ONLINE DATA SUPPLEMENT

TITLE

Whole Genome Sequencing Identifies Novel Functional Loci Associated with Lung Function in Puerto Rican Youth

AUTHORS

Eunice Y. Lee, Angel C.Y. Mak, Donglei Hu, Satria Sajuthi, Marquitta White, Kevin L. Keys, Walter Eckalbar, Luke Bonser, Scott Huntsman, Cydney Urbanek, Celeste Eng, Deepti Jain, Gonçalo Abecasis, Hyun M. Kang, Soren Germer, Michael C. Zody, Deborah A. Nickerson, David Erle, Elad Ziv, Jose Rodriguez-Santana, Max A. Seibold, Esteban G. Burchard

Study Subjects

Subjects were eligible to participate if they were 8-21 years of age and identified all four grandparents as Latino. Participants were excluded if they had: (1) 10 or more pack-years of smoking; (2) any smoking within 1 year of recruitment date; (3) pregnancy in the third trimester; or (4) history of one of the following conditions: sickle cell disease, cystic fibrosis, sarcoidosis, cerebral palsy, or history of heart or chest surgery. A full description about the study design and recruitment has been previously described elsewhere[1-3]. Subjects who met the eligibility criteria completed in-person questionnaires, which provided medical, asthma, allergic, social, environmental and demographic information, and provided blood sample for DNA analysis. In addition, nasal brushing samples were collected for RNA sequencing. Asthma case status was self-reported physician-diagnosis further verified based on reports of asthma medication use and symptoms of coughing, and wheezing or shortness of breath in the 2 years preceding enrollment.

Local and Global Ancestry Estimation

Reference genotypes for European and African ancestries were obtained from the Axiom® Genotype Data Set. The Axiom® Genome-Wide LAT 1 array was used to generate the Native American ancestry reference genotypes, which came from the 71 Native Americans (17 Zapotec, 2 Mixe and 11 Mixtec from Oaxaca, 44 Nahua from Central Mexico). We describe in more detail elsewhere[4]. To estimate local ancestry tracts, we joined the biallelic SNPs that PASS filter from our WGS data with the corresponding sites on the Axiom® Genome-Wide LAT 1 array. We merged these data with our European (CEU), African (YRI), and Native American (NAM) reference panels, which overlapped with 402,838 markers with less than 10% missing data. We phased all ancestral and study samples using BEAGLE 5.0 prior to calling local ancestry tracts. Phased genotypes and a GRCh38 genetic map distributed with BEAGLE 5.0 were used to estimate local ancestry using RFMIX version 2 assuming three ancestral populations: African (HapMap YRI), European (HapMap CEU), and Native American (reference genotypes came from the 71 Native Americans described above). To estimate the global ancestry, ADMIXTURE was used with the reference genotypes in a supervised analysis assuming three ancestral populations.

Whole Genome Sequencing Data Generation, Processing and Quality Control

DNA was isolated from whole blood using the Wizard Genomic DNA Purification kits (Promega, Fitchburg, WI). DNA samples were quantified by fluorescence using the Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific, Waltham, MA) on a Spectramax fluorometer (Molecular Devices, Sunnyvale, CA). DNA samples were sequenced as part of the Trans-Omics for Precision Medicine (TOPMed) whole genome sequencing (WGS) program[5]. WGS was performed at the New York Genome Center and the Northwest Genomics Center on a HiSeqX system (Illumina, San Diego, CA) using a paired-end read length of 150 base pairs, to a minimum of 30X mean genome coverage. Details on DNA sample handling, quality control, library construction, clustering and sequencing, read processing, and sequence data quality control were previously described elsewhere[6]. Variant calls were obtained from TOPMed data freeze 8 variant call format files aligned to the GRCh38 genome reference. In the common variant analysis of this study, we excluded indels and focused only on biallelic SNPs. SNPs with a minimal depth of coverage of 10 reads that had a value of PASS in the VCF FILTER column were included in our analyses. Data quality control was performed in PLINK 1.9[7]. We purged SNPs that had less than 95% call rate, that showed an allele frequency less than 1%, or that failed the Hardy-Weinberg test at 0.0001 significance threshold level. We removed individuals with more than 5% missing genotypes and pairs of individuals that showed genetic relatedness equivalent to second cousin or closer (PLINK flags --genome --min 0.025). After quality control of WGS data, there were 76,988 SNPs within the 5.4 Mb at 1q32 and 12,032 number of SNPs within the 2.6 Mb at 5q35.1 candidate regions.

Admixture Mapping

There were 402,838 loci with locus-specific ancestry estimates, which were used for our genome-wide admixture mapping. We adopted a similar approach as Shriner and our previous studies to determine the effective number of tests[8-10]. We accomplished this by fitting an autoregression model to the summary statistics using the coda package in R to determine the effective number of tests (n=595). This approach corrects for the effective number of tests, which is often smaller than the empirical number of tests due to local genetic correlation between the SNPs used in admixture mapping. A Bonferroni corrected threshold of genome-wide significance at alpha level 0.05 was then defined as $\alpha=0.05/595 = 8.40 \times 10^{-5}$.

Fine Mapping: Conditional Analysis

To further understand the genetic variation underlying the association between local ancestry with lung function, fine mapping of the regions identified using admixture mapping was performed with the genome sequence data described previously. In order to define the boundary of the genomic region of interest, we used the threshold (p-value < 8.40×10^{-5}) to find the base positions of the start and end SNPs that defined a region. We extracted the sequenced data within the boundary of admixture mapping region of interest. The allelic association tests were then conducted adjusting for the same covariates in the local ancestry regression model.

The joint effects of multiple SNPs on lung function were evaluated by a stepwise regression model using a forward selection procedure adjusting for the same covariates mentioned above and the lead locus-specific ancestry signal (SNP with the smallest p-value from admixture mapping)[11]. SNPs were first rank-ordered by their strength of association (p-values) and added to the regression model one SNP at a time until the effect of the lead locus-specific ancestry signal (SNP) reached p-value > 0.05. To account for the local linkage disequilibrium (LD) structure, SNPs were filtered based on their pairwise correlation (r^2>0.8). The final model included the minimum number of independent and significant SNPs associated with lung function.

Region-based association analysis

Region-based association analyses of rare variants (MAF < 0.01) conditioned on the top admixture mapping SNP were performed in 1 Kb sliding windows with 500 bp increments within the admixture mapping peaks (defined above) using SKAT-O from the SKAT R package v1.3.2.1[12]. Since SKAT imputes missing genotypes by default using mean genotype values (impute.method="fixed"), we chose to use low coverage genotypes instead of SKAT imputation. Consequently, TOPMed freeze 8 DP0 variants with a VCF FILTER of PASS were included in the analysis. The function effectiveSize in the R package CODA was used to estimate the effective number of independent hypothesis tests for accurate Bonferroni multiple testing corrections[13]. P-value thresholds for statistical significance and suggestive significance were defined as 0.05 and 1 divided by the effective number of tests, respectively. To study the contribution of individual variant to a region-based association p-value, drop-one variant analyses were performed by repeating the region-based analysis excluding one variant at a time.

