1 **TITLE**

Novel *KITLG* regulatory variants are associated with lung function in African American children with asthma

**AUTHORS**

Angel CY Mak¹,*, Satria Sajuthi²,^, Jaehyun Joo³,^, Shujie Xiao⁴,^, Patrick M Sleiman⁵,⁶,^, Marquitta J White¹,^, Eunice Y Lee¹, Benjamin Saef³, Donglei Hu¹, Hongsheng Gui⁴, Kevin L Keys¹,⁷, Fred Lurmann⁸, Deepi Jain⁹, Gonçalo Abecasis¹⁰, Hyun Min Kang¹⁰, Deborah A. Nickerson¹¹,¹²,¹³, Soren Germer¹⁴, Michael C Zody¹⁴, Lara Winterkorn¹⁴, Catherine Reeves¹⁴, Scott Huntsman¹, Celeste Eng¹, Sandra Salazar¹, Sam S Oh¹, Frank D Gilliland¹⁵, Zhanghua Chen¹⁵, Rajesh Kumar¹⁶, Fernando D Martinez¹⁷, Ann Chen Wu¹⁸, Elad Ziv¹, Hakon Hakonarson⁵,⁶,#, Blanca E Himes³,#, L Keoki Williams⁴,#, Max A Seibold¹,#, Esteban G. Burchard¹,²⁰

¹ Department of Medicine, University of California San Francisco, San Francisco, CA, USA ¹² Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA ¹³ Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA ¹⁴ Center for Individualized and Genomic Medicine Research, Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA ¹⁵ Center for Applied Genomics, Children’s Hospital of Philadelphia, Philadelphia, PA, USA ¹⁶ Department of Human Genetics, Children’s Hospital of Philadelphia, Philadelphia, PA, USA ¹⁷ Berkeley Institute for Data Science, University of California, Berkeley, CA, USA ¹⁸ Sonoma Technology Inc, Petaluma, CA, USA ¹⁹ Department of Biostatistics, University of Washington, Seattle, WA, USA ²⁰ Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA ²¹ Department of Genome Sciences, University of Washington, Seattle, WA, USA ²² Northwest Genomics Center, Seattle, WA, USA ²³ Brotman Baty Institute, Seattle, WA, USA
New York Genome Center, New York, NY, USA

Department of Preventive Medicine, Division of Environmental Health, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, USA

Asthma and Airway Disease Research Center, University of Arizona, Tucson, AZ, USA

Precision Medicine Translational Research (PRoMoTeR) Center, Department of Population Medicine, Harvard Medical School and Pilgrim Health Care Institute, Boston, MA, USA

Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA

^, # These authors contributed equally to this work

* Corresponding author

Angel CY Mak

Address: Box 2911, 1550 4th Street, San Francisco, CA 94112

Tel: 415-514-9931

Email: angel.mak@ucsf.edu

AUTHOR CONTRIBUTIONS

ACYM, SSaj, MJW, SSO, FG, ZC, RK, FDM, ACW, EZ, and EGB designed and supervised the study.

ACYM, SSaj, JJ, SX, PS, MJW, EYL, BS, DH, SH, BEH, HH, LKW, and MAS supervised/perform analyses and/or interpreted results. CE processed biospecimens for whole genome sequencing.

FL estimated air pollution exposure. DJ, GA, HMK, DAN, and SG contributed to TOPMed WGS data generation and/or analysis. MZ, LW, and CR contributed to CCDG WGS data generation.

SSal coordinated study recruitment and processed phenotype data ACYM, SSaj, JJ, SX, PS, MJW, EYL, BS, HG, KLK, SG, BEH, HH, LKW, MAS, and EGB wrote and/or critically reviewed the manuscript.
**SHORT TITLE**

*KITLG* associated with FEV$_1$ in AA youths

**KEYWORDS**

GWAS, African American, FEV$_1$, gene-by-environment interaction, air pollution

**ABSTRACT**

Baseline lung function, quantified as forced expiratory volume in the first second of exhalation (FEV$_1$), is a standard diagnostic criterion used by clinicians to identify and classify lung diseases. Using whole genome sequencing data from the National Heart, Lung, and Blood Institute TOPMed project, we identified a novel genetic association with FEV$_1$ on chromosome 12 in 867 African American children with asthma ($p = 1.26 \times 10^{-8}$, $\beta = 0.302$). Conditional analysis within 1 Mb of the tag signal (rs73429450) yielded one major and two other weaker independent signals within this peak. We explored statistical and functional evidence for all variants in linkage disequilibrium with the three independent signals and yielded 9 variants as the most likely candidates responsible for the association with FEV$_1$. Hi-C data and eQTL analysis demonstrated that these variants physically interacted with *KITLG* and their minor alleles were associated with increased expression of *KITLG* gene in nasal epithelial cells. Gene-by-air-pollution interaction analysis found that the candidate variant rs58475486 interacted with past-year SO$_2$ exposure ($p = 0.003$, $\beta = 0.32$). This study identified a novel protective genetic association with FEV$_1$, possibly mediated through *KITLG*, in African American children with asthma.
INTRODUCTION

Asthma, a chronic pulmonary condition characterized by reversible airway obstruction, is one of the hallmark diseases of childhood in the United States (World Health Organization 2017). Asthma is also the most disparate common disease in the pediatric clinic, with significant variation in prevalence, morbidity, and mortality among U.S. racial/ethnic groups (Oh et al. 2016). Specifically, African American children carry a higher asthma disease burden compared to their European American counterparts (Akinbami et al. 2014; Akinbami 2015). Forced expiratory volume in the first second (FEV₁), a measurement of lung function, is a vital clinical trait used by physicians to assess overall lung health and diagnose pulmonary diseases such as asthma (Johnson and Theurer. 2014). We have previously shown that genetic ancestry plays an important role in FEV₁ variation and that African Americans have lower FEV₁ compared to European Americans regardless of asthma status (Kumar et al. 2010; Pino-Yanes et al. 2015).

The disparity in lung function between populations may explain disparities in asthma disease burden. Understanding the factors that influence FEV₁ variation among individuals with asthma could lead to improved patient care and therapeutic interventions.

Twin and family-based studies estimate that the heritability of FEV₁ is as high as 81%, supporting a strong contribution by genetic factors in FEV₁ variation (Chatterjee and Das. 1995; Hukkinen et al. 2011; Palmer et al. 2001; Sillanpaa et al. 2017; Tian et al. 2017; Yamada et al. 2015). Genome-wide association studies (GWAS) of FEV₁, including among individuals with asthma, have identified several variants that contribute to lung function (Li et al. 2013; Liao et al. 2014; Repapi et al. 2010; Soler Artigas et al. 2011; Soler Artigas et al. 2015; Wain et al. 2013).
Most of these previous GWAS, however, were performed in adult populations of European descent, and their results may not generalize across populations or across the life span of an individual (Carlson et al. 2013; Martin, A. R. et al. 2017; Wojcik et al. 2019). Previous GWAS results are also limited due to their reliance on genotyping arrays. In particular, variation in non-coding regions of the genome is not adequately covered by many genotyping arrays because they were not designed while taking into account the population-specific genetic variability of all populations (Kim, M. S. et al. 2018; Zhang and Lupski. 2015). Whole genome sequencing (WGS) is a newer technology that captures nearly all common variation from coding and non-coding regions of the genome and is unencumbered by genotype array design constraints and differences in linkage disequilibrium patterns among populations. To date, no large-scale WGS studies of lung function have been performed in African American children with asthma (Martin et al. 2017).

