PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Michigan Genomics Initiative Freeze 5 Star Allele and Activity Phenotype Inferences

Brett Vanderwerff[1*], Matthew Zawistowski[1], Lars G. Fritsche[1], Emily Bertucci-Richter[1], Snehal Patil[1,2], Michael Boehnke[1], Xiang Zhou[1], and Sebastian Zöllner[1,3]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. [3]Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA

*To whom correspondence regarding data preparation should be addressed: brettva@umich.edu

1 Overview

Pharmacogenomics (PGx) studies the interaction between genetics and pharmaceuticals and provides insights into variations in drug efficacy and toxicity among individuals. The information obtained from PGx studies can inform personalized drug treatment approaches. In PGx studies, the star allele nomenclature system is used to summarize genetic variation into alleles for genes that interact with pharmaceuticals, also known as pharmacogenes. Many star alleles have been annotated with phenotype information in the form of pharmacogene activity levels through various curation efforts.[1]

Biobanks such as the Michigan Genomics Initiative (MGI) and the UK Biobank are large cohort studies that link genetic and electronic health record data that may include medication history, clinical notes, quantitative laboratory tests, and diagnostic billing codes.[2,3] The availability of electronic health record data and activity phenotype inferences based on star alleles for biobank participants provides the unique opportunity for performing large scale retrospective PGx studies in biobanks. For example, inferred activity phenotypes were recently used in the UK Biobank to test for association with drug maintenance dose and drug response[4] and in both MGI and the UK Biobank to predict drug gene interactions among participants.[5,6]

In this MGI Freeze 5 data release, we offer star allele and activity phenotype inferences for 70,266 participants in the MGI cohort. We provide these inferences for the transporter pharmacogenes ABCG2 and SLCO1B1 and the enzyme pharmacogenes CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A4, CYP3A5, DPYD, TPMT, UGT1A1, UGT1A4, and UGT2B15 using genotypes imputed in MGI from the TOPMed reference panel and the star allele and phenotype inference software PyPGx[7] (Figure 1).

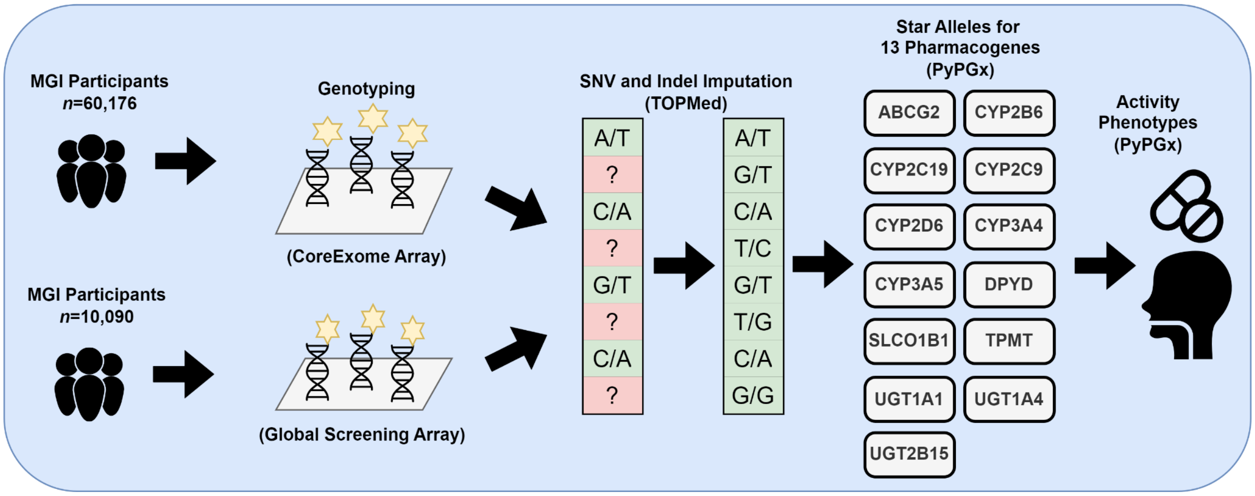

Figure 1. Schematic of Star Allele and Phenotype Inference in MGI. We used genotypes obtained from the CoreExome array or Global Screening Array as target and the TOPMed panel as reference to impute single nucleotide variants (SNV) and indels in MGI. We translated imputed SNV and indels to star alleles and corresponding activity phenotype inferences for 13 pharmacogenes using the software PyPGx.

2 Methods

2.1 Genetic Data

We describe genotyping, genotype imputation, and sample and variant quality control for participants of the MGI elsewhere.[2] Briefly, participants were genotyped on one of 3 synthesis batches of a customized Illumina Infinium CoreExome v1.0, v1.1, or v1.3 (n=60,176) array or an Illumina Infinium Global Screening Array v1.3 (n=10,090). We excluded sites with Hardy Weinberg p < 1x10^{-4}, Gentrain score < 0.15, Cluster Separation score < 0.3, or call rate < 99%. We performed haplotype phasing and genotype imputation in separate batches for samples assayed on the CoreExome arrays and the Global Screening Array and imputed short insertion deletion (indel) and single nucleotide variants (SNV) in MGI using the TOPMed reference panel. We filtered the imputed genotypes to exclude sites with minor allele frequency < 0.01% or estimated imputation quality (Rsq) < 0.3.

