nature

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41586-019-1310-4

In the format provided by the authors and unedited.

Genetic analyses of diverse populations improves discovery for complex traits

Genevieve L. Wojcik[1,35], Mariaelisa Graff[2,35], Katherine K. Nishimura[3,35], Ran Tao[4,5,35], Jeffrey Haessler[3,35], Christopher R. Gignoux[1,6,35], Heather M. Highland[2,35], Yesha M. Patel[7,35], Elena P. Sorokin[1], Christy L. Avery[2], Gillian M. Belbin[8,9], Stephanie A. Bien[3], Iona Cheng[10], Sinead Cullina[8,9], Chani J. Hodonsky[2], Yao Hu[3], Laura M. Huckins[11], Janina Jeff[8,9], Anne E. Justice[2], Jonathan M. Kocarnik[3], Unhee Lim[12], Bridget M. Lin[2], Yingchang Lu[9], Sarah C. Nelson[13], Sung-Shim L. Park[7], Hannah Poisner[8,9], Michael H. Preuss[9], Melissa A. Richard[14], Claudia Schurmann[9,15,16], Veronica W. Setiawan[7], Alexandra Sockell[1], Karan Vahi[17], Marie Verbanck[9], Abhishek Vishnu[9], Ryan W. Walker[9], Kristin L. Young[2], Niha Zubair[3], Victor Acuña-Alonso[18], Jose Luis Ambite[17], Kathleen C. Barnes[6], Eric Boerwinkle[19], Erwin P. Bottinger[9,15,16], Carlos D. Bustamante[1], Christian Caberto[12], Samuel Canizales-Quinteros[20], Matthew P. Conomos[13], Ewa Deelman[17], Ron Do[9,11], Kimberly Doheny[21], Lindsay Fernández-Rhodes[2,22], Myriam Fornage[14], Benyam Hailu[23], Gerardo Heiss[2], Brenna M. Henn[24], Lucia A. Hindorff[25], Rebecca D. Jackson[26], Cecelia A. Laurie[13], Cathy C. Laurie[13], Yuqing Li[10,27], Dan-Yu Lin[2], Andres Moreno-Estrada[28], Girish Nadkarni[9], Paul J. Norman[6], Loreall C. Pooler[7], Alexander P. Reiner[13], Jane Romm[21], Chiara Sabatti[1], Karla Sandoval[28], Xin Sheng[7], Eli A. Stahl[11], Daniel O. Stram[7], Timothy A. Thornton[13], Christina L. Wassel[29], Lynne R. Wilkens[12], Cheryl A. Winkler[30], Sachi Yoneyama[2], Steven Buyske[31,36], Christopher A. Haiman[32,36], Charles Kooperberg[3,36], Loic Le Marchand[12,36], Ruth J. F. Loos[9,11,36], Tara C. Matise[33,36], Kari E. North[2,36], Ulrike Peters[3,36], Eimear E. Kenny[8,9,11,34,36]* & Christopher S. Carlson[3,36]*

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Division of Public Health Science, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [4]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA. [5]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [6]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [7]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [8]The Center for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [9]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [10]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA. [11]Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [12]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA. [13]Department of Biostatistics, University of Washington, Seattle, WA, USA. [14]Brown Foundation Institute for Molecular Medicine, The University of Texas Health Science Center, Houston, TX, USA. [15]Hasso-Plattner-Institute for Digital Engineering, Digital Health Center, Potsdam, Germany. [16]Hasso-Plattner-Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [17]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA. [18]Escuela Nacional de Antropologia e Historia, Mexico City, Mexico. [19]Human Genetics Center, School of Public Health, The University of Texas Health Science Center, Houston, TX, USA. [20]Instituto Nacional de Medicina Genómica, Mexico City, Mexico. [21]Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD, USA. [22]Department of Biobehavioral Health, The Pennsylvania State University, University Park, PA, USA. [23]NIH National Institute on Minority Health and Health Disparities, Bethesda, MD, USA. [24]Department of Anthropology, University of California Davis, Davis, CA, USA. [25]NIH National Human Genome Research Institute, Bethesda, MD, USA. [26]Center for Clinical and Translational Science, Ohio State Medical Center, Columbus, OH, USA. [27]Cancer Prevention Institute of California, Fremont, CA, USA. [28]National Laboratory of Genomics for Biodiversity (UGA-LANGEBIO), Irapuato, Mexico. [29]College of Medicine, University of Vermont, Burlington, VT, USA. [30]Basic Science Program, Frederick National Laboratory, Frederick, MD, USA. [31]Department of Statistics, Rutgers University, New Brunswick, NJ, USA. [32]Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [33]Department of Genetics, Rutgers University, New Brunswick, NJ, USA. [34]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [35]These authors contributed equally: Genevieve L. Wojcik, Mariaelisa Graff, Katherine K. Nishimura, Ran Tao, Jeffrey Haessler, Christopher R. Gignoux, Heather M. Highland, Yesha M. Patel. [36]These authors jointly supervised this work: Steven Buyske, Christopher A. Haiman, Charles Kooperberg, Loic Le Marchand, Ruth J. F. Loos, Tara C. Matise, Kari E. North, Ulrike Peters, Eimear E. Kenny, Christopher S. Carlson. *e-mail: eimear.kenny@mssm.edu; ccarlson@fredhutch.org

Supplementary Information

1. Detailed PAGE Study Descriptions

BioMe Biobank: The Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center (MSMC), BioMe™ BioBank (BioMe) is an EMR-linked bio-repository drawing from Mount Sinai Medical Center consented patients which were drawn from a population of over 70,000 inpatients and 800,000 outpatients annually. [1] The MSMC serves diverse local communities of upper Manhattan, including Central Harlem (86% African American), East Harlem (88% Hispanic/Latino), and Upper East Side (88% Caucasian/White) with broad health disparities. BioMe™ enrolled over 26,500 participants from September 2007 through August 2013, with 25% African American, 36% Hispanic/Latino (primarily of Caribbean origin), 30% Caucasian, and 9% of Other ancestry. The BioMe™ population reflects community-level disease burdens and health disparities with broad public health impact. Biobank operations are fully integrated in clinical care processes, including direct recruitment from clinical sites waiting areas and phlebotomy stations by dedicate Biobank recruiters independent of clinical care providers, prior to or following a clinician standard of care visit. Recruitment currently occurs at a broad spectrum of over 30 clinical care sites. Study participants of self-reported European ancestry were not included in this analysis. (dbGaP study accession number: phs000925).

HCHS/SOL: HCHS/SOL: The Hispanic Community Health Study / Study of Latinos (HCHS/SOL) is a multi-center study of Hispanic/Latino populations with the goal of determining the role of acculturation in the prevalence and development of diseases, and to identify other traits that impact Hispanic/Latino health. [2] The study is sponsored by the National Heart, Lung, and Blood Institute (NHLBI) and other institutes, centers, and offices of the National Institutes of Health (NIH). Recruitment began in 2006 with a target population of 16,000 persons of Cuban, Puerto Rican, Dominican, Mexican or Central/South American origin. Household sampling was employed as part of the study design. Participants were recruited through four sites affiliated with San Diego State University, Northwestern University in Chicago, Albert Einstein College of Medicine in Bronx, New York, and the University of Miami. Researchers from seven academic centers provided scientific and logistical support. Study participants who were self-identified Hispanic/Latino and aged 18-74 years underwent extensive psycho-social and clinical assessments during 2008-2011. A re-examination of the HCHS/SOL cohort is conducted during 2015-2017. Annual telephone follow-up interviews are ongoing since study inception to determine health outcomes of interest. (dbGaP study accession number: phs000555).

MEC: The Multiethnic Cohort (MEC) is a population-based prospective cohort study including approximately 215,000 men and women from Hawaii and California. All participants were 45-75 years of age at baseline, and primarily of 5 ancestries: Japanese Americans, African Americans, European Americans, Hispanic/Latinos, and Native Hawaiians. [3,4] MEC was funded by the National Cancer Institute in 1993 to examine lifestyle risk factors and genetic susceptibility to cancer. All eligible cohort members completed baseline and follow-up questionnaires. Within the PAGE II investigation, MEC proposes to study: 1) diseases for which we have DNA available for large numbers of cases and controls (breast, prostate, and colorectal cancer, diabetes, and obesity); 2) common traits that are risk factors for these diseases (e.g., body mass index / weight, waist-to-hip ratio, height), and 3) relevant disease-associated biomarkers (e.g., fasting insulin and lipids, steroid hormones). The specific aims are: 1) to determine the population-based epidemiologic profile (allele frequency, main effect, heterogeneity by disease characteristics) of putative causal variants in the five racial/ethnic groups in MEC; 2) for variants displaying effect heterogeneity across ethnic/racial groups, we will utilize differences in LD to identify a more complete spectrum of associated variants at these loci; 3) investigate gene x gene and gene x environment interactions to identify modifiers; 4) examine the associations of putative causal variants with already measured intermediate phenotypes (e.g., plasma insulin, lipids, steroid hormones); and 5) for variants that do not fall within known genes, start to investigate their relationships with gene expression and epigenetic patterns in small genomic studies. For this project, MEC contributed African American, Japanese American, and Native Hawaiian samples. (dbGaP study accession number: phs000220).

PAGE Global Reference Panel: The Global Reference Panel (GRP) was created by Stanford-contributed samples that can act as a population reference dataset across the globe. Therefore, this dataset includes reference individuals, without phenotypes, chosen to help infer ancestry that will aid in understanding the

diverse samples available in PAGE. The complete dataset comprises individuals of European, African, Asian, Oceanian, and Native American descent, from a total of over 50 populations. A subset of these individuals from Puno, Peru and Easter Island (Rapa Nui), Chile, are included in the PAGE samples that were whole genome sequenced in 2015. The Global Reference Panel comprises 6 sample sets: (1) a population sample of Andean individuals primarily of Quechuan/Aymaran ancestry from Puno, Peru; (2) a population sample of Easter Island (Rapa Nui), Chile; (3) individuals of indigenous origin from Oaxaca, Mexico; (4) individuals of indigenous origin from Honduras; (5) individuals of indigenous origin from Colombia; (6) individuals of indigenous origin from the Nama and Khomani KhoeSan populations of the Northern Cape, South Africa. PAGE also used samples from the Human Genome Diversity Project (HGDP) [5], a subset of the Maasai from HapMap, as well as individuals sampled by the Bustamante Lab and their collaborators. The dataset comprises individuals of European, African, Asian, Oceanian, and Native American descent, from over 50 populations. Study participants were selected to reflect a family history of living in the region. The data are currently available through dbGaP (dbGaP study accession number: phs001033).

WHI: The Women's Health Initiative (WHI) is a long-term, prospective, multi-center cohort study investigating post-menopausal women's health in the US. [6] WHI was funded by the National Institutes of Health and the National Heart, Lung, and Blood Institute to study strategies to prevent heart disease, breast cancer, colon cancer, and osteoporotic fractures in women 50-79 years of age. WHI involves 161,808 women recruited between 1993 and 1998 at 40 centers across the US. The study consists of two parts: the WHI Clinical Trial which was a randomized clinical trial of hormone therapy, dietary modification, and calcium/Vitamin D supplementation, and the WHI Observational Study, which focused on many of the inequities in women's health research and provided practical information about incidence, risk factors, and interventions related to heart disease, cancer, and osteoporotic fractures. For this project, women who self-identified as European were excluded from the study sample (dbGaP study accession number: phs000227).

2. Phenotype Harmonization and Modeling

The phenotypes included in this study were previously harmonized across the PAGE studies.

Anthropometry: The following anthropometric traits were analyzed: height, body mass index (BMI), and waist-to-hip ratio (WHR). Weight in kilograms and height in centimeters were measured by trained clinic staff in the SOL and WHI studies at the time of enrollment. Waist and hip were also measured in SOL and WHI to the nearest centimeter. In MEC and BioMe weight and height were self-reported by questionnaire and in MEC waist and hip were also self-reported. BMI was then calculated as the ratio of weight to height squared. Individuals <18 years of age and women who were pregnant were also excluded. For GWAS analysis, measurements outside of 6 standard deviations from the mean (based on sex and race) were removed. Then we created sex-specific residuals for each trait adjusted for age (and BMI for waist-to-hip ratio), then inverse normally transformed these residuals. These inverse normally transformed residuals were used in the final analysis and further adjustment was made for self-identified ancestry, study, study center (for MEC and SOL only), and 10 principal components.

