# The Phenotypic Consequences of Genetic Divergence between Admixed Latin American Populations: Antioquia and Chocó, Colombia

Aroon T. Chande[1,2,3], Lavanya Rishishwar[2,3], Dongjo Ban[1,3], Shashwat D. Nagar[1,3], Andrew B. Conley[2,3], Jessica Rowell[1], Augusto E. Valderrama-Aguirre[3,4,5], Miguel A. Medina-Rivas[3,6], and I. King Jordan[1,2,3,]*

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia

[2]IHRC-Georgia Tech Applied Bioinformatics Laboratory, Atlanta, Georgia

[3]PanAmerican Bioinformatics Institute, Valle del Cauca, Cali, Colombia

[4]Biomedical Research Institute (COL0082529), Cali, Colombia

[5]Universidad Santiago de Cali, Colombia

[6]Centro de Investigación en Biodiversidad y Hábitat, Universidad Tecnológica del Chocó, Quibdó, Colombia

*Corresponding author: E-mail: king.jordan@biology.gatech.edu.

**Accepted:** 12 July 2020

## Abstract

Genome-wide association studies have uncovered thousands of genetic variants that are associated with a wide variety of human traits. Knowledge of how trait-associated variants are distributed within and between populations can provide insight into the genetic basis of group-specific phenotypic differences, particularly for health-related traits. We analyzed the genetic divergence levels for 1) individual trait-associated variants and 2) collections of variants that function together to encode polygenic traits, between two neighboring populations in Colombia that have distinct demographic profiles: Antioquia (*Mestizo*) and Chocó (Afro-Colombian). Genetic ancestry analysis showed 62% European, 32% Native American, and 6% African ancestry for Antioquia compared with 76% African, 10% European, and 14% Native American ancestry for Chocó, consistent with demography and previous results. Ancestry differences can confound cross-population comparison of polygenic risk scores (PRS); however, we did not find any systematic bias in PRS distributions for the two populations studied here, and population-specific differences in PRS were, for the most part, small and symmetrically distributed around zero. Both genetic differentiation at individual trait-associated single nucleotide polymorphisms and population-specific PRS differences between Antioquia and Chocó largely reflected anthropometric phenotypic differences that can be readily observed between the populations along with reported disease prevalence differences. Cases where population-specific differences in genetic risk did not align with observed trait (disease) prevalence point to the importance of environmental contributions to phenotypic variance, for both infectious and complex, common disease. The results reported here are distributed via a web-based platform for searching trait-associated variants and PRS divergence levels at http://map.chocogen.com (last accessed August 12, 2020).

**Key words:** polygenic, traits, disease, health, genetic ancestry, population genomics.

## Significance

An understanding of how trait-associated genetic variants are distributed within and between populations can provide insight into the genetic underpinnings of human phenotypic diversity, particularly for health-related traits that show disparate impacts. We addressed this issue by analyzing the distributions of trait-associated variants in diverse Colombian populations: Antioquia (*Mestizo*) and Chocó (Afro-Colombian). We found that genetic ancestry differences between the two Colombian populations affected the presence of trait-associated variants in a way that largely reflected observable anthropometric differences and reported disease prevalences. The importance of environmental contributions to human phenotypic variance—for both infectious and complex, common disease—was underscored by cases where genetically predicted trait differences between populations did not align with observed phenotypic differences.

## Introduction

The genetic basis of human phenotypic diversity is both an issue of fundamental evolutionary interest and critical to a deeper understanding of health disparities. Early genetic linkage analyses, and more recent genome-wide association studies (GWAS), have uncovered thousands of genetic variants that are associated with a wide variety of human traits (Amberger et al. 2015; MacArthur et al. 2017). Investigations of how trait-associated genetic variants are distributed within and between populations have the potential to shed light on the genetic architecture of human phenotypic diversity, particularly as related to disease prevalence disparities (Corona et al. 2013; Chande et al. 2018).

The power of this approach has long been apparent for single locus traits. Population-specific distributions of rare and highly penetrant variants that cause Mendelian diseases are responsible for a wide variety of population health disparities, such as sickle cell anemia (OMIM: 603903), cystic fibrosis (OMIM: 219700), and Tay–Sachs disease (OMIM: 272800). Of course, the vast majority of human traits are encoded by multiple loci, each of which contributes only a small fraction of the total trait variance (Visscher et al. 2017). Individuals' genomic predispositions to such multilocus traits can be captured by polygenic risk scores (PRS)—also known as polygenic trait scores, genome-wide risk scores, or genetic risk scores—which are calculated as (weighted) sums of the total number of trait-associated or trait-increasing alleles present in the genome (Chatterjee et al. 2016; Lambert et al. 2019). Changes in PRS distributions across populations have been taken as evidence of polygenic selection on a number of anthropometric (Turchin et al. 2012; Racimo et al. 2018; Berg et al. 2019), neurological (Beiter et al. 2017), and disease-related traits (Berg and Coop 2014).

Despite their apparent potential for discovering genetic changes that underlie phenotypic divergence among populations, recent studies have underscored a number of challenges entailed by cross-population comparisons of PRS. Systematic differences in allele frequencies, proportions of ancestral versus derived alleles, and patterns of linkage disequilibrium (LD) can yield large shifts in PRS distributions that do not necessarily reflect observed phenotypic differences among populations (Martin et al. 2017; Kim et al. 2018; Novembre and Barton 2018). Furthermore, the fact that the vast majority of GWAS have been conducted on cohorts of European ancestry (Need and Goldstein 2009; Bustamante et al. 2011; Popejoy and Fullerton 2016) yields PRS that are far more accurate for European populations compared with other, less-studied global population groups (Martin et al. 2019). In light of these challenges, the goals of this study were to: 1) characterize the genetic ancestry patterns for diverse populations from within a single Latin American country, 2) evaluate the impact of ancestry differences between these populations on the genetic variants associated with anthropometric and disease traits, and 3) consider observed differences in the frequencies of trait-associated variants in light of known phenotypic differences between the populations.

Recently admixed populations hold great promise for studies aimed at characterizing the genetic basis of phenotypic divergence (Winkler et al. 2010), but studies of cross-population PRS have yet to focus explicitly on admixed populations. Furthermore, studies of this kind have not focused on diverse populations that often coexist in close physical proximity in the modern world. Our research group is focused on the study of admixed American populations, with the broad aim of relating differences in ancestry to genetic determinants of health-related phenotypes (Rishishwar, Conley, et al. 2015, Rishishwar et al. 2015; Norris et al. 2018, 2019; Jordan et al. 2019; Nagar et al. 2019). Latin American populations are particularly interesting for studies of this kind given their high levels of genetic admixture among ancestral African, European, and Native American population groups (Bryc et al. 2010; Moreno-Estrada et al. 2013; Ruiz-Linares et al. 2014; Homburger et al. 2015). Populations within and between Latin American countries are characterized by different levels of continental and regional ancestry. We have been studying two neighboring populations from Colombia—Antioquia and Chocó—that are distinguished by a combination of close proximity and divergent demographic profiles. We previously found that sample donors from Antioquia show primarily European genetic ancestry, whereas donors from Chocó show majority African ancestry (Medina-Rivas et al. 2016; Conley et al. 2017), and we showed that this divergent genetic ancestry, and the allele frequency differences that it entails, lead to an increase in the predicted risk of type 2 diabetes (T2D) in Chocó compared with Antioquia (Chande et al. 2017). T2D is an intensively studied disease, and this pattern of greater predicted T2D risk in Chocó holds irrespective of the ancestry of the GWAS cohorts used for risk allele discovery (Chande et al. 2020). For this study, we performed a broader survey of the genetic divergence levels for trait-associated variants and differences in PRS for these two admixed Colombian populations, and we considered the results of these comparisons in light of known (observable) anthropometric and disease prevalence profiles for these two populations.

