A Genome-Wide Association Study in Hispanics/Latinos Identifies Novel Signals for Lung Function. The Hispanic Community Health Study/Study of Latinos

Kristin M Burkart, MD, MSc1*, Tamar Sofer, PhD2*, Stephanie J. London, MD, DrPH3, Ani Manichaikul, PhD4, Fernando P Hartwig, MSc5, Qi Yan, PhD6, María Soler Artigas, PhD7, Lydiana Avila, MD8, Wei Chen, PhD6, Sonia Davis Thomas, DrPH9, Alejandro A. Diaz, MD, MPH10, Ian P Hall, Prof.11, Bernardo L Horta, MD, PhD5, Robert C Kaplan, PhD12, Cathy C. Laurie, PhD13, Ana M Menezes, MD, PhD5, Jean V. Morrison14, Elizabeth C Oelsner, MD, MPH1, Deepa Rastogi, MBBS, MS15, Stephen S Rich, PhD4, Manuel Soto-Quiros, PhD6, Adrienne M. Stilp, PhD13, Martin D Tobin, Prof., PhD7.16, Louise V Wain, PhD7, Juan C. Celedón, MD, DrPH6*, R. Graham Barr, MD, DrPH1,17*

*These authors contributed equally to this work

Author’s affiliations:

1 Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY

2 Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Division of Sleep and Circadian Disorders, Boston, MA

Department of Biostatistics, University of Washington School of Public Health, Seattle, WA

3 Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC
4 Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA
5 Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil
6 Division of Pulmonary Medicine, Allergy and Immunology, Department of Pediatrics, Children’s Hospital of Pittsburgh of UPMC, University of Pittsburgh, Pittsburgh, PA
7 Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, UK
8 Division of Pediatric Pulmonology, Hospital Nacional de Niños, San José, Costa Rica
9 Department of Biostatistics, University of North Carolina, Chapel Hill, NC
10 Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA
11 Division of Respiratory Medicine, Queen’s Medical Centre, University of Nottingham, Nottingham NG7 2RD, UK
12 Department of Epidemiology & Population Health, Albert Einstein College of Medicine, New York, NY
13 Department of Biostatistics, University of Washington School of Public Health, Seattle, WA
14 Department of Human Genetics and Statistics, University of Chicago, Chicago, IL
15 Department of Pediatrics, Albert Einstein College of Medicine, New York, NY
16 National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, UK
17 Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY
Correspondence:  
Kristin M Burkart, MD, MSc  
Columbia University Medical Center  
630 West 168th Street, PH 8 East - Room 101  
New York, NY 10032  
Fax: 212-305-8464  
E-mail: kb2319@cumc.columbia.edu

Word Count: 3,464  
Abstract Word Count: 250  
Short title: Genome-wide study of lung function in Hispanics/Latinos

Funding: HCHS/SOL: We thank the participants and staff of the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) for their contributions to this study. The baseline examination of HCHS/SOL was carried out as a collaborative study supported by contracts from the NHLBI to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following institutes, centers, and offices contributed to the first phase of 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 (NIDCR), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Neurological Disorders and Stroke, and NIH Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). Additional analysis support was provided by 1R01DK101855-01and13GRNT16490017. Genotyping efforts were supported by the NIH Department of Health and Human Services (HSN26220/20054C), National Center for Advancing Translational Science.
Clinical Translational Science Institute (UL1TR000124), and NIDDK Diabetes Research Center (DK063491). This manuscript has been reviewed by the HCHS/SOL Publications Committee for scientific content and consistency of data interpretation with previous HCHS/SOL publications. TS is supported by NHLBI 1R35HL135818. AAD is supported by K01HL118714 and the Brigham and Women’s Hospital Minority Faculty Career Development Award. SJL is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. RGB is supported by R01-HL077612 and R01-HL092081.

COPD-Costa Rica, MESA Lung, 1982 Pelotas Birth Cohort Study, and SpiroMeta:
See online supplement

AUTHOR CONTRIBUTION

Study design: KMB, JCC, RGB

Data collection: FPH, IPH, BLH, RCK, LA, DR, SSR, MS-Q, AMS, AMM, MDT, RCK, SJL, JCC, RGB

Data analysis: TS, AM, WC, FPH, MSA, LVW

Obtaining funding: RCK, CCL, SSR, RGB

Drafting Manuscript: KMB

Critical revision of manuscript: KMB, TS, AM, WC, FPH, MSA, LA, AD, IPH, BLH, RCK, CCL, SJL, AMM, JVM, ECO, DR, SSR, MS-Q, AMS, MDT, LVW, QY, JCC, RGB
AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

Lung function and chronic obstructive pulmonary disease (COPD) are heritable traits. Prior genome-wide association studies (GWAS) of lung function have identified numerous genetic variants associated with COPD risk; however, much of the individual variance in lung function remains unexplained. Furthermore, most of these studies have been limited to participants of European ancestry. Including multi-ethnic populations in GWAS research may identify novel variants that contribute to the etiology of lung function.

What this Study Adds to the Field

This study, the largest GWAS of lung function and COPD to exclusively include Hispanic/Latino participants, identified eight novel signals of which three replicated in independent populations. A novel locus for FEV₁ (rs4791658; ZSWIM7) replicated in a cohort of European ancestry. A locus for FEV₁/FVC (rs145174011; HAL); in a previously identified region for FEV₁/FVC in individuals of European ancestry and for percent emphysema in a Hispanic population remained independent in conditional analyses but did not replicate. Admixture mapping identified a novel region associated with Amerindian ancestry and FEV₁ which included a candidate variant (rs4133185) in the AGMO gene that replicated. Additionally, we identified a SNP (rs7709630) for COPD, which replicated in individuals of European ancestry. Several loci previously identified in European samples were also associated with lung function traits among Hispanic/Latino participants in HCHS/SOL at the genome-wide significance level. These results emphasize the importance of including admixed populations when performing genetic
studies to identify variants that may contribute to the genetic etiology of pulmonary function and COPD.
ABSTRACT

Rationale: Lung function and chronic obstructive pulmonary disease (COPD) are heritable traits. Genome-wide association studies (GWAS) have identified numerous pulmonary function and COPD loci, primarily in cohorts of European ancestry.

