TITLE

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

AUTHORS

Angel CY Mak[1,*], Satria Sajuthi[2,^], Jaehyun Joo[3,^], Shujie Xiao[4,^], Patrick M Sleiman[5,6,^], Marquitta J White[1,^], Eunice Y Lee[1], Benjamin Saef[3], Donglei Hu[1], Hongsheng Gui[4], Kevin L Keys[1,7], Fred Lurmann[8], Deepti Jain[9], Gonçalo Abecasis[10], Hyun Min Kang[10], Deborah A. Nickerson[11,12,13], Soren Germer[14], Michael C Zody[14], Lara Winterkorn[14], Catherine Reeves[14], Scott Huntsman[1], Celeste Eng[1], Sandra Salazar[1], Sam S Oh[1], Frank D Gilliland[15], Zhanghua Chen[15], Rajesh Kumar[16], Fernando D Martínez[17], Ann Chen Wu[18], Elad Ziv[1], Hakon Hakonarson[5,6,#], Blanca E Himes[3,#], L Keoki Williams[4,#], Max A Seibold[3,#], Esteban G. Burchard[1,20]

[1] Department of Medicine, University of California San Francisco, San Francisco, CA, USA

[2] Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA

[3] Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[4] Center for Individualized and Genomic Medicine Research, Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA

[5] Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

[6] Department of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

[7] Berkeley Institute for Data Science, University of California, Berkeley, CA, USA

[8] Sonoma Technology Inc, Petaluma, CA, USA

[9] Department of Biostatistics, University of Washington, Seattle, WA, USA

[10] Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA

[11] Department of Genome Sciences, University of Washington, Seattle, WA, USA

[12] Northwest Genomics Center, Seattle, WA, USA

[13] Brotman Baty Institute, Seattle, WA, USA

[14] New York Genome Center, New York, NY, USA

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

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

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

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

[19] 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/performed 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 (FEV_1), 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 FEV_1 variation and that African Americans have lower FEV_1 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 FEV_1 variation among individuals with asthma could lead to improved patient care and therapeutic interventions.

Twin and family-based studies estimate that the heritability of FEV_1 is as high as 81%, supporting a strong contribution by genetic factors in FEV_1 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 FEV_1, 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.*

91 2017). Most of these previous GWAS, however, were performed in adult populations of
92 European descent, and their results may not generalize across populations or across the life
93 span of an individual (Carlson *et al.* 2013; Martin, A. R. *et al.* 2017; Wojcik *et al.* 2019). Previous
94 GWAS results are also limited due to their reliance on genotyping arrays. In particular, variation
95 in non-coding regions of the genome is not adequately covered by many genotyping arrays
96 because they were not designed while taking into account the population-specific genetic
97 variability of all populations (Kim, M. S. *et al.* 2018; Zhang and Lupski. 2015). Whole genome
98 sequencing (WGS) is a newer technology that captures nearly all common variation from coding
99 and non-coding regions of the genome and is unencumbered by genotype array design
100 constraints and differences in linkage disequilibrium patterns among populations. To date, no
101 large-scale WGS studies of lung function have been performed in African American children
102 with asthma (Martin *et al.* 2017).

103 In addition to genetics, FEV_1 is a complex trait that is significantly influenced by both genetic
104 variation and environmental factors, such as air pollution (Chatterjee and Das. 1995; Hukkinen
105 *et al.* 2011; Palmer *et al.* 2001; Sillanpaa *et al.* 2017; Tian *et al.* 2017; Yamada *et al.* 2015).
106 Exposure to ambient air pollution has been consistently associated with poor respiratory
107 outcomes, including reduced FEV_1 (Barraza-Villarreal *et al.* 2008; Brunekreef and Holgate. 2002;
108 Ierodiakonou *et al.* 2016; Wise 2019). We previously showed that exposure to sulfur dioxide
109 (SO_2), an air pollutant emitted by the burning of fossil fuels, is significantly associated with
110 reduced FEV_1 in African American children with asthma in the SAGE II study (Neophytou *et al.*
111 2016). Because the genetic variants associated with FEV_1 thus far do not account for the
112 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 FEV_1 in African American children with asthma in SAGE II and investigated the effect of gene-by-air-pollution (SO_2) interactions on FEV_1 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 (FEV_1) 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).

FEV_1 GWAS

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

155 = 1.05 x 10^{-8} for $\log_{10}$ FEV_1), FEV_1 was regressed on all covariates (Age, sex, height, controller
156 medications, sequencing centers, and the first 5 genetic PCs) and the residuals were inverse-
157 normalized. These inverse-normalized residuals (FEV_1.res.rnorm) were the main outcome of the
158 discovery GWAS. The controller medication covariate included the use of inhaled
159 corticosteroids (ICS), long-acting beta-agonists (LABA), leukotriene inhibitors and/or an
160 ICS/LABA combo in the 2 weeks prior to the recruitment date.

161 Genome-wide single variant analysis was performed on the ENCORE server (https://github.com
162 /statgen/encore) using the linear Wald test (q.linear) originally implemented in EPACTS
163 (https://genome.sph.umich.edu /wiki/EPACTS) and TOPMed freeze 8 data (DP0 PASS) with a
164 MAF filter of 0.1%. All pairwise relationships with degree 3 or more relatedness (kinship values
165 > 0.044) were identified, and one participant of the related pair was subsequently chosen at
166 random and removed prior to analysis. All covariates used to obtain FEV_1.res.rnorm were also
167 included as covariates in the GWAS as recommended in a recent publication (Sofer *et al.* 2019).
168 The association analysis was repeated using untransformed FEV_1 and FEV_1 percent predicted
169 (FEV_1.perc.predicted). FEV_1 percent predicted was defined as the percentage of measured FEV_1
170 relative to predicted FEV_1 estimated by the Hankinson lung function prediction equation for
171 African Americans (Hankinson *et al.* 1999). A secondary analysis that included smoking-related
172 covariates (smoking status and number of smokers in the family) was performed in PLINK 1.9
173 (version 1p9_2019_0304_dev) (Chang *et al.* 2015; Purcell and Chang. 2013). Regional
174 association results were plotted using LocusZoom 1.4 (Pruim *et al.* 2010) with a 500 kilobase
175 (Kb) flanking region. Linkage disequilibrium (R^2) was estimated in PLINK 1.9.

176 The function effectiveSize in the R package CODA was used to estimate the actual effective
177 number of independent tests and CODA-adjusted statistical and suggestive significance p-value
178 thresholds were defined as 0.05 and 1 divided by the effective number of tests, respectively
179 (Duggal *et al.* 2008). We compared the CODA-adjusted statistical significance threshold and the
180 widely used 5 x 10^{-8} GWAS genome-wide significance threshold (Pe'er *et al.* 2008) and selected
181 the more stringent threshold for genome-wide significance.

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

188 To determine if the rs73429450 association with FEV_1 was only identifiable using whole genome
189 sequencing data, we repeated the linear regression association analysis on signals that passed
190 the genome-wide significance threshold using PLINK 1.9 and genotype data generated with
191 Axiom Genome-Wide LAT 1 array (Affymetrix, Santa Clara, CA, dbGaP phs000921.v1.p1). These
192 array genotype data were imputed into the following reference panels: 1000 Genomes phase 3
193 version 5, Haplotype Reference Consortium (HRC) r1.1, the Consortium on Asthma among
194 African-ancestry Populations in the Americas (CAAPA) and the TOPMed phase 5 panels on the
195 Michigan Imputation Server (Das *et al.* 2016). It should be noted that 500 SAGE II subjects were
196 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 1×10^{-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

218 coverage genotypes instead of SKAT imputation, and hence, TOPMed freeze 8 DP0 variants with
219 a VCF FILTER of PASS were included in the analysis. The function effectiveSize in the R package
220 CODA (Plummer *et al.* 2006) was used to estimate the effective number of independent
221 hypothesis tests for accurate Bonferroni multiple testing corrections. P-value thresholds for
222 statistical significance and suggestive significance were defined as 0.05 and 1 divided by the
223 effective number of tests, respectively (Duggal *et al.* 2008). If a region was suggestively
224 significant, region-based analyses were repeated with functional variants and/or rare variants
225 (MAF <= 0.01) to assess contribution of common, rare and/or functional variants. Region-based
226 analyses using rare variants only were performed using SKAT-O (Lee *et al.* 2012). The WGSA
227 annotation filters used to define functional variants are provided in File S1 (Supplementary Text
228 1). To study the contribution of individual variants to a region-based association p-value, drop-
229 one variant analysis was performed by repeating the region-based analysis multiple times and
230 dropping one variant only at a time.