C-Reactive Protein (CRP): Serum CRP was reported in mg/L. CRP outliers (+/- 4 standard deviations) were dropped, and CRP was +1 and then natural log transformed. Those who were pregnant at blood draw were excluded from the analysis. There were 28,537 individuals in the final sample. Models were adjusted by age at CRP measurement, sex, BMI, current smoking status, self-identified race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Cigarettes per Day (CPD): The number of cigarettes smoked per day (CPD) was estimated among ever smokers (n=15,8672) based on self-report and electronic health record data. To normalize the distribution of CPD, we added one to the reported CPD and then log transformed this variable. Models were adjusted for age, sex, study, study center (for MEC and SOL only), self-reported race/ethnicity, and the first 10 principal components.

Chronic Kidney Disease (CKD): CKD was defined as an eGFR (estimated by the CKD Epi Equation) <=60 ml/min/1.73m^2 or ICD-9 codes 585.1-585.6, or 585.9, or ICD-10 codes N18.1-N18.5, or N18.9. Participants with end stage-renal disease (ESRD) were excluded from the analysis. CKD was modeled as a binary outcome, and models were adjusted for age, sex, race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Coffee Consumption: The coffee analysis included 35,902 subjects with coffee consumption measured by number of cups per day which was natural log transformed. Models were adjusted age, sex, study, study center (for MEC and SOL only), and first 10 principal components.

Diastolic Blood Pressure (DBP): Diastolic blood pressure was measured as the average of resting measurements in mmHg. Diastolic blood pressure was adjusted by 10 mmHg for the self-reported use of any antihypertensive medication. We winsorized outliers by setting measurements +/- 6 standard deviations from the overall mean to that value. Models for diastolic blood pressure adjusted for age, sex, BMI, self-identified race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Electrocardiogram – PR interval: PR interval is a heritable electrocardiographic measure of atrial and atrioventricular nodal conduction. Resting, supine, or semi-recumbent ECGs were digitally recorded in each study at baseline by certified technicians using standard 12-lead ECGs using either Marquette MAC12 or MAC PC machines (GE Healthcare, Milwaukee, WI, USA; Supplemental Table 1). Comparable procedures were used for preparing participants, placing electrodes, recording, transmitting, processing, and controlling the quality of the ECGs. The PR interval was measured electronically using the Marquette 12SL algorithm. Exclusion criteria included pregnancy, poor ECG quality, non-sinus rhythm including atrial fibrillation and atrial flutter on ECG, pacemaker implantation, second or third degree heart block, extreme PR values (PR ≤ 80 ms or ≥ 320 ms), prevalent heart failure or myocardial infarction, and Wolff-Parkinson-White syndrome on ECG. All models were adjusted for age, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, systolic blood pressure, height, body mass index, the use of beta-adrenergic blocking agents, and the first 10 principal components.

Electrocardiogram – QRS interval: QRS interval, from the beginning of the Q wave to the end of the S wave on an electrocardiogram, reflects ventricular depolarization and conduction time. Resting, supine, or semi-recumbent ECGs were digitally recorded in each study at baseline by certified technicians using standard 12-lead ECGs using either Marquette MAC12 or MAC PC machines (GE Healthcare, Milwaukee, WI, USA; Supplemental Table 1). Comparable procedures were used for preparing participants, placing electrodes, recording, transmitting, processing, and controlling the quality of the ECGs. The QRS interval was measured electronically using the Marquette 12SL algorithm. Exclusion criteria included pregnancy, poor ECG quality, non-sinus rhythm including atrial fibrillation and atrial flutter on ECG, pacemaker implantation, second- or third-degree heart block, QRS duration > or equal 120 ms, use of antiarrhythmic medications, prevalent heart failure or myocardial infarction, and Wolff-Parkinson-White syndrome on ECG. All models were adjusted for age, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, heart rate, systolic blood pressure, height, body mass index, and the first 10 principal components.

Electrocardiogram (ECG) measures – QT interval: QT interval is a measurement of ventricular depolarization and repolarization. Resting, supine, or semi-recumbent ECGs were digitally recorded in each study at baseline by certified technicians using standard 12-lead ECGs using either Marquette MAC12 or MAC PC machines (GE Healthcare, Milwaukee, WI, USA; Supplemental Table 1). Comparable procedures were used for preparing participants, placing electrodes, recording, transmitting, processing, and controlling the quality of the ECGs. The QT interval was measured electronically using the Marquette 12SL algorithm. Exclusion criteria included pregnancy, poor ECG quality, non-sinus rhythm including atrial fibrillation and atrial flutter on ECG, pacemaker implantation, QRS duration > or equal 120ms, and prevalent heart failure. All models were adjusted for age, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, heart rate, and the first 10 principal components.

End-Stage Renal Disease (ESRD): ESRD was defined as an eGFR (by the CKD-Epi Equation) of <=15

ml/min/1.73m^2 or ICD-9 code 585.6 or ICD-10 code N18.6. Participants with chronic kidney disease (CKD) were excluded from the analysis. ESRD was modeled as a binary outcome, and models were adjusted for age, sex, race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Estimated Glomerular Filtration Rate (eGFR) eGFR by CKD Epi Equation: Continuous eGFR in ml/mon/1.73m^2 was estimated by the serum creatinine-based CKD-Epi equation [7], which has been validated for Hispanics [8]. Non-Hispanic White equations were used to estimate GFR in Hispanic and Asian participants. eGFR was not transformed, and models were adjusted for age, sex, race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Fasting Glucose (FG): FG was reported in mmol/L. Individuals that were pregnant, had a fasting glucose greater than 7 mmol/L, had Type 2 Diabetes, or were non-fasting at measurement were excluded from analysis. Rank normalized residuals were calculated after adjusting for age, sex, age*sex, study, smoking status, and BMI. Association models were adjusted for self-identified race/ethnicity, 10 principal components, and study center (for MEC and SOL only).

Fasting Insulin (FI): FI was reported in pmol/L. Individuals that were pregnant, had a fasting glucose greater than 7 mmol/L, had Type 2 Diabetes, or were non-fasting at measurement were excluded from analysis. Insulin levels were log-transformed. Rank normalized residuals were calculated after adjusting for age, sex, ageXsex, study, smoking status, and BMI. Association models were adjusted for self-identified race/ethnicity, 10 principal components, and study center (for MEC and SOL only).

Glycated Hemoglobin (HbA1c): Glycated Hemoglobin was reported in mmol/mol. Individuals that were pregnant, had a fasting glucose greater than 7 mmol/L, or had Type 2 Diabetes were excluded from the analysis. Rank normalized residuals were calculated after adjusting for age, sex, age by sex, study, smoking status, and BMI. Association models were adjusted for self-identified race/ethnicity, 10 principal components, and study center (for MEC and SOL only).

High-Density Lipoprotein (HDL): HDL measurements were reported in mg/dL, were untransformed, and were adjusted for each individual's medication use by adding a constant based on the type of medication used. If multiple medications were reported, only the correction factor with the largest effect was applied. The constant used for adjustment was based on effects observed in previous publications, and included adjustments for statins [9], fibrates [9], bile acid sequestrants [10], niacin [9], and cholesterol absorption inhibitors [11,12]. An individual's raw HDL measurement was adjusted by the following values if the participant was taking one of these medications: statins: -2.3; fibrates: -5.9; bile acid sequestrants: -1.9; niacin: -9.9, cholesterol absorption inhibitors: +0.0. Those who were pregnant at blood draw, or who had fasted less than 8 hours prior to lipid blood draw were excluded from the study sample. There were 33,063 individuals in the final study sample. Models were adjusted by age at lipid measurement, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, and 10 principal components.

Hypertension (HT): Hypertension cases were defined based on any of the following criteria: 1) measured systolic blood pressure ≥140 mmHg, 2) measured diastolic blood pressure ≥90 mmHg, 3) reported use of any antihypertensive medication, or 4) ICD-9 codes 401.x-405.x or ICD-10 codes I10.x-I15.x. Individuals not meeting any of these criteria were considered normotensive (controls). Models for hypertension were adjusted for age, sex, BMI, study, study center (for MEC and SOL only), self-identified race/ethnicity, and 10 principal components.

Low-Density Lipoprotein (LDL): LDL measurements were reported in mg/dL, and were calculated using the Friedewald Equation [13], which subtracts the HDL measurement and the Triglyceride measurement (divided by 5) from the Total Cholesterol value. LDL was not calculated if the triglyceride value was greater than 400 mg/dL. LDL values were then adjusted for each individual's medication use by adding a constant based on the type of medication used. If multiple medications were reported, only the correction factor with the largest effect was applied. The constant used for adjustment was based on effects observed in previous publications, and included adjustments for statins [9], fibrates [9], bile acid sequestrants [10], niacin [9], and cholesterol absorption inhibitors [11,12]. An individual's raw LDL measurement was adjusted by the following values if the participant was taking one of these medications: statins: +49.9; fibrates: +40.1; bile acid

sequestrants: +40.5; niacin: +24.7; cholesterol absorption inhibitors: +40.5. Those who were pregnant at blood draw, or who had fasted less than 8 hours prior to lipid blood draw were excluded from the study sample. There were 32,221 individuals in the final study sample. Models were adjusted by age at lipid measurement, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, and 10 principal components.

Mean Corpuscular Hemoglobin Concentration (MCHC): Mean corpuscular hemoglobin concentration was reported in g/dL and was untransformed. MCHC is calculated using the formula: 100*hemoglobin/hematocrit. MCHC outliers (+/- 4 standard deviations) were dropped, along with observations for HIV+ individuals, participants with a reported hereditary anemia, and women pregnant at time of blood draw. The MCHC model was adjusted for age at blood draw, sex, current smoking status, self-identified race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Platelet Count (PLT): Platelet count was reported as cells x 10^9/L and was untransformed. PLT outliers (+/- 4 standard deviations) were dropped, along with observations for HIV+ individuals, and women pregnant at time of blood draw. The PLT model was adjusted for age at blood draw, sex, current smoking status, self-identified race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Systolic Blood Pressure (SBP): Systolic blood pressure was measured as the average of resting measurements in mmHg. Systolic blood pressure was adjusted by 15 mmHg for the self-reported use of any antihypertensive medication. We winsorized outliers by setting measurements +/- 6 standard deviations from the overall mean to that value. Models for systolic blood pressure adjusted for age, sex, BMI, self-identified race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

Total Cholesterol (TC): Total Cholesterol measurements were reported in mg/dL, were untransformed, and were adjusted for each individual's medication use by adding a constant based on the type of medication used. If multiple medications were reported, only the correction factor with the largest effect was applied. The constant used for adjustment was based on effects observed in previous publications, and included adjustments for statins [9], fibrates [9], bile acid sequestrants [10], niacin [9], and cholesterol absorption inhibitors [11,12]. An individual's raw rotal cholesterol measurement was adjusted by the following values if the participant was taking one of these medications: statins: +52.1; fibrates: +46.1; bile acid sequestrants: +0.0; niacin: +34.6; cholesterol absorption inhibitors: +40.5. Those who were pregnant at blood draw, or who had fasted less than 8 hours prior to lipid blood draw were excluded from the study sample. There were 33,185 individuals in the final study sample. Models were adjusted by age at lipid measurement, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, and 10 principal components.

Type II Diabetes (T2D): The Type 2 Diabetes analysis included 14,046 cases and 31,695 controls with complete covariate data after excluding individuals who were pregnant at blood draw and those who were classified as cases for Type 1 Diabetes. Controls with glucose values greater than 7 mmol/L were excluded, as well as any cases that were younger than 20 years of age. The models were adjusted by age at T2D diagnosis, sex, study, study center (for MEC and SOL only), BMI, self-identified race/ethnicity, and first 10 principal components.

Triglycerides (TG): Triglyceride measurements were reported in mg/dL, were adjusted for each individual's medication use by adding a constant based on the type of medication used, and then natural log transformed. If multiple medications were reported, only the correction factor with the largest effect was applied. The constant used for adjustment was based on effects observed in previous publications, and included adjustments for statins [9], fibrates [9], bile acid sequestrants [10], niacin [9], and cholesterol absorption inhibitors [11,12]. An individual's raw Triglyceride measurement was adjusted by the following values if the participant was taking one of these medications: statins: +18.4; fibrates: +57.1; bile acid sequestrants: +0.0; niacin: +89.4; cholesterol absorption inhibitors: +0.0. Those who were pregnant at blood draw, or who had fasted less than 8 hours prior to lipid blood draw were excluded from the study sample. Individuals with a Triglyceride value greater than 3000 mg/dL were dropped from the analysis (n=1). There were 33,096 individuals in the final study sample. Models were adjusted by age at lipid measurement, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, and 10 principal components.