## Materials and Methods

### Genomic Data

The sources of genomic data used for this study are shown in supplementary table 1, Supplementary Material online. Whole-genome genotype data for the population of Chocó, Colombia were taken from the ChocoGen research project https://www.chocogen.com (last accessed August 12, 2020) (Medina-Rivas et al. 2016; Conley et al. 2017). The ChocoGen project was conducted with the approval of the Ethics Committee of the Universidad Tecnológica del Chocó

(ACTA No. 01-v1) following the Helsinki ethical principles for medical research involving human subjects. All sample donors signed informed consent documents. Whole-genome sequence data for the population of Antioquia, Colombia were taken from the phase 3 data release of the 1000 Genomes Project Consortium (2015). The 1000 Genomes Project human genome sequence data are deidentified and made publicly available for research use without restriction.

Whole-genome sequence and genotype data for continental reference populations from Africa, the Americas, and Europe were taken from the 1000 Genomes Project and from a collection of previously characterized Native American populations (Reich et al. 2012). The Native American genotype data are deidentified and made publicly available for research according to the terms of a data use agreement from the Universidad de Antioquia. A list of all bioinformatics programs and databases used for the analyses is shown in supplementary table 2, Supplementary Material online.

## Genetic Ancestry Analysis

Whole-genome genotype and sequence variant data were merged using PLINK version 1.9 (Chang et al. 2015), with single nucleotide polymorphisms (SNPs) common to all three data sources retained for subsequent analysis and SNP strand orientations corrected as needed. The merged SNP set was phased using ShapeIT version 2.r837 with the 1000 Genomes Project haplotype reference panel (Delaneau et al. 2013, 2014), and PLINK was used to prune linked SNPs from the phased genotype data set with an $r^2$ threshold of 0.1. The merged and pruned SNP set was used to infer three-way continental ancestry ($f_{\mathrm{African}}$, $f_{\mathrm{European}}$, and $f_{\mathrm{NativeAmerican}}$) for Antioquia and Chocó using the program ADMIXTURE version 1.3.0 (Alexander et al. 2009) run in unsupervised mode, with $K = 3$ continental ancestral groups corresponding to the African, European, and Native American reference populations shown in supplementary table 1, Supplementary Material online. SNP allele frequency differences and fixation index ($F_{\mathrm{ST}}$) values between Antioquia and Chocó were computed from the merged SNP set using PLINK. $F_{\mathrm{ST}}$ values were calculated using the Weir and Cockerham estimator (Weir and Cockerham 1984). Ternary plots were constructed using the inferred global ancestry fractions for each individual and the position of each individual (point) within the triangle is a composition of the individual's three ancestry components: $\left(\frac{1}{2}\cdot\frac{2A+N}{E+A+N}, \frac{\sqrt{3}}{2}\cdot\frac{N}{E+A+N}\right)$, where $E$, $A$, and $N$ are the inferred European, African, and Native American ancestry components.

## SNP Trait Associations and Polygenic Scores

SNP trait associations were taken from the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/; last accessed August 12, 2020) (Buniello et al. 2019), with the SNP rsID number, effect allele, effect size, and study population recorded for all associations. Effect alleles are operationally defined as the allele for any given SNP that is associated with cases, for case–control GWAS, or with an increase in the trait under consideration for quantitative trait GWAS. The SNP associations used here are limited to biallelic variants, do not include SNP interactions, and are all significant at $P < 1 \times 10^{-5}$ (number of SNPs = 107,784). SNP associations were grouped into polygenic traits using the NHGRI-EBI GWAS Catalog trait terms (number of traits = 2,382), which are derived from the EBI Experimental Factor Ontology (https://www.ebi.ac.uk/efo/; last accessed August 12, 2020) (Malone et al. 2010). After filtering, 65,283 (60.5%) SNPs remained. Of the 42,501 (39.5%) associations excluded: 25,305 (23.5%) had an unknown or unreported effect allele (effect allele = "?"); 14,615 (13.5%) had multiple reported effect alleles for the same trait and reported effect alleles were not strand flips (i.e., A and C); and 2,581 (2.4%) had no associated rsID (i.e., the variant is given by chromosomal location, chr1:2345).

Whole-genome genotype data from Chocó were imputed up to 1000 Genomes phase 3 variant calls using the program IMPUTE2 version 2.3.2 (Howie et al. 2011, 2012) and the 1000 Genomes Project haplotype reference panel. Imputed sites were retained for subsequent analysis if they had a 95% imputation rate across samples and an INFO score >0.4. The imputed data from Chocó were merged with the whole-genome sequence variant data from Antioquia using PLINK.

PRSs, also referred to as polygenic trait scores, were computed for each GWAS trait $i$ as the sum of the effect alleles across all trait-associated SNPs as previously described (Chande et al. 2018):

$$\mathrm{PRS}_i = \frac{\sum_{j=1}^{n} EA_j}{\sum_{j=1}^{n} A_j},$$

where $EA_j \in \{0, 1, 2\}$ corresponds to homozygous absent, heterozygous present or homozygous present effect alleles at each SNP, and $A_j \in \{0, 1, 2\}$ corresponds to the total number of alleles with base calls at each SNP.

Our approach to PRS calculation and comparison between populations is characterized by three important choices: 1) the use of only significantly associated SNPs ($P < 10^{-5}$) for PRS calculation, 2) the calculation of PRS that are unweighted by SNP effect sizes, and 3) the calculation of PRS without the use of LD pruning or clumping. PRS were calculated in this way to facilitate comparisons of PRS distributions between divergent populations with distinct ancestry profiles and LD structures. 1) The use of a relatively small number of significantly associated SNPs, albeit at the relaxed threshold of $P < 10^{-5}$ used by the NHRI-EBI GWAS database, is known as the "top-SNP" approach, in contrast to the use of far more liberal $P$-value thresholds that allow for the inclusion of thousands or even millions of variants for PRS calculation. The top-SNP approach has been shown to mitigate the effects of population structure, particularly compared with approaches that use many thousands or millions of SNPs, which are essentially guaranteed to recapitulate population structure (Duncan et al. 2019). Furthermore, the top-SNP approach to PRS calculation has been shown to work almost as well or better compared with the approach using many thousands or even millions of SNPs (Khera et al. 2018). For example, a top-SNP approach to T2D PRS calculation using only 72 SNPs yielded an accuracy (area under the curve) of 0.70 compared with an average accuracy of 0.71 when more than 6.9 million SNPs were used. 2) Unweighted PRS were used to allow for combining SNP trait associations across multiple studies, each with distinct effect size estimates (Chande et al. 2018). Effect sizes from different studies cannot be readily combined due to differences in study cohorts, including cohort size, allele frequencies, and population structure. Furthermore, as effect sizes represent SNP heritability estimates, which are dependent on the particular cohort that is being studied, it does not make sense to attempt to normalize effect sizes across studies. 3) We opted not to use LD pruning for PRS calculation to facilitate direct comparison of PRS between populations with divergent LD structures. In particular, the top-SNP approach means that we are using a relatively small number of SNPs per population and the divergent LD structure means that different subsets of this small number of SNPs would likely be removed from each population if LD pruning were used. Thus, our approach to PRS calculation without LD pruning provides for both additional resolution, in terms of the numbers of SNPs available for analysis, and more direct comparisons between populations with divergent LD structures. Furthermore, several studies, including our own work, have shown that PRS calculated with and without LD pruning do not show big differences (De La Vega and Bustamante 2018; Chande et al. 2020; Elliott et al. 2020). An extended discussion of the rationale that underlies our PRS calculation method can be found in the supplementary methods section, Supplementary Material online.

