two racial groups. *Ann Am Thorac Soc* 2017;14:1280–1287.
42. Pulit SL, de With SAJ, de Bakker PIW. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genet Epidemiol* 2017;41:145–151.