Objectives: Perform a GWAS of COPD-phenotypes in Hispanic/Latino populations to identify loci not previously detected in European populations.

Methods: GWAS of lung function and COPD in Hispanic/Latino participants from a population-based cohort. We performed replication studies of novel loci in independent studies.

Measurements and Main Results: Among 11,822 Hispanic/Latino participants, we identified eight novel signals; three replicated in independent populations of European Ancestry. A novel locus for forced expiratory volume in one second (FEV₁) in ZSWIM7 (rs4791658; p=4.99×10⁻⁹) replicated. A rare variant (MAF=0.002) in HAL (rs145174011) was associated with FEV₁ to forced vital capacity (FEV₁/FVC) (p=9.59×10⁻⁹) in a region previously identified for COPD-related phenotypes; it remained significant in conditional analyses but did not replicate. Admixture mapping identified a novel region, with a variant in AGMO (rs41331850), associated with Amerindian ancestry and FEV₁, which replicated. A novel locus for FEV₁ identified among ever smokers (rs291231; p=1.92×10⁻⁸) approached statistical significance for replication in admixed populations of African ancestry and a novel SNP for COPD in PDZD2 (rs7709630; p=1.56×10⁻⁸) regionally replicated. Additionally, loci previously identified for lung function in European samples were associated in Hispanic/Latino participants in HCHS/SOL at the genome-wide significance level.
**Conclusions:** We identified novel signals for lung function and COPD in a Hispanic/Latino cohort. Including admixed populations when performing genetic studies may identify variants contributing to genetic etiologies of COPD.

**Word Count:** 250

**Key Words:** Hispanic/Latino, genome-wide association study, single nucleotide polymorphisms, lung function, airflow limitation, chronic obstructive pulmonary disease
INTRODUCTION

The Hispanic/Latino population is the largest and fastest-growing minority population in the United States (U.S.), accounting for 17% of the U.S. population currently and an estimated 33% by 2060(1). The Hispanic/Latino population in the U.S. is genetically diverse, with a mixture of European, African and Amerindian genetic ancestries(2-4).

Chronic obstructive pulmonary disease (COPD), characterized by persistent airflow obstruction(5, 6), is the third-leading cause of death in the U.S.(7). The prevalence of COPD in some Hispanic groups is lower than in non-Hispanic whites and African-Americans in the U.S.(8-10). In the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), the largest and best characterized cohort of Hispanics/Latinos in the US, the prevalence of COPD varied substantially by Hispanic/Latino heritage with Puerto Ricans and Cubans having a much higher prevalence of COPD than other Hispanic/Latino groups(10). This difference in HCHS/SOL was explained by smoking and asthma history; however, other investigators have noted that Hispanic ethnicity is inversely associated with COPD-phenotypes compared to whites(11, 12) and have hypothesized that genetic differences may lower risk of COPD in some Hispanic groups(11, 13).

Examination of genetic risk among non-European populations may reveal novel variants that yield new pathways to treatments, such as PCSK9(14). We are aware of only two genome-wide association studies (GWAS) of lung function or COPD-related phenotypes that analyzed Hispanics separately(13, 15). One identified a novel variant near MAN2B1 that was associated with percent emphysema on computed tomography (CT)(16) and the other identified two loci that approached genome wide statistical
significance for COPD(13). This literature contrasts with that among persons of European ancestry, in whom GWAS have identified multiple loci in genes for lung function and COPD(17-31)

To improve our current understanding of the genetic architecture of lung function-related traits in Hispanics/Latinos, we performed a meta-analysis for FEV₁, FEV₁/FVC, airflow limitation and COPD among six Hispanic/Latino groups in the HCHS/SOL cohort(10, 32). Findings were replicated in European, Hispanic and African populations. Some of the results have been previously reported in abstract form(33).

METHODS

Study Sample

HCHS/SOL is a community-based cohort study of 16,415 self-identified Hispanic/Latino persons aged 18-74 years recruited from four U.S. communities. The study design, cohort recruitment(10, 32), and baseline clinical examination(34) have been previously described. Institutional Review Boards at each field center approved study protocols and written informed consent was obtained from all participants.

For genetic analyses, HCHS/SOL participants were classified in six genetic analysis groups: Central American, Cuban, Dominican, Mexican, Puerto Rican, and South American. The genetic ancestry structure of these groups, including principal components (PCs) plots for all individuals, have been previously reported(35). Henceforward we will refer to the six genetic analysis groups as Hispanic/Latino ancestry groups.
Exclusion criteria were lack of valid spirometric or genetic data, missing covariates, near exclusive Asian ancestry and participants who were not classified in the Hispanic/Latino ancestry groups. Details are provided in the online supplement.

**Phenotypic and Outcome Measures**

Spirometry was conducted in accordance with American Thoracic Society/European Respiratory Society guidelines (36) using a dry rolling sealed spirometer as previously described (10). Participants with pre-bronchodilator FEV₁/FVC ratios less than 70% or the lower limit of normal were selected for post-bronchodilator spirometry, the former defined by NHANES reference equations (37) since those from HCHS/SOL (38) were not available during data collection.

Airflow limitation was defined as a pre-bronchodilator FEV₁/FVC ratio < 70% and COPD as a post-bronchodilator FEV₁/FVC ratio < 70% (5). Controls had normal lung function defined as a pre-bronchodilator FEV₁/FVC ratio > 70% and FVC > 80 percent predicted.

**Genotyping, Quality Control and Imputation**

Consenting HCHS/SOL subjects were genotyped at Illumina on the HCHS/SOL custom 15041502 B3 array. We applied standard quality-assurance and quality-control methods (39) as previously described (35). Genome-wide imputation was carried out with the full, cosmopolitan 1000 Genomes Project phase 1 reference panel (40); associations for novel loci were confirmed with 1000 Genome phase 3 imputation panel. Details are
provided in the online supplement.

