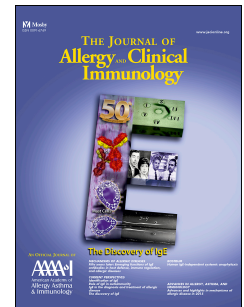


Accepted Manuscript

An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos

Chris R. Gignoux, Ph.D., Dara G. Torgerson, Ph.D., Maria Pino-Yanes, Ph.D., Lawrence H. Uricchio, Ph.D., Joshua Galanter, M.D., M.A.S., Lindsey A. Roth, M.A., Celeste Eng, B.S., Donglei Hu, Ph.D., Elizabeth A. Nguyen, B.S., Scott Huntsman, M.S., Rasika A. Mathias, ScD., Rajesh Kumar, M.D., M.S.P.H., Jose Rodriguez-Santana, M.D., Neeta Thakur, M.D., M.P.H., Sam S. Oh, Ph.D., Meghan McGarry, M.D., Andres Moreno-Estrada, M.D., Ph.D., Karla Sandoval, Ph.D., Cheryl A. Winkler, Ph.D., Max A. Seibold, Ph.D., Badri Padhukasahasram, Ph.D., David V. Conti, Ph.D., Harold J. Farber, M.D., M.S.P.H., Pedro Avila, M.D., Emerita Brigino-Buenaventura, M.D., Michael Lenoir, M.D., Kelley Meade, M.D., Denise Serebrisky, M.D., Luisa N. Borrell, D.D.S., Ph.D., William Rodriguez-Cintron, M.D., Shannon Thyne, M.D., Bonnie R. Joubert, Ph.D., Isabelle Romieu, M.D., Albert M. Levin, Ph.D., Juan-Jose Sienra-Monge, M.D., Blanca Estela del Rio-Navarro, M.D., Weiniu Gan, Ph.D., Benjamin A. Raby, M.D., M.P.H., Scott T. Weiss, M.D., Eugene Bleecker, M.D., Deborah A. Meyers, Ph.D., Fernando J. Martinez, M.D., W. James Gauderman, Ph.D., Frank Gilliland, M.D., Stephanie J. London, M.D., Ph.D., Carlos D. Bustamante, Ph.D., Dan L. Nicolae, Ph.D., Carole Ober, Ph.D., Saunak Sen, Ph.D., Kathleen Barnes, Ph.D., L. Keoki Williams, M.D., M.A.S., Ryan D. Hernandez, Ph.D., Esteban G. Burchard, M.D., M.P.H.



PII: S0091-6749(18)31274-0

DOI: [10.1016/j.jaci.2016.08.057](https://doi.org/10.1016/j.jaci.2016.08.057)

Reference: YMAI 13601

To appear in: *Journal of Allergy and Clinical Immunology*

Received Date: 13 October 2015

Revised Date: 20 August 2016

Accepted Date: 29 August 2016

Please cite this article as: Gignoux CR, Torgerson DG, Pino-Yanes M, Uricchio LH, Galanter J, Roth LA, Eng C, Hu D, Nguyen EA, Huntsman S, Mathias RA, Kumar R, Rodriguez-Santana J, Thakur N, Oh SS, McGarry M, Moreno-Estrada A, Sandoval K, Winkler CA, Seibold MA, Padhukasahasram B, Conti DV, Farber HJ, Avila P, Brigino-Buenaventura E, Lenoir M, Meade K, Serebrisky D, Borrell LN, Rodriguez-

Cintron W, Thyne S, Joubert BR, Romieu I, Levin AM, Sienna-Monge J-J, del Rio-Navarro BE, Gan W, Raby BA, Weiss ST, Bleecker E, Meyers DA, Martinez FJ, Gauderman WJ, Gilliland F, London SJ, Bustamante CD, Nicolae DL, Ober C, Sen S, Barnes K, Williams LK, Hernandez RD, Burchard EG, An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos, *Journal of Allergy and Clinical Immunology* (2018), doi: 10.1016/j.jaci.2016.08.057.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos

Chris R Gignoux, Ph.D.^{1,2*ab}, Dara G Torgerson, Ph.D.^{3*}, Maria Pino-Yanes, Ph.D.^{3,4}, Lawrence H. Uricchio, Ph.D.^{2c}, Joshua Galanter, M.D., M.A.S.,^{2,3}, Lindsey A Roth, M.A.³, Celeste Eng, B.S.³, Donglei Hu, Ph.D.³, Elizabeth A Nguyen, B.S.³, Scott Huntsman, M.S.³, Rasika A Mathias, ScD.⁵, Rajesh Kumar, M.D., M.S.P.H.⁶, Jose Rodriguez-Santana, M.D.⁷, Neeta Thakur, M.D., M.P.H.³, Sam S. Oh, Ph.D.³, Meghan McGarry, M.D.⁸, Andres Moreno-Estrada, M.D., Ph.D.⁹, Karla Sandoval, Ph.D.⁹, Cheryl A. Winkler, Ph.D.¹⁰, Max A Seibold, Ph.D.¹¹, Badri Padhukasahasram, Ph.D.¹², David V Conti, Ph.D.¹³, Harold J. Farber, M.D., M.S.P.H.¹⁵, Pedro Avila, M.D.¹⁶, Emerita Brigino-Buenaventura, M.D.¹⁷, Michael Lenoir, M.D.¹⁸, Kelley Meade, M.D.¹⁴, Denise Serebrisky, M.D.¹⁹, Luisa N. Borrell, D.D.S., Ph.D.²⁰, William Rodriguez-Cintron, M.D.²¹, Shannon Thyne, M.D.³, Bonnie R. Joubert, Ph.D.²², Isabelle Romieu, M.D.²³, Albert M Levin, Ph.D.¹², Juan-Jose Sienra-Monge, M.D.²⁴, Blanca Estela del Rio-Navarro, M.D.²⁴, Weiniu Gan, Ph.D.²⁵, Benjamin A. Raby, M.D., M.P.H.²⁶, Scott T. Weiss, M.D.²⁶, Eugene Bleecker, M.D.²⁷, Deborah A. Meyers, Ph.D.²⁷, Fernando J. Martinez, M.D.²⁸, W. James Gauderman, Ph.D.¹³, Frank Gilliland, M.D.¹³, Stephanie J. London, M.D., Ph.D.²², Carlos D Bustamante, Ph.D.⁹, Dan L. Nicolae, Ph.D.²⁹, Carole Ober, Ph.D.³⁰, Saunak Sen, Ph.D.³¹, Kathleen Barnes, Ph.D.⁵, L. Keoki Williams, M.D., M.A.S.^{12,32}, Ryan D. Hernandez, Ph.D., Esteban G Burchard, M.D., M.P.H.^{1,2,3}

1. Program in Pharmaceutical Sciences and Pharmacogenomics, University of California, San Francisco, San Francisco, CA
2. Department of Bioengineering & Therapeutic Sciences, University of California, San Francisco, San Francisco, CA
3. Department of Medicine, University of California, San Francisco, San Francisco, CA
4. CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain.
5. Department of Medicine, Johns Hopkins University, Baltimore, Maryland
6. The Ann and Robert H. Lurie Children's Hospital of Chicago; Feinberg School of Medicine, Northwestern University, Chicago, IL
7. Centro de Neumologia Pediatrica, San Juan, Puerto Rico
8. Department of Pediatrics, University of California, San Francisco, San Francisco, CA
9. Department of Genetics, Stanford University, Palo Alto, CA
10. Molecular Genetics Epidemiology Section, Frederick National Lab for Cancer Research, Frederick MD
11. Integrated Center for Genes, Environment, and Health; Department of Pediatrics; Division of Pulmonary and Critical Care Medicine, National Jewish Health, Denver, CO
12. Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, MI
13. Department of Preventative Medicine, University of Southern California, Los Angeles CA
14. Children's Hospital and Research Center Oakland, Oakland, CA
15. Department of Pediatrics, Section of Pulmonology, Baylor College of Medicine and Texas Children's Hospital, Houston, TX

16. Division of Allergy-Immunology, Feinberg School of Medicine, Northwestern University, Chicago, IL
17. Department of Allergy & Immunology, Kaiser Permanente-Vallejo Medical Center, Vallejo, CA
18. Bay Area Pediatrics, Oakland, CA
19. Pediatric Pulmonary Division, Jacobi Medical Center, Bronx, NY
20. Department of Health Sciences, Graduate Program in Public Health, Lehman College, City University of New York, Bronx, NY
21. Veterans Caribbean Health Care System, San Juan, Puerto Rico
22. National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC
23. Nutritional Epidemiology Group, International Agency for Research on Cancer, Lyon, France
24. Departamento de Alergia e Inmunologia, Clinica Hospital Infantil de Mexico Federico Gomez, Mexico City, Mexico
25. Division of Lung Diseases, National Heart, Lung, and Blood Institute, Bethesda, MD
26. Department of Medicine, Harvard Medical School, Boston, MA
27. Center for Genomics & Personalized Medicine Research, Wake Forest University, Winston-Salem, NC
28. BIO5 Institute, The University of Arizona, Tucson, AZ
29. Physical Sciences Division, Department of Statistics, University of Chicago, Chicago, IL
30. Department of Human Genetics, University of Chicago, Chicago, IL
31. Department of Preventive Medicine, University of Tennessee Health Sciences Center, Memphis, TN
32. Department of Internal Medicine, Henry Ford Health System, Detroit MI

*These authors contributed equally

^aCorresponding author: cgignoux@stanford.edu

^bCurrent address: Stanford University, Department of Genetics, Stanford Center for Computational, Evolutionary, and Human Genomics, Stanford, CA.

^cCurrent address: Stanford University, Department of Biology, Stanford Center for Computational, Evolutionary, and Human Genomics, Stanford, CA.

Abstract

Background: Asthma is a common but complex disease with racial/ethnic differences in prevalence, morbidity, and response to therapies.

Objective: Perform an analysis of genetic ancestry to identify new loci that contribute to asthma susceptibility.

Methods: We leveraged the mixed ancestry of 3,902 Latinos and performed an admixture mapping meta-analysis for asthma susceptibility. We replicated associations in an independent study of 3,774 Latinos, performed targeted sequencing for fine mapping, and tested for disease correlations with gene expression in the whole blood of >500 individuals from 3 racial/ethnic groups.

Results: We identified a genome-wide significant admixture mapping peak at 18q21 in Latinos ($p=6.8 \times 10^{-6}$), where Native American ancestry was associated with increased risk of asthma (OR=1.20, 95% CI=1.07-1.34, $p=0.002$) and European ancestry with protection (OR=0.86, 95% CI=0.77-0.96, $p=0.008$). Our findings replicated in an independent childhood asthma study in Latinos ($p=5.3 \times 10^{-3}$, combined $p=2.6 \times 10^{-7}$). Fine mapping of 18q21 in 1,978 Latinos identified a significant association with multiple variants 5' of *SMAD2* in Mexicans, whereas a single rare variant in the same window was the top association in Puerto Ricans. Low versus high *SMAD2* blood expression was correlated with case status (13.4% lower expression, OR=3.93, 95% CI 2.12-7.28, $p<0.001$). In addition, lower expression of *SMAD2* was associated with more frequent exacerbations among Puerto Ricans with asthma.

Conclusion: Ancestry at 18q21 was significantly associated with asthma in Latinos, and implicated multiple ancestry-informative non-coding variants upstream of *SMAD2* with asthma susceptibility. Furthermore, decreased *SMAD2* expression in blood was strongly associated with increased asthma risk and increased exacerbations.

Key Messages

- Genetic ancestry at 18q21 is associated with asthma susceptibility in Latinos
- Targeted sequencing of the admixture peak at 18q21 identified multiple non-coding variants associated with asthma. These variants are at different frequencies in the ancestral populations of Latinos.
- Low *SMAD2* expression correlates with increased susceptibility to asthma, and more frequent asthma exacerbations.

Capsule Summary: Genetic ancestry and non-coding variation at 18q21 are associated with asthma in Latinos.

Keywords

Asthma, asthma exacerbations, admixture mapping, meta-analysis, Latinos, *SMAD2*, gene expression, targeted sequencing, rare variation

Abbreviations

bp: base-pair

GWAS: Genome-Wide Association Study

LD: Linkage Disequilibrium

SKAT: Sequence Kernel Association Test

SMAD2: SMAD family member 2

SNP: single nucleotide polymorphism

ZBTB7C: Zinc finger and BTB domain containing 7C

Introduction

Asthma prevalence varies dramatically by race/ethnicity. In the U.S., childhood asthma prevalence is highest among Puerto Ricans (24.8%), intermediate among African Americans (16.3%) and lowest among European Americans (7.8%) and Mexican Americans (7.8%).¹ Substantial evidence supports both environmental and genetic contributions to asthma with heritability estimates ranging to upwards of 75%.² Genome-wide association studies (GWAS) have identified >25 genetic risk factors for asthma.³ Nonetheless, known genetic associations account for only a small proportion of the genetic basis of asthma, and have provided limited insight into racial/ethnic disparities in its prevalence and severity. This may partially be due to the presence of population-specific genetic risk factors, and the limited number of GWAS in non-European populations.⁴⁻⁶ For example, many asthma-associated variants identified in European Americans demonstrate significant effect heterogeneity or have failed to replicate in non-European populations.^{7, 8} Additionally, rare genetic polymorphisms, which are more likely to be population specific,⁹ can contribute to the genetic underpinnings of complex diseases.¹⁰ These unexplored genetic factors may contribute to disparities in asthma susceptibility between racial/ethnic groups.

Latinos are primarily admixed descendants of Native American, European, and sub-Saharan African ancestors,^{11, 12} and African Americans primarily of sub-Saharan African and European ancestors.¹³ This variation in genetic ancestry at the locus-specific level can be leveraged to identify novel loci associated with asthma through admixture mapping^{14, 15, 16} – an approach that is complementary to traditional GWAS. Specifically, admixture mapping compares genetic ancestry between cases and controls, rather than

genotypes at individual single nucleotide polymorphisms (SNPs) as in a GWAS. The variation in ancestry patterns stems from the unique patterns of history encoded in the genes of admixed populations. Admixture mapping identifies disease-associated loci via ancestry linkage disequilibrium (ancestry-LD) with causal variation, rather than SNP-LD, and thus may better capture risk loci where multiple rare variants are more strongly linked to genetic ancestry than to a single common variant. Importantly these longer blocks of ancestry result in a decreased multiple testing burden. Admixture mapping is especially relevant when the disease etiology may be different in different ancestral populations as in asthma. Recently we demonstrated that admixture mapping and ancestry deconvolution can identify associations with asthma in African Americans¹⁷ and Latinos^{18, 19}, total IgE²⁰, bronchodilator response²¹, and with lung function in Mexicans and Mexican Americans²².

Here, we extend the trans-ethnic meta-analysis framework using to study asthma in the EVE Asthma Genetics Consortium^{23, 24} by performing an admixture mapping meta-analysis of 7,606 Latino and 3,106 African American asthma cases and controls. We identified a novel locus where ancestry is associated with asthma at 18q21.1 in Latinos, and implicate multiple non-coding variants 5' of *SMAD2* with asthma susceptibility. Furthermore, we correlate the expression of *SMAD2* with asthma susceptibility and exacerbations, further implicating *SMAD2* as playing an important role in asthma.

Materials and Methods:

Our study approach is outlined in Figure 1.

Study Subjects

Discovery. Self-identified Latino and African American subjects from eight independent studies participating in the EVE Asthma Genetics Consortium were included in our analysis (Table 1, total of 3,902 Latinos and 3,106 African-Americans). EVE is a large, multi-ethnic assembly of asthma studies with existing genome-wide SNP genotypes from nine U.S. institutions. Detailed descriptions of each study have been published previously²³. Briefly, three case-control (GALA I MX, GALA I PR, and CHS) and two trio (CARE, MCCAS) studies were included in the discovery analysis. Additional details can be seen in Table 1. In the current project, we used all genotyped, autosomal SNPs passing prior quality control standards²³ (imputed genotypes were not included in ancestry estimation). Local institutional review boards approved the studies.

Replication. We replicated a genome-wide significant association with ancestry in the Genes-environments & Admixture in Latino Americans (GALA II) Study,¹⁸ a large, multi-center case-control study of Latino children between the ages of 8-21 years with and without asthma (see Supplementary Methods in the Online Repository). Individuals were genotyped on the Affymetrix Axiom[®] Genome-Wide LAT1 Array (World Array 4)²⁵ as described in Galanter et al.²⁶, and see Supplementary Methods in the Online Repository. Local institutional review boards approved the study and all subjects and legal guardians provided written informed assent/consent.

Targeted Sequencing

Sequencing of a 342 Kb region at 18q21.1 was performed on 1,978 Latinos from the replication study (GALA II), including 986 Puerto Ricans (495 cases, 491 controls) and 992 Mexicans (495 cases, 497 controls), through the NHLBI's Resequencing and Genotyping Service (RS&G) (chr18:45321283-45663680, hg19 coordinates). Target enrichment was performed using a custom NimbleGen SeqCap EZ library, for which 87.3% of the target could be uniquely captured. Sequencing was performed to an average of 257x coverage (with 97% of the target having >20x coverage) on a HiSeq using paired-end reads of 150 bp in length. Reads were mapped to the human genome build hg19 and variants called using GATK best practices²⁷.

Gene Expression Measurements

We measured the expression of *SMAD2*, *SMAD3*, *ZBTB7C*, and the housekeeping gene β -glucuronidase (*GUSB*) in peripheral blood using TaqMan[®] RT-PCR assays. Total RNA was isolated from PAXgene[™] Blood RNA tubes, and RNA integrity was assessed with Aligent's BioAnalyzer. Samples with RNA integrity below 6 were excluded from further analysis. *SMAD2* expression was measured in GALA II Puerto Ricans (107 cases/54 controls), Mexicans (122 cases/56 controls), and African Americans from SAGE^{28 29} (114 cases/51 controls, see Supplementary Methods in the Online Repository). Lastly, we downloaded *SMAD2* expression data from isolated CD4⁺ lymphocytes in the Gene Expression Omnibus³⁰ from Mobini et al.³¹ The study

included individuals from Germany with and without asthma, seasonal allergic rhinitis, and eczema, totaling 15 asthma cases and 8 asthma controls.

Statistical Analyses

Ancestry Estimation: We estimated local ancestry at every SNP on our genotyping for each study individually using two LAMP algorithms: LAMP³² for case-control studies and LAMP-HAP³³ for family-based studies. We ran LAMP assuming three ancestral populations consistent with prior studies, and with reference panels designed with each platform in mind. Extensive details on platform choice, rationale, and methods are available in the Online Repository.

Admixture Mapping: Briefly, we treat ancestry estimates as an additive count (0,1,2) of ancestries per locus. We used a 2-degree of freedom likelihood ratio test to evaluate the total local effect of the three primary ancestries in Latinos on asthma susceptibility (see Supplementary Methods in the Online Repository). We performed separate versions of generalized linear models for each study design: logistic regression for case/control (GALA I MX, GALA I PR, CHS) and an extended TDT poisson regression for trio-based studies (MCCAS, CARE). P-values were then meta-analyzed across studies to evaluate the joint association using Fisher's method. We determined a study-specific significance threshold using empirical autoregression to estimate the effective number of tests using the *coda* package in R (see Supplementary Methods in the Online Repository).

