bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TITLE

Whole Genome Sequencing of Pharmacogenetic Drug Response in Racially and Ethnically Diverse Children with Asthma

AUTHORS

Angel C.Y. Mak[1#*], Marquitta J. White[1#], Zachary A. Szpiech[2#], Walter L. Eckalbar[1], Sam S. Oh[1], Maria Pino-Yanes[3,4], Donglei Hu[1], Scott Huntsman[1], Joshua Galanter[1], Dara G. Torgerson[1], Ann Chen Wu[5,6], Blanca E. Himes[7], Soren Germer[8], Julia M. Vogel[8], Karen L. Bunting[8], Celeste Eng[1], Sandra Salazar[1], Kevin L. Keys[1], Thomas A. Nguyen[1], Pui-Yan Kwok[9,10], Albert M. Levin[11], Juan C. Celedón[12], Erick Forno[12], Hakon Hakonarson[13,14], Patrick M. Sleiman[13,14], Amber Dahlin[5], Kelan G. Tantisira[5], Scott T. Weiss[5], Denise Serebrisky[15], Emerita Brigino-Buenaventura[16], Harold J. Farber[17], Kelley Meade[18], Michael A. Lenoir[19], Pedro C. Avila[20,+], Saunak Sen[1], Shannon M. Thyne[21], William Rodriguez-Cintron[22], Cheryl A. Winkler[23], Andrés Moreno-Estrada[24], Karla Sandoval[24], Jose R. Rodriguez-Santana[25], Rajesh Kumar[26,27], L. Keoki Williams[28,29], Nadav Ahituv[1], Elad Ziv[1], Max A. Seibold[30], Robert B. Darnell[8,31,32], Noah Zaitlen[1], Ryan D. Hernandez[2,10,33], Esteban G. Burchard[1] and the Trans-Omics for Precision Medicine Whole Genome Sequencing Program (TOPMed) Team

[#] These authors contributed equally to this work

[+] Current address: Allergy & ENT Associates. 10847 Kuykendahl Rd, Suite 100. The Woodlands, TX 77382.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

* Correspondence and material requests should be addressed to Angel C.Y. Mak (angel.mak@ucsf.edu)

AUTHOR AFFILIATIONS

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

[2] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA.

[3] Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain.

[4] CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain.

[5] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.

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

[7] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

[8] New York Genome Center, New York, New York, USA.

[9] Cardiovascular Research Institute, University of California San Francisco, San Francisco, California, USA.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[10] Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA.

[11] Department of Public Health Sciences, Henry Ford Health System, Detroit Michigan, USA.

[12] Division of Pediatric Pulmonary Medicine, Allergy and Immunology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA.

[13] Center for Applied Genomics, The Children's Hospital of Philadelphia Research Institute, Philadelphia, Pennsylvania, USA.

[14] Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

[15] Pediatric Pulmonary Division, Jacobi Medical Center, Bronx, New York, USA.

[16] Department of Allergy and Immunology, Kaiser Permanente Vallejo Medical Center, Vallejo, California, USA.

[17] Department of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, Houston, Texas, USA.

[18] Children's Hospital and Research Center, Oakland, California, USA.

[19] Bay Area Pediatrics, Oakland, California, USA.

[20] Department of Medicine, Northwestern University, Chicago, Illinois, USA.

[21] Department of Pediatrics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA.

[22] Veterans Caribbean Health Care System, San Juan, Puerto Rico.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[23] Basic Science Laboratory, Center for Cancer Research, National Cancer Institute, Leidos Biomedical Research, Frederick National Laboratory, Frederick, Maryland, USA.

[24] National Laboratory of Genomics for Biodiversity (UGA-LANGEBIO), CINVESTAV, Irapuato, Guanajuato, Mexico.

[25] Centro de Neumologia Pediatrica, San Juan, Puerto Rico.

[26] Feinberg School of Medicine's Division of Allergy and Immunology, Northwestern University, Chicago, Illinois, USA.

[27] Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois, USA.

[28] Department of Internal Medicine, Henry Ford Health System, Detroit, Michigan, USA.

[29] Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, Michigan, USA.

[30] Center for Genes, Environment and Health, Department of Pediatrics, National Jewish Health, Denver, Colorado, USA.

[31] Laboratory of Molecular Neuro-Oncology, The Rockefeller University, New York, New York, USA.

[32] Howard Hughes Medical Institute, The Rockefeller University, New York, New York, USA.

[33] Quantitative Biosciences Institute, University of California San Francisco, San Francisco, California, USA.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ACKNOWLEDGEMENTS

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung, and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Genes-environments & Admixture in Latino Americans (GALA II) Study" (phs000920) and "NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environments (SAGE)" (phs000921) was performed at the New York Genome Center (3R01HL117004-01S3). We acknowledge New York Genome Center investigators and teams for whole genome sequencing sample preparation, quality control, data generation, data processing and initial joint genotyping. Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. This work was supported in part by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, Harry Wm. and Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, National Heart, Lung, and Blood Institute R01HL117004, R01HL128439, R01HL135156, X01HL134589, National Institute of Environmental Health Sciences R01ES015794, R21ES24844, the National Institute on Minority Health and Health Disparities P60MD006902, U54MD009523, R01MD010443, and the Tobacco-Related Disease

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

105 Research Program under Award Number 24RT-0025. M.P.-Y. was supported by the
106 grant AC15/00015 by Instituto de Salud Carlos III within the ERACoSysMed 1st Joint
107 Transnational Call (SysPharmPedia) from the European Union, under the Horizon
108 2020. R.B.D., S.G., K.L.B. and J.M.V. were supported by the National Human Genome
109 Research Institute of the National Institutes of Health under Award Number
110 UM1HG008901. C.A.W. was funded in part with federal funds from the National
111 Cancer Institute, National Institutes of Health, under contract HHSN26120080001E,
112 and by the Intramural Research Program of the NIH, National Cancer Institute, Center
113 for Cancer Research. The content of this publication does not necessarily reflect the
114 views or policies of the Department of Health and Human Services, nor does mention
115 of trade names, commercial products, or organizations imply endorsement by the U.S.
116 Government.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AUTHOR CONTRIBUTIONS

A.C.Y.M., M.J.W., Z.A.S., D.H., M.P.-Y., J.G. and S.H. analyzed data. R.D.H., N.Z., M.A.S., E.G.B., D.G.T., M.J.W., A.C.Y.M., N.A., E.Z., R.B.D. and P.-Y.K. designed experiments or supervised research. A.C.Y.M., M.J.W., Z.A.S., S.S.O., M.P.-Y., E.G.B. and K.L.K. prepared tables and figures, and wrote the manuscript. C.E., S.G., K.L.B. and J.M.W. generated data and supervised data generation. W.E., T.N. and B.E.H. performed validation experiments and analyses. M.J.W., M.P.-Y., L.K.W., A.M.L., J.C.C., E.F., H.H., P.M.S., A.C.W., A.D., K.G.T. and S.T.W. performed replication analyses. E.G.B., D.S., E.B.B., H.J.F., J.R.R.-S., K.M., M.A.L., P.C.A., R.K., S. Sen, S.T. and W.R.-C. recruited study subjects. S. Salazar coordinated recruitment and maintained phenotype data quality. C.A.W., A.M.-E. and K.S. contributed Native American samples. TOPMed team provided variants annotation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Albuterol, a bronchodilator medication, is the first-line therapy for asthma treatment worldwide. However it has a wide variation of drug response among different racial/ethnic groups. We performed the largest pharmacogenetics study to date, using whole genome sequencing data from 1,441 minority children with asthma from the extremes of bronchodilator drug response (BDR) to albuterol. We identified population-specific and shared pharmacogenetic variants associated with BDR, including genome-wide significant and suggestive loci near genes previously associated with lung capacity (*DNAH5*), immunity (*NFKB1* and *PLCB1*), and β-adrenergic signaling pathways (*ADAMTS3* and *COX18*). Functional assays revealed that the BDR-associated SNP within *NFKB1* is in linkage disequilibrium with SNPs in a functionally active enhancer and is also associated with the expression of a neighboring gene *SLC39A8*. Our study expands the understanding of pharmacogenetic analyses in racially and ethnically diverse populations and advances the foundation for precision medicine in at-risk and understudied minority populations.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Asthma is a chronic inflammatory disorder of the airways characterized by recurrent respiratory symptoms and reversible airway obstruction. Asthma affects 5% of the world population[1] and is the most common chronic diseases among children[2, 3]. In the United States (U.S.), asthma is the most racially disparate health condition among common diseases[4, 5]. Specifically, U.S. asthma prevalence is highest among Puerto Ricans (36.5%), intermediate among African Americans (13.0%) and European Americans (12.1%), and lowest among Mexican Americans (7.5%)[6]. These disparities also extend to asthma mortality, which is four-fold higher in Puerto Ricans and African Americans compared to European and Mexican Americans[7].

Current asthma guidelines recommend inhaled β_2-agonists (e.g., albuterol) for treatment of acute asthma symptoms. Albuterol is a short-acting β_2-adrenergic receptor (β_2AR) agonist and it produces bronchodilation by causing rapid smooth muscle relaxation in the airways. Albuterol is the most commonly prescribed asthma medication in the world and is the mainstay of acute asthma management across all ethnic groups[8, 9]. Among low income and minority populations in the U.S., albuterol is often the only medication used for asthma regardless of asthma severity[10, 11]. Response to albuterol is quantified based on bronchodilator drug response (BDR) using spirometry. We and others have demonstrated that there is significant variability in BDR among individuals and between populations[12, 13]. Specifically, the populations with the highest asthma prevalence and mortality also have the lowest drug response to albuterol: Puerto Rican and African American children have

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significantly lower BDR than European and Mexican American children[13, 14]. This variation in drug response across ethnic groups may contribute to the observed disparities in asthma morbidity and mortality[15-19].

BDR is a complex trait, influenced by environmental and genetic factors, with heritability estimates ranging from 47% to 92%[20-22]. Genome-wide association studies (GWAS) have identified several common single nucleotide polymorphisms (SNPs) associated with BDR in populations of European descent[23-25]. To date, only one GWAS of BDR has been conducted among African Americans[26]. While that study identified a novel genic association with BDR, it did not replicate known associations discovered in populations of European descent, suggesting that BDR may be determined in part by population-specific variants. Our previous study of genetic predictors of BDR in Latino populations identified a significant contribution of population-specific rare variants to BDR[27].

GWAS studies were designed to identify common variants associated with disease through the use of genotyping arrays that relied on linkage disequilibrium to tag/represent variants not explicitly genotyped on the array itself. Early GWAS arrays were optimized for performance in populations of European origin and lacked the ability to capture race-/ethnic-specific genetic variation due to differences in linkage disequilibrium (LD) across racially/ethnically diverse populations[28]. Recent generations of arrays have attempted to tailor genotyping panels for major HapMap populations (Affymetrix Axiom® World Arrays[29]), or to design array that includes population-specific and trans-ethnic tag SNPs derived from 1000 Genomes Project

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and other consortia studying admixed populations to statistically infer genotypes not directly genotyped in diverse populations (Illumina Infinium® Multi-Ethnic Genotyping Array). However, imputation accuracy decreases significantly with variant frequency[30, 31], making it difficult to use genotyping arrays to study rare and/or population-specific variants. The most striking weakness of GWAS studies was their inability to adequately capture rare variation.

Whole exome sequencing (WES), and other forms of targeted sequencing, were developed to address the inability of genotyping arrays to capture rare variation. WES only allows for the capture of common and rare variants within coding and flaking regions, but studies have shown that a large number of variants associated with complex disease lie within non-coding regions of the genome (reviewed in Zhang and Lupski, 2015[32]). Also, the target capture procedures result in uneven sequence coverage, limiting the reliability of SNP calling for loci close to the boundary of targeted regions. WES also has limited usage for the detection of structure variation, which depends heavily on uniform coverage across the genome.

Whole genome sequencing (WGS) is the ideal technology for identifying disease-causing variants that are rare and/or population-specific. Unlike GWAS genotyping arrays or targeted sequencing technologies, WGS allows the detection of both common and rare variants in both coding and non-coding regions. WGS is the only technology capable of a truly comprehensive and agnostic evaluation of genetic sequence variation in the context of complex disease. The persistent lack of large-scale genetic studies conducted in populations of non-European descent further

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exacerbates racial/ethnic disparities in clinical and biomedical research[33-35]. The application of WGS to the evaluation of genetic factors within a racially/ethnically diverse study population is a necessary step towards eliminating health disparities in BDR and other complex phenotypes.