Statistical Analyses

See the online supplement for details.

Quantitative Lung Function Analyses

Analyses of FEV₁ and FEV₁/FVC ratio employed linear mixed models, stratified by Hispanic/Latino ancestry group, and adjusted for age, age², sex, height, height², study center, smoking status, pack-years, sampling weights, and the first 5 PCs, as fixed effects. We used random effects for genetic relatedness (kinship), and household and community (block unit) to account for environmental correlation. Results from each Hispanic/Latino ancestry group were then meta-analyzed using the MetaCor method(41). Genome-wide significance threshold was defined as p<5x10⁻⁸, the Bonferroni adjustment for 1 million independent tests (42). A priori, we planned additional meta-analyses stratified by never vs ever smoking.

Analyses for Airflow Limitation and COPD

We analyzed airflow limitation and COPD in pooled analyses implementing the GMMAT software(43) to fit a logistic mixed model adjusting for age, sex, study center, smoking status, pack-years, sampling weights, the first 5 PCs and Hispanic/Latino ancestry groups, as fixed effects and with random effects for kinship and block unit. Participants younger than 45 years were excluded.

Admixture Mapping and Analyses
Local ancestry estimates were previously inferred in the HCHS/SOL(44). A genome-wide admixture mapping scan was performed using a linear mixed model, as described above, testing European, African, and Amerindian ancestries at each available local ancestry interval (LAI). Statistical significance for admixture mapping was set at p<5.68x10^{-5} based on previous simulation results for HCHS/SOL(45). After discovering a genome-wide significant LAI association (at the admixture mapping level), we identified a candidate variant and performed conditional admixture mapping analysis(46).

*Replication of Lung Function SNPs*

We pursued replication of novel loci associated with FEV\textsubscript{1} and FEV\textsubscript{1}/FVC in populations with Hispanics(13, 47-51), African-American(47, 48), and European representation(17). We performed a look-up in each individual study then meta-analyzed across studies. For airflow limitation and COPD, we performed regional replication in publically available GWAS results of airflow obstruction in participants of European ancestry(23). For admixture mapping signals, we performed a look-up in the UK Biobank results for individuals of European ancestry (https://sites.google.com/broadinstitute.org/ukbbgwasresults).

*Generalization analyses of previously reported SNPs*

We looked-up previously reported SNPs, their effects, standard errors and p-value in HCHS/SOL(52).
RESULTS

The mean age of the 11,822 HCHS/SOL participants with valid lung function and genetic data was 46±14 years. Thirty-nine percent had ever smoked cigarettes with median pack-years of 7.5. The COPD analysis included 363 and 5,253 individuals with and without COPD, respectively. Characteristics of the participants are shown in table 1. Manhattan and Quantile-Quantile (Q-Q) plots for FEV$_1$, FEV$_1$/FVC, airflow limitation and COPD are shown in figures E1, E2, E3 and E4. Genomic inflation factors ranged from 1.020–1.026 in quantitative lung function analyses and 0.988–1.011 in analyses of COPD.

GWAS of Lung Function across all Hispanic/Latino ancestry groups

Across all Hispanic/Latino ancestry groups, seven signals achieved genome-wide statistical significance, of which two were novel for lung function (table 2). The lead novel SNP, rs4791658, in the locus associated with FEV$_1$ (figure 1a) was an intron variant in the gene ZSWIM7 on chromosome 17 (Beta=33.4, p=4.99×10$^{-9}$). Its minor allele frequency (MAF) ranges from 0.39–0.49 across Hispanic/Latino ancestry groups (figure 1b). SNP, rs4791658 neared genome-wide significance for FEV$_1$ among ever smokers (Beta=46.8, p=1.53×10$^{-6}$). Effect size and p-values for rs4791658 across all lung function and COPD-related traits are shown in Table E1.

The lead novel SNP, rs145174011, in the locus associated with FEV$_1$/FVC (figure 1c) was an intronic variant in the gene HAL on chromosome 12 (Beta=4.86, p=9.59×10$^{-9}$). The SNP is rare with a MAF that ranges from 0.001–0.008 across Hispanic/Latino ancestry groups (figure 1d). The minor allele does not exhibit an outlier
effect (figure E5). Sensitivity analysis using rank-normalized residuals is shown in supplement. This locus is near the gene CCDC38, which was previously identified for FEV₁/FVC in individuals of European Ancestry (17) and the gene SNRPF, which was previously identified for percent emphysema on CT in racially/ethnically diverse participants (16). In conditional analysis adjusting for rs1036429 (CCDC38) and rs7957346 (SNRPF) in HCHS/SOL, rs145174011 (HAL SNP) remained associated with FEV₁/FVC (Beta=4.92 p=6.4×10⁻⁹), suggesting its association with FEV₁/FVC is independent of previously identified variants. The findings for rs145174011 among ever and never smokers are concordant and trended towards significance (table E1). Regional association plots and forest plots of previously reported genome wide significant associations for FEV₁ and FEV₁/FVC meta-analyzed across all Hispanic/Latino ancestry groups are shown in figure E6.

Stratification by Smoking Status

In the meta-analysis across all Hispanic/Latino ancestry groups stratified by smoking status, we identified one novel locus among ever smokers for FEV₁ and one novel locus among never smokers for FEV₁/FVC (table 3). The lead variant in the locus associated with FEV₁ in ever smokers was an Indel variant (p=1.66×10⁻⁸) on chromosome 11. The top SNP, rs291231 (Beta=57.28, p=1.92×10⁻⁸), in this locus has a MAF that ranges from 0.28 – 0.42 across Hispanic/Latino ancestry groups. This SNP lies between the genes EED and CCDC81 (figure 2a and 2b). The SNP-by-pack-years interaction trended towards significance (p-interaction=0.06) and the genotype effect became greater at higher pack-years (figure E7).
The top SNP (rs7228593) in the locus associated with FEV₁/FVC in never smokers (Beta=1.34, p=3.47×10⁻⁸) is located between the genes SMIM21 and LCOC339298 (figure 2c) on chromosome 18. The findings in never smokers for airflow limitation (OR=0.46, p=2.63×10⁻⁴) and COPD (OR=0.36, p=0.005) were directionally concordant and trended towards significance (table E1). Its MAF varies significantly across different Hispanic/Latino ancestry groups (figure 2d) with the highest in participants from the Dominican ancestry group (MAF=0.16) and lowest in Mexican ancestry group (MAF=0.02). Forest plots and regional association plot of remaining genome-wide significant associations for FEV₁ and FEV₁/FVC stratified by smoking status are shown in figure E8.