To evaluate the effect of a specific ancestry, we coded ancestry at each position as a biallelic state (e.g., African vs non-African), and evaluated the additive case of 0,1 or 2 copies of that ancestry at any locus. Similar to GWAS, we then used logistic regression for case-control studies using R,³⁴ and the transmission disequilibrium test (TDT)³⁵ for trio studies using custom Python³⁶ scripts available upon request. Single ancestry tests were combined using fixed effects models in PLINK³⁷ to evaluate statistical significance, magnitude of effect, and study heterogeneity (I^2). Sites with an I^2 value of heterogeneity > 50% were inspected to determine whether a random effects model was warranted (to incorporate between-study heterogeneity).

Imputation and Fine Mapping: To identify genetic variants driving the association with local ancestry, we performed genotype imputation using a standard pipeline (Online Repository) across a ~5Mb region around *SMAD2*. We analyzed each study using a similar framework as described above (i.e., logistic regression and TDT). A fixed effects meta-analysis was performed in PLINK, and a random effects meta-analysis using the R package *metafor*.³⁸

Targeted Sequencing Analysis: Pooled rare variant association testing was performed using the optimized Sequence Kernel Association Test (SKAT-O, available at <http://www.hsph.harvard.edu/skat/download/>)^{39, 40} as implemented in R⁴¹, and on individual variants using logistic regression. The kernel provides an efficient method for combining evidence across individual variants, where pooling provides additional power compared to single rare variant tests. We performed tests of association with asthma as

a case-control test for variants within coding exons of *SMAD2* and *ZBTB7C* separately, and agnostically across the joint 342 Kb region using a sliding window of 5Kb in length and a step size of 500 bp (total 676 windows) to evaluate all possible loci within the region of ancestry association. We also performed a similar scan of overall expression-quantitative trait loci (eQTL) adjusting for asthma as a covariate. A total of 10,000 permutations were run at $\alpha=0.05$ to determine statistical significance, and from this we could determine a false discovery rate in R to adjust for multiple comparisons (method="BH"). For individual variants at 18q21.1, we performed additional imputation using IMPUTE2 for the remaining 1,778 Latinos from the GALA II study who were not directly sequenced using filters described above. Reference haplotypes included individuals from the GALA II study that were sequenced (n=1,978), and individuals from the 1000 Genomes Project (phase 1). All tests of association were adjusted for local and global ancestry, age, sex, and recruitment area.

Gene Expression Analysis: To provide additional lines of evidence for our associations, we performed gene expression at the genes of interest using qPCR on whole blood for a subset of GALA II and SAGE II individuals with available RNA. We rescaled expression to fold-change levels ($2^{-\Delta CT}$). We maximized goodness-of-fit of a cutpoint of gene expression using Bayes Factors calculated from generalized linear models, adjusting for age, sex, recruitment center and ancestry. These were used downstream for regression testing with observed exacerbations and as outcome for single variant and pooled (SKAT-O) association tests. Additional details of exacerbation

calculations, specific models used, estimation of cell composition, and other details are available in the Online Repository.

Results

Admixture Mapping Meta-Analysis and Replication

We performed local ancestry estimation, ancestry interpolation, and admixture mapping in five different studies of childhood asthma from EVE, comprising 3,902 Latino individuals. The multiple testing threshold, using autocorrelation, was determined to be $p < 4 \times 10^{-5}$, a value similar to prior analyses leveraging permutations (e.g. ²⁶). We identified a genome-wide significant admixture mapping peak in Latinos at 18q21 using the likelihood ratio test ($p = 6.8 \times 10^{-6}$, Figure 2A and B). This peak was driven by multiple ancestry signals at the locus: Native American was found to increase risk (OR=1.20, 95% CI=1.07-1.34, $p=0.002$), whereas European ancestry was found to be protective (OR=0.86, 95% CI=0.77-0.96, $p=0.008$, Figure 2C); African ancestry did not appear to play a significant role ($p=0.4$) (Figure 2C), but appears protective in the trio-based studies. We note that Native and European ancestries are negatively correlated and therefore these effects are not independent. The admixture peak is bounded by two recombination hotspots and overlaps two protein-coding genes: *SMAD2* and *ZBTB7C* (Figure 2B).

We replicated the association between asthma and ancestry at 18q21 in an independent sample of 3,774 Latinos from the GALA II Study (min $p = 5.3 \times 10^{-3}$, Table 2). The direction of the effect in the discovery and replication populations was

homogeneous for both Native American and European ancestry (OR=1.09, $p=6.3 \times 10^{-3}$; and OR=0.87, $p=5.8 \times 10^{-3}$ respectively; Table 2). In African Americans we found no significant association with ancestry at 18q21 (see Figure S1 in the Online Repository, $p=0.7$). In European Americans we are unable to perform admixture mapping, however no SNP within the 18p21 region was previously identified as being significantly associated with asthma in the EVE meta-analysis.²³

In addition to the likelihood ratio test, we performed a meta-analysis of single ancestry admixture mapping across Latino individuals and identified two genome-wide significant peaks for European ancestry at 9q22 and 12p12 ($p < 4 \times 10^{-5}$, see Table S1 and Figure S2 in the Online Repository). Both of these peaks failed to replicate in GALA II (lowest p-value = 0.17 and 0.21, respectively). We also note that while genome-wide admixture mapping peaks remain rare, we were able to provide the replication to a peak found on 6p21 for asthma²⁶. A comprehensive set of suggestive peaks, the relevant underlying ancestry test, and the genes within each peak are available in Table S1 in the Online Repository.

Fine Mapping of Genetic Associations

We imputed SNP genotypes from the 1000 Genomes Project (phase I) within a 5Mb region centered on the admixture peak at 18q21 in all of the discovery populations, however we found no significant associations at SNPs within 18q21 (see Figure S3 in the Online Repository).

In our replication study, we performed targeted sequencing of 352 Kb of the admixture peak to augment genotype imputation, and to better enable tests of

association with pooled variants in our populations of interest. We observed a significant association with pooled variants in Mexicans within a 5Kb window upstream of *SMAD2* at chr18:45495471-45500349 (hg19) ($p=2.0 \times 10^{-4}$, FDR=0.045, Figure 3). When we stratified based on local ancestry, we found the association at pooled variants to be strongest on haplotypes of non-Native American ancestry and individuals that were homozygous for African ancestry in Mexicans (see Figures S4 and S5 in the Online Repository). Consistent with these findings the top SKAT-O window harbored 7 individual variants at frequency $<1\%$ and with $p < 0.05$ were protective and African in origin (Table 3). Similarly, an additional variant with $p < 0.05$ within the significant SKAT-O window was at a frequency of 6% in Mexicans, was more common on Native haplotypes, and was a risk factor for asthma (rs59002988, Table 3). While the function of these variants is unknown, four of these variants are within ENCODE transcription factor binding sites and/or DNase peaks⁴² (Figure 3, see Table S3 in the Online Repository).

We found no significant association with pooled variants in Puerto Ricans only, however the top associated individual variants were either directly within the window implicated in Mexicans (four variants, $\text{min}p=4.6 \times 10^{-4}$ at rs59002988), or nearby (three variants, $\text{min}p=3.8 \times 10^{-4}$ at rs7238092 [137 bp from window start], Figure 3).

Furthermore, rs59002988 was locus-wide significantly associated with asthma in all of GALA II (OR=1.67, $p=1 \times 10^{-5}$), and contributes to the SKAT-O signal overlapping that same window.

Within chr18:45495283-45500283, we identified 56 variants in Mexicans and 61 variants in Puerto Ricans through direct sequencing, of which 30 were shared between

populations. A stratified analysis partitioning shared vs. private variants (i.e. variants only found in one population) identified a marginal association in both classes of variation in Mexicans (see Figure S6 in the Online Repository), suggesting that both private and shared variants are contributing to the association. No significant associations were observed at individual variants in Mexicans, and no significant associations were observed with pooled variants in the coding exons of *SMAD2* or *ZBTB7C* in either Mexicans or Puerto Ricans (see Table S3 in the Online Repository).

eQTL analysis and Gene Expression Associations

We measured expression of *SMAD2*, *SMAD3* (the cystolic hetero-dimeric partner of *SMAD2*), and *ZBTB7C* via rt-PCR from whole blood RNA in a subset of 161 Puerto Ricans from the replication study (GALA II). We found no significant association between genotypes at individual or pooled variants and gene expression in whole blood of 177 Mexicans (*SMAD2*) and 144 Puerto Ricans (*SMAD2* and *ZBTB7C*) from the GALA II study that were sequenced in this study (see Figure S7 in the Online Repository). However, asthma cases had significantly lower mean levels of *SMAD2* gene expression ($p < 0.001$). In contrast neither *SMAD3* nor *ZBTB7C* demonstrated any difference in expression between cases and controls ($p = 0.8$ and 0.9 , respectively).

We then investigated *SMAD2* gene expression in whole blood in all 503 African and Mexican Americans alongside the Puerto Ricans previously measured. We determined the best-fit cut-point at 127% of mean expression in controls across all populations when partitioning by low-vs.-high *SMAD2* expression (see Figure S8 in the Online Repository). Each population has a different cut-point, here the trans-ethnic

value is provided to maximize the lower bounds on the odds ratio across the entire sample. *SMAD2* expression below this cut-point was associated with a 4-fold increase in the odds of asthma (OR=3.93, 95% CI=2.12-7.28, $p<0.001$), with no ethnic heterogeneity (Cochran's Q $p=0.34$). In addition, the CD4+ lymphocyte dataset replicated the association between asthma and low *SMAD2* expression in individuals from Germany (OR=1.9, 95% CI=1.3-2.8, $p=0.02$, see Table S4 and Figure S9 in the Online Repository). While the sample size of this additional source of replication is small, we observe an overlapping effect size with our larger set of samples from SAGE II and GALA II.

In the subset of individuals with estimated blood cell counts, the proportion of cell types was not significantly associated with *SMAD2* expression (min marginal p : 0.35 for the proportion of CD4+ T cells), nor was the joint effect of all types (likelihood ratio test $p=0.93$). Thus, *SMAD2* expression was not associated with nor confounded by peripheral blood cell proportions, consistent with previous findings⁴³. The association between asthma status and *SMAD2* expression was strongest in those with an age of asthma onset <5 years (see Figure S10 in the Online Repository), consistent with prior discoveries of a stronger genetic effect in early onset asthma⁴⁴.

Lower *SMAD2* gene expression was associated with higher exacerbation scores across the 342 cases (OR per 10% decrease in expression: 1.11, 95% CI=1.02-1.20, $p=0.01$; see Table 4 and Figure S11 in the Online Repository for the distribution of exacerbation scores). There was no significant heterogeneity (Cochran's Q $p=0.08$), and the p -value remained significant ($p=0.04$) even when modeling the overall association via random effects meta-analysis in METASOFT⁴⁵. As with the case-control

association, incorporating peripheral blood cell counts did not attenuate the significance of the findings. Oral steroids could affect gene expression, however only 5 of the participants report taking oral steroids: including a variable for steroid usage was not in itself significant nor did it affect the association between gene expression and asthma status. While we found a highly significant association with exacerbations, we did not observe associations with other measures of severity, including asthma control or spirometric values.

We then built logistic regression models to test the ability of *SMAD2* gene expression to explain asthma exacerbations in Latinos and African Americans. We incorporated age, population, medication use, and spirometry in the full models (see Table S5 and Figure S12 in the Online Repository). Individuals were dichotomized into low and high exacerbation scores. We applied three prediction models: use of controller medication (any long-term asthma medication), response to albuterol, and *SMAD2* gene expression. Incorporating all three predictors (along with age and ethnicity) minimized the AIC, demonstrating the additional power of *SMAD2* expression to predict high-end healthcare utilizers beyond standard clinical measurements. Increasing the cut-point of high-vs.-low exacerbation resulted in improved specificity and prediction accuracy (Figure 4B).

Discussion

Our novel investigation of admixture mapping and asthma among 3,902 Latinos from the EVE Asthma Genetics Consortium identified a genome-wide significant association between asthma and ancestry at 18q21, centered on the *SMAD2-ZBTB7C* locus. We replicated this finding among 3,774 individuals in the GALA II study, and identified a significant association with multiple, non-coding rare variants 5' upstream of *SMAD2* in Mexicans. While we stratified by population for the sake of analysis, the strong fixed effects analysis is consistent with a Hispanic/Latino-wide association rather than one that is geographically restricted. We do note that the original forest plots may look somewhat heterogeneous, the results at our top hit are overall consistent and reflect the need for large sample sizes to uncover accurate effects. In addition, we demonstrated that low *SMAD2* expression is associated with increased risk of asthma, and more frequent asthma exacerbations in whole blood. Genetic variation at 18q21 was not identified as a risk factor for asthma in a prior meta-analysis utilizing traditional GWAS methods in the identical studies.^{23, 46} This demonstrates the power of admixture mapping to detect genetic associations that are different from those identified using traditional GWAS, and for identifying loci where rare genetic variation contributes to asthma susceptibility.

An important and unique contribution offered by admixture mapping is its potential for increased coverage of untyped variants due to ancestry linkage disequilibrium (ancestry-LD) as compared with genotype-LD.⁴⁷ Our admixture mapping approach identified Native American ancestry at 18q21 as a risk factor for asthma in Latinos. Indeed, within the 5Kb window showing a significant association with multiple non-coding variants in Mexicans, variants showing a marginal association on their own

includes one that is risky for asthma and more common on Native American haplotypes, and seven that are protective for asthma and more common on African haplotypes. These observations are consistent with the admixture signal, and may explain why no association was identified with individual variants in the original EVE meta-analysis, as the effect appears to be driven predominantly by rare and low frequency variation. Admixture mapping with followup sequencing therefore provides the capability to discover additional genetic variation important for disease susceptibility using existing GWAS data.

Our admixture mapping peak at 18q21 was fairly broad. However, fine mapping of this peak implicated non-coding variants directly between *SMAD2* and *ZBTB7C*. *ZBTB7C* has limited functional characterization and no known role in asthma pathophysiology. Furthermore, we found no significant difference in the expression of *ZBTB7C* in the whole blood of asthma cases vs. controls. In contrast, *SMAD2* was found to exhibit decreased expression in asthma cases vs. controls, and is a well-characterized cofactor involved in the *TGF- β* signaling pathway often studied in asthma and other inflammatory diseases. The *TGF- β* pathway has been implicated multiple components of asthma biology: negative regulation of allergic airway inflammation,⁴⁸ airway remodeling,⁴⁹ and drug response.⁵⁰ Ligation of *TGF- β* receptors activates the proximal transcription factors *SMAD2* and *SMAD3*; these translocate to the nucleus to regulate transcription of several hundred target genes along with a complex of DNA binding cofactors. Lower levels of *SMAD2* would then be expected to correlate with lower levels of *TGF- β* -mediated signaling effect,⁵¹ consistent with our observation that lower *SMAD2* expression is associated with asthma. We note that effects are not

consistent in overall inflammation levels and locally in the airway, rather, these levels would broadly represent inflammatory response. Among cases, we would expect that limited *TGF- β* signaling in the blood would result in more airway inflammation, thereby leading to increased asthma exacerbations.

Although *TGF- β* pathway genes are known to play a functional role in asthma, they have rarely been identified via GWAS, and to our knowledge variation within or nearby *SMAD2* has not been previously associated with asthma. Interestingly, two previous meta-analyses^{23, 46} identified an association between *SMAD3*, the cytosolic hetero-dimeric partner of *SMAD2*, and asthma in Europeans and European Americans. Here, variation in *SMAD3* was not significantly associated with asthma through admixture mapping in either Latinos or African Americans, nor did we find that *SMAD3* was differentially expressed in the whole blood of asthma cases vs. controls in Puerto Ricans.

Fine mapping implicated multiple non-coding variants 5' of *SMAD2* in asthma susceptibility, including several variants in the proximity of regulatory elements. However, none of the variants were correlated with the expression of either *SMAD2* or *ZBTB7C* in peripheral blood, and thus we find no evidence these variants affect the transcription of either of these genes in *cis*-. This may indicate the variants contribute to asthma susceptibility through a mechanism independent of *SMAD2* or *ZBTB7C* expression, a lack of statistical power to identify eQTLs in the subset of individuals examined, or that the variants identified regulate the expression of *SMAD2* and/or *ZBTB7C* in a tissue other than peripheral blood.

Despite our inability to identify *cis*-eQTLs, our findings support the role of differential regulation of *SMAD2* gene expression in the blood in asthma cases, with a four-fold increase in odds of asthma for low-vs.-high *SMAD2* expression. The disconnect between our genetic and gene expression associations at 18q21 may thus reflect a more complex, multi-layered contribution of *SMAD2* on asthma susceptibility. The effects of *SMAD2* expression are seen jointly when pooled across three ethnic groups, with the largest effect observed in Puerto Ricans and those with early asthma onset (5 years or younger). The potential for using *SMAD2* expression as a biomarker for asthma and asthma severity in peripheral blood is attractive due to its ease of collection as compared to other tissues. Including peripheral blood *SMAD2* expression levels improved the explanatory power of statistical models of asthma exacerbations beyond the use of traditional variables collected in the clinic. By estimating white blood cell counts for a subset of our data, we ruled out potential confounding by cell types, another attractive feature of a biomarker measured in peripheral blood. Furthermore, we found that *SMAD2* expression was also significantly lower in atopic asthma cases vs. controls in isolated CD4+ lymphocytes in German individuals. This provides additional support that the difference we observe is not due to differences in white blood cell counts, and that *SMAD2* expression may additionally be a useful biomarker in individuals of European ancestry.

Our results uncover a novel asthma locus in Hispanic/Latinos that replicates, is the heterodimeric partner of another well-replicated asthma locus and provides additional insight via gene expression in whole blood. However this study is not without shortcomings: first, we deployed our admixture mapping framework in a highly diverse

collection of studies, with unique ancestries and environmental exposures. We cannot adequately measure the full scope of environmental exposures that may affect expression patterns. In addition, as our gene expression measurements came from studies of childhood asthma, we only have access to one relevant tissue type reflecting the inflammatory nature of asthma (whole blood). While we did not identify a consensus set of expression-QTL loci, the intrinsic heterogeneity of exposures across individuals can conceal associations. We note that the large Genotype-Tissue Expression Consortium⁵² (GTEx) was also unable to identify eQTLs for *SMAD2* in whole blood. The small subset of individuals may not reflect the entire spectrum of expression patterns. However our fine mapping does result in a suggestion of transcription factor binding sites and/or DNase peaks underlying the signal suggesting a regulatory component reflecting the noncoding nature of our top hits. The novel methods, large-scale computational approach, numerous independent study designs and platforms, fine mapping and biological relevance to disease provide a new window into the pathophysiology of asthma in Hispanic/Latinos.

In conclusion, an admixture mapping meta-analysis across 3,902 Latinos identified a novel genome-wide significant association with ancestry at chr18q21, centered on *SMAD2-ZBTB7C*, which replicated in an additional 3,774 Latinos. We further implicated multiple non-coding variants 5' of *SMAD2* and asthma susceptibility, which vary in frequency between the ancestral haplotypes of Latinos. Low peripheral blood *SMAD2* expression was associated with a four-fold increase in odds of asthma. In individuals with asthma, low *SMAD2* expression in peripheral blood is associated with increased asthma exacerbations in Latinos.