In this study, we performed WGS on 1,441 minority children with asthma from the tails of the BDR distribution. Our study included high and low drug responders from three ethnic groups: Puerto Ricans (PR) (n=483), Mexicans (MX) (n=483), and African Americans (AF) (n=475). We identified multiple BDR-associated common and rare variants that are population-specific or shared among populations. This study is part of the National Heart, Lung, and Blood Institute's Trans-Omics for Precision Medicine Whole Genome Sequencing (TOPMed) program and represents the largest WGS study thus far to investigate genetic variants important for bronchodilator drug response in racially and ethnically diverse children with asthma.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

Descriptive characteristics of study subjects

Descriptive characteristics for all study subjects (n=1,441, including 483 Puerto Ricans, 483 Mexicans and 475 African Americans) are summarized in **Table 1**. Covariates and demographic variables were assessed for significant differences between high and low drug responders for each racial/ethnic group. Significant differences were only found for age (Mexicans, $p<0.001$) and baseline lung function (pre-FEV_1 % predicted, $p<0.001$), defined as the percentage of observed FEV_1 relative to the expected population average FEV_1 estimated using the Hankinson lung function prediction equations[36].

We estimated genetic ancestry for all participants (see **Methods**) and found that the largest ancestry proportions in Puerto Ricans, Mexicans and African Americans are European, Native American and African ancestries, respectively (**Table 1**, **Supplementary Fig. 1**). Analysis of genetic substructure of the three admixed populations by principal component analysis (PCA) demonstrated that the three populations displayed the characteristic spectrum of ancestry found in admixed populations (**Supplementary Fig. 2**).

Variant summary statistics

Genetic variant summary statistics revealed that the average number of variants by population corresponded to the proportion of African ancestry: the most variants

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were found among African Americans, followed by Puerto Ricans and Mexicans (**Fig. 1a**, **Table 2**).

The majority of observed variants (>90%) were SNPs. The union of biallelic SNPs from all individuals in each population varied from 28.1M among Mexicans, 29.2M among Puerto Ricans to 36.3M among African Americans. Approximately 65% of biallelic SNPs were rare (non-reference allele frequency < 1%, **Fig. 1b**, **Table 2**). Biallelic SNPs that were population-specific (i.e., the SNPs were only found in one population) accounted for 23% (6.68M / 29.2M in Puerto Ricans) to 39% (14.1M / 36.3M in African Americans) of the biallelic SNPs observed in each population. Over 99% of the population-specific SNPs had a non-reference allele frequency less than 5% and a majority of these population-specific SNPs (69% to 82%) were also singletons (**Fig. 1c**, **Table 2**). Based on dbSNP build 147, an average of 20% of biallelic SNPs were novel (**Fig. 1d**, **Table 2**).

In all three populations, 99% of the biallelic SNPs were observed in noncoding regions. Based on the Combined Annotation Dependent Depletion (CADD) score[37], which estimates a variant's deleteriousness, over 99% of highly deleterious biallelic SNPs (CADD score ≥ 25) were observed in coding regions, regardless of ethnicity (**Table 3**). This may be due to the relatively limited availability of functional annotations as positive training data for CADD to estimate deleteriousness in non-coding regions[38]. The percentage of singletons in these highly deleterious biallelic SNPs varied from 51% (Puerto Ricans) to 70% (Mexicans).

BDR association testing with common variants

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We performed genome-wide association testing of common variants with BDR (dichotomized as high/low drug responders from the extremes of BDR distribution) for each population, adjusting by age, sex, body mass index (BMI) categories, and the first ten principal components (PCs) (see **Methods** section "Single locus BDR association testing on common variants" for rationale on including these covariates). We then performed a trans-ethnic meta-analysis on these results across all three populations. A universal *p*-value threshold of 5.00×10^{-8} is often used to determine significance in GWAS studies. This statistical threshold was calculated based on the Bonferroni correction under the assumption of 1,000,000 independent tests and may not be appropriate for WGS studies. It has also been shown, using both GWAS and WGS data, that the number of effective independent tests varies by LD patterns, which in turn vary by race/ethnicity[39]. We calculated the effective number of independent tests for each population, and for our trans-ethnic meta-analysis, and generated appropriately adjusted genome-wide significance thresholds (see **Methods**). Population-specific genome-wide significance thresholds after correcting for the number of effective tests (adjusted genome-wide significance level) were 1.57×10^{-7} for Puerto Ricans, 2.42×10^{-7} for Mexicans, and 9.59×10^{-8} for African Americans (see **Methods**). These numbers are highly concordant with WGS significance thresholds derived from the African America (ASW), Mexican (MXL), and Puerto Rican (PUR) 1000 Genomes sequencing data[39]. For our trans-ethnic meta-analysis, the adjusted genome-wide significance level was 3.53×10^{-7}.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Two SNPs, rs17834628 and rs35661809, located on chromosome 5 were significantly associated with BDR (p = 1.18 x 10^{-8} and 3.33 x 10^{-8}, respectively, **Fig. 2a and Table 4**). **Figure 2b** shows that the direction of effect of the top associated SNP (rs17834628) is the same in all three populations. **Figure 2c** displays a LocusZoom plot of rs17834628 with 400kb flanking regions. We attempted to replicate this SNP in our replication cohorts (GALA I, SAGE I, HPR, SAPPHIRE and CHOP) separately and via meta-analysis. However, the SNP did not show significant association with BDR (replication $p > 0.05$).

It has been shown that functionally relevant variants do not always display the lowest p-values in association studies[40]. As this is the first study of this scope to be performed in diverse study populations, we decided to also investigate "suggestive" loci ($p \leq 7.06 \times 10^{-6}$, see **Methods** for definition of suggestive significance level) to identify variants with functional relevance. After merging SNPs with high LD ($r^2 \geq 0.8$) into unique loci, 11 other unique loci were identified at a suggestive significance level ($p \leq 7.06 \times 10^{-6}$) (**Fig. 2a**). The top 10 unique loci (represented by 27 SNPs) and their nearest genes are listed in **Table 4**, with the most significant SNP (lowest p-value) for each locus marked by asterisk. All but three of these SNPs are intergenic; two SNPs mapped to introns 3 and 5 of *NFKB1* (chromosome 4, rs28450894 and rs4648135), and another mapped to intron 7 of *PLCB1* (chromosome 20, rs16995064).

Among the *NFKB1* SNPs, the low BDR-associated T allele of rs28450894 is found predominantly among African populations (minor allele frequency [MAF] 8.8% – 28.7%), followed by European populations (MAF 3.7% - 7.6%) and Puerto Ricans

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

313 (MAF 6.2%), and is relatively rare in Mexicans (MAF 1.5%) based on 1000 Genomes
314 data (**Supplementary Fig. 3**). Using ChIP-Seq, we identified potential regulatory
315 regions marked by H3K27ac peaks in primary bronchial smooth muscle cells
316 (BSMCs). We found that SNPs in moderate to high LD (r^2 = 0.47 to 0.82) with
317 rs28450894 overlapped with H3K27ac ChIP-Seq peaks identified in BSMCs, implying
318 that they may have regulatory functions (**Supplementary Fig. 4a**). The regulatory
319 function of two H3K27ac peak regions overlapping variants in LD ($r^2 \geq 0.8$) with
320 rs28450894 was then functionally validated using luciferase enhancer assays. These
321 sequences were cloned into the pGL4.23 enhancer assay vector (Promega), which
322 contains a minimal promoter and a luciferase reporter gene. The ApoE liver enhancer
323 was used as a positive control, and pGL4.23 empty vector as a negative control. All
324 constructs were tested for their enhancer activity in BSMCs. One enhancer, *NFKB1*
325 Region 2, showed significantly increased enhancer activity over empty vector (2.24
326 fold increase, $p = 8.70 \times 10^{-6}$, unpaired t-test; **Supplementary Fig. 4b**).
327 Given the relevance of *NFKB1* in immune pathways and asthma, we used RNA
328 sequencing (RNA-seq) to verify whether the identified intronic *NFKB1* SNPs regulate
329 gene expression of neighboring genes. Among genes within 1Mb of rs28450894
330 meeting expression reliability cutoffs (see **Methods**), we found that the low BDR-
331 associated T allele of rs28450894 is significantly associated with decreased
332 expression of *SLC39A8* in blood (**Supplementary Fig. 5**, $p = 0.0066$, FDR-adjusted p
333 = 0.0856, $\log_2(\beta) = -0.327$).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We did not find SNPs associated with BDR at genome-wide significance levels in any of the population-specific analyses (**Supplementary Fig. 6**).

Previously reported GWAS-based BDR-associated common variants in GALA II did not replicate in the current study[27]. To distinguish whether the discrepancy between findings was due to data type (imputed array-based vs. WGS-based) or study design (continuous trait vs. extreme phenotype), the common variant analysis in the current analysis was repeated among the subset of samples that had both array-based and WGS data (n = 1,414 out of 1,441). Based on the top 1000 BDR-associated SNPs, there was a high correlation between association *p*-values generated from imputed array-based and WGS-based genotypes (Spearman correlation = 1.0), indicating that data type is not the cause of the observed discrepancy (**Supplementary Fig. 7a**). Nearly all SNPs with high imputation r^2 exhibited low genotype discordance between array-based and WGS-based genotypes, confirming high imputation quality of most common SNPs (≥ 99.7%). (**Supplementary Fig. 7b and 7c**). Using imputed array-based data, we also performed linear regression using BDR (ΔFEV_1) as a continuous trait. Both of the genome-wide significant BDR-associated SNPs have the same direction of effect in the meta-analysis (β=0.51) as the extreme phenotype study design (OR=1.67, **Table 4**), although the *p*-value was not genome-wide significant. This indicates that the discrepancy between findings may be due to the power of different study designs (continuous trait vs. extreme phenotype) for individual SNP to reach genome-wide significant level.

BDR association testing using rare variants

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We tested the combined effects of rare variants on BDR using SKAT-O[41] by grouping common and rare variants into 1kb sliding windows, which moved across the genome in 500bp increments. The same covariates used for common variant association testing were applied.

After identifying the effective number of tests and adjusting for multiple comparisons on each population separately (see **Methods**), we identified three population-specific loci associated with BDR at genome-wide significance levels ($p \leq 8.15 \times 10^{-8}$ for Puerto Ricans, $p \leq 8.60 \times 10^{-8}$ for Mexicans and $p \leq 6.94 \times 10^{-8}$ for African Americans) (**Fig. 3a-c**). The associated windows in these loci and their nearest genes are listed in **Table 5**.

We also performed association testing using subjects across all three populations in a single analysis. Pooling subjects increased the sample size and thereby maximized the power of the association test. To minimize any potential effect of confounding by population substructure, we included local genetic ancestry, defined as the proportions of Native American and African ancestries for the window under testing, as additional covariates in the association test. Two loci on chromosomes 4 and 8 were found to be genome-wide significant ($p \leq 1.53 \times 10^{-7}$) (**Fig. 3d, Table 5**).

We examined alternative grouping strategies for rare variants, including grouping (1) by genes from transcription start to end sites with or without 50kb flanking regions, (2) by transcription start site with 20kb flanking regions, and (3) by H3K27ac peaks from chromatin immunoprecipitation sequencing analysis (ChIP-Seq) on

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

377 airway epithelial cells and airway smooth muscle cells. Association tests with these

378 alternate grouping strategies identified no further significant associations.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

In this study, we have identified population-specific and shared common and rare variants that are associated with bronchodilator drug response in three ethnically diverse populations of children with asthma. WGS, unlike GWAS genotyping arrays and targeted sequencing, provides a comprehensive detection of both common and rare variants in both coding and non-coding regions. African Americans, Latinos, and other minorities have been dramatically underrepresented in GWAS studies[33-35]. The present study presents an important initiative and investment from the NIH/NHLBI to include underrepresented populations in large whole genome sequencing efforts and to improve racial/ethnic diversity in clinical and biomedical research.

