Automating the Interpretation of Pharmacogenetic Data

Seung-been Lee

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Deborah A. Nickerson, Chair

Evan E. Eichler

Kenneth E. Thummel

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Automating the Interpretation of Pharmacogenetic Data

Seung-been Lee

Chair of the Supervisory Committee:
Professor Deborah A. Nickerson
Department of Genome Sciences

Genetic polymorphism contributes significantly to the wide inter-individual variability in drug response, affecting both efficacy and toxicity. It has been estimated that more than 90% of the United States population has at least one clinically actionable pharmacogenetic (PGx) variant that affects their response to medication. Variation in the enzymatic activity of pharmacogenes is defined by star alleles (haplotypes) comprised of single nucleotide variants, small insertion-deletions, and large structural variants (SVs). In this dissertation, I detail the development and application of Stargazer, a novel SV-aware algorithm that can call star alleles in various polymorphic pharmacogenes from next-generation sequencing (NGS) data. When developing Stargazer, I selected the clinically important *CYP2D6* gene as a starting point because the enzyme it encodes metabolizes approximately 25% of drugs, and it is one of the most difficult genes to genotype in the human genome. To assess the performance of Stargazer, I utilized targeted sequencing data of 32 ethnically diverse trios that were genotyped for *CYP2D6* by multiple orthogonal methods. Next, I applied Stargazer to targeted sequencing data from human liver tissues (N>300) with deep phenotyping data to evaluate Stargazer's predictive power for CYP2D6 mRNA expression, protein abundance, and enzyme activity. Finally, I extended Stargazer to 28 key pharmacogenes using whole genome sequencing data from 70 reference samples that were robustly characterized by several PGx testing assays. Taken together, this

dissertation demonstrates that the combination between NGS and Stargazer offers a feasible path for accurate PGx analysis and prediction of individual drug responses. This approach will be increasingly useful in clinical practice, particularly as whole genome sequencing and targeted panel sequencing become more widely available.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my parents and sister in South Korea. Being on the opposite side of the world from your family never gets easier, and I simply would not have made it this far without them. Along the way, I have been blessed with wonderful friends whose support, mentorship, and inspiration have been priceless. I have also enjoyed the company of great people in the Nickerson lab, especially Marsha M. Wheeler who has been instrumental in my PhD career. To everyone who has shared in my journey of learning and discovery: thank you. Finally, I would like to express my deepest gratitude to my advisor, Deborah A. Nickerson. She allowed me to work on a fantastic project, provided unwavering support, and challenged me to solve difficult but important problems. I am truly proud to have been her student for the past four years.

Chapter 1 Introduction

1.1 Clinical and Economic Burden of Adverse Drug Reactions

Drug treatment is the most common form of therapy advocated for patients by physicians, and its prevalence is still rapidly growing [**1**]. By 2020, it is forecasted that global annual spending on prescription medications will reach $1.4 trillion [**2**]. However, there exists considerable inter-individual heterogeneity in drug response, affecting both efficacy and toxicity [**3**]. It has been reported that the proportion of patients who respond beneficially to the first drug offered in the treatment of a wide range of diseases is typically just 50-75% [**4**]. In the United States alone, adverse drug reactions cause 1.3 million emergency department visits and 350,000 hospitalizations each year, imposing an economic burden of at least $30.1 billion annually [**5**,**6**]. These findings indicate that inter-individual drug variability leads to patient harm and the excessive and inefficient use of limited healthcare resources.

1.2 Pharmacogenetics: Parsing Inter-individual Variability of Drug Response

Pharmacogenetics focuses on the identification of genetic variants that influence drug effects, typically through changes in pharmacokinetics (i.e., the way in which drugs move through the body during absorption, distribution, metabolism, and excretion) and pharmacodynamics (i.e., the effect that drugs have on the body) [**7**]. These pharmacogenetic (PGx) variants are often found in genes encoding drug-metabolizing enzymes, drug transporters, and drug targets, as well as disease- and treatment-modifying genes (**Table 1.1**). As of April 2019, there are 359 gene/drug relationships (N=127 for genes and N=226 for drugs) described by the Clinical Pharmacogenetics Implementation Consortium with accompanying levels of recommendation for changing drug choice and dosing decisions [**8**].

