Population structure and pharmacogenomic risk stratification in the United States

Shashwat Deepali Nagar, [1,2] Andrew B. Conley, [1,2,3] and I. King Jordan [1,2,3]

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, United States; [2]PanAmerican Bioinformatics Institute, Cali, Colombia; [3]IHRC-Georgia Tech Applied Bioinformatics Laboratory, Atlanta, GA, United States

*Correspondence: king.jordan@biology.gatech.edu

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, and Hispanic. 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. Nevertheless, 97% of PGx variation is found within rather than between groups. Examples of highly differentiated PGx variants illustrate how SIRE partitions PGx variation based on group-specific ancestry patterns and contains valuable information for risk stratification. Finally, we show that individuals who identify as Black or Hispanic benefit more when SIRE is considered for treatment decisions than individuals from the majority White population.

Introduction

Pharmacogenomic (PGx) variants are associated with inter-individual differences in drug exposure and response, affecting medication dosage, efficacy and toxicity[1; 2]. A number of studies have shown racial and/or ethnic differences in drug response[3-7], based in part on group-specific differences in the frequencies of PGx variants[8]. 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[7]. 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[9-12].

Critiques of the use of racial and ethnic categories in biomedical research point to the appalling history of race science[13-15] and stress the potential of such research to reify outmoded notions of racial difference[16-18]. This school of thought holds that race is a primarily a social construct with little or no biological (genetic) meaning[19-23]. 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[24]. 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[25-30].

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[31], and restricted our study to the three largest racial/ethnic groups in the US: White, Black (or African-American), and Hispanic/Latino[32]. Our study cohort is made up of 8,629 participants from the Health and Retirement Study (HRS)[33], 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

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)[33]. 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.

Genetic Ancestry (GA) Analysis

HRS participants were previously genotyped at ~2,381,000 genomic sites using the Illumina Omni2.5 BeadChip[33]. 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 (Supplementary Table 1)[34]. Reference populations were taken from (i) the 1000 Genomes Project (648)[35], (ii) the Human Genome Diversity Project (110)[36], and (iii) 21 Native American populations from across the Americas (90)[37]. A custom script that employs PLINK version 1.9[38] 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[39]. 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[34; 40] 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[41] 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[25]. 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.

Measurement of PGx Variation

Single nucleotide variants (SNVs) associated with pharmacogenomic response – *i.e.* PGx variants – were mined from the Pharmacogenomic Knowledgebase (PharmGKB)[2]. 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. PharmGKB annotates the specific effect alleles that are associated with inter-individual differences in drug dosage, efficacy, and toxicity. The direction of effect (higher or lower) is specific to individual PGx variants for dosage and efficacy, 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, $\bar{H}_S = \frac{1}{2}\sum_i 2(p_i)(1-p_i)(\frac{count_i}{total\ count})$, 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, $H_T = 2(\bar{p})(1-\bar{p})$, 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

$F_{ST} = 1 - \frac{\bar{H}_S}{H_T}$ [42]. 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[43]. 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 the formula $OR = \frac{P_E/Q_E}{Q_N/Q_N}$, with confidence intervals calculated as $CI = \exp\left(log(OR) \pm Z_{\alpha/2} * SE_{log(OR)}\right)$, where α is 0.05, $Z_{\alpha/2}$ is 1.6, and $SE_{log(OR)} = \sqrt{\frac{1}{P_E} + \frac{1}{P_N} + \frac{1}{Q_E} + \frac{1}{Q_N}}$. Similarly, using group-specific PGx effect counts the absolute risk increase (ARI) was calculated as $ARI = \frac{P_E}{P_E + P_A} - \frac{Q_E}{Q_E + Q_A}$, with confidence intervals calculated as $CI = ARI \pm Z_{\alpha/2} \times SE_{ARI}$, where α is 0.05, $Z_{\alpha/2}$ is 1.96, and $SE_{ARI} = \sqrt{P_E P_A + Q_E Q_A}$ [44]. Group-specific genotype prediction accuracy values were calculated as $Accuracy = (TP + TN)/(TP + TN + FP + FN)$, 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 $CI = Accuracy \pm Z_{\alpha/2} \times \sqrt[2]{\frac{Error_{prediction}}{1-Error_{prediction}}/N}$, where $Error_{prediction} = \frac{FP+FN}{TP+TN+FP+FN}$ and $N = TP + TN + FP + FN$. 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): $Prevalence_{overall} = Count_{EA}/Count_{Total}$, 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 calculate as: $PPV_A = Count^A_{EA}/Count^A_{Total}$, where $Count^A_{EA}$ is the count of the effect allele/genotype in population *A* and $Count^A_{Total}$ is the total count of alleles/genotypes at that locus in the population *A*. Information gain is then calculated as: $InfoGain_A = |PPV_A - Prevalence_{overall}|$.

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.

Table 1. **Demographic description for the cohort used in this study.**

	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)

[1]Number (Percentage)
[2]Median age in years (Confidence intervals)

Results

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 1). 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[45]. 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%[45].

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[46]. Permutation analysis was used to confirm the stability of the resulting GA groups and their robustness to changes in sample size (Supplementary Figure 1). The distributions of continental ancestry fractions were compared for the three SIRE groups – White, Black, and Hispanic – and the three GA groups (Figure 1).