Our trans-ethnic common variants meta-analysis identified one locus on chromosome 5 that was associated with BDR at a genome-wide significance level ($p < 5.00 \times 10^{-8}$). The proximity of this BDR-associated locus to *DNAH5* and *LINC01194* is of particular interest. A SNP in *DNAH5* has been associated with total lung capacity in white subjects with chronic obstructive pulmonary disease[42]. In a separate GWAS, the *DNAH5*/*LINC01194* locus was reported among Europeans to be associated with levels of IgE[43, 44], a biomarker associated with asthma endotypes. Baseline lung function (FEV_1) and total IgE levels are associated with asthma severity and can predispose an individual to lower bronchodilator drug responsiveness[13, 14, 45]. Two *NFKB1* intronic variants on chromosome 4 were associated with BDR at a suggestive significance level. The NFκB protein has a known role in allergic response, and various studies have demonstrated that the NFκB pathway is activated in patients

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with asthma, as reviewed by Edwards *et al.*[46]. Bronchodilator drug responsiveness may thus be modulated by an individual's baseline airway allergic status. ChIP-seq and functional enhancer assays in BSMCs suggest these *NFKB1* intronic variants may regulate expression of nearby genes. This was in fact supported by our RNA-seq data, which showed that individuals with the low BDR associated genotype displayed reduced expression of the neighboring gene *SLC39A8*, a gene which has previously been found to be responsive to cytokine treatment in airway epithelial cells[47] and had reduced expression in mice with allergic airway inflammation[48]. Recent studies have also shown that SLC39A8 is unique among other zinc transporters in that upregulation of *SLC39A8* is sufficient to protect lung epithelia against TNF-α-induced cytotoxicity[49]. Additionally, the higher frequency of the low BDR-associated allele (T allele of rs28450894 in *NFKB1*) in African populations suggests that the low BDR-associated allele tracks with African ancestry. This may explain why admixed populations with higher proportions of African ancestry, i.e., African Americans and Puerto Ricans, have lower bronchodilator drug responsiveness, and by extension may shed light on the higher asthma morbidity and mortality in these populations[14]. Another intronic variant (chromosome 20, rs16995064, *PLCB1* intron 7) was associated with BDR at a suggestive significance level. *PLCB1* is highly relevant, as this gene has been reported to be differentially expressed in therapy-resistant childhood asthma compared to controlled persistent asthma or age-matched healthy control subjects in a Swedish cohort[50]. Functional studies also reported that silencing *PLCB1* inhibited the effect of lipopolysaccharide-induced endothelial cell inflammation

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

423 through inhibiting expression of proinflammatory cytokines[51]. Further functional
424 studies are necessary to establish the role of *NFKB1* and *PLCB1* on BDR.

425 Apart from assessing the individual effect of common variants on BDR, we also
426 identified various combined effects of rare variants that were population-specific or
427 shared across populations. This includes the chromosome 1 locus in intron 9 of *MAGI3*
428 and the chromosome 11 intergenic locus between *LOC105376671* and *LIN7C* for
429 Mexicans, the chromosome 19 intron 3 locus in *FDX1L* for African Americans, and the
430 chromosome 4 intergenic locus between *ADAMTS3* and *COX18* and the chromosome
431 8 locus in intron 4 of *CPQ* shared across populations.

432 While some of the genes are uncharacterized or have no known function related to
433 BDR (*MAGI3*, *LOC105376671*, *LIN7C*, *FDX1L* and *CPQ*), there appears to be functional
434 relevance for the locus between *ADAMTS3* and *COX18*. The *ADAMTS3* and *COX18* loci
435 were associated with β-adrenergic responses in cardiovascular-related traits in
436 mice[52]. This locus was significantly associated with cardiac atrial weight in mice
437 treated with the β blocker atenolol; the association also replicated in mice treated
438 with the β agonist isoproterenol. These findings suggest that SNPs found in this locus
439 may modify adrenergic signaling pathways in BDR. In the present study, we also
440 identified BDR association with rare variants within the *CPQ* gene, which encodes a
441 protein from the carboxypeptidase family. Although no previous association has been
442 identified for CPQ, another member of the carboxypeptidase family,
443 carboxypeptidase A3 (CPA3), is known to be expressed at higher levels in the airway

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

444 epithelium among subjects with T_H2-high asthma[53, 54]. Further studies are necessary
445 to determine the role of CPQ in BDR.

446 GWAS-based BDR-associated common variants in GALA II have previously been
447 reported[27]. However, these variants did not replicate in the current study, likely due
448 to two major differences between the previous and current investigations: study
449 design and data type. The previous BDR GWAS used an array-based genotyping panel
450 to examine children with asthma across the entire BDR spectrum, i.e., BDR (ΔFEV_1)
451 was used as a continuous variable. In contrast, the current study sequenced the entire
452 genome to investigate only the extremes of the BDR (ΔFEV_1) distribution (i.e.,
453 high/low drug responders). By repeating our current analysis using a subset of
454 individuals that had both array and WGS data, we confirmed that the major
455 discrepancy between the two studies is due to study design instead of differences in
456 data type. The contrast in results between GWAS and WGS due to differences in study
457 design implies that varied study designs are necessary for a comprehensive
458 understanding of variants associated with asthma-related phenotypes and drug
459 response. Studying samples from the extreme tail of drug response distributions has
460 been recognized as one of the success factors in the study design of pharmacogenomic
461 GWAS[55]. Furthermore, it was recently demonstrated that the power gain from
462 studying extreme phenotypes is much greater in rare variant studies[56]. Since cost is
463 often a limiting factor for WGS studies, choosing an extreme phenotypic study design
464 may be beneficial for the study of rare variants and the discovery of common variant

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associations that may otherwise be missed when sampling across the entire phenotypic spectrum.

In this study, we did not identify BDR-associated variants from β_2AR signaling pathways. Instead, most of the BDR-associated genes identified in this study are related to lung function and allergic response, including total IgE levels and cytokine production in mast cells. This suggests that at least part of BDR may be due to the predisposition or intrinsic state of airway smooth muscle cells. Individuals may have different intrinsic expression levels of candidate genes determined by genetic variants, which determine whether their response to albuterol is beneficial.

A higher percentage of African ancestry often implies a higher degree of genetic variation[57]. Although Puerto Ricans have higher proportions of African ancestry than Mexicans (**Table 1**), they have fewer population-specific SNPs, an observation that is consistent with findings from the 1000 Genomes Project[58]. This is likely due to the fact that our Puerto Rican subjects, who were mostly recruited from Puerto Rico, have gone through recent population bottlenecks. We have demonstrated that Puerto Ricans may be more genetically related than expected[59], suggesting that our current relatedness filters may be too conservative for Puerto Ricans.

Including admixed populations in whole genome sequencing studies has important scientific implications. First, it allows for discovery of genetic variation of multiple ancestral populations in a single study. Second, it is extremely useful to study admixed populations with ancestries that are currently underrepresented in existing genetic repositories. For example, the widely popular PCSK9 inhibitors used to treat

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hypercholesterolemia were discovered by studying the genetics of African Americans but the biology and final drug development have benefited all patients regardless of race/ethnicity[60]. On a genetic research level, studying admixed populations such as Mexicans will enhance the understanding of genetic variation in Native American ancestry, an area that is currently lacking in all major sequencing efforts.

Although an extensive effort was made to replicate the top BDR-associated variants, we were unable to replicate our results because few studies of non-European populations exist, as we and others have documented[33-35, 61]. Our efforts to perform replication of rare BDR-associated variants were further hindered by the lack of studies with whole genome sequencing data. These challenges highlight the need to include more racially/ethnically diverse populations in all clinical and biomedical research.

In an era of precision medicine, addressing questions about the impact of genetic factors on therapeutic drug response in globally diverse populations is essential for making precision medicine socially and scientifically precise[4]. This study advances our understanding of genetic analysis in admixed populations and may play an important role in advancing the foundation of precision medicine for understudied and racial/ethnically diverse populations.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ONLINE METHODS

Data availability

TOPMed whole genome sequencing data is available to download by submitting data access request through dbGaP. The dbGaP study of GALA II and SAGE are phs000920.v1.p1 and phs000921.v1.p1.

Study cohorts and sample details

This study examined a subset of subjects with asthma from the Study of African Americans, Asthma, Genes & Environments (SAGE II)[62-65] and the Genes-Environments & Admixture in Latino Americans (GALA II) study[27]. SAGE II recruited African American subjects from the San Francisco Bay area. GALA II recruited Latino subjects from Puerto Rico and the mainland United States (Bronx, NY; Chicago, IL; Houston, TX; San Francisco Bay Area, CA). Ethnicity of the subjects was self-reported and all four of the participant's biological grandparents must have reported the same ethnicity.

A total of 1,484 individuals from three ethnic groups (494 Puerto Ricans, 500 Mexicans and 490 African Americans), representing the extremes of the bronchodilator response (BDR, see below) distribution were selected for whole genome sequencing. Genomic DNA was extracted and purified from whole blood using Wizard® Genomic DNA Purification Kits (Promega, Fitchburg, WI).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bronchodilator response measurements

Spirometry was performed and BDR (ΔFEV_1) was calculated as previously described[27]. In brief, BDR was calculated as the percent change in FEV_1 after 2 doses of albuterol (post-FEV_1) compared with baseline values before administration of albuterol (pre-FEV_1).

Analysis on descriptive data of study subjects

Dichotomous variables were tested for association with BDR using Fisher's exact test. Continuous variables were tested for normality using the Shapiro-Wilk test. Normally and non-normally distributed continuous variables were tested using Student's t-test and the Wilcoxon rank sum test, respectively.

Sample quality control and whole genome sequencing

DNA samples were quantified by fluorescence using the Quant-iT PicoGreen dsDNA assay (ThermoFisher Scientific, Waltham, MA, USA) on a Spectramax fluorometer (Molecular Devices, Sunnyvale, CA, USA). Sample integrity was ascertained using the Fragment Analyzer™ (Advanced Analytical Technologies, Inc., Ankeny, IA, USA). Samples passing QC were genotyped using the HumanCoreExome-24 array (Illumina®, San Diego, CA, USA). Genotyping results were analyzed using VerifyIDintensity[66] to flag sample contamination. Sequencing libraries were constructed using the TruSeq PCR-free DNA HT Library Preparation Kit (Illumina®, San Diego, CA, USA) with 500ng DNA input. Briefly, genomic DNA was sheared using a Covaris sonicator (Covaris, Woburn, MA), followed by end-repair and bead-based

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

size selection of fragmented molecules. Selected fragments were then A-tailed, and sequence adaptors were ligated onto the fragments, followed by a final bead purification of the libraries. Final libraries were reviewed for size distribution using Fragment Analyzer and quantified by qPCR (Kapa Biosystems, Wilmington, MA, USA). Libraries were sequenced on HiSeq X system (Illumina®, San Diego, CA, USA) with v2 chemistry, using a paired-end read length of 150 bp, to a minimum of 30x mean genome coverage.

WGS data processing and quality control

Sequencing data were demultiplexed using bcl2fastq version 2.16.0.10 (Illumina), and aligned to the human reference hs37d5 with decoy sequences using BWA-MEM v0.7.8[67]. Data were further processed using the GATK best-practices v3.2-2 pipeline[68]. Quality control procedures included marking of duplicate reads using Picard tools v1.83 (http://picard.sourceforge.net), realignment around indels, and base quality recalibration using 1000 Genomes Phase 1 high confidence SNPs, HapMap v3.3, dbSNP v137, 1000 Genomes omni2.5, 1000 Genomes Phase 1 indels, and, Mills and 1000 Genomes gold standard indels. Single-sample genotypes were called using GATK HaplotypeCaller followed by joint genotyping of all subjects. The resulting multi-sample Variant Call Format (VCF) file was used for variant quality score recalibration (VQSR). A 99.8% truth sensitivity tranche level was used for SNPs and 99.0% for indel variants. SNP calls were used to check for sample contamination using VerifyBAMId[66], and sample identity was confirmed by requiring > 99.5% concordance with SNP array (HumanCoreExome-24 array) genotypes.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As part of NIH's Trans-Omics for Precision Medicine (TOPMed) Program, BAM files were submitted to the Informatics Resource Center (IRC) at the University of Michigan. All 1,484 samples sequenced passed TOPMed's IRC quality control metrics (mean genome coverage >30x; >95% of genome covered at >10x; and <3% contamination).