For each of the three continental ancestry components ($f_{\mathrm{African}}$, $f_{\mathrm{European}}$, and $f_{\mathrm{NativeAmerican}}$), individuals' continental ancestry fractions were regressed against their PRS using unweighted ordinary least squares regression (OLS) as follows:

$$\mathrm{PRS_i} = \alpha + \beta \mathrm{x_i} + \varepsilon_\mathrm{i},$$

where $\mathrm{PRS}_i$ is the predicted polygenic risk score for individual $i$; $\alpha$ and $\beta$ are constants describing the intercept and slope, respectively; $x_i$ is the ancestry fraction for individual $i$; and $\varepsilon_i$ is an error term describing the deviation from the fitted line. The resulting OLS produces: $\beta_0$, the model $\beta$ or slope; the standard error of the model; the $r^2$ value describing the model's fit; the model $t$-statistic; and a two-tailed $P$ value.

Trait-associated SNPs were mapped to the nearest genes for pathway enrichment analysis using the ENSEMBL rsID to HGNC mapping API (getBM) provided as part of the biomaRt R package (attributes = refsnp_id, ensemble_gene_stable_id, hgnc_symbol, entrezgene_id; filter = snp_filter & ensembl_gene_id; values = GWAS Catalog SNP rsIDs). SNPs that did not return an HGNC mapping were discarded. Genes were assigned population-specific effect allele frequency difference values ($\Delta f = f(EA_{\mathrm{Ant}}) - f(EA_{\mathrm{Cho}})$) based on the SNP with the maximum effect allele frequency difference: $\max|\Delta f_{g,i}|$, where $g$ is a trait-associated gene and $i$ is $i$th SNP in gene $g$. The $\Delta f$ values for all mapped trait-associated genes were used to create population-specific gene lists for pathway overrepresentation analysis using the hypergeometric test implemented in the "enricher" function from the clusterProflier version 3.14.0 R package (Yu et al. 2012). Briefly, for each gene, the sign on $\Delta f$ was used to assign a gene to the Antioquia (positive) or Chocó (negative) gene lists. For each population-specific gene list and for each gene set, a hypergeometric test was performed using the following equation: $\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$, where $m$ is the number of population-specific genes, $k$ is the number of population-specific genes in gene set, $n$ is the number of genes in gene set, and $N$ is number of genes in the background. Gene sets from the KEGG, MSigDB (http://software.broadinstitute.org/gsea/msigdb/; last accessed August 12, 2020), and PID(http://www.ndexbio.org/#/user/301a91c6-a37b-11e4-bda0-000c29202374; last accessed August 12, 2020) were used in the enrichment analysis.

The relative predicted disease risk and observed disease prevalence for Antioquia and Chocó were computed as the $\log_2$ odds ratio for the effect allele frequencies and the reported age-adjusted disease prevalence values for Chocó/Antioquia. For each disease-associated SNP, its log odds ratio is computed as: $\log_2 \frac{p_{\mathrm{Cho}}/q_{\mathrm{Cho}}}{p_{\mathrm{Ant}}/q_{\mathrm{Ant}}}$, where $p_{\mathrm{pop}}$ is the population-specific frequency of the effect allele and $q_{\mathrm{pop}}$ is the population-specific frequency of the noneffect allele. The log odds ratio values for all associated SNPs were summed for each disease. The log odds ratio for disease prevalence is computed as follows: $\log_2 \frac{\mathrm{Disease_{Cho}/No\ \ disease_{Cho}}}{\mathrm{Disease_{Ant}/No\ \ disease_{Ant}}}$. Disease prevalence ($\mathrm{Disease_{pop}}$ and $\mathrm{No\ \ disease_{pop}}$) was defined as the population- and age-adjusted prevalence per 100,000 and $(100,000 - \mathrm{prevalence})$ reported for each department in 2017 and were taken from Colombian governmental and nongovernmental resources (see Demographic, Lifestyle and Disease Prevalence Data).

### Demographic, Lifestyle and Disease Prevalence Data

A variety of sources was used to curate demographic, lifestyle, and disease prevalence data for Antioquia and Chocó. The 2005 general census published by the Colombian Departamento Administrativo Nacional de Estadística (DANE) was used for demographic and socioeconomic status data (Uribe Vélez et al. 2006). Disease prevalence data were taken from three epidemiological databases: 1) Cuenta de Alto Costo (https://cuentadealtocosto.org/; last accessed

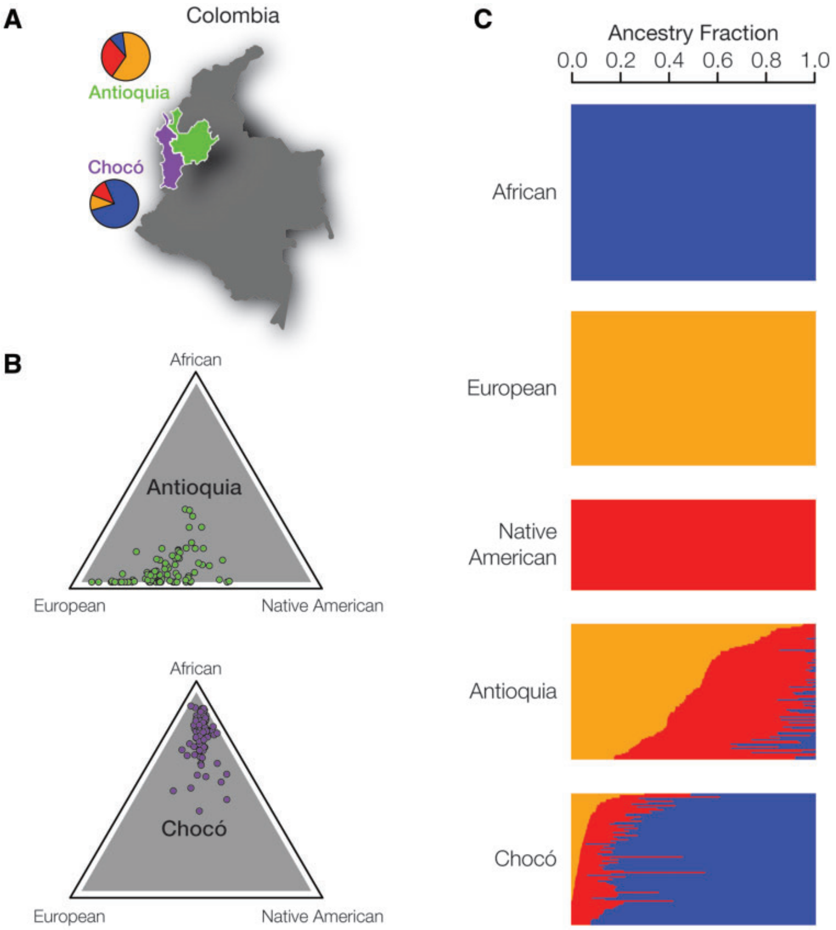

Fig. 1.—Genetic ancestry in Antioquia and Chocó. (*A*) The locations of the Colombian administrative departments of Chocó (purple) and Antioquia (green) are shown along with pie charts indicating the average continental ancestry fractions: African (blue), European (orange), and Native American (red). (*B*) Ternary plots showing the relative contributions of African, European, and Native American ancestry to individuals from Antioquia (green) and Chocó (purple). (*C*) ADMIXTURE plot showing the continental ancestry fractions for African (blue), European (orange), and Native American (red) reference populations together with Antioquia and Chocó.

August 12, 2020), 2) Observatorio de Diabetes de Colombia (http://www.odc.org.co/; last accessed August 12, 2020), and 3) the Sistema Integral de Información de la Protección Social (https://www.minsalud.gov.co/salud/Paginas/SistemaIntegraldeInformaciónSISPRO.aspx; last accessed August 12, 2020). Diet and lifestyle data were taken from the Colombian national nutritional survey (Alvarez 2006).