Table 1.1 Examples of genes whose polymorphisms influence drug effects through various pharmacological pathways

Gene	Medication	Pathway	Example of Altered Drug Effects
CYP2C9	Warfarin	Drug metabolism	Increased anticoagulant effects of warfarin
SLC6A4	Clozapine	Drug transport	Decreased antidepressant response
ALOX5	Leukotriene receptor antagonists	Drug target	Lower changes in forced expiratory volume
HLA	Abacavir	Treatment modification	Increased risk of hypersensitivity reaction to abacavir
MGMT	Carmustine	Disease modification	Increased response of glioma to carmustine

An important pattern to recognize when understanding pharmacogenetics is that many of the pharmacogenes are members of gene families such as cytochromes P450 (e.g., *CYP3A4*), human leukocyte antigens (e.g., *HLA-B*), solute carriers (e.g., *SLCO1B1*), and Uridine 5'-diphospho-glucuronosyltransferase (e.g., *UGT2B15*). These gene families are characterized by a few subfamilies with large numbers of genes and many subfamilies with one or fewer genes. This skewed distribution of gene family sizes can be well described by the power law, a behavior that is typical of families of paralogous genes in eukaryotes [**9**].

For example, the human genome contains a total of 57 genes encoding cytochrome P450 (CYP) enzymes that are responsible for the metabolism of 70-80% of all drugs in clinical use [**10**]. These *CYP* genes are classified into 18 subfamilies on the basis of sequence similarities, distributed in clusters over 15 chromosomes (**Table 1.2**) [**11**]. All 57 human *CYP* genes show conservation to primates, mammals, and vertebrates and many of them are thought to have arisen from the two rounds of whole genome duplications in the ancestral vertebrate, as well as other evolutionarily independent segmental duplications (**Table 1.2**) [**11**,**12**,**13**,**14**,**15**,**16**,**17**,**18**,**19**]. Notably, the *CYP* genes that have resulted from segmental duplication form four distinctive clusters CYP2ABFGST, CYP2C, CYP4ABXZ, and CYP4F on chromosomes 19, 10, 1, 19, respectively. Interestingly, when both micro- and macro-synteny were used to identify genes that coexisted near *CYP* genes in the animal ancestor, the 'cytochrome P450 genesis locus' was found [**12**]. Estimated to be more than 3 billion years old, the genesis locus contains one

progenitor *CYP* gene duplicated to create a tandem set of genes that were precursors of the 11 animal CYP clans: 2, 3, 4, 7, 19, 20, 26, 46, 51, 74, and mitochondrial.

