Comprehensive Allele Genotyping in Critical Pharmacogenes Reduces Residual Clinical Risk in Diverse Populations

Shishi Luo[1], Ruomu Jiang[1], Joseph J. Grzymski[2,3], William Lee[1], James T. Lu[1] and Nicole L. Washington[1,*]

Genomic-guided pharmaceutical prescribing is increasingly recognized as an important clinical application of genetics. Accurate genotyping of pharmacogenomic (PGx) genes can be difficult, owing to their complex genetic architecture involving combinations of single-nucleotide polymorphisms and structural variation. Here, we introduce the Helix PGx database, an open-source star allele, genotype, and resulting metabolic phenotype frequency database for *CYP2C9*, *CYP2C19*, *CYP2D6,* and *CYP4F2*, based on short-read sequencing of >86,000 unrelated individuals enrolled in the Helix DNA Discovery Project. The database is annotated using a pipeline that is clinically validated against a broad range of alleles and designed to call *CYP2D6* structural variants with high (98%) accuracy. We find that *CYP2D6* has greater allelic diversity than the other genes, manifest in both a long tail of low-frequency star alleles, as well as a disproportionate fraction (36%) of all novel predicted loss-of-function variants identified. Across genes, we observe that many rare alleles (<0.1% frequency) in the overall cohort have 10 times higher frequency in one or more subgroups with non-European genetic ancestry. Extending these PGx genotypes to predicted metabolic phenotypes, we demonstrate that >90% of the cohort harbors a high-risk variant in one of the four pharmacogenes. Based on the recorded prescriptions for >30,000 individuals in the Healthy Nevada Project, combined with predicted PGx metabolic phenotypes, we anticipate that standard-of-care screening of these 4 pharmacogenes could impact nearly half of the general population.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Pharmacogenetic variants and drug-genotype interaction reference datasets are available for research use. However, comprehensive population frequencies for known variants, including structural variants, and genotypes are not widely available, limiting their clinical utility.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ We aim to quantify the frequency of known alleles in key pharmacogenes using a clinically validated annotation pipeline and to assess the potential clinical impact that standard-of-care pharmacogenomic (PGx) testing would have on prescribing behavior in a health system over a 10-year period.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ We provide a comprehensive open-source database of allele, genotype, and metabolic phenotype frequencies, and find that nearly half the general population carries a high-risk variant for a drug they are prescribed.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Reporting allele frequencies can help prioritize future functional studies and guide drug efficacy study design, ensuring that subpopulations with a higher prevalence of high-risk metabolic phenotypes are represented. Furthermore, incorporating PGx testing into routine medical care could greatly affect prescribing guidance.

Population genomics studies have shown that nearly every individual has at least one variant affecting known pharmacogenes. Genomic-guided individualized drug therapies, thus, are ripe for standard-of-care clinical integration.[1,2] Among important pharmacogenes are the cytochrome P450 (CYP450) enzymes, which are essential for the production of cholesterol, steroids, and prostacyclins, and are necessary for the metabolism of most drugs. More than 30% of major adverse drug events are found to be related to underlying polymorphisms in *CYP2C9* (HGNC:2623), *CYP2C19* (HGNC:2621), and *CYP2D6* (HGNC:2625) alone. As a result, the negative consequences of ineffective and/or inadequate treatment are numerous for both individual patients and the healthcare system as a whole, which bears the resulting cost burden.[3–6] For pharmacogenomic (PGx) testing to be deployed

[1]Helix, San Mateo, California, USA; [2]Desert Research Institute, Reno, Nevada, USA; [3]Renown Institute of Health Innovation, Reno, Nevada, USA.
*Correspondence: Nicole L. Washington (nicole.washington@helix.com)

Received February 1, 2021; accepted April 19, 2021. doi:10.1002/cpt.2279

at scale in the clinic, accurate tests combined with comprehensive reference databases are required to hasten clinical interpretation.

Accurately genotyping PGx genes can be difficult, owing to their complex genetic architecture. *CYP2D6* is a prime example; it has many known variant haplotypes involving not just combinations of single-nucleotide polymorphisms, but also structural variation (SV), including copy number variation, complex rearrangements, and/or gene conversion events (**Table 1**). Clinical applications have typically used panels that report only known alleles with greatest population-level impact, because they can be quickly deployed. These applications, however, may lead to mis-genotyping and produce erroneous therapeutic recommendations.[7,8] Off-the-shelf genotype arrays, whole exome sequencing, and whole genome sequencing methods are unable to comprehensively call *CYP2D6* genotypes[1] without specialized chemistry and/or custom bioinformatic algorithms. Similar customizations are needed to identify whole gene duplications and deletions in *CYP2C9*, *CYP2C19*, and *CYP4F2* (HGNC:2645).[9] As with Mendelian phenotypes, incorporation of previously unidentified rare predicted loss-of-function (pLoF) variants that may impart high effect, are needed to improve phenotypic prediction accuracy.

Clinical interpretation of PGx test reports rely on drug-gene-variant interactions that are curated by several international organizations. These include: the Clinical Pharmacogenetics Implementation Consortium (CPIC), the Pharmacogenomics KnowledgeBase (PharmGKB), Pharmacogene Variation Consortium (PharmVar), and the US Food and Drug Administration (FDA). As of December 2020, there were 72 drug-gene combinations listed on the CPIC for *CYP2D6* alone, over half with sufficient evidence to require PGx testing, affect treatment action, or to inform a prescribing physician with possible recommendations based on the genotype found.[10] To aid panel design and interpretation, accurate reference databases of both allele frequency and resulting metabolic function are required. Furthermore, these resources need to be as comprehensive as possible, illuminating population-specific differences and/or deficiencies.