538

539 **Acknowledgments**

540 The authors acknowledge the families and patients for their participation, and thank the
541 numerous health care providers and community clinics for their support and
542 participation in all the EVE Consortium and replication studies. In particular, the authors
543 thank GALA I, GALA II and SAGE study coordinator Sandra Salazar; the recruiters who
544 obtained the data: Duanny Alva, MD, Gaby Ayala-Rodriguez, Lisa Caine, Elizabeth
545 Castellanos, Jaime Colon, Denise DeJesus, Blanca Lopez, Brenda Lopez, MD, Louis
546 Martos, Vivian Medina, Juana Olivo, Mario Peralta, Esther Pomares, MD, Jihan
547 Quraishi, Johanna Rodriguez, Shahdad Saeedi, Dean Soto, Emmanuel Viera, Ana
548 Taveras. Resequencing services were kindly provided through the RS&G Service by the
549 Northwest Genomics Center at the University of Washington, Department of Genome
550 Sciences, under U.S. Federal Government contract number HHSN268201100037C
551 from the National Heart, Lung, and Blood Institute. Some computations were performed
552 using the UCSF Biostatistics High Performance Computing System. The authors would
553 like to thank Dean Sheppard, MD, David Erle, MD, and Amy J. Markowitz for helpful
554 edits, comments and advice on early versions of manuscript.

555

556 **Declaration of funding**

557 This research was supported in part by National Institutes of Health (AI061774,
558 AI077439, AI077439, AI079139, CA113710, DK064695, ES015794, ES015794,
559 HL078885, HL079055, HL087699, HL088133, HL088133, HL104608, M01-RR00188,
560 MD006902); ARRA grant RC2 HL101651; Flight Attendant Medical Research Institute

(FAMRI); UCSF Chancellor's Research Fellowship, Dissertation Year Fellowship, and in part by NIH Training Grant T32GM007175 and T32HG000044 (to CRG); RWJF Amos Medical Faculty Development Award (to EGB); the Sandler Foundation; the American Asthma Foundation (to EGB). MPY was funded by a Postdoctoral Fellowship from Fundación Ramón Areces; KCB was supported in part by the Mary Beryl Patch Turnbull Scholar Program; RAM was supported in part by the MOSAIC initiative of Johns Hopkins University; Awards from American Asthma Foundation and the Fund for Henry Ford Hospital (to LKW); Ernest S. Bazley Grant (to PCA); NHLBI K23 to RK (K23HL093023); GCRC M01 RR00188 to HJF; NHLBI K23 (K23HL111636) and NCATS KL2 (KL2TR000143) to JMG. This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number KL2TR000143. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN26120080001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

ReferencesSources

1. National Health Interview Survey (NHIS) Data. 2011.] Available from <http://www.cdc.gov/asthma/nhis/2011/table4-1.htm>.
2. Willemssen G, van Beijsterveldt TCEM, van Baal CGCM, Postma D, Boomsma DI. Heritability of self-reported asthma and allergy: A study in adult Dutch twins, siblings and parents. *Twin Research and Human Genetics* 2008; 11:132-42.
3. A Catalog of Published Genome-Wide Association Studies, Accessed 03/2013.; 2013.] Available from <http://www.genome.gov/gwastudies>.
4. Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. *Nature* 2011; 475:163-5.
5. Ober C, Hoffjan S. Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun* 2006; 7:95-100.
6. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. *Trends Genet* 2009; 25:489-94.
7. Galanter JM, Torgerson D, Gignoux CR, Sen S, Roth LA, Via M, et al. Cosmopolitan and ethnic-specific replication of genetic risk factors for asthma in 2 Latino populations. *The Journal of allergy and clinical immunology* 2011; 128:37-43.e12.
8. Wu H, Romieu I, Shi M, Hancock DB, Li H, Sienra-Monge JJ, et al. Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *J Allergy Clin Immunol* 2010; 125:321-7 e13.
9. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* 2011; 108:11983-8.
10. Cohen JC, Boerwinkle E, Mosley TH, Jr., Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* 2006; 354:1264-72.
11. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* 2010; 107 Suppl 2:8954-61.
12. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, et al. Population stratification confounds genetic association studies among Latinos. *Human Genetics* 2006; 118:652-64.
13. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America* 2010; 107:786-91.

14. Choudhry S, Taub M, Mei R, Rodriguez-Santana J, Rodriguez-Cintron W, Shriver MD, et al. Genome-wide screen for asthma in Puerto Ricans: evidence for association with 5q23 region. *Human Genetics* 2008; 123:455-68.
15. Cheng C-Y, Kao WHL, Patterson N, Tandon A, Haiman CA, Harris TB, et al. Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. *PLoS genetics* 2009; 5:e1000490.
16. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences of the United States of America* 2006; 103:14068-73.
17. Torgerson DG, Capurso D, Ampleford EJ, Li X, Moore WC, Gignoux CR, et al. Genome-wide ancestry association testing identifies a common European variant on 6q14.1 as a risk factor for asthma in African American subjects. *J Allergy Clin Immunol* 2012; 130:622-9 e9.
18. Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, et al. GWAS and admixture mapping identify asthma-associated loci in Latinos: The GALA II Study. *J Allergy Clin Immunol* 2013; accepted.
19. Torgerson DG, Gignoux CR, Galanter JM, Drake KA, Roth LA, Eng C, et al. Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *The Journal of allergy and clinical immunology* 2012.
20. Pino-Yanes M, Gignoux CR, Galanter JM, Levin AM, Campbell CD, Eng C, et al. Genome-wide association study and admixture mapping reveal new loci associated with total IgE levels in Latinos. *J Allergy Clin Immunol* 2014.
21. Drake KA, Torgerson DG, Gignoux CR, Galanter JM, Roth LA, Huntsman S, et al. A genome-wide association study of bronchodilator response in Latinos implicates rare variants. *J Allergy Clin Immunol* 2014; 133:370-8.
22. Moreno-Estrada A, Gignoux CR, Fernandez-Lopez JC, Zakharia F, Sikora M, Contreras AV, et al. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 2014; 344:1280-5.
23. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nature Genetics* 2011; 43:887-92.
24. Myers RA, Himes BE, Gignoux CR, Yang JJ, Gauderman WJ, Rebordosa C, et al. Further replication studies of the EVE Consortium meta-analysis identifies 2 asthma risk loci in European Americans. *J Allergy Clin Immunol* 2012; 130:1294-301.
25. Hoffmann TJ, Zhan Y, Kvale MN, Hesselton SE, Gollub J, Iribarren C, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* 2011; 98:422-30.

- 661 26. Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, et al. Genome-wide
662 association study and admixture mapping identify different asthma-associated loci
663 in Latinos: the Genes-environments & Admixture in Latino Americans study. *J*
664 *Allergy Clin Immunol* 2014; 134:295-305.
- 665 27. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A,
666 et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit
667 best practices pipeline. *Curr Protoc Bioinformatics* 2013; 11:11 0 1- 0 33.
- 668 28. Nishimura KK, Galanter JM, Roth LA, Oh SS, Thakur N, Nguyen EA, et al. Early-life air
669 pollution and asthma risk in minority children. The GALA II and SAGE II studies. *Am*
670 *J Respir Crit Care Med* 2013; 188:309-18.
- 671 29. Thakur N, Oh SS, Nguyen EA, Martin M, Roth LA, Galanter J, et al. Socioeconomic
672 status and childhood asthma in urban minority youths. The GALA II and SAGE II
673 studies. *Am J Respir Crit Care Med* 2013; 188:1202-9.
- 674 30. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression
675 and hybridization array data repository. *Nucleic Acids Res* 2002; 30:207-10.
- 676 31. Mobini R, Andersson BA, Erjefalt J, Hahn-Zoric M, Langston MA, Perkins AD, et al. A
677 module-based analytical strategy to identify novel disease-associated genes shows
678 an inhibitory role for interleukin 7 Receptor in allergic inflammation. *BMC Syst Biol*
679 2009; 3:19.
- 680 32. Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific
681 ancestry in closely related populations. *Bioinformatics* 2009; 25:i213-21.
- 682 33. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and
683 accurate inference of local ancestry in Latino populations. *Bioinformatics* 2012;
684 28:1359-67.
- 685 34. R Core Team. R: A language and environment for statistical computing. R
686 Foundation for Statistical Computing, Vienna, Austria, 2012.
- 687 35. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium:
688 the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum*
689 *Genet* 1993; 52:506-16.
- 690 36. Python Software Foundation. <http://www.python.org/>. 2012.
- 691 37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a
692 tool set for whole-genome association and population-based linkage analyses. *Am J*
693 *Hum Genet* 2007; 81:559-75.
- 694 38. Schwartz AG, Wenzlaff AS, Bock CH, Ruterbusch JJ, Chen W, Cote ML, et al.
695 Admixture mapping of lung cancer in 1812 African-Americans. *Carcinogenesis*
696 2011; 32:312-7.
- 697 39. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal
698 unified approach for rare-variant association testing with application to small-
699 sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;
700 91:224-37.

40. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012; 13:762-75.
41. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2013.
42. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489:57-74.
43. Palmer C, Diehn M, Alizadeh AA, Brown PO. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* 2006; 7:115.
44. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007; 448:470-3.
45. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American journal of human genetics* 2011; 88:586-98.
46. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010; 363:1211-21.
47. Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, et al. PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum Biol* 2012; 84:343-64.
48. Hansen G, McIntire JJ, Yeung VP, Berry G, Thorbecke GJ, Chen L, et al. CD4(+) T helper cells engineered to produce latent TGF-beta1 reverse allergen-induced airway hyperreactivity and inflammation. *J Clin Invest* 2000; 105:61-70.
49. Sagara H, Okada T, Okumura K, Ogawa H, Ra C, Fukuda T, et al. Activation of TGF-beta/Smad2 signaling is associated with airway remodeling in asthma. *J Allergy Clin Immunol* 2002; 110:249-54.
50. Burchard EG, Avila PC, Nazario S, Casal J, Torres A, Rodriguez-Santana JR, et al. Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am J Respir Crit Care Med* 2004; 169:386-92.
51. Hough C, Radu M, Dore JJ. Tgf-beta induced Erk phosphorylation of smad linker region regulates smad signaling. *PLoS ONE* 2012; 7:e42513.
52. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; 348:648-60.
53. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res* 2002; 12:996-1006.

Tables

Table 1. Baseline characteristics of studies. Latinos from the EVE Asthma Genetics Consortium were used for discovery, and GALA II was used for replication. GALA II and SAGE II were used for gene expression associations. Characteristics include origin of study and study type. Other descriptions from the EVE studies can be found in detail in Torgerson et al.²³

- a. CARE: Childhood Asthma Research Network (University of Arizona). CHS: Children's Health Study (University of Southern California). GALA I: Genetics of Asthma in Latino Americans (University of California, San Francisco). MCCAS: Mexico City Childhood Asthma Study (National Institute of Environmental Health Sciences). GALA II: Genes-Environments and Admixture in Latino Americans (University of California, San Francisco). SAGE II: The Study of African Americans, Asthma, Genes and Environment (University of California, San Francisco).
- b. Values given as % of predicted based on age, sex, height, and ethnicity. For studies with bronchodilator response, reported values are pre-albuterol administration, for other studies, the spirometric values are baseline from the time of recruitment. Spirometry values are only reported for studies in EVE in which the majority of individuals had reliable tests.

Table 2. Ancestry associations at the 18q21 peak centered on *SMAD2*. Summary characteristics of admixture mapping findings at the top hit in the chr18q21 region. Meta-analysis of discovery and replication studies was performed assuming fixed effects of effect size estimates, and Fisher's method was used for the overall likelihood ratio test meta-analysis.

Table 3. Individual allelic association testing with $p < 0.05$ in Mexicans (and Puerto Ricans) within the top SKAT-O window. The ancestral frequency of A1 for Native ancestry was inferred from GALA II individuals who were homozygous for Native American ancestry at 18q21, for African and European ancestry from YRI and CEU from the 1000 Genomes Project, respectively.

Table 4. Associations of *SMAD2* gene expression with asthma and morbidity outcomes. Dichotomous outcomes use a best-fit cut-point of 127% of healthy control *SMAD2* gene expression to estimate the odds ratio. Associations with exacerbations are per 10% increase in *SMAD2* expression compared to controls.

- a. adjusting for age, gender, recruitment site and ancestry
- b. p-value given from Fisher exact test, but odds ratio and confidence interval given from logistic regression ($p=0.01$) to be more robust. Both tests used the same data.
- c. Levels 0-9, ordered logistic regression, p-value from likelihood ratio test
- d. $p < 0.05$ for random effects meta-analysis as well.

Figure Legends

Figure 1. Outline of the study approach. Data generation steps are in green, analysis steps are in blue. Numbers of individuals available for analyses can be found in Table 1.

Figure 2. Results of the admixture mapping meta-analysis. **A)** Admixture mapping Manhattan plot. **B)** Summary of admixture mapping findings across the 18q21 locus. **C)** Forest plots for each ancestry including odds ratios and confidence intervals with the size of the square inversely proportional to the standard error. Meta-analysis estimates via fixed effects models are given as diamonds, with values in Table 2.

Figure 3. Fine mapping 18q21: pooled (SKAT-O) and individual (GWAS) variants in **A)** Mexicans and **B)** Puerto Ricans. **C)** Regulatory elements from ENCODE⁴², with darker boxes having stronger signals, and numbers indicating hypersensitive cell lines. Also transcription factor bindings site via ChIP-seq in ENCODE⁴². Green lines indicate the highest scoring binding motif for the corresponding factor and most associated cell type, from the UCSC Genome Browser⁵³.

Figure 4. Analysis of *SMAD2* expression in whole blood. **A)** Mean/95% CI of whole blood *SMAD2* expression, lower in cases ($p=0.002$). *ZBTB7C* nor *SMAD3* showed any differential expression. **B)** ROC curves of high exacerbation scores in individuals with asthma using clinical variables, demographic data, and *SMAD2* gene expression accounting for age, population, spirometry, and controller medication (CM) use. AUCs/10-fold cross-validated prediction errors are in legend.

Study Name ^a	Genotyping platform	Study type	cases/controls/trios	Analytical Method	Recruitment Location	% male	Age at recruitment	Age of onset	FEV ₁ ^b	FEV ₁ ^c
<i>EVE Hispanic/Latino Studies</i>										
CARE	Affymetrix 6.0	Trios	42	TDT/Poisson GLM	USA-wide	78.8	4.8±3.2	1.4±1.6	90.0±5.5	101.0±5.5
CHS	Illumina 550K, 610K	Case/Control	606/792	Logistic GLM	Southern California	55.9	7.4±1.7	6.8±4.8	101.0±12.6	100.5±12.6
GALA I Mexicans	Affymetrix 6.0	Case/Control	252/151	Logistic GLM	Mexico City/ Bay Area	44.3	15.9±8.3	8.3±7.7	89.7±18.8	98.0±18.8
GALA I Puerto Ricans	Affymetrix 6.0	Case/Control	277/191	Logistic GLM	Puerto Rico/ New York	42.1	13.8±7.0	3.4±4.8	85.6±15.9	98.6±15.9
MCCAS	Illumina 550K	Trios	492	TDT/Poisson GLM	Mexico City	58.7	9.0±2.4	3.3±1.4	90.5±16.8	93.1±16.8
<i>EVE African American Studies</i>										
Barbados	Illumina 650K	Pedigrees	382/461	Pedigree-aware case-control	Barbados	48.2	20.8±12.9	8.2±10.6	NA	NA
GRAAD	Illumina 650K	Case/Control	606/792	Logistic GLM	Maryland	44.4	24.4±17.6	11.9±13.2	NA	NA
CAG/CSGA/SARP	Illumina 1M	Case/Control	541/451	Logistic GLM	USA-wide	38.9	26.4±15.8	9.8±12.4	64.5±20.5	98.0±20.5
SAPPHIRE	Affymetrix 6.0	Case/Control	150/131	Logistic GLM	Detroit, MI	21.4	32.3±13.3	10.5±11.6	NA	NA
<i>GALA II Hispanic/Latino Replication</i>										
GALA II Mexicans	Affymetrix Axiom World Array 4	Case/Control	586/659	Logistic GLM/SKAT	USA-wide	48.4	12.8±3.3	5.3±3.7	96.9±13.3	101.0±13.3
GALA II Puerto Ricans	Affymetrix Axiom World Array 4	Case/Control	896/890	Logistic GLM/SKAT	USA-wide	50.7	12.4±3.2	2.6±2.9	86.0±16.4	89.7±16.4
GALA II Mixed/Other	Affymetrix Axiom World Array 4	Case/Control	402/324	Logistic GLM/SKAT	USA-wide	48.2	12.5±3.2	4.4±3.9	92.4±15.0	97.7±15.0
<i>Study Populations with Gene Expression</i>										
SAGE II African Americans	TaqMan	Case/Control	114/51	Logistic GLM	Bay Area	47.3	13.4±3.7	3.8±3.9	83.3±10.4	87.8±10.4

GALA II Mexicans	TaqMan	Case/ Control	122/56	Logistic GLM	USA-wide	51.7	12.5±3.4	5.7±4.3	95.9±13.5	100.5
GALA II Puerto Ricans	TaqMan	Case/ Control	107/54	Logistic GLM	USA-wide	58.1	12.6±2.9	1.9±2.4	89.2±13.6	92.8

18q21	EVE	GALA II (average)	Combined
African <i>p</i>	0.42	0.16	0.27
African OR	0.91 (0.73-1.14)	1.05 (0.98-1.13)	1.04 (0.97-1.11)
European <i>p</i>	8.35x10 ⁻³	5.83x10 ⁻³	1.36x10 ⁻⁴
European OR	0.86 (0.77-0.96)	0.87 (0.78-0.96)	0.86 (0.80-0.93)
Native American <i>p</i>	1.63x10 ⁻³	6.26x10 ⁻³	9.15x10 ⁻⁵
Native American OR	1.20 (1.07-1.34)	1.09 (1.02-1.16)	1.11 (1.05-1.17)
Overall <i>p</i>	6.80x10 ⁻⁶	0.017 (min=5.3x10 ⁻³)	2.6x10 ⁻⁷

				Mexicans:				Puerto Ricans:				Ancestral Allele Frequency (A1):		
SNP	Position	A1	A2	Freq A1	OR	SE	P-value	Freq A1	OR	SE	P-value	Native American	African (YRI)	European (CEU)
rs76402589	45497831	A	T	0.00429	0.076	1.08	0.0164	0.0255	0.71	0.223	0.125	0	0.15	0
rs113382450	45499287	G	T	0.00429	0.076	1.08	0.0164	0.0255	0.71	0.223	0.125	0	0.15	0
rs79405582	45499892	T	G	0.00429	0.076	1.08	0.0164	0.0255	0.71	0.223	0.125	0	0.15	0
rs112042427	45496169	A	G	0.00429	0.076	1.08	0.0164	0.0257	0.72	0.222	0.147	0	0.15	0
rs75478387	45496266	T	C	0.00429	0.076	1.08	0.0164	0.0257	0.72	0.222	0.147	0	0.15	0
rs8099232	45496784	T	G	0.00429	0.076	1.08	0.0164	0.0257	0.72	0.222	0.147	0	0.15	0
rs7236490	45498260	G	T	0.00468	0.082	1.07	0.0194	0.0255	0.71	0.223	0.125	0	0.15	0
rs59002988	45496882	T	G	0.0616	1.4	0.172	0.0488	0.0282	2.2	0.227	4.6×10^{-4}	0.102	0.07	0.01

