UNIVERSITY OF CALGARY

Evaluating the Impact of Quality Control Filters and Imputation Reference Panels on Pharmacogenomic Variant Calling

by

Courtney Lenz

A THESIS

SUBMITTED TO THE CUMMING SCHOOL OF MEDICINE  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE  
DEGREE OF BACHELOR OF HEALTH SCIENCES HONOURS

Bachelor of Health Sciences  
Cumming School of Medicine  
University of Calgary  
Calgary, AB

March 2024

© Courtney Lenz 2024

# Abstract

Pharmacogenomics (PGx) has the potential to revolutionize drug treatment by leveraging individual genetic profiles. Despite the growing use of genome-wide genotyping arrays in PGx research, limited attention has been given to optimizing data preprocessing methods, including quality control (QC) and imputation. The aim of this project is to evaluate the impact of standard and modified QC filtering procedures and popular imputation panels from the 1000 Genomes (1000G) Project and the Trans-Omics for Precision Medicine (TOPMed) on PGx variant calling. PGx variant coverage was compared between preprocessing methods for *CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A4, CYP3A5, CYP4F2, NUDT15,* and *SLCO1B1*, with a focus on the Tier 1 (minimum set) and Tier 2 (optional set of alleles) recommended for PGx testing by the Association for Molecular Pathology. It was found that modified QC filtering retained significantly more Tier 1 and Tier 2 variants than standard QC filtering, however, there was little difference between the standard and modified datasets after imputation. Both imputation methods provided similar overall gene coverage with the exception of CYP2D6, in which 1000G provided significantly more coverage. Additionally, 1000G outperformed TOPMed for Tier 1 and Tier 2 allele coverage. Finally, it was discovered that both imputation methods struggled to impute indels, while TOPMed also struggled to impute more common variants. This study highlights that imputation has a more significant impact on PGx variant calling than QC; but given its limitations, it may be dangerous to rely on imputation to accurately “fill in” missing PGx variants. With the ever-increasing amount of GWAS data available for use in secondary analyses, it is important to consider that certain arrays used for genotyping may not be optimally suited for PGx applications regardless of the preprocessing methods used.

# Acknowledgements

I would like to express my sincere thanks to my supervisors Dr. Chad Bousman and Dr. Ankita Narang for their invaluable support and guidance throughout my time with the Psychiatric Pharmacogenomics (PsychPGx) Lab.

I am very grateful to Dr. Bousman for welcoming me into his lab and for his dedication to teaching me about the fascinating and albeit confusing world of pharmacogenomics. Thank you for your attentiveness and insight throughout this project.

I am exceptionally thankful to Dr. Narang who was instrumental to my success with this project. Thank you for your troubleshooting expertise, for helping me work through all the wrinkles in the pipeline, and for your dedication to this project. Your support and patience have helped me overcome numerous challenges and grow my skills as a bioinformatician.

Finally, I’d like to thank my family and friends, who have helped me push through during the toughest parts of this project and supported me throughout my degree. Thank you for listening to all my frustrations, for reading my papers, and for learning so much about bioinformatics against your will.

To my supervisors, family, and friends, you are all *stars*!

# Reflection

Overall, my project has been very rewarding. Although I knew going into the project that progress wouldn't be linear, I had to practice patience both with the programs I was working with and with myself. Troubleshooting code is one of the most frustrating and sometimes torturous tasks. But at the same time, finally finding that little bug and having the program run as expected is one of the best feelings ever. Its like solving a Sunday NYT crossword! I’ve learned a lot about myself throughout this project. Most notably, when faced with an issue, I become determined to solve it alone. While knowing how to independently work through problems is an important skill, knowing when to stop and ask for help is equally as valuable. This project has helped me recognize my limitations and take a more balanced approach to problem solving.

One of my favourite parts of this experience was learning how to use the high-performance computing cluster, ARC. Working with such large genomic datasets can be quite a challenge as the resources on the typical laptop are not nearly sufficient. While I had some experience with TALC in previous classes, it was very guided; assignments were highly structured, and commands were often provided. For this project, I had to determine how to troubleshoot program installation, run various programs, write shell and batch scripts, and estimate the amount of resources required per job, all without the guidance of an instructor. I also learned how to parallelize jobs on ARC, which in combination with access to immense computational resources, reduced runtime from over an hour to less than 10 minutes for my first parallel job! I believe learning to harness the power of parallel computing and knowing how to work effectively with HPCs will be the most valuable takeaway from this project for future applications.

Table of Contents

[Abstract ii](#_Toc162075100)

[Acknowledgements iii](#_Toc162075101)

[Reflection iv](#_Toc162075102)

[List of Tables vi](#_Toc162075103)

[List of Figures vi](#_Toc162075104)

[1. Introduction 1](#_Toc162075105)

[1.1 Background & Rationale 1](#_Toc162075106)

[1.2 Significance 3](#_Toc162075107)

[1.3 Literature Review 4](#_Toc162075108)

[2. Research Objectives 7](#_Toc162075109)

[2.1 Primary Objective 7](#_Toc162075110)

[2.2 Research Question 7](#_Toc162075111)

[2.3 Specific Aims 7](#_Toc162075112)

[2.4 Hypothesis 7](#_Toc162075113)

[3. Methods 8](#_Toc162075114)

[3.1 Data Sources 8](#_Toc162075115)

[3.2 Quality Control Filtering 8](#_Toc162075116)

[3.3 Imputation 9](#_Toc162075117)

[3.4 Assessing Coverage of PGx Variants in Unimputed and Imputed datasets 9](#_Toc162075118)

[4. Results 11](#_Toc162075119)

[4.1 Variant Retention Post QC Filtering 11](#_Toc162075120)

[4.2 PGx Variant Coverage Post Imputation 11](#_Toc162075121)

[4.3 Tier 1 & Tier 2 Star Allele Coverage 12](#_Toc162075122)

[4.4 Global Pharmacogene Coverage 13](#_Toc162075123)

[4.5 Imputation Run Time 14](#_Toc162075124)

[5. Discussion 14](#_Toc162075125)

[6. Conclusion 20](#_Toc162075126)

[Tables 21](#_Toc162075127)

[Figures 25](#_Toc162075128)

[References 28](#_Toc162075129)

[Supplementary Materials 32](#_Toc162075130)

# List of Tables

Table 1 **–** Number of samples and SNPs at each stage of quality control

Table 2 – Total number of PGx variants present in each dataset

Table 3 – Variant information for Tier 1 & Tier 2 indels missing from 1000G and TOPMed datasets.

Table 4 – Variant information for CYP2C19 Tier 2 variants missing from 1000G and TOPMed datasets.

Table 5 – Variant information for Tier 1 & Tier 2 SNPs missing in TOPMed datasets only.

Table 6 – Percent full coverage per gene for imputed datasets.

Table 7 – Comparison of imputation panels

Supplementary Table 1 – Variant information and presence for Tier 1 & Tier 2 variants across 7 datasets.

# List of Figures

Figure 1 – Quality control, imputation, and analysis flowchart

Figure 2 – Comparative coverage of Tier 1 & Tier 2 alleles identified from PharmVar variants

Figure 3 – Percent coverage of PharmVar genes by dataset

Figure 4 – Gene-specific coverage of PharmVar genes

Supplementary Figure 1 **–** Gene-specific coverage of PharmVar genes.

Supplementary Figure 2 **–** Code snippet for ‘find\_stars.sh’ script

Supplementary Figure 3 **–** Code snippet for ‘extract.sh’ script

Supplementary Figure 4 **–** Code snippet for ‘get\_stars.py’ script

Supplementary Figure 5 **–** Code snippet for ‘condense.py’ script

Supplementary Figure 6 **–** Code snippet for ‘haplotype\_coverage.java’ script

# 1. Introduction

## 1.1 Background & Rationale

As precision medicine transitions from theory to practice, the field of pharmacogenomics (PGx) facilitates this shift by exploring the genetic factors that influence drug response. PGx aims to improve drug safety and efficacy by investigating clinically actionable pharmacogenes. These genes play roles in pharmacokinetic functions, such as drug absorption, distribution, metabolism, and excretion (ADME), as well as in pharmacodynamic processes, such as the molecular interactions between a drug and its target.1 The pharmacogene haplotypes are typically represented using star allele nomenclature. Star alleles are comprised of one or more PGx variants and may alter the metabolizer phenotype (e.g., ultrarapid, normal, intermediate, or poor) for a given gene.2 By identifying genotype-phenotype correlations for pharmacogenes, PGx research can be used to predict an individual’s response to a medication.1

One of the most powerful tools for detecting genotype-phenotype associations is genome-wide association studies (GWAS). GWAS scans for genetic variants across populations to identify statistical correlations with a given phenotype. Due to the vast improvements in sequencing technology, single nucleotide polymorphism (SNP)-based genotyping arrays have become the preferred tool for use in GWAS as they are highly accurate and considerably more cost-effective than whole genome sequencing.3 Genome-wide genotyping arrays provide more extensive variant coverage in comparison to traditional genotyping arrays.4

As PGx analysis using genome-wide genotyping arrays has become increasingly common, the Association for Molecular Pathology (AMP) has identified a minimum set of alleles (Tier 1) for PGx testing and an optional extended panel of alleles (Tier 2).5 Tier 1 alleles have a significant impact on protein and/or gene function resulting in a noticeable effect on drug response, have an appreciable minor allele frequency (MAF) in a given population, and have available reference materials.5,6 Tier 2 alleles only need to satisfy at least one of the characteristics defined above for Tier 1 alleles.5,6 Due to the characterization and considerable frequency of these variants, they are commonly screened for in PGx testing; however, the majority of variants in pharmacogenes are rare (MAF < 1%).7,8,9 Moreover, calling common and rare PGx variants for genome-wide arrays can be complicated when applying standard quality control filtering and selecting imputation reference panels.