Human bronchial epithelial cell culture and H3K27ac ChIP-seq assay

Human bronchial epithelial cells (HBECs) were cultured at an air-liquid interface as previously described[14]. The use of HBECs isolated from lungs not used for transplantation was approved by the UCSF Committee on Human Research; written consent was not required since materials were leftover clinical samples obtained from de-identified individuals.

H3K27ac ChIP-seq was performed as previously reported for BSMCs[4]. Briefly, PFA-fixed HBECs were lysed, then chromatin was sheared using a Covaris S2 sonicator. Chromatin immunoprecipitation was performed with a H3K27ac antibody (Abcam) using the Diagenode Low Cell Chip kit as per the manufacturer's instructions. Libraries were prepared using the Rubicon DNA-Seq kit and then sequenced.

Functional Annotation and Validation of Genetic Variants

Genomic annotation was conducted for the lead ancestry SNP, and the SNPs identified from fine-mapping analyses, as well as the SNPs in linkage disequilibrium with them ($r^2 > 0.8$). Functional annotation was performed using public data (ENCODE, HUGIn, and RegulomeDB)[15-17]. Each SNP was investigated for its biological function and impact using DNase I hypersensitive site sequencing, histone modification ChIP-seq, transcription factor binding ChIP-seq, and Hi-C data. We used gene expression and eQTL data derived from nasal epithelial cells (see TEXT E8) and H3K27ac ChIP-Seq peaks in bronchial smooth muscle cells (BSMCs, published)[4] and human bronchial epithelial cells (HBECs, see TEXT E6) to provide additional evidence for the admixture signal and delineate the potential biological links between genetic variants and lung function. SNPs with one or more supportive functional evidence were prioritized as candidate SNPs contributing to the lung function genetic association signals.

RNA sequencing and Expression Quantitative Trait Loci (eQTL) analysis

This manuscript leverages nasal airway epithelial expression data on 3 genes and eQTL data for the *TMEM9* gene, which are part of manuscripts in preparation. Generation of this expression and eQTL data is summarized below. Whole-transcriptome RNA-seq libraries of 681 nasal brushings from GALA II Puerto Rican children with and without asthma (434 with asthma and 247 controls) were constructed using Kapa Biosystems mRNA-seq library kits (catalog number KK8421) in conjunction with the Beckman Coulter Biomek FX^P automation system (Beckman Coulter, Fullerton, CA). Libraries were sequenced with the Illumina HiSeq 2500 system. Raw RNA-Seq reads were trimmed using Skewer[18] and mapped to human reference genome hg38 using Hisat2[19]. Reads mapped to genes were counted with htseq-count using the UCSC hg38 GTF file as reference[20]. Genome-wide genetic variation data used for the cis-expression quantitative trait locus (eQTL) analysis were generated from whole genome sequencing data as described above. Cis-eQTL analysis was performed as described in the Genotype-Tissue Expression (GTEx) project version 7 protocol using age, sex, BMI, global African and European ancestries and 60 Probabilistic Estimation of Expression Residuals (PEER) factors as covariates[21]. eQTLs are defined as genetic variants that were associated with gene expression at False Discovery Rate (FDR) of beta-approximated p-value < 0.05 using fastQTL[22]. In order to determine whether the prioritized variants (TEXT E7) are associated with gene expression, we examined their overlap with all eQTLs within an 1 Mb flanking region. Further, we investigated correlation between phenotype and gene expression level using normalized gene counts. As a supplementary analysis, we tested previously reported relationships between *TMEM9*, *IL-6*, and *IL-1β* in the Wnt inflammatory pathway by examining their gene expression correlation[23,24].

REFERENCES

1. Neophytou AM, White MJ, Oh SS, et al. Air Pollution and Lung Function in Minority Youth with Asthma in the GALA II (Genes-Environments and Admixture in Latino Americans) and SAGE II (Study of African Americans, Asthma, Genes, and Environments) Studies. Am J Respir Crit Care Med 2016;193:1271-80.

2. Nishimura KK, Galanter JM, Roth LA, et al. Early-life air pollution and asthma risk in minority children. The GALA II and SAGE II studies. Am J Respir Crit Care Med 2013;188:309-18.

3. Thakur N, Oh SS, Nguyen EA, et al. Socioeconomic status and childhood asthma in urban minority youths. The GALA II and SAGE II studies. Am J Respir Crit Care Med 2013;188:1202-9.

4. Mak ACY, White MJ, Eckalbar WL, et al. Whole-Genome Sequencing of Pharmacogenetic Drug Response in Racially Diverse Children with Asthma. Am J Respir Crit Care Med 2018;197:1552-64.

5. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv 2019:563866.

6. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084-97.

7. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.

8. Gignoux CR, Torgerson DG, Pino-Yanes M, et al. An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos. J Allergy Clin Immunol 2019;143:957-69.

9. Pino-Yanes M, Thakur N, Gignoux CR, et al. Genetic ancestry influences asthma susceptibility and lung function among Latinos. J Allergy Clin Immunol 2015;135:228-35.

10. Shriner D. Overview of admixture mapping. Curr Protoc Hum Genet 2013;Chapter 1:Unit 1 23.

11. Fejerman L, Chen GK, Eng C, et al. Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. Hum Mol Genet 2012;21:1907-17.

12. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 2012;91:224-37.

13. Plummer M, Best, K., Cowles and K. Vines. CODA: convergence diagnosis and output analysis for MCMC. 2006.

14. Bonser LR, Zlock L, Finkbeiner W, Erle DJ. Epithelial tethering of MUC5AC-rich mucus impairs mucociliary transport in asthma. J Clin Invest 2016;126:2367-71.

15. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012;22:1790-7.

16. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57-74.

17. Martin JS, Xu Z, Reiner AP, et al. HUGIn: Hi-C Unifying Genomic Interrogator. Bioinformatics 2017;33:3793-5.

18. Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. Bmc Bioinformatics 2014;15:182.

19. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357-60.

20. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166-9.

21. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, et al. Genetic effects on gene expression across human tissues. Nature 2017;550:204-13.

22. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics 2016;32:1479-85.

23. Jevnikar Z, Ostling J, Ax E, et al. Epithelial IL-6 trans-signaling defines a new asthma phenotype with increased airway inflammation. J Allergy Clin Immunol 2019;143:577-90.

24. Wei W, Jiang F, Liu XC, Su Q. TMEM9 mediates IL-6 and IL-1beta secretion and is modulated by the Wnt pathway. Int Immunopharmacol 2018;63:253-60.