GWAS of COPD
We identified two previously unreported SNPs for COPD (table E2 and figure E9). SNP rs7709630 is an intron in the PDZD2 gene on chromosome 5 (p=1.56×10⁻⁸; MAF=0.11) and SNP rs2286351 is a noncoding transcript variant in CDRT15P1 on chromosome 17 (p=1.97×10⁻⁸, MAF=0.09). Lung function association analyses of these COPD SNPs are reported in table E3.

Admixture Mapping:
Manhattan plots from admixture mapping analyses for FEV₁, and FEV₁/FVC testing Amerindian, African, and European ancestry counts individually in LAIs are shown in figures E10 and E11. We identified one novel significant local ancestry association
region of Amerindian ancestry for FEV\textsubscript{1} (figure E10a) on chromosome 7 (Beta = -37.51, p=2.86x10\textsuperscript{-6}). There were no significant associations in the European versus others or African versus others analyses.

We identified a candidate variant (rs4133185) within this novel LAI region (Beta = -31.44, p= 7.43x10\textsuperscript{-7}; table E4 and figure 3) with considerable differences in ancestry-specific allele frequencies (figure 4). The lead SNP, rs4133185, in this candidate variant was an intron variant in the gene \textit{AGMO}. The admixture mapping signal is less significant after conditional analyses adjusting for rs4133185 (table E5, figures 4 and E12).

\textit{Replication}

Characteristics of the participants in the replication cohorts are shown in table E4. Of the two novel loci identified for lung function, only the FEV\textsubscript{1} locus on chromosome 17 successfully replicated (replication p=3.37x10\textsuperscript{-5}; adjusted threshold for statistical significance was p<9.2x10\textsuperscript{-4}) but only in the SpiroMeta consortium look-up (n=94,612)(17). The candidate variant (rs4133185), identified through admixture mapping, associated with Amerindian ancestry and FEV\textsubscript{1} replicated in UK Biobank (Beta = -8.3, p=0.004). SNP rs291239, a genotyped SNP in the novel locus on chromosome 11 identified for FEV\textsubscript{1} among ever smokers, approached statistical significance for replication in the meta-analysis across all cohorts with admixed populations of African ancestry (replication p=0.007); however, the direction of effect was discordant in the individual replication cohorts (table E7).
The novel SNP (rs7709630; PDZD2) associated with COPD regionally replicated in publically available results from the SpiroMeta-CHARGE GWAS meta-analysis of airflow obstruction among participants of European ancestry(23). The smallest p-value in the replication region was for rs409839 (p=8.9x10^{-4}) and the adjusted threshold for statistical significance was p< 4.2x10^{-3}. The other loci were not successfully replicated.

**Generalization Analyses of Previously Identified Lung Function Loci**

Using the FDR adjusted look-up threshold, several of the previously reported SNPs for FEV₁, FEV₁/FVC and COPD identified in European populations were generalized to Hispanic/Latino populations (tables E8, E9, E10 and Figure E13). See online supplement for details.

**DISCUSSION**

In this GWAS of 11,822 Hispanic/Latino HCHS/SOL participants, we identified eight novel signals, of which three replicated in independent cohorts of European ancestry. Three novel signals were associated with lung function. A novel locus in ZSWIM7 (rs4791658) was associated with FEV₁ replicated and a rare variant in HAL (rs145174011) was associated with FEV₁/FVC did not replicate; it is in a previously identified region but remained significant in conditional analyses, suggesting an independent effect. Admixture mapping identified a novel LAI region, with a candidate variant in the AGMO gene (rs4133185), associated with Amerindian ancestry and FEV₁, which replicated. In smoking stratified analyses, a novel locus (rs291231) associated with FEV₁ among 4,595 ever smokers approached statistical significance for replication
in cohorts with admixed populations of African ancestry and a novel SNP for COPD in *PDZD2* (rs7709630) regionally replicated in individuals of European ancestry. In addition, we confirmed loci previously identified in European samples [*KCNE2*(17) and *GPR126*(27)] for FEV\(_1\)/FVC in this Hispanic/Latino sample.

Several SNPs in *ZSWIM7* were associated with FEV\(_1\) at genome-wide significance and replicated in independent cohorts of European ancestry. *ZSWIM7* mutations were associated colorectal adenomatous polyposis(53) and in a GWAS of SNP-by-smoking interaction, *ZSWIM7* neared significance for systolic blood pressure(54). *ZSWIM7* is a highly conserved eukaryotic regulator of homologous recombination and plays an important role in error-free DNA repair processes for DNA double-strand breaks (DSBs)(55-57). Unrepaired DSBs trigger cell senescence, apoptosis, and pro-inflammatory responses, all of which are established mechanisms in the pathogenesis of COPD(58-60). Oxidative stress and cigarette smoke cause DSBs(61, 62) and DSBs have been implicated in the pathogenesis of COPD(62). Furthermore, COPD patients have a greater number of unrepaired DSBs which are associated with a higher expression of markers for senescence, apoptosis, and pro-inflammatory phenotypes compared to asymptomatic smokers and nonsmokers(62). In our study, the effect size of *ZSWIM7* for FEV\(_1\) was of greater magnitude among ever smokers which is consistent with known effects of cigarette smoke causing DSBs. Thus, *ZSWIM7* through its regulation of homologous recombination and role in error-free DNA repair processes for DSBs is relevant to the pathogenesis of COPD.