White Blood Cell Count (WBC): Total WBC count was measured in 10^9 cells/L. WBC count outliers (+/- 4 standard deviations) were dropped, and total WBC was natural log transformed. Those who were pregnant a blood draw were excluded from the analysis. There were 28,534 individuals in the final sample. Models were adjusted by age at WBC measurement, sex, BMI, study, current smoking status, self-identified race/ethnicity, study, study center (for MEC and SOL only), and 10 principal components.

3. Genotyping and Imputation

One major challenge in multi-ethnic studies is the limited availability of genotyping arrays that comparably tag variation in multiple genetic ancestries, especially in those with African ancestry. To address this, a collaboration among PAGE, Illumina Inc., the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) [14], and other academic partners developed the Multi-Ethnic Genotyping Array (MEGA), which includes a GWAS scaffold designed to tag both common and low frequency variants in global populations. [15] **(Extended Data Figure 2)** Additionally, it contains enhanced tagging in exonic regions, hand-curated content to interrogate clinically relevant variants, and enriched coverage to fine-map known GWAS loci. [16]

DNA was isolated from blood (SOL, BioMe, GRP), buffy coat (WHI, MEC), mouthwash/saliva (MEC, GRP), or lymphoblastoid cell line (GRP). There were 548 HapMap genotyping control samples, and 1,001 blind study duplicate samples. Samples were genotyped on complete or partial 96-well plates, over three batches. MEC, BioMe, SOL and WHI samples were distributed across the three batches in a proportion that represented the size of each study. Sex, race/ethnicity, recruitment site, and DNA source were not stratified, but samples were randomly selected within each stratification level. To better understand the rich genetic diversity within PAGE, particularly in underrepresented ancestries from the Africa and the Americas, we genotyped an additional 1,553 samples on MEGA from a Global Reference Panel (GRP), which were drawn from the Human Genome Diversity Project [5] and supplemented with previously sampled populations from the Americas and Africa [14,17–20]. Some GRP samples were included on the plates in Batches 1 and 2, but the majority of these samples were included on 12 separate plates. Each plate contained one or two duplicates. Duplicate samples were place on a different plate than the original sample. Each plate contained an average of one HapMap sample. There were also 110 investigator controls previously genotyped that were included with at most one of these controls per plate.

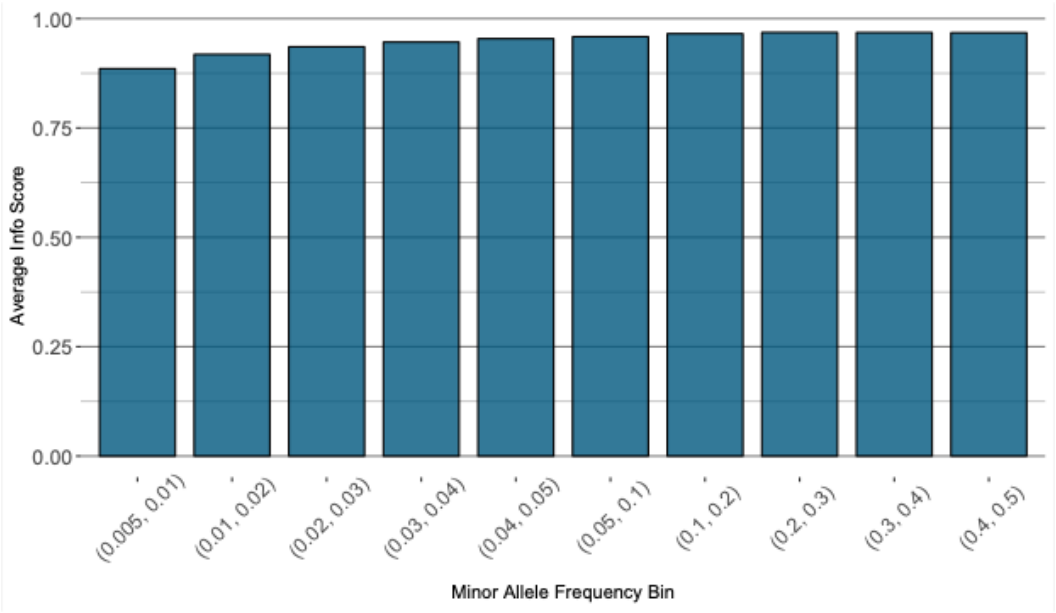
A total of 53,426 samples were genotyped at the Center for Inherited Disease Research (CIDR) using the Multi-Ethnic Genotyping Array (MEGA), Consortium version. MEGA was designed through a collaboration between PAGE, University of Michigan, CAAPA, and Illumina to provide broad coverage for globally diverse populations, as well as provide enhanced exomic, functional, and clinically-relevant content. Genotypes were called by the Center for Inherited Disease Research (CIDR) using the GenomeStudio version 2001.1, Genotyping Module 1.9.4, and GenTrain version 1.0.

Genotyping data that passed initial quality control at CIDR, including sex discrepancies, Mendelian inconsistencies, unexpected duplication, unexpected non-duplication, poor performance, or DNAmixture observed were released to the Quality Assurance / Quality Control (QA/QC) analysis team at the University of Washington Genetic Coordinating Center (UWGAC), the study investigator's team, and dbGaP. The UWGAC QA/QC team used quality control methods previously described by Laurie et al. [21] The UWGAC QA/QC team further removed samples with identity issues, restricted consent, and duplicate scans to return data for 51,520 subjects.

A total of 1,705,969 variants were genotyped on MEGA. Variant-level quality control (QC) was completed by were filtered through various criteria, including the exclusion of (1) CIDR technical filters, (2) variants with missing call rate >= 2%, (3) variants with more than 6 discordant calls in 988 study duplicates, (4) SNPs with greater than 1 Mendelian errors in 282 trios and 1,439 duos, (5) variants with a Hardy-Weinberg p-value less than 10^{-4}, (6) variants with sex difference in allele frequency >= 0.2 for autosomes/XY, and (7) variants with sex difference in heterozygosity > 0.3 for autosomes/XY. After variant QC, a total of 1,438,399 variants remained.

Imputation was conducted at the UWGAC. Sites were further restricted to variants with (1) known chromosome and position; (2) located on chromosomes 1-22, X, or XY (pseudo-autosomal); (3) with unique positions, which involved removing redundant and duplicate sites; and (4) sites with available

strand annotation. After these restrictions, a total of 1,402,653 sites remained. The study samples were phased with SHAPEIT2 [22] and imputed with IMPUTE2 [23] to the 1000 Genomes Project Phase 3 data release [24]. Reference panel variants were filtered to exclude all monomorphs and singletons (i.e. restricting to minor allele count (MAC) >= 2 across all 1000 Genomes Phase 2 samples). Imputed variants were excluded if the IMPUTE2 info score was less than 0.4.

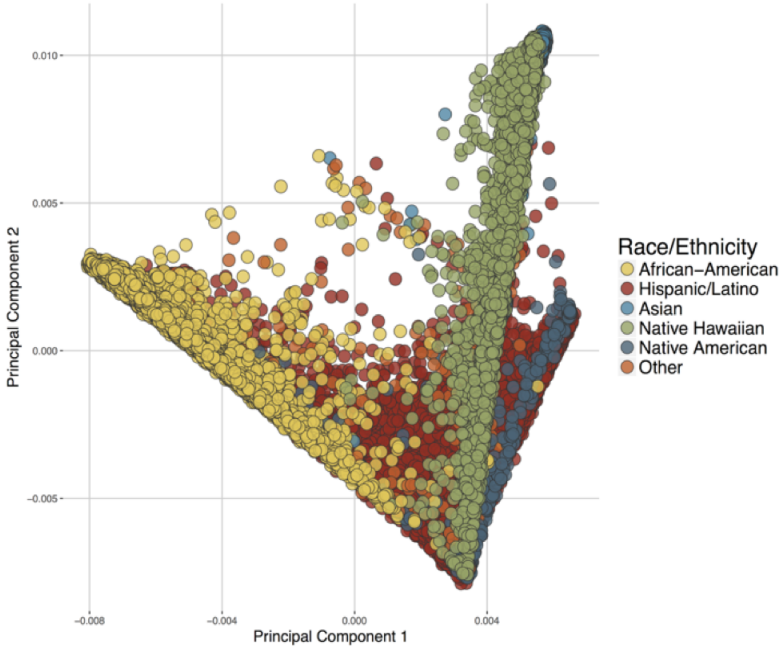


Supplementary Figure 1: Average info from imputed data from MEGA to 1000Genomes Project within PAGE by minor allele frequency.
We show high imputation quality across all minor allele frequency bins from 0.5-50%.

4. Population Substructure

Historically, analyses have been stratified by self-identified race/ethnicity to account for confounding by genetic ancestry. In PAGE, we conducted principal component analysis to evaluate population substructure and mapped self-identified racial/ethnic groups (Hispanic/Latino, African American, Asian, Native Hawaiian, Native American, and Other) onto the estimated principal components (PCs). The selection of unrelated individuals was essential for accurate estimation of the principal components within the global study population. Kinship coefficients were estimated using PC-Relate, as implemented in the R package GENESIS (Conomos et al. 2015; Conomos, Reiner, et al. 2016). The SNPRelate (Zheng et al. 2012) package in R was then used for principal components analysis using unrelated individuals, defined as pairwise kinship coefficients less than 2^(-9/2). Since principal components are required for unbiased kinship estimation in admixed populations, the two estimation procedures were iterated to ensure that the principal components were computed over unrelated individuals. Principal component scores were then estimated for all remaining individuals by projection. [25]

Most notably in Hispanics/Latinos, but evident to a lesser extent in all populations, genetic ancestry reveals greater demographic complexity compared with culturally assigned labels. Genetic ancestry appears as a continuum, demonstrating that it is not categorical in diverse populations that have varying degrees of admixture (**Supp Figure 1**). Stratifying by self-reported race/ethnicity would fail to separate groups with similar patterns of genetic ancestry and therefore would still require adjustment of PCs with reduced statistical power in a smaller sample size. For this reason, we pooled all samples in a single mega-analysis.



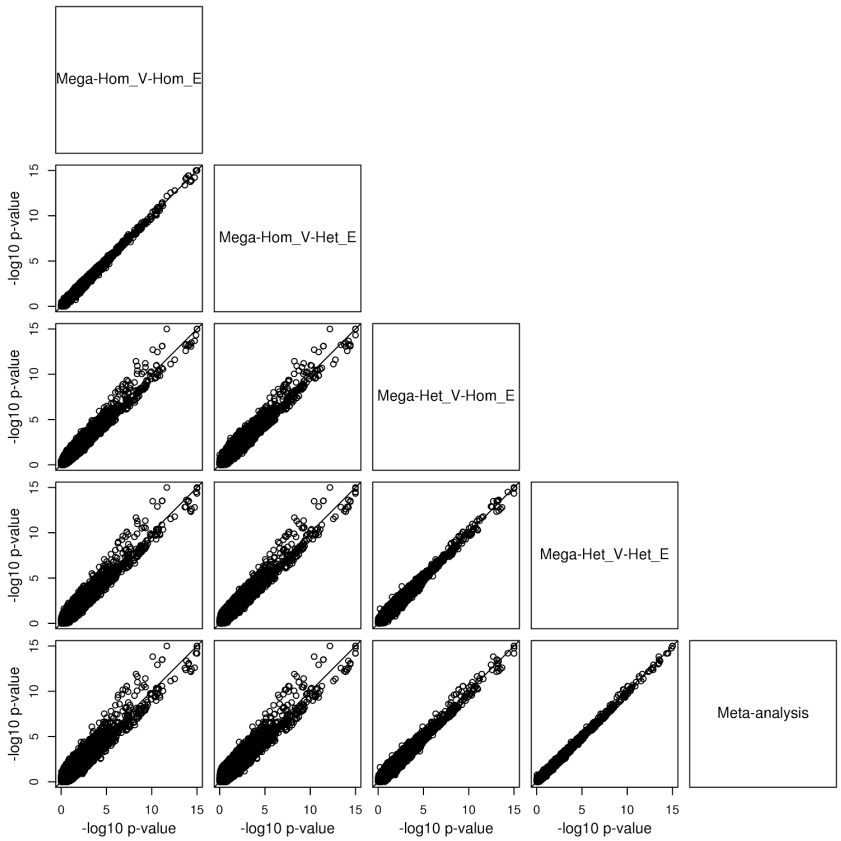
Supplementary Figure 2: Principal Component Analysis of PAGE Populations.
Scatterplot of PCs for PAGE racial/ethnic groups. Each point represents one individual, color-coded by self-identified race/ethnicity. Global variation for all PAGE participants for principal component (PC) 1 versus PC2. Genotyped individuals self-identified as Hispanic/Latino (N=22,216), African American (N=17,299), Asian (N=4,680), Native Hawaiian (N=3,940), Native American (N=652), or Other (N=1,052).