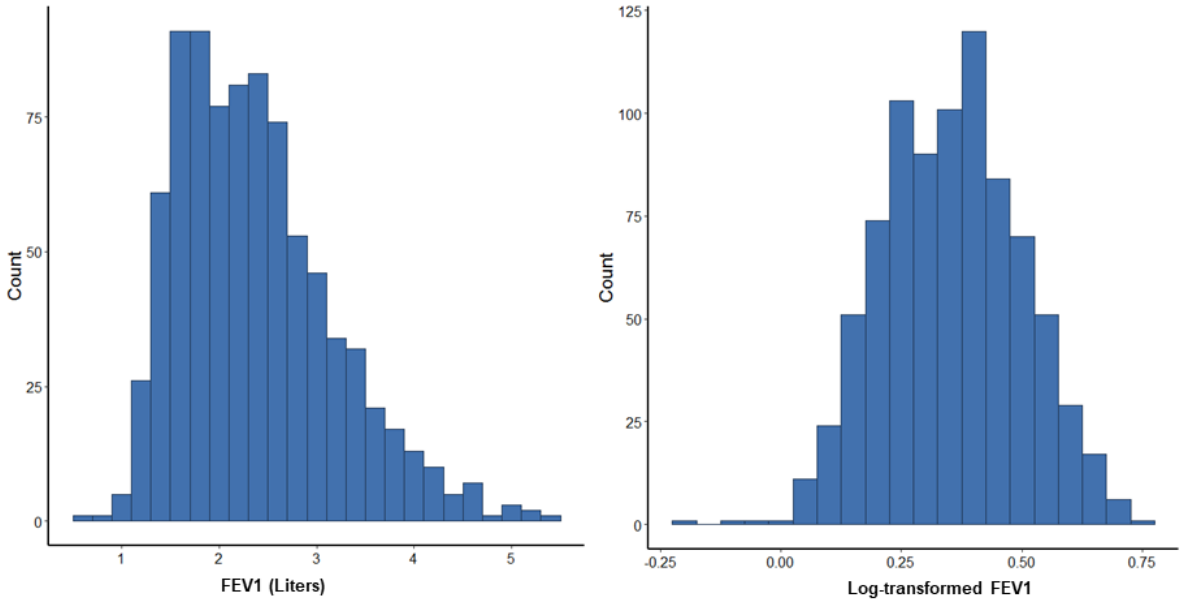
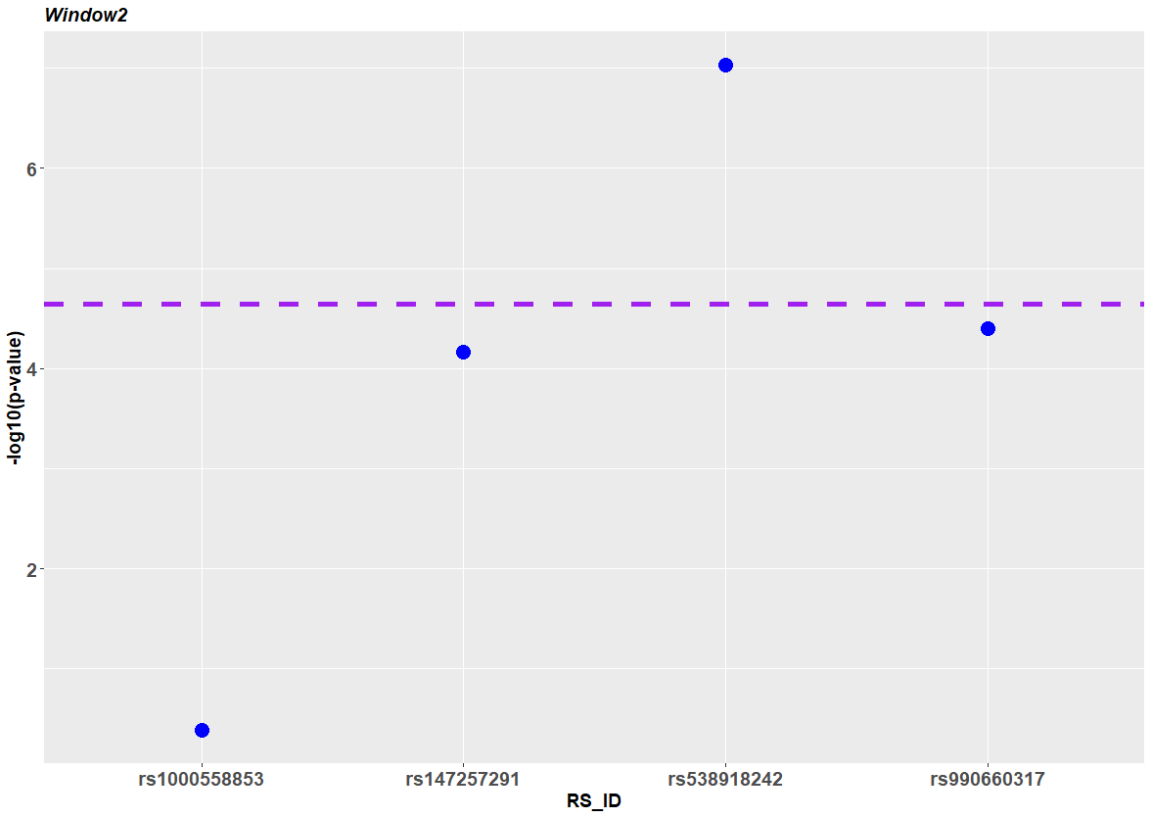
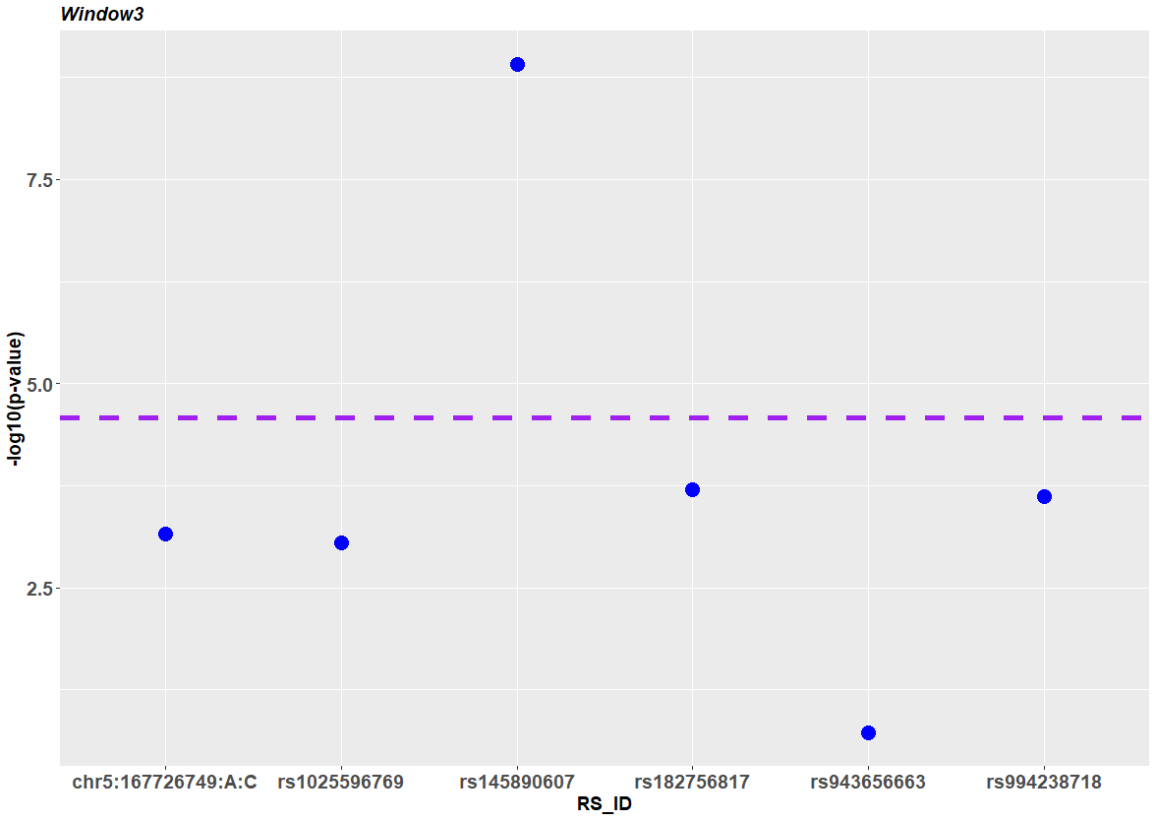


