The Journal of Molecular Diagnostics, Vol. 24, No. 4, April 2022

ELSEVIER

the Journal of Molecular Diagnostics

jmdjournal.org

# *CYP2C8, CYP2C9,* and *CYP2C19* Characterization Using Next-Generation Sequencing and Haplotype Analysis

## *A GeT-RM Collaborative Project*

Andrea Gaedigk,[*†] Erin C. Boone,[*] Steven E. Scherer,[‡] Seung-been Lee,[§] Ibrahim Numanagić,[¶] Cenk Sahinalp,[‖] Joshua D. Smith,[**] Sean McGee,[**] Aparna Radhakrishnan,[**] Xiang Qin,[‡] Wendy Y. Wang,[*] Emily G. Farrow,[†,††] Nina Gonzaludo,[‡‡] Aaron L. Halpern,[‡‡] Deborah A. Nickerson,[**] Neil A. Miller,[†,††] Victoria M. Pratt,[§§] and Lisa V. Kalman[¶¶]

*From the Division of Clinical Pharmacology, Toxicology and Therapeutic Innovation ,* and the Center for Genomic Medicine,[††] Children's Mercy Kansas City, Kansas City, Missouri; the University of Missouri—Kansas City School of Medicine,[†] Kansas City, Missouri; the Human Genome Sequencing Center,[‡] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas; the Precision Medicine Institute,[§] Macrogen Inc., Seongnam, Republic of Korea; the Department of Computer Science,[¶] University of Victoria, Victoria, British Columbia, Canada; the Cancer Data Science Laboratory,[‖] National Cancer Institute, National Institutes of Health, Bethesda, Maryland; the Department of Genome Sciences,[**] University of Washington, Seattle, Washington; Medical Genomics Research,[‡‡] Illumina Inc., San Diego, California; the Department of Medical and Molecular Genetics,[§§] Indiana University School of Medicine, Indianapolis, Indiana; and the Informatics and Data Science Branch,[¶¶] Division of Laboratory Systems, Centers for Disease Control and Prevention, Atlanta, Georgia*

Accepted for publication December 28, 2021.

Address correspondence to Lisa V. Kalman, Ph.D., Informatics and Data Science Branch, Division of Laboratory Systems, Office of Surveillance, Epidemiology, and Laboratory Services, Centers for Disease Control and Prevention, 1600 Clifton Rd., Mailstop V24-3, Atlanta, GA 30333. E-mail: LKalman@cdc.gov.

Pharmacogenetic tests typically target selected sequence variants to identify haplotypes that are often defined by star (*) allele nomenclature. Due to their design, these targeted genotyping assays are unable to detect novel variants that may change the function of the gene product and thereby affect phenotype prediction and patient care. In the current study, 137 DNA samples that were previously characterized by the Genetic Testing Reference Material (GeT-RM) program using a variety of targeted genotyping methods were recharacterized using targeted and whole genome sequencing analysis. Sequence data were analyzed using three genotype calling tools to identify star allele diplotypes for *CYP2C8, CYP2C9,* and *CYP2C19*. The genotype calls from next-generation sequencing (NGS) correlated well to those previously reported, except when novel alleles were present in a sample. Six novel alleles and 38 novel suballeles were identified in the three genes due to identification of variants not covered by targeted genotyping assays. In addition, several ambiguous genotype calls from a previous study were resolved using the NGS and/or long-read NGS data. Diplotype calls were mostly consistent between the calling algorithms, although several discrepancies were noted. This

Supported by the Intramural Research Program of the National Cancer Institute, NIH (C.S.) and National Human Genome Research Institute grant U01 HG010245 (V.M.P.).

This article is dedicated to the memory of Dr. Deborah A. Nickerson, who passed away December 24, 2021. Debbie was an admired colleague, researcher, and mentor dedicated to bringing the latest technologies to bear on understanding human variation and its impact on human health. She was always a voice of reason as well as a tireless promoter of her trainees and women in science — we deeply mourn her loss.

Disclosures: V.M.P.'s institution, Indiana University, is a fee-for-service clinical PGx laboratory, and V.M.P. does consulting for LabCorp; A.L.H. is employed by Illumina Inc. and owns stock; E.F.G. has previously served on a clinical expert panel for whole genome sequencing for Illumina, Inc.; N.G. is an employee of Pacific Biosciences, and a shareholder of Pacific Biosciences and Illumina, Inc.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention/the Agency for Toxic Substances and Disease Registry. Use of tradenames and commercial sources is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention, the Public Health Service, or the US Department of Health and Human Services.

Current address of N.G., Pacific Biosciences, Menlo Park, CA; of N.A.M., Bionano Genomics, San Diego, CA.


https://doi.org/10.1016/j.jmoldx.2021.12.011

study highlights the utility of NGS for pharmacogenetic testing and demonstrates that there are many novel alleles that are yet to be discovered, even in highly characterized genes such as *CYP2C9* and *CYP2C19*. (J Mol Diagn 2022, 24: 337—350; https://doi.org/10.1016/j.jmoldx.2021.12.011)

Patients often respond differently to drugs. Some individuals benefit, whereas others fail to respond or experience an adverse reaction to a given dose of the same drug. These responses may be predicted or explained using pharmacogenetic (PGx) tests that identify variant alleles of genes known to affect drug absorption, distribution, metabolism, and excretion (ADME) or the target of drug action. These genes are often referred to as ADME genes or pharmacogenes.

According to an extensive review,[1] *CYP2C8*, *CYP2C9,* and *CYP2C19* are collectively major contributors to the metabolism of many drugs approved by the Food and Drug Administration. Although the role of *CYP2C8* is less clear (there are currently no guidelines supporting clinical use of *CYP2C8* genetic tests due to limited data), clinical guidelines for genotype-guided drug therapy have been developed for *CYP2C9* and *CYP2C19* (PharmGKB, *https://www.pharmgkb.org/prescribingInfo*, last accessed March 18, 2021). For *CYP2C9*, these include several widely prescribed medications such as antifungals and phenytoin,[2] whereas many antidepressants,[3] anitfungals,[4] proton pump inhibitors,[5] and the antiplatelet medication clopidogrel[6] are metabolized by *CYP2C19*.

Many pharmacogene haplotypes, including those for *CYP2C8*, *CYP2C9,* and *CYP2C19,* are defined using the star (*) allele nomenclature, where **1* is designated as the normal or wild-type allele, which often corresponds to the gene's reference sequence. The Pharmacogene Variation Consortium (PharmVar, *https://www.pharmvar.org*, last accessed March 18, 2021) assigns star allele designations and systematically catalogs allelic variation to provide the pharmacogenetic community with a standardized nomenclature system.[7–9] PharmVar displays clinical allele function as assigned by the Clinical Pharmacogenetics Implementation Consortium (CPIC; *https://cpicpgx.org*, last accessed May 19, 2021). Each allele is assigned a predicted enzyme activity that ranges from no function to increased function, leading to a broad phenotypic range between individuals and populations. Activity of an allele may also be substrate dependent. Accurate genotype analysis helps predict a patient's phenotype (or metabolic capacity), which can be utilized, together with other pertinent information, by physicians to practice individualized drug therapy for their patients. CPIC has developed guidelines providing recommendations based on gene—drug pairs to guide drug choice and dose when a patient's genotype information is available.[10]

*CYP2C8, CYP2C9,* and *CYP2C19* have numerous known star alleles (PharmVar, *https://www.pharmvar.org*, last accessed March 18, 2021). Some of the star alleles have only one defining single nucleotide variant, whereas others have more. Also, not every variant is unique to a haplotype; some may occur on more than one star allele, which may complicate genotype calling.

Most pharmacogenetic assays use locus-specific methods designed to identify known variants that allow star allele identification. Rare and novel variants and alleles, which may impact how individuals metabolize and respond to drugs, are, however, not detected using traditional genotyping methods due to assay design. It has been shown that rare and novel variants likely explain some of the interindividual variability of drug response that remains unaccounted for by routine pharmacogenetic testing.[11]

Next-generation sequencing (NGS) technology may be used as a comprehensive pharmacogenetic genotyping platform. Various NGS approaches can be used to detect both common and novel sequence variants.[12] The discovery of novel haplotypes and assignment of their star allele designation by PharmVar lay the groundwork for subsequent functional characterization and eventual inclusion in clinical implementation.

The Centers for Disease Control and Prevention's Genetic Testing Reference Material (GeT-RM) program has previously characterized 137 publicly available genomic DNA reference materials for 28 clinically relevant pharmacogenes using a variety of genotyping and haplotype assignment methods.[13] In the current study, DNA sequence from the same samples was generated using targeted and whole genome sequencing (WGS) methods.