2.2 Genetic Ancestry

We inferred the genetic ancestry of MGI participants using the software ADMIXTURE and a reference panel of Human Genome Diversity Project samples and super-population labels for Central/South Asian, East Asian, West Asian, Native American, African, European, and Oceanian.[8,9] We labeled MGI participants as European genetic ancestry if the European Q value (global ancestry fraction) reported by ADMIXTURE was > 0.9.

2.3 Star Allele and Activity Phenotype Inference

We inferred star alleles and activity phenotypes for pharmacogenes ABCG2, CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A4, CYP3A5, DPYD, SLCO1B1, TPMT, UGT1A1, UGT1A4, and UGT2B15 using the "run-chip-pipeline" command from PyPGx v0.19.0 using phased, "best guess" genotypes imputed in MGI from the TOPMed reference panel as input.[7] We ran PyPGx with the argument "--assembly" set to GRCh38, the genome build of the TOPMed reference panel. Our star allele and phenotype inference approach thus leveraged SNVs and indels that can be imputed using this reference panel. However, this data provides no structural variation information for CYP2D6 or any other pharmacogene we evaluated.

2.4 Diplotype Concordance

We compared star allele inferences made in MGI by PyPGx to commercially available "ground truth" pharmacogenomics tests for DPYD (Mayo Clinic Laboratories, Rochester, MN, USA), TPMT (Prometheus Laboratories, San Diego CA, USA), UGT1A1 (Mlabs, Ann Arbor MI, USA), CYP2C19 (ARUP Laboratories, Salt Lake City UT, USA or Assurex Health, Mason OH, USA) and CYP2B6, CYP2C9, CYP2D6, CYP3A4, UGT1A4, and UGT2B15 (Assurex Health).

Methodological differences between the ground truth tests and PyPGx affected the comparisons we could make between diplotypes in our concordance evaluation. Each ground truth test used polymerase chain reaction to evaluate a pre-selected set of core variants which were then translated into the star alleles that each test covers. The set of star alleles that we were able to infer with PyPGx was limited to those star alleles where all the core variants used to define each allele were imputed with high quality in MGI. In addition, differences exist in the methods used by the ground truth tests and PyPGx for the reporting of ”default calls" (calls that are reported in the absence of any tested star alleles) among a subset of the pharmacogenes we evaluated. Given the differences in coverage and default calling between the ground truth tests and PyPGx we expect some differences in the results, thus we made some exceptions for which alleles can be compared (see Supplementary Methods).

As phase information is not reported from the ground truth tests, we restricted our comparisons between PyPGx and the ground truth test to the diplotype level. To account for coverage differences, we further restricted comparisons between PyPGx and the ground truth test to cases where both haplotypes of each diplotype were covered by both methods with the exception that we always considered CYP2D6 duplication and deletion calls from the ground truth test discordant with PyPGx, which lacks the ability to call structural variants from arrays. We considered diplotypes concordant if the alleles of both haplotypes matched between PyPGx and the ground truth test. We expressed the diplotype concordance rate as the percentage of concordant diplotypes out of the total number of compared diplotypes.

2.5 Star Allele and Activity Phenotype Frequency Resources

We calculated star allele frequency in MGI using the star alleles in the main diplotype reported by PyPGx. In addition to the top ranked star alleles that are included in the main diplotype call, PyPGx also outputs lower ranked candidate star alleles for which the constellation of SNVs and indels on each haplotype also fit. We did not include candidate star alleles in the frequency calculation. We obtained estimated star allele and activity phenotype frequency information from the "PGx Gene-specific Information Tables" available from the Pharmacogenomics Knowledgebase (PharmGKB, accessed 4/3/2023).[10] Due to the limited sample size of participants with non-European genetic ancestry in MGI, we limited our comparisons to the "European" biogeographical group in PharmGKB and MGI participants with inferred European genetic ancestry by ADMIXTURE.[11]

Several caveats exist to using frequency data compiled from PharmGKB. PharmGKB notes that allele frequency estimates are often based on small samples that are biased towards individuals with a specific phenotype or drug exposure and almost no studies use genetics to describe ancestry. Also, PharmGKB does not summarize activity phenotype frequency directly from observed phenotypes, but rather uses star allele frequency estimates based on literature sources and the Hardy Weinberg Equilibrium equation to estimate diplotype and activity phenotype frequency.[12] Error at each of these steps will compound and may be high for less studied pharmacogenes. As such, we consider our comparisons of allele and phenotype frequency to sources compiled by PharmGKB to be only general estimates of inferred star allele accuracy in MGI.[13]

3 Data Quality Evaluation

3.1 Diplotype Concordance

We inferred star alleles for 13 pharmacogenes: ABCG2, CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A4, CYP3A5, DPYD, SLCO1B1, TPMT, UGT1A1, UGT1A4, and UGT2B15 for 70,266 participants included as a part of MGI Genetic Data Freeze 5. We evaluated the accuracy of star allele inferences at the diplotype level for pharmacogenes CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A4, DPYD, TPMT, UGT1A1, UGT1A4, and UGT2B15 in a subset of MGI participants by the retrospective analysis of star allele calls from commercially available "ground truth" pharmacogenomics tests.

