Received: 2 August 2024 | Revised: 2 August 2024 | Accepted: 10 August 2024

DOI: 10.1111/cts.70017

ARTICLE

Patterns of pharmacogenetic variation in nine biogeographic groups

Sophia Hernandez | **Lucia A. Hindorff** | **Joannella Morales** | **Erin M. Ramos** | **Teri A. Manolio**

National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA

Correspondence
Teri A. Manolio, Division of Genomic Medicine, National Human Genome Research Institute, 6700B Rockledge Drive, Room 3118, MSC 6908, Bethesda, MD 20892-6908, USA.
Email: manolio@nih.gov

Abstract

Frequencies of pharmacogenetic (PGx) variants are known to differ substantially across populations but much of the available PGx literature focuses on one or a few population groups, often defined in nonstandardized ways, or on a specific gene or variant. Guidelines produced by the Clinical Pharmacogenetic Implementation Consortium (CPIC) provide consistent methods of literature extraction, curation, and reporting, including comprehensive curation of allele frequency data across nine defined "biogeographic groups" from the PGx literature. We extracted data from 23 CPIC guidelines encompassing 19 genes to compare the sizes of the populations from each group and allele frequencies of altered function alleles across groups. The European group was the largest in the curated literature for 16 of the 19 genes, while the American and Oceanian groups were the smallest. Nearly 200 alleles were detected in nonreference groups that were not reported in the largest (reference) group. The genes *CYP2B6* and *CYP2C9* were *more* likely to have higher frequencies of altered function alleles in nonreference groups compared to the reference group, while the genes *CYP4F2*, *DPYD*, *SLCO1B1*, and *UGT1A1* were *less* likely to have higher frequencies in nonreference groups. PGx allele frequencies and function differ substantially across nine biogeographic groups, all but two of which are underrepresented in available PGx data. Awareness of these differences and increased efforts to characterize the breadth of global PGx variation are needed to ensure that implementation of PGx-guided drug selection does not further widen existing health disparities among populations currently underrepresented in PGx data.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Current knowledge includes the recognition of differences in frequencies in pharmacogenetic sequence variants across populations, though much of the current literature focuses on one or a few population groups defined in different ways, or on one or a few specific genes or variants.

WHAT QUESTION DID THIS STUDY ADDRESS?

This study compiled and compared pharmacogenetic allele frequencies curated systematically by CPIC and PharmGKB from the published literature and large biobanks across nine defined population groups in a consistent and comprehensive way, to identify differences across groups that may be clinically significant.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

This study sheds light on the representation of diverse population groups within the curated guidelines available for pharmacogenomic reference. The study highlights certain genes and populations where lack of information could potentially lead to clinically important outcomes.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

Recognizing the pharmacogenomic differences across populations that can have clinical implications is critical to applying pharmacogenetic guidelines in an accurate and equitable way, at least until guidelines can be derived from genomic data with adequate representation of diverse groups. If these differences go unrecognized there is a possibility of the widening of health disparities that already exist within medicine.

INTRODUCTION

More than 2.2 million serious adverse drug reactions (ADRs) are estimated to occur in hospitalized patients, leading to over 100,000 deaths annually.[1] In addition to the toll of ADRs, an estimated 10%–45% of patients will experience limited or no efficacy from their treatment.[2] It is hoped that pharmacogenetic (PGx) implementation into clinical practice will improve drug safety and efficacy and create more customizable and patient-tailored health care.[3] Although many ADRs can be difficult to predict, genetic variation in pharmacogenes, a subset of genes influencing drug metabolism and response,[3] has been implicated in a significant number of them. Guidelines derived from peer-reviewed published studies are available for drug dosing based on the presence of these variants,[4] but those studies were conducted primarily in populations of European ancestry, despite frequencies of these variants being known to differ widely across ancestral groups.[5] Important allele frequency differences or population-specific variants remaining unrecognized could reduce the accuracy of dosing algorithms in these populations and worsen health disparities already experienced by these groups as PGx guidelines become more widely adopted.

Marked differences across population groups in allele frequencies of variations in pharmacogenes have been identified and several have been well studied. *HLA-B*1502*, for example, is associated with carbamazepine and oxcarbazepine hypersensitivity and can result in Stevens–Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN) on exposure to these drugs.[2] The *HLA-B*1502* haplotype has an allele frequency of roughly 22% in Asian populations while being very rare outside of Asia.[6] Along with variation among continental populations, variation within continental ancestry groups has also been recognized. This same *HLA-B*1502* haplotype is found at a frequency of approximately 50% in populations from the Philippines while having a <0.1% frequency in Japan.[2] While differences in frequencies in PGx genes across multiple populations are widely recognized and acknowledged, much of the current literature focuses on one or a few population groups defined in different ways, or on a specific gene or variant. Several analyses of pharmacogenomic allele frequencies across populations within large biobanks have been conducted.[7–9] However, a comprehensive comparison of allele frequencies derived from the primary literature and large databases such as the Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org/) using consistent methods across a large number of pharmacogenes and population groups has not to our knowledge been published.

The Clinical Pharmacogenetic Implementation Consortium (CPIC)[10] guideline development process includes the estimation of allele frequencies extracted systematically from the published literature[11] using standardized "biogeographic" groupings covering the worldwide populations.[12] These groupings are defined primarily by geographic origin and are used here to be consistent with CPIC curation and reporting; below they are described for ease of reference as simply "groups." We compiled the most recent CPIC allele frequency tables

available as of June 20, 2023, for 19 pharmacogenes with published CPIC guidelines to compare frequencies of altered function alleles across these groups. In this paper, we will: (1) document and compare sizes of the populations from each group used to generate allele frequency data for each of 19 pharmacogenes; (2) quantify the allele frequencies of altered function alleles and compare them across groups; and (3) consider the impact this could have for populations and PGx-directed clinical care.