Quality control (QC) measures are required when processing raw SNP data from genotyping experiments, as variant calling relies on the clustering of signals.10 Poor-quality data and noise signals can cause misclustering, which can result in false positive and false negative results (type I and type II errors, respectively).11 In the context of GWAS analysis, type I errors may lead to the incorrect identification of genetic associations, while type II errors may result in genuine associations being overlooked. To improve the accuracy of variant-calling from GWAS data, genome-wide analysis applications, such as PLINK, are used to filter out poor-quality data.12

Following QC, imputation is commonly applied in GWAS analyses to “fill in” missing genotypic data, as SNP arrays have lesser genomic density in comparison to whole genome sequencing.13 The quality, size, and diversity of the reference panel directly influences the accuracy of imputation.13 Therefore, globally focused reference panels, such as the 1000 Genomes (1000G) Project and Trans-Omics for Precision Medicine (TOPMed) are often used for genomic imputation due to their robust size and diversity.14,15 Currently, it is unclear how the aforementioned data processing methods impact the ability to call either common or rare PGx variants.

## 1.2 Significance

Precision medicine has the potential to revolutionize healthcare in many ways, most notably by allowing treatments to be tailored to the individual based on their genetic makeup. Every year in the United States, approximately 2,216,000 individuals are hospitalized due to serve adverse drug reactions (ADRs), resulting in 106,000 deaths annually.16 Globally, ADRs are estimated to be between the 4th and 6th most common cause of death, disproportionately affecting pediatric and geriatric populations and low-income communities.17 Individualized treatment based on genetic factors could help significantly reduce the global burden of ADRs by predicting how an individual will metabolize and respond to specific medications.

The progression of precision medicine depends on accurate genetic data interpretation; however, there lacks a recommended pipeline for processing GWAS data specifically for PGx applications. This leaves potential for inaccuracies and errors in research results, which may pose risks when applying such research in a clinical setting. While there has been extensive research focused on data preprocessing methods and techniques for GWAS, few have focused their efforts towards PGx applications specifically. Additionally, since the establishment of biobanks, the availability of genotype data for use in secondary GWAS analyses has increased exponentially. Given the widespread use of genome-wide genotyping arrays in GWAS and the growing popularity of secondary GWAS analyses for PGx applications, understanding the impact of data preprocessing methods on PGx data quality is crucial to ensure the precision and reliability of PGx research. To address this knowledge gap, this project seeks to investigate how GWAS-based QC filters and the choice of imputation panel influence PGx variant calling, with the aim to provide guidelines for repurposing GWAS data for PGx research.

## 1.3 Literature Review

Effective QC for GWAS data requires strong genetics, statistics, and bioinformatics knowledge. To simplify this task, Turner et al. (2011) developed a protocol based on the standardized QC strategies established by the eMERGE (electronic MEdical Records and GEnomics).18 Current standard QC procedure involves filtering out data based on the following sample-based and SNP-based factors:

Sample-Level Filters:

1. Individual missingness: High levels of missing data for a particular individual often indicates poor quality or genotyping errors.18,19 Including this data could result in aberrant genotype calling.18

2. Sex discrepancies: Differences between the assigned sex and genotyped sex of an individual indicate potential sample handling errors.18,19

3. Heterozygosity rate: High levels of heterozygosity may indicate low sample quality or contamination, while low levels may indicate inbreeding.19

4. Sample relatedness: Given that GWAS operates on the assumption of unrelated subjects, including related individuals without proper adjustment can lead to biased estimations.18,19

5. Population stratification: The presence of multiple subpopulations. Given that allele frequencies might differ among ethnic groups, not accounting for potential population-specific variations may introduce type I and type II errors.18,19

SNP-Level Filters:

1. SNP missingness: Same as individual missingness but for SNP positions.18,19

2. Minor allele frequency (MAF): The frequency at which the second most common allele occurs. SNPs with low MAFs are less reliable, and SNP-phenotype associations are unlikely to be detected.18,19

3. Hardy-Weinberg equilibrium (HWE): Genetic variation in one population will remain constant from one generation to the next. Deviations from HWE are likely to be a result of genotyping errors; however, this may indicate evolutionary selection.18,19

Although QC is required in GWAS to ensure the accuracy and relevance of identified associations, it inherently leads to data loss. QC methods that are too strict could result in the loss of valuable data and may introduce type II errors. Previous studies have found that excluding low-quality SNPs instead of assigning them a reduced quality score weight can diminish the power of locus-based methods if the causal variant is of good quality.20,21,22 In 2021, Charon et al. examined the potential consequences of QC filtering and the subsequent data loss on imputation quality.22 This study found that without pre-filtration, which excluded SNPs with MAFs < 0.01, imputation results were reliable, and the absence of pre-filtration improved imputation for all classes of MAFs.22 Additionally, raising the imputation quality score r2 – a measure of imputation accuracy ranging from 0 for an unimputed variant to 1 for a perfectly imputed variant – from the standard threshold of 0.3 to 0.8 resulted in a 2.5-fold reduction of rare SNPs with a mean MAF < 0.001 and halved the number of very rare SNPs (MAF < 0.0005).22 Therefore, a less conservative approach to QC filtering was recommended to reduce information loss for rare and very rare variants.22

Genotype imputation is a statistical method that utilizes linkage disequilibrium (LD) to predict unobserved genotypes, which can be used to identify potential genetic associations that may otherwise be missed if relying solely on directly assayed variants.13 LD refers to the non-random association of alleles at different genetic loci within a population.23 This occurs because genetic variants are often inherited in groups due to their close proximity on a chromosome.13 Imputation algorithms can detect patterns of LD in a study sample by comparing the observed genetic variants with those in a reference panel.13 When a matching pattern is identified in the reference panel, the genotype of unobserved variants in LD with those observed in the study sample can be predicted.13 However, LD patterns can vary significantly between populations, even for populations with similar geographical origins.24 If LD patterns differ significantly between the reference population and the sample population, imputation accuracy may decrease.13

The accuracy of genotype imputation is primarily impacted by two key factors: the reference panel and the imputation algorithm.25 Historically, investigation into imputation accuracy has focused on the algorithm used.25 However, in a recent preprint, Li et al. (2023) examined the extent to which errors in the reference panel impact imputation performance.25 This study found that introducing perturbations into the reference panel decreased imputation accuracy.25 As MAF decreases, perturbation effects are amplified, while the imputation at common variants remains consistent regardless of the magnitude of the perturbation.25 Genotyping error rates tend to increase as MAF decreases, therefore, this study highlights the importance of accurate variant calls for rare variants within the reference panel.25

Considering that a majority of pharmacogenes are rare, the studies by Charon et al. (2021) and Li et al. (2023) suggest that strict QC measures and errors in the reference panel are likely to impact the imputation accuracy of PGx variants.

# 2. Research Objectives

## 2.1 Primary Objective

The primary objective of this project is to evaluate the influence of quality control filters and imputation reference panels on pharmacogenomic variant calling.

## 2.2 Research Question

Do commonly used quality control filters and imputation reference panels impact PGx variant calling when processing genome-wide association array data?

## 2.3 Specific Aims

1. Apply standard and modified quality control filtering to genome-wide association array data, followed by genotype imputation using commonly used reference panels.
2. Cross-reference pre-QC and post-QC imputed datasets with characterized pharmacogene variants.
3. Calculate the PGx variant coverage for each dataset and compare star allele calls across the datasets.

## 2.4 Hypothesis

It is hypothesized that standard quality control filters will negatively impact variant calling for some pharmacogenes, but not all. Additionally, PGx variant calling will differ depending on the reference panels used.

# 3. Methods

## 3.1 Data Sources

Genome-wide association data was sourced from 804 participants enrolled in the CLOZapine INternational (CLOZIN) consortium ([www.clozinstudy.com](http://www.clozinstudy.com)), a multi-center study to detect associations of clozapine response and clozapine-associated adverse drug reactions using extensive phenotypic and genetic data.26 The participants were between the ages of 18-25 and were primarily of European ancestry from the Netherlands, Germany, Austria, Finland, Turkey, and Australia. Blood samples from each participant were collected for DNA extraction.26 All CLOZIN participants were genotyped using Illumina’s Infinium Global Screening Array (GSA) V3.0.26 This genomic dataset was used in all of the subsequent processing detailed below.