In addition to genetics, FEV\textsubscript{1} is a complex trait that is significantly influenced by both genetic variation and environmental factors, such as air pollution (Chatterjee and Das. 1995; Hukkinen et al. 2011; Palmer et al. 2001; Sillanpaa et al. 2017; Tian et al. 2017; Yamada et al. 2015). Exposure to ambient air pollution has been consistently associated with poor respiratory outcomes, including reduced FEV\textsubscript{1} (Barraza-Villarreal et al. 2008; Brunekreef and Holgate. 2002; Ierodiakonou et al. 2016; Wise 2019). We previously showed that exposure to sulfur dioxide (SO\textsubscript{2}), an air pollutant emitted by the burning of fossil fuels, is significantly associated with reduced FEV\textsubscript{1} in African American children with asthma in the SAGE II study (Neophytou et al. 2016). Because the genetic variants associated with FEV\textsubscript{1} thus far do not account for the majority of its estimated heritability, considering gene-environment interactions, specifically...
gene-by-air-pollution, may improve our understanding of lung function genetics (Moore 2005; Moore and Williams. 2009). Here, we performed a genome-wide association analysis using WGS data to identify common genetic variants associated with FEV₁ in African American children with asthma in SAGE II and investigated the effect of gene-by-air-pollution (SO₂) interactions on FEV₁ associations.

METHODS

Study population

This study examined African American children between 8-21 years of age with physician-diagnosed asthma from the Study of African Americans, Asthma, Genes & Environments (SAGE II). All SAGE II participants were recruited from the San Francisco Bay Area. The inclusion and exclusion are previously described in detailed (Oh et al. 2012; White et al. 2016). Briefly, participants were eligible if they were 8-21 years of age and self-identified as African American and had four African American grandparents. Study exclusion criteria included the following: 1) any smoking within one year of the recruitment date; 2) 10 or more pack-years of smoking; 3) pregnancy in the third trimester; 4) history of lung diseases other than asthma (for cases) or chronic illness (for cases and controls). Baseline lung function defined as forced expiratory volume in the first second (FEV₁) was measured by spirometry prior to administering albuterol as previously described (Oh et al. 2012).

TOPMed whole genome sequencing data

SAGE II DNA samples were sequenced as part of the Trans-Omics for Precision Medicine (TOPMed) whole genome sequencing (WGS) program (Taliun et al. 2019). WGS was performed
at the New York Genome Center and Northwest Genomics Center on a HiSeq X system (Illumina, San Diego, CA) using a paired-end read length of 150 base pairs (bp), with a minimum of 30x mean genome coverage. DNA sample handling, quality control, library construction, clustering and sequencing, read processing and sequence data quality control are described in detail in the TOPMed website (TOPMed 2019). Variant calls were obtained from TOPMed data freeze 8 VCF files corresponding to the GRCh38 assembly. Variants with a minimal read depth of 10 (DP10) were used for analysis unless otherwise stated.

**Genetic principal components, global ancestry, and kinship estimation**

Genetic principal components (PCs), global ancestry, and kinship estimation on genetic relatedness were computed using biallelic single nucleotide polymorphisms (SNPs) with a PASS flag from TOPMed freeze 8 DP10 data. PCs and kinship estimates were computed using the PC-Relate function from the GENESIS R package (Conomos et al. 2015; Conomos et al. 2016) using a workflow available from the Summer Institute in Statistical Genetics Module 17 course website (Summer Institute in Statistical Genetics 2019). African global ancestry was computed using the ADMIXTURE package (Alexander et al. 2009) in supervised mode using European (CEU), African (YRI) and Native American (NAM) reference panels as previously described (Mak, A. C. Y. et al. 2018).

**FEV₁ GWAS**

Non-normality of the distribution of FEV₁ values was tested with the Shapiro-Wilk test in R using the shapiro.test function. Since FEV₁ was not normally distributed ($p = 1.41 \times 10^{-8}$ for FEV₁ and $p$
\[ \log_{10}(FEV_1) = 1.05 \times 10^{-8} \]

FEV\(_1\) was regressed on all covariates (Age, sex, height, controller medications, sequencing centers, and the first 5 genetic PCs) and the residuals were inverse-normalized. These inverse-normalized residuals (FEV\(_1\).res.rnorm) were the main outcome of the discovery GWAS. The controller medication covariate included the use of inhaled corticosteroids (ICS), long-acting beta-agonists (LABA), leukotriene inhibitors and/or an ICS/LABA combo in the 2 weeks prior to the recruitment date.

Genome-wide single variant analysis was performed on the ENCORE server ([https://github.com/statgen/encore](https://github.com/statgen/encore)) using the linear Wald test (q.linear) originally implemented in EPACTS ([https://genome.sph.umich.edu/wiki/EPACTS](https://genome.sph.umich.edu/wiki/EPACTS)) and TOPMed freeze 8 data (DP0 PASS) with a MAF filter of 0.1%. All pairwise relationships with degree 3 or more relatedness (kinship values > 0.044) were identified, and one participant of the related pair was subsequently chosen at random and removed prior to analysis. All covariates used to obtain FEV\(_1\).res.rnorm were also included as covariates in the GWAS as recommended in a recent publication (Sofer et al. 2019).

The association analysis was repeated using untransformed FEV\(_1\) and FEV\(_1\) percent predicted (FEV\(_1\).perc.predicted). FEV\(_1\) percent predicted was defined as the percentage of measured FEV\(_1\) relative to predicted FEV\(_1\) estimated by the Hankinson lung function prediction equation for African Americans (Hankinson et al. 1999). A secondary analysis that included smoking-related covariates (smoking status and number of smokers in the family) was performed in PLINK 1.9 (version 1p9_2019_0304_dev) (Chang et al. 2015; Purcell and Chang. 2013). Regional association results were plotted using LocusZoom 1.4 (Pruim et al. 2010) with a 500 kilobase (Kb) flanking region. Linkage disequilibrium (R\(^2\)) was estimated in PLINK 1.9.
The function `effectiveSize` in the R package CODA was used to estimate the actual effective number of independent tests and CODA-adjusted statistical and suggestive significance p-value thresholds were defined as 0.05 and 1 divided by the effective number of tests, respectively (Duggal et al. 2008). We compared the CODA-adjusted statistical significance threshold and the widely used $5 \times 10^{-8}$ GWAS genome-wide significance threshold (Pe'er et al. 2008) and selected the more stringent threshold for genome-wide significance.

The following WGS quality control steps were applied to all reported variants from ENCORE to ensure WGS variant quality: (1) The variant had `VCF FILTER = PASS`; (2) Variant quality was confirmed via manual inspection on the BRAVO server based on TOPMed freeze 5 data (University of Michigan and NHLBI TOPMed. 2018); (3) Variants were reanalyzed with linear regression using PLINK 1.9 by applying the arguments `--mac 5 --geno 0.1 --hwd 0.0001` using TOPMed freeze 8 DP10 PASS data.

To determine if the rs73429450 association with FEV$_1$ was only identifiable using whole genome sequencing data, we repeated the linear regression association analysis on signals that passed the genome-wide significance threshold using PLINK 1.9 and genotype data generated with Axiom Genome-Wide LAT 1 array (Affymetrix, Santa Clara, CA, dbGaP phs000921.v1.p1). These array genotype data were imputed into the following reference panels: 1000 Genomes phase 3 version 5, Haplotype Reference Consortium (HRC) r1.1, the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) and the TOPMed phase 5 panels on the Michigan Imputation Server (Das et al. 2016). It should be noted that 500 SAGE II subjects were part of the TOPMed freeze 5 reference panel.
Ninety-eight GWAS FEV$_1$-associated loci were retrieved from GWAS catalog version 1.0.2-associations_e93_r2019-01-31 (Buniello et al. 2019) using the trait name “Lung function (FEV$_1$)”. Overlap between these 98 loci with 20 Kb flanking regions and the discovery GWAS FEV$_1$ associations (p < 0.001) were detected using the BEDTools intersect tool (Quinlan and Hall. 2010).

**Conditional analysis**

Conditional analysis was performed to identify all independent signals in a GWAS peak using PLINK 1.9. All TOPMed freeze 8 DP10 variants within 1 megabase (Mb) of the tag association signal and with association p-value of 1 x 10$^{-4}$ or smaller in the discovery GWAS were included in the analysis. Variants were first ordered by ascending p-value. A variant was considered to be an independent signal if the association p-value after conditioning (conditional p-value) on the tag signal was smaller than 0.05. Newly identified independent signals were included with the tag signal for conditioning on the next variant.