The primary goal of this investigation was to determine whether the previously characterized samples harbor allelic variants that eluded detection by traditional genotyping assays and understand how these changes affected the predicted diplotype and phenotype. This study also examined whether NGS-based sequencing methods could reliably reproduce the prior genotype calls. To that end, results from various sequencing methods and genotype calling tools were compared with each other and with the original star allele calls for these three genes from the previous GeT-RM study.[13]

## Materials and Methods

### DNA Sequence Data and Participating Laboratories

Sequence analysis was performed on DNA derived from 137 cell lines selected from the National Institute of General Medical Sciences and the National Human Genome Research Institute repositories at the Coriell Institute for Medical Research that had been characterized using a variety of different genotyping platforms for 28 pharmacogenetic genes in a previous GeT-RM study.[13]

**Table 1** Overview of Investigator Groups, Data Sets, and Bioinformatic Tools

| | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|
| Institutions | Children's Mercy Research Institute | Baylor College of Medicine, Human Genome Sequencing Center | University of Washington, Genome Sciences and Macrogen Inc., Precision Medicine Institute | University of Victoria, Department of Computer Science and National Cancer Institute (NIH) |
| Investigators | A Gaedigk, NA Miller, EC Boone, WY Wang, EG Farrow | S Scherer, X Qin | D Nickerson, JD Smith, S McGee, A Radhakrishnan, SB Lee | I Numanagić, C Sahinalp |
| Targeted NGS gene panel (sample number) | ADMEseq ($n$ = 137)<br>PGx-seq‡ ($n$ = 137) | PGx-seq‡ ($n$ = 137) | PGRNseq v1§ ($n$ = 134) | PGx-seq ($n$ = 137) |
| WGS | WGS-1* "HiSeqX PGx Cohort" ($n$ = 70 + 26) | WGS-1* "HiSeqX PGx Cohort" ($n$ = 70) | WGS-2† ($n$ = 137) | WGS-1* "HiSeqX PGx Cohort" ($n$ = 70) |
| Analysis tools | Astrolabe software version 0.8.7.2 | Stargazer software version 1.08 | Stargazer software version 1.08 | Aldy software version 3.0 |

*WGS-1 ($n$ = 70) data available at GitHub (*https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort*, last accessed March 21, 2021).
†WGS-2 ($n$ = 137) data generated by Group 3.
‡PGx-seq ($n$ = 137) data generated by Group 2.
§PGRNseq v1 ($n$ = 134) data generated by Group 3.

Volunteer laboratories were selected for this study to maximize the variety of sequencing methods and diplotype calling tools used to characterize the samples. The laboratories involved in this study and the tools and assays used are shown in Table 1.

All transcript and genomic reference sequences (RefSeqs) utilized in this project are according to the National Center for Biotechnology Information (NCBI) Reference Sequence Database (*https://www.ncbi.nlm.nih.gov/refseq*, last accessed September 20, 2021).

## DNA Sequencing and Characterization Protocols

Participating laboratories generated sequence data and performed pharmacogenetic allele calling on the samples using their current laboratory methods as described below. Each laboratory performed allele calling and reported their results to Group 1 (A.G. and E.C.B.) who examined the data for quality and discrepancies.

Sequencing methods and analyses followed two major protocols: three targeted capture sequencing panels for

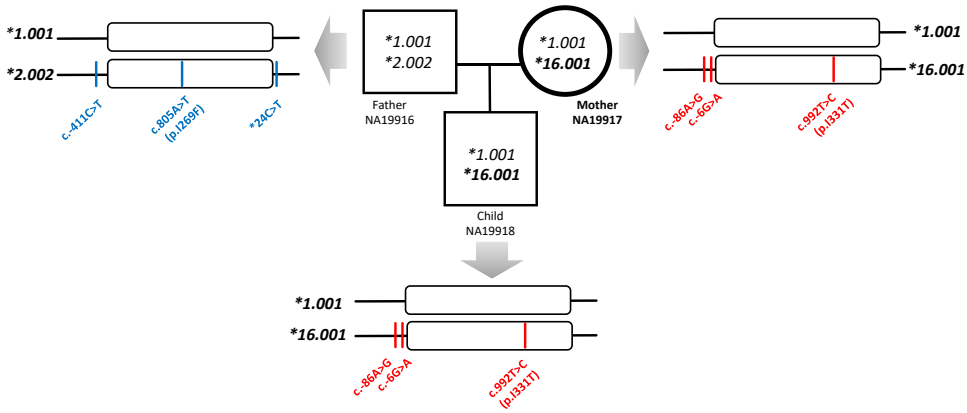


**Figure 1** *CYP2C8* haplotype not recognized by the calling tools. Next-generation sequencing revealed NM_000770.3:c.992T>A (rs146806199) in NA19917 (**bold outline** in pedigree). This missense variant causes a p.Ile331Thr change in exon 7. The haplotype has two additional variants in the 5′ untranslated region (NM_000770.3:c.-6G>A and NM_000770.3:c.-86A>G). The function of this allele is unknown. As shown in the pedigree, the novel allele was inherited by the offspring (NA19918). The phase of the *CYP2C8*16* allele in NA19917 was further corroborated by 10x Linked-Read technology. Because this allele is not part of any of the allele calling tools, it was called as *CYP2C8*1/*1*. The *CYP2C8*2.002* suballele in NA19916 was also only recently designated by PharmVar. Variants inherited together from mother to the child are shown in red, whereas those present on the father (shown in blue) were not passed to the child. Transcript and genomic reference sequences (RefSeqs) are available from the National Center for Biotechnology information (NCBI) Reference Sequence Database (*https://www.ncbi.nlm.nih.gov/refseq*, last accessed September 20, 2021).

genes known to be involved in drug transport and metabolism, and WGS performed on a subset of the samples by two laboratories (Table 1).

### Targeted Sequencing Panels

Three panels capturing different sets of pharmacogenes were utilized for this study:

#### *ADMEseq*

This custom gene panel from Integrated DNA Technologies (IDT, Coraville, IA) targets 289 ADME genes for a total of 660 kb. The amount of upstream and downstream regions covered vary among genes. The regions covered by the panel included 2 kb upstream of the ATG start codon for *CYP2C9* and *CYP2C19,* and 0.5 kb for *CYP2C8,* as well as 250 bp downstream for all three genes. This panel was used by Group 1 for their analysis.

#### *PGRNseq v1*

This custom capture panel (Roche-NimbleGen, Madison, WI), was conceived and characterized by the Pharmacogenetics Research Network (PGRN).[14] This test targets 84 PGx genes including exons, 2 kb upstream and 1 kb downstream of each gene, together with the genotyping targets for the Affymetrix DMET Plus (Affymetrix/Thermo Fisher Scientific, Santa Clara, CA) and Illumina VeraCode ADME (Illumina, San Diego, CA) targeted array platforms[14] for a total of 968 kb. Group 3 performed this test and utilized the generated data for genotype calls.

#### *PGx-seq*

This custom capture panel is an extensively modified version of PGRNseq v1 (Roche-NimbleGen). This test targets 77 genes, including a subset of the PGRNseq v1 gene targets and all the genotyping sites present in PGRNseq v1. A notable difference is that the upstream and downstream regions have been shortened to promote greater multiplexing and reduce costs resulting in a target totaling 458 kb. Group 2 used this method to generate data and genotype calls. Data from this method were also shared with Groups 1 and 4 for independent analyses.

Further details are presented by each collaborating laboratory below.

### WGS

Two independently generated sets of WGS data were utilized for this study as follows:

#### *WGS-1 HiSeqX PGx Cohort*

Briefly, sequencing libraries were prepared from 96 of the 137 Coriell GeT-RM samples using an Illumina TruSeq