Characterized star alleles for 9 genes known to influence drug response phenotypes were sourced from the Pharmacogene Variation (PharmVar) Consortium. PharmVar was specifically chosen due to the availability of star allele definitions for both genome builds GRCh37 and GRCh38.27,28 Haplotype tables for genes - *CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A4, CYP3A5, CYP4F2, NUDT15,* and *SLCO1B1* were collected and sub-allele variants were removed.28 GRCh37.p13 (Release 111) and GRCh38.p14 (Release 45) human reference genome assemblies were downloaded from Ensembl and Gencode, respectively.29,30

## 3.2 Quality Control Filtering

PLINK 2.0, an open-source whole genome association analysis toolset, was used to perform quality control filtering on the raw CLOZIN data.12 All individuals who failed sample-level QC filtering in the CLOZIN study were excluded from the dataset. These QC filters included the removal of samples with high individual missingness, sex discrepancies, extreme levels of heterozygosity or homozygosity, and samples of non-European ancestry.

To understand the impact of SNP-level QC on PGx variant calling, two variant filtering procedures, labelled as “standard” and “modified”, were applied. All variants with > 10% missingness rate were excluded in both standard and modified QC filtering methods. For standard filtering, 2 additional GWAS-based QC filters were applied to remove variants that deviate from Hardy-Weinberg equilibrium (HWE > 1.00 x 10-4) and variants with MAF < 0.01. BCFtools, was used to normalize the filtered datasets with the GRCh37 reference genome and convert the dataset files into VCF format.31

## 3.3 Imputation

Output files from the standard and modified QC filtering protocols were imputed using the 1000G Phase 3 v5 (GRCh37/hg19) and TOPMed r3 (GRCh38/hg38) reference panels.14,15 Imputation was conducted using the Michigan Imputation and TOPMed Imputation Servers.32,33 Both 1000G and TOPMed imputation methods were performed using Minimac4, version 1.7.4 and 1.8.0-beta4, respectively.34 Eagle v2.4 was used for phasing the imputed data.

To ensure the quality of imputed datasets, only variants with r2 > 0.5 were included post-imputation using BCFtools. After variant normalization, the VCF files were used for further analysis.

## 3.4 Assessing Coverage of PGx Variants in Unimputed and Imputed datasets

The impact of quality control filtering and imputation reference panels on PGx variant calling was assessed by cross-referencing the raw, post-QC unimputed, and post-QC imputed datasets with the PharmVar dataset. PGx variants were identified in these datasets by mapping positions with reference PGx variants in the PharmVar database for GRCh37 and GRCh38 builds. This analysis was performed using a custom Python script (version 3.11.3)and BCFtools (**Supp. Figures 2-5**).35 The minor allele frequency, Hardy-Weinberg equilibrium p-value, and SNP missingness frequency was calculated for each PGx variant in each dataset using PLINK 2.0.

Tier 1 (essential panel), Tier 2 (extended panel), and gene-specific coverage was determined by comparing the number of variants present in the dataset for a given star allele to the total number of variants needed for the star allele as defined by PharmVar. This was accomplished using a custom Java script (version 19.0.1) (**Supp. Figure 6**).36 Total variant coverage was calculated by dividing the number of detected variants by the total number of variants in the PharmVar dataset. Copy number variants (CNVs) such as *CYP2D6* \*5 and \*xN were excluded as they were not included on the Infinium Global Screening Array V3.0. Heatmaps were created to visualize the coverage of star alleles in each dataset using ggplot2 in R (version 4.3.3).37 An overview of the methodology and analysis pipeline is presented in **Figure 1**.

All analysis was conducted using the Advanced Research Computing (ARC) cluster at the University of Calgary.

*Sex & Gender Consideration*

Neither sex nor gender is anticipated to influence the impact of quality control filters or imputation reference panels on PGx variant calling. All individuals, regardless of the participant's sex, were uniformly subjected to the study's protocol. Furthermore, for transparency and to support reproducibility in subsequent research, the sex of the participants will be reported in all publications resulting from this study.

# 4. Results

## 4.1 Variant Retention Post QC Filtering

The original CLOZIN dataset included 839 individuals and 725,831 SNPs (**Table 1**). After the removal of individuals who failed sample-level QC, 687 samples remained. 860 SNP positions had > 10% missingness and were removed in both filtering procedures. In the standard filtering process, 198,378 SNPs with a MAF of < 0.01 were removed, along with 3,819 SNPs that deviated from HWE. Finally, 25,117 and 37,825 SNPs were removed with the exclusion of the sex chromosomes in the standard and modified filtering procedures, respectively. In total, 228,174 SNPs and 38,685 SNPs were removed during the standard filtering and modified filtering procedures, respectively. In total, the modified filtering procedure retained 38.2% more variants, preserving 94.8% of the original variants, compared to 68.6% in the standard dataset.

From PharmVar, 775 and 809 variants corresponding to GRCh37 and GRCh38 were compiled (**Table 2**).Of the 775 GRCh37 variants, 69 are Tier 1 and Tier 2. Of the 809 GRCh38 variants, 98 are Tier 1 and Tier 2.Within the raw (unfiltered) dataset, 325/775 PGx variants were identified, with 34 belonging to Tier 1 and Tier 2. The standard dataset retained 90/775 (11.6%) PGx variants and 9/69 (13.0.5%) Tier 1 and Tier 2 variants, while the modified dataset retained 141/775 (18.2%) PGx variants and 19/69 (27.5%) Tier 1 and Tier 2 variants. In total, the modified filtering procedure retained 57.8% more PGx variants and 111.1% more Tier 1 and Tier 2 variants than the standard filtering procedure.

## 4.2 PGx Variant Coverage Post Imputation

After 1000G imputation, 368/775 (47.5%) PGx variants were identified in the standard dataset and 378/775 (48.8%) PGx variants were identified in the modified dataset, reflecting a relative difference of 2.7% (**Table 2**). After TOPMed imputation, 165/809 (20.4%) PGx variants were identified in the standard dataset and 166/809 (20.5%) PGx variants were identified in the modified dataset, reflecting a relative difference of 0.2%.

## 4.3 Tier 1 & Tier 2 Star Allele Coverage

More than half of Tier 1 and Tier 2 variants were identified in the 1000G standard dataset (39/69; 56.5%) and the 1000G modified dataset (41/69, 59.4%) (**Supp. Table 1**). Whereas only one-third (33/98, 33.7%) of Tier 1 and Tier 2 variants were identified in both TOPMed standard and modified datasets. Only 1 of the 14 indels, *CYP3A5*\*7 (rs41303343), was imputed by either 1000G or TOPMed datasets. Most of the missed indels were present in *CYP2D6* (**Table 3**).

For Tier 1, both 1000G datasets had full coverage for 15/23 (65.2%) star alleles and partial coverage for 2/23 (8.7%) star alleles (**Figure 2**). Both TOPMed datasets had full coverage for 12/23 (52.2%) star alleles and no partial coverage. Notably, the TOPMed datasets did not provide coverage for any of the 9 *CYP2D6* alleles, while both 1000G datasets provided either full or partial coverage for 8/9 alleles.

For Tier 2, the 1000G standard dataset fully covered 6/30 (20%) star alleles and partially covered 5/30 (16.7%) star alleles. The 1000G modified provided slightly better coverage, with 8/30 (26.7%) star alleles fully covered and 4/30 (13.3%) partially covered. Both TOPMed datasets fully covered 7/30 (23.3%) star alleles and partially covered 7/30 (23.3%) star alleles. Notably, TOPMed provided full or partial coverage for all 8 *CYP2C19* alleles, while 1000G only covered 2/8. Variants missing from both 1000G and TOPMed for *CYP2C19* star alleles can be found in **Table 4**.

All Tier 1 and Tier 2 star alleles retained in the unimputed modified dataset but lost in unimputed standard were regained through imputation with TOPMed (**Figure 2**). Interestingly, TOPMed excelled in imputing rare SNPs (42.4% identified variants < 0.05 MAF), however, struggled to impute some much more common variants, such as *CYP2D6*\*10 rs1065852 (MAF = 0.21839) and *CYP2D6*\*49 rs1065852 (MAF = 0.21839) (**Supp. Table 1, Table 5**).

## 4.4 Global Pharmacogene Coverage

Global gene coverage was relatively similar across all imputed datasets. The 1000G standard and modified datasets had full coverage for 30.6% and 31.5% of all PharmVar genes, respectively (**Figure 3**). Similarly, the TOPMed standard and modified datasets had full coverage for 32.8% and 33.1% of all PharmVar genes, respectively.

Of the PharmVar genes analyzed, only 2 genes, *CYP2C19* and *CYP2D6*, had notable differences in overall coverage (full + partial) between the 1000G and TOPMed imputed datasets (**Supp. Figure 1**). The TOPMed datasets showed overall coverage for 90.9% of *CYP2C19* star alleles, compared to 33.3% for the 1000G datasets. In contrast, the 1000G datasets showed overall coverage for ~75% of *CYP2D6* star alleles, compared to 6.3% coverage for the TOPMed datasets. When comparing full coverage for *CYP2C19*, 1000G and TOPMed are more similar, with 24.2% and 27.3% full coverage, respectively (**Table 6**). For *CYP2D6*, both 1000G and TOPMed datasets offer very limited full coverage. However, 1000G outperforms TOPMed, with ~15% coverage compared to just 2.5%, respectively.