**Region-based association analysis**

Region-based association analyses were performed in 1 Kb sliding windows with 500 bp increments in a 1 Mb flanking region of the tag GWAS signal using the SKAT_CommonRare function from the SKAT R package v1.3.2.1 (Ionita-Laza et al. 2013). Default settings were used with method = "C" and test.type = "Joint". A minor allele frequency (MAF) threshold of 0.01 was used as the cutoff to distinguish rare and common variants. Variants were annotated in TOPMed using the WGSA pipeline (Liu et al. 2016). Since SKAT imputes missing genotypes by default by assigning mean genotype values (impute.method="fixed"), we chose to use low
coverage genotypes instead of SKAT imputation, and hence, TOPMed freeze 8 DPO variants with a VCF FILTER of PASS were included in the analysis. The function effectiveSize in the R package CODA (Plummer et al. 2006) was used to estimate the effective number of independent hypothesis tests for accurate Bonferroni multiple testing corrections. P-value thresholds for statistical significance and suggestive significance were defined as 0.05 and 1 divided by the effective number of tests, respectively (Duggal et al. 2008). If a region was suggestively significant, region-based analyses were repeated with functional variants and/or rare variants (MAF <= 0.01) to assess contribution of common, rare and/or functional variants. Region-based analyses using rare variants only were performed using SKAT-O (Lee et al. 2012). The WGSA annotation filters used to define functional variants are provided in File S1 (Supplementary Text 1). To study the contribution of individual variants to a region-based association p-value, drop-one variant analysis was performed by repeating the region-based analysis multiple times and dropping one variant only at a time.

Functional annotations and prioritization of genetic variants

The Hi-C Unifying Genomic Interrogator (HUGIN) (Ay et al. 2014; Martin, J. S. et al. 2017; Schmitt et al. 2016) was used to assign potential gene targets to each variant. ENCODE annotations (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012) were based on overlap of the variants with functional data downloaded from the UCSC Table Browser (Karolchik et al. 2004). These data included DNAase I hypersensitivity peak clusters (hg38 wgEncodeRegDnaseClustered table), transcription factor ChIP-Seq clusters (hg38 encRegTfbsClustered table) and histone modification ChIP-Seq peaks (hg19...
wgEncodeBroadHistone<cell type><histone>StdPk tables). For DNase I hypersensitivity and transcription factor binding sites, we focused on blood, bone marrow, lung and embryonic cells. For histone modification ChIP-Seq, we focused on H3K27ac and H3K4me3 modifications in human blood (GM12878), bone marrow (K562), lung fibroblast (NHLF), and embryonic stem cells (H1-hESC). LiftOver tool (Hinrichs et al. 2006) was used to convert genomic coordinates from hg19 to hg38. Candidate cis-regulatory elements (ccREs) were a subset of representative DNase hypersensitivity sites with epigenetic activity further supported by histone modification (H3K4me3 and H3K27ac) or CTCF-binding data from the ENCODE project. Overlap of variants with ccREs were detected using the Search Candidate cis-Regulatory Elements by ENCODE (SCREEN) web interface (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012).

Prioritization of genetic variants was based on the presence of statistical, functional and/or bioinformatic evidence as described in the Diverse Convergent Evidence (DiCE) prioritization framework (Ciesielski et al. 2014). The priority score of each variant was obtained by summing the number of evidence present for that variant.

**Replication of GWAS associations**

All replication analyses were performed in subjects with asthma. Replication of GWAS FEV\textsubscript{1} associations was attempted on TOPMed whole genome sequencing data generated from four cohorts. These cohorts included Puerto Rican (n=1,109) and Mexican American (n=649) children in the Genes-Environments and Admixture in Latino Americans (GALA II) study (Oh et al. 2012), African American adults in the Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE, n=3,428) (Levin et al. 2014) and African American children in
Genetics of Complex Pediatric Disorders (GCPD-A, n=1,464) study (Ong et al. 2013). Age, sex, height, controller medications and the first 5 PCs were used as covariates.

Additionally, replication of GWAS FEV$_1$ associations was attempted using data of black UK Biobank subjects who had asthma (n=627) while adjusting for age, sex, height and the first 5 principal components. Asthma status was defined by ICD code or self-reported asthma. UK Biobank genotype data was generated on Affymetrix UK BiLEVE axiomatic or UK Biobank Axiom array and imputed into the Haplotype Reference Consortium, 1000G and UK 10K projects (Bycroft et al. 2018; Canela-Xandri et al. 2018). Additional details on the UK Biobank study and the replication procedures are available in File S1 (Supplementary Text 2).

**RNA sequencing and expression quantitative trait loci (eQTL) analysis**

Whole-transcriptome libraries of 370 nasal brushings from GALA II Puerto Rican children with asthma were constructed by using the Beckman Coulter FX automation system (Beckman Coulter, Fullerton, CA). Libraries were sequenced with the Illumina HiSeq 2500 system. Raw RNA-Seq reads were trimmed using Skewer (Jiang et al. 2014) and mapped to human reference genome hg38 using Hisat2 (Kim, D. et al. 2015). Reads mapped to genes were counted with htseq-count and using the UCSC hg38 GTF file as reference (Anders et al. 2015). Cis-expression quantitative trait locus (eQTL) analysis of KITLG was performed as described in the Genotype-Tissue Expression (GTEx) project version 7 protocol (GTEx Consortium et al. 2017) using age, sex, BMI, global African and European ancestries and 60 PEER factors as covariates.
Gene-by-air-pollution interaction analysis

We hypothesized that the effect of genetic variation on lung function in our study population may differ by the levels of exposure to SO₂ (Neophytou et al. 2016). To test for an interaction between a genetic variant and SO₂, an additional multiplicative interaction term (variant x SO₂ exposure) was included in the original GWAS model (see Method Section “FEV₁ GWAS”). The SO₂ estimates used in the interaction analysis were first-year, past-year, and lifetime exposure to ambient SO₂, which were estimated as described previously (Neophytou et al. 2016). Residuals of FEV₁ were plotted against exposure to SO₂ and stratified by the number of copies of the minor allele. Residuals of FEV₁ were obtained as described in the Methods Section “FEV₁ GWAS”.

Data availability

Local institutional review boards approved the studies (IRB# 10-02877). All subjects and legal guardians provided written informed consent. TOPMed whole genome sequencing data are available on dbGaP with the accession number: phs000921.v4.p1. The remaining sequence data are available upon request. Supplemental materials available at figshare.

RESULTS

Novel lung function associations

Subject characteristics of the 867 African American children with asthma included in this study are shown in Table 1, and the distribution of their FEV₁ values is in Figure S1. The CODA-adjusted statistical significance threshold (2.10 x 10⁻⁸) was selected as the genome-wide
significant level. According to this threshold, one SNP in chromosome 12 (chr12:88846435, rs73429450, G>A) was associated with FEV₁.res.rnorm (Figure 1, \( p = 9.01 \times 10^{-9} \), \( \beta = 0.801 \)). The association between rs73429450 and lung function remained statistically significant when the association was repeated using untransformed FEV₁ (\( p = 1.26 \times 10^{-8} \), \( \beta = 0.302 \)) as the outcome variable. The association between rs73429450 and lung function was suggestive using FEV₁.perc.predicted (\( p = 1.69 \times 10^{-7} \), \( \beta = 0.100 \)). Thirty-nine variants with association \( p < 0.001 \) overlapped with 14 out of 98 previously reported FEV₁-associated loci and their 20 Kb flanking regions (Table S1). When including all variants with association \( p < 0.05 \), we found overlap with all 98 previously reported loci. Our top FEV₁ association, rs73429450, did not overlap with any previously reported loci in both scenarios and it is a novel association with FEV₁ in this study population.

Secondary analysis that included covariates correcting for smoking status and number of smokers in the family showed that smoking-related factors were not significantly associated with FEV₁ in our pediatric SAGE cohort: using 657 out of 867 individuals with available smoking-related covariates, the FEV₁.res.rnorm association p-values before and after including the smoking-related covariates were \( 2.01 \times 10^{-6} \) and \( 1.89 \times 10^{-6} \). Both p-values of the covariates smoking status (\( p = 0.27 \)) and number of smokers in the family (\( p = 0.54 \)) were not significant.