A rare variant in *HAL* (rs145174011) was associated with FEV\(_1\)/FVC among all participants. The minor allele was associated with decreases in FEV\(_1\)/FVC of 4.9% and
6.3% (all participants and ever smokers, respectively), and with decreases in FEV₁ of 100 mL and 173 mL (all participants and ever smokers, respectively). It is possible that this variant was not successfully replicated due to lack of power, given that it is a rare variant in African and Hispanic populations and is not polymorphic in European populations (63).

The HAL SNP (rs145174011) is near two previously identified loci for pulmonary traits (CCDC38 and SNRPF). CCDC38 has been associated with FEV₁/FVC in European ancestry individuals (17), and SNRPF has been associated with percent emphysema on CT in racially/ethnically diverse participants in the Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study (16). In the MESA Lung Study, linkage disequilibrium (LD) between SNRPF and CCDC38 accounted for much of the effect seen in Whites, and less so in African Americans and Chinese; however, the observed effect of SNRF on percent emphysema on CT in Hispanics appeared to be independent from CCDC38 (16). In our study, the association of SNP rs145174011 (HAL locus) with FEV₁/FVC was maintained in analyses conditioning on previously reported loci. Collectively, these data suggest there are three independent loci associated with COPD-related phenotypes in this region.

Although the HAL locus contained numerous SNPs, they were all imputed rare variants and did not replicate; therefore, further investigation, such as whole-genome sequencing of this region and functional studies, is required to refine these loci, confirm association, and understand their biological implications for COPD.

The HAL gene encodes the enzyme histidine ammonia-lyase (64, 65) - the first step in histidine catabolism (66). Low levels of histidine have been associated with
increased inflammation and oxidative stress(67)(68) and decreased concentrations of histidine were associated with advanced COPD and emphysema(69). Recently, whole-exome sequencing in African-Americans identified three rare loss-of-function variants in \textit{HAL} that were associated with increased histidine levels and were inversely related to coronary heart disease risk in African-American and European-American populations. Thus, a functional \textit{HAL} variant may decrease histidine levels and increase susceptibility to COPD and emphysema through its effects on oxidative stress.

The region associated with Amerindian ancestry and FEV\textsubscript{1} on chromosome 7 contained a candidate variant (rs4133185) within the \textit{AGMO} gene. This SNP, rs4133185, has large differences in ancestry-specific allele frequencies (~0.8 for Amerindian and ≤0.2 for European and African) and is associated with obstructive physiology, as measured by FEV\textsubscript{1} (p=7.43x10\textsuperscript{-7}). In the conditional analysis, rs4133185 explains part but not all of the admixture signal. SNP rs4133185 may not be the causal variant and the presence of multiple SNPs within the LAI region may explain the residual signal. A prior report showed an association of \textit{AGMO} with FVC in a multi-ethnic population(70). Future whole-genome sequencing studies are needed for better fine mapping of this region.

The chromosome 11 locus (rs291231) associated with FEV\textsubscript{1} among ever smokers nominally replicated in the meta-analysis across all cohorts with admixed populations of African ancestry. However, the direction of effect was discordant in some replication cohorts. Different directions of associations may be due different patterns of LD with the causal variant between populations. The minor allele in rs291231 might be
protective in ever smokers as demonstrated by a per-allele increase in FEV\textsubscript{1} of 57.28 mL.

The chromosome 18 locus associated with FEV\textsubscript{1}/FVC among never smokers contained multiple SNPs. The lead SNP (rs7228593) varies in MAF across different Hispanic/Latino ancestry groups in HCHS/SOL (0.02-0.16), is common in African populations (MAF=0.39) and is not polymorphic in European populations (MAF=0)(71) These variable allele frequencies and small sample size for replication may have contributed to lack of replication. Differing environmental exposures and LD may explain the observed genetic association of FEV\textsubscript{1}/FVC among never smokers in HCHS/SOL and lack of replication in independent cohorts of different ancestry backgrounds.

In the COPD analysis, SNP rs7709630, a genotyped intron variant in PDZD2, was a single SNP association with COPD. This genomic region replicated for airflow obstruction in an independent population of European ancestry(23) suggesting a true association. Whereas the CDRT15P1 locus only contained imputed SNPs and did not replicate, thus reducing the probability that this is a true association with COPD.

We hypothesize that novel loci did not replicate due to small sample size of the Hispanic/Latino replication cohorts along with differing environmental exposures, differing LD, and the complex racial admixture(2) that varies by different Hispanic/Latino ancestry groups. Furthermore, the novel locus for FEV\textsubscript{1}/FVC in HAL and the chromosome 18 locus for FEV\textsubscript{1}/FVC among never smokers have marked differences in MAF between race and ethnic populations suggesting genetic diversity between races in these loci. The highest MAF for the lead SNPs in the HAL and in the chromosome 18 locus were observed in African populations (MAF = 0.005 and 0.38, respectively)
compared with virtually no variability in European populations for these loci (MAF=0) (63). The lack of variance of these loci in European populations limits the ability to replicate these loci as the vast majority of GWAS of lung function have been performed in studies of European ancestry. Evidence for replication of the novel locus on chromosome 11 for FEV₁ in ever smokers was stronger in the meta-analysis of cohorts with admixed populations that include African ancestry compared with the meta-analysis of all cohorts (p=0.007 and p=0.04, respectively) supporting the need for additional studies on admixed populations to identify possible ethnicity- or race-specific variants that may elucidate novel pathways in the pathogenesis of COPD.

The study has several potential limitations. This population-based cohort of diverse Hispanic/Latino participants is subject to population stratification. Multiple loci demonstrated varying MAF across Hispanic/Latino groups highlighting the complex racial admixture of Hispanic/Latino populations and raises the possibility of residual population stratification. However, we adjusted our analysis for PCs and our inflation factors ranged from 0.988 to 1.026 which indicates good control of population stratification with small inflation. The COPD analysis consisted of 363 and 5,253 participants with and without COPD (respectively), which is a relatively small sample size for a GWAS. We optimized COPD phenotype by limiting the analysis to age greater than or equal to 45 and used post-bronchodilator spirometry measurements to define COPD.