In this study, we introduce the Helix PGx database (https://github.com/myhelix/helix-pgxdb), which reports PGx star allele frequencies, genotype frequencies, and their metabolic phenotype for *CYP2C9*, *CYP2C19*, *CYP2D6*, and *CYP4F2* across 86,490 unrelated individuals enrolled in the Helix DNA Discovery Project (HDDP). By using a specialized whole exome sequencing assay and custom variant calling approach, the clinically validated Helix PGx pipeline achieves a high level of accuracy (**Tables S1, S2**), which provides comprehensive coverage for rare and less-assayed alleles. We also examine the population-level impact by connecting predicted PGx phenotypes to medications with strong evidence of interaction effects. Last, we analyze the real-world prescription history of individuals in the Healthy Nevada Project (HNP) in a retrospective analysis to demonstrate the clinical utility of routine PGx testing in the general population.

METHODS

Human subject research

All subjects provided informed consent to participate in either one or both of these institutional review board (IRB)-approved research studies: Helix DNA Discovery Project (HDDP; WIRB Protocol #20170748) or the HNP (University of Nevada Reno IRB protocol #7701703417). Consent was obtained either in-person or electronically, consistent with the IRB-approved protocols. All individuals resided in the United States at the time of providing their saliva sample for sequencing. Importantly, these individuals have not been sequenced based on the presence or absence of any medical phenotype (i.e., there are no inclusion or exclusion criteria in the registration process based on any medical phenotype). Most, but not all of the HNP participants were also part of the HDDP; ~ 26% of HNP participants did not consent for research under the HDDP protocol, which would have represented only ~ 8% of the combined (HNP only / (HNP only + HDDP)) cohort.

Annotation pipeline

Samples in the HDDP and HNP were processed on Helix's Exome+ platform and annotated with Helix's PGx pipeline. This pipeline builds upon existing next generation sequencing methods Aldy[11] and Stargazer.[12] Briefly, reads are mapped for each gene and ambiguous or mismapped reads are removed. Likelihoods for exon-level copy number and single nucleotide variations/indels are then generated from read counts and allele depths. The combination of star alleles that achieve the highest likelihood is reported, along with a phred-scaled quality score (see **Supplementary Methods** for details). Star allele combinations with a quality score of twenty or higher (i.e., they are a hundred times more likely than the next candidate allele combination) are considered passing calls.

In some cases, a sample has a high-quality star allele call but there are defining variants in the sample that are not supported by the star allele combination; these are considered incomplete matches. There are also novel variants, identified with snpEff[13] to annotate predicted loss-of-function variants (start lost, stop gained, frameshift, splice donor, and splice acceptor variants) that are not associated with any existing star allele. These incomplete matches and novel pLoF variants are not phased, so the variant cannot be assigned to a particular star allele and are therefore associated with an allele combination (genotype), rather than a star allele (haplotype).

Samples with insufficient read depth for any defining variant or low overall read count are flagged as low-quality and removed. To avoid inflating population frequencies for rare alleles, we removed samples that appeared to be first- or second-degree relatives. Kinship coefficients were calculated using KING[14] on 184,445 variants (as described[15]).

Genotype functional annotation

Star allele functional annotations were obtained from CPIC allele functionality tables for *CYP2D6*, *CYP2C19*, and *CYP2C9*.[16–18] Activity scores for *CYP2D6* and *CYP2C9* star alleles were used as provided. Activity scores were inferred for *CYP2C19* star alleles for computational convenience by transforming each "Allele Clinical Functional Status" from the guideline as follows: no function = 0; decreased function = 0.5; normal function = 1.0; and increased function = 1.5. The final genotype-to-metabolizer status mapping is identical with the standard practice for interpreting *CYP2C19* (see **Table S4**). Phenotype assignment for alleles of *CYP4F2* was interpreted from the Warfarin dosing guidelines,[19] which lists *CYP4F2*3* as a decreased function allele and indicates reporting carrier status only. Increased function variants, and their associated Electronic Health Record (EHR) Priority Result Notation (risk), have not been reported for *CYP2C9* or *CYP4F2*,[20] so alleles with full gene duplications were considered unknown. For all genes, novel loss of function and full gene deletion variants were assigned an activity score of zero.

Activity scores for *CYP2D6*, *CYP2C9*, and *CYP2C19* genotypes were prepared by summing activity scores for individual alleles. This method was preferred over using the CPIC diplotype-phenotype reference tables[21–23] directly as a lookup for an "off-the-shelf" implementation because some common genotypes (such as the *1/*10+*36

genotype found at >16% in our East Asia [EAS] population) were absent from the CPIC diplotype-phenotype tables, and it was difficult to match complex SVs to corresponding elements in the CPIC tables. For genotypes with full-gene duplications, all allele copies are summed. Metabolic function and associated risk profile for each genotype was assigned using the activity score thresholds set in each guideline together with the EHR Priority Result Notation, respectively, and summarized in **Table S4**. The following rules were also applied for risk determination: if there was an allele in the genotype with an unknown activity score/function, the full genotype was assigned an unknown risk; for genotypes with a novel pLoF allele, if the remaining alleles impart a likely intermediate metabolizer (LIM), intermediate metabolizer (IM), likely poor metabolizer (LPM), poor metabolizer (PM) phenotype, or is a *CYP4F2*3* carrier, then the metabolizer status is unlikely to improve with the inclusion of the pLoF allele, and is therefore left as the same phenotype and high risk, otherwise it is changed to unknown phenotype and risk. We treat *CYP2C19* LIM and LPM identically to IM or PM phenotypes, respectively, because the EHR Priority Result Notation is identical for each (abnormal/priority/high risk), and our analysis does not need to distinguish between LIM and IM (or LPM and PM) phenotypes.