Table 1.2 Evolution of the human cytochrome P450 gene family

No.	Subfamily	Gene	Chromosome	Evolutionary History
1		*CYP1A1*	15	Present before 2R WGD; SD
2	*CYP1*	*CYP1A2*	15	SD
3		*CYP1B1*	2	-
4		*CYP2A6*	19	SD; CYP2ABFGST cluster; 6 genes plus 7 pseudogenes
5		*CYP2A7*	19	SD; CYP2ABFGST cluster; 6 genes plus 7 pseudogenes
6		*CYP2A13*	19	SD; CYP2ABFGST cluster; 6 genes plus 7 pseudogenes
7		*CYP2B6*	19	SD; CYP2ABFGST cluster; 6 genes plus 7 pseudogenes
8		*CYP2F1*	19	SD; CYP2ABFGST cluster; 6 genes plus 7 pseudogenes
9		*CYP2S1*	19	SD; CYP2ABFGST cluster; 6 genes plus 7 pseudogenes
10		*CYP2C8*	10	SD; CYP2C cluster
11	*CYP2*	*CYP2C9*	10	SD; CYP2C cluster
12		*CYP2C18*	10	SD; CYP2C cluster
13		*CYP2C19*	10	SD; CYP2C cluster
14		*CYP2D6*	22	2R WGD from a CYP3 ancestor
15		*CYP2E1*	10	-
16		*CYP2J2*	1	-
17		*CYP2R1*	11	-
18		*CYP2U1*	4	-
19		*CYP2W1*	7	2R WGD from a CYP3 ancestor
20		*CYP3A4*	7	Present before 2R WGD
21	*CYP3*	*CYP3A5*	7	Present before 2R WGD
22		*CYP3A7*	7	Present before 2R WGD
23		*CYP3A43*	7	Present before 2R WGD
24		*CYP4A11*	1	SD; CYP4ABXZ cluster
25		*CYP4A22*	1	SD; CYP4ABXZ cluster
26		*CYP4X1*	1	SD; CYP4ABXZ cluster
27		*CYP4Z1*	1	SD; CYP4ABXZ cluster
28		*CYP4B1*	1	2R WGD from a CYP4 ancestor; CYP4ABXZ cluster
29	*CYP4*	*CYP4F2*	19	SD; CYP4F cluster
30		*CYP4F3*	19	SD; CYP4F cluster
31		*CYP4F8*	19	SD; CYP4F cluster
32		*CYP4F11*	19	SD; CYP4F cluster
33		*CYP4F12*	19	SD; CYP4F cluster
34		*CYP4F22*	19	2R WGD from a CYP4 ancestor; CYP4F cluster
35		*CYP4V2*	4	The oldest CYP4 gene
36	*CYP5*	*CYP5A1*	7	-
37	*CYP7*	*CYP7A1*	8	Present before 2R WGD
38		*CYP7B1*	8	-
39	*CYP8*	*CYP8A1*	20	Originated between lancelet and jawless fishes
40		*CYP8B1*	3	Originated between lancelet and jawless fishes; 2R WGD from a CYP7A ancestor
41		*CYP11A1*	15	Present before 2R WGD
42	*CYP11*	*CYP11B1*	8	2R WGD from CYP11A1
43		*CYP11B2*	8	2R WGD from CYP11A1
44	*CYP17*	*CYP17A1*	10	Present before 2R WGD
45	*CYP19*	*CYP19A1*	15	Present before 2R WGD; a possible member of the animal CYP genesis locus
46	*CYP20*	*CYP20A1*	2	2R WGD from a CYP20 ancestor
47	*CYP21*	*CYP21A2*	6	SD; one gene plus one pseudogene; in the HLA locus
48	*CYP24*	*CYP24A1*	20	Originated from an mitochondrial CYP ancestor
49		*CYP26A1*	10	2R WGD from a CYP26B ancestor
50	*CYP26*	*CYP26B1*	2	Present before 2R WGD; a possible member of the animal CYP genesis locus
51		*CYP26C1*	10	2R WGD from a CYP26B ancestor
52		*CYP27A1*	2	2R WGD from a CYP27B ancestor
53	*CYP27*	*CYP27B1*	12	Present before 2R WGD; originated from an mitochondrial CYP ancestor
54		*CYP27C1*	2	2R WGD from a CYP27B ancestor
55	*CYP39*	*CYP39A1*	6	A possible member of the animal CYP genesis locus
56	*CYP46*	*CYP46A1*	14	-
57	*CYP51*	*CYP51A1*	7	Originated from a bacterial ancestor

WGD, whole genome duplication; SD, segmental duplication.