Across all imputed datasets, *CYP2C9* and *NUDT15* had the worst overall (average 14.3% and 14.5%, respectively) and *CYP4F2* and *CYP3A5* had the best coverage (average 91.7% and 70.0%, respectively) (**Supp. Figure 1**).

## 4.5 Imputation Run Time

Imputation run time was assessed for both the 1000G and TOPMed datasets. The 1000G Phase 3 v5 imputation panel consist of 2504 samples and a total of 81.2M polymorphic markers (**Table 7**).39 The imputation run time for the standard and modified datasets with 1000G was 1:05:48 and 1:04:44, respectively. The TOPMed r3 imputation panel consists of 133,597 reference samples and more than 445.6M variants (**Table 7**).40 The imputation run time for the standard and modified datasets with TOPMed was 1:37:46 and 1:35:16, respectively. A breakdown of the reference demographics for both imputation panels can be found in **Table 7**.

# 5. Discussion

In this study, we evaluated the impact of various quality control (QC) and imputation methods on PGx variant calling by analyzing the coverage of PGx star alleles across seven datasets. This analysis primarily assesses the collective presence of all, some, or none of the PGx variants defined for a given star allele within each dataset, rather than sample-level allele presence. This approach focuses the analysis on the retention and imputation of PGx variants required for star allele calling, regardless of their distribution among samples within the dataset.

The modified filtering procedure demonstrated a significant improvement in retaining PGx variants, particularly for Tier 1 and Tier 2 variants. Compared to the standard filtering procedure, the modified method retained 57.8% more PGx variants and 111.1% more Tier 1 and Tier 2 variants. The removal of MAF and HWE filters significantly improved the retention of rare PGx variants, confirming that the currently recommended filtering methods for GWAS data preparation are too strict for use in PGx applications.

Errors in genotyping are more likely for variants with low MAFs because smaller and less distinct clusters are formed by these alleles during genotype calling.41 This makes it challenging for both manual and automatic clustering methods to accurately identify and separate these clusters, increasing the risk for error.41 However, given the low frequencies of many PGx variants, it is more common for such variants to be incorrectly removed from the dataset, potentially resulting in false negatives if not successfully imputed. In the context of star allele calling, false negatives could result in samples defaulting to the \*1 allele for a given gene, which indicates normal or “wild-type” function. Defaulting to \*1 could obscure true genotype-phenotype associations in a GWAS and therefore may result in inappropriate clinical recommendations. Additionally, GWAS are inherently prone to false negatives due to their low power to detect small effects.42 Given that the application of less strict filtering methods led to the retention of far more PGx variants in this study, reducing or removing MAF or HWE filtering thresholds could help reduce the potential for false negatives in downstream analysis.

However, in combination with imputation, QC filtering may have a less significant impact on PGx calling than initially observed. Intrasample concordance for the number of identified PGx variants between standard and modified datasets was high across both imputation methods, with a relative difference of only 2.7% between the 1000G datasets and 0.2% between the TOPMed datasets. In combination with the observation that imputation with the TOPMed panel effectively compensates for the loss of Tier 1 and Tier 2 variants by standard filtering, the choice of imputation panel appears to mitigate the impact of QC filtering methods on PGx variant coverage.

In terms of overall PGx gene coverage, the 1000G and TOPMed imputed datasets were comparable, with full coverage across all PGx genes ranging from 30-33% and total coverage ranging from 48-50%. However, the Tier 1/Tier 2 coverage analysis revealed some notable differences between the imputation methods, particularly for the *CYP2C19* and *CYP2D6* variants. For *CYP2C19*, the discrepancy in Tier 2 coverage between the 1000G and TOPMed datasets (33.3% vs 90.9% overall coverage, respectively) can be attributed to differences in the dataset’s genome builds (GRCh37 vs. GRCh38) and corresponding variations in star allele definitions between builds. The star alleles *CYP2C19* \*5, \*6, \*7, \*9, and \*10 include an extra variant, rs3758581, for the GRCh38 allele definitions that is not present in the GRCh37 definitions. Therefore, both 1000G and TOPMed are missing the same variants for these star alleles and can be considered “equal” in coverage. As such, there is only one true difference in *CYP2C19* coverage between imputation methods, as TOPMed provides full coverage of *CYP2C19*\*4 rs28399504 (MAF = 0.003268) while 1000 Genomes does not.

Similarly, the discrepancy in coverage for *CYP2D6* between the 1000G and TOPMed datasets (~75% vs 6.3% total coverage, respectively) can be primarily attributed to differences in the GRCh37 and GRCh38 allele definitions. When comparing the GRCh37 star allele definitions to the GRCh38 definitions, there are 2 variants, rs1135840 and rs16947, whose inclusion in various star allele definition frequently differ between builds.43 Star alleles that included these variants in their definitions for GRCh37 did not include these variants in the corresponding star allele definitions for GRCh38, and vice versa.43 The partial allele coverage for *CYP2D6* \*3, \*6, \*9, and \*15 observed for the 1000G datasets is attributed to the presence of these 2 variants, which are not present in the definitions for these star alleles in GRCh38. Therefore, the coverage in the 1000G datasets can be considered “equal” to the coverage for the TOPMed datasets as both 1000G and TOPMed lack the required indels included in the definitions for these alleles.

While disparities in build definitions account for certain discrepancies in *CYP2D6* coverage, the 1000G datasets exhibit superior coverage compared to TOPMed for several additional CYP2D6 star alleles: \*4, \*10, \*17, \*29, \*40, \*41, \*49, and \*59. Surprisingly, some of the absent variants are relatively common, with MAF > 0.1. Considering that TOPMed boasts over 364 million more variants and over 131,000 additional reference samples compared to 1000G, it was unexpected for 1000G to outperform TOPMed. Additionally, 14 out of 33 variants (42.4%) identified in the TOPMed datasets have MAFs < 0.05, indicating TOPMed has the capacity to impute rare PGx variants. Furthermore, TOPMed's panel consists of 40% European ancestry, compared to 20% in the 1000G panel. Given that the majority of individuals in the CLOZIN dataset are of European descent, this suggests that TOPMed should be better suited for imputation within this population.

In a recent study by Lau et al. that assesses the hazards of imputation using the TOPMed panel, they found that imputation is biased towards the reference panel, favouring allelic calls that are more prevalent in the reference panel than in the case-control study.44 Additionally, while TOPMed showed improved imputation quality for rare SNPs (<0.01 MAF) across the genome, the imputation quality for more common variants (MAF > 0.05) is much more variable.44 This variation was observed in the TOPMed datasets, resulting in potential false-negatives for several common variants.

Both 1000G and TOPMed struggled to impute indels, with 13/14 Tier 1/Tier 2 indels absent from all 4 imputed datasets. In an evaluation of imputation performance with 1000G, Liu et al. identified several factors that contribute to a region having “low imputability”, which include low variant heterozygosity, high sequence similarity to other regions, high GC content, and segmental duplication.45 The factors identified by Liu et al. may explain the poor imputation of *CYP* indels by 1000G, which account for 9 of the 13 unimputed indels, as *CYP* genes have been recognized for their high sequence similarity to other pseudogenes, high GC content, and presence of copy number variants.46

In a similar study, Hanks et al. assessed the extent to which imputation with TOPMed could approximate whole-genome sequencing.47 They determined specific MAF thresholds for bi-allelic and multi-allelic variants that could be reliably imputed with high accuracy (r2 > 0.8).47 These thresholds were determined separately for 4 genotyping arrays examined in their study.47 For bi-allelic and multi-allelic indels, the average MAF thresholds across the 4 datasets in European populations was 0.016 and 0.015, respectively.47 Given that 13/14 of the Tier 1/Tier 2 indels have a MAF < 0.014, which falls below the MAF thresholds for reliable imputation with TOPMed, the rarity of these variants likely explains the poor imputation of indels by TOPMed.

*Study Strengths & Limitations*

One of the most significant strengths of this study is its comprehensive comparative approach. This study conducted an in-depth assessment of how various QC filtering and imputation procedures may influence PGx variant calling. Additionally, the study relied on commonly used and well-established bioinformatics tools, ensuring that the results are directly applicable to a wide range of PGx analyses.