Case-Control Phenotypes: <127% Mean Control Expression^a	N (case vs control)	OR (95% CI)	p-value
All Three Populations	343 vs 161	3.93 (2.12-7.28)	<0.001
All Populations, Heterogeneity	343 vs 161	-	0.34
Subset, adjusting for blood cell counts	74 vs 32	22.4 (1.7-294)	0.018
CD4+ Lymphocytes ^b	15 vs 8	1.9 (1.3-2.8)	0.02

Exacerbation Score: per 10% decrease in expression compared to controls^c	N	OR (95% CI)	p-value
All Three Populations	333	1.06 (1.01-1.12)	0.011^d
All Populations, Heterogeneity	333	-	0.08
Subset, adjusting for blood cell counts	74	1.16 (1.01-1.34)	0.036

Discovery: 5 studies of asthma in Latinos from the EVE Asthma Genetics Consortium (3,902 in total)

Reference data for local ancestry estimation: Sub-Saharan African, European & Native American genotypes

Local ancestry estimation, imputation, admixture mapping

Meta-analysis of the 5 EVE studies

Replication: Genes, Environments and Ancestry in Latino Americans (GALA II) study (3,774 individuals)

GALA II (Latino Americans) and SAGE II (African Americans) with RNA extracted (total: 342 with asthma, 161 controls)

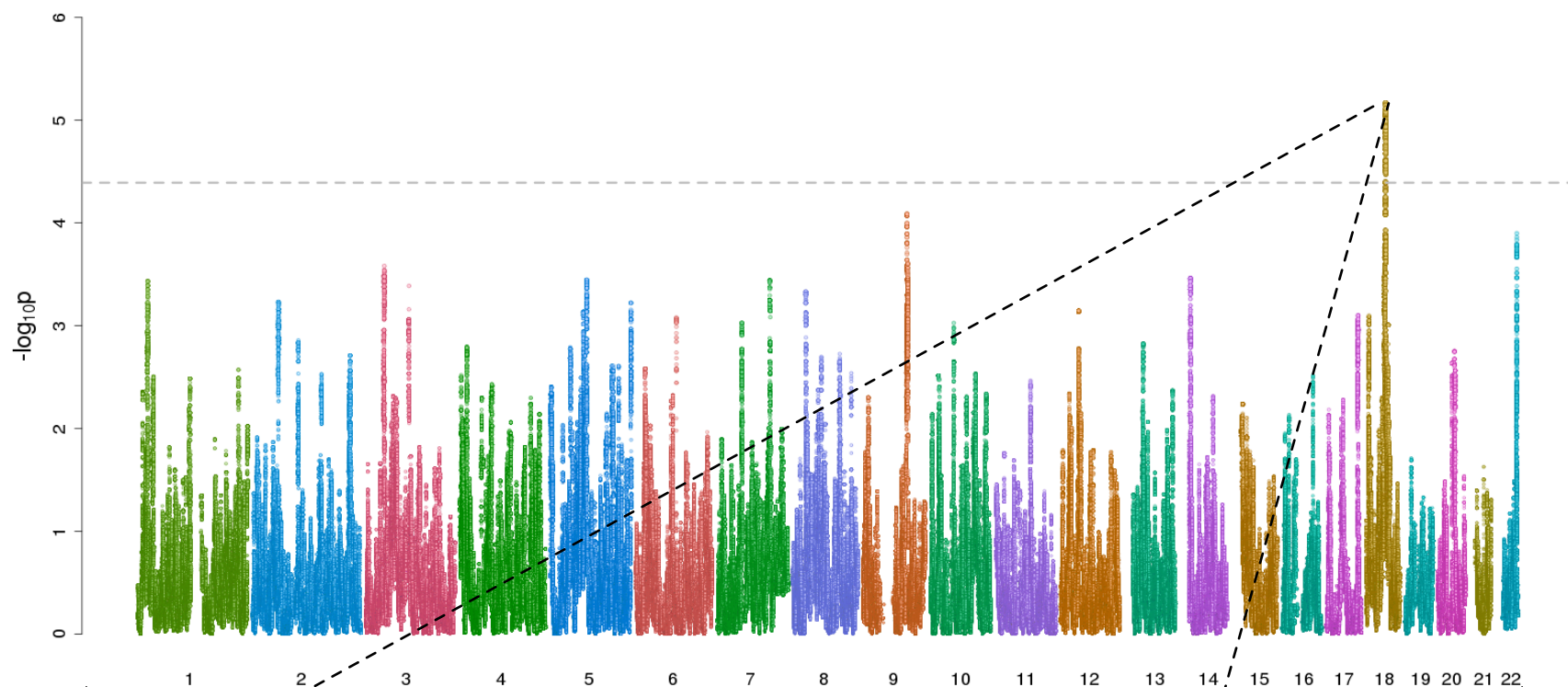
Fine of 18q21 via resequencing in 1,978 individuals from the GALA II study

Association of gene expression with asthma

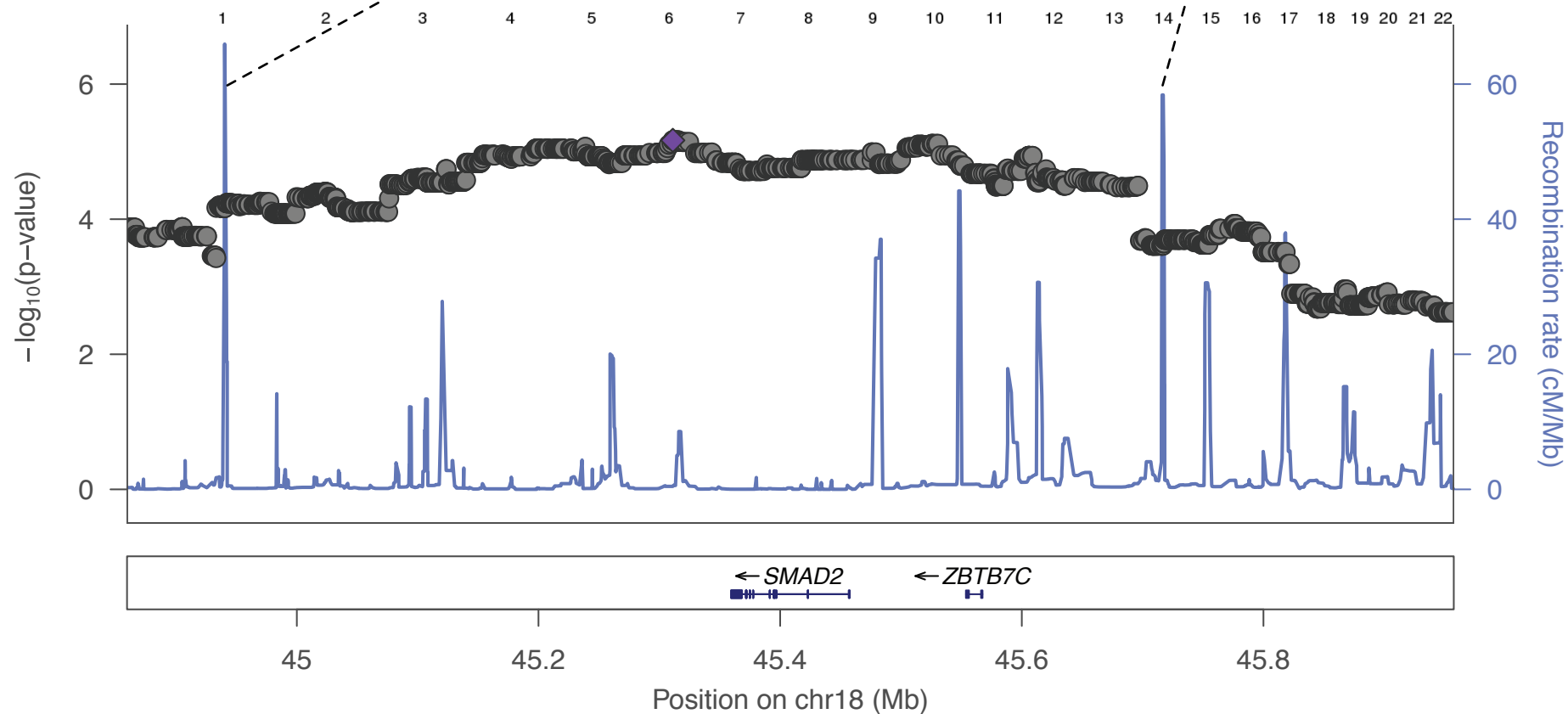
Pooled variant (SKAT) and single variant testing across 18q21

Association of gene expression with secondary phenotypes: Exacerbations, ER visits, Oral Steroid Use

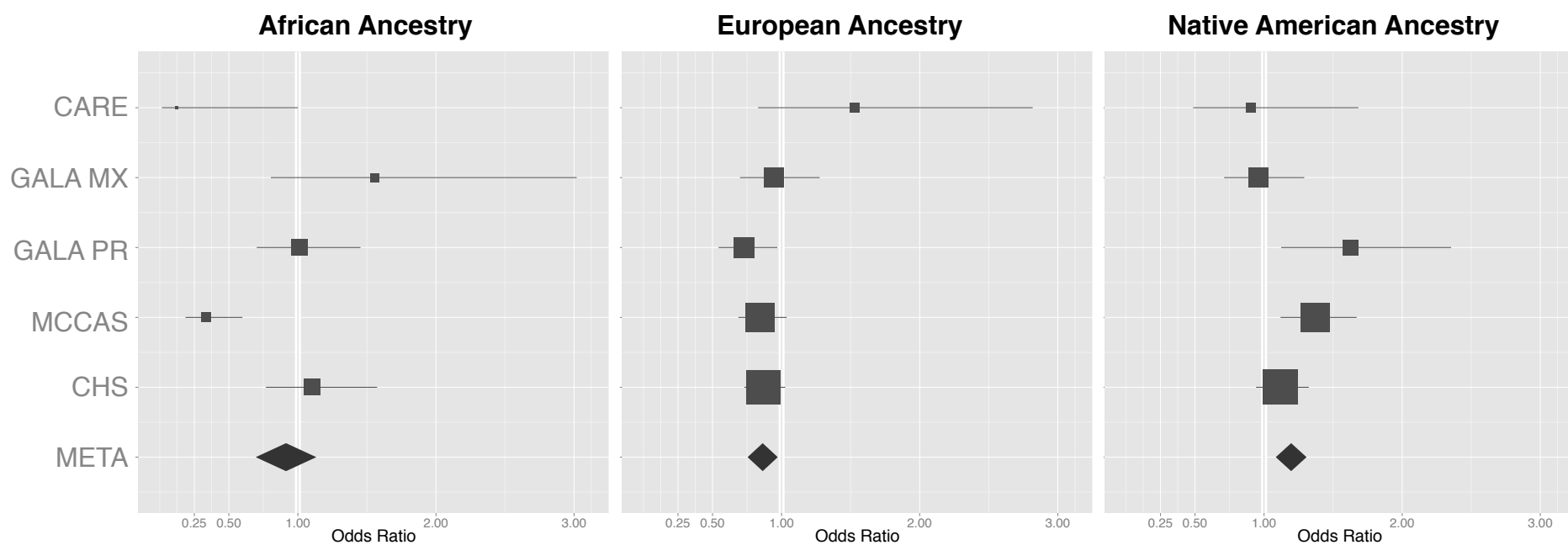
A



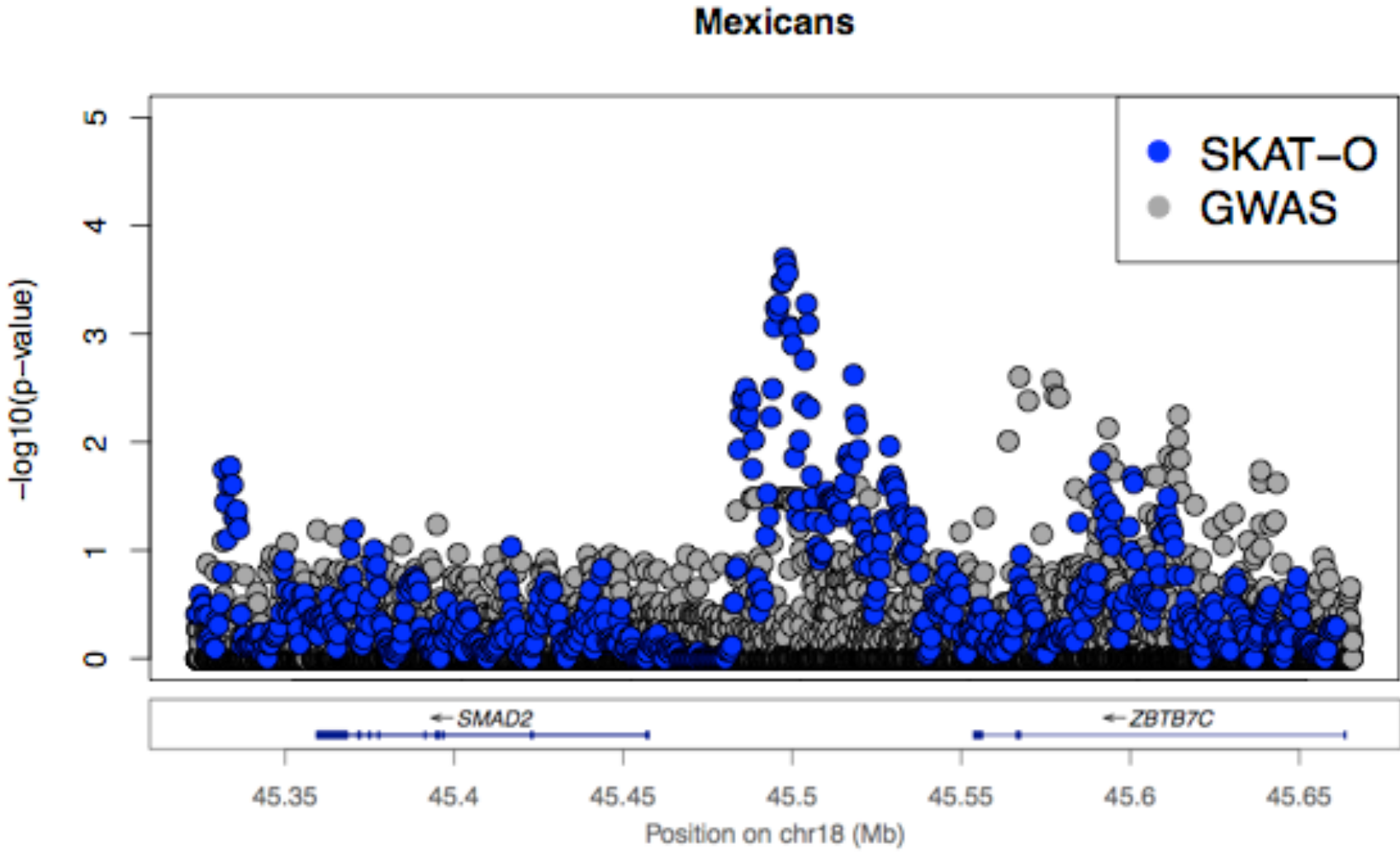
B



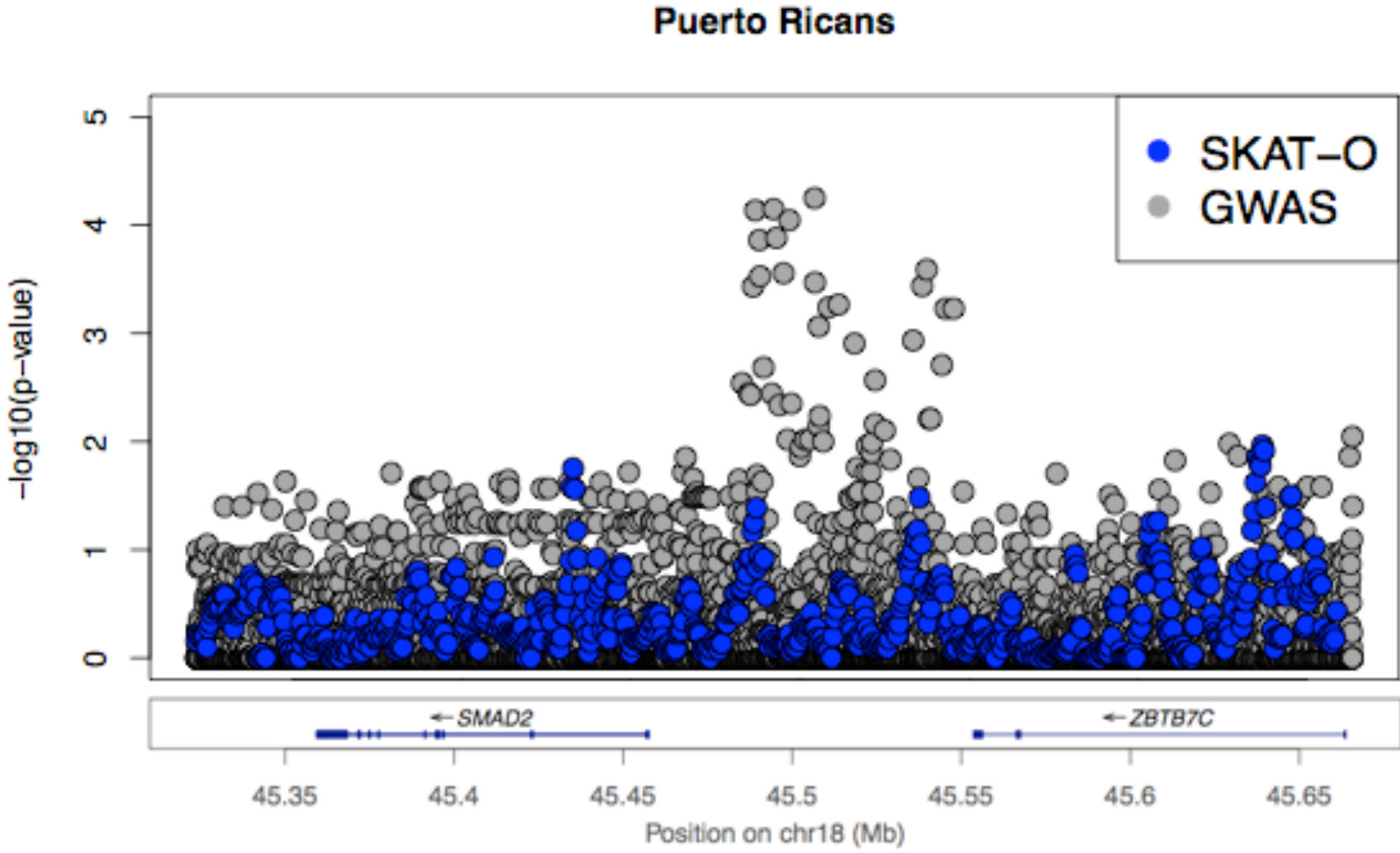
C



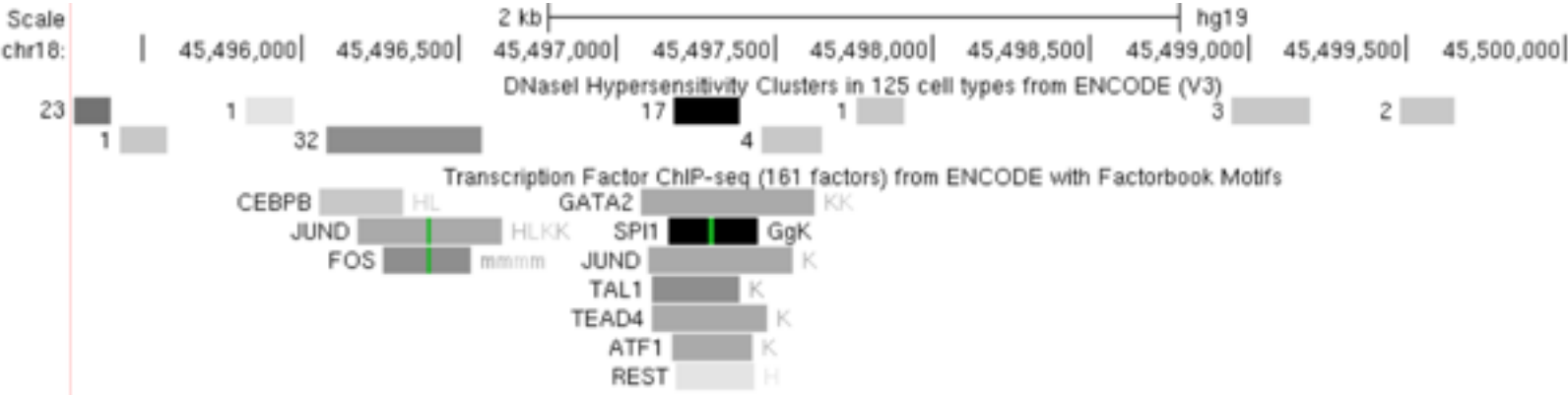
A.