Genetic ancestry assignment

Because participants' self-reported ancestry was unavailable, we assigned ancestry labels based on genetic data. For each cohort participant, we used exome-wide variant calls to calculate Admixture coefficients[12,24] for five reference populations – Africa (AFR), America (AMR), EAS, Europe (EUR), and South Asia (SAS) -- from phase III data of the 1000 Genomes project.[25] For simplicity, we used a single threshold to assign participants to a reference population, such that samples with a single admixture coefficient of greater than 0.8 were assigned that population's label (see **Figure S1**). Samples that did not exceed 0.8 in any population were labeled "other."

These labels are crude approximations for the geographically based semi-isolation that the human species has experienced for part of its existence in the regions of AFR, AMR, EUR, EAS, and SAS. Our goal for performing this ancestry assignment was to identify variants and star alleles, which, due to population bottlenecks and founder effects throughout human existence, now have heterogeneous distributions in different subpopulations. Given that the composition of the US population is much more diverse than the composition of our cohort (**Figure S1**), this ancestry assignment was important for uncovering variants of functional importance that are not represented among variants found in the EUR group.

Medication analysis

Prescribing records from the Renown Health System were provided for 31,165 participants in the HNP for years 2009–2019. Each medication was mapped to its target PGx gene from the evidence-based drug-gene list shown in **Table 2**. Prescribing rates were derived from the unique set of individuals who were prescribed at least one drug in the list associated with a given PGx gene divided by all participants who were prescribed at least one drug, either in a given year or averaged across the 2009–2019 time period as a proxy for "lifetime." A list of 59 drug-gene pairs derived from the CPIC,[10] PharmGKB,[26] and the FDA[27] curated lists was used to filter and analyze the prescribing history for HNP participants (see **Supplementary Methods** for details).

RESULTS

The Helix PGx database: A pharmacogenomics reference database

We used the Helix PGx pipeline to annotate 86,490 unrelated individuals in the HDDP and created a database of star allele and genotype frequencies for *CYP2C9*, *CYP2C19*, *CYP2D6*, and

Table 1 Summary of PGx star alleles and genotypes in HDDP

Gene	Star alleles (known[a])			Star alleles (known, in HDDP[b])			Novel pLoF (in HDDP[c])		Genotypes (in HDDP)	
	SNP	SV	w/o normal func	SNP	SV	w/o normal func	Total	$n > 1$	Unique*	w/unknown pheno/risk[d]
CYP2C19	34	1	28	23	4	22	21	7	157	71 (1.2%)
CYP2C9	70	0	68	39	4	41	28	4	159	66 (0.23%)
CYP2D6	128	10	127	67	45	102	49	16	1288	588 (3.4%)
CYP4F2	3	0	2	2	2	3	38	11	56	24 (0.07%)

HDDP, Helix DNA Discovery Project; PGx, pharmacogenomic; PharmGKB, Pharmacogenomics KnowledgeBase; pLoF, predicted loss-of-function; SNP, single-nucleotide polymorphism; SV, structural variation.
* Includes novel pLoF and incomplete matches.
[a]Curated by PharmVar as per API query. SNP or SV columns indicate number of uniquely defined star alleles for each gene. Note: SV here does not include whole-gene duplications. [b]Alleles found in the HDDP cohort. Note: SV here has a higher value than the previous SV column as it includes whole-gene duplications of known star alleles (e.g., CYP2D6*14 × 2 is not explicitly listed in PharmVar, but we consider this a known allele instead of a novel one). [c]Tally of unique novel putative LoF variants observed, overall or only in more than one unrelated individual ($n > 1$; nonsingleton). [d]Phenotype/risk based on CPIC and PharmGKB allele functionality tables. Percentage in parentheses is frequency in the HDDP cohort (out of 86,490 individuals).

CYP4F2 (see Methods). The pipeline has been validated using 355 samples carrying up to 875 known alleles per gene, achieving 100% accuracy for *CYP2C9*, *CYP2C19*, and *CYP4F2*. For *CYP2D6*, accuracy was 99.6% for non-SV alleles and 98.0% for SV alleles (**Tables S1, S2**). The pipeline also reports novel pLoF variants and incomplete matches (i.e., novel combinations of known mutations).

The database is hosted on github (https://github.com/myhelix/helix-pgxdb) and comprises two files: (i) star allele frequencies for each gene; and (ii) genotype (allele combinations)-phenotype frequencies. For each table, frequencies are reported for the overall cohort, as well as by inferred genetic ancestry (AFR, AMR, EAS, EUR, SAS, and other; see Methods for details on group assignment). For convenience, we also include PharmVar-assigned unique identifiers and functional annotations, where known.

Allele frequencies

Overall, we observed 186 unique star alleles across the four PGx genes in our HDDP cohort. Of these, 140 exactly match alleles documented in PharmVar, and 46 are duplications/deletions of these alleles. An additional 106 unique star alleles documented in PharmVar were not detected in our cohort, most likely because they are rare, or they were recently uploaded and thus not reported by our pipeline (see **Supplementary Methods**). In addition, we identified 136 new pLoF variants across all genes, 38 of which (28%) were present in more than one unrelated individual (**Table 1**).

CYP2D6 exhibits higher allelic diversity and greater prevalence of structural variants compared with the other genes tested here. The 10 most common star alleles for *CYP2D6* account for < 95% of the total allele calls. In contrast, for each of the non-*CYP2D6* genes, they account for > 99.5% of the total allele calls (**Figure 1a**). Furthermore, SVs account for > 11% of all *CYP2D6* alleles called, whereas for each of *CYP2C9*, *CYP2C19*, and *CYP4F2*, structural variants make up less than 0.05% of alleles called (**Figure 1b**). *CYP2D6* novel pLoF variants also account for a disproportionate fraction (36%) of all novel pLoF variants identified, suggesting that this locus continues to accumulate new mutations (**Table 1**). These characteristics demonstrate that a comprehensive assay is critical for the accurate annotation of *CYP2D6* star alleles in order to avoid mis-genotyping a nontrivial fraction of the population.