A significant limitation of this study is the predominant representation of individuals of European descent in the dataset. This lack of diversity restricts the generalizability of the study’s findings as the QC and imputation methods may perform differently in populations with greater genetic diversity. In the context of imputation, the reference panels used may not be ideal for certain populations, such as Hispanic/Latino or Asian population, which may lead to inaccuracies in imputation due to the limited coverage of rare variants specific to the population within the reference panel. While the study findings suggest that QC filtering methods play a less significant role in PGx variant calling as lost variants were recaptured through imputation, this dynamic may differ in less well-represented populations. Additionally, the GWAS data used in this project was sequenced on the Infinium GSA V3.0 array. Lenz et al. examined the PGx variant coverage available on several genome-wide genotyping arrays and found that some arrays are better suited for PGx usage than others.48 The GSA array was identified as one of the lowest-performing arrays, lacking coverage for several tier 1 and tier 2 alleles, including alleles for CYP2C9, CYP2C19, CYP2D6, and NUDT15.48 The limited coverage of PGx variants on the Infinium GSA array may compromise imputation accuracy, as imputation algorithms may struggle to accurately infer PGx genotypes due to the absence of many key variants.48

*Future Directions*

To extend the scope of this study, future research could include GWAS data from other studies with more diverse demographics, allowing the impact of QC and imputation methods to be assessed in the context of various genetic backgrounds. Moreover, future research would benefit from analyzing a wider range of pharmacogenes, such as *ABCG2, CFTR, DPYD, G6PD, RYR1, TPMT, UGT1A1,* and *VKORC1,* which have prescribing guidelines associated with them. Both of these approaches would collectively enhance generalizability and real-world applicability. Finally, future studies could utilize long-read sequencing technologies to sequence pharmacogenes to establish a "truth" dataset against which QC and imputation results can be benchmarked, allowing imputation accuracy to be assessed along with its performance. This approach would enable the identification of potential false-negatives and false-positives introduced by each QC and imputation method.

# 6. Conclusion

This study aimed to evaluate the impact of quality control filters and imputation reference panels on pharmacogenomic variant calling. While the less strict modified filtering procedure did retain more PGx variants, the method of imputation had a far more influential impact on PGx variant calling. This study suggests that the 1000G imputation panel is the most suitable for processing GWAS data for PGx applications, given the superior coverage of Tier 1 and Tier 2 alleles in 1000G compared to TOPMed, along with TOPMed's limitations in imputing common PGx variants. However, Lau et al. warns that imputation, regardless of the software utilized, tends to favor more frequent alleles in the reference panel when making allele calls, leading to false-negative errors.44 Therefore, regardless of the QC or imputation methods chosen, some genome-wide genotyping arrays, such as Infinium’s GSA, may not be suitable for PGx applications. Due to the extensive genomic data collected by large consortia using the GSA array, researchers should approach secondary PGx analyses with caution.

# Tables

**Table 1.** Number of samples and SNPs at each stage of quality control.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Number of Samples** | **Number of SNPs (Standard)** | | **SNPs Removed (Standard)** | **Number of SNPs (Modified)** | **SNPs Removed (Modified)** |
| Start | 839 | | 725,831 | - | 725,831 | - |
| Removal of individuals who failed QC | 687 | | 725,831 | 0 | 725,831 | 0 |
| Removal of SNPs with missingness > 10% | 687 | | 724,971 | 860 | 724,971 | 860 |
| Removal of SNPs with MAF < 1% | 687 | | 526,593 | 198,378 | - | - |
| Removal of SNPs with HWE > 1x10-4 | 687 | | 522,774 | 3,819 | - | - |
| Removal of sex chromosomes | 687 | | 497,657 | 25,117 | 688,006 | 37,825 |

**Table 2.** Total number of PGx variants present in each dataset. The 'PharmVar' column represents the total number of PGx variants compiled from PharmVar. The values in the subsequent columns indicate the number of variants from the PharmVar set identified within each respective dataset.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **PharmVar** | **Raw** | **Unimputed Standard** | **Unimputed Modified** | **1000G Standard** | **1000G Modified** | **TOPMed Standard** | **TOPMed Modified** |
| All PGx (GRCh37) | 775 | 325 | 90 | 141 | 368 | 378 | - | - |
| Tier 1 & Tier 2 (GRCh37) | 69 | 34 | 9 | 19 | 39 | 41 | - | - |
| All PGx (GRCh38) | 809 | - | - | - | - | - | 165 | 166 |
| Tier 1 & Tier 2 (GRCh38) | 98 | - | - | - | - | - | 33 | 33 |
| % PGx | - | 41.9% | 11.6% | 18.2% | 47.5% | 48.8% | 20.4% | 20.4% |
| % Tier 1 & Tier 2 | - | 42.3% | 13.0% | 27.5% | 56.5% | 59.4% | 33.7% | 33.7% |

**Table 3.** Variant information for Tier 1 & Tier 2 indels missing from 1000G and TOPMed datasets. Function refers to the function of the star allele as defined by PharmVar. MAF values were sourced from the ALFA project.27,38

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Star Allele** | **Gene** | **SNP** | **Chr** | **Position (GRCh38)** | **Position (GRCh38)** | **Tier** | **Function** | **MAF EUR** | **MAF All** | **Nature of Variant** | **Type** |
| \*6 | CYP2C9 | rs9332131 | 10 | 96709040 | 94949283 | Tier 1 | no function | 0 | 0.00038 | indel | SV - multiallelic |
| \*3 | CYP2D6 | rs35742686 | 22 | 42524244 | 42128242 | Tier 1 | no function | 0.01361 | 0.01142 | indel | SV |
| \*6 | CYP2D6 | rs5030655 | 22 | 42525086 | 42129084 | Tier 1 | no function | 0.00167 | 0.00177 | indel | SV |
| \*9 | CYP2D6 | rs5030656 | 22 | 42524178 | 42128176 | Tier 1 | decreased function | 0.02212 | 0.01875 | indel | SV |
| \*15 | CYP2D6 | rs774671100 | 22 | 42526657 | 42130655 | Tier 2 | no function | 0.00019 | 0.00022 | indel | SV |
| \*21 | CYP2D6 | rs72549352 | 22 | 42524220 | 42128218 | Tier 2 | no function | 0.00006 | 0.00009 | indel | SV - multiallelic |
| \*40 | CYP2D6 | rs72549356 | 22 | 42524944 | 42128942 | Tier 2 | no function | 0 | 0.00005 | indel | SV - multiallelic |
| \*42 | CYP2D6 | rs72549346 | 22 | 42523535 | 42127533 | Tier 2 | no function | 0 | 0.00019 | indel | SV |
| \*20 | CYP3A4 | rs67666821 | 7 | 99355811 | 99758188 | Tier 2 | not assigned | 0.00007 | 0.00006 | indel | SV - multiallelic |
| \*2 | NUDT15 | rs746071566 | 13 | 48611937 | 48037801 | Tier 2 | no function | 0.0008 | 0.00062 | indel | SV - multiallelic |
| \*6 | NUDT15 | rs746071566 | 13 | 48611937 | 48037801 | Tier 2 | uncertain function | 0.0008 | 0.00062 | indel | SV - multiallelic |
| \*9 | NUDT15 | rs746071566 | 13 | 48611932 | 48037796 | Tier 2 | no function | 0.0008 | 0.00062 | indel | SV - multiallelic |
| \*14 | NUDT15 | rs777311140 | 13 | 48611962 | 48037826 | Tier 2 | not assigned | 0.00049 | 0.00038 | indel | SV |

**Table 4.** Variant information for *CYP2C19* Tier 2 variants missing from 1000G and TOPMed datasets. Function refers to the function of the star allele as defined by PharmVar. MAF values were sourced from the ALFA project.27,38

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Star Allele** | **Gene** | **SNP** | **Chr** | **Position (GRCh38)** | **Position (GRCh38)** | **Tier** | **Function** | **MAF EUR** | **MAF All** | **Nature of Variant** | **Type** |
| \*5 | CYP2C19 | rs56337013 | 10 | 96612495 | 94852738 | Tier 2 | no function | 0.000045 | 0.000041 | substitution | biallelic |
| \*6 | CYP2C19 | rs72552267 | 10 | 96535210 | 94775453 | Tier 2 | no function | 0.0004 | 0.00038 | substitution | biallelic |
| \*7 | CYP2C19 | rs72558186 | 10 | 96541756 | 94781999 | Tier 2 | no function | 0.00001 | 0.00001 | substitution | biallelic |
| \*9 | CYP2C19 | rs17884712 | 10 | 96535246 | 94775489 | Tier 2 | decreased function | 0.00002 | 0.00194 | substitution | biallelic |
| \*10 | CYP2C19 | rs6413438 | 10 | 96541615 | 94781858 | Tier 2 | decreased function | 0.000244 | 0.000412 | substitution | biallelic |

**Table 5.** Variant information for Tier 1 & Tier 2 variants missing in TOPMed datasets only. Function refers to the function of the star allele as defined by PharmVar. MAF values were sourced from the ALFA project.27,38