METHODS

CPIC provides peer-reviewed and evidence-based guidelines for actionable PGx variants in a standardized format for each gene-drug (or gene-drug class) pair. Allele frequency and functionality tables are provided for each corresponding gene. CPIC working groups for each gene follow a standard protocol to ensure consistency across the multiple guidelines. Tables are regularly updated, particularly when updated versions of existing guidelines are released. The guidelines are open to the public and were accessed under the "Guidelines" tab on the CPIC website: https://cpicpgx.org/guidelines/.

Of note, PGx variants are typically distinguished by haplotypes using a "star allele" nomenclature such as "*CYP2C19*2*." These alleles may be defined by multiple single nucleotide variants (SNVs) inherited as a block, or by structural variants such as duplications or deletions, or even by an individual SNV.[13] For our purposes, all of these types of variants, whether defined as star allele haplotypes or SNVs, are referred to as "alleles." Also of note, simple compilation of SNVs for pharmacogenes that are defined by star allele haplotypes, such as *CYP2C19* and *CYP2D6*, can thus be quite misleading.

We extracted data on population sizes, allele frequency, and allele functionality from 23 CPIC guidelines that encompassed 19 genes, with some published guidelines having multiple genes combined.[14–28] Genes in CPIC guidelines that did not have allele frequency tables or a corresponding allele functionality table, or those without defined risk alleles as described below, were excluded. Allele functionality classifications are based on the CPIC term-standardization project[29] as ***normal*** (fully functional or encode little activity), ***increased*** (greater than normal function), ***decreased*** (less than normal function), ***no*** (nonfunctional), or ***unknown*** (no literature describing function or allele is novel) function. These terms as used by CPIC and here refer strictly to their pharmacogenetic function specific to a given drug and not to other biochemical or pleotropic functions. We combined the CPIC classification ***uncertain*** (conflicting or weak literature supporting function) with ***unknown*** for our analyses. Alleles annotated as having increased, decreased, or no function were considered jointly as an **altered function** group. Alleles without functionality classifications but annotated as associated with increased risk for adverse events were included in the altered function group as described below.

CPIC biogeographic groups are defined as[12]:

- "***European*** (primarily European descent, including European Americans)"
- "***Latino*** (Mestizo descent, from Latin America, and self-identified Latino persons in the US),"
- "***American*** (from the Americas with ancestors predating European colonization) ancestries,"
- ***"African American/Afro-Caribbean*** (admixture between African, European, and Indigenous ancestries)"
- ***"Central/South Asian*** (Populations from Pakistan, Sri Lanka, Bangladesh, India, and ranges from Afghanistan to western border of China)"
- ***"East Asian*** (Populations from Japan, Korea, and China Stretching from the mainland through islands of Southeast Asia. Includes portions of central Asia, and Russia east of Ural Mountains)"
- ***"Near Eastern*** (Northern Africa, the Middle East, and the Caucasus, including Turkey and African nations north of Saharan Desert)"
- ***"Oceanian*** (Precolonial populations of the Pacific Islands, including Hawaii, Australia, New Zealand, and Papua New Guinea)"
- ***"Sub-Saharan African*** (all regions in Sub-Saharan Africa, including Madagascar)"

Of note, population definitions extracted by CPIC are as defined by publication or database authors, typically by geographic origin or self-identified race-ethnicity rather than by genetic analyses. CPIC groups these populations as defined above and uses them consistently throughout their guidelines. Throughout this paper, we apply these descriptors consistently, as recommended by the recent National Academies Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors.[30]

Allele frequencies for 11 genes with data in one or more of the nine defined groups for alleles classified as having increased, decreased, or no function; normal function; and unknown function were extracted from the "allele frequency" tab of the gene's downloadable CPIC "frequency table" linked to PharmGKB "Gene-specific Information Tables" within each guideline. These were: *ABCG*, *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *CYP3A5*, *DPYD*, *NUDT15*, *SLCO1B1*, *TPMT*, and *UGT1A1*.

CPIC did not assign functionality for two other genes, *CYP4F2* and *VKORC1*, but instead described a reference allele and one or more variants.[17] The variants in these

genes were associated with both altered warfarin sensitivity and warfarin dose and are considered here as altered function alleles, with the reference allele as normal function. CPIC-assigned alleles in *CACNA1S* and *RYR1* both an optional biochemical functional status and a required clinical functional status[25]; alleles with the clinical functional status of "malignant hyperthermia associated" were considered here as "altered function" while those assigned normal and uncertain function are considered here as "normal" and "unknown" function.

Allele frequencies for four additional genes with differing functional classifications but having alleles known to be associated with high risk for ADRs (*HLA-A*, *HLA-B*, *G6PD*, and *MT-RNR1*) were extracted in a similar way (Table S1). Functionality classifications for *MT-RNR1* were described only in terms of risk of aminoglycoside-induced hearing loss[23]; normal risk was included with the normal function group, increased risk with the altered function group, and uncertain risk with the unknown function group. CPIC described only the *HLA-A*31:01* and *HLA-B*15:02* alleles as associated with risk for Stevens–Johnson syndrome and toxic epidermal necrolysis (SJS/TEN)[24]; both are considered here as altered function. Similarly, *HLA-B*5701* and *HLA-B*5801* were described as associated with the risk of ADRs from allopurinol[27] and abacavir,[28] respectively, and are considered here as altered function. *G6PD* alleles were defined functionally as classes I-III/Deficient, class IV/Normal, or uncertain function,[26] these were considered here as altered function, normal function, or unknown function. The genes *COMT*, *HTR2A*, *OPRM1*, and *SLC6A4* have no CPIC recommendations and therefore have no allele frequency tables and were excluded from our analyses. The genes *CFTR* and *IFNL3* were also excluded from our analyses as their population definitions differed from the defined biogeographic groups described above.

For each gene, we estimated total population sizes from the "References" tab of the CPIC "frequency table" for each gene. The "References" tab of each of these spreadsheets lists all the sources used including author, PMID, and "population group" classified by PharmGKB curators based primarily on geographic region of origin as one of nine groups, as described above. This tab also lists "*N* subjects genotyped" for each study or allele set (in the case of gnomAD or its predecessor, the Exome Aggregation Consortium, ExAC). To determine the total number of samples genotyped across all alleles in a gene, we summed the *N* subjects genotyped for each group. However, not all subjects were genotyped for all alleles. When CPIC derived allele frequencies from databases such as gnomAD or ExAC, multiple entries were made in the spreadsheet when the number of subjects reported to be genotyped in these databases differed by allele (e.g., *DPYD*). To account for the different, yet likely overlapping, sample size by allele, we took the largest sample size for the database. While this approach may not be exact, it was the best approximation possible with the available information.