The use of SNV-based testing by the ground truth tests and the use of imputed genotypes to infer star alleles in MGI leads to coverage differences that affect the comparisons that can be made when evaluating concordance between these sources (see Supplementary Methods). To express these coverage differences, we computed the percentage of comparable diplotypes as the number of diplotypes that are sufficiently covered between these sources for use in our concordance evaluation out of the total number of diplotypes we observed. The percentage of comparable diplotypes ranged from 17.3% for UGT1A1 to 100% for TPMT and UGT2B15 (Table S1).

Diplotype concordance between star allele inferences in MGI and the ground truth test was generally high and ranged from 84% at CYP2D6 to 100% at CYP2B6, CYP2C19, CYP3A4, DPYD, TPMT, UGT1A1, and UGT1A4 (Table 1). Five of the seven discordant diplotypes for

CYP2D6 involved structural variants reported by the ground truth test: three CYP2D6 *5 calls (whole gene deletion) and two CYP2D6 gene duplications.

Gene	# of Diplotypes	# of Concordant Diplotypes	% Concordant
CYP2B6	31	31	100
CYP2C19	8	8	100
CYP2C9	42	41	98
CYP2D6	44	37	84
CYP3A4	43	43	100
DPYD	1	1	100
TPMT	67	67	100
UGT1A1	9	9	100
UGT1A4	43	43	100
UGT2B15	44	43	98

Table 1. Diplotype Concordance. The numbers of concordant and compared diplotypes for each pharmacogene where we compared PyPGx inferences to commercial star allele calling platforms. Concordance is calculated as the percentage of concordant diplotypes out of all diplotypes evaluated for each pharmacogene.

3.2 Star Allele Frequency

We evaluated how closely the frequency of star allele inferences in MGI participants agreed with literature sources. We calculated the frequencies of inferred star alleles for ABCG2, CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A5, SLCO1B1, and TPMT in 53,966 European genetic ancestry MGI participants and compared to literature sources compiled by Pharmacogenomics Knowledgebase (PharmGKB) for the European biogeographical group.[10,14,15] We compared frequencies between MGI and PharmGKB at star alleles that were represented in both datasets with a frequency > 0.1%. 56 of the 112 distinct star alleles observed at least once in MGI across the eight pharmacogenes intersected the PharmGKB set and had an allele frequency > 0.1% in either data set.

The frequency of star alleles inferred in MGI generally agreed with sources compiled from PharmGKB. The square of the Pearson correlation coefficient (R^2) between star allele frequency in MGI and PharmGKB within the European population ranged from .584 in CYP2B6 to .999 in CYP3A5 (Figure 2). While the agreement between star allele frequency in MGI and PharmGKB was typically the highest among star alleles with frequency > ~1%, some examples of large discordance exist. The 2 largest absolute differences in star allele frequency were for

CYP2B6 *9 (~23% in MGI compared to ~1% in PharmGKB) and SLCO1B1 *37 (~6% in MGI compared to ~25% in PharmGKB).

Frequency discordance between MGI and PharmGKB for CYP2B6 stems from coverage differences. CYP2B6 *9 is defined by 19:41006936G>T, a core variant for 16 distinct star alleles including CYP2B6 *6 which is defined by the core variants 19:41006936G>T and 19:41009358A>G.[16] 19:41009358A>G is not currently in the TOPMed reference panel and thus is not imputed in MGI resulting in an allele frequency of 0% for CYP2B6 *6 in MGI compared to ~23% in PharmGKB. Participants in MGI that are true carriers of both 19:41006936G>T and 19:41009358A>G on the same haplotype will likely be inferred CYP2B6 *9 by PyPGx, inflating the frequency of this star allele in MGI relative to studies with coverage for both CYP2B6 *6 and *9. Both CYP2B6 *6 and *9 are “decreased function” star alleles and are handled similarly by the PyPGx phenotyping approach.