5. Meta-analysis versus Mega-analysis in Multi-ethnic Studies

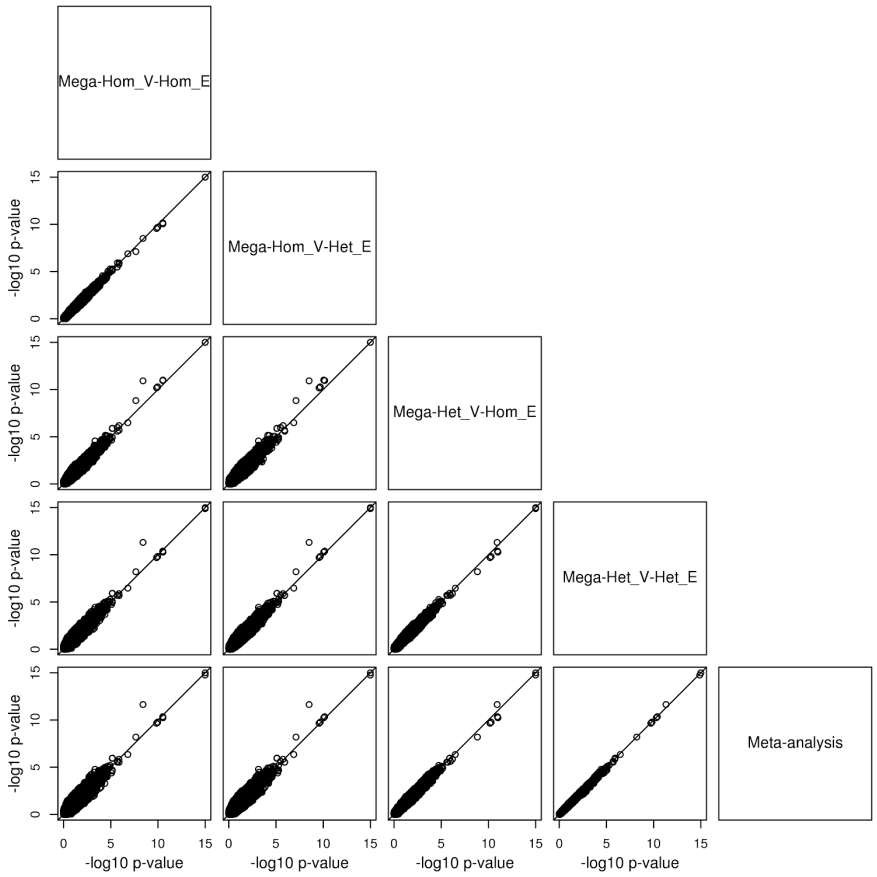
It has been shown, both theoretically and numerically, that meta-analysis and mega-analysis of independent studies are (asymptotically) equivalent, if mega-analysis allows nuisance parameters (i.e., trait variances and covariate effects) to be different among studies [26,27]. The comparison of meta and pooled analysis has some subtle aspects that are important. Considering a simple scenario with a variant with MAF=50% in population 1 and MAF=1% in population 2, and the same sample size and same effect size in each population, if allele frequency is the only difference between the two studies then it is actually slightly more powerful to perform a mega analysis (with no adjustment for study) rather than meta-analysis. In general, however, we expect that allele frequency will not constitute the only difference between studies, and we therefore always include study indicators in pooled analyses. Inclusion of study as a covariate ensures that the mega-analysis estimator for the genetic effect can be expressed as an inverse-variance weighted estimator, with weights that are asymptotically equivalent to the weights in the meta-analysis estimator; see Example 1 in Lin & Zeng (2010). [26] **Supplementary Figure 2** compares the results of meta-analysis to four types of mega-analysis (assuming heterogeneous vs homogeneous trait variances and heterogeneous vs homogeneous covariate effects) for MCHC. The mega-analysis that allows both trait variances and covariate effects to be different among studies (i.e., ethnicities in our case) fits the same model as meta-analysis does, so the *p*-values from the two methods are virtually identical; the *p*-values can be quite different when trait variances are allowed to be different across ethnicities vs when they are assumed to the same across ethnicities; and the results are fairly similar when covariate effects are allowed to be different across ethnicities vs when they are assumed to be the same across ethnicities. **Supplementary Figure 3** shows that the same conclusions hold for variants whose ethnic-specific MAF differences are greater than 0.4. We chose mega-analysis over meta-analysis *because the former allows related individuals across studies (whereas the latter does not) and provides greater*

flexibilities in modelling the effects of covariates.

SUGEN and GENESIS allow trait variances to be different among studies (i.e., ethnicities in our case) whereas other LMM methods do not. We chose to focus on methods adjusting for global ancestry as these were more stable across the extremely heterogeneous mix of populations in PAGE, where local ancestry estimation could be challenging to reconcile. Further, as we have previously shown, local ancestry adjustment is expected to impair statistical power for discovery when compared to global ancestry adjustment. [28] **Supplementary Figures 4-8** display the *p*-values for the five traits with the most severe trait-variance heterogeneity, when the trait variances are assumed to be homogeneous vs heterogeneous across ethnicities in the analysis. These results confirm the theoretical expectation that allowing heterogeneous variances yields better control of type I error and higher power in association tests.

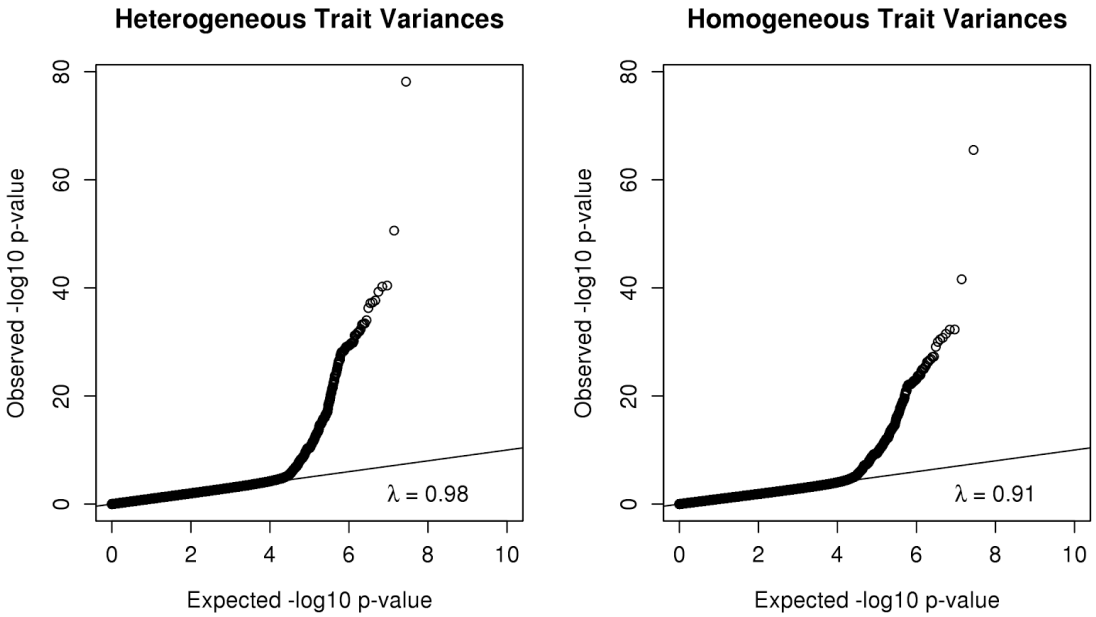


Supplementary Figure 3. Comparisons of the *p*-values between meta-analysis and four types of mega-analysis for MCHC (N=19,803).
Het_V and Hom_V denote, respectively, heterogeneous and homogeneous trait variances among ethnic groups; Het_E and Hom_E denote, respectively, heterogeneous and homogeneous covariate effects among ethnic groups. The p-values are estimated from Wald test, with values less than 110^{-15} winsorized at 110^{-15}.

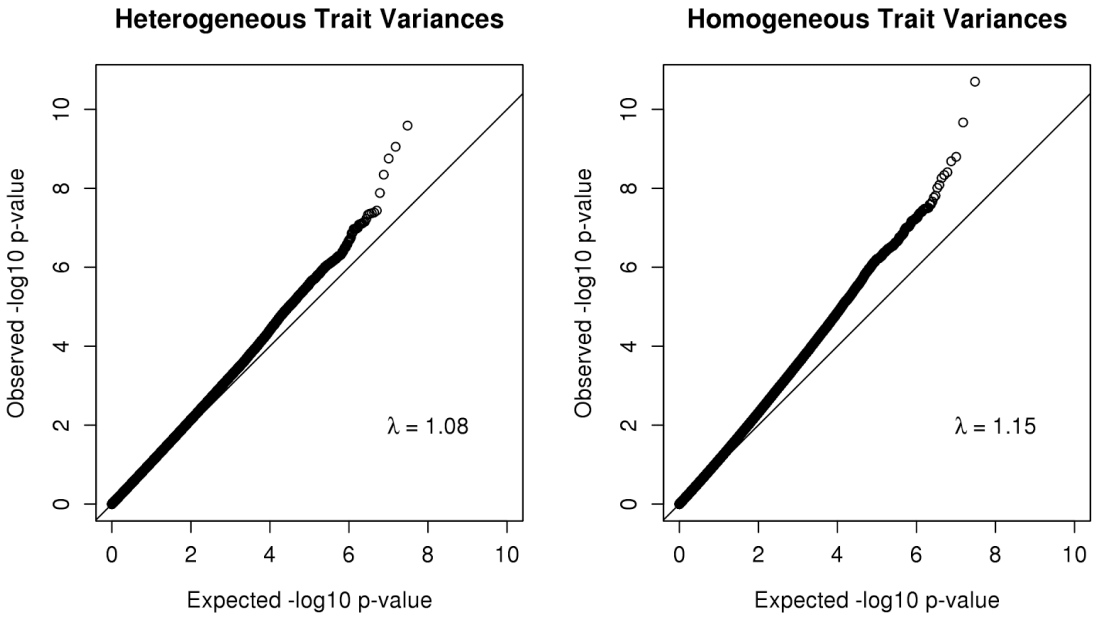


Supplementary Figure 4. Comparisons of the *p*-values between meta-analysis and four types of mega-analysis for MCHC with variants whose ethnic-specific minor allele frequency (MAF) differences are greater than 0.4 (N=19,803).

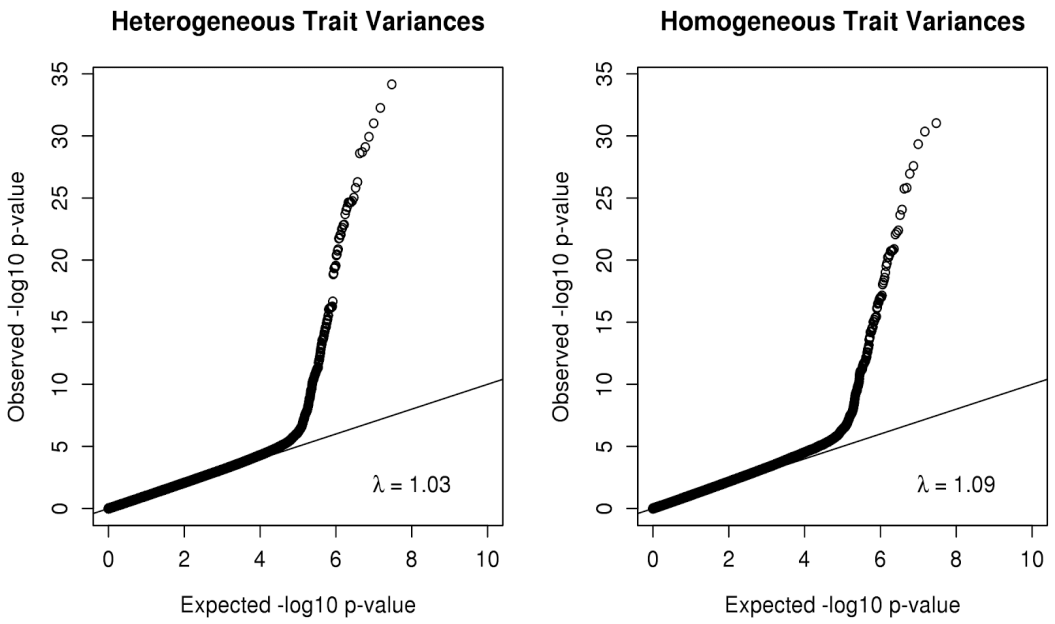
Het_V and Hom_V denote, respectively, heterogeneous and homogeneous trait variances among ethnic groups; Het_E and Hom_E denote, respectively, heterogeneous and homogeneous covariate effects among ethnic groups. The p-values are estimated from Wald test, with values less than 110^{-15} winsorized at 110^{-15}.