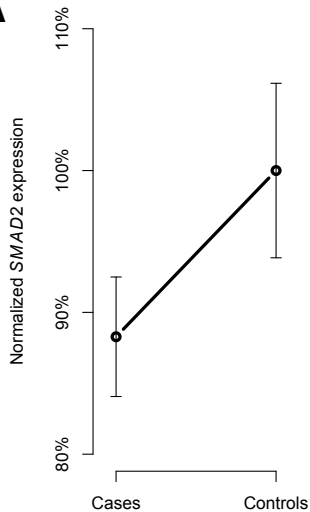
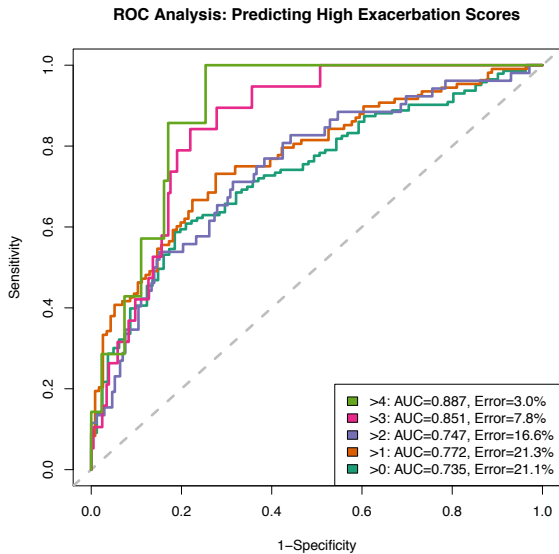


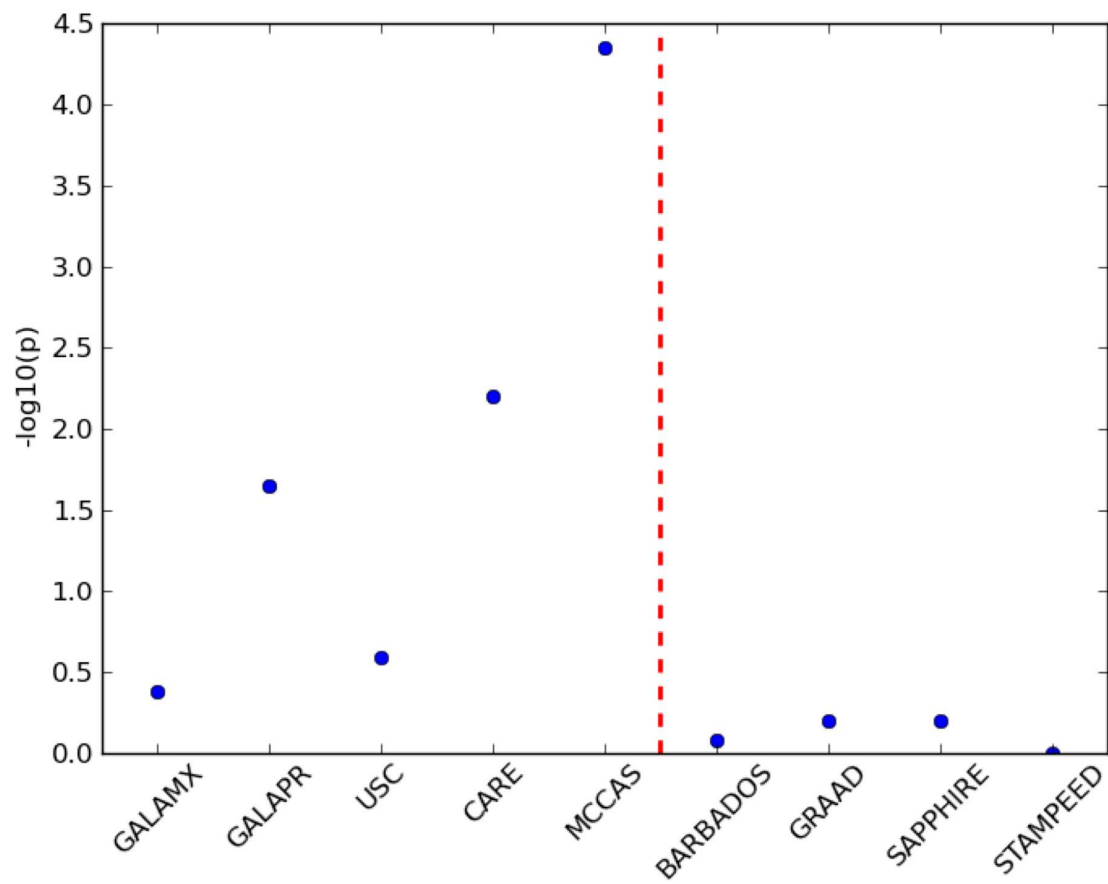
B.

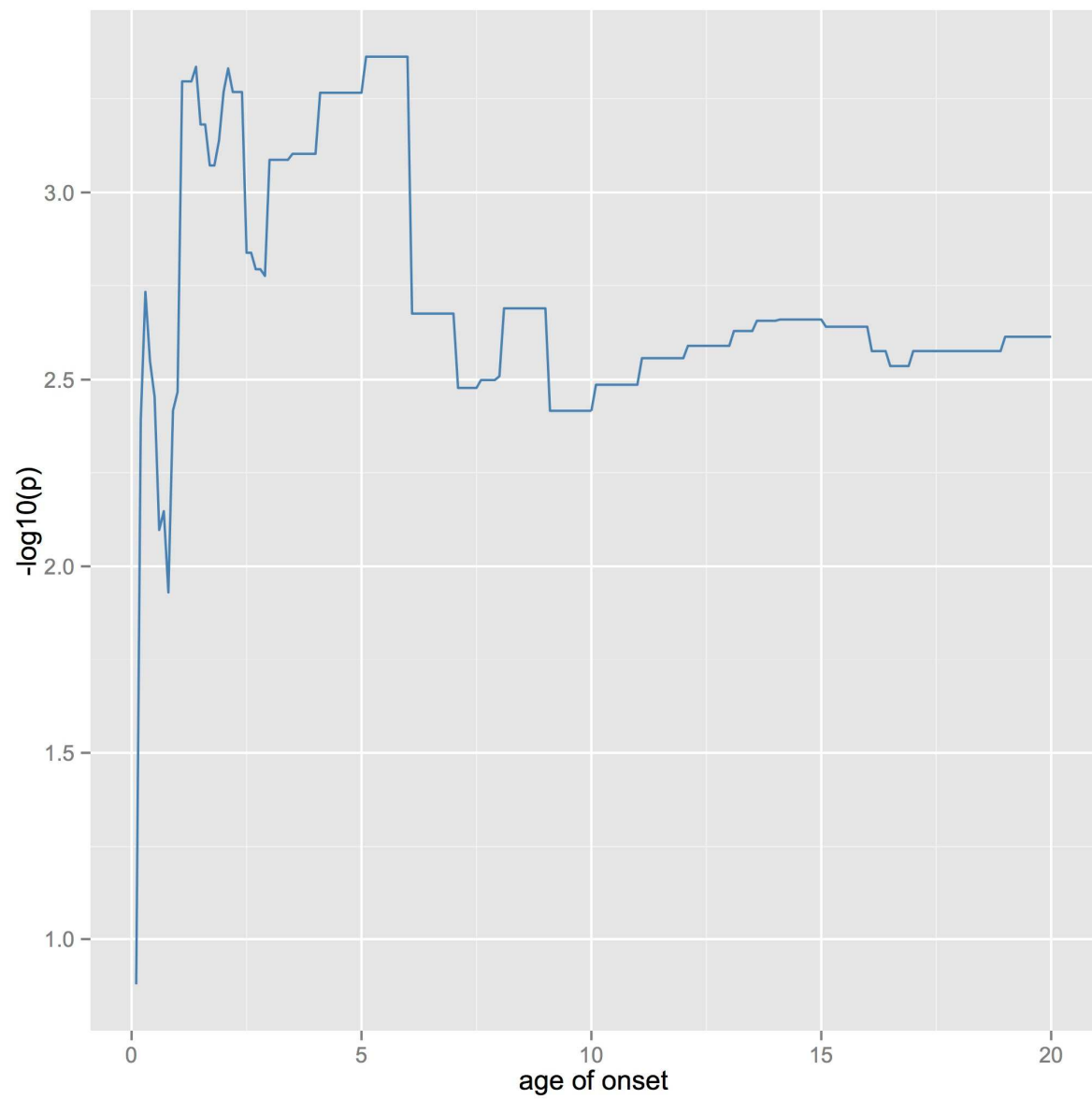


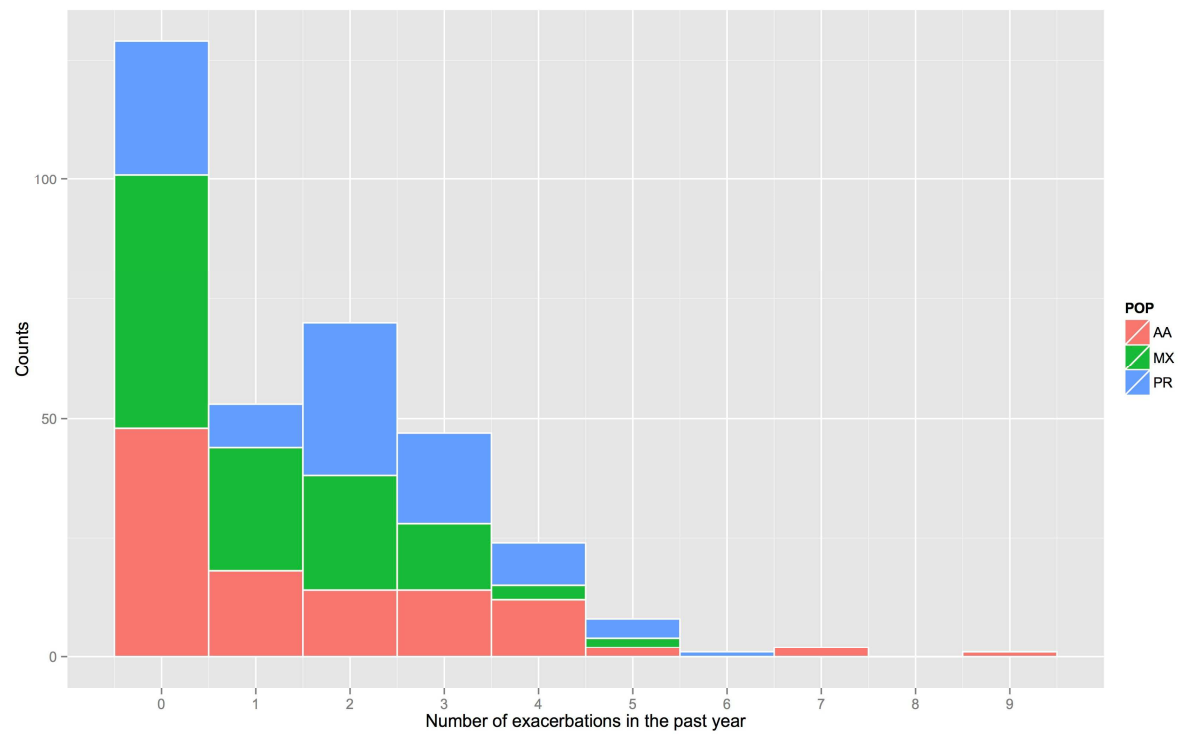
C.