231 ***Functional annotations and prioritization of genetic variants***

232 The Hi-C Unifying Genomic Interrogator (HUGIN) (Ay *et al.* 2014; Martin, J. S. *et al.* 2017;
233 Schmitt *et al.* 2016) was used to assign potential gene targets to each variant. ENCODE
234 annotations (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012) were based
235 on overlap of the variants with functional data downloaded from the UCSC Table Browser
236 (Karolchik *et al.* 2004). These data included DNAase I hypersensitivity peak clusters (hg38
237 wgEncodeRegDnaseClustered table), transcription factor ChIP-Seq clusters (hg38
238 encRegTfbsClustered table) and histone modification ChIP-Seq peaks (hg19

239 wgEncodeBroadHistone<cell type><histone>StdPk tables). For DNase I hypersensitivity and
240 transcription factor binding sites, we focused on blood, bone marrow, lung and embryonic cells.
241 For histone modification ChIP-Seq, we focused on H3K27ac and H3K4me3 modifications in
242 human blood (GM12878), bone marrow (K562), lung fibroblast (NHLF), and embryonic stem
243 cells (H1-hESC). LiftOver tool (Hinrichs *et al.* 2006) was used to convert genomic coordinates
244 from hg19 to hg38. Candidate cis-regulatory elements (ccREs) were a subset of representative
245 DNase hypersensitivity sites with epigenetic activity further supported by histone modification
246 (H3K4me3 and H3K27ac) or CTCF-binding data from the ENCODE project. Overlap of variants
247 with ccREs were detected using the Search Candidate cis-Regulatory Elements by ENCODE
248 (SCREEN) web interface (ENCODE Project Consortium 2011; ENCODE Project Consortium 2012).

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

253 *Replication of GWAS associations*

254 All replication analyses were performed in subjects with asthma. Replication of GWAS FEV_1
255 associations was attempted on TOPMed whole genome sequencing data generated from four
256 cohorts. These cohorts included Puerto Rican (n=1,109) and Mexican American (n=649) children
257 in the Genes-Environments and Admixture in Latino Americans (GALA II) study (Oh *et al.* 2012),
258 African American adults in the Study of Asthma Phenotypes and Pharmacogenomic Interactions
259 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 axiom 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 SO_2 (Neophytou *et al.* 2016). To test for an interaction between a genetic variant and SO_2, an additional multiplicative interaction term (variant x $S0_2$ exposure) was included in the original GWAS model (see Method Section "FEV_1 GWAS"). The SO_2 estimates used in the interaction analysis were first-year, past-year, and lifetime exposure to ambient of SO_2, which were estimated as described previously (Neophytou *et al.* 2016). Residuals of FEV_1 were plotted against exposure to SO_2 and stratified by the number of copies of the minor allele. Residuals of FEV_1 were obtained as described in the Methods Section "FEV_1 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 FEV_1 values is in Figure S1. The CODA-adjusted statistical significance threshold (2.10×10^{-8}) was selected as the genome-wide

299 significance level. According to this threshold, one SNP in chromosome 12 (chr12:88846435,
300 rs73429450, G>A) was associated with FEV_1.res.rnorm (Figure 1, $p = 9.01 \times 10^{-9}$, $\beta = 0.801$). The
301 association between rs73429450 and lung function remained statistically significant when the
302 association was repeated using untransformed FEV_1 ($p = 1.26 \times 10^{-8}$, $\beta = 0.302$) as the outcome
303 variable. The association between rs73429450 and lung function was suggestive using
304 FEV_1.perc.predicted ($p = 1.69 \times 10^{-7}$, $\beta = 0.100$). Thirty-nine variants with association $p < 0.001$
305 overlapped with 14 out of 98 previously reported FEV_1-associated loci and their 20 Kb flanking
306 regions (Table S1). When including all variants with association $p < 0.05$, we found overlap with
307 all 98 previously reported loci. Our top FEV_1 association, rs73429450, did not overlap with any
308 previously reported loci in both scenarios and it is a novel association with FEV_1 in this study
309 population.

310 Secondary analysis that included covariates correcting for smoking status and number of
311 smokers in the family showed that smoking-related factors were not significantly associated
312 with FEV_1 in our pediatric SAGE cohort: using 657 out of 867 individuals with available smoking-
313 related covariates, the FEV_1.res.rnorm association p-values before and after including the
314 smoking-related covariates were 2.01×10^{-6} and 1.89×10^{-6}. Both p-values of the covariates
315 smoking status ($p = 0.27$) and number of smokers in the family ($p = 0.54$) were not significant.

316 Conditional analysis was performed on 45 variants with association $p < 1 \times 10^{-4}$ located within 1
317 Mb of the strongest association signal (rs73429450). Two weaker independent signals
318 (rs11312747, rs58475486) were identified (Table S2).

319 The minor allele frequency of rs73429450 in continental populations from the 1000 Genomes
320 Project (1000G) is 3% in Africans (AFR) and < 1% in Admixed Americans (AMR), Europeans (EUR)
321 and Asians (EAS and SAS) (1000 Genomes Project Consortium *et al.* 2015). Rs73429450 was not
322 included on the Affymetrix LAT1 genotyping array where SAGE participants were previously
323 genotyped. To determine if the rs73429450 association with FEV_1 was only identifiable using
324 whole genome sequencing data, we attempted to reproduce our results by imputing the
325 genotype of rs73429450 in 851 SAGE participants with available array data using 1000G phase 3
326 (n = 2,504), HRC r1.1 (n = 32,470), CAAPA (n = 883) and TOPMed freeze 5 (n = 62,784) reference
327 panels. Our results remained statistically significant when using the 1000G phase 3 ($p = 4.97 \times$
328 10^{-8}, $\beta = 0.79$, imputation $R^2 = 0.95$) and TOPMed freeze 5 ($p = 1.22 \times 10^{-8}$, $\beta = 0.80$, imputation
329 $R^2 = 0.98$) reference panels, but lost statistical significance when rs73429450 genotypes were
330 imputed using the HRC ($p = 4.35 \times 10^{-7}$, $\beta = 0.68$, imputation $R^2 = 0.94$) and CAAPA ($p = 1.95 \times$
331 10^{-7}, $\beta = 0.80$, imputation $R^2 = 0.71$) reference panels.

332 Region-based association analysis including all variants conditioned on the association signal
333 from rs73429450 was performed in its 1 Mb flanking region (chr12:87846435-89846435). No
334 windows were significantly associated after Bonferroni multiple testing correction ($p < 2.81 \times$
335 10^{-4}, Figure S2), but 20 windows were suggestively associated with FEV_1.res.rnorm ($p < 5.62 \times$
336 10^{-3}, Table S3). Two of 20 windows re-tested using only functional variants were suggestively
337 significant (region 4 and 16). Both of these windows were no longer suggestively significant
338 after removing the common variants, indicating that association signal from these regions was
339 mostly driven by common variants. Further investigation on region 16 using drop-one analysis
340 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 FEV_1-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

361 ancestry, and therefore could share eQTLs. All five eQTLs corresponded to one signal in a
362 region with strong linkage disequilibrium ($r^2 > 0.8$, markers 25, 26, 28, 29 and 32 in Figure 2).

363 *Replication of genetic association with FEV_1*

364 Subject Characteristics of our four replication cohorts (SAPPHIRE, GCPD-A, UK Biobank and
365 GALA II) are shown in Table S7. We attempted to replicate the association of the 45 SNPs in our
366 primary FEV_1 GWAS in each cohort. We used 0.05 as the suggestive p-value threshold and
367 0.0167 as the Bonferroni-corrected p-value threshold after correcting for 3 independent signals
368 (see conditional analysis in Results Section). A total of 20 variants were replicated at $p < 0.05$
369 with consistent direction of effect in black UK Biobank participants; 14 variants in SAPPHIRE and
370 2 variants in GCPD-A were significant but had an opposite direction of effect (Table S8).

371 We attempted to replicate the FEV_1.res.rnorm association in Mexican American (n = 649) and
372 Puerto Rican (n = 1,109) children with asthma from the GALA II study. In Mexican Americans,
373 we excluded 19 variants with MAF < 0.1% and associations for the remaining 26 variants did not
374 replicate (Table S9). In Puerto Ricans, we observed the same protective effect in 38 of the 45
375 variants in the locus, but the associations were not statistically significant (Table S9).

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

378 We combined and summarized all functional evidence for the top 45 variants, along with eQTL
379 findings from nasal epithelial RNA-Seq and replication results (Figure 2, Table 2). To facilitate
380 interpretation of the variant association with FEV_1, the effect sizes and p-values of both FEV_1 (β

and p) and FEV_1.res.rnorm (β_{norm} and p_{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 SO_2 were associated with 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 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

423 found to be suggestively significant using data imputed from the CAAPA reference panel ($p =$
424 1.95×10^{-7}, $\beta = 0.80$). Of the imputation reference panels that we assessed, CAAPA is one of the
425 more relevant reference panels for our study population because it is based on African
426 populations in the Americas. However, we note that the effect size estimated from CAAPA-
427 imputed data was comparable to that generated from WGS data. While whole genome
428 sequencing data is usually praised for enabling analysis of rare-variant contributions to
429 phenotype variability, our results show the utility of whole genome sequencing data for the
430 reliable analysis of common variants as well in the absence of relevant imputation panels.

431 Although rs73429450 had the lowest p-value from our whole genome sequencing association
432 analysis, we did not find the required amount of functional evidence to prioritize this marker
433 for inclusion in downstream gene-by-air-pollution analyses. Another variant, rs73440122, was
434 in moderate to strong linkage disequilibrium ($r^2 = 0.76$) with rs7349450 and had a similar MAF
435 (0.027) in our study population, but was only suggestively associated with FEV_1 in our
436 association analysis ($p = 2.08 \times 10^{-7}$, Table2). In contrast to rs7349450, there were multiple lines
437 of evidence suggesting the functional relevance of rs73440122: rs73440122 received the
438 highest priority score, its association with FEV_1 replicated in black UK Biobank participants, and
439 it was one of the most likely drivers of FEV_1 variability among individuals, possibly mediated
440 through *KITLG*.