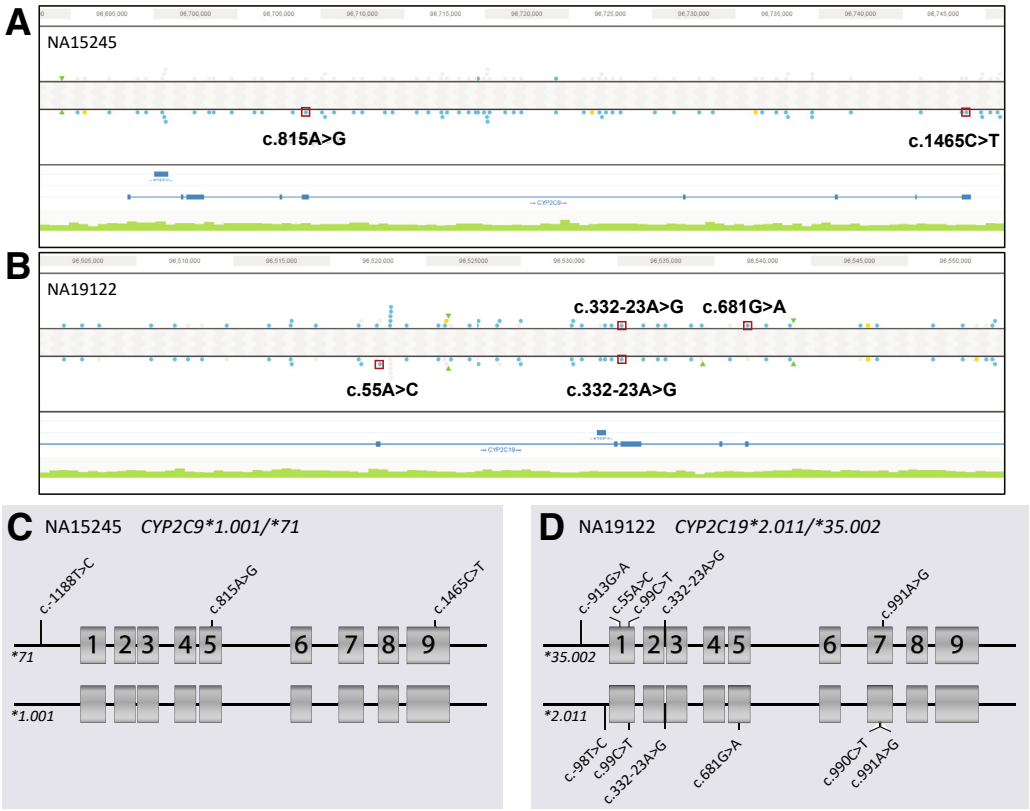


**Figure 2** Novel *CYP2C9* and *CYP2C19* alleles. 10x Genomics Linked-Read data were utilized to phase observed sequence variants across respective genes. **A:** A Loupe screenshot showing that the core variants are in *cis* and thus form a novel *CYP2C9* haplotype (*CYP2C9*71*) in NA15245. **B:** A Loupe screenshot showing two haplotypes, one corresponding to the *CYP2C19*2.011* suballele, whereas the second allele represents the novel *CYP2C19*35.002* suballele in NA19122. **C** and **D:** All variants found on respective *CYP2C9* and *CYP2C19* haplotypes of samples NA15245 and NA19122, respectively, are shown.

DNA PCR-Free kit per the manufacturer's instructions and sequenced on Illumina HiSeq X instruments by the Illumina Clinical Service Laboratory. Samples were sequenced to >30× coverage using a 2 × 150-bp paired-end protocol. Sequence data for 70 of the samples that are consented for public release can be obtained from the European Nucleotide Archive (*https://www.ebi.ac.uk/ena/browser/view/PRJEB19931*, last accessed March 18, 2021). The WGS data set for these 70 samples was obtained through GitHub (*https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort*, last accessed March 24, 2021) by Groups 1, 2, and 4 (Table 1) and is referred to as the HiSeqX PGx Cohort.

### *WGS-2*

WGS was performed by Group 3 on 137 GeT-RM samples using the Illumina TruSeq DNA PCR-Free kit per the manufacturer's instructions. Samples were sequenced on an Illumina NovaSeq 6000 instrument to an average depth of >30× using 2 × 150-bp paired-end sequencing and processed using the Illumina Dynamic Read Analysis for GENomics (DRAGEN) pipeline (3.4.12) for data generation.

Read coverage data and quality metrics for all sequencing tests used are summarized in Supplemental Table S1.

## Star Allele Calling, Group 1

### *ADMEseq*

Group 1 performed sequencing on $n = 137$ Coriell GeT-RM samples using a custom NGS gene panel that includes 287 pharmacogenetic genes. The samples were prepared using an Illumina TruSeq PCR-free library preparation kit with 10 cycles of PCR, followed by enrichment with the custom PGx gene panel from Integrated DNA Technologies to select for the targeted loci. Samples were sequenced on the Illumina MiSeq instrument to an average read depth of approximately 530× over the panel target of 660 kb. Total data were approximately 355 MB with 2 × 200-nt reads. Reads were aligned and variants detected using the Dynamic Read Analysis for GENomics (DRAGEN) Bio-IT platform v2.0.4 − v2.5.3 (Illumina). Variants were called with positions downsampled to 2000 reads using bases with sequence quality ≥10, with mapping quality ≥20, and with a minimum phred-scaled confidence score of 20.0. Read coverage is summarized in Supplemental Table S1.

Star alleles were called from the ADMEseq, PGx-seq, and WGS-1 HiSeqX PGx Cohort data using Astrolabe software version 0.8.7.2 (*https://www.childrensmercy.org/genomesoftwareportal*) with default parameters as described previously.[15,16] Briefly, based on simulation of all theoretical diplotypes, Astrolabe determines the most likely diplotype from a NGS-derived variant call format file using a probabilistic scoring system. The version, v0.8.7.2, utilized for this project contained *CYP2C8, CYP2C9,* and *CYP2C19* allele definitions as defined by PharmVar v4.1.4 (February 14, 2020) (Supplemental Table S2). Astrolabe was run against 137 samples sequenced with the ADMEseq

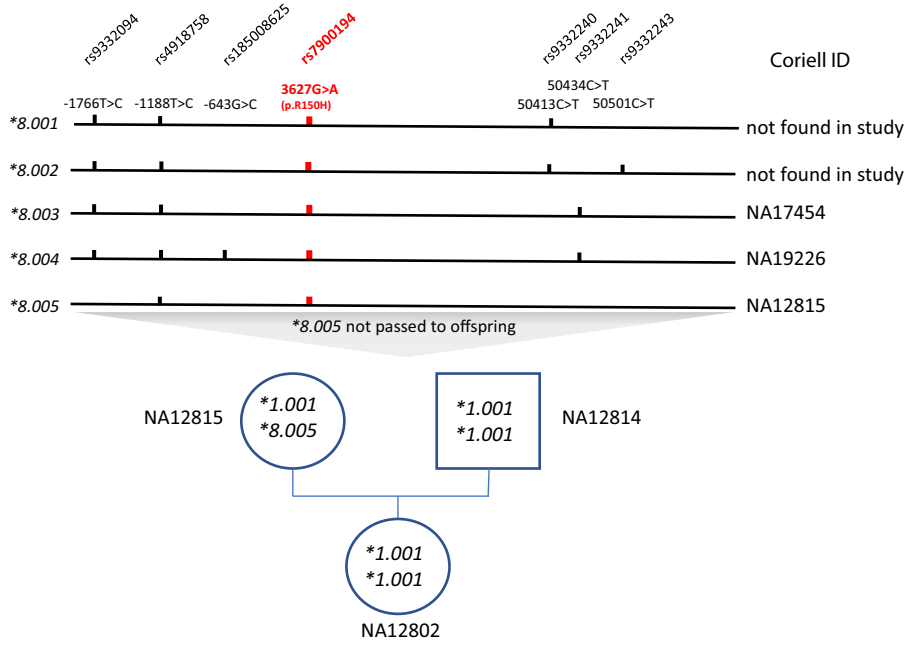


**Figure 3** Discovery of novel *CYP2C9*8* suballeles. The top three lines represent the *CYP2C9*8.001, CYP2C9*8.002,* and *CYP2C9*8.003* suballeles that were defined by PharmVar before the start of the investigation. Of those, only *CYP2C9*8.003* was found among the study samples (the presence of *CYP2C8*8.003* was inferred; no 10x Genomics data were available to confirm this allele call). Two novel *CYP2C9*8* suballeles, designated *CYP2C9*8.004* and *CYP2C9*8.005,* were identified. The latter was discovered in NA12815 and the phase of the two variants informed by inheritance in a trio for which data were obtained from the 1000 Genomes Project; the subject in question is a member of a large pedigree. Although this novel allele has NM_000771.4:c.449G>A, p.Arg150His, it lacked NM_000771.4:c.-1766T>C (rs9332094). The core variant of the *CYP2C9*8* allele is highlighted in red. Transcript and genomic reference sequences (RefSeqs) are available from the National Center for Biotechnology information (NCBI) Reference Sequence Database (*https://www.ncbi.nlm.nih.gov/refseq*, last accessed September 20, 2021).