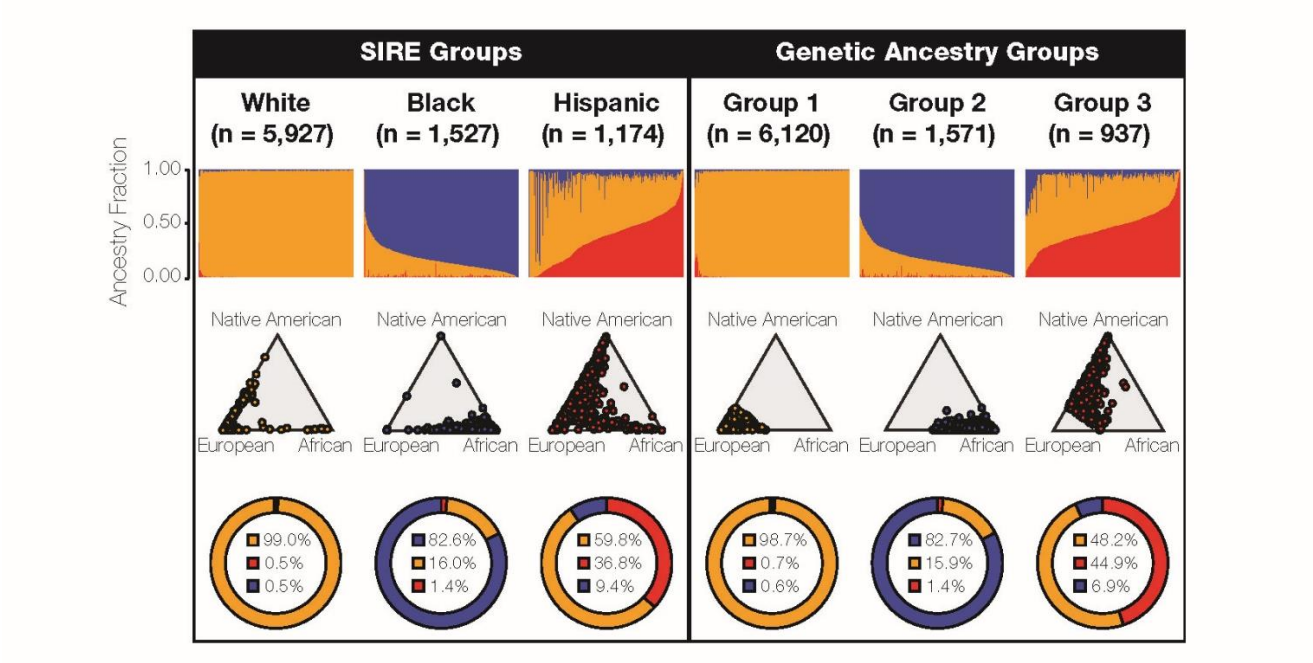


Figure 1. **Race, ethnicity, and genetic ancestry in the US**. Continental genetic ancestry patterns are shown for self-identified race/ethnicity (SIRE) and genetic ancestry (GA) groups: European ancestry (orange), African ancestry (blue), and Native American ancestry (red). HRS cohort participants are 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.

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 1 – 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 Supplementary Figure 2. 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 1 – 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 (Supplementary Figure 3). 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.

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 2, and results for the analogous comparison of GA groups are shown in Supplementary Figure 4. 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 2A-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-Hispanic group comparisons (Figure 2D-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 2G). 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 2H). Genome-wide patterns of PGx variation characterized in this way show 96.1% correspondence to SIRE group labels (Figure 2I).

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 3). 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 3A). 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 3B and 3C). 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[43]. 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. Detailed results for all PGx variants analyzed here can be found in Supplementary Table 2.

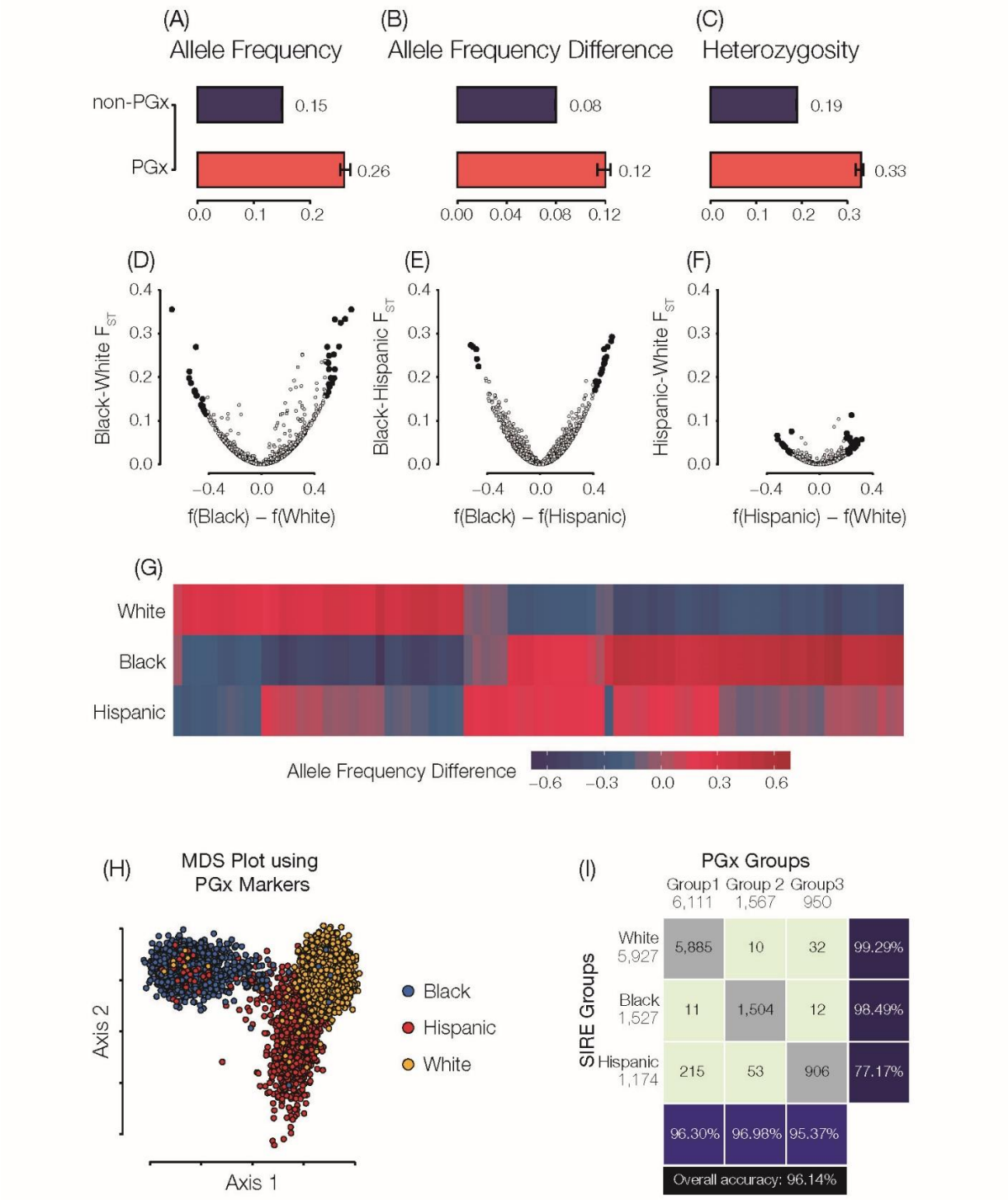


Figure 2. **Pharmacogenomic variation in the US.** Genome-wide average allele frequencies (A), group-specific allele frequency differences (B), and heterozygosity fractions (C) are shown for PGx variants (red) compared to non-PGx variants (blue). (D-F) Fixation index (FST; y-axis) and allele frequency 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 Supplementary Figure 4.