441 Bioinformatic interrogation of rs73440122 revealed that the variant overlapped with a ccRE
442 (SCREEN accession EH37E0279310), DNase I hypersensitivity site, and SPI1 ChIP-Seq clusters
443 that were indicative of a candidate open chromatin gene regulatory region. The transcription

444 factor SPI1, which is expressed in lung, spleen and whole blood, was previously linked to fibrotic
445 diseases and tissues (Wohlfahrt *et al.* 2019) and *KITLG* in erythroid progenitor cells (Quang *et al.*
446 1995), but there is no existing evidence of its role in lung function. Although rs73440122 itself is
447 not an eQTL of *KITLG*, it is located in a region that physically interacted with *KITLG* based on Hi-
448 C data in fetal lung fibroblast cells. Additionally, five neighboring FEV_1 associated variants were
449 identified as eQTLs of *KITLG*, although they also appeared to be an independent signal ($r^2 < 0.2$).
450 Overall, these results support regulatory interactions between this region and *KITLG*.

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

460 Gene-by-environment interactions likely account for a portion of the "missing" heritability of
461 many complex phenotypes (Moore and Williams. 2009). We hypothesized that a significant
462 portion of the heritability of lung function was due in part to gene-by-air-pollution (SO_2)
463 interaction effects. The interaction between rs58475486 and past year exposure to SO_2 that
464 was significantly associated with lung function supports our hypothesis. The T allele of

465 rs58475486 is common (8-14%) in African populations and showed a protective effect on lung
466 function in the presence of past year SO_2 exposure. SNP rs58475486 is located in a ccRE
467 (SCREEN accession EH37E0279296) and a *FOXA1* binding site in the A549 lung adenocarcinoma
468 cell line. *Foxa1* has a known compensatory role with *Foxa2* during lung morphogenesis in mice
469 (Wan *et al.* 2005). Deletion of both *Foxa1* and *Foxa2* inhibited cell proliferation, epithelial cell
470 differentiation, and branching morphogenesis in fetal lung tissue. Further functional validation
471 on the effect of rs58475486 on binding affinity of FOXA1 is necessary to confirm whether the
472 role of *FOXA1* in this ccRE is important for *KITLG* regulatory and lung function.

473 The higher frequency of the protective alleles of both rs73440122 and rs58465486 in African
474 populations appears to contradict previous findings that African ancestry was associated with
475 lower lung function (Kumar *et al.* 2010). One possible explanation for this seeming
476 inconsistency is that FEV_1 is a complex trait whose variation is influenced by many genetic
477 variants of small to moderate effect sizes whose influence on lung function may vary by
478 exposure to environmental factors. We found suggestive evidence that the interaction between
479 rs73440122 and first year exposure to SO_2 reverses the positive association of rs73440122 with
480 lung function to a negative one (Table 3). When assessed independently, our genetic
481 association analysis showed that the protective A allele of rs73440122 was associated with
482 higher lung function. However, with increasing levels of SO_2 exposure in the first year of life,
483 increasing copies of the A allele of rs73440122 were associated with decreased lung function.
484 Air pollution is known to negatively impact lung function, and we have previously shown that
485 the deleterious effects of air pollution on lung phenotypes may be significantly increased in
486 African American children compared to other populations experiencing the same amount of

487 exposure (Nishimura *et al.* 2013). It has also been reported that Latino and African American
488 populations often live in neighborhoods with high levels of air pollution (Mott 1995). The
489 increased susceptibility to negative pulmonary effects from air pollution exposure coupled with
490 the disproportionate exposure to air pollution experienced by the African American population
491 may also contribute to the lower lung function seen in this population despite the presence of
492 protective alleles.

493 One limitation of this study is that the FEV_1 genetic association and the eQTL analyses with
494 *KITLG* were performed in different populations due to data availability constraints. Although we
495 did not have RNA-Seq data from lung tissues from our study subjects, we previously
496 demonstrated that there is a high degree of overlap in gene expression profiles between nasal
497 and bronchial epithelial cells (Poole *et al.* 2014). The direction of effect of the association was
498 the same in GALA II Puerto Rican children with asthma but not statistically significant. This may
499 in part due to the significantly lower African Ancestry in Puerto Ricans compared to African
500 Americans.

501 We replicated 20 of 45 variants in black UK Biobank subjects and observed conflicting "flip-flop"
502 associations in African Americans from the SAPPHIRE and GCPD-A studies. In the past, flip-flop
503 associations were deemed as spurious results. Traditional association testing approach studies
504 the effect of each variant on phenotype independently and increases the chance of flip-flop
505 associations detected between studies. Differences in study design, sampling variation that
506 leads to variation in LD patterns, and lack of consideration of other disease influencing genetic
507 and/or environmental factors are all potential causes of flip-flop associations (Kraft *et al.* 2009;

508 Lin *et al.* 2007). Hence, it is not surprising to observe flip-flop associations when gene and
509 environment interactions were detected at our FEV_1 GWAS locus. It was previously shown that
510 flip-flop associations can occur between and within populations even in the presence of a
511 genuine genetic effect (Kraft *et al.* 2009; Lin *et al.* 2007). Further functional analysis is thus
512 required to validate the relationship between the candidate variants, *KITLG* and FEV_1. This may
513 include reporter assays to validate potential enhancer or repressor activity and CRISPR-based
514 editing assays to validate the regulatory role of the candidate variants on *KITLG*. Although
515 literature exists describing KIT signaling for lung function in mice (Lindsey *et al.* 2011),
516 additional knockout experiments in a model animal system may be necessary to confirm the
517 causal effect of *KITLG* on lung function.

518 The average concentration of ambient SO_2 exposure in our participants (Table 1) was lower
519 than the National Ambient Air Quality Standards. It is possible that SO_2 acted as a surrogate for
520 other unmeasured toxic pollutants emitted from local point sources. Major sources of SO_2 in
521 San Francisco Bay Area during the recruitment years of 2006 to 2011 include airports,
522 petroleum refineries, gas and oil plants, calcined petroleum coke plants, electric power plants,
523 cement manufacturing factories, chemical plants, and landfills (United States Environmental
524 Protection Agency 2008; United States Environmental Protection Agency 2011). The
525 Environmental Protection Agency's national emissions inventory data also showed that these
526 facilities emit Volatile Organic Compounds, heavy metals (lead, mercury, chromium, arsenic),
527 formaldehyde, ethyl benzene, acrolein, 1,3-butadiene, 1,4-dichlorobenzene, and
528 tetrachloroethylene into the air along with SO_2. These chemicals are highly toxic and inhaling

529 even a small amount may contribute to poor lung function. Another possibility is that exposure
530 to SO_2 captured unmeasured confounding socioeconomic factors.

531 This study identified a novel protective allele for lung function in African American children with
532 asthma. The protective association with lung function intensified with increased past year
533 exposure to SO_2. Our findings showcases the complexity of the relationship between genetic
534 and environmental factors impacting variation in FEV_1, highlights the utility of WGS data for
535 genetic research of complex phenotypes, and underscores the importance of including diverse
536 study populations in our exploration of the genetic architecture underlying lung function.

537 ACKNOWLEDGEMENTS

538 The Genes-Environments and Admixture in Latino Americans (GALA II) Study, the Study of
539 African Americans, Asthma, Genes and Environments (SAGE) Study and E.G.B. were supported
540 by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical
541 Faculty Development Program, the Harry Wm. and Diana V. Hind Distinguished Professor in
542 Pharmaceutical Sciences II, the National Heart, Lung, and Blood Institute (NHLBI)
543 [R01HL117004, R01HL128439, R01HL135156, X01HL134589]; the National Institute of
544 Environmental Health Sciences [R01ES015794]; the National Institute on Minority Health and
545 Health Disparities (NIMHD) [P60MD006902, R01MD010443], the National Human Genome
546 Research Institute [U01HG009080] and the Tobacco-Related Disease Research Program [24RT-
547 0025]. MJW was supported by the NHLBI [K01HL140218]. JJ and BEH were supported by the
548 NHLBI [R01HL133433, R01HL141992]. KLK was supported by the NHLBI [R01HL135156-S1], the
549 UCSF Bakar Institute, the Gordon and Betty Moore Foundation [GBMF3834], and the Alfred P.

550 Sloan Foundation [2013-10-27] grant to UC Berkeley through the Moore-Sloan Data Science
551 Environment Initiative. ACW was supported by the Eunice Kennedy Shriver National Institute of
552 Child Health and Human Development [1R01HD085993-01].

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

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

559 Part of this research was conducted using the UK Biobank Resource under Application Number
560 40375. We would like to thank UK Biobank participants and researchers who contributed or
561 collected data.

562 Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed)
563 program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for
564 "NHLBI TOPMed: Gene-Environment, Admixture and Latino Asthmatics Study" (phs000920) and
565 "NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environments" (phs000921)
566 was performed at the New York Genome Center (3R01HL117004-02S3) and the University of
567 Washington Northwest Genomics Center (HHSN268201600032I). WGS for "NHLBI TOPMed:
568 Study of Asthma Phenotypes & Pharmacogenomic Interactions by Race-Ethnicity" (phs001467)
569 and "Genetics of Complex Pediatric Disorders - Asthma" (phs001661) was performed at the

570 University of Washington Northwest Genomics Center (HHSN268201600032I). Centralized read
571 mapping and genotype calling, along with variant quality metrics and filtering were provided by
572 the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract
573 HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and
574 general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-
575 120393-02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and
576 participants who provided biological samples and data for TOPMed.