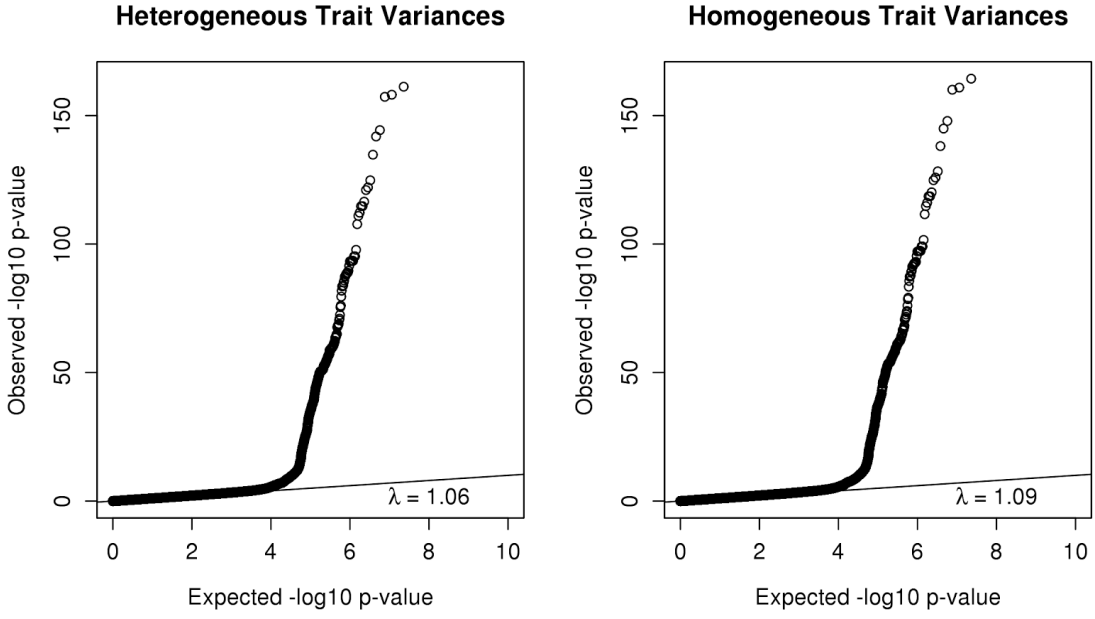
Supplementary Figure 5. Quantile-quantile plots of p-values for MCHC when trait values are assumed to be heterogeneous versus homogeneous. P-values estimated from Wald test. (N=19,803)



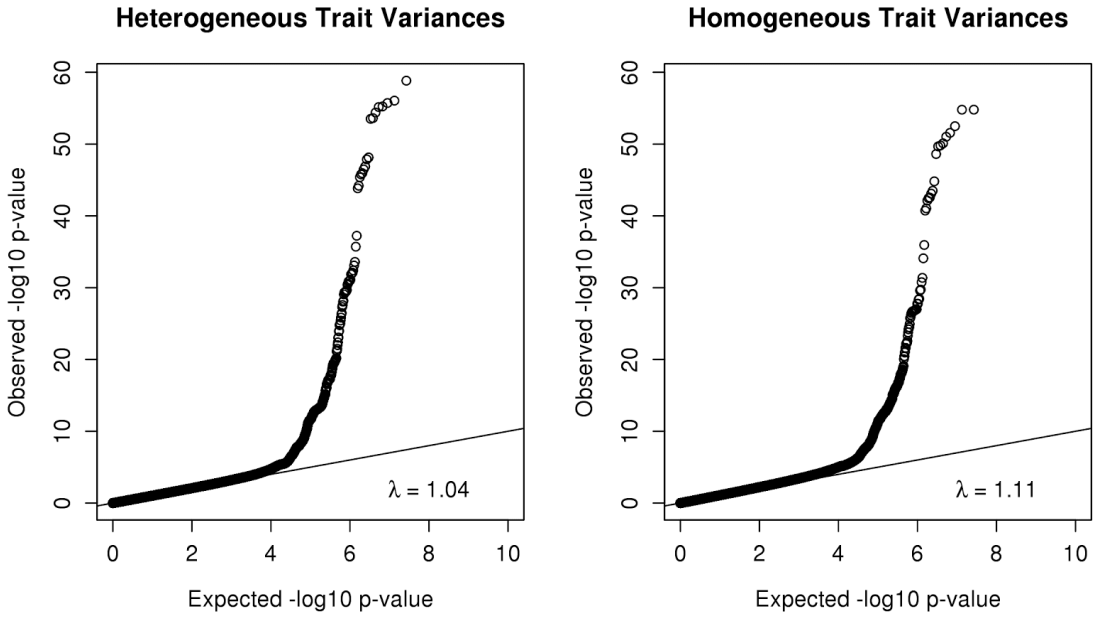
Supplementary Figure 6. Quantile-quantile plots of p-values for eGFR when trait values are assumed to be heterogeneous versus homogeneous. P-values estimated from Wald test. (N=27,900)



Supplementary Figure 7. Quantile-quantile plots of p-values for FG when trait values are assumed to be heterogeneous versus homogeneous. P-values estimated from Wald test. (N=23,963)



Supplementary Figure 8. Quantile-quantile plots of p-values for HbA1c when trait values are assumed to be heterogeneous versus homogeneous. P-values estimated from Wald test. (N=11,178)



Supplementary Figure 9. Quantile-quantile plots of p-values for PR interval when trait values are assumed to be heterogeneous versus homogeneous. P-values estimated from Wald test. (N=17,428)

6. Selecting Principal Components of Ancestry for Use as Covariates

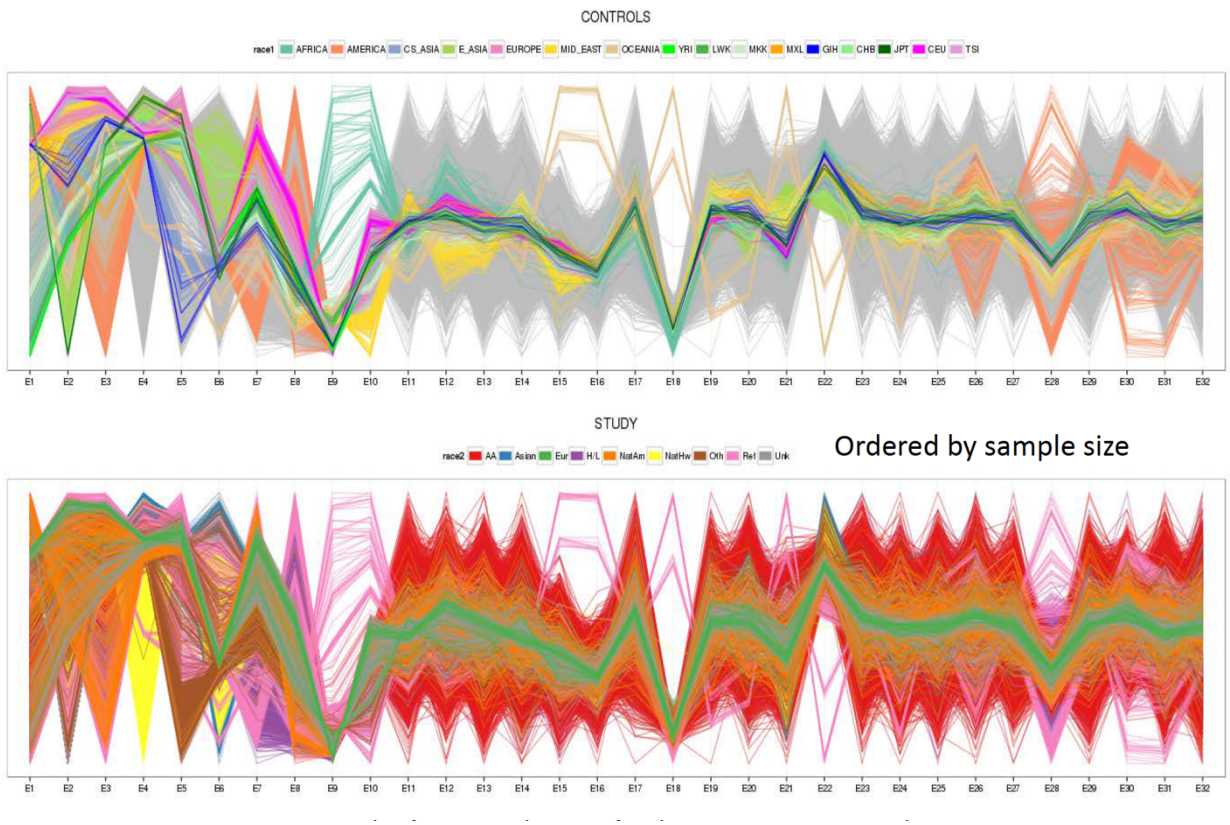
Previous GWAS analyses have generally used between 3 and 10 principal components (PCs) for adjustment for population structure, often simply determined by post-hoc interpretation of results. In our study, we used two additional criteria to calibrate the number of PCs to include in the analysis: reference population specificity (**Supp Fig 9**) and chromosome specificity (**Supp Fig 10**).

While PC1 through PC10 showed clear population specificity in the reference samples, most of the higher PCs showed much weaker population-specificity. PC9 and PC10 were clearly African in origin in the reference data (top panel), although these two PC did not vary tremendously within the PAGE study population. A few higher PC showed population specificity in the reference samples (PC15, 16, 18 and 22), but we felt it was more appropriate to specify a single threshold (exclude all PC past PC10), rather than cherry picking a PC less-relevant to the samples with phenotypic data.

An alternative approach to assessing which principal components to include in an analysis uses chromosome-specificity. Principal components which load uniformly across the genome are likely to represent ancestral populations, while principal components that load heavily onto specific chromosomes are frequently artifacts in the data (e.g. large polymorphic inversions on chromosomes 8 and 17). A loading analysis is shown in the new **Supp Fig 10**, which shows the correlation between SNP genotype and PC1 through PC20. Again, PC1 through PC8 are clearly consistent across the genome, so their inclusion as covariables is justified. In contrast, PC11 through PC20 show significant chromosome

specificity, so we chose to exclude PC11 and higher due to the lack of discernible population specificity in either the reference or study samples.

PC9 and PC10 posed a unique challenge. Population specificity (**Supp Fig 9**) suggests that these two might be important in African populations, arguing for inclusion, but the chromosome loading (**Supp Fig 10**) suggests that these two could be artifactual. We conservatively chose to retain these two PC in our analyses, but analyses with either eight or twelve PCs are almost entirely consistent with the ten PC analysis, so PC9 and PC10 were not influential in the results that are presented.



Supplementary Figure 10: Standardized principal components by population.
After standardizing the ranges of principal component 1 (PC1) through PC32, we plotted the value for each individual as a line (N=49,839). The top panel shows individuals within the reference population color coded by population, with study samples in grey. The bottom panel shows PAGE participants colored by their self-reported ancestry with the reference populations in pink. This allows us to see the distribution of different race/ethnicity groups across the different principal components. For example, in the top panel we see orange lines at the outer ranges of the distribution, indicating that principal component 3 represents the spectrum of Native American ancestry, as orange denotes reference groups from the Americas.

Supplementary Figure 11: Correlation between SNP genotype and PC, by chromosome.

Genome position is shown on the X axis, and the correlation between genotype and PC is shown on the Y axis, range (0,1) in all panels. The first eight PCs are clearly consistent across the genome, while higher PCs tend to be more chromosome-specific. (N=49,839)

7. Genome-wide Association Analysis

Analysis:

We based our analysis on generalized linear models of form

M1: $g(E(Y)) = X\alpha + G\beta$

Where Y is the vector (of length n, the number of participants used in a given analysis) of observed outcomes which may be continuous or binary, E() denotes expectation, g is a link function, X denotes a matrix of adjustment variables (of size n x p , with p the number of adjustment variables) including age, sex, principal components (PCs), self-identified ethnicity as a proxy for cultural background, and other relevant variables, and G is a vector of length n of observed or imputed allele counts for a given variant of interest.

For binary traits we used the logit link function g(x)=log(x)-log(1-x) so that it is the log odds that is linear in the G and X variables, while for continuous traits the identity function, g(x)=x, was used. Some continuous traits were inverse-normal transformed before model M1 (and other models) were applied to the sorted trait values. In particular, if Y(1),…Y(n) are the sorted trait values, then Y(i) is replaced by Φ-1(i/(n+1)), where Φ-1 is the inverse cumulative normal distribution function.