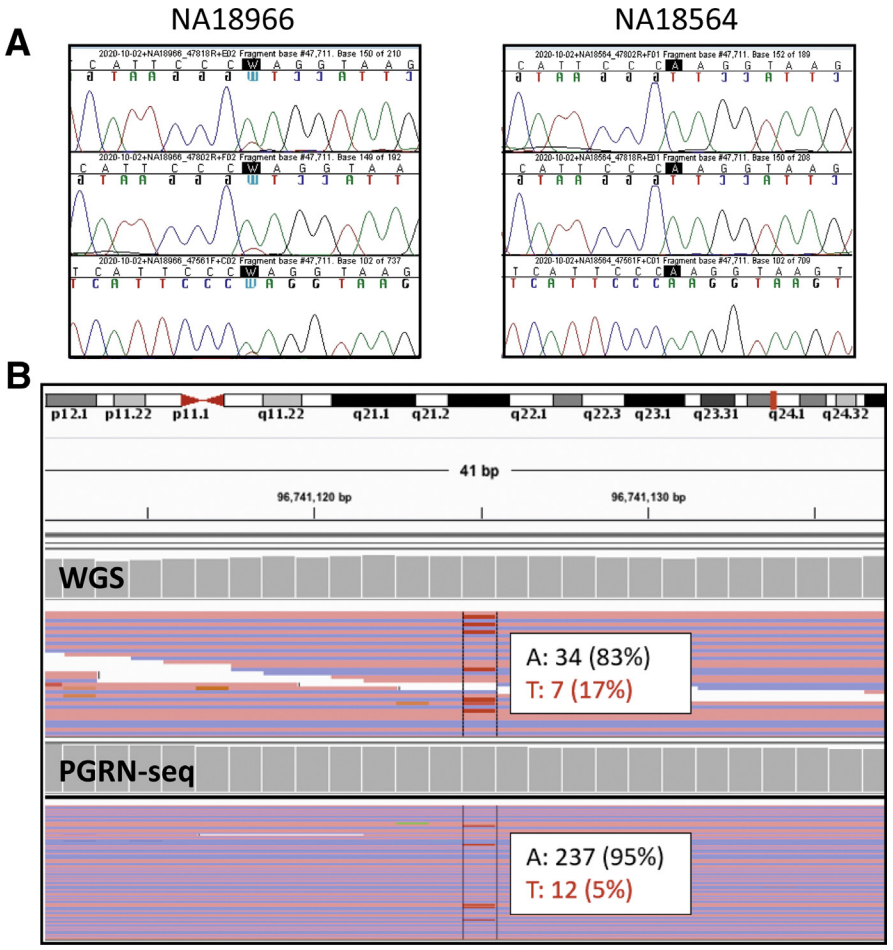


**Figure 4** *CYP2C9* missense variant NM_000771.4:c.1147A>T. A missense variant was discovered in NA18966 at NM_000771.4:c.1147A>T, which introduces a stop codon (p.Lys383Ter). **A:** A forward Sanger sequence trace for NA18966 with the reference c.1147A being the dominant peak. The trace for a *CYP2C9*1/*1* control sample, NA18564, is shown for comparison. **B:** Selected WGS-1 and PGRNseq read alignments with most reads having the reference c.1147A. Read distributions for the variant T were 4.8% (PGRNseq v1, shown), 11% (WGS-2), 17.1% (WGS-1, shown), and 18% (ADMEseq) reads. The variant is visualized by **red horizontal bars** and % reads shown in **red font**. Transcript and genomic reference sequences (RefSeqs) are available from the National Center for Biotechnology Information (NCBI) Reference Sequence Database (*https://www.ncbi.nlm.nih.gov/refseq*, last accessed September 20, 2021).

panel, as well as the PGx-seq data provided by Group 2. In addition, Astrolabe calls were generated using the HiSeqX PGx Cohort WGS data ($n = 70$) and WGS data from 26 additional samples from the GeT-RM project made available by Illumina via direct download through Amazon Web Services (Seattle, WA) for this project.

### Star Allele Calling, Group 2

#### *PGx-seq*

DNA from all 137 GeT-RM Coriell samples was used to prepare the paired-end pre-capture libraries by sonication and ligation to Illumina paired-end adapters. The adapter-ligated DNA was PCR-amplified using primers containing sequencing barcodes (indexes) to enable sample multiplexing. For the target enrichment capture procedure, the pre-capture libraries were enriched by solution hybridization to biotinylated probes (Roche NimbleGen) using a 47-plex format. Sequencing was performed with the Illumina HiSeq 2500 platform using a 94-plex format generating 2 × 101-bp paired-end reads, and reads were mapped to the human reference using Burrow-Wheeler Aligner.[17] To remain compliant with downstream file input requirements, variant call format files were generated for both targeted and whole genome data sets using the Genome Analysis Toolkit (GATK)-Haplotype Caller software version 3.8.0 (Broad Institute, Cambridge, MA; *https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller*).

*CYP2C8*, *CYP2C9,* and *CYP2C19* star alleles were determined from PGx-seq and the WGS-1 HiSeqX PGx Cohort data using the Stargazer genotyping pipeline as described previously.[18,19] Briefly, single nucleotide and insertion/deletion variants for these three genes from the variant call format file were phased using the program Beagle[20] and the 1000 Genomes Project haplotype reference panel. Phased variants and insertion/deletions were then matched to star alleles in a lookup table. Alleles covered by Stargazer are detailed in Supplemental Table S2.

### Star Allele Calling, Group 3

Group 3 performed WGS on 137 samples as described above (WGS-2).

#### *PGRNseq v1*

Custom capture probes for PGRNseq v1 (Roche-NimbleGen) were used on 134 of the Coriell GeT-RM samples. Target enrichment capture using a 24-plex format were sequenced with the Illumina HiSeq 2500 platform with 2 × 100-bp paired-end reads to an average mean coverage of 230×.

Allele calls were made from the PGRNseq v1 and WGS-2 data using Stargazer software version 1.0.8 (*https://stargazer.gs.washington.edu/stargazerweb*) as described above for Group 2.

### Star Allele Calling, Group 4

Star alleles were called using Aldy software version 3.0 (*http://aldy.csail.mit.edu*).[21,22] Aldy was run on data from the 70 publicly available samples from the HiSeqX PGx Cohort and on 137 samples sequenced with the PGx-seq panel (data provided by Group 2). Briefly, Aldy calls star alleles by first enumerating the possible copy number and gene fusion configurations. Because copy number events are either rare or nonexistent in *CYP2C8*, *CYP2C9,* and *CYP2C19*, this step was omitted. Aldy attempts to find the optimal major star alleles directly from a SAM/BAM file in a combinatorial fashion via integer linear programming.[21] Each optimal major star allele solution is later evaluated and refined with suballele data. The solution with the lowest error score is reported as the final star allele call. If there are multiple equally likely solutions to the optimization problem (a rare, but not an impossible, event), Aldy will report all such solutions. Alleles covered by Aldy are detailed in Supplemental Table S2.

### Long-Distance Phasing with 10x Genomics

Ninety-six Coriell GeT-RM samples were prepared using 10x Chromium Single Cell 3′ libraries (10x Genomics, Pleasanton, CA) and sequenced on Illumina HiSeq4000 instruments by Illumina, Inc. Each sample was sequenced across two lanes on the HiSeq4000, resulting in approximately 2 billion reads per sample, or 42.6× coverage. 10x Genomics Linked-Read data from the Illumina HiSeqX-PGx Cohort were used to inform haplotypes, that is, to determine whether variants are in *cis* or *trans*. Data for the 96 sequenced samples were obtained from Illumina collaborators and analyzed with Long Ranger software version 2.2.2 and Loupe software version 2.1 (10x Genomics, Pleasanton, CA) against the GRCh37 reference genome. These 10x phases were cross-validated using an alternative 10x Genomics phasing pipeline (EMA and HapTree-X).[23,24] 10x Genomics data for 70 of the publicly available samples with proper consent are available through GitHub (*https://github.com/Illumina/Polaris/wiki/HiSeqX-PGx-Cohort*, last accessed March 24, 2021).

### Sanger Sequencing (Group 1)

A 1590-bp–long *CYP2C9* PCR product was generated using forward primer 5′-AGAAACCGGAGCCCCTGCAT-3′ and reverse primer 5′-AGAAGGCCAGTTCATCTC-TATGTGC-3′. The resulting 1590-bp PCR product was sequenced in both directions with the 5′-AGAAACCG-GAGCCCCTGCAT-3′ (forward) and 5′-AGTTATG-CACTTCTCTCACCCG-3′ (reverse) primers and aligned to NG_008385.2 to confirm the presence of g.48211A>T (NM_000771.4:c.1147A>T; p.Lys383Ter). Sequencing was performed using BigDye chemistry and a 3730 XL-DNA analyzer instrument (Applied Biosystems, Foster City, CA).

### Variant Summary Lists

A list of sequence variants identified in each sample was compiled for each of the three interrogated genes (Supplemental Table S3). This list was used to facilitate comparisons between sequencing platforms and to identify variants and haplotypes not covered by PharmVar allele designations.

Group 1 created their variant list using WGS-1 and ADMEseq data for $n = 96$ samples, and ADMEseq data for $n = 41$ samples. Of note, although known allelic variants in upstream regions are specifically targeted by the ADMEseq panel, the entire region is not targeted, leading to variations in coverage. The list was created using a combination of bcftools software version 1.9 (a set of utilities that manipulate variant calls in the variant call format), the Genome Analysis Toolkit (GATK) software version 3.8 and Variant Effect Predictor software version 88_37.[25–27]

Group 2 created their variant list using the Genome Analysis Toolkit (GATK)-HaplotypeCaller.[26]

Group 3 created their variant list from the WGS-2 data. The list was created using bcftools software version 1.9 and a Browser Extensible Data (BED) file that annotates the regions into upstream, exon, and downstream regions. Intronic variants known to have a functional impact (eg, *CYP2C19*2* NM_000769.4:c.332-23A>G) causing alternative splicing) were listed in the exon category.[25] Differences among the lists created by each of the three groups were identified and resolved by manual data inspection.

Race and ethnic origins of the samples in Supplemental Table S3 are according to those provided by the Coriell Institute for Medical Research (Camden, NJ).