## Results and Discussion

### Demography and Genetic Ancestry in Antioquia and Chocó

Antioquia and Chocó are Colombian administrative departments (i.e., states) that are located in the northwestern part of the country and share a common border (fig. 1*A*). Chocó runs along the Pacific coast and borders Panamá to the north; it is the only department in Colombia with Pacific and Atlantic coasts. Antioquia is situated due east of Chocó, in the interior of the country, and also has a short Atlantic coastline. Despite their close proximity, the two departments have very distinct geography and climate as well as distinct historic and demographic profiles. Antioquia occupies the mountainous Andean region of the country and is traversed by the Western and Central Andes mountain ranges. According to the 2005 census, ~89% of the Antioquia population identifies as white or mestizo compared with 11% black or Afro-Colombian and <1% Indigenous. Chocó lies along the lowland Pacific coastal region and is almost entirely covered by dense tropical rainforest. The climate is hot and humid, and the region receives some of the highest rainfall totals in the world. The population of Chocó identifies as 82% Afro-Colombian, 13% Indigenous, and 5% white or mestizo.

Genome-wide variant data from Antioquia and Chocó were compared with data from African, European, and Native American continental reference populations to infer the patterns of genetic ancestry and admixture in the two Colombian populations. The genetic ancestry of Antioquia

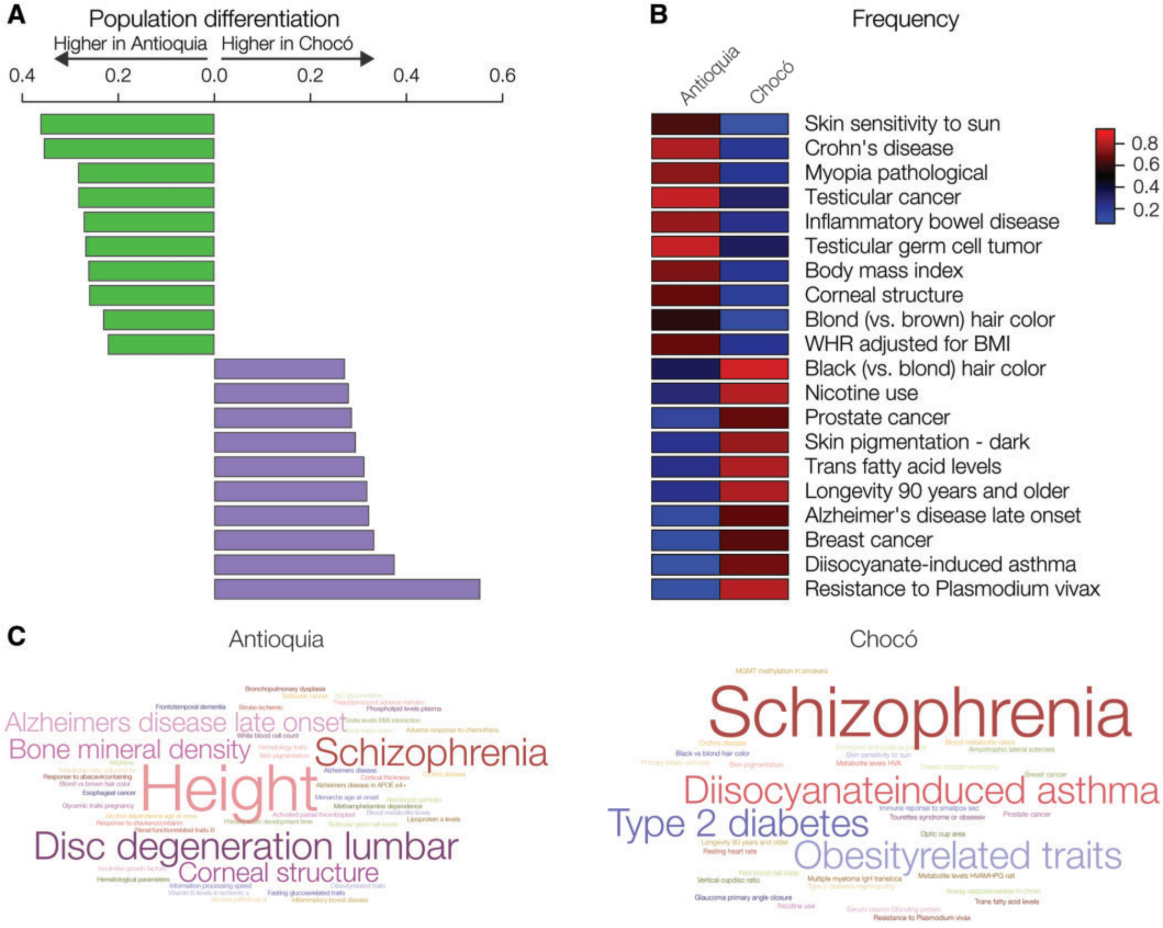

**Fig. 2.**—Single nucleotide variant phenotype associations. (*A*) Polarized fixation index ($F_{ST}$) values for divergent trait-associated SNP effect alleles: higher effect allele frequency in Antioquia (left, green) and higher effect allele frequency Chocó (right, purple). The corresponding SNP associations are shown in panel B (see supplementary table 3, Supplementary Material online, for details). (*B*) Heatmap of effect allele frequencies in Antioquia and Chocó (see key) and their SNP associations. (*C*) Word clouds showing the enrichment of SNP-associated traits for each population. Word clouds were generated by counting the occurrences of SNP trait-annotations for SNPs with an $F_{ST}$ value >0.2, 98 for Chocó and 61 for Antioquia (all SNPs significantly divergent at $P \ll 0.001$; supplementary table 3, Supplementary Material online), and words are scaled by number of times they appear in the trait association list.

and Chocó reflect their distinct historical founding populations, physical and cultural barriers to migration, and current demographic profiles (fig. 1*B* and *C*). Antioquia shows predominantly European genetic ancestry (average ± standard error; 62% ± 1.55) followed by Native American (32% ± 1.24) and then African (6% ± 0.83) components, whereas Chocó has primarily African genetic ancestry (76% ± 1.65) with approximately equal parts Native American (14% ± 0.83) and European (10% ± 1.03) ancestry.

## Single-Variant Divergence and Phenotypic Associations

The potential impact of ancestry differences between Antioquia and Chocó on the genetic architecture of phenotype and function was assessed for individual SNP trait associations (fig. 2). A total of 47,398 SNP trait associations were curated and evaluated with respect to the extent and direction of differentiation between Antioquia and Chocó. Population differentiation was measured by effect allele $F_{ST}$ values and frequency differences between the two populations (fig. 2*A* and *B* and supplementary table 3, Supplementary Material online). The top 20 most extreme values correspond to both known phenotype and disease prevalence differences between the two populations as well as novel differences (supplementary fig. 1, Supplementary Material online). Pigmentation-associated variants for both skin and hair show expected differences with lighter skin and hair effect alleles found in higher frequency in Antioquia compared with Chocó. Antioquia also shows higher frequencies of Crohn's and inflammatory bowel disease SNP effect alleles than Chocó, whereas Chocó shows higher frequencies of variants associated with prostate and breast cancer along with Alzheimer's and asthma, consistent with known health disparities around the world. Chocó also