We used a combination of methods to account for hidden population structure (e.g. admixture) and relatedness among study participants. For both continuous and binary traits, we included leading PCs of the genotype matrix as part of the adjustment variables in M1 to ensure that large scale population structure would not induce false positive associations. In all our analyses, we included ten PCs in M1. Based on our assessment, ten PCs was sufficient to account for all major ethnic variation, while not including too many PCs to negatively affect the power of the analyses. Limited experimentation (not shown) suggested that adding a few more PCs did not noticeably influence the results.

For continuous traits, we adopted a linear mixed models (LMM) and a generalized estimating equations (GEE) approach to correct for the effects of relatedness between individuals. For binary traits, we only used the GEE approach. Two programs, GENESIS and SUGEN, were developed by PAGE collaborators to implement these approaches for GWAS studies of populations with genetic admixture and known or cryptic relatedness.

GENESIS: The GENESIS program [29–31] is available as a Bioconductor package made available in R, and uses a LMM to test for SNP - phenotype associations. For continuous traits, the regression models were fit assuming a variance matrix model for the variance-covariance matrix of the outcomes Y of form $\sigma^2 I + \gamma^2 K$.

Here I is the identity matrix and K is the genetic relatedness matrix computed from the available SNP data, once for all of PAGE II. Score tests of β=0 were computed by replacing M1 with the null model

M0: $E(Y) = X\alpha$

and then in M1 performing a test for $\beta = 0$ with σ^2 and γ^2 held at their estimated values from the fit of M0. The elements of G (in turn) are simply the observed allele counts (0, 1, or 2), or for imputed data the estimated allele counts (taking values from 0-2) for each of the variants of interest. Using the variance model V1 corrects the score tests of $\beta = 0$ for both close and more distant relationships between individuals in the dataset. Estimation of α and the variance parameters σ^2, γ^2 only needs to be performed once which provides a great savings in computer time needed to use GENESIS.

Both GENESIS and SUGEN rely upon the estimated relationship matrix *K*. The GENESIS package includes the program PC-Relate, which uses a principal component analysis (PCA) based method to infer genetic relatedness in samples with unspecified and unknown population structure. By using individual-specific allele frequencies estimated with sample principal component eigenvectors, it provides estimates of kinship coefficients and identity by descent (IBD) sharing probabilities in samples with population structure, admixture, and HWE departures. It does not require additional reference population panels or prior specification of the number of ancestral subpopulations.

SUGEN: The SUGEN program [32] is a command-line software program developed for genetic association analysis with complex survey sampling and relatedness patterns. It implements the generalized estimating equation (GEE) approach, which empirically accounts for within-family correlations without modeling the correlation structures of complex pedigrees.

Association analysis in SUGEN requires the study subjects to be grouped into "independent" families. There is a complex pattern of relatedness in HCHS/SOL: individuals in the same household are related, and there is endogamous mating within the Hispanic/Latino community, such that some households are connected into large pedigrees. To address this challenge, we first used the genetic relationship matrix K to identify pairs of individuals who are first-degree or second-degree relatives. We then formed extended families by connecting the households who share first-degree relatives or either first- or second-degree relatives. The trait values are assumed to be correlated within families but independent between families. In our dataset, we found it sufficient to account for first-degree relatedness in association analysis.

The GEE approach uses a "sandwich" variance estimator to empirically estimate within-family correlations. SUGEN adopts a modified version of the sandwich variance estimator, which replaces the empirical covariance matrix of the score vectors by the Fisher information matrix for unrelated subjects. This modified variance estimator is more accurate than the original sandwich variance estimator for low-frequency variants.

SUGEN can perform Wald and score tests. We used the Wald test because it yields slightly better control of the type I error than the score test. SUGEN can accommodate binary, continuous, and age-at-onset traits. When analyzing continuous traits, we allowed the trait variances to be different among different ethnic groups.

The GEE assumption of independence between families is more restrictive than the covariance model assumed in the LMM. However, the GEE approach does not rely on the normality assumption and is robust to model misspecification. While the original SUGEN version had general methodology, the software has been extended to handle heterogeneous variances for the PAGE analyses (https://github.com/dragontaoran/SUGEN#). In our dataset, SUGEN provides very accurate control of the type I error, as judged by the QQ-plots and genomic control parameter. We used SUGEN as the primary method for association testing and ran GENESIS in parallel for comparison purposes (**Supplemental Table 2, Supplemental Figure 5**).

Genetic Ancestry Interactions: Besides simple tests of association, we were also interested in whether allelic effects differ according to ethnicity or ancestral makeup. These presence of these effects can be estimated by adding SNPxPC interaction effects into model M1 to form

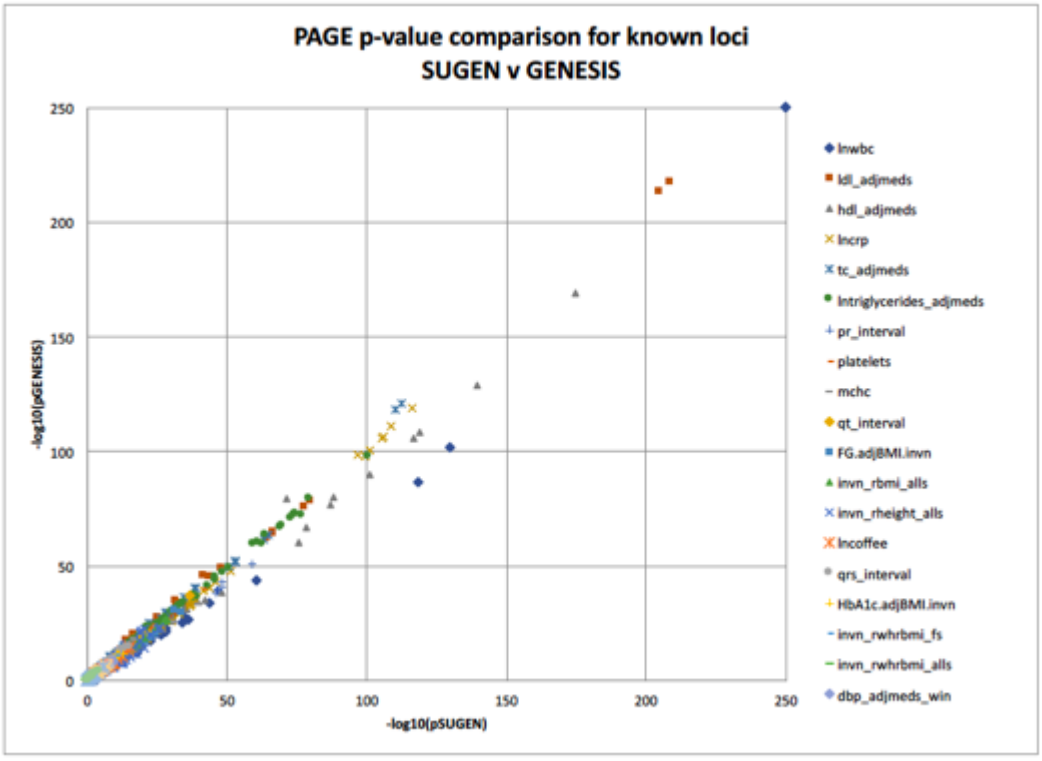
M2: $g(E(Y)) = X\alpha + G\beta + (Gx)\theta$

and then testing that $\theta = 0$, generally this will yield a multi degree of freedom test (F-test) for heterogeneity of effects depending upon how many PCs are included in the hypothesis tests. We choose to include interaction terms for all 10 PCs to account for the sub-continental differences that were differentiated in the higher PCs. We explored using a smaller number of PCs in M2, but found that the P-values for the F-test obtaned with 5 PCs were extremely similar to those with 10 PCs for the vast majority of SNPs (results not shown).

Assessing Single Variant Results: For each phenotype, QQ plots and genomic inflation factors (λ) were used to assess inflation, using the full set of results, and results omitting previously known loci. Inflation values ranged from 0.98 to 1.15 for all traits. Analyses were restricted to SNPs with an imputation quality score greater than 0.4 and an effective sample size (effN) greater than 30 for continuous traits, and greater than 50 for binary traits. The effN was calculated based on previous publications[24]:

effN = 2*MAF*(1-MAF)*N*info

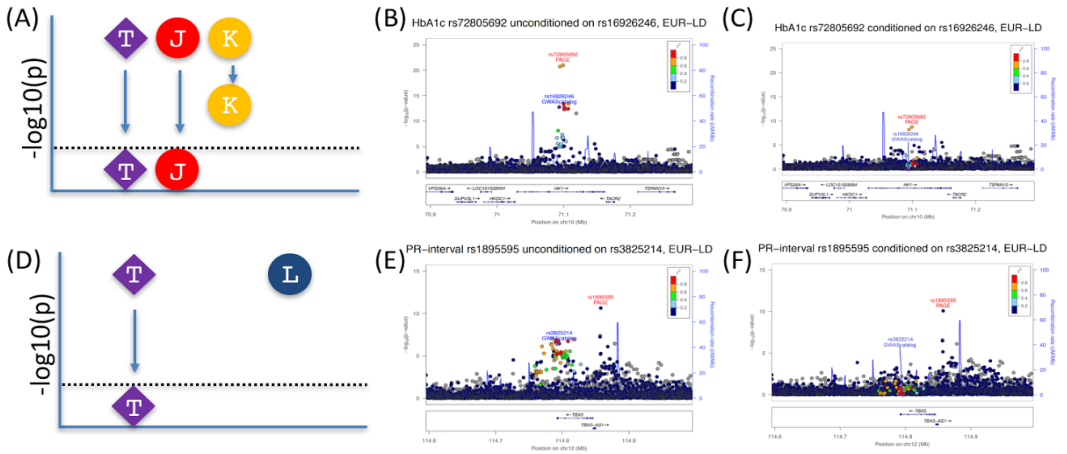
where MAF is the minor allele frequency among the set of individuals included in a phenotype-specific model, N is the total sample size for a given phenotype, and info is the impute2 info quality score. The SNP with the smallest p-value in a 1Mb region was considered the Lead SNP. A Lead SNP was considered to be a Novel loci if it met the following criteria: 1) it was located greater than +/- 500 Kb away from a previously known loci (per the phenotype-specific Known Loci list); 2) it had a SUGEN p-value less than 5E-08; 3) it had a SUGEN conditional p-value less than 5E-08 after adjustment for all previously known loci on the same chromosome; and 4) it had 2 or more neighboring SNPs (within +/- 500 Kb) with a p-value less than 1E-05. A Lead SNP was considered to be a Residual signal in a previously known loci if it met the following criteria: 1) it was located within +/- 500 Kb of a previously known loci; 2) it had a SUGEN p-value less than 5E-08; and 3) it had a SUGEN conditional p-value less than 5E-08 after adjustment for all previously known loci on the same chromosome. Full results for all Novel and Residual findings are included in **Supplemental Tables 2-3**. Additionally, minor allele frequency-dependent thresholds were used for genome-wide significance, as per guidelines in Fadista et al (2016). [33] For common variants with MAF>5%, the standard $P<5x10^{-8}$ threshold was used to determine significance. For low frequency and rare variants with MAF<5%, a more stringent $P<3x10^{-9}$ was utilized. This is reflected in the 16 novel genome-wide significant trait-variant associations and the 11 low-frequency loci with suggestive associations ($3x10^{-9}>P>5x10^{-8}$).



Supplementary Figure 12: Comparison of P-values from GWAS for SUGEN (Wald test) vs. GENESIS across all traits. (N_{max}=49,781; see Extended Data Table 1)

8. Secondary Signals versus Fine-mapping

To further illustrate the difference in mechanism between fine-mapping and secondary independent signals, we highlight two examples **(Supplementary Figure 12).** The first is a refinement of the association between hexokinase 1 (*HK1*) and HbA1c. The residual signal at rs72805692 (P_{unadj}=9.22x10^{-22}, N=11,178, CAF=0.061) is in moderate LD in European (r^2=0.61) and Hispanic/Latino (r^2=0.63) populations with the previously implicated SNP (rs16926246) 5.7kb away. Therefore, after adjustment, the signal is greatly diminished but remains statistically significant (P_{cond}=3.05x10^{-9}). This represents the refinement of a known locus (fine-mapping), as the high LD present in this area results in an attenuated, but still statistically significant, signal, and may represent only one underlying fSNP. In contrast, we found a residual signal for PR interval at rs1895595, upstream of *TBX5* (P_{unadj}=2.16x10^{-11}, N=17,428, CAF=0.17). After adjustment for 5 known tagSNPs in this region (rs3825214, rs7312625, rs7135659, rs1895585, rs1896312), the signal remains largely unchanged (P_{cond}=1.99x10^{-11}). This secondary signal at rs1895595 is independent of all 5 conditioned SNPs, with extremely low LD (r^2<0.03) across all global populations, and therefore likely represents an independent fSNP. Both fine-mapping of primary findings and knowledge of independent, secondary alleles are important to comprehensively characterize GWAS loci, particularly in diverse populations, thereby improving genetic risk prediction.