Frequencies of alleles categorized as having "increased, decreased, and no function" were summed and considered as an "altered function alleles" group for comparison to the sum of allele frequencies categorized as "normal function." Alleles with "unknown function" (including "uncertain function," as defined above) were uncommon, typically with frequencies well under 1%. These are often recently discovered alleles or alleles occurring predominantly in non-European groups that may not yet be assayed on commercial PGx tests nor reported in all published studies. These were also summed and reported separately as "unknown function." Of note, the *UGT1A1*80* allele is listed as unknown function and is frequent (>30%) in African American/Afro-Caribbean, European, and Latino groups. It is however in very high linkage disequilibrium (>99%) with known decreased function alleles **28* and **37*,[20] from which it is tabulated separately, so for this analysis **80* was presumed to have decreased function.

For the purposes of this analysis, we considered alleles that were assayed and reported as "absent" together with those that were "not reported," presumably because they were not assayed in a particular report or had not yet been discovered. Although these categories have different implications, since "not reported" alleles may actually be present in a given group, this is currently the information available to clinicians working to apply CPIC guidelines across groups and is the best approximation possible with the information available in the curated literature.

Using the largest-sized population as the reference for each gene, significance of differences in allele frequencies was estimated using standard chi-square tests with one degree of freedom for each gene in each comparison group and plotted in a "heat map" format by gene and group. The heat map shows the allele frequency values for each gene within each group and its significance level; increasing color intensity denotes an increasing significance of the *p*-value either above or below the reference group. Assigning the largest group as the reference ensured the most robust and stable estimates of allele frequency for comparisons while avoiding arbitrarily choosing one group as the standard against which all others were compared.

RESULTS

Group sizes

The European group had the largest population/sample size for 16 of the 19 genes and so served as the reference

population for those genes, while the East Asian group was the largest for *CYP2C9* and *CYP3A5* and the Central/South Asian group was the largest for *CYP4F2* (Figure 1, Figure S1, Table S2). European samples comprised the majority of the total sample for 12 of the 19 genes—for three of those (*DPYD*, *MT-RNR1*, and *RYR1*), European groups constituted more than 70% of the total sample. Allele frequency data were available for only nine genes each for the American and Oceanian groups and only 12 genes each for the African-American/Afro-Caribbean and Sub-Saharan groups. Only the Central/South Asian, East Asian, and European groups had data for all 19 genes, and only six genes (*ABCG2*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *HLA-A*, and *HLA-B*) had data in all nine groups. Sample sizes were 12,000 or less for Oceanian, Sub-Saharan African, American, Near/Middle Eastern, and African-American/Afro-Caribbean for all genes except *HLA-A* and *HLA-B*. The largest total sample size across all groups was 7.99 million, for *G6PD*.

Allele frequencies

A total of 191 alleles, regardless of function, were absent or not reported in the reference group but reported in at least one nonreference group in 14 of the genes (*CACNA1S*, *CYP2B6*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *CYP3A5*, *DPYD*, *G6PD*, *MT-RNR1*, *NUDT1*, *RYR1*, *SLCO1B1*, *TPMT*, and *UGT1A1*, Table S3). For example, the no function **18* allele of *CYP2B6*, discussed below, has a frequency of 0.0577 in the Sub-Saharan African group and 0.0330 in the African American/Afro-Caribbean group but is absent in the European reference group, while the no function **2* allele of *NUDT15*, also absent in the European reference group, has frequencies of 0.0350 and 0.0365 in the East Asian and Latino groups, respectively. *CYP2D6*, with the European group as the reference, had at least one allele reported in each nonreference group that was not reported within the reference; of the nonreference groups for *CYP2D6*, the African-American/Afro-Caribbean group had the most at 10. Alleles not reported in the reference group were reported frequently in the Latino and East Asian groups, with 12 and 9, respectively, of the 14 genes having an allele in these groups that was not reported in the reference group. The total numbers of altered function alleles reported in each group for these 14 genes also varied widely across groups and genes (Table S4).

Frequencies of altered function alleles (considered in aggregate by gene and group) differed substantially across groups with few discernible between-group patterns, though several between-gene differences were notable (Figure 2, heat map). Genes that had greater frequencies of altered function alleles in most or all groups compared to the reference group included *CYP2B6* (except for East Asian) and *CYP2C9* (except American and Oceanian). In contrast, the reference group had higher frequencies of altered function alleles than nearly every other group for *CYP4F2*, *DPYD* (except African-American/Afro-Caribbean and Central/South Asian), *SLCO1B1* (except for American), and *UGT1A1* (except for African-American/Afro-Caribbean). When considered by groups, there were no consistent trends toward one group being consistently higher or lower than the reference group in frequencies of altered function alleles across the 19 genes. Due to the large differences in sample sizes across groups, in which substantial differences in allele frequencies among some underrepresented groups might not be significant due to very small sample sizes, these calculations were repeated assuming all groups having the same size of 500 each, but results were essentially the same (data not shown).

As previously noted, frequencies of unknown function alleles were typically <1%, though *CYP2D6* and *MT-RNR1* exceeded 1% in multiple groups, reaching as high as 28% for *MT-RNR1* in Near Eastern groups (Table S5). *MT-RNR1* had many variants proposed as risk alleles for aminoglycoside-induced hearing loss that were judged by CPIC to have insufficient evidence as they were only described in single studies,[23] contributing to the high unknown function allele frequency for this gene. Unknown function alleles ranged from 1.7% to 10.4% frequency in two groups each for *CYP2B6*, *NUDT15*, and *TPMT*.

Sample sizes and allele frequency differences for each of the 19 genes are described in narrative form in Data S1.