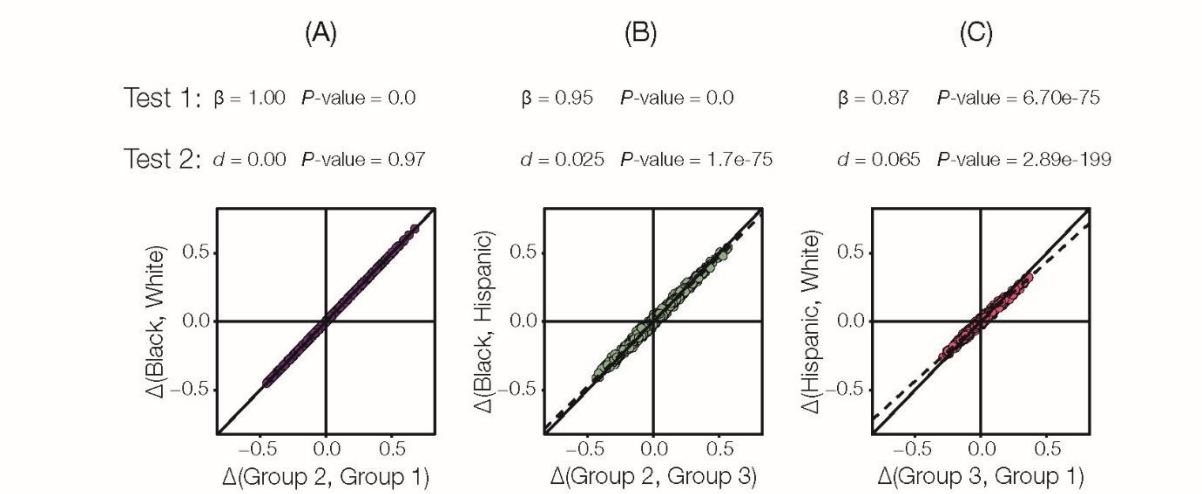


Figure 3. **Self-identified race/ethnicity (SIRE) versus genetic ancestry (GA) for partitioning pharmacogenomic (PGx) variation.** (A-C) Regression of pairwise PGx variant effect allele frequency differences calculated using SIRE (y-axis) versus the corresponding GA groups (x-axis). Results of two statistical 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.

Examples of highly differentiated PGx variants are shown in Table 2 and Figure 4. The relative percentages of PGx effect (above) and non-effect (below) alleles across SIRE groups reveal the extent of differentiation for these variants (Figure 4A), and the observed allele frequency differences are associated with SIRE group-specific continental ancestry fractions (Figure 4B-D). Nevertheless, as described above and shown in Figure 2, even highly differentiated PGx variants show levels of F_{ST} that indicate substantially more within than between group variation (see pie charts in Figure 4B-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) encoding gene is associated with a decreased fentanyl opioid dose requirement[47] (Figure 4B). 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[48] (Figure 4C). 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 4D). 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 [49]. 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.

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 2G and Figure 4) 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 4C), 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 antipsycohtic toxicity (Figure 4D). 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 5). 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 2A) 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).

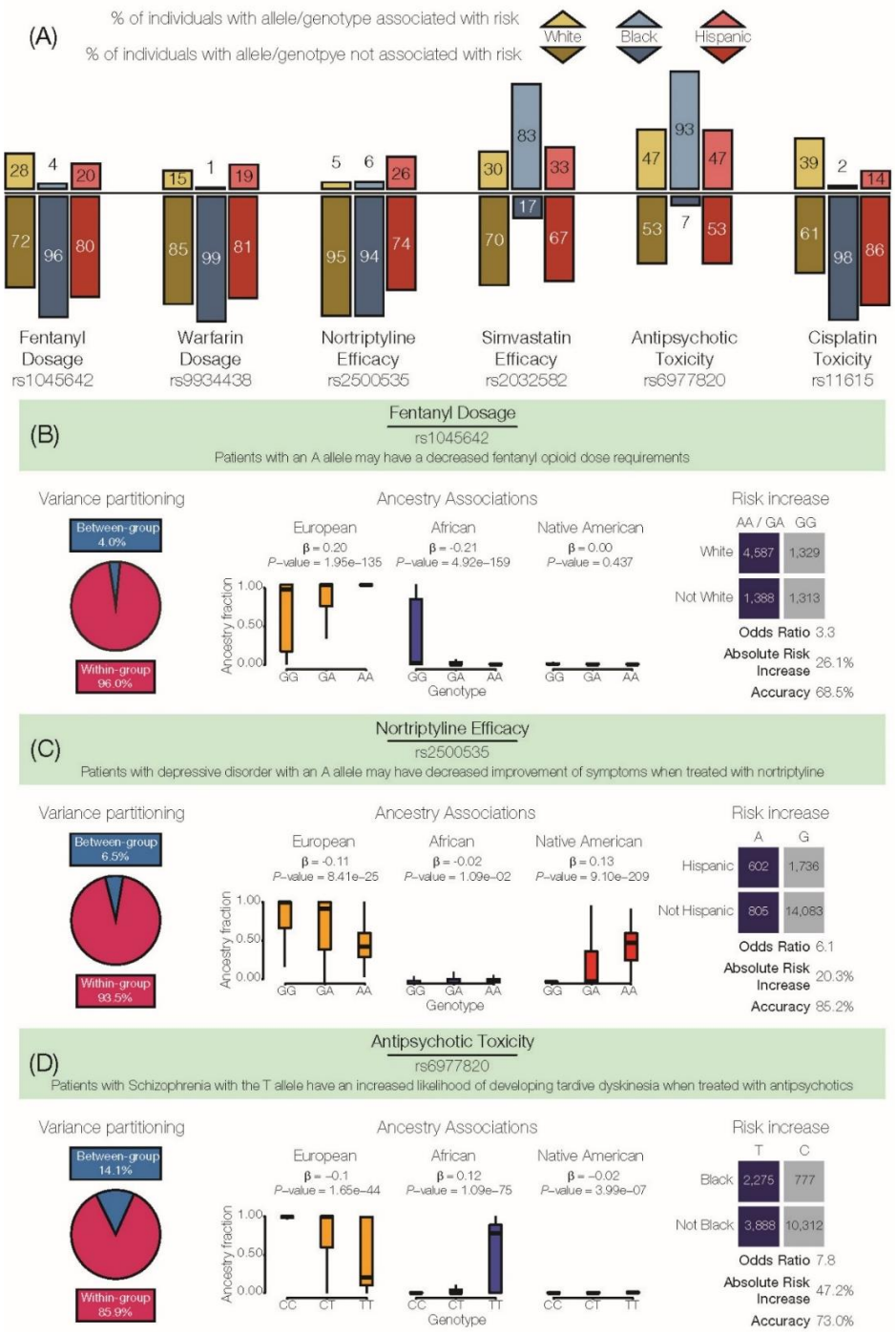


Figure 4. **Examples of highly differentiated pharmacogenomic (PGx) variants.** (A) SIRE group percentages of effect (above axis) versus non-effect (below axis) alleles/genotypes are shown for six highly differentiated PGx variants. Allele counts are used for the additive PGx effect mode, and genotype 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