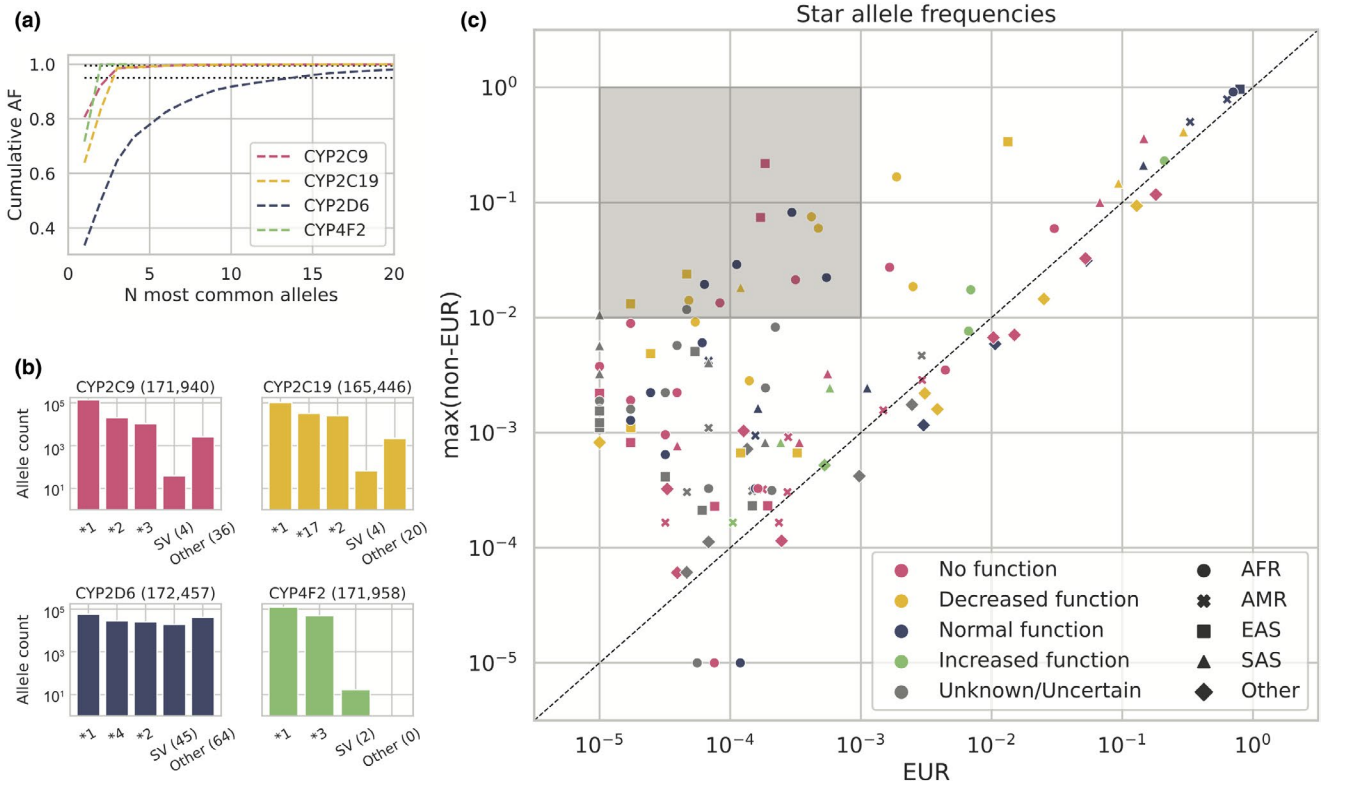

Figure 1 Helix PGx database star allele frequencies. (**a**) Cumulative star allele frequencies. Dotted black lines indicate 95% and 99.5%. (**b**) For each gene, allele counts are shown for the most common alleles, alleles with structural variants (SVs; includes deletions and duplications), and the everything else ("other"). Total allele counts are displayed in parentheses next to the gene name. For SV and other, the number of unique alleles in that category are displayed in parentheses. (**c**) Allele frequencies (excluding novel variants and novel combinations) plotted by their frequencies in different subpopulations. Each point is one of the 186 star alleles, across all 4 genes, from **b**. The X-coordinate is its allele frequency in the EUR subset of the cohort and the Y-coordinate is its maximum frequency across the non-EUR subsets of the cohort. Marker shape corresponds to the non-EUR ancestral group in which the maximum is achieved. Marker color corresponds to the allele function as annotated by CPIC. See Methods for details on assignment of genetic ancestry groups and **Table S3** for allele name, frequency, and functional annotation for each of the points in the shaded region. AF, allege frequency; AFR, African; AMR, American; CPIC, Clinical Pharmacogenetics Implementation Consortium; EAS, East Asian; EUR, European; SAS, Southeast Asian.

Previous reports on population frequencies of CYP genes have indicated ancestry-specific differences in allele frequencies.[2,28,29] With the benefit of a more comprehensive panel here, we can further quantify ancestry-specific differences for rare alleles. We compared allele frequencies calculated for the EUR subset of the cohort to the maximum of the allele frequencies calculated for each of the non-EUR subsets (**Figure 1c**). Although many alleles are present at similar orders of magnitude in all subpopulations, a substantial portion of alleles that are low-frequency in the EUR group are prevalent in a non-EUR subpopulation. Specifically, of the 152 alleles that occur with frequency < 0.1% in the EUR subset of the cohort, 16 alleles are at least 10 times more prevalent (i.e., > 1%) in one or more non-EUR groups (shaded box, **Figure 1c**). To name a few, *CYP2C19*3*, a nonfunctional allele, has a frequency of 7.4% among the EAS group, but occurs with frequency 0.02% in EUR. *CYP2C9*8* and *CYP2D6*29*, both decreased function alleles, occur with frequencies 6.0% and 7.5%, respectively, in the AFR group, but occur with frequencies < 0.05% in EUR (full details in **Table S3**). These striking differences between EUR and non-EUR allele frequencies reinforce the notion that panels which do not assay for an array of alleles from diverse populations are at risk of missing alleles that are of high prevalence and consequence in non-EUR populations.