Another key characteristic of pharmacogenetics is that many of the pharmacogenes are genetically complex and highly polymorphic at the population level, which has important implications for personalized drug therapy and healthcare programs. For example, many of the

CYP genes mentioned above harbor a large repertoire of single nucleotide variants (SNVs), small insertion-deletions (indels), and/or large structural variants (SVs). These variants provide a wide spectrum of CYP enzymatic activity ranging from poor to ultrarapid metabolism. This functional diversity is thought to be based on low evolutionary constraints due to the lack of essential endogenous function of the encoded gene products and genetic drift [**20**]. Importantly, the distribution of *CYP* alleles differs considerably across the globe (**Figure 1.1**), which is not surprising given that throughout the history of mankind, CYP enzymes have detoxified many kinds of compounds in the diet of human populations living in different environments. There is also some evidence that suggests that several *CYP* alleles have been under selective pressures in human evolution [**21**,**22**,**23**,**24**,**25**].

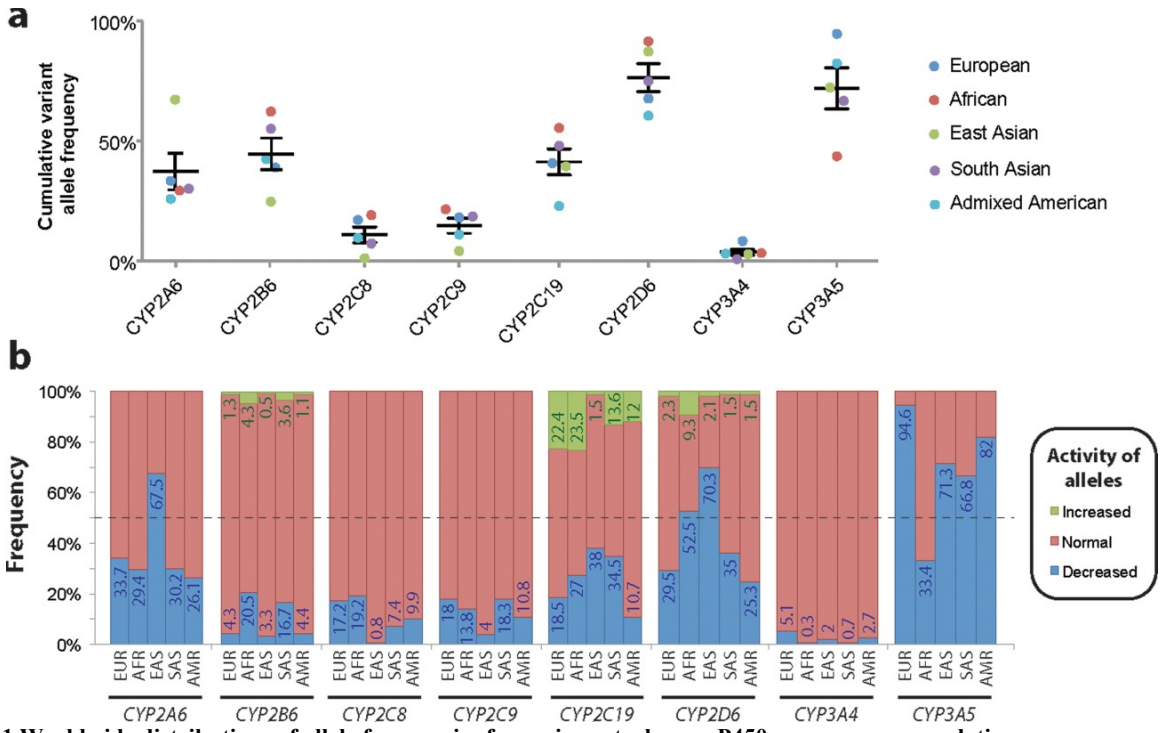


Figure 1.1 Worldwide distributions of allele frequencies for major cytochrome P450 genes across populations
(A) Cumulative frequencies of all major variant alleles in Europeans (EUR), Africans (AFR), East Asians (EAS), South Asians (SAS), and admixed Americans (AMR) are shown for each major *CYP* gene. Overall, *CYP2D6* constitutes the most variable gene, whereas *CYP3A4* is the most conserved. (B) The expected functional consequences of allelic distributions across world populations are shown. This figure is reproduced from Figure 2 in [**20**].