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

583 The TOPMed imputation panel was supported by the NHLBI and TOPMed study investigators
584 who contributed data to the reference panel. The panel was constructed and implemented by
585 the TOPMed Informatics Research Center at the University of Michigan (3R01HL-117626-02S1;
586 contract HHSN268201800002I). The TOPMed Data Coordinating Center (3R01HL-120393-02S1;
587 contract HHSN268201800001I) provided additional data management, sample identity checks,
588 and overall program coordination and support. We gratefully acknowledge the studies and
589 participants who provided biological samples and data for TOPMed.

590 The authors wish to acknowledge the following GALA II and SAGE study collaborators: Shannon
591 Thyne, UCSF; Harold J. Farber, Texas Children's Hospital; Denise Serebrisky, Jacobi Medical
592 Center; Rajesh Kumar, Lurie Children's Hospital of Chicago; Emerita Brigino-Buenaventura,
593 Kaiser Permanente; Michael A. LeNoir, Bay Area Pediatrics; Kelley Meade, UCSF Benioff
594 Children's Hospital, Oakland; William Rodríguez-Cintrón, VA Hospital, Puerto Rico; Pedro C.
595 Ávila, Northwestern University; Jose R. Rodríguez-Santana, Centro de Neumología Pediátrica;
596 Luisa N. Borrell, City University of New York; Adam Davis, UCSF Benioff Children's Hospital, Oakland;
597 Saunak Sen, University of Tennessee.

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

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

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

609

610 LITERATURE CITED

611 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al*,
612 2015 A global reference for human genetic variation. Nature **526:** 68-74.

613 Akinbami, L. J., 2015 *Asthma Prevalence, Health Care use and Mortality: United States, 2003-05*.
614 [Online] Available at: http://www.cdc.gov/nchs/data/hestat/asthma03-05/asthma03-05.htm.
615 [Accessed 2020 Jan 8].

616 Akinbami, L. J., J. E. Moorman, A. E. Simon and K. C. Schoendorf, 2014 Trends in racial
617 disparities for asthma outcomes among children 0 to 17 years, 2001-2010. J. Allergy Clin.
618 Immunol. **134:** 547-553.e5.

619 Alexander, D. H., J. Novembre and K. Lange, 2009 Fast model-based estimation of ancestry in
620 unrelated individuals. Genome Res. **19:** 1655-1664.

621 Anders, S., P. T. Pyl and W. Huber, 2015 HTSeq--a python framework to work with high-
622 throughput sequencing data. Bioinformatics **31:** 166-169.

623 Ay, F., T. L. Bailey and W. S. Noble, 2014 Statistical confidence estimation for hi-C data reveals
624 regulatory chromatin contacts. Genome Res. **24:** 999-1011.

625 Barraza-Villarreal, A., J. Sunyer, L. Hernandez-Cadena, M. C. Escamilla-Nunez, J. J. Sienra-Monge
626 *et al*, 2008 Air pollution, airway inflammation, and lung function in a cohort study of mexico city
627 schoolchildren. Environ. Health Perspect. **116:** 832-838.

628 Bhattacharya, S., S. Srisuma, D. L. Demeo, S. D. Shapiro, R. Bueno *et al*, 2009 Molecular
629 biomarkers for quantitative and discrete COPD phenotypes. Am. J. Respir. Cell Mol. Biol. **40:**
630 359-367.

631 Brunekreef, B., and S. T. Holgate, 2002 Air pollution and health. Lancet **360:** 1233-1242.

632 Buniello, A., J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst *et al*, 2019 The NHGRI-EBI
633 GWAS catalog of published genome-wide association studies, targeted arrays and summary
634 statistics 2019. Nucleic Acids Res. **47:** D1005-D1012.

635 Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott *et al*, 2018 The UK biobank resource
636 with deep phenotyping and genomic data. Nature **562:** 203-209.

637 Canela-Xandri, O., K. Rawlik and A. Tenesa, 2018 An atlas of genetic associations in UK biobank.
638 Nat. Genet. **50:** 1593-1599.

639 Carlson, C. S., T. C. Matise, K. E. North, C. A. Haiman, M. D. Fesinmeyer *et al*, 2013
640 Generalization and dilution of association results from european GWAS in populations of non-
641 european ancestry: The PAGE study. PLoS Biol. **11:** e1001661.

642 Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al*, 2015 Second-generation
643 PLINK: Rising to the challenge of larger and richer datasets. Gigascience **4:** 7-8. eCollection
644 2015.

645 Chatterjee, S., and N. Das, 1995 Lung function in indian twin children: Comparison of genetic
646 versus environmental influence. Ann. Hum. Biol. **22:** 289-303.

647 Ciesielski, T. H., S. A. Pendergrass, M. J. White, N. Kodaman, R. S. Sobota *et al*, 2014 Diverse
648 convergent evidence in the genetic analysis of complex disease: Coordinating omic, informatic,
649 and experimental evidence to better identify and validate risk factors. BioData Min. **7:** 10-10.
650 eCollection 2014.

651 Conomos, M. P., M. B. Miller and T. A. Thornton, 2015 Robust inference of population structure
652 for ancestry prediction and correction of stratification in the presence of relatedness. Genet.
653 Epidemiol. **39:** 276-293.

654 Conomos, M. P., A. P. Reiner, B. S. Weir and T. A. Thornton, 2016 Model-free estimation of
655 recent genetic relatedness. Am. J. Hum. Genet. **98:** 127-148.

656 Das, S., L. Forer, S. Schonherr, C. Sidore, A. E. Locke *et al*, 2016 Next-generation genotype
657 imputation service and methods. Nat. Genet. **48:** 1284-1287.

658 Duggal, P., E. M. Gillanders, T. N. Holmes and J. E. Bailey-Wilson, 2008 Establishing an adjusted
659 p-value threshold to control the family-wide type 1 error in genome wide association studies.
660 BMC Genomics **9:** 516-516.

661 ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human
662 genome. Nature **489:** 57-74.

663 ENCODE Project Consortium, 2011 A user's guide to the encyclopedia of DNA elements
664 (ENCODE). PLoS Biol. **9:** e1001046.

665 GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)-Analysis Working
666 Group, Statistical Methods groups-Analysis Working Group, Enhancing GTEx (eGTEx) groups,
667 NIH Common Fund *et al*, 2017 Genetic effects on gene expression across human tissues. Nature
668 **550:** 204-213.

669 Hankinson, J. L., J. R. Odencrantz and K. B. Fedan, 1999 Spirometric reference values from a
670 sample of the general U.S. population. Am. J. Respir. Crit. Care Med. **159:** 179-187.

671 Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano *et al*, 2006 The UCSC genome
672 browser database: Update 2006. Nucleic Acids Res. **34:** 590.

673 Hukkinen, M., J. Kaprio, U. Broms, A. Viljanen, D. Kotz *et al*, 2011 Heritability of lung function: A
674 twin study among never-smoking elderly women. Twin Res. Hum. Genet. **14:** 401-407.

675 Huusko, J. M., M. Mahlman, M. K. Karjalainen, T. Kaukola, R. Haataja *et al*, 2015 Polymorphisms
676 of the gene encoding kit ligand are associated with bronchopulmonary dysplasia. Pediatr.
677 Pulmonol. **50:** 260-270.

678 Ierodiakonou, D., A. Zanobetti, B. A. Coull, S. Melly, D. S. Postma *et al*, 2016 Ambient air
679 pollution, lung function, and airway responsiveness in asthmatic children. J. Allergy Clin.
680 Immunol. **137:** 390-399.

681 Ionita-Laza, I., S. Lee, V. Makarov, J. D. Buxbaum and X. Lin, 2013 Sequence kernel association
682 tests for the combined effect of rare and common variants. Am. J. Hum. Genet. **92:** 841-853.

683 Jiang, H., R. Lei, S. W. Ding and S. Zhu, 2014 Skewer: A fast and accurate adapter trimmer for
684 next-generation sequencing paired-end reads. BMC Bioinformatics **15:** 182-182.

685 Johnson, J. D., and W. M. Theurer, 2014 A stepwise approach to the interpretation of
686 pulmonary function tests. Am. Fam. Physician **89:** 359-366.

687 Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet *et al*, 2004 The UCSC table
688 browser data retrieval tool. Nucleic Acids Res. **32:** 493.

689 Kim, D., B. Langmead and S. L. Salzberg, 2015 HISAT: A fast spliced aligner with low memory
690 requirements. Nat. Methods **12:** 357-360.

691 Kim, M. S., K. P. Patel, A. K. Teng, A. J. Berens and J. Lachance, 2018 Genetic disease risks can be
692 misestimated across global populations. Genome Biol. **19:** 179-7.

693 Kraft, P., E. Zeggini and J. P. Ioannidis, 2009 Replication in genome-wide association studies.
694 Stat. Sci. **24:** 561-573.