In conclusion, we identified novel biologically plausible signals associated with clinically important pulmonary measures with evidence for replication in ZSWIM7, AGMO, and PDZD2; confirmed previous reports of association with FEV₁/FVC in
KCNE2 and GPR126, and established that loci previously identified in European populations are generalizable in Hispanic/Latino populations. There is a paucity of genetic studies for lung function that include under-represented minority populations, such as Hispanic/Latino(2, 72) and African populations(14, 26, 73-75). Our findings emphasize the importance of including admixed and multi-racial populations when performing genetic studies of complex diseases and have the potential to advance our understanding of genetic risks for lung disease affecting Hispanic/Latino populations.

ACKNOWLEDGEMENTS

The HCHS/SOL investigators thank the National Heart, Lung, and Blood Institute and the research institutions, study investigators, and field staff for their support in creating this study. We also thank the study participants, without whom this endeavor would not have been possible. The authors thank the other investigators, staff, and participants of HCHS/SOL for their valuable contributions.
References:


Figure 1 Regional Association Plots and Forest Plots of genome wide significant loci associated with FEV₁ and FEV₁/FVC meta-analyzed across all Hispanic/Latino ancestry groups.

For each locus, we provide regional association plots with correlations between the reference SNP (the SNP with the lowest p-value) and other SNPs in the region. The reference SNPs are purple diamonds (genotyped SNP) or upside down triangle (imputed SNP). Other SNPs in the region are depicted as circles (genotyped SNPs) and “X”s (imputed SNPs). The correlations ($r^2$) are calculated from the group of interest and are indicated by the colors shown on the plot. For each locus, we also provide a forest plot comparing the SNP-trait association testing results across the Hispanic/Latino ancestry groups. (a) Regional association plot for the FEV₁ locus (lead SNP rs4791658) on chromosome 17. (b) Forrest Plot for the FEV₁ locus (lead SNP rs4791658) on chromosome 17. (c) Regional association plot for the FEV₁/FVC locus (lead SNP rs145174011) on chromosome 12. (d) Forrest Plot for the FEV₁/FVC locus (lead SNP rs192375903) on chromosome 12.
**Figure 2** Regional Association Plots and Forest Plots of genome wide significant loci associated with FEV₁ and FEV₁/FVC stratified by smoking status and meta-analyzed across all Hispanic/Latino Ancestry groups.

For each locus, we provide regional association plots with correlations between the reference SNP (the SNP with the lowest p-value) and other SNPs in the region. The reference SNPs a purple diamond (genotyped SNP) or upside down triangle (imputed SNP). Other SNPs in the region are depicted as circles (genotyped SNPs) and “X”s (imputed SNPs). The correlations ($r^2$) are calculated from the group of interest and are indicated by the colors shown on the plot. For each locus, we also provide a forest plot comparing the SNP-trait association testing results across the Hispanic/Latino ancestry in smoking stratified group of interest. (a) Regional association plot for the FEV₁/FVC locus among never smokers (lead SNP rs7228593) on chromosome 18. (b) Forrest Plot for the FEV₁/FVC locus among never smokers (lead SNP rs7228593) on chromosome 18. (c) Regional association plot for the FEV₁ locus among ever smokers (lead SNP rs291231) on chromosome 11. (d) Forrest Plot for the FEV₁ locus among ever smokers (lead SNP rs291231) on chromosome 11.
Figure 3 Regional Association Plot and Forest Plot of candidate variant identified in local ancestry interval region associated with Amerindian ancestry and FEV$_1$ meta-analyzed across all Hispanic/Latino ancestry groups.

For each locus, we provide regional association plots with correlations between the reference SNP (the SNP with the lowest p-value) and other SNPs in the region. The reference SNPs a purple diamond (genotyped SNP) or upside-down triangle (imputed SNP). Other SNPs in the region are depicted as circles (genotyped SNPs) and “X”s (imputed SNPs). The correlations ($r^2$) are calculated from the group of interest and are indicated by the colors shown on the plot. We also provide a forest plot comparing the SNP-trait association testing results across the Hispanic/Latino ancestry groups. (a) Regional association plot for the FEV$_1$ locus (lead SNP rs4133185) on chromosome 7. (b) Forrest Plot for the FEV$_1$ locus (lead SNP rs4133185) on chromosome 7.
**Figure 4** Amerindian Admixture Mapping Region on Chromosome 7 for FEV$_1$

The left panel provides the admixture mapping results as two lines with the blue line representing results from the primary analysis, the green line representing results from the conditional analysis and the association results in the same region are represented as circles. The genome-wide significant threshold for admixture mapping in HCHS/SOL data set is the horizontal grey dashed line. The blue and green lines and circles are given as $-\log(p$-value,10) against genomic positions. The red-filled triangle corresponds to the SNP used in the conditional analysis (rs4133185). The right panel provides the ancestry-specific effect allele frequencies for the SNP used in the conditional analysis (rs4133185), as estimated by ASAFE applied on HCHS/SOL data set(46).
Table 1: Characteristics of Participants in HCHS/SOL Study