VCF-level variants were filtered by GATK version 3.4.46 and VCFtools version 0.1.14[69]. Variants were filtered according to the following procedures: (1) remove variants that were not indicated as "PASS" in the VCF FILTER column, (2) remove variants in low complexity regions[70] (downloaded from https://github.com/lh3/varcmp/tree/master/scripts/LCR-hs37d5.bed.gz), and (3) keep sample genotypes that have minimum read depths of 10 and genotype qualities of 20 (DP ≥ 10 and GQ ≥ 20). The ratio of heterozygous to homozygous variants (hom/het), ratio of transitions to transversions (Ti/Tv), and other variant summary statistics were generated using GATK VariantEval. VCF files were converted into PLINK format using PLINK v1.9 software[71] according to recommended best practices[72]. Genotype consistency between WGS data and previously published Axiom® Genome-Wide LAT 1 array (Affymetrix, Santa Clara, CA) genotype data (dbGaP phs000920.v1.p1 and phs000921.v1.p1) were assessed using VCFtools[69]. Individuals with percentage consistency three S.D. below the mean (< 96.3%) were removed (n=7, **Supplementary Fig. 8**). Cryptic relatedness was detected using REAP[73]. Global ancestry and allele frequency used by REAP were estimated using ADMIXTURE in supervised mode[74]. Related individuals (kinship coefficient > 0.044, corresponding to

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

589 a third degree relationship[75]) were excluded from further analysis (n=36), yielding a
590 final sample size of 1,441 for downstream analysis. Downstream analyses were only
591 performed on biallelic SNPs that passed all quality filters mentioned above and had
592 less than 10% of genotype missingness. The 10% genotype missingness filter was
593 applied per population instead of across all three populations except for the rare
594 variant analysis performed with all three populations combined (see **Methods**
595 section "Multi-variant analyses of combined effects of rare variants on BDR").

596 **Principal component analysis**

597 Principal component analysis (PCA) was performed to control for hidden population
598 substructure using EIGENSTRAT's smartpca program[76]. After removing biallelic SNPs
599 with low minor allele frequency (MAF ≤ 0.05) and in linkage disequilibrium ($r^2 > 0.5$
600 in a 50-SNP window with a shift size of 5 SNPs) using PLINK1.9, 710,256 variants
601 were used as input to smartpca.

602 **Local ancestry estimation**

603 Reference genotypes for European and African ancestries were obtained from the
604 Axiom® Genotype Data Set (http://www.affymetrix.com/support/technical/
605 sample_data/axiom_db/axiomdb_data.affx) and SNPs with less than a 95% call rate
606 were removed. Since no Native American reference samples are available in the
607 HapMap database, reference genotypes for Native American ancestry were generated
608 from 71 Native American individuals previously genotyped on the Axiom® Genome-
609 Wide LAT 1 array[27].

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To call local ancestry tracts, we first created a subset of our WGS data corresponding to sites found on the Axiom® Genome-Wide LAT 1 array, leaving 765,321 markers. Using PLINK1.9, we merged these data with our European (CEU), African (YRI), and Native American (NAM) reference panels, which overlapped at 434,145 markers. After filtering multi-allelic SNPs and SNPs with > 10% missing data, we obtained a final merged dataset of 428,644 markers. We phased all samples using SHAPEIT2[77] and called local ancestry tracts jointly with RFMix[78] under a three-way admixture model based on the African, European, and Native American reference genotypes described above.

Variant annotation

TOPMed freeze 2 and 3 variants were annotated using the WGSA annotation pipeline[79]. Annotated VCF files were downloaded from the TOPMed Data Coordinating Center SFTP sites.

Single locus BDR association testing on common variants

An additive logistic regression model was used to evaluate the association of biallelic common variants (MAF > 1%) with BDR using PLINK 1.9 separately for each population. Throughout this study, high drug responders were assigned as cases. Logistic regression models included the covariates age, sex and body mass index (BMI) categories to account for previously reported confounders of asthma and BDR[80-87], and the first ten principal components (PCs) to correct for population substructure in admixed populations. BMI and age- and sex-specific BMI percentiles

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(BMI-pct) were calculated as previously described[65] and used for assignment to BMI categories. For subjects aged 20 years and over, BMI categories were defined as follows: underweight (BMI < 18), normal (18 ≤ BMI < 25), overweight (25 ≤ BMI < 30) and obese (BMI ≥ 30). For subjects under 20 years of age, BMI categories were defined as follows: underweight (BMI-pct < 5), normal (5 ≤ BMI-pct < 85), overweight (85 ≤ BMI-pct < 95) and obese (BMI-pct ≥ 95). Baseline lung function (pre-FEV_1) has a significant impact on how much further a participant can respond to albuterol. Nevertheless, we did not explicitly include pre-FEV_1 as a covariate since the variation in pre-FEV_1 was captured by the ten PCs already included in the model (**Supplementary Table 1**). In addition, the BDR (ΔFEV_1) calculation has already adjusted for pre-FEV1 (see **Methods** section "Bronchodilator response measurements"). Correlation and test for association between pre-FEV_1 and PCs were based on Spearman correlation. Population-specific genome-wide significance thresholds for the single locus analyses were calculated based on genotypes using the autocorrelation-based *effective size()* function in the R package '*coda'* as published by Sobota *et al.*[39]. Population-specific genome-wide significance thresholds after adjusting for the effective number of effective tests (adjusted genome-wide significance) were 1.57×10^{-7} for Puerto Ricans, 2.42×10^{-7} for Mexicans, and 9.59×10^{-8} for African Americans. Suggestive significance thresholds were calculated by one divided by the effective number of tests[88]. The local linkage disequilibrium pattern (Genome build: hg19/1000 Genomes Nov 2014 AMR) of the flanking regions of BDR-associated SNPs was visualized using LocusZoom[89]. Quantile-quantile (q-q) plots

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were generated using a uniform distribution as the expected *p*-value distribution (**Supplementary Fig. 9a-c**). The genomic inflation factor (λ_{GC}) was calculated using the R package *'gap'*.

Trans-ethnic meta-analysis of common variant effects on BDR

A meta-analysis of the effects of common variants on BDR across the three populations was performed using METASOFT[90]. We used the Han and Eskin's random effects model optimized for detecting associations under heterogeneous genetic effects from different study conditions[90]. The number of effective tests was estimated using the R package *'coda*' as described above, yielding an adjusted genome-wide significance threshold of 3.53×10^{-7} and a suggestive significance threshold of 7.06×10^{-6}. Suggestive significance thresholds were calculated by one divided by the effective number of tests[88]. Allele frequency variation in the world population was visualized using the Geography of Genetic Variants Browser (GGV) beta v0.92 (http://www.popgen.uchicago.edu/ggv, accessed 9/30/2016). The q-q plot and λ_{GC} were generated in the same way as described above (**Supplementary Fig. 9d**, see **Methods** section, "Single locus BDR association testing on common variants").

Multi-variant analyses of combined effects of rare variants on BDR

Combined effects of rare variants on BDR was analyzed using SKAT-O[41]. Rare variants were collapsed into 1kb windows sliding across the GRCh37 genome in steps of 500 base pairs. A total of 5.3 million windows were analyzed and the R package *'coda'* was used to determine the number of effective tests using autocorrelation of the

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

association *p*-value, as described above. Adjusted genome-wide significance thresholds for Puerto Ricans (8.15×10^{-8}), Mexicans (8.60×10^{-8}), African Americans (6.94×10^{-8}) and for all three populations combined (1.53×10^{-7}) were used to determine windows of variants with genome-wide significant association with BDR. The same covariates used for common variant association testing were used for analyses of individual populations. For analyses of individuals combined across all three populations, we avoided confounding from population substructure by including local ancestry as additional covariates, defined as the proportions of Native American and African ancestries for the window under testing. The q-q plots and λ_{GC} were generated in the same way as described above (**Supplementary Fig. 10**, see **Methods** section, "Single locus BDR association testing on common variants").

Single locus BDR association and trans-ethnic meta-analysis of array data

To address the discrepancy between our current common variant analysis results with previously published BDR GWAS results[27], we used 1,414 of the 1,441 individuals who had both Axiom® Genome-Wide LAT 1 array (see **Methods** section, "WGS data processing and QC") and WGS data available to rerun the single locus BDR association testing and trans-ethnic meta-analysis. Array data were imputed to the Haplotype Reference Consortium[91] (HRC release 1) panel using the Michigan Imputation Server[92]. We used the top 1000 BDR-associated SNPs to examine the relationship between array-based and WGS-based association *p*-values, genotype discordance and imputation r^2. Correlation between the array-based and WGS-based

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

association *p*-values was determined by Spearman correlation. We also performed single locus BDR association testing and trans-ethnic meta-analysis by applying linear regression on 1,122 Puerto Ricans, 662 Mexicans and 1,105 African Americans using BDR (ΔFEV_1) as a continuous trait. HRC imputed array-based data and the same covariates as described above were used for the analysis.

Replication of top BDR-associated common variants

Replication cohorts included the Genetics of Asthma in Latino Americans Study (GALA I), the Study of African Americans, Asthma, Genes & Environments (SAGE I), a case-control study of childhood asthma in Puerto Ricans (HPR)[93], the Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE)[94] and the cohort from the Children's Hospital of Philadelphia (CHOP)[95].

Association with BDR was tested by two different approaches. We first treated BDR as a continuous outcome and used linear regression to assess the effect of each SNP on BDR. The second approach dichotomized BDR based on the cutoffs observed in the current study to identify high responders (ΔFEV_1 > 16.29 for Puerto Ricans, > 8.55 for Mexicans and > 11.81 for African Americans) and low resonders (ΔFEV_1 < 7.23 for Puerto Ricans, < 6.05 for Mexicans and < 5.53 for African Americans); logistic regression was used to test the association between high/low responder status and each SNP. In both analyses, an additive genetic model was assumed for each SNP tested. SNP associations were adjusted for sex, current age, BMI categories, and the first 10 principal components in both analyses.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

716 The sample size for the GALA I and SAGE I replication cohorts included 172 Puerto
717 Ricans, 253 Mexicans and 191 African Americans with BDR measurements and
718 complete data for all the covariates (age, sex, BMI categories and the first ten PCs).
719 Genotype data were phased with SHAPE-IT[96] and imputation was performed with
720 IMPUTE2[97] using all populations from 1000 Genomes Project Phase 3 as reference[58].
721 In the HPR replication cohort, 523 Puerto Ricans subjects were included. Spirometry
722 data were collected as previously described[98]. Genome-wide genotyping was
723 performed using the Illumina HumanOmni2.5 BeadChip platform (Illumina Inc., San
724 Diego, CA) and processed as previously described[99]. Phasing and imputation of HPR
725 participant genotype data were done as described for GALA I and SAGE I. In the
726 SAPPHIRE replication cohort, 1,325 African American subjects with asthma were
727 included. Genome-wide genotyping was performed using the Axiom® Genome-Wide
728 AFR 1 array (Affymetrix Inc., Santa Clara, CA) as previously described[26]. Genotype
729 data were imputed to the cosmopolitan 1000 Genomes Phase 1 version haplotypes
730 using the Michigan Imputation server. The CHOP replication cohort included 378
731 African Americans. Genotyping was performed as described[95], and genotype data
732 were imputed to HRC panel using the Sanger Imputation server[91]. Dichotomizing BDR
733 yielded 18 high and 107 low responding (18/107) Puerto Ricans, 105/97 Mexicans,
734 and 57/84 African Americans for GALA I and SAGE I, 43/272 Puerto Ricans for HPR,
735 465/577 African Americans for SAPPHIRE, and 155/120 African Americans for
736 CHOP.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identification of nearest genes to BDR-associated loci

The four nearest transcripts to BDR-associated loci were identified by using the "closest" command in BEDTools with the parameters "-d -k 4" and the RefSeq gene annotations (Feb.2009 [GRCh37/hg19] assembly) downloaded in refFlat format from the UCSC Table Browser[100]. Genes with multiple transcripts were reported once only.

Primary bronchial smooth muscle cell culture

Cryopreserved primary human bronchial smooth muscle from two donors (from Lonza catalog number CC-2576, lot number 0000212076 and from ATCC catalog number PCS-130-011, lot number 62326179) were thawed and expanded in Lonza Smooth Muscle Growth Media (SmGM; catalogue number CC-3182) on T75 flasks (E&K Scientific Products, catalog number 658175).

H3K27ac ChIP-seq assay

Once BSMCs reached 80% confluency, BSMCs were serum-starved by replacing SmGM with smooth muscle basal media (SmBM) for 24 hours. After serum starvation, BSMCs were grown in SmBM containing 5% FBS for 4 hours, then fixed in 1% formaldehyde for 10 min and quenched with 0.125 M glycine for 5 minutes. Cells were removed from the T75 flasks by scraping in cold PBS containing sodium butyrate (20 mM, Diagenode, catalog number C12020010). Chromatin sheering was carrying out using a Covaris S2 sonicator. Sheared chromatin was used for immunoprecipitation with antibodies against active chromatin marks (H3K27ac; Abcam, ab4729) using the Diagenode LowCell# ChIP kit (CAT#C01010072), following the manufacture's

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

758 protocol. Libraries were prepared using the Rubicon DNA-Seq kit (CAT#R400406)
759 following the manufacturer's protocol and sequenced on an Illumina HiSeq 4000
760 instrument using single-end 50-bp reads to a sequencing depth of at least 25 million
761 reads (submitted under BioProject PRJNA369271). Uniquely mapping raw reads
762 were aligned using Bowtie[101] with default settings. Peak regions for each individual
763 were called using MACS2[102, 103] and reproducible peaks identified using the ENCODE
764 IDR pipeline[104].