695 Kumar, R., M. A. Seibold, M. C. Aldrich, L. K. Williams, A. P. Reiner *et al*, 2010 Genetic ancestry
696 in lung-function predictions. N. Engl. J. Med. **363:** 321-330.

697 Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder *et al*, 2012 Optimal unified
698 approach for rare-variant association testing with application to small-sample case-control
699 whole-exome sequencing studies. Am. J. Hum. Genet. **91:** 224-237.

700 Levin, A. M., Y. Wang, K. E. Wells, B. Padhukasahasram, J. J. Yang *et al*, 2014 Nocturnal asthma
701 and the importance of race/ethnicity and genetic ancestry. Am. J. Respir. Crit. Care Med. **190:**
702 266-273.

703 Li, X., G. A. Hawkins, E. J. Ampleford, W. C. Moore, H. Li *et al*, 2013 Genome-wide association
704 study identifies TH1 pathway genes associated with lung function in asthmatic patients. J.
705 Allergy Clin. Immunol. **132:** 313-20.e15.

706 Liao, S. Y., X. Lin and D. C. Christiani, 2014 Genome-wide association and network analysis of
707 lung function in the framingham heart study. Genet. Epidemiol. **38:** 572-578.

708 Lin, P. I., J. M. Vance, M. A. Pericak-Vance and E. R. Martin, 2007 No gene is an island: The flip-
709 flop phenomenon. Am. J. Hum. Genet. **80:** 531-538.

710 Lindsey, J. Y., K. Ganguly, D. M. Brass, Z. Li, E. N. Potts *et al*, 2011 C-kit is essential for alveolar
711 maintenance and protection from emphysema-like disease in mice. Am. J. Respir. Crit. Care
712 Med. **183:** 1644-1652.

713 Liu, X., S. White, B. Peng, A. D. Johnson, J. A. Brody *et al*, 2016 WGSA: An annotation pipeline
714 for human genome sequencing studies. J. Med. Genet. **53:** 111-112.

715 Mak, A. C. Y., M. J. White, W. L. Eckalbar, Z. A. Szpiech, S. S. Oh *et al*, 2018 Whole-genome
716 sequencing of pharmacogenetic drug response in racially diverse children with asthma. Am. J.
717 Respir. Crit. Care Med. **197:** 1552-1564.

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

721 Martin, A. R., C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale *et al*, 2017 Human
722 demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum.
723 Genet. **100:** 635-649.

724 Martin, J. S., Z. Xu, A. P. Reiner, K. L. Mohlke, P. Sullivan *et al*, 2017 HUGIn: Hi-C unifying
725 genomic interrogator. Bioinformatics **33:** 3793-3795.

726 Moore, J. H., 2005 A global view of epistasis. Nat. Genet. **37:** 13-14.

727 Moore, J. H., and S. M. Williams, 2009 Epistasis and its implications for personal genetics. Am. J.
728 Hum. Genet. **85:** 309-320.

729 Mott, L., 1995 The disproportionate impact of environmental health threats on children of
730 color. Environ. Health Perspect. **103 Suppl 6:** 33-35.

731 Neophytou, A. M., M. J. White, S. S. Oh, N. Thakur, J. M. Galanter *et al*, 2016 Air pollution and
732 lung function in minority youth with asthma in the GALA II (genes-environments and admixture
733 in latino americans) and SAGE II (study of african americans, asthma, genes, and environments)
734 studies. Am. J. Respir. Crit. Care Med. **193:** 1271-1280.

735 Nishimura, K. K., J. M. Galanter, L. A. Roth, S. S. Oh, N. Thakur *et al*, 2013 Early-life air pollution
736 and asthma risk in minority children. the GALA II and SAGE II studies. Am. J. Respir. Crit. Care
737 Med. **188:** 309-318.

738 Oh, S. S., M. J. White, C. R. Gignoux and E. G. Burchard, 2016 Making precision medicine socially
739 precise. take a deep breath. Am. J. Respir. Crit. Care Med. **193:** 348-350.

740 Oh, S. S., H. Tcheurekdjian, L. A. Roth, E. A. Nguyen, S. Sen *et al*, 2012 Effect of secondhand
741 smoke on asthma control among black and latino children. J. Allergy Clin. Immunol. **129:** 1478-
742 83.e7.

743 Ong, B. A., J. Li, J. M. McDonough, Z. Wei, C. Kim *et al*, 2013 Gene network analysis in a
744 pediatric cohort identifies novel lung function genes. PLoS One **8:** e72899.

745 Palmer, L. J., M. W. Knuiman, M. L. Divitini, P. R. Burton, A. L. James *et al*, 2001 Familial
746 aggregation and heritability of adult lung function: Results from the busselton health study. Eur.
747 Respir. J. **17:** 696-702.

748 Pe'er, I., R. Yelensky, D. Altshuler and M. J. Daly, 2008 Estimation of the multiple testing burden
749 for genomewide association studies of nearly all common variants. Genet. Epidemiol. **32:** 381-
750 385.

751 Pino-Yanes, M., N. Thakur, C. R. Gignoux, J. M. Galanter, L. A. Roth *et al*, 2015 Genetic ancestry
752 influences asthma susceptibility and lung function among latinos. J. Allergy Clin. Immunol. **135:**
753 228-235.

754 Plummer, M., N. Best, K. Cowles and K. Vines, 2006 CODA: Convergence diagnosis and output
755 analysis for MCMC. R News **6:** 7-11.

756 Poole, A., C. Urbanek, C. Eng, J. Schageman, S. Jacobson *et al*, 2014 Dissecting childhood asthma
757 with nasal transcriptomics distinguishes subphenotypes of disease. J. Allergy Clin. Immunol.
758 **133:** 670-8.e12.

759 Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines *et al*, 2010 LocusZoom: Regional
760 visualization of genome-wide association scan results. Bioinformatics **26:** 2336-2337.

761 Purcell, S., and C. Chang, 2013 *Plink 1.9*. [Online] Available at: www.cog-
762 genomics.org/plink/1.9/. [Accessed 2019 Mar].

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

766 Quinlan, A. R., and I. M. Hall, 2010 BEDTools: A flexible suite of utilities for comparing genomic
767 features. Bioinformatics **26:** 841-842.

768 Repapi, E., I. Sayers, L. V. Wain, P. R. Burton, T. Johnson *et al*, 2010 Genome-wide association
769 study identifies five loci associated with lung function. Nat. Genet. **42:** 36-44.

770 Schmitt, A. D., M. Hu, I. Jung, Z. Xu, Y. Qiu *et al*, 2016 A compendium of chromatin contact maps
771 reveals spatially active regions in the human genome. Cell. Rep. **17:** 2042-2059.

772 Sillanpaa, E., S. Sipila, T. Tormakangas, J. Kaprio and T. Rantanen, 2017 Genetic and
773 environmental effects on telomere length and lung function: A twin study. J. Gerontol. A Biol.
774 Sci. Med. Sci. **72:** 1561-1568.

775 Sofer, T., X. Zheng, S. M. Gogarten, C. A. Laurie, K. Grinde *et al*, 2019 A fully adjusted two-stage
776 procedure for rank-normalization in genetic association studies. Genet. Epidemiol. **43:** 263-275.

777 Soler Artigas, M., L. V. Wain, S. Miller, A. K. Kheirallah, J. E. Huffman *et al*, 2015 Sixteen new
778 lung function signals identified through 1000 genomes project reference panel imputation. Nat.
779 Commun. **6:** 8658.

780 Soler Artigas, M., D. W. Loth, L. V. Wain, S. A. Gharib, M. Obeidat *et al*, 2011 Genome-wide
781 association and large-scale follow up identifies 16 new loci influencing lung function. Nat.
782 Genet. **43:** 1082-1090.

783 Summer Institute in Statistical Genetics, 2019 *PC-Relate*. [Online] Available at: https://uw-
784 gac.github.io/SISG_2019/pc-relate.html. [Accessed 2019 Jul 25].

785 Taliun, D., D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech *et al*, 2019 Sequencing of 53,831
786 diverse genomes from the NHLBI TOPMed program. bioRxiv 563866.

787 Tian, X., C. Xu, Y. Wu, J. Sun, H. Duan *et al*, 2017 Genetic and environmental influences on
788 pulmonary function and muscle strength: The chinese twin study of aging. Twin Res. Hum.
789 Genet. **20:** 53-59.

790 TOPMed, 2019 *TOPMed Whole Gneome Sequencing Methods: Freeze 8*. [Online] Available at:
791 https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8. [Accessed
792 2019 Dec 13].

793 United States Environmental Protection Agency, 2011 *National Emissions Inventory (NEI) 2011*
794 *Data*. [Online] Available at: https://www.epa.gov/air-emissions-inventories/2011-national-
795 emissions-inventory-nei-data. [Accessed 2020 Jan 8].

796 United States Environmental Protection Agency, 2008 *National Emissions Inventory (NEI) 2008*
797 *Data*. [Online] Available at: https://www.epa.gov/air-emissions-inventories/2008-national-
798 emissions-inventory-nei-data. [Accessed 2020 Jan 8].

799 University of Michigan, and NHLBI TOPMed, 2018 *BRAVO Variant Browser*. [Online] Available
800 at: https://bravo.sph.umich.edu/freeze5/hg38/. [Accessed 2019 Aug].