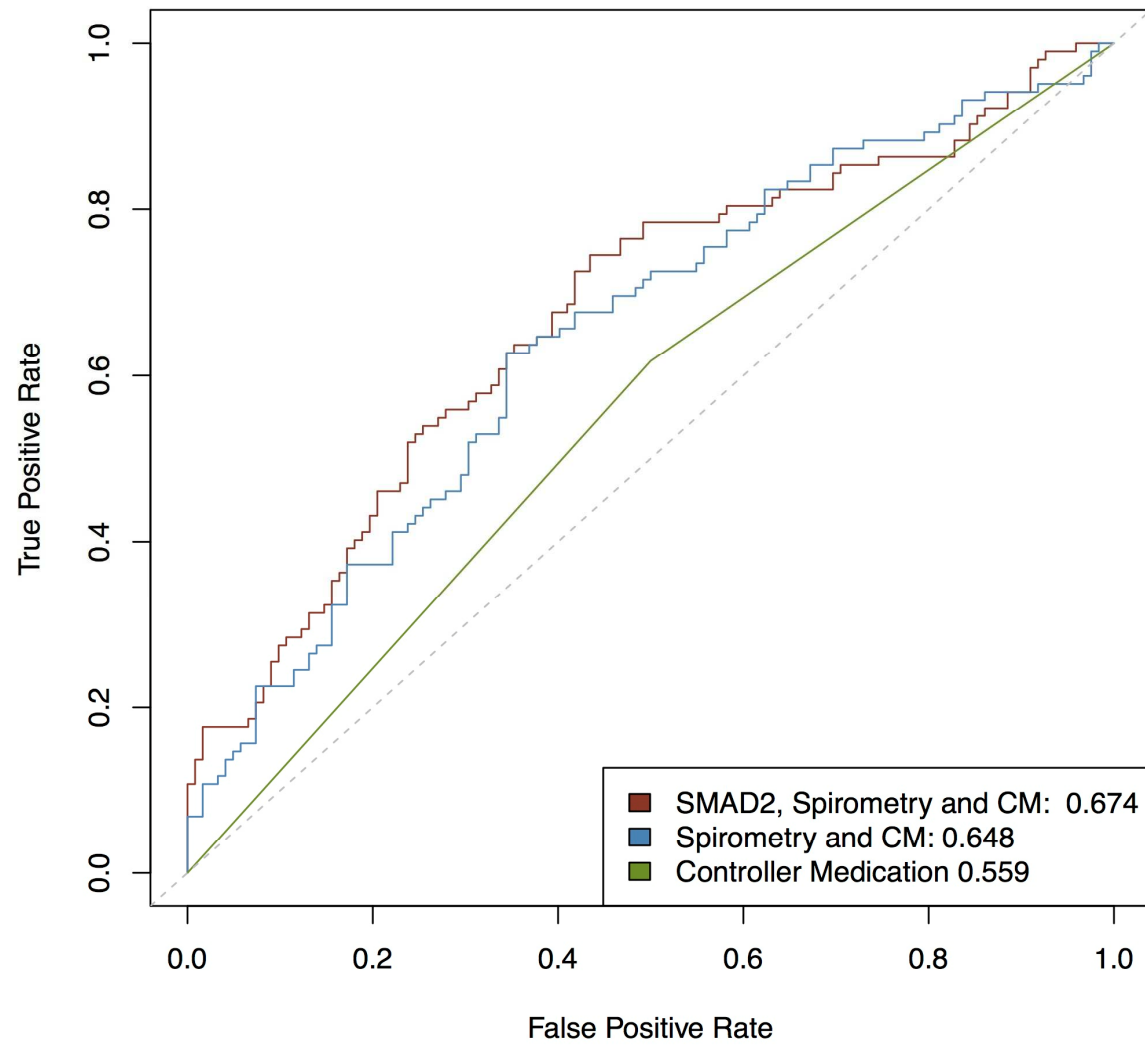


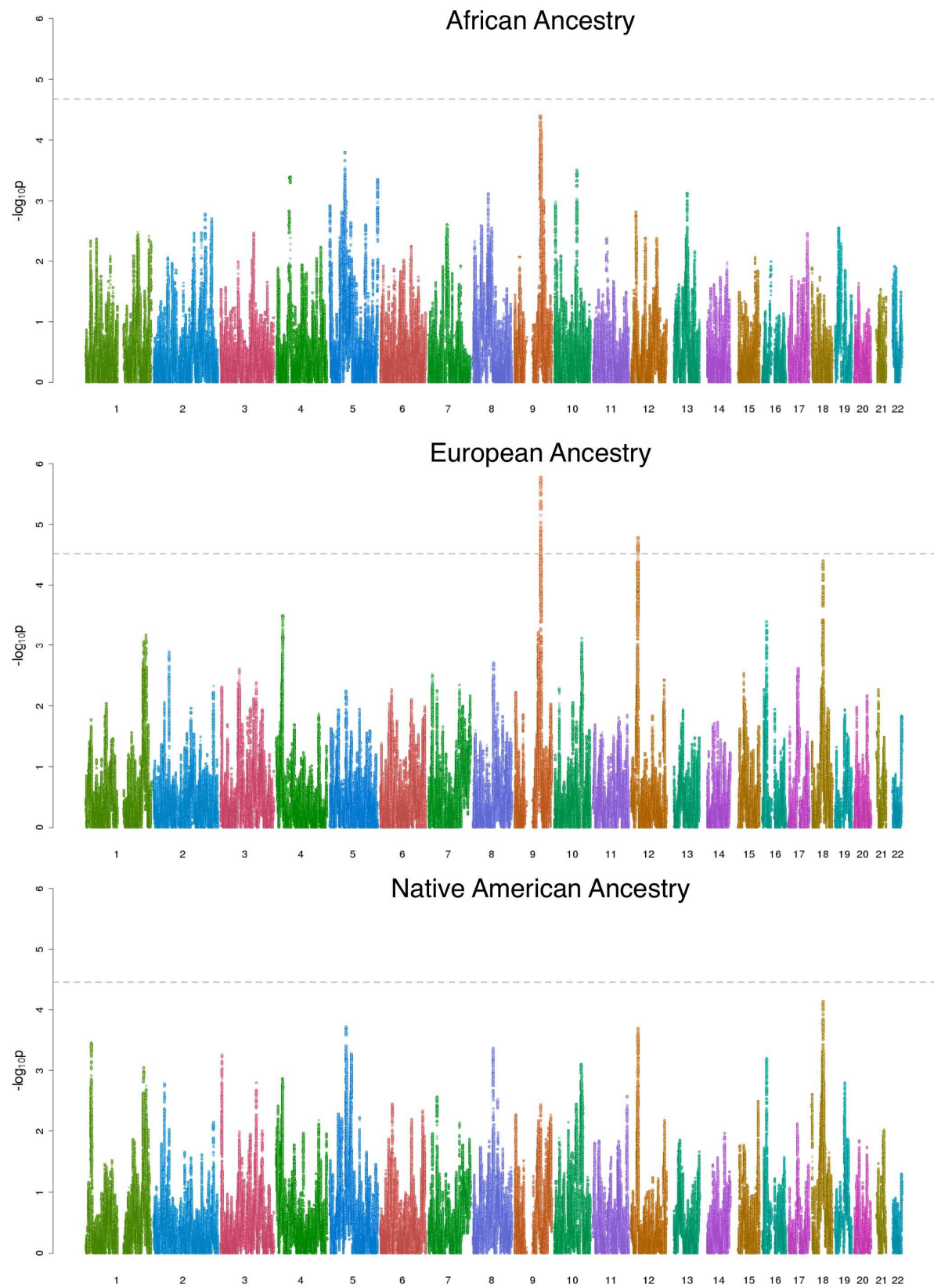
A**B**

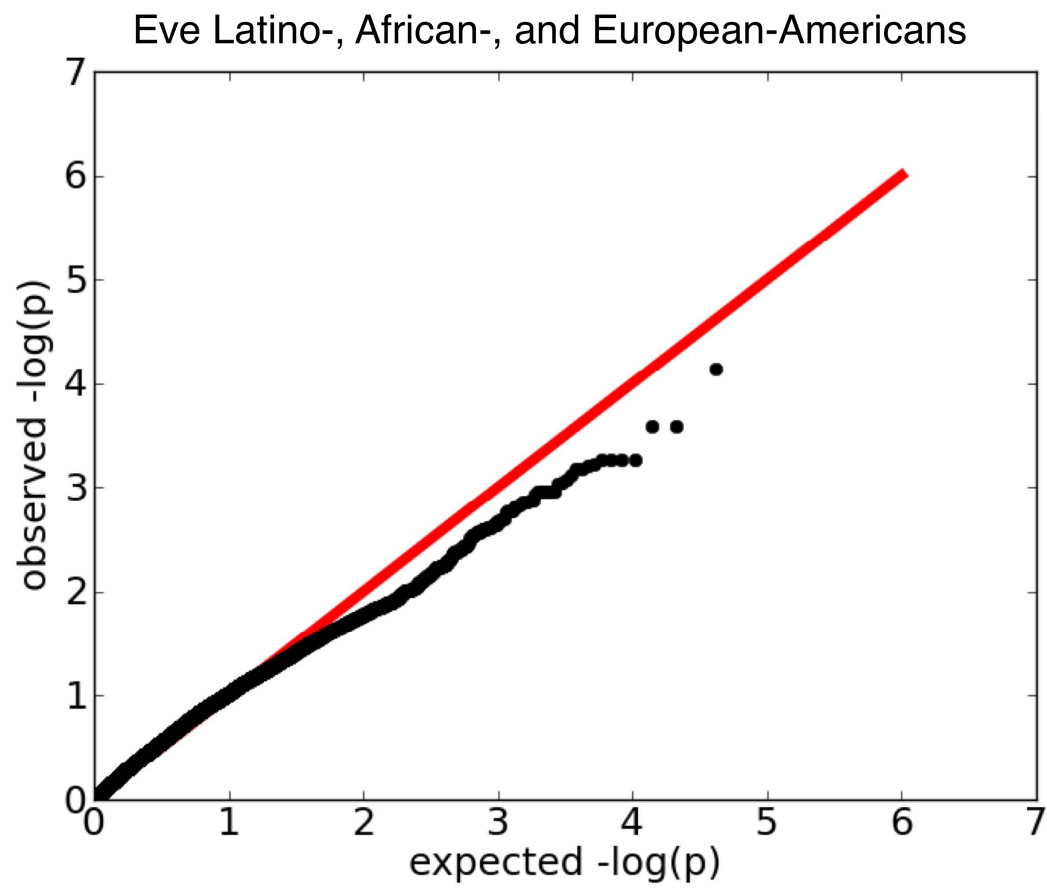


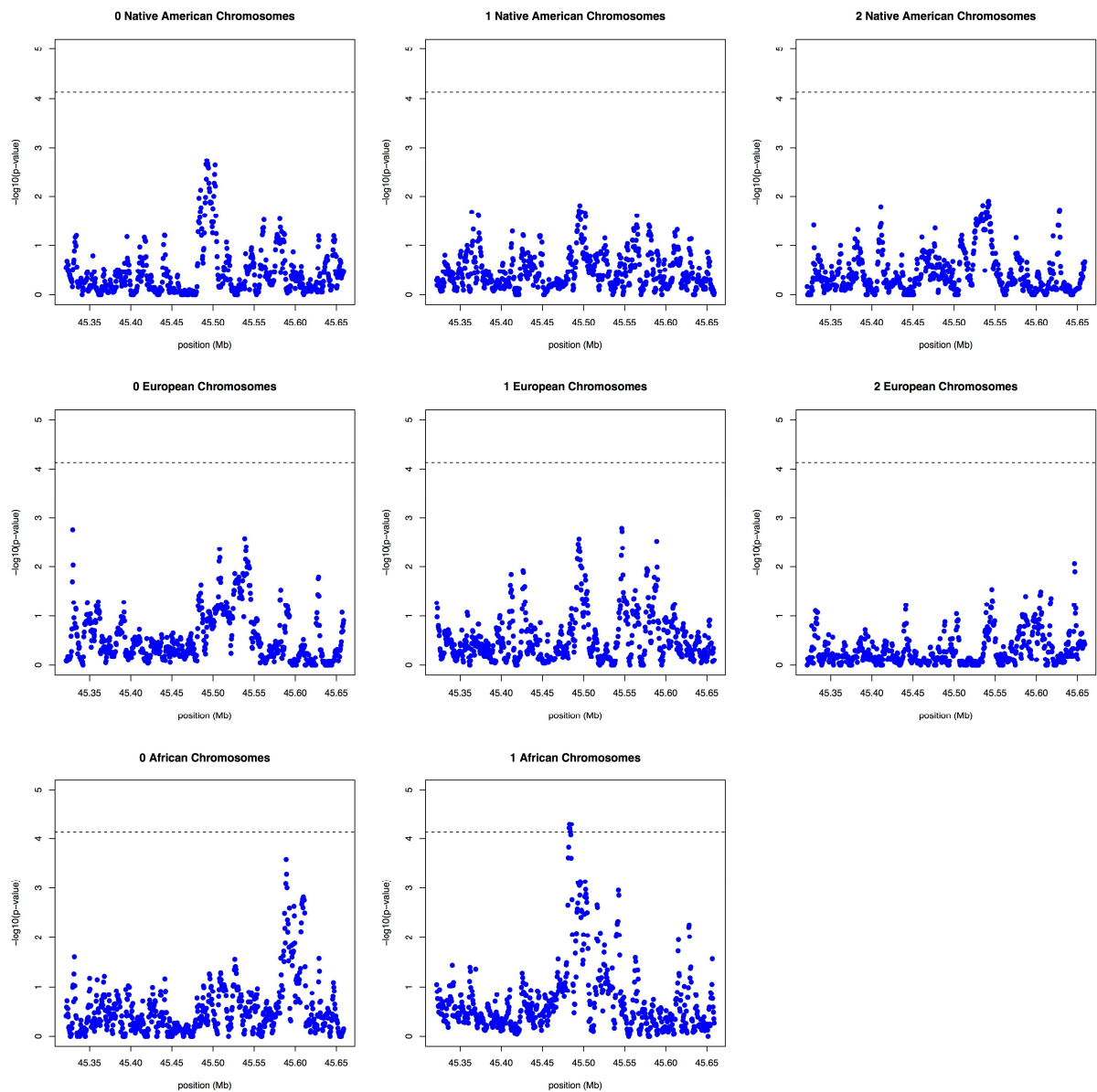


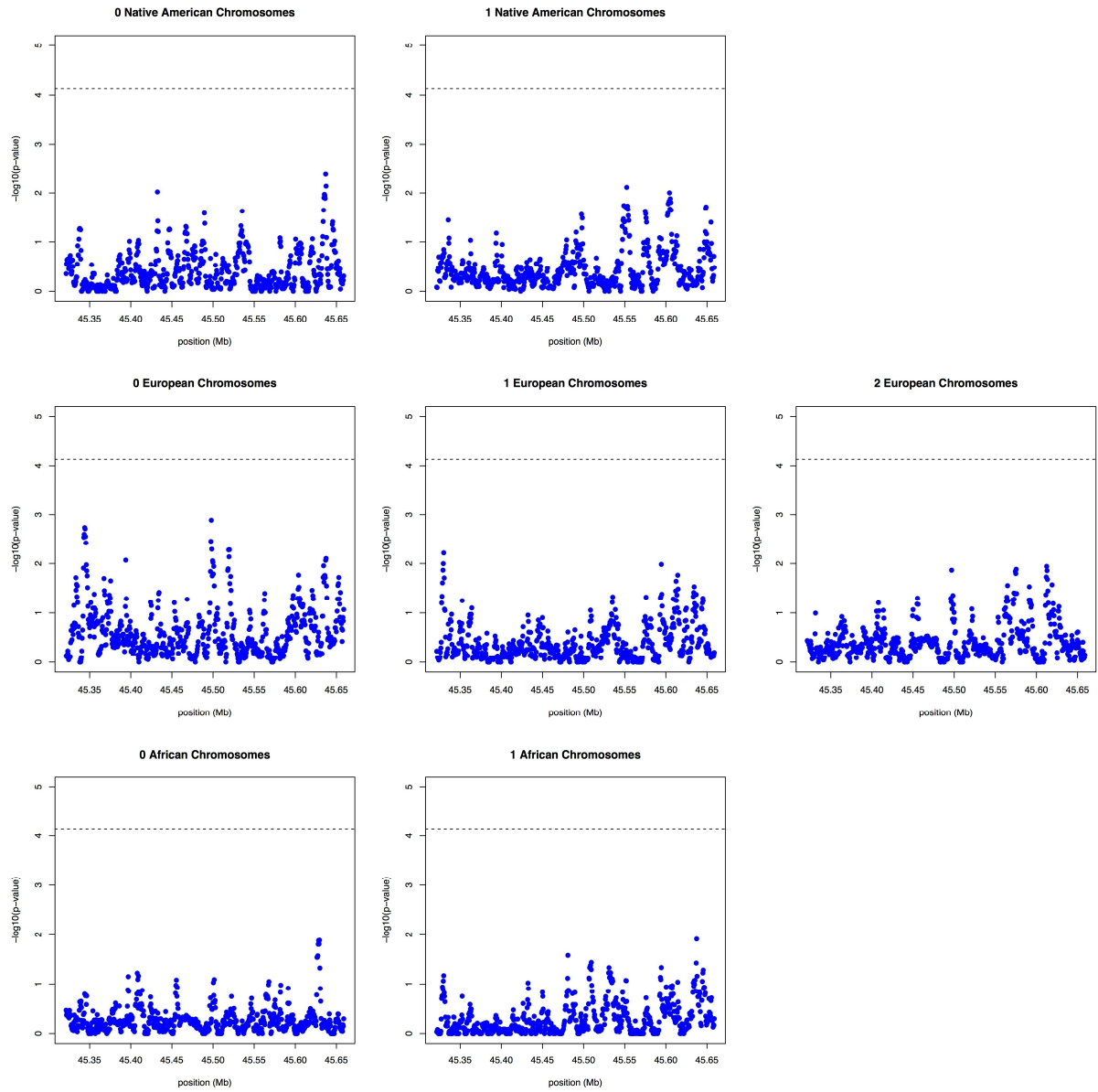


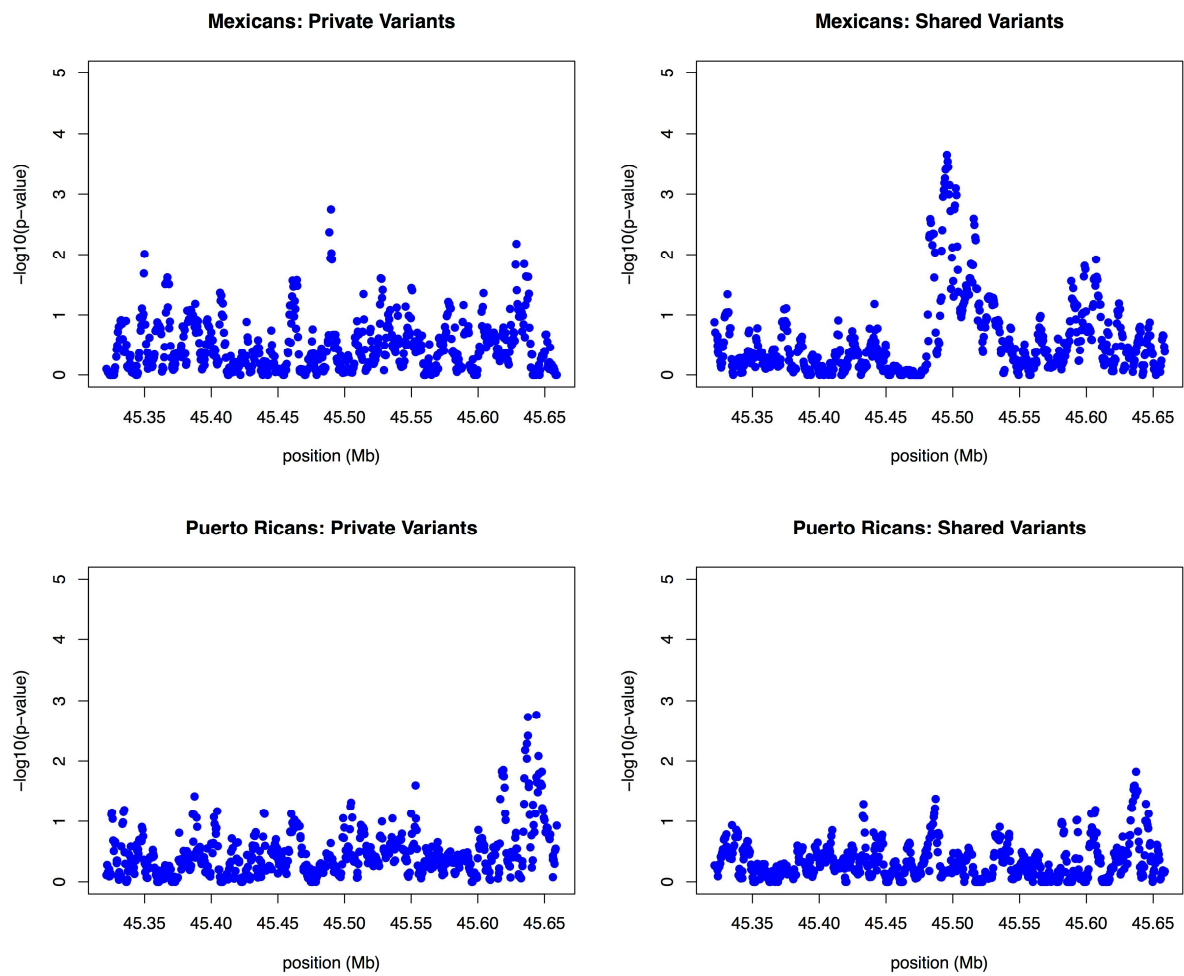




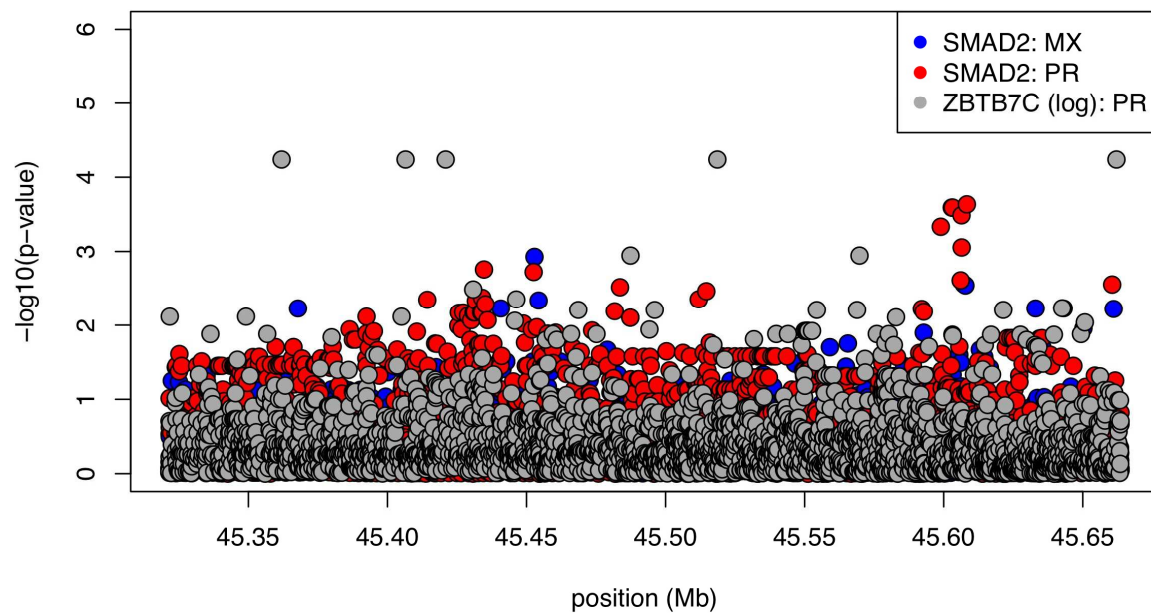




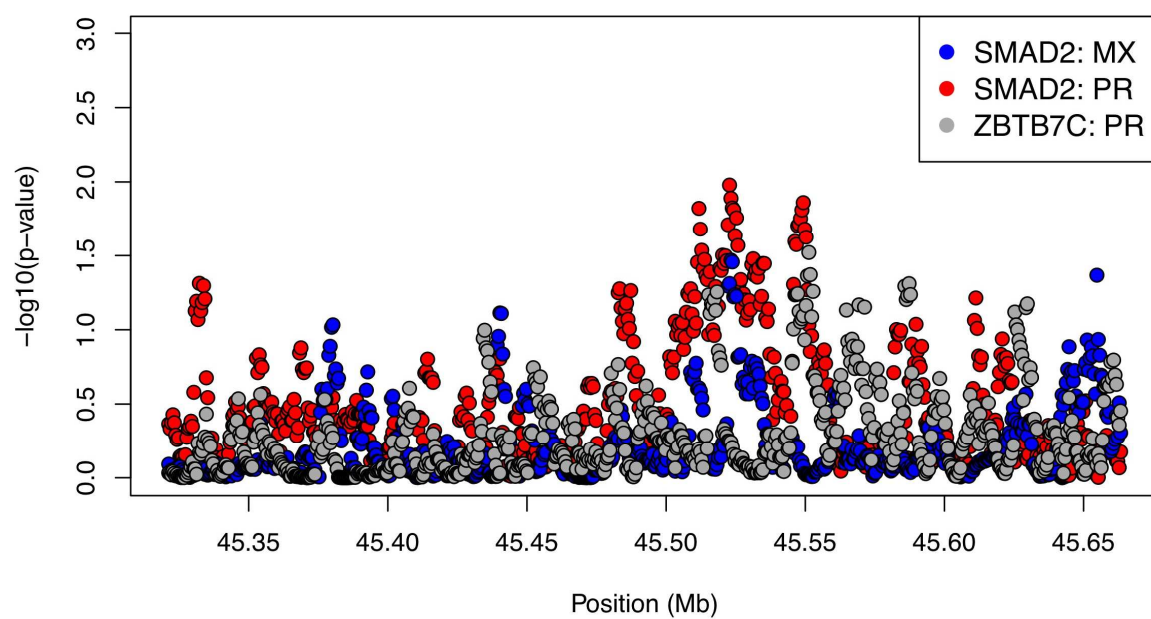


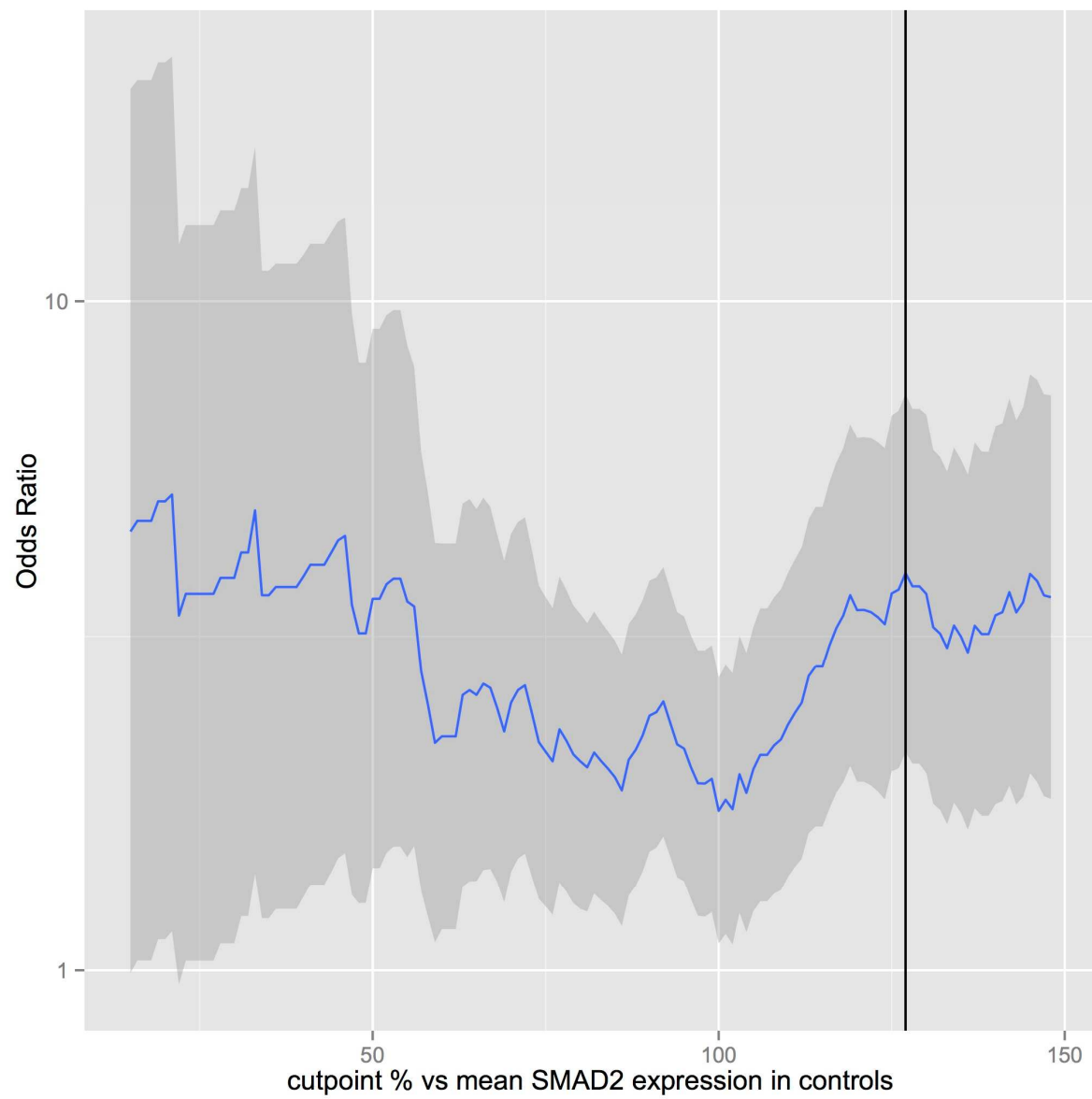


eQTL tests, Individual Variants



SKAT eQTL tests, 5 Kb sliding window, stepsize=500bp





1 **Supplementary Material**

2

3 **An admixture mapping meta-analysis in Latinos implicates rare genetic**
4 **variation at 18q21 with asthma susceptibility**

5

6

7 Christopher R Gignoux, Dara G Torgerson, Maria Pino-Yanes et al.

8

9

10

Supplementary Methods – the GALA II study

Study criteria: To be eligible for participation in the GALA II study, each participant or parent must have identified all four grandparents as Latino. A total of 3,774 children were included in the current study (1,893 participants with asthma and 1,881 healthy controls) from five centers (Chicago, Bronx, Houston, San Francisco Bay Area, and Puerto Rico) using a combination of community and clinic-based recruitment. Participants with asthma reported a physician diagnosis of asthma and at least two symptoms (shortness of breath, wheezing, or cough not associated with upper respiratory illness) or chronic use of controller medication (inhaled corticosteroids, leukotriene modifiers, theophylline or oral steroids) within the two years preceding recruitment. Eligible controls had no reported history of asthma, lung disease or chronic illness, atopic disease (eczema, hives, hay fever, allergic rhinitis), no reported use of medication for allergies, and no reported symptoms of coughing, wheezing, or shortness of breath in their lifetime. Controls were 1:1 frequency matched within each region by catchment area and age (within 1 year) and recruited from the same hospitals or community clinics as the cases. All GALA II participants were subject to extensive phenotype characterization, including pulmonary function testing using standard reference equations.¹

Medication Use: Subjects were asked to list their asthma prescriptions and responses were grouped into five different treatment categories: (1) no medications, (2) rescue inhalers/short-acting beta agonists (SABA) only, (3)

controller monotherapy, (4) combination therapy, and (5) oral corticosteroids (OCS). Monotherapy included subjects who were taking an inhaled corticosteroid (ICS), leukotriene receptor antagonist (LTRA), or theophylline to control their asthma. The combination therapy group included subjects using two or more controller medications with or without long-acting beta agonists (LABA).

Composite Exacerbation Score: Information regarding self-reported asthma-related hospitalizations, emergency department visit, and oral steroid use over the 12 months prior to recruitment was collected through the administered questionnaire. With these data, a self-reported exacerbation score was assigned to each subject based on the American Thoracic Society and European Respiratory Society consensus statement from 2009.² One point was given for each report of hospitalization or emergency department visit in the last 12 months. For reported history of oral steroid use, one point was assigned for a report of any oral steroid use in the last 12 months and two points were assigned if the subject reported greater than two continuous weeks of oral steroid use over the 12 months prior to recruitment. Histograms of exacerbation scores for all three populations for the individuals with gene expression data is available in Figure S8. For statistical tests, we used multiple ordered logistic regression to test for the association between scores using the MASS package in R. Significance for ordered logistic regression terms was determined where possible via likelihood ratio test.

Pulmonary Function Testing: Subjects with asthma were instructed to withhold their bronchodilator medications for at least 8 hours before lung function testing. Spirometry was performed according to the American Thoracic Society standards (1995). Standard measurements of airway obstruction included Forced Expiratory Volume in one second (FEV₁), Forced Expiratory Flow between 25-75% of vital capacity (FEF₂₅₋₇₅) and Forced Vital Capacity (FVC). FEV₁/FVC, and FEF₂₅₋₇₅ are all adjusted for age, sex and height² as covariates, as percent predicted equations are not available for Puerto Ricans. Upon reception of spirometry data at UCSF, a pulmonologist reviewed every spirometry curve for the individuals involved in the present study. In addition, one individual with good values but with a change in FEV₁ > 60% was removed to minimize outlier effects in linear regression.

Genome-wide SNP genotyping: Individuals were genotyped on the Affymetrix Axiom® Genome-Wide LAT1 Array (World Array 4). Genotype calling and quality control was performed according to standard recommendations by Affymetrix. Briefly, we filtered out single nucleotide polymorphisms (SNPs) with >5% missing data and those with poor platform specific clustering (n=63,328), along sites deviating strongly from Hardy-Weinberg equilibrium (n=1845; $p < 10^{-6}$) within their respective populations (African American, Puerto Rican, Mexican, and other Latino). Subjects were required to have >97% call rates and no gender discrepancies, and run through a pairwise identity by descent (IBD) scan to identify related individuals to remove (IBD>0.3). The QC process yielded 3,774 individuals and 747,129 variable sites for further analysis.

Supplementary Methods – the GALA II study

The Study of African Americans, Asthma, Genes & Environments (SAGE) is a study of childhood asthma in African Americans from the San Francisco Bay area that was run in parallel with the GALA II study using identical protocols and questionnaires. Local institutional review boards approved the study and all subjects and legal guardians provided written informed assent/consent.

Supplementary Methods – Statistical Analyses

Local Ancestry Estimation: Like other mapping methods, admixture mapping requires accurate input data for high-powered studies to map associations with biological relevance. Genomic locus-specific, or local, ancestry can be estimated using models that leverage the recent mixture of ancestries present in populations such as Hispanic/Latinos, using high-density genotype data. Each method is tailored specifically for the mentioned method. Numerous papers in both population and medical genetics have demonstrated that Hispanic/Latino populations are largely comprised of three different continental ancestries, represented by African, Native American, and European parental populations, e.g.^{3,4}.

reference included the CEU and YRI from HapMap⁵ (European and African reference haplotypes, respectively), and Indigenous Mexican individuals from HGDP⁶ that were genotyped on the Illumina 650Y, and Pima and Maya individuals kindly provided by Drs. Mark Shriver and Abigail Bigham that were genotyped on

the Affymetrix 6.0^{7,8} As the available Illumina and Affymetrix arrays do not overlap much in their genotypic content we chose the reference populations genotyped on each platform. For each study we used the intersection of genotyped markers with available reference data: allele frequencies for each population for LAMP, and ancestral haplotypes for LAMP-HAP phased using Beagle⁹.

Given the relatively sparse intersection of platforms, we chose to estimate local ancestry first then merge ancestry estimates on a consistent background. Local ancestry estimation may be sensitive to genotyping errors, and thus we did not estimate local ancestry from imputed genotypes. Rather, we created a consensus set of sites across different genotyping platforms and inferred ancestry at untyped sites using linear (genomic distance) interpolation. In this framework, an untyped site was assigned the average of its neighboring genotyped sites, weighted by the genetic distance in bp from each. We used this framework to interpolate local ancestry for each study at 3,192,437 HapMap II SNPs. While other platforms are available this common framework gave a union of sites that would balance SNP overlap across all platforms with computational efficiency. Given that admixture in the Americas is a recent phenomenon we observed few ancestry switches per chromosome per individual, and thus the majority of untyped sites were within blocks of contiguous ancestry. As such we chose the HapMap sites more as a background reference for association tests rather than a set of unique sites, and we would anticipate interpolating to larger-or-smaller panels to give consistent results.