Conditional analysis was performed on 45 variants with association \( p < 1 \times 10^{-4} \) located within 1 Mb of the strongest association signal (rs73429450). Two weaker independent signals (rs11312747, rs58475486) were identified (Table S2).
The minor allele frequency of rs73429450 in continental populations from the 1000 Genomes Project (1000G) is 3% in Africans (AFR) and < 1% in Admixed Americans (AMR), Europeans (EUR) and Asians (EAS and SAS) (1000 Genomes Project Consortium et al. 2015). Rs73429450 was not included on the Affymetrix LAT1 genotyping array where SAGE participants were previously genotyped. To determine if the rs73429450 association with FEV1 was only identifiable using whole genome sequencing data, we attempted to reproduce our results by imputing the genotype of rs73429450 in 851 SAGE participants with available array data using 1000G phase 3 (n = 2,504), HRC r1.1 (n = 32,470), CAAPA (n = 883) and TOPMed freeze 5 (n = 62,784) reference panels. Our results remained statistically significant when using the 1000G phase 3 (p = 4.97 x 10^-8, β = 0.79, imputation R^2 = 0.95) and TOPMed freeze 5 (p = 1.22 x 10^-8, β = 0.80, imputation R^2 = 0.98) reference panels, but lost statistical significance when rs73429450 genotypes were imputed using the HRC (p = 4.35 x 10^-7, β = 0.68, imputation R^2 = 0.94) and CAAPA (p = 1.95 x 10^-7, β = 0.80, imputation R^2 = 0.71) reference panels.

Region-based association analysis including all variants conditioned on the association signal from rs73429450 was performed in its 1 Mb flanking region (chr12:87846435-89846435). No windows were significantly associated after Bonferroni multiple testing correction (p < 2.81 x 10^-4, Figure S2), but 20 windows were suggestively associated with FEV1.res.rnorm (p < 5.62 x 10^-3, Table S3). Two of 20 windows re-tested using only functional variants were suggestively significant (region 4 and 16). Both of these windows were no longer suggestively significant after removing the common variants, indicating that association signal from these regions was mostly driven by common variants. Further investigation on region 16 using drop-one analysis on the 2 rare and 1 common function variants confirmed the major contribution by the
common variant, rs1895710, as shown by the major increase in p-value (Table S4). The signal
was also slightly driven by the singleton, rs990979778. Drop-one analysis was not performed on
region 4 because there were only 1 common and 1 rare variants.

A Hi-C assay couples a chromosome conformation capture (3C) assay with next-generation
sequencing to capture long-range interactions in the genome. We identified a statistically
significant long-range chromatin interaction between the GWAS peak and the KIT ligand (KITLG,
also known as stem cell factor, SCF) gene in human fetal lung fibroblast cell line IMR90 (Table
S5). The long-range interaction detected in human primary lung tissue was not significant,
implying that the potential long-range interactions are specific to tissue type or developmental
stage.

**Potential regulatory role of FEV₁-associated variants on KITLG**

expression

To further elucidate potential regulatory relationships between the GWAS association peak and
KITLG, we analyzed whether variants in the peak were eQTL of KITLG in previously published
whole blood RNA-Seq data available from the same study participants (Mak, Angel CY et al.
2016). The whole blood RNA-Seq data, however, did not yield evidence of expressed KITLG,
consistent with results in GTEx. We subsequently used RNA-Seq data from nasal epithelial cells
of 370 Puerto Rican children with asthma from the GALA II study, and found that five out of 45
variants were eQTL of KITLG (Table S6). While Puerto Ricans are a different population than
African Americans, they are both admixed populations with substantial African genetic
ancestry, and therefore could share eQTLs. All five eQTLs corresponded to one signal in a region with strong linkage disequilibrium ($r^2 > 0.8$, markers 25, 26, 28, 29 and 32 in Figure 2).

**Replication of genetic association with FEV$_1$**

Subject Characteristics of our four replication cohorts (SAPPHIRE, GCPD-A, UK Biobank and GALA II) are shown in Table S7. We attempted to replicate the association of the 45 SNPs in our primary FEV$_1$ GWAS in each cohort. We used 0.05 as the suggestive p-value threshold and 0.0167 as the Bonferroni-corrected p-value threshold after correcting for 3 independent signals (see conditional analysis in Results Section). A total of 20 variants were replicated at $p < 0.05$ with consistent direction of effect in black UK Biobank participants; 14 variants in SAPPHIRE and 2 variants in GCPD-A were significant but had an opposite direction of effect (Table S8).

We attempted to replicate the FEV$_1$.res.rnorm association in Mexican American (n = 649) and Puerto Rican (n = 1,109) children with asthma from the GALA II study. In Mexican Americans, we excluded 19 variants with MAF < 0.1% and associations for the remaining 26 variants did not replicate (Table S9). In Puerto Ricans, we observed the same protective effect in 38 of the 45 variants in the locus, but the associations were not statistically significant (Table S9).

**Incorporating statistical and functional evidence for candidate variant prioritization**

We combined and summarized all functional evidence for the top 45 variants, along with eQTL findings from nasal epithelial RNA-Seq and replication results (Figure 2, Table 2). To facilitate interpretation of the variant association with FEV$_1$, the effect sizes and p-values of both FEV$_1$ ($\beta$...
and \( \beta_{\text{norm}} \) and \( p_{\text{norm}} \) associations are reported in Table 2. CADD functional prediction score and ENCODE histone modification ChIP-Seq peaks in embryonic, blood, bone marrow, and lung-related tissues were also examined but not reported because none of the variants had a CADD score greater than 10 and none overlapped with histone modification sites. After summing up the number of evidence present for each variant to obtain its priority score, rs73440122 received the highest functional score of 3 based on replication in the UK Biobank, overlap with a DNase I hypersensitivity site in B-lymphoblastoid cells (GM12865) and overlap with an SPI1 binding site in acute promyelocytic leukemia cells. Eight other variants were prioritized with score > 2 or evidence of being an eQTL for KITLG in nasal epithelial cells (Table 2, score marked with * or # respectively). These nine candidate variants were selected for gene-by-air-pollution interaction analyses.

**Gene-by-air-pollution interaction of rs58475486**

*We previously found that first year of life and lifetime exposure to \( \text{SO}_2 \) were associated with \( \text{FEV}_1 \) in African American children (Neophytou et al. 2016). We investigated whether the effect of the nine prioritized genetic variants associated with lung function varied by \( \text{SO}_2 \) exposure (first year of life, past year, and lifetime exposure). Since the nine variants represent three independent signals (see conditional analysis in the Data availability)*

Local institutional review boards approved the studies (IRB# 10-02877). All subjects and legal guardians provided written informed consent. TOPMed whole genome sequencing data are
available on dbGaP with the accession number: phs000921.v4.p1. The remaining sequence data are available upon request. Supplemental materials available at figshare.

R Section), the Bonferroni-corrected p-value threshold was set to $p = 0.0056$ (correction for nine tests; three signals and three exposure periods to SO$_2$). We observed a single statistically significant interaction between the T allele of rs58475486 and past year exposure to SO$_2$ that was positively associated with FEV$_1$ ($p = 0.003$, $\beta = 0.32$, Table 3, Figure 3A). Interestingly, six of the remaining eight variants also displayed interaction effects with past year exposure to SO$_2$ that were suggestively associated ($p < 0.05$) with FEV$_1$ (Table 3). We also found a suggestive interaction of the C allele of rs73440122 with first year exposure to SO$_2$ that was associated with decreased FEV$_1$ ($p = 0.045$, $\beta = -0.32$, Figure 3B).