Frequency discordance between MGI and PharmGKB for SLCO1B1 is likely related to limited sample size in PharmGKB and coverage differences. SLCO1B1 *37 is defined by 12:21176804A>G, a core variant for 19 distinct star alleles, 9 of which are observed at least once in the European sample of MGI (SLCO1B1 *14, *15, *20, *27, *28, *30, *31, *37, and *43).[17] Since we compute star allele frequency from the top-ranked alleles called by PyPGx, carriers of 12:21176804A>G will be distributed among all 9 star alleles containing this core variant, 8 of which will effectively “absorb” calls that would have otherwise gone to *37. PharmGKB summarizes the frequency of SLCO1B1 *37 in Europeans across 12 studies which range in their reporting of SLCO1B1 *37 frequency from 5.3%[18] to 34.5%.[19] Half of these PharmGKB studies only evaluated SLCO1B1 *37 and *15 among the set of possible alleles containing 12:21176804A>G. Ramsey et al. previously discussed limitations related to sample size and coverage of PharmGKB’s frequency reports for SLCO1B1 alleles (including SLCO1B1 *37).[20]

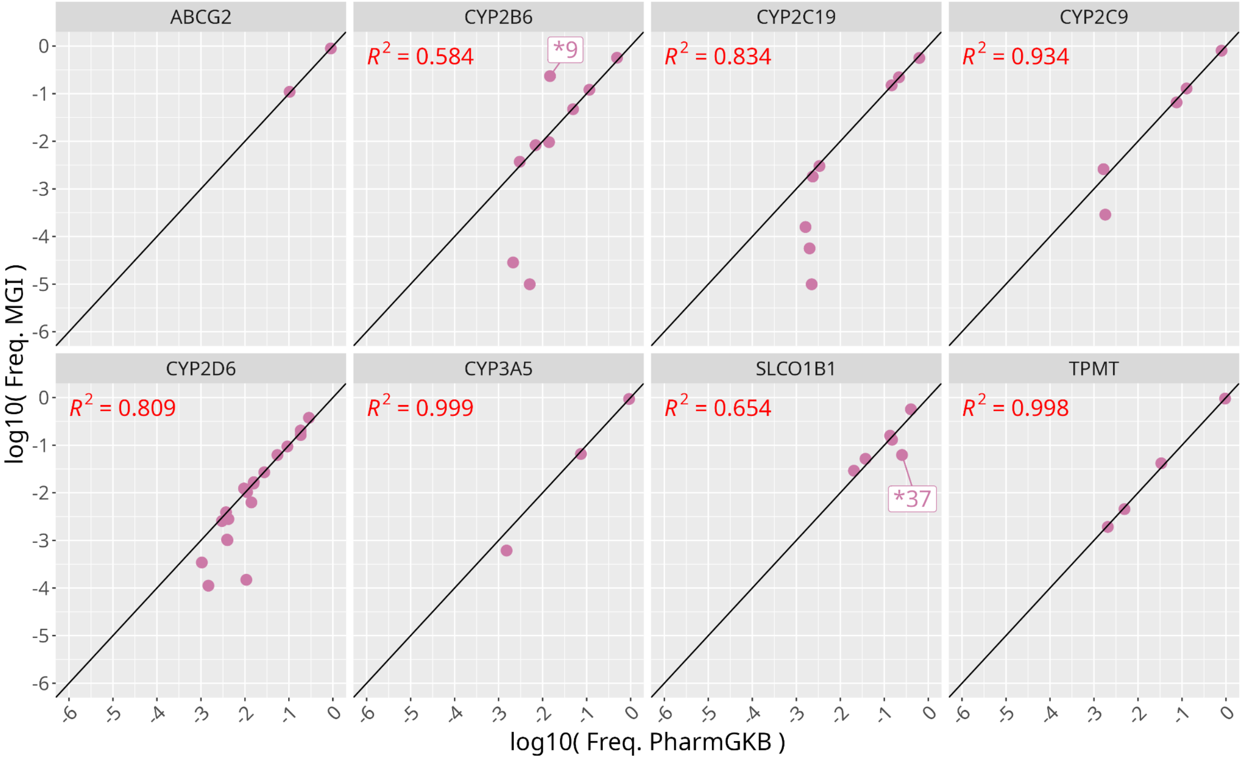

Figure 2. Star Allele Frequency in MGI and PharmGKB. Only star alleles that are both called in MGI and reported by PharmGKB with a frequency > 0.1% are plotted. R^2 is the square of the Pearson correlation coefficient between log transformed star allele frequency observed in MGI and summarized by PharmGKB. Select star alleles with high frequency discordance between each data set are labeled. Pharmacogene ABCG2 had an insufficient number of star alleles to calculate R^2 . Freq: frequency.

3.3 Activity Phenotype Frequency

We evaluated how closely the frequency of inferred activity phenotypes in MGI participants agreed with literature sources. We calculated the frequencies of inferred activity phenotypes for ABCG2, CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A5, SLCO1B1, and TPMT in 53,966 European genetic ancestry MGI participants and compared to literature sources compiled by PharmGKB for the European biogeographical group.[10,14,15] Activity phenotypes follow the consensus terms of increased, normal, decreased, and poor function for transporters and ultrarapid, rapid, normal, intermediate, and poor metabolizer for enzymes.[1] We compared only the activity phenotypes that were both inferred in MGI and reported in PharmGKB with a non-zero frequency (Figure 3).

