POPULATION GENOMICS AND ANCESTRAL ORIGINS FOR HEALTH DISPARITIES RESEARCH

A Dissertation
Presented to
The Academic Faculty

by

Shashwat Deepali Nagar

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
December 2021

POPULATION GENOMICS AND ANCESTRAL ORIGINS FOR HEALTH DISPARITIES RESEARCH

Approved by:

Dr. I. King Jordan, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Peng Qiu
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Joseph L. Lachance
School of Biological Sciences
Georgia Institute of Technology

Dr. Leonardo Mariño-Ramírez
Division of Intramural Research
National Institute on Minority Health and Health Disparities

Dr. Gregory Gibson
School of Biological Sciences
Georgia Institute of Technology

Date Approved: August 13, 2021

To my mother

ACKNOWLEDGEMENTS

I will be eternally grateful to Prof. I. King Jordan for his mentorship, guidance, and support throughout my time at Georgia Tech. From him, I have learned a lot about life, the world, and (of course) science. Along with being a fantastic research advisor with an infectious love for science, he is also an impressive raconteur with a large bag of gripping, contextually appropriate references. I am fortunate that I have had the honor of his mentorship and friendship. I will miss the hours we spent discussing ideas over coffee, sushi, and beer.

I would like to thank Prof. Joseph Lachance for his mentorship and his anecdotes, which made this journey through graduate school a little less frightening. Also, for his recommendations of literature, great haunts in Atlanta, and endless conversations about academia. I would also like to thank Prof. Greg Gibson for engaging scientific discussions in hallways and for his training in literature evaluation. I will never read papers the same way. I would also like to thank Prof. Peng Qiu whose course in machine learning helped me better understand this toolkit of black boxes. Finally, I would like to thank Dr. Leonardo Mariño-Ramírez for being an incredible collaborator and for engaging discussions about complex matters.

I have immense gratitude for my family of friends at Georgia Tech who made me feel right at home in foreign lands. I would like to thank Anurag Sethi, Manasa Kadiri, and Sheetal Chauhan for never letting distance come in the way of friendship. Especially to Shareef Khalid for always taking out time to talk about life and science, and for looking

past national origins. To my friends Dr. Aroon Tushar Chande[1] and Anna Gaines for countless nights filled with movies, music, and delightful conversation – thank you for making drab moments light up. To Dr. Evan Allen Clayton[2], thank you for being an incredible friend and a source of reliable, balanced advice and to Dongjo Ban for being a great friend and roommate throughout graduate school. I will be forever grateful to Dr. Hector Fabio Espitia Navarro[3] for his continued friendship, and for answering all my persistent questions about Spanish and music. *Muchisimas gracias* to Dr. Luz Karime Medina Cordoba[4] for animated conversations and *chisme* in the lab and beyond, and for always being available to talk. A large debt of gratitude to Dr. Camila Medrano Trochez[5] for being an incredible friend – my time at Georgia Tech would not be the same without your friendship. Finally, I would like to thank other friends and colleagues from the Jordan lab and EBB's second floor who have helped me along the way.

I would be remiss if I did not thank Lisa Redding for being an extraordinarily helpful and very patient in answering all my inane questions about the Bioinformatics program and Troy Hilley for rapidly resolving any server issues that came up.

I am extremely grateful to Sharlene Elaine Fernandes for her love and support, for engaging discussions about the world around us, and for keeping me sane during the pandemic.

I am forever indebted to my late grandmother Jeevan Lata Kapoor and my grandfather Jagdish Chandra Kapoor for their love and support. Last, but not least I would like to thank my mother Deepali Nagar for being my most ardent supporter. Her strength and passion for life are a constant source of inspiration for me. I owe everything to her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

1KGP	**1000 Genomes Project**
CRP	**C-reactive protein**
GA	**Genetic ancestry**
PGx	**Pharmacogenomics**
SED	**Socioeconomic deprivation**
SIRE	**Self-identified race/ethnicity**
T2D	**Type 2 diabetes**
UK	**United Kingdom**
US	**United States**

SUMMARY

Ameliorating health disparities – avoidable differences in health outcomes between population groups – is both a social imperative and a pressing scientific challenge. The relative importance of genetic versus environmental effects for health disparities i.e., the enduring question of nature versus nurture, particularly for complex common diseases that have multifactorial etiologies, has long been debated[6,7]. Nevertheless, the reality is that health outcomes are influenced by a combination of genetic and environmental factors as well as myriad interactions among them. This thesis aimed to study both genetic and environmental contributions to health disparities by leveraging population biobanks and large genomic datasets.

This thesis investigates the relationship between ancestral origins, environmental factors, and health disparities. The importance of social and environmental determinants of health disparities is well established, whereas the role of genetics is more controversial. Nevertheless, the two classes of effects are not mutually exclusive; genes are expressed and function in the context of specific environmental conditions. Thus, it is reasonable to consider the influence of genetic and environmental factors on health disparities together. Indeed, the importance of interactions between genetic and environmental factors for shaping health outcomes has recently been recognized and emphasized as a promising avenue for health disparities research[8-10]. Biobank datasets of the kind that were be analyzed here, which include collections of genetic data together with rich clinical, phenotypic, and environmental data for thousands of individuals, are ideally suited for this purpose. This thesis leverages biobank data analysis to decipher how ancestral origins and

environmental factors jointly impact health disparities. Few studies have investigated this possibility and this project aims to combine these two worldviews in health disparities research.

This thesis is split into two parts: (1) population pharmacogenomics spanning chapters 2 and 3, and (2) complex common health disparities covered in chapters 5, 6, and 7. Chapters 2 and 3 investigate the partitioning of pharmacogenomic variation between populations in different geographic and socioeconomic locales (Colombia and the United States) to study differences in predicted therapeutic response among populations. Chapters 5, 6, and 7 illustrate the use of a large biobank – the UK Biobank – to understand health disparities and their complex relationship to genetic, environmental, and social factors.

Research advance 1: Chapter 2 explores the application of the precision public health paradigm for ancestrally-guided pharmacogenomics in Colombia. This study focuses on two neighboring populations with distinct ancestry profiles: Antioquia (with primarily European genetic ancestry) and Chocó (with primarily African genetic ancestry). This study was a result of working with collaborators from Colombia to identify and prioritize pharmacogenomic variants in Antioquia and Chocó. In addition to pharmacogenomic alleles related to increased toxicity risk, this investigation also identified evidence that alleles related to dosage and metabolism have large frequency differences between the two populations, which are associated with their specific majority genetic ancestries. This fruitful collaboration has also led to the development of cost-effective PCR-based assays that avoid the prohibitively high cost of sequencing/genotyping while also bringing the promise of precision medicine to Colombia.

Research advance 2: Chapter 3 describes a study which shows that self-identified race/ethnicity (SIRE) information in older Americans is useful for partitioning pharmacogenomic variation, and that SIRE carries clinically valuable information for stratifying pharmacogenomic risk among US populations. Perhaps more interesting is the illustration that people who identify as Black or Hispanic stand to gain more from the consideration of SIRE in treatment decisions compared to those belonging to the majority White population.

Taken together, research advancements 1 and 2 highlight that population genomics can be a powerful tool for clinical decision-making especially in settings where resources are limited (e.g. Colombia) or where resources are unequally distributed between population groups (e.g. USA). This is in support of the precision public health paradigm which shifts the focus from individuals to populations to identify interventions that work best at the population level. This allows for uniform priors for treatment to be adjusted based on population membership.

Research advance 3: Chapter 4 explores the landscape of health disparities in the United Kingdom (UK) by leveraging data from the UK Biobank. The chapter describes an online web browser that catalogs the prevalence of disease phenotypes in groups defined by the following population attributes: age, country of residence, ethnic group, socioeconomic deprivation, and sex. This online browser will enable researchers to explore the landscape of health disparities and direct their attention to areas of research which might have the most impact.

Research advance 4: Chapter 5 sheds light on the interaction between genetic ancestry and socioeconomic deprivation (SED) – a proxy for a large family of environmental exposures and lifestyle factors – for type 2 diabetes (T2D) in the United Kingdom (UK). Leveraging multivariable logistic regression, this study finds that genetic ancestry and SED show significant interaction effects on T2D, with SED being a relatively greater T2D risk factor for individuals with South Asian and African ancestry, compared to those with European ancestry. The interactions between SED and GA underscore how the effects of environmental risk factors can differ among ancestry groups, suggesting the need for group-specific interventions.

Research advance 5: Chapter 6 compares genetic and socioeconomic contributions to ethnic differences in C-reactive protein (CRP) – a routinely used inflammation blood biomarker – in the UK. CRP is associated with response to infection, risk for a number of complex common diseases, and psychosocial stress. Using structural equation modeling, the study shows that socioeconomic deprivation (SED) explains more than twice the variation in CRP levels than genetic ancestry, and the effect of ethnicity on CRP is mediated by SED but not by genetic ancestry. Taken together, these results indicate that socioenvironmental factors contribute more to CRP ethnic differences than genetics. The study also finds that differences in CRP are associated with ethnic disparities for a number of chronic diseases, including type 2 diabetes, essential hypertension, sarcoidosis, and lupus erythematosus. These results indicate that ethnic differences in CRP are linked to both socioeconomic deprivation and numerous ethnic health disparities.

Together, research advancements 3, 4, and 5 demonstrate the massive potential of employing biobanks – large data repositories with genetic, environmental, and clinical data – to study and decompose health disparities.

Beyond these specific research advances, this thesis also takes a step towards addressing the lack of diversity in genomics research. Genomics research is currently biased towards European ancestry cohorts, and results from these studies may not transfer to more diverse ancestry groups. This genomics research gap has the potential to exacerbate existing health disparities. The focus on ancestrally diverse populations, both in developing countries and for underrepresented minority groups in the US and the UK, has the potential to support health equity through ancestrally-guided insights and interventions.

INTRODUCTION

1.1 Health disparities

Health disparities are avoidable differences in health outcomes among population groups, where populations can be defined in a variety of ways, such as by race and ethnicity, socioeconomic status, or gender[11]. Health equity entails the elimination of such avoidable health differences, a top priority for the US Department of Health and Human Services[12].

The movement towards more equitable health outcomes includes both moral and scientific dimensions; health disparities are fundamentally unjust and their causes are multifactorial and complex[13,14]. Unequal access to healthcare, socioeconomic factors, environmental exposures, diet and lifestyle, along with biological and genetic factors, all contribute to population health disparities[9,15]. Accordingly, efforts to promote health equity must embrace sociologically, environmentally, and biologically informed interventions.

1.2 Biobanks

Biobanks have been defined as "an organized collection of human biological material and associated information stored for one or more research purposes"[16,17], and population biobanks combine biological material with associated lifestyle, environmental, and clinical data for many thousands of participants. As such, biobanks provide an unprecedented

opportunity to jointly analyze genetic and environmental contributions to health disparities at a high level of resolution, in support of promoting health equity for currently underserved populations[18]. For this thesis, I aim to use population biobanks to characterize and analyze the effects of genetic and environmental factors on health disparities at two levels – drug response and disease prevalence.

1.3 Pharmacogenomics

Pharmacogenomics (PGx) refers to the link between human genetic variation and drug response[19,20]. PGx variants are specific genetic variants that mediate how individuals respond to a wide variety of medications, and PGx variant effects on drug response can be categorized with respect to dosage, efficacy, or toxicity/adverse drug reactions. Information on PGx can be accessed from the Pharmacogenomics Knowledgebase (PharmGKB), which is an NIH-funded resource that provides a manually curated set of clinical annotations with information about PGx variants and their corresponding drug responses.

1.4 Precision medicine and precision public health

Precision medicine refers to an approach to treatment that considers the genetic variants harbored by an individual to tailor clinical decision making. Despite holding great promise, this treatment paradigm requires genetic characterization of each individual patient and can thus be prohibitively expensive for the developing world and for

underserved patients from developed nations who do not have equitable access to healthcare.

A recently articulated alternative to the precision medicine model is referred to as precision public health[21-24]. The focus of precision public health is populations, instead of individuals, and the idea is to leverage modern healthcare technologies for more precise population-level interventions. The *mantra* for precision public health is "the right intervention, to the right population, at the right time." This population-centered model of healthcare delivery, which moves away from the need to genetically characterize individual patients in favor of developing population profiles, provides one way for the technological innovations underlying precision medicine to realize their potential in developing countries and in underserved minority populations in the developed world.

1.5 Ancestral origins

1.5.1 Race and ethnicity in the US

Race and ethnicity are treated as separate concepts in the US, as defined by the Office of Management and Budget (OMB). According to the OMB, race and ethnicity data are collected for many Federal programs and are critical to making policy decisions (especially for civil rights). These data are particularly important for assessing disparities in health and environmental exposures. Individuals in the US self-identify with respect to both race and ethnicity, using the socially defined categories defined below.

1.5.1.1 Race in the US

Race designations in the US are based on individuals self-identifying as having origins in population groups that correspond to broad geographic regions: "White" (Europe, the Middle East, or North Africa), "Black" (sub-Saharan Africa), "American Indian or Alaska Native" (Indigenous peoples of North and South America), "Asian" (East Asia, Southeast Asia, or Indian subcontinent), "Native Hawaiian or Other Pacific Islander" (Indigenous peoples of Hawaii, Guam, Samoa, or other Pacific Islands)[25]. According to the OMB, racial categories on the US census "… generally reflect a social definition of race recognized in this country and not an attempt to define race biologically, anthropologically, or genetically. In addition, it is recognized that the categories of the race item include racial and national origin or sociocultural groups."

1.5.1.2 Ethnicity in the US

Ethnicity in the US census is narrowly defined as either having a Hispanic origin or not. The OMB defines "Hispanic or Latino" as a person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin regardless of race[26].

1.5.2 Ethnic groups in the UK

In the UK, ethnic groups are used in a way that is analogous to racial groups in the US. Individuals in the UK self-identify as belonging to a single ethnic group – White,

Mixed (Multiple), Asian, Black, or Other – and choose a single ethnic background within each group. According to the Office for National Statistics in the UK, there is no consensus on what constitutes an ethnic group and membership is something that is self-defined and subjectively meaningful to the person concerned. Elements that shape someone's ethnicity include common ancestry and elements of culture, identity, religion, language, and physical appearance. What seems to be generally accepted, however, is that ethnicity includes all these aspects, and others, in combination[27].

1.5.3 Genetic ancestry

Genetic ancestry refers to genetic similarities derived from common ancestors[28]. Since genetic ancestry reflects distinct allele frequency patterns found in different populations, it can be indicative not only of biogeographic origins, born out of physical separation between populations, but also of ethno-cultural groups owing to reproductive isolation and endogamy.

Genetic ancestry is a characteristic of the genome, which can be objectively defined with confidence and precision. It can be represented as a categorical variable, capturing the discrete aspects of human genetic variation, or as a continuous variable reflecting the range of variation between discrete groups.

As recent ancient DNA research has shown, the history of humanity shows repeating cycles of population isolation and divergence followed by interaction and mixture[29]. The process of genetic exchange between previously diverged lineages is called admixture. As

barriers to long-distance travel are reducing in the modern world, the process of admixture, which was historically limited to geographically proximate populations, is increasing. Admixture, i.e. contributions of different ancestral source populations to modern individual genomes, can be quantified through the construct of genetic ancestry.

1.6 Using ancestral origins for health disparities research

The relationship between race, ethnicity, and genetic ancestry is complex and requires nuanced understanding and analysis. Race and ethnicity are markers of membership in social groups that influence, and are associated with social interactions, access to societal resources, and other socioenvironmental factors[30]. Genetic ancestry on the other hand, is a characteristic of the genome and serves as a proxy for genetic diversity among human populations. It serves as a marker for the likely presence of certain genetic variants in individuals descending from different ancestral populations[31]. Given that all these constructs – race, ethnicity, and genetic ancestry – are related to one's ancestral origins, they are correlated[32-34]. Importantly, however, the specifics of the information carried by these related concepts differ in important ways. Self-identified race and ethnicity carry information about the social experiences of an individual[35] and serve as epidemiologic proxies, while genetic ancestry carries information about probability of carrying certain genetic variants. Genetic ancestry also allows us to account for genetic admixture between ancestral populations – information that is not captured by categorical race and ethnicity labels. Like a recent article on the use of race and genetic ancestry in

medicine[30], I propose that all of these pieces of information – race, ethnicity, and genetic ancestry – be used to understand, characterize, and address health disparities.

It should be made clear that the relevance of genetics to racial and ethnic health disparities remains a matter of contention[36,37]. On the one hand, there are concerns that a focus on genetics will distract from more important socioeconomic determinants of health outcomes, and possibly reinforce stereotypes, which themselves contribute to disparities[38-40]. On the other hand, there is a growing sense that the insights being provided by genomics research on the genetic architecture of complex common diseases and cancer should be harnessed to enhance health equity[8,41]. Bioethicists have begun to consider how genetics and genomics research can be used to reduce health disparities. Specific recommendations include expanding the focus of genomic research to underrepresented minority populations[42-44], and an emphasis on gene-environment interactions that can be used to tailor population-level interventions[9,10]. Given their large sample sizes and the inclusion of a plethora of environmental and lifestyle factors, population biobanks are ideally suited to support both of these recommendations. Additionally, the inclusion of underrepresented populations in genetic research can support the identification of tailored diagnostics and interventions.

CHAPTER 2. POPULATION PHARMACOGENOMICS FOR PRECISION PUBLIC HEALTH IN COLOMBIA

2.1 Abstract

While genomic approaches to precision medicine hold great promise, they remain prohibitively expensive for developing countries. The precision public health paradigm, whereby healthcare decisions are made at the level of populations as opposed to individuals, provides one way for the genomics revolution to directly impact health outcomes in the developing world. Genomic approaches to precision public health require a deep understanding of local population genomics, which is still missing for many developing countries. We are investigating the population genomics of genetic variants that mediate drug response in an effort to inform healthcare decisions in Colombia. Our work focuses on two neighboring populations with distinct ancestry profiles: Antioquia and Chocó. Antioquia has primarily European genetic ancestry followed by Native American and African components, whereas Chocó shows mainly African ancestry with lower levels of Native American and European admixture. We performed a survey of the global distribution of pharmacogenomic variants followed by a more focused study of pharmacogenomic allele frequency differences between the two Colombian populations. Worldwide, we found pharmacogenomic variants to have both unusually high minor allele frequencies and high levels of population differentiation. A number of these pharmacogenomic variants also show anomalous effect allele frequencies within and between the two Colombian populations, and these differences were found to be associated with their distinct genetic ancestry profiles. For example, the C allele of the SNP

rs4149056 (SLCO1B1*5), which is associated with an increased risk of toxicity to a commonly prescribed statin, is found at relatively high frequency in Antioquia and is associated with European ancestry. In addition to pharmacogenomic alleles related to increased toxicity risk, we also have evidence that alleles related to dosage and metabolism have large frequency differences between the two populations, which are associated with their specific ancestries. Using these findings, we have developed and validated an inexpensive allele-specific PCR assay to test for the presence of such population-enriched pharmacogenomic SNPs in Colombia. These results serve as an example of how population-centered approaches to pharmacogenomics can help to realize the promise of precision medicine in resource-limited settings.

2.2 Introduction

The precision medicine approach to healthcare entails a customized model whereby medical decisions and treatments are specifically tailored to individual patients[45,46]. Currently, precision medicine is most commonly implemented via pharmacogenomic methods, which account for how individuals' genetic makeup affects their response to drugs[47,48]. Pharmacogenomic knowledge of genetic variant-to-drug response interactions provides a means to optimize individual patients' treatment regimes, simultaneously maximizing drug efficacy while minimizing adverse reactions. Indeed, the essence of precision medicine has been described as "the right treatment, to the right patient, at the right time". While the precision medicine paradigm promises to revolutionize healthcare delivery, its prohibitive costs put it out of reach for the developing world. In particular,

the need to characterize genomic information for each individual patient in a given population can place a tremendous burden on healthcare systems that may be struggling to provide basic services. For the moment, precision medicine as a standard of care is still very much limited to the Global North.

A recently articulated alternative to the precision medicine model is referred to as precision public health[21-23]. The focus of precision public health is populations, instead of individuals, and the idea is to leverage modern healthcare technologies for more precise population-level interventions. The mantra for precision public health is "the right intervention, to the right population, at the right time". This population-centered model of healthcare delivery provides one way for the technological innovations underlying precision medicine to realize their potential in developing countries. With respect to pharmacogenomics, knowledge regarding population genomic distributions of the genetic variants that mediate drug response can be used to focus resources and efforts where they will be most effective[49]. Under the precision public health model, population genomic profiles, as opposed to genomic information for each individual patient, can be employed to guide pharmacogenomic interventions; this is a far more cost-effective and realistic approach for the developing world[50]. For this study, we applied the precision public health paradigm using a survey of the distribution of pharmacogenomic variants in diverse Colombian populations. The major aim of this work was to tailor pharmacogenomic testing and interventions to the specific populations for which they will realize the greatest benefit.

Colombia is home to a highly diverse, multi-ethnic society. The modern population of Colombia is made up of individuals with genetic ancestry contributions from ancestral source populations in Africa, the Americas, and Europe[51-55]. Colombia is also known to contain a number of unique regional identities. There are at least five distinct recognized regions in Colombia, each of which has its own defining demographic contours[56,57]. In fact, owing to historical barriers to migration, Colombian populations with very different genetic ancestry profiles can be found in close geographic proximity. This is very much the case for the two populations characterized for this study: Antioquia and Chocó[58,59]. Despite the fact that these neighboring administrative departments share a common border, their populations show clearly distinct genetic ancestries. Antioquia has primarily European ancestry, whereas Chocó is mainly African, and both populations also show varying levels of Native American admixture.

Previous studies have shown that the frequencies of pharmacogenomic variants can vary across populations with divergent genetic ancestries. This includes variation in pharmacogenomic variant allele frequencies among distantly related populations worldwide[60,61] as well as marked frequency differences among populations sampled from within the same country[62,63]. We hypothesized that pharmacogenomic allele frequencies should differ between the Colombian populations of Antioquia and Chocó, given their distinct ancestry profiles. If this was indeed the case, it would have direct implications for the development of pharmacogenomic approaches in the country. In this way, we hoped that a survey of the population pharmacogenomic patterns for Antioquia and Chocó could serve as an exemplar for the implementation of precision public health in the developing world.

Colombia's first clinical genomics laboratory – GenomaCES from Universidad CES in Antioquia (*https://www.genomaces.com/*) – is currently working to develop genomic diagnoses that are tailored to the local population, and members of the ChocoGen Research Project (*https://www.chocogen.com/*) are exploring the connections between genetic ancestry and health disparities in the understudied Colombian population of Chocó. Here, these two groups have joined forces in an effort to (i) discover pharmacogenomic variants with special relevance for these two Colombian populations and (ii) develop cost-effective and rapid pharmacogenomic assays for those variants, which can be readily deployed in resource-limited settings.

2.3 Materials and Methods

2.3.1 Pharmacogenomic (PGx) variants

Pharmacogenomic single nucleotide polymorphisms (PGx variants), *i.e.* human genetic variants associated with specific drug responses, were mined from the Pharmacogenomic Knowledgebase (PharmGKB *https://www.pharmgkb.org/* accessed April 2018)[64]. PharmGKB provides a manually curated set of clinical annotations with information about PGx variants and their corresponding drug responses. The PharmGKB clinical annotations were downloaded and filtered to extract all individual PGx variant clinical annotations. Data on PGx variant clinical annotations were parsed and stored, including information about the direction and nature of the variant associated drug responses, the identity of each PGx variant effect and non-effect allele, the genes wherein PGx variants are located, and the drug interaction evidence levels.

2.3.2 PGx variant genetic variation

Data on human genome sequence variation were taken from the phase 3 data release of the 1000 Genomes Project[65]. For the 1000 Genomes Project, genome-wide SNPs were characterized via whole genome sequencing for 2,504 individuals from 26 global populations, including the Colombian population of Antioquia (CLM - Colombian in Medellín, Colombia *https://www.coriell.org/0/Sections/Collections/NHGRI/1000Clm.aspx*). All of the PGx variants from PharmGKB were found to be present in 1000 Genomes Project phase 3 variant calls. Genome sequence variation for the Colombian population of Chocó was characterized as part of the ChocoGen Research Project (*https://www.chocogen.com/*) as previously described[58,59,66].

Genome sequence variation data were used to calculate the average minor allele frequency (MAF) and fixation index (F_{ST}) for a genome-wide set of n=28,137,656 pruned SNPs and for the set of n=1,995 PGx variants using the program PLINK[67]. Linkage disequilibrium pruning was performed to yield the genome-wide background SNP set with the PLINK indep command, using an r^2 threshold of 0.5 with a sliding window of 50nt and a step size of 5nt. MAF (p) values for each SNP were calculated across all populations as described as:

$$p = \frac{number\ of\ variant\ sites}{total\ number\ of\ sites} \quad (1)$$

F_{ST} values for each SNP were calculated among populations as described as:

$$F_{ST} = \frac{\sigma^2}{\bar{p} \times (1 - \bar{p})} \quad (2)$$

where $\bar{p}$ is the average MAF across all 26 global populations and σ^2 is the observed MAF variation. Pairwise genomic distances were computed as 1-*identity-by-state/Hamming distances* between genomes using the PLINK distance command with the --distance-matrix option. The resulting high-dimensional pairwise genomic distance matrix was projected in two dimensions using multi-dimensional scaling (MDS) method implemented in the base package of the R statistical language[68]. The program ADMIXTURE was used to characterized genetic ancestry components based on the genome-wide and PGx variant sets using *K*=3 clusters[69].

The differences in PGx variant effect allele frequencies (*f*) between Antioquia (*ANT*) and Chocó (*CHO*) were measured as (1) the log-transformed ratio of the population-specific allele frequencies as:

$$\log_2(f_{ANT}/f_{CHO}) \quad (3)$$

and (2) as the population-specific allele frequency difference as:

$$\Delta = f_{ANT} - f_{CHO} \quad (4)$$

These two effect allele difference metrics were plotted orthogonally and the Euclidean distance from the origin was calculated for each PGx variant to yield a composite difference.

2.3.3 PGx variant ancestry associations

The influence of genetic ancestry on PGx variant genotype frequencies was measured via ancestry association analysis. To do this, individuals' genetic ancestry fractions – African, European, and Native American – inferred using ADMIXTURE with the genome-wide SNP set, were regressed against their individual PGx variant genotypes. The strength of the resulting ancestry × PGx variant associations were quantified using a linear regression model: $y = \beta x + \epsilon$, where $x \epsilon \{0, 1, 2\}$, corresponding to the number of PGx variant effect alleles, y is the ancestry fraction for a given ancestral group (African, European, or Native American), and β quantifies the strength of the association. The significance of the ancestry association is measured as the *P*-value obtained from a *t*-test, where $t = \beta / SE_{\beta}$.

2.3.4 Exome sequence analysis

Whole exome sequence (WES) analysis was conducted on a cohort of 132 de-identified patients characterized for the purposes of genetic testing by the GenomaCES laboratory[70]. The study was carried out in accordance with article 11 of resolution 8430 of 1993 of Colombian law, which states that for every investigation in which a human being is the study subject, respect for their dignity and the protection for their rights should always be present. The study protocol was reviewed and approved by the ethics committee and the research committee of Universidad CES, and all subjects gave written informed consent authorizing use of their biological samples and genetic information obtained through exome sequencing for research and academic training in accordance with the

Declaration of Helsinki. Patient DNA was extracted from peripheral blood using the salting out method[71]. Exon enrichment was performed using the Integrated DNA Technologies xGen capture kit, and exome sequencing was performed on the Illumina HiSeq 4000, generating 150bp paired end reads at 100X coverage. Read quality was assessed using the FastQC program with a threshold of Q≥30[72]. Sequence reads were mapped to the hs37d5 (1000 Genomes Phase II) human genome reference sequence using SAMtools[73], and variants were called using VarScan 2[74]. The resulting VCF files were surveyed for the presence of PGx variant alleles using the VCFtools package[75]. Manual inspection of the mapped sequence reads in support of PGx variant variant calls was performed using the Integrative Genomics Viewer (IGV)[76].

2.3.5 Allele-specific PCR assay

The identity of PGx variant allelic variants was assayed in the same 132 patients using custom-designed allele-specific PCR assays following the Web-based Allele-Specific PCR (WASP) primer design protocol[77]. Both the WASP and Primer-BLAST[78] tools were used to design pairs of allele-specific forward primers that overlap with the PGx variants of interest and their corresponding single reverse primers. PCR assays were performed using the Thermo Scientific™ Taq DNA Polymerase kit, with 25 μL final reagent volume, on the Bio-Rad thermocycler (C1000 Touch™ Thermal Cycler). PCR products were visualized and scored as homozygous non-effect allele, heterozygous, or homozygous effect allele using electrophoresis performed with 2.5% agarose gels stained

with ethidium bromide (10 μL) with a running time of 60 minutes at 70V in 1X TBE buffer. UV light was used to visualize the gel-separated PCR products.

2.4 Results

2.4.1 Pharmacogenomic SNP variation worldwide

We operationally define pharmacogenomic single nucleotide polymorphisms (PGx variants) as human nucleotide variants that are known to affect how individuals respond to medications. The Pharmacogenomics Knowledgebase (PharmGKB *https://www.pharmgkb.org/*) provides a catalog of PGx variants together with information regarding their known impacts on drug response. PharmGKB categorizes PGx variants with respect to their specific effects on drug efficacy, dosage, or toxicity/adverse drug reactions as well as the level of evidence for their role in drug response: (1) high, (2) moderate, (3) low, or (4) preliminary. We mined the PharmGKB database for PGx variants across all four evidence levels, yielding a total of 1,995 SNPs genome-wide.

We evaluated the global patterns of PGx variant variation using whole genome sequence data for 26 populations from 5 continental (super) population groups characterized as part of the 1000 Genomes Project[65]. Levels and patterns of variation for PGx variants were compared to a genome-wide background set of >28 million SNPs. Across all 26 global populations, PGx variants show a very high average minor allele frequency (avg. MAF=0.25) compared to genome-wide SNPs (avg. MAF=0.02) (Figure 1A). PGx variants also show significantly higher levels of the fixation index (avg.

F_{ST}=0.07), a measure of between-population differentiation, for global populations compared to genome-wide SNPs (avg. F_{ST}=0.01) (Figure 1B).

It should be noted that the higher average minor allele frequency observed for PGx variants compared to genome-wide SNPs could reflect an ascertainment bias owing to a relative excess of rare variants in the 1000 Genomes Project sequence data. However, no such bias is expected for the F_{ST} values as calculated here, which are largely unaffected by the presence of rare variants in the 1000 Genomes Project data[79].

Given the high levels of variation and between-population discrimination shown by PGx variants, we also evaluated the extent to which they carry information about genetic ancestry and admixture, particularly for the Colombian populations of Antioquia and Chocó. Pairwise genomic distances were computed for the Colombian populations together with a set of global reference populations from Africa, the Americas, and Europe, using both PGx variants and the genome-wide SNP set. Pairwise genomic distances computed using both sets of SNPs were used to reconstruct the evolutionary relationships among human populations worldwide. The results for the genome-wide (Figure 1C) and PGx variant (Figure 1D) sets are highly similar. The genome-wide SNP set does provide higher resolution and tighter groupings than the PGx variants, but the nature of the relationships among global populations does not change between the two SNP sets. The African, European, and Native American populations occupy the three poles of the MDS plot, with Antioquia falling along the axis between the European and Native American groups and Chocó grouping more closely with the African populations. Both Colombian

populations show evidence of substantial admixture compared to the global reference populations.

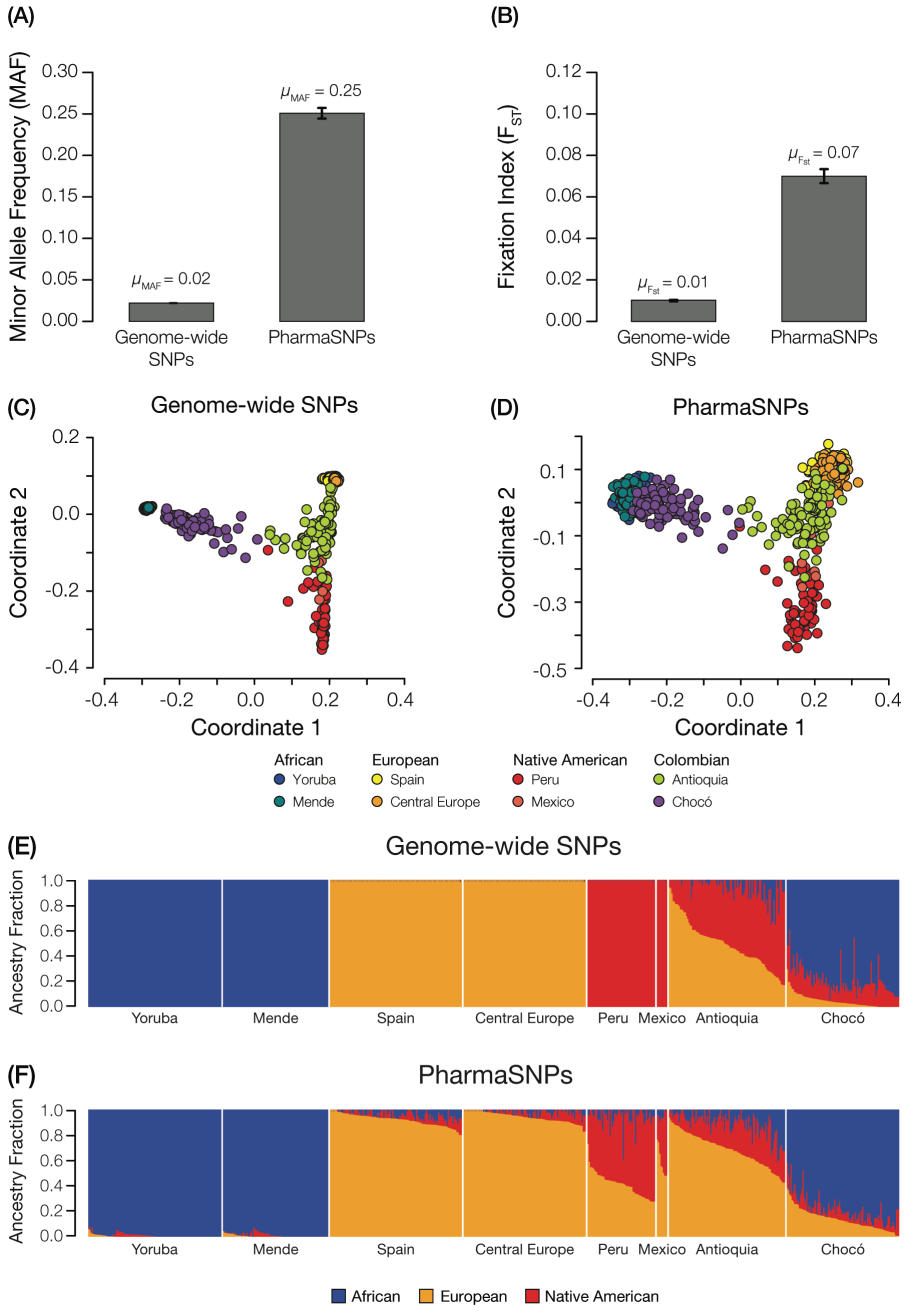


Figure 1. Patterns of variation for pharmacogenomic SNPs worldwide

Average (A) minor allele frequency (MAF) and (B) fixation index (FST) values for all genome-wide SNPs (n= 28,137,656) and all PGx variants (n= 1995) across the 26 1KGP

populations studied here. Multi-dimensional scaling (MDS) plots showing the inter-individual genetic distances of admixed Colombian individuals (Antioquia and Chocó) in relation to global reference populations from Africa, Europe, and the Americas for (C) genome-wide SNPs and (D) PGx variants. ADMIXTURE plots showing the genome-wide continental ancestry fractions using (E) all genome-wide SNPs and (F) only PGx variants for admixed Colombian populations (Antioquia and Chocó) and reference African (blue), European (orange), and Native American (red) populations.

Given the high levels of variation and between-population discrimination shown by PGx variants, we also evaluated the extent to which they carry information about genetic ancestry and admixture, particularly for the Colombian populations of Antioquia and Chocó. Pairwise genomic distances were computed for the Colombian populations together with a set of global reference populations from Africa, the Americas, and Europe, using both PGx variants and the genome-wide SNP set. Pairwise genomic distances computed using both sets of SNPs were used to reconstruct the evolutionary relationships among human populations worldwide. The results for the genome-wide (Figure 1C) and PGx variant (Figure 1D) sets are highly similar. The genome-wide SNP set does provide higher resolution and tighter groupings than the PGx variants, but the nature of the relationships among global populations does not change between the two SNP sets. The African, European, and Native American populations occupy the three poles of the MDS plot, with Antioquia falling along the axis between the European and Native American groups and Chocó grouping more closely with the African populations. Both Colombian populations show evidence of substantial admixture compared to the global reference populations.

We performed a similar comparison of the ability PGx variants to quantify patterns of genetic ancestry compared to genome-wide SNPs using the program ADMIXTURE. Using K=3 ancestry components, genome-wide SNPs clearly distinguish the reference

African, European, and Native American populations, and characterize the Colombian populations of Antioquia and Chocó as distinct mixtures of all three ancestries (Figure 1E). Consistent with previous results[58], Antioquia shows an average of 61% European, 32% Native American, and 7% African ancestry, whereas Chocó shows primarily African ancestry (76%) followed by 13% Native American, and 11% European fractions. PGx variants show qualitatively similar results albeit with lower resolution compared to the genome-wide SNP set (Figure 1F). Using PGx variants, the global reference populations are not quite as distinct, and the European component of ancestry appears to be over-estimated in both the Native American reference populations as well as Antioquia and Chocó. Nevertheless, the clear distinction between the patterns of ancestry and admixture for the Colombian populations, whereby Antioquia is primarily European and Chocó is mostly African, is captured when only the PGx variants are used.

2.4.2 Pharmacogenomic SNP variation in Colombia: Antioquia versus Chocó

Despite the fact that the Colombian administrative departments of Antioquia and Chocó are located in close proximity, their populations have distinct global origins (Figure 2A). As discussed in the previous section and elsewhere[53,58,59], the population of Antioquia shows mainly European genetic ancestry with substantial Native American admixture, whereas Chocó has primarily African ancestry with lower levels of Native American and European admixture. In light of the high levels of global variation seen for PGx variants (Figure 1), we expected to see pronounced differences in the distributions of PGx variant alleles between Antioquia and Chocó. Such differences should have implications for

public health strategies in the country, particularly with respect to the allocation of resources for pharmacogenomic testing.

We compared the frequencies of PGx variant effect alleles between Antioquia and Chocó to test this hypothesis. PGx variant effect alleles are operationally defined for this purpose as the allelic variants that increase the observed effect for a given drug-gene interaction, *i.e.* the alleles that increase the efficacy, dosage, or risk of toxicity/adverse drug responses for a drug. To ensure maximum relevance of our results for public health in Colombia, we focused on PGx variants corresponding to the highest evidence levels in PharmGKB (levels 1 and 2; n=155 PGx variants). PGx variant effect allele frequency differences between Antioquia and Chocó were measured in two ways – (1) as the log transformed ratio of allele frequencies Antioquia/Chocó and (2) as the allele frequency differences between Antioquia and Chocó – in order to capture both high relative differences at low allele frequencies and high absolute differences at high allele frequencies (Figure 2B). When these two dimensions of PGx variant effect allele frequency differences are plotted orthogonally, the Euclidean distance from the origin captures the overall between-population difference seen for each SNP (Figure 2C).

As expected, numerous PGx variant effect alleles show large frequency differences between Antioquia and Chocó (Figure 2). We sought to quantify the role that the distinct genetic ancestry profiles of these two populations plays in these PGx variants effect allele frequency differences. To do so, we developed and applied an ancestry association method whereby individuals' genetic ancestry fractions – African, European, and Native American – are regressed against their genotypes for any given PGx variant. This approach allows

us to visualize and quantify the influence of genetic ancestry on PGx variants genotype frequencies in these two diverse Colombian populations. Figure 3 shows examples of ancestry associations for three PGx variants with high levels of effect allele (and genotype) divergence between Antioquia and Chocó; ancestry associations for nine additional PGx variants of interest to Colombia can be seen in Appendix A: Figure 25. Table 1 shows the results of ancestry association analyses for 13 PGx variants of interest to Colombia, based on high levels of divergence between Antioquia and Chocó, and Appendix A: Table 9 contains the ancestry association results for all level 1 and 2 PharmGKB SNPs showing PGx variant effect allele Euclidean distances >0.5 (as shown in Figure 2C).

Tacrolimus: The T allele of the PGx variant rs776746 (*CYP3A5**3) is found at higher frequency in Chocó and is positively correlated with African ancestry and negatively correlated with both European and Native American ancestry (Figure 3A). This PGx variants is a splice site acceptor variant located within an intron of the *CYP3A5* (Cytochrome P450 Family 3 Subfamily A Member 5) encoding gene. The T allele is associated with increased metabolism of Tacrolimus, an immunosuppressive drug often used to treat transplant patients, and thus individuals with T containing genotypes may require relatively higher dosages of this drug. Consistent with these observations, physicians in Cali, Colombia have anecdotally reported that Afro-Colombian transplant patients do not respond well to standard doses of Tacrolimus.

Warfarin: The C allele of the PGx variant rs9923231 (*VKORC1**2) shows a similar pattern with higher frequency in Chocó, a positive correlation with African ancestry, and negative correlations with both European and Native American ancestry (Figure 3B). This

PGx variant is one of several variants of the *VKORC1* (Vitamin K Epoxide Reductase Complex Subunit 1) encoding gene that have been associated with warfarin sensitivity. The SNP is located in the upstream, regulatory region of the gene, and individuals with the C allele may require an increased dosage of warfarin.

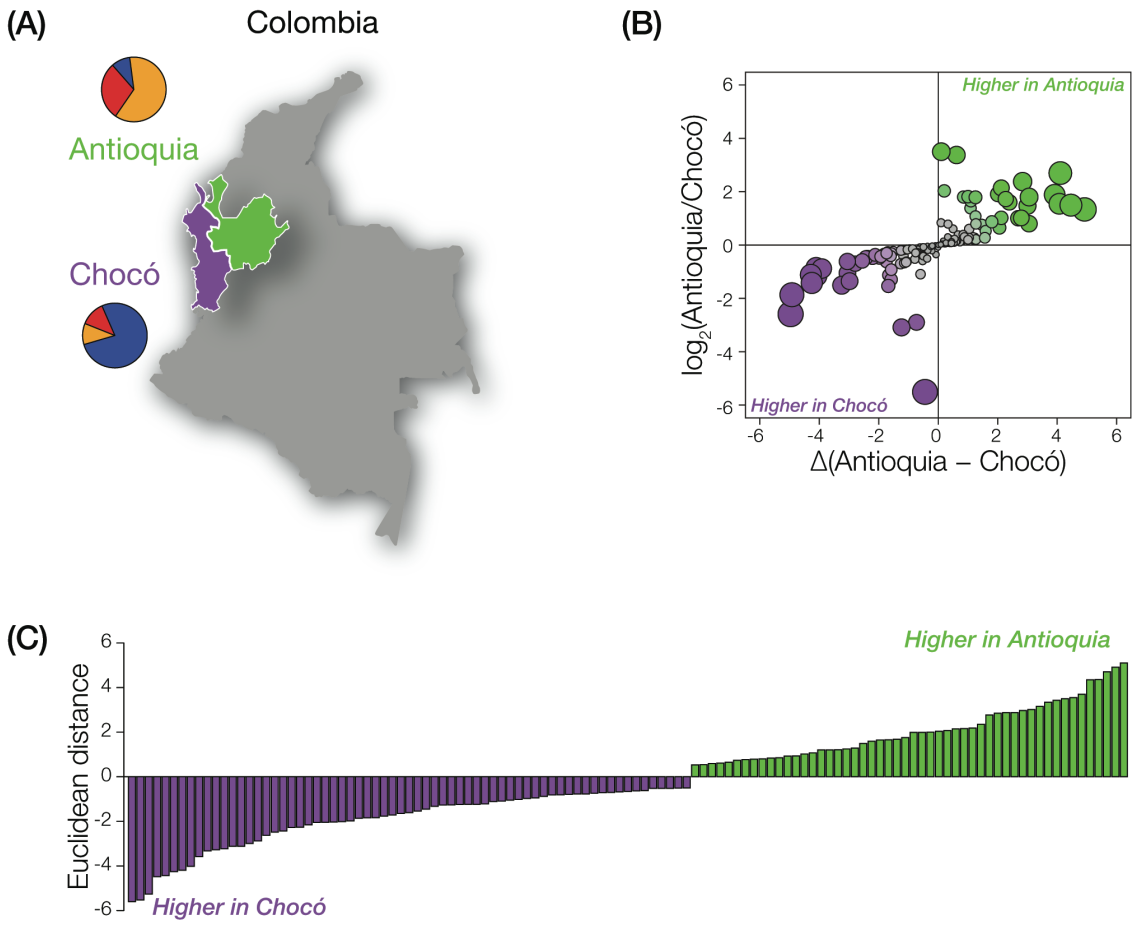


Figure 2. PGx variants with population-specific effect allele frequency differences in Colombia.

(A) Map of Colombia, highlighting Antioquia in green and Chocó in purple. Population-specific mean ancestry fractions are shown as pie charts: African (blue), European (orange), and Native American (red). (B) Comparison of the ratio of PGx variant effect allele frequency differences between Antioquia and Choco (y-axis) to the magnitude of the frequency differences (x-axis). Circles are scaled according to their Euclidean distance

(distance from the origin) and are colored to indicate the direction of their difference (green – higher effect allele frequency in Antioquia; purple – higher effect allele frequency in Chocó). (C) Distribution of PGx variants with Euclidean distance>0.5. Green indicates that the PGx variant effect allele is more frequent in Antioquia, while purple indicates that the effect allele is more frequent in Chocó.

Warfarin: The C allele of the PGx variant rs9923231 (*VKORC1**2) shows a similar pattern with higher frequency in Chocó, a positive correlation with African ancestry, and negative correlations with both European and Native American ancestry (Figure 3B). This PGx variant is one of several variants of the *VKORC1* (Vitamin K Epoxide Reductase Complex Subunit 1) encoding gene that have been associated with warfarin sensitivity. The SNP is located in the upstream, regulatory region of the gene, and individuals with the C allele may require an increased dosage of warfarin.