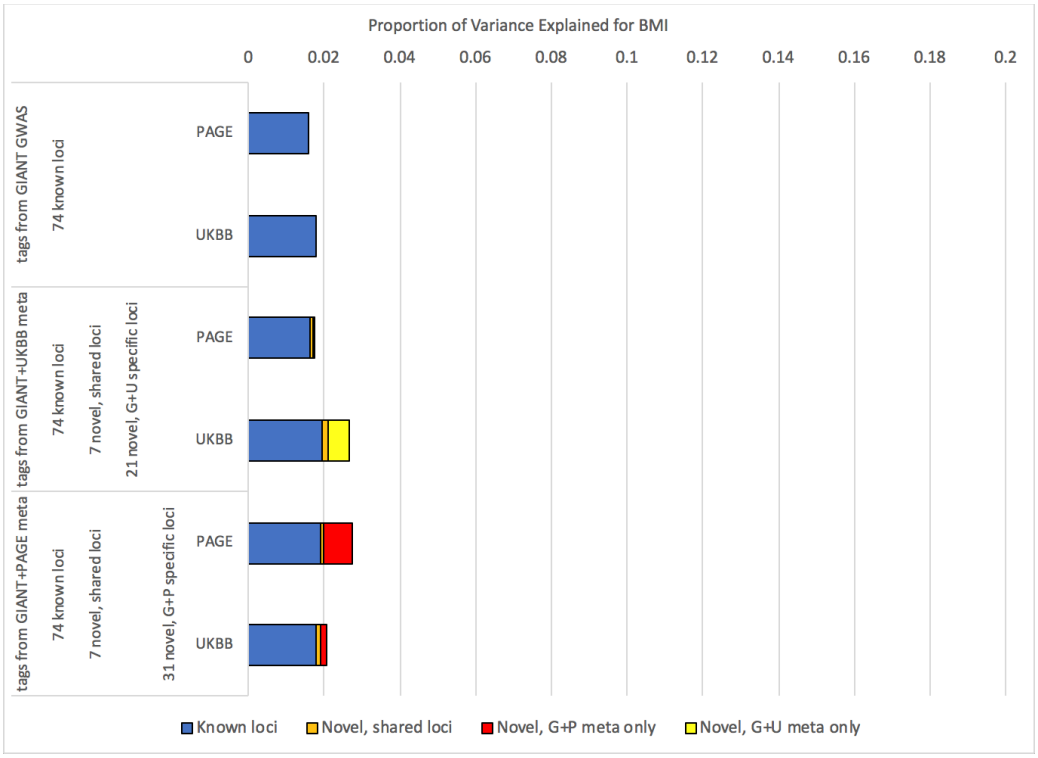


Supplementary Figure 13: Residual signals can represent either refinement of signal or secondary alleles.
*(**A**) Fine-mapping: -log10 p values from SUGEN Wald test are plotted against position for a GWAS catalog tagSNP T, as well as two tagged SNPs: J is strongly tagged by T (r^2=1) in all populations, and K is variably tagged across populations. After adjustment, signal at T and J is no longer significant, but residual signal at K indicates that the original association has been fine-mapped. Unadjusted (**B**) and adjusted (**C**) results for trait HbA1c (N=11,178), showing weakened signal at residual SNP rs72805692 after adjusting for GWAS catalog tagSNP rs16926246, consistent with signal refinement. This tagSNP was first reported from a study of 46,368 Europeans[34], so LD with the tagSNP is shown from a European reference panel, illustrating how the set of strongly tagged SNPs (red/orange) is fine-mapped to the two strongest (residual) signals in the multi-ethnic population. (**D**) Secondary alleles are independent of known loci, so L is not in significant LD with T (r^2 ~ 0). After adjustment for T, signal at L is unchanged. Unadjusted (**E**) and adjusted (**F**) results for trait PR interval (N=17,428), showing no change in signal at residual SNP rs1895595 after adjusting for GWAS catalog tagSNP rs3825214, consistent with the residual signal being an independent secondary allele. Again, LD shown is from a European population, as the GWAS catalog report [35] was from 12,670 Europeans. P-values estimated from SUGEN Wald test.*

9. Meta-analysis and Finemapping with GIANT, UK Biobank

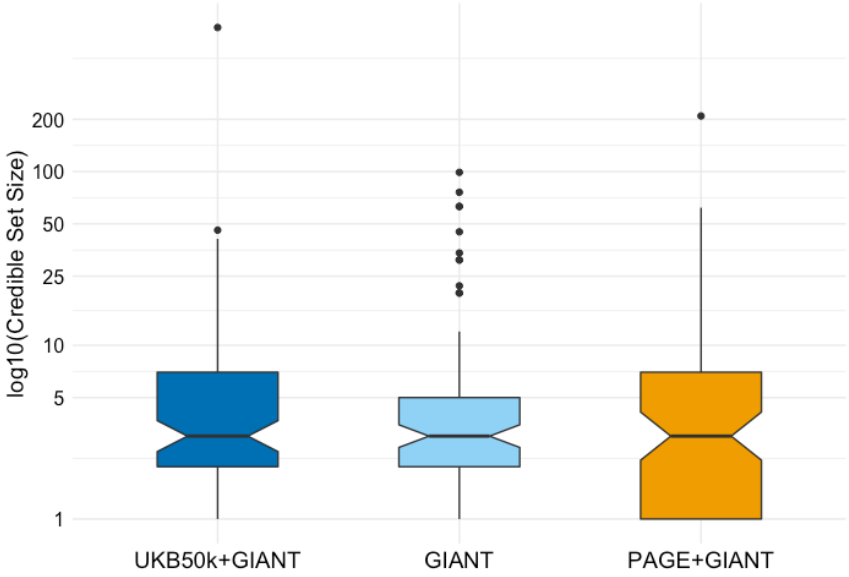
For height [36], GIANT imputed their GWAS data to the HapMap in roughly 250,000 individuals, yielding 2.5M variants that overlapped with the PAGE dataset. All of these are common (MAF>5%) in at least one ancestry, so the traditional threshold of statistical significance ($P<5x10^{-8}$) is appropriate. In the GIANT BMI manuscript [37], the GWAS data were augmented with metabochip data (a focused platform targeting specific regions of the genome) from ~80,000 additional individuals. The previously published manuscripts used a more relaxed definiton of "locus" than we have in this manuscript (associations less than 1Mbp apart were merged into a single locus, where we have used 500kbp), and also reported results from multiple analytic approaches (by subset, or conditioned on known loci). For clarity in our comparison, we use the same locus definition as we used earlier in this manuscript, and we limit the comparison to a single approach: the sex-combined joint analysis of all European individuals in GIANT, meta-analyzed with the sex-combined SUGEN results from either PAGE or the UK Biobank.

To create a comparable sample size as PAGE, a total of 50,000 "White British" individuals were randomly subset from the larger UK Biobank (UKB50k). Height and BMI was adjusted for both sex and age identically to PAGE procedures. The linkage disequilibrium information for GIANT was generated from 9,700 ARIC individuals who were part of the larger GIANT consortium. For PAGE, correlation between sites were calculated separately for each race/ethnicity group. These matrices were then combined, weighted by the inverse sample size to create a combined weighted correlation matrix of sites. For the meta-analyses, the correlation matrices were again combined, weighted by the inverse of the sample size proportional to 250k GIANT, 50k PAGE, and 50k "White British" UK Biobank participants.

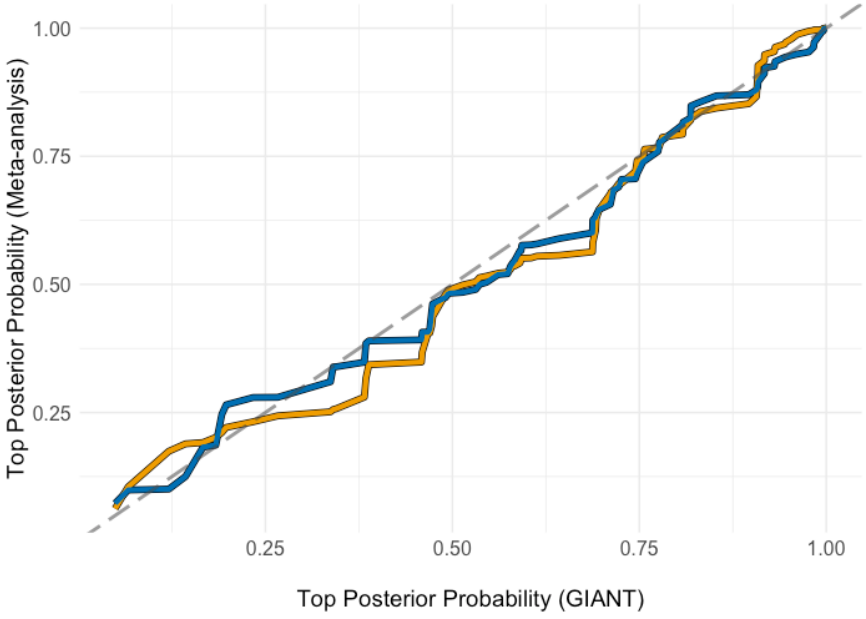


Supplementary Figure 14: BMI PVE.
Although less of the variance is explained for BMI than for height, results are broadly consistent: meta-analysis with more Europeans (GIANT+UKB) exacerbates existing disparities in PVE between Europeans and a multi-ethnic cohort (center pair of bars), while in this case, meta-analysis with the multi-ethnic cohort (GIANT+PAGE) actually yields improved PVE in the multi-ethnic cohort.

A



B

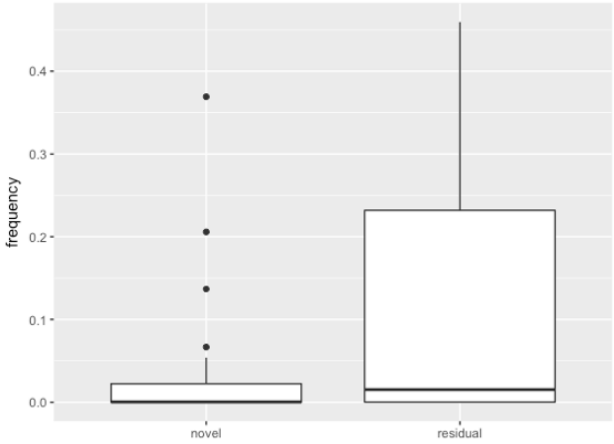


Supplementary Figure 15: Finemapping for BMI.

(A) Comparison of 95% credible sets for height, comparing GIANT alone (N=253,288) to UKB50k+GIANT (N=303,288; paired sample t-test *P*=0.60) and PAGE+GIANT (N=303,069; paired sample t-test *P*=0.50). Boxplots show the median at the notch, with the top and bottom of the box indicating the interquartile range (IQR). Whiskers extend to either the minimum value or 1.5*IQR. Notches indicate the 95% confidence interval of the medians. (B) Top posterior probability from each 95% credible set for height, comparing GIANT (N=253,288) to UKB50k+GIANT (N=303,288) and PAGE+GIANT (N=303,069).