Figure E1. Distributinon of lung function measurements (FEV_1) in regular scale on the left and log-transformed scale on the right.


Window2
6
4
2
-log10(p-value)
rs1000558853
rs147257291
rs538918242
rs990660317
RS_ID


Window3
7.5
5.0
2.5
-log10(p-value)
chr5:167726749:A:C
rs1025596769
rs145890607
rs182756817
rs943656663
rs994238718
RS_ID

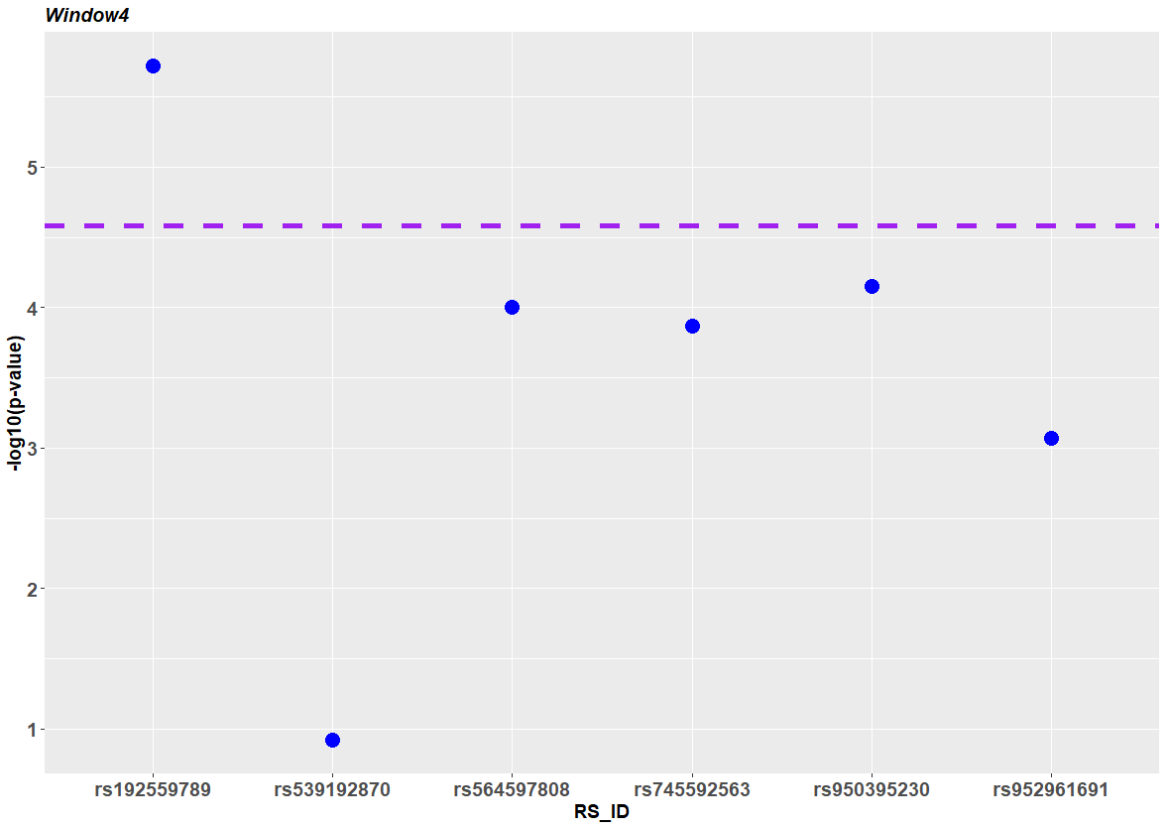


Figure E2. Region-based association tests. For each suggestive window, a drop-one variant analysis was conducted to examine the effect of each rare variant on lung function (log-transformed FEV_1). The p-value associated with each variant represents a statistical significance of the association between a suggestive window and lung function when that particular variant was removed from the window.