Simvastatin: The C allele of the PGx variant rs4149056 (*SLCO1B1**5) is found in higher frequency in Antioquia, showing a negative correlation with African ancestry and a positive correlation with European ancestry (Figure 3C). The correlation with Native American ancestry for this SNP is not significant. This SNP is a missense variant in the *SLCO1B1* (Solute Carrier Organic Anion Transporter Family Member 1B1) encoding gene. The C allele is associated with simvastatin toxicity, and individuals with this allele may be at higher risk for simvastatin-related myopathy. These results agree very well with observations of physicians from the Universidad CES clinic in Antioquia, who have observed that ~30% of patients treated with Simvastatin show evidence of adverse drug reactions.

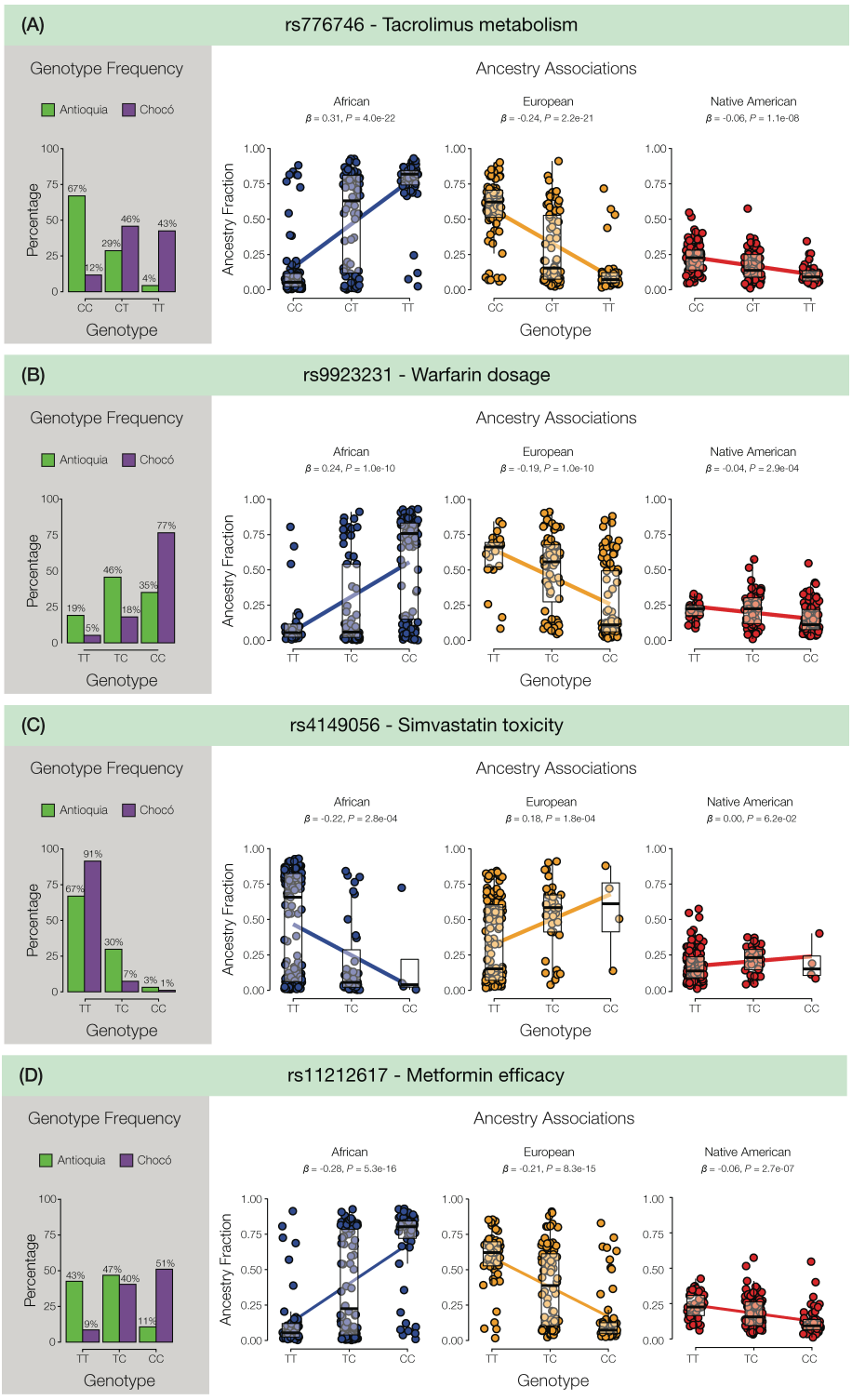


Figure 3. Ancestry associations for PGx variants in Colombia.

For each panel in the figure, PGx variant genotype frequencies are shown for Antioquia (green) and Chocó (purple) followed by the ancestry association plots. For each genetic

*ancestry component – African (blue), European (orange), and Native American (red) – individuals' ancestry fractions (y-axis) are regressed against their PGx variant genotypes (x-axis). Ancestry associations are quantified by the slope of the regression (β) and its significance level (P). Results are shown for (A) the tacrolimus metabolism-associated SNP rs776746 (CYP3A5*3), (B) the warfarin dosage-associated SNP rs9923231 (VKORC1*2), (C) the simvastatin toxicity-associated SNP rs4149056 (SLCO1B1*5), and (D) the metformin efficacy-associated SNP rs11212617.*

Metformin: The C allele of the PGx variant rs11212617 is found at substantially higher frequency in Chocó compared to Antioquia, and it is positively correlated with African ancestry and negatively correlated with both European and Native American ancestry (Figure 3D). This PGx variant shows an interaction with the type 2 diabetes drug Metformin; the C effect allele was found to be associated with greater treatment success[80]. Interestingly, Metformin was subsequently proven to have higher efficacy for the reduction of blood glucose levels reduction in African-Americans compared to European-Americans[81,82]. Ergo, this ancestry associated PGx variant shows a direct connection between genetic ancestry differences and differential drug response.

Table 1. Colombian ancestry-associated PGx variants of interest.

The PGx variants marked with an asterisk (*) are shown in Figure 3.

PGx variant	Effect	Frequency		African Ancestry Correlation		European Ancestry Correlation		Native American Ancestry Correlation	
		Antioquia	Chocó	β value	*P*-value	β value	*P*-value	β value	*P*-value
rs776746 *	Tacrolimus metabolism	0.81	0.32	0.31	4.00e-22	-0.24	2.20e-21	-0.06	1.10e-08
rs1799853	Warfarin dosage	0.88	0.97	-0.25	4.20e-04	0.21	2.50e-04	0.04	8.40e-02
rs9923231 *	Warfarin dosage	0.57	0.88	0.24	1.00e-10	-0.19	1.00e-10	-0.04	2.90e-04
rs4149056 *	Simvastatin toxicity	0.18	0.05	-0.22	2.80e-04	0.18	1.80e-04	0.00	6.20e-02
rs4244285	Clopidogrel efficacy	0.90	0.84	0.09	1.36e-01	-0.06	1.87e-01	-0.02	1.67e-01
rs2740574	Tacrolimus metabolism	0.10	0.59	0.32	1.20e-22	-0.24	2.20e-21	-0.06	1.00e-09
rs11615	Platin toxicity	0.48	0.07	-0.30	9.80e-18	0.25	2.50e-19	0.04	1.10e-04
rs11212617	Metformin efficacy	0.33	0.74	0.28	5.30e-16	-0.21	8.30e-15	-0.06	2.70e-07
rs6977820	Antipsychotic drug toxicity	0.25	0.68	0.28	5.30e-16	-0.19	1.70e-12	-0.07	1.20e-11
rs3812718	Antiepileptic treatment resistance	0.55	0.27	-0.27	2.00e-06	0.21	8.40e-06	0.05	7.74e-04
rs7793837	Salbutamol efficacy	0.69	0.24	0.21	4.30e-17	0.21	2.90e-16	0.05	2.80e-07
rs1954787	Antidepressant efficacy	0.62	0.21	-0.24	3.00e-13	0.18	6.80e-12	0.05	4.20e-07
rs1719247	Simvastatin adverse reaction	0.54	0.27	-0.21	3.90e-08	0.17	3.50e-08	0.04	2.00e-03

2.4.3 Cost-effective PGx variant genotyping in Colombia with allele-specific PCR

The results from the analysis of PGx variant variation in Colombia uncovered a number of SNPs with specific relevance to the country, in terms of anomalous effect allele frequencies within specific populations, associations with different genetic ancestry groups, and broad relevance to public health. We reasoned that such population genomic

profiling can be used to focus efforts to develop precision medicine in the country and to maximize the return on investment for pharmacogenomic testing in resource-limited settings. To this end, GenomaCES developed and validated three custom allele-specific PCR assays to genotype PGx variants of special relevance to these Colombian populations.

The criteria for the selection of PGx variants that were interrogated with our custom allele-specific PCR assays included the PharmGKB evidence level along with a combination of population genomic and clinical information. Pharmacogenomic assays were only developed for PGx variants from the PharmGKB evidence level 1A. This is the highest evidence level and corresponds to PGx variants that are included in medical society-endorsed pharmacogenomics guidelines and/or implemented in major health systems. The additional criteria used to prioritize PGx variants for the development of allele-specific PCR assays were: (i) observations of population-specific allele frequencies in Colombia along with related ancestry-associations, (ii) pharmacogenomic associations with drugs that are widely prescribed in Colombia and used to treat common conditions, and (iii) pharmacogenomic associations with drugs for which GenomaCES investigators have anecdotal information from collaborating physicians that pharmacogenomic tests would be of use to the local population, based on their observations of anomalous drug responses in their patients. It should be noted that the population and clinical criteria are not mutually exclusive; indeed, physicians' observations of anomalous drug responses in their patient populations are almost certainly related to the population-specific allele frequencies of the relevant PGx variants

An example of an allele-specific PCR assay developed for the simvastatin-associated PGx variant rs4149056 (*SLCO1B1**5), located with an exon of the *SLCO1B1* protein coding gene on the short arm of chromosome 12, is shown in Figure 4A. The PGx variant variant detection assay relies on the use of two forward primers – one to capture the non-effect allele T and one to capture the effect allele C – and a single reverse primer. Use of these two primer-pairs results in allele-specific amplicons, depending on the presence of each allele in an individual patient's genome. PCR results are shown for four patients: Patient-132 homozygous TT, Patient-44 heterozygous TC, Patient-17 and Patient-26 homozygous CC (Figure 4B). We visualized the results of exome sequence analysis, with respect to the quality and coverage of mapped reads along with the counts of the different variant calls, to manually confirm the results of the allele-specific PCR assays (Figure 4C).

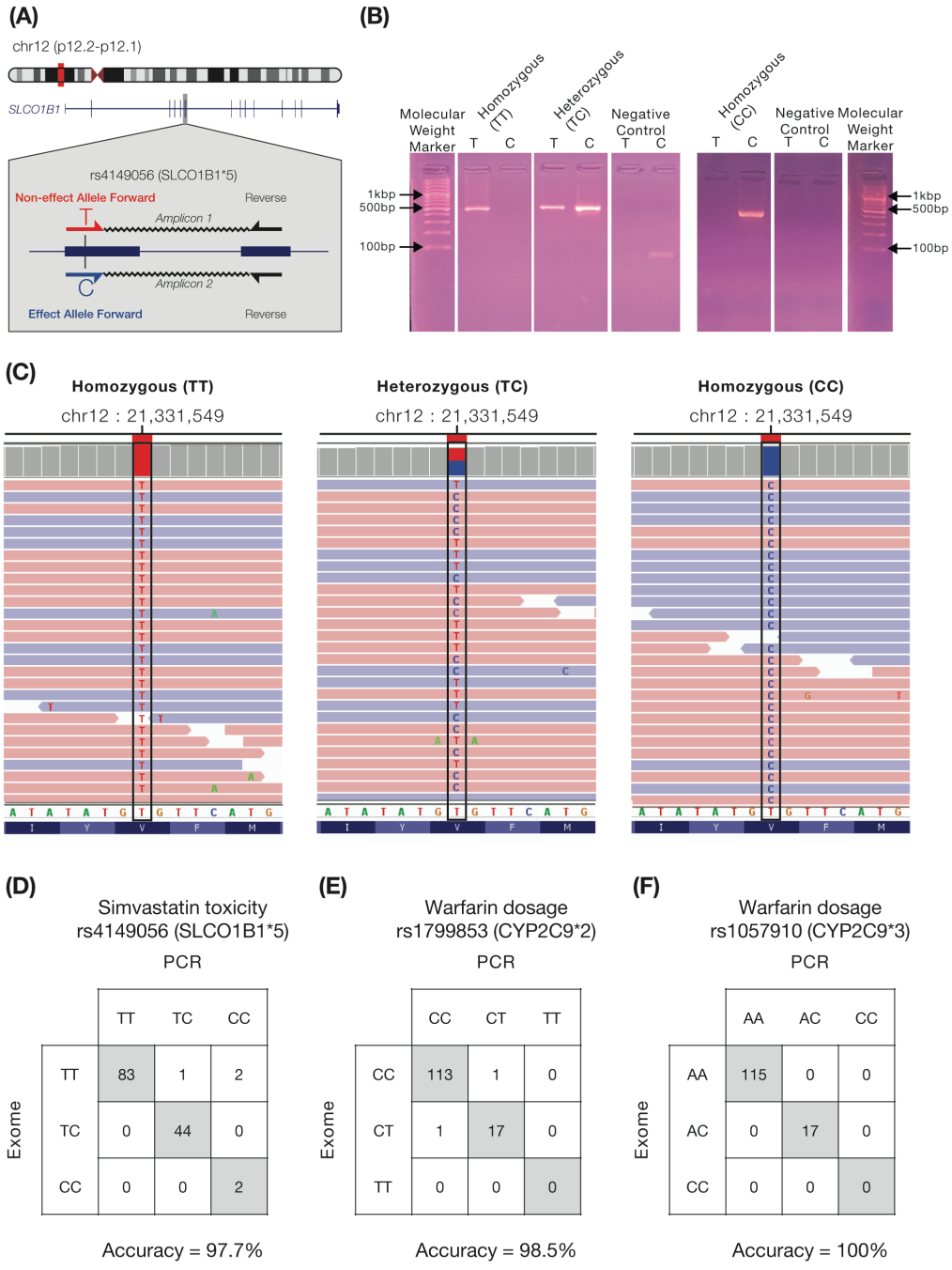


Figure 4. Allele-specific PCR assay for PGx variants.

*(A) Schema depicting the design of the allele-specific PCR assay for the PGx variant rs4149056 (SLCO1B1*5) on chromosome 12. Two allele-specific forward primers are designed for the PGx variant of interest and paired with a single reverse primer, yielding allele-specific amplicons. (B) Allele-specific PCR results for four individuals are shown. PCR gel lanes are labeled with the allele used for the forward primer – T or C. (C) Results*

*of exome sequence analysis used to confirm the results of the allele-specific PCR assays. Sequence reads (red – forward, blue – reverse) mapped to the genomic position for the SNP rs4149056, coverage levels (gray boxes above), and the identity of the called nucleotide variants at that same position are shown along with the reference nucleotide and amino acid sequences for the corresponding region of the SLCO1B1gene (protein). Images were taken from the Integrative Genomics Viewer (IGV). Confusion matrices showing comparisons between the PGx variant variant calls made via exome sequence analysis and the allele-specific PCR assays are shown for (D) the simvastatin toxicity SNP rs4149056 (SLCO1B1*5), and the warfarin dosage SNPs (E) rs1799853 (CYP2C9*2) and (F) rs1057910 (CYP2C9*3). Identical variant calls are shown along the diagonal, whereas off-diagonal calls show discrepancies between the exome and PCR variant calls; accuracy levels for each test are shown.*

Having confirmed the accuracy of the rs4149056 (*SLCO1B1**5) variant detection assay, we then ran it on a cohort of 132 de-identified patients from the GenomaCES laboratory, all of whom have exome sequences available for confirmatory analysis. The results of the allele-specific PCR and exome analyses are highly similar; taking the exome results as the ground truth against which to compare the PCR assay yields an overall accuracy of 97.7% for this test (Figure 4D). Two additional allele-specific PCR assays for SNPs associated with warfarin dosage – rs1799853 (*CYP2C9**2) and rs1057910 (*CYP2C9**3) – were tested on the same patient set and confirmed via exome sequence analysis. These two allele-specific PCR genotyping assays show even higher accuracies of 98.5% and 100%, respectively. We calculated a number of additional performance metrics for all three of these tests, breaking down each assay into its three constituent genotypes, the results of which are shown in Appendix A: Figure 26.

2.5 Discussion

2.5.1 Caveats and limitations

We would like to point out some of the caveats and limitations of the current study as they relate to the accuracy and utility of pharmacogenomic tests in understudied populations. The reach of our analysis is somewhat limited by the focus on PGx variants, *i.e.* single nucleotide variants, as opposed to all possible genetic variants that may impact drug response. PharmGKB contains annotations of gene-to-drug response interactions that are mediated by a number of different kinds of variants, including larger scale structure variants such as insertion/deletion events and copy number variations[83,84]. Furthermore, there are a number of pharmacogenomic tests that rely on the characterization of combinations of linked SNPs, *i.e*. haplotypes or star-alleles. For example, the most reliable warfarin sensitivity assays utilize multiple SNPs (haplotypes) across two genes in order to arrive at specific dosage recommendations[85,86]. Our survey of PGx variant variation will not capture these complex classes of pharmacogenomic variants and interactions.

Our focus on PGx variants can be primarily attributed to the availability and the reliability of SNP data at our disposal, as opposed to other more complex genetic variants, particularly for the population of Chocó, which was characterized using a genome-wide SNP array[58,59,66]. Nevertheless, it is important to note that (i) there are numerous documented cases of individual SNPs that show demonstrable and reproducible effects on drug response[87] and (ii) there are many more PGx variants available for analysis compared to the other variant classes[64]. For example, ~93% of PharmGKB variant annotations correspond to individual PGx variants (1,995 out of 2,144 total variants). Accordingly, we

are confident that our study design captures the majority of the pharmacogenomically relevant human genetic variation based on current knowledge in the field.

Another limitation relates to the fact that we compared PGx variant allele frequencies among populations with distinct ancestries compared to the cohorts where they were originally characterized. As with other classes of clinical genetics studies[43,88], there remains a very strong bias whereby the majority of pharmacogenetic clinical trials have been conducted in developed countries on cohorts with European ancestry[89,90]. Thus, it is formally possible that the PGx variants we analyzed may have different effects on drug response in our populations of interest. Of course, the most rigorous way to assess the population-specific role of genetic variation in drug response would be to conduct clinical trials in all populations of interest. Currently however, the high cost and complexity of performing clinical trials across multiple populations, particularly for variants with already well documented effects on drug response, renders this approach prohibitive. In addition, it is important to point out that the associations between PGx variants and drug response that our study relies on are far more likely to be causal than associations uncovered by genome-wide association studies (GWAS), many of which do not replicate across populations with distinct ancestry profiles[91]. This is because GWAS SNPs do not correspond to causal variants *per se*; rather, they are tag variants that mark haplotypes wherein the causal SNPs lie, and haplotype structure is known to vary widely across populations[92]. PGx variants, on the other hand, correspond to the specific causal variants for which there is direct evidence of an impact on drug response. This is particularly the case for the narrower set of 155 PGx variants deemed to be most confident by PharmGKB, which we used for our comparison of Antioquia and Chocó. The strong clinical and

experimental evidence of these high confidence PGx variants effects on drug response gives us confidence with respect to their potential relevance for our populations of interest.

2.5.2 The underlying complexity of so-called Hispanic/Latino populations

As briefly mentioned in the previous section, a number of recent studies have underscored the major sampling bias that currently exists for human clinical genomic studies and emphasized the corollary importance of extending clinical trials to currently understudied populations. These studies rely on a variety of labels related to "Hispanic/Latino" to describe understudied populations from Latin America, or individuals and communities with origins in Latin America. For example, in a survey of the ancestry of study participants in GWAS cohorts, the authors used the label "Hispanic and Latin American ancestry", showing that members of this group made up a mere 0.06% of GWAS study participants in 2009 and 0.54% in 2016[43]. Another study, which demonstrated the importance of using matched ancestry samples for clinical variant interpretation, employed the category "Latino ethnicity" to classify exome variants into a single control group[88]. The widely used Exome Aggregation Consortium (ExAC) database uses the term "Latino" as a population category for exome sequence variants[93], and the 1000 Genomes Project uses the super population code "Ad Mixed American (AMR)" to group genetically diverse populations from Colombia, Mexico, Peru, and Puerto Rico[65].

It is interesting to note that the origins of the term Hispanic/Latino as a catch-all phrase to describe an extraordinarily diverse set of populations can be traced to decisions imposed by activists and bureaucrats of the US Census Bureau, motivated by the

opportunity to create a politically influential interest group[94]. The results of our study highlight the artificial nature, and the lack of practical utility, of the Hispanic/Latino label as it pertains to clinical genetic studies. Our two populations of interest – Antioquia and Chocó – would both be considered Hispanic/Latino, and in fact they are both from the same country within Latin America, but they have very distinct patterns of genetic ancestry and admixture. Furthermore, we show here that the differences in genetic ancestry have specific implications for the pharmacogenomic profiles of each population. The same thing will certainly hold true for many other sets of populations both within and between different Latin American countries. In light of this realization, we would like to emphasize that the stratification of so-called Hispanic/Latino populations for clinical genetic studies should be performed using their distinct genetic ancestry profiles as opposed to a politically imposed pan-ethnic label.

2.5.3 Population-guided approaches to pharmacogenomics in the developing world

We hope that the population pharmacogenomic approach we applied to Colombian populations in this study can serve as model for their broader application in the developing world. Currently, genomic approaches to precision medicine are prohibitively expensive for many developing countries owing to their reliance on deep genetic characterization of individual patients. Precision public health, on the other hand, entails population-level interventions, and the focus on populations can provide a more cost-effective means for the implementation of novel genomic approaches to healthcare[21-23]. Population-guided approaches to pharmacogenomics allow healthcare providers to allocate resources and

efforts where they will be most effective by uncovering pharmacogenomic variants with special relevance to specific populations[49,50].

Here, we report a number of examples of pharmacogenomic variants with anomalously high effect allele frequencies in distinct Colombian populations. For example, the T allele of the PGx variant rs776746 is associated with African ancestry and found at a relatively high frequency in Chocó (Figure 4 and Table 1). Since this variant is associated with the need for a higher dosage of the immunosuppressive drug Tacrolimus, Afro-Colombians may be particularly prone to organ rejection following allogeneic transplant. Accordingly, the local deployment of a pharmacogenomic test for this particular SNP in Chocó would simultaneously focus limited resources for genetic testing while also ensuring an outsized impact for Afro-Colombian patients. As another example, the population of Antioquia shows an elevated frequency of the C allele of the PGx variant rs4149056, which is associated with increased risk of simvastatin toxicity (Figure 4 and Table 1). The development of a pharmacogenomic assay for this SNP, which is currently underway at GenomaCES in Antioquia, could help to mitigate the risk of adverse drug reactions to this commonly prescribed medication in the local population.

2.6 Conclusion

This is an auspicious moment for the development of pharmacogenomic approaches to public health in Colombia. The Colombian biomedical community is simultaneously faced with a combination of great opportunities and profound challenges, both with respect to genomic medicine overall and for pharmacogenomics in particular[95]. In all of Latin

America, Colombia is one of only two countries, together with Argentina, with nationalized healthcare systems that guarantee comprehensive coverage for all of its citizens. In 2015, the terms of this guarantee were updated, via the Ministry of Health and Social Protection resolution 5592, to cover broadly defined molecular genetic and genomic tests. This change resulted in a far more comprehensive coverage policy for these kinds of tests than currently exists in the United States, where many precision medicine treatments are still directly paid by patients[96]. This resolution reflects great foresight on the part of Colombian policy makers and represents a tremendous opportunity for local biomedical researchers, clinicians, and the patients that they serve. Furthermore, a very strong case has been made for how genome-enabled approaches to precision medicine should ultimately lead to substantial cost savings for the national healthcare system over the long term[97,98].

On the other hand, the costs of many of the tests covered by this policy are so expensive in Colombia that the sustainability of the policy has been called into serious question. For example, the molecular biology reagents needed for tests of this kind can often cost three-times as much or more in Colombia, compared to the United States, owing to taxes and tariffs. We firmly believe that key solutions to this economic challenge will be to (i) build the local capacity needed to perform such tests and (ii) develop genomic assays that are specifically tailored to the needs of Colombian populations. To these ends, Universidad CES has invested substantially in the development of local capacity in genomic medicine via the establishment of GenomaCES, which is Colombia's first homegrown genomic medicine laboratory. As we have shown here, GenomaCES is working to develop inexpensive and rapid pharmacogenetic genotyping tests based on relatively simple allele-specific PCR assays. Developing local tests of this kind can help to ensure that variants of

specific relevance to the country are prioritized for testing and to avoid the prohibitively high costs of commercially available tests and/or kits.

CHAPTER 3. POPULATION STRUCTURE AND PHARMACOGENOMIC RISK STRATIFICATION IN THE UNITED STATES

3.1 Abstract

Pharmacogenomic (PGx) variants mediate how individuals respond to medication, and response differences among racial/ethnic groups have been attributed to patterns of PGx diversity. We hypothesized that genetic ancestry (GA) would provide higher resolution for stratifying PGx risk, since it serves as a more reliable surrogate for genetic diversity than self-identified race/ethnicity (SIRE), which includes a substantial social component. We analyzed a cohort of 8,628 individuals from the United States (US), for whom we had both SIRE information and whole genome genotypes, with a focus on the three largest SIRE groups in the US: White, Black (African-American), and Hispanic (Latino). Our approach to the question of PGx risk stratification entailed the integration of two distinct methodologies: population genetics and evidence-based medicine. This integrated approach allowed us to consider the clinical implications for the observed patterns of PGx variation found within and between population groups. Whole genome genotypes were used to characterize individuals' continental ancestry fractions – European, African, and Native American – and individuals were grouped according to their GA profiles. SIRE and GA groups were found to be highly concordant. Continental ancestry predicts individuals' SIRE with >96% accuracy, and accordingly GA provides only a marginal increase in resolution for PGx risk stratification. PGx variants are highly diverged compared to the genomic background; 82 variants show significant frequency

differences among SIRE groups, and genome-wide patterns of PGx variation are almost entirely concordant with SIRE. The vast majority of PGx variation is found within rather than between groups, a well-established fact for all genetic variants, which is often taken to argue against the clinical utility of population stratification. Nevertheless, analysis of highly differentiated PGx variants illustrates how SIRE partitions PGx variation based on groups' characteristic ancestry patterns. These cases underscore the extent to which SIRE carries clinically valuable information for stratifying PGx risk among populations, albeit with less utility for predicting individual-level PGx alleles (genotypes), supporting the concept of population pharmacogenomics. Perhaps most interestingly, we show that individuals who identify as Black or Hispanic stand to gain far more from the consideration of race/ethnicity in treatment decisions than individuals from the majority White population.

3.2 Introduction

Pharmacogenomic (PGx) variants are associated with inter-individual differences in drug exposure and response, affecting medication dosage, efficacy and toxicity[19,20]. A number of studies have shown racial and/or ethnic differences in drug response[99-103], based in part on group-specific differences in the frequencies of PGx variants[49]. A 2015 review found that 20% of drugs approved over the previous six years showed response differences among racial/ethnic groups, and these differences are often translated into group-specific prescription recommendations that are issued on FDA-approved drug labels[103]. Examples of such recommendations include contraindication of Rasburicase, a medication used to

clear uric acid from the blood in patients undergoing chemotherapy, for individuals of African or Mediterranean ancestry, and a toxicity warning for the anticonvulsant Carbamazepine in Asian patients. A higher dosage of the immunosuppressive drug Tacrolimus is indicated for African-American transplant patients, whereas a lower initial dose of Rosuvastatin is recommended for Asians. Despite the inclusion group-specific recommendations in a number of drug labels, the utility of racial and ethnic categories in biomedical research, and their relevance to clinical decision making, remain a matter of substantial controversy[104-107].

Critiques of the use of racial and ethnic categories in biomedical research point to the appalling history of race science[108-110] and stress the potential of such research to reify outmoded notions of racial difference[111-113]. This school of thought holds that race is a primarily a social construct with little or no biological (genetic) meaning[114-118]. As it relates to clinically relevant PGx variation across groups, the extent to which racial and ethnic categories serve as a reliable proxy for genetic diversity has also been called into question. The authors of the recent commentary 'Taking race out of human genetics' make a compelling case for eliminating the use of race as a category in genetic research, asserting that race and ethnicity are taxonomic (*i.e.* categorical) labels that by definition cannot capture the full complexity of individuals' genetic ancestry[119]. They suggest that genetics research should instead focus on biogeographically defined populations and genetic ancestry, as opposed to racial categories, and for this study we hypothesized that genetic ancestry should better partition PGx variation than SIRE. We posit that genetic ancestry provides a number of advantages over racial/ethnic categories for biomedical research: (i) it can be characterized independent of the social and environmental dimensions of

race/ethnicity, (ii) it can be measured objectively and with precision, and (iii) it can be quantified as a continuous variable, as opposed to categorical racial/ethnic labels. Indeed, a number of recent studies have focused on PGx variation among populations defined by genetic ancestry rather than racial and ethnic groups[60-63,120-122].

The goal of this study was to compare the relative utility of race/ethnicity versus genetic ancestry for partitioning PGx variation among populations in the United States (US). We focused on individuals aged 50 and older, 75% of whom take prescription medication on a regular basis[123], and restricted our study to the three largest racial/ethnic groups in the US: White, Black (or African-American), and Hispanic/Latino[124]. Our study cohort is made up of 8,629 participants from the Health and Retirement Study (HRS)[125], for whom we had both SIRE information and whole genome genotypes. We first compared the relationship between self-identified race/ethnicity (SIRE) and genetic ancestry (GA), characterized via analysis of whole genome genotype data, and we then measured the extent to which PGx variation is partitioned by SIRE versus GA. We provide a number of examples of PGx variants that are highly differentiated among groups and discuss the implications of these findings in light of population genetics and clinical decision-making.

Materials and methods

3.2.1 Study Cohort

Self-identified race and ethnicity (SIRE) information and whole genome genotypes for Americans over the age of 50 and their spouses were collected as part of a nationally-representative longitudinal panel study called the Health and Retirement Study (HRS) [125].

For the current study, only HRS participants with both SIRE and genotype information were considered (8,912 participants). The 284 participants who did not identify with one of the three largest racial/ethnic categories in the HRS data – non-Hispanic White (5,927), non-Hispanic Black (1,527), and Hispanic/Latino of any race (1,174) – were excluded from this analysis. This yielded a total of 8,628 individuals in our final analysis cohort.

3.2.2 Genetic Ancestry (GA) Analysis

HRS participants were previously genotyped at ~2,381,000 genomic sites using the Illumina Omni2.5 BeadChip[125]. Whole genome genotype data from HRS participants were compared to reference populations from Europe, Africa, and the Americas in order to infer their continental genetic ancestry patterns as previously described[126] (see Additional file 1: Table S1) [65,127,128]. Reference populations were taken from (i) the 1000 Genomes Project (648)[65], (ii) the Human Genome Diversity Project (110)[127], and (iii) 21 Native American populations from across the Americas (90)[128]. A custom script that employs PLINK version 1.9[129] was used to harmonize the HRS and reference population variant calls. The variant call data were merged by identifying the set of variants common to both datasets, with strand flips and variant identifier inconsistencies corrected as needed. The initial merged and cleaned variant data set was filtered for variants with >1% missingness and <1% minor allele frequency among samples. The final harmonized genotype data contains 228,190 genomic sites. The harmonized genotype dataset was phased using ShapeIT version 2.r837[130]. ShapeIT was run without reference haplotypes, and all individuals were

phased at the same time. Individual chromosomes were phased separately, and the X chromosome was phased with the additional '-X' flag.

A modified version of the RFMix program[126,131] was used to characterize the continental genetic ancestry patterns for the HRS participants, with European, African, and Native American populations used as reference populations. RFMix was run in the `PopPhased` mode with a minimum node size of five, using 12 generations and the "—use-reference-panels-in-EM" for two rounds of EM, to assign continental ancestry for haplotypes genome-wide. Contiguous regions of ancestral assignment, "ancestry tracts," were created where RFMix ancestral certainty was at least 95%, and genome-wide continental ancestry estimates for HRS participants were obtained by averaging across confidently assigned ancestry tracts.

Non-overlapping genetic ancestry (GA) groups were defined from individual participants' continental ancestry estimates obtained via RFMix analysis using *k*-means clustering implemented in the Python package Scikit-learn[132] with k=3. Each participant was represented as a point in three-dimensional (3-D) space, parameterized by their three continental ancestry fractions. Formally, the position of a participant (i) in this genetic ancestry space was defined by (E_i, A_i, N_i), where E_i, A_i, and N_i are the European, African, and Native American ancestry fractions. *K*-means clustering using Euclidean distances between all pairs of individual participants in this 3-D genetic ancestry space to yield three non-overlapping clusters. Given that *k*-means clustering can be unstable, the algorithm was run on these data 100 times and the most probable group membership was assigned to each participant. This method allowed us to define three non-overlapping groups of HRS

participants informed entirely by their genetic ancestry and free from the social dimensions of SIRE.

The association between GA and PGx variant genotypes was measured using our previously described method[120]. To obtain the strength of association (β) between continental ancestry proportions and genotypes, continental ancestry fractions were regressed against the observed PGx variant genotypes. Formally, the genetic ancestry fraction $y = \beta x + \varepsilon$, where $x \in \{0, 1, 2\}$ refers to the number of PGx variant effect alleles. The significance of these ancestry associations was quantified using a t-test.

3.2.3 Measurement of PGx Variation

Single nucleotide variants (SNVs) associated with pharmacogenomic response – *i.e.* PGx variants – were mined from the Pharmacogenomic Knowledgebase (PharmGKB)[20]. This online database is a source of manually curated clinical variant annotations for PGx variants and their associated drug-response phenotypes. Data on the chromosomal locations of PGx variants, the identity of PGx effect (risk) alleles, PGx variants' mode of effect (additive or dominant), clinical annotations, and clinical evidence levels were parsed and taken for analysis. A total of 2,351 PGx variants were accessed from PharmGKB, 989 of which were genotyped for the HRS cohort. Only directly genotyped PGx variants were used for analysis. PharmGKB annotates the specific effect alleles that are associated with inter-individual differences in drug dosage, efficacy, metabolism, and toxicity. The direction of effect (higher or lower) is specific to individual

PGx variants for dosage, efficacy, and metabolism whereas toxicity effect alleles always correspond to increased toxicity.

PGx allele frequencies for SIRE and GA groups were computed as the group-specific counts of effect alleles normalized by the total number of typed individuals for each group. Pairwise between group fixation index (F_{ST}) values for each variant were computed by calculating two components: (i) the mean expected heterozygosity within subpopulations as:

$$\bar{H}_S = \frac{1}{2}\sum_i 2(p_i)(1-p_i)(\frac{count_i}{total\ count}) \tag{5}$$

where p_i is the frequency of risk allele in population i, and $count_i$ is the number of individuals in population i, and $total\ count$ refers to the total number of individuals in both populations and (ii) the expected heterozygosity in the total population as:

$$H_T = 2(\bar{p})(1-\bar{p}) \tag{6}$$

where $\bar{p}$ is the mean effect allele frequency in both populations under consideration. The fixation index was computed by combining the two computed metrics as described as[133]:

$$F_{ST} = 1 - \frac{\bar{H}_S}{H_T} \tag{7}$$

PGx variants were used to calculate pairwise inter-individual distances for all HRS participants using PLINK, and the resulting distance matrix was projected into two dimensions using multi-dimensional scaling (MDS) with the mds function in R. *K*-means

clustering of the participants in MDS space was used to generate three non-overlapping PGx variant groups in the same way as described for the GA groups.

Odds ratios (ORs) were calculated for group-specific PGx effect allele counts[134]. In a contingency table for the counts of effect allele in population P_A with the four values: P_E (Effect allele count in P_A), P_N (Non-effect allele count in P_A), Q_E (Effect allele count in non-P_A individuals), Q_N (Non-effect allele count in non-P_A individuals), this was done using:

$$OR = \frac{P_E/Q_E}{Q_N/Q_N} \tag{8}$$

with confidence intervals calculated as:

$$CI = \exp\left(log(OR) \pm Z_{\alpha/2} * SE_{log(OR)}\right) \tag{9}$$

where α is 0.05, $Z_{\alpha/2}$ is 1.6, and SE per:

$$SE_{log(OR)} = \sqrt{\frac{1}{P_E} + \frac{1}{P_N} + \frac{1}{Q_E} + \frac{1}{Q_N}} \tag{10}$$

Similarly, using group-specific PGx effect counts the absolute risk increase (ARI) was calculated as:

$$ARI = \frac{P_E}{P_E + P_A} - \frac{Q_E}{Q_E + Q_A} \tag{11}$$

with confidence intervals calculated as:

$$CI = ARI \pm Z_{\alpha/2} \times SE_{ARI} \quad (12)$$

where α is 0.05, $Z_{\alpha/2}$ is 1.96, and SE per[135]:

$$SE_{ARI} = \sqrt{P_E P_A + Q_E Q_A} \quad (13)$$

Group-specific genotype prediction accuracy values were calculated as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (14)$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. TP, TN, FP, and FN designations are assigned based on the SIRE group that shows enrichment for PGx effect allele (or genotype). The presence of the PGx effect allele in the implicated SIRE group is counted as a true positive, whereas its presence in the other groups is counted as a false positive. Conversely, the presence of the PGx non-effect allele in the implicated SIRE group is counted as a false negative, whereas its presence in the other groups is counted as a true negative. Accuracy confidence intervals are calculated as:

$$CI = Accuracy \pm Z_{\alpha/2} \times \sqrt[2]{\frac{Error_{prediction}}{1 - Error_{prediction}}/N} \quad (15)$$

where error is calculated using:

$$Error_{prediction} = \frac{FP + FN}{TP + TN + FP + FN} \quad (16)$$

$$N = TP + TN + FP + FN \quad (17)$$

As noted before, when α is 0.05, $Z_{\alpha/2}$ is 1.96.

Pre- and post-test probabilities were compared in order to compute the amount of information gained per 100 individuals based on PGx stratification with SIRE. For any given PGx variant, the pre-test probability is calculated as the overall population prevalence of the PGx effect allele (additive mode) or genotype (dominant mode):

$$Prevalence_{overall} = Count_{EA}/Count_{Total} \quad (18)$$

where $Count_{EA}$ is the count of the effect allele/genotype in the cohort and $Count_{Total}$ is the total count of alleles/genotypes at that locus in the cohort. The post-test probability is calculated as the group-specific positive predictive values (PPVs) for the PGx effect allele or genotype. PPV is calculated using:

$$PPV_A = Count_{EA}^{A}/Count_{Total}^{A} \quad (19)$$

where $Count_{EA}^{A}$ is the count of the effect allele/genotype in population A and $Count_{Total}^{A}$ is the total count of alleles/genotypes at that locus in the population A. Information gain is then calculated as:

$$InfoGain_A = |PPV_A - Prevalence_{overall}|. \quad (20)$$

3.2.4 Comparison of SIRE and GA

To test whether PGx variant allele frequencies were correlated between SIRE and GA, pairwise PGx variant allele frequency differences calculated for SIRE groups were

regressed against allele frequency differences calculated for GA groups. Here, the null hypothesis is $H_0: \beta = 0$, while the alternate hypothesis is $H_A: \beta \neq 0$. The significance of this correlation was testing using a t-test where $t = (\beta_{obs} - \beta_{exp})/SE$ and $P = P(T_{DF} \leq \beta_{exp})$. Next, we tested whether GA groups partition PGx variation more than SIRE groups using the same regression. For this test, the null hypothesis is $H_0: \beta = 1$, while the alternate hypothesis is $H_A: \beta < 0$. An underlying assumption for this one-tailed test is that GA groups should hold more information about PGx allele frequency differences when compared to SIRE groups. We calculated the difference in the expected (unity line) and observed (SIRE versus GA) regression slopes, $d = (\beta_{exp} - \beta_{obs})/2$ to quantify the magnitude of the effect. A denominator of 2 was chosen to reflect the entire range of possible slopes that the data may take – going from -1, where SIRE groups reflect exactly the opposite difference in allele frequencies, to 1, where SIRE groups faithfully and completely capture the allele frequency differences observed in GA groups. The statistical significance was tested using a t-test as described above.

Results

3.2.5 Self-identified race/ethnicity (SIRE) and Genetic Ancestry (GA) in the US

We compared SIRE to GA for a cohort of 8,628 individuals characterized as part of the Health and Retirement Study (HRS), for whom both SIRE information and whole genome genotypes were available (Table 2). HRS participants self-identified according to racial and ethnic labels defined by the US Government Office of Management and Budget

(OMB). OMB defines five racial groups and two ethnic groups to assess disparities in health and environmental risks[136]. HRS participants were asked to select one or more race category and a single ethnic designation as Hispanic/Latino or not. We considered the race and ethnicity selections together and focused on the three largest categories in the HRS cohort: non-Hispanic White (5,927; 68.7%), non-Hispanic Black (1,527; 17.7%), and Hispanic/Latino of any race (1,174; 13.6%). We refer to these three groups here as White, Black, and Hispanic. The percentages of each SIRE group in the HRS cohort resemble the demographics of the US: White=72.4%, Black=12.6%, and Hispanic=16.3%[136].

Table 2. Demographic description for the cohort used in this study.

[1]*Number (Percentage)*

[2]*Median age in years (Confidence intervals)*

	All participants	White	Black	Hispanic
All[1]	8,628 (100.0)	5,927 (68.7)	1,527 (17.7)	1,174 (13.6)
Sex[1]				
Male	3,544 (41.1)	2,499 (42.2)	568 (37.2)	488 (41.6)
Female	5,084 (58.9)	3,428 (57.8)	959 (62.8)	697 (59.4)
Age[2]	57.5 (57.0, 58.0)	60.0 (60.0, 60.5)	54.5 (54.5, 55.0)	54 (53.5, 54.0)

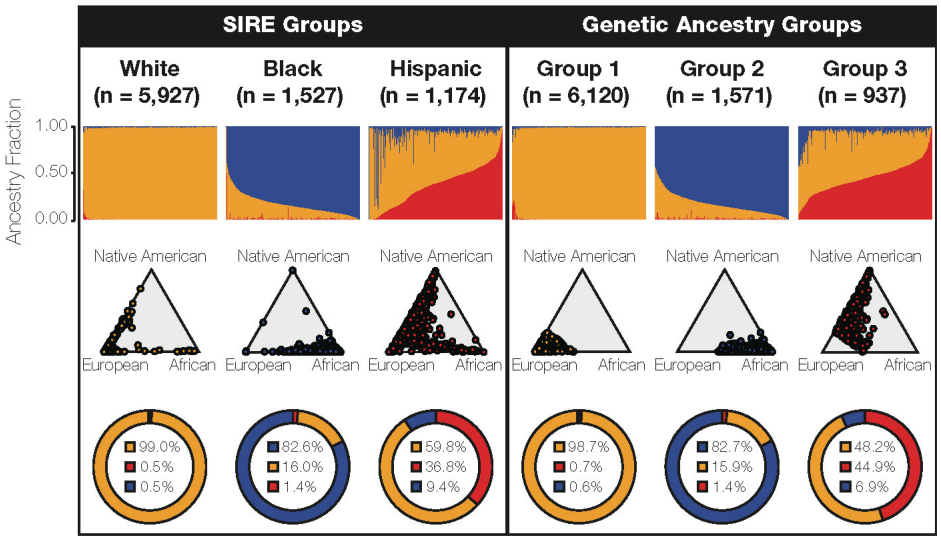


Figure 5. Race, ethnicity, and genetic ancestry in the US.

Continental genetic ancestry patterns are shown for self-identified race/ethnicity (SIRE) and genetic ancestry (GA) groups: European ancestry (orange), African ancestry (blue), and Native American ancestry (red). HRS cohort participants are grouped by SIRE and GA, as described in the text, and continental ancestry fractions are compared for each grouping system. Top row: continental ancestry fractions for individuals organized into the three SIRE and three GA groups. Each column represents an individual genome, and the three continental ancestry fractions are shown for each individual column. Middle row: ternary plots showing the continental ancestry fractions for the SIRE and GA groups, as illustrated by the relative proximity to each of the three ancestry poles. Bottom row: average continental ancestry percentages for the SIRE and GA groups.

Continental ancestry profiles were inferred for members of the HRS cohort by comparing their whole genome genotypes to whole genome sequence and genotype data for reference populations from Europe, Africa, and the Americas as described in the Materials and Methods. Each HRS participant was assigned European, African, and Native American ancestry proportions, and the resulting ancestry profiles were then clustered into three distinct (non-overlapping) GA groups using *k*-means clustering. GA groups were defined without reference to SIRE group labels, using unsupervised clustering on continental ancestry fractions alone, and the choice to cluster ancestry profiles into three

groups was made to allow for direct comparison with the three SIRE groups and in light of known patterns of continental ancestry in the US[137]. Permutation analysis was used to confirm the stability of the resulting GA groups and their robustness to changes in sample size (see Appendix B: Figure 27). The distributions of continental ancestry fractions were compared for the three SIRE groups – White, Black, and Hispanic – and the three GA groups (Figure 5).

The three objectively defined GA groups appear to correspond well to the SIRE groups, with respect to the distributions of individuals' continental ancestry fractions (Figure 5 – top row). GA groups 1, 2, and 3 correspond to the White, Black, and Hispanic SIRE groups, respectively. The distributions of continental ancestry fractions for the SIRE and their corresponding GA groups are compared in Appendix B: Figure 28. Despite the apparent similarity between SIRE and GA, ternary plots underscore the broader distribution of ancestry fractions within SIRE groups compared to the non-overlapping GA groups delineated by *k*-means clustering (Figure 5 – middle row). This is especially true for the Hispanic group, consistent with the fact that it may include individuals who identify as any race. Overall, SIRE and the GA groups show similar average continental ancestry percentages: White/Group 1 show ~99% European ancestry, Black/Group2 have ~82% African ancestry, and Hispanic/Group 3 show predominantly European ancestry (~60%) with the highest levels of Native American ancestry (~37%) and the greatest variance in continental ancestry for any of the three groups.

The correspondence between the SIRE and GA groups was quantified by characterizing the overlap of membership assignments across the two groupings (see

Appendix B: Figure 29). Overall, individuals' membership in the three SIRE and corresponding GA groups show 96.2% concordance. The highest concordance is seen for the White/Group 1 pair, followed by Black/Group 2, with Hispanic/Group 3 showing the lowest concordance. The levels of concordance vary according to which grouping system is taken as the reference for comparison. This distinction is most obvious for the Hispanic/Group 3 pairing: 96.6% of Group 3 members self-identify as Hispanic, while only 77.1% of self-identified Hispanics fall into Group 3.