**DISCUSSION**

Variant rs73429450 (MAF = 0.030) was identified as the strongest association signal with FEV$_1$. Each additional copy of the protective A allele of rs73429450 was associated with a 0.3 L increase of FEV$_1$. We did not find any statistically significant contribution of rare variants to the association signal from a 1 Kb sliding window analyses in the 1 MB flanking region centered on rs73429450. We were surprised to identify a novel common variant (MAF = 0.030) associated with lung function using whole genome sequence data in a population that was previously analyzed for associations with lung function using genotype array data. Further investigation revealed that our discovered variant, rs73429450, was not captured by the LAT 1 genotyping array, and the association with lung function depended on the reference panel used to impute the variant into our population. More surprisingly, our statistically significant finding was only...
found to be suggestively significant using data imputed from the CAAPA reference panel (p = 1.95 x 10^{-7}, \beta = 0.80). Of the imputation reference panels that we assessed, CAAPA is one of the more relevant reference panels for our study population because it is based on African populations in the Americas. However, we note that the effect size estimated from CAAPA-imputed data was comparable to that generated from WGS data. While whole genome sequencing data is usually praised for enabling analysis of rare-variant contributions to phenotype variability, our results show the utility of whole genome sequencing data for the reliable analysis of common variants as well in the absence of relevant imputation panels.

Although rs73429450 had the lowest p-value from our whole genome sequencing association analysis, we did not find the required amount of functional evidence to prioritize this marker for inclusion in downstream gene-by-air-pollution analyses. Another variant, rs73440122, was in moderate to strong linkage disequilibrium ($r^2 = 0.76$) with rs7349450 and had a similar MAF (0.027) in our study population, but was only suggestively associated with FEV$_1$ in our association analysis (p = 2.08 x 10^{-7}, Table2). In contrast to rs7349450, there were multiple lines of evidence suggesting the functional relevance of rs73440122: rs73440122 received the highest priority score, its association with FEV$_1$ replicated in black UK Biobank participants, and it was one of the most likely drivers of FEV$_1$ variability among individuals, possibly mediated through KITLG.

Bioinformatic interrogation of rs73440122 revealed that the variant overlapped with a ccRE (SCREEN accession EH37E0279310), DNase I hypersensitivity site, and SPI1 ChIP-Seq clusters that were indicative of a candidate open chromatin gene regulatory region. The transcription
factor SPI1, which is expressed in lung, spleen and whole blood, was previously linked to fibrotic
diseases and tissues (Wohlfahrt et al. 2019) and KITLG in erythroid progenitor cells (Quang et al.
1995), but there is no existing evidence of its role in lung function. Although rs73440122 itself is
not an eQTL of KITLG, it is located in a region that physically interacted with KITLG based on Hi-
C data in fetal lung fibroblast cells. Additionally, five neighboring FEV1 associated variants were
identified as eQTLs of KITLG, although they also appeared to be an independent signal ($r^2 < 0.2$).
Overall, these results support regulatory interactions between this region and KITLG.

KITLG, a ligand of the KIT tyrosine kinase receptor, is expressed in lung and its decreased
expression was previously observed in patients with chronic obstructive pulmonary disease
(Bhattacharya et al. 2009). Inactivation of KIT signaling led to airspace enlargement and
contributed to declining lung function in mice (Lindsey et al. 2011). Polymorphisms of KITLG
have been previously associated with susceptibility to moderate-to-severe bronchopulmonary
dysplasia, a chronic inflammatory lung disease that affects preterm infants (Huusko et al. 2015).
These lines of evidence support the potential contribution of our novel locus to lung function
phenotype variation mediated through KITLG, especially in patients with inflammatory
conditions like asthma.

Gene-by-environment interactions likely account for a portion of the “missing” heritability of
many complex phenotypes (Moore and Williams. 2009). We hypothesized that a significant
portion of the heritability of lung function was due in part to gene-by-air-pollution (SO2)
interaction effects. The interaction between rs58475486 and past year exposure to SO2 that
was significantly associated with lung function supports our hypothesis. The T allele of
rs58475486 is common (8-14%) in African populations and showed a protective effect on lung function in the presence of past year SO₂ exposure. SNP rs58475486 is located in a ccRE (SCREEN accession EH37E0279296) and a FOXA1 binding site in the A549 lung adenocarcinoma cell line. Foxa1 has a known compensatory role with Foxa2 during lung morphogenesis in mice (Wan et al. 2005). Deletion of both Foxa1 and Foxa2 inhibited cell proliferation, epithelial cell differentiation, and branching morphogenesis in fetal lung tissue. Further functional validation on the effect of rs58475486 on binding affinity of FOXA1 is necessary to confirm whether the role of FOXA1 in this ccRE is important for KITLG regulatory and lung function.

The higher frequency of the protective alleles of both rs73440122 and rs58465486 in African populations appears to contradict previous findings that African ancestry was associated with lower lung function (Kumar et al. 2010). One possible explanation for this seeming inconsistency is that FEV₁ is a complex trait whose variation is influenced by many genetic variants of small to moderate effect sizes whose influence on lung function may vary by exposure to environmental factors. We found suggestive evidence that the interaction between rs73440122 and first year exposure to SO₂ reverses the positive association of rs73440122 with lung function to a negative one (Table 3). When assessed independently, our genetic association analysis showed that the protective A allele of rs73440122 was associated with higher lung function. However, with increasing levels of SO₂ exposure in the first year of life, increasing copies of the A allele of rs73440122 were associated with decreased lung function. Air pollution is known to negatively impact lung function, and we have previously shown that the deleterious effects of air pollution on lung phenotypes may be significantly increased in African American children compared to other populations experiencing the same amount of
exposure (Nishimura et al. 2013). It has also been reported that Latino and African American populations often live in neighborhoods with high levels of air pollution (Mott 1995). The increased susceptibility to negative pulmonary effects from air pollution exposure coupled with the disproportionate exposure to air pollution experienced by the African American population may also contribute to the lower lung function seen in this population despite the presence of protective alleles.

One limitation of this study is that the FEV$_1$ genetic association and the eQTL analyses with $KITLG$ were performed in different populations due to data availability constraints. Although we did not have RNA-Seq data from lung tissues from our study subjects, we previously demonstrated that there is a high degree of overlap in gene expression profiles between nasal and bronchial epithelial cells (Poole et al. 2014). The direction of effect of the association was the same in GALA II Puerto Rican children with asthma but not statistically significant. This may in part due to the significantly lower African Ancestry in Puerto Ricans compared to African Americans.

We replicated 20 of 45 variants in black UK Biobank subjects and observed conflicting “flip-flop” associations in African Americans from the SAPPHIRE and GCPD-A studies. In the past, flip-flop associations were deemed as spurious results. Traditional association testing approach studies the effect of each variant on phenotype independently and increases the chance of flip-flop associations detected between studies. Differences in study design, sampling variation that leads to variation in LD patterns, and lack of consideration of other disease influencing genetic and/or environmental factors are all potential causes of flip-flop associations (Kraft et al. 2009;
Lin et al. 2007). Hence, it is not surprising to observe flip-flop associations when gene and environment interactions were detected at our FEV\textsubscript{1} GWAS locus. It was previously shown that flip-flop associations can occur between and within populations even in the presence of a genuine genetic effect (Kraft et al. 2009; Lin et al. 2007). Further functional analysis is thus required to validate the relationship between the candidate variants, KITLG and FEV\textsubscript{1}. This may include reporter assays to validate potential enhancer or repressor activity and CRISPR-based editing assays to validate the regulatory role of the candidate variants on KITLG. Although literature exists describing KIT signaling for lung function in mice (Lindsey et al. 2011), additional knockout experiments in a model animal system may be necessary to confirm the causal effect of KITLG on lung function.

The average concentration of ambient SO\textsubscript{2} exposure in our participants (Table 1) was lower than the National Ambient Air Quality Standards. It is possible that SO\textsubscript{2} acted as a surrogate for other unmeasured toxic pollutants emitted from local point sources. Major sources of SO\textsubscript{2} in San Francisco Bay Area during the recruitment years of 2006 to 2011 include airports, petroleum refineries, gas and oil plants, calcined petroleum coke plants, electric power plants, cement manufacturing factories, chemical plants, and landfills (United States Environmental Protection Agency 2008; United States Environmental Protection Agency 2011). The Environmental Protection Agency’s national emissions inventory data also showed that these facilities emit Volatile Organic Compounds, heavy metals (lead, mercury, chromium, arsenic), formaldehyde, ethyl benzene, acrolein, 1,3-butadiene, 1,4-dichlorobenzene, and tetrachloroethylene into the air along with SO\textsubscript{2}. These chemicals are highly toxic and inhaling
even a small amount may contribute to poor lung function. Another possibility is that exposure
to SO$_2$ captured unmeasured confounding socioeconomic factors.