Likelihood Ratio Test for Admixture Mapping: To estimate the combined effect of all ancestries in Latinos, we used a 2-degrees of freedom (df) likelihood ratio test comparing regression models with and without local ancestry terms. We used logistic regression for case-control studies and Poisson regression models of counts of transmitted/untransmitted ancestry states for trio studies. All logistic regression models were adjusted for genomic ancestry as determined using ADMIXTURE¹⁰ on the full autosomal data. Our likelihood ratio test for case-control analyses consisted of comparing the likelihoods of two nested generalized linear models:

$$full : \log\left(\frac{p}{1-p}\right) \sim \beta_0 + \beta_1 local_1 + \beta_2 local_2 + \beta_3 global_1 + \beta_4 global_2 + \varepsilon$$

$$restricted : \log\left(\frac{p}{1-p}\right) \sim \beta_0 + \beta_1 global_1 + \beta_2 global_2 + \varepsilon$$

The statistic $2 \log\left(\frac{likelihood(full)}{likelihood(restricted)}\right)$ then follows a χ^2 distribution with 2df (equal to the number of additional terms in the full model).

Similarly, for the trio studies we compared likelihoods of two generalized linear models, although the framework is more complicated. We modeled counts of transmitted/untransmitted ancestry pairs at each locus using a mixture of Poisson regression terms for each ancestry (e.g., transmitted African/untransmitted Native American, six different combinations from three different ancestries in total). Each specific ancestry contributes to the counts of the observed transmitted/untransmitted pairs using an indicator function:

$$\Phi_{local} = \{1 \in transmitted, -1 \in untransmitted, 0 \in otherwise\}$$

Each Φ_{local} is then a six-term vector. To account for the investigation of multiple transmitted/untransmitted ancestries at every locus, we adjusted our analyses using a 6-term factor Ξ that stratified counts into corresponding pairs of transmitted/untransmitted ancestries. Then, the joint effects of incorporating multiple ancestry terms were modeled by comparing the likelihoods of two nested generalized linear models:

$$\begin{aligned} full : \log(counts) &\sim \beta_0 + \beta_1\Phi_{African} + \beta_2\Phi_{European} + \beta_3\Phi_{NativeAmerican} + \beta_4\Xi + \varepsilon \\ restricted : \log(counts) &\sim \beta_0 + \beta_1\Xi + \varepsilon \end{aligned}$$

The statistic $2\log\left(\frac{likelihood(full)}{likelihood(restricted)}\right)$ then follows a χ^2 distribution with 2df as the 3 ancestries are collinear (and one ancestry term is dropped). This approach is similar to a McNemar's test of symmetry, however our implementation saves a degree of freedom given the inherent correlation structure of the 3 ancestries.

Scripts and functions were written in R and Python and are available upon request.

Multiple test correction using empirical autoregression: We determined the effective number of tests across the genome in the full meta-analysis by fitting an autoregressive model to the summary statistic data using the *coda* package in R, using an approach similar to Shriner et al.,¹¹ for both odds ratios and p-values. Overall correlation patterns were determined by estimating the correlation sequentially along the chromosome; each chromosome was run independently then the number of independent tests was summed. While similar to that obtained from odds ratios, the number of effective tests determined from p-values was slightly

more conservative, so we chose to use those. Our criterion for genome-wide significance was then 0.05 divided by the total number of effective tests across the genome. This method provided an estimate for the familywise error rate (controlling the total number of false positives), and thus more stringent than applying a false discovery rate.

Determining Gene Expression Cut-Point: Given that gene expression is continuous, we wanted to find a cut-point that would best determine high-vs-low gene expression to evaluate its association with asthma. We estimated the *maximum a posteriori* value for a cut-point of gene expression by calculating likelihoods of a logistic regression fit across the continuum of *SMAD2* gene expression levels, holding all other covariates constant. The total likelihood is defined then the sum of all the possible cut-point likelihoods. The logistic regression model including relevant covariates as a generative model for likelihoods, and then tested the hypothesis that expression affects the odds of disease, vs. the null hypothesis of no expression effect. In modeling both hypotheses, Bayes Factors provide an evidence-based rationale for determining the model that best fits the model of association. For each percentage point in our scale of normalized gene expression, we evaluated the odds ratio, confidence interval, and Bayes Factor. We chose the best-fitting model to differentiate high vs low expression as the cutpoint with the maximum Bayes Factor. Using the AIC or another likelihood-based statistic would be expected to give similar results in this scenario, particularly for the finding the best-fit cutpoint.

For the CD4+ Lymphocyte study, we had a limited number of individuals and a different set of units as expression was measured on a microarray. Therefore we used the extrapolated cutpoint estimated from the whole blood expression measurements in GALA II and SAGE II for our analyses rather than estimating a cutpoint from the CD4 data. Both continuous and cutpoint analyses are presented in Table S5 and a logistic regression plot is shown in Figure S9.

Exacerbation Score Prediction: We used Receiver Operator Characteristic (ROC) curves to calculate the performance of multiple logistic regression models. As these models can be prone to overfitting, we chose regression models with a small number of nesting terms, starting with use of controller medications, then bronchodilator drug response spirometric values, and finally adding gene expression. To create our outcome, we regressed out the effects of age and population of origin in the exacerbation scores using a linear model (those being significantly associated with outcome in the original model, see Table S6). Then, our outcome for logistic regression became a product of those residuals, separating out the individuals into high exacerbation and low exacerbation groups. We used the AIC to evaluate the fit of increased terms. We then evaluated the possibility of modifying the cutpoint between high exacerbation and low exacerbation, while keeping track of the AUC of the model. In addition, we chose to model the error rate in prediction from the model using out-of-sample cross-validation. The sample was split into K=10 pools, and each time 9 were used for training and one for testing. An

overall out-of-sample prediction error can then be estimated to avoid overfitting.

We performed cross-validation using the `cv.glm` function in the 'boot' package in R¹².

Estimation of white blood cell count proportions: *SMAD2* expression data was obtained from whole blood, a tissue that contains multiple distinct cells. It has been shown that these functionally distinct populations have unique expression signatures.¹³ This can lead to potential confounding when investigating expression differences between cases and controls, if the distribution of cell types also differs between the cases and controls. To address this, we used epigenome wide methylation data measured using the Illumina Infinium HD Methylation Assay (Illumina) according to the manufacturer's instructions in a subset of 62 Mexican and 44 Puerto Rican samples in GALA II. We estimated cell proportions using a statistical algorithm that uses reference information on cell-specific methylation signatures to estimate cell proportions from genome-scale methylation data¹⁴. In this method, the distribution of cell types for each sample was inferred based on DNA methylation signatures of the constituent cell types. A total of six different cell types, including CD4+ and CD8+ T cells, NK cells, B cells, monocytes and granulocytes, were estimated, and we tested to see whether *SMAD2* expression correlated with any cell types. We repeated the analyses of asthma susceptibility and exacerbations using the subset of individuals for whom we had overlapping expression and methylation data, accounting for the cell types estimated using generalized linear models, investigating individual blood cell count associations with *SMAD2* expression (adjusting for disease status, age, sex, gender, recruitment

center and population) and overall associations of the suite of cell counts using likelihood ratio tests.