DISCUSSION

Our review of the CPIC guidelines revealed that the European group had the highest representation in the curated literature for 16 of the 19 genes, while representation of the American and Oceanian groups was the most limited (only nine genes having data for these groups). For nearly all reported genes, altered function alleles were identified in one or more nonreference groups that were not reported in the reference group. This was especially seen within the Latino and East Asian groups. When looking specifically at altered function alleles, we found no consistent cross-group trends, with frequencies being both greater than the reference groups in some genes and lower in others. The genes *CYP2B6* and *CYP2C9* were *more* likely to have higher frequencies of altered function alleles in nonreference groups compared to the reference group, while the genes *CYP4F2*, *DPYD*, *SLCO1B1*, and *UGT1A1* were *less* likely to have higher frequencies of altered function alleles in nonreference groups.

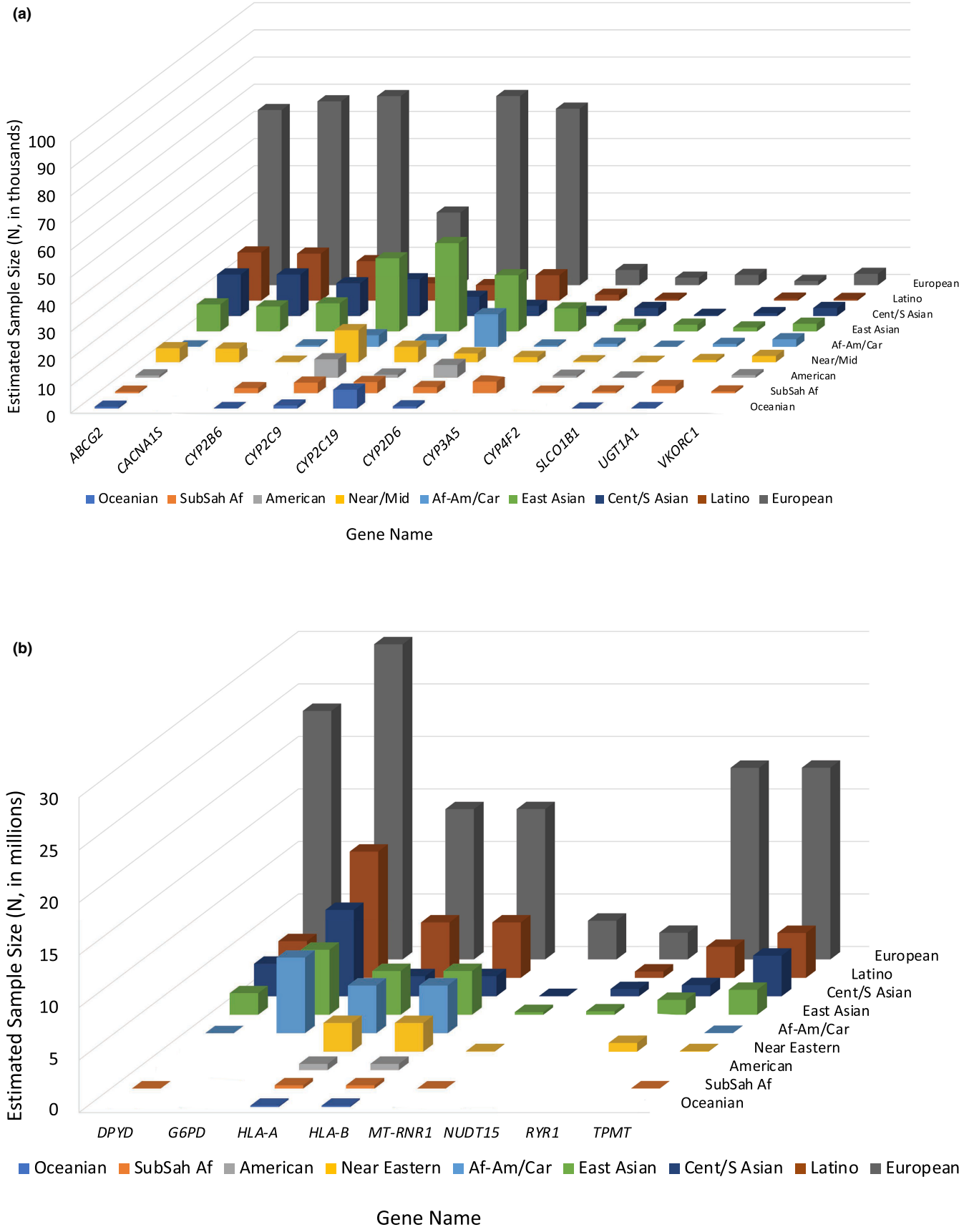

FIGURE 1 Sample sizes by gene and group. (a) Genes with sample sizes <70,000 in all groups. (b) Genes with sample sizes >250,000 in at least one group. Cent/S Asian, Central/South Asian; Af-Am/Car, African American/Afro-Caribbean; SubSah Afr, Sub-Saharan African.