3.2.6 Pharmacogenomic variation in the US

PGx variants that influence drug response were mined from the PharmGKB database, and levels of PGx variation were compared within and between the SIRE and GA groups defined for the HRS cohort. Results for SIRE group comparisons are shown in Figure 6, and results for the analogous comparison of GA groups are shown in Appendix B: Figure 30. PGx variants show higher allele frequencies, higher allele frequency differences between groups, and higher levels of heterozygosity compared to non-PGx variants genome-wide (Figure 6A-C). We considered group-specific differences in PGx variation in terms of the fixation index (F_{ST}), a commonly employed measure of population differentiation, and effect allele frequency differences. PGx F_{ST} and effect allele frequency difference values are highly correlated, as can be expected, and the largest differences are seen for the Black-White and Black-

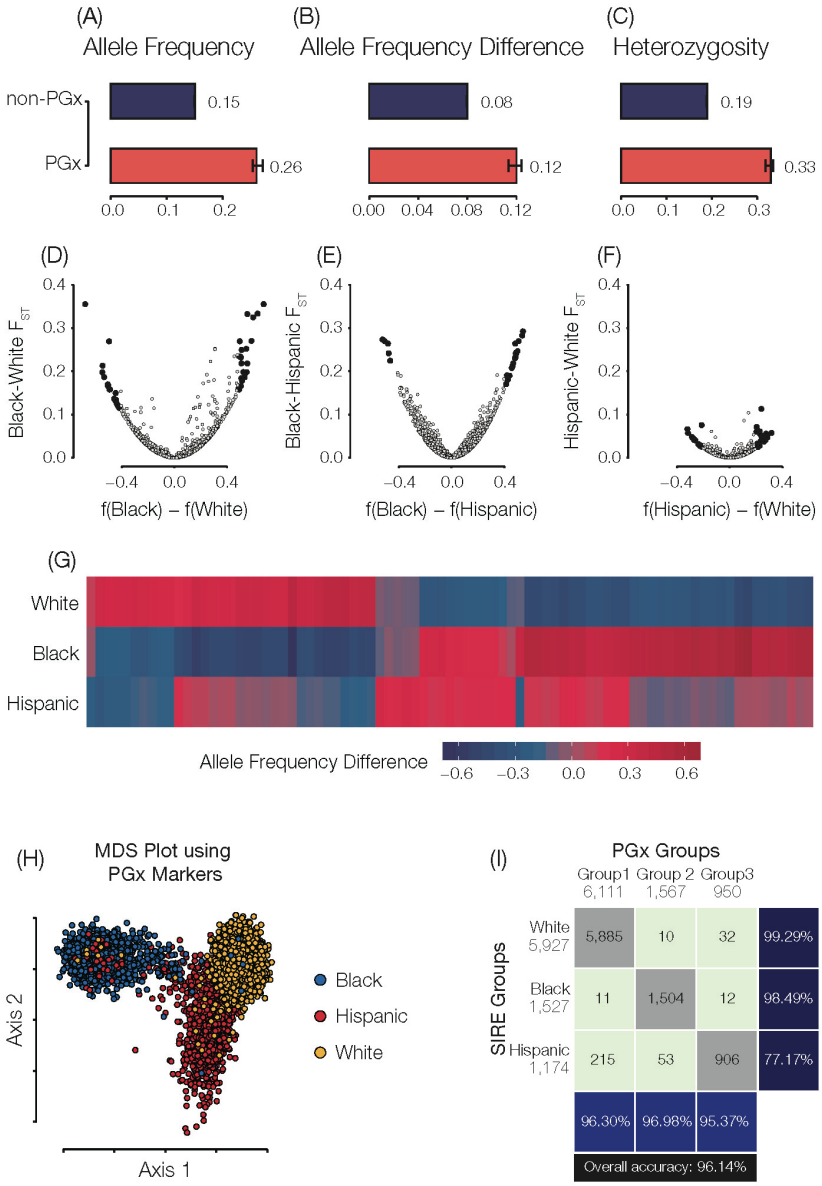


Figure 6. Pharmacogenomic variation in the US.

Genome-wide average allele frequencies (A), group-specific allele frequency differences (B), and heterozygosity fractions (C) are shown for PGx variants (red) compared to non-PGx variants (blue). (D-F) Fixation index (F_{ST}; y-axis) and allele frequency differences (x-axis) for pairs of SIRE groups. Statistically significant PGx allele frequency differences are highlighted in black. (G) Heatmap showing group-specific allele frequencies for significantly diverged PGx variants. (H) Multi-dimensional scaling (MDS) plot showing the relationship among individual genomes as measured by PGx variants alone. Each dot is an individual HRS participant genome, and genomes are color-coded by participants

SIRE. (I) The correspondence between SIRE groups and PGx groups defined by K-means clustering on the results of the MDS analysis. Data shown here correspond to SIRE groups; analogous results for GA groups are shown in Appendix B: Figure 30.

Hispanic group comparisons (Figure 6D-F). Notably, even the most extreme values of F_{ST} fall well below 0.5, indicating the most PGx variation is found within rather than between SIRE groups. Nevertheless, there are 82 PGx variants that show statistically significant (FDR q<0.05) values of allele frequency differentiation between any individual SIRE group and the other two groups, *i.e.* their complements (Figure 6G). The significantly diverged PGx variants show an average F_{ST} value of 0.15 compared to 0.05 for the remaining variants (see Appendix B: Figure 31). All-against-all pairwise distances for HRS participants were calculated using PGx variants and projected into two-dimensions with multi-dimensional scaling (MDS). *K*-means clustering was used to create three groups based on the PGx MDS distances, and individuals were labeled according to their SIRE (Figure 6H). Genome-wide patterns of PGx variation characterized in this way show 96.1% correspondence to SIRE group labels (Figure 6I).

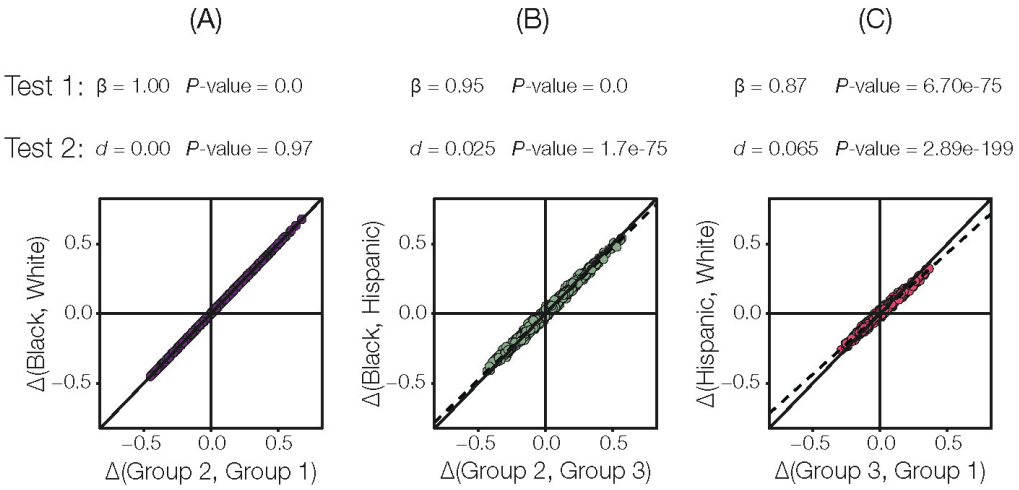


Figure 7. Self-identified race/ethnicity (SIRE) versus genetic ancestry (GA) for partitioning pharmacogenomic (PGx) variation.

(A-C) Regression of pairwise PGx variant effect allele frequency differences calculated using SIRE (y-axis) versus the corresponding GA groups (x-axis). Results of two statistical tests are shown for each of three pairwise group regressions. Test 1 evaluates whether SIRE and GA PGx allele frequencies are correlated, and test 2 evaluates that amount of additional resolution on PGx variant divergence that is provided by GA compared to SIRE. Details on each test are provided in the text.

3.2.7 SIRE versus GA for Partitioning Pharmacogenomic Variation

Given the overall correspondence, and group-specific differences, seen for SIRE and GA, we wanted to compare the utility of SIRE versus GA for partitioning pharmacogenomic variation in the US. Here, we asked two questions regarding PGx variation between groups: (1) are PGx allele frequencies correlated between SIRE and GA groups, and (2) do GA groups partition PGx variation more so than SIRE groups? The first question was addressed by regressing PGx frequency differences between grouping systems (SIRE vs. GA groups), and the second question was addressed by considering the

deviation of the regression from the unity line (*i.e.* the expected value under perfect correlation). As expected given the observed similarities between SIRE and GA groups, PGx allele frequency differences are highly correlated when corresponding group pairs are compared (Figure 7). The highest correlation is seen when the Black and White SIRE groups are compared to their corresponding GA groups. Comparisons that include the Hispanic SIRE group show lower levels of correlation.

With respect to the second question regarding the partitioning of PGx variation, allele frequency differences between the Black/White SIRE groups and their corresponding GA groups fall almost entirely along the unity line; in this case, genetic ancestry does not provide any additional information regarding PGx variation (Figure 7A). For both comparisons that include the Hispanic group however, the slope of the regression is less than one, indicating greater PGx allele frequency differences between GA groups compared to their corresponding SIRE groups (Figure 7B and 7C). Thus, GA does provide more information than SIRE when ethnicity is considered, but the effect size of this difference is small (d=2.5% for Black/Group 2 vs. Hispanic/Group 3 and d=6.5% for Hispanic/Group 3 vs. White/Group 1).

Thus far, we have shown that SIRE and GA groups are highly concordant for the HRS cohort and that PGx allele frequency differences are similar for both classification systems. Since SIRE labels are routinely collected as patient provided information, and are also readily available as part of electronic health records, we focused on PGx variation between SIRE groups to explore the potential clinical utility of race and ethnicity. We wanted to know whether PGx effect allele frequency differences of the magnitude observed

here have any utility for guiding medication prescription decisions in light of the fact that the majority of PGx variation is found within rather than between SIRE groups. We considered the odds ratios for the apportionment of PGx risk alleles among individual SIRE groups and their complements as an indicator of SIRE groups' predictive utility, given that odds ratios are widely used to associate categorical risk factors with health outcomes [134]. We also computed absolute risk increase values to account for the population frequency of PGx risk alleles when considering the magnitude of between group differences as well as the accuracy with which SIRE group membership predicts PGx alleles or genotypes.

Table 3. Examples of highly differentiated PGx variants.

This table lists some examples of highly diverged PGx variants in the three SIRE groups under consideration. In the table, 'Ref. Pop.' refers to Reference Population, OR refers to Odds Ratios, ARI refers to the Absolute Risk Increase percentage. Values in brackets specify the 95% confidence intervals for each computation.

			Effect allele frequency						
rsID	**Drug**	**Effect**	**White**	**Black**	**Hispanic**	**Ref. Pop.**	**OR**	**ARI**	**Accuracy**
rs1045642	Fentanyl	Dosage	0.78	0.37	0.70	White	3.26 (2.96, 3.60)	26.1 (24, 28)	68.5 (67.0, 69.9)
rs9934438	Warfarin	Dosage	0.38	0.83	0.33	Black	8.27 (7.18, 9.54)	45.93 (44, 48)	66.53 (65.03, 68.03)
rs2884737	Warfarin	Dosage	0.27	0.04	0.18	Black	8.99 (7.43, 10.87)	36.0 (34, 38)	52.5 (50.5, 54.5)
rs2500535	Nortriptyline	Efficacy	0.05	0.06	0.26	Hispanic	6.1 (5.40, 6.82)	20.3 (18, 22)	85.2 (84.6, 85.9)
rs11615	Platinum compounds	Efficacy	0.37	0.88	0.64	Black	9.90 (8.85, 11.09)	45.95 (45, 47)	63.5 (62.4, 64.6)
rs20455	Atorvastatin	Efficacy	0.36	0.79	0.40	Black	14.2 (11.11, 18.17)	35.71 (34, 37)	50.01 (47.9, 52.1)
rs1048943	Capecitabine, Docetaxel	Efficacy	0.04	0.02	0.27	Hispanic	12.74 (11.14, 14.79)	39.4 (37, 42)	87.3 (86.5, 88.1)
rs4646450	Tacrolimus	Metabolism	0.16	0.84	0.33	Black	66.80 (49.17, 90.88)	63.15 (62, 65)	71.5 (70.2, 7.2)
rs6977820	Antipsychotics	Toxicity	0.04	0.28	0.05	Black	14.8 (12.13, 18.14)	45.96 (44, 48)	60.09 (58.4, 6.1)
rs1801394	Methotrexate	Toxicity	0.46	0.72	0.67	White	2.82 (2.63, 3.02)	24.68 (23, 26)	59.40 (58.2, 60.1)
rs16969968	Nicotine	Toxicity	0.66	0.95	0.80	Black	8.17 (6.97, 9.59)	26.6 (26, 28)	43.17 (41.4, 44.9)

Examples of highly differentiated PGx variants are shown in Table 3 and Figure 8. These examples were chosen as variants that had relatively high odds ratio values across different PGx effect types (dosage, efficacy, metabolism, and toxicity), highlighting instances for each of the three SIRE groups. The relative percentages of PGx effect (above) and non-effect (below) alleles across SIRE groups reveal the extent of differentiation for these variants (Figure 8A), and the observed allele frequency differences are associated with SIRE group-specific continental ancestry fractions (Figure 8B-D). Nevertheless, as described above and shown in Figure 6, even highly differentiated PGx variants show levels of F_{ST} that indicate substantially more within than between group variation (see pie charts in Figure 6B-D). Despite the relatively high levels of within group PGx variation, these variants show high group-specific odds ratios and substantial absolute risk increase values. In other words, HRS cohort members' racial and ethnic self-identities carry substantial information that can be used to stratify pharmacogenomic risk at the population level. However, the accuracy levels with which group affiliations predict specific risk alleles or genotypes are only marginally high, indicating that SIRE has relatively less utility for individual-level risk prediction compared to risk stratification.

For example, the A allele of the PGx variant (rs1045642) in the ATP Binding Cassette Subfamily B Member 1 (*ABCB1*) gene is associated with a decreased fentanyl opioid dose requirement [138] (Figure 8B). This PGx variant has a dominant mode of effect, such that patients with either the AA or GA genotype tend to metabolize fentanyl slower than patients with the GG genotype and will therefore require a lower dosage. 96.0% of variation for this PGx variant is partitioned within SIRE groups compared to 4.0% variation between groups. However, the dosage-associated genotypes are far more common in

individuals who identify as White (OR=3.3, CI=3.0-3.6; ARI=26.1%, CI=24.0%-28.3%), and from the ancestry association plot, it can be seen that the effect allele (A) is highly correlated with European genetic ancestry (β=0.20, P=1.95e-35). Self-identification as White predicts dosage-associated genotypes with 68.5% accuracy.

Similarly, a PGx variant (rs2500535) in the Uronyl 2-Sulphotransferase (*UST*) gene has been found to be associated with the efficacy of nortriptyline – an antidepressant – in patients with major depressive disorder [139] (Figure 8C). This PGx variant has a dominant mode of effect; patients with the A allele are associated with a decreased improvement of depression symptoms when prescribed nortriptyline. These lower efficacy genotypes are more common in individuals who identify as Hispanic. Even though the variation at this genomic site is far higher within (93.5%) compared to between (6.5%) groups, the odds ratio for having risk-associated genotypes is high for the Hispanic population (OR=6.07, CI=5.44-6.82) along with a high absolute risk increase (ARI=20.3%, CI=18.5%-22.2%). Hispanic ethnicity predicts nortriptyline efficacy-associated genotypes with 85.2% accuracy.

Another PGx variant (rs6977820) found in the Dipeptidyl Peptidase Like 6 (*DPP*) gene has been associated with adverse response to antipsychotic drugs (Figure 8D). This PGx variant has an additive effect mode, whereby the T allele is positively correlated with African ancestry and associated with tardive dyskinesia among Schizophrenia patients treated with antipsychotics [140]. When individuals that self-identify as Black are compared to the other two SIRE groups, most variation at this variant is found within (85.9%) rather than between (14.1%) groups. However, the odds ratio for the presence of the risk allele

for adverse reaction to antipsychotics is high (OR=7.7, 95% CI=7.1-8.49), as is the absolute risk increase (ARI=47.2%, 95% CI=45.4%-48.9%), consistent with a substantially elevated risk of adverse drug reaction for the Black SIRE group compared to the others. Individuals who self-identify as Black can be predicted to have the effect-associated allele with 73.0% accuracy.

3.2.8 Clinical Value of Pharmacogenomic Stratification by SIRE

We quantified the clinical utility of SIRE for partitioning PGx variation by comparing the ability to predict PGx effect alleles/genotypes before (pre) and after (post) stratification of the population by SIRE. The approach we used is equivalent to the comparison of pre- and post-test probabilities for diagnostic tests, where the test in this case is patient stratification by SIRE. For any given PGx variant, the pre-test probability is the overall population prevalence of the PGx effect allele/genotype, and the post-test probabilities are the group-specific positive predictive values (PPVs) for the PGx effect allele or genotype. Allele counts were used to compute these probabilities for PGx variants that show an additive effect mode, and genotype counts were used for the dominant effect mode. The absolute difference of the pre- and post-test probabilities calculated in this way was taken as a measure of the amount of information that is gained, with respect to PGx variant prediction for each specific group, when SIRE is used for patient stratification.

When highly differentiated PGx variants (Figure 6G and Figure 8) are analyzed in this way, the SIRE groups that show the highest effect allele frequencies for any given variant provide substantial additional information for PGx prediction. Considering the

PGx variant (rs2500535) that is associated with Nortriptyline efficacy (Figure 8C), stratification by Hispanic identity yields an additional 14 individuals, for every 100 patients to be treated, who are predicted to show decreased improvement of symptoms related to depressive disorder. The information gain is even more extreme for the PGx variant (rs6977820) that is associated with antipsychotic toxicity (Figure 8D). For this variant, stratification of individuals that self-identify as Black will yield an additional 39 out of every 100 patients that are counter-indicated for the antipsychotic medications owing to toxic side effects. The overall levels of information gained via stratification by SIRE differ widely by group. Individuals that self-identify as Black show the highest levels of information gain for PGx variant prediction followed the Hispanic and White groups, respectively (Figure 9). This pattern can be attributed to the relative numbers of individuals in each SIRE group together with the extent of genetic diversification seen between groups. The relatively high frequency of PGx effect alleles (Figure 6A) also contributes to the amount information gain observed here, given the fact that PPVs depend on the prevalence of the condition that is being tested (i.e. the presence of PGx effect alleles/genotypes).

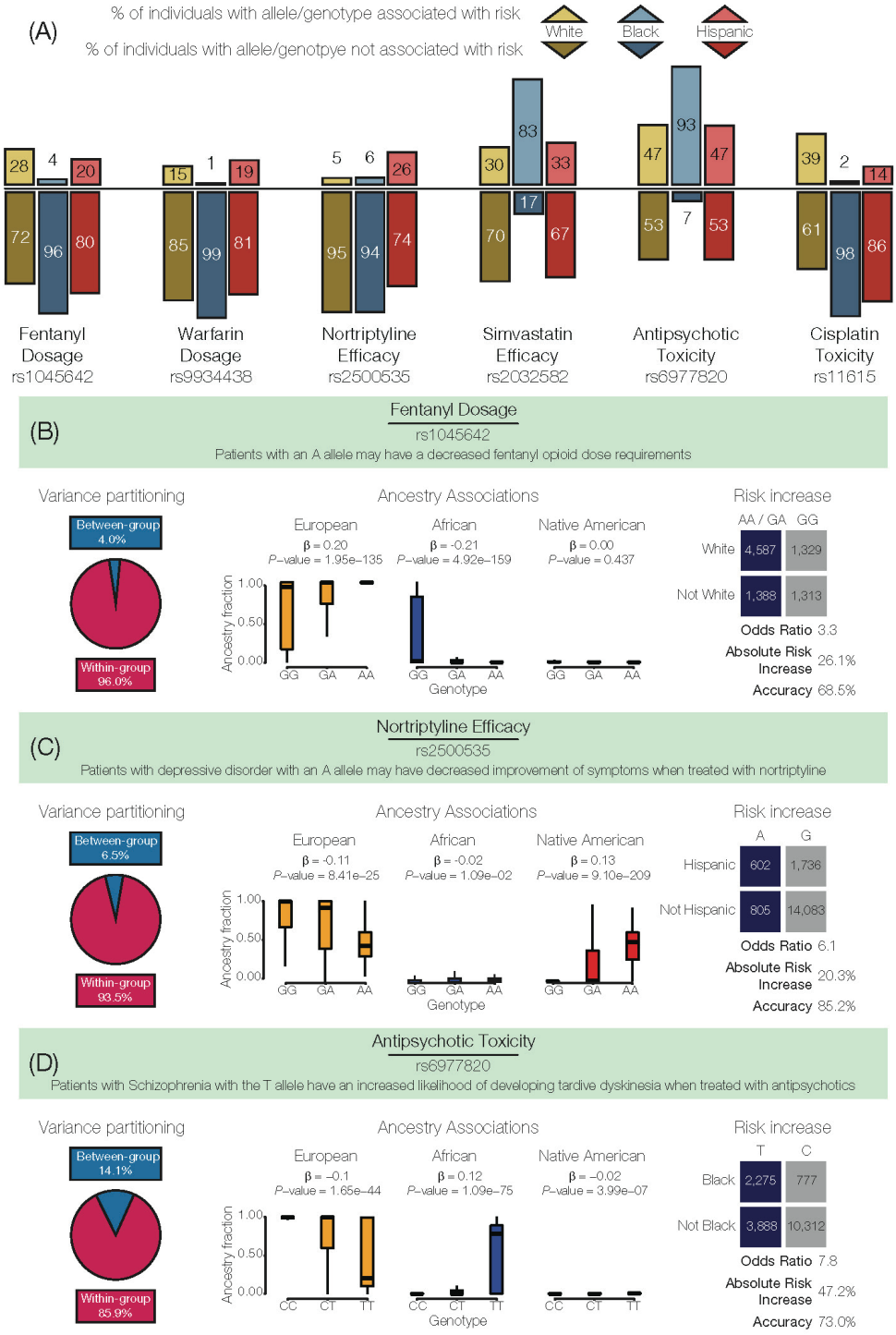


Figure 8. Examples of highly differentiated pharmacogenomic (PGx) variants.

(A) SIRE group percentages of effect (above axis) versus non-effect (below axis) alleles/genotypes are shown for six highly differentiated PGx variants. Allele counts are used for the additive PGx effect mode, and genotype counts are used for the dominant effect mode. (B-C) The extent of within versus between group variation, ancestry associations, and PGx stratification/risk by SIRE groups are shown for three examples. Ancestry associations relate the ancestry fractions for individuals that bear distinct PGx genotypes: European (orange), African (blue), and Native American (red). Effect (blue) versus non-effect (gray) allele/genotype counts are compared for the group enriched for a specific PGx variant compared to the other two groups. Allele counts are shown for the additive PGx effect mode, and genotype counts are shown for the dominant mode. Group-specific allele/genotype counts were used to compute odds ratios and absolute risk increase values (risk stratification) along with group-specific prediction accuracy values (risk prediction) as shown.

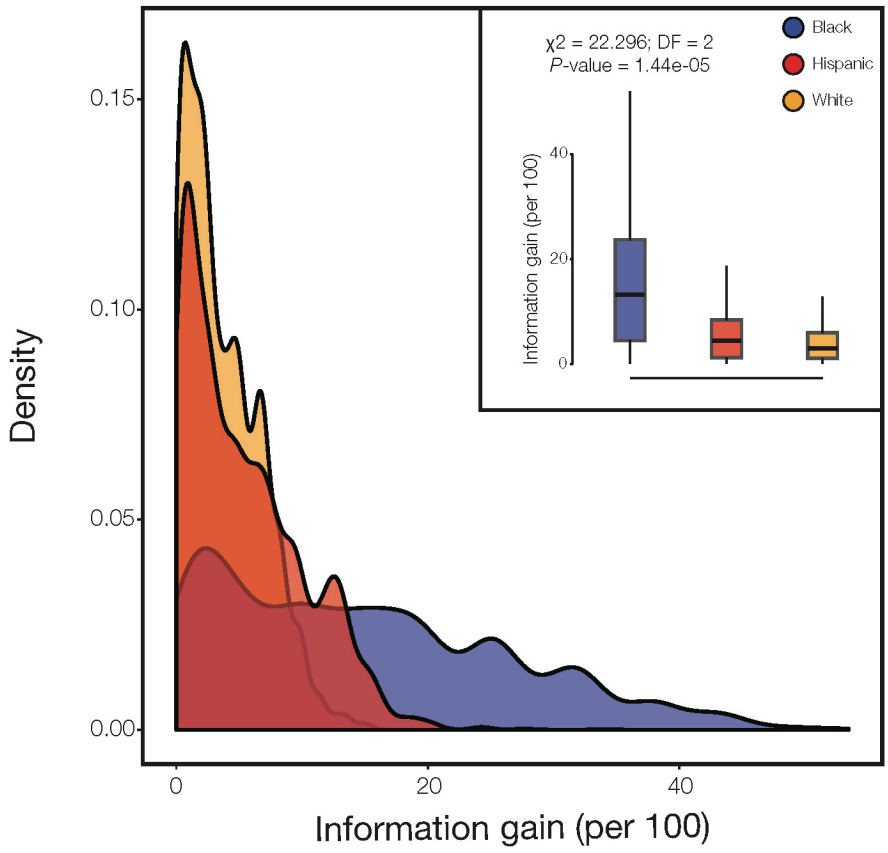


Figure 9. Information gained when SIRE is used for PGx stratification.

The amount of information gained per 100 individuals is the number additional correct PGx variant predictions made when SIRE is used to stratify the population. Information gain is calculated for all PGx variants in each SIRE group, as described in the text, and the group-specific distributions are shown as density distributions and box-plots (inset): White (orange), Black (blue), and Hispanic (red).

3.3 Discussion

3.3.1 Concordance Between SIRE and GA in the US

The SIRE and GA groups from the US analyzed here show >96% overall concordance (Figure 5, also see Appendix B: Figures 28 and 29). It must be stressed that these results only apply to the three major racial/ethnic groups covered by the ~8,600 individual HRS cohort; nevertheless, the concordance between SIRE and GA seen for the HRS cohort is very much consistent with a number of previous studies of the US population. In 2005, investigators showed a 99.9% concordance between SIRE and genetically derived clusters for 3,636 individuals from four racial/ethnic groups[141], and a 2007 study reported 100% classification accuracy of individuals from geographically separated population groups when thousands of genetic variants were used for clustering[34]. More recently, a study of >11,000 cancer patients from The Cancer Genome Atlas found an 95.6% concordance between self-reported race (not ethnicity) and GA[33], and a study of >200,000 individuals from the Million Veterans Program found >99.4% concordance between SIRE and GA[32]. The latter two studies relied on machine learning classifiers powered by vectors of 7 and 30 ancestry principal components, respectively, whereas our clustering algorithm uses vectors of only three continental ancestry components to classify individual genomes. Additionally, the distribution of GA fractions observed here for the HRS cohort SIRE groups is consistent with previous studies[51,126,137,142,143]. Taken together, our results and others underscore the extent to which continental ancestry patterns can distinguish SIRE groups in the US.

Genetic differences accumulate among populations when they are reproductively isolated, and isolation by distance[144] best accounts for the apportionment of human genetic diversity among global populations[145]. Populations that are physically distant, or separated by major geographic barriers, are more genetically diverged than nearby populations[146]. It follows that the appearance of population structure, *i.e*. distinct clusters of genetically related individuals, can represent an artifact of uneven sampling of human populations at extremes of distance[147]. For instance, isolation by distance can explain much of the apparent genetic structure observed for major genome sequencing projects such as the 1000 Genomes Project[65,148] and the Human Genome Diversity Project[127,149]. Conversely, when human populations are sampled more evenly across a range of distances, and in the absence of major geographical barriers, genetic diversity appears to be continuously distributed as a cline of variation[150,151].

Isolation by distance can be taken to explain the concordance of the SIRE and GA groups observed for the HRS cohort, since the three major US SIRE groups are made up of individuals with ancestry from continental population groups – European, African, and Native American – that were isolated at great distances for tens-of-thousands of years before coming back together over the last 500 years[126,137]. Since each SIRE group contains distinct patterns of continental ancestry, they correspond well to objectively defined clusters formed based on the partitioning of GA (Figure 5, also see Appendix B: Figure 28 and 29). In addition, despite the fact that these population groups are currently co-located within the US, assortative mating based on culture stands as an ongoing reproductive barrier among groups[152,153] (but see below for an important caveat regarding this fact). It is nevertheless important to note that most of the SIRE and GA groups analyzed here are

not composed of individuals with highly coherent ancestry patterns. Only the White/Cluster 1 groups show coherent ancestry patterns, whereas the Black/Cluster 2 and Hispanic/Cluster 3 groups are made of up of individuals that vary along a range of continental ancestry fractions (Figure 5 and see Appendix B: Figure 28). This is especially true for the Hispanic group, consistent with the fact Hispanic is an intentionally broad label that covers individuals from different races and with very distinct ancestry patterns[94].

An important caveat with respect to the high concordance between SIRE and GA observed here relates to the age of the individuals in the HRS cohort (Table 2). We chose to focus on older Americans given their disproportionate use of prescription medications[123], and HRS recruited participants aged 50 and over starting in 1992. The average age of the HRS cohort analyzed here is 57.5 years (CI: 57.0-58.0), and all of the study participants were born before 1965, when there were still "anti-miscegenation" laws in nineteen states[154]. Rates of intermarriage among SIRE groups have increased substantially since that era[155], and as admixture continues to increase over time, the ancestral coherence of SIRE groups is expected to fall precipitously. Increased rates of immigration, coupled with the arrival of more globally diverse immigrant groups, will also blur boundaries between SIRE groups, potentially rendering the current labels clinically uninformative. Indeed, the most widely used SIRE labels in the US are mandated by the OMB, and they will likely be revised in the near future to better capture the increasing diversity of the US population. As such, the clinical relevance of SIRE will almost certainly decrease over time.

3.3.2 Within versus between group genetic divergence

It has long been appreciated that the vast majority of human genetic variation is found within rather than between populations. This fundamental result was first reported for worldwide racial groups, based on analysis of a handful of (surrogate) genetic markers[156], and has since been confirmed by numerous studies of populations defined by GA using larger-scale analyses[149,157-161]. The distinction between this fundamental result and the high concordance seen for SIRE and GA, as well as the ability to cluster human population groups at various levels of relatedness, can be explained by the difference between univariate methods for variance partitioning versus multivariate classification methods[162,163]. The analysis of PGx variation reported here is univariate, since we focus on the apportionment of variation for individual PGx variants, and we confirm that the majority of PGx variation is found within the HRS cohort groups (Figure 6 and 8).

We used a standard evidence based medicine analytical framework[134,135] in an effort to understand the clinical relevance of PGx variation that is partitioned among SIRE groups in this way. In particular, we asked how the observed PGx differences between groups could be clinically relevant when the majority of variation falls within population groups, even for the most divergent variants found here. Despite the observed pattern of within versus between group PGx variation, we found numerous cases of high odds ratios and high absolute risk increases for the group-specific prevalence of PGx variants (Table 3 and Figure 8). In other words, membership in any given SIRE group can entail substantially greater odds, and far higher risk, of carrying clinically relevant PGx variants compared to members of other groups. Information of this kind should be an important consideration for clinicians charged with making treatment decisions and could also be of value for well-informed patients.

Finally, it should be emphasized that humans are far more similar than they are different at the genomic level, both within and between population groups. As of August 2019, there were 674 million annotated single nucleotide variants among the ~3 billion sites in the human genome[164]. Thus, more than 75% of genomic positions are conserved among all human population groups, and for those positions that do vary, the majority are rare variants that segregate at <1% frequency worldwide[65]. Nevertheless, the results reported here underscore the potential clinical relevance for those genetic variants that do show relatively high levels of between-group divergence.

3.3.3 Caveats and limitations

It is important to note that in this study we measure the frequency of PGx variants across different SIRE and GA groups, rather than drug response differences *per se*. Even though the penetrance of PGx variants is generally high[20], clinical interpretations of variant frequency differences should be considered in light of variable penetrance levels as well. In cases of low penetrance, the magnitude of drug response differences between groups will be dampened. Furthermore, if PGx variants have different magnitudes of effect for different groups, *i.e.* group-specific effect sizes, then differences in drug response cannot be directly inferred from PGx variant frequency differences alone. However, since the majority of PGx variants are causative protein coding variants[20], the likelihood of group-specific effect sizes is far lower than would be expected for non-coding variants discovered by genome-wide association studies, which are typically tag markers that are linked to nearby causative variants. Finally, the focus on single nucleotide variants (SNVs) is

another limitation of the study, given the fact structural variants and multi-variant haplotypes have also been associated with inter-individual drug response differences. Nevertheless, the vast majority of PGx variants annotated in the PharmGKB database are SNVs[20], suggesting that our analytical approach captures most of the known variant-drug associations.

3.4 Conclusions

As previously noted, demographic trends in the US suggest that the clinical relevance of SIRE, including its predictive utility for PGx variation, is expected to continuously decrease over time. The increasing adoption of routine genetic testing for precision medicine could also render SIRE obsolete for stratifying PGx variation[165]. This is because genotyping of specific PGx variants will obviously provide far more accurate risk prediction than SIRE. For example, even a highly divergent PGx variant, like the antipsychotic toxicity associated variant rs6977820 (Figure 8D), will yield a mis-prediction of the PGx risk allele 27% of the time if SIRE alone were used as a predictor. In this sense, the high group-specific PGx odds ratios and absolute risk increases observed in this study are best considered as surrogate guides to inform the optimal choice of prescribed medication, rather than precise diagnostic tools. In other words, SIRE categories provide valuable information for stratifying PGx risk at the population level but not for predicting individual-level PGx variants. Having said that, and despite the promise of population scale genomic screening initiatives and biobanks[18], such as the NIH All of Us project[166], the day when all Americans will have ready access to their genetic profiles remains far in

the future. Unfortunately, this is likely to be even more so for minority communities that are vastly underrepresented among clinical genetic cohorts[43,88]. Until that time, SIRE will remain an important feature for clinicians to consider when making treatment decisions.

Perhaps most importantly, the current utility of SIRE is most apparent for groups who are underrepresented in biomedical research. Individuals who self-identify as Black or Hispanic stand to gain far more information with respect to precision treatment decisions than those who identify as White (Figure 9). This finding can be attributed to the relative frequencies of individuals in each of the three SIRE groups analyzed here, which closely mirror the current demography of the US, and the extent of genetic divergence among groups. If a 'one size fits all' approach to drug prescription is used, patients who identify as White are more likely to receive the most appropriate treatment, since their PGx variant frequencies will be closest to the overall population mean. Conversely, individuals who identify as Black or Hispanic have the most to lose if SIRE is not considered when making treatment decisions.

CHAPTER 4. THE LANDSCAPE OF HEALTH DISPARITIES IN THE UK BIOBANK

4.1 Abstract

Publicly accessible resources that catalog health disparities in the UK Biobank do not exist. Such resources can enable researchers to explore the landscape of health disparities and direct their attention to areas of research which might have the most impact. Here, we developed the UK Biobank Health Disparities Browser for groups defined by age, country of residence, ethnic group, sex, and socioeconomic deprivation. We defined disease cohorts by mapping ICD-10 diagnosis codes from the UK Biobank to phenotype codes (phecodes). Phecodes aggregate one or more related ICD-10 codes into distinct diseases, and they use both inclusion and exclusion criteria to define disease case and control cohorts. For each of the population attributes used to define groups, disease percent prevalence values were computed for all groups, and the magnitude of the disparities were calculated by both the difference and ratio of the range of disease prevalence values among groups in an attempt to identify high and low prevalence disparities, respectively.

We identify several disease phenotypes with disparate prevalence values across population attributes and have deployed an interactive web browser accessible at https://ukbatlas.health-disparities.org. The interactive browser includes prevalence data for 1,513 diseases based on a cohort of >500,000 participants from the UK Biobank. Researchers can browse and sort by disease prevalence and differences to visualize health

disparities for each of these five population attributes, and users can search for diseases of interest by disease names or codes.

4.2 Introduction

Health disparities can be defined as differences in health outcomes between groups of people, where the groups can be delineated in a variety of ways. These differences in outcomes are often multifactorial and can be attributed to a combination of biological, social, and environmental factors[14]. Easy availability of information on health disparities can allow researchers and policy makers in identifying areas of research and/or interventions where possible.

Biobanks, being repositories of large amounts of demographic and clinical data[16,17], are ideally suited for characterizing health disparities. The UK Biobank is arguably the largest and most mature biobank that is available to researchers. Accordingly, the UK Biobank[167,168] offers an unprecedented opportunity to characterize the landscape of health disparities in the United Kingdom. It should be noted that participants of the UK Biobank are generally healthier and wealthier than the general population[169], and this "healthy volunteer" bias might dampen the extent of some of the disparities identified here. Additionally, given the diverse, cosmopolitan nature of the population of the UK[170], characterizing disparities using the UK Biobank can support of health equity for underserved minority populations.

We developed the UK Biobank Health Disparities Browser as a means for researchers to explore the landscape of health disparities in the United Kingdom, for groups defined by age, country of residence, ethnicity, sex, and socioeconomic deprivation. The browser includes prevalence data for 1,513 diseases based on a cohort of >500,000 participants from the UK Biobank. Users can browse and sort by disease prevalence and differences to visualize health disparities for each of these four groups, and users can search for diseases of interest by disease names or codes.

4.3 Methods

4.3.1 Study cohort

We used participant data from the UK Biobank, a prospective cohort study set up to investigate the lifestyle, environmental, and genetic determinants of a wide variety of diseases of adulthood. The study recruited over 500,000 participants aged between 40 and 70 years between 2006 and 2010[167]. Participant data includes completed questionnaires, nurse-led interviews, medical assessments, and biological samples.

4.3.2 Population attributes and comparison groups

We used the following fields from UK Biobank data: (1) age (Field 21003: Age when attended assessment center)[171], (2) assessment center (Field 54: UK Biobank assessment center)[172], (3) ethnic group and background (Field 21000: Ethnic

background)[173], (4) ICD-10 codes (Field 41270: Diagnoses – ICD10)[174], (5) sex (Field 31: Sex)[175], and (6) Townsend deprivation index (Field 189: Townsend deprivation index at recruitment)[176].

Investigators from the UK Biobank invited participants who lived within 25 miles of one of the 22 recruitment centers located across England. Scotland, and Wales. Accordingly, we used the location of a participant's assessment center to determine their country of residence.

We used the Townsend index of deprivation as a measure of socioeconomic deprivation. The Townsend index is a composite metric the incorporates (1) unemployment, (2) non-car ownership, (3) non-home ownership, and (4) household overcrowding in a given area[177]. A higher value of the Townsend index indicated higher material deprivation while lower values indicate relative affluence.

Comparison groups were defined for each of the five population attributes studied here: age, country of residence, ethnic group, sex, and socioeconomic deprivation. Participants were partitioned into four groups based on their age at recruitment (35-44, 45-54, 55-64, and 65-74 years old). Three groups were created for country of residence (England, Scotland, and Wales; The UK Biobank did not have recruitment centers in Northern Ireland). The initial questionnaire for recruitment asked participants to identify as one of six ethic groups (Asian, Black, Chinese, Mixed, White, or Other), and a distinct ethnic background within each group. Participants' self-identified ethnic groups were used to create groups for comparison. For socioeconomic deprivation, the participants were

divided into five equal groups using the Townsend index of deprivation (as quintiles). For sex, males and females were compared.

4.3.3 Phenotype case/control cohorts

We used the UK Biobank participants' ICD-10 diagnosis codes to define case/control cohorts using the phecode scheme defined be the PheWAS consortium[178,179]. The phecode scheme provides phenotype-specific inclusion and exclusion criteria ICD-10 codes for generating case/control cohorts from electronic health records. These phecodes are manually curated and validated by physicians and experts. This approach allows investigators to define clearly distinct case and control cohorts that can be compared confidently. Phecode case/control cohorts were curated for a total of 1,513 diseases or health-related conditions.

4.3.4 Disease prevalence and quantifying disparities

The prevalence for each of the 1,513 disease was calculated for the overall cohort and each individual group defined by the population attributes under consideration. The prevalence was calculated as:

$$Prevalence = \frac{N_{cases}}{N_{cases} + N_{controls}} \tag{21}$$

where N_{cases} refers to number of cases and $N_{controls}$ refers to number of controls.

For each population attribute under consideration, we calculated the range of prevalence values for each of the constituent groups as:

$$Range\ difference = Max(Prev_{Disease}) - Min(Prev_{Disease}) \tag{22}$$

where $Prev_{Disease} = \left[Prev_{Disease}^{Group1},\ Prev_{Disease}^{Group2},\ Prev_{Disease}^{Group3}, \ldots\right]$ along with calculating the ratio of the range of prevalence values as:

$$Range\ ratio = log_2\left(\frac{Max(Prev_{Disease})}{Min(Prev_{Disease})}\right) \tag{23}$$

Taken together, these two metrics enable the identification of health disparities for high prevalence diseases (using the $Range\ difference$) and for those diseases with low overall prevalence values (using $Range\ ratio$). On plotting these two metrics orthogonally, we computed a unified disparity score defined as the Euclidean distance from the origin as:

$$Disparity\ score = \sqrt{(Range\ difference)^2 + (Range\ ratio)^2} \tag{24}$$

Within a population attribute, a relative disease burden was calculated for each group as:

$$RDB_{Group} = 1 - \frac{NMax_{Group}}{NullAvg} \tag{25}$$

where, RDB_{Group} refers to group-specific relative disease burden, $NMax_{Group}$ refers to the number of phenotypes where $Group$ has the highest prevalence, and $NullAvg$ refers to the null expectation calculated as $\frac{1{,}513}{N_{Groups}}$ (N_{Groups} is the number of groups for that population attribute). An RDB_{Group} value of 0 would mean that the $Group$ in question has the highest prevalence for exactly $NullAvg$ diseases. A high positive value would represent a disproportionately high burden of disease for the sub-population $Group$, while a negative value would indicate a disproportionately low burden of disease.

4.3.5 Interactive web server

Data processing and analysis were done using the Pandas library in Python[180]. Plots were made using the ggplot2 library[181] in the R statistical language v3.6.1[68]. The interactive webserver was developed using the Plotly Dash framework[182].

4.4 Results

4.4.1 Health disparities across population attributes

Overall, we had information on the following population attributes for 500,428 participants from the UK Biobank: age, country of residence, ethnic group, sex, and socioeconomic deprivation (SED) (Table 4). Most of our analysis cohort falls primarily between the ages of 55 and 64 (42.3%), resides in England (88.7%), identify as belonging

to the White ethnic group (94.3%), and is Female (54.4%) (Table 4). Leveraging the phecode schema[179], which specifies ICD-10 diagnosis codes inclusion and exclusion criteria for phenotypes, we generated 1,513 case/control cohorts. For each of these case/control cohorts, we calculated the prevalence of disease in groups defined by the five population attributes under consideration. Next, health disparities were quantified as the difference and ratio of the range of disease prevalence among groups defined by population attributes under consideration (Figure 10; Appendix C: Figures 32-35). The two metrics employed – range difference and range ratio – were combined into a single, comparable metric by computing the Euclidean distance from the origin in a space parametrized by these two parameters. On comparing different population attributes, we find that ethnic groups show that most disparity (median disparity score: 2.00), followed by age (median disparity score: 1.32), country of residence (median disparity score: 0.70), SED (median disparity score: 0.68), and sex (median disparity score: 0.58) (Figure 11).

Table 4. Cohort table.

Characteristic	**Number (%)**
Complete cohort	500,428
Age	
35-44	51,473 (10.3)
45-54	141,781 (28.3)
55-64	211,518 (42.3)
65-74	95,656 (19.1)
Country of residence	
England	444,061 (88.7)
Scotland	35,655 (7.1)
Wales	20,712 (4.1)
Ethnic group	
Asian	9,823 (2.0)
Black	8,019 (1.6)
Chinese	1,569 (0.3)
Mixed	2,908 (0.6)

	Other	6,422 (1.3)
	White	471,687 (94.3)
Sex		
	Female	272,368 (54.4)
	Male	228,060 (45.5)

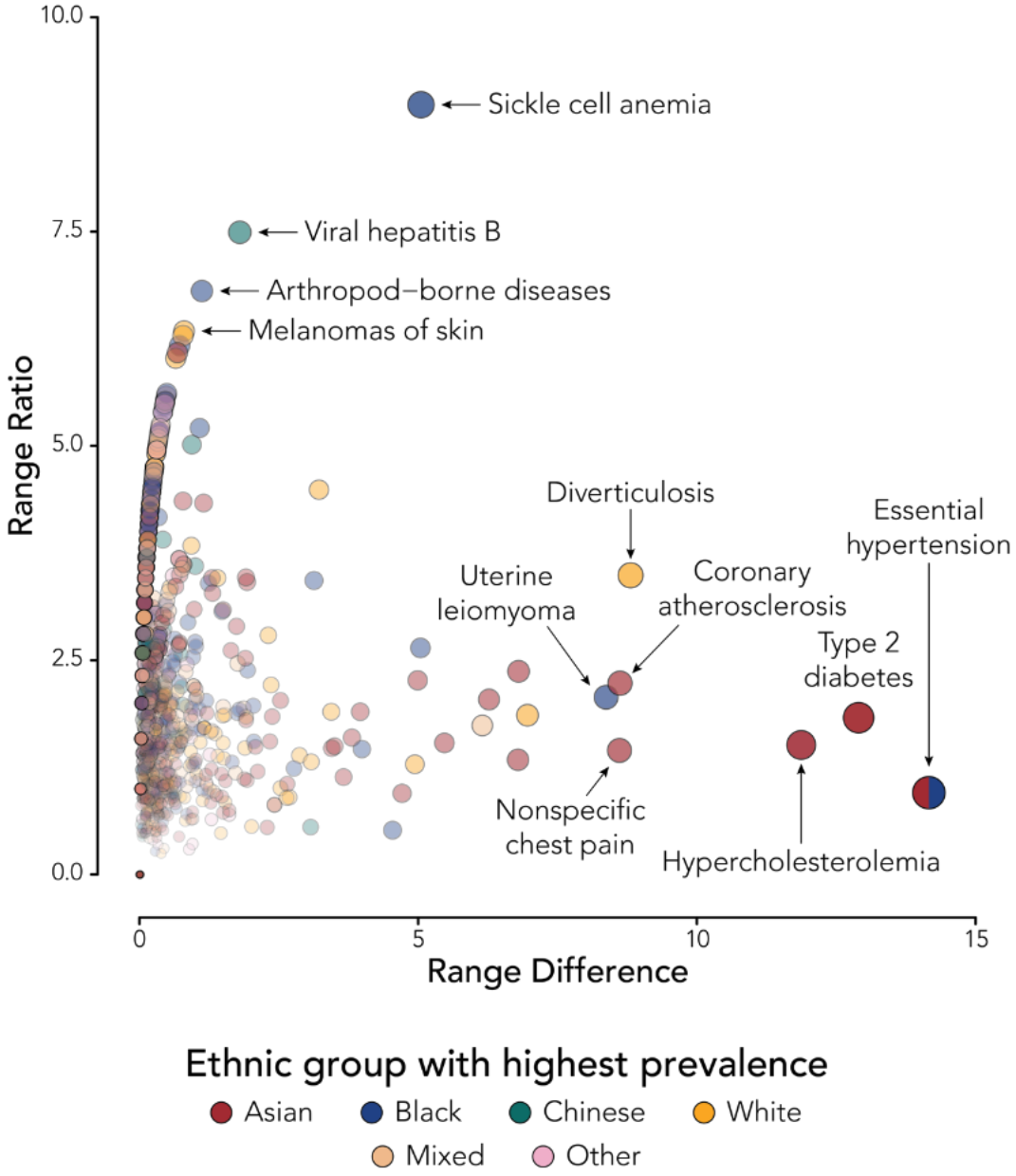


Figure 10. Disease phenotype disparities for ethnic groups.