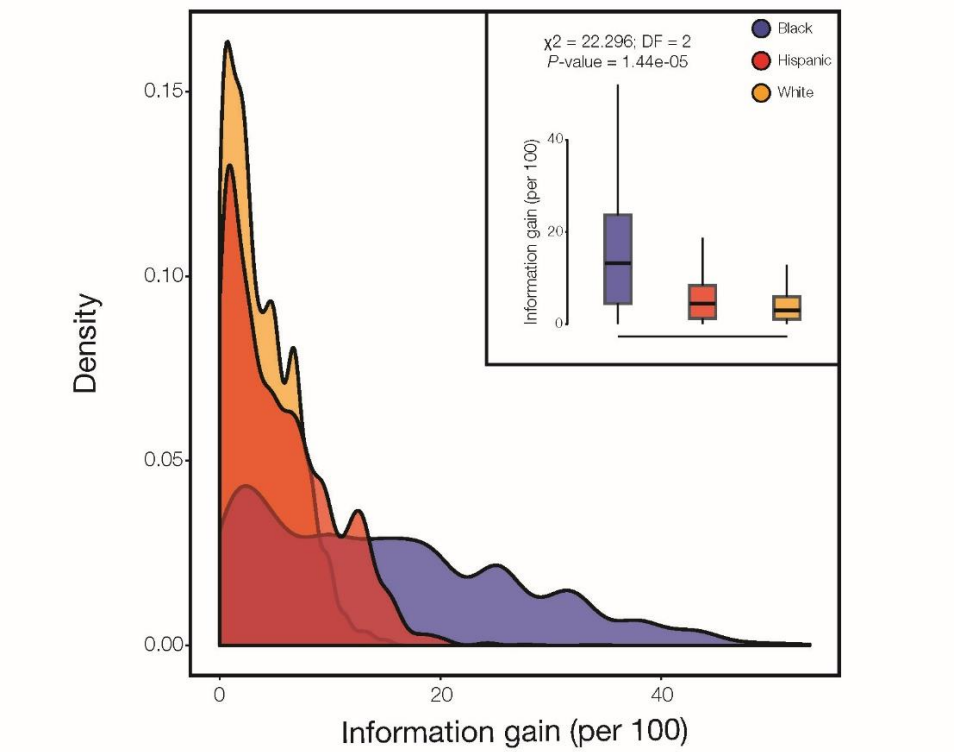


Figure 5. Information gained when SIRE is used for PGx stratification. The amount of information gained per 100 individuals is the number additional correct PGx variant predictions made when SIRE is used to stratify the population. Information gain is calculated for all PGx variants in each SIRE group, 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).

Discussion

Concordance Between SIRE and GA in the US

The SIRE and GA groups from the US analyzed here show >96% overall concordance (Figure 1, Supplementary Figures 2 and 3). 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 groups[50], and a 2007 study reported 100% classification accuracy of individuals from geographically separated population groups when thousands of genetic variants were used for clustering[51]. 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[52], and a massive study of >200,000 individuals from the Million Veterans Program found >99.4% concordance between SIRE and GA[53]. 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[34; 46; 54-56]. 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[57] best accounts for the apportionment of human genetic diversity among global populations[58]. Populations that are physically distant, or separated by major geographic barriers, are more genetically diverged than nearby populations[59]. 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[60]. 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[35; 61] and the Human Genome Diversity Project[36; 62]. 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[63; 64].

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[34; 46]. 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 1, Supplementary Figures 2 and 3). 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[65; 66] (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 1 and Supplementary Figure 2). 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[67].

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 1). We chose to focus on older Americans given their disproportionate use of prescription medications[31], 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[68]. Rates of intermarriage among SIRE groups have increased substantially since that era[69], 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.

Table 2. **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)
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)
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)
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)

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[70], and has since been confirmed by numerous studies of populations defined by GA using larger-scale analyses[62; 71-75]. 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[76; 77]. 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 2 and 4).

We used a standard evidence based medicine analytical framework[43; 44] 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 2 and Figure 4). 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[78]. 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[35]. Nevertheless, the results reported here underscore the potential clinical relevance for the small the minority of genetic variants that show relatively high levels of between-group divergence.

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[2], 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[2], 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[2], suggesting that our analytical approach captures most of the known variant-drug associations.

The current and future utility of race and ethnicity in pharmacogenomics

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[79]. 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 4D), 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[80], such as the NIH All of Us project[81], 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[82; 83]. 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 5). 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.

Supplemental Data

Supplemental data included two tables and four figures.

Declaration of Interests

The authors declare no competing interests

Acknoweldgements

The authors thank Dr. Greg Gibson for comments on a manuscript draft and Dr. Joe Lachance for helful discussion.