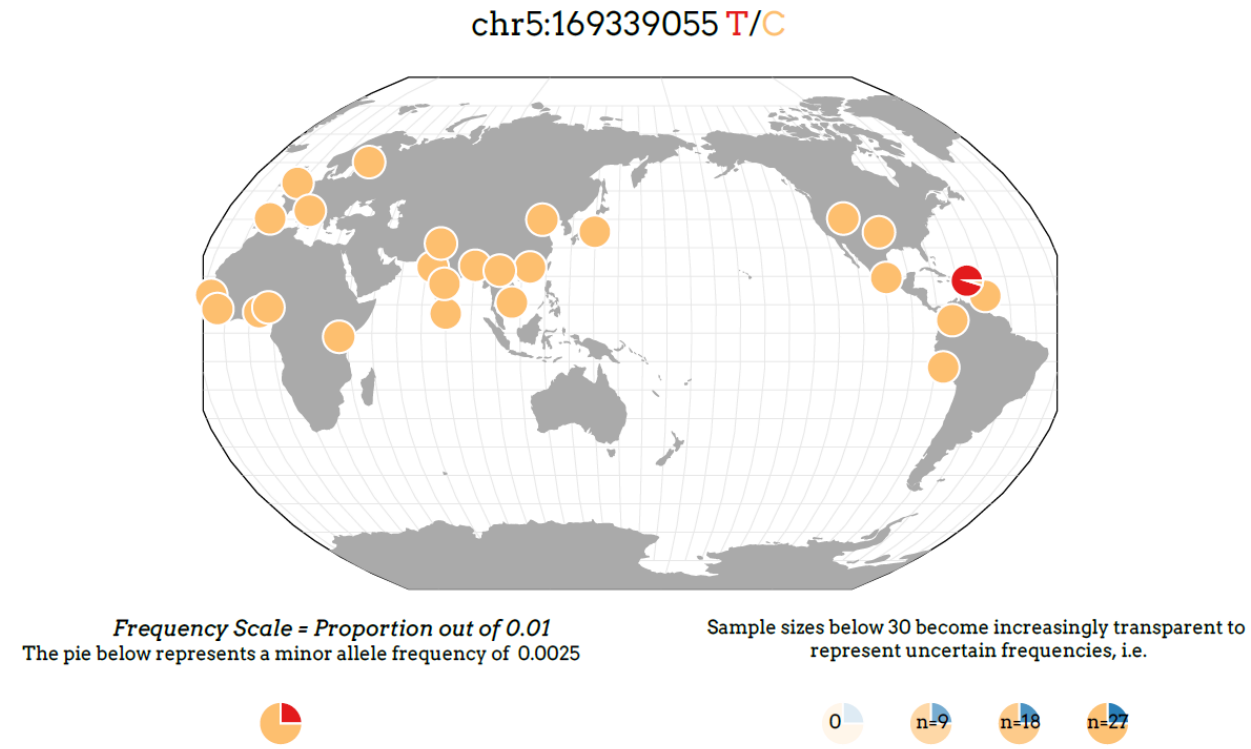
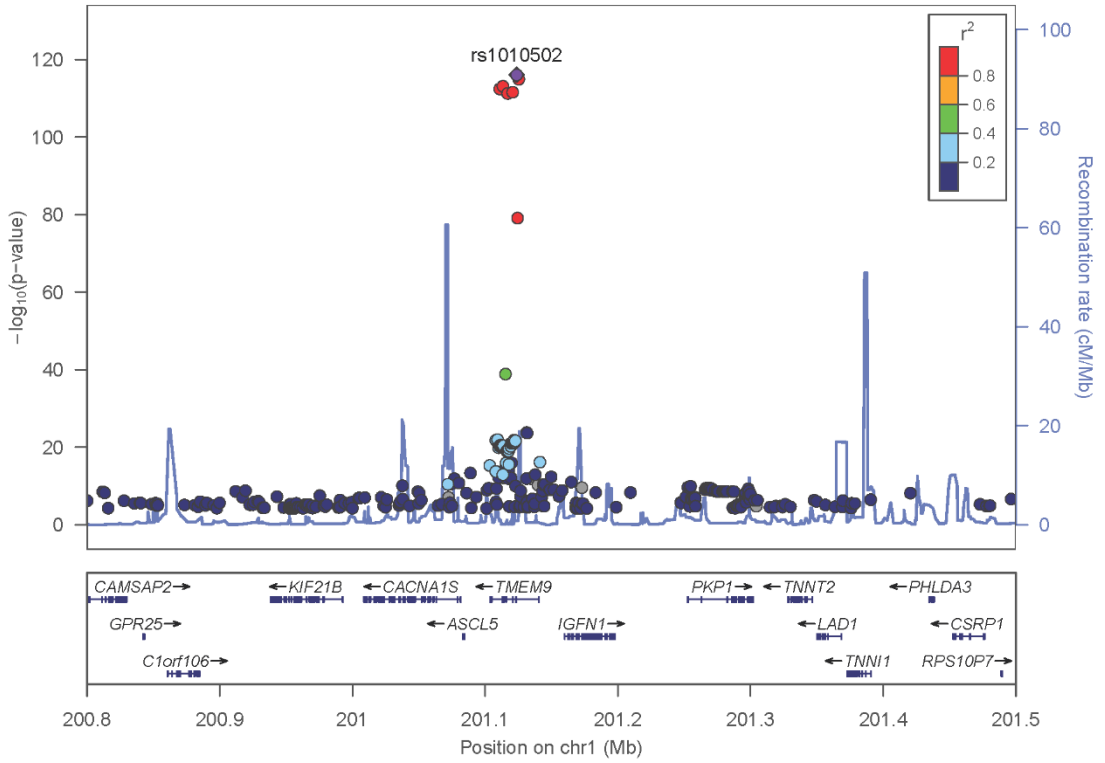
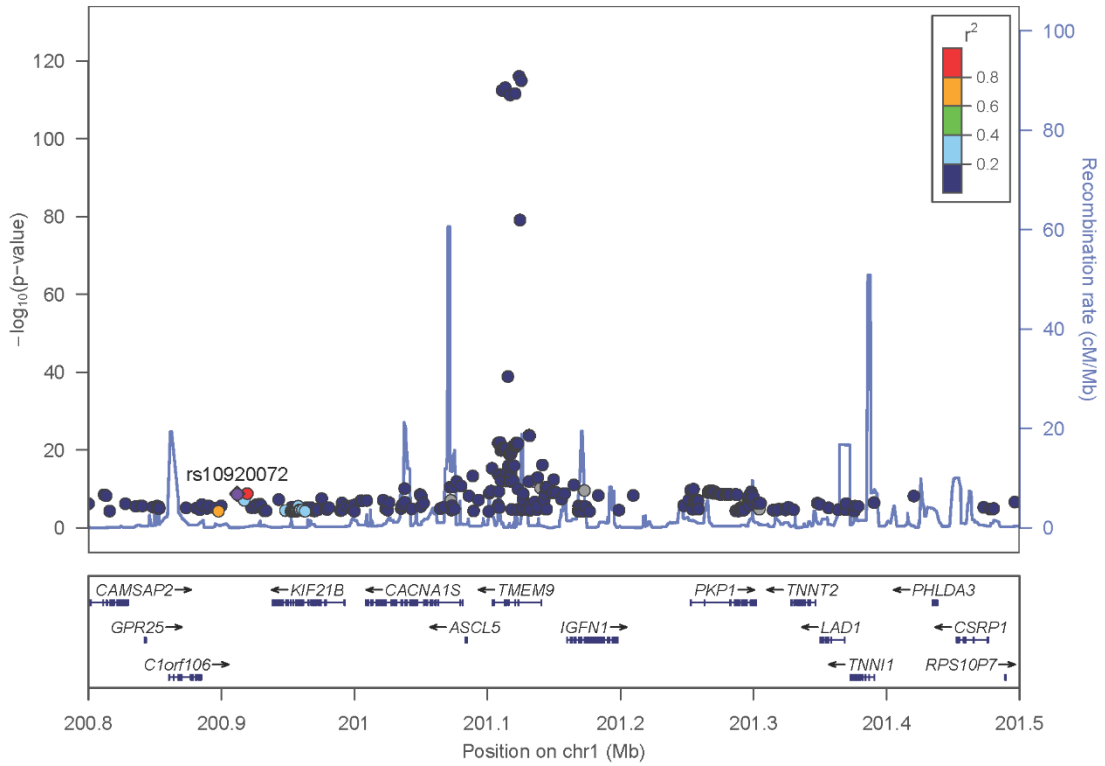