This study identified a novel protective allele for lung function in African American children with
asthma. The protective association with lung function intensified with increased past year
exposure to SO$_2$. Our findings showcases the complexity of the relationship between genetic
and environmental factors impacting variation in FEV$_1$, highlights the utility of WGS data for
genetic research of complex phenotypes, and underscores the importance of including diverse
study populations in our exploration of the genetic architecture underlying lung function.

ACKNOWLEDGEMENTS

The Genes-Environments and Admixture in Latino Americans (GALA II) Study, the Study of
African Americans, Asthma, Genes and Environments (SAGE) Study and E.G.B. were supported
by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical
Faculty Development Program, the Harry Wm. and Diana V. Hind Distinguished Professor in
Pharmaceutical Sciences II, the National Heart, Lung, and Blood Institute (NHLBI)
[R01HL117004, R01HL128439, R01HL135156, X01HL134589]; the National Institute of
Environmental Health Sciences [R01ES015794]; the National Institute on Minority Health and
Health Disparities (NIMHD) [P60MD006902, R01MD010443], the National Human Genome
Research Institute [U01HG009080] and the Tobacco-Related Disease Research Program [24RT-
0025]. MJW was supported by the NHLBI [K01HL140218]. JJ and BEH were supported by the
NHLBI [R01HL133433, R01HL141992]. KLK was supported by the NHLBI [R01HL135156-S1], the
UCSF Bakar Institute, the Gordon and Betty Moore Foundation [GBMF3834], and the Alfred P.
Sloan Foundation [2013-10-27] grant to UC Berkeley through the Moore-Sloan Data Science Environment Initiative. ACW was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development [1R01HD085993-01].

The SAPPHIRE study was supported by the Fund for Henry Ford Hospital, the American Asthma Foundation, the NHLBI [R01HL118267, R01HL141485, X01HL134589], the National Institute of Allergy and Infectious Diseases [R01AI079139], and the National Institute of Diabetes and Digestive and Kidney Diseases [R01DK113003].

The GCPD-A study was supported by an Institutional award from the Children’s Hospital of Philadelphia and by the NHLBI [X01HL134589].

Part of this research was conducted using the UK Biobank Resource under Application Number 40375. We would like to thank UK Biobank participants and researchers who contributed or collected data.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Gene-Environment, Admixture and Latino Asthmatics Study" (phs000920) and "NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environments" (phs000921) was performed at the New York Genome Center (3R01HL117004-02S3) and the University of Washington Northwest Genomics Center (HHSN268201600032I). WGS for "NHLBI TOPMed: Study of Asthma Phenotypes & Pharmacogenomic Interactions by Race-Ethnicity" (phs001467) and “Genetics of Complex Pediatric Disorders - Asthma” (phs001661) was performed at the
University of Washington Northwest Genomics Center (HHSN268201600032I). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

WGS of part of GALA II was performed by New York Genome Center under The Centers for Common Disease Genomics of the Genome Sequencing Program (GSP) Grant (UM1 HG008901). The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. GSP is funded by the National Human Genome Research Institute, the National Heart, Lung, and Blood Institute, and the National Eye Institute.

The TOPMed imputation panel was supported by the NHLBI and TOPMed study investigators who contributed data to the reference panel. The panel was constructed and implemented by the TOPMed Informatics Research Center at the University of Michigan (3R01HL-117626-02S1; contract HHSN268201800002I). The TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I) provided additional data management, sample identity checks, and overall program coordination and support. We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.
The authors wish to acknowledge the following GALA II and SAGE study collaborators: Shannon Thyne, UCSF; Harold J. Farber, Texas Children's Hospital; Denise Serebrisky, Jacobi Medical Center; Rajesh Kumar, Lurie Children's Hospital of Chicago; Emerita Brigino-Buenaventura, Kaiser Permanente; Michael A. LeNoir, Bay Area Pediatrics; Kelley Meade, UCSF Benioff Children's Hospital, Oakland; William Rodríguez-Cintrón, VA Hospital, Puerto Rico; Pedro C. Ávila, Northwestern University; Jose R. Rodríguez-Santana, Centro de Neumología Pediátrica; Luisa N. Borrell, City University of New York; Adam Davis, UCSF Benioff Children's Hospital, Oakland; Saunak Sen, University of Tennessee.

The authors acknowledge the families and patients for their participation and thank the numerous health care providers and community clinics for their support and participation in GALA II and SAGE. In particular, the authors thank the recruiters who obtained the data: Duanny Alva, MD; Gaby Ayala-Rodríguez; Lisa Caine, RT; Elizabeth Castellanos; Jaime Colón; Denise DeJesus; Blanca López; Brenda López, MD; Louis Martos; Vivian Medina; Juana Olivo; Mario Peralta; Esther Pomares, MD; Jihan Quraishi; Johanna Rodríguez; Shahdad Saeedi; Dean Soto; and Ana Taveras.

The authors thank María Pino-Yanes for providing feedback on this study and Thomas W Blackwell for providing critical review on this manuscript.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
LITERATURE CITED


eCollection 2014.


GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)-Analysis Working Group, Statistical Methods groups-Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund et al, 2017 Genetic effects on gene expression across human tissues. Nature 550: 204-213.


Mak, A. C., M. J. White, C. Eng, D. Hu, S. Huntsman et al, 2016 Whole Genome Sequencing to Identify Genetic Variation Associated with Bronchodilator Response in Minority Children with Asthma.


Quang, C. T., M. Pironin, M. von Lindern, H. Beug and J. Ghysdael, 1995 Spi-1 and mutant p53 regulate different aspects of the proliferation and differentiation control of primary erythroid progenitors. Oncogene 11: 1229-1239.


Wise, J., 2019 Air pollution is linked to infant deaths and reduced lung function in children. BMJ 366: i5772.


Figure 1.

A.