Each point is a disease phenotype and is colored to indicate the ethnic groups with the highest prevalence for that phenotype. The size and opacity of each point is scaled by the distance from origin.

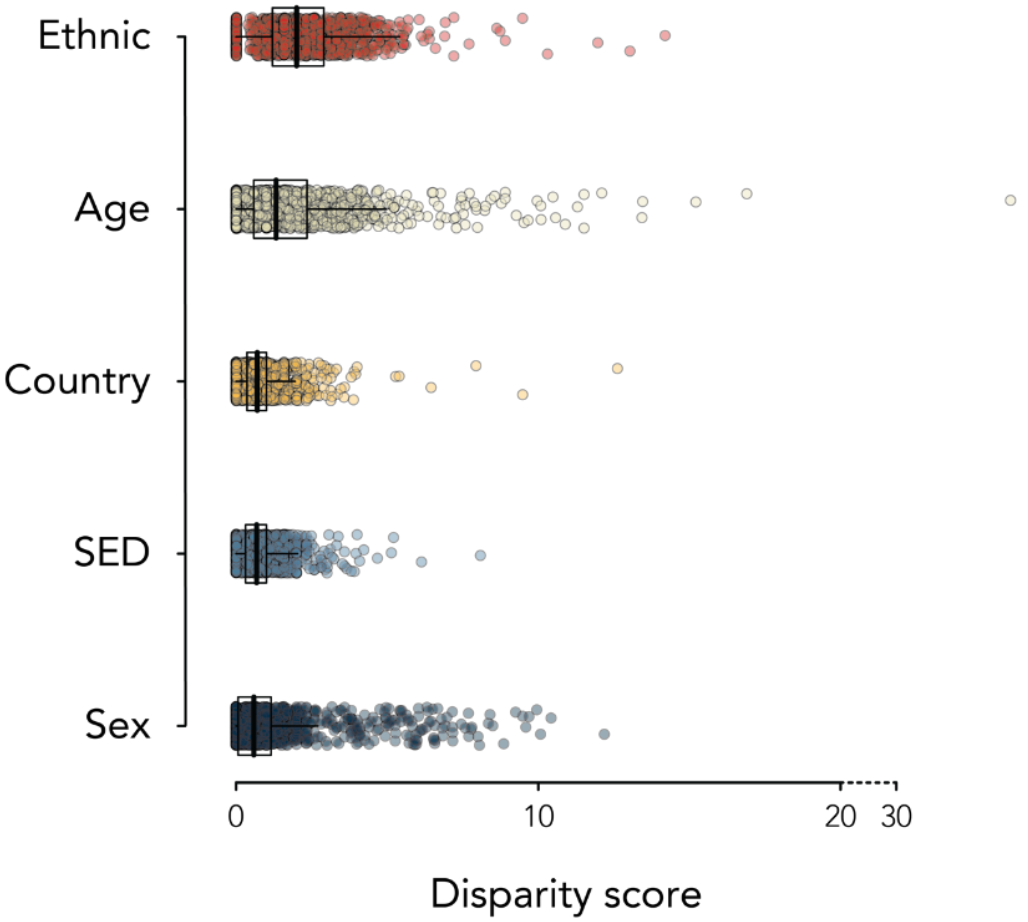


Figure 11. Distribution of disease-level disparity score per population attribute.

Each point is a disease phenotype plotted with its disparity score among the groups defined by each population attribute under consideration.

We evaluated pairwise correlations between the different disease phenotype disparity scores for each population attribute and found that country of residence and SED had the highest correlation (Spearman r = 0.95) and ethnic group and sex had the lowest correlation (Spearman r = 0.27). (Appendix C: Figure 36).

4.4.2 Health disparities among groups defined by population attributes

To identify groups with disproportionately high disease prevalence across phenotypes, we quantified the relative disease burden for groups defined by each population attribute (Figure 12). This was done by calculating the deviation from the number of times a group had the highest prevalence of disease phenotypes compared to the null hypothesis of equally distributed disease prevalence. Among the groups defined by age, we find that participants aged between 65 and 74 years of age had the highest relative burden of disease (1.27) while those aged between 45 and 54 seemed to have the lowest burden of disease (-0.71) in our analysis cohort. For groups defined by country of residence, those residing in England had the highest relative burden of disease (0.91) and those residing in Scotland had the lowest (-0.51). Those identifying as belonging to the Asian ethnic group had the highest relative burden of disease (0.52) while those identifying as Chinese had the lowest (-0.47). We see that the most socioeconomically deprived quintile of participants (Q5) has the highest relative burden of disease (2.00) while those in the third quintile seem to have the lowest (-0.69). We also find that females have a higher relative burden of disease (0.05) compared to males (-0.05).

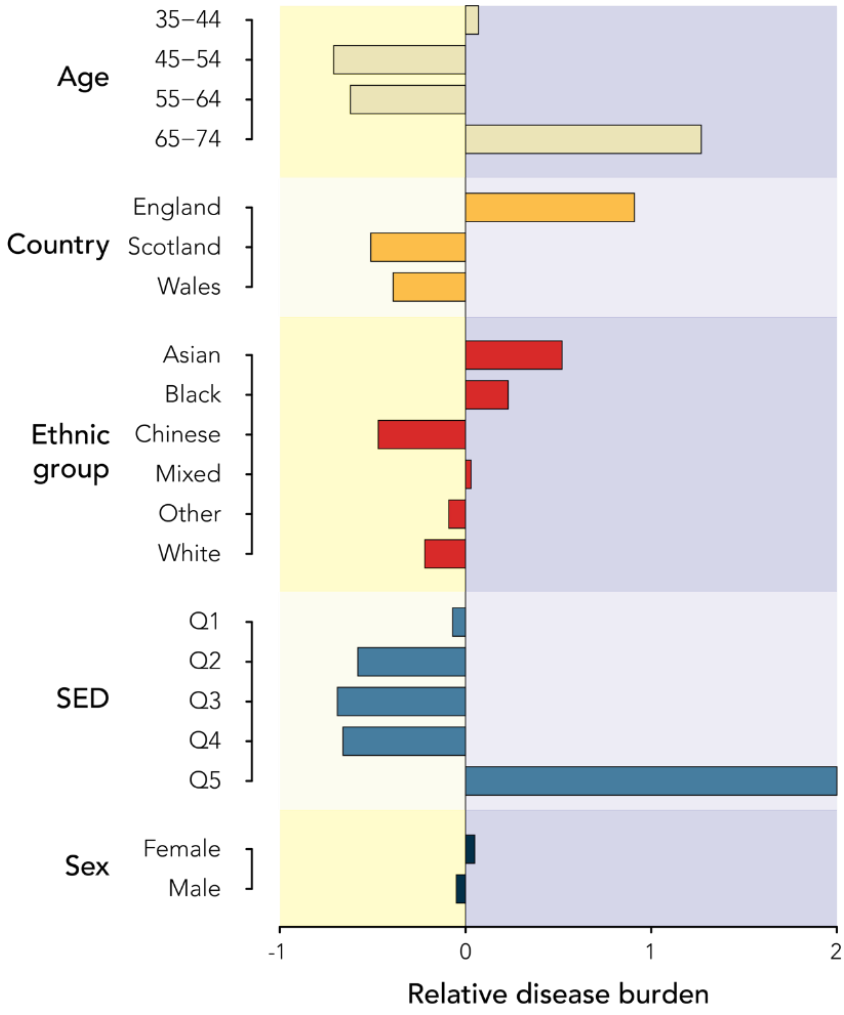


Figure 12. Relative disease burden across groups defined by population attributes.

Each bar shows the relative burden of disease among all groups defined by a population attribute.

We identified the most disparate disease for groups defined by each population attribute under consideration (Table 5). We find that Essential hypertension is a large health disparity across four out of the five population attributes studied here. Prevalence

values for each group defined by the different population attributes under consideration, along with the disparity metric can be accessed using the interactive browser.

Table 5. Most disparate diseases for groups defined by each population attribute.

No.	Phenotype	Range difference	Log_2(range ratio)	Overall prevalence	Group with maximum prevalence
			Age		
1	Essential hypertension	33.67	2.72	22.51	65-74
2	Hypercholesterolemia	16.60	3.14	9.98	65-74
3	Diverticulosis	14.89	3.11	9.40	65-74
4	Obstetrical/birth trauma	9.17	9.84	1.39	35-44
5	Cataract	12.70	4.36	5.45	65-74
			Country		
1	Essential hypertension	12.61	1.11	22.51	England
2	Other mental disorder	9.54	3.82	9.40	England
3	Hypercholesterolemia	7.91	1.94	9.98	Wales
4	Arthropathy NOS	6.44	2.19	7.70	England
5	Allergy/adverse effect of penicillin	5.39	2.37	4.90	Wales
			Ethnic group		
1	Essential hypertension	14.16	0.96	22.51	Asian
2	Type 2 diabetes	12.90	1.83	5.95	Asian
3	Hypercholesterolemia	11.87	1.51	9.98	Asian
4	Sickle cell anemia	5.05	8.98	0.10	Black
5	Diverticulosis	8.81	3.49	9.40	White
			SED		
1	Tobacco use disorder	7.88	1.80	5.78	Q5
2	Essential hypertension	6.12	0.38	22.51	Q5
3	Type 2 diabetes	5.08	1.15	9.24	Q5
4	Other mental disorder	5.08	0.76	12.41	Q5
5	Hypercholesterolemia	4.63	0.64	12.87	Q5
			Sex		
1	Hyperplasia of prostate	7.55	9.56	3.36	Male
2	Uterine leiomyoma	5.21	9.02	2.81	Female
3	Postmenopausal bleeding	4.74	8.89	2.47	Female
4	Excessive or frequent menstruation	4.57	8.84	2.37	Female
5	Cancer of prostate	4.22	8.72	1.84	Male

4.4.3 Interactive health disparities browser

The interactive browser was developed using the Model-View-Controller software design paradigm[183], which divides the program logic into three interconnected elements: the *Model*, the *View*, and the *Controller*. This separation allows for easier management of front and backend components of the browser. In the Model-View-Controller framework, the *Model* represents the data structures and databases which are queried, the *View* represents the user interface, and the *Controller* represents the mediator between these two components (Figure 13).

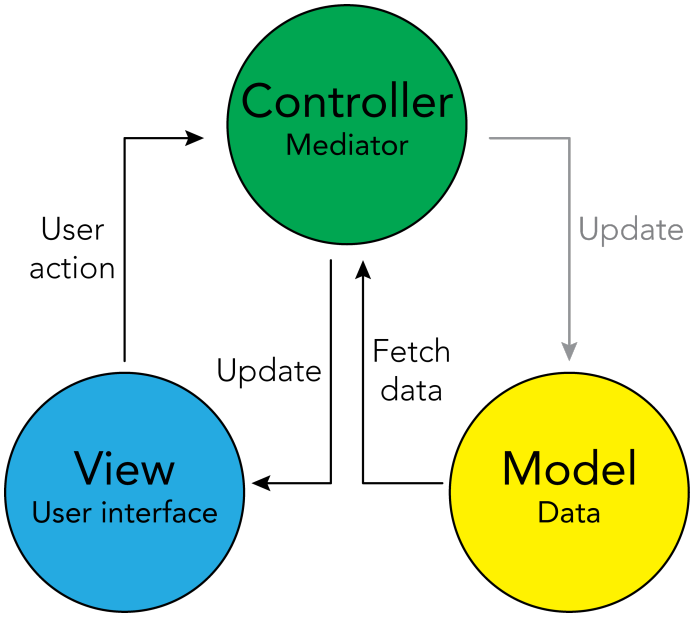
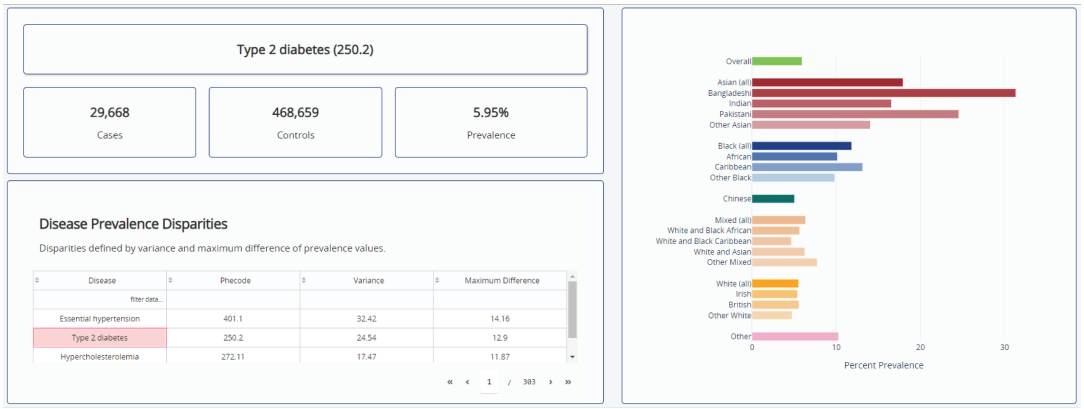


Figure 13. Model-View-Controller software design pattern used for the UK Biobank Health Disparities browser.

Schematic showing the Model-View-Controller (MVC) software design pattern used to develop the interactive webserver. Parts of the pattern that are not applicable to the current browser are greyed out.

The browser allows researchers to identify health disparities among groups based on the population attribute of their choosing. The browser displays disease prevalence values for each group defined using the chosen population attribute, sorted by the disparity score (Figure 14A). There is another table that will help users select disease phenotypes by prevalence in groups (Figure 14B). The tables with information on disease prevalence can be sorted using any of its columns and also allows for keyword searches.

(A)



(B)

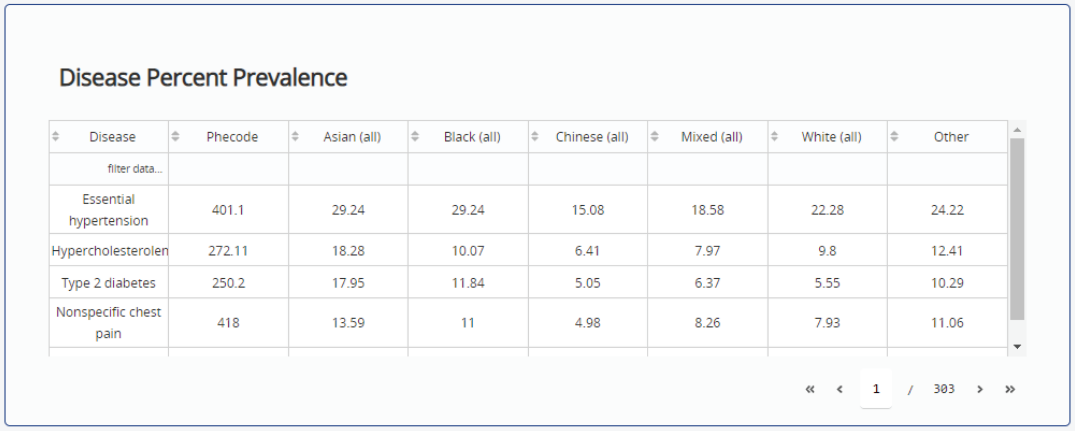


Figure 14. Screenshot of the UK Biobank Health Disparity Browser.

Screenshots of the UK Biobank Health Disparity Browser showing (A) disease phenotype prevalence for different ethnic groups sorted by disparity score and (B) table of disease prevalence for each ethnic group.

4.5 Conclusion

For this study of the UK Biobank, we describe the landscape of health disparities in the analysis cohort. We find that there are several disease phenotypes which exhibit high levels of disparity among the groups defined by the population attributes studied here. An interactive browser documenting the prevalence of disease phenotypes and disparity metrics is accessible at *https://ukbatlas.health-disparities.org*.

CHAPTER 5. SOCIOECONOMIC DEPRIVATION AND GENETIC ANCESTRY INTERACT TO MODIFY TYPE 2 DIABETES ETHNIC DISPARITIES IN THE UNITED KINGDOM

5.1 Abstract

Type 2 diabetes (T2D) is a complex common disease that disproportionately impacts minority ethnic groups in the United Kingdom (UK). Socioeconomic deprivation (SED) is widely considered as a potential explanation for T2D ethnic disparities in the UK, whereas the effect of genetic ancestry (GA) on such disparities has yet to be studied. We leveraged data from the UK Biobank prospective cohort study, with participants enrolled between 2006 to 2010, to model the relationship between SED (Townsend index), GA (clustering principal components of whole genome genotype data), and T2D status (ICD-10 codes) across the three largest ethnic groups in the UK – Asian, Black, and White – using multivariable logistic regression. The Asian group shows the highest T2D prevalence (17.9%), followed by the Black (11.7%) and White (5.5%) ethnic groups. We find that both SED (OR: 1.11, 95% CI: 1.10-1.11) and non-European GA (OR South Asian versus European: 4.37, 95% CI: 4.10-4.66; OR African versus European: 2.52, 95% CI: 2.23-2.85) are significantly associated with the observed T2D disparities. GA and SED show significant interaction effects on T2D, with SED being a relatively greater risk factor for T2D for individuals with South Asian and African ancestry, compared to those with European ancestry. The significant interactions between SED and GA underscore how the effects of environmental risk factors can differ among ancestry groups, suggesting the need for group-specific interventions.

5.2 Introduction

Diabetes is rapidly becoming a global pandemic, largely due to increasing rates of obesity[184]. It is estimated that by 2030, diabetes will impact ~5.5 million individuals in the United Kingdom (UK), with type 2 diabetes (T2D) accounting for ~90% of all cases[185]. T2D is a health disparity that disproportionately impacts minority ethnic groups[186]. Asian and Black ethnic groups in the UK have approximately two to four times the T2D prevalence compared to White and other ethnic groups[185]. Efforts to mitigate health disparities of this kind are both a social imperative and a pressing scientific challenge.

It should be noted that studies of health disparities in the UK often rely on the ethnicity categories used by the National Health Service (NHS)[187]. NHS ethnic categories include six ethnic groups – Asian, Black, Chinese, Mixed, White, and Other – and a distinct ethnic background within each group. UK ethnic group classifications make no distinction between the related concepts of race and ethnicity[188]. Accordingly, the ethnic group labels used in the UK may correspond to racial group labels used in other countries, such as the United States.

T2D is a complex common disease caused by a multifactorial interplay between social, environmental, and genetic factors, all of which contribute to T2D health disparities[189,190]. Accordingly, efforts to elucidate the risk factors associated with T2D ethnic disparities require an integrated approach that considers social, environmental, and genetic components together. An integrated approach of this kind is further distinguished by its potential to characterize how interactions between genetic and environmental factors

contribute to disparate health outcomes. Indeed, gene-by-environment interactions have been prioritized for health disparities research[9,10].

Socioeconomic deprivation (SED) is widely considered an important risk factor for T2D ethnic health disparities[191-193]. Lifestyle conditions associated with higher SED – psychosocial stress, restricted autonomy, and limited access to healthy food, exercise facilities, and health services – have been shown to modify risk for T2D[194-196]. In the UK, SED has been associated with a greater T2D prevalence among minority Asian and Black populations than among those identifying as White[197,198]. Genetic differences between ethnic groups, owing to their different ancestral origins, have also been associated with T2D disparities[192]. In the US and Latin America, both African and Native American genetic ancestry (GA) have been associated with T2D disparities in Black and Hispanic populations[66,199-201]. However, the inclusion of SED has been shown to attenuate the effect of GA on T2D status in these populations[192,200,201]. To our knowledge, there have been no studies that simultaneously consider the impact of GA and SED on T2D ethnic disparities in the UK.

GA provides a number of advantages for health disparities research. Ethnic groups are socially constructed and co-vary with both socioenvironmental and genetic factors. GA inference can be used to stratify populations based on evolutionary genetic diversity alone. A focus on GA can thereby allow for the disambiguation of the genetic and socioenvironmental dimensions of ethnic health disparities. Joint consideration of GA, SED, and their interactions can be used to tailor population-level interventions aimed at mitigating health disparities[9,10].

The objective of this study was to investigate the joint effects of SED and GA on T2D ethnic disparities in the UK. Leveraging the UK Biobank, a large prospective cohort study with genetic and environmental data from more than 500,000 participants, we modeled the relationship between SED, GA and T2D across the three largest ethnic groups in the UK – Asian, Black, and White – using multivariable logistic regression[167]. GA groups were delineated by clustering genetic principal components analysis data, yielding discrete and coherent groups that capture the genetic diversity of the study cohort, thereby isolating genetic from socioenvironmental effects on T2D.

5.3 Materials and methods

5.3.1 Study cohort

The cohort for this study was obtained from the UK Biobank, a prospective cohort study set up to investigate the lifestyle, environmental, and genetic determinants of a range of important diseases of adulthood for participants aged between 40 and 70 years collected between 2006 and 2010[167]. The UK Biobank database contains phenotypic and genotypic information on more than 500,000 participants over multiple waves of collection. Participants provided information in the form of completed questionnaires, nurse-led interviews, medical assessments, and biological samples. Participant DNA was extracted from 850 μL buffy coat aliquots, derived from 10 ml of whole blood, and participant whole genome genotypes were characterized using the UK Biobank Axiom Array or UK BiLEVE Array as previously described[202]. The study adheres to RECORD reporting guidelines.

5.3.2 Population attributes and data filtering

We extracted the following information for UK Biobank participants: (1) age (Field 21003: Age when attended assessment center)[171], (2) sex (Field 31: Sex)[175], (3) Townsend deprivation index (Field 189: Townsend deprivation index at recruitment)[176], (4) ethnic group and background (Field 21000: Ethnic background)[173], (5) ICD-10 codes (Fields 41270: Diagnoses – ICD10)[174], and (6) genetic principal components (Field 22009: Genetic principal components)[203]. As not all of these data fields were available for all participants, the final analysis cohort was constructed by merging these datasets (Appendix D: Figure 37).

UK Biobank participants self-identified as belonging to one of six ethnic groups (Asian, Black, Chinese, Mixed, White, or Other), and a distinct ethnic background within each group, at the time of enrollment. We consider the three largest ethnic groups for analysis: Asian, Black, and White. The corresponding ethnic backgrounds for the ethnic groups considered for our analyses were: Asian (Indian, Pakistani, Bangladeshi, Any other Asian background), Black (Caribbean, African, Any other Black background), and White (British, Irish, Any other white background).

To study levels of SED, we use the Townsend index of deprivation, a widely used measure of SED that is known to be associated with worse health outcomes[204]. The Townsend index is a composite metric that incorporates (1) unemployment, (2) non-car ownership, (3) non-home ownership, and (4) household overcrowding in a given area[177]. Higher (positive) values of the index indicate high material deprivation, whereas lower

(negative) values indicate relative affluence. The cutoff values for the SED quintiles were -3.95, -2.80, -1.37, and 1.23, while those for the SED terciles were -3.17 and -0.68.

5.3.3 Type 2 diabetes prevalence

UK Biobank participants' case or control status for type 2 diabetes (T2D) was determined using ICD-10 diagnosis codes curated following the phecode scheme defined by the PheWAS consortium[179]. The phecode scheme provides disease-specific inclusion and exclusion criteria ICD-10 codes for generating case/control cohorts from electronic health records. This approach allows investigators to define clearly distinct case and control cohorts that can be compared confidently. For example, when studying participants with T2D, participants with type 1 diabetes are removed from the control cohort to avoid any overlapping environmental/genetic signals that might be common to both. This improves power to detect any signals for a condition of interest. The phecode scheme to define case and control cohorts using ICD-10 codes was validated by investigating phenotype reproducibility with the gold standard ICD-9-CM phecode map and by conducting a PheWAS to replicate older, well-known results[179]. Here, inclusion ICD-10 codes were first used to generate the T2D case cohort, and exclusion codes were subsequently used to remove individuals with related conditions from the remaining control cohort. The T2D phecode (250.2) inclusion and exclusion ICD-10 codes can be found at *https://phewascatalog.org/phecodes_icd10*. Participants T2D case and control status were used to calculate crude T2D prevalence values for ethnic groups and

backgrounds as the percent of cases in each group. Crude prevalence values were used owing to the fact that age and sex were included as covariates in all T2D models.

5.3.4 Genetic ancestry inference

UK Biobank participants self-identify as belonging to ethnic groups based on shared culture and heritage. In other words, ethnic groups are socially constructed and thus may not serve as reliable proxies for genetic diversity[119]. Patterns of genetic diversity among UK Biobank participants were characterized by principal components analysis (PCA) of whole genome genotypes as previously described[167]. Genetic ancestry groups were defined by clustering the first three principal component values from the genetic PCA data. Two different clustering approaches were used to generate (1) continuous genetic ancestry groups and (2) coherent genetic ancestry groups. Continuous genetic ancestry groups were characterized using the k-means clustering algorithm, implemented in the function `kmeans` in R v3.6.1[68], using $k = 3$. The value of k was set to three ($k = 3$) to identify three clusters in the PCA data to match the three self-identified ethnic groups under consideration. The resulting groups included all individuals (and are therefore dubbed 'continuous genetic ancestry groups'). Coherent genetic ancestry groups were characterized using the density-based clustering algorithm HDBSCAN[205] implemented in the python module 'hdbscan'. The clustering function was run with a minimum cluster size (`min_cluster_size`) set to 1,000 individuals to extract large, coherent clusters from the data. Density-based clustering only categorizes a subset of participants into ancestry clusters, while marking the rest as uncategorized. In excluding participants that are not

tightly clustered, we were able to obtain coherent and highly distinct genetic ancestry clusters. The resulting GA groups are distinguished by systematic (correlated) allele frequency differences arising from ancestral source populations with distinct biogeographical origins.

5.3.5 Statistical analyses

All statistical analyses were performed using the R statistical language v3.6.1[68]. T2D odds of prevalence were modeled using multivariable logistic regression computed using the 'glm' function in R. Age was standard normalized when included in logistic models. Two logistic regression models were used for analysis – Model 1: T2D ~ GA + SED + Age + Sex + GA*SED and Model 2: T2D ~ GA-SED + Age + Sex. It should be noted that Model 1 includes SED as a continuous variable and its interaction with GA, while Model 2 includes a categorical variable whose levels are given by the combination of GA categories and SED tercile categories, yielding a total of 9 categories with European Low SED as the reference. Odds ratios (ORs) and 95% confidence intervals were calculated for each term in the models by exponentiating the estimated coefficients. Forest plots were generated using the forestmodel R package[206]. The importance of predictors in the multivariable logistic regression was determined using dominance analysis[207] implemented in the R 'dominanceanalysis' v2.0.0 package. Dominance analysis estimates R^2 values for all possible values of predictors and is used to measure the relative importance of predictors by running pairwise comparisons of all predictors in the model as they relate to the outcome variable. Linear regression equations and plots were generated using the R

'ggplot' v3.3.3 library. Slopes of linear regression models were compared by calculating a z statistic as described here[208].

5.3.6 Ethics approval

Ethics approval for the UK Biobank was obtained from the North West Multi-centre Research Ethics Committee (MREC) for the United Kingdom, the Patient Information Advisory Group (PIAG) for England and Wales, and the Community Health Index Advisory Group (CHIAG) for Scotland (see https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics).

5.4 Results

5.4.1 Type 2 diabetes ethnic disparities and socioeconomic deprivation

We generated type 2 diabetes (T2D) case/control cohorts from the UK Biobank using participants' ICD-10 diagnosis codes, with the phecode scheme inclusion and exclusion criteria[179]. Our final analysis cohort had 27,748 T2D cases and 446,436 controls (Table 6). Participant case/control status was used to calculate T2D prevalence for the three largest ethnic groups in the UK – Asian, Black, and White – and for different levels of socioeconomic deprivation (SED). SED is measured using the Townsend index of deprivation, where lower values indicate less deprivation and higher values indicate more deprivation. It can be seen that T2D prevalence varies greatly among different ethnic

groups and backgrounds in the UK (Figure 15A). The Asian group shows the highest prevalence (17.86%) followed by the Black (11.71%) and White (5.51%) groups, respectively. The Asian group also shows the greatest variance of T2D prevalence among constituent ethnic backgrounds. Within the Asian group, the Bangladeshi ethnic background shows the highest T2D prevalence by far (31.65%), with the Indian (16.51%) and Other (14.04%) backgrounds showing prevalence values approximately half as high. Along with the ethnic disparity in the prevalence of T2D, we also see a marked disparity in SED among the three groups under consideration (Figure 15B). The Black group shows the highest level of median SED (2.93) followed by the Asian (0.25) and White (-2.27) groups, respectively. Consistent with what is known about the relationship between SED and T2D, we also find that T2D prevalence increases monotonically with an increase in social deprivation (Figure 15C)[209,210].

Table 6. Characteristics of the T2D analysis cohort.

†SED = Socioeconomic deprivation as measured with the Townsend index. Higher (positive) values of the index indicate high material deprivation, whereas lower (negative) values indicate relative affluence.

Characteristic	Full cohort (n = 474,184)	Asian cohort (n = 9,361)	Black cohort (n = 7,541)	White cohort (n = 457,282)
Age – no. (Cohort share %)				
<45	47,697 (10.06)	1,810 (19.34)	1,611 (21.36)	44,276 (9.68)
45-54	133,102 (28.07)	3,434 (36.68)	3,359 (44.54)	126,309 (27.62)
55-64	201,760 (42.55)	2,920 (31.19)	1,807 (23.96)	197,033 (43.09)
>65	91,625 (19.32)	1,197 (12.79)	764 (10.13)	89,664 (19.61)
Mean age – yr	56.62	53.32	51.90	56.77
Sex – no. (%)				
Female	257,015 (54.20)	4,306 (46.00)	4,309 (57.14)	248,400 (54.32)
Male	217,169 (45.80)	5,055 (54.00)	3,323 (42.86)	208,882 (45.68)
Median SED †	-2.19	0.25	2.93	-2.27
T2D cases – no. (%)	27.748 (6.22)	1.672 (17.86)	883 (11.71)	25,193 (5.51)

To further interrogate the relationship between SED and T2D ethnic disparities, we compared the T2D prevalence with the mean SED for each ethnic group and background (Figure 15D). We see that a strong relationship does exist between group specific T2D prevalence and SED, but the disparity is not completely explained by SED. Participants who identify as Black have higher average SED but a much lower T2D prevalence

compared to participants who identify as Asian, who have lower average SED compared to Black participants but far higher T2D prevalence. Furthermore, on plotting T2D prevalence per ethnic group for each SED quintile, we find that the ethnic disparities remain within each strata of SED, indicating that other factors also contribute to the T2D ethnic disparities (Figure 15E).

5.4.2 Genetic ancestry groups

Principal components analysis (PCA) of participants' whole genome genotypes were used to generate discrete and coherent genetic ancestry (GA) groups. Overall, participants' self-identified ethnicity co-varies with GA groups defined using PCA (Appendix D: Figure 38). Nevertheless, there are numerous cases where participants' self-identified ethnicity does not align with GA groups. Accordingly, we rely on GA group analysis to more precisely measure genetic differences that may be associated with T2D ethnic disparities.

GA groups were delineated by performing density-based clustering of participant genetic PCA data, yielding three coherent groups: African (n = 5,176), European (n = 448,446), and South Asian (n = 6,969) (Figure 16). The ancestral origins for these groups are based on the majority self-identification of member participants. These GA groups represent non-overlapping, discrete, and genetically diverse cohorts.

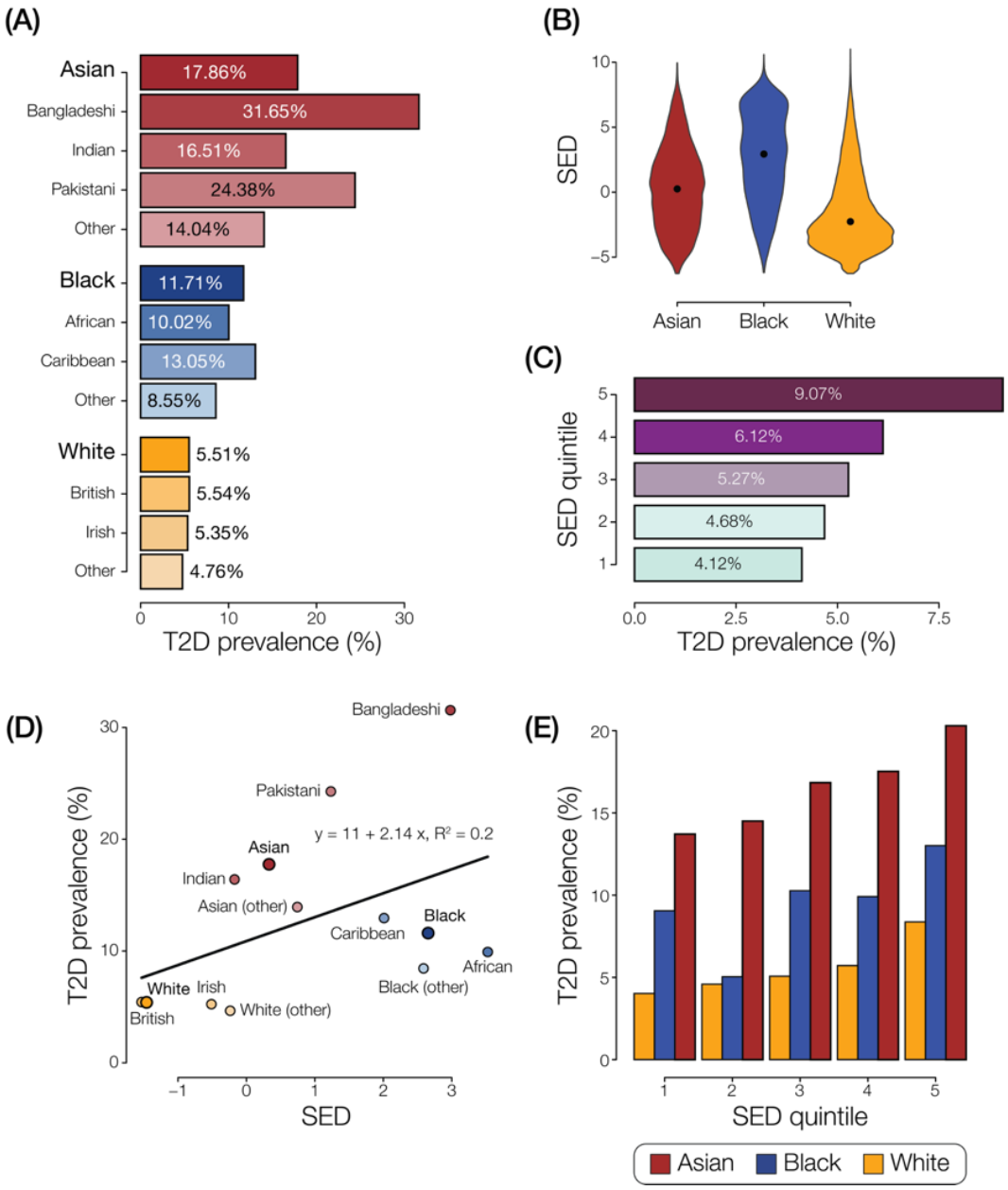


Figure 15. T2D ethnic health disparities and SED.

(A) T2D prevalence for ethnic groups and backgrounds. (B) SED distributions for ethnic groups. (C) T2D for SED quintiles, 1-least deprivation to 5-highest deprivation. (D) Relationship between T2D prevalence (y-axis) and mean SED (x-axis) for ethnic groups and backgrounds. (E) T2D ethnic prevalence disparities across SED quintiles.

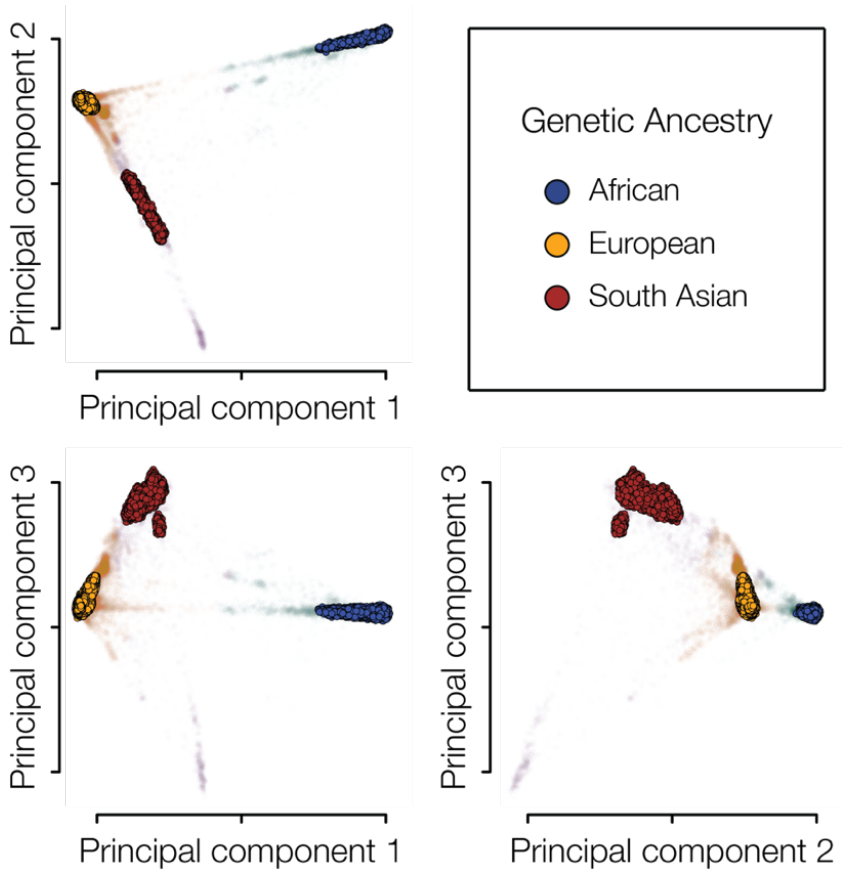


Figure 16. GA groups.

Clustering of genetic PCA data was used to generate continuous and coherent GA groups: African (blue), European (orange), and South Asian (red). Participants that fall into coherent ancestry groups are prominently colored, and participants that fall into the continuous groups are shown as faded points.

5.4.3 Genetic ancestry, socioeconomic deprivation, and type 2 diabetes

We modeled T2D case/control status using GA and SED along with the covariates age and sex using multivariable logistic regression (Model 1; Appendix D: Table 12). Model 1 includes SED as a continuous variable and its interaction with GA. This analysis shows that being part of the South Asian GA group compared to being in the European GA group had the highest impact on modifying T2D risk (OR: 4.37, 95% CI: 4.10 - 4.66),

followed by being in the African GA group (OR: 2.52, 95% CI: 2.23 - 2.85). As would be expected, age (OR: 1.78, 95% CI: 1.75 - 1.80), being male (OR: 1.86, 95% CI: 1.81 - 1.90), and SED (OR: 1.11, 95% CI: 1.10 - 1.11) are all significantly associated with T2D risk. Dominance analysis shows that the most important predictors to explain T2D status in this model are age, sex, GA group, and SED. However, we found the GA-SED interaction terms – South Asian-SED and African-SED – to be statistically significant (p-values of 0.001 and 0.016, respectively), suggesting that the impact of SED on T2D varies among GA groups. The full model that includes the interaction term has a significantly higher log likelihood than a reduced model with no interaction term, further supporting the presence of GA-SED interactions (likelihood ratio χ^2=15.96 P=3.4 x 10^{-4}; Appendix D: Table 13). Given the observed GA-SED interactions, it is not possible to make any firm conclusions regarding the relative importance GA versus SED on T2D outcomes.

Next, we used another logistic regression model (Model 2; Appendix D: Table 14), which includes a categorical variable whose levels are given by the combination of GA categories and SED tercile categories, yielding a total of 9 categories with European Low SED as the reference. As seen for Model 1, age, sex, GA and SED all show significant associations with T2D status with Model 2 (Figure 17). The relative impact of GA groups on T2D status is the same: European has the lowest effect sizes, followed by African, and South Asian showing the highest effect sizes. For each GA group, increasing SED is consistently associated with greater effect sizes, thereby confirming the GA-SED interactions detected in Model 1.

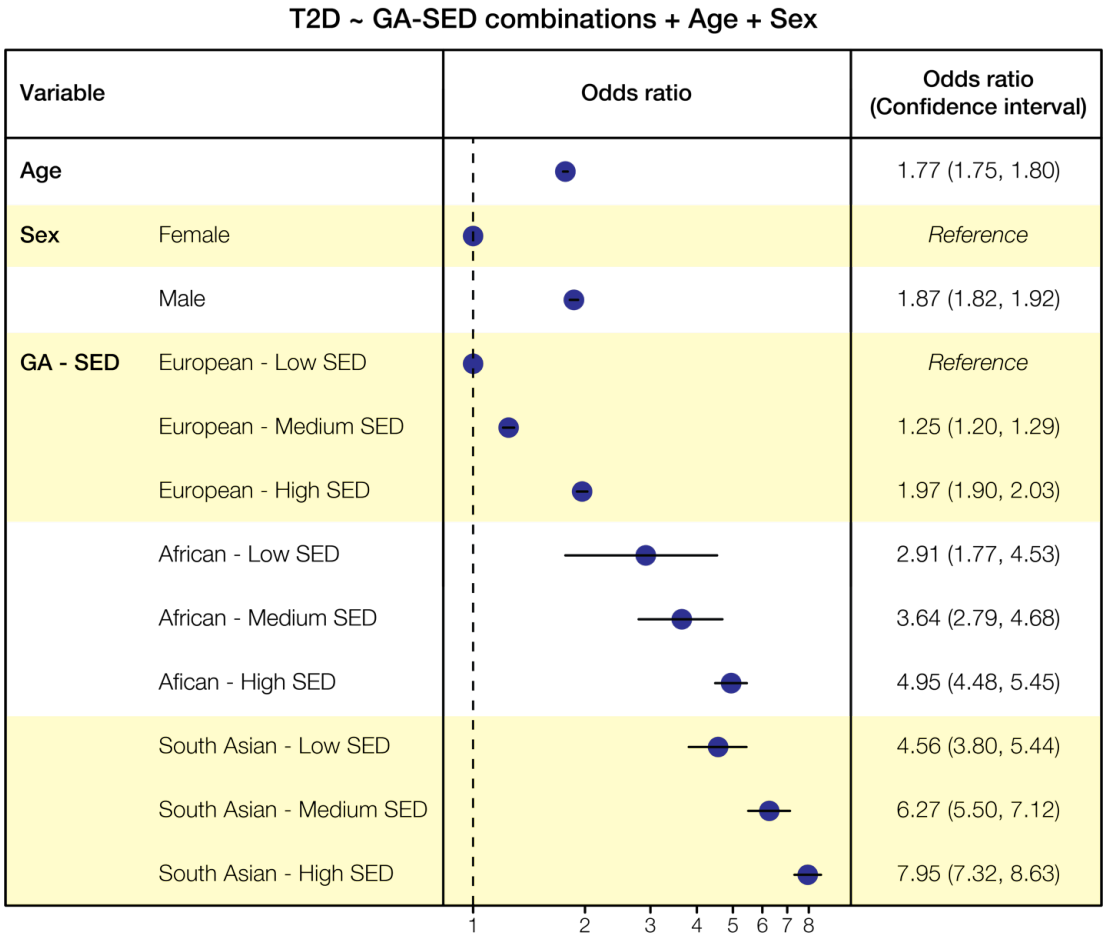


Figure 17. T2D multivariable logistic regression model with GA-SED tercile combinations (Model 2).

Model 2 includes terms for GA groups combined with low, medium, and high SED terciles, age, and sex. The forest plot shows odds ratios and 95% confidence intervals along with the statistical significance for each variable used to model T2D status. Details of the estimated coefficients, their standard errors, and p-values are shown in Appendix C: Table 14.

To further characterize the impact of GA-SED interactions on T2D status, we modeled T2D case/control status using SED, along with the covariates age and sex, using multivariable logistic regression and then stratified the results by GA groups. For each individual GA group, the logistic model was used to calculate the probability of predicted T2D per participant. T2D observed prevalence values increase monotonically for each GA

group across T2D model prediction quintiles, and the relative T2D prevalence values for each GA group stay the same within each quintile (Figure 18A). On regressing T2D observed prevalence against T2D model predictions per GA group and fitting a linear trend for each group separately, we found that the slopes for each GA group differed substantially (Figure 18B). The magnitude of association between SED and T2D for the South Asian group is ~2.5 times higher than the European group, and the African group association is ~1.5 times higher than the European group. Differences in the slopes are all statistically significant – confirming the interaction effects between GA and SED (African – European slope p-value=7.28×10^{-07}, African – South Asian slope p-value=1.39×10^{-05}, and European – South Asian slope p-value=7.33×10^{-25}). Furthermore, the intercept of this fitted line is higher for the South Asian group implying that even at the lowest possible SED level recorded in these data, the risk for T2D is relatively high in this group.