Figure E3. Global distribution of minor allele frequency. Minor allele of the rare variant, rs539192870, was observed only in Puerto Ricans and almost never found in other populations.


rs1010502
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
CAMSAP2
KIF21B
CACNA1S
TMEM9
PKP1
TNNT2
PHLDA3
GPR25
ASCL5
IGFN1
LAD1
CSRP1
C1orf106
TNNI1
RPS10P7
Position on chr1 (Mb)


rs10920072
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
CAMSAP2
KIF21B
CACNA1S
TMEM9
PKP1
TNNT2
PHLDA3
GPR25
ASCL5
IGFN1
LAD1
CSRP1
C1orf106
TNNI1
RPS10P7
Position on chr1 (Mb)

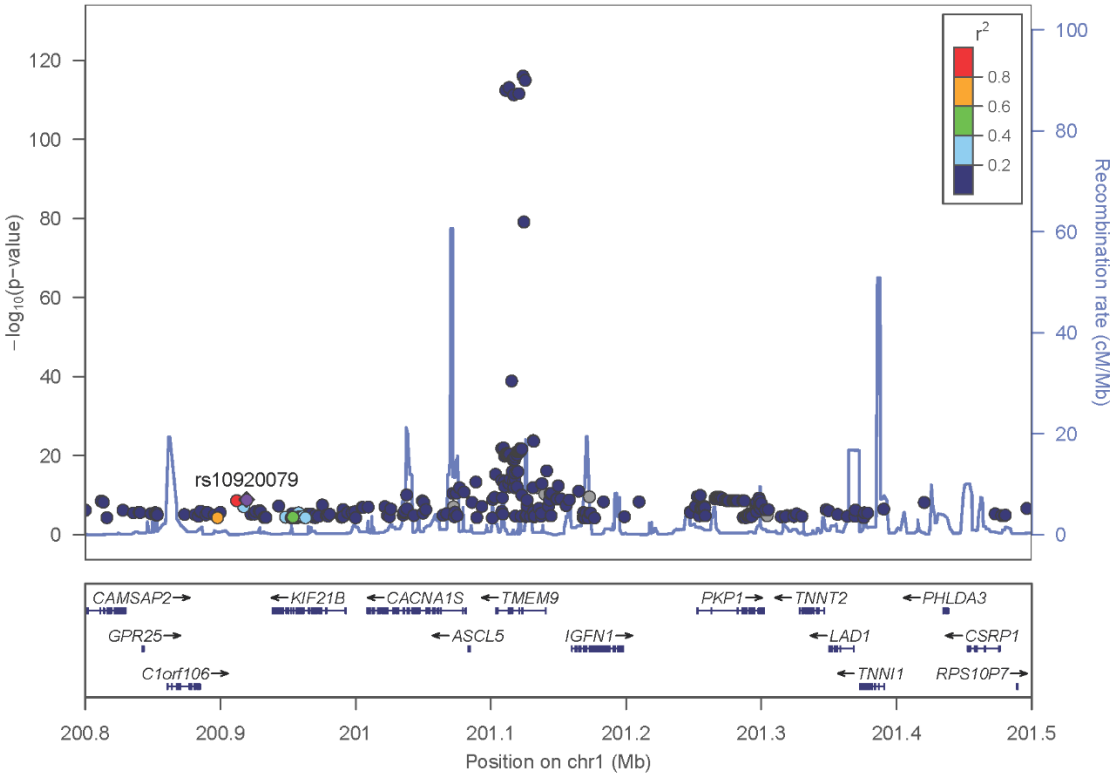


Figure E4. LocusZoom plot displaying the eQTLs for *TMEM9*. SNPs rs10920072 and rs10920079 are weakly correlated with the main eQTLs for *TMEM9*.

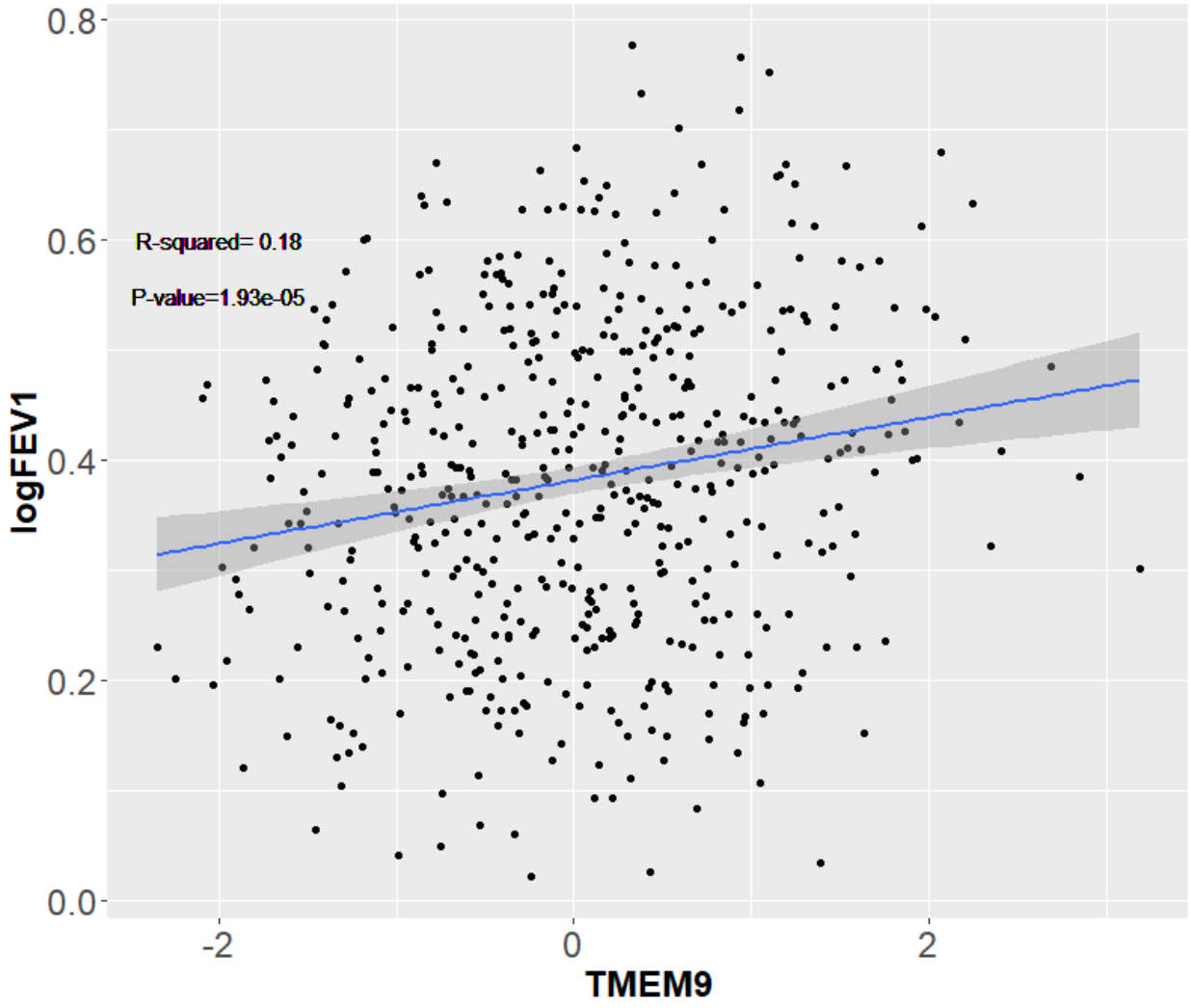


Figure E5. Relationship between *TMEM9* gene expression on the x-axis and lung function on the y-axis. The p-value represents statistical significance of the univariate regression coefficient.