765 **Luciferase assays**

766 *NFKB1* candidate enhancer sequences were amplified from human genomic DNA
767 (Roche) using oligonucleotides designed in Primer3 with 18 and 20 bp overhangs for
768 forward and reverse primers, respectively, (5′-GGCCTAACTGGCCGGTAC -3′ and
769 5′CGCCGAGGCCAGATCTTGAT -3′) complementary to sequences flanking the KpnI
770 and EcoRV sites in the pGL4.23 Gate A vector (Promega) using Phusion High-Fidelity
771 PCR kit (NEB, catalog number M0531S). PCR products where then cloned into the
772 pGL4.23 vector using the Gibson Assembly method (NEB, catalog number E2611S).
773 Smooth muscle cells were plated at 50-70% confluency in 24-well cell culture plates
774 (Falcon, catalog number 353047) and grown to 80% confluency in SmGM.
775 Transfections were carried out by combining polyethyleneimine (PEI) with DNA
776 vectors at a 1:1 ratio by weight in opti-MEM (Life Technologies, catalog number
777 31985070). Transfection mixture consisted of 225 ng of enhancer assay vectors and
778 25ng of pGL4.24 (Renilla transfection efficiency control) with 250 ng of PEI in 50 μL

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

779 of opti-MEM. After 15 minutes, 500 μL of SmBM was added to the transfection
780 mixture and the combination added to cell culture. Cells were incubated for 4 hours
781 in SmBM plus the transfection mixture, then media was replaced with SmGM for 24
782 hours. Cells were then washed with PBS and enhancer assay cells were lysed with 100
783 μL of Passive Lysis Buffer (Promega). Reporter activity was measured using the Dual-
784 Luciferase Reporter Assay System (Promega) and measured on the Glomax 96 well
785 plate luminometer (Promega).

786 **RNA extraction, library preparation and sequencing**

787 Among the African American subjects with WGS data in our study, 39 samples were
788 selected for RNA-seq based on BDR status and the number of copies of low-BDR
789 associated alleles at rs28450894. The number of samples in each category is shown
790 in **Supplementary Table 2**. Peripheral blood samples were collected into PAXgene
791 Blood RNA tubes (PreAnalytiX, Hombrechtikon, Switzerland). Total RNA of 39
792 samples was extracted from PAXgene Blood RNA tubes using MagMAX™ for
793 Stabilized Blood Tubes RNA Isolation Kit (CAT#4451894, Thermo Fisher Scientific,
794 Waltham, MA, USA) according to manufacturer's protocols. RNA integrity and yield
795 were assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara,
796 CA, USA). Globin depletion was performed using GLOBINclear™ kit (CAT#AM1980,
797 Thermo Fisher Scientific, Waltham, MA, USA). Library preparation and ribosomal
798 depletion were performed using KAPA Stranded RNA-Seq Kit with RiboErase
799 (CAT#KK8483, Kapa Biosystems, Wilmington, MA, USA) according to the

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

manufacturer's protocols. Each sample was uniquely barcoded with NEXTflex™ DNA Barcodes (CAT#514104, Bioo Scientific®, Austin, TX, USA). Barcoded libraries were pooled and sequenced on 4 lanes on a HiSeq 4000 sequencing system (Illumina®, San Diego, CA, USA) with a paired-end read length of 100 bp at the University of California, San Francisco's Center for Advanced Technology.

RNA-seq data processing and analysis

Raw sequencing reads were aligned to the human reference genome (hg19) using STAR[105] (v2.4.2a). Gene read counts were obtained from uniquely mapped reads based on Ensembl annotation[106] (v75). Read counts were analyzed for differential gene expression changes between genotypes, including an interaction term with genotype and sex (genotype * sex), with DESeq2[107] using a linear model to account for covariates of sex, age and library prep batch, and a custom model matrix to correct for GC content difference between genes. After normalization for sequencing depth and GC percentage, genes with fewer than an average of five normalized read counts per sample and fewer than 20 samples with at least one read count were removed. This filtering process kept 19,592 Ensembl genes for analysis to obtain fold changes and raw *p*-values for both the genotype and genotype * sex terms. Genes were then further filtered to specifically analyze the locus surrounding rs28450894 for differential gene expression, by only including all genes with a transcriptional start site within 1Mbp of rs28450894. *P*-values were then corrected using the false discovery rate method to account for the 13 genes in this locus.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

821 **REFERENCES**

822 1. Vos, T. *et al*. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases
823 and injuries 1990-2010: a systematic analysis for the Global Burden of Disease
824 Study 2010. *Lancet* **380**, 2163-2196 (2012).

825 2. Asthma [Internet]. World Health Organization. 2013 Nov [last updated 2013 Nov;
826 accessed: 2017 Jan 28]. Available from:
827 http://www.who.int/mediacentre/factsheets/fs307/en/.

828 3. World Health Organization. **Global surveillance, prevention and control of**
829 **chronic respiratory diseases: a comprehensive approach**. (2007). Available
830 from: http://www.who.int/gard/publications/GARD_Manual/en/.

831 4. Oh, S. S., White, M. J., Gignoux, C. R. & Burchard, E. G. Making Precision Medicine
832 Socially Precise. Take a Deep Breath. *Am. J. Respir. Crit. Care Med.* **193**, 348-350
833 (2016).

834 5. Burchard, E. G. Medical research: Missing patients. *Nature* **513**, 301-302 (2014).

835 6. Barr, R. G. *et al*. Pulmonary Disease and Age at Immigration among Hispanics.
836 Results from the Hispanic Community Health Study/Study of Latinos. *Am. J. Respir.*
837 *Crit. Care Med.* **193**, 386-395 (2016).

838 7. Akinbami, L. Asthma Prevalence, Health Care Use and Mortality: United States,
839 2003-05 [Internet]. Centers for Disease Control and Prevention. 2010 Apr 6 [last

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

updated 2015 Nov 6; accessed: 2017 Jan 29]. Available from: http://www.cdc.gov/nchs/data/hestat/asthma03-05/asthma03-05.htm.

8. Palmer, L. J., Silverman, E. S., Weiss, S. T. & Drazen, J. M. Pharmacogenetics of asthma. *Am. J. Respir. Crit. Care Med.* **165**, 861-866 (2002).

9. Nelson, H. S. Beta-adrenergic bronchodilators. *N. Engl. J. Med.* **333**, 499-506 (1995).

10. Eggleston, P. A. *et al*. Medications used by children with asthma living in the inner city. *Pediatrics* **101**, 349-354 (1998).

11. Finkelstein, J. A., Lozano, P., Farber, H. J., Miroshnik, I. & Lieu, T. A. Underuse of controller medications among Medicaid-insured children with asthma. *Arch. Pediatr. Adolesc. Med.* **156**, 562-567 (2002).

12. Drazen, J. M., Silverman, E. K. & Lee, T. H. Heterogeneity of therapeutic responses in asthma. *Br. Med. Bull.* **56**, 1054-1070 (2000).

13. Burchard, E. G. *et al*. Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *Am. J. Respir. Crit. Care Med.* **169**, 386-392 (2004).

14. Naqvi, M. *et al*. Ethnic-specific differences in bronchodilator responsiveness among African Americans, Puerto Ricans, and Mexicans with asthma. *J. Asthma* **44**, 639-648 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

858 15. Wechsler, M. E. *et al*. Impact of race on asthma treatment failures in the asthma
859 clinical research network. *Am. J. Respir. Crit. Care Med.* **184**, 1247-1253 (2011).

860 16. Martinez, F. D. Safety of long-acting beta-agonists--an urgent need to clear the
861 air. *N. Engl. J. Med.* **353**, 2637-2639 (2005).

862 17. Dixon, A. E. Long-acting beta-agonists and asthma: the saga continues. *Am. J.*
863 *Respir. Crit. Care Med.* **184**, 1220-1221 (2011).

864 18. Nelson, H. S. *et al*. The Salmeterol Multicenter Asthma Research Trial: a
865 comparison of usual pharmacotherapy for asthma or usual pharmacotherapy plus
866 salmeterol. *Chest* **129**, 15-26 (2006).

867 19. Kramer, J. M. Balancing the benefits and risks of inhaled long-acting beta-
868 agonists--the influence of values. *N. Engl. J. Med.* **360**, 1592-1595 (2009).

869 20. McGeachie, M. J. *et al*. Polygenic heritability estimates in pharmacogenetics:
870 focus on asthma and related phenotypes. *Pharmacogenet Genomics* **23**, 324-328
871 (2013).

872 21. Nieminen, M. M., Kaprio, J. & Koskenvuo, M. A population-based study of
873 bronchial asthma in adult twin pairs. *Chest* **100**, 70-75 (1991).

874 22. Fagnani, C. *et al*. Heritability and shared genetic effects of asthma and hay fever:
875 an Italian study of young twins. *Twin Res. Hum. Genet.* **11**, 121-131 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

876 23. Himes, B. E. *et al*. Genome-wide association analysis in asthma subjects identifies
877 SPATS2L as a novel bronchodilator response gene. *PLoS Genet.* **8**, e1002824 (2012).

878 24. Duan, Q. L. *et al*. A genome-wide association study of bronchodilator response in
879 asthmatics. *Pharmacogenomics J.* **14**, 41-47 (2014).

880 25. Israel, E. *et al*. Genome-wide association study of short-acting beta2-agonists. A
881 novel genome-wide significant locus on chromosome 2 near ASB3. *Am. J. Respir. Crit.*
882 *Care Med.* **191**, 530-537 (2015).

883 26. Padhukasahasram, B. *et al*. Gene-based association identifies SPATA13-AS1 as a
884 pharmacogenomic predictor of inhaled short-acting beta-agonist response in
885 multiple population groups. *Pharmacogenomics J.* **14**, 365-371 (2014).

886 27. Drake, K. A. *et al*. A genome-wide association study of bronchodilator response
887 in Latinos implicates rare variants. *J. Allergy Clin. Immunol.* **133**, 370-378 (2014).

888 28. Conrad, D. F. *et al*. A worldwide survey of haplotype variation and linkage
889 disequilibrium in the human genome. *Nat. Genet.* **38**, 1251-1260 (2006).

890 29. Hoffmann, T. J. *et al*. Design and coverage of high throughput genotyping arrays
891 optimized for individuals of East Asian, African American, and Latino race/ethnicity
892 using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-430
893 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

894 30. Zheng, H. F. *et al*. Performance of genotype imputation for low frequency and
895 rare variants from the 1000 genomes. *PLoS One* **10**, e0116487 (2015).

896 31. Huang, J. *et al*. Improved imputation of low-frequency and rare variants using
897 the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).

898 32. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol.*
899 *Genet.* **24**, R102-10 (2015).

900 33. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world.
901 *Nature* **475**, 163-165 (2011).

902 34. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**,
903 161-164 (2016).

904 35. Oh, S. S. *et al*. Diversity in Clinical and Biomedical Research: A Promise Yet to Be
905 Fulfilled. *PLoS Med.* **12**, e1001918 (2015).

906 36. Hankinson, J. L., Odencrantz, J. R. & Fedan, K. B. Spirometric reference values
907 from a sample of the general U.S. population. *Am. J. Respir. Crit. Care Med.* **159**, 179-
908 187 (1999).

909 37. Kircher, M. *et al*. A general framework for estimating the relative pathogenicity
910 of human genetic variants. *Nat. Genet.* **46**, 310-315 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

911 38. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating
912 probabilities of fitness consequences for point mutations across the human genome.
913 *Nat. Genet.* **47**, 276-283 (2015).

914 39. Sobota, R. S. *et al*. Addressing population-specific multiple testing burdens in
915 genetic association studies. *Ann. Hum. Genet.* **79**, 136-147 (2015).

916 40. Zaykin, D. V. & Zhivotovsky, L. A. Ranks of genuine associations in whole-genome
917 scans. *Genetics* **171**, 813-823 (2005).

918 41. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing
919 association studies. *Biostatistics* **13**, 762-775 (2012).