References

1. Evans, W.E., and Relling, M.V. (1999). Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286, 487-491.
2. Barbarino, J.M., Whirl-Carrillo, M., Altman, R.B., and Klein, T.E. (2018). PharmGKB: A worldwide resource for pharmacogenomic information. Wiley Interdiscip Rev Syst Biol Med 10, e1417.
3. Yasuda, S.U., Zhang, L., and Huang, S.M. (2008). The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. Clin Pharmacol Ther 84, 417-423.
4. Huang, S.M., and Temple, R. (2008). 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.
5. Chen, M.L. (2006). Ethnic or racial differences revisited: impact of dosage regimen and dosage form on pharmacokinetics and pharmacodynamics. Clin Pharmacokinet 45, 957-964.
6. Bjornsson, T.D., Wagner, J.A., Donahue, S.R., Harper, D., Karim, A., Khouri, M.S., Murphy, W.R., Roman, K., Schneck, D., Sonnichsen, D.S., et al. (2003). A review and assessment of potential sources of ethnic differences in drug responsiveness. J Clin Pharmacol 43, 943-967.
7. Ramamoorthy, A., Pacanowski, M.A., Bull, J., and Zhang, L. (2015). Racial/ethnic differences in drug disposition and response: review of recently approved drugs. Clin Pharmacol Ther 97, 263-273.
8. Bachtiar, M., and Lee, C.G. (2013). Genetics of population differences in drug response. Curr Genet Med Rep 1, 162-170.
9. Risch, N., Burchard, E., Ziv, E., and Tang, H. (2002). Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3, comment2007.
10. Cooper, R.S., Kaufman, J.S., and Ward, R. (2003). Race and genomics. N Engl J Med 348, 1166-1170.
11. Caulfield, T., Fullerton, S.M., Ali-Khan, S.E., Arbour, L., Burchard, E.G., Cooper, R.S., Hardy, B.J., Harry, S., Hyde-Lay, R., Kahn, J., et al. (2009). Race and ancestry in biomedical research: exploring the challenges. Genome Med 1, 8.
12. Burchard, E.G., Ziv, E., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Perez-Stable, E.J., Sheppard, D., and Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 348, 1170-1175.
13. Montagu, A. (1997). Man's most dangerous myth: The fallacy of race.(Lanham: Rowman & Littlefield).
14. Graves Jr, J.L. (2003). The emperor's new clothes: Biological theories of race at the millennium.(New Brunswick: Rutgers University Press).
15. Saini, A. (2019). Superior: the return of race science.(Boston: Beacon Press).
16. Lee, S.S., Mountain, J., and Koenig, B.A. (2001). The meanings of "race" in the new genomics: implications for health disparities research. Yale J Health Policy Law Ethics 1, 33-75.
17. Braun, L. (2006). Reifying human difference: the debate on genetics, race, and health. Int J Health Serv 36, 557-573.
18. Gannett, L. (2004). The biological reification of race. Brit J Philos Sci 55, 323-345.
19. Ackerman, R., Athreya, S., Bolnick, D., Fuentes, A., Lasisi, T., Lee, S.H., McLean, S.A., and Nelson, R. (1996). AAPA statement on biological aspects of race. American Journal of Physical Anthropology 101, 569-570.
20. Graves Jr, J.L. (2011). Evolutionary Versus Racial Medicine: Why It Matters. In Race and the Genetic Revolution: Science, Myth, and Culture, S. Krimsky and K. Sloan, eds. (New York, NY, Columbia University Press), pp 142-170.
21. Graves Jr, J.L. (2015). Why the nonexistence of biological races does not mean the nonexistence of racism. American Behavioral Scientist 59, 1474-1495.
22. Graves Jr, J.L. (2015). 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.
23. Graves Jr, J.L. (2018). Biological theories of race beyond the millenium. In Reconsidering Race: Social Science Perspectives on Racial Categories in the Age of Genomics, K. Suzuki and D.A. Von Vacano, eds. (Oxford, UK, Oxford University Press), pp 21-31.
24. Yudell, M., Roberts, D., DeSalle, R., and Tishkoff, S. (2016). Taking race out of human genetics. Science 351, 564-565.
25. 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.

26. Ahsan, T., Urmi, N.J., and Sajib, A.A. (2020). Heterogeneity in the distribution of 159 drug-response related SNPs in world populations and their genetic relatedness. PLoS One 15, e0228000.
27. Hariprakash, J.M., Vellarikkal, S.K., Keechilat, P., Verma, A., Jayarajan, R., Dixit, V., Ravi, R., Senthivel, V., Kumar, A., Sehgal, P., et al. (2018). Pharmacogenetic landscape of DPYD variants in south Asian populations by integration of genome-scale data. Pharmacogenomics 19, 227-241.
28. Lakiotaki, K., Kanterakis, A., Kartsaki, E., Katsila, T., Patrinos, G.P., and Potamias, G. (2017). Exploring public genomics data for population pharmacogenomics. PLoS One 12, e0182138.
29. Bonifaz-Pena, V., Contreras, A.V., Struchiner, C.J., Roela, R.A., Furuya-Mazzotti, T.K., Chammas, R., Rangel-Escareno, C., Uribe-Figueroa, L., Gomez-Vazquez, M.J., McLeod, H.L., et al. (2014). Exploring the distribution of genetic markers of pharmacogenomics relevance in Brazilian and Mexican populations. PLoS One 9, e112640.
30. Ramos, E., Doumatey, A., Elkahloun, A.G., Shriner, D., Huang, H., Chen, G., Zhou, J., McLeod, H., Adeyemo, A., and Rotimi, C.N. (2014). Pharmacogenomics, ancestry and clinical decision making for global populations. Pharmacogenomics J 14, 217-222.
31. Kirzinger, A., Neuman, T., Cubanski, J., and Brodie, M. (2019). Prescription drugs and older adults. In. (San Francisco, Kaiser Family Foundation.
32. Bureau, U.C. (2010). Quick facts: United States. In. (
33. Sonnega, A., Faul, J.D., Ofstedal, M.B., Langa, K.M., Phillips, J.W., and Weir, D.R. (2014). Cohort Profile: the Health and Retirement Study (HRS). Int J Epidemiol 43, 576-585.
34. Jordan, I.K., Rishishwar, L., and Conley, A.B. (2019). Native American admixture recapitulates population-specific migration and settlement of the continental United States. PLoS Genet 15, e1008225.
35. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
36. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104.
37. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature 488, 370-374.
38. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.
39. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10, 5-6.
40. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet 93, 278-288.
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V.J.J.o.m.l.r. (2011). Scikit-learn: Machine learning in Python. 12, 2825-2830.
42. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics 132, 583-589.
43. Bland, J.M., and Altman, D.G. (2000). Statistics notes. The odds ratio. BMJ 320, 1468.
44. Altman, D.G., and Andersen, P.K. (1999). Calculating the number needed to treat for trials where the outcome is time to an event. BMJ 319, 1492-1495.
45. Humes, K.R., Jones, N.A., and Ramirez, R.R. (2011). Overview of Race and Hisapnic Origin. In. (Washington, DC, US Census Bureau.
46. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet 96, 37-53.
47. Lotsch, J., von Hentig, N., Freynhagen, R., Griessinger, N., Zimmermann, M., Doehring, A., Rohrbacher, M., Sittl, R., and Geisslinger, G. (2009). Cross-sectional analysis of the influence of currently known pharmacogenetic modulators on opioid therapy in outpatient pain centers. Pharmacogenet Genomics 19, 429-436.
48. Uher, R., Perroud, N., Ng, M.Y., Hauser, J., Henigsberg, N., Maier, W., Mors, O.,

Placentino, A., Rietschel, M., Souery, D., et al. (2010). Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. Am J Psychiatry 167, 555-564.