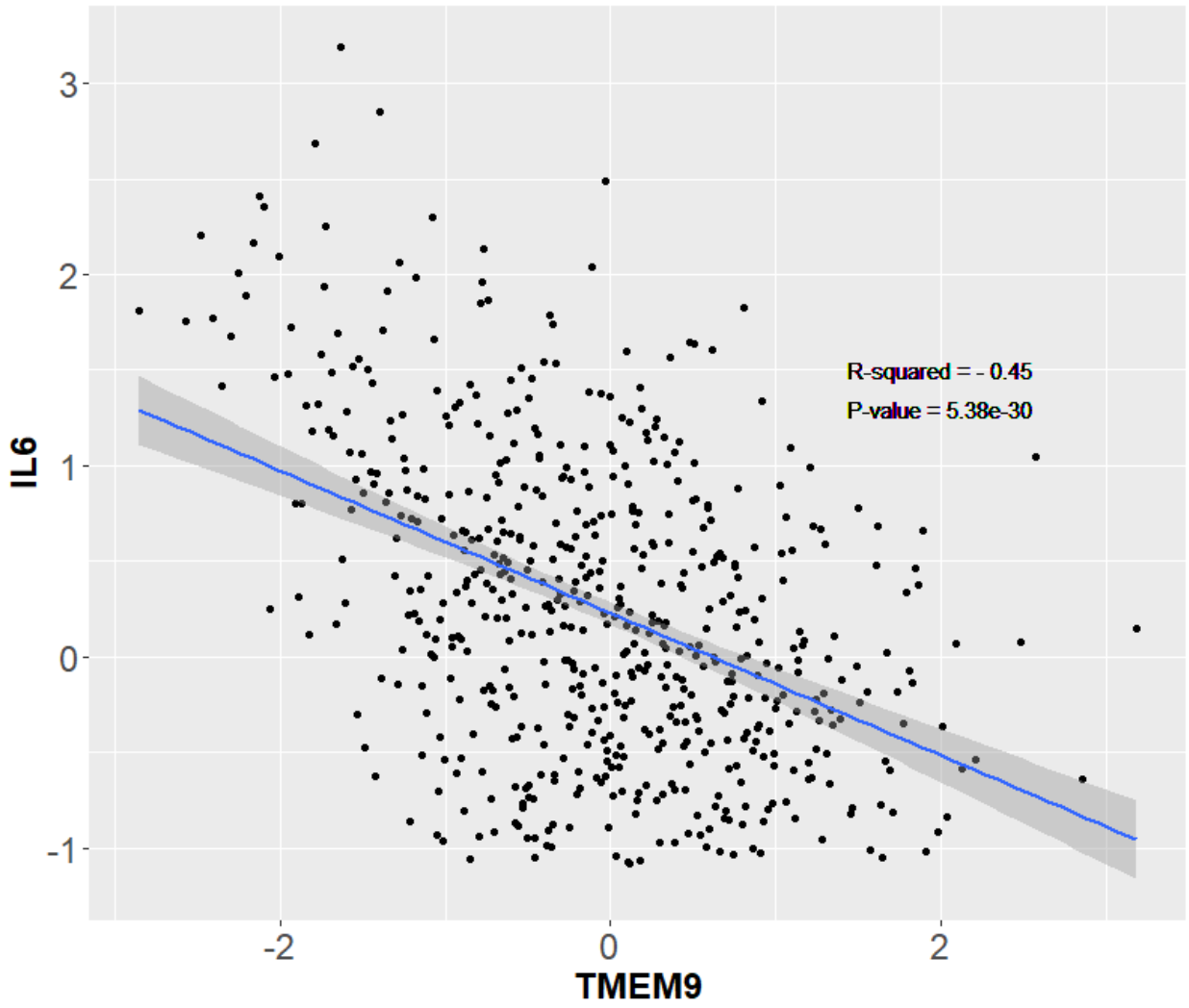


Figure E6. Relationship between *TMEM9* gene expression on the x-axis and *IL-6* expression levels on the y-axis. The p-value represents statistical significance of the univariate regression coefficient.