Functional impact of allele combinations

In order to understand the potential clinical impact of the 1,660 unique allele combinations (genotypes) identified in our population for the 4 genes studied here, we predicted the metabolic function for each genotype using an activity score-based algorithm as described by the CPIC-DWG standardization working group guideline for *CYP2D6*[30] (see Methods). The genotype-based activity scores were then used to annotate each individual's overall gene-based metabolizer status (poor, intermediate, normal, ultrarapid, or unknown) and PGx risk profile (normal/routine/low risk, abnormal/priority/high risk, or unknown) from the appropriate guideline (see Methods, **Table S4**). The genotypes, their predicted metabolizer status, PGx risk, overall frequency, and ancestry-specific frequencies are included in the genotype-phenotype table of the Helix PGx database.

We were able to assign metabolic status and PGx risk to 55% (911/1660) of the observed genotypes, representing > 95% of the population. For those genotypes where we could assign function, we found that 92.4% of our cohort has an "abnormal/priority/high risk" genotype in at least one of the four pharmacogenes (**Figure 2a**). The majority of the genotypes for which we were unable to assign function (588/749 or 78%) were from the highly polymorphic *CYP2D6* locus. With its many incomplete matches and novel predicted LoF alleles, 3.4% of the population carry a *CYP2D6* genotype with an unknown phenotype and risk (**Table 1**, **Figure 2a**).

When more closely analyzing the metabolic differences that contribute to the abnormal/priority/high risk impact for each gene (**Figure 2a**), we find that poor metabolizers (PMs) comprise the smallest fraction of the population, whereas ultrarapid metabolizers (UMs) are the greatest contributors for *CYP2C19*, and intermediate metabolizers (IMs) have the greatest impact for *CYP2C9* and *CYP2D6*. Presently, there are no guidelines that assign a PM or UM status to any *CYP4F2* alleles, so all high-risk phenotypes here are assigned based on the carrier status of *3 allele. We anticipate future functional studies will assign these putative LoF alleles to be PMs with high risk, but here we assign an unknown function and risk.

We further divided the collection of high-risk phenotypes for each gene by dominant ancestry and compared each to understand if there were differences in phenotypic outcome by ancestry (**Figure 2b**). We observed that ancestry was a significant contributor to differences in high-risk phenotypes observed in *CYP2C19*, *CYP2C9*, and *CYP4F2* (χ^2, $P < 0.01$), but not for *CYP2D6*. There was a notable increase in the fraction of SAS with high-risk genotypes in *CYP4F2* ($P < 10^{-5}$), and a decrease of *CYP2C9* high-risk genotypes in EAS ($P < 10^{-4}$), *CYP2C19* in AMR ($P < 0.001$), and *CYP4F2* in AFR ($P < 10^{-5}$). Ancestry did not significantly contribute to differences observed for unknown phenotypes (χ^2, $P = 0.25$).

Predicted clinical impact of PGx testing on real-world prescription drug usage

Next, we wanted to understand the impact that the observed variation in these four pharmacogenes could have on the prescribing behavior of a real-world cohort. We Exome+ sequenced and called the PGx alleles for the pharmacogenes for 31,165 participants in the HNP[31] who had at least one prescription record (average medication prescription history was 7 years, range 1–10 years.) We analyzed the phenotypic risk for each participant against their medication history in the context of a list of 59 drug-gene pairs for which there was strong evidence for either prescribing changes or PGx-guided dosing information based on guidelines and/or curated information from the CPIC, PharmGKB, and/or the FDA (see **Table S5** for the full list of drug-gene pairs). **Table 2** summarizes the PGx genes reported in this study, their PGx-associated medication(s), and the percentage of individuals enrolled in HNP who were prescribed one or more of the indicated medications. We find that > 75% of the individuals have been prescribed one or more of these medications in their medical record lifetime, with an average of 46% prescribed one or more medications in any given year. In addition, > 40% of the population is at high risk for PGx effects for one or more of the medications they were prescribed in their EHR-lifetime. On a yearly basis, this translates to nearly a quarter of the population being prescribed a medication for which the individual is at high risk for a PGx effect. Although there are more medications indicated for *CYP2D6* than for the other genes studied, most of the PGx risk interactions in any given time period are due to just a few drugs (such as ondansetron, oxycodone, and codeine, see **Table S6**), meaning that the medication count does not contribute appreciably to the overall higher impact of *CYP2D6* risk.

DISCUSSION

Personalized genomic medicine will require comprehensive PGx tests and accurate PGx-guided prescriptions for all patients, not just those who carry high-impact or common variants. The Helix PGx Database provides a fine-scale accounting,