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Star Allele** | **Gene** | **SNP** | **Chr** | **Position (GRCh38)** | **Position (GRCh38)** | **Tier** | **Function** | **MAF EUR** | **MAF All** | **Nature of Variant** | **Type** |
| \*4 | CYP2D6 | rs3892097 | 22 | 42524947 | 42128945 | Tier 1 | no function | 0.19181 | 0.17694 | substitution | biallelic |
| \*10 | CYP2D6 | rs1065852 | 22 | 42526694 | 42130692 | Tier 1 | decreased function | 0.21839 | 0.21154 | substitution | biallelic |
| \*17 | CYP2D6 | rs28371706 | 22 | 42525772 | 42129770 | Tier 1 | decreased function | 0.00226 | 0.01308 | substitution | multiallelic |
| \*29 | CYP2D6 | rs59421388 | 22 | 42523610 | 42127608 | Tier 1 | decreased function | 0.000493 | 0.004656 | substitution | biallelic |
| \*29 | CYP2D6 | rs61736512 | 22 | 42525134 | 42129132 | Tier 1 | decreased function | 0.000493 | 0.003274 | substitution | biallelic |
| \*41 | CYP2D6 | rs28371725 | 22 | 42523805 | 42127803 | Tier 1 | decreased function | 0.10421 | 0.09107 | substitution | biallelic |
| \*40 | CYP2D6 | rs28371706 | 22 | 42525772 | 42129770 | Tier 2 | no function | 0.00226 | 0.01308 | substitution | multiallelic |
| \*49 | CYP2D6 | rs1065852 | 22 | 42526694 | 42130692 | Tier 2 | decreased function | 0.21839 | 0.21154 | substitution | biallelic |
| \*59 | CYP2D6 | rs79292917 | 22 | 42523854 | 42127852 | Tier 2 | decreased function | 0.00459 | 0.00428 | substitution | biallelic |

**Table 6.** Percent full coverage per gene for imputed datasets.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Gene** | **1000G Standard** | **1000G Modified** | **1000G Average** | **TOPMed Standard** | **TOPMed Modified** | **TOPMed Average** | **Top Performer** |
| CYP2B6 | 40.0% | 40.0% | 40.0% | 22.2% | 24.4% | 23.3% | 1000G |
| CYP2C19 | 24.2% | 24.2% | 24.2% | 27.3% | 27.3% | 27.3% | TOPMed |
| CYP2C9 | 4.8% | 7.1% | 6.0% | 10.7% | 10.7% | 10.7% | TOPMed |
| CYP2D6 | 13.9% | 17.1% | 15.5% | 2.5% | 2.5% | 2.5% | 1000G |
| CYP3A4 | 14.7% | 14.7% | 14.7% | 20.6% | 20.6% | 20.6% | TOPMed |
| CYP3A5 | 40.0% | 40.0% | 40.0% | 60.0% | 60.0% | 60.0% | TOPMed |
| CYP4F2 | 83.3% | 83.3% | 83.3% | 100.0% | 100.0% | 100.0% | TOPMed |
| NUDT15 | 5.3% | 10.5% | 7.9% | 10.5% | 10.5% | 10.5% | TOPMed |
| SLCO1B1 | 48.8% | 46.3% | 47.6% | 41.5% | 41.5% | 41.5% | 1000G |

**Table 7.** Comparison of imputation panels.14,33,39,40

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **1000G Phase 3 v5** | | **TOPMed r3** |
| Number of Reference Samples | 2,504 | 133,597 | |
| Number of Polymorphic Markers | 81.2M | 445.6M | |
| % European Ancestry | 20% | 40% | |
| % African Ancestry | 26% | 29% | |
| % Hispanic/Latino Ancestry | 14% | 19% | |
| % Asian Ancestry | – | 8% | |
| % East Asian Ancestry | 20% | – | |
| % South Asian Ancestry | 20% | – | |
| % Other/Multiple/Unknown Ancestry | – | 4% | |
| Time to Impute Standard QC'd Data (hr:min:sec) | 1:05:48 | 1:37:46 | |
| Time to Impute Modified QC'd Data (hr:min:sec) | 1:04:44 | 1:35:16 | |

# Figures

A diagram of a flowchart

Description automatically generated

**Figure 1:** Quality control, imputation, and analysis flowchart. Orange and pink arrows represent standard and modified QC’ed datasets, respectively.

**A screenshot of a graph

Description automatically generatedFigure 2.** Comparative coverage of Tier 1 & Tier 2 alleles identified from PharmVar variants. **(A)** Displays the coverage for the minimum set of Tier 1 alleles recommended for PGx testing in each dataset. **(B)** Displays the variant coverage for the optional set of Tier 2 alleles in each dataset. Raw, unimputed standard, unimputed modified, 1000G standard, and 1000G modified were assessed for GRCh37 variants from PharmVar (69 total variants). TOPMed standard and TOPMed modified were assessed for GRCh38 variants from PharmVar (98 total variants). Dark blue, light blue, and white boxes indicate full coverage, partial coverage, and absence of star allele variants, respectively.

A group of blue and gray bars

Description automatically generated

**Figure 3.** Percent coverage of PharmVar genes by dataset. Dark blue, light blue, and grey bars represent full coverage, partial coverage, and absence of star alleles, respectively.

# References

1. Adams, J. Pharmacogenomics and Personalized Medicine. Nat. Edu. 2008. 1(1):194
2. Kane M. CYP2D6 Overview: Allele and Phenotype Frequencies. In: Medical Genetics Summaries [Internet]. National Center for Biotechnology Information (US); 2021. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK574601/>
3. Zhao S, Jiang L, Yu H, Guo Y. GTQC: Automated Genotyping Array Quality Control and Report. J Genomics. 2022 Feb 14;10:39–44.
4. Bejjani BA, Shaffer LG. Application of Array-Based Comparative Genomic Hybridization to Clinical Diagnostics. J Mol Diagn. 2006 Nov;8(5):528–33.
5. PharmGKB. AMP’s Minimum Sets of Alleles for PGx Testing. [Internet]. PharmGKB. [cited 2023 Sep 30]. Available from: <https://www.pharmgkb.org/ampAllelesToTest>
6. Pratt VM, Cavallari LH, Del Tredici AL, et al. Recommendations for Clinical CYP2C9 Genotyping Allele Selection. J Mol Diagn. 2019;21(5):746-55.
7. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. Genetics in Medicine. 2017 Jan 1;19(1):20–9.
8. Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science. 2012 Jul 6;337(6090):100–4.
9. Lakiotaki K, Kanterakis A, Kartsaki E, Katsila T, Patrinos GP, Potamias G. Exploring public genomics data for population pharmacogenomics. PLOS ONE. 2017 Aug 3;12(8):e0182138.
10. Lamy P, Andersen CL, Wikman FP, Wiuf C. Genotyping and annotation of Affymetrix SNP arrays. Nucleic Acids Res. 2006;34(14):e100.
11. Pavan S, Delvento C, Ricciardi L, Lotti C, Ciani E, D’Agostino N. Recommendations for Choosing the Genotyping Method and Best Practices for Quality Control in Crop Genome-Wide Association Studies. Front Genet. 2020 Jun 5;11:447.
12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
13. Li Y, Willer C, Sanna S, Abecasis G. Genotype Imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
14. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.
15. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021 Feb;590(7845):290–9.
16. Lazarou J, Pomeranz BH, Corey PN. Incidence of Adverse Drug Reactions in Hospitalized Patients A Meta-analysis of Prospective Studies. JAMA. 1998 Apr 15;279(15):1200–5.
17. Bates DW. Drugs and Adverse Drug Reactions: How Worried Should We Be? JAMA. 1998 Apr 15;279(15):1216–7.
18. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality Control Procedures for Genome-Wide Association Studies. Current Protocols in Human Genetics. 2011;68(1):1.19.1-1.19.18.
19. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie‐Claire C, et al. A tutorial on conducting genome‐wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Feb 27;27(2):e1608.
20. Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. Hum Mol Genet. 2013 Oct 15;22(R1):R16–21.
21. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: Testing for an Accumulation of Rare Variants Using Next-Generation Sequencing Data. Hum Hered. 2012;73(2):84–94.
22. Charon C, Allodji R, Meyer V, Deleuze JF. Impact of pre- and post-variant filtration strategies on imputation. Sci Rep. 2021 Mar 18;11:6214.
23. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun;9(6):477–85.
24. Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, et al. Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet. 2005 May;13(5):677–86.
25. Li JH, Liu A, Buerkle CA, Palmer W, Belbin GM, Ahangari M, et al. The effects of reference panel perturbations on the accuracy of genotype imputation. bioRxiv [Internet]. 2023 [cited 2023 Oct 12]. Available from: <https://www.biorxiv.org/content/10.1101/2023.08.10.552684v1>
26. Okhuijsen-Pfeifer C, van der Horst MZ, Bousman CA, Lin B, van Eijk KR, Ripke S, et al. Genome-wide association analyses of symptom severity among clozapine-treated patients with schizophrenia spectrum disorders. Transl Psychiatry. 2022 Apr 7;12:145.
27. Gaedigk A, Whirl-Carrillo M, Pratt VM, Miller NA, Klein TE. PharmVar and the Landscape of Pharmacogenetic Resources. Clin Pharmacol Ther. 2020 Jan;107(1):43–6.
28. PharmVar [Internet]. Pharmacogene Variation Consortium; [cited 2024 Mar 12]. Available from: <https://www.pharmvar.org/>
29. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023 Nucleic Acids Res. 2023 Jan 6;51( D1):D933–41.
30. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D766–73.
31. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008.
32. Michigan Imputation Server [Internet]. University of Michigan; [cited 2024 Mar 22]. Available from: [https://imputationserver.sph.umich.edu/index.html#](https://imputationserver.sph.umich.edu/index.html)!
33. TOPMed Imputation Server [Internet]. National Heart, Lung, and Blood Institute (U.S.); [cited 2024 Mar 22]. Available from: [https://imputation.biodatacatalyst.nhlbi.nih.gov/#](https://imputation.biodatacatalyst.nhlbi.nih.gov/)!
34. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.
35. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
36. Arnold K, Gosling J, Holmes D. The Java programming language. Addison Wesley Professional; 2005.
37. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2016. Available from: <https://www.R-project.org/>
38. L. Phan, Y. Jin, H. Zhang, W. Qiang, E. Shekhtman, D. Shao, et al. ALFA: Allele Frequency Aggregator [Internet]. National Center for Biotechnology Information, National Library of Medicine (U.S); 2020 Mar 10 [cited 2024 Mar 22]. Available from: <https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/>
39. 1000G Phase3 v5 Reference [Internet]. University of Michigan; [cited 2024 Mar 23]. Available from: <https://csg.sph.umich.edu/abecasis/mach/download/1000G.Phase3.v5.html>
40. About TOPMed [Internet]. National Heart, Lung, and Blood Institute (US); [cited 2024 Mar 23]. Available from: <https://topmed.nhlbi.nih.gov/>
41. Coleman JRI, Euesden J, Patel H, Folarin AA, Newhouse S, Breen G. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. Brief Funct Genomics. 2016 Jul;15(4):298–304.
42. Marigorta UM, Rodríguez JA, Gibson G, Navarro A. Replicability and Prediction: lessons and challenges from GWAS. Trends Genet. 2018 Jul;34(7):504–17.
43. Gene Info CYP2D6 [Internet]. Pharmacogene Variation Consortium; [cited 2024 Mar 19]. Available from: <https://www.pharmvar.org/gene/CYP2D6>
44. Lau W, Ali A, Maude H, Andrew T, Swallow DM, Maniatis N. The hazards of genotype imputation when mapping disease susceptibility variants. Genome Biology. 2024 Jan 3;25(1):7.
45. Liu Q, Cirulli ET, Han Y, Yao S, Liu S, Zhu Q. Systematic assessment of imputation performance using the 1000 Genomes reference panels. Brief Bioinform. 2015 Jul;16(4):549–62.
46. Drögemöller BI, Wright GEB, Niehaus DJH, Emsley R, Warnich L. Next-generation sequencing of pharmacogenes: a critical analysis focusing on schizophrenia treatment. Pharmacogenetics and Genomics. 2013 Dec;23(12):666.
47. Hanks SC, Forer L, Schönherr S, LeFaive J, Martins T, Welch R, et al. Extent to which array genotyping and imputation with large reference panels approximate deep whole-genome sequencing. Am J Hum Genet. 2022 Sep 1;109(9):1653–66.
48. Lenz C, Narang A, Bousman CA. Pharmacogenomic allele coverage of genome-wide genotyping arrays: a comparative analysis. Pharmacogenet Genomics. 2024 Feb 8;