49. Tanaka, S., Syu, A., Ishiguro, H., Inada, T., Horiuchi, Y., Ishikawa, M., Koga, M., Noguchi, E., Ozaki, N., Someya, T., et al. (2013). DPP6 as a candidate gene for neuroleptic-induced tardive dyskinesia. Pharmacogenomics J 13, 27-34.
50. Tang, H., Quertermous, T., Rodriguez, B., Kardia, S.L., Zhu, X., Brown, A., Pankow, J.S., Province, M.A., Hunt, S.C., Boerwinkle, E., et al. (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 76, 268-275.
51. Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A., and Jorde, L.B. (2007). Genetic similarities within and between human populations. Genetics 176, 351-359.
52. Yuan, J., Hu, Z., Mahal, B.A., Zhao, S.D., Kensler, K.H., Pi, J., Hu, X., Zhang, Y., Wang, Y., Jiang, J., et al. (2018). Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. Cancer Cell 34, 549-560 e549.
53. Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T.L., Huang, J., Vujkovic, M., Damrauer, S.M., Pyarajan, S., Gaziano, J.M., et al. (2019). Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. Am J Hum Genet 105, 763-772.
54. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc Natl Acad Sci U S A 107 Suppl 2, 8954-8961.
55. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.M., Wambebe, C., Tishkoff, S.A., et al. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A 107, 786-791.
56. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., et al. (2016). The Great Migration and African-American Genomic Diversity. PLoS Genet 12, e1006059.
57. Wright, S. (1943). Isolation by distance. Genetics 28, 114-138.
58. Cavalli-Sforza, L.L. (1994). The history and geography of human genes.(Princeton: Princeton University Press).
59. Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., and Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1, e70.
60. Serre, D., and Paabo, S. (2004). Evidence for gradients of human genetic diversity within and among continents. Genome Res 14, 1679-1685.
61. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
62. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science 298, 2381-2385.
63. Prugnolle, F., Manica, A., and Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. Curr Biol 15, R159-160.
64. Handley, L.J., Manica, A., Goudet, J., and Balloux, F. (2007). Going the distance: human population genetics in a clinal world. Trends Genet 23, 432-439.
65. Domingue, B.W., Fletcher, J., Conley, D., and Boardman, J.D. (2014). Genetic and educational assortative mating among US adults. Proc Natl Acad Sci U S A 111, 7996-8000.
66. Schwartz, C.R. (2013). Trends and variation in assortative mating: Causes and consequences. Annual Review of Sociology 39, 451-470.
67. Mora, G.C. (2014). Making Hispanics: How Activists, Bureaucrats, and Media Constructed a New American.(Chicago: University of Chicago Press).
68. Newbeck, P. (2008). Virginia Hasn't Always Been for Lovers: Interracial Marriage Bans and the Case of Richard and Mildred Loving.(Carbondale: Southern Illinois University Press).
69. Wang, W. (2012). The rise of intermarriage. In. (Washingtoin, DC, Pew Research Center.
70. Lewontin, R.C. (1972). The apportionment of human diversity. In Evolutionary Biology, T.H. Dobzhansky, M.K. Hecht, andW.C. Steere, eds. (New York, NY, Springer), pp 381-398.

71. Barbujani, G., and Di Benedetto, G. (2001). Genetic variances within and between human groups. Genes, Fossils and Behaviour, 63-77.
72. Brown, R.A., and Armelagos, G.J. (2001). Apportionment of racial diversity: a review. Evolutionary Anthropology 10, 34-40.
73. Excoffier, L., and Hamilton, G. (2003). Comment on "Genetic structure of human populations". Science 300, 1877; author reply 1877.
74. Long, J.C., and Kittles, R.A. (2003). Human genetic diversity and the nonexistence of biological races. Hum Biol 75, 449-471.
75. Ruvolo, M., and Seielstad, M. (2001). The apportionment of human diversity: 25 years later. In Thinking about Evolution: Historical, Philosophical, and Political Perspectives, R.S. Singh, C.B. Krimbas, D.B. Paul, andJ. Beatty, eds. (Cambridge: Cambridge University Press), pp 141-151.
76. Edwards, A.W. (2003). Human genetic diversity: Lewontin's fallacy. Bioessays 25, 798-801.
77. Rosenberg, N.A. (2018). Variance-partitioning and classification in human population genetics. In Phylogenetic Inference, Selection Theory, and History of Science: Selected Papers of AWF Edwards with Commentaries. (Cambridge, UK, Cambridge University Press), pp 399-403.
78. Team, d. (2019). NCBI dbSNP. In. (
79. Ng, P.C., Zhao, Q., Levy, S., Strausberg, R.L., and Venter, J.C. (2008). Individual genomes instead of race for personalized medicine. Clin Pharmacol Ther 84, 306-309.
80. Abul-Husn, N.S., and Kenny, E.E. (2019). Personalized Medicine and the Power of Electronic Health Records. Cell 177, 58-69.
81. All of Us Research Program, I., Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The "All of Us" Research Program. N Engl J Med 381, 668-676.
82. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature 538, 161-164.
83. Petrovski, S., and Goldstein, D.B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol 17, 157.