The history of pharmacogenetics stretches as far back as 510 BC when the Greek philosopher Pythagoras noticed that ingestion of fava beans caused a fatal reaction in some, but not all, individuals [**26**]. Over two thousand years later, in 1956, Paul E. Carson and his colleagues attributed this reaction to an inherited deficiency of an enzyme called glucose-6-phosphate dehydrogenase or G6PD [**27**]. In 1957, Arno G. Motulsky further refined the concept that inherited defects of metabolism may explain individual differences in drug response [**28**]. In 1959, Friedrich Vogel was the first to use the term 'pharmacogenetics' [**29**,**30**]. Between the 1960s and 1980s, multiple landmark studies were conducted to document patterns of inheritance in twins or families for many drug effects, which eventually led to molecular studies that revealed the inherited determinants for many of the traits [**30**]. For example, in 1968, Elliott S. Vesell and John G. Page showed that the pharmacokinetic profile of the pain-relieving drug antipyrine was more similar in monozygotic twins than in dizygotic twins [**31**]. In 1987, *CYP2D6* became the first polymorphic human drug-metabolizing gene to be cloned and characterized when Frank J. Gonzalez and his colleagues were investigating the cause of poor metabolism phenotype for the antihypertensive drug debrisoquine [**32**]. By the 1990s, the potential clinical utility of pharmacogenetics was clear [**33**].

As in most areas of genetics, the rate of PGx discovery has been accelerated by the Human Genome Project [**34**]—which was completed in 2003—and by improved technologies for the genome-wide interrogation of variation. This shortened the timeline for discovery and enabled genome-wide studies of populations of patients with specific drug-related phenotypes, often leading to the identification of unanticipated genetic variants that were statistically associated with drug effects [**35**]. This convergence of pharmacogenetics and human genomics in recent years also brought the term 'pharmacogenomics' into the pharmacology lexicon [**36**]. The

distinction between pharmacogenetics and pharmacogenomics is arbitrary, however, and both terms can be used interchangeably [**37**].

1.3 The Opportunities and Challenges of Pharmacogenetic Testing in the Clinic

It has been estimated that more than 90% of the United States population has at least one clinically actionable PGx variant that affects their response to medication [**38**]. Hence, PGx tests offer great opportunities for personalized medicine, by combining genetic information and corresponding phenotypes. PGx tests are more likely to be useful in treatment with those drugs that have a narrow therapeutic window (**Figure 1.2**). For example, the high efficacy of the anticoagulant drug warfarin has been challenged by the high risk of adverse drug reactions (e.g., spontaneous bleeding events) due to its narrow therapeutic window, requiring careful monitoring and strict compliance to balance efficacy and toxicity [**39**]. In the early 2000s, polymorphisms in two enzymes were reported to be associated with risk of adverse drug reactions of warfarin: cytochrome P450 2C9 (CYP2C9) and vitamin K epoxide reductase complex subunit 1 (VKORC1) [**40**,**41**]. In 2010, a prospective study demonstrated a reduction in hospitalizations when starting doses of warfarin were informed by patient genotypes for *VKORC1* and *CYP2C9* [**42**]. The Food and Drug Administration provides additional guidance for warfarin and other medications, by requiring that applicable PGx test information be included in drug labeling [**43**]. However, broad implementation of PGx testing has met several challenges, and only a few tests are currently routinely used in the clinic [**44**].