The frequency of activity phenotypes inferred in MGI generally agreed with sources compiled from PharmGKB, particularly for phenotypes with a frequency > ~0.1%. R^2 between activity phenotype frequency in MGI and PharmGKB within the European population ranged from .938 for CYP2B6 to 1 for ABCG2, CYP2C9, and CYP3A5. The 3 largest fold change differences in star allele frequency were for CYP2B6 ultrarapid metabolizers (~.01% in MGI compared to ~.3% in PharmGKB), CYP2C19 likely intermediate metabolizers (~.01% in MGI compared to ~0.1% in PharmGKB), and CYP2B6 rapid metabolizers (~1.4% in MGI compared to ~7.2% in PharmGKB).

Notable phenotypes that were missing in MGI or PharmGKB were CYP2D6 ultrarapid metabolizers (0% in MGI compared to ~2.3% in PharmGKB) and CYP2C19 indeterminate metabolizers (~.026% in MGI compared to 0% in PharmGKB). CYP2D6 normal metabolizers are inferred to occur at higher frequency in MGI (~56%) compared to PharmGKB (~49%). While we observed some of the highest frequency discordance between MGI and PharmGKB among rare CYP2B6 and CYP2C19 phenotypes, particularly among those with a frequency < ~0.1%, many of the PGx studies used by PharmGKB to estimate phenotype frequency from observed star allele frequency are typically small with much less than 1000 subjects, so our ability to evaluate the frequency of these phenotypes in MGI by comparing to PharmGKB is limited.

We emphasize that we do not currently infer CYP2D6 structural variants such as duplications in MGI and thus are not able to infer structural variant-dependent CYP2D6 ultrarapid metabolizers. Similarly, we do not currently infer the loss-of-function whole gene deletion CYP2D6 *5, which occur at a frequency of ~2.9% in Europeans.[21] We expect CYP2D6 deletion and duplication star alleles in MGI will often be replaced by the normal function CYP2D6 *1 by PyPGx and a relatively higher frequency of CYP2D6 normal metabolizers in MGI compared to PharmGKB is expected.