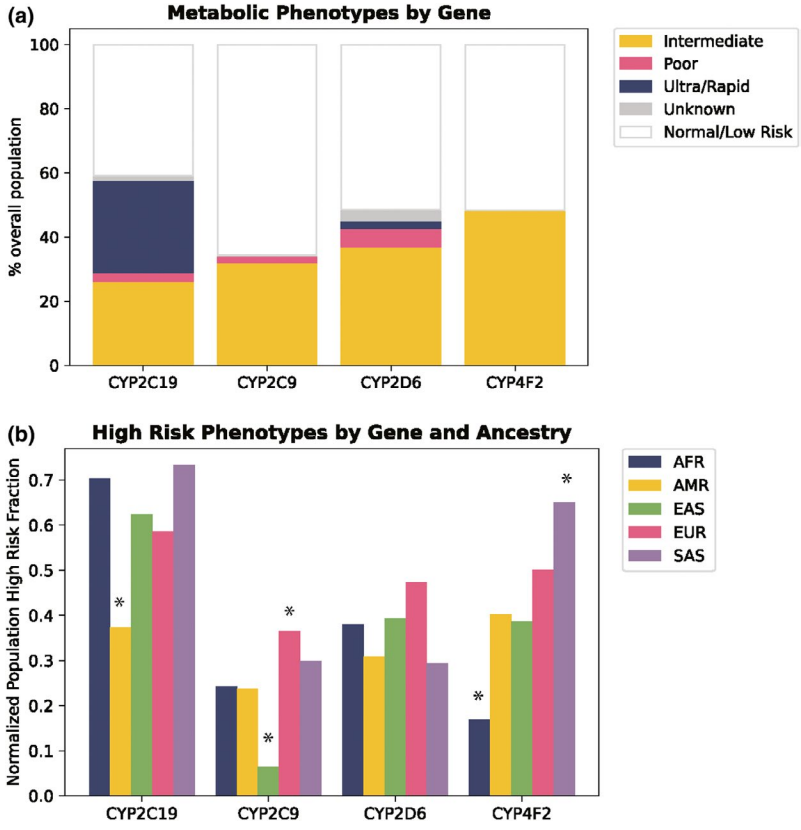

Figure 2 Metabolic risk phenotypes for *CYP2C19*, *CYP2C9*, *CYP2D6*, and *CYP4F2* applied from CPIC guidelines. (**a**) Overall risk profile for each gene. Normal/low risk is shown in white, Unknown risk in grey. Abnormal/high risk metabolic phenotypes indicated by color and combined for simplicity as follows: intermediate = IM + LIM; poor = PM + LPM; ultra/rapid = RM + UM. Approximately 50% of the population for each gene has an affected gene; overall 92.6% of the population has at least one high risk genotype. (**b**) Abnormal/High Risk metabolizer phenotypes segmented by dominant continental ancestry (Admix > 0.8) represent the fraction of individuals with any predicted high risk metabolic phenotype for the indicated gene and ancestry. The *P* values in B are calculated by Fisher-Exact test of frequency for a given population vs. combined mean of remaining populations for the specified gene, and represent $P < 1E\text{-}3$. Other (no dominant ancestry) removed from B for clarity (IM, intermediate metabolizer; LIM, likely intermediate metabolizer; LPM, likely poor metabolizer; PM, poor metabolizer; RM, rapid metabolizer; UM, ultrarapid metabolizer). AFR, African; AMR, American; CPIC, Clinical Pharmacogenetics Implementation Consortium; EAS, East Asian; EUR, European; SAS, Southeast Asian.

Table 2 Population-level prescriptions and risk by PGx gene target over EHR lifetime or yearly average

		Lifetime		Yearly (mean ± SD)	
Gene	Drugs[a]	% pres[b]	% high risk[c]	% pres	% high risk
CYP2C19	amitriptyline, brivaracetam, **citalopram**, clobazam, clopidogrel, escitalopram, flibanserin, pantoprazole, sertraline, voriconazole	20.3	10.7	10.2 ± 1.4	5.4 ± 0.7
CYP2C9	celecoxib, flurbiprofen, **ibuprofen**, meloxicam, phenytoin, piroxicam, warfarin	40.0	12.9	14.6 ± 1.2	4.8 ± 0.3
CYP2D6	amitriptyline, aripiprazole, atomoxetine, codeine, doxepin, fluvoxamine, imipramine, nortriptyline, **ondansetron**, oxycodone, paroxetine, propafenone, tamoxifen, thioridazine, tramadol, venlafaxine, vortioxetine	69.4	27.8	37.1 ± 1.8	14.8 ± 0.9
CYP4F2	warfarin	2.55	1.2	1.4 ± 0.2	0.6 ± 0.1
Any	**any**	**75.4**	**41.7**	**46.0 ± 0.8**	**22.5 ± 0.3**

EHR, electronic health record; PGx, pharmacogenomic.
[a]Bolded drugs contribute the most to the prescribed fraction for the given gene. Breakdown of individual drug frequency is available in **Supplementary Table 5**. [b]% pres is the fraction of the population who has been prescribed one or more of the drugs listed that interacts with the indicated PGx gene. [c]High risk indicates the overall percent of the population who are at high risk for PGx effects who have also taken one or more of the indicated medications.

in some cases for the first time, of the prevalence of rare and/or less-assayed alleles and genotypes. Although allele frequency data for PGx genes from large cohorts already exist,[2,29] the Helix database has the benefit of being generated from a clinically validated sequencing pipeline for a comprehensive panel of star alleles. Already, the star allele frequency database fills gaps in knowledge for common CYP2D6 structural variants: *CYP2D6*36*, which is reported in PharmGKB as having a prevalence of 1.2% in the PharmGKB-designated "East Asian" group, is found in the EAS group here at 22%. Similarly, PharmGKB does not report the frequency of *CYP2D6*68*, but we find it at a prevalence of 5.2% in the EUR group.

The frequencies reported here underline the unique challenges posed by *CYP2D6* genotyping. With its long tail of low frequency alleles and relatively high prevalence of structural variants and novel pLoF variants, comprehensive *CYP2D6* panels will be necessary but not sufficient for the clinical application of PGx testing. Compendiums of curated genotype-phenotype associations such as that at PharmVar are extremely useful to build personalized predictions of metabolic phenotype for known alleles. However, interpreting novel alleles will require a library of functional assay results for noninvasive prediction of metabolizer status. Rapid screening methods to increase the library of functional evidence for all possible alleles, similar to that as developed for *CYP2C9*,[32–34] CYP2C19,[34] and *NUDT15* (HGNC:23063),[35] will be necessary for on-demand comprehensive *CYP2D6* functional annotation in the clinical setting. Burgeoning machine learning approaches to predict function from *CYP2D6* haplotype[36] also offer a promising direction for addressing *CYP2D6* allelic diversity.