920 42. Lee, J. H. *et al*. DNAH5 is associated with total lung capacity in chronic
921 obstructive pulmonary disease. *Respir. Res.* **15**, 97-014-0097-y (2014).

922 43. Ortiz, R. A. & Barnes, K. C. Genetics of allergic diseases. *Immunol. Allergy Clin.*
923 *North. Am.* **35**, 19-44 (2015).

924 44. Ramasamy, A. *et al*. A genome-wide meta-analysis of genetic variants associated
925 with allergic rhinitis and grass sensitization and their interaction with birth order. *J.*
926 *Allergy Clin. Immunol.* **128**, 996-1005 (2011).

927 45. Naqvi, M. *et al*. Association between IgE levels and asthma severity among
928 African American, Mexican, and Puerto Rican patients with asthma. *J. Allergy Clin.*
929 *Immunol.* **120**, 137-143 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

930 46. Edwards, M. R. *et al*. Targeting the NF-kappaB pathway in asthma and chronic
931 obstructive pulmonary disease. *Pharmacol. Ther.* **121**, 1-13 (2009).

932 47. Zhen, G. *et al*. IL-13 and epidermal growth factor receptor have critical but
933 distinct roles in epithelial cell mucin production. *Am. J. Respir. Cell Mol. Biol.* **36**, 244-
934 253 (2007).

935 48. Murgia, C. *et al*. Apical localization of zinc transporter ZnT4 in human airway
936 epithelial cells and its loss in a murine model of allergic airway inflammation.
937 *Nutrients* **3**, 910-928 (2011).

938 49. Besecker, B. *et al*. The human zinc transporter SLC39A8 (Zip8) is critical in zinc-
939 mediated cytoprotection in lung epithelia. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **294**,
940 L1127-36 (2008).

941 50. Persson, H. *et al*. Transcriptome analysis of controlled and therapy-resistant
942 childhood asthma reveals distinct gene expression profiles. *J. Allergy Clin. Immunol.*
943 **136**, 638-648 (2015).

944 51. Lin, Y. J. *et al*. Genetic variants in PLCB4/PLCB1 as susceptibility loci for
945 coronary artery aneurysm formation in Kawasaki disease in Han Chinese in Taiwan.
946 *Sci. Rep.* **5**, 14762 (2015).

947 52. Hersch, M. *et al*. Mapping genetic variants associated with beta-adrenergic
948 responses in inbred mice. *PLoS One* **7**, e41032 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

949 53. Dougherty, R. H. *et al*. Accumulation of intraepithelial mast cells with a unique
950 protease phenotype in T(H)2-high asthma. *J. Allergy Clin. Immunol.* **125**, 1046-
951 1053.e8 (2010).

952 54. Woodruff, P. G. *et al*. Genome-wide profiling identifies epithelial cell genes
953 associated with asthma and with treatment response to corticosteroids. *Proc. Natl.*
954 *Acad. Sci. U. S. A.* **104**, 15858-15863 (2007).

955 55. Motsinger-Reif, A. A. *et al*. Genome-wide association studies in
956 pharmacogenomics: successes and lessons. *Pharmacogenet Genomics* **23**, 383-394
957 (2013).

958 56. Peloso, G. M. *et al*. Phenotypic extremes in rare variant study designs. *Eur. J. Hum.*
959 *Genet.* **24**, 924-930 (2016).

960 57. Tishkoff, S. A. *et al*. The genetic structure and history of Africans and African
961 Americans. *Science* **324**, 1035-1044 (2009).

962 58. 1000 Genomes Project Consortium *et al*. A global reference for human genetic
963 variation. *Nature* **526**, 68-74 (2015).

964 59. Zou, J. Y. *et al*. Genetic and socioeconomic study of mate choice in Latinos reveals
965 novel assortment patterns. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13621-13626 (2015).

966 60. Hall, S. S. Genetics: a gene of rare effect. *Nature* **496**, 152-155 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

967 61. Burchard, E. G., Oh, S. S., Foreman, M. G. & Celedon, J. C. Moving toward true
968 inclusion of racial/ethnic minorities in federally funded studies. A key step for
969 achieving respiratory health equality in the United States. *Am. J. Respir. Crit. Care*
970 *Med.* **191**, 514-521 (2015).

971 62. Borrell, L. N. *et al*. Childhood obesity and asthma control in the GALA II and SAGE
972 II studies. *Am. J. Respir. Crit. Care Med.* **187**, 697-702 (2013).

973 63. Nishimura, K. K. *et al*. Early-life air pollution and asthma risk in minority
974 children. The GALA II and SAGE II studies. *Am. J. Respir. Crit. Care Med.* **188**, 309-318
975 (2013).

976 64. Thakur, N. *et al*. Socioeconomic status and childhood asthma in urban minority
977 youths. The GALA II and SAGE II studies. *Am. J. Respir. Crit. Care Med.* **188**, 1202-
978 1209 (2013).

979 65. White, M. J. *et al*. Novel genetic risk factors for asthma in African American
980 children: Precision Medicine and the SAGE II Study. *Immunogenetics* **68**, 391-400
981 (2016).

982 66. Jun, G. *et al*. Detecting and estimating contamination of human DNA samples in
983 sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839-848 (2012).

984 67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
985 transform. *Bioinformatics* **25**, 1754-1760 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

986 68. DePristo, M. A. *et al*. A framework for variation discovery and genotyping using
987 next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011).

988 69. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-
989 2158 (2011).

990 70. Li, H. Toward better understanding of artifacts in variant calling from high-
991 coverage samples. *Bioinformatics* **30**, 2843-2851 (2014).

992 71. Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and
993 richer datasets. *Gigascience* **4**, 7-015-0047-8. eCollection 2015 (2015).

994 72. Best practice for converting VCF files to plink format [Internet]. Genetics For
995 Fun. 2014 Nov 25 [last updated 2014 Nov 25; accessed: 2015 Oct 1]. Available from:
996 http://apol1.blogspot.nl/2014/11/best-practice-for-converting-vcf-files.html.

997 73. Thornton, T. *et al*. Estimating kinship in admixed populations. *Am. J. Hum. Genet.*
998 **91**, 122-138 (2012).

999 74. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of
1000 ancestry in unrelated individuals. *Genome Res.* **19**, 1655-1664 (2009).

1001 75. Manichaikul, A. *et al*. Robust relationship inference in genome-wide association
1002 studies. *Bioinformatics* **26**, 2867-2873 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1003 76. Price, A. L. *et al*. Principal components analysis corrects for stratification in
1004 genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).

1005 77. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing
1006 for disease and population genetic studies. *Nat. Methods* **10**, 5-6 (2013).

1007 78. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative
1008 modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*
1009 **93**, 278-288 (2013).

1010 79. Liu, X. *et al*. WGSA: an annotation pipeline for human genome sequencing
1011 studies. *J. Med. Genet.* **53**, 111-112 (2016).

1012 80. Yao, T. C. *et al*. Associations of age, gender, and BMI with prevalence of allergic
1013 diseases in children: PATCH study. *J. Asthma* **48**, 503-510 (2011).

1014 81. Nicolai, T., Pereszlenyiova-Bliznakova, L., Illi, S., Reinhardt, D. & von Mutius, E.
1015 Longitudinal follow-up of the changing gender ratio in asthma from childhood to
1016 adulthood: role of delayed manifestation in girls. *Pediatr. Allergy Immunol.* **14**, 280-
1017 283 (2003).

1018 82. Joseph, M. *et al*. Racial disparity in the association between body mass index and
1019 self-reported asthma in children: a population-based study. *J. Asthma* **53**, 492-497
1020 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1021 83. Dixon, A. E. *et al*. Effect of obesity on clinical presentation and response to
1022 treatment in asthma. *J. Asthma* **43**, 553-558 (2006).

1023 84. Ullah, M. I., Newman, G. B. & Saunders, K. B. Influence of age on response to
1024 ipratropium and salbutamol in asthma. *Thorax* **36**, 523-529 (1981).

1025 85. Mohamed, M. H., Lima, J. J., Eberle, L. V., Self, T. H. & Johnson, J. A. Effects of
1026 gender and race on albuterol pharmacokinetics. *Pharmacotherapy* **19**, 157-161
1027 (1999).

1028 86. Carroll, C. L., Bhandari, A., Zucker, A. R. & Schramm, C. M. Childhood obesity
1029 increases duration of therapy during severe asthma exacerbations. *Pediatr. Crit.*
1030 *Care. Med.* **7**, 527-531 (2006).

1031 87. Forno, E. *et al*. Decreased response to inhaled steroids in overweight and obese
1032 asthmatic children. *J. Allergy Clin. Immunol.* **127**, 741-749 (2011).

1033 88. Duggal, P., Gillanders, E. M., Holmes, T. N. & Bailey-Wilson, J. E. Establishing an
1034 adjusted p-value threshold to control the family-wide type 1 error in genome wide
1035 association studies. *BMC Genomics* **9**, 516-2164-9-516 (2008).

1036 89. Pruim, R. J. *et al*. LocusZoom: regional visualization of genome-wide association
1037 scan results. *Bioinformatics* **26**, 2336-2337 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1038 90. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in
1039 meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586-598
1040 (2011).

1041 91. McCarthy, S. *et al*. A reference panel of 64,976 haplotypes for genotype
1042 imputation. *Nat. Genet.* **48**, 1279-1283 (2016).

1043 92. Das, S. *et al*. Next-generation genotype imputation service and methods. *Nat.*
1044 *Genet.* **48**, 1284-1287 (2016).

1045 93. Han, Y. Y. *et al*. Diet, interleukin-17, and childhood asthma in Puerto Ricans. *Ann.*
1046 *Allergy Asthma Immunol.* **115**, 288-293.e1 (2015).

1047 94. Levin, A. M. *et al*. Nocturnal asthma and the importance of race/ethnicity and
1048 genetic ancestry. *Am. J. Respir. Crit. Care Med.* **190**, 266-273 (2014).

1049 95. Ong, B. A. *et al*. Gene network analysis in a pediatric cohort identifies novel lung
1050 function genes. *PLoS One* **8**, e72899 (2013).

1051 96. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for
1052 thousands of genomes. *Nat. Methods* **9**, 179-181 (2011).

1053 97. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype
1054 imputation method for the next generation of genome-wide association studies.
1055 *PLoS Genet.* **5**, e1000529 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1056 98. Miller, M. R. *et al*. Standardisation of lung function testing: the authors' replies to
1057 readers' comments. *Eur. Respir. J.* **36**, 1496-1498 (2010).

1058 99. Chen, W. *et al*. Expression quantitative trait loci (eQTL) mapping in Puerto Rican
1059 children. *PLoS One* **10**, e0122464 (2015).

1060 100. Karolchik, D. *et al*. The UCSC Table Browser data retrieval tool. *Nucleic Acids*
1061 *Res.* **32**, D493-6 (2004).

1062 101. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-
1063 efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**,
1064 R25-2009-10-3-r25. Epub 2009 Mar 4 (2009).

1065 102. Zhang, Y. *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137-
1066 2008-9-9-r137. Epub 2008 Sep 17 (2008).

1067 103. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment
1068 using MACS. *Nat. Protoc.* **7**, 1728-1740 (2012).

1069 104. Landt, S. G. *et al*. ChIP-seq guidelines and practices of the ENCODE and
1070 modENCODE consortia. *Genome Res.* **22**, 1813-1831 (2012).

1071 105. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**,
1072 15-21 (2013).