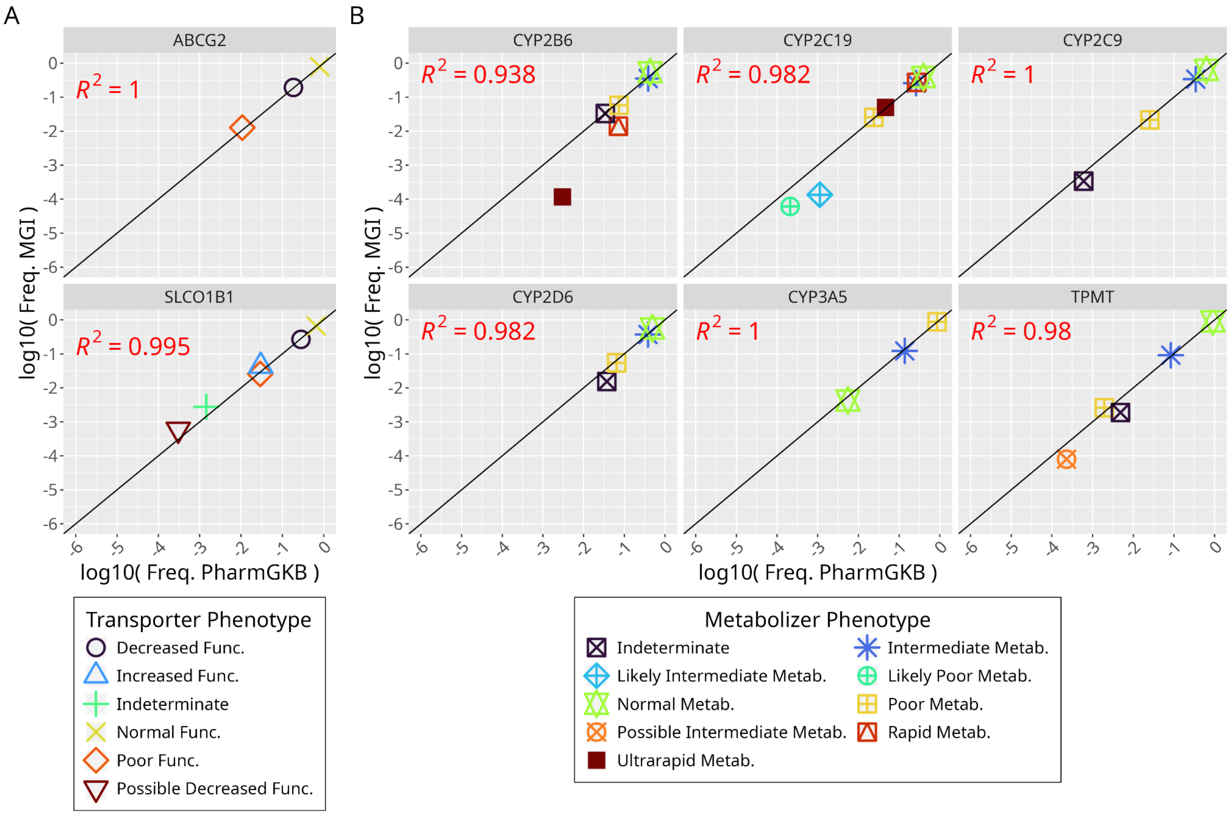

Figure 3. Activity Phenotype Frequency in MGI and PharmGKB. R^2 is the square of the Pearson correlation coefficient between log transformed **A** transporter phenotype or **B** metabolizer phenotype frequency observed in MGI and summarized by PharmGKB. Freq: frequency; Metab: Metabolizer.

4 Data Format

These data are available in a tab delimited text file format with rows for participants and columns describing the genetic and phenotypic data. Here we provide a data dictionary (adapted from the PyPGx documentation) to aid in results interpretation.[22]

Column Name	**Description**	**Example Data**
IID	DeID_PatientID, unique identifier for the participant	xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx
Genotype	The diplotype for the participant, which is generated from selecting the highest ranked candidate star alleles from each haplotype	*1/*5
Phenotype	The phenotype for the participant (derived from the genotype)	Indeterminate
Haplotype1	The candidate star allele(s) for the first haplotype	*1;
Haplotype2	The candidate star allele(s) for the second haplotype	*5;*2;
AlternativePhase	Other possible star allele(s) if phasing error is assumed	*4;
VariantData	Relevant star allele core definitions	*5:4-68647129-T-G;*2:default;
Gene	Pharmacogene for which the calls are for	UGT2B15

5 Supplementary Tables

Gene	# Incomparable Diplotypes	# Comparable Diplotypes	% Comparable Diplotypes
CYP2B6	13	31	70.45
CYP2C19	38	8	17.39
CYP2C9	3	42	93.33
CYP2D6	1	44	97.78
CYP3A4	1	43	97.73
DPYD	1	1	50
TPMT	0	67	100
UGT1A1	41	9	18
UGT1A4	1	43	97.73
UGT2B15	0	44	100

Table S1. Comparable Diplotypes. The number (and percentage) of diplotypes that were comparable in the concordance evaluation between the ground truth commercial tests and PyPGx inference in MGI out of all diplotypes we observed.

6 Supplementary Methods

6.1 Diplotype Concordance

We made exceptions to comparisons where either PyPGx or the ground truth test did not interrogate the full set of core variants required for a perfect match or where default alleles differed between the sources. Descriptions of these exceptions on a per gene basis are as follows:

CYP2B6

We compared CYP2B6 *9 (19:41006936:G>T) calls from PyPGx to CYP2B6 *6 (19:41006936:G>T,19:41009358:A>G) calls from the Assurex Health test based on the shared SNV 19:41006936:G>T. 19:41009358:A>G is not imputed in MGI and thus CYP2B6 *6 is not interrogated by PyPGx. The CYP2B6 *6 calls from the Assurex Health test would have been CYP2B6 *9 calls had the test not interrogated 19:41009358:A>G. Both CYP2B6 *6 and *9 are phenotyped by PyPGx as “decreased function” alleles.

CYP2C19

We did not allow comparisons of diplotypes where either the Assurex Health test or PyPGx diplotype contained a CYP2C19 *1 call or where PyPGx called CYP2C19 *38. The Assurex Health test defaults to CYP2C19 *1 in the absence of any other tested variation. PyPGx evaluates CYP2C19 *1 as 10:94842866A>G and defaults to CYP2C19 *38 (10:94842866G>A) in the absence of any other tested variation as “A” is the reference allele for build GRCh38 at position 10:94842866. Unlike PyPGx, the Assurex Health test does not directly interrogate variation at position 10:94842866, thus comparisons including CYP2C19 *1 or *38 are ambiguous. Both CYP2C19 *1 and *38 are phenotyped by PyPGx as “normal function” alleles.