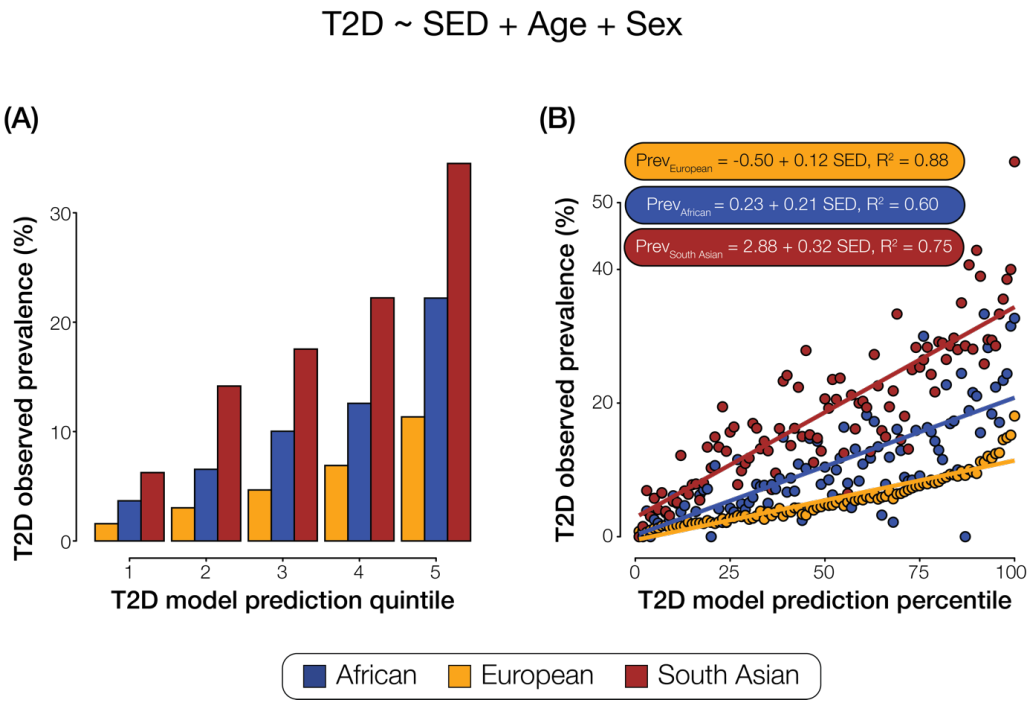


Figure 18. Interaction between genetic ancestry and socioeconomic deprivation.

T2D was predicted using a multivariable logistic regression model using SED, age, and sex as terms. T2D prevalence per GA group partitioned by quintiles (A) and percentiles (B) of SED model predictions. Linear equations and model fits are shown for each ancestry group in panel B. Ancestry groups are color coded as shown.

5.5 Discussion

We make a crucial distinction between GA and self-identified ethnicity in this study. As part of the UK Biobank enrollment survey, participants are asked to identify their ethnic group followed by their ethnic background (i.e., subgroup). For example, participants that identify with the Asian ethnic group are then prompted to choose from Bangladeshi, Indian, Pakistani, or Other Asian backgrounds. These self-identified ethnic group and background identities are social constructs based on shared heritage and culture, whereas GA reflects genetic differences among populations with distinct biogeographic origins. The approach

of forming coherent clusters from genetic PCA data allowed us to generate discrete, non-overlapping GA groups, which can be used to help us disambiguate socioenvironmental factors from genetic factors that might contribute to T2D ethnic disparities. It should be noted that the GA groups delineated here and the participant self-identified ethnic groups assess different constructs and are not entirely concordant (Appendix D: Figure 38). There are a number of cases where participants' self-identified ethnicity does not coincide with their GA, but the majority of participants' ethnic identities correspond to their GA. This reflects the fact that social determinants of ethnicity are strongly informed by notions of ancestral origins and may correlate with phenotypic characteristics.

SED is used here as a proxy for lifestyle factors and environmental exposures that might exacerbate or ameliorate risk for T2D. The implications of a significant interaction between SED and GA groups can be attributed to a number of different factors. Lifestyle and exposures that co-vary with higher SED may have a disproportionately higher impact on T2D risk in certain populations owing to their genetics, and/or higher SED may lead to different lifestyle and exposures among different populations. The latter possibility could include influences on SED-related experiences of structural oppression that differ among GA groups. In any case, targeted group-specific interventions that are informed by such differences can help to decrease T2D health disparities.

There are several potential limitations to our observational study of T2D health disparities. Some cultural attributes like diet and lifestyle factors might co-vary with GA, SED and self-reported ethnicity, especially for recent immigrants. Thus, the observed GA effects on T2D could be attributed to unmeasured confounders. Co-variation between GA

and environmental factors might change over time, with second and third generation immigrants becoming acculturated and changing their dietary habits. It has been shown that second generation Asians in England are more likely to be obese compared to the first generation of immigrants [211]. It is also known that the risk for T2D in South Asians increases for a BMI >23 compared to a BMI of >25 in Europeans[212]. We did not account for generation of immigration in our GA analyses.

SED was measured here using the Townsend Index, which is a composite metric of four different variables, each of which may reflect different kinds of adverse exposures. This measure of SED may miss important indicators such as household income and education level. As this is an observational study, albeit with a large sample size, it is hard to completely disentangle the effects of different contributing factors on the observed health disparity. In addition, the UK Biobank recruited participants who are healthier, on average, compared to the general population and live in less socioeconomically deprived areas compared to non-participants (also referred to as a 'healthy volunteer bias). Regardless, disease-exposure relationships in the UK Biobank are thought to be generalizable, irrespective of the healthy volunteer bias[169].

Finally, it should be noted that the PCA clustering approach used for GA inference yields groups that are largely concordant with continental ancestry. Accordingly, there is a substantial overlap between the GA groups analyzed here and participants' ethnic self-identification (Appendix D: Figure 38). A more nuanced approach that includes quantitative GA estimates, i.e. percent ancestry contributions from ancestral source populations, could help to further disambiguate genetic from socioenvironmental effects

on T2D. Furthermore, the use of GA poses operational difficulties in targeting the impacted communities since this information is not readily available to policymakers and physicians. However, once a gene-by-environmental interaction is identified, as is the case here for the interaction between GA and SED, population-specific interventions and policies can be targeted at the closest corresponding ethnic groups where there exists a high concordance between GA and ethnic groups (Appendix D: Figure 38).

5.6 Conclusion

For this study on the UK Biobank, we confirmed previously observed T2D ethnic disparities and found that SED is indeed a significant risk factor for T2D. We report for the first time that T2D is associated with significant interactions between GA and SED. In particular, SED is a relatively greater risk factor for T2D for individuals with South Asian and African ancestry, compared to those with European ancestry. This finding suggests that more ancestry-specific interventions need to be taken at the policy level to ameliorate health disparities, channeling resources to communities which are at highest risk.

CHAPTER 6. COMPARING GENETIC AND SOCIOENVIRONMENTAL CONTRIBUTIONS TO ETHNIC DIFFERENCES IN C-REACTIVE PROTEIN

6.1 Abstract

C-reactive protein (CRP) is a routinely measured blood biomarker for inflammation. Elevated levels of circulating CRP are associated with response to infection, risk for a number of complex common diseases, and psychosocial stress. The objective of this study was to compare genetic and socioenvironmental contributions to ethnic differences in C-reactive protein levels. We modeled the effects of demography, genetics, and socioeconomic status on CRP blood serum levels using the UK Biobank (UKBB) prospective cohort study. CRP serum levels are significantly associated with ethnicity, age, and sex in the UKBB cohort. Study participants who identify as Black have higher average CRP than those who identify as White, CRP increases with age, and females have higher average CRP than males. Ethnicity and sex show a significant interaction effect on CRP. Black females have higher average CRP levels than White females, whereas White males have higher average CRP than Black males. Socioeconomic deprivation explains more than twice the variation in CRP levels than genetic ancestry, and the effect of ethnicity on CRP is mediated by socioeconomic deprivation but not by genetic ancestry. Taken together, these results indicate that socioenvironmental factors contribute more to CRP ethnic differences than genetics. Differences in CRP are associated with ethnic disparities for a number of chronic diseases, including type 2 diabetes, essential hypertension,

sarcoidosis, and lupus erythematosus. Our results indicate that ethnic differences in CRP are linked to both socioeconomic deprivation and numerous ethnic health disparities.

6.2 Introduction

C-reactive protein (CRP) is synthesized by hepatocytes and secreted to the bloodstream in response to inflammation. CRP is employed as a serum biomarker for both acute and chronic inflammation, with important implications for immune response and overall health[213,214]. Elevated levels of CRP have been shown to be associated with an increased risk of diabetes[215], cardiovascular disease[216], psychological stress[217], and all-cause mortality[218].

CRP blood serum levels vary across ethnic groups[219], with a number of studies showing that Black patients have higher average levels of circulating CRP than White patients[220-227]. Ethnic differences of this kind are likely to have multifactorial causes, including contributions from genetic, socioeconomic, and environmental factors. Given the fact that ethnicity co-varies with all of these classes of risk factors, it is difficult to tease apart the genetic and socioenvironmental contributions to ethnic health disparities. This is further complicated by the fact that socially defined ethnicity is an imprecise proxy for genetic diversity.

We use genetic ancestry inference as a means to disambiguate genetic and socioenvironmental effects on ethnic health disparities. Genetic ancestry refers to patterns of genetic diversity that are linked to the geographical origins of human populations[28].

Individuals who share common ancestors have genetic similarities, and distinct ancestry groups show correlated allele frequency differences[31,65]. Genetic ancestry can be defined objectively, using comparative genomic analysis, without relying on socially defined ethnic groups [119]. Patterns of genetic ancestry can be compared to self-identified ethnicity to understand the extent to which they overlap and how they may differ[32,141,228]. Modelling of health outcomes with genetic ancestry and socioenvironmental factors as independent (predictor) variables can be used to assess how each contribute to health disparities and how they may interact[229,230].

The objective of this study was to characterize the effects of genetic ancestry and socioeconomic deprivation on ethnic differences in CRP serum levels. Participants from the UK Biobank (UKBB) prospective cohort study who self-identified as belonging to Black or White ethnic groups were characterized with respect to CRP levels, genome-wide genotypes, and socioeconomic deprivation. Multivariable modeling of study participants' CRP levels was conducted using genetic ancestry and socioeconomic deprivation as independent (predictor) variables, and structural equation models were used to quantify the mediating effects of each on CRP ethnic differences. Age and sex were considered as covariates in all models given their known associations with CRP serum levels.

6.3 Materials and methods

6.3.1 Study cohort

Study participants and data were taken from the UK Biobank (UKBB), a prospective cohort study on the effects of demography, environment, and genetics on health and disease[167]. The UKBB database contains phenotypic, clinical, and genetic information on more than 500,000 participants between the ages of 40 and 70, enrolled from 2006 to 2010. Ethics approval for the UKBB was obtained from the North West Multi-centre Research Ethics Committee (MREC) for the United Kingdom, the Patient Information Advisory Group (PIAG) for England and Wales, and the Community Health Index Advisory Group (CHIAG) for Scotland[a]. UKBB participants self-identified as belonging to a single ethnic group upon enrollment[b], and we included participants who identified as Black or White for this study. It should be noted that the UKBB ethnic group labels used here correspond directly to racial group labels from the United States.

6.3.2 Participant data

UKBB participants completed questionnaires, nurse-led interviews, and medical assessments upon enrollment and provided access to their electronic health records. We accessed participant information on ethnicity (Field 21000: Ethnic background), age (Field 21003: Age when attended assessment center), sex (Field 31: Sex), Townsend deprivation

[a] https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics

[b] https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=21000

index (Field 189: Townsend deprivation index at recruitment), and ICD-10 disease diagnosis codes (Fields 41270: Diagnoses – ICD10) from the UKBB data portal.

UKBB participants provided whole blood samples for characterization of protein biomarkers and DNA as previously described[168]. C-reactive protein (CRP) blood serum levels were measured as mg/L units using the immuno-turbidimetric method with the Beckman Coulter AU5800 clinical chemistry analyzer[c]. This procedure corresponds to the high-sensitivity (hs) CRP test. DNA was extracted from 850μL buffy coat blood aliquots[d], and participant genome-wide genotypes were characterized using the UKBB Axiom Array or UK BiLEVE Array[202].

6.3.3 Disease case/control cohorts

Disease (or health condition) diagnoses for study participants were taken from UKBB ICD-10 diagnosis codes, which were then converted into disease-specific phenotype codes (phecodes) using the scheme developed by the PheWAS consortium[178]. Phecodes have been manually curated and validated by disease experts, and they are widely used for the analysis of electronic health record data[179]. The phecode scheme provides ICD-10 code inclusion and exclusion criteria for each individual disease in order to define disease-specific case/control cohorts that can be confidently compared. For example, when studying participants with type 2 diabetes, participants with type 1 diabetes are removed

[c] https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/serum_biochemistry.pdf

[d] https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ukb_dna_processing.pdf

from the control cohort to avoid any overlapping genetic or environmental signals that might be common to both. This approach improves power for the detection of disease-specific association signals when modelling case/control status. Phecode case/control cohorts were curated for a total of 1,537 diseases or health-related conditions.

6.3.4 Genetic ancestry inference

UKBB participant genome-wide genotypes were merged and harmonized with whole genome sequence data from global reference populations characterized as part of the 1000 Genomes Project (1KGP) and the Human Genome Diversity Project (HGDP)[31,65]. Global reference populations were grouped into six regional ancestry groups based on their genetic and geographic affinity, including African (sub-Saharan) and European reference population groups (Appendix E: Table 15).

UKBB, 1KGP, and HGDP genomic variant data were merged to include variants present in all three data sets. Minor allele frequency >1% and variant sample missingness <5% filters were used for merging, with variant strand flips and identifier inconsistencies corrected as needed. The merged genome variant data set was pruned for linkage disequilibrium using the program PLINK v2[129].

Principal component analysis (PCA) of the harmonized UKBB, 1KGP, and HGDP genome variant dataset was performed using the FastPCA program implemented in PLINK v2[231]. PCA data were used to infer UKBB participant genetic ancestry fractions for African, European, and other regional ancestry groups. Our PCA-based genetic ancestry

inference approach compares PCA data from UKBB participants to PCA data from reference population individuals using non-negative least squares to assign genetic ancestry fractions for regional ancestry groups as previously described[58,126]. Participants showing >5% non-African or non-European ancestry fractions were excluded from the study cohort.

6.3.5 Statistical modelling

All statistical analyses were performed using the R statistical language v3.6.1[68]. Forest plots were generated using the forestmodel R package[232]. Other plots were generated using the ggplot R package[181].

Linear regression: CRP blood serum levels, measured in mg/L units, were modeled as the dependent (outcome) variable with multivariable linear regression models using the 'lm' function in R. Independent (predictor) variables included ethnicity, age, sex, genetic ancestry, and socioeconomic deprivation. Ethnicity was modeled as a binary variable (Black or White), age was modeled as increments of ten years, and sex was modeled as a binary variable (female or male). Socioeconomic deprivation was modeled using the Townsend deprivation index, a widely used measure of socioeconomic deprivation known to be associated with poor health outcomes[204]. It combines four variables – unemployment, non-car ownership, non-home ownership, and household overcrowding – to generate a numerical score[177], which ranges from -6.26 to 11.0 in the UKBB study cohort. Negative values indicate less socioeconomic deprivation, and relative affluence, whereas higher

scores indicate greater socioeconomic deprivation. African ancestry was modeled as increments of ten percent African genetic ancestry.

Logistic regression: The odds of prevalence of specific diseases or health conditions were modeled as the dependent (outcome) variable with multivariable logistic regression computed using the 'glm' function in R. Independent (predictor) variables for disease models included ethnicity, age, sex, and CRP levels.

Details of all statistical models are provided in the Appendix E. For each model, we provide the regression equation and model coefficients, along with effect size estimates, standard errors, z-values, and P-values for each model coefficient.

6.4 Results

6.4.1 C-reactive protein, ethnicity, age, and sex

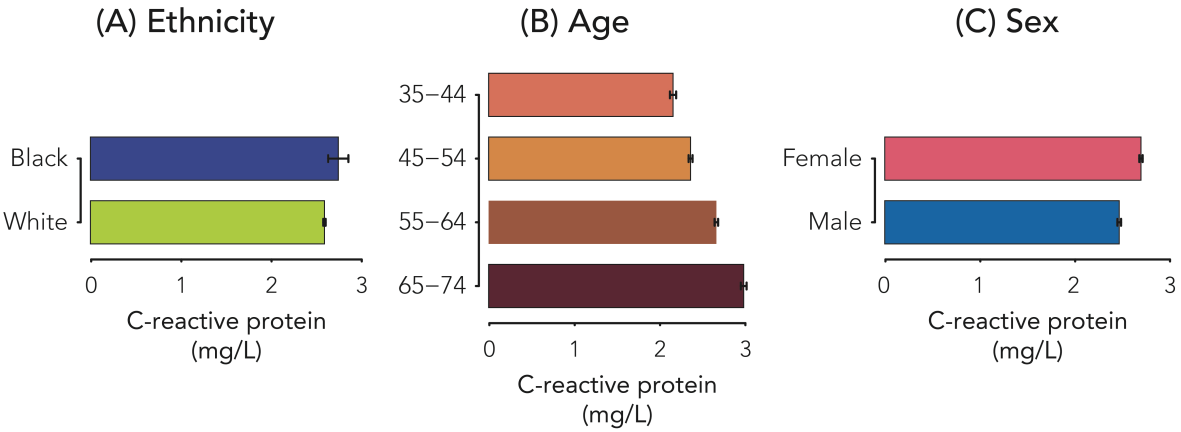
The study cohort is made up of 433,298 UK Biobank participants who self-identify as Black (n=6,456) or White (n=426,842) (Table 7). Males make up 45.7% of the cohort compared to 54.3% females, and the mean age of cohort participants is 57. C-reactive protein (CRP) blood serum levels vary by ethnicity, age, and sex. Black participants show a mean CRP level of 2.75 mg/L, and White participants show mean CRP level of 2.59 mg/L (Table 7 and Figure 19A). Participant CRP levels increase with increasing age (Figure 19B), and females show higher mean CRP levels than males (Figure 19C). When CRP levels are modeled by ethnicity, age, and sex, Black ethnicity and age show significant

positive associations with CRP, whereas male sex shows a significant negative association with CRP (Figure 19D; Appendix E: Table 16).

Table 7. Characteristics of the UK Biobank participant cohort

Characteristic	Full cohort	Black	White
N	433,298	6,456	426,842
Mean age (yrs)	56.71 (56.69 – 56.73)	52.08 (51.88 – 52.28)	56.78 (56.58 – 56.98)
Sex – no. (%)			
Female	235,318 (54.31)	3,721 (57.64)	231,597 (54.26)
Male	197,980 (45.69)	2,735 (42.36)	195,245 (45.74)
Mean C-reactive protein (mg/L)	2.60 (2.59 – 2.61)	2.75 (2.64 – 2.86)	2.59 (2.48 – 2.70)
Mean European ancestry (%)	98.45 (98.42 – 98.48)	11.71 (11.45 – 11.97)	99.76 (99.74 – 99.78)
Mean African ancestry (%)	1.33 (1.30 – 1.36)	87.81 (87.55 – 88.07)	0.03 (0.02 – 0.04)
Mean Townsend index	-1.42 (-1.43 – -1.41)	2.63 (2.55 – 2.71)	-1.48 (-1.55 – -1.41)

Inclusion of interaction terms in the CRP linear regression model revealed a significant interaction between ethnicity and sex ($P<2\times10^{-16}$; Appendix E: Table 17). A likelihood ratio test showed a significantly better fit for a model with the ethnicity-sex interaction term compared to the model with no interaction term, providing additional support for the interaction ($P=1.82\times10^{-16}$; Appendix E: Table 18). The observed ethnicity-sex interaction results from higher CRP for Black female participants compared to White female participants and lower CRP for Black male participants compared to White male participants (Figure 20).



(D) C-reactive protein ~ Ethnicity + Age + Sex

Variable		N	Estimate (95% CI)	*P*-value
Ethnicity	White	426,842	*Reference*	
	Black	6,456	0.30 (0.20, 0.41)	2.7×10^{-08}
Age		433,298	0.34 (0.32, 0.36)	$<2 \times 10^{-16}$
Sex	Female	235,318	*Reference*	
	Male	197,980	-0.24 (-0.27, -0.22)	$<2 \times 10^{-16}$

Figure 19. C-reactive protein (CRP), ethnicity, age, and sex.

Average CRP serum levels (±95% CI) are shown for (A) Black and White participants (B) age ranges, and (C) female and male participants. (D) Forest plot showing the results of the multivariable linear regression model of participant CRP serum levels. Effect sizes, 95% CIs, and P-values are shown for ethnicity, age, and sex.

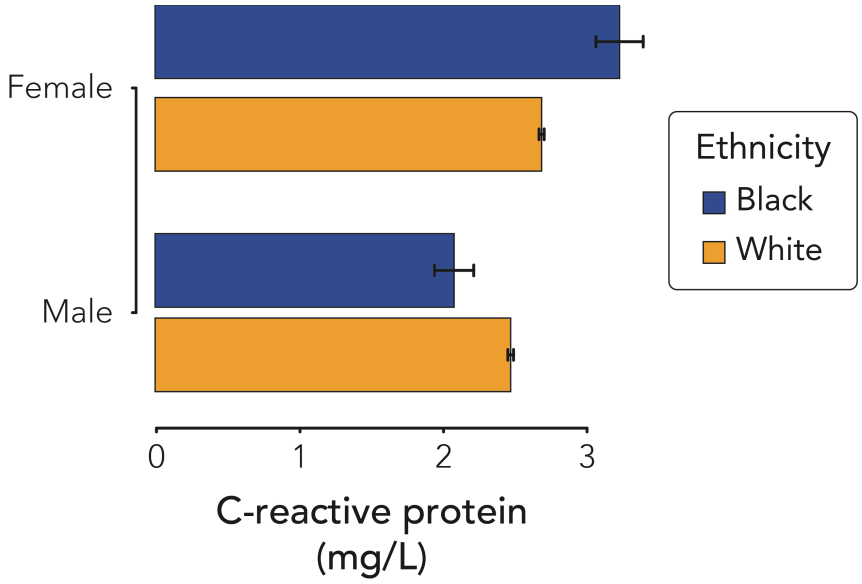


Figure 20. C-reactive protein (CRP) interaction effect of sex and ethnicity.

Average CRP serum levels (±95% CI) are shown for Black and White, female and male participants.

6.4.2 Ethnicity, genetic ancestry, and socioeconomic deprivation

Black and White participants differ with respect to mean levels of genetic ancestry and socioeconomic deprivation (Table 7). Black participants show averages of 87.8% African ancestry and 11.7% European ancestry compared to averages of 99.8% European and 0.03% African ancestry for White participants. Black participants have an average Townsend deprivation index of 2.63 compared to -1.48 for White participants, where higher (positive) values indicate greater socioeconomic deprivation.

The relationship between ethnicity and genetic diversity for Black and White participants is shown via principal components analysis of genome-wide genotype data

(Figure 21A). Principal components one and two separate participants by ethnicity along a continuum of genetic diversity, whereas principal component two alone shows more within ethnic group differences among participants. Black participants show a range of admixture between African and European genetic ancestry fractions, whereas White participants show almost entirely European ancestry (Figure 21B). The probability of participant self-identification as Black or White shifts in the range of 23-44% African ancestry (Figure 21C). Participants are more likely to identify as Black, and less likely to identify as White, if they have ≥29% African ancestry.

Structural equation modelling was used to evaluate the mediating effect of genetic ancestry and socioeconomic deprivation on ethnic differences in CRP serum levels. When African genetic ancestry is modelled as a potential mediator, the total effect of ethnicity on CRP is significant but the indirect mediating effect of African ancestry is non-significant (Figure 22A). When socioeconomic deprivation, as measured by the Townsend deprivation index, is modeled as a potential mediator, both the indirect effect of socioeconomic deprivation and the total effect of ethnicity on CRP levels are significant (Figure 22B). Socioeconomic deprivation explains 87.9% of the total effect of ethnicity on CRP levels. Details of the structural equation models are shown in Appendix E: Tables 19 and 20.

When the effect of genetic ancestry and socioeconomic deprivation on CRP serum levels are modeled separately, socioeconomic deprivation explains more than twice as much of the variation in CRP levels ($R^2=9.32\times10^{-3}$) compared to genetic ancestry ($R^2=4.58\times10^{-3}$) (Appendix E: Table 21).

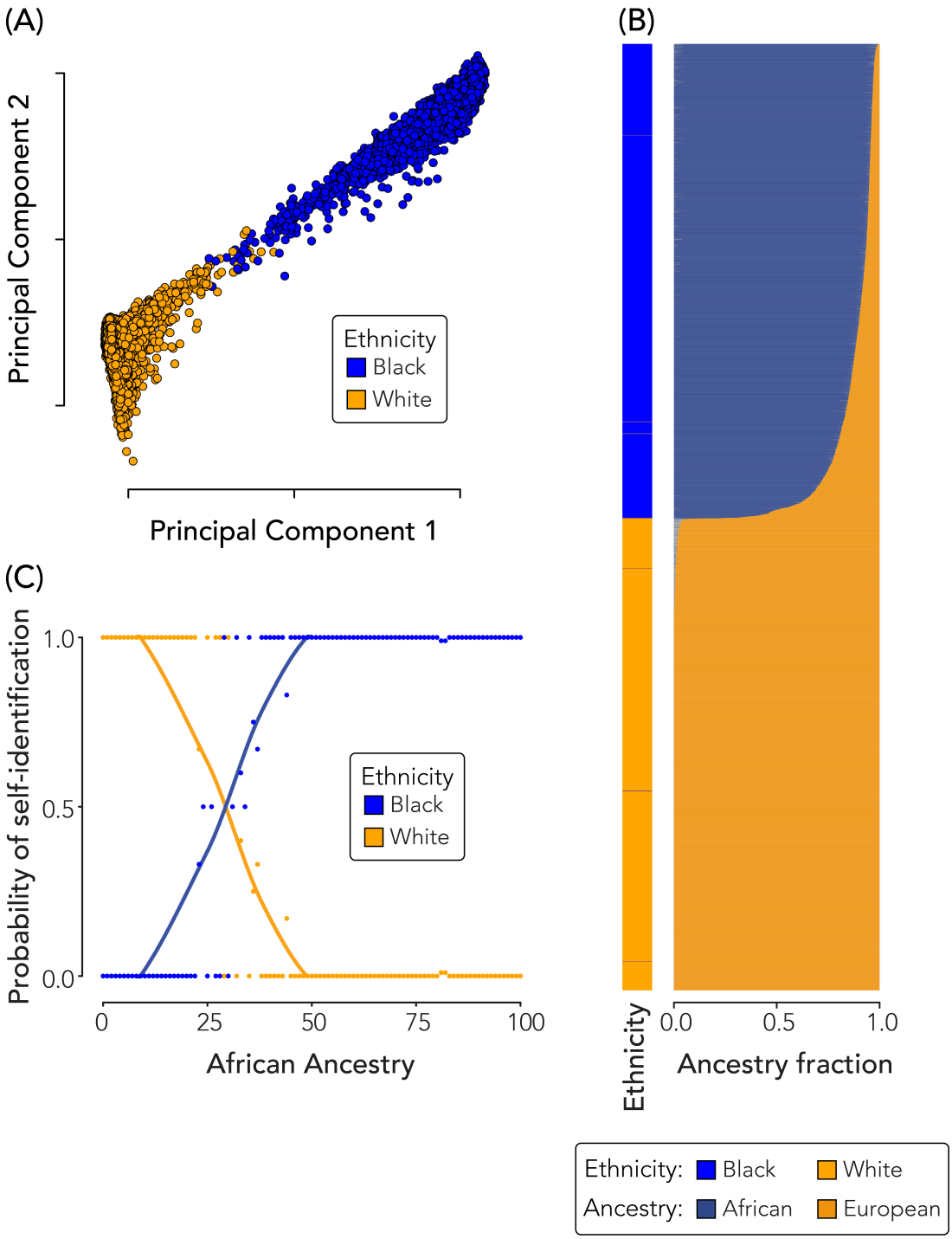


Figure 21. Ethnicity and genetic ancestry.

(A) Principal components analysis showing the relationship between participant ethnicity and genetic diversity. (B) Participant ethnicity (left) compared to participant genetic ancestry fractions (right). (C) Probability of participant ethnic self-identity (y-axis) compared to African genetic ancestry (x-axis). Dots are color-coded by participant

ethnicity and each dot shows the probability of self-identification as Black or White across one hundred bins of African ancestry. Black and White self-identification probability trend lines were fit using loess regression.

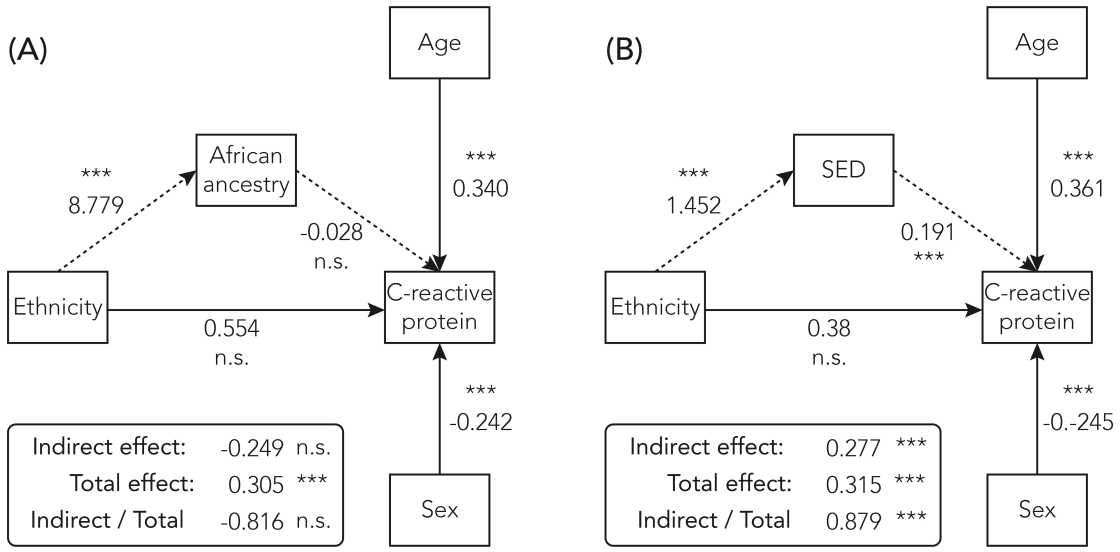


Figure 22. Mediating effects of genetic ancestry and socioeconomic deprivation on C-reactive protein ethnic differences

*Structural equation path models are shown for (A) African ancestry and (B) socioeconomic deprivation (SED). Effects of ethnicity, age, and sex on CRP serum levels are shown with solid arrows, and indirect effects of ethnicity mediated by African ancestry and SED are shown with dashed arrows. Effect sizes (β-values) and significance levels (*** = P<0.01 and n.s. = P>0.05) are shown for each modelled relationship. The indirect effects of African ancestry and SED are shown for each model along with the total effect of ethnicity and the ratio between the two.*

6.4.3 C-reactive protein and ethnic health disparities

The relationship between CRP serum levels and ethnic health disparities was evaluated by independently modeling the effect of CRP and the effect of ethnicity on disease outcomes and comparing the results. There are 109 out of 1,537 diseases analyzed

where both CRP and ethnicity showed significant associations with disease status, after correcting for multiple tests using the Bonferroni correction (Figure 23). The effect size estimates for all diseases with significant CRP and ethnicity associations were evaluated to identify diseases where differences in CRP serum levels are implicated in ethnic health disparities (Figure 24). The combined effects of CRP and ethnicity on disease outcomes were quantified by summing the ranks of the individual effect sizes. The top 20 diseases ordered via descending effect size rank sums are shown in Table 8. The top ranked diseases include examples of infectious disease (tuberculosis and HIV), metabolic diseases (type 2 diabetes and hypoglycemia), circulatory system diseases (hypertensive chronic kidney disease, hypertensive heart disease, and essential hypertension), mental disorders (schizophrenia and substance addiction), genitourinary diseases (nephrotic syndrome and chronic kidney disease), and dermatologic diseases (lupus erythematosus and sarcoidosis).

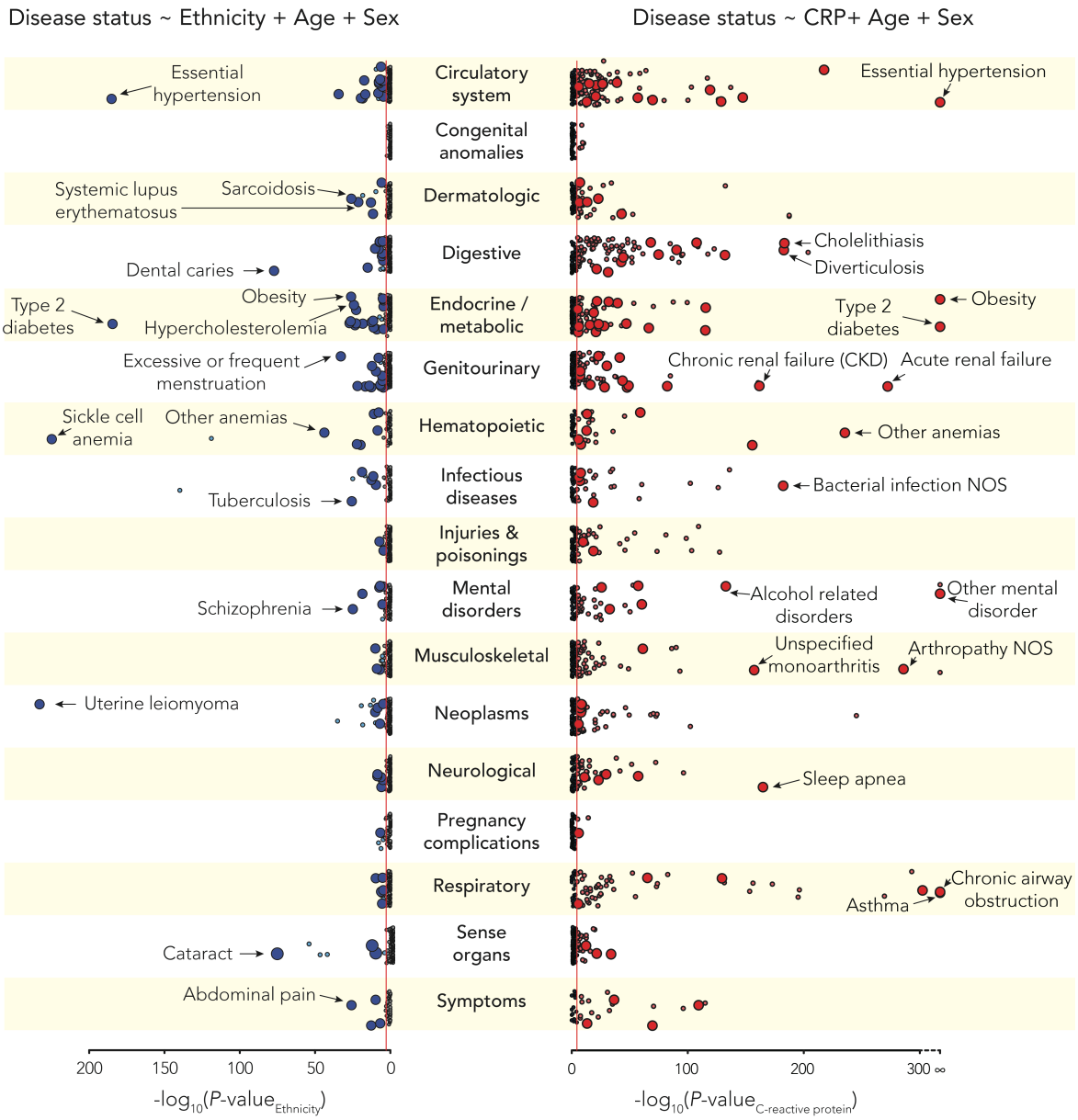


Figure 23. C-reactive protein (CRP) and ethnic health disparities.

Manhattan plots showing the statistical significance levels (-log$_{10}$P-value) for associations of ethnicity and disease outcomes (left, blue circles) and associations of CRP and disease (right, red circles). Diseases are categorized as shown using the phecode scheme. Bonferroni corrected P-value thresholds are shown with red lines. Significant associations are indicated with larger circles and select disease examples are annotated as shown.

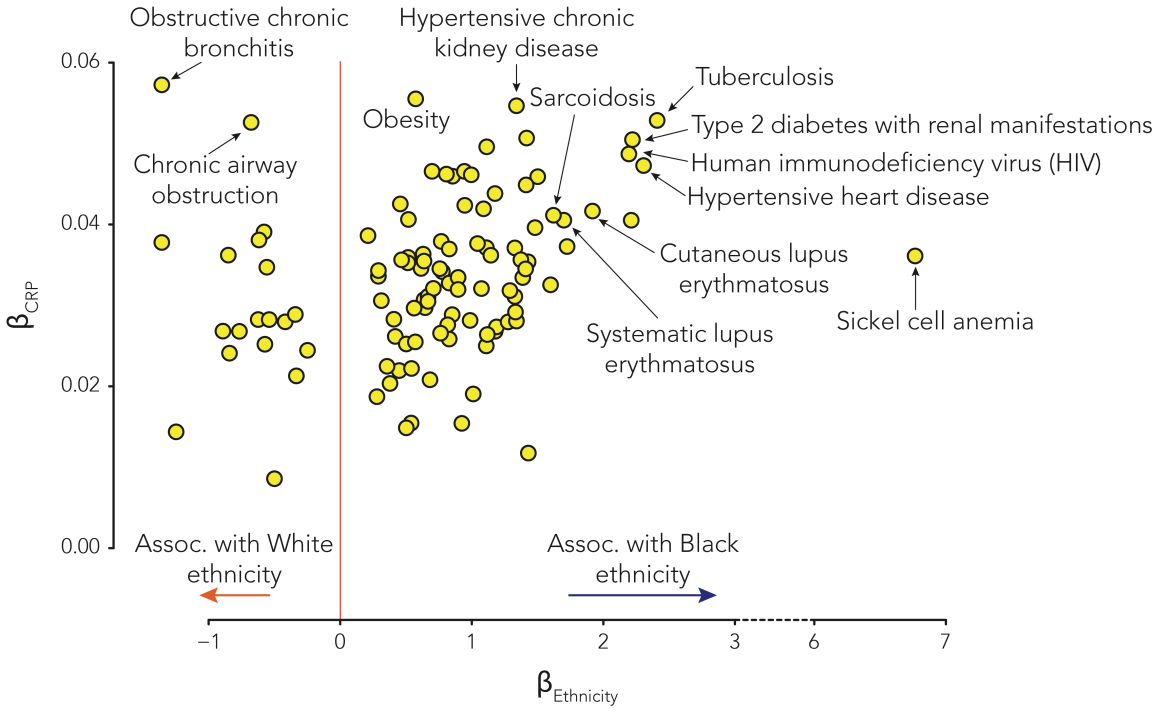


Figure 24. Effects of C-reactive protein (CRP) and ethnicity on disease.

Effect sizes for statistically significant CRP-disease associations (β_{CRP}, y-axis) and significant ethnicity-disease associations ($\beta_{Ethnicity}$, x-axis) associations. $\beta_{Ethnicity}>0$ shows diseases that are positively associated with Black ethnicity, and $\beta_{Ethnicity}<0$ shows diseases that are positively associated with White ethnicity. Select disease examples are annotated as shown.

Table 8. Top 20 diseases implicated for CRP-associated ethnic health disparities.

Phecode	Phenotype	β_{CRP}	$\beta_{Ethnicity}$	*P*-value$_{CRP}$	*P*-value$_{Ethnicity}$	Disease Category
10.0	Tuberculosis	0.0529	2.4087	3.01E-19	2.38E-26	Infectious diseases
250.22	Type 2 diabetes with renal manifestations	0.0505	2.2211	2.17E-24	1.35E-23	Endocrine / metabolic
71.0	Human immunodeficiency virus (HIV) disease	0.0487	2.1939	1.05E-06	4.17E-12	Infectious diseases
401.21	Hypertensive heart disease	0.0473	2.3056	1.16E-13	3.66E-19	Circulatory system
401.22	Hypertensive chronic kidney disease	0.0547	1.3395	1.76E-129	2.69E-20	Circulatory system
583.31	Renal dialysis	0.0507	1.4168	1.06E-49	1.12E-14	Genitourinary
295.1	Schizophrenia	0.0459	1.5011	2.05E-33	1.15E-25	Mental disorders
316.0	Substance addiction and disorders	0.0496	1.1132	2.00E-26	4.72E-08	Mental disorders
580.2	Nephrotic syndrome without mention of glomerulonephritis	0.0449	1.413	4.23E-29	2.15E-13	Genitourinary
695.41	Cutaneous lupus erythematosus	0.0417	1.9174	3.55E-07	1.27E-13	Dermatologic
250.21	Type 2 diabetes with ketoacidosis	0.0405	2.2118	7.31E-07	4.97E-19	Endocrine / Metabolic
697.0	Sarcoidosis	0.0411	1.6206	1.31E-23	1.15E-26	Dermatologic
250.2	Type 2 diabetes	0.0438	1.1764	~0	3.16E-185	Endocrine / metabolic
695.42	Systemic lupus erythematosus	0.0405	1.6996	5.28E-14	5.43E-22	Dermatologic
585.3	Chronic renal failure (CKD)	0.0461	0.9958	2.01E-162	1.48E-22	Genitourinary
580.14	Chronic glomerulonephritis	0.0466	0.9437	2.02E-48	2.22E-05	Genitourinary
250.42	Other abnormal glucose	0.0396	1.4808	2.59E-16	3.74E-13	Endocrine / Metabolic
401.1	Essential hypertension	0.046	0.8543	~0	7.15E-186	Circulatory system
251.1	Hypoglycemia	0.0419	1.0895	8.47E-48	3.80E-12	Endocrine / metabolic
585.2	Renal failure	0.0462	0.8071	5.08E-83	5.85E-07	Genitourinary

6.5 Discussion

6.5.1 Interaction between ethnicity and sex

UKBB participant CRP blood serum levels vary by ethnicity, age, and sex. Modeling CRP levels with all of these factors reveled a highly significant interaction effect between ethnicity and sex. Black females show higher CRP levels than White females, whereas Black males have lower CRP than White males. Thus, Black females are at the highest risk of chronic inflammation, suggesting the possibility of exposure to particularly high levels of stress for this group. This finding is consistent with previous studies showing that Black women can experience worse health outcomes than Black men, White women, or White men owing to their relatively subordinate position in both ethnic and gender hierarchies[233,234]. This perspective underscores the importance of an ethnic health disparities analysis framework that includes multiple, interacting demographic, genetic, and socioenvironmental factors[235-238].

6.5.2 Inflammation and ethnic health disparities

We related inflammation and ethnic health disparities by independently modeling the effect of CRP and ethnicity on disease status and then looking for diseases that showed significant associations with both factors. There were 109 out of 1,537 diseases that showed significant associations with both CRP and ethnicity, and we explored the diseases that showed the strongest effects for both. This approach uncovered a number of diseases linked to immune response and inflammation, including infectious diseases and complex,

common diseases. This suggests the possibility that ethnic differences in inflammation, related to environmental exposures and psychosocial stress, could be broadly related to ethnic health disparities.

6.5.3 Caveats and limitations

It is important to note, however, that our observational study design and statistical modelling do not allow for unambiguous causal inference regarding the relationship between CRP and disease[239,240]. For infectious diseases, CRP levels are expected to be elevated after infection, which would entail a kind of reverse causality with respect to how our regression models are specified. For chronic diseases, systemic inflammation could precede disease or contribute to disease progression, but it could also reflect the presence of disease. Our models cannot distinguish between these possibilities, and it is not known whether participant CRP levels measured at recruitment precede or follow the diagnosis and course of disease. Thus, it is possible that the observed ethnic differences in CRP reflect a higher overall burden of disease for ethnic minorities in the UKBB, linked to higher levels of socioeconomic deprivation, rather than a causal risk factor for ethnic health disparities.

6.6 Conclusion

C-reactive protein (CRP) is a widely used clinical marker of inflammation. Our study of the UKBB found that CRP blood serum levels differ according to participants'

self-identification as belonging to Black or White ethnic groups, and ethnicity is highly correlated with genetic ancestry. Given these results, it could be expected that genetic ancestry would mediate ethnic differences in CRP levels, thereby pointing to a potential role for genetic factors in the observed disparity. However, we found that socioeconomic deprivation, and not genetic ancestry, mediates the observed ethnic differences in CRP levels. This indicates that the environment plays a more important role than genetics in shaping ethnic disparities in inflammation for this cohort. Possible environmental factors leading to higher levels of CRP observed for Black participants could include psychosocial stress linked to racial discrimination and poverty[241-244]. Aspects of diet and lifestyle associated with socioeconomic deprivation could also be linked to ethnic differences in inflammation[245-247].

CHAPTER 7. CONCLUSIONS AND NEXT STEPS

Despite the fact that genes and environmental factors work together to affect health outcomes, health disparity research programs tend to be siloed with a narrow focus on either genetic or environmental contributions to health differences among groups. There exists a tremendous and unmet opportunity to overcome these biases via a more inclusive research design that is focused on the diverse genomic cohorts, many of which are overlooked in existing biobanks, and addresses the role of genetic and the environment together in shaping health outcomes. Utilizing any and all approaches to understand health disparities is the first step towards health equity.

This thesis first focuses on disparities in drug response, which are more directly explained by genetics, and then moves to more complex disparities that result from the interplay of genetics and the environment. In chapters 2 and 3, the thesis illustrates the role that genetic ancestry can play in improving therapeutic outcomes for diverse populations in two very different countries in the New World (Colombia and the US). The inclusion of information on ancestral origins can improve outcomes for different population groups, especially those who are underrepresented in research. Starting with chapter 4, the focus becomes broader and shifts to more complex traits. In chapters 5 and 6, the thesis demonstrates that using information on ancestral origins in addition to environmental factors helps us derive insights about different health disparities.

The thesis also documents the development of a free and interactive online resource for identifying health disparities in the UK Biobank, a routinely used dataset for genomics and

epidemiology research. This will allow researchers to identify areas of research which might have the most impact in alleviating health disparities.

An underlying theme of this thesis is that we can learn more about disparities and health outcomes if genomics research is more diverse. The thesis presents different studies which involve participants who are diverse in different ways – genetic ancestry, racial/ethnic self-identification, and socioeconomic context. The studies discussed in chapter 2 and 3 highlight how genomic insights derived from one population group may not always be transferable to other populations; chapters 5 and 6 illustrate how genetics and the environment seem to interact to cause dissimilar outcomes in different groups. It should be noted that for chapters 4, 5, and 6, the thesis leveraged data from diverse participants characterized as part of populations biobanks, which were already available to researchers but often overlooked.