Gene \ Group	African American/ Afro-Caribbean	American	Central/South Asian	East Asian	European	Latino	Near Eastern	Oceanian	Sub-Saharan African
ABCG2	0.035 6.5 x 10^{-5}	0.20 5.2 x 10^{-38}	0.093 6.2 x 10^{-8}	0.31 p = 0	0.10 ref	0.22 p = 0	0.065 1.2 x 10^{-35}	0.20 2.0 x 10^{-33}	0.0058 9.8 x 10^{-25}
CACNA1S	N/A	N/A	3.3 x 10^{-5} 0.51	1.1 x 10^{-4} 0.02	1.5 x 10^{-5} ref	8.7 x 10^{-5} 0.03	0 0.7	N/A	N/A
CYP2B6	0.42 1.5 x 10^{-9}	N/A	0.34 1.7 x 10^{-4}	0.29 6.3 x 10^{-26}	0.33 ref	0.41 1.0 x 10^{-160}	0.46 3.9 x 10^{-8}	0.64 8.1 x 10^{-34}	0.52 6.6 x 10^{-191}
CYP2C9	0.13 1.2 x 10^{-44}	0.088 0.0354	0.23 p = 0	0.082 ref	0.21 p = 0	0.14 1.7 x 10^{-81}	0.22 p = 0	0.045 1.3 x 10^{-10}	0.15 1.3 x 10^{-71}
CYP2C19	0.43 7.2 x 10^{-16}	0.21 2.4 x 10^{-57}	0.46 1.4 x 10^{-89}	0.38 6.9 x 10^{-7}	0.37 ref	0.28 1.0 x 10^{-91}	0.33 6.6 x 10^{-20}	0.81 p = 0	0.39 1.2 x 10^{-4}
CYP2D6	0.52 6.7 x 10^{-181}	0.23 1.1x 10^{-264}	0.38 3.0 x 10^{-7}	0.58 p = 0	0.42 ref	0.34 9.5 x 10^{-86}	0.48 1.7 x 10^{-25}	0.22 8.5 x 10^{-61}	0.64 1.2 x 10^{-213}
CYP3A5	0.55 2.9 x 10^{-42}	N/A	0.67 3.8 x 10^{-17}	0.75 ref	0.93 p = 0	0.83 1.2 x 10^{-28}	0.88 7.3 x 10^{-70}	N/A	0.52 3.2 x 10^{-286}
CYP4F2	0.076 p = 0	0.41 6.0 x 10^{-28}	0.57 ref	0.32 1.1 x 10^{-154}	0.43 2.9 x 10^{-49}	0.43 4.0 x 10^{-22}	0.56 0.58	N/A	0.33 1.7 x 10^{-45}
DPYD	0.025 0.0207	N/A	0.033 0.988	0.0023 p = 0	0.036 ref	0.010 p = 0	N/A	N/A	0.028 0.0002
G6PD	0.12 p = 0	N/A	0.035 p = 0	0.026 p = 0	4.7 x 10^{-3} ref	0.0072 7.2 x 10^{-3}	N/A	N/A	N/A

Color Key	Reference Group	Below Reference Group, p <				Above Reference Group, p <				No Data or No Diff
		5 x 10^{-4}	5 x 10^{-8}	5 x 10^{-16}	5 x 10^{-32}	5 x 10^{-4}	5 x 10^{-8}	5 x 10^{-16}	5 x 10^{-32}	

Gene \ Group	African American/ Afro-Caribbean	American	Central/South Asian	East Asian	European	Latino	Near Eastern	Oceanian	Sub-Saharan African
HLA-A	0.0099 0.93	0.062 p = 0	0.033 2.9 x 10^{-122}	0.035 p = 0	0.026 ref	0.045 p = 0	0.011 p = 0	0.011 1.3 x 10^{-51}	0.0080 1.2 x 10^{-179}
HLA-B	0.045 2.0 x 10^{-9}	0.039 2.3 x 10^{-18}	0.14 p = 0	0.12 p = 0	0.044 ref	0.028 p = 0	0.042 8.4 x 10^{-7}	0.038 6.4 x 10^{-6}	0.050 2.6 x 10^{-12}
MT-RNR1	N/A	N/A	0.0024 2.2 x 10^{-4}	0.020 2.9 x 10^{-43}	0.0021 ref	N/A	0.0014 5.7 x 10^{-4}	N/A	0.003 6.8 x 10^{-2}
NUDT15	N/A	N/A	0.068 p = 0	0.096 p = 0	0.0038 ref	0.044 p = 0	N/A	N/A	N/A
RYR1	N/A	N/A	2.9 x 10^{-4} 1.0 x 10^{-4}	2.5 x 10^{-4} 7.1 x 10^{-8}	4.8 x 10^{-4} ref	1.2 x 10^{-4} 2.6 x 10^{-36}	2.0 x 10^{-4} 1.9 x 10^{-7}	N/A	N/A
SLCO1B1	0.01 9.8 x 10^{-10}	0.24 0.014	0.07 1.5 x 10^{-58}	0.13 3.7 x 10^{-162}	0.34 ref	N/A	0.20 8.2 x 10^{-8}	0 6.4 x 10^{-8}	0.028 1.3 x 10^{-105}
TPMT	0.037 0.07	N/A	0.017 0	0.017 0	0.044 ref	0.054 2.4 x 10^{-306}	0.034 3.2 x 10^{-6}	N/A	0.055 9 x 10^{-4}
UGT1A1	0.52 5.5 x 10^{-168}	N/A	0.46 2.0 x 10^{-4}	0.29 2.1 x 10^{-77}	0.33 0.01	0.41 2.2 x 10^{-18}	0.32 8.8 x 10^{-42}	0.043 4.0 x 10^{-59}	0.51 ref
VKORC1	0.10 p = 0	0.62 5.9 x 10^{-57}	0.16 1.2 x 10^{-235}	0.87 p = 0	0.41 ref	0.39 0.19	0.45 8.3 x 10^{-6}	N/A	0.11 9.0 x 10^{-96}

Color Key	Reference Group	Below Reference Group, p <				Above Reference Group, p <				No Data or No Diff
		5 x 10^{-4}	5 x 10^{-8}	5 x 10^{-16}	5 x 10^{-32}	5 x 10^{-4}	5 x 10^{-8}	5 x 10^{-16}	5 x 10^{-32}	

FIGURE 2 Frequency and significance of altered function alleles in nonreference compared to reference group. Significance levels below the analytical level of estimation shown as "$p = 0$." Shades of red indicate genes and groups with frequencies above the reference group, while shades of blue indicate frequencies below the reference group. N/A, not available; ref, reference.

Group sizes

Available data on PGx variants in the CPIC guidelines, like other genomic data, are heavily over-represented by European and Asian populations. As of 2018, 78% of individuals were of European ancestry and 10% were of Asian ancestry across genome-wide association studies (GWAS).[31] Groups such as Native American and

Oceanian made up less than 1% of individuals included. This massive under-representation of five-sixths of the world's population[32] in curated PGx data impedes both critical scientific progress, by limiting the proportion of global human PGx variation that is recognized and investigated; and vital clinical applications, by failing to capture and incorporate clinically actionable variation into PGx guidelines.[32] Efforts to improve this situation are underway[33–35] as are efforts to standardize terminology describing population groups,[12,31,36] but bringing population representation in PGx data anywhere near the distribution of global populations will require monumental effort.