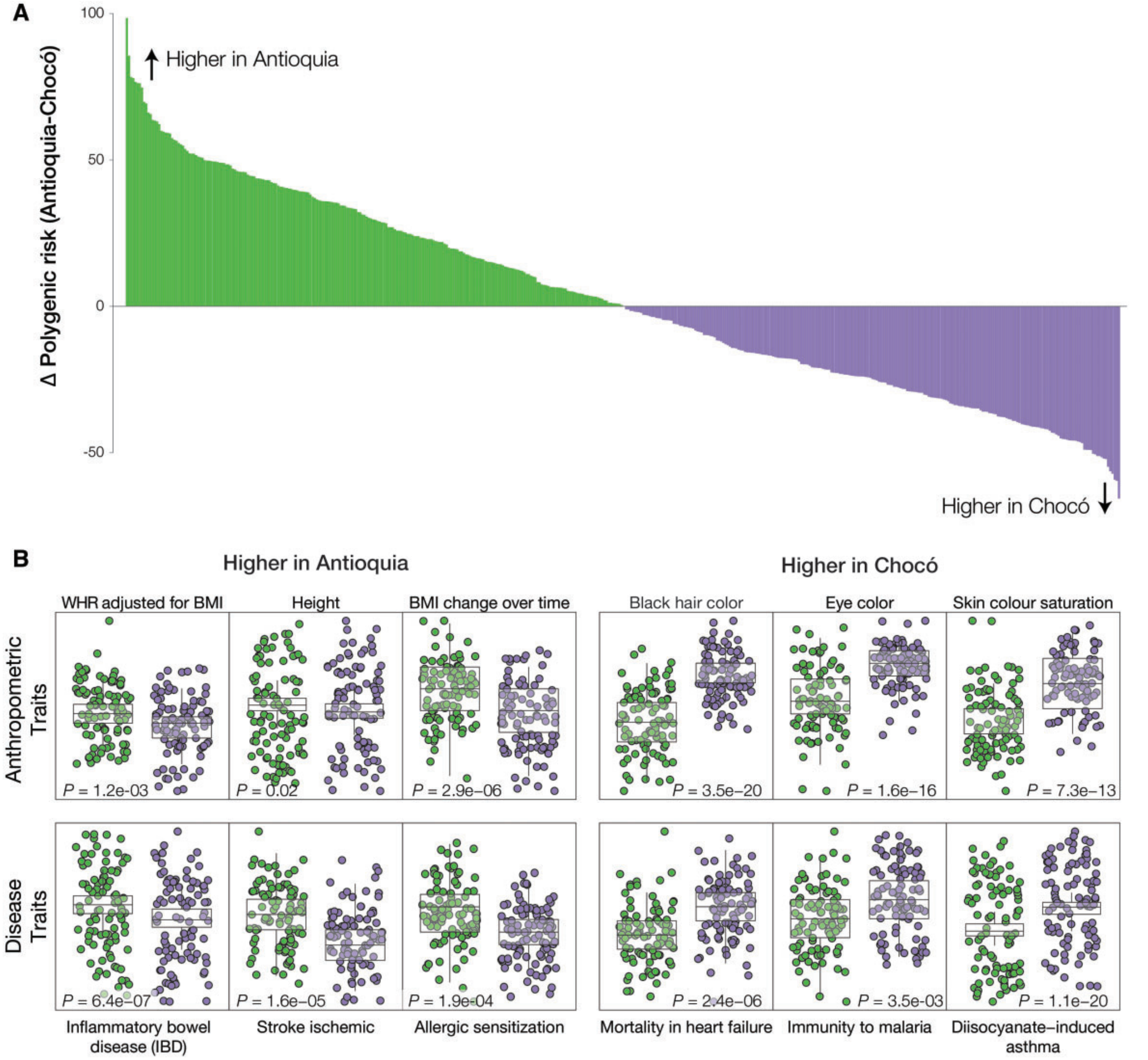

FIG. 3.—Polygenic risk divergence. (*A*) Distribution of the differences in population-average PRS are shown for significantly divergent traits: higher in Antioquia (above, green) and higher in Chocó (below, purple). (*B*) Population-specific PRS distributions for examples of anthropometric and disease traits are shown for Antioquia (green) and Chocó (purple) along with the significance levels for the distribution differences. Traits with increased prevalence/risk in Antioquia are shown on the left, traits with increased prevalence/risk in Chocó are shown on the right.

showed a substantially higher frequency of variants linked to resistance to the malaria parasite *Plasmodium vivax*. Unexpected results include the higher frequency of nicotine use associated SNP effect alleles in Chocó, as tobacco use is known to be lower in Chocó compared with Antioquia, the greater waist–hip ratio in Antioquia, and the increased longevity in Chocó.

Word clouds provide a visual sense of the overall between-population divergence for all trait-associated SNPs, with the most enriched traits highlighted for each population (fig. 2*C*). The word clouds were generated using all trait-associated SNPs that showed $F_{ST}$ >0.2, 61 SNPs for Antioquia and 98 for Chocó, and therefore provide additional resolution on the divergence of single-variant associations between populations. For example, schizophrenia appears in the word clouds for both populations (fig. 3*B*), with more weight in Chocó, although it was not present in the top 20 divergent associations shown in figure 2, panels A and B. Obesity-related traits appears as overrepresented in Chocó in the word cloud (fig. 2*C*), despite the fact that the most diverged body mass index SNP shows higher frequency in Antioquia (fig. 2*A* and *B*). This is due to a preponderance of obesity-associated SNPs among the total set of variants with $F_{ST} > 0.2$ and is consistent with what is seen via polygenic trait divergence analysis (see Polygenic Trait Divergence section and fig. 3). Overall, the population divergence observed for single-variant associations are consistent with reported health disparities and demographic data in Colombia and Latin American (supplementary table 4, Supplementary Material online).

## Polygenic Trait Divergence

Most human phenotypes are encoded by multiple loci across the genome, each of which contributes to a small fraction of the overall trait variance, that is, they are polygenic. The relationship between genetic ancestry and polygenic trait architecture in Antioquia and Chocó was assessed by comparing distributions of PRS between the two populations (fig. 3 and supplementary table 5, Supplementary Material online). A total of 1,983 PRS were compared between the two populations, and the overall distribution of ΔPRS (Ant − Choc) is symmetrically distributed around −0.01 (supplementary fig. 2, Supplementary Material online), indicating that the differences in genetic ancestry between the populations are slightly biased toward increased predicted risk in Chocó in cross-population PRS inference ($P < 0.001$). This is consistent with theoretical results showing that the divergence of neutral polygenic traits between populations is expected to be small, no different from the expectation for single-gene traits and symmetrically distributed around zero (Edge and Rosenberg 2015a,b). ΔPRS (Ant − Choc) values for traits that show significantly different mean PRS (Holm–Bonferroni corrected $P < 0.05$) are shown in figure 3*A* (column D in supplementary table 5, Supplementary Material online), and population-specific PRS distributions for individual traits of interest are shown in figure 3*B*. The specific traits of interest were chosen based on their highly divergent PRS values and their relevance to Colombia due to the reported public health burden in the country and as reflected by their descriptions in epidemiological and/or census databases.

The individual PRS distributions shown in figure 3*B* are organized into anthropometric and disease traits, most of which correspond to the top SNPs from figure 2. For anthropometric traits, Antioquia has a higher predicted height and body mass index, whereas Chocó has higher predicted values for several pigmentation-related traits: hair, eye, and skin color. For disease traits, Antioquia has greater predicted risk for inflammatory bowel disease, ischemic stroke, and allergic sensitization, whereas Chocó has a higher predicted risk for mortality in heart failure, immunity to malaria, and environmentally (diisocyanate) induced asthma. We also explored the impact of GWAS discovery and replication population ancestry on PRS differences for four selected traits from figures 2 and 3 for which multiple GWAS using different ancestry populations were available: asthma, ischemic stroke, myopia, and T2D (supplementary fig. 3 and table 6, Supplementary Material online). In all cases, significant differences in predicted population risk profiles were robust to discovery population ancestry, suggesting a shared genetic architecture of risk. In addition, predicted population-specific disease risk profiles are consistent with what has been observed in Colombia (supplementary table 4, Supplementary Material online) as well as with known ancestry–disease associations worldwide: for example, asthma (Moorman et al. 2007; Nyenhuis et al. 2017), heart failure (Bahrami et al. 2008; Bibbins-Domingo et al. 2009), irritable bowel disease (Nguyen et al. 2014; Park and Jeen 2019), malaria (Tishkoff et al. 2001; Shriner and Rotimi 2018; Yao et al. 2018), and stroke (Zweifler et al. 1995).