Supplementary Material

Population structure and pharmacogenomic risk stratification in the United States

Shashwat Deepali Nagar, Andrew B. Conley, and I. King Jordan

Table of Contents

Supplementary Table S1. **Global reference populations used for genetic ancestry inference.**

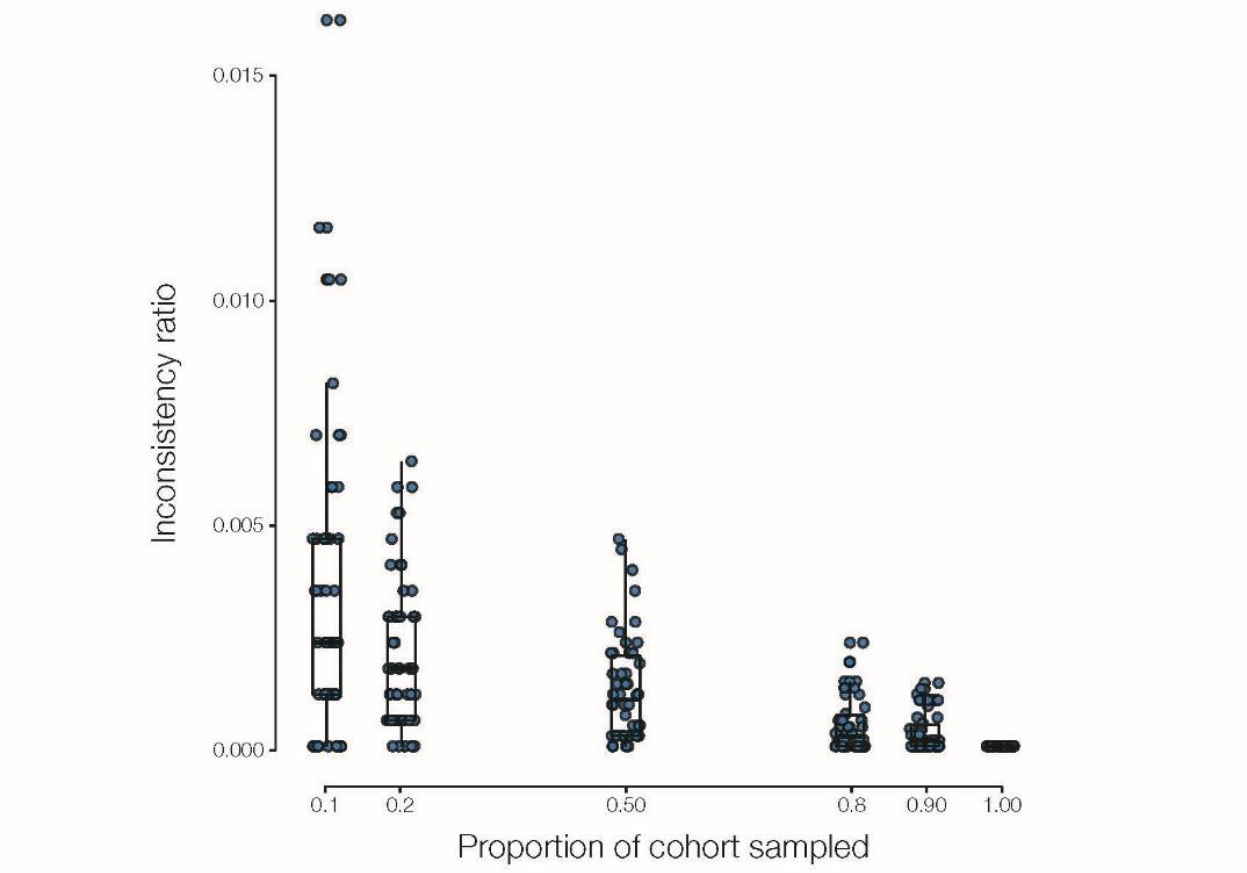
Population[1]	N[2]	Continental ancestry[3]	Source[4]
African Caribbean in Barbados	94	African	1KGP
Algonquin	5	Native American	Reich et al.
Americans of African ancestry from SW USA	51	African	1KGP
Utah Residents with Northern and Western European Ancestry	99	European	1KGP
Chipewyan	13	Native American	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