Comprehensive calling of PGx alleles may identify the presence of those star alleles that lack interpretation, akin to variants of unknown significance (VUS) for Mendelian disease. Although many clinical panels either do not assay for or exclude VUSs from their reportable range, it is important to identify these in the context of PGx. By excluding these variants from detection or reporting, individuals with these variants are often misclassified as *1 with high-confidence and reported as a normal metabolizer (and therefore low risk) when the true classification remains undetermined and interpretation should not be provided with any confidence. This is especially problematic for genotypes that include novel pLoFs that occur at low frequencies, which will be increasingly uncovered as cohort sizes increase. Functional assays will one day provide an interpretation for these unknown-function alleles, and necessitate updating those individuals' metabolic profiles. For those individuals who would be high risk after incorporating new functional assay results, reporting them initially as unknown rather than as no/low risk is more accurate and may hedge against misdosing and potentially dangerous adverse side effects. Although the overall population that is affected by these rare alleles is small, they are no less important for that individual's treatment.

Whereas comprehensive *CYP2D6* annotation is needed to address its high genetic diversity, assaying for all known alleles of the other CYP genes is no less important. *CYP2C9* and *CYP2C19* alleles make up more than half of the 16 alleles that were found to be rare in the EUR group but not rare in non-EUR groups (**Table S3**). It is worth pointing out that not all of these alleles are in the panels used to generate gold standard PGx calls by the Genetic Testing Reference Materials Coordination Program[37] (GeT-RM). Compounding this shortcoming of gold-standard assays is our finding that these ancestry-specific alleles span multiple non-EUR groups. Comprehensive panels and unbiased sequencing methods are thus crucial for the equitable clinical application of PGx testing.

Although the potential clinical impact of PGx screening has been estimated to be high based on predicted metabolic phenotype status alone,[2] it was unclear how this would translate to actual prescribing behavior in the general population. Here, we find that in at least one health system, more than 25% of the population may be impacted annually. This is likely an underestimate given that our analysis does not include self-medication behavior using over-the-counter drugs (such as over-the-counter ibuprofen). There are also potentially significant public health benefits to population screening for PGx status: with an estimated > 30% of patients with PM or UM *CYP2D6* experiencing adverse outcomes after being prescribed codeine, tramadol, oxycodone, or hydrocodone (Sauver *et al*. 2017), our observation that nearly ~ 40% of those having been prescribed codeine or oxycodone are at high risk for PGx effects is consistent and suggestive of one of many immediate improvements to be had. Furthermore, we anticipate that using PGx-guided drug therapy would only improve overall patient outcomes because the "right" drug would be delivered sooner and trigger less adverse outcomes.

As with any population study, there are limitations to acknowledge. Although our cohort is based on "all comers" with no medical or other specific targeted recruitment strategies, we do recognize the disproportionate distribution of ancestry represented in our cohort with an enrichment of EUR as compared to the general US population (see Methods, **Figure S1**), and the likely enrichment of more-educated individuals who tend to enroll in research studies who may bias medically associated outcomes.[38,39] Although our pipeline reports full-gene duplications and deletions—known or unknown—for all the genes, we only report known complex SVs. We also do not report here on novel rare missense variants that may have a strong effect on function, including gain-of-function, as these are more challenging to identify without functional evidence. Last, the medication data analyzed here for HNP was for prescriptions written, not filled, so it is possible the actual drug "taking" behavior differed from that prescribed. We do not believe that these biases undermine or diminish the results of this study; rather, we anticipate that improvements to any of these biases in future studies would likely only increase the population incidence of abnormal PGx phenotypes and more strongly argue for more comprehensive and accurate genotyping and reporting.

With the possibility of every person's PGx metabolic phenotype being easily integrated into their medical record, it is certain that individuals will receive improved medical care through truly personalized drug therapies. The Helix PGx database provides an extensive view into PGx variant, genotype, and resulting phenotypic

frequencies to help prioritize the functional studies of VUSs, design population-based clinical trials, as well as improve clinical interpretation of PGx results.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors would like to thank the generous contributions of individuals who enrolled in the Helix DNA Discovery Project and/or the Healthy Nevada Project; the Desert Research Institute staff who helped acquire and process the relevant medical records: Jim Metcalf and Andrew Joros; the Helix Laboratory, Bioinformatics, and Research teams for their ongoing support and efforts to build and maintain our analytical platform, and who reviewed this manuscript, particularly: Alex Bolze, Liz Cirulli, Andrew Dei Rossi, Magnus Isaksson, Sharoni Jacobs, and Simon White.

FUNDING

This work was paid for by Helix and/or by the Nevada Governor's Office of Economic Development Knowledge Fund (no. 14685) and by Renown Health (no. GR09183).

CONFLICT OF INTEREST

S.L., R.J., W.L., J.T.L., and N.L.W. are all employees of Helix. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

S.L., R.J., J.J.G., W.L., J.T.L., and N.L.W. wrote the manuscript. S.L., J.J.G., W.L., J.T.L., and N.L.W. designed the research. S.L. and N.L.W. performed the research. S.L. and N.L.W. analyzed the data. S.L., R.J., and N.L.W. contributed new reagents/analytical tools.