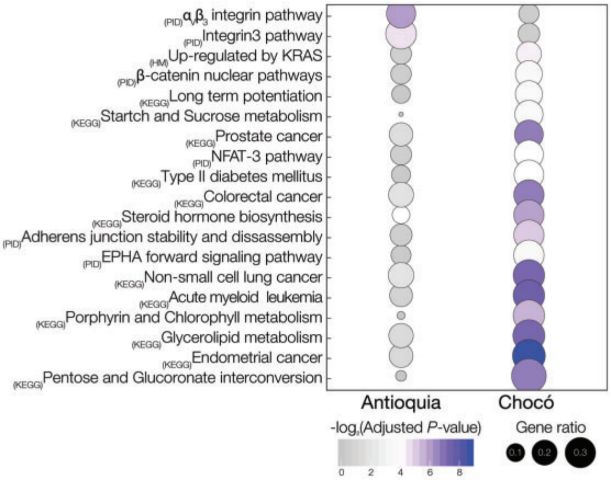

**Fig. 4.**—Population-specific differences in trait endophenotypes: pathways and biochemical functions. Gene set enrichment was used uncover pathways and functional gene sets that are enriched for divergent associated SNPs in each population. For each pathway or function, circles are scaled to the relative number of implicated genes for each population and colored according to the population-specific levels of enrichment.

We also explored population-specific differences in endophenotypes, with respect to specific pathways and/or biochemical functions that underlie the observed trait differences, using pathway enrichment analysis (fig. 4). Antioquia shows enrichment for integrin pathways implicated in a number of cancers and inflammatory bowel disease. Chocó shows enrichment for a number of cancer-related pathways, including prostate cancer, which is known to be more prevalent in men of African ancestry (Toles 2008; Mahal et al. 2018), as well as T2D and related glycerolipid metabolism pathways.

Given the differences in genetic ancestry seen for Antioquia and Chocó (fig. 1), we evaluated the relationship between individuals' continental genetic ancestry fractions and their PRS for each trait considered here. It should be noted that despite the clear differences in the overall ancestry seen for the two Colombian populations, almost all individuals analyzed here show substantial admixture, with varying fractions of African, European, and Native American ancestry.

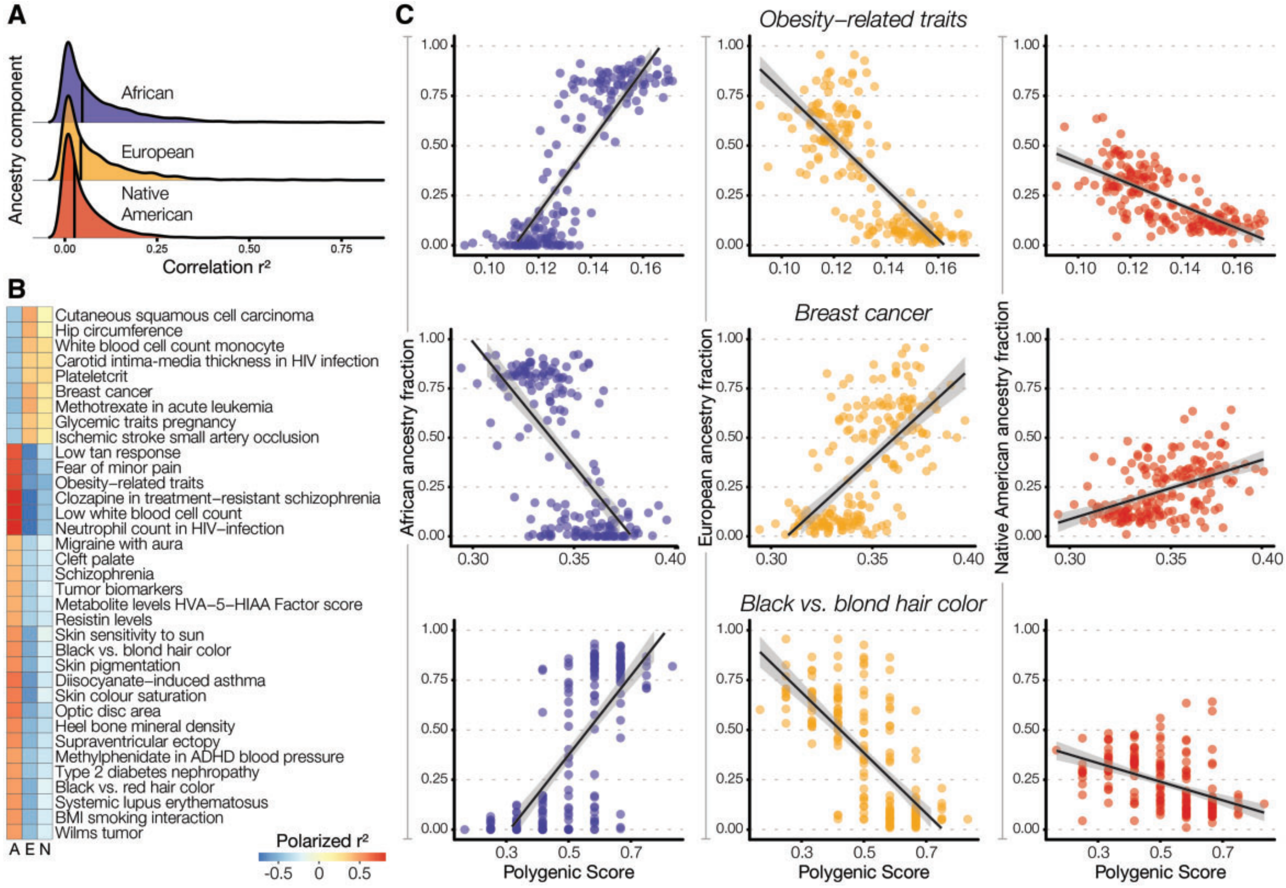

FIG. 5.—Genetic ancestry and polygenic trait divergence. (*A*) Distributions of the correlations ($r^2$) between individuals' genetic ancestry fractions—African (blue), European (orange), Native American (red)—and their PRS for all traits analyzed here. Vertical lines show the median for each distribution. (*B*) Ancestry × PRS correlations ($r^2$) polarized by the direction of the correlation (positive or negative) are shown for all traits where $r^2 > 0.4$ for at least one ancestry component—African (A), European (E), and Native American (N). (*C*) Examples of polygenic traits with high correlations between ancestry and PRS are shown. Ancestry components are color coded as in panel A, and for each scatter plot, ancestry fractions (*y* axis) are regressed against PRS (*x* axis). Linear trend lines with 95% confidence intervals are shown for each regression.

This fact allowed us to correlate genetic ancestry and PRS along a continuum of continental ancestry fractions (fig. 5). There are significant differences in the magnitude of the PRS correlations among the three ancestry components ($F = 4.79$, $P = 8.3 \times 10^{-3}$); African ancestry shows the highest overall correlation with the PRS values of all traits analyzed here, as shown by the median of the distribution, followed by the European and then the Native American ancestry components (fig. 5*A*). All three populations show a number of apparent cases of high correlations between ancestry and PRS. All traits that show $r^2 > 0.4$ for any of the three ancestry components are shown in figure 5*B*, and individual examples of ancestry × PRS regressions are shown in figure 5*C*. Breast cancer PRS is positively associated with European ancestry and negatively associated with African ancestry (fig. 5*C*), in contrast to what was seen for an individual breast cancer–associated variant found at higher frequency in Chocó (fig. 2*B*). This difference is best explained by the analysis of individual SNPs shown in figure 2 and the PRS based on multiple SNPs, which are likely to be more reliable, shown in figures 3 and 5. All ancestry × PRS $r^2$ values are shown in supplementary table 7, Supplementary Material online.