<table>
<thead>
<tr>
<th></th>
<th>Central American</th>
<th>Cuban</th>
<th>Dominican</th>
<th>Mexican</th>
<th>Puerto Rican</th>
<th>South American</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>1,315</td>
<td>2,138</td>
<td>1,063</td>
<td>4,438</td>
<td>2,000</td>
<td>868</td>
<td>11,822</td>
</tr>
<tr>
<td>Age (years)</td>
<td>44 ± 13</td>
<td>49 ± 13</td>
<td>45 ± 14</td>
<td>44 ± 14</td>
<td>48 ± 14</td>
<td>46 ± 13</td>
<td>46 ± 14</td>
</tr>
<tr>
<td>Male Sex (%)</td>
<td>40</td>
<td>47</td>
<td>35</td>
<td>39</td>
<td>43</td>
<td>40</td>
<td>41</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>160 ± 9</td>
<td>164 ± 9</td>
<td>162 ± 9</td>
<td>161 ± 9</td>
<td>163 ± 9</td>
<td>161 ± 9</td>
<td>162 ± 9</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>30 ± 6</td>
<td>29 ± 6</td>
<td>29 ± 6</td>
<td>30 ± 6</td>
<td>31 ± 7</td>
<td>29 ± 5</td>
<td>30 ± 6</td>
</tr>
<tr>
<td>Smoking (never) %</td>
<td>68</td>
<td>52</td>
<td>77</td>
<td>65</td>
<td>48</td>
<td>66</td>
<td>61</td>
</tr>
<tr>
<td>Smoking (former) %</td>
<td>19</td>
<td>19</td>
<td>13</td>
<td>19</td>
<td>20</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Smoking (current) %</td>
<td>13</td>
<td>29</td>
<td>10</td>
<td>16</td>
<td>32</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Pack years of smoking*</td>
<td>4.5 (1.7 - 14)</td>
<td>18.5 (6 - 36)</td>
<td>8.2 (3.4 - 18.3)</td>
<td>4.0 (1.3 - 10.5)</td>
<td>11.0 (4 - 25.5)</td>
<td>4.9 (1.5 - 16)</td>
<td>7.5 (2.4 - 21)</td>
</tr>
<tr>
<td>Pre-bronchodilator FEV₁ (liters)</td>
<td>2.88 ± 0.74</td>
<td>2.80 ± 0.81</td>
<td>2.70 ± 0.75</td>
<td>2.97 ± 0.78</td>
<td>2.70 ± 0.83</td>
<td>2.95 ± 0.79</td>
<td>2.86 ± 0.80</td>
</tr>
<tr>
<td>Pre-bronchodilator FEV₁/FVC (%)</td>
<td>81.5 ± 5.9</td>
<td>78.7 ± 8.3</td>
<td>81.8 ± 7.2</td>
<td>80.8 ± 6.1</td>
<td>79.3 ± 7.9</td>
<td>80.1 ± 6.5</td>
<td>80.3 ± 7</td>
</tr>
<tr>
<td>COPD n (%)†</td>
<td>24 (1.8)</td>
<td>124 (5.8)</td>
<td>24 (2.3)</td>
<td>85 (1.9)</td>
<td>86 (4.3)</td>
<td>20 (2.3)</td>
<td>363 (3.1)</td>
</tr>
<tr>
<td>Airflow Limitation n (%)‡</td>
<td>60 (4.6)</td>
<td>260 (12.2)</td>
<td>62 (5.8)</td>
<td>207 (4.7)</td>
<td>216 (10.8)</td>
<td>49 (5.7)</td>
<td>854 (7.2)</td>
</tr>
</tbody>
</table>

Data presented as mean ± standard deviation for continuous measures and percentage for binary measures

**Definition of abbreviations:** FEV₁ = forced expiratory volume in one second, FEV₁/FVC = ratio of forced expiratory volume in one second over forced vital capacity %

* Pack years of smoking presented as median ± Inter Quartile Range (IQR) in former or current smokers

† Chronic Obstructive Pulmonary Disease defined as post-bronchodilator FEV₁/FVC ratio < 0.70

‡ Defined as pre-bronchodilator FEV₁/FVC ratio < 0.70
### Table 2: Genome wide significant loci associated with FEV₁ and FEV₁/FVC

<table>
<thead>
<tr>
<th>SNP rsID</th>
<th>Trait</th>
<th>Position Build 37 Chr: basepair</th>
<th>Gene / nearest Gene(s) (function)</th>
<th>Effect allele</th>
<th>EAF</th>
<th>Genotyped or Imputed SNP</th>
<th>oevar*</th>
<th>N</th>
<th>Beta</th>
<th>SE</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs76656601</td>
<td>FEV₁/FVC</td>
<td>6: 168428548</td>
<td><strong>KIF25</strong> (intron)</td>
<td>G</td>
<td>0.996</td>
<td>Imputed</td>
<td></td>
<td>11822</td>
<td>4.70</td>
<td>0.77</td>
<td>1.31 x 10⁻⁹</td>
</tr>
<tr>
<td>rs4791658</td>
<td>FEV₁</td>
<td>17:15884792</td>
<td><strong>ZSWIM7</strong>&lt;sup&gt;†&lt;/sup&gt; (intron)</td>
<td>G</td>
<td>0.541</td>
<td>Imputed&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>0.996</td>
<td>11822</td>
<td>33.35</td>
<td>5.70</td>
<td>4.99 x 10⁻⁹</td>
</tr>
<tr>
<td>rs145174011</td>
<td>FEV₁/FVC</td>
<td>12: 96378700</td>
<td><strong>HAL</strong>&lt;sup&gt;†&lt;/sup&gt; (Intron)</td>
<td>T</td>
<td>0.997</td>
<td>Imputed</td>
<td>0.98</td>
<td>11822</td>
<td>4.85</td>
<td>0.85</td>
<td>9.59 x 10⁻⁹</td>
</tr>
<tr>
<td>rs28593428</td>
<td>FEV₁/FVC</td>
<td>21: 35632170</td>
<td><strong>KCNE2</strong></td>
<td>C</td>
<td>0.902</td>
<td>Imputed</td>
<td>0.996</td>
<td>11821</td>
<td>0.80</td>
<td>0.14</td>
<td>1.45 x 10⁻⁸</td>
</tr>
<tr>
<td>rs262113</td>
<td>FEV₁/FVC</td>
<td>6: 142824950</td>
<td><strong>GPR126</strong></td>
<td>T</td>
<td>0.812</td>
<td>Imputed&lt;sup&gt;†&lt;/sup&gt;</td>
<td>0.999</td>
<td>11822</td>
<td>- 0.57</td>
<td>0.10</td>
<td>2.83 x 10⁻⁸</td>
</tr>
<tr>
<td>rs74444778</td>
<td>FEV₁/FVC</td>
<td>7: 42553708</td>
<td><strong>LOC105375250</strong> (intron)</td>
<td>G</td>
<td>0.999</td>
<td>Imputed</td>
<td>0.914</td>
<td>11822</td>
<td>6.17</td>
<td>1.12</td>
<td>3.61 x 10⁻⁸</td>
</tr>
<tr>
<td>rs115745680</td>
<td>FEV₁</td>
<td>5: 154407079</td>
<td><strong>GEMIN5 KIF4B</strong></td>
<td>A</td>
<td>0.997</td>
<td>Imputed</td>
<td>0.981</td>
<td>11822</td>
<td>313.24</td>
<td>57.32</td>
<td>4.63 x 10⁻⁸</td>
</tr>
</tbody>
</table>