The importance of exploring alleles that may be absent in typical reference groups but present in other groups is evident, for example, from the detection of 27 potentially novel *CYP2D6* alleles in 961 high-depth whole-genome sequences across multiple Sub-Saharan African populations, including two novel alleles likely to render the genes nonfunctional.[37] Further study of the function of these variants is needed to characterize them definitively and incorporate them into guidelines; investigators should be encouraged to pursue functional investigations beyond allelic discovery. *CYP2D6* is by far the most polymorphic of the *CYP* genes and is involved in metabolism of over 20% of drugs approved by the US Food and Drug Administration.[38] Prevalence of poor and ultrarapid metabolizers based on *CYP2D6* diplotype-to-phenotype conversion standards has major implications for prescribing and varies markedly across global populations—from 0.4% to 10% for poor metabolizers and 1% to 21% for ultrarapid metabolizers—across population groups.[37] Current clinical panels are not standardized and only genotype a small number of *CYP2D6* alleles, tending to focus on those most commonly found in reference populations which are strongly biased toward European ancestry groups. Failure to detect, characterize, and report these altered function alleles in clinical testing of other groups could have serious implications for drug efficacy and safety.

Frequencies of altered function alleles

The high frequency of altered function alleles in *CYP2B6*, *CYP2C9*, and *NUDT15* in nonreference groups could have important therapeutic implications if PGx testing is limited or if available testing does not assay the variants more likely to be carried by persons in these groups. For example, the frequency of *CYP2B6* altered function alleles in the Oceanian group was nearly twice that in the European reference group (0.64 vs. 0.33, $p < 8 \times 10^{-34}$) and more than 50% greater in the Sub-Saharan African group (0.52, $p < 6.6 \times 10^{-191}$), with the decreased function *6 allele being particularly common in these two groups at 0.62 and 0.37, respectively, compared to the European group at 0.23. The no function *18 allele also has a 0.06 frequency in the Sub-Saharan African group and is undetected in the reference group. The antiretroviral drug efavirenz is metabolized predominantly by *CYP2B6* and is a cornerstone of treatment for HIV-1 infection, particularly in resource-limited settings.[15] Reduced doses are recommended for intermediate metabolizers (carriers of one normal function and one reduced or one no function alleles) and poor metabolizers (carriers of two reduced or no function alleles), and patients with these phenotypes are at increased risk for central nervous system toxicity and discontinuation of treatment,[15] making PGx testing a particular priority for these groups.

In a similar way, the high frequency of *CYP2C9* altered function alleles in nearly all groups except the Oceanian compared to the East Asian reference group (0.14–0.23 vs. 0.08, p for all groups $<10 \times 10^{-43}$), places them at higher risk of adverse effects or lack of efficacy from a variety of pharmacological agents, including warfarin,[17] statins,[18] phenytoin,[39] and nonsteroidal anti-inflammatory agents.[40] *CYP2C9* also illustrates the important variability in frequency of altered alleles that can be seen within broadly defined groups. In a study looking at the frequency of the decreased or no function *CYP2C9* alleles *2, *3, *4, *5, and *6 among two Mexican and one Mestizo-Mexican-American population, a much lower frequency of *2 alleles was found in the Mexican-Tepehuano (0.01) compared to Mexican-Mestizo (0.07) or Mexican-American samples (0.08).[41] These seven- to eight-fold differences in frequencies demonstrate the inadequacy of broadly defined biogeographic groups and the importance of fully characterizing PGx variation in a wide range of population subgroups.

The high numbers of alleles not present in reference groups reported in East Asian ($n = 47$) and Latino ($n = 41$) groups is somewhat surprising, particularly in comparison to the African-American/Afro-Caribbean ($n = 27$) and Sub-Saharan African ($n = 26$) groups. This was particularly true for *CYP2C19* in the East Asian group, with a **18* allele present at nearly 2% frequency, and a variety of alleles in both the East Asian and Latino groups in *G6PD*. African ancestry populations have been shown to have higher numbers of variants per individual genome than Asian or Latino populations in the 1000 Genome Project[42]; the lower numbers reported from CPIC curation of the PGx literature may be due to smaller sample sizes (particularly for the Sub-Saharan African group) and less dense sequencing or genotyping methods in these populations.

Limitations

We acknowledge that races/ethnicities identified by investigators or by participants themselves are an imperfect proxy for genetic diversity or genetic similarity, but this is the information currently available to curators developing the CPIC guidelines. As recommended by the NASEM report,[30] we have defined and used these terms consistently throughout. We recognize that a significant limitation of this analysis is the self-identification of race/ethnicity and subsequent clustering into CPIC-defined biogeographic groups. A more accurate representation would be using ancestral genetic markers to accurately differentiate between ancestries, but such information is not available to CPIC curators. It is often not available to clinicians or patients either, nor is it clear how they would use it. A balancing advantage is that the representations are applied and reported in a consistent manner across all the genes and drugs examined here. These findings serve as a further indication that future work using ancestral genetic markers within populations is needed to determine to what extent these allele frequencies differ.

Calculated allele frequencies provided by CPIC are similarly limited by the dependence on published literature, including that most studies test for only a subset of alleles (with some alleles rarely or never assayed or reported), some studies do not test for structural variants, and some studies default to a specific allele assignment (or even the reference allele), if the subset of alleles tested does not include all the haplotype-defining variants.[43] These weaknesses in the curated literature can lead to incorrect allele assignment, nondetection, or mis-estimation of allele frequencies. Thus, while allele frequency estimates based on CPIC are likely the best and most systemically curated estimates available, they are still only estimates. In addition, CPIC-defined clinical function based on literature reviews may not correspond to the biochemical function of the allele, though in some cases (e.g., *CACNA1S*, *RYR1*) both may be reported.