The high correlations observed between ancestry and PRS could be attributed to artifacts related to uneven cohort sampling in GWAS, as previously discussed, or they could represent actual ancestry-related phenotypic differences between the two populations. The small overall systematic bias in PRS for the two populations (supplementary fig. 2, Supplementary Material online), considered together with the fact that most of these ancestry associations conform to observable anthropometric features and/or known disease prevalence differences between populations suggest that these associations reflect real phenotypic differences. However, definitive proof for this would require individual-level phenotype data, as opposed to the population-level data used here, as well as the use of trait-associated variants that replicate across ancestry-specific GWAS. It should also be noted that these regressions could be confounded by a number of other variables including

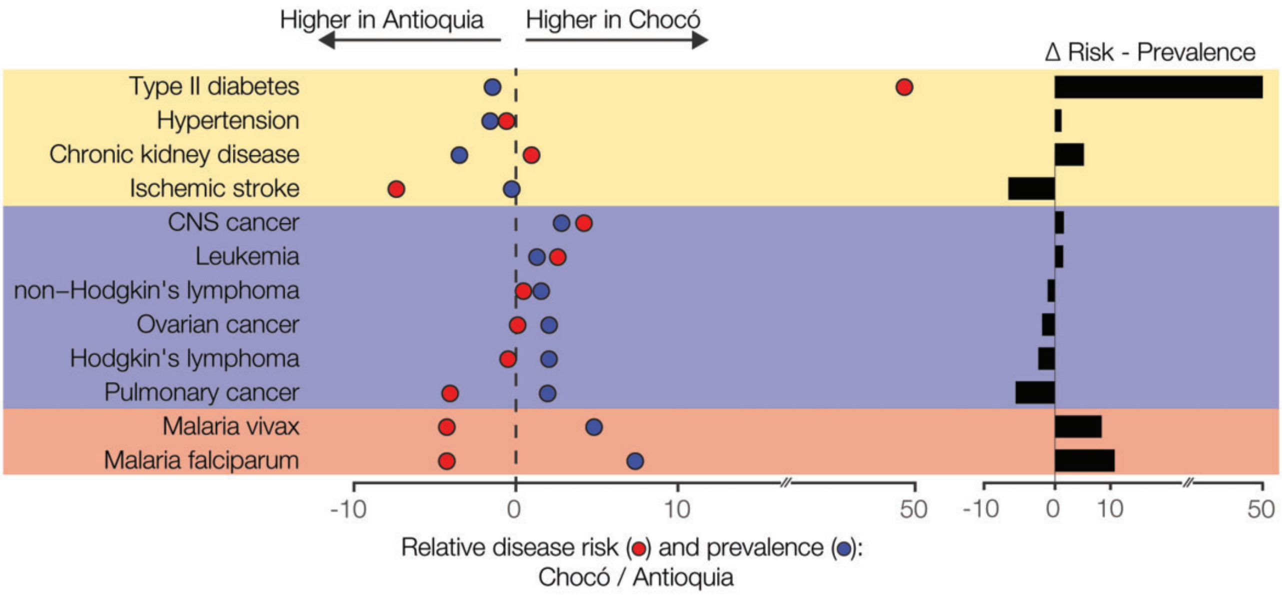

FIG. 6.—Predicted versus observed disease risk. Left: For each disease, the predicted genetic risk difference for Antioquia compared with Chocó (red circles) is compared to the observed prevalence of the disease (blue circles). Right: The differences between predicted disease risk minus observed prevalence. Diseases are grouped into bands as complex common diseases (yellow), cancer (blue), and infectious disease (red). The *x* axis values are log odds ratios for population-specific disease risk allele frequencies and observed disease prevalence values, as described in Materials and Methods.

sex, age, and socioeconomic status that are not available for this study, and which would need to be simultaneously modeled to ensure that the correlations between ancestry and PRS observed here are robust.

### Predicted versus Observed Disease Risk Profiles

Population-specific differences for trait-associated variants, both for single SNP associations and polygenic traits, showed an overall concordance between genetic risk predictions and observed anthropometric and epidemiological profiles for Antioquia and Chocó (figs. 2 and 3). We quantified the relationship between predicted disease risk and observed prevalence for twelve high impact diseases that have been prioritized by the Colombian Ministry of Health via the 'Cuenta de Alto Costo' (http://www.cuentadealtocosto.org/ ). This analysis was done for complex common diseases, cancers, and infectious diseases (fig. 6). T2D shows the largest difference between predicted disease risk versus observed disease prevalence for Antioquia and Chocó. We previously showed that this difference can be attributed to higher genetic risk associated with African genetic ancestry and T2D protective environmental factors associated with socioeconomic status in Chocó (Chande et al. 2017). In Colombia, environmental factors associated with differences in development across the country appear to have a high impact on the risk of complex common diseases like T2D. A similar, although not nearly as extreme, difference can be seen for chronic kidney disease; Chocó has a higher predicted genetic risk but lower prevalence compared with Antioquia. Higher risk for chronic kidney disease has been observed for Afro-descendant populations in other countries (Crews et al. 2010; Kaze et al. 2018), consistent with the higher genetic risk for Chocó seen here. Thus, it may be the case that similar environmental protective factors, with respect to diet and lifestyle, also serve to mitigate the risk of chronic kidney disease in Chocó. Finally, there are large differences in predicted risk (susceptibility) versus observed prevalence for malaria caused by both *Plasmodium vivax* and *P. falciparum*. The population of Chocó has lower predicted risk for malaria infections, consistent with previous studies on Afro-descendant populations (Tishkoff et al. 2001; Shriner and Rotimi 2018; Yao et al. 2018), but both *P. vivax* and *P. falciparum* are far more prevalent in Chocó compared with Antioquia (Battle et al. 2019; Nosten and Phyo 2019; Weiss et al. 2019), thereby explaining the higher malaria prevalence in Chocó.

## Conclusions

Results on the population divergence of trait-associated variants reported here should be interpreted with caution in light of the previously discussed challenges to cross-population genetic risk inference (Martin et al. 2017, 2019; Kim et al. 2018; Novembre and Barton 2018). This is particularly true for populations that have strikingly different ancestry profiles, as is the case for Antioquia and Chocó. However, for this study, the general concordance seen between genetically inferred (predicted) phenotypic differences and the observed differences for anthropometric traits, or known prevalence differences in the case of disease traits, supports the approach taken here (supplementary table 4, Supplementary Material online). It should be stressed that both trait-associated variant allele frequencies and PRS distributions overlap substantially between Antioquia and Chocó; in other words, predicted phenotypic differences vary along a continuum, with distinct group-specific averages in a minority of cases, as opposed to showing discrete values between populations. This is consistent with the expectation that the majority of genetic variation is found within rather than between human populations (Lewontin 1972; Li et al. 2008).

Finally, it is important to note that detailed individual-level phenotypic information will be needed to more rigorously evaluate the implications of genetic divergence at trait-associated variants in diverse populations of the kind studied here. Fortunately, data of this kind are increasingly being generated by biobank collections around the world, via the combination of genetic profiles and detailed phenotypic information gleaned from participant surveys and electronic health records. Many of these biobanks—for example, All of Us, BioMe, and the UK Biobank—include the kind of ancestrally diverse participant cohorts that can facilitate detailed investigations on the genetic basis of group-specific trait differences and health disparities.

The results reported here are distributed via a web-based platform that allows users to explore the extent of between-population divergence for individual trait-associated variants and for PRS: http://map.chocogen.com

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgment

This research and publication costs were funded by IHRC-Georgia Tech Applied Bioinformatics Laboratory (ATC, LR, ACB, IKJ). AV-A was supported by Fulbright Colombia. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

## Literature Cited

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9):1655–1664.

Alvarez MC. 2006. Encuesta nacional de la situación nutricional en Colombia. Bogota (CO): Instituto Colombiano de Bienestar Familiar.

Amberger J S, Bocchini C A, Schiettecatte F, Scott A F, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research. 43(D1):D789–D798. 10.1093/nar/gku1205

Bahrami H, et al. 2008. Differences in the incidence of congestive heart failure by ethnicity: the multi-ethnic study of atherosclerosis. Arch Intern Med. 168(19):2138–2145.

Battle KE, et al. 2019. Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. Lancet 394(10195):332–343.

Beiter ER, et al. 2017. Polygenic selection underlies evolution of human brain structure and behavioral traits. bioRxiv. doi:10.1101/164707

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. PLoS Genet. 10(8):e1004412.