**Definition of abbreviations:** SNP = single nucleotide polymorphism, FEV₁/FVC = ratio of forced expiratory volume in one second over forced vital capacity,

Chr = chromosome, nearest gene = the nearest genes on each side of the intergenic marker, EAF = effect allele frequency, oevar = observed / expected variance, N = sample size used in testing the given variant, Beta (FEV₁) = per-allele change in FEV₁ (mL), Beta (FEV₁/FVC) = per-allele change in FEV₁/FVC %, SE = standard error
Gene abbreviations: KIF25 = (kinesin family member 25), ZSWIM7 = (zinc finger, SWIM-type containing 7), HAL = (histidine ammonia-lyase), GPR126 = (G protein-coupled receptor 126), KCNE2 = (potassium channel, voltage gated subfamily E regulatory beta subunit 2), GEMNIN5 = gem nuclear organelle associated protein 5, KIF4B = kinesin family member 4B

Analyses were performed using data imputed to 1000 Genome phase 1 data and stratified by Hispanic/Latino ancestry group and were adjusted for age, age$^2$, sex, height, height$^2$, smoking status, pack-years, sampling weights, first 5 principal components, Hispanic/Latino ancestry groups, kinship and block unit. Within subgroup estimates were then meta-analyzed. Previously unreported associations are in bold. Retrospectively, association analyses using genetic data imputed to 1000 Genome phase 3 imputation panel were performed in novel loci.

* oevar is provided by MACH software as a measure of imputation quality;
† Results for analyses using genetic data imputed to 1000 Genome phase 3 imputation panel
‡ Genotyped SNP in locus
<table>
<thead>
<tr>
<th>SNP rsID</th>
<th>Trait</th>
<th>Position Build 37 Chr: basepair</th>
<th>Nearest Gene(s)</th>
<th>Effect allele</th>
<th>EAF</th>
<th>Genotyped or imputed SNP</th>
<th>oevar*</th>
<th>N</th>
<th>Beta</th>
<th>SE</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9974878</td>
<td>FEV₁/FVC</td>
<td>21: 35642446</td>
<td>KCNE2</td>
<td>G</td>
<td>0.887</td>
<td>Imputed†</td>
<td>0.994</td>
<td>4594</td>
<td>1.28</td>
<td>0.23</td>
<td>3.58 x 10⁻⁵</td>
</tr>
<tr>
<td>rs291231</td>
<td>FEV₁</td>
<td>11: 86007090</td>
<td>EED CCDC81</td>
<td>G</td>
<td>0.341</td>
<td>Imputed†</td>
<td>0.999</td>
<td>4595</td>
<td>57.28</td>
<td>10.19</td>
<td>1.92 x 10⁻⁸</td>
</tr>
<tr>
<td>Indel (no rs#)</td>
<td>FEV₁</td>
<td>11: 86006803</td>
<td>EED CCDC81</td>
<td>GACA</td>
<td></td>
<td>Imputed†</td>
<td>1.017</td>
<td>56.95</td>
<td>10.09</td>
<td>1.66 x 10⁻⁸</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNP rsID</th>
<th>Trait</th>
<th>Position Build 37 Chr: basepair</th>
<th>Nearest Gene(s)</th>
<th>Effect allele</th>
<th>EAF</th>
<th>Genotyped or imputed SNP</th>
<th>oevar*</th>
<th>N</th>
<th>Beta</th>
<th>SE</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs116726860</td>
<td>FEV₁/FVC</td>
<td>2: 155486042</td>
<td>KCNJ3 GALNT13</td>
<td>T</td>
<td>0.992</td>
<td>Imputed</td>
<td>0.897</td>
<td>7227</td>
<td>3.49</td>
<td>0.60</td>
<td>6.04 x 10⁻⁹</td>
</tr>
<tr>
<td>rs7228593</td>
<td>FEV₁/FVC</td>
<td>18: 73575704</td>
<td>SMIM21 LOC339298</td>
<td>G</td>
<td>0.948</td>
<td>Imputed†</td>
<td>0.997</td>
<td>7227</td>
<td>1.34</td>
<td>0.24</td>
<td>3.47 x 10⁻⁸</td>
</tr>
</tbody>
</table>

**Definition of abbreviations:** SNP= single nucleotide polymorphism, FEV₁/FVC= ratio of forced expiratory volume in one second over forced vital capacity,

Chr = chromosome, nearest gene = the nearest genes on each side of the intergenic marker, EAF = effect allele frequency, oevar = observed / expected variance, N = sample size used in testing the given variant, Beta (FEV₁) = per-allele change in FEV₁ (mL), Beta (FEV₁/FVC) = per-allele change in FEV₁/FVC %, SE = standard error, Indel = Indel mutation
Gene abbreviations: KCNJ3 = potassium voltage-gated channel subfamily J member 3, GALNT13 = polypeptide N-acetylgalactosaminyltransferase 13, SMIM21 = small integral membrane protein 21, EED = genes embryonic ectoderm development, KCNE2 = (potassium channel, voltage gated subfamily E regulatory beta subunit 2), MROH2A = (maestro heat-like repeat family member 2A)

Analyses were stratified by Hispanic/Latino ancestry group and smoking status. Model adjusted for age, age², sex, height, height², smoking status, pack-years, sampling weights, first 5 principal components, Hispanic/Latino ancestry groups, kinship and block unit. Within subgroup estimates were then meta-analyzed. Previously unreported associations are in bold.

* oevar is provided by MACH software as a measure of imputation quality; it is a measure of imputation quality
† Genotyped SNP in locus