801 Wain, L. V., N. Shrine, M. S. Artigas, A. M. Erzurumluoglu, B. Noyvert *et al*, 2017 Genome-wide
802 association analyses for lung function and chronic obstructive pulmonary disease identify new
803 loci and potential druggable targets. Nat. Genet. **49:** 416-425.

804 Wan, H., S. Dingle, Y. Xu, V. Besnard, K. H. Kaestner *et al*, 2005 Compensatory roles of Foxa1
805 and Foxa2 during lung morphogenesis. J. Biol. Chem. **280:** 13809-13816.

806 White, M. J., O. Risse-Adams, P. Goddard, M. G. Contreras, J. Adams *et al*, 2016 Novel genetic
807 risk factors for asthma in african american children: Precision medicine and the SAGE II study.
808 Immunogenetics **68:** 391-400.

809 Wise, J., 2019 Air pollution is linked to infant deaths and reduced lung function in children. BMJ
810 **366:** l5772.

811 Wohlfahrt, T., S. Rauber, S. Uebe, M. Luber, A. Soare *et al*, 2019 PU.1 controls fibroblast
812 polarization and tissue fibrosis. Nature **566:** 344-349.

813 Wojcik, G. L., M. Graff, K. K. Nishimura, R. Tao, J. Haessler *et al*, 2019 Genetic analyses of
814 diverse populations improves discovery for complex traits. Nature **570:** 514-518.

815 World Health Organization, 2017 *Asthma*. [Online] Available at:
816 http://www.who.int/mediacentre/factsheets/fs307/en/. [Accessed 2020 Jan 8].

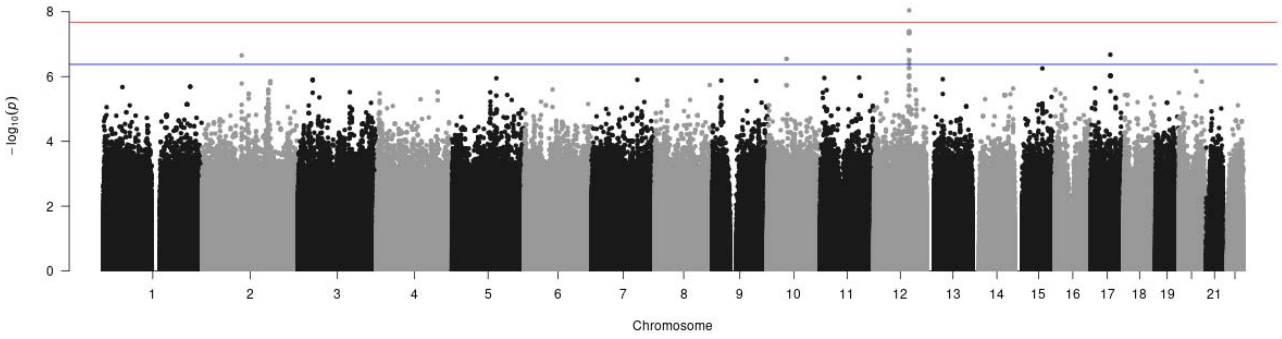
817 Yamada, H., Y. Yatagai, H. Masuko, T. Sakamoto, H. Iijima *et al*, 2015 Heritability of pulmonary
818 function estimated from genome-wide SNPs in healthy japanese adults. Respir. Investig. **53:** 60-
819 67.

820 Zhang, F., and J. R. Lupski, 2015 Non-coding genetic variants in human disease. Hum. Mol.
821 Genet. **24:** 102.

822

Figure 1.

A.


8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
21
Chromosome

B.

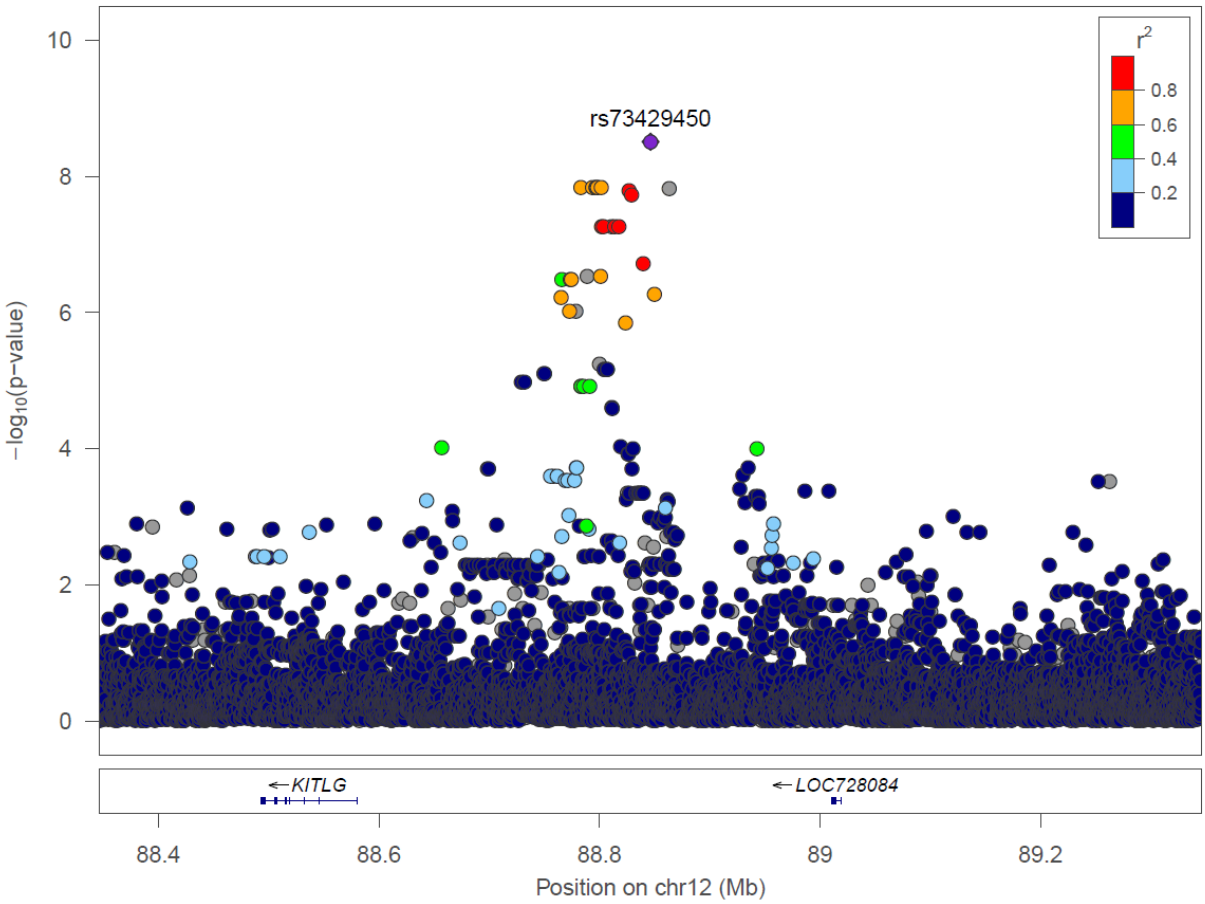

rs73429450
r^2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
$-\log_{10}$(p-value)
KITLG
LOC728084
88.4
88.6
88.8
89
89.2
Position on chr12 (Mb)

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 2.10×10^{-8}. Blue horizontal line: CODA-adjusted suggestive significance p-value of 4.19×10^{-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.