Additional Gene Expression Details: Expression levels were calibrated using the $2^{-\Delta CT}$ transformation (yielding relative-fold expression), and rescaled to fold-change in expression versus controls. This transformation should not affect statistical significance of results but should be more easily interpretable. We screened for *cis*-eQTLs using the relative expression ($2^{-\Delta CT}$) of *SMAD2* and *ZBTB7C* with pooled variants using SKAT-O, and at individual variants using linear regression in R. The expected fold-change in expression due to asthma status was compared using linear regression with fold-change expression as the outcome, adjusting for age, sex, recruitment center and genetic ancestry to account for potential population stratification. We evaluated the goodness-of-fit for varying cutpoints of high- versus low-gene expression as a predictor of asthma status using Bayes Factors, however other measures of goodness-of-fit would yield similar results. We estimated the prediction power of multiple models on a composite asthma exacerbation score (combining information on hospitalizations, emergency department visits, and oral steroid usage in the past year determined by survey), and performed model selection in generalized linear models by estimating the AUC of Receiver Operating Characteristic (ROC) curves and the Aikake Information Criterion (AIC), along with 10-fold cross-validated estimates of prediction error (see Supplementary Methods in the Online Repository). Lastly, we sought to account for potential changes in expression due to changes in white blood cell compositions as our samples derived

from whole blood extractions. As methylation can track cell composition we estimated white blood cell count proportions from a subset of individuals with epigenome-wide methylation data to test whether *SMAD2* expression was correlated to cell counts.

Imputation: To perform dense fine mapping with sparse data we performed a joint phasing/imputation step using standard methods from the field of genetic epidemiology. This requires statistical haplotype inference using models that leverage population genetic theory to provide maximum-likelihood estimates of the two haplotypes in the autosomes for a given individual, a process known as phasing. This haplotypic inference allows for genome-level imputation using a secondary algorithm, providing a statistical estimation of untyped variants and the level of posterior certainty at any given site. For this project, we implemented a standard pipeline in genetic epidemiology: we first phased each study using SHAPEIT¹⁵ and performed genotype imputation using IMPUTE2¹⁶ using haplotypes from the 1000 Genomes Project (phase I) as a reference. Imputation was carried out using the recommended settings in IMPUTE2 for prephased data. Imputed genotypes with information scores >0.3 were used for downstream analysis as recommended by the IMPUTE2 authors. We performed the phasing across the entire chromosome, followed by imputation for a small (~5Mb) region around *SMAD2*.

ACCEPTED MANUSCRIPT

Table S1. Top associations with ancestry (admixture mapping peaks), as defined by a minimum p-value < 0.001, out to 0.01 on each side. Joint refers to the 2-df likelihood ratio test for all ancestries. If a region was identified in multiple ancestry scans, the ancestry with the smallest p-value was used and the other ancestry information is given in parentheses. Coordinates are in hg18.

Associated Ancestry	Chr	Start	End	minimum p-value	Genes in Peak
Native	1	20716967	21885179	3.48x10 ⁻⁴	<i>USP48, HP1BP3, SH2D5, CDA, EIF4G3, RAP1GAP, NBPFF3, KIF17, ECE1, ALPL, PINK1, LOC100506801, DDOST, FAM43B</i>
Joint	1	23046750	24398722	3.67x10 ⁻⁴	<i>None</i>
Native (European)	1	218900613	219027538	8.97x10 ⁻⁴	<i>MARC2, MARK1, C1orf115, MARC1</i>
European	1	222843773	223116557	9.93x10 ⁻⁴	<i>CNIH3</i>
European	1	226684711	227897382	6.67x10 ⁻⁴	<i>DUSP5P, ACTA1, NUP133, RAB4A, TAF5L, RNF187, CCSAP, RHOU, ABCB10, HIST3H2BB, URB2, MIR4666A, HIST3H2A, SPHAR</i>
Joint	2	57596338	58186261	5.85x10 ⁻⁴	<i>None</i>
Native	3	4015247	4556887	5.48x10 ⁻⁴	<i>ITPR1, SUMF1, SETMAR</i>
Joint	3	40060302	41086194	2.62x10 ⁻⁴	<i>None</i>
Joint	3	95606242	96292801	4.09x10 ⁻⁴	<i>None</i>
European	4	23534331	23968123	3.19x10 ⁻⁴	<i>None</i>
African	4	48077300	52634909	4.05x10 ⁻⁴	<i>OCIAD2, OCIAD1, SLC10A4, ZAR1, SPATA18, SGCB, CWH43, SLAIN2, FRYL, DCUN1D4, LRRC66</i>
African	5	55678536	57344829	1.60x10 ⁻⁴	<i>SETD9, GPBP1, MAP3K1, ACTBL2, MIER3</i>
Native	5	60547184	61124131	1.94x10 ⁻⁴	<i>ZSWIM6, C5orf64</i>
Native	5	62620986	63502012	6.70x10 ⁻⁴	<i>HTR1A, RNF180</i>
Joint	5	73710967	74034961	7.35x10 ⁻⁴	<i>None</i>

Associated Ancestry	Chr	Start	End	minimum p-value	Genes in Peak
Joint (Native)	5	80682142	81813860	3.57x10 ⁻⁴	<i>ATG10, ACOT12, SSBP2, RPS23, ATP6AP1L (CKMT2, LOC100131067, RASGRF2, ZCCHC9)</i>
African (Joint)	5	178998248	180340825	4.43x10 ⁻⁴	<i>FLT4, CANX, C5orf60, MAML1, OR2Y1, LOC729678, RNF130, LTC4S, LOC100859930, MAPK9, CNOT6, RASGEF1C, MGAT1, MIR1229, C5orf45, CBY3, SQSTM1, BTNL8, SCGB3A1, ZFP62, TBC1D9B, MGAT4B, MIR340, GFPT2</i>
Joint	6	90202813	90311362	8.35x10 ⁻⁴	<i>None</i>
Joint	7	55129505	55621885	9.39x10 ⁻⁴	<i>None</i>
Joint	7	117211513	117400523	3.64x10 ⁻⁴	<i>None</i>
Joint	7	118344280	118830576	7.12x10 ⁻⁴	<i>None</i>
Joint	8	28677088	29253448	4.64x10 ⁻⁴	<i>None</i>
African	8	56410052	56871124	7.63x10 ⁻⁴	<i>XKR4, TGS1, SBF1P1, TMEM68</i>
Native	8	75215069	75733856	4.27x10 ⁻⁴	<i>FLJ39080, MIR5681A, MIR5681B, JPH1, GDAP1</i>
European	9	88694862	90508804	6.04x10 ⁻⁴	<i>LOC392364, LOC286238, CTSL1P8, LOC100506834, FAM75C2, LOC494127, FAM75C1, C9orf170, LOC440173, CTSL1, CTSL3, CDK20, GAS1, NXNL2, SPIN1, FAM75E1, DAPK1</i>

Associated Ancestry	Chr	Start	End	minimum p-value	Genes in Peak
European (African, Joint)	9	96446960	100419073	1.67x10 ⁻⁶	LOC100507346, LOC100499484, TMOD1, NCBP1, FOXE1, MIR27B, HABP4, MIR23B, C9orf3, MIR2278, LOC158434, LOC158435, TSTD2, GABBR2, TRIM14, HIATL2, ZNF782, FANCC, LOC340508, ZNF510, HSD17B3, LOC286359, C9orf174, LINC00092, XPA, TDRD7, ZNF367, LOC441454, HEMGN, AAED1, LOC441455, CORO2A, NANS, ANP32B, MIR24, LOC100132781, FAM22G, CDC14B, MIR3074, C9orf156, ERCC6L2, LINC00476, CTSL2, SLC35D2, PTCH1, TBC1D2 (ANKS6, COL15A1, FBP1, FBP2, GALNT12)
African	9	108352743	108738041	9.81x10 ⁻⁴	ZNF462
Joint	10	51824523	52200679	9.38x10 ⁻⁴	None
African	10	83972038	84362276	3.10x10 ⁻⁴	NRG3
Native	10	100396273	101144430	7.93x10 ⁻⁴	CNNM1, HPSE2
European (Native)	10	102072780	102969197	7.40x10 ⁻⁴	MIR608, PDZD7, KAZALD1, NDUFB8, SCD, PAX2, C10orf2, WNT8B, TLX1, MRPL43, LINC00263, SFXN3, PKD2L1, HIF1AN, TLX1NB, LZTS2, SEC31B, FAM178A, SEMA4G (ACTR1A, ARL3, BTRC, C10orf76, C10orf95, CUEDC2, DPCD, ELOVL3, FBXL15, FBXW4, FGF8, FLJ41350, GBF1, HPS6, KCNIP2, LBX1, LDB1,

Associated Ancestry	Chr	Start	End	minimum p-value	Genes in Peak
					<i>LOC100289509, LOC100505761, MGEA5, MIR146B, MIR3158, NFKB2, NOLC1, NPM3, PITX3, POLL, PPRC1, PSD, SUFU, TMEM180, TRIM8)</i>
European (Native)	12	21952123	25112841	1.67x10 ⁻⁵	<i>MIR920, C12orf77, BCAT1, ST8SIA1, CMAS, ABCC9, ETNK1, LINC00477, LRMP, KIAA0528, SOX5</i>
Joint	12	40400256	41129450	7.16x10 ⁻⁴	<i>None</i>
African	13	67676298	68122400	7.40x10 ⁻⁴	<i>None</i>
Joint	14	22233894	23109399	3.47x10 ⁻⁴	<i>None</i>
Joint	14	23958088	24164696	8.49x10 ⁻⁴	<i>None</i>
European (Native)	16	15398985	16537826	3.98x10 ⁻⁴	<i>MYH11, KIAA0430, MPV17L, NOMO3, NDE1, FOPNL, MIR484, C16orf45, PKD1P1, ABCC6, MIR3179, ABCC1, MIR3180</i>
Joint	17	69510566	71397400	7.96x10 ⁻⁴	<i>None</i>
Joint	18	6230094	6746661	7.96x10 ⁻⁴	<i>None</i>
Native	18	38783085	39434024	4.92x10 ⁻⁴	<i>SYT4, RIT2</i>
Joint (European, Native) ¹	18	40530823	44316091	1.71x10 ⁻⁴	<i>C18orf25, SLC14A2, SLC14A1, ATP5A1, HAUS1, LOXHD1, RNF165, MIR4319, SIGLEC15, SETBP1, PSTPIP2, EPG5, HDHD2, SMAD2, TCEB3C, TCEB3B, ZBTB7C, TCEB3CL, IER3IP1, KATNAL2, ST8SIA5, LOC100506888, PIAS2</i>
Joint	18	49677502	49903778	9.83x10 ⁻⁴	<i>None</i>
Joint	22	46376748	46640808	1.26x10 ⁻⁴	<i>None</i>

¹ Peak had a gap in joint analysis from chr18:42207149-42218476, where the p-values were above 0.01 to a maximum of 0.0128. However given how close the two peaks were, we combined them into a single entry in the table.

Table S2. Functional annotation of top asthma-associated variants from RegulomeDB¹⁷. TF = transcription factor.

SNP	Position	RegulomeDB Score	Supporting Data
rs76402589	45497831	3a	TF binding + any motif + DNase peak
rs113382450	45499287	No Data	
rs79405582	45499892	No Data	
rs112042427	45496169	4	TF binding + DNase peak
rs75478387	45496266	3a	TF binding + any motif + DNase peak
rs8099232	45496784	6	Other
rs7236490	45498260	6	Other
rs59002988	45496882	5	TF binding or DNase peak

Table S3. P-values for tests of association with pooled variants in coding exons of *SMAD2* and *ZBTB7C* using SKAT. A total of 9 variants were identified in the coding exons of *SMAD2*, and 47 in the coding exons of *ZBTB7C*.

SMAD2:	Unadjusted	Adjusted for Global Ancestry	Adjusted for Global and Local Ancestry
Mexicans	0.32	0.096	0.15
Puerto Ricans	0.25	0.32	0.27
Combined*	0.51	0.67	0.59
ZBTB7C:	Unadjusted	Adjusted for Global Ancestry	Adjusted for Global and Local Ancestry
Mexicans	0.77	0.90	0.89
Puerto Ricans	0.66	0.75	0.27
Combined*	0.84	0.80	0.74

*adjusted for ethnicity

Table S4. Association testing for *SMAD2* expression in isolated CD4+ Lymphocytes in individuals from Germany (GEO accession number GDS267).

CD4+ Lymphocytes: <127% Control Expression	N	OR (95% CI)	p-value
Logistic Regression	15 vs 8	1.87 (1.25 - 2.79)	<0.01
Fisher's Exact Test	15 vs 8	15.48 (1.08 - 974)	0.02

CD4+ Lymphocytes: per 10% decrease compared to controls	N	OR (95% CI)	p-value
Logistic Regression	15 vs 8	1.61 (1.00-2.60)	<0.01

Table S5. Full covariate ordered logistic regression modeling of exacerbation scores, with odds ratios for single-level transitions between exacerbations.

Exacerbation score

	Value	Std.	p	OR (L95-U95)
SMAD2 expression^a	0.6028	0.24003	0.011	1.11 (1.02-1.20)
age	-0.1159	0.03114	<0.001	0.89 (0.84-0.95)
Sex	0.3097	0.20504	0.13	1.36 (0.91-2.04)
Mexicans ^b	-0.2589	0.25464	0.15	0.77 (0.54-1.10)
Puerto Ricans^b	0.6106	0.26533	0.011	1.84 (1.15-2.95)
African Americans ^{b,c}	-0.1258	0.22852	0.29	0.88 (0.70-1.11)

- per 10% decrease in expression compared to the average level in controls
- Population significance calculated from t-statistic, rather than likelihood ratio, as populations are intercorrelated.
- African American parameters estimated using a second ordered logistic model, with African American ancestry listed as yes/no rather than a population factor.

Figure S1. Omnibus tests of admixture mapping in both the Latino (left, 2-df) and African-American (right, 1-df) studies in the discovery sample in EVE at the top hit on 18q21. The meta-analysis p-value in Latinos is 6.8×10^{-6} while in African-Americans it is 0.7.

Figure S2. Manhattan plots of single-ancestry admixture mapping for African, European, and Native American ancestry, respectively. Peaks on chr9p31 and chr12p12 show up as genome-wide significant although they do not replicate in GALA II. Peaks encompassing *SMAD2* approach genome-wide significance in both the European- and Native American-specific admixture mapping meta-analyses.

Figure S3. QQplot of allelic association in the *SMAD2* region following imputation of genotypes from the phase I 1000 Genomes Project for all studies included in the EVE meta-analysis (including the African American and European American studies). Associations were performed using logistic regression, TDT, or mixed model analysis (EMMAX) depending on study design.

Figure S4: Results of association testing for asthma in Mexicans separated by local ancestry across sliding windows of 5Kb in length, and a stepsize of 500 bp using SKAT-O. The dashed line represents a bonferroni correction for 675 tests.

Figure S5: Results of association testing for asthma in Puerto Ricans separated by local ancestry across sliding windows of 5Kb in length, and a stepsize of 500 bp using SKAT-O. The dashed line represents a bonferroni correction for 675 tests.

Figure S6: Results of association testing for asthma for private and shared variants in Mexicans and Puerto Ricans across sliding windows of 5Kb in length, and a stepsize of 500 bp using SKAT-O.

Figure S7. Results of eQTL analysis for individual (linear regression) and pooled variants (SKAT-O) with expression of *SMAD2* and *ZBTB7C* in Mexicans (MX, n=177) and Puerto Ricans (PR, n=144).

Figure S8. Evaluation of Odds Ratios and cutpoints along the spectrum of possible values for SMAD2 expression, adjusting for population, age, sex, and recruitment center. Best-fit cut-point as discussed in the main text corresponds to the black line at 127% of average control SMAD2 expression.

Figure S9. Association of SMAD2 expression in CD4+ Lymphocytes with Asthma (GEO accession number GDS267). Distributions of cases and controls (status 1 and 0 respectively), are given as dots and boxplots, with the logistic regression line ($p=0.01$) fit to the data in red. The predetermined cutpoint of 127% expression in controls from Figure S9 and used in the association testing in the main text is given at the vertical line. Plot created from a modification of the popbio package¹⁸ in R.

Figure S10. Testing for a significant effect of asthma age of onset and *SMAD2* expression. For each age value on the x axis, we test whether we see a strong effect only including individuals whose asthma began early than that date. For example, the point at age 10 is the test of association between controls and anyone whose asthma began before the age of 10. As can be seen, the most significant test can be found at an age of 5.1-5.5, therefore we expect the association of *SMAD2* with asthma status to be strongest in those who develop asthma at 5 or younger.

385 **Figure S11.** Distribution of exacerbation scores of individuals in GALA II with
386 measured *SMAD2* expression.
387
388

Figure S12. Baseline ROC curve and AUCs of SMAD2 expression, spirometry, and long-term medication use on the exacerbation score residuals, given an equal number of high- and low-exacerbations, after accounting for age and population. All models are nested and introducing additional terms each time remains significant via likelihood ratio test/AIC. These then provide the justification for including SMAD2 expression in the logistic regression models seen in Figure 3B. Improved AUC performance can be achieved by including demographic parameters, and varying the threshold between high and low exacerbations.

References

1. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med* 1999;159:179-87.
2. Reddel HK, Taylor DR, Bateman ED, et al. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am J Respir Crit Care Med* 2009;180:59-99.
3. Galanter JM, Gignoux CR, Torgerson DG, et al. Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. *J Allergy Clin Immunol* 2014;134:295-305.
4. Bryc K, Velez C, Karafet T, et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* 2010;107 Suppl 2:8954-61.
5. Consortium IH, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851-61.
6. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008;319:1100-4.
7. Bigham A, Bauchet M, Pinto D, et al. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet* 2010;6.
8. Bigham A, Bauchet M, Pinto D, et al. Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. *PLoS genetics* 2010;6:e1001116.
9. Browning S, Browning B. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies *The American Journal of Human Genetics* 2007.
10. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009;19:1655-64.
11. Shriner D, Adeyemo A, Rotimi CN. Joint ancestry and association testing in admixed individuals. *PLoS Computational Biology* 2011;7:e1002325.
12. Canty A, Ripley B. *boot: Bootstrap R (S-Plus) Functions*. 2013.
13. Palmer C, Diehn M, Alizadeh AA, Brown PO. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* 2006;7:115.
14. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012;13:86.
15. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2012;9:179-81.
16. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;5:e1000529.
17. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;22:1790-7.

- 444 18. Stubben C, Milligan B. Estimating and analyzing demographic models using
445 the popbio package in R. Journal of Statistical Software 2007;22:1-23.
446