Another limitation is that we chose to consider all altered function variants as departing from the reference and potentially affecting the relevance of the CPIC guidelines in nonreference groups. This may not be the case, particularly for individuals with one increased function allele and one decreased function allele, but teasing out the potential functional impact of such heterozygotes for all gene-drug pairs was beyond the scope of our analysis. Other limitations to our review of CPIC guidelines were that all sample sizes from each study were combined to generate the total sample size, though not all subjects were tested for the same alleles in a particular gene, and sample sizes for databases such as ExAC, which vary by allele, were estimated using the largest sample size.

CONCLUSION

The CPIC guidelines have been developed to facilitate the use of gene/drug interactions in clinical care. This implementation of PGx-directed treatment, while it has the potential to help many individuals receive better treatment and care, also has the potential to widen existing health disparities among populations other than those of European ancestry if they do not undergo genotyping or are not included in PGx studies in adequate numbers. As our findings demonstrate there is great variability within allele frequencies that can influence drug dosage calculations, adverse effect likelihood, and medicinal efficacy.

The goal here is to not further separate individuals by socially constructed divisions or extend the misconception that health is ancestrally or racially determined or dependent. The goal is to acknowledge and find the small differences within our genetic make-up that can have big clinical implications. More research that encompasses the true genetic diversity of our world is needed to find these important altered allele frequencies. Precision medicine must not benefit only the few who are heavily represented in available data, but all. Understanding and implementing these important distinctions can be expected to improve treatment, lower adverse event frequency, and improve healthcare.

AUTHOR CONTRIBUTIONS

Sophia M. Hernandez, Lucia A. Hindorff, Joannella Morales, Erin M. Ramos, and Teri A. Manolio wrote the manuscript. Teri A. Manolio and Sophia M. Hernandez designed the research. Teri A. Manolio and Sophia M. Hernandez performed the research. Sophia M. Hernandez analyzed the data.

ACKNOWLEDGMENTS

We thank Teri E. Klein, Ph.D., Stanford University, for helpful review and comments on earlier drafts of this manuscript.

FUNDING INFORMATION

No funding was received for this work.

CONFLICT OF INTEREST STATEMENT

The authors declared no competing interests in this work.

ORCID

Teri A. Manolio https://orcid.org/0000-0001-5844-4382

REFERENCES

1. Lazarou J, Pomeranz B, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*. 1998;279:1200-1205.

2. Zhou Y, Lauschke VM. Population pharmacogenomics: an update on ethnogeographic differences and opportunities for precision public health. *Hum Genet*. 2022;141:1113-1136. doi:10.1007/s00439-021-02385-x
3. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. *N Engl J Med*. 2011;364(12):1144-1153.
4. Clinical Pharmacogenomics Implementation Consortium (CPIC). Guidelines. Accessed 14 February 2024. https://cpicpgx.org/guidelines/
5. Ramos E, Doumatey A, Elkahloun AG, et al. Pharmacogenomics, ancestry and clinical decision making for global populations. *Pharmacogenomics J*. 2014;14(3):217-222.
6. Zhou Y, Krebs K, Milani L, Lauschke VM. Global frequencies of clinically important HLA alleles and their implications for the cost-effectiveness of preemptive pharmacogenetic testing. *Clin Pharmacol Ther*. 2021;109(1):160-174.
7. Li B, Sangkuhl K, Whaley R, et al. Frequencies of pharmacogenomic alleles across biogeographic groups in a large-scale biobank. *Am J Hum Genet*. 2023;110(10):1628-1647.
8. Verma SS, Keat K, Li B, et al. Evaluating the frequency and the impact of pharmacogenetic alleles in an ancestrally diverse biobank population. *J Transl Med*. 2022;20(1):550.
9. McInnes G, Lavertu A, Sangkuhl K, Klein TE, Whirl-Carrillo M, Altman RB. Pharmacogenetics at scale: an analysis of the UK biobank. *Clin Pharmacol Ther*. 2021;109(6):1528-1537.
10. Relling MV, Klein TE. CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research. *Network Clin Pharmacol Ther*. 2011;89(3):464-467.
11. Relling MV, Klein TE, Gammal RS, Whirl-Carrillo M, Hoffman JM, Caudle KE. The clinical Pharmacogenetics implementation consortium: 10 years later. *Clin Pharmacol Ther*. 2020;107(1):171-175.
12. Huddart R, Fohner AE, Whirl-Carrillo M, et al. Standardized biogeographic grouping system for annotating populations in pharmacogenetic research. *Clin Pharmacol Ther*. 2019;105(5):1256-1262.
13. Kalman LV, Agúndez J, Appell ML, et al. Pharmacogenetic allele nomenclature: international workgroup recommendations for test result reporting. *Clin Pharmacol Ther*. 2016;99:172-185.
14. Lima JJ, Thomas CD, Barbarino J, et al. Clinical Ph4armacogenetics implementation consortium (CPIC) guideline for CYP2C19 and proton pump inhibitor dosing. *Clin Pharmacol Ther*. 2021;109(6):1417-1423.
15. Desta Z, Gammal RS, Gong L, et al. Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2B6 and Efavirenz-containing antiretroviral therapy. *Clin Pharmacol Ther*. 2019;106(4):726-733.
16. Bousman CA, Stevenson JM, Ramsey LB, et al. Clinical pharmacogenetics implementation consortium (CPIC) guideline for *CYP2D6*, *CYP2C19*, *CYP2B6*, *SLC6A4*, and *HTR2A* genotypes and serotonin reuptake inhibitor antidepressants. *Clin Pharmacol Ther*. 2023;114(1):51-68.
17. Johnson JA, Caudle KE, Gong L, et al. Clinical Pharmacogenetics implementation consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. *Clin Pharmacol Ther*. 2017;102(3):397-404.
18. Cooper-DeHoff RM, Niemi M, Ramsey LB, et al. The clinical pharmacogenetics implementation consortium guideline for SLCO1B1, ABCG2, and CYP2C9 genotypes and statin-associated musculoskeletal symptoms. *Clin Pharmacol Ther*. 2022;111(5):1007-1021.
19. Relling MV, Schwab M, Whirl-Carrillo M, et al. Clinical pharmacogenetics implementation consortium guideline for Thiopurine dosing based on TPMT and NUDT15 genotypes: 2018 update. *Clin Pharmacol Ther*. 2019;105(5):1095-1105.
20. Gammal RS, Court MH, Haidar CE, et al. Clinical pharmacogenetics implementation consortium (CPIC) guideline for UGT1A1 and Atazanavir prescribing. *Clin Pharmacol Ther*. 2016;99(4):363-369.
21. Birdwell KA, Decker B, Barbarino JM, et al. Clinical pharmacogenetics implementation consortium (CPIC) guidelines for CYP3A5 genotype and tacrolimus dosing. *Clin Pharmacol Ther*. 2015;98(1):19-24.
22. Amstutz U, Henricks LM, Offer SM, et al. Clinical pharmacogenetics implementation consortium (CPIC) guideline for Dihydropyrimidine dehydrogenase genotype and Fluoropyrimidine dosing: 2017 update. *Clin Pharmacol Ther*. 2018;103(2):210-216.
23. McDermott JH, Wolf J, Hoshitsuki K, et al. Clinical Pharmacogenetics implementation consortium guideline for the use of aminoglycosides based on MT-RNR1 genotype. *Clin Pharmacol Ther*. 2022;111(2):366-372.
24. Phillips EJ, Sukasem C, Whirl-Carrillo M, et al. Clinical pharmacogenetics implementation consortium guideline for HLA genotype and use of carbamazepine and oxcarbazepine: 2017 update. *Clin Pharmacol Ther*. 2018;103(4):574-581.
25. Gonsalves SG, Dirksen RT, Sangkuhl K, et al. Clinical pharmacogenetics implementation consortium (CPIC) guideline for the use of potent volatile anesthetic agents and succinylcholine in the context of RYR1 or CACNA1S genotypes. *Clin Pharmacol Ther*. 2019;105(6):1338-1344.
26. Gammal RS, Pirmohamed M, Somogyi AA, et al. Expanded clinical pharmacogenetics implementation consortium guideline for medication use in the context of G6PD genotype. *Clin Pharmacol Ther*. 2023;113(5):973-985.
27. Saito Y, Stamp LK, Caudle KE, et al. Clinical pharmacogenetics implementation consortium. Clinical pharmacogenetics implementation consortium (CPIC) guidelines for human leukocyte antigen B (HLA-B) genotype and allopurinol dosing: 2015 update. *Clin Pharmacol Ther*. 2016;99(1):36-37.
28. Martin MA, Hoffman JM, Freimuth RR, et al. Clinical pharmacogenetics implementation consortium. Clinical pharmacogenetics implementation consortium guidelines for HLA-B genotype and Abacavir dosing: 2014 update. *Clin Pharmacol Ther*. 2014;95(5):499-500.
29. Caudle KE, Dunnenberger HM, Freimuth RR, et al. Standardizing terms for clinical pharmacogenetic test results: consensus terms from the clinical pharmacogenetics implementation consortium (CPIC). *Genet Med*. 2017;19(2):215-223.
30. National Academies of Sciences, Engineering, and Medicine. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. The National Academies Press; 2023. doi:10.17226/26902
31. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell*. 2019;177(1):26-31.
32. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584-591.