1. Cohn, I. *et al*. Genome sequencing as a platform for pharmacogenetic genotyping: a pediatric cohort study. *NPJ Genom. Med.* **2**, 19 (2017).
2. McInnes, G. *et al*. Pharmacogenetics at Scale: an analysis of the UK Biobank. *Clin. Pharmacol. Ther.* **109**, 1528–1537 (2021).
3. Verbeurgt, P., Mamiya, T. & Oesterheld, J. How common are drug and gene interactions? Prevalence in a sample of 1143 patients with CYP2C9, CYP2C19 and CYP2D6 genotyping. *Pharmacogenomics* **15**, 655–665 (2014).
4. Sauver, J.S. *et al*. CYP2D6 phenotypes are associated with adverse outcomes related to opioid medications. *Pharmacogenomics Pers. Med.* **10**, 217–227 (2017).
5. Chou, W.H. *et al*. Extension of a pilot study: impact from the cytochrome P450 2D6 polymorphism on outcome and costs associated with severe mental illness. *J. Clin. Psychopharmacol.* **20**, 246–251 (2000).
6. Verbelen, M., Weale, M.E. & Lewis, C.M. Cost-effectiveness of pharmacogenetic-guided treatment: are we there yet? *Pharmacogenomics J.* **17**, 395–402 (2017).
7. Cavallari, L.H. *et al*. Multi-site investigation of strategies for the clinical implementation of CYP2D6 genotyping to guide drug prescribing. *Genet. Med.* **21**, 2255–2263 (2019).
8. Hoshitsuki, K. *et al*. Challenges in clinical implementation of CYP2D6 genotyping: choice of variants to test affects phenotype determination. *Genet. Med.* **22**, 232–233 (2020).
9. Dalton, R. *et al*. Interrogation of CYP2D6 structural variant alleles improves the correlation between CYP2D6 genotype and CYP2D6-mediated metabolic activity. *Clin. Transl. Sci.* **13**, 147–156 (2020).
10. CPIC Genes-Drugs. CPIC (2020). <https://cpicpgx.org/genes-drugs/>.
11. Numanagić, I. *et al*. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nat. Commun.* **9**, 828 (2018).
12. Lee, S.-B. *et al*. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet. Med.* **21**, 361–372 (2019).
13. Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
14. Manichaikul, A. *et al*. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
15. Cirulli, E.T. *et al*. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* **11**, 542 (2020).
16. CPIC & PharmGKB CYP2C19 Allele Functionality Reference. (2020). <https://api.pharmgkb.org/v1/download/file/attachment/CYP2C19_allele_functionality_reference.xlsx>.
17. CPIC & PharmGKB CYP2D6 Allele Functionality Reference. CYP2D6 Allele Functionality Reference (2020). <https://api.pharmgkb.org/v1/download/file/attachment/CYP2D6_allele_functionality_reference.xlsx>.
18. CPIC & PharmGKB CYP2C9 Allele Functionality Reference. (2019). <https://api.pharmgkb.org/v1/download/file/attachment/CYP2C9_allele_functionality_reference.xlsx>.
19. Johnson, J.A. *et al*. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. *Clin. Pharmacol. Ther.* **102**, 397–404 (2017).
20. CPIC & PharmGKB CYP4F2 Allele Functionality Reference. (2019). <https://api.pharmgkb.org/v1/download/file/attachment/CYP2D6_allele_functionality_reference.xlsx>.
21. CPIC CYP2C9 Diplotype-Phenotype Table. (2019). <https://api.pharmgkb.org/v1/download/file/attachment/CYP2C9_Diplotype_Phenotype_Table.xlsx>.
22. CPIC CYP2C19 Diplotype-Phenotype Table. (2020). <https://api.pharmgkb.org/v1/download/file/attachment/CYP2C19_Diplotype_Phenotype_Table.xlsx>.
23. CPIC CYP2D6 Diplotype-Phenotype Table. (2019). <https://api.pharmgkb.org/v1/download/file/attachment/CYP2D6_Diplotype_Phenotype_Table.xlsx>.
24. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
25. Consortium, G.P. *et al*. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
26. PharmGKB Downloads. PharmGKB (2020). <https://www.pharmgkb.org/downloads>.
27. FDA Table of Pharmacogenetic Associations. (2020). <https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-associations>.
28. Zhou, Y., Ingelman-Sundberg, M. & Lauschke, V.M. Worldwide distribution of cytochrome P450 alleles: a meta-analysis of population-scale sequencing projects. *Clin. Pharmacol. Ther.* **102**, 688–700 (2017).
29. Del Tredici, A.L. *et al*. Frequency of CYP2D6 alleles including structural variants in the United States. *Front. Pharmacol.* **9**, 305 (2018).
30. Caudle, K.E. *et al*. Standardizing CYP2D6 genotype to phenotype translation: consensus recommendations from the clinical pharmacogenetics implementation consortium and Dutch Pharmacogenetics Working Group. *Clin. Transl. Sci.* **13**, 116–124 (2020).
31. Grzymski, J.J. *et al*. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. *Nat. Med.* **26**, 1235–1239 (2020).

32. Lynch, T. & Price, A. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physician* **76**, 391–396 (2007).
33. Amorosi, C.J. Massively parallel functional profiling of CYP2C9 variants using a yeast activity assay. bioRxiv preprint https://doi.org/10.1101/2021.03.12.435209.
34. Zhang, L. *et al*. CYP2C9 and CYP2C19: deep mutational scanning and functional characterization of genomic missense variants. *Clin. Transl. Sci.* **13**, 727–742 (2020).
35. Suiter, C.C. *et al*. Massively parallel variant characterization identifies alleles associated with thiopurine toxicity. *Proc. Natl. Acad. Sci. USA* **117**, 5394–5401 (2020).
36. McInnes, G. *et al*. Transfer learning enables prediction of CYP2D6 haplotype function. *PLoS Comput. Biol.* **16**, e1008399 (2020).
37. Reference materials for pharmacogenetics. (2019) <https://www.cdc.gov/labquality/get-rm/inherited-genetic-diseases-pharmacogenetics/pharmacogenetics.html>.
38. Spitzer, S. Biases in health expectancies due to educational differences in survey participation of older Europeans: It's worth weighting for. *Eur. J. Health Econ.* **21**, 573–605 (2020).
39. Cutler, D.M. & Lleras-Muney, A. *Education and Health: Evaluating Theories and Evidence*, NBER Working Paper Series. (National Bureau of Economic Research, Cambridge, MA, 2006).