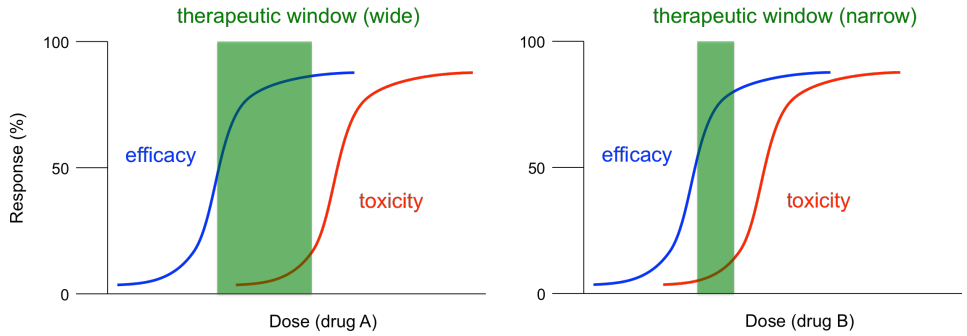


Figure 1.2 Illustration of dose curves of drug response
A dose-dependent increase in drug therapeutic effect and toxicity is shown. Pharmacogenetic tests are more likely to be useful in treatment with those drugs with a narrow therapeutic window (i.e., drug B).

A major barrier has been the complexity of many pharmacogenes. Several genes require that PGx testing include a large number of genetic variants to provide accurate predictions of enzymatic activity [**45**,**46**]. For example, the clinically important *CYP2D6* gene has more than 100 haplotype patterns known as 'star alleles' that are defined by tagging variants (SNVs, indels, SVs, or a combination of these) [**47**]. These *CYP2D6* alleles encode enzymes with normal, decreased, increased, or no function, which translate to inferred clinical phenotypes that range from ultrarapid to poor metabolism [**48**]. Importantly, the frequency of star alleles and phenotypes can vary across different populations (**Figure 1.3**) [**49**], highlighting the need for comprehensive variant testing.

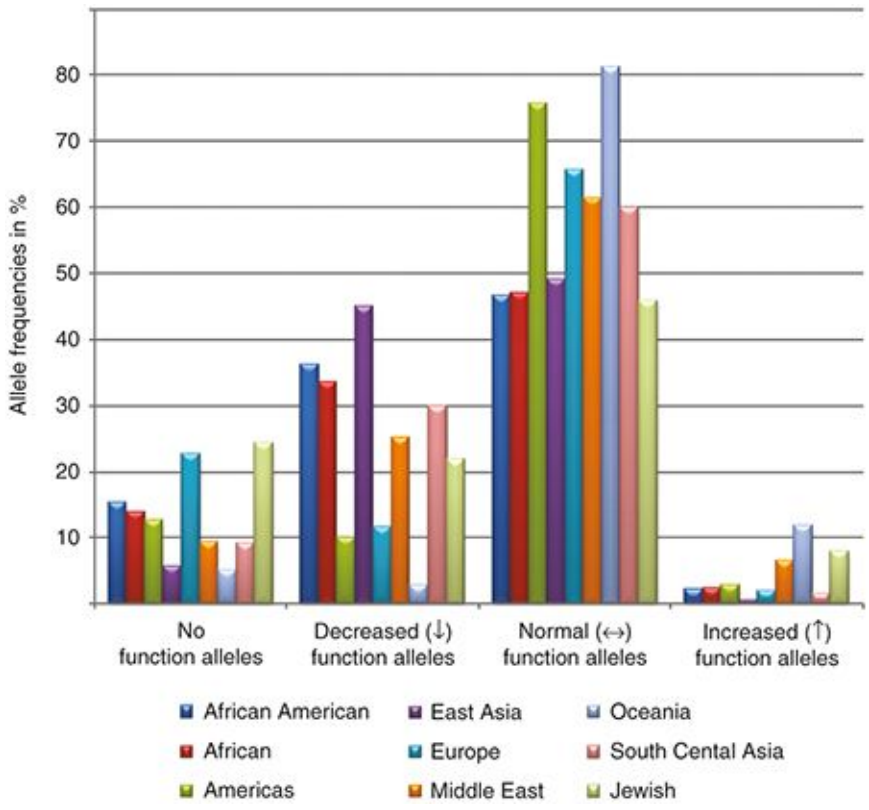


Figure 1.3 *CYP2D6* allele frequencies across world populations
This figure is reproduced from Figure 2 in [**49**].