We compared calls for the “decreased function” allele CYP2C19 *2 from PyPGx and the Assurex Health test based on the shared SNV 10:94781859G>A, although the core variant definitions from each of these sources differ. CYP2C19 *2 is defined by PyPGx as 10:94775367A>G,10:94781859G>A and by the Assurex Health test as 10:94781859G>A. While both 10:94775367A>G and 10:94781859G>A are annotated as variants that alter the function of CYP2C19 by the Pharmacogene Variation Consortium (PharmVar) Version 5.2.22, 10:94781859G>A is unique to CYP2C19 *2 and not included in the core variant definition for any other CYP2C19 allele, thus we consider its presence sufficient to distinguish this allele in these comparisons.[23–26]

CYP2D6

We compared between any of CYP2D6 *2 (22:42126611C>G,22:42127941G>A), CYP2D6 *2A (22:42126611C>G,22:42127941G>A,22:42132375G>C), CYP2D6 *35 (22:42126611C>G,22:42127941G>A,22:42130761C>T), and CYP2D6 *45 (22:42126611C>G,22:42127941G>A,22:42129075C>T) calls based on the shared SNVs 22:42126611C>G and 22:42127941G>A. PyPGx does not interrogate CYP2D6 *2A and the Assurex Health test does not interrogate CYP2D6 *35 or CYP2D6 *45. CYP2D6 *2, *35, and *45 are phenotyped by PyPGx as “normal function” alleles. Similarly, PharmVar annotates the legacy labeled CYP2D6 *2A allele as “normal function”.[27]

We compared CYP2D6 *4 calls from PyPGx and the Assurex Health test based on the shared SNV 22:42128945C>T, although the core variant definitions from each of these sources differ. CYP2D6 *4 is defined by PyPGx as 22:42128945C>T and by the Assurex Health test as 22:42126611C>G,22:42128945C>T,22:42130692G>A. Both sources interrogate 22:42128945C>T, the single core variant for CYP2D6 *4 defined by PharmVar and annotated as “no function”.[27]

We compared calls for the “decreased function” allele CYP2D6 *17 from PyPGx and the Assurex Health test based on the shared SNV 22:42129770G>A, although the core variant definitions from each of these sources differ. CYP2D6 *17 is defined by PyPGx as 22:42129770G>A and by the Assurex Health test as 22:42126611C>G,22:42127941G>A,22:42129770G>A. 22:42126611C>G,22:42127941G>A are the core variants for CYP2D6 *2 which is phenotyped by PyPGx as a “normal function” allele. The developers of PyPGx chose to define CYP2D6 *17 without 22:42126611C>G,22:42127941G>A with the assumption that these variants do not affect enzymatic activity and to guard against phenotype misclassification from defaulting to CYP2D6

*1 when calls are missed at 22:42126611C>G or 22:42127941G>A (Seung-been Lee, personal communication, May 3, 2023).

DPYD

We compared calls for the "decreased function" allele DPYD HapB3 from PyPGx and the Mayo Clinic Laboratories test, although the core variant definitions from each of these sources differ. DPYD HapB3 is defined by PyPGx as 1:97573863C>T,1:97579893G>C and by the Mayo Clinic Laboratories test as 1:97579893G>C. 1:97573863C>T is a synonymous tag SNV in complete linkage with 1:97579893G>C, the SNV thought to be responsible for the reduced function of HapB3 via aberrant splicing, thus we consider the presence of 1:97579893G>C sufficient to distinguish this allele in these comparisons.[28,29]

UGT1A4

We compared UGT1A4 *3A (2:233718602C>T,2:233718658G>A,2:233718962T>G) calls from PyPGx to UGT1A4 *3 (2:233718962T>G) calls from the Assurex Health test based on the shared SNV 2:233718962T>G. The Assurex Health test does not interrogate 2:233718602C>T and 2:233718658G>A. Both UGT1A4 *3A and *3 (2:233718962T>G, named *3B by PyPGx) are phenotyped by PyPGx as "decreased function" alleles.

UGT2B15

We compare UGT2B15 *1 calls from PyPGx to UGT2B15 *1 calls from the Assurex Health test, although their designation as default alleles differ between these sources. The Assurex Health test evaluates a single allele UGT2B15 *2 (4:68670366C>A) and defaults to UGT2B15 *1 when UGT2B15 *2 is not detected. PyPGx evaluates UGT2B15 *1 as 4:68670366A>C and defaults to *2 in the absence of any other tested variation as "A" is the reference allele for build GRCh38 at position 4:68670366. In this scenario both sources are directly interrogating SNV at position 4:68670366

We compared UGT2B15 *5 (4:68647129T>G) calls from PyPGx to UGT2B15 *2 (4:68670366C>A) calls from the Assurex Health test. The SNV 4:68647129T>G is not interrogated by the Assurex Health test. The UGT2B15 *5 calls from PyPGx would default to UGT2B15 *2 had PyPGx not interrogated 4:68647129T>G. Both UGT2B15 *5 and UGT2B15 *2 are phenotyped by PyPGx as "unknown function" alleles.

We compared UGT2B15 *4 (4:68647129T>G,4:68670366A>C) calls from PyPGx to UGT2B15 *1 (4:68670366A>C) calls from the Assurex Health test based on the shared SNV 4:68670366A>C. The variant 4:68647129T>G is not interrogated by the Assurex Health test. The UGT2B15 *4 calls from PyPGx would have been UGT2B15 *1 calls had PyPGx not interrogated 4:68647129T>G. UGT2B15 *4 is phenotyped by PyPGx as "unknown function" and UGT2B15 *1 is phenotyped by PyPGx as "normal function".