Future research that leverages information about ancestral origins – race, ethnicity, and genetic ancestry – while incorporating more detailed, "deep phenotyping" data on environment and lifestyle from diverse populations will be key in a societal move towards health equity. New biobank initiatives like the All of Us Research Program are already incorporating data from fitness trackers, but other more granular data like macronutrient intake can be extremely useful in understanding the complex disease burden on minority populations and then in devising alleviation strategies.

Until we reach a point where everyone in society has equitable access to basic resources like healthcare, healthy food, and safe open spaces for exercise, such research that tries to understand specific risk factors – lifestyle, environmental, and biological features – that

have a disproportionate impact on health outcomes can help alleviate some of the burden of health disparities we observe today.

APPENDIX A.

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

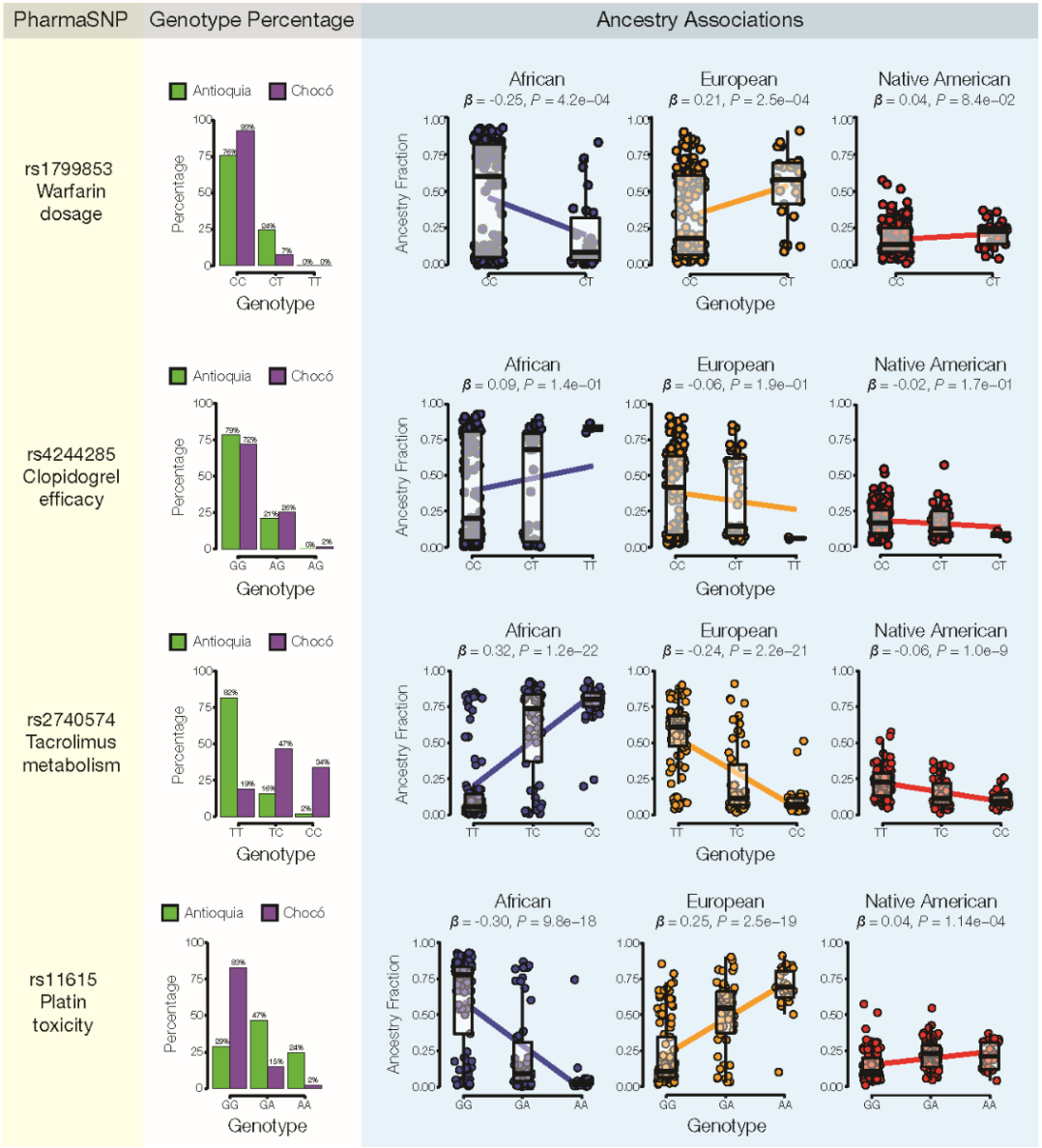
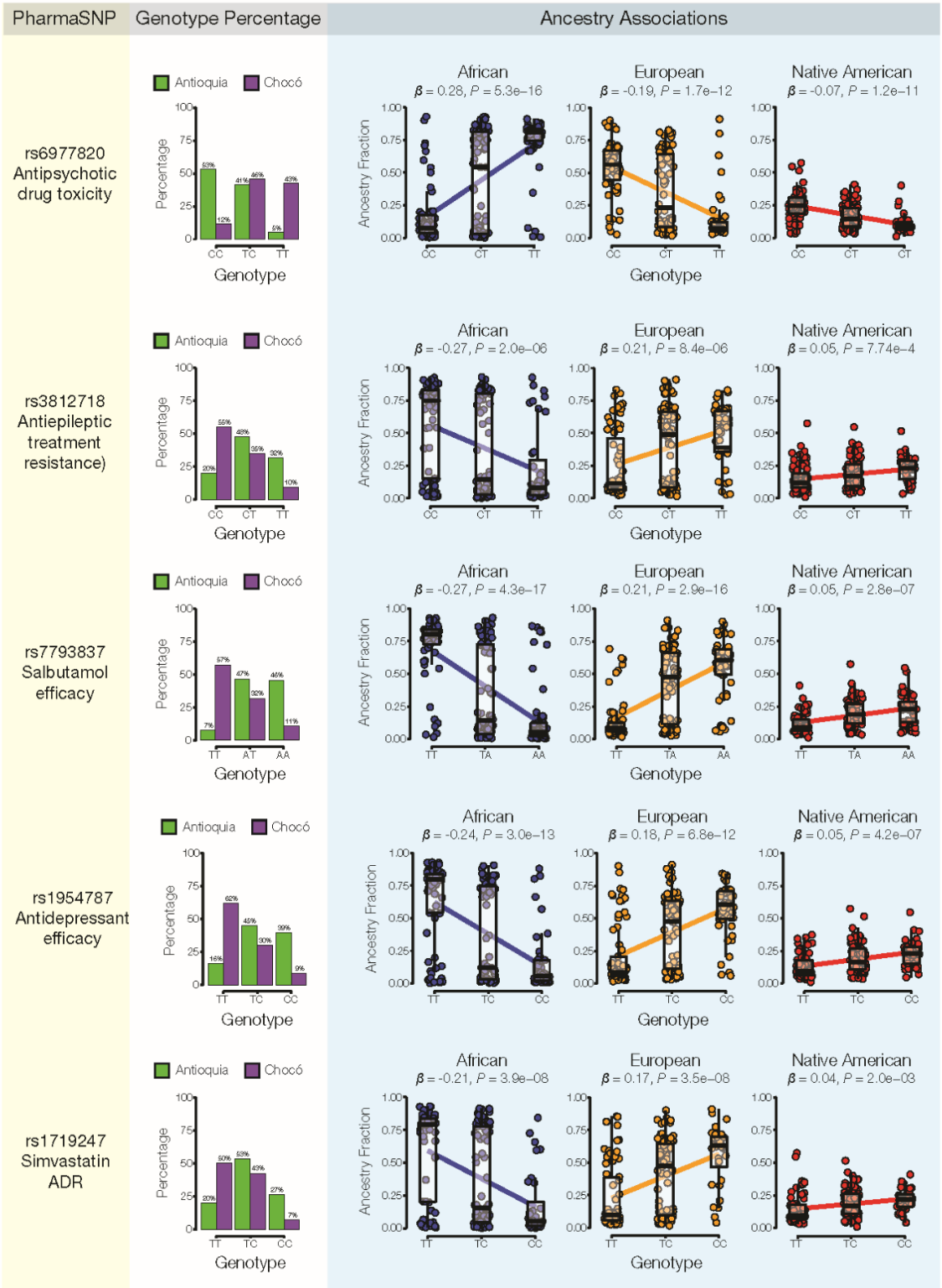


Figure 25. Ancestry associations for PGx variants in Colombia.

[Continued on next page] For each panel in the figure, PGx variant genotype percentages are shown for Antioquia (green) and Chocó (purple) followed by the ancestry association plots. For each genetic ancestry component – African (blue), European (orange), and

Native American (red) – individuals' ancestry fractions (y-axis) are regressed against their PGx variant genotypes (x-axis). Ancestry associations are quantified by the slope of the regression (β) and its significance level (P).



rs4149056 (SLCO1B1*5)

Homozygous Non-effect Allele TT

		Exome		
		YES	**NO**	
PCR	**YES**	83	0	83
	NO	3	46	49
		86	46	132

Sensitivity	96.5	95% CI	(93.4, 99.6)
Specificity	100	95% CI	(100, 100)
PPV	100	95% CI	(100, 100)
NPV	93.9	95% CI	(89.8, 98.0)

Heterozygous TC

		Exome		
		YES	**NO**	
PCR	**YES**	44	1	45
	NO	0	87	87
		44	88	132

Sensitivity	100	95% CI	(100, 100)
Specificity	98.9	95% CI	(97.1, 100)
PPV	97.8	95% CI	(95.3, 100)
NPV	100	95% CI	(100, 100)

Homozygous Effect Allele CC

		Exome		
		YES	**NO**	
PCR	**YES**	2	2	4
	NO	0	128	128
		2	130	132

Sensitivity	100	95% CI	(100, 100)
Specificity	98.5	95% CI	(96.4, 100)
PPV	50	95% CI	(41.5, 58.5)
NPV	100	95% CI	(100, 100)

rs1799583 (CYP2C9*2)

Homozygous Non-effect Allele CC

		Exome		
		YES	**NO**	
PCR	**YES**	113	1	114
	NO	1	17	18
		114	18	132

Sensitivity	99.1	95% CI	(97.5, 100)
Specificity	94.4	95% CI	(90.5, 98.4)
PPV	99.1	95% CI	(97.5, 100)
NPV	94.4	95% CI	(90.5, 98.4)

Heterozygous CT

		Exome		
		YES	**NO**	
PCR	**YES**	17	1	18
	NO	1	113	114
		18	114	132

Sensitivity	94.4	95% CI	(90.5, 98.4)
Specificity	99.1	95% CI	(97.5, 100)
PPV	94.4	95% CI	(90.5, 98.4)
NPV	99.1	95% CI	(97.5, 100)

rs1057910 (CYP2C9*3)

Homozygous Non-Effect Allele AA

		Exome		
		YES	**NO**	
PCR	**YES**	115	0	115
	NO	0	17	17
		115	17	132

Sensitivity	100	95% CI	(100, 100)
Specificity	100	95% CI	(100, 100)
PPV	100	95% CI	(100, 100)
NPV	100	95% CI	(100, 100)

Heterozygous AC

		Exome		
		YES	**NO**	
PCR	**YES**	17	0	17
	NO	0	115	115
		17	115	132

Sensitivity	100	95% CI	(100, 100)
Specificity	100	95% CI	(100, 100)
PPV	100	95% CI	(100, 100)
NPV	100	95% CI	(100, 100)

Figure 26. Comparison of the allele-specific PCR PGx variant genotyping assay results and the exome sequencing results.

Three PGx variants were genotyped in a 132 patient cohort from the GenomaCES laboratory in Medellín, Colombia using the custom allele-specific PCR assay described in the manuscript. Each individual PCR assay was validated via comparison with the exome sequence analysis results for these same patients. Taking the exome sequences as the ground truth for the presence of the PGx variant alleles in these patients, PCR results were scored as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), and the following metrics were computed to validate the PCR genotyping assays for each individual genotype assayed:

$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = TN/(TN + FP)$$

$$Positive\ predictive\ value\ (PPV) = \frac{TP}{TP + FP}$$
$$Negative\ predictive\ value\ (NPV) = TN/(TN + FN)$$

The 95% confidence intervals for these metrics were computed as:
$$x \pm 1.96 * \sqrt{(x * (1 - x))/n}$$

where x is the value of the metric and n is the total number of genotype assays conducted.

Table 9. PGx variant effect allele frequencies and ancestry associations for Colombian populations.

			African Ancestry Associations		European Ancestry Associations		Native American Ancestry Associations		Effect allele frequencies	
rsID	**Effect**	**Level of evidence**	**β Value**	***P* Value**	**β Value**	***P* Value**	**β Value**	***P* Value**	**Antioquia**	**Chocó**
rs776746	Tacrolimus Dosage	1A	0.31	2.7E-24	-0.24	2.9E-23	-0.06	1.2E-09	0.19	0.68
rs1057910	Warfarin Toxicity/ADR	1A	-0.32	1.3E-03	0.30	1.3E-04	0.01	6.7E-01	0.07	0.01
rs9923231	Warfarin Dosage	1A	0.25	8.0E-12	-0.19	7.3E-12	-0.04	9.6E-05	0.57	0.88
rs4149056	Simvastatin Toxicity/ADR	1A	-0.22	9.0E-05	0.18	5.5E-05	0.03	4.3E-02	0.18	0.05
rs3892097	Antidepressant Dosage	1A	0.23	7.7E-05	-0.19	3.1E-05	-0.03	5.4E-02	0.84	0.96
rs1799853	Warfarin Dosage	1A	0.25	4.2E-04	-0.21	2.5E-04	-0.04	8.4E-02	0.88	0.97
rs887829	Atazanavir Toxicity/ADR	1A	0.07	8.1E-02	-0.05	7.8E-02	-0.01	3.0E-01	0.34	0.41
rs4244285	Clopidogrel Toxicity/ADR	1A	0.09	1.4E-01	-0.06	1.9E-01	-0.02	1.7E-01	0.10	0.16
rs9923231	Anticoagulant Dosage	1B	0.25	8.0E-12	-0.19	7.3E-12	-0.04	9.6E-05	0.57	0.88
rs9934438	Warfarin Dosage	1B	0.25	8.0E-12	-0.19	7.3E-12	-0.04	9.6E-05	0.57	0.88
rs2108622	Warfarin Dosage	1B	-0.23	5.6E-07	0.18	5.0E-07	0.04	4.0E-03	0.27	0.07
rs8099917	Antiviral Efficacy	1B	0.17	1.1E-04	-0.13	1.2E-04	-0.03	2.5E-02	0.73	0.91
rs1800497	Bupropion Efficacy	1B	0.12	2.9E-03	-0.11	7.9E-04	-0.01	4.0E-01	0.20	0.33
rs116855232	Azathioprine, Mercaptopurine Toxicity/ADR	1B	-0.14	3.5E-01	0.08	4.9E-01	0.05	3.1E-01	0.02	0.01
rs7294	Warfarin Dosage	1B	0.06	9.7E-02	-0.05	1.1E-01	-0.01	2.2E-01	0.36	0.43
rs3745274	Efavirenz Dosage	1B	-0.03	4.2E-01	0.01	6.2E-01	0.01	2.2E-01	0.64	0.57
rs2740574	Tacrolimus Dosage	2A	0.32	4.1E-25	-0.24	2.8E-23	-0.06	9.2E-11	0.10	0.60
rs7900194	Warfarin Toxicity/ADR	2A	0.44	8.0E-04	-0.32	2.0E-03	-0.10	8.2E-03	0.00	0.05
rs776746	Sirolimus Dosage	2A	0.31	2.7E-24	-0.24	2.9E-23	-0.06	1.2E-09	0.19	0.68
rs1801133	Cyclophosphamide Efficacy	2A	-0.27	9.5E-16	0.21	5.1E-16	0.05	7.5E-06	0.54	0.15
rs1057910	Acenocoumarol Toxicity/ADR	2A	-0.32	1.3E-03	0.30	1.3E-04	0.01	6.7E-01	0.07	0.01

rs9934438	Acenocoumarol, phenprocoumon Dosage	2A	0.25	8.0E-12	-0.19	7.3E-12	-0.04	9.6E-05	0.57	0.88
rs2108622	Phenprocoumon Dosage	2A	-0.23	5.6E-07	0.18	5.0E-07	0.04	4.0E-03	0.27	0.07
rs2228570	Peginterferon alfa-2B, Ribavirin Efficacy	2A	-0.17	9.3E-06	0.12	1.0E-04	0.05	9.2E-05	0.42	0.21
rs2359612	Warfarin Dosage	2A	0.18	6.2E-07	-0.14	6.6E-07	-0.03	3.5E-03	0.56	0.78
rs4149056	Cerivastatin Toxicity/ADR	2A	-0.22	9.0E-05	0.18	5.5E-05	0.03	4.3E-02	0.18	0.05
rs1045642	Methotrexate Toxicity/ADR	2A	-0.16	1.8E-05	0.12	6.9E-05	0.04	5.7E-04	0.45	0.26
rs4148323	Irinotecan Toxicity/ADR	2A	-0.23	1.3E-01	0.17	1.5E-01	0.04	3.4E-01	0.03	0.01
rs10509681	Rosiglitazone Metabolism/PK	2A	-0.21	2.3E-03	0.18	8.8E-04	0.02	2.6E-01	0.12	0.03
rs8050894	Warfarin Dosage	2A	0.16	3.4E-05	-0.12	2.3E-05	-0.03	2.3E-02	0.55	0.75
rs1041983	Ethambutol, Isoniazid, Pyrazinamide, Rifampin Toxicity/ADR	2A	0.13	4.9E-04	-0.10	5.4E-04	-0.02	3.6E-02	0.32	0.50
rs17708472	Warfarin Dosage	2A	-0.20	4.6E-04	0.13	3.6E-03	0.06	2.7E-04	0.18	0.07
rs2884737	Warfarin Dosage	2A	0.24	6.6E-06	-0.19	4.3E-06	-0.04	1.4E-02	0.81	0.96
rs1799978	Risperidone Efficacy	2A	-0.20	3.1E-04	0.14	1.3E-03	0.05	2.2E-03	0.94	0.82
rs1695	Cyclophosphamide, Epirubicin Toxicity/ADR	2A	0.10	8.4E-03	-0.08	1.3E-02	-0.02	6.2E-02	0.36	0.48
rs3892097	Tamoxifen Dosage	2A	0.23	7.7E-05	-0.19	3.1E-05	-0.03	5.4E-02	0.84	0.96
rs7294	Acenocoumarol, Phenprocoumon Dosage	2A	0.06	9.7E-02	-0.05	1.1E-01	-0.01	2.2E-01	0.36	0.43
rs1800566	Alkylatingagents, Anthracyclines and related substances, Fluorouracil, Platinum compounds Efficacy	2A	0.10	3.3E-02	-0.07	4.2E-02	-0.02	1.3E-01	0.72	0.80
rs28399499	Efavirenz Metabolism/PK	2A	-0.26	2.3E-03	0.21	1.5E-03	0.04	1.4E-01	0.99	0.92
rs3745274	Methadone Dosage	2A	-0.03	4.2E-01	0.01	6.2E-01	0.01	2.2E-01	0.64	0.57
rs7412	Atorvastatin Efficacy	2A	0.06	3.7E-01	-0.05	3.5E-01	-0.01	6.7E-01	0.09	0.13
rs4680	Nicotine Efficacy	2A	-0.06	1.5E-01	0.06	4.5E-02	-0.01	5.9E-01	0.37	0.31

rs17244841	HMG-CoA reductase inhibitors, Pravastatin, Simvastatin Efficacy	2A	-0.16	2.0E-02	0.11	4.3E-02	0.05	3.0E-02	0.95	0.90
rs2279345	Efavirenz Metabolism/PK	2A	0.06	2.4E-01	-0.05	2.0E-01	-0.01	6.3E-01	0.80	0.85
rs4149015	Pravastatin Efficacy	2A	0.23	1.7E-02	-0.21	5.5E-03	-0.02	5.9E-01	0.93	0.98
rs776746	Cyclosporine Dosage	2B	0.31	2.7E-24	-0.24	2.9E-23	-0.06	1.2E-09	0.19	0.68
rs11615	Carboplatin, Cisplatin, Oxaliplatin, Platinum, Platinum compounds Toxicity/ADR	2B	-0.30	2.3E-19	0.25	4.1E-21	0.05	3.0E-05	0.49	0.07
rs7793837	Salbutamol, selectivebeta-2-adrenoreceptoragonists Efficacy	2B	-0.27	1.1E-18	0.21	8.5E-18	0.05	4.0E-08	0.69	0.25
rs20455	Atorvastatin Efficacy	2B	0.29	7.7E-20	-0.21	3.6E-17	-0.07	3.6E-11	0.37	0.80
rs1954787	Antidepressants Efficacy	2B	-0.24	1.6E-14	0.18	4.0E-13	0.05	6.3E-08	0.62	0.21
rs11212617	Metformin Efficacy	2B	0.28	1.9E-17	-0.21	3.5E-16	-0.06	3.8E-08	0.33	0.74
rs7997012	Antidepressants, citalopram, selective serotonin reuptake inhibitors Efficacy	2B	-0.27	3.2E-11	0.20	2.0E-09	0.07	1.6E-07	0.35	0.07
rs489693	Amisulpride, Aripiprazole, Clozapine, Haloperidol, Olanzapine, Paliperidone, Quetiapine, Risperidone, Ziprasidone Toxicity/ADR	2B	0.24	1.8E-11	-0.18	5.2E-10	-0.06	6.4E-07	0.18	0.50
rs339097	Warfarin Dosage	2B	0.33	2.0E-06	-0.26	3.2E-06	-0.06	3.0E-03	0.02	0.14
rs6988229	Salbutamol Efficacy	2B	0.22	5.7E-10	-0.17	6.0E-09	-0.05	3.4E-06	0.19	0.49
rs1042522	Antineoplastic agents, Cisplatin, Cyclophosphamide, Fluorouracil, Paclitaxel Toxicity/ADR	2B	0.23	2.9E-10	-0.16	9.0E-09	-0.05	7.8E-07	0.28	0.59
rs4713916	Antidepressants, Citalopram, Fluoxetine,	2B	-0.29	2.9E-09	0.23	1.3E-09	0.05	1.3E-03	0.27	0.06

	Mirtazapine, Paroxetine, Selective serotonin reuptake inhibitors, Venlafaxine Efficacy									
rs3812718	Carbamazepine Dosage	2B	-0.18	3.8E-07	0.13	1.8E-06	0.04	2.9E-04	0.55	0.27
rs1051730	Nicotine Toxicity/ADR	2B	-0.20	7.4E-07	0.17	6.4E-08	0.02	7.0E-02	0.36	0.12
rs1719247	HMG-CoA reductase inhibitors, Simvastatin Toxicity/ADR	2B	-0.21	4.9E-09	0.17	4.2E-09	0.04	8.9E-04	0.54	0.27
rs924607	Vincristine Toxicity/ADR	2B	-0.24	7.5E-08	0.18	2.5E-07	0.05	2.1E-04	0.33	0.10
rs2108622	Acenocoumarol Dosage	2B	-0.23	5.6E-07	0.18	5.0E-07	0.04	4.0E-03	0.27	0.07
rs4444903	Cetuximab Efficacy	2B	0.18	3.3E-07	-0.14	2.2E-07	-0.03	3.8E-03	0.50	0.76
rs25487	Carboplatin, Cisplatin, Oxaliplatin, Platinum, Platinum compounds Efficacy	2B	0.25	5.2E-09	-0.19	1.9E-08	-0.05	9.9E-05	0.63	0.87
rs2232228	Anthracyclines Toxicity/ADR	2B	0.20	2.8E-07	-0.16	1.5E-07	-0.03	4.9E-03	0.64	0.88
rs7779029	Irinotecan Toxicity/ADR	2B	0.24	2.4E-06	-0.17	2.0E-05	-0.06	9.6E-05	0.09	0.26
rs1933437	Sunitinib Toxicity/ADR	2B	-0.14	1.2E-04	0.12	2.0E-05	0.01	2.1E-01	0.57	0.36
rs3753380	Latanoprost Efficacy	2B	0.25	8.5E-08	-0.18	6.2E-07	-0.06	5.1E-05	0.73	0.94
rs2231142	Allopurinol Dosage	2B	-0.25	2.9E-04	0.15	4.9E-03	0.08	3.9E-05	0.14	0.04
rs17782313	Antipsychotics Toxicity/ADR	2B	0.20	2.1E-05	-0.15	6.9E-05	-0.05	1.4E-03	0.11	0.27
rs2298383	Caffeine Toxicity/ADR	2B	0.16	6.5E-06	-0.11	7.2E-05	-0.04	5.7E-05	0.50	0.69
rs16960228	Hydrochlorothiazide Efficacy	2B	0.17	1.2E-04	-0.13	3.7E-04	-0.04	2.2E-03	0.14	0.31
rs738409	Asparaginase, Cyclophosphamide, Daunorubicin, Prednisolone, Vincristine Toxicity/ADR	2B	-0.17	3.7E-05	0.13	7.5E-05	0.03	6.9E-03	0.40	0.22
rs1799971	Ethanol Metabolism/PK	2B	-0.22	2.3E-04	0.17	2.5E-04	0.04	2.3E-02	0.16	0.05

rs1532624	HMG-CoA Efficacy	2B	0.19	8.4E-06	-0.15	1.3E-05	-0.04	3.5E-03	0.69	0.87
rs2234922	Carbamazepine Dosage	2B	0.14	5.0E-04	-0.10	2.0E-03	-0.04	3.0E-03	0.18	0.33
rs1876828	Budesonide, Corticosteroids, Fluticasonepropionate, Fluticasone/Salmeterol, Triamcinolone Efficacy	2B	0.38	5.4E-10	-0.32	1.8E-11	-0.05	1.6E-02	0.82	0.99
rs578776	Nicotine Toxicity/ADR	2B	-0.11	3.4E-03	0.10	3.2E-04	0.00	1.0E+00	0.52	0.36
rs730012	Aspirin Toxicity/ADR	2B	-0.16	9.4E-04	0.11	2.5E-03	0.04	6.7E-03	0.24	0.11
rs11598702	Gemcitabine Metabolism/PK	2B	-0.11	4.8E-03	0.09	4.8E-03	0.02	8.6E-02	0.36	0.21
rs4803419	Efavirenz Metabolism/PK	2B	0.11	8.8E-03	-0.07	2.7E-02	-0.03	1.1E-02	0.67	0.82
rs7297610	Hydrochlorothiazide Efficacy	2B	-0.20	3.3E-05	0.14	1.1E-04	0.05	1.7E-03	0.91	0.76
rs746647	Nevirapine Toxicity/ADR	2B	-0.15	9.6E-04	0.11	2.0E-03	0.03	1.3E-02	0.30	0.17
rs1800497	Ethanol Efficacy	2B	0.12	2.9E-03	-0.11	7.9E-04	-0.01	4.0E-01	0.20	0.33
rs1800497	Antipsychotics, Clozapine, Olanzapine, Risperidone Efficacy	2B	0.12	2.9E-03	-0.11	7.9E-04	-0.01	4.0E-01	0.20	0.33
rs264588	Radiotherapy Toxicity/ADR	2B	0.12	4.7E-03	-0.10	2.4E-03	-0.02	2.3E-01	0.19	0.31
rs7582141	Radiotherapy Toxicity/ADR	2B	0.10	4.0E-02	-0.09	2.6E-02	-0.01	4.1E-01	0.19	0.30
rs1056892	Anthracyclines Toxicity/ADR	2B	-0.08	2.5E-02	0.07	2.4E-02	0.01	2.1E-01	0.72	0.60
rs10306114	Aspirin Efficacy	2B	0.15	3.3E-02	-0.14	1.2E-02	-0.01	7.8E-01	0.05	0.11
rs1695	Cisplatin Toxicity/ADR	2B	0.10	8.4E-03	-0.08	1.3E-02	-0.02	6.2E-02	0.36	0.48
rs6432512	Radiotherapy Toxicity/ADR	2B	0.12	5.1E-03	-0.10	2.8E-03	-0.02	2.1E-01	0.20	0.30
rs12777823	Warfarin Dosage	2B	-0.17	1.1E-03	0.12	3.3E-03	0.04	5.6E-03	0.90	0.78
rs1051740	Carbamazepine Dosage	2B	-0.11	1.3E-02	0.07	4.4E-02	0.04	6.7E-03	0.31	0.20
rs13306278	Selective Efficacy	2B	0.38	1.8E-06	-0.33	1.1E-07	-0.04	1.1E-01	0.89	1.00
rs1801274	Trastuzumab Efficacy	2B	-0.06	1.0E-01	0.05	1.0E-01	0.01	2.8E-01	0.60	0.50
rs4961	Furosemide, Spironolactone Efficacy	2B	0.19	1.9E-03	-0.15	1.4E-03	-0.03	9.7E-02	0.83	0.93

rs7853758	Anthracyclinesandrelatedsubstances Toxicity/ADR	2B	-0.11	1.1E-02	0.08	1.4E-02	0.02	9.5E-02	0.78	0.68
rs2072661	Nicotine Toxicity/ADR	2B	0.08	8.4E-02	-0.08	3.8E-02	0.00	9.1E-01	0.18	0.26
rs716274	Etoposide, Platinum compounds Toxicity/ADR	2B	-0.05	1.7E-01	0.07	3.0E-02	-0.01	2.9E-01	0.46	0.37
rs2952768	Buprenorphine, Fentanyl, Meperidine, Morphine, Opioids, Pentazocine Dosage	2B	-0.08	5.9E-02	0.05	1.3E-01	0.03	3.8E-02	0.35	0.27
rs885004	Anthracyclines and related substances Toxicity/ADR	2B	0.13	2.3E-02	-0.09	4.2E-02	-0.03	4.1E-02	0.83	0.91
rs7270101	Peginterferon alfa-2B, Ribavirin Toxicity/ADR	2B	-0.13	9.9E-02	0.12	6.1E-02	0.01	6.7E-01	0.08	0.05
rs1517114	Irinotecan Toxicity/ADR	2B	0.06	9.3E-02	-0.05	6.3E-02	-0.01	5.4E-01	0.36	0.44
rs1076560	Cocaine Toxicity/ADR	2B	-0.06	2.5E-01	0.03	4.5E-01	0.03	8.7E-02	0.15	0.10
rs1801394	Methotrexate Toxicity/ADR	2B	-0.06	1.7E-01	0.06	6.2E-02	-0.01	6.4E-01	0.28	0.22
rs1801019	Capecitabine, Fluorouracil, Leucovorin, Tegafur Toxicity/ADR	2B	0.06	1.5E-01	-0.06	7.4E-02	0.00	9.6E-01	0.21	0.28
rs4693075	Atorvastatin, HMG-CoA reductase inhibitors, Rosuvastatin Toxicity/ADR	2B	0.06	1.4E-01	-0.04	2.3E-01	-0.02	1.2E-01	0.36	0.43
rs4880	Cyclophosphamide Efficacy	2B	0.05	1.8E-01	-0.04	1.6E-01	-0.01	5.5E-01	0.52	0.58
rs6065	Aspirin Efficacy	2B	0.06	2.6E-01	-0.04	3.1E-01	-0.01	3.4E-01	0.15	0.20
rs4149601	Diuretics, Hydrochlorothiazide Efficacy	2B	-0.08	5.1E-02	0.05	1.5E-01	0.03	1.2E-02	0.73	0.67
rs1346268	HMG-CoA reductase inhibitors, Simvastatin Toxicity/ADR	2B	0.05	2.4E-01	-0.02	5.3E-01	-0.03	2.7E-02	0.58	0.63
rs10455872	HMG-CoA Efficacy	2B	0.28	5.3E-03	-0.23	3.1E-03	-0.04	2.2E-01	0.94	1.00
rs61750900	Nicotine Metabolism/PK	2B	0.16	6.6E-02	-0.16	1.9E-02	0.00	8.6E-01	0.93	0.98

Table 10. PGx variant effect allele frequencies and ancestry associations for Colombian populations.

		Frequency Differences			
rsID	**Effect**	**$\log_2$(Antioquia/ Chocó)**	**Δ(Antioquia-Chocó)**	**Euclidean Distance from Origin**	**Effect Allele**
rs776746	Tacrolimus Dosage	-1.86	-0.49	5.26	T
rs1057910	Warfarin Toxicity/ADR	3.37	0.06	3.43	C
rs9923231	Warfarin Dosage	-0.62	-0.31	3.12	C
rs4149056	Simvastatin Toxicity/ADR	1.79	0.13	2.18	C
rs3892097	Antidepressant Dosage	-0.19	-0.12	1.22	C
rs1799853	Warfarin Dosage	-0.15	-0.10	0.97	C
rs887829	Atazanavir Toxicity/ADR	-0.29	-0.08	0.82	T
rs4244285	Clopidogrel Toxicity/ADR	-0.59	-0.05	0.79	A
rs9923231	Anticoagulant Dosage	-0.62	-0.31	3.12	C
rs9934438	Warfarin Dosage	-0.62	-0.31	3.12	G
rs2108622	Warfarin Dosage	1.90	0.20	2.77	T
rs8099917	Antiviral Efficacy	-0.31	-0.17	1.76	T
rs1800497	Bupropion Efficacy	-0.70	-0.13	1.45	A
rs116855232	Azathioprine, Mercaptopurine Toxicity/ADR	0.84	0.01	0.84	T
rs7294	Warfarin Dosage	-0.27	-0.07	0.77	T
rs3745274	Efavirenz Dosage	0.15	0.06	0.65	G
rs2740574	Tacrolimus Dosage	-2.59	-0.50	5.60	C
rs7900194	Warfarin Toxicity/ADR	-5.51	-0.04	5.52	A
rs776746	Sirolimus Dosage	-1.86	-0.49	5.26	T
rs1801133	Cyclophosphamide Efficacy	1.88	0.39	4.35	A

rs1057910	Acenocoumarol Toxicity/ADR	3.37	0.06	3.43	C
rs9934438	Acenocoumarol, phenprocoumon Dosage	-0.62	-0.31	3.12	G
rs2108622	Phenprocoumon Dosage	1.90	0.20	2.77	T
rs2228570	Peginterferon alfa-2B, Ribavirin Efficacy	1.02	0.21	2.35	A
rs2359612	Warfarin Dosage	-0.48	-0.22	2.26	G
rs4149056	Cerivastatin Toxicity/ADR	1.79	0.13	2.18	C
rs1045642	Methotrexate Toxicity/ADR	0.83	0.20	2.15	A
rs4148323	Irinotecan Toxicity/ADR	2.03	0.02	2.04	A
rs10509681	Rosiglitazone Metabolism/PK	1.80	0.09	1.99	C
rs8050894	Warfarin Dosage	-0.43	-0.19	1.98	C
rs1041983	Ethambutol, Isoniazid, Pyrazinamide, Rifampin Toxicity/ADR	-0.63	-0.17	1.86	T
rs17708472	Warfarin Dosage	1.38	0.11	1.76	A
rs2884737	Warfarin Dosage	-0.25	-0.15	1.54	A
rs1799978	Risperidone Efficacy	0.21	0.13	1.28	T
rs1695	Cyclophosphamide, Epirubicin Toxicity/ADR	-0.40	-0.12	1.24	G
rs3892097	Tamoxifen Dosage	-0.19	-0.12	1.22	C
rs7294	Acenocoumarol, Phenprocoumon Dosage	-0.27	-0.07	0.77	T
rs1800566	Alkylatingagents, Anthracyclines and related substances, Fluorouracil, Platinum compounds Efficacy	-0.14	-0.07	0.74	G
rs28399499	Efavirenz Metabolism/PK	0.11	0.07	0.74	T

rs3745274	Methadone Dosage	0.15	0.06	0.65	G
rs7412	Atorvastatin Efficacy	-0.51	-0.04	0.63	T
rs4680	Nicotine Efficacy	0.23	0.05	0.60	A
rs17244841	HMG-CoA reductase inhibitors, Pravastatin, Simvastatin Efficacy	0.08	0.05	0.54	A
rs2279345	Efavirenz Metabolism/PK	-0.09	-0.05	0.52	C
rs4149015	Pravastatin Efficacy	-0.08	-0.05	0.51	G
rs776746	Cyclosporine Dosage	-1.86	-0.49	5.26	T
rs11615	Carboplatin, Cisplatin, Oxaliplatin, Platinum, Platinum compounds Toxicity/ADR	2.70	0.41	4.91	A
rs7793837	Salbutamol, selectivebeta-2-adrenoreceptoragonists Efficacy	1.49	0.45	4.71	A
rs20455	Atorvastatin Efficacy	-1.12	-0.43	4.43	G
rs1954787	Antidepressants Efficacy	1.55	0.41	4.36	C
rs11212617	Metformin Efficacy	-1.16	-0.41	4.25	C
rs7997012	Antidepressants, citalopram, selective serotonin reuptake inhibitors Efficacy	2.38	0.28	3.70	A
rs489693	Amisulpride, Aripiprazole, Clozapine, Haloperidol, Olanzapine, Paliperidone, Quetiapine, Risperidone, Ziprasidone Toxicity/ADR	-1.51	-0.32	3.58	A
rs339097	Warfarin Dosage	-3.09	-0.12	3.32	G
rs6988229	Salbutamol Efficacy	-1.35	-0.30	3.27	T
rs1042522	Antineoplastic agents, Cisplatin,	-1.05	-0.31	3.23	G

rs4713916	Cyclophosphamide, Fluorouracil, Paclitaxel Toxicity/ADR Antidepressants, Citalopram, Fluoxetine, Mirtazapine, Paroxetine, Selective serotonin reuptake inhibitors, Venlafaxine Efficacy	2.14	0.21	3.01	A
rs3812718	Carbamazepine Dosage	1.02	0.28	2.97	T
rs1051730	Nicotine Toxicity/ADR	1.60	0.24	2.88	A
rs1719247	HMG-CoA reductase inhibitors, Simvastatin Toxicity/ADR	1.00	0.27	2.88	C
rs924607	Vincristine Toxicity/ADR	1.72	0.23	2.85	T
rs2108622	Acenocoumarol Dosage	1.90	0.20	2.77	T
rs4444903	Cetuximab Efficacy	-0.59	-0.26	2.62	G
rs25487	Carboplatin, Cisplatin, Oxaliplatin, Platinum, Platinum compounds Efficacy	-0.47	-0.24	2.48	C
rs2232228	Anthracyclines Toxicity/ADR	-0.46	-0.24	2.43	A
rs7779029	Irinotecan Toxicity/ADR	-1.53	-0.17	2.27	C
rs1933437	Sunitinib Toxicity/ADR	0.65	0.21	2.16	A
rs3753380	Latanoprost Efficacy	-0.37	-0.21	2.15	C
rs2231142	Allopurinol Dosage	1.81	0.10	2.07	T
rs17782313	Antipsychotics Toxicity/ADR	-1.29	-0.16	2.04	C
rs2298383	Caffeine Toxicity/ADR	-0.49	-0.20	2.04	C
rs16960228	Hydrochlorothiazide Efficacy	-1.13	-0.17	2.01	A

rs738409	Asparaginase, Cyclophosphamide, Daunorubicin, Prednisolone, Vincristine Toxicity/ADR	0.87	0.18	2.00	G
rs1799971	Ethanol Metabolism/PK	1.66	0.11	1.99	G
rs1532624	HMG-CoA Efficacy	-0.34	-0.18	1.84	C
rs2234922	Carbamazepine Dosage	-0.92	-0.16	1.83	G
rs1876828	Budesonide, Corticosteroids, Fluticasonepropionate, Fluticasone/Salmeterol, Triamcinolone Efficacy	-0.27	-0.17	1.71	C
rs578776	Nicotine Toxicity/ADR	0.53	0.16	1.68	G
rs730012	Aspirin Toxicity/ADR	1.08	0.13	1.65	C
rs11598702	Gemcitabine Metabolism/PK	0.76	0.15	1.65	C
rs4803419	Efavirenz Metabolism/PK	-0.31	-0.16	1.61	C
rs7297610	Hydrochlorothiazide Efficacy	0.27	0.16	1.59	C
rs746647	Nevirapine Toxicity/ADR	0.79	0.13	1.49	G
rs1800497	Ethanol Efficacy	-0.70	-0.13	1.45	A
rs1800497	Antipsychotics, Clozapine, Olanzapine, Risperidone Efficacy	-0.70	-0.13	1.45	A
rs264588	Radiotherapy Toxicity/ADR	-0.67	-0.11	1.32	A
rs7582141	Radiotherapy Toxicity/ADR	-0.65	-0.11	1.26	T
rs1056892	Anthracyclines Toxicity/ADR	0.27	0.12	1.25	G
rs10306114	Aspirin Efficacy	-1.09	-0.06	1.25	G
rs1695	Cisplatin Toxicity/ADR	-0.40	-0.12	1.24	G

rs6432512	Radiotherapy Toxicity/ADR	-0.62	-0.11	1.23	T
rs12777823	Warfarin Dosage	0.21	0.12	1.21	G
rs1051740	Carbamazepine Dosage	0.60	0.10	1.20	C
rs13306278	Selective Efficacy	-0.17	-0.11	1.09	C
rs1801274	Trastuzumab Efficacy	0.27	0.10	1.07	A
rs4961	Furosemide, Spironolactone Efficacy	-0.17	-0.10	1.05	G
rs7853758	Anthracyclinesandrelated substances Toxicity/ADR	0.20	0.10	1.02	G
rs2072661	Nicotine Toxicity/ADR	-0.53	-0.08	0.95	A
rs716274	Etoposide, Platinum compounds Toxicity/ADR	0.31	0.09	0.93	G
rs2952768	Buprenorphine, Fentanyl, Meperidine, Morphine, Opioids, Pentazocine Dosage	0.39	0.08	0.93	C
rs885004	Anthracyclines and related substances Toxicity/ADR	-0.14	-0.09	0.88	G
rs7270101	Peginterferon alfa-2B, Ribavirin Toxicity/ADR	0.79	0.03	0.85	C
rs1517114	Irinotecan Toxicity/ADR	-0.27	-0.08	0.81	C
rs1076560	Cocaine Toxicity/ADR	0.61	0.05	0.79	A
rs1801394	Methotrexate Toxicity/ADR	0.39	0.07	0.77	G
rs1801019	Capecitabine, Fluorouracil, Leucovorin, Tegafur Toxicity/ADR	-0.39	-0.07	0.77	C
rs4693075	Atorvastatin, HMG-CoA reductase inhibitors,	-0.24	-0.07	0.71	G

	Rosuvastatin Toxicity/ADR				
rs4880	Cyclophosphamide Efficacy	-0.18	-0.07	0.70	A
rs6065	Aspirin Efficacy	-0.42	-0.05	0.66	T
rs4149601	Diuretics, Hydrochlorothiazide Efficacy	0.12	0.06	0.61	G
rs1346268	HMG-CoA reductase inhibitors, Simvastatin Toxicity/ADR	-0.12	-0.05	0.52	T
rs10455872	HMG-CoA Efficacy	-0.08	-0.05	0.52	A
rs61750900	Nicotine Metabolism/PK	-0.08	-0.05	0.50	G

APPENDIX B.

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Table 11. Global reference populations used for genetic ancestry inference.

[1]Population name

[2]Number of samples

[3]Population continental ancestry

[4]Data source: 1000 Genomes Project (1KGP) [65], Human Genome Diversity Project (HGDP) [127], Collection of Native American Samples (Reich et al.)[128].

Population[1]	N[2]	Continental ancestry[3]	Source[4]
African Caribbean in Barbados	94	African	1KGP
Algonquin	5	Native American	Reich et al.
Americans of African ancestry from SW USA	51	African	1KGP
Utah Residents with Northern and Western European Ancestry	99	European	1KGP
Chipewyan	13	Native American	Reich et al.
Cree	4	Native American	Reich et al.
Finnish in Finland	99	European	1KGP
French	28	European	HGDP
British in England and Scotland	91	European	1KGP
Iberian Population in Spain	107	European	1KGP
Mixe	17	Native American	Reich et al.
Mixtec	5	Native American	Reich et al.
Ojibwa	5	Native American	Reich et al.
Orcadian	15	European	HGDP
Piapoco	7	Native American	Reich et al.
Pima	14	Native American	HGDP
Russian	25	European	HGDP
Sardinian	28	European	HGDP
Tepehuano	25	Native American	Reich et al.
Teribe	3	Native American	Reich et al.
Ticuna	6	Native American	Reich et al.
Toscani in Italia	107	European	1KGP

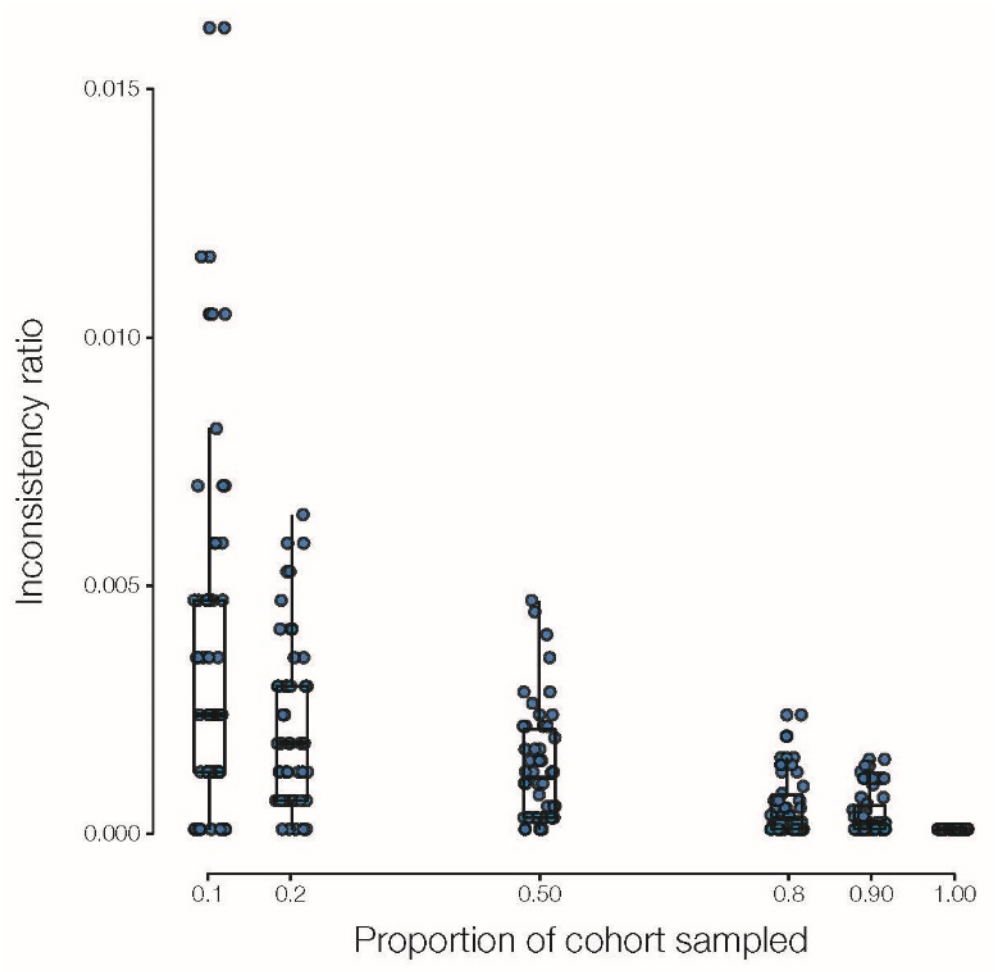


Figure 27. Permutation analysis to evaluate the stability of k-means genetic ancestry (GA) clusters.