## Results

Four groups (Groups 1 to 4) participated in this study as described in Table 1. The groups analyzed NGS data obtained by two independent WGS data sets and three targeted NGS gene panels using three allele calling algorithms: Aldy software version 3.0, Astrolabe software version 0.8.7.2, and Stargazer software version 1.0.8 (from hereon referred

to as "tools"). A comprehensive summary of all data from this study is provided in Supplemental Table S3.

## Identification of Novel Alleles

To systematically identify all novel variants not currently defined by the PharmVar database and to determine whether these variants are part of known or novel haplotypes, lists summarizing the variants found in each sample were generated for the three genes by Groups 1 to 3 using their respective data sets. These three lists were then used to create a consensus list for each gene (Supplemental Table S3). Variants are shown in separate columns in Supplemental Table S3 based on their location (upstream, coding including exon/intron junctions, and downstream); those novel to PharmVar are highlighted. For selected samples, the unequivocal phase of variants in a haplotype was determined using 10x Genomics Linked-Read data and/or by inheritance using family trio data by Groups 1 and 4. The novel haplotypes for which variant phase was established were submitted to PharmVar for designation. Novel haplotypes and method(s) used to establish haplotype for all three genes are summarized in Supplemental Table S4.

All sequencing methods, except for the PGx-seq panel, covered the regions required by PharmVar for allele definitions (*CYP2C9* and *CYP2C19*, 2 kb of upstream region; *CYP2C8,* 0.5 kb of upstream region, and 250 bp of the 3′UTR for each gene). Sequencing coverage metrics are provided in Supplemental Table S1. Data were available for at least two sequencing methods for each sample covering the regions of interest. Allele frequencies cited below are according to those reported by dbSNP (*https://www.ncbi.nlm.nih.gov/snp*, last accessed February 2, 2021). For alleles without a unique identifying variant, estimated frequencies are provided.

### CYP2C8

NGS consensus calls for *CYP2C8* are shown in Supplemental Table S3. Four novel alleles, *CYP2C8*15-*18*, and 15 novel *CYP2C8*1* suballeles, *CYP2C8*1.004-*1.018,* were identified among the 137 GeT-RM samples (Supplemental Table S4). *CYP2C8*15* ($n$ = 1 Caucasian) has a single variant [NM_000770.3:c.541G>A, p.Val181Ile (rs41286886)]; its frequency ranges between 0.2% and 1.1%. *CYP2C8*16* ($n$ = 1 African American) was characterized using inheritance information (Figure 1). This allele has three variants, one of which, NM_000770.3:c.992T>C, causes an amino acid change [p.Ile331Thr (rs146806199)]. Based on dbSNP frequency data, this allele is rare (<0.1%) and may predominantly be observed in Asians. *CYP2C8*17* ($n$ = 2 Yoruban) has two variants, one of which is nonsynonymous [NM_000770.3:c.730A>G, p.Ile244Val (rs11572102)]. This allele is also rare at frequencies of <0.2% across populations. Finally, *CYP2C8*18* ($n$ = 1 Caucasian) was discovered in NA07048 in this study. Because there were no 10x Genomics Linked-Read or pedigree data available for this sample, the haplotype was defined using an unrelated trio that was identified via the allele's core variant [NM_000770.3:c.1081C>T, p.Leu361Phe (rs45438799)]. This allele also appears to be rare with a frequency of 0.003%. Of note, all samples identified as having *CYP2C8*15, *16, *17,* or **18* alleles were consistently called as *CYP2C8*1/*1* by the allele calling tools, and only one of the observed differing calls (Supplemental Table S5) was caused by the presence of a novel haplotype, *CYP2C8*1.010*.

### CYP2C9

NGS consensus calls for *CYP2C9* are shown in Supplemental Table S3. One novel allele, *CYP2C9*71,* seven novel *CYP2C9*1* suballeles (**1.007-*1.013*), and two novel *CYP2C9*8* suballeles, *CYP2C9*8.004* and *CYP2C9*8.005*, were identified among the 137 GeT-RM samples (Supplemental Table S4).

*CYP2C9*71* ($n$ = 1, race/ethnicity unknown) has two nonsynonymous variants, [NM_000771.4:c.815A>G, p.Glu272Gly (rs9332130) and NM_000771.4:c.1464C>T p.Pro489Ser (rs9332239)], which are the defining variants for *CYP2C9*12* and *CYP2C9*10*, respectively (Figure 2). This haplotype, identified in NA15245, caused inconsistent genotype calls among the tools (Supplemental Table S6). Because this allele does not have a single unique variant, its frequency is estimated to be under 0.004% based on the rarer of the two variants in this haplotype. Ambiguous calls for NA15245 (Supplemental Table S6) were resolved with 10x Genomics data showing that the *CYP2C9*10* and **12* core variants are indeed in *cis* as predicted by Stargazer (this novel haplotype was designated *CYP2C9*71* by PharmVar).

One of the two novel *CYP2C9*8* suballeles, *CYP2C9*8.004,* was found in NA19226 (Yoruban). This allele has an additional variant in the upstream region, NM_000771.4:c.-643G>C (rs185008625) (Figure 3). The second novel *CYP2C9*8* suballele, *CYP2C9*8.005*, was found in NA12815. Of note, this is the first *CYP2C9*8* allele identified in a Caucasian subject. This allele not only lacks variants in exon 9, but also lacks NM_000771.4: c.-1766T>C (rs9332094). The designation of this haplotype caused the PharmVar *CYP2C9* expert panel to reverse the core variant status for c.-1766T>C, which allowed this haplotype to be categorized as a novel *CYP2C9*8* suballele instead of designating it as a novel major allele. There is evidence suggesting that c.-1766T>C decreases expression levels, but the data were deemed inconclusive upon re-evaluation.[28] Finally, of the novel *CYP2C9*1* suballeles, all but **1.009* have multiple variants in the upstream region, and each of *CYP2C9*1.007, CYP2C9*1.009, CYP2C9*1.011,* and *CYP2C9*1.013* also contain one synonymous variant.

NA17290 (Caucasian) has two novel *CYP2C9* variants that are on the same allele, NM_000771.4:c.295A>C

(rs750662900) and NM_000771.4:c.296T>A (rs763302345). These two variants are adjacent to each other and were found on the same NGS reads indicating that they are in *cis*. The presence of this variant combination (NM_000771.4:c.295_296CAdelins) causes a p.Ile99His amino acid change, whereas each variant on its own would cause p.Ile99Leu and p.Ile99Asn changes, respectively. However, because the sample also has the *CYP2C9*3*-defining variant NM_000771.4:c.1075A>C (as well as several variants in the upstream region), it remains unknown whether this haplotype is a novel *CYP2C9*3* suballele or rather represents a novel haplotype. Unfortunately, no 10x Genomics Linked-Read or pedigree data were available for this sample.

Lastly, a single variant, NM_000771.4:c.1147A>T, p.Lys383Ter, was found in sample NA18966 (Japanese). This nonsense variant was observed by all sequencing platforms including confirmatory Sanger sequencing; however, there was consistent allele imbalance (4.8%, PGRNseq v1; 11%, WGS-2; 17.1%, WGS-1; 18%, ADMEseq) (Figure 4 and Supplemental Table S6). This variant was first described (in the same sample) by Lee et al[19] and termed as **S1* in Stargazer. This allele was not submitted to PharmVar for naming due to concerns that the variant may be a cell line–specific mutation.

### CYP2C19

NGS consensus calls for *CYP2C19* are shown in Supplemental Table S3. One novel star allele, *CYP2C19*39*, and 14 novel *CYP2C19* suballeles were identified (Supplemental Table S4). This novel *CYP2C19*39* allele, found in two Yoruban samples (NA19143 and NA19213), is characterized by three nonsynonymous variants [NM_000769.1:c.55A>C, p.Ile19Leu (rs17882687); NM_000769.1:c.365A>C, p.Glu122Ala (rs17885179), and NM_000769.1:c.991A>G, p.Ile331Val (rs3758581)]. Of particular interest is c.55A>C (p.Ile331Val), which is part of two other star allele definitions: *CYP2C19*15* and *CYP2C19*28*. The *CYP2C19*39* allele is rare at a global frequency of 0.062% but varies across populations.

For sample NA19122 (Yoruban), the novel *CYP2C19*35.002* suballele includes the shared variant with *CYP2C19*2*, but also contains c.55A>C, which is part of three other haplotypes, *CYP2C19*15, CYP2C19*28,* and the novel *CYP2C19*39* allele (Figure 2). Due to the presence of c.55A>C and c.332-23A>G, phasing data were required to call this haplotype. NA19122 (Yoruban) also possessed a novel *CYP2C19*2* suballele, *CYP2C19*2.011* (Supplemental Table S4).