B.
Figure 1. Manhattan and LocusZoom plots from genome-wide association study of lung function*. (A) Manhattan plot from genome-wide association study of lung function* using linear regression in ENCORE. Red horizontal line: CODA-adjusted genome-wide significance p-value of $2.10 \times 10^{-8}$. Blue horizontal line: CODA-adjusted suggestive significance p-value of $4.19 \times 10^{-7}$. (B) LocusZoom plot of rs73429450 (chr12: 88846435) and 500 Kb flanking region. Colors show linkage disequilibrium in the study population. * FEV$_1$.res.rnorm was used as the phenotype for the association testing.
Figure 2. Integration of statistical and functional evidence for variant prioritization. Numbers and different shades of black in the LD plot represent LD in $R^2$. The three independent signals identified in the conditional analysis are marked with *. Nasal eQTL, variants eQTL of $KITLG$ in nasal epithelial cells. ccREs, candidate cis-regulatory elements in SCREEN registry. ENCORE, DNase I hypersensitivity site and/or transcription factor ChIP-Seq overlapping with the variants. UK Biobank, SAPPHIRE, GCPD-A, replication results using Blacks in UK Biobank and African Americans in the SAPPHIRE and GCPD-A cohorts (R = replicated at $p < 0.05$; F = flip-flop association at $p < 0.05$). Candidate, candidate variants prioritized because of presence of two or more evidence or is nasal eQTL. + indicates presence of evidence.
Figure 3. Gene-by-environment interaction analysis on FEV₁. FEV₁ residuals, residuals after FEV₁ was regressed on the covariates age, sex, height, controller medications, sequencing centers and the first 5
genetic PCs. FEV\textsubscript{1} residuals was plotted against (A) past year exposure to SO\textsubscript{2} stratified by the number of copies of T allele of rs58475486, (B) first year of life exposure to SO\textsubscript{2} stratified by the number of copies of C allele of rs73440122.
Table 1. Descriptive characteristics of 867 African American children with asthma included in this study.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>African American (n=867)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>14.1 (3.64)</td>
</tr>
<tr>
<td>Median [Min, Max]</td>
<td>13.8 [8.01, 21.8]</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>439 (50.6%)</td>
</tr>
<tr>
<td>Female</td>
<td>428 (49.4%)</td>
</tr>
<tr>
<td><strong>Height (m)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>1.58 (0.145)</td>
</tr>
<tr>
<td>Median [Min, Max]</td>
<td>1.60 [1.14, 1.92]</td>
</tr>
<tr>
<td><em><em>Any control medications</em> in last 2 weeks</em>*</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>543 (62.6%)</td>
</tr>
<tr>
<td>Yes</td>
<td>324 (37.4%)</td>
</tr>
<tr>
<td><strong>ICS in last 2 weeks</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>211 (24.3%)</td>
</tr>
<tr>
<td>Yes</td>
<td>306 (35.3%)</td>
</tr>
<tr>
<td>Missing</td>
<td>350 (40.4%)</td>
</tr>
<tr>
<td><strong>LABA in last 2 weeks</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>5 (0.6%)</td>
</tr>
<tr>
<td>Yes</td>
<td>94 (10.8%)</td>
</tr>
<tr>
<td>Missing</td>
<td>768 (88.6%)</td>
</tr>
<tr>
<td><strong>Leukotriene inhibitor in last 2 weeks</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>11 (1.3%)</td>
</tr>
<tr>
<td>Yes</td>
<td>68 (7.8%)</td>
</tr>
<tr>
<td>Missing</td>
<td>788 (90.9%)</td>
</tr>
<tr>
<td><strong>African ancestry</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>0.792 (0.129)</td>
</tr>
<tr>
<td>Median [Min, Max]</td>
<td>0.826 [0.118, 1.00]</td>
</tr>
<tr>
<td><strong>Smoking status</strong></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>793 (91.5%)</td>
</tr>
<tr>
<td>Past</td>
<td>72 (8.3%)</td>
</tr>
<tr>
<td>Current</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Missing</td>
<td>2 (0.2%)</td>
</tr>
<tr>
<td><strong>Number of smokers in family</strong></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>469 (54.1%)</td>
</tr>
<tr>
<td>1</td>
<td>137 (15.8%)</td>
</tr>
<tr>
<td>2</td>
<td>42 (4.8%)</td>
</tr>
<tr>
<td>3+</td>
<td>10 (1.2%)</td>
</tr>
<tr>
<td>Missing</td>
<td>209 (24.1%)</td>
</tr>
<tr>
<td><strong>SO2 first year exposure (ppb)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>1.59 (0.961)</td>
</tr>
<tr>
<td></td>
<td>Median [Min, Max]</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>SO2 past year exposure (ppb)</strong></td>
<td>1.50 [0.0, 14.5]</td>
</tr>
<tr>
<td><strong>Mean (SD)</strong></td>
<td>1.10 (0.302)</td>
</tr>
<tr>
<td><strong>Median [Min, Max]</strong></td>
<td>1.08 [0.120, 2.27]</td>
</tr>
<tr>
<td><strong>Missing</strong></td>
<td>206 (23.8%)</td>
</tr>
<tr>
<td><strong>SO2 lifetime exposure (ppb)</strong></td>
<td>1.50 [0.371]</td>
</tr>
<tr>
<td><strong>Mean (SD)</strong></td>
<td>1.10 (0.302)</td>
</tr>
<tr>
<td><strong>Median [Min, Max]</strong></td>
<td>1.47 [0.964, 7.39]</td>
</tr>
<tr>
<td><strong>Missing</strong></td>
<td>206 (23.8%)</td>
</tr>
</tbody>
</table>

**TOPMed sequencing center and phase**

- **phase 1, CAAPA**: 6 (0.7%)
- **phase 1, NYGC**: 460 (53.1%)
- **phase 3, NW**: 401 (46.3%)

Control medications include inhaled corticosteroid (ICS), long-acting beta-agonist (LABA), leukotriene inhibitor and/or ICS/LABA combo. SO$_2$ exposure are hourly exposure averaged over the specified time period before spirometry testing as previously described in Neophytou et al. 2016. ppb, parts per billion or $\mu$g/m$^3$.
### Table 2. Integration of statistical and function evidence for variant prioritization.

<table>
<thead>
<tr>
<th>rsID</th>
<th>Alt</th>
<th>β</th>
<th>p</th>
<th>MAF</th>
<th>ALL</th>
<th>AFR</th>
<th>AMR</th>
<th>EUR</th>
<th>ccRE</th>
<th>DNaseI</th>
<th>TFBS</th>
<th>Nasal eQTL</th>
<th>UKBB</th>
<th>SAPPHIRE</th>
<th>GCPD- A</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11835305</td>
<td>T</td>
<td>0.126</td>
<td>1.93E-05</td>
<td>0.320</td>
<td>3.69E-05</td>
<td>0.104</td>
<td>0.036</td>
<td>0.119</td>
<td>0.012</td>
<td>0.001</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>rs117015979</td>
<td>T</td>
<td>0.127</td>
<td>1.45E-05</td>
<td>0.323</td>
<td>2.81E-05</td>
<td>0.105</td>
<td>0.037</td>
<td>0.123</td>
<td>0.012</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>rs17015979</td>
<td>C</td>
<td>0.126</td>
<td>1.93E-05</td>
<td>0.320</td>
<td>3.69E-05</td>
<td>0.104</td>
<td>0.036</td>
<td>0.120</td>
<td>0.012</td>
<td>0.001</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>rs17015979</td>
<td>C</td>
<td>0.245</td>
<td>1.63E-06</td>
<td>0.654</td>
<td>1.06E-06</td>
<td>0.033</td>
<td>0.011</td>
<td>0.030</td>
<td>0.006</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>2*</td>
</tr>
<tr>
<td>rs58475486</td>
<td>T</td>
<td>0.252</td>
<td>8.19E-07</td>
<td>0.761</td>
<td>5.46E-07</td>
<td>0.033</td>
<td>0.010</td>
<td>0.030</td>
<td>0.006</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>rs17015979</td>
<td>T</td>
<td>0.127</td>
<td>1.45E-05</td>
<td>0.323</td>
<td>2.81E-05</td>
<td>0.105</td>
<td>0.037</td>
<td>0.123</td>
<td>0.012</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>rs58475486</td>
<td>T</td>
<td>0.252</td>
<td>8.19E-07</td>
<td>0.761</td>
<td>5.46E-07</td>
<td>0.033</td>
<td>0.010</td>
<td>0.030</td>
<td>0.006</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>rs17015979</td>
<td>T</td>
<td>0.127</td>
<td>1.45E-05</td>
<td>0.323</td>
<td>2.81E-05</td>
<td>0.105</td>
<td>0.037</td>
<td>0.123</td>
<td>0.012</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>rs17015979</td>
<td>T</td>
<td>0.127</td>
<td>1.45E-05</td>
<td>0.323</td>
<td>2.81E-05</td>
<td>0.105</td>
<td>0.037</td>
<td>0.123</td>
<td>0.012</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>rs17015979</td>
<td>T</td>
<td>0.127</td>
<td>1.45E-05</td>
<td>0.323</td>
<td>2.81E-05</td>
<td>0.105</td>
<td>0.037</td>
<td>0.123</td>
<td>0.012</td>
<td>0.001</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