# Supplementary Materials

**Supplementary Table 1.** Variant information and presence for Tier 1 & Tier 2 variants across 7 datasets. The last 7 columns represent datasets, with '1' denoting presence and '0' denoting absence of star allele variants for each respective dataset. Variant positions are provided for both GRCh37 and GRCh38 builds; '-' indicates that the variant is not included in the star allele definition for the corresponding build. Function refers to the function of the star allele as defined by PharmVar. MAF values were sourced from the ALFA project. Only Tier 1 & Tier 2 variants found in at least one of the datasets are included in the table.27,48

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Star Allele** | **Gene** | **SNP** | **Chr** | **GRCh37 Position** | **GRCh38 Position** | **Tier** | **Function** | **MAF EUR** | **MAF All** |
| \*2 | CYP2C19 | rs12769205 | 10 | 96535124 | 94775367 | Tier 1 | no function | 0.14225 | 0.14643 |
| \*2 | CYP2C19 | rs4244285 | 10 | 96541616 | 94781859 | Tier 1 | no function | 0.14702 | 0.14966 |
| \*2 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 1 | no function | 0.06598 | 0.05936 |
| \*3 | CYP2C19 | rs4986893 | 10 | 96540410 | 94780653 | Tier 1 | no function | 0.00561 | 0.00722 |
| \*3 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 1 | no function | 0.06598 | 0.05936 |
| \*17 | CYP2C19 | rs12248560 | 10 | 96521657 | 94761900 | Tier 1 | increased function | 0.23135 | 0.22229 |
| \*17 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 1 | no function | 0.06598 | 0.05936 |
| \*2 | CYP2C9 | rs1799853 | 10 | 96702047 | 94942290 | Tier 1 | decreased function | 0.12126 | 0.10802 |
| \*3 | CYP2C9 | rs1057910 | 10 | 96741053 | 94981296 | Tier 1 | no function | 0.06666 | 0.06389 |
| \*5 | CYP2C9 | rs28371686 | 10 | 96741058 | 94981301 | Tier 1 | decreased function | 0.00003 | 0.00024 |
| \*8 | CYP2C9 | rs7900194 | 10 | 96702066 | 94942309 | Tier 1 | decreased function | 0 | 0.00001 |
| \*11 | CYP2C9 | rs28371685 | 10 | 96740981 | 94981224 | Tier 1 | decreased function | 0.00223 | 0.00289 |
| \*3 | CYP2D6 | rs35742686 | 22 | 42524244 | 42128242 | Tier 1 | no function | 0.01361 | 0.01142 |
| \*3 | CYP2D6 | rs1135840 | 22 | 42522613 | - | Tier 1 | no function | 0.43263 | 0.42619 |
| \*3 | CYP2D6 | rs16947 | 22 | 42523943 | - | Tier 1 | no function | 0.31818 | 0.32159 |
| \*4 | CYP2D6 | rs3892097 | 22 | 42524947 | 42128945 | Tier 1 | no function | 0.19181 | 0.17694 |
| \*6 | CYP2D6 | rs5030655 | 22 | 42525086 | 42129084 | Tier 1 | no function | 0.00167 | 0.00177 |
| \*6 | CYP2D6 | rs16947 | 22 | 42523943 | - | Tier 1 | no function | 0.31818 | 0.32159 |
| \*9 | CYP2D6 | rs1135840 | 22 | 42522613 | - | Tier 1 | decreased function | 0.43263 | 0.42619 |
| \*9 | CYP2D6 | rs16947 | 22 | 42523943 | - | Tier 1 | decreased function | 0.31818 | 0.32159 |
| \*10 | CYP2D6 | rs1065852 | 22 | 42526694 | 42130692 | Tier 1 | decreased function | 0.21839 | 0.21154 |
| \*10 | CYP2D6 | rs16947 | 22 | 42523943 | - | Tier 1 | decreased function | 0.31818 | 0.32159 |
| \*17 | CYP2D6 | rs28371706 | 22 | 42525772 | 42129770 | Tier 1 | decreased function | 0.00226 | 0.01308 |
| \*29 | CYP2D6 | rs59421388 | 22 | 42523610 | 42127608 | Tier 1 | decreased function | 0.00049 | 0.00466 |
| \*29 | CYP2D6 | rs61736512 | 22 | 42525134 | 42129132 | Tier 1 | decreased function | 0.00049 | 0.00327 |
| \*41 | CYP2D6 | rs28371725 | 22 | 42523805 | 42127803 | Tier 1 | decreased function | 0.10421 | 0.09107 |
| \*22 | CYP3A4 | rs35599367 | 7 | 99366316 | 99768693 | Tier 1 | not assigned | 0.04628 | 0.04201 |
| \*3 | CYP3A5 | rs776746 | 7 | - | 99672916 | Tier 1 | no function | 0.07004 | 0.11263 |
| \*6 | CYP3A5 | rs10264272 | 7 | 99262835 | 99665212 | Tier 1 | no function | 0.0011 | 0.00748 |
| \*6 | CYP3A5 | rs776746 | 7 | 99270539 | - | Tier 1 | no function | 0.07004 | 0.11263 |
| \*7 | CYP3A5 | rs41303343 | 7 | 99250393 | 99652770 | Tier 1 | no function | 0.00024 | 0.00474 |
| \*7 | CYP3A5 | rs776746 | 7 | 99270539 | - | Tier 1 | no function | 0.07004 | 0.11263 |
| \*3 | NUDT15 | rs116855232 | 13 | 48619855 | 48045719 | Tier 1 | no function | 0.00287 | 0.00794 |
| \*4 | CYP2C19 | rs28399504 | 10 | 96522463 | 94762706 | Tier 2 | no function | 0.00327 | 0.00312 |
| \*4 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*5 | CYP2C19 | rs56337013 | 10 | 96612495 | 94852738 | Tier 2 | no function | 0.00005 | 0.00004 |
| \*5 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*6 | CYP2C19 | rs72552267 | 10 | 96535210 | 94775453 | Tier 2 | no function | 0.0004 | 0.00038 |
| \*6 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*7 | CYP2C19 | rs72558186 | 10 | 96541756 | 94781999 | Tier 2 | no function | 0.00001 | 0.00001 |
| \*7 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*8 | CYP2C19 | rs41291556 | 10 | 96535173 | 94775416 | Tier 2 | no function | 0.00267 | 0.00249 |
| \*8 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*9 | CYP2C19 | rs17884712 | 10 | 96535246 | 94775489 | Tier 2 | decreased function | 0.00002 | 0.00194 |
| \*9 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*10 | CYP2C19 | rs6413438 | 10 | 96541615 | 94781858 | Tier 2 | decreased function | 0.00024 | 0.00041 |
| \*10 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*35 | CYP2C19 | rs12769205 | 10 | 96535124 | 94775367 | Tier 2 | no function | 0.14225 | 0.14643 |
| \*35 | CYP2C19 | rs3758581 | 10 | - | 94842866 | Tier 2 | no function | 0.06598 | 0.05936 |
| \*12 | CYP2C9 | rs9332239 | 10 | 96748777 | 94989020 | Tier 2 | decreased function | 0.0026 | 0.00237 |
| \*13 | CYP2C9 | rs72558187 | 10 | 96701715 | 94941958 | Tier 2 | no function | 0 | 0.00002 |
| \*7 | CYP2D6 | rs5030867 | 22 | 42523858 | 42127856 | Tier 2 | no function | 0.00057 | 0.00059 |
| \*7 | CYP2D6 | rs1135840 | 22 | 42522613 | - | Tier 2 | no function | 0.43263 | 0.42619 |
| \*7 | CYP2D6 | rs16947 | 22 | 42523943 | - | Tier 2 | no function | 0.31818 | 0.32159 |
| \*8 | CYP2D6 | rs5030865 | 22 | 42525035 | 42129033 | Tier 2 | no function | 0 | 0.00004 |
| \*12 | CYP2D6 | rs5030862 | 22 | 42526670 | 42130668 | Tier 2 | no function | 0.00001 | 0.00009 |
| \*14 | CYP2D6 | rs5030865 | 22 | 42525035 | 42129033 | Tier 2 | decreased function | 0 | 0.00004 |
| \*15 | CYP2D6 | rs1135840 | 22 | 42522613 | - | Tier 2 | no function | 0.43263 | 0.42619 |
| \*15 | CYP2D6 | rs16947 | 22 | 42523943 | - | Tier 2 | no function | 0.31818 | 0.32159 |
| \*31 | CYP2D6 | rs267608319 | 22 | 42522751 | 42126749 | Tier 2 | no function | 0.00038 | 0.00034 |
| \*40 | CYP2D6 | rs28371706 | 22 | 42525772 | 42129770 | Tier 2 | no function | 0.00226 | 0.01308 |
| \*49 | CYP2D6 | rs1065852 | 22 | 42526694 | 42130692 | Tier 2 | decreased function | 0.21839 | 0.21154 |
| \*49 | CYP2D6 | rs16947 | 22 | 42523943 | - | Tier 2 | decreased function | 0.31818 | 0.32159 |
| \*56 | CYP2D6 | rs72549347 | 22 | 42523592 | 42127590 | Tier 2 | no function | 0.00006 | 0.00046 |
| \*59 | CYP2D6 | rs79292917 | 22 | 42523854 | 42127852 | Tier 2 | decreased function | 0.00459 | 0.00428 |
| \*3 | CYP4F2 | rs2108622 | 19 | 15990431 | 15879621 | Tier 2 | not assigned | 0.29399 | 0.2834 |
| \*2 | NUDT15 | rs116855232 | 13 | 48619855 | 48045719 | Tier 2 | no function | 0.00287 | 0.00794 |
| \*4 | NUDT15 | rs147390019 | 13 | 48619856 | 48045720 | Tier 2 | uncertain function | 0.00006 | 0.00022 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Star Allele** | **Gene** | **Nature** | **Type** | **Raw** | **Unimputed Modified** | **Unimputed Standard** | **1000G Modified** | **1000G Standard** | **Topmed Modified** | **Topmed Standard** |