Finally, *CYP2C19*38* is an allele that was designated by PharmVar while this investigation was underway.[29] This allele was called by the tools as *CYP2C19*1*, but unlike *CYP2C19*1, CYP2C19*38* lacks NM_000769.1:c.991A>G, p.Ile331Val, (rs3758581). Two novel *CYP2C19*38* suballeles (**38.002* and **38.003*) were identified in study samples (Supplemental Table S4). Sequence information showed that 13 (8.4%) of the 155 alleles initially called as *CYP2C19*1* are in fact *CYP2C19*38*. The *CYP2C19*38* allele was found in Caucasians ($n = 7$), Han Chinese ($n = 2$), Japanese ($n = 2$), Mexican/American ($n = 1$), and unknown ($n = 1$).

## Aldy, Astrolabe, and Stargazer ("Tool") Diplotype Calls

Alleles called by the tools (Supplemental Table S3) correspond to those described by PharmVar at the outset of the study (see Supplemental Table S2 for alleles covered by each tool). Therefore, the tool-generated diplotype calls did not include any of the novel haplotypes discovered in this investigation. Supplemental Table S5 (*CYP2C8*), Supplemental Table S6 (*CYP2C9*), and Supplemental Table S7 (*CYP2C19*) are derived from Supplemental Table S3 and highlight ambiguous calls or calls that were inconsistent among the tools. Brief explanations are provided for each observed inconsistency within the respective tables.

Overall, diplotype calls were consistent among the tools for the vast majority of samples (Supplemental Table S3). Many of the inconsistent and ambiguous calls could be explained by the presence of novel alleles or suballeles. It is important to note, that the presence of novel alleles did not necessarily lead to call inconsistencies and that several novel alleles were found in samples that were consistently called as **1/*1* by all tools for all sequencing methods. One example is NA07048, which was called as *CYP2C8*1/*1* by all tools even though this sample harbors the novel *CYP2C8*18* allele.

## NGS Consensus Calls Versus Previous GeT-RM Calls and Impact on Phenotype Prediction

NGS-based consensus calls (Supplemental Table S3) include the novel alleles identified in this study; these calls may differ from the tool calls. Phenotype predictions for *CYP2C9* and *CYP2C19* are according to those provided by the PharmGKB reference tables for genotype to phenotype translation (PharmGKB, *https://www.pharmgkb.org/page/pgxGeneRef*, last accessed April 6, 2021); there is no genotype-to-phenotype translation table for *CYP2C8*.

### CYP2C8

NGS consensus calls differed from the previous GeT-RM consensus calls[13] for five samples. All were called as *CYP2C8*1/*1* in the previous study and were reassigned as *CYP2C8*1/*15* ($n = 1$), *CYP2C8*1/*16* ($n = 1$), *CYP2C8*1/*17* ($n = 2$), and *CYP2C8*1/*18* ($n = 1$) (Supplemental Table S3). The function of the novel alleles is unknown, and therefore, it is impossible to predict the impact on phenotype.

### CYP2C9

Twelve samples were assigned an ambiguous genotype, *CYP2C9*3* (**18*), in the previous GeT-RM study.[13]

*CYP2C9*18* has an additional variant [NM_000771.4: c.1190A>C, p.Asp397Ala, (rs72558193)], which was not interrogated by the methods used in that study, and thus, *CYP2C9*3* and *CYP2C9*18* could not be differentiated. *CYP2C9*18* was not found in any of the samples using NGS. This revision did not impact phenotype prediction.

NA17102 was initially called *CYP2C9*1/*5* and revised to *CYP2C9*5/*36,* which changes the phenotype prediction from intermediate metabolizer (IM) to indeterminate (Supplemental Table S3). This NGS consensus call assumes that NM_000771.4:c.1080C>G, p.Asp360Glu and NM_000771.4:c.1A>G, p.Met1Val are in *trans* per current allele definitions. Stargazer, as detailed in Supplemental Table S6, suggests that these variants may occur in *cis* in this sample. Unfortunately, this could not be substantiated because 10x Genomics data were not available for NA17102.

The diplotype for HG01190 was revised from *CYP2C9*1/*2* to *CYP2C9*2/*61*; the *CYP2C9*61* allele was not tested in the previous study. The presence of the *CYP2C9*61* allele did not impact the IM phenotype prediction.

Finally, NA15245 was revised from *CYP2C9*10/*12* to *CYP2C9*1/*71,* which left phenotype prediction as indeterminate.

There was also one sample, NA17290, for which the diplotype could not be resolved because no 10x Genomics Linked-Read data were available. Depending on the phase of the novel variation (NM_000771.4:c.295_296CAdelins), the p.Ile99His change may be located on the *CYP2C9*3* allele giving rise to a novel suballele or represent a novel haplotype. NA17290 was reported as *CYP2C9*1/*3 (*18)* in the previous study.

### CYP2C19

Two samples, NA19143 and NA19213, were found to have a novel *CYP2C19* allele. Both samples were reported as *CYP2C19*1/*15* in the previous study and were revised to *CYP2C19*1/*39*, which changes their phenotype assignment from normal metabolizer (NM) to indeterminate (Supplemental Table S3).

Five samples had ambiguous calls in the previous GeT-RM study.[13] Sequencing confirmed the presence of a *CYP2C19*12* allele in NA17074, which allowed us to update the genotype from *CYP2C19*1(*12)/*17* (rapid metabolizer or indeterminate phenotype) to *CYP2C19*12/*17* (indeterminate). NA19122 had a *CYP2C19*1 (*15)/*2* assignment in the previous study, which was revised to *CYP2C19*2/*35*, changing the predicted phenotype from IM or indeterminate to poor metabolizer. NA19700 was initially reported as *CYP2C19*1/*12* (indeterminate phenotype); because NGS did not detect a *CYP2C19*12* allele in this sample its genotype was revised to *CYP2C19*1/*1* changing the phenotype prediction from indeterminate to normal metabolizer. NA19917 was reported as *CYP2C19*1 (*15; *28)/*2* in the previous study. This ambiguous call was revised to *CYP2C19*2/*15,* which changed the phenotype prediction from IM or indeterminate to IM. Lastly, NA23878 was previously described as *CYP2C19*1/*4B* with a possible alternate diplotype of *CYP2C19*4/*17*.[30] Because no 10x Genomics data were available, we were not able to determine the sample's diplotype with certainty. However, the predicted phenotype is the same for both possible diplotypes.

NA17074 (Puerto Rican) was previously reported as *CYP2C19*1(*12)/*17,* suggesting the possible presence of a rare *CYP2C19*12* allele. Although the presence of the *CYP2C19*12*-identifying variant NM_000769.1:c.1473A>C, p.Ter491Cys (rs55640102) was confirmed by NGS, the ambiguous Stargazer call (*CYP2C19*1/*2 [*17]*) raised concerns regarding the phase of the variants. Unfortunately, because no 10x Genomics data were available for this sample, the sample's diplotype could not be determined with certainty.

Although the NGS consensus call matches that of the previous study for NA07439 (African American), the Stargazer call also raises concerns regarding variant phasing for this sample. Unfortunately, no 10x Genomics data were available for this sample.

Alleles reported as *CYP2C19*27* in the previous GeT-RM study[13] were changed to *CYP2C19*1* to reflect changes in star allele definitions,[29] which were made while this investigation was underway. In addition, *CYP2C19*1* allele calls for 13 samples were revised to *CYP2C19*38* (Supplemental Table S3). Because *CYP2C9*1* and **38* are considered normal function alleles, this change does not affect phenotype prediction.

## Discussion

The previous GeT-RM study[13] utilized a variety of commercial and laboratory-developed genotyping platforms to characterize the 137 samples that were reexamined in the current study. The genotyping platforms were designed to distinguish the presence or absence of specific variants defining a limited set of star alleles. As genotyping assays typically include the more commonly found variants, rare or novel variants that may also affect protein structure, function, and phenotype prediction would not be detected. The goal of this study was to recharacterize *CYP2C8, CYP2C9,* and *CYP2C19* in the 137 samples using WGS and targeted NGS gene panels to assess the differences between sequence-based genotyping and to provide a more robust characterization of these previously studied samples.

For this study, the authors examined *CYP2C8*, *CYP2C9,* and *CYP2C19*. Of those, *CYP2C9* and *CYP2C19* are well-characterized, widely tested, and have guidelines to support clinical utility (CPIC, *https://cpicpgx.org/guidelines*, last accessed June 1, 2021). Although several drugs are metabolized by *CYP2C8*, there are currently no clinical guidelines, and genotyping is not routinely performed.

Because *CYP2C8* has not been as well characterized as the other two genes, this study offered the opportunity to assess the extent of variation and close this knowledge gap.

The use of sequence-based data generated with different NGS technologies allowed detection of novel alleles, resolution of ambiguous genotypes, and reaffirmation or modification of phenotype assignment for several samples. The changes in predicted phenotype, such as from CYP2C9 IM to indeterminate (Supplemental Table S3), or CYP2C19 IM (or indeterminate) to poor metabolizer (Supplemental Table S3) would have an impact on clinical management based on CPIC and/or DPWG recommendations. As with any clinical testing scheme, if there is a strong clinical suspicion that a patient may have rare no-function variants that were not interrogated by a targeted panel test, additional testing such as WGS or targeted NGS may be indicated. In such cases, the clinician must balance identifying variants of unknown or uncertain clinical significance in NGS assays versus testing a panel of known variants.