7 References

1. Caudle, K. E. *et al.* Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genetics in Medicine* **19**, 215–223 (2017).
2. Zawistowski, M. *et al.* The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genomics* **3**, 100257 (2023).
3. Bycroft, C. *et al. Genome-wide genetic data on ~500,000 UK Biobank participants*. 166298 https://www.biorxiv.org/content/10.1101/166298v1 (2017) doi:10.1101/166298.
4. McInnes, G. & Altman, R. B. Drug Response Pharmacogenetics for 200,000 UK Biobank Participants. *Pac Symp Biocomput* **26**, 184–195 (2021).
5. McInnes, G. *et al.* Pharmacogenetics at Scale: An Analysis of the UK Biobank. *Clinical Pharmacology & Therapeutics* **109**, 1528–1537 (2021).
6. Pasternak, A. L. *et al.* Identifying the prevalence of clinically actionable drug-gene interactions in a health system biorepository to guide pharmacogenetics implementation services. *Clinical and Translational Science* **16**, 292–304 (2023).
7. Lee, S., Shin, J.-Y., Kwon, N.-J., Kim, C. & Seo, J.-S. ClinPharmSeq: A targeted sequencing panel for clinical pharmacogenetics implementation. *PLOS ONE* **17**, e0272129 (2022).
8. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
9. Li, J. Z. *et al.* Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319**, 1100–1104 (2008).
10. PGx Gene-specific Information Tables. *PharmGKB* https://www.pharmgkb.org/page/pgxGeneRef.
11. Huddart, R. *et al.* Standardized Biogeographic Grouping System for Annotating Populations in Pharmacogenetic Research. *Clin Pharmacol Ther* **105**, 1256–1262 (2019).

12. PGx Gene-specific Information Tables. *PharmGKB* https://www.pharmgkb.org/page/pgxGeneRef.

13. Gene-specific Information Tables for CYP2D6. *PharmGKB* https://www.pharmgkb.org/page/cyp2d6RefMaterials.

14. Relling, M. V. & Klein, T. E. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther* **89**, 464–467 (2011).

15. Whirl-Carrillo, M. *et al.* An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin Pharmacol Ther* **110**, 563–572 (2021).

16. Gene-specific Information Tables for CYP2B6. *PharmGKB* https://www.pharmgkb.org/page/cyp2b6RefMaterials.

17. Gene-specific Information Tables for SLCO1B1. *PharmGKB* https://www.pharmgkb.org/page/slco1b1RefMaterials.

18. Boivin, A.-A. *et al.* Organic anion transporting polypeptide 1B1 (OATP1B1) and OATP1B3: genetic variability and haplotype analysis in white Canadians. *Drug Metab Pharmacokinet* **25**, 508–515 (2010).

19. Aklillu, E. *et al.* Frequency of the SLCO1B1 388A>G and the 521T>C polymorphism in Tanzania genotyped by a new LightCycler®-based method. *Eur J Clin Pharmacol* **67**, 1139–1145 (2011).

20. Ramsey, L. B. *et al.* PharmVar GeneFocus: SLCO1B1. *Clinical Pharmacology & Therapeutics* **113**, 782–793 (2023).

21. Gene-specific Information Tables for CYP2B6. *PharmGKB* https://www.pharmgkb.org/page/cyp2b6RefMaterials.

22. Welcome to PyPGx's documentation! — pypgx documentation. https://pypgx.readthedocs.io/en/latest/.

23. PharmVar. https://www.pharmvar.org/gene/CYP2C19.

24. Gaedigk, A. *et al.* The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the

Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clin Pharmacol Ther* **103**, 399–401 (2018).

25. Gaedigk, A., Whirl-Carrillo, M., Pratt, V. M., Miller, N. A. & Klein, T. E. PharmVar and the Landscape of Pharmacogenetic Resources. *Clin Pharmacol Ther* **107**, 43–46 (2020).
26. Gaedigk, A., Casey, S. T., Whirl-Carrillo, M., Miller, N. A. & Klein, T. E. Pharmacogene Variation Consortium: A Global Resource and Repository for Pharmacogene Variation. *Clin Pharmacol Ther* **110**, 542–545 (2021).
27. PharmVar. https://www.pharmvar.org/gene/CYP2D6.
28. Froehlich, T. K., Amstutz, U., Aebi, S., Joerger, M. & Largiadèr, C. R. Clinical importance of risk variants in the dihydropyrimidine dehydrogenase gene for the prediction of early-onset fluoropyrimidine toxicity. *International Journal of Cancer* **136**, 730–739 (2015).
29. van Kuilenburg, A. B. P. *et al.* Intragenic deletions and a deep intronic mutation affecting pre-mRNA splicing in the dihydropyrimidine dehydrogenase gene as novel mechanisms causing 5-fluorouracil toxicity. *Hum Genet* **128**, 529–538 (2010).