Another major challenge has been that a large fraction of existing star alleles cannot be accurately assessed with a single methodology. As stated above, *CYP2D6* alleles include SVs such as deletions, duplications, and complex gene hybrids. Many of these structural variants are difficult to detect due to high sequence homology (>95%) with a nearby non-functional paralog called *CYP2D7* (**Figure 1.4**) [**50**]. Thus, several orthogonal genotyping methods including TaqMan assays, long-range polymerase chain reaction, quantitative multiplex polymerase chain reaction, High Resolution Melt analysis, and Sanger sequencing are required to accurately call all SVs in *CYP2D6* [**51**,**52**]. These methods do reliably detect the star alleles needed for clinical application, but can be time consuming and biased toward the detection of known SVs. Hence, a new approach for PGx genotyping is needed that is more robust and capable of higher throughput.

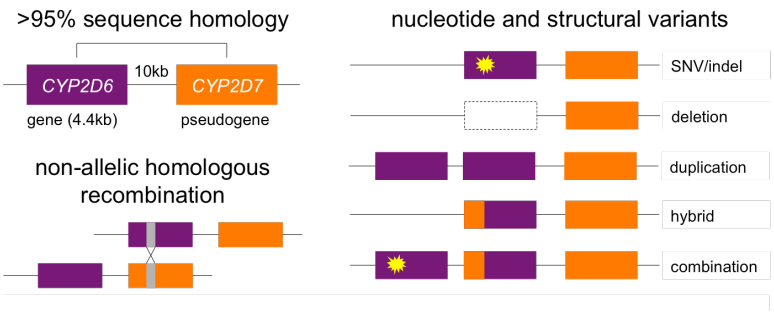


Figure 1.4 Complex *CYP2D6* genomics
CYP2D6 structural variants (deletions, duplications, and hybrids) are difficult to detect due to high sequence homology with a nearby non-functional paralog known as *CYP2D7*. These variants are thought to be caused primarily by non-allelic homologous recombination as shown above.

1.4 Next-generation Sequencing as a Pharmacogenetic Genotyping Platform

Next-generation sequencing (NGS), also known as massively parallel sequencing, has demonstrated the capacity to sequence DNA at an unprecedented scale, thereby enabling previously unattainable scientific innovations and novel biological discoveries [**53**]. Over the past decade, NGS technologies have continued to advance—increasing capacity by a factor of 100 to 1,000—and have brought the cost to sequence a human genome down to ~$1,000 [**54**]. Several sequencing platforms have been developed to date, but Illumina's short-read sequencing system (50 to 300 base pairs) accounts for the largest market share worldwide for sequencing instruments compared to other platforms [**54**]. In this approach, template DNA is amplified to form clonal clusters on a solid surface and many millions of DNA molecules are sequenced simultaneously. Throughout this dissertation, unless otherwise stated, NGS is assumed to refer to this Illumina platform.

NGS is a powerful platform for variant detection because of its high-throughput data generation, comprehensive genotyping capabilities, and ever-decreasing cost [**55**]. Furthermore, cost-effectiveness can be increased for variant discovery by applying custom capture panels. Until recently, the identification of PGx alleles from NGS data has involved the 'manual' assignment of haplotypes by comparing the variants found in a sample to those listed in PGx

haplotype definition tables. However, the large amount of data obtained via NGS makes this approach cumbersome and prone to error, highlighting the need for novel bioinformatics tools. In addition, NGS-based PGx analysis has been limited to SNVs and indels due to the lack of algorithms that can reliably call SVs from NGS data, particularly in genes with one or more paralogs. Stargazer, a new bioinformatics tool developed during my dissertation research meets these needs. Stargazer assesses star alleles from NGS data by combining SNV/indel calls with SV data to improve the accuracy of diplotype calls and automates the process of calling complex star allele [**56**].