| \*2 | CYP2C19 | substitution | biallelic | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| \*2 | CYP2C19 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| \*2 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*3 | CYP2C19 | substitution | biallelic | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| \*3 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*17 | CYP2C19 | substitution | biallelic | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| \*17 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*2 | CYP2C9 | substitution | biallelic | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| \*3 | CYP2C9 | substitution | multiallelic | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| \*5 | CYP2C9 | substitution | multiallelic | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| \*8 | CYP2C9 | substitution | multiallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*11 | CYP2C9 | substitution | biallelic | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| \*3 | CYP2D6 | indel | SV | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*3 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*3 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*4 | CYP2D6 | substitution | biallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*6 | CYP2D6 | indel | SV | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*6 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*9 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*9 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*10 | CYP2D6 | substitution | biallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*10 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*17 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*29 | CYP2D6 | substitution | biallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*29 | CYP2D6 | substitution | biallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*41 | CYP2D6 | substitution | biallelic | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| \*22 | CYP3A4 | substitution | biallelic | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| \*3 | CYP3A5 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*6 | CYP3A5 | substitution | biallelic | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| \*6 | CYP3A5 | substitution | biallelic | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| \*7 | CYP3A5 | indel | SV | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| \*7 | CYP3A5 | substitution | biallelic | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| \*3 | NUDT15 | substitution | biallelic | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| \*4 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*4 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*5 | CYP2C19 | substitution | biallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*5 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*6 | CYP2C19 | substitution | biallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*6 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*7 | CYP2C19 | substitution | biallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*7 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*8 | CYP2C19 | substitution | biallelic | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| \*8 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*9 | CYP2C19 | substitution | biallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*9 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*10 | CYP2C19 | substitution | biallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*10 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*35 | CYP2C19 | substitution | biallelic | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| \*35 | CYP2C19 | substitution | biallelic | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| \*12 | CYP2C9 | substitution | biallelic | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| \*13 | CYP2C9 | substitution | biallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*7 | CYP2D6 | substitution | biallelic | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| \*7 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*7 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*8 | CYP2D6 | substitution | multiallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*12 | CYP2D6 | substitution | biallelic | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| \*14 | CYP2D6 | substitution | multiallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*15 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*15 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*31 | CYP2D6 | substitution | biallelic | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| \*40 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*49 | CYP2D6 | substitution | biallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*49 | CYP2D6 | substitution | multiallelic | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*56 | CYP2D6 | substitution | biallelic | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| \*59 | CYP2D6 | substitution | biallelic | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| \*3 | CYP4F2 | substitution | multiallelic | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| \*2 | NUDT15 | substitution | biallelic | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| \*4 | NUDT15 | substitution | biallelic | 1 | 1 | 0 | 1 | 0 | 1 | 1 |

A blue and black grids

Description automatically generated with medium confidence**Supplementary Figure 1.** Gene-specific coverage of PharmVar genes. Heatmaps **A**, **B**, **C**, **D**, **E**, **F**, and **G** display the haplotype coverage for CYP2B6, CYP2C9, CYP2C19, CYP3A5, CYP4F2, CYP2D6, CYP3A4, SLCO1B1, and NUDT15, respectively. Raw, unimputed standard, unimputed modified, 1000G standard, and 1000G modified were assessed for GRCh37 variants and TOPMed standard and TOPMed modified were assessed for GRCh38 variants. Dark blue, light blue, and white boxes indicate full presence, partial presence, and absence of star allele variants, respectively.

A screenshot of a computer program

Description automatically generated

**Supplementary Figure 2.** Code snippet for ‘find\_stars.sh’ script. This program executes ‘extract.sh’, ‘find\_stars.py’, and ‘condense.py’ to identify all star allele variants within a dataset.

A screenshot of a computer program

Description automatically generated

**Supplementary Figure 3.** Code snippet for ‘extract.sh’ script. This program extracts all variants from a given vcf file that exist in the star allele definition list for the corresponding chromosome.

A screenshot of a computer program

Description automatically generated

**Supplementary Figure 4.** Code snippet of ‘get\_stars.py’ script. This program appends the star allele names corresponding to each extracted variant identified by ‘extract.sh’.

A screen shot of a computer program

Description automatically generated

**Supplementary Figure 5.** Code snippet from ‘condense.py’ script. This program condenses output from ‘get\_stars.sh’ by grouping star alleles by variant position into a single row.

A screenshot of a computer code

Description automatically generated

**Supplementary Figure 6.** Code snippet for ‘haplotype\_coverage.java’ script. This program reads the identified star alleles for each data set and the star allele definition tables for GRCh37 and GRCh38 into hashmaps (not pictured). For each haplotype, the variants are extracted from the dataset (dataVariants) and the corresponding allele definition table (sourceTable) into lists. The lists are compared to determine coverage for the haplotype.