Overall, NGS-based genotype calls correlated well with variant-based genotyping for *CYP2C8*, *CYP2C9*, and *CYP2C19* except for the identification of novel alleles. In other words, the original GeT-RM calls were correct, considering the constraints of limited testing and the catalog of defined star alleles available at that time. However, it was not surprising that the current study revealed several rare and novel variants not detected by genotype approaches (Supplemental Table S3).

Although only a relatively small number of samples ($n$ = 137) were examined, several rare or novel haplotypes were identified (*CYP2C8*, $n$ = 4, *CYP2C9*, $n$ = 1, and *CYP2C19*, $n$ = 1) for all three genes as well as numerous novel suballeles (Supplemental Table S4). Finding novel haplotypes was not surprising, given that variation in human *CYP* genes is extensive.[31] This highlights the need to identify and fully characterize novel alleles and submit them to PharmVar. A more complete inventory of genetic variation of these genes allows better understanding of whether sequence-based or targeted variant-based genotyping approaches adequately predict a patient's phenotype, regardless of race or ethnicity.

One novel variant, NM_000771.4:c.1147A>T (annotated as **S1* by Stargazer), was detected in a single sample, NA18966. Of concern, this variant consistently presented with severe allele imbalance across all NGS-based data and even Sanger sequencing (Figure 4). In fact, the extreme allele imbalance in both PGx-seq and PGRNseq data sets caused the variant to be filtered out during variant calling. The variant is a stop-gain mutation that has no rsID but is reported in gnomAD v3.1.1 (10-94981368-A-T) (gnomAD, *https://gnomad.broadinstitute.org*, last accessed September 20, 2021) as a singleton in 152,116 counts. This sample is part of the 1000 Genomes Project database and shows an imbalance similar to the one described here. One can speculate that this is not a germline variant but rather the result of a cell line–specific variant or mosaicism in the donor. The possibility of DNA contamination was excluded because allele fractions of other variants in the vicinity of NM_000771.4:c.1147A>T were within expected ratios. Additionally, NA18966 was previously shown to contain a duplication in chromosome Y that is most likely to be a cell line artifact.[32,33] Because it remains uncertain whether this variant is a mutation that arose in the cell line, this haplotype was not submitted to PharmVar for allele designation.

Novel variants or haplotypes are often defaulted to a **1* allele assignment, which is common practice if none of the tested variants are identified. This was the case for NA19143, which was called *CYP2C19*1/*15* by all three tools despite the presence of the novel *CYP2C19*39* allele, defined by NM_000769.1:c.365A>C, p.Glu122Ala (Supplemental Tables S3 and S4). Because *CYP2C19*39* was novel and not defined in the calling algorithms, this haplotype was not called by the tools used in the study. In addition, calling algorithms may not always be completely up to date with the most recent version of alleles available in PharmVar, and thus can miss calling recently added alleles. This is exemplified by *CYP2C9*61,* which was accurately called in HG01190 by Aldy and Astrolabe but defaulted to *CYP2C9*1/*2* by the Stargazer version utilized for this study. Also, depending on specific reporting features of each tool, the presence of novel variant(s) may be reported separately from the diplotype call and would require manual follow-up by the user.

Some novel alleles were identified indirectly by the tools (Supplemental Table S6). For example, a novel *CYP2C9*71* haplotype in sample NA15245 did not default to a **1* assignment, but caused inconsistent calls among the tools (Aldy, *CYP2C9*10/*12*; Astrolabe, *CYP2C9*1/*10* or **1/*12,* and Stargazer, **1/*12 [*10]*). The ambiguous and differing calls made by the tools were caused by the novel haplotype having variants including those that were otherwise found on *CYP2C9*10* and *CYP2C9*12*, respectively. In this case, a tool's variant output (list of variants present) would not signal the presence of a novel haplotype because all variants are part of other allele definitions. It remains to be seen if expanding the allele inventory of these tools to include *CYP2C9*71* would indeed produce an accurate *CYP2C9*1/*71* genotype call for this sample.

One limitation of the current sequencing by synthesis approach is that haplotype phasing may be uncertain; however, once the presence of a novel variant is identified, the full haplotype may be resolved using a variety of approaches. In this study, long-read NGS data were used to determine or validate the phase of variants. This information was invaluable to fully characterize novel alleles (ie, determine which variants are on each chromosome) and confirm existing allele definitions. The utility of such data is well illustrated by sample NA15245 (described in the preceding paragraph and shown in Figure 2), which conclusively showed that the two core variants defining *CYP2C9*10* and **12* are not in *trans* in this sample, but in *cis*, forming a novel haplotype (Supplemental Table S6). This finding does, however, raise

some concerns regarding the accuracy of the current definitions of *CYP2C9*10* and *12, which were first described in a subject having a *CYP2C9*10/*12* genotype[34]; to the best of our knowledge, there have been no other reports validating these allele definitions. A complex *CYP2C19* diplotype was also resolved with long-read NGS data for sample NA19122, which would have otherwise remained ambiguous (Figure 2).

Samples for which pedigree information from the 1000 Genomes Project was available to infer the phase of variants are also presented (Figures 1 and 3, and Supplemental Table S4). If neither data are available, one may also search for other samples within the 1000 Genomes Projects or other databases that have the variants corresponding to those found in the proband. This approach was taken to complement the characterization of novel *CYP2C8*1* and *CYP2C9*1* suballeles. The predicted *CYP2C8*1.005* haplotype in sample NA23878 was found to be homozygous in HG03740, which was not part of this study, and the predicted *CYP2C9*1.010* haplotype found in NA18861 matched that of a trio in the CMH data warehouse (data accessible to A.G., E.C.B, and N.A.M.). In the absence of long-read (phased) data for a patient of interest, variant phases (haplotypes) may be determined using pedigree analysis.

Despite the relatively small sample size, this follow-up investigation demonstrates the importance of accurate and complete star allele definitions so that calling tools produce accurate diplotype calls. It also underscores that efforts need to continue to discover and catalog star alleles, and that tools need to be updated as the catalog of star alleles continues to grow.[15] Clinical laboratories will need to validate any updates made to software or calling tools in accordance with accrediting agencies, and prescribed by their process or policy for updating tools; equivalency can be documented by reanalyzing files with the updated tool(s). Updating the tools used in this study was beyond its scope, because this requires each tool to be independently revised by their respective developers.

PharmVar does not currently include intronic variants in allele definitions unless they have been demonstrated to cause aberrant splicing or cause altered activity through different mechanisms. NM_000769.1:c.332-23A>G found in *CYP2C9*2* and *35 is a prime example of a variant causing a splice defect. Alleles with synonymous variants are cataloged by PharmVar as suballeles, assuming they do not impact activity (eg, several *CYP2C8*1* and *CYP2C9*1* suballeles have synonymous variants) but may be assigned their own star allele if evidence arises that a synonymous variant alters activity.

Overall, calls made from NGS (both WGS and targeted panels) data provided accuracy on par with, or superior to, the results from genotyping methods. Furthermore, given that remaining errors in star allele calling from both NGS and genotyping data were more often a consequence of incomplete catalogs of star alleles and suballeles than errors in variant detection, it is significant that the ability of NGS data to accurately detect novel star alleles and suballeles was demonstrated. Also, the use of WGS offers the advantage of examining sequence in non-coding genomic regions and provides better performance of structural variant characterization when compared with targeted NGS. Targeted NGS represents a more cost-effective approach that can still discover novel and rare coding variants as a halfway step between genotyping and WGS, especially for genes that do not require routine testing for gene copy number variation, such as *CYP2C8, CYP2C9,* and *CYP2C19* for which copy number variations are rarely observed.[28,29] The accuracy of WGS for complex gene loci, such as *CYP2D6*, needs to be more systematically evaluated (side-by-side comparisons of tools on data sets that include a variety of reference materials with gene copy number variation). Although emerging data on *CYP2D6* are promising,[19,35–37] *CYP2D6* analysis remains challenging owing to its highly polymorphic nature and the presence of gene deletions, duplications and multiplications, and rearrangements with pseudogenes that give rise to hybrid genes in various configurations.[38]

The results of this and other studies demonstrate that there are many novel alleles that are yet to be discovered, even in highly characterized genes such as *CYP2C9* and *CYP2C19*. This highlights the need for continued development of reference materials for pharmacogenetic testing, particularly in under-represented populations, that can be used to develop and validate allele calling algorithms, develop and validate new assays, provide quality control, and enable further research. Information about these and other reference materials is available on the GeT-RM website (*https://www.cdc.gov/labquality/get-rm/index.html*, last accessed March 26, 2021).

## Acknowledgments

We acknowledge Dr. Richard Gibbs, Donna Muzny, and other members of the BCM-HGSC production group for their contributions, and thank the *All of Us* project and Illumina, Inc. for the contribution of whole genome sequences from GeT-RM samples. Some 10x Genomics data were provided by the Emerging Applications group at Illumina.

## Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2021.12.011*.

## References

1. Saravanakumar A, Sadighi A, Ryu R, Akhlaghi F: Physicochemical properties, biotransformation, and transport pathways of established and newly approved medications: a systematic review of the top 200 most prescribed drugs vs. the FDA-approved drugs between 2005 and 2016. Clin Pharmacokinet 2019, 58:1281–1294
2. Daly AK, Rettie AE, Fowler DM, Miners JO: Pharmacogenomics of CYP2C9: functional and clinical considerations. J Pers Med 2017, 8:1

3. Hicks JK, Bishop JR, Sangkuhl K, Müller DJ, Ji Y, Leckband SG, Leeder JS, Graham RL, Chiulli DL, LLerena A, Skaar TC, Scott SA, Stingl JC, Klein TE, Caudle KE, Gaedigk A; Clinical Pharmacogenetics Implementation Consortium: Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 and CYP2C19 genotypes and dosing of selective serotonin reuptake inhibitors. Clin Pharmacol Ther 2015, 98:127–134
4. Moriyama B, Obeng AO, Barbarino J, Penzak SR, Henning SA, Scott SA, Agundez J, Wingard JR, McLeod HL, Klein TE, Cross SJ, Caudle KE, Walsh TJ: Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP2C19 and voriconazole therapy. Clin Pharmacol Ther 2017, 102:45–51
5. Li XQ, Andersson TB, Ahlström M, Weidolf L: Comparison of inhibitory effects of the proton pump-inhibiting drugs omeprazole, esomeprazole, lansoprazole, pantoprazole, and rabeprazole on human cytochrome P450 activities. Drug Metab Dispos 2004, 32: 821–827
6. Scott SA, Sangkuhl K, Stein CM, Hulot J-S, Mega JL, Roden DM, Klein TE, Sabatine MS, Johnson JA, Shuldiner AR; Clinical Pharmacogenetics Implementation Consortium: Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. Clin Pharmacol Ther 2013, 94: 317–323
7. Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, Klein TE; PharmVar Steering Committee: The Pharmacogene Variation (PharmVar) Consortium: incorporation of the human cytochrome P450 (CYP) allele nomenclature database. Clin Pharmacol Ther 2018, 103:399–401
8. Orringer MB, Bluett M, Deeb GM: Aggressive treatment of chylothorax complicating transhiatal esophagectomy without thoracotomy. Surgery 1988, 104:720–726
9. Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Twist GP, Klein TE, Miller NA; PharmVar Steering Committee: The evolution of PharmVar. Clin Pharmacol Ther 2019, 105:29–32
10. Relling MV, Klein TE, Gammal RS, Whirl-Carrillo M, Hoffman JM, Caudle KE: The Clinical Pharmacogenetics Implementation Consortium: 10 years later. Clin Pharmacol Ther 2020, 107:171–175
11. Lauschke VM, Ingelman-Sundberg M: Precision medicine and rare genetic variants. Trends Pharmacol Sci 2016, 37:85–86
12. Schwarz UI, Gulilat M, Kim RB: The role of next-generation sequencing in pharmacogenetics and pharmacogenomics. Cold Spring Harb Perspect Med 2019, 9:a033027
13. Pratt VM, Everts RE, Aggarwal P, Beyer BN, Broeckel U, Epstein-Baak R, Hujsak P, Kornreich R, Liao J, Lorier R, Scott SA, Smith CH, Toji LH, Turner A, Kalman LV: Characterization of 137 genomic DNA reference materials for 28 pharmacogenetic genes: a GeT-RM collaborative project. J Mol Diagn 2016, 18:109–123
14. Gordon AS, Fulton RS, Qin X, Mardis ER, Nickerson DA, Scherer S: PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. Pharmacogenet Genomics 2016, 26: 161–168
15. Twist GP, Gaedigk A, Miller NA, Farrow EG, Willig LK, Dinwiddie DL, Petrikin JE, Soden SE, Herd S, Gibson M, Cakici JA, Riffel AK, Leeder JS, Dinakarpandian D, Kingsmore SF: Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. NPJ Genom Med 2016, 1:15007
16. Twist GP, Gaedigk A, Miller NA, Farrow EG, Willig LK, Dinwiddie DL, Petrikin JE, Soden SE, Herd S, Gibson M, Cakici JA, Riffel AK, Leeder JS, Dinakarpandian D, Kingsmore SF: Erratum: Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. NPJ Genom Med 2017, 2:16039
17. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754–1760
18. Lee S-B, Wheeler MM, Patterson K, McGee S, Dalton R, Woodahl EL, Gaedigk A, Thummel KE, Nickerson DA: Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. Genet Med 2019, 21:361–372
19. Lee S-B, Wheeler MM, Thummel KE, Nickerson DA: Calling star alleles with Stargazer in 28 pharmacogenes with whole genome sequences. Clin Pharmacol Ther 2019, 106:1328–1337
20. Loka TP, Tausch SH, Renard BY: Reliable variant calling during runtime of Illumina sequencing. Sci Rep 2019, 9:16502
21. Numanagić I, Malikić S, Ford M, Qin X, Toji L, Radovich M, Skaar TC, Pratt VM, Berger B, Scherer S, Sahinalp SC: Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. Nat Commun 2018, 9:828
22. Numanagić I, Malikić S, Pratt VM, Skaar TC, Flockhart DA, Sahinalp SC: Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. Bioinformatics 2015, 31: i27–i34
23. Shajii A, Numanagić I, Whelan C, Berger B: Statistical binning for barcoded reads improves downstream analyses. Cell Syst 2018, 7: 219–226.e5
24. Berger E, Yorukoglu D, Zhang L, Nyquist SK, Shalek AK, Kellis M, Numanagić I, Berger B: Improved haplotype inference by exploiting long-range linking and allelic imbalance in RNA-seq datasets. Nat Commun 2020, 11:4662
25. Li H: A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011, 27: 2987–2993
26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297–1303
27. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F: The Ensembl Variant Effect Predictor. Genome Biol 2016, 17:122
28. Sangkuhl K, Claudio-Campos K, Cavallari LH, Agundez JAG, Whirl-Carrillo M, Duconge J, Del Tredici AL, Wadelius M, Botton MR, Woodahl EL, Scott SA, Klein TE, Pratt VM, Daly AK, Gaedigk A: PharmVar GeneFocus: CYP2C9. Clin Pharmacol Ther 2021, 110:662–676
29. Botton MR, Whirl-Carrillo M, Del Tredici AL, Sangkuhl K, Cavallari LH, Agúndez JAG, Duconge J, Lee MTM, Woodahl EL, Claudio-Campos K, Daly AK, Klein TE, Pratt VM, Scott SA, Gaedigk A: PharmVar GeneFocus: CYP2C19. Clin Pharmacol Ther 2021, 109:352–366
30. Scott SA, Martis S, Peter I, Kasai Y, Kornreich R, Desnick RJ: Identification of CYP2C19*4B: pharmacogenetic implications for drug metabolism including clopidogrel responsiveness. Pharmacogenomics J 2012, 12:297–305
31. Fujikura K, Ingelman-Sundberg M, Lauschke VM: Genetic variation in the human cytochrome P450 supergene family. Pharmacogenet Genomics 2015, 25:584–594
32. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al: Global variation in copy number in the human genome. Nature 2006, 444:444–454
33. Capes-Davis A, Theodosopoulos G, Atkin I, Drexler HG, Kohara A, MacLeod RAF, Masters JR, Nakamura Y, Reid YA, Reddel RR, Freshney RI: Check your cultures! A list of cross-contaminated or misidentified cell lines. Int J Cancer 2010, 127:1–8
34. Blaisdell J, Jorge-Nebert LF, Coulter S, Ferguson SS, Lee S-J, Chanas B, Xi T, Mohrenweiser H, Ghanayem B, Goldstein JA: Discovery of new potentially defective alleles of human CYP2C9. Pharmacogenetics 2004, 14:527–537

35. Twesigomwe D, Wright GEB, Drogemoller BI, da Rocha J, Lombard Z, Hazelhurst S: A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping. NPJ Genom Med 2020, 5:30
36. Twesigomwe D, Drogemoller BI, Wright GEB, Siddiqui A, da Rocha J, Lombard Z, Hazelhurst S: StellarPGx: a Nextflow pipeline for calling star alleles in cytochrome P450 genes. Clin Pharmacol Ther 2021, 110:741−749
37. Chen X, Shen F, Gonzaludo N, Malhotra A, Rogert C, Taft RJ, Bentley DR, Eberle MA: Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. Pharmacogenomics J 2021, 21: 251−261
38. Nofziger C, Turner AJ, Sangkuhl K, Whirl-Carrillo M, Agúndez JAG, Black JL, Dunnenberger HM, Ruano G, Kennedy MA, Phillips MS, Hachad H, Klein TE, Gaedigk A: PharmVar GeneFocus: CYP2D6. Clin Pharmacol Ther 2020, 107:154−170