[1]Population name

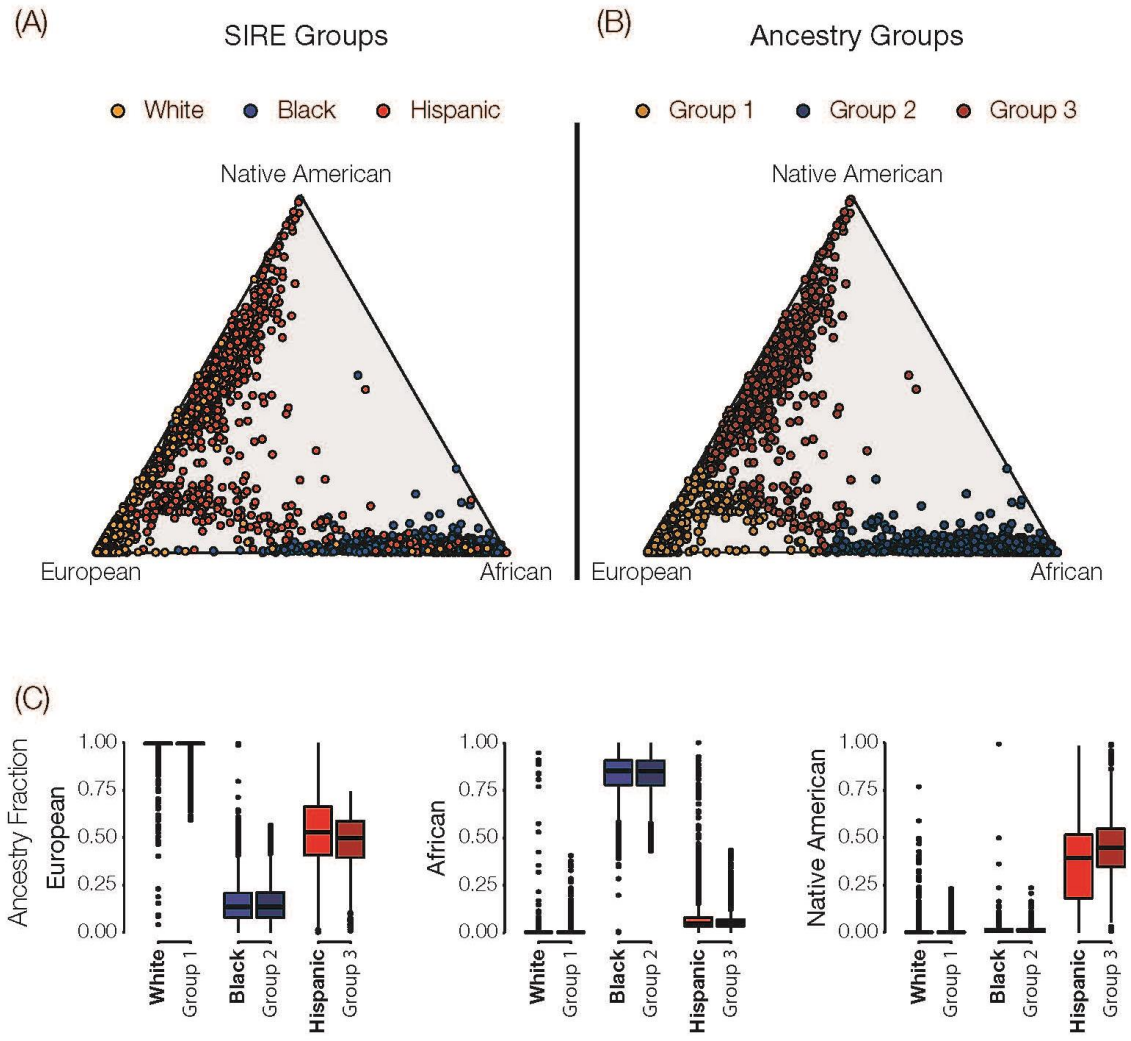
[2]Number of samples

[3]Population continental ancestry

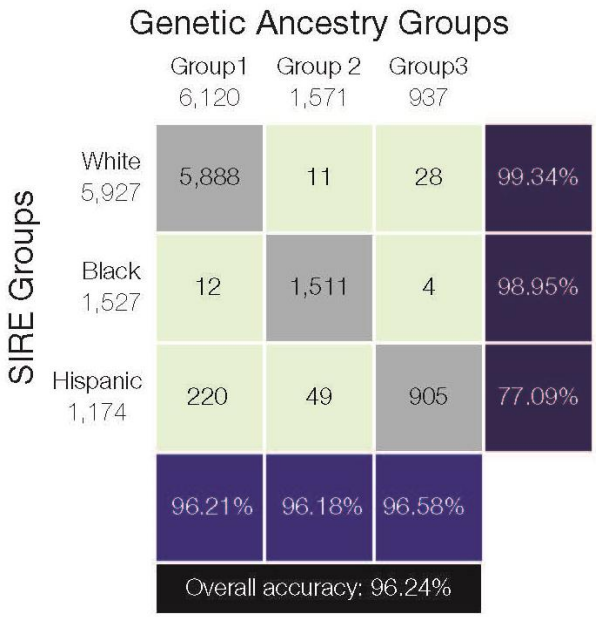
[4]Data source: 1000 Genomes Project (1KGP) [1], Human Genome Diversity Project (HGDP) [2], Collection of Native American Samples (Reich et al.)[3].



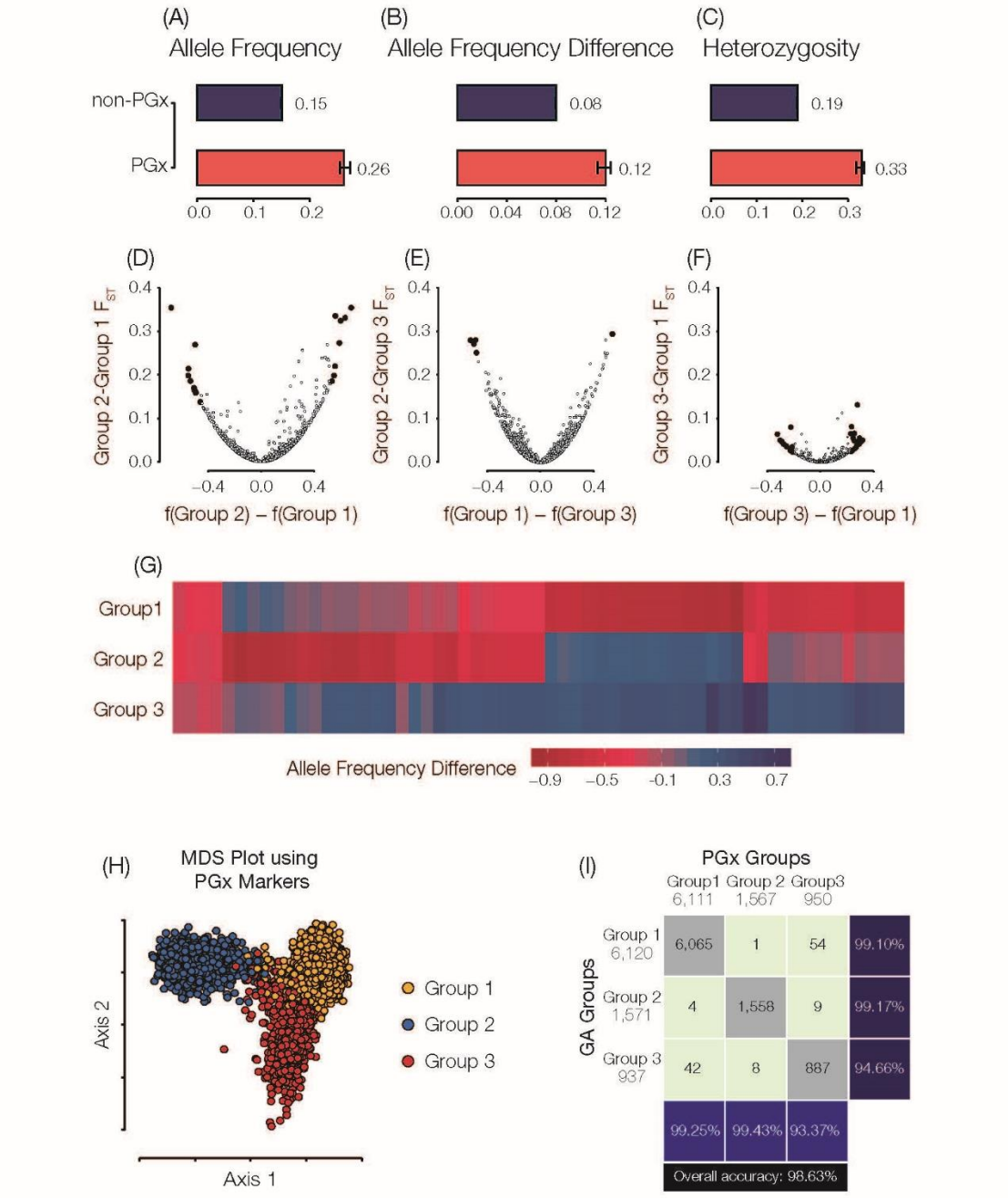
Supplementary Figure 1. **Permutation analysis to evaluate the stability of *k*-means genetic ancestry (GA) clusters.** The HRS cohort was randomly sampled at different proportions, where the proportion of the cohort sampled = the number of participants in the random sample / the total number of participants 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.



Supplementary Figure 2. **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.



Supplementary Figure 3. **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.



Supplementary Figure 4. **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.

References

1. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
2. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104.
3. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature 488, 370-374.