The HRS cohort was randomly sampled at different proportions, where the proportion of the cohort sampled = the number of participants in the random sample / the total number of participants in the cohort. For each random sample, k-means clustering was run 50 times and an inconsistency ratio was calculated for each independent run, where the inconsistency ratio is the number of mismatches between the random sample group assignments / the number of participants in the random sample. In other words, the inconsistency ratio measures the error in k-means cluster assignments due to sampling bias. As can be expected, error is higher for smaller random cohort proportions and decreases monotonically as the proportion of the random cohorts increases. Nevertheless, the error level, even at the smallest sampling proportions, is extremely low. The mean error at a sampling proportion of 0.1 is 0.4%, and when the entire cohort is sampled (i.e. cohort proportion=1) k-means clustering is 100% consistent.

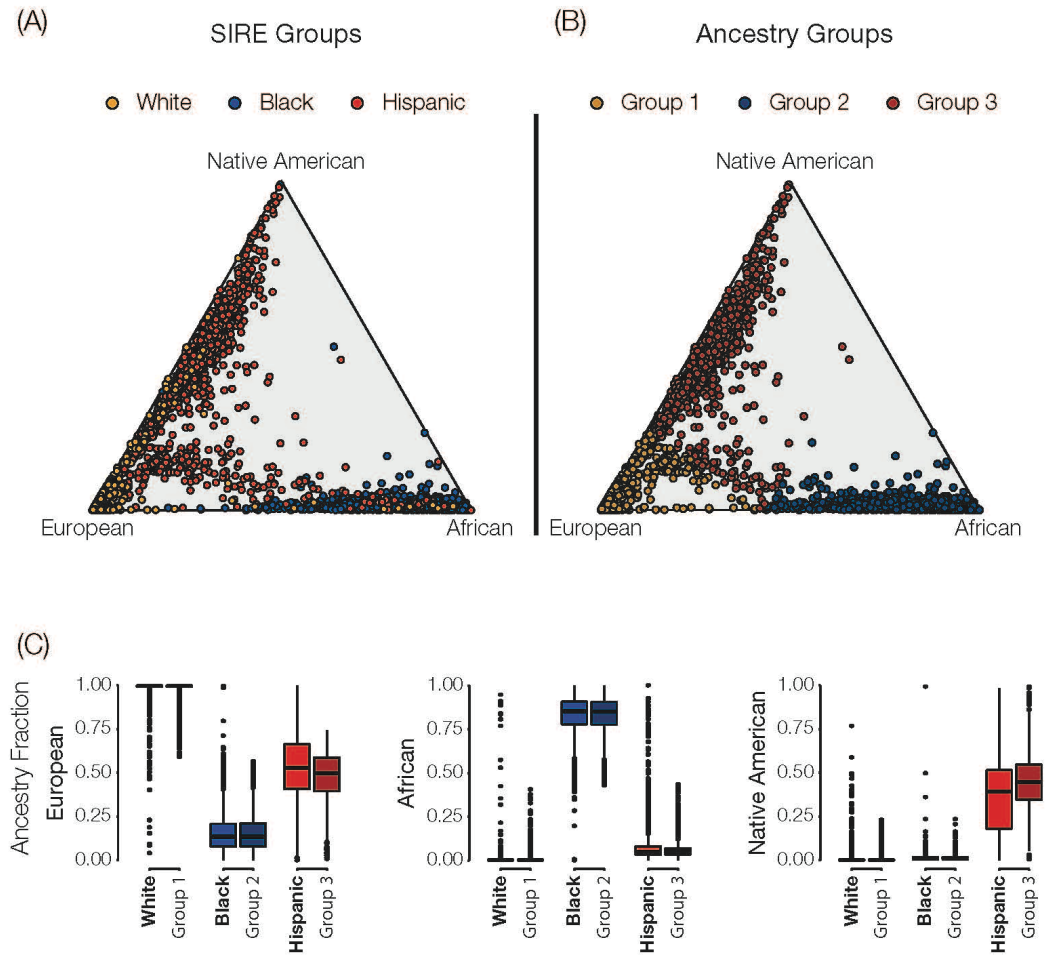


Figure 28. Comparison of self-identified race/ethnicity (SIRE) versus genetic ancestry (GA) groups in the US.

Ternary plots showing the relative continental ancestry fractions for HRS participants are shown with individuals color coded by SIRE (A) or genetic ancestry (B). SIRE and their corresponding GA groups are coded as White/Group 1 (orange), Black/Group 2 (blue), and Hispanic/Group 3 (red). (C) Distributions of continental ancestry fractions – European, African, and Native American – for HRS participants are shown corresponding SIRE and GA groups.

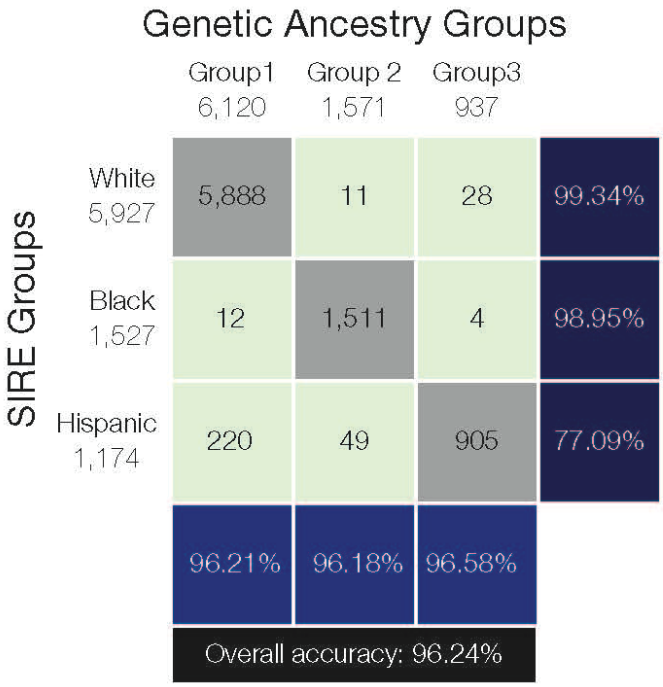


Figure 29. Correspondence between self-identified race/ethnicity (SIRE) versus genetic ancestry (GA) groups in the US.

Numbers of HRS participants that fall into each combination of SIRE and GA groups is shown along with the percentage correspondence. Individual percent correspondence values are calculated as the number of individuals along the diagonal, i.e. that fall into the corresponding SIRE and GA groups, divided by the total number of individuals in each SIRE group (right) or each GA group (bottom), times 100. The overall percent correspondence is calculated as the number of individuals along the diagonal divided by the total number of individuals in the HRS cohort, times 100.

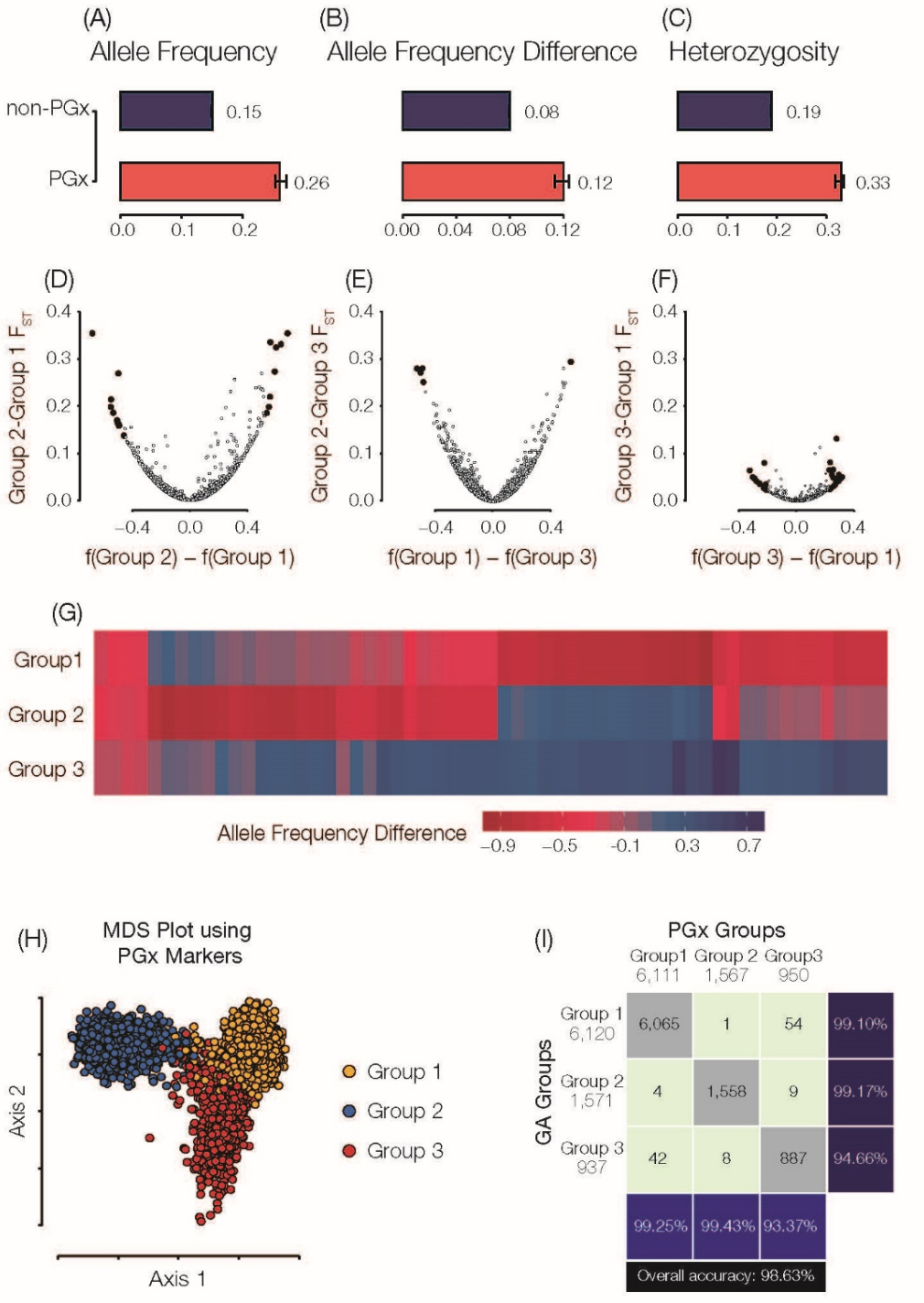


Figure 30. Pharmacogenomic variation in the US: genetic ancestry (GA).

Data shown here correspond to GA groups; analogous results for SIRE groups shown in Figure 2. Genome-wide average allele frequencies (A), group-specific allele frequency differences (B), and heterozygosity fractions (C) are shown for PGx variants (red) compared to non-PGx variants (blue). (D-F) Fixation index (F_{ST}; y-axis) and allele

frequency differences (x-axis) for pairs of GA groups. Statistically significant PGx allele frequency differences are highlighted in black. (G) Heatmap showing group-specific allele frequencies for significantly diverged PGx variants. (H) Multi-dimensional scaling (MDS) plot showing the relationship among individual genomes as measured by PGx variants alone. Each dot is an individual HRS participant genome, and genomes are color-coded by participants GA groups. (I) The correspondence between GA groups and PGx groups defined by K-means clustering on the results of the MDS analysis.

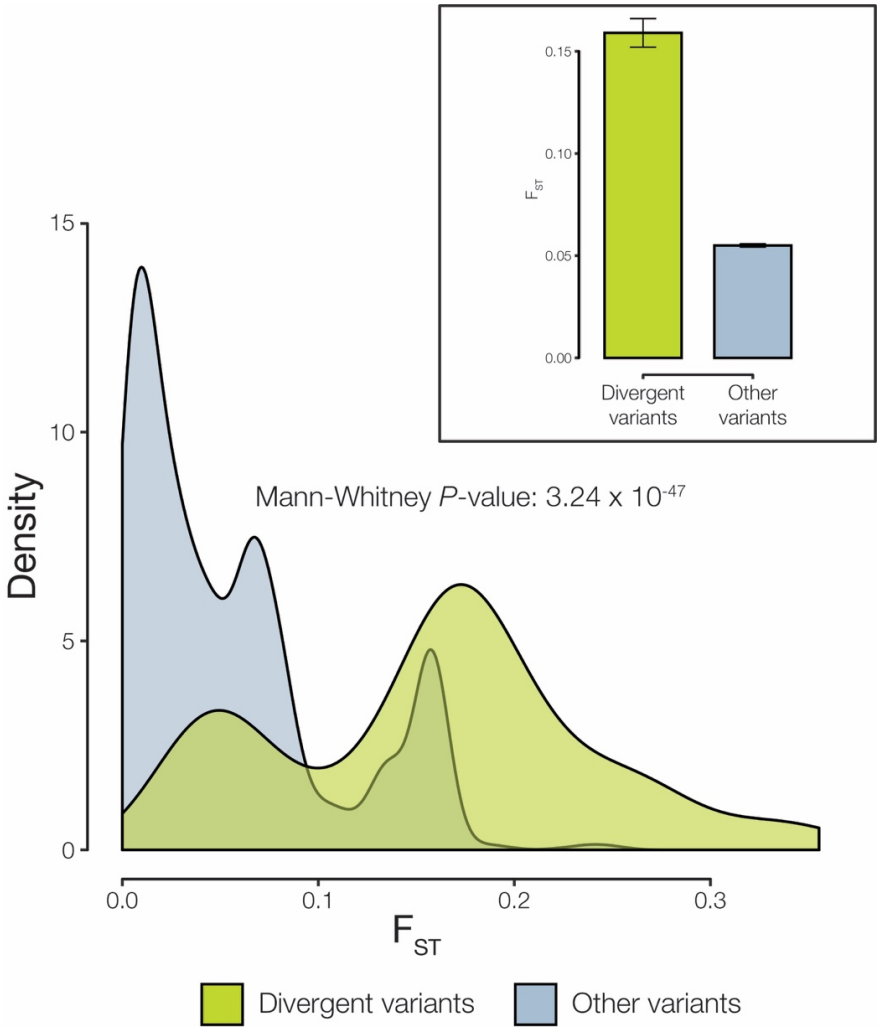


Figure 31. F_{ST} distribution of divergent PGx variants.

Data shown here correspond to the F_{ST} distribution of the diverged variants shown in Figure 2G. The inset shows mean values plotted as barplots and standard error plotted as error bars.

APPENDIX C. SUPPLEMENTARY INFORMATION FOR CHAPTER 4

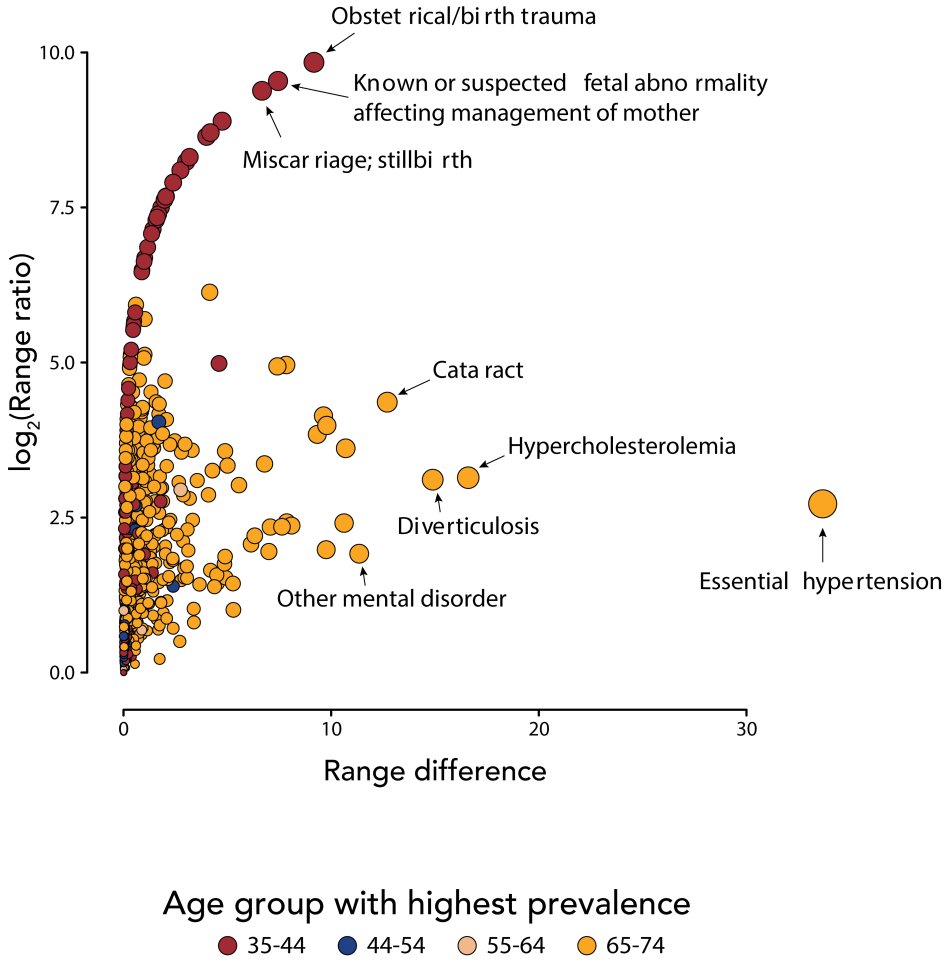


Figure 32 Disease phenotype disparities for groups defined by age.

Each point is a disease phenotype and is colored to indicate the groups defined by age with the highest prevalence for that phenotype. The size and opacity of each point is scaled by the distance from origin.

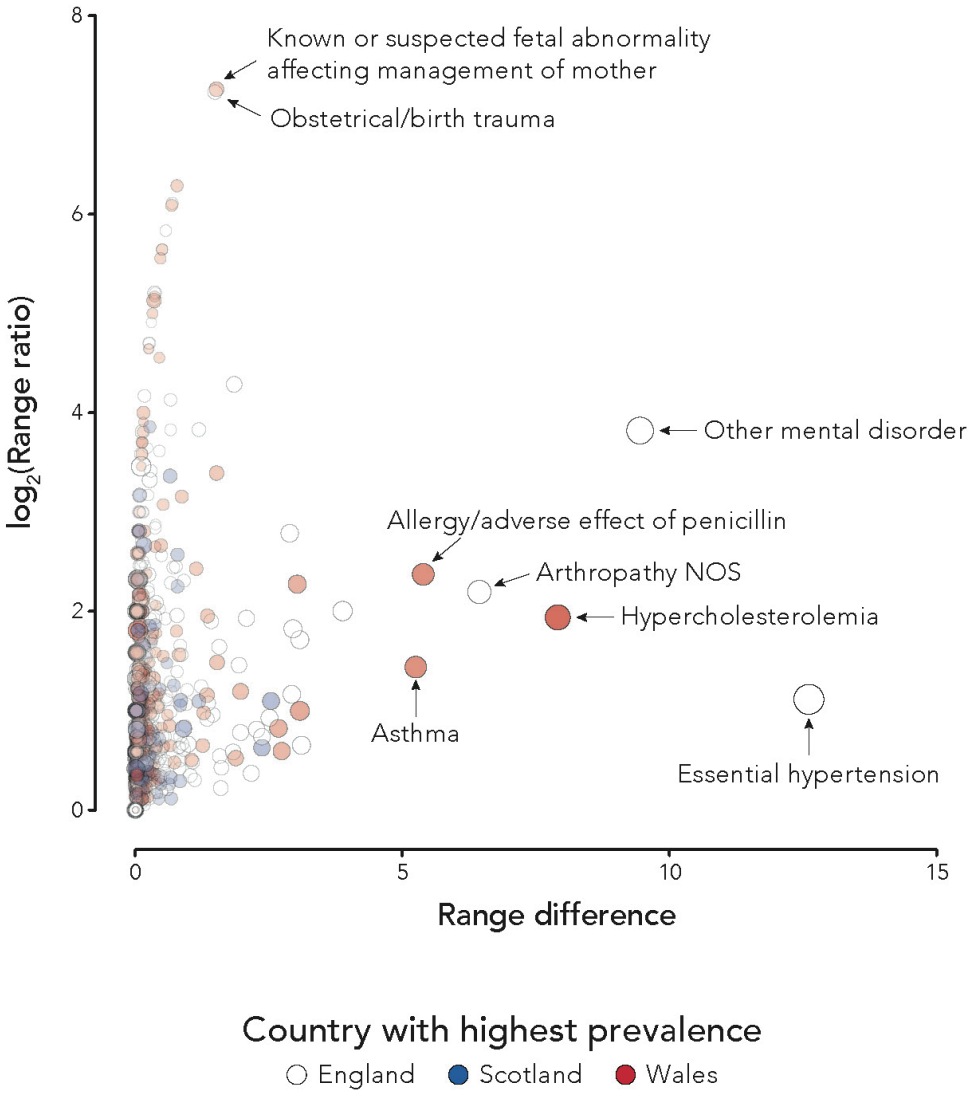


Figure 33. Disease phenotype disparities for groups defined by country of residence.

Each point is a disease phenotype and is colored to indicate the groups defined by country of residence with the highest prevalence for that phenotype. The size and opacity of each point is scaled by the distance from origin.

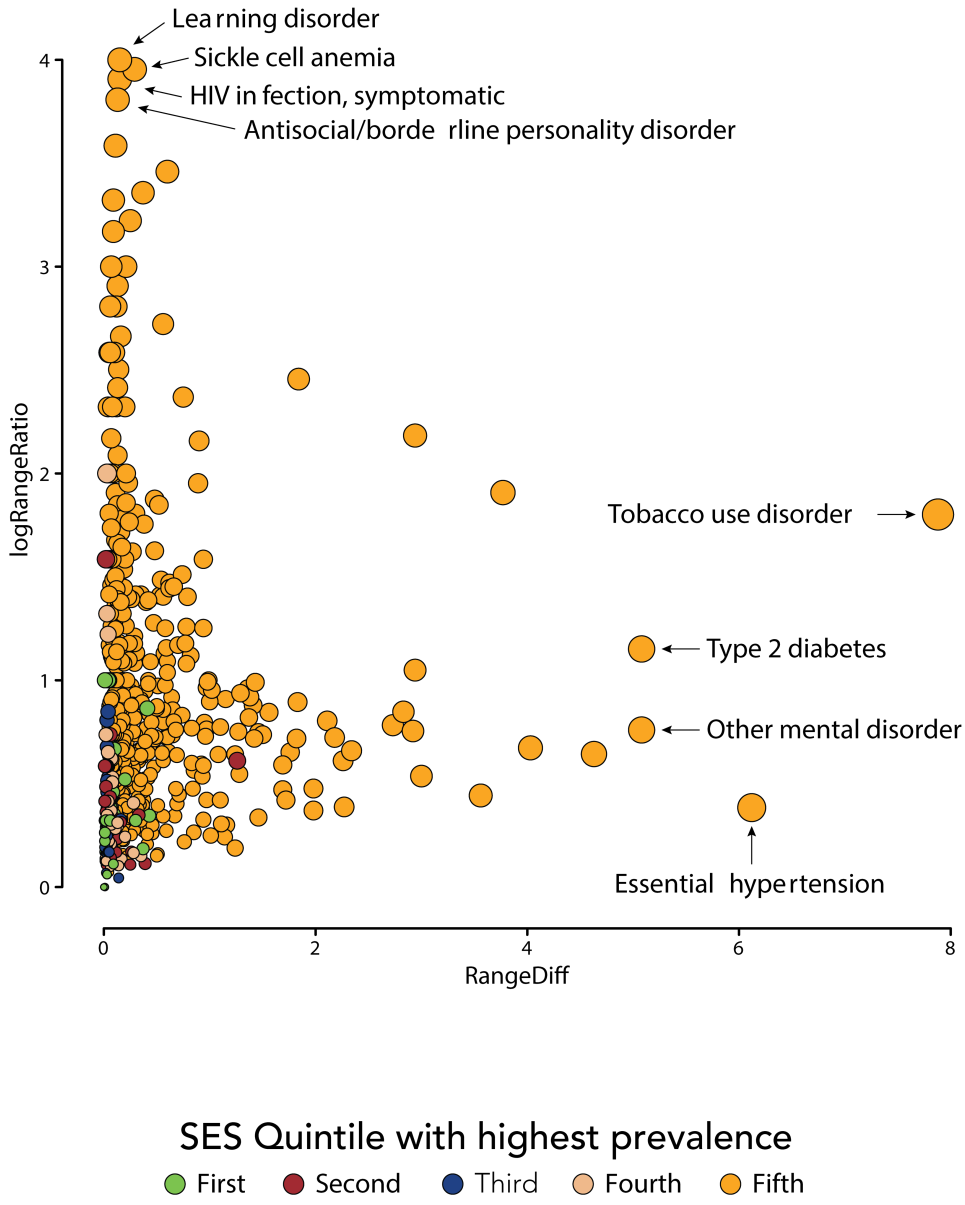


Figure 34 Disease phenotype disparities for groups defined by socioeconomic deprivation.

Each point is a disease phenotype and is colored to indicate the groups defined by socioeconomic deprivation with the highest prevalence for that phenotype. The size and opacity of each point is scaled by the distance from origin.

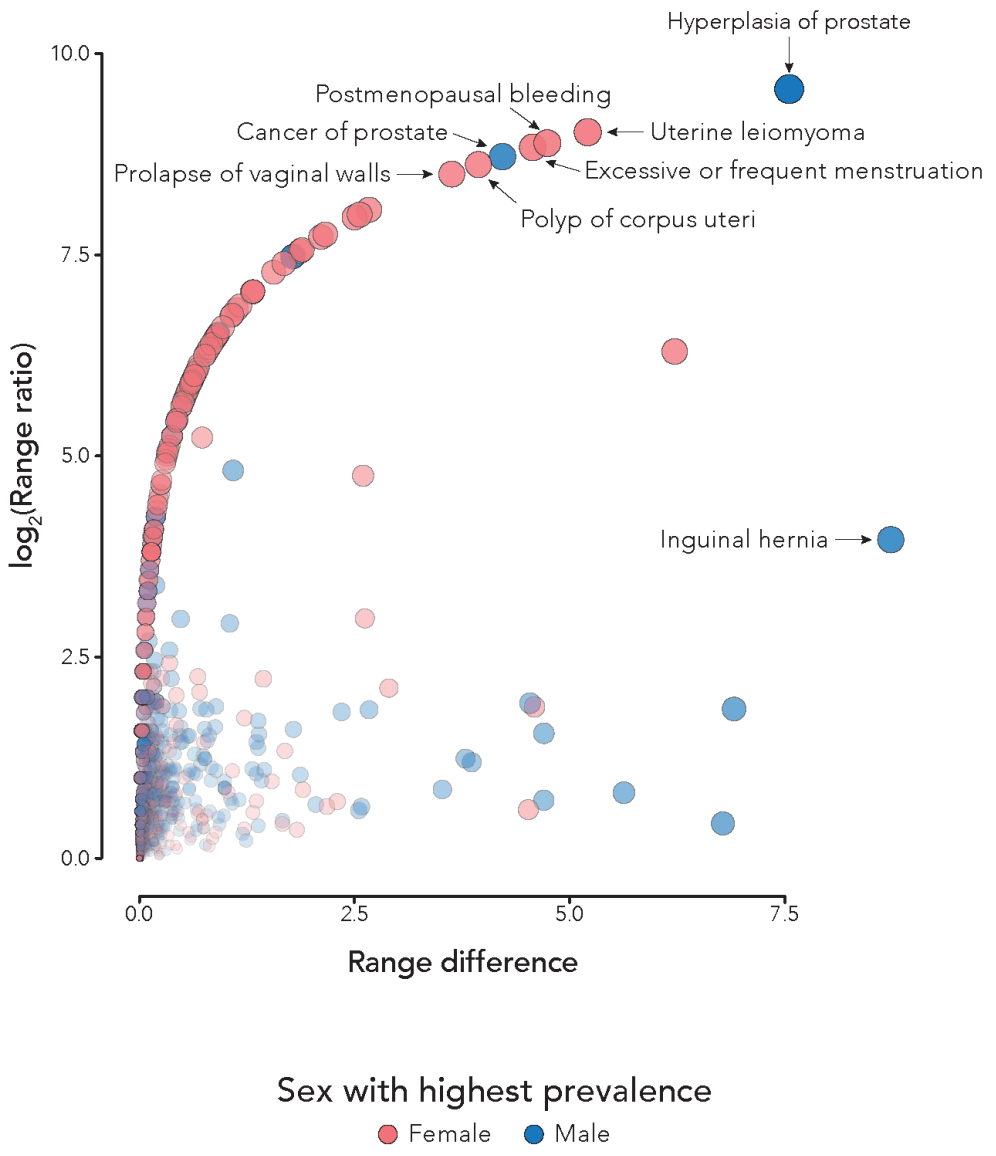


Figure 35. Disease phenotype disparities for groups defined by sex.

Each point is a disease phenotype and is colored to indicate the groups defined by sex with the highest prevalence for that phenotype. The size and opacity of each point is scaled by the distance from origin.

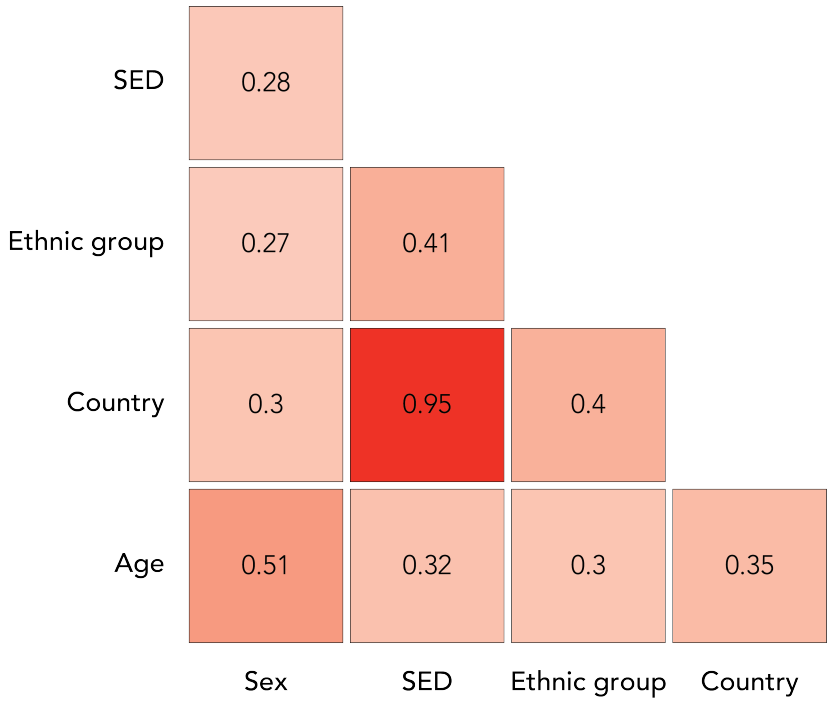


Figure 36. Pairwise correlation between disease disparity scores across population attributes.

Heatmap showing pairwise correlations between population attributes. Each bix shows the Spearman r value for the correlation. Darker red indicates higher correlation.

APPENDIX D.

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

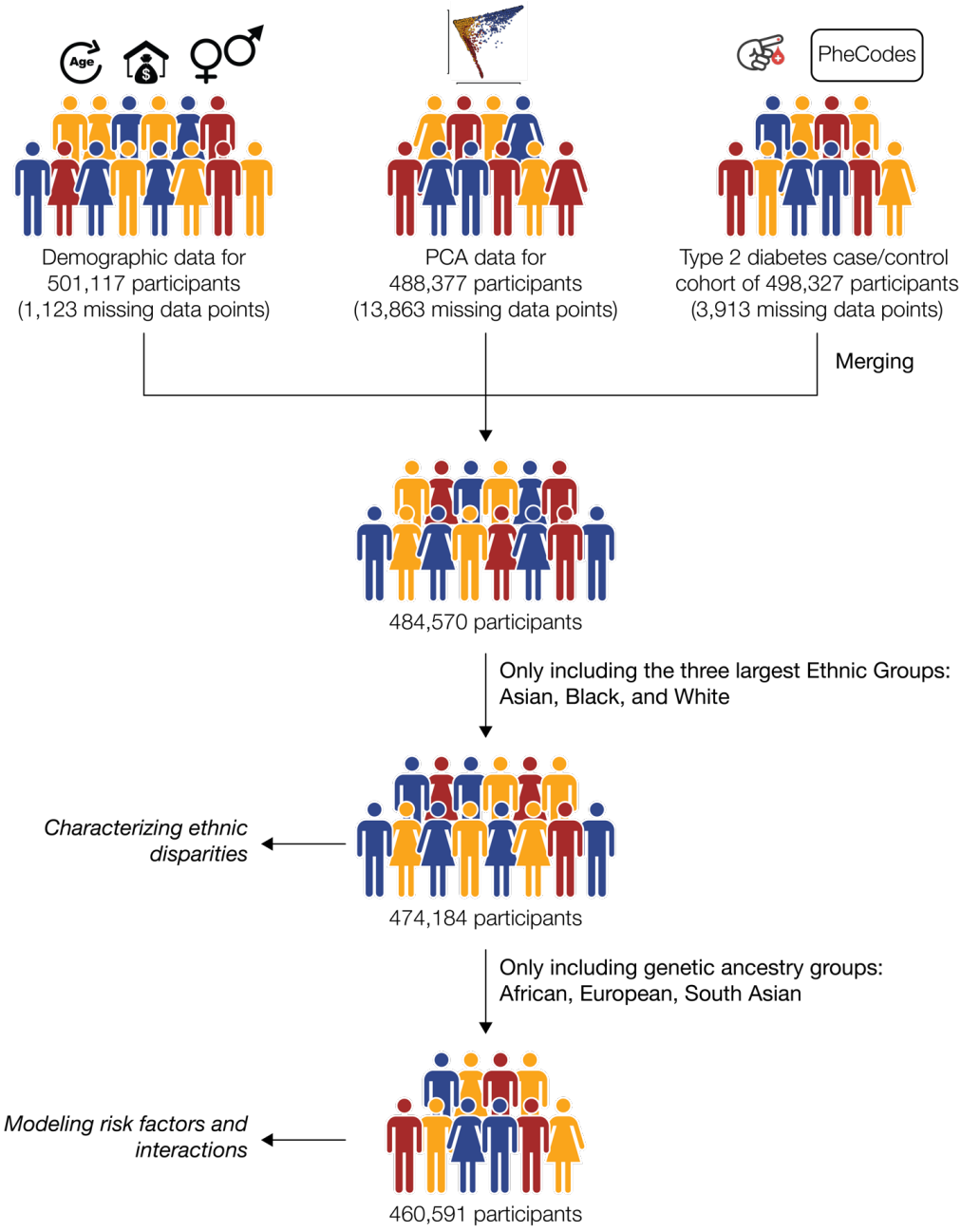


Figure 37. Generation of study cohort.

Overview of data inclusion/exclusion criteria from UK Biobank to final analysis cohort.

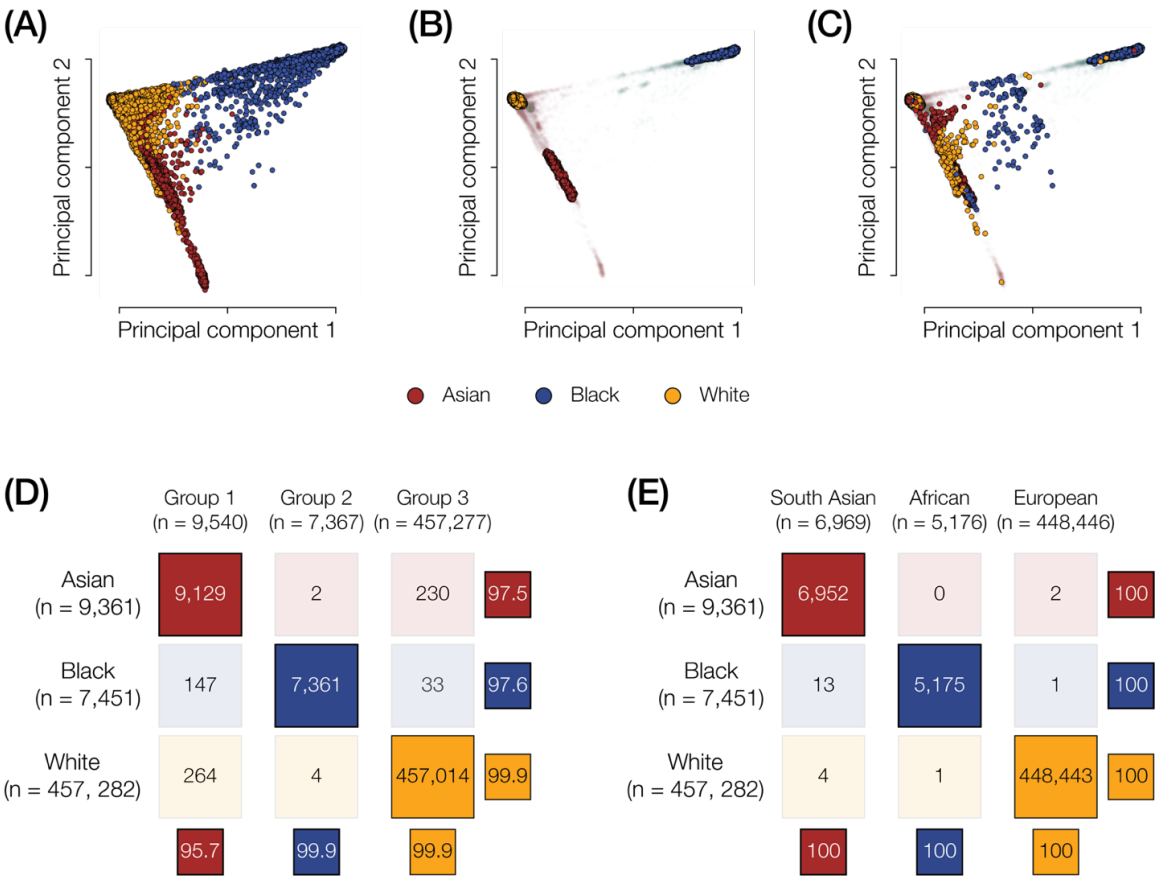


Figure 38. **Ethnic group and genetic ancestry.**

Individuals in a space defined by the first two genetic principal components colored by ethnic group – Asian (red), Black (blue), and White (orange). (A) All individuals shown prominently. (B) Individuals part of genetic ancestry groups highlighted. (C) Individuals part of genetic ancestry groups and outliers highlighted. (D) Concordance between self-identified ethnic group and continuous genetic ancestry groups defined using genetic data. (E) Concordance table between discrete genetic ancestry groups defined using genetic data.

Table 12. T2D multivariable logistic regression model with interaction terms (Model 1).

Details of the model used are shown below along with model coefficient estimates, standard errors, z values, p-values, and dominance analysis R^2 values and ranks.

*logit(p) = b0 + b1*Age + b2*Sex + b3*African + b4*SouthAsian + b5*SED + b6*African*SED + b7*SouthAsian*SED*

	Name	Estimate	Std. Error	z value	P-value	Dominance analysis
b0	Intercept	-3.198359	0.01082	-295.61	~0	
b1	Age	0.576009	0.007392	77.925	~0	3.21 (1)
b2	Sex (Male)	0.618593	0.013116	47.162	~0	1.20 (2)
b3	African	0.925661	0.062244	14.871	5.06×10^{-50}	0.81 (3)
b4	SouthAsian	1.475287	0.033195	44.443	~0	
b5	SED	0.102774	0.002018	50.925	~0	0.62 (4)
b6	African*SED	-0.043717	0.013307	-3.285	1.02×10^{-2}	
b7	SouthAsian*SED	-0.025288	0.01048	-2.413	1.58×10^{-2}	

Table 13. Likelihood ratio test for T2D multivariable logistic regression model with and without interaction terms.

The models and the test statistic are shown below.

Model A: T2D ~ Age + Sex + GA + SED

*Model B: T2D ~ Age + Sex + GA + SED + GA*SED*

Models compared	χ^2	P-value	Δdf
B vs. A	15.96	3.418×10^{-4}	2

Table 14. T2D multivariable logistic regression model with GA combined with SED terciles (Model 2).

Details of the model used are shown below along with model coefficient estimates, standard errors, z values, and p-values.

*logit(p) = b0 + b1*Age + b2*Sex + b3*EuropeanMediumSED + b4*EuropeanHighSED + b5*AfricanLowSED + b6*AfricanMediumSED + b7*AfricanHighSED + b8*SouthAsianLowSED + b9*SouthAsianMediumSED + b10*SouthAsianHighSED*

	Name	Estimate	Std. Error	z value	P-value
b0	Intercept	-3.636372	0.015566	-233.617	~0
b1	Age	0.571606	0.007382	77.43	~0
b2	Sex (Male)	0.6246	0.013104	47.664	~0
b3	EuropeanMediumSED	0.21993	0.017502	12.566	3.25×10^{-36}
b4	EuropeanHighSED	0.675504	0.01646	41.039	~0
b5	AfricanLowSED	1.06971	0.238512	4.485	7.29×10^{-06}
b6	AfricanMediumSED	1.292777	0.132121	9.785	1.31×10^{-22}
b7	AfricanHighSED	1.598551	0.049888	32.043	2.77×10^{-225}
b8	SouthAsianLowSED	1.517856	0.09115	16.652	2.92×10^{-62}
b9	SouthAsianMediumSED	1.834989	0.066043	27.785	6.61×10^{-170}
b10	SouthAsianHighSED	2.073268	0.042197	49.133	~0

APPENDIX E.

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table 15. Global reference populations used for genetic ancestry inference of UK Biobank participants.

Geographic region	Population Name (Abbreviation)	Source[1]
African	Yoruba in Ibadan, Nigeria (YRI)	1KGP
	Esan in Nigeria (ESN)	1KGP
	Luhya in Webuye, Kenya (LWK)	1KGP
	Gambian in Western Divisions in the Gambia (GWD)	1KGP
	Mbuti in Democratic Republic of Congo (MBU)	HGDP
	Biaka in Central African Republic (BIA)	HGDP
European	Finnish in Finland (FIN)	1KGP
	British in England and Scotland (GBR)	1KGP
	Iberian Population in Spain (IBS)	1KGP
	Toscani in Italia (TSI)	1KGP
	French in France (FRE)	HGDP
	Basque in France (BAS)	HGDP
	Bergamo Italian in Bergamo, Italy (BER)	HGDP
	Tuscan in Italy (TUS)	HGDP

[1] *1KGP – 1000 Genomes Project, HGDP – Human Genome Diversity Project*

Table 16. CRP multivariable linear regression (Model 1).

Model equation, coefficient estimates, standard errors, z values, and P-values are shown.

*CRP = b0 + b1*Ethnicity + b2*Age + b3*Sex*

Coefficient	Name	Estimate	Std. Error	z value	P-value
b0	Intercept	0.7768	0.0476	16.314	$< 2 \times 10^{-16}$
b1	Ethnicity (Black)	0.3049	0.0548	5.562	2.67×10^{-16}
b2	Age	0.3396	0.0083	41.124	$< 2 \times 10^{-16}$
b3	Sex (Male)	-0.2423	0.0133	-18.214	$< 2 \times 10^{-16}$

Table 17. CRP multivariable linear regression with interaction term (Model 2).

Model equation, coefficient estimates, standard errors, z values, and P-values are shown.

*CRP = b0 + b1*Ethnicity + b2*Age + b3*Sex + b4*(Ethnicity * Sex)*

Coefficient	Name	Estimate	Std. Error	z value	P-value
b0	Intercept	0.7729	0.0476	16.234	$< 2 \times 10^{-16}$
b1	Ethnicity (Black)	0.6912	0.0722	5.579	2.67×10^{-16}
b2	Age	0.3392	0.0083	41.078	$< 2 \times 10^{-16}$
b3	Sex (Male)	-0.2290	0.0134	-17.083	$< 2 \times 10^{-16}$
b4	Ethnicity * Sex	-0.9111	0.1106	-8.233	$< 2 \times 10^{-16}$

Table 18. Likelihood ratio test for CRP multivariable linear regression models with and without ethnicity-sex interaction term.

The model equations and the likelihood test statistic values are shown below.

*Model 1: CRP = b0 + b1*Ethnicity + b2*Age + b3*Sex*

*Model 2: CRP = b0 + b1*Ethnicity + b2*Age + b3*Sex + b4*(Ethnicity * Sex)*

Models compared	χ^2	P-value	Δdf
1 vs. 2	67.78	1.822×10^{-16}	1

Table 19. CRP structural equation modeling with African ancestry mediation (Model 3). Path diagram for the model is shown in Figure 4A.

The top table shows all models evaluated along with each effect size estimate, standard error, z value, and P-value. The bottom table shows effect size estimates, standard errors, z values, and P-values for the indirect effect and total effect ethnicity on CRP levels.

Models evaluated (Dependent ~ Independent variables)	Path Branch Label	Effect size estimate	Standard error	z value	P-value
CRP ~					
Ethnicity	c	0.5542	0.4316	1.2842	0.1991
Age		0.3396	0.0082	41.1193	~0
Sex		-0.2423	0.0133	-18.2107	~0
African ancestry ~					
Ethnicity	a	8.7786	0.0017	5152.4389	~0
CRP ~					
African ancestry	b	-0.0284	0.0488	-0.5823	0.5603

			Effect size estimate	Standard error	z value	P-value
Indirect effect	:=	a*b	-0.2493	0.4281	-0.5823	0.5603
Total effect	:=	c + (a*b)	0.3049	0.0548	5.5618	2.67×10^{-08}

Table 20. CRP structural equation modeling with socioeconomic deprivation (SED) mediation (Model 4).

Path diagram for the model is shown in Figure 4B. The top table shows all models evaluated along with each effect size estimate, standard error, z value, and P-value. The bottom table shows effect size estimates, standard errors, z values, and P-values for the indirect effect and total effect ethnicity on CRP levels.