Berg JJ, Zhang X, Coop G. 2019. Polygenic adaptation has impacted multiple anthropometric traits. bioRxiv. doi:10.1101/167551

Bibbins-Domingo K, et al. 2009. Racial differences in incident heart failure among young adults. N Engl J Med. 360(12):1179–1190.

Bryc K, et al. 2010. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc Natl Acad Sci U S A. 107(Suppl 2):8954–8961.

Buniello A, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47(D1):D1005–D1012.

Bustamante CD, Burchard EG, De la Vega FM. 2011. Genomics for the world. Nature 475(7355):163–165.

Chande AT, Rishishwar L, Conley AB, Valderrama-Aguirre A, Medina-Rivas MA, Jordan IK. 2020. Ancestry effects on type 2 diabetes genetic risk inference in Hispanic/Latino populations. BMC Med Genet. 21(Suppl 2):132.

Chande AT, et al. 2017. Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Choco and Antioquia. Sci Rep. 7(1):17127.

Chande AT, et al. 2018. GlobAl distribution of GEnetic traits (GADGET) web server: polygenic trait scores worldwide. Nucleic Acids Res. 46(W1):W121–W126.

Chang CC, et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci. 4(1):7.

Chatterjee N, Shi J, Garcia-Closas M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 17(7):392–406.

Conley AB, et al. 2017. A comparative analysis of genetic ancestry and admixture in the Colombian populations of Choco and Medellin. G3 (Bethesda). 7:3435–3447.

Corona E, et al. 2013. Analysis of the genetic basis of disease in the context of worldwide human relationships and migration. PLoS Genet. 9(5):e1003447.

Crews DC, Charles RF, Evans MK, Zonderman AB, Powe NR. 2010. Poverty, race, and CKD in a racially and socioeconomically diverse urban population. Am J Kidney Dis. 55(6):992–1000.

Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. 2013. Haplotype estimation using sequencing reads. Am J Hum Genet. 93(4):687–696.

Delaneau O, Marchini J, 1000 Genomes Project Consortium. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun. 5(1):3934.

De La Vega FM, Bustamante CD. 2018. Polygenic risk scores: a biased prediction? Genome Med. 10(1):100.

Duncan L, et al. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 10(1):3328.

Edge MD, Rosenberg NA. 2015a. A general model of the relationship between the apportionment of human genetic diversity and the apportionment of human phenotypic diversity. Hum Biol. 87:313–337.

Edge MD, Rosenberg NA. 2015b. Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. Stud Hist Philos Biol Biomed Sci. 52:32–45.

Elliott J, et al. 2020. Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. JAMA 323(7):636–645.

Homburger JR, et al. 2015. Genomic insights into the ancestry and demographic history of South America. PLoS Genet. 11(12):e1005602.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 44(8):955–959.

Howie B, Marchini J, Stephens M. 2011. Genotype imputation with thousands of genomes. G3 (Bethesda). 1:457–470.

Jordan IK, Rishishwar L, Conley AB. 2019. Native American admixture recapitulates population-specific migration and settlement of the continental United States. PLoS Genet. 15(9):e1008225.

Kaze AD, Ilori T, Jaar BG, Echouffo-Tcheugui JB. 2018. Burden of chronic kidney disease on the African continent: a systematic review and meta-analysis. BMC Nephrol. 19(1):125.

Khera AV, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 50(9):1219–1224.

Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. 2018. Genetic disease risks can be misestimated across global populations. Genome Biol. 19(1):179.

Lambert SA, Abraham G, Inouye M. 2019. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 28(R2):R133–R142.

Lewontin RC. 1972. Evolutionary biology. In: Dobzhansky TH, Hecht MK, Steere WC, editors. The apportionment of human diversity. New York: Springer. p. 381–398.

Li JZ, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104.

MacArthur J, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45(D1):D896–D901.

Mahal BA, Berman RA, Taplin ME, Huang FW. 2018. Prostate cancer-specific mortality across Gleason scores in Black vs Nonblack men. JAMA 320(23):2479–2481.

Malone J, et al. 2010. Modeling sample variables with an experimental factor ontology. Bioinformatics 26(8):1112–1118.

Martin AR, et al. 2017. Human demographic history impacts genetic risk prediction across diverse populations. Am J Hum Genet. 100(4):635–649.

Martin AR, et al. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 51(4):584–591.

Medina-Rivas MA, et al. 2016. Choco, Colombia: a hotspot of human biodiversity. Rev Biodivers Neotrop. 6(1):45–54.

Moorman JE, et al. 2007. National surveillance for asthma: United States, 1980–2004. MMWR Surveill Summ. 56(8):1–54.

Moreno-Estrada A, et al. 2013. Reconstructing the population genetic history of the Caribbean. PLoS Genet. 9(11):e1003925.

Nagar SD, et al. 2019. Population pharmacogenomics for precision public health in Colombia. Front Genet. 10:241.

Need AC, Goldstein DB. 2009. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 25(11):489–494.

Nguyen GC, Chong CA, Chong RY. 2014. National estimates of the burden of inflammatory bowel disease among racial and ethnic groups in the United States. J Crohns Colitis. 8(4):288–295.

Norris ET, et al. 2019. Assortative mating on ancestry-variant traits in admixed Latin American populations. Front Genet. 10:359.

Norris ET, et al. 2018. Genetic ancestry, admixture and health determinants in Latin America. BMC Genomics 19(S8):861.

Nosten FH, Phyo AP. 2019. New malaria maps. Lancet 394(10195):278–279.

Novembre J, Barton NH. 2018. Tread lightly interpreting polygenic tests of selection. Genetics 208(4):1351–1355.

Nyenhuis SM, et al. 2017. Race is associated with differences in airway inflammation in patients with asthma. J Allergy Clin Immunol. 140(1):257–265.e211.

Park SC, Jeen YT. 2019. Genetic studies of inflammatory bowel disease-focusing on Asian patients. Cells 8(5):404.

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. Nature 538(7624):161–164.

Racimo F, Berg JJ, Pickrell JK. 2018. Detecting polygenic adaptation in admixture graphs. Genetics 208(4):1565–1584.

Reich D, et al. 2012. Reconstructing native American population history. Nature 488(7411):370–374.

Rishishwar L, Conley AB, Vidakovic B, Jordan IK. 2015. A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians. Gene 574(2):345–351.

Rishishwar L, et al. 2015. Ancestry, admixture and fitness in Colombian genomes. Sci Rep. 5(1):12376.

Ruiz-Linares A, et al. 2014. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. PLoS Genet. 10(9):e1004572.

Shriner D, Rotimi CN. 2018. Whole-genome-sequence-based haplotypes reveal single origin of the sickle allele during the Holocene wet phase. Am J Hum Genet. 102(4):547–556.

Tishkoff SA, et al. 2001. Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. Science 293(5529):455–462.

Toles CA. 2008. Black men are dying from prostate cancer. ABNF J. 19:92–95.

Turchin MC, et al. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet. 44(9):1015–1019.

Uribe Vélez A, Maldonado Gómez H, Fernández Ayala PJ, Vargas Bad A, Serna Ríos C. 2006. Censo General 2005. Bogota (CO): Departamento Administrativo Nacional de Estadística (DANE).

Visscher PM, et al. 2017. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 101(1):5–22.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38(6):1358–1370.

Weiss DJ, et al. 2019. Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. Lancet 394(10195):322–331.

Winkler CA, Nelson GW, Smith MW. 2010. Admixture mapping comes of age. Annu Rev Genom Hum Genet. 11(1):65–89.

Yao S, et al. 2018. Genetic ancestry and population differences in levels of inflammatory cytokines in women: role for evolutionary selection and environmental factors. PLoS Genet. 14(6):e1007368.

Yu G, Wang LG, Han Y, He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 16(5):284–287.

Zweifler RM, Lyden PD, Taft B, Kelly N, Rothrock JF. 1995. Impact of race and ethnicity on ischemic stroke. The University of California at San Diego Stroke Data Bank. Stroke 26(2):245–248.

**Associate editor:** Emilia Huerta-Sanchez