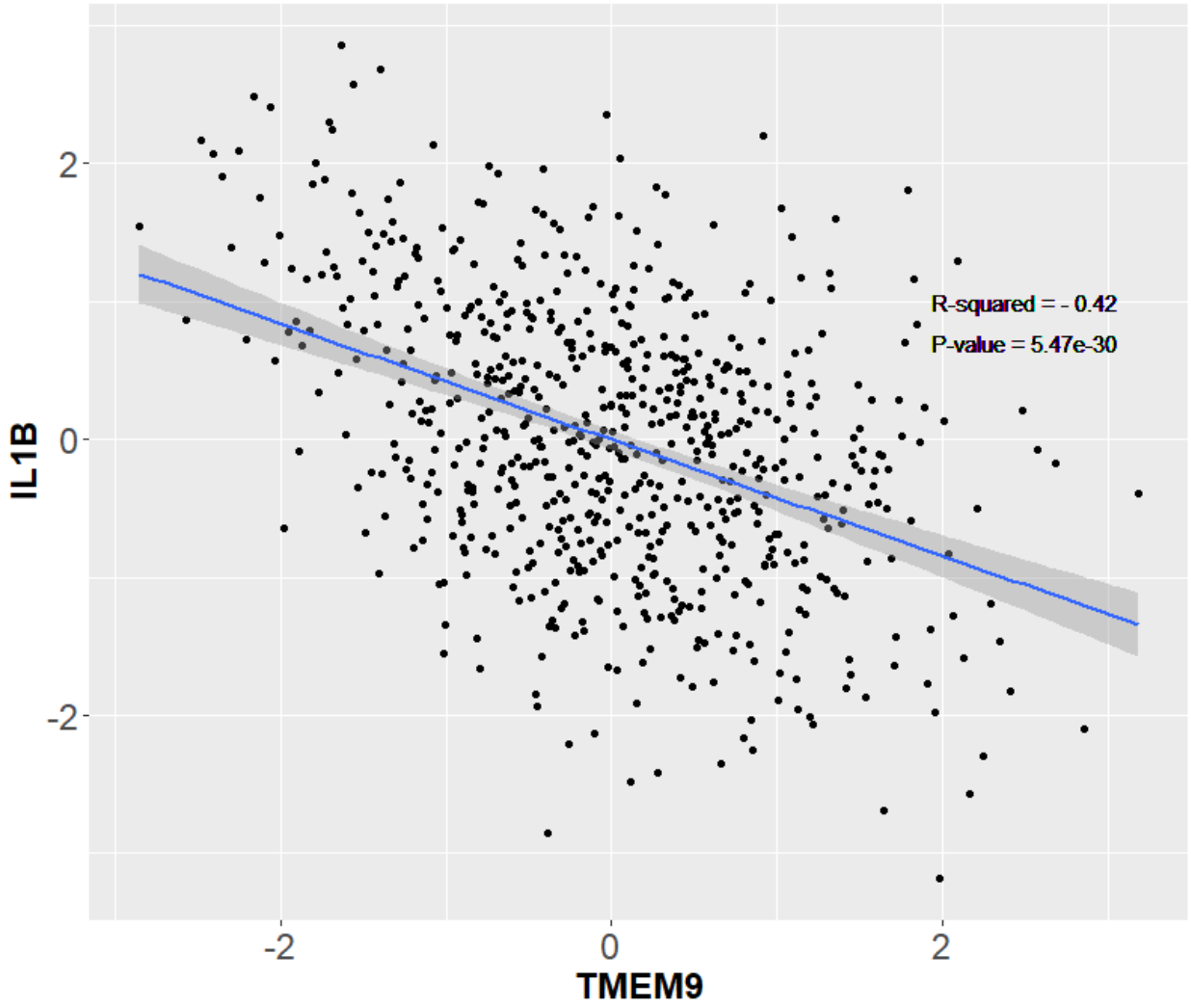


Figure E7. Relationship between *TMEM9* gene expression on the x-axis and *IL-1β* expression levels on the y-axis. The p-value represents statistical significance of the univariate regression coefficient.

SNP ID	BP (hg38)	Alleles	Allele Frequency			βeta	P-value	Gene
			AFR	EUR	PUR[¶]			
African Ancestry								
rs16847664	201074574	G/A	0.07	0	0.03	-0.08	1.07×10^{-6}	*CACNA1S*
rs6679485	194642520	G/T	0.31	0.04	0.26	-0.02	3.72×10^{-6}	.
rs116270262	200713581	T/A	0.18	0.02	0.11	-0.04	4.11×10^{-6}	.
rs10920079	200950242	G/A	0.66	0.06	0.42	-0.02	4.23×10^{-6}	*MROH3P*
Native American Ancestry								
rs187841563	169794681	C/T	0	0.01	0.03	-0.07	1.09×10^{-5}	*DOCK2*
rs11134502	168279460	T/C	0.30	0.62	0.70	-0.01	5.21×10^{-5}	.
rs73320187	169764658	G/A	0.11	0.04	0.32	0.02	6.81×10^{-5}	*DOCK2*
rs4323254	167438477	T/C	0.86	0.98	0.14	0.03	8.19×10^{-5}	*TENM2*
rs74677211	170180072	G/A	0.04	0.05	0.06	0.05	8.87×10^{-5}	.
rs62384405	167937881	C/T	0.02	0.10	0.13	-0.03	9.24×10^{-5}	*TENM2*
rs12522512	168493664	G/A	0.03	0.01	0.14	0.03	1.19×10^{-4}	*RARS*

Table E1. Individual allelic association tests between lung function (log-transformed FEV_1) and SNPs discovered from fine-mapping conditional analysis with forward selection regression models. All regression models were adjusted for age, sex, height, asthma status, and both African and Native American ancestries.

[¶]Minor allele frequency computed from the Puerto Rican study participants (n=836)

Window 2

Chr	Position	RS_ID	MAF	MAC	P-value	Overlapping gene
1	199360074	rs1000558853	0.0006	1	0.41	*LINC02789*
1	199360172	rs147257291	0.0006	1	6.88e-05	*LINC02789*
1	199359644	rs538918242	0.001	2	9.31e-08	*LINC02789*
1	199360216	rs990660317	0.0006	1	4.00e-05	*LINC02789*

Window 3

Chr	Position	RS_ID	MAF	MAC	P-value	Overlapping gene
5	167726749	chr5:167726749:A:C	0.0006	1	6.93e-04	*TENM2*
5	167726353	rs1025596769	0.0006	1	9.04e-04	*TENM2*
5	167727034	rs145890607	0.002	4	1.23e-09	*TENM2*
5	167726803	rs182756817	0.0006	1	1.97e-04	*TENM2*
5	167727054	rs943656663	0.0006	1	0.19	*TENM2*
5	167726347	rs994238718	0.0006	1	2.41e-04	*TENM2*

Window 4

Chr	Position	RS_ID	MAF	MAC	P-value	Overlapping gene
5	169912125	rs192559789	0.005	8	1.91e-06	*DOCK2, INSYN2B*
5	169912051	rs539192870	0.005	8	0.12	*DOCK2, INSYN2B*
5	169911794	rs564597808	0.0006	1	9.99e-05	*DOCK2, INSYN2B*
5	169912033	rs745592563	0.0006	1	1.35e-04	*DOCK2, INSYN2B*
5	169911527	rs950395230	0.001	2	7.03e-05	*DOCK2, INSYN2B*
5	169911530	rs952961691	0.001	2	8.58e-04	*DOCK2, INSYN2B*

Table E2. Region-based association tests with drop-one variant from each suggestive window.

MAF: Minor Allele Frequency

MAC: Minor Allele Count

African Ancestry			
Variables	***n***	***βeta***	***P-value***
Age[¶]	836	0.01	$<2.00 \times 10^{-16}$ *
Age × rs17696752[‡]	836	0.0002	0.88
Age of asthma onset[†]	715	-0.002	0.10
Age of asthma onset × rs17696752[§]	715	-0.002	0.37
Native American Ancestry			
Age[¶]	836	0.01	$<2.00 \times 10^{-16}$ *
Age × rs12153426[‡]	836	0.003	0.11
Age of asthma onset[†]	715	-0.002	0.07
Age of asthma onset × rs12153426[§]	715	0.008	0.001*

Table E3. The effects of age and age of asthma onset (among the participants with asthma) and interactions with the lead ancestry SNPs on lung function (log-transformed FEV_1). Both age and age of asthma onset were considered as continuous variables.

[¶]Regression models were adjusted for lead ancestry SNP, sex, height, asthma status, and African and Native American ancestries.

[‡]Regression models were adjusted for lead ancestry SNP, age, sex, height, asthma status, and African and Native American ancestries. Multiplicative interaction term (lead ancestry SNP × age) was added to the regression models.

[†]Regression models were adjusted for lead ancestry SNP, age, sex, height, and African and Native American ancestries.

[§]Regression models were adjusted for lead ancestry SNP, age, sex, height, age of asthma onset, and African and Native American ancestries. Multiplicative interaction term (lead ancestry SNP × age of asthma onset) was added to the regression models.

[*]P-value < 0.05