*All rights reserved. No reuse allowed without permission. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. https://doi.org/10.1101/2020.02.20.20019588 doi: medRxiv preprint*
Candidate variants were prioritized if they had a functional score of greater than 2 (*) or if they are eQTL of KITLG in nasal epithelial cells (#). Indels were marked with the superscript i near the rsID. $\beta$ (p) and $\beta_{\text{norm}}$ ($p_{\text{norm}}$) are the effect sizes (p-values) of the genetic associations of the alternate allele (alt) with FEV$_1$ and FEV$_1$-res.rnorm respectively. MAF minor allele frequency. ALL/AFR/AMR/EUR, 1000 Genomes minor allele frequency from all/African/American/European populations. ccRE, candidate cis-regulatory elements derived from ENCODE data and stored in the SCREEN web interface. DNasel, ENCODE DNase I hypersensitivity clusters derived from assays in 95 cell (UCSC hg38 wgEncodeRegDnaseClustered table) in embryonic, blood, bone marrow or lung-related cell lines. TFBS, ENCODE transcription factor ChIP-Seq clusters representing 340 transcription factors and 129 cell types (UCSC hg38 encRegTfbsClustered table). Nasal eQTL, KITLG eQTLs in nasal epithelial cells. UKBB/SAPPHIRE/GCPD-A, replication association analysis of FEV$_1$ in UK Biobank Blacks and African Americans in SAPPHIRE and GCPD-A (details results in Supplementary Table 3). Score, priority score. GM12865, a B-lymphoblastoid cell line. HL-60, an acute promyelocytic
leukemia cell line. A549, a lung adenocarcinoma cell line. -, not available; FF, flip-flop association defined as replication association ($p < 0.05$) of the opposite allele instead of the original allele identified in the discovery cohort.
<table>
<thead>
<tr>
<th>Variant</th>
<th>Exposure</th>
<th>β.variant</th>
<th>β.env</th>
<th>β.gxe</th>
<th>p.variant</th>
<th>p.env</th>
<th>p.gxe</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs58475486_T</td>
<td>SO₂ first year</td>
<td>0.18</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0.119</td>
<td>0.009</td>
<td>0.658</td>
</tr>
<tr>
<td>rs57692452_C</td>
<td>SO₂ first year</td>
<td>0.61</td>
<td>-0.04</td>
<td>-0.24</td>
<td>0.006</td>
<td>0.007</td>
<td>0.091</td>
</tr>
<tr>
<td>rs111289668_G</td>
<td>SO₂ first year</td>
<td>0.68</td>
<td>-0.04</td>
<td>-0.27</td>
<td>0.003</td>
<td>0.007</td>
<td>0.079</td>
</tr>
<tr>
<td>rs73440122_C</td>
<td>SO₂ first year</td>
<td>0.73</td>
<td>-0.04</td>
<td>-0.32</td>
<td>0.002</td>
<td>0.008</td>
<td>0.045 **</td>
</tr>
<tr>
<td>rs17016065_G</td>
<td>SO₂ first year</td>
<td>0.23</td>
<td>-0.04</td>
<td>-0.08</td>
<td>0.004</td>
<td>0.027</td>
<td>0.108</td>
</tr>
<tr>
<td>rs17016066_A</td>
<td>SO₂ first year</td>
<td>0.23</td>
<td>-0.04</td>
<td>-0.08</td>
<td>0.004</td>
<td>0.027</td>
<td>0.108</td>
</tr>
<tr>
<td>rs1398303_A</td>
<td>SO₂ first year</td>
<td>0.23</td>
<td>-0.04</td>
<td>-0.09</td>
<td>0.004</td>
<td>0.030</td>
<td>0.082</td>
</tr>
<tr>
<td>rs61924868_T</td>
<td>SO₂ first year</td>
<td>0.22</td>
<td>-0.04</td>
<td>-0.08</td>
<td>0.004</td>
<td>0.029</td>
<td>0.088</td>
</tr>
<tr>
<td>rs73429415_A</td>
<td>SO₂ first year</td>
<td>0.23</td>
<td>-0.04</td>
<td>-0.09</td>
<td>0.004</td>
<td>0.031</td>
<td>0.069</td>
</tr>
<tr>
<td>rs58475486_T</td>
<td>SO₂ past year</td>
<td>-0.21</td>
<td>-0.04</td>
<td>0.32</td>
<td>0.084</td>
<td>0.478</td>
<td>0.003 **</td>
</tr>
<tr>
<td>rs57692452_C</td>
<td>SO₂ past year</td>
<td>-0.09</td>
<td>0.01</td>
<td>0.29</td>
<td>0.680</td>
<td>0.839</td>
<td>0.100</td>
</tr>
<tr>
<td>rs111289668_G</td>
<td>SO₂ past year</td>
<td>-0.18</td>
<td>0.01</td>
<td>0.41</td>
<td>0.444</td>
<td>0.869</td>
<td>0.037 *</td>
</tr>
<tr>
<td>rs73440122_C</td>
<td>SO₂ past year</td>
<td>-0.16</td>
<td>0.01</td>
<td>0.39</td>
<td>0.482</td>
<td>0.825</td>
<td>0.051</td>
</tr>
<tr>
<td>rs17016065_G</td>
<td>SO₂ past year</td>
<td>-0.11</td>
<td>-0.03</td>
<td>0.20</td>
<td>0.299</td>
<td>0.583</td>
<td>0.026 *</td>
</tr>
<tr>
<td>rs17016066_A</td>
<td>SO₂ past year</td>
<td>-0.11</td>
<td>-0.03</td>
<td>0.20</td>
<td>0.299</td>
<td>0.583</td>
<td>0.026 *</td>
</tr>
<tr>
<td>rs1398303_A</td>
<td>SO₂ past year</td>
<td>-0.12</td>
<td>-0.04</td>
<td>0.21</td>
<td>0.231</td>
<td>0.521</td>
<td>0.021 *</td>
</tr>
<tr>
<td>rs61924868_T</td>
<td>SO₂ past year</td>
<td>-0.12</td>
<td>-0.04</td>
<td>0.21</td>
<td>0.248</td>
<td>0.534</td>
<td>0.023 *</td>
</tr>
<tr>
<td>rs73429415_A</td>
<td>SO₂ past year</td>
<td>-0.12</td>
<td>-0.04</td>
<td>0.20</td>
<td>0.246</td>
<td>0.539</td>
<td>0.026 *</td>
</tr>
<tr>
<td>rs58475486_T</td>
<td>SO₂ lifetime</td>
<td>-0.24</td>
<td>-0.13</td>
<td>0.26</td>
<td>0.390</td>
<td>0.001</td>
<td>0.173</td>
</tr>
<tr>
<td>rs57692452_C</td>
<td>SO₂ lifetime</td>
<td>-0.44</td>
<td>-0.13</td>
<td>0.47</td>
<td>0.436</td>
<td>0.001</td>
<td>0.221</td>
</tr>
<tr>
<td>rs111289668_G</td>
<td>SO₂ lifetime</td>
<td>-0.19</td>
<td>-0.13</td>
<td>0.32</td>
<td>0.764</td>
<td>0.001</td>
<td>0.444</td>
</tr>
<tr>
<td>rs73440122_C</td>
<td>SO₂ lifetime</td>
<td>-0.15</td>
<td>-0.13</td>
<td>0.29</td>
<td>0.809</td>
<td>0.001</td>
<td>0.489</td>
</tr>
<tr>
<td>rs17016065_G</td>
<td>SO₂ lifetime</td>
<td>0.40</td>
<td>-0.10</td>
<td>-0.19</td>
<td>0.043</td>
<td>0.013</td>
<td>0.143</td>
</tr>
<tr>
<td>rs17016066_A</td>
<td>SO₂ lifetime</td>
<td>0.40</td>
<td>-0.10</td>
<td>-0.19</td>
<td>0.043</td>
<td>0.013</td>
<td>0.143</td>
</tr>
<tr>
<td>rs1398303_A</td>
<td>SO₂ lifetime</td>
<td>0.36</td>
<td>-0.11</td>
<td>-0.17</td>
<td>0.064</td>
<td>0.012</td>
<td>0.184</td>
</tr>
<tr>
<td>rs61924868_T</td>
<td>SO₂ lifetime</td>
<td>0.35</td>
<td>-0.11</td>
<td>-0.17</td>
<td>0.070</td>
<td>0.012</td>
<td>0.199</td>
</tr>
<tr>
<td>rs73429415_A</td>
<td>SO₂ lifetime</td>
<td>0.34</td>
<td>-0.11</td>
<td>-0.16</td>
<td>0.078</td>
<td>0.012</td>
<td>0.207</td>
</tr>
</tbody>
</table>

*, p < 0.05; **, p < bonferroni p value of 0.0056.