1073 106. Flicek, P. *et al*. Ensembl 2014. *Nucleic Acids Res.* **42**, D749-55 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1074 107. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and

1075 dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

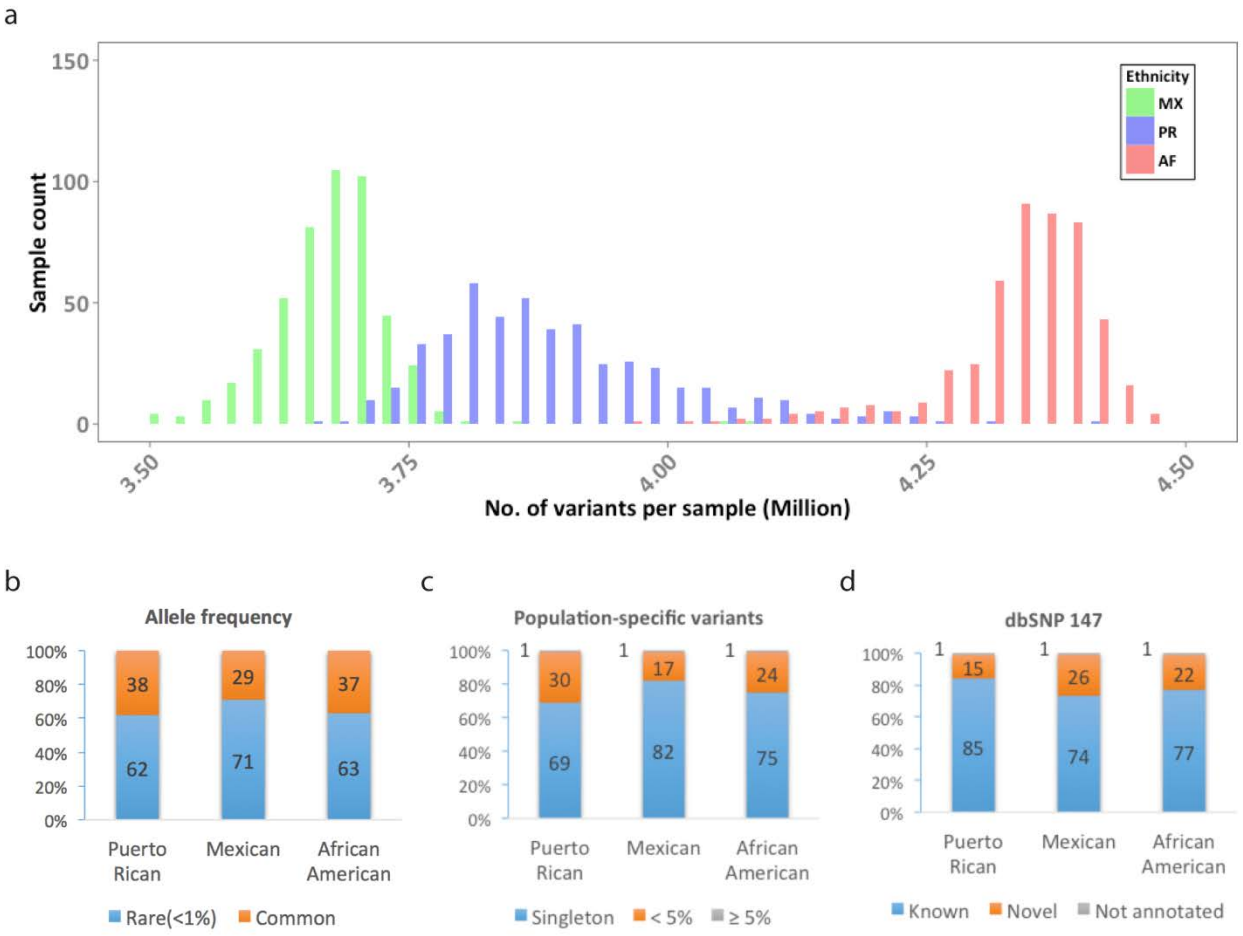
Figure 1. (**a**) Number of variants per sample. The bin size is 0.025M variants. (**b**) Allele frequency of biallelic SNPs (relative to hg19). (**c**) Allele frequency of population-specific biallelic SNPs. (d) Novel biallelic SNPs based on dbSNP build147.

Figure 2. (a) Manhattan plot of the trans-ethnic meta-analysis of single locus BDR association testing. The top ten BDR-associated loci are circled. SNPs with high LD (r^2 > 0.8) are considered as a single locus. The horizontal lines are colored black for universal genome-wide significance threshold (5.00×10^{-8}), red for adjusted genome-wide significance threshold (3.53×10^{-7}) and blue for suggestive significance threshold (7.06×10^{-6}). **(b)** Forest plot of the population-specific and joint effect of the top BDR-associated SNP, rs17834628. The direction of effect is the same in all three populations. **(c)** The top BDR-associated SNP (rs17834628) is plotted together with 400kb flanking regions on either side. Multiple SNPs in high LD (r^2 > 0.8, red) reached a suggestive significance level.

Figure 3. Manhattan plot of SKAT-O analysis of biallelic common and rare SNPs grouped by 1kb windows sliding across chromosome 1 to 22 in **(a)** Puerto Ricans **(b)** African Americans, **(c)** Mexicans, and **(d)** all populations combined. Bonferroni-corrected genome-wide and suggestive significance levels are marked by red and blue lines, respectively.

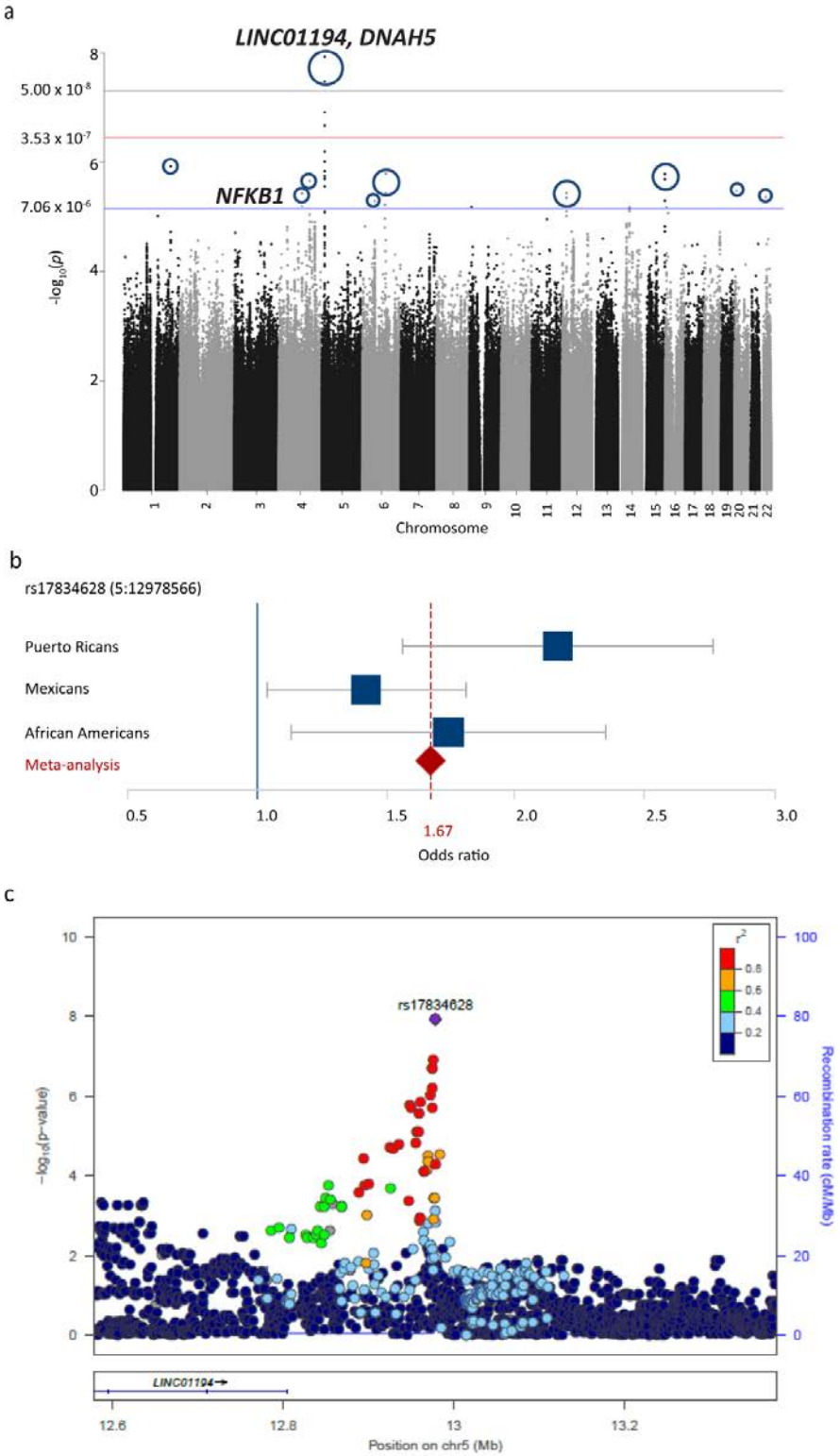
bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3

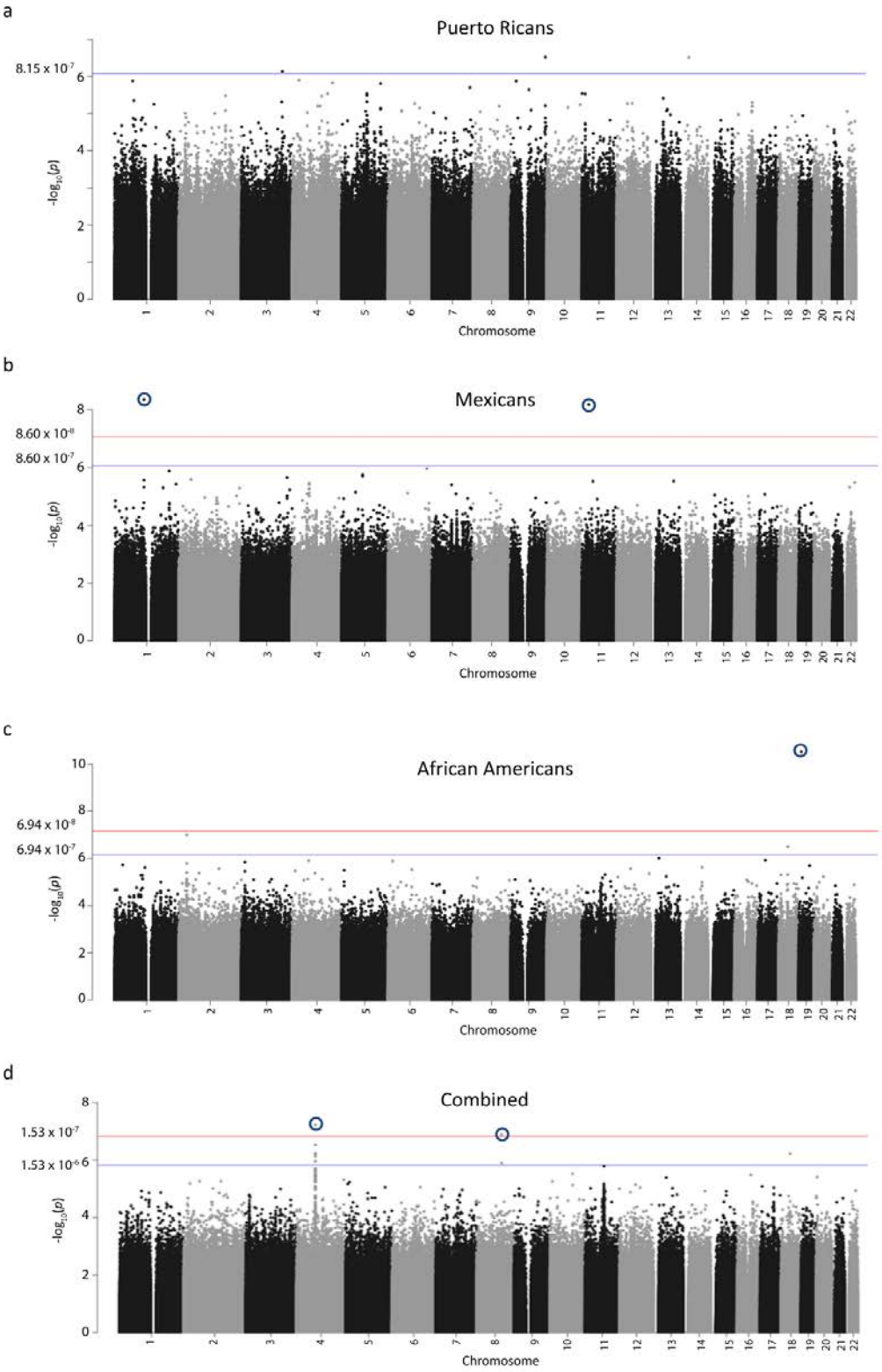


Table 1. Study Population Description (N=1,441).