10. Comparison of novel and secondary variant allele frequencies in European populations

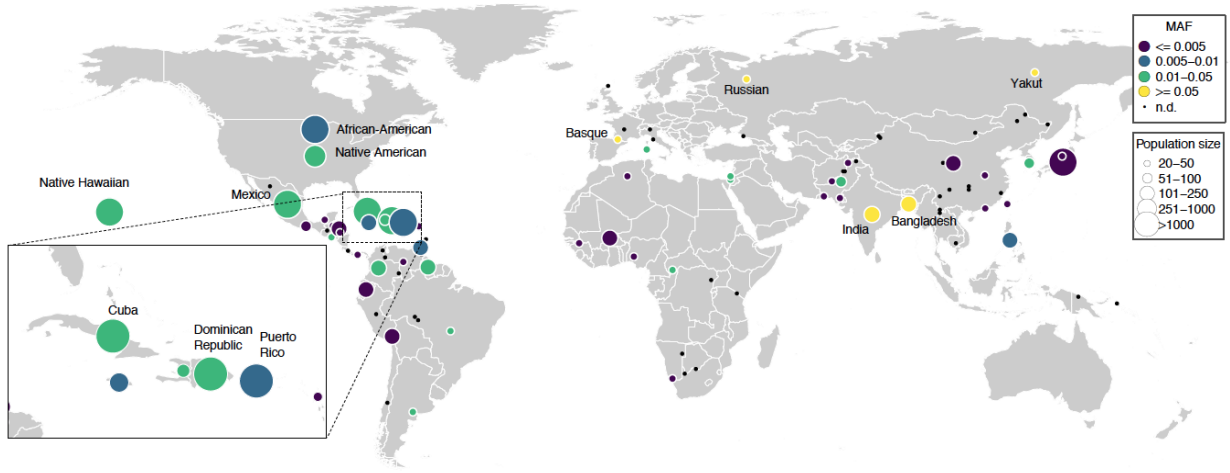
To interrogate the possibility of discovery of our novel and secondary findings in a European ancestry sample, we downloaded the vcfs from the gnomAD browser (https://gnomad.broadinstitute.org) and extracted the sites that matched our PAGE hits for the non-Finnish European group (NFE), as the largest public repository of European-derived allele frequencies. Of these, we identified 24 novel and 35 residual sites that were biallelic, did not contain repeat motifs, and within the callability mask. As can be seen in the boxplots below, novel sites had significantly lower allele frequencies than secondary sites (median minor allele frequency: 0.0050 vs 0.015, Wilcoxon p: 0.03). We also observed a weakly significant increase in sites with a measured MAF of 0 in NFE: 7/24 vs 3/35 in novel and secondary, respectively: logistic regression *P*=0.05, OR=4.4, 95% CI: 1.1-22, reflecting the small sample size of novel and secondary sites. However as can be seen in the plot, even in the novel findings there are common variants (maximum MAF: rs6730558, MAF~37%), indicating that lead variants still require fine mapping to uncover the causal signals as described above.



Supplementary Figure 16: European allele frequencies of novel and secondary findings in PAGE.
Here we show the distribution of allele frequencies in the Non-Finnish European group in the gnomAD browser (N=63,369) for our novel and secondary findings, demonstrating the preponderance of low frequency variants in European populations which are now adequately powered in PAGE groups. The median is denoted in bold with the top and bottom indicating the interquartile range (IQR). The whiskers denote 1.5*IQR or the minimum/maximum value, with outliers displayed as dots.

11. Clinically-relevant variants and their distribution in PAGE

We also investigated the HLA-B*57:01 allotype, which interacts with the HIV drug abacavir to trigger a potentially life-threatening immune response [38–41] and therefore is recommended by the FDA for screening prior to treatment initiation [42]. The rs2395029 (G) variant in *HCP5* is used to screen for abacavir hypersensitivity [43], as it is a near perfect tag of *HLA-B*57:01* in Europeans and has utility (r ~ 0.92, [44]) across globally diverse populations in the 1000 Genomes Project. Using PAGE and Global Reference Panel samples, we show that risk allele frequencies for rs2395029 rise above 5% in multiple large South Asian populations, and above 1% within some admixed populations with Native American ancestry (**Figure 4**). PAGE allele frequencies can therefore aid in expanding the reach of precision medicine to encompass individuals of diverse ancestry, particularly when combined with other studies. [17,45]



Supplementary Figure 17: World map of *HCP5*-G frequencies within PAGE groups.
The histocompatibility protein variant HLA-B*57:01 interacts with the HIV drug abacavir to stimulate a hypersensitivity response. A variant in a gene near HLA-B, HCP5 rs2395029 (G allele), can be used to genotype for the -B*57:01 allele as it is in high linkage disequilibrium (correlation ~0.92 in 1000 Genomes Phase 1).[43,44,46,47]. This HCP5 tag-SNP segregates within all continental populations of the PAGE study, providing increased resolution of the global haplotype frequency, particularly within Latin America. Above, minor allele (G) frequency is shown. Population size is indicated by the radius of the circle. Black dot (MAF not displayed): population has less than twenty individuals or the variant is a singleton in that population.

12. Additional Acknowledgements

BioMe Biobank: Samples and data of The Charles Bronfman Institute for Personalized Medicine (IPM) BioMe Biobank used in this study were provided by The Charles Bronfman Institute for Personalized Medicine at the Icahn School of Medicine at Mount Sinai (New York). Phenotype data collection was supported by The Andrea and Charles Bronfman Philanthropies.

HCHS/SOL: The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.

PAGE Global Reference Panel: The specific datasets are:

Mexico: Samples of indigenous origin in Oaxaca were kindly provided by co-authors, and Samuel Canizales Quinteros and Victor Acuña Alonzo. Peru: Individuals from a primarily Quechuan and Aymaran-speaking community in Puno were kindly provided with funding support from the Burroughs Welcome Fund. Rapa Nui (Easter Island, Chile): Samples were kindly provided with funding from the Charles Rosenkranz Prize for Health Care Research in Developing Countries and the International Center for Genetic Engineering and Biotechnology (ICGEB) Grant CRP/MEX15-04_EC awarded to A.M.-E. South Africa: Samples of KhoeSan individuals from the ‡Khomani and Nama communities were kindly provided with funding from the Morrison Institute for Population and Resource Studies. Honduras and Colombia: Samples from communities in Honduras and Colombia were kindly provided by co-authors, Edwin Herraro-Paz (Universidad Católica de Honduras, San Pedro Sula, Honduras), Alvaro Mayorga (Universidad Católica de Honduras, San Pedro Sula, Honduras), Luis Caraballo (University of Cartagena), Javier Marrugo (university of Cartagena) Additional global samples: The following datasets are open access and available through the lab website of Carlos Bustamante (https://bustamantelab.stanford.edu/). The Human Genome Diversity Panel (HGDP-CEPH) is a group of cell lines maintained by the Centre d'Étude du Polymorphisme Humain, Fondation Jean Dausset (Paris, France) comprising 52 diverse populations across the world (Africa, Near East, Europe, South Asia, Central Asia, East Asia, Oceania and the Americas). Additional information on these datasets can be found on the CEPH website (http://www.cephb.fr/en/hgdp_panel.php), or originally at http://www.ncbi.nlm.nih.gov/pubmed/11954565 and http://www.ncbi.nlm.nih.gov/pubmed/12493913, with numerous subsequent publications. Samples were filtered to include the H952 unrelated individuals as published here: http://www.ncbi.nlm.nih.gov/pubmed/17044859. Also available on the Bustamante Lab website is genotype data for the Maasai from Kinyawa, Kenya (MKK) samples maintained by the Coriell Institute for Medical Research (https://catalog.coriell.org/1/NHGRI/Collections/HapMap-Collections/Maasai-in-Kinyawa-Kenya-MKK) and genotyped as part of the International HapMap Project Phase 3(http://hapmap.ncbi.nlm.nih.gov/, http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html) . We have genotyped a subset of unrelated individuals using the filters recommended in http://www.ncbi.nlm.nih.gov/pubmed/20869033.

WHI: The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A listing of WHI investigators can be found at: https://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf

Supplementary Information Bibliography

1. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15,** 761–771 (2013).
2. Sorlie, P. D. *et al.* Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* **20,** 629–641 (2010).
3. Lim, U. *et al.* Asian women have greater abdominal and visceral adiposity than Caucasian women with similar body mass index. *Nutr. Diabetes* **1,** e6 (2011).
4. Kolonel, L. N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151,** 346–357 (2000).
5. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–262 (2002).
6. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control. Clin. Trials* **19,** 61–109 (1998).
7. Levey, A. S. *et al.* A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150,** 604–612 (2009).
8. Stevens, L. A. *et al.* Evaluation of the Chronic Kidney Disease Epidemiology Collaboration equation for estimating the glomerular filtration rate in multiple ethnicities. *Kidney Int.* **79,** 555–562 (2011).
9. Kono, T. *et al.* Differential effects of cyclosporine A on ornithine decarboxylase activity induced by ultraviolet-B and PUVA in mouse skin. *J. Invest. Dermatol.* **96,** 871–874 (1991).
10. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106,** 3143–3421 (2002).
11. Sweeney, M. E. & Johnson, R. R. Ezetimibe: an update on the mechanism of action, pharmacokinetics and recent clinical trials. *Expert Opin. Drug Metab. Toxicol.* **3,** 441–450 (2007).
12. Eaton, C. Selective cholesterol absorption inhibitors have effects on LDL-C very similar to bile sequestrants. (2013).
13. Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18,** 499–502 (1972).
14. Galanter, J. M. *et al.* Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. *J. Allergy Clin. Immunol.* **134,** 295–305 (2014).
15. Wojcik, G. L. *et al.* Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies. *G3 (Bethesda)* (2018). doi:10.1534/g3.118.200502
16. Bien, S. A. *et al.* Strategies for enriching variant coverage in candidate disease loci on a multiethnic

genotyping array. *PLoS ONE* **11,** e0167758 (2016).
17. Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA* **108,** 5154–5162 (2011).
18. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
19. Drake, K. A. *et al.* A genome-wide association study of bronchodilator response in Latinos implicates rare variants. *J. Allergy Clin. Immunol.* **133,** 370–378 (2014).
20. Uren, C. *et al.* Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. *Genetics* **204,** 303–314 (2016).
21. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34,** 591–602 (2010).
22. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9,** 179–181 (2011).
23. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
24. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
25. Zhu, X., Li, S., Cooper, R. S. & Elston, R. C. A unified association analysis approach for family and unrelated samples correcting for stratification. *Am. J. Hum. Genet.* **82,** 352–365 (2008).
26. Lin, D. Y. & Zeng, D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epidemiol.* **34,** 60–66 (2010).
27. Lin, D. Y. & Zeng, D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97,** 321–332 (2010).
28. Zhang, J. & Stram, D. O. The role of local ancestry adjustment in association studies using admixed populations. *Genet. Epidemiol.* **38,** 502–515 (2014).
29. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* **39,** 276–293 (2015).
30. Conomos, M. P. *et al.* Genetic diversity and association studies in US hispanic/latino populations: applications in the hispanic community health study/study of latinos. *Am. J. Hum. Genet.* **98,** 165–184 (2016).
31. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *Am. J. Hum. Genet.* **98,** 127–148 (2016).
32. Lin, D.-Y. *et al.* Genetic association analysis under complex survey sampling: the Hispanic Community Health Study/Study of Latinos. *Am. J. Hum. Genet.* **95,** 675–688 (2014).
33. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24,** 1202–1205 (2016).

34. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathways. *Diabetes* **59,** 3229–3239 (2010).
35. Holm, H. *et al.* Several common variants modulate heart rate, PR interval and QRS duration. *Nat. Genet.* **42,** 117–122 (2010).
36. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
37. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).
38. Mallal, S. *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358,** 568–579 (2008).
39. Sousa-Pinto, B. *et al.* Pharmacogenetics of abacavir hypersensitivity: A systematic review and meta-analysis of the association with HLA-B*57:01. *J. Allergy Clin. Immunol.* **136,** 1092–4.e3 (2015).
40. Hetherington, S. *et al.* Hypersensitivity reactions during therapy with the nucleoside reverse transcriptase inhibitor abacavir. *Clin. Ther.* **23,** 1603–1614 (2001).
41. Illing, P. T. *et al.* Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature* **486,** 554–558 (2012).
42. Dean, L. in *Medical Genetics Summaries* (eds. Pratt, V., McLeod, H., Dean, L., Malheiro, A. & Rubinstein, W.) (National Center for Biotechnology Information (US), 2012).
43. Martin, M. A. *et al.* Clinical Pharmacogenetics Implementation Consortium Guidelines for HLA-B Genotype and Abacavir Dosing: 2014 update. *Clin. Pharmacol. Ther.* **95,** 499–500 (2014).
44. Pappas, D. J. *et al.* Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. *Pharmacogenomics J.* **18,** 367–376 (2018).
45. Baker, J. L., Shriner, D., Bentley, A. R. & Rotimi, C. N. Pharmacogenomic implications of the evolutionary history of infectious diseases in Africa. *Pharmacogenomics J.* **17,** 112–120 (2017).
46. Colombo, S. *et al.* The HCP5 single-nucleotide polymorphism: a simple screening tool for prediction of hypersensitivity reaction to abacavir. *J. Infect. Dis.* **198,** 864–867 (2008).
47. Sanchez-Giron, F. *et al.* Association of the genetic marker for abacavir hypersensitivity HLA-B*5701 with HCP5 rs2395029 in Mexican Mestizos. *Pharmacogenomics* **12,** 809–814 (2011).