33. Popejoy AB. Diversity in precision medicine and pharmacogenetics: methodological and conceptual considerations for broadening participation. *Pharmgenomics Pers Med*. 2019;12:257-271.
34. Zhang H, De T, Zhong Y, Perera MA. The advantages and challenges of diversity in pharmacogenomics: can minority populations bring us closer to implementation? *Clin Pharmacol Ther*. 2019;106(2):338-349.
35. Madden EB, Hindorff LA, Bonham VL, et al. Advancing genomics to improve health equity. *Nat Genet*. 2024;56:752-757.
36. Popejoy AB, Crooks KR, Fullerton SM, et al. Clinical genome resource (ClinGen) ancestry and diversity working group. Clinical genetics lacks standard definitions and protocols for the collection and use of diversity measures. *Am J Hum Genet*. 2020;107(1):72-82.
37. Twesigomwe D, Drögemöller BI, Wright GEB, et al. Characterization of CYP2D6 pharmacogenetic variation in sub-Saharan African populations. *Clin Pharmacol Ther*. 2023;113(3):643-659.
38. Kehinde O, Ramsey LB, Gaedigk A, Oni-Orisan A. Advancing CYP2D6 pharmacogenetics through a Pharmacoequity lens. *Clin Pharmacol Ther*. 2023;114(1):69-76.
39. Karnes JH, Rettie AE, Somogyi AA, et al. Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2C9 and HLA-B genotypes and phenytoin dosing: 2020 update. *Clin Pharmacol Ther*. 2021;109(2):302-309.
40. Theken KN, Lee CR, Gong L, et al. Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2C9 and nonsteroidal anti-inflammatory drugs. *Clin Pharmacol Ther*. 2020;108(2):191-200.
41. Dorado P, Sosa-Macias M, Peñas-LLedó E, et al. CYP2C9 allele frequency differences between populations of Mexican-mestizo, Mexican-Tepehuano, and Spaniards. *Pharmacogenomics J*. 2011;11:108-112. doi:10.1038/tpj.2010.29
42. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.
43. Pratt VM, Cavallari LH, Del Tredici AL, et al. Recommendations for clinical CYP2D6 genotyping allele selection: a joint consensus recommendation of the Association for Molecular Pathology, College of American Pathologists, Dutch Pharmacogenetics working Group of the Royal Dutch Pharmacists Association, and the European Society for Pharmacogenomics and Personalized Therapy. *J Mol Diagn*. 2021;23(9):1047-1064.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hernandez S, Hindorff LA, Morales J, Ramos EM, Manolio TA. Patterns of pharmacogenetic variation in nine biogeographic groups. *Clin Transl Sci*. 2024;17:e70017. doi:10.1111/cts.70017