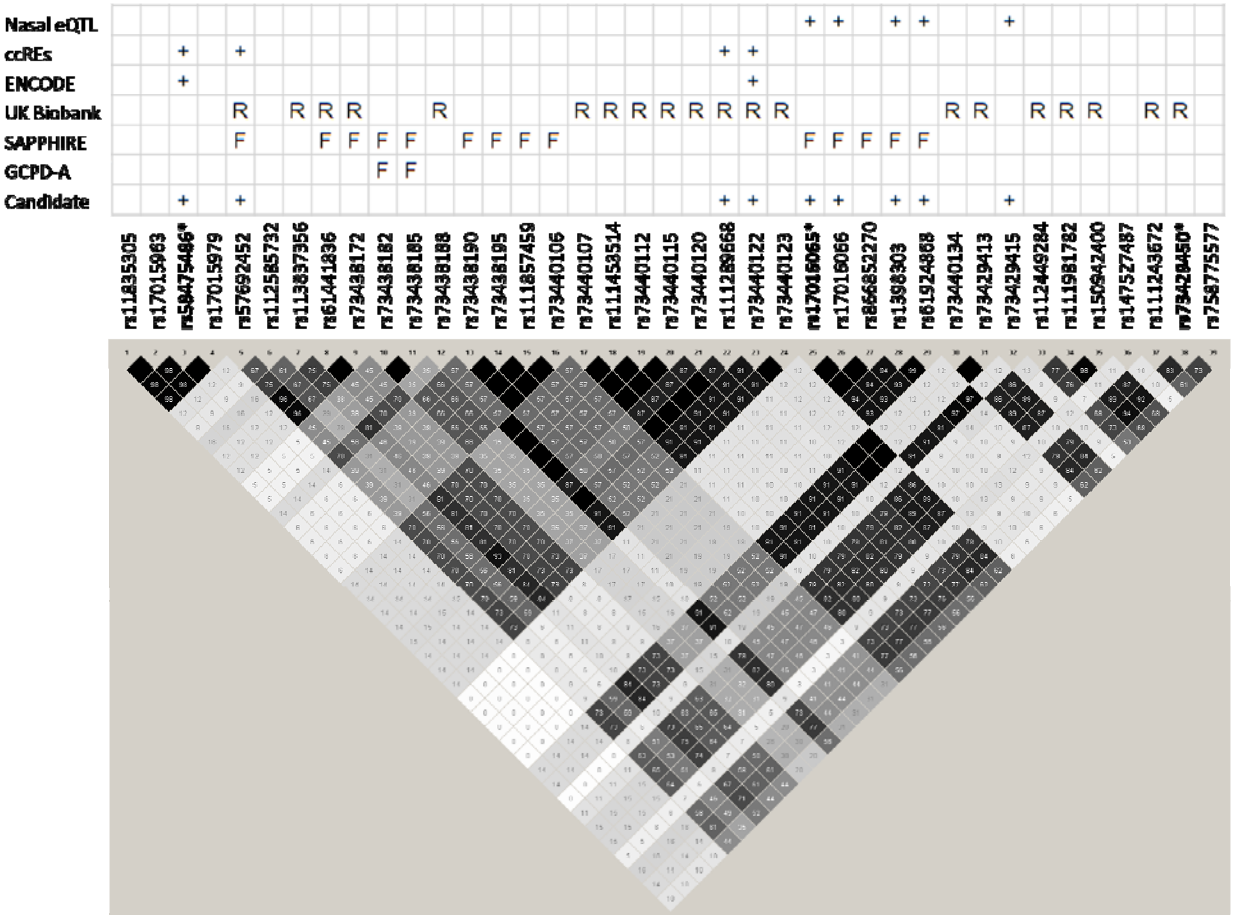
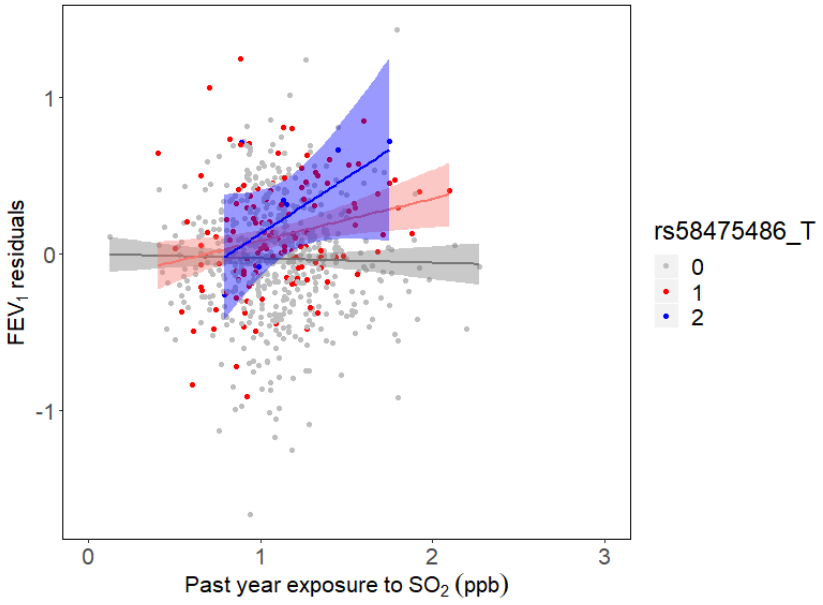


Figure 2. Integration of statistical and functional evidence for variant prioritization. Numbers and different shades of black in the LD plot represent LD in R^2. The three independent signals identified in the conditional analysis are marked with *. Nasal eQTL, variants eQTL of *KITLG* in nasal epithelial 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.

A



B

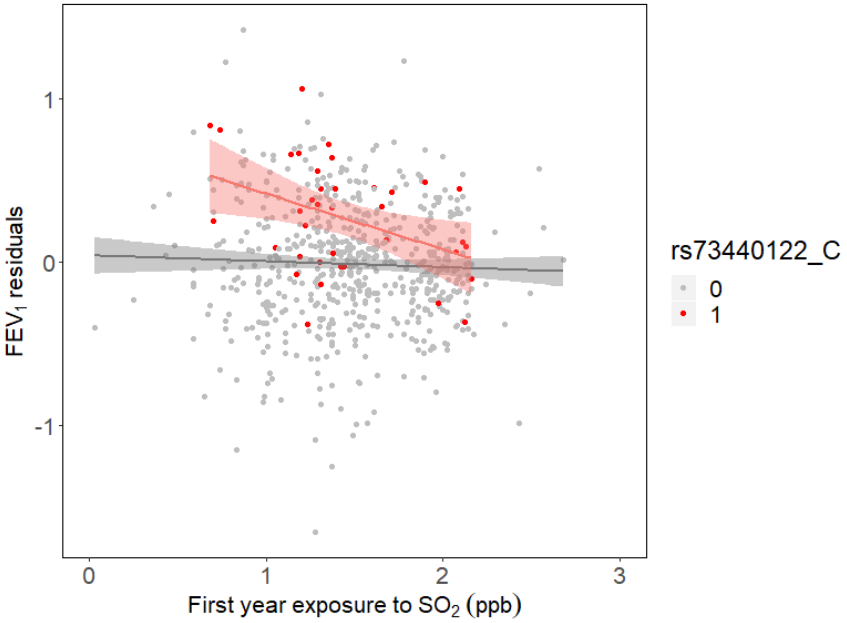


Figure 3. Gene-by-environment interaction analysis on FEV_1. FEV_1 residuals, residuals after FEV_1 was regressed on the covariates age, sex, height, controller medications, sequencing centers and the first 5

genetic PCs. FEV_1 residuals was plotted against (A) past year exposure to SO_2 stratified by the number of copies of T allele of rs58475486, (B) first year of life exposure to SO_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.

	African American (n=867)
Age	
Mean (SD)	14.1 (3.64)
Median [Min, Max]	13.8 [8.01, 21.8]
Sex	
Male	439 (50.6%)
Female	428 (49.4%)
Height (m)	
Mean (SD)	1.58 (0.145)
Median [Min, Max]	1.60 [1.14, 1.92]
Any control medications* in last 2 weeks	
No	543 (62.6%)
Yes	324 (37.4%)
ICS in last 2 weeks	
No	211 (24.3%)
Yes	306 (35.3%)
Missing	350 (40.4%)
LABA in last 2 weeks	
No	5 (0.6%)
Yes	94 (10.8%)
Missing	768 (88.6%)
Leukotriene inhibitor in last 2 weeks	
No	11 (1.3%)
Yes	68 (7.8%)
Missing	788 (90.9%)
African ancestry	
Mean (SD)	0.792 (0.129)
Median [Min, Max]	0.826 [0.118, 1.00]
Smoking status	
Never	793 (91.5%)
Past	72 (8.3%)
Current	0 (0%)
Missing	2 (0.2%)
Number of smokers in family	
0	469 (54.1%)
1	137 (15.8%)
2	42 (4.8%)
3+	10 (1.2%)
Missing	209 (24.1%)
SO2 first year exposure (ppb)	
Mean (SD)	1.59 (0.961)

Median [Min, Max]	1.50 [0.0300, 14.5]
Missing	227 (26.2%)
SO2 past year exposure (ppb)	
Mean (SD)	1.10 (0.302)
Median [Min, Max]	1.08 [0.120, 2.27]
Missing	206 (23.8%)
SO2 lifetime exposure (ppb)	
Mean (SD)	1.50 (0.371)
Median [Min, Max]	1.47 [0.964, 7.39]
Missing	206 (23.8%)
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.