Models evaluated (Dependent ~ Independent variables)	Path Branch Label	Effect size estimate	Standard error	z value	P-value
CRP ~					
Ethnicity	c	0.0382	0.0552	0.6929	0.4884
Age		0.3607	0.0082	43.7560	~0
Sex		-0.2452	0.0133	-18.4653	~0
SED ~					
Ethnicity	a	1.4516	0.0175	83.0066	~0
CRP ~					
SED	b	0.1905	0.0047	40.1743	~0

			Effect size estimate	Standard error	z value	P-value
Indirect effect	:=	a*b	0.2765	0.0076	36.1616	~0
Total effect	:=	c + (a*b)	0.3148	0.0548	5.7410	9.41×10^{-09}

Table 21. CRP multivariable linear regression models with (1) ethnicity, (2) African ancestry, and (3) socioeconomic deprivation as independent (predictor) variables.

Model equations, coefficient estimates, standard errors, z values, and P-values are shown.

CRP ~ Age + Sex + African ancestry | Adjusted R^2 = 0.004576					
Coefficient	**Name**	**Estimate**	**Std. Error**	**z value**	**P-value**
b0	Intercept	0.7771	0.0476	16.321	$< 2 \times 10^{-16}$
b1	Age	0.3396	0.0083	41.116	$< 2 \times 10^{-16}$
b2	Sex (Male)	-0.2424	0.0133	-18.218	$< 2 \times 10^{-16}$
b3	African ancestry	0.0337	0.0062	5.443	5.24×10^{-08}
CRP ~ Age + Sex + SED | Adjusted R^2 = 0.009315					
Coefficient	**Name**	**Estimate**	**Std. Error**	**z value**	**P-value**
b0	Intercept	0.7837	0.0473	16.56	$< 2 \times 10^{-16}$
b1	Age	0.3649	0.0082	44.28	$< 2 \times 10^{-16}$
b2	Sex (Male)	-0.2504	0.0133	-18.86	$< 2 \times 10^{-16}$
b3	SED	0.1004	0.0022	45.85	$< 2 \times 10^{-16}$

PUBLICATIONS

1. **Nagar, S.D.**, Mariño-Ramírez, L., and Jordan, I.K. The landscape of health disparities in the UK Biobank. *In preparation*.

2. **Nagar, S.D.**, Gupta, S., Pellebon, J.M., Augenbroe, A.M., and Jordan, I.K. Population-specific differences in the burden of tier 1 genomics disease. *In preparation*.

3. **Nagar, S.D.**, Conley, A.B., Sharma, S., Rishishwar, L., Jordan, I.K., and Mariño-Ramírez, L. Comparing genetic and socioenvironmental contributions to ethnic differences in C-reactive protein. *In review*.

4. **Nagar, S.D.**, Pemu, P., Zuchner, S., Jordan, I.K., and Meller, R. Investigation of Hypertension and Diabetes as Risk Factors for Dementia: An Investigation using the All of Us Research Workbench. *In review*.

5. **Nagar, S.D.**, Conley, A.B., Chande, A.T., Rishishwar, L., Sharma, S., Mariño-Ramírez, L., Aguinaga-Romero, G., González-Andrade, F., and Jordan, I.K. Genetic ancestry and ethnic identity in Ecuador. *In review*.

6. **Nagar, S.D.**, Nápoles, A.M., Jordan, I.K., and Mariño-Ramírez, L. (2021). Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the United Kingdom. EClinicalMedicine 37, 100960.

7. **Nagar, S.D.**, Conley, A.B., and Jordan, I.K. (2020). Population structure and pharmacogenomic risk stratification in the United States. BMC Biol 18, 140.

8. **Nagar, S.D.**, Moreno, A.M., Norris, E.T., Rishishwar, L., Conley, A.B., O'Neal, K.L., Velez-Gomez, S., Montes-Rodriguez, C., Jaraba-Alvarez, W.V., Torres, I., et al. (2019). Population Pharmacogenomics for Precision Public Health in Colombia. Front Genet 10, 241.

9. **Nagar, S.D.**, Conley, A.B., and Jordan, I.K. (2019). Deconvolving Human Evolutionary History: Using Network-Based Approaches to Better Understand Our Past. In. (Systems Medicine: Integrative Qualitative and Computational Approaches, Elsevier.

10. **Nagar, S.D.**, Aggarwal, B., Joon, S., Bhatnagar, R., and Bhatnagar, S. (2016). A Network Biology Approach to Decipher Stress Response in Bacteria Using *Escherichia coli* As a Model. OMICS 20, 310-324.

11. Martínez, B., Conley, A.B., Rishishwar, L., **Nagar, S.D.**, Gusmão, L., Daya, M., Gignoux, C.R., Gravel, S., O'Connor, T.D., Mariño-Ramírez, L., et al. Genetic ancestry, language, and African origins of San Basilio de Palenque: the Americas first free town. *In preparation*.

12. Lee, K.K., Rishishwar, L., Ban, D., **Nagar, S.D.**, Mariño-Ramírez, L., McDonald, J.F., and Jordan, I.K. The effects of genetic ancestry and molecular signatures on cancer survival disparities: a pan-cancer analysis. *In review.*

13. Chande, A.T., **Nagar, S.D.**, Rishishwar, L., Mariño-Ramírez, L., Medina-Rivas, M.A., Valderrama-Aguirre, A.E., Jordan, I.K., and Gallo, J.E. The impact of ethnicity and genetic ancestry on disease prevalence and risk in Colombia. *In review*.

14. Mariño-Ramírez, L., Sharma, S., Rishishwar, L., Conley, A.B., **Nagar, S.D.**, and Jordan, I.K. African genetic ancestry mediates ethnic differences in serum creatinine levels. *In review*.

15. Espitia-Navarro, H.F., Chande, A.T., **Nagar, S.D.**, Smith, H., Jordan, I.K., and Rishishwar, L. (2020). STing: accurate and ultrafast genomic profiling with exact sequence matches. Nucleic Acids Res 48, 7681-7689.

16. Chande, A.T., Rishishwar, L., Ban, D., **Nagar, S.D.**, Conley, A.B., Rowell, J., Valderrama-Aguirre, A.E., Medina-Rivas, M.A., and Jordan, I.K. (2020). The Phenotypic Consequences of Genetic Divergence between Admixed Latin American Populations: Antioquia and Choco, Colombia. Genome Biol Evol 12, 1516-1527.

17. Romano, J.D., Bernauer, M., McGrath, S.P., **Nagar, S.D.**, and Freimuth, R.R. (2019). A Decade of Translational Bioinformatics: A Retrospective Analysis of "Year-in-Review" Presentations. AMIA Jt Summits Transl Sci Proc 2019, 335-344.

REFERENCES

1 Chande, A. T. *Bioinformatic platforms and methods for worldwide polygenic risk scores*, Georgia Institute of Technology, (2020).

2 Clayton, E. *Global dysregulation of gene expression and tumorigenesis: Data science for cancer*, Georgia Institute of Technology, (2019).

3 Espitia Navarro, H. F. *Efficient alignment-free software applications for next generation sequencing-based molecular epidemiology*, Georgia Institute of Technology, (2020).

4 Medina Cordoba, L. K. *BIOFERTILIZERS FOR SUSTAINABLE AGRICULTURE: ISOLATION AND GENOMIC CHARACTERIZATION OF NITROGEN-FIXING BACTERIA FROM SUGARCANE*, Georgia Institute of Technology, (2020).

5 Medrano Trochez, C. *Transcriptomic profiling of human cells destined for therapeutic applications*, Georgia Institute of Technology, (2021).

6 Chakravarti, A. & Little, P. Nature, nurture and human disease. *Nature* **421**, 412-414, doi:10.1038/nature01401 (2003).

7 Tremblay, J. & Hamet, P. Environmental and genetic contributions to diabetes. *Metabolism* **100S**, 153952, doi:10.1016/j.metabol.2019.153952 (2019).

8 Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat Rev Genet* **19**, 175-185, doi:10.1038/nrg.2017.89 (2018).

9 Fullerton, S. M., Knerr, S. & Burke, W. Finding a place for genomics in health disparities research. *Public Health Genomics* **15**, 156-163, doi:10.1159/000334717 (2012).

10 Smith, C. E. *et al.* Using genetic technologies to reduce, rather than widen, health disparities. *Health Affairs* **35**, 1367-1373 (2016).

11 Braveman, P. Health disparities and health equity: concepts and measurement. *Annu Rev Public Health* **27**, 167-194, doi:10.1146/annurev.publhealth.27.021405.102103 (2006).

12 health, E.-U. D. o., services, h., Control, C. f. D., Prevention & Statistics, N. C. f. H. *Healthy people 2010: Final review*. (US Government Printing Office, 2012).

13 Jones, C. M. The moral problem of health disparities. *Am J Public Health* **100 Suppl 1**, S47-51, doi:10.2105/AJPH.2009.171181 (2010).

14 Adler, N. E. & Rehkopf, D. H. U.S. disparities in health: descriptions, causes, and mechanisms. *Annu Rev Public Health* **29**, 235-252, doi:10.1146/annurev.publhealth.29.020907.090852 (2008).

15 Braveman, P. & Gottlieb, L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep* **129 Suppl 2**, 19-31, doi:10.1177/00333549141291S206 (2014).

16 Hewitt, R. E. Biobanking: the foundation of personalized medicine. *Curr Opin Oncol* **23**, 112-119, doi:10.1097/CCO.0b013e32834161b8 (2011).

17 Kauffmann, F. & Cambon-Thomsen, A. Tracing biological collections: between books and clinical trials. *JAMA* **299**, 2316-2318, doi:10.1001/jama.299.19.2316 (2008).

18 Abul-Husn, N. S. & Kenny, E. E. Personalized Medicine and the Power of Electronic Health Records. *Cell* **177**, 58-69, doi:10.1016/j.cell.2019.02.039 (2019).

19 Evans, W. E. & Relling, M. V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487-491, doi:10.1126/science.286.5439.487 (1999).

20 Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdiscip Rev Syst Biol Med* **10**, e1417, doi:10.1002/wsbm.1417 (2018).

21 Khoury, M. J. *et al.* From public health genomics to precision public health: a 20-year journey. *Genet Med* **20**, 574-582, doi:10.1038/gim.2017.211 (2018).

22 Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision Public Health for the Era of Precision Medicine. *Am J Prev Med* **50**, 398-401, doi:10.1016/j.amepre.2015.08.031 (2016).

23 Weeramanthri, T. S. *et al.* Editorial: Precision Public Health. *Front Public Health* **6**, 121, doi:10.3389/fpubh.2018.00121 (2018).

24 Khoury, M. J. *Precision Public Health: What Is It?*, <https://blogs.cdc.gov/genomics/2018/05/15/precision-public-health-2/> (2018).

25 Bureau, U. C. *About Race*, <https://www.census.gov/topics/population/race/about.html> (2020).

26 Bureau, U. C. *About Hispanic Origin*, <https://www.census.gov/topics/population/hispanic-origin/about.html> (2020).

27 Service, G. S. *Ethnicity Harmonized Standard*, <https://gss.civilservice.gov.uk/policy-store/ethnicity/> (2021).

28 Mathieson, I. & Scally, A. What is ancestry? *PLoS Genet* **16**, e1008624, doi:10.1371/journal.pgen.1008624 (2020).

29 Yang, M. A. & Fu, Q. J. T. i. G. Insights into modern human prehistory using ancient genomes. **34**, 184-196 (2018).

30 Borrell, L. N. *et al.* Race and Genetic Ancestry in Medicine - A Time for Reckoning with Racism. *N Engl J Med* **384**, 474-480, doi:10.1056/NEJMms2029562 (2021).

31 Bergstrom, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, doi:10.1126/science.aay5012 (2020).

32 Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet* **105**, 763-772, doi:10.1016/j.ajhg.2019.08.012 (2019).

33 Yuan, J. *et al.* Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell* **34**, 549-560 e549, doi:10.1016/j.ccell.2018.08.019 (2018).

34 Witherspoon, D. J. *et al.* Genetic similarities within and between human populations. *Genetics* **176**, 351-359, doi:10.1534/genetics.106.067355 (2007).

35 Breathett, K. *et al.* The Groundwater of Racial and Ethnic Disparities Research: A Statement From Circulation: Cardiovascular Quality and Outcomes. *Circ Cardiovasc Qual Outcomes* **14**, e007868, doi:10.1161/CIRCOUTCOMES.121.007868 (2021).

36 Frank, R. What to make of it? The (Re) emergence of a biological conceptualization of race in health disparities research. *Social science & medicine* **64**, 1977-1983 (2007).

37 Fine, M. J., Ibrahim, S. A. & Thomas, S. B. (American Public Health Association, 2005).

38 West, K. M., Blacksher, E. & Burke, W. Genomics, Health Disparities, and Missed Opportunities for the Nation's Research Agenda. *JAMA* **317**, 1831-1832, doi:10.1001/jama.2017.3096 (2017).

39 Shields, A. E. *et al.* The use of race variables in genetic studies of complex traits and the goal of reducing health disparities: a transdisciplinary perspective. *American Psychologist* **60**, 77 (2005).

40 Sankar, P. *et al.* Genetic research and health disparities. *JAMA* **291**, 2985-2989, doi:10.1001/jama.291.24.2985 (2004).

41 Rotimi, C., Shriner, D. & Adeyemo, A. Genome science and health disparities: a growing success story? *Genome Med* **5**, 61, doi:10.1186/gm465 (2013).

42 Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163-165, doi:10.1038/475163a (2011).

43 Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161-164, doi:10.1038/538161a (2016).

44 Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet* **25**, 489-494, doi:10.1016/j.tig.2009.09.012 (2009).

45 Jameson, J. L. & Longo, D. L. Precision medicine--personalized, problematic, and promising. *N Engl J Med* **372**, 2229-2234, doi:10.1056/NEJMsb1503104 (2015).

46 Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* **372**, 793-795, doi:10.1056/NEJMp1500523 (2015).

47 Ma, Q. & Lu, A. Y. Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacol Rev* **63**, 437-459, doi:10.1124/pr.110.003533 (2011).

48 Weinshilboum, R. M. & Wang, L. Pharmacogenetics and pharmacogenomics: development, science, and translation. *Annu Rev Genomics Hum Genet* **7**, 223-245, doi:10.1146/annurev.genom.6.080604.162315 (2006).

49 Bachtiar, M. & Lee, C. G. Genetics of population differences in drug response. *Curr Genet Med Rep* **1**, 162-170 (2013).

50 Nordling, L. How the genomics revolution could finally help Africa. *Nature* **544**, 20-22, doi:10.1038/544020a (2017).

51 Bryc, K. *et al.* Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl 2**, 8954-8961, doi:10.1073/pnas.0914618107 (2010).

52 Moreno-Estrada, A. *et al.* Reconstructing the population genetic history of the Caribbean. *PLoS Genet* **9**, e1003925, doi:10.1371/journal.pgen.1003925 (2013).

53 Rishishwar, L. *et al.* Ancestry, admixture and fitness in Colombian genomes. *Sci Rep* **5**, 12376, doi:10.1038/srep12376 (2015).

54 Ruiz-Linares, A. *et al.* Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet* **10**, e1004572, doi:10.1371/journal.pgen.1004572 (2014).

55 Wang, S. *et al.* Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* **4**, e1000037, doi:10.1371/journal.pgen.1000037 (2008).

56 Appelbaum, N. P. *Mapping the Country of Regions: The Chorographic Commission of Nineteenth-Century Colombia*. (University of North Carollina Press, 2016).

57 Wade, P. in *Degrees of Mixture, Degrees of Freedom* Ch. 4, 99-121 (Duke University Press, 2017).

58 Conley, A. B. *et al.* A Comparative Analysis of Genetic Ancestry and Admixture in the Colombian Populations of Choco and Medellin. *G3 (Bethesda)* **7**, 3435-3447, doi:10.1534/g3.117.1118 (2017).

59 Medina-Rivas, M. A. *et al.* Choco, Colombia: a hotspot of human biodiversity. *Rev Biodivers Neotrop* **6**, 45-54, doi:10.18636/bioneotropical.v6i1.341 (2016).

60 Lakiotaki, K. *et al.* Exploring public genomics data for population pharmacogenomics. *PLoS One* **12**, e0182138, doi:10.1371/journal.pone.0182138 (2017).

61 Ramos, E. *et al.* Pharmacogenomics, ancestry and clinical decision making for global populations. *Pharmacogenomics J* **14**, 217-222, doi:10.1038/tpj.2013.24 (2014).

62 Hariprakash, J. M. *et al.* Pharmacogenetic landscape of DPYD variants in south Asian populations by integration of genome-scale data. *Pharmacogenomics* **19**, 227-241, doi:10.2217/pgs-2017-0101 (2018).

63 Bonifaz-Pena, V. *et al.* Exploring the distribution of genetic markers of pharmacogenomics relevance in Brazilian and Mexican populations. *PLoS One* **9**, e112640, doi:10.1371/journal.pone.0112640 (2014).

64 Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* **92**, 414-417, doi:10.1038/clpt.2012.96 (2012).

65 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

66 Chande, A. T. *et al.* Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Choco and Antioquia. *Sci Rep* **7**, 17127, doi:10.1038/s41598-017-17380-4 (2017).

67 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).

68 Team, R. C. (Vienna, Austria, 2013).

69 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).

70 O'Donnell-Luria, A. H. & Miller, D. T. A Clinician's perspective on clinical exome sequencing. *Hum Genet* **135**, 643-654, doi:10.1007/s00439-016-1662-x (2016).

71 Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215 (1988).

72 Andrews, S. *FastQC: a quality control tool for high throughput sequence data*, <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).

73 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

74 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

75 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

76 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

77 Wangkumhang, P. *et al.* WASP: a Web-based Allele-Specific PCR assay designing tool for detecting SNPs and mutations. *BMC Genomics* **8**, 275, doi:10.1186/1471-2164-8-275 (2007).

78 Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134, doi:10.1186/1471-2105-13-134 (2012).

79 Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res* **23**, 1514-1521, doi:10.1101/gr.154831.113 (2013).

80 GoDarts *et al.* Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. *Nat Genet* **43**, 117-120, doi:10.1038/ng.735 (2011).

81 Zhang, C. & Zhang, R. More effective glycaemic control by metformin in African Americans than in Whites in the prediabetic population. *Diabetes Metab* **41**, 173-175, doi:10.1016/j.diabet.2015.01.003 (2015).

82 Williams, L. K. *et al.* Differing effects of metformin on glycemic control by race-ethnicity. *J Clin Endocrinol Metab* **99**, 3160-3168, doi:10.1210/jc.2014-1539 (2014).

83 He, Y., Hoskins, J. M. & McLeod, H. L. Copy number variants in pharmacogenetic genes. *Trends Mol Med* **17**, 244-251, doi:10.1016/j.molmed.2011.01.007 (2011).

84 Roden, D. M. *et al.* Pharmacogenomics: challenges and opportunities. *Ann Intern Med* **145**, 749-757 (2006).

85 Fung, E. *et al.* Effect of genetic variants, especially CYP2C9 and VKORC1, on the pharmacology of warfarin. *Semin Thromb Hemost* **38**, 893-904, doi:10.1055/s-0032-1328891 (2012).

86 Johnson, J. A. *et al.* Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. *Clin Pharmacol Ther* **90**, 625-629, doi:10.1038/clpt.2011.185 (2011).

87 Lauschke, V. M., Milani, L. & Ingelman-Sundberg, M. Pharmacogenomic Biomarkers for Improved Drug Therapy-Recent Progress and Future Developments. *AAPS J* **20**, 4, doi:10.1208/s12248-017-0161-x (2017).

88 Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol* **17**, 157, doi:10.1186/s13059-016-1016-y (2016).

89 Karlberg, J. P. Globalization of sponsored clinical trials. *Nature Reviews Drug Discovery* **7**, 458 (2008).

90 Thiers, F. A., Sinskey, A. J. & Berndt, E. R. (Nature Publishing Group, 2008).

91 Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635-649, doi:10.1016/j.ajhg.2017.03.004 (2017).

92 Conrad, D. F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**, 1251-1260, doi:10.1038/ng1911 (2006).

93 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

94 Mora, G. C. *Making Hispanics: How activists, bureaucrats, and media constructed a new American*. (University of Chicago Press, 2014).

95 De Castro, M. & Restrepo, C. M. Genetics and genomic medicine in Colombia. *Mol Genet Genomic Med* **3**, 84-91, doi:10.1002/mgg3.139 (2015).

96 Szabo, L. in *Kaiser Health News* (Kaiser Family Foundation, Menlo Park, California, 2018).

97 Gibson, G. Going to the negative: genomics for optimized medical prescription. *Nat Rev Genet*, doi:10.1038/s41576-018-0061-7 (2018).

98 Gallo, J. E. Current state of cardiovascular genomics in Colombia. *Revista Colombiana de Cardiología* **24**, 1-2 (2017).

99 Yasuda, S. U., Zhang, L. & Huang, S. M. The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. *Clin Pharmacol Ther* **84**, 417-423, doi:10.1038/clpt.2008.141 (2008).

100 Huang, S. M. & Temple, R. Is this the drug or dose for you? Impact and consideration of ethnic factors in global drug development, regulatory review, and clinical practice. *Clin Pharmacol Ther* **84**, 287-294, doi:10.1038/clpt.2008.144 (2008).

101 Chen, M. L. Ethnic or racial differences revisited: impact of dosage regimen and dosage form on pharmacokinetics and pharmacodynamics. *Clin Pharmacokinet* **45**, 957-964, doi:10.2165/00003088-200645100-00001 (2006).

102 Bjornsson, T. D. *et al.* A review and assessment of potential sources of ethnic differences in drug responsiveness. *J Clin Pharmacol* **43**, 943-967, doi:10.1177/0091270003256065 (2003).

103 Ramamoorthy, A., Pacanowski, M. A., Bull, J. & Zhang, L. Racial/ethnic differences in drug disposition and response: review of recently approved drugs. *Clin Pharmacol Ther* **97**, 263-273, doi:10.1002/cpt.61 (2015).

104 Risch, N., Burchard, E., Ziv, E. & Tang, H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* **3**, comment2007, doi:10.1186/gb-2002-3-7-comment2007 (2002).

105 Cooper, R. S., Kaufman, J. S. & Ward, R. Race and genomics. *N Engl J Med* **348**, 1166-1170, doi:10.1056/NEJMsb022863 (2003).

106 Caulfield, T. *et al.* Race and ancestry in biomedical research: exploring the challenges. *Genome Med* **1**, 8, doi:10.1186/gm8 (2009).

107 Burchard, E. G. *et al.* The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* **348**, 1170-1175, doi:10.1056/NEJMsb025007 (2003).

108 Montagu, A. *Man's most dangerous myth: The fallacy of race*. (Rowman & Littlefield, 1997).

109 Graves Jr, J. L. *The emperor's new clothes: Biological theories of race at the millennium*. (Rutgers University Press, 2003).

110 Saini, A. *Superior: the return of race science*. (Beacon Press, 2019).

111 Lee, S. S., Mountain, J. & Koenig, B. A. The meanings of "race" in the new genomics: implications for health disparities research. *Yale J Health Policy Law Ethics* **1**, 33-75 (2001).

112 Braun, L. Reifying human difference: the debate on genetics, race, and health. *Int J Health Serv* **36**, 557-573, doi:10.2190/8JAF-D8ED-8WPD-J9WH (2006).

113 Gannett, L. The biological reification of race. *Brit J Philos Sci* **55**, 323-345, doi:DOI 10.1093/bjps/55.2.323 (2004).

114 Ackerman, R. *et al.* AAPA statement on biological aspects of race. *American Journal of Physical Anthropology* **101**, 569-570 (1996).

115 Graves Jr, J. L. in *Race and the Genetic Revolution: Science, Myth, and Culture* (eds S. Krimsky & K. Sloan) 142-170 (Columbia University Press, 2011).

116 Graves Jr, J. L. Why the nonexistence of biological races does not mean the nonexistence of racism. *American Behavioral Scientist* **59**, 1474-1495 (2015).

117 Graves Jr, J. L. Great is their sin: Biological determinism in the age of genomics. *The Annals of the American Academy of Political and Social Science* **661**, 24-50 (2015).

118 Graves Jr, J. L. in *Reconsidering Race: Social Science Perspectives on Racial Categories in the Age of Genomics* (eds K. Suzuki & D.A. Von Vacano) 21-31 (Oxford University Press, 2018).

119 Yudell, M., Roberts, D., DeSalle, R. & Tishkoff, S. Taking race out of human genetics. *Science* **351**, 564-565, doi:10.1126/science.aac4951 (2016).

120 Nagar, S. D. *et al.* Population Pharmacogenomics for Precision Public Health in Colombia. *Front Genet* **10**, 241, doi:10.3389/fgene.2019.00241 (2019).

121 Ahsan, T., Urmi, N. J. & Sajib, A. A. Heterogeneity in the distribution of 159 drug-response related SNPs in world populations and their genetic relatedness. *PLoS One* **15**, e0228000, doi:10.1371/journal.pone.0228000 (2020).

122 Mizzi, C. *et al.* A European Spectrum of Pharmacogenomic Biomarkers: Implications for Clinical Pharmacogenomics. *PLoS One* **11**, e0162866, doi:10.1371/journal.pone.0162866 (2016).

123 Kirzinger, A., Neuman, T., Cubanski, J. & Brodie, M. *Prescription drugs and older adults*, <https://www.kff.org/health-reform/issue-brief/data-note-prescription-drugs-and-older-adults/> (2019).

124 Bureau, U. C. (2010).

125 Sonnega, A. *et al.* Cohort Profile: the Health and Retirement Study (HRS). *Int J Epidemiol* **43**, 576-585, doi:10.1093/ije/dyu067 (2014).

126 Jordan, I. K., Rishishwar, L. & Conley, A. B. Native American admixture recapitulates population-specific migration and settlement of the continental United States. *PLoS Genet* **15**, e1008225, doi:10.1371/journal.pgen.1008225 (2019).

127 Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104, doi:10.1126/science.1153717 (2008).

128 Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370-374, doi:10.1038/nature11258 (2012).

129 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

130 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).

131 Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* **93**, 278-288, doi:10.1016/j.ajhg.2013.06.020 (2013).

132 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. **12**, 2825-2830 (2011).

133 Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583-589 (1992).

134 Bland, J. M. & Altman, D. G. Statistics notes. The odds ratio. *BMJ* **320**, 1468, doi:10.1136/bmj.320.7247.1468 (2000).

135 Altman, D. G. & Andersen, P. K. Calculating the number needed to treat for trials where the outcome is time to an event. *BMJ* **319**, 1492-1495, doi:10.1136/bmj.319.7223.1492 (1999).

136 Humes, K. R., Jones, N. A. & Ramirez, R. R. *Overview of Race and Hisapnic Origin*, <https://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf> (2011).

137 Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet* **96**, 37-53, doi:10.1016/j.ajhg.2014.11.010 (2015).

138 Lotsch, J. *et al.* Cross-sectional analysis of the influence of currently known pharmacogenetic modulators on opioid therapy in outpatient pain centers. *Pharmacogenet Genomics* **19**, 429-436, doi:10.1097/fpc.0b013e32832b89da (2009).

139 Uher, R. *et al.* Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. *Am J Psychiatry* **167**, 555-564, doi:10.1176/appi.ajp.2009.09070932 (2010).

140 Tanaka, S. *et al.* DPP6 as a candidate gene for neuroleptic-induced tardive dyskinesia. *Pharmacogenomics J* **13**, 27-34, doi:10.1038/tpj.2011.36 (2013).

141 Tang, H. *et al.* Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet* **76**, 268-275, doi:10.1086/427888 (2005).

142 Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 786-791, doi:10.1073/pnas.0909559107 (2010).

143 Baharian, S. *et al.* The Great Migration and African-American Genomic Diversity. *PLoS Genet* **12**, e1006059, doi:10.1371/journal.pgen.1006059 (2016).

144 Wright, S. Isolation by distance. *Genetics* **28**, 114-138 (1943).

145 Cavalli-Sforza, L. L. *The history and geography of human genes*. (Princeton University Press, 1994).

146 Rosenberg, N. A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* **1**, e70, doi:10.1371/journal.pgen.0010070 (2005).

147 Serre, D. & Paabo, S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res* **14**, 1679-1685, doi:10.1101/gr.2529604 (2004).

148 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

149 Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381-2385, doi:10.1126/science.1078311 (2002).

150 Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15**, R159-160, doi:10.1016/j.cub.2005.02.038 (2005).

151 Handley, L. J., Manica, A., Goudet, J. & Balloux, F. Going the distance: human population genetics in a clinal world. *Trends Genet* **23**, 432-439, doi:10.1016/j.tig.2007.07.002 (2007).

152 Domingue, B. W., Fletcher, J., Conley, D. & Boardman, J. D. Genetic and educational assortative mating among US adults. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 7996-8000, doi:10.1073/pnas.1321426111 (2014).

153 Schwartz, C. R. Trends and variation in assortative mating: Causes and consequences. *Annual Review of Sociology* **39**, 451-470 (2013).

154 Newbeck, P. *Virginia Hasn't Always Been for Lovers: Interracial Marriage Bans and the Case of Richard and Mildred Loving*. (Southern Illinois University Press, 2008).

155 Wang, W. *The rise of intermarriage*, <https://www.pewsocialtrends.org/2012/02/16/the-rise-of-intermarriage/> (2012).

156 Lewontin, R. C. in *Evolutionary Biology* (eds T.H. Dobzhansky, M.K. Hecht, & W.C. Steere) 381-398 (Springer, 1972).

157 Barbujani, G. & Di Benedetto, G. Genetic variances within and between human groups. *Genes, Fossils and Behaviour*, 63-77 (2001).

158 Brown, R. A. & Armelagos, G. J. Apportionment of racial diversity: a review. *Evolutionary Anthropology* **10**, 34-40 (2001).

159 Excoffier, L. & Hamilton, G. Comment on "Genetic structure of human populations". *Science* **300**, 1877; author reply 1877, doi:10.1126/science.1083411 (2003).

160 Long, J. C. & Kittles, R. A. Human genetic diversity and the nonexistence of biological races. *Hum Biol* **75**, 449-471, doi:10.1353/hub.2003.0058 (2003).

161 Ruvolo, M. & Seielstad, M. in *Thinking about Evolution: Historical, Philosophical, and Political Perspectives* (eds R.S Singh, C.B. Krimbas, D.B. Paul, & J. Beatty) 141-151 (Cambridge: Cambridge University Press, 2001).

162 Edwards, A. W. Human genetic diversity: Lewontin's fallacy. *Bioessays* **25**, 798-801, doi:10.1002/bies.10315 (2003).

163 Rosenberg, N. A. in *Phylogenetic Inference, Selection Theory, and History of Science: Selected Papers of AWF Edwards with Commentaries* 399-403 (Cambridge University Press, 2018).

164 Team, d. *NCBI dbSNP*, <https://www.ncbi.nlm.nih.gov/feed/rss.cgi?ChanKey=dbsnpnews> (2019).

165 Ng, P. C., Zhao, Q., Levy, S., Strausberg, R. L. & Venter, J. C. Individual genomes instead of race for personalized medicine. *Clin Pharmacol Ther* **84**, 306-309, doi:10.1038/clpt.2008.114 (2008).

166 All of Us Research Program, I. *et al.* The "All of Us" Research Program. *N Engl J Med* **381**, 668-676, doi:10.1056/NEJMsr1809937 (2019).

167 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

168 Elliott, P., Peakman, T. C. & Biobank, U. K. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* **37**, 234-244, doi:10.1093/ije/dym276 (2008).

169 Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026-1034, doi:10.1093/aje/kwx246 (2017).

170 Statistics, O. f. N., Scotland, N. R. o., Statistics, N. I. & Agency, R. 2011 Census aggregate data. UK Data Service (Edition: June 2016). (2016).

171 showcase, U. B. d. *Data-Field 21003: Age when attended assessment centre*, <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21003> (2020).

172 showcase, U. B. d. *Data-Field 54: UK Biobank assessment centre*, <https://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=54> (2020).

173 showcase, U. B. d. *Data-Field 21000: Ethnic background*, <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21000> (2020).

174 showcase, U. B. d. *Data-Field 41270: Diagnoses - ICD10*, <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41270> (2020).

175 showcase, U. B. d. *Data-Field 31: Sex*, <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=31> (2020).

176 showcase, U. B. d. *Data-Field 189: Townsend deprivation index at recruitment*, <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=189> (2020).

177 Townsend, P., Phillimore, P. & Beattie, A. *Health and deprivation: inequality and the North*. (Routledge, 1988).

178 Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375-2376, doi:10.1093/bioinformatics/btu197 (2014).

179 Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325, doi:10.2196/14325 (2019).

180 McKinney, W. in *Proceedings of the 9th Python in Science Conference.* 51-56 (Austin, TX).

181 Wickham, H. Elegant graphics for data analysis. *Media* **35**, 10.1007 (2009).

182 Hossain, S., Calloway, C., Lippa, D., Niederhut, D. & Shupe, D. in *Proceedings of the 18th Python in Science Conference.* 126-133.

183 Leff, A. & Rayfield, J. T. in *Proceedings fifth ieee international enterprise distributed object computing conference.* 118-127 (IEEE).

184 Hu, F. B., Satija, A. & Manson, J. E. Curbing the Diabetes Pandemic: The Need for Global Policy Solutions. *JAMA* **313**, 2319-2320, doi:10.1001/jama.2015.5287 (2015).

185 Whicher, C. A., O'Neill, S. & Holt, R. I. G. Diabetes in the UK: 2019. *Diabet Med* **37**, 242-247, doi:10.1111/dme.14225 (2020).

186 Goff, L. M. Ethnicity and Type 2 diabetes in the UK. *Diabet Med* **36**, 927-938, doi:10.1111/dme.13895 (2019).

187 Gov.uk. *Ethnicity facts and figures*, <https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups> (

188 Kertzer, D., Arel, D. in *The Politics of Race, Ethnicity, and Language in National Censuses* (ed D.I.; Arel Kertzer, D.) 1-42 (Cambridge University Press, 2002).

189 Cuschieri, S. The genetic side of type 2 diabetes - A review. *Diabetes Metab Syndr* **13**, 2503-2506, doi:10.1016/j.dsx.2019.07.012 (2019).

190 Hill-Briggs, F. *et al.* Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care*, doi:10.2337/dci20-0053 (2020).

191 Piccolo, R. S., Subramanian, S. V., Pearce, N., Florez, J. C. & McKinlay, J. B. Relative Contributions of Socioeconomic, Local Environmental, Psychosocial, Lifestyle/Behavioral, Biophysiological, and Ancestral Factors to Racial/Ethnic Disparities in Type 2 Diabetes. *Diabetes Care* **39**, 1208-1217, doi:10.2337/dc15-2255 (2016).

192 Piccolo, R. S., Pearce, N., Araujo, A. B. & McKinlay, J. B. The contribution of biogeographical ancestry and socioeconomic status to racial/ethnic disparities in type 2 diabetes mellitus: results from the Boston Area Community Health Survey. *Ann Epidemiol* **24**, 648-654, 654 e641, doi:10.1016/j.annepidem.2014.06.098 (2014).

193 Link, C. L. & McKinlay, J. B. Disparities in the prevalence of diabetes: is it race/ethnicity or socioeconomic status? Results from the Boston Area Community Health (BACH) survey. *Ethn Dis* **19**, 288-292 (2009).

194 Spencer Bonilla, G., Rodriguez-Gutierrez, R. & Montori, V. M. What We Don't Talk About When We Talk About Preventing Type 2 Diabetes-Addressing Socioeconomic Disadvantage. *JAMA Intern Med* **176**, 1053-1054, doi:10.1001/jamainternmed.2016.2952 (2016).

195 Volaco, A., Cavalcanti, A. M., Filho, R. P. & Precoma, D. B. Socioeconomic Status: The Missing Link Between Obesity and Diabetes Mellitus? *Curr Diabetes Rev* **14**, 321-326, doi:10.2174/1573399813666170621123227 (2018).

196 Espelt, A. *et al.* Socioeconomic position and type 2 diabetes mellitus in Europe 1999-2009: a panorama of inequalities. *Curr Diabetes Rev* **7**, 148-158, doi:10.2174/157339911795843131 (2011).

197 Thomas, C. *et al.* Socio-economic position and type 2 diabetes risk factors: patterns in UK children of South Asian, black African-Caribbean and white European origin. *PLoS One* **7**, e32619, doi:10.1371/journal.pone.0032619 (2012).

198 Nishino, Y., Gilmour, S. & Shibuya, K. Inequality in diabetes-related hospital admissions in England by socioeconomic deprivation and ethnicity: facility-based cross-sectional analysis. *PLoS One* **10**, e0116689, doi:10.1371/journal.pone.0116689 (2015).

199 Chande, A. T. *et al.* The Phenotypic Consequences of Genetic Divergence between Admixed Latin American Populations: Antioquia and Choco, Colombia. *Genome Biol Evol* **12**, 1516-1527, doi:10.1093/gbe/evaa154 (2020).

200 Campbell, D. D. *et al.* Amerind ancestry, socioeconomic status and the genetics of type 2 diabetes in a Colombian population. *PLoS One* **7**, e33570, doi:10.1371/journal.pone.0033570 (2012).

201 Signorello, L. B. *et al.* Comparing diabetes prevalence between African Americans and Whites of similar socioeconomic status. *Am J Public Health* **97**, 2260-2267, doi:10.2105/AJPH.2006.094482 (2007).

202 Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* **18**, 26, doi:10.1186/s12864-016-3391-x (2017).

203 showcase, U. B. d. *Data-Field 22009: Genetic principal components*, <https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=22009> (2020).

204 Foster, H. M. E. *et al.* The effect of socioeconomic deprivation on the association between an extended measurement of unhealthy lifestyle factors and health outcomes: a prospective analysis of the UK Biobank cohort. *Lancet Public Health* **3**, e576-e585, doi:10.1016/S2468-2667(18)30200-7 (2018).

205 Campello, R. J., Moulavi, D. & Sander, J. in *Pacific-Asia conference on knowledge discovery and data mining.* 160-172 (Springer).

206 forestmodel: Forest Plots from Regression Models. v. 0.6.2 (2020).

207 Azen, R. & Traxel, N. Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics* **34**, 319-347 (2009).

208 Clogg, C. C., Petkova, E. & Haritou, A. Statistical methods for comparing regression coefficients between models. *American journal of sociology* **100**, 1261-1293 (1995).

209 Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T. & Sidorchuk, A. Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. *Int J Epidemiol* **40**, 804-818, doi:10.1093/ije/dyr029 (2011).

210 van Zon, S. K., Snieder, H., Bultmann, U. & Reijneveld, S. A. The interaction of socioeconomic position and type 2 diabetes mellitus family history: a cross-sectional analysis of the Lifelines Cohort and Biobank Study. *BMJ Open* **7**, e015275, doi:10.1136/bmjopen-2016-015275 (2017).

211 Smith, N. R., Kelly, Y. J. & Nazroo, J. Y. The effects of acculturation on obesity rates in ethnic minorities in England: evidence from the Health Survey for England. *Eur J Public Health* **22**, 508-513, doi:10.1093/eurpub/ckr070 (2012).

212 Misra, A. Ethnic-Specific Criteria for Classification of Body Mass Index: A Perspective for Asian Indians and American Diabetes Association Position Statement. *Diabetes Technol Ther* **17**, 667-671, doi:10.1089/dia.2015.0007 (2015).

213 Pepys, M. B. & Hirschfield, G. M. C-reactive protein: a critical update. *J Clin Invest* **111**, 1805-1812, doi:10.1172/JCI18921 (2003).

214 Black, S., Kushner, I. & Samols, D. C-reactive Protein. *J Biol Chem* **279**, 48487-48490, doi:10.1074/jbc.R400025200 (2004).

215 Dehghan, A. *et al.* Genetic variation, C-reactive protein levels, and incidence of diabetes. *Diabetes* **56**, 872-878, doi:10.2337/db06-0922 (2007).

216 Danesh, J. *et al.* C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease. *N Engl J Med* **350**, 1387-1397, doi:10.1056/NEJMoa032804 (2004).

217 Wium-Andersen, M. K., Orsted, D. D., Nielsen, S. F. & Nordestgaard, B. G. Elevated C-reactive protein levels, psychological distress, and depression in 73, 131 individuals. *JAMA Psychiatry* **70**, 176-184, doi:10.1001/2013.jamapsychiatry.102 (2013).

218 Zacho, J., Tybjaerg-Hansen, A. & Nordestgaard, B. G. C-reactive protein and all-cause mortality--the Copenhagen City Heart Study. *Eur Heart J* **31**, 1624-1632, doi:10.1093/eurheartj/ehq103 (2010).

219 Nazmi, A. & Victora, C. G. Socioeconomic and racial/ethnic differentials of C-reactive protein levels: a systematic review of population-based studies. *BMC Public Health* **7**, 212, doi:10.1186/1471-2458-7-212 (2007).

220 Wong, N. D., Pio, J., Valencia, R. & Thakal, G. Distribution of C-reactive protein and its relation to risk factors and coronary heart disease risk estimation in the National Health and Nutrition Examination Survey (NHANES) III. *Prev Cardiol* **4**, 109-114, doi:10.1111/j.1520-037x.2001.00570.x (2001).

221 Matthews, K. A. *et al.* Ethnic differences in cardiovascular risk factor burden among middle-aged women: Study of Women's Health Across the Nation (SWAN). *Am Heart J* **149**, 1066-1073, doi:10.1016/j.ahj.2004.08.027 (2005).

222 Ford, E. S. Does exercise reduce inflammation? Physical activity and C-reactive protein among U.S. adults. *Epidemiology* **13**, 561-568, doi:10.1097/00001648-200209000-00012 (2002).

223 Danner, M., Kasl, S. V., Abramson, J. L. & Vaccarino, V. Association between depression and elevated C-reactive protein. *Psychosom Med* **65**, 347-356, doi:10.1097/01.psy.0000041542.29808.01 (2003).

224 Alley, D. E. *et al.* Socioeconomic status and C-reactive protein levels in the US population: NHANES IV. *Brain Behav Immun* **20**, 498-504, doi:10.1016/j.bbi.2005.10.003 (2006).

225 Abramson, J. L., Weintraub, W. S. & Vaccarino, V. Association between pulse pressure and C-reactive protein among apparently healthy US adults. *Hypertension* **39**, 197-202, doi:10.1161/hy0202.104270 (2002).

226 Lakoski, S. G. *et al.* Gender and C-reactive protein: data from the Multiethnic Study of Atherosclerosis (MESA) cohort. *Am Heart J* **152**, 593-598, doi:10.1016/j.ahj.2006.02.015 (2006).

227 Khera, A. *et al.* Race and gender differences in C-reactive protein levels. *J Am Coll Cardiol* **46**, 464-469, doi:10.1016/j.jacc.2005.04.051 (2005).

228 Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-1295, doi:10.1534/genetics.115.178616 (2015).

229 Nagar, S. D., Nápoles, A. M., Jordan, I. K. & Mariño-Ramírez, L. Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the United Kingdom. *EClinicalMedicine*, 100960 (2021).

230 Choudhry, S. *et al.* Dissecting complex diseases in complex populations: asthma in latino americans. *Proc Am Thorac Soc* **4**, 226-233, doi:10.1513/pats.200701-029AW (2007).

231 Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* **98**, 456-472, doi:10.1016/j.ajhg.2015.12.022 (2016).

232 Kennedy, N. Forestmodel: forest plots from regression models. *R Package Version 0.6* **2** (2020).

233 Woods-Giscombe, C. L. Superwoman schema: African American women's views on stress, strength, and health. *Qual Health Res* **20**, 668-683, doi:10.1177/1049732310361892 (2010).

234 Farmer, H. R., Wray, L. A. & Haas, S. A. Race, Gender, and Socioeconomic Variations in C-Reactive Protein Using the Health and Retirement Study. *J Gerontol B Psychol Sci Soc Sci* **76**, 583-595, doi:10.1093/geronb/gbaa027 (2021).

235 Bauer, G. R. Incorporating intersectionality theory into population health research methodology: challenges and the potential to advance health equity. *Soc Sci Med* **110**, 10-17, doi:10.1016/j.socscimed.2014.03.022 (2014).

236 Bowleg, L. The problem with the phrase women and minorities: intersectionality-an important theoretical framework for public health. *Am J Public Health* **102**, 1267-1273, doi:10.2105/AJPH.2012.300750 (2012).

237 Richardson, L. J. & Brown, T. H. (En)gendering Racial Disparities in Health Trajectories: A Life Course and Intersectional Analysis. *SSM Popul Health* **2**, 425-435, doi:10.1016/j.ssmph.2016.04.011 (2016).

238 Brown, T. H. & Hargrove, T. W. Multidimensional approaches to examining gender and racial/ethnic stratification in health. *Women, Gender, and Families of Color* **1**, 180-206 (2013).

239 Pingault, J. B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nat Rev Genet* **19**, 566-580, doi:10.1038/s41576-018-0020-3 (2018).

240 Davey Smith, G. *et al.* Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* **366**, 1484-1498, doi:10.1016/S0140-6736(05)67601-5 (2005).

241 Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044, doi:10.1126/science.1172257 (2009).

242 Sanders-Phillips, K., Settles-Reaves, B., Walker, D. & Brownlow, J. Social inequality and racial discrimination: risk factors for health disparities in children of color. *Pediatrics* **124 Suppl 3**, S176-186, doi:10.1542/peds.2009-1100E (2009).

243 Lewis, T. T., Aiello, A. E., Leurgans, S., Kelly, J. & Barnes, L. L. Self-reported experiences of everyday discrimination are associated with elevated C-reactive protein levels in older African-American adults. *Brain Behav Immun* **24**, 438-443, doi:10.1016/j.bbi.2009.11.011 (2010).

244 Beatty, D. L., Matthews, K. A., Bromberger, J. T. & Brown, C. Everyday Discrimination Prospectively Predicts Inflammation Across 7-Years in Racially Diverse Midlife Women: Study of Women's Health Across the Nation. *J Soc Issues* **70**, 298-314, doi:10.1111/josi.12061 (2014).

245 Inglis, V., Ball, K. & Crawford, D. Why do women of low socioeconomic status have poorer dietary behaviours than women of higher socioeconomic status? A qualitative exploration. *Appetite* **45**, 334-343, doi:10.1016/j.appet.2005.05.003 (2005).

246 Feinstein, J. S. The relationship between socioeconomic status and health: a review of the literature. *Milbank Q* **71**, 279-322 (1993).

247 Govil, S. R., Weidner, G., Merritt-Worden, T. & Ornish, D. Socioeconomic status and improvements in lifestyle, coronary risk factors, and quality of life: the Multisite Cardiac Lifestyle Intervention Program. *Am J Public Health* **99**, 1263-1270, doi:10.2105/AJPH.2007.132852 (2009).