Descriptive Statistics		**Puerto Ricans** (n=483)			**Mexicans** (n=483)			**African Americans** (n=475)		
		High BDR	**Low BDR**	***P***	**High BDR**	**Low BDR**	***P***	**High BDR**	**Low BDR**	***P***
Number of Subjects		239	244	-	243	240	-	233	242	-
Percent Male		53.6%	53.3%	1.0	60.1%	52.1%	0.08	55.4%	47.9%	0.12
Median Age, yr (IQR)		11.6 (9.7 - 14.8)	12.2 (10.1 - 15.2)	0.18	11.7 (9.6 - 14.0)	13.3 (10.6 - 16.0)	<0.001	13.8 (11.0 - 16.8)	13.8 (10.9 - 17.1)	0.48
Mean Global Ancestry Proportions	AFR	0.24	0.22	0.44	0.05	0.05	0.37	0.79	0.79	0.80
	EUR	0.63	0.64	0.27	0.37	0.36	0.84	0.19	0.20	0.70
	NAM	0.13	0.13	0.93	0.58	0.59	0.90	0.02	0.02	0.90
BMI Category, N	Obese	76	67	0.32	100	96	0.85	82	83	0.85
	Non-Obese	163	177		143	144		151	159	
Pre-FEV_1 % Predicted, N	< 80%	149	56	<0.001	43	7	<0.001	47	6	<0.001
	≥ 80%	90	188		200	233		186	236	
Median ΔFEV_1, % (IQR)		21.2 (18.2 - 25.7)	5.0 (2.9 - 6.3)	-	12.7 (10.3 - 16.8)	3.6 (2.0 - 4.9)	-	15.5 (13.3 - 20.3)	3.3 (2.0 - 4.4)	-

BDR: bronchodilator drug response. IQR: interquartile range. Pre-FEV_1 % predicted: percentage of measured FEV_1 relative to predicted FEV_1 estimated by the Hankinson lung function prediction equations[36], prior to administration to albuterol. ΔFEV_1: a quantitative measure of BDR, measured as the percent change in baseline FEV_1 after administration of albuterol. High and low drug responders were chosen from the extremes of the BDR (ΔFEV_1) distribution.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Summary statistics of variants

	Puerto Ricans	Mexicans	African Americans
Range in variants, per individual (mean)			
All	3.7M – 4.4 M (4.0 M)	3.5M – 4.1M (3.7M)	4.0M – 4.5M (4.3M)
SNPs	3.4M – 4.0 M (3.6M)	3.2M - 3.7M (3.4M)	3.6M – 4.1M (4.0M)
Indels	284,067 – 344,493 (305,618)	272,635 – 321,778 (289,096)	311,997 – 354,646 (339,570)
Others*	21,233 – 29,177 (24,728)	19,703 – 26,772 (23,190)	22,014 – 30,993 (27,911)
No. of biallelic SNPs, union**	29.2M	28.1M	36.3M
By allele frequency of data			
Rare (< 1%)	18,169,292 (62%)	20,029,291 (71%)	22,847,938 (63%)
Common (≥ 1%)	11,007,125 (38%)	8,092,157 (29%)	13,459,004 (37%)
Population-specific***	6,680,909	9,687,651	14,114,142
Singleton	4,616,383 (69%)	7,983,950 (82%)	10,574,879 (75%)
< 5%	6,676,286 (> 99%)	9,675,271 (> 99%)	14,096,844 (> 99%)
≥ 5%	4,623 (< 1%)	12,380 (< 1%)	17,298 (< 1%)
By novelty (dbSNP147)			
Known	24,654,288 (85%)	20,673,589 (74%)	28,122,310 (77%)
Novel	4,463,442 (15%)	7,384,900 (26%)	8,119,236 (22%)
Not annotated^	58,687 (< 1%)	62,959 (< 1%)	65,396 (< 1%)
By protein impact****			
Coding	284,269 (1%)	289,055 (1%)	362,823 (1%)
Nonsynonymous	157,541	164,312	203,029
Stopgain / stoploss	3,194	3,531	4,158
Splicing	2,111	2,214	2,620
Other coding	121,423	118,998	153,016
Noncoding	28,833,461 (99%)	27,769,434 (99%)	35,878,723 (99%)
Not annotated^	58,687 (< 1%)	62,959 (< 1%)	65,396 (< 1%)

* Includes multi-nucleotide polymorphism (MNPs), complex, symbolic and mixed variants as defined by GATK VariantEval.

** Biallelic SNPs with less than 10% genotype missingness per population were included.

*** Biallelic SNPs that are present in only one of the three studied populations.

**** Other coding variants include those annotated as exonic and synonymous in ANNOVAR.

^ **Not annotated**: biallelic SNPs not included in the annotation pipeline because they are not present in TOPMed freeze 2 and 3 data releases.

bioRxiv preprint doi: https://doi.org/10.1101/128116; this version posted April 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3. CADD score summary statistics of biallelic SNPs*.

Population	CADD score	No. of SNPs*	% Coding	% Singletons
Puerto Ricans	0-9	26,936,285	0.5%	29%
	10-19	2,027,392	4%	32%
	20-24	123,481	32%	37%
	≥ 25	30,572	>99%	51%
Mexicans	0-9	25,932,006	0.5%	44%
	10-19	1,967,096	4%	49%
	20-24	124,173	34%	55%
	≥ 25	35,214	>99%	70%
African Americans	0-9	33,468,245	0.5%	36%
	10-19	2,573,523	4%	40%
	20-24	158,967	32%	46%
	≥ 25	40,811	>99%	64%

*Biallelic SNPs with less than 10% genotype missingness per population were included.

Table 4. Results from trans-ethnic BDR association tests for common variants

Chr	Start	LD SNP	rsID	Effect allele	OR (95%CI)	p	Effect allele frequency			Nearest genes
							PR	MX	AA	
5	12978566	*	rs17834628	A	1.67 (1.29-2.16)	1.18E-08	0.32	0.42	0.17	*LINC01194* (173k), *MIR4454* (311k), *CTNND2* (1074k); *DNAH5* (712k)
5	12968341		rs35661809	G	1.59 (1.20-2.10)	3.33E-08	0.34	0.43	0.24	*LINC01194* (163k), *MIR4454* (300k), *CTNND2* (1064k); *DNAH5* (722k)
5	12975934		rs17237639	G	1.61 (1.30-2.00)	1.22E-07	0.31	0.43	0.16	*LINC01194* (171k), *MIR4454* (308k), *CTNND2* (1072k); *DNAH5* (715k)
5	12975187		rs1017452	G	1.60 (1.31-1.96)	2.11E-07	0.31	0.43	0.16	*LINC01194* (170k), *MIR4454* (307k), *CTNND2* (1071k); *DNAH5* (715k)
5	12975322		rs1017454	A	1.60 (1.31-1.96)	2.11E-07	0.31	0.43	0.16	*LINC01194* (170k), *MIR4454* (307k), *CTNND2* (1071k); *DNAH5* (715k)
5	12975265		rs1017453	C	1.56 (1.25-1.95)	6.40E-07	0.31	0.42	0.16	*LINC01194* (170k), *MIR4454* (307k), *CTNND2* (1071k); *DNAH5* (715k)
5	12972636		rs17237443	C	1.59 (1.28-1.97)	9.85E-07	0.29	0.42	0.11	*LINC01194* (170k), *MIR4454* (307k), *CTNND2* (1071k); *DNAH5* (715k)
1	209324294	*	rs10746419	T	1.29 (0.75-2.25)	1.19E-06	0.49	0.54	0.53	*MIR205HG* (278k), *MIR205* (281k), *CAMK1G* (433k), *LAMB3* (464k)
5	12961545		rs17833938	A	1.56 (1.28-1.91)	1.45E-06	0.30	0.42	0.12	*LINC01194* (156k), *MIR4454* (294k), *CTNND2* (1057k); *DNAH5* (729k)
6	104240500	*	rs13437006	C	1.56 (1.21-2.02)	1.61E-06	0.22	0.24	0.32	*HACE1* (935k), *LINC00577* (1144k), *LIN28B* (1164k)
15	101230457	*	rs1565749	A	1.66 (1.18-2.32)	1.64E-06	0.18	0.15	0.18	*ASB7* (39k), *LINS1* (88k), *PRKXP1* (131k)
5	12948369		rs34845041	T	1.56 (1.26-1.92)	1.77E-06	0.30	0.42	0.12	*LINC01194* (143k), *MIR4454* (280k), *CTNND2* (1044k); *DNAH5* (742k)
5	12975108		rs1017451	T	1.55 (1.24-1.93)	1.96E-06	0.30	0.42	0.13	*LINC01194* (170k), *MIR4454* (307k), *CTNND2* (1071k); *DNAH5* (715k)
5	12950432		rs62347395	G	1.55 (1.26-1.92)	2.02E-06	0.30	0.42	0.12	*LINC01194* (145k), *MIR4454* (282k), *CTNND2* (1046k); *DNAH5* (740k)
15	101231049		rs57924834	A	1.59 (1.25-2.03)	2.04E-06	0.23	0.20	0.20	*ASB7* (39k), *LINS1* (89k), *PRKXP1* (132k)
4	137382142	*	rs17048684	A	1.8 (1.06-3.05)	2.20E-06	0.11	0.14	0.18	*LINC00613* (547k), *LINC00613* (573k); *PCDH18* (1058k)
5	12959598		rs1438293	G	1.55 (1.24-1.93)	2.73E-06	0.29	0.42	0.11	*LINC01194* (154k), *MIR4454* (292k), *CTNND2* (1055k); *DNAH5* (731k)
20	8635168	*	rs16995064	G	1.96 (1.12-3.43)	3.30E-06	0.12	0.13	0.05	***PLCB1* (intron 7)**; *PLCB4* (415k)
12	19821401	*	rs66544720	T	0.66 (0.55-0.78)	3.66E-06	0.33	0.37	0.16	*AEBP2* (146k), *PLEKHA5* (292k)
6	104235591		rs6926020	C	1.57 (1.25-1.97)	3.68E-06	0.19	0.23	0.27	*HACE1* (940k), *LINC00577* (1149k), *LIN28B* (1169k)
4	103453535	*	rs28450894	T	0.47 (0.34-0.64)	3.75E-06	0.06	0.03	0.12	*SLC39A8* (187k); ***NFKB1* (intron 3)**; *MANBA* (99k)
4	103461559		rs4648006	T	0.47 (0.34-0.64)	3.75E-06	0.06	0.03	0.12	*SLC39A8* (195k); ***NFKB1* (intron 5)**; *MANBA* (91k)
22	27826429	*	rs60163793	G	2.01 (1.20-3.38)	4.30E-06	0.04	0.14	0.15	*MN1* (318k), *PITPNB* (421k)
12	19824386		rs7313907	C	0.66 (0.55-0.79)	4.35E-06	0.33	0.37	0.16	*AEBP2* (149k), *PLEKHA5* (295k)
12	19820677		rs11044754	A	0.66 (0.55-0.79)	4.54E-06	0.33	0.37	0.16	*AEBP2* (146k), *PLEKHA5* (291k)
15	101233236		rs55638658	A	1.61 (1.13-2.30)	5.08E-06	0.18	0.15	0.18	*ASB7* (41k), *LINS1* (91k), *PRKXP1* (134k)
6	54581204	*	rs13200833	A	0.66 (0.48-0.90)	5.15E-06	0.32	0.24	0.22	*TINAG* (326k), *MLIP* (450k); *FAM83B* (130k)

LD SNP: An asterisk indicates SNPs that are not in linkage disequilibrium (not in LD defined as $R^2 < 0.8$ in any population) with other SNPs that have more significant association *p*-values; **Nearest genes**: Distance to the nearest RefSeq genes are indicated in parentheses. Genes that overlap with BDR-associated SNPs are bold. High drug responders were assigned as cases in analyses throughout this study.

Table 5. Results from association testing on combined effects of rare variants on BDR

Chr	Start	Stop	*p*	Population	Nearest genes
1	114177000	114178000	4.40E-09	MX	***MAGI3* (intron 9)**, *PHTF1* (62k), *RSBN1* (126k)
11	27507000	27508000	6.59E-09	MX	*LOC105376671* (3k), *LGR4* (13k), *LIN7C* (8k)
19	10424000	10425000	3.12E-11	AA	*ZGLP1* (4k), *ICAM5* (17k), ***FDX1L* (intron 3)**, *RAVER1* (2k)
4	73478000	73479000	6.25E-08	Combined	*ADAMTS3* (43k), *COX18* (441k)
8	97926000	97927000	1.32E-08	Combined	*SDC2* (302k), ***CPQ* (intron 4)**, *LOC101927066* (37k), *TSPYL5* (359k)

chr: chromosome; AA: African Americans; MX: Mexicans; Combined: all individuals in all three populations. Nearest genes: Distance to the nearest RefSeq genes is indicated in parentheses. Genes that overlap with BDR-associated SNPs are bold.