rsID	Alt	β	p	β_{norm}	p_{norm}	MAF	1000 Genomes ALL	AFR	AMR	EUR	ccRE	DNaseI	TFBS	Nasal eQTL	UKBB	SAPPHIRE	GCPD-A	Score
rs11835305	T	0.126	1.93E-05	0.320	3.69E-05	0.104	0.036	0.119	0.012	0.001	N	N	N	N	N	N	N	0
rs17015963	C	0.126	1.93E-05	0.320	3.69E-05	0.104	0.036	0.120	0.012	0.001	N	N	N	N	N	N	N	0
rs58475486	T	0.127	1.45E-05	0.323	2.81E-05	0.105	0.037	0.123	0.012	0.001	Y	N	FOXA1 (A549)	N	N	N	N	2*
rs17015979	T	0.127	1.45E-05	0.323	2.81E-05	0.105	0.037	0.123	0.012	0.001	N	N	N	N	N	N	N	0
rs57692452	C	0.245	1.63E-06	0.654	1.06E-06	0.033	0.011	0.030	0.006	0.001	Y	N	N	N	Y	FF	N	2*
rs112585732	T	0.235	4.35E-07	0.625	3.06E-07	0.041	0.016	0.050	0.006	0.001	N	N	N	N	N	N	N	0
rs113837356	T	0.270	3.44E-06	0.719	2.49E-06	0.025	0.008	0.027	0.006	0.001	N	N	N	-	Y	N	N	1
rs61441836	G	0.252	8.19E-07	0.671	5.46E-07	0.033	0.010	0.030	0.006	0.001	N	N	N	N	Y	FF	N	1
rs73438172	A	0.252	8.19E-07	0.671	5.46E-07	0.033	0.010	0.030	0.006	0.001	N	N	N	N	Y	FF	N	1
rs1044043958[i]	A	0.270	3.44E-06	0.719	2.49E-06	0.025	-	-	-	-	N	N	N	-	N	N	N	0
rs73438182	G	0.138	1.68E-04	0.378	9.18E-05	0.064	0.020	0.068	0.006	0.001	N	N	N	N	N	FF	FF	0
rs73438185	A	0.138	1.68E-04	0.378	9.18E-05	0.064	0.020	0.068	0.006	0.001	N	N	N	N	N	FF	FF	0
rs73438188	A	0.297	6.97E-08	0.792	4.42E-08	0.028	0.010	0.027	0.006	0.001	N	N	N	N	Y	N	N	1
rs73438190	C	0.181	1.86E-05	0.486	1.22E-05	0.048	0.016	0.047	0.006	0.001	N	N	N	N	N	FF	N	0
rs73438195	A	0.181	1.86E-05	0.486	1.22E-05	0.048	0.016	0.047	0.006	0.001	N	N	N	N	N	FF	N	0
rs111857459	T	0.181	1.86E-05	0.486	1.22E-05	0.048	0.016	0.047	0.006	0.001	N	N	N	N	N	FF	N	0
rs144369986[i]	T	0.285	1.21E-06	0.756	9.44E-07	0.025	0.008	0.026	0.006	0.001	N	N	N	-	N	N	-	0
rs73440106	G	0.181	1.86E-05	0.486	1.22E-05	0.048	0.016	0.047	0.006	0.001	N	N	N	N	N	FF	N	0
rs73440107	A	0.297	6.97E-08	0.792	4.42E-08	0.028	0.010	0.027	0.006	0.001	N	N	N	N	Y	N	N	1
rs111453514	C	0.297	6.97E-08	0.792	4.42E-08	0.028	0.010	0.027	0.006	0.001	N	N	N	N	Y	N	N	1
rs73440112	T	0.297	6.97E-08	0.792	4.42E-08	0.028	0.010	0.027	0.006	0.001	N	N	N	N	Y	N	N	1
rs73440115	G	0.297	6.97E-08	0.792	4.42E-08	0.028	0.011	0.028	0.006	0.001	N	N	N	N	Y	N	N	1
rs11312747	A	0.133	1.43E-05	0.357	8.51E-06	0.100	0.036	0.121	0.010	0.001	N	N	N	-	N	N	N	0
rs73440120	A	0.285	1.21E-06	0.756	9.44E-07	0.025	0.008	0.026	0.006	0.001	N	N	N	-	Y	N	N	1
rs111289668	G	0.297	6.97E-08	0.792	4.42E-08	0.028	0.010	0.027	0.006	0.001	Y	N	N	N	Y	N	N	2*
rs73440122	C	0.292	2.08E-07	0.775	1.55E-07	0.027	0.011	0.030	0.006	0.001	Y	GM12865	SPI1 (HL-60)	N	Y	N	N	3*
rs73440123	G	0.292	2.08E-07	0.775	1.55E-07	0.027	0.011	0.030	0.006	0.001	N	N	N	N	Y	N	N	1

rs17016065	G	0.112	3.19E-06	0.296	2.54E-06	0.177	0.075	0.217	0.017	0.009	N	N	N	Y	N	FF	N	1#
rs17016066	A	0.112	3.19E-06	0.296	2.54E-06	0.177	0.075	0.217	0.017	0.009	N	N	N	Y	N	FF	N	1#
rs147400083[i]	T	0.112	3.19E-06	0.296	2.54E-06	0.177	-	-	-	-	N	N	N	-	N	-	N	0
rs866852270	T	0.112	3.19E-06	0.296	2.54E-06	0.177	-	-	-	-	N	N	N	-	N	FF	N	0
rs141293300	C	0.292	2.08E-07	0.775	1.55E-07	0.027	0.011	0.030	0.006	0.001	Y	N	N	-	N	N	N	1
rs1398303	A	0.104	1.22E-05	0.274	1.12E-05	0.186	0.077	0.223	0.020	0.009	N	N	N	Y	N	FF	N	1#
rs61924868	T	0.104	1.24E-05	0.275	1.14E-05	0.185	0.078	0.223	0.020	0.009	N	N	N	Y	N	FF	N	1#
rs73440134	T	0.292	2.08E-07	0.775	1.55E-07	0.027	0.011	0.030	0.006	0.001	N	N	N	N	Y	N	N	1
rs73429413	G	0.292	2.08E-07	0.775	1.55E-07	0.027	0.011	0.030	0.006	0.001	N	N	N	N	Y	N	N	1
rs73429415	A	0.096	5.13E-05	0.253	4.84E-05	0.189	0.078	0.225	0.022	0.009	N	N	N	Y	N	N	N	1#
rs112449284	T	0.242	4.64E-06	0.640	3.87E-06	0.031	0.012	0.035	0.007	0.001	N	N	N	N	Y	N	N	1
rs111981782	C	0.296	5.78E-08	0.786	4.09E-08	0.029	0.012	0.033	0.006	0.001	N	N	N	N	Y	N	N	1
rs150942400	T	0.293	6.01E-08	0.780	4.01E-08	0.029	0.012	0.034	0.006	0.002	N	N	N	N	Y	N	N	1
rs147527487	C	0.086	8.49E-05	0.226	8.14E-05	0.205	0.095	0.249	0.016	0.005	N	N	N	N	N	N	N	0
rs111243672	A	0.258	6.41E-07	0.690	3.99E-07	0.032	0.014	0.037	0.007	0.004	N	N	N	N	Y	N	N	1
rs73429450	A	0.302	1.26E-08	0.801	9.01E-09	0.031	0.012	0.033	0.009	0.002	N	N	N	N	Y	N	N	1
rs758775577	C	0.217	2.22E-06	0.574	1.85E-06	0.041	-	-	-	-	N	N	N	N	N	N	N	0
rs142679473[i]	C	0.285	6.30E-08	0.756	4.62E-08	0.031	0.012	0.033	0.009	0.002	N	N	N	-	N	N	N	0

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. β (p) and β_{norm}(p_{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. DNaseI, 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 3. Gene-and-environment analysis on FEV_1

Variant	Exposure	β.variant	β.env	β.gxe	p.variant	p.env	p.gxe	
rs58475486_T	SO_2 first year	0.18	-0.04	-0.03	0.119	0.009	0.658	
rs57692452_C	SO_2 first year	0.61	-0.04	-0.24	0.006	0.007	0.091	
rs111289668_G	SO_2 first year	0.68	-0.04	-0.27	0.003	0.007	0.079	
rs73440122_C	SO_2 first year	0.73	-0.04	-0.32	0.002	0.008	0.045	*
rs17016065_G	SO_2 first year	0.23	-0.04	-0.08	0.004	0.027	0.108	
rs17016066_A	SO_2 first year	0.23	-0.04	-0.08	0.004	0.027	0.108	
rs1398303_A	SO_2 first year	0.23	-0.04	-0.09	0.004	0.030	0.082	
rs61924868_T	SO_2 first year	0.22	-0.04	-0.08	0.004	0.029	0.088	
rs73429415_A	SO_2 first year	0.23	-0.04	-0.09	0.004	0.031	0.069	
rs58475486_T	SO_2 past year	-0.21	-0.04	0.32	0.084	0.478	0.003	**
rs57692452_C	SO_2 past year	-0.09	0.01	0.29	0.680	0.839	0.100	
rs111289668_G	SO_2 past year	-0.18	0.01	0.41	0.444	0.869	0.037	*
rs73440122_C	SO_2 past year	-0.16	0.01	0.39	0.482	0.825	0.051	
rs17016065_G	SO_2 past year	-0.11	-0.03	0.20	0.299	0.583	0.026	*
rs17016066_A	SO_2 past year	-0.11	-0.03	0.20	0.299	0.583	0.026	*
rs1398303_A	SO_2 past year	-0.12	-0.04	0.21	0.231	0.521	0.021	*
rs61924868_T	SO_2 past year	-0.12	-0.04	0.21	0.248	0.534	0.023	*
rs73429415_A	SO_2 past year	-0.12	-0.04	0.20	0.246	0.539	0.026	*
rs58475486_T	SO_2 lifetime	-0.24	-0.13	0.26	0.390	0.001	0.173	
rs57692452_C	SO_2 lifetime	-0.44	-0.13	0.47	0.436	0.001	0.221	
rs111289668_G	SO_2 lifetime	-0.19	-0.13	0.32	0.764	0.001	0.444	
rs73440122_C	SO_2 lifetime	-0.15	-0.13	0.29	0.809	0.001	0.489	
rs17016065_G	SO_2 lifetime	0.40	-0.10	-0.19	0.043	0.013	0.143	
rs17016066_A	SO_2 lifetime	0.40	-0.10	-0.19	0.043	0.013	0.143	
rs1398303_A	SO_2 lifetime	0.36	-0.11	-0.17	0.064	0.012	0.184	
rs61924868_T	SO_2 lifetime	0.35	-0.11	-0.17	0.070	0.012	0.199	
rs73429415_A	SO_2 lifetime	0.34	-0.11	-0.16	0.078	0.012	0.207	

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