1.5 Dissertation Aims

In this dissertation, I detail the Stargazer algorithm and its utility. Chapter 2 explains how Stargazer identifies star alleles from NGS data using the clinically important *CYP2D6* gene as a model. When developing Stargazer, I purposely chose *CYP2D6* as a starting point because it is one of the most difficult genes to genotype among the clinically important pharmacogenes. Additionally, chapter 2 highlights Stargazer's use of paralog-specific copy number to reliably detect complex SVs including *CYP2D6/CYP2D7* hybrids. Finally, chapter 2 shows that genotyping by Stargazer was 99.0% concordant with data obtained by multiple orthogonal methods.

In Chapter 3, I evaluate the predictive power of Stargazer for various CYP2D6 phenotypes in human liver tissues. The work described in this chapter was a collaborative project, which involved targeted sequencing to uncover genetic variants, RNAseq for estimating mRNA expression, mass spectrometry for measuring protein content, and probe drugs for assessing enzymatic activity. Our results indicate that the accuracy of CYP2D6 activity prediction was significantly improved using Stargazer that included SV data compared to manual genotype

assignment by SNV/indel data alone. Chapter 3 also describes the identification and functional characterization of novel coding variants, one of which, A449D, exhibited significantly decreased catalytic activity.

Chapter 4 presents the utility of extending Stargazer to call star alleles for 28 pharmacogenes. For validation, I applied Stargazer to WGS data from 70 ethnically diverse reference samples from the Genetic Testing Reference Materials Coordination Program (GeT-RM). These samples were extensively characterized by GeT-RM using multiple PGx testing assays. In all 28 genes, Stargazer recalled 100% of star alleles present in GeT-RM's consensus genotypes. Stargazer also detected star alleles not previously reported by GeT-RM including complex SVs.

Chapter 5 reflects on lessons learned from my studies of pharmacogenomics, presents preliminary data on the application of Stargazer to large-scale and ethnically diverse sample sets, and offers a perspective on future development of Stargazer. Supplementary materials for Chapters 2-4 are provided in Appendices A-C. Ultimately, this work demonstrates that combining NGS data and Stargazer offers a feasible path for accurate PGx analysis and prediction of individual drug responses.

Chapter 2 Stargazer: A Software Tool for Calling Star Alleles from Next-generation Sequencing Data Using *CYP2D6* as a Model

This chapter has been published: Lee, S.B., Wheeler, M.M., Patterson, K., McGee, S., Dalton, R., Woodahl, E.L., Gaedigk, A., Thummel, K.E. & Nickerson, D.A. *Genet. Med.* **21**, 361-372 (2019).

2.1 Summary

Genotyping *CYP2D6* is important for precision drug therapy because the enzyme it encodes metabolizes approximately 25% of drugs, and its activity varies considerably among individuals. Genotype analysis of *CYP2D6* is challenging due to its highly polymorphic nature. Over 100 haplotypes (star alleles) have been defined for *CYP2D6*, some involving a gene hybrid with its nearby nonfunctional but highly homologous paralog *CYP2D7*. We present Stargazer, a new bioinformatics tool that uses next-generation sequencing (NGS) data to call star alleles for *CYP2D6* (https://stargazer.gs.washington.edu/stargazerweb/). Stargazer is currently being extended for other pharmacogenes. Stargazer identifies star alleles from NGS data by detecting single nucleotide variants, insertion-deletion variants, and structural variants. Stargazer detects structural variation, including gene deletions, duplications, and hybrids, by calculating paralog-specific copy numbers from read depths. We applied Stargazer to the NGS data of 32 ethnically diverse HapMap trios that were genotyped by TaqMan assays, long-range polymerase chain reaction, quantitative multiplex polymerase chain reaction, high-resolution melting analysis, and/or Sanger sequencing. *CYP2D6* genotyping by Stargazer was 99.0% concordant with the data obtained by these methods, and showed that 28.1% of the samples had structural variation including *CYP2D6/CYP2D7* hybrids. Accurate genotyping of pharmacogenes with NGS and subsequent allele calling with Stargazer will aid the implementation of precision drug therapy.

2.2 Introduction

Many cytochrome P450 enzymes play a role in pharmacological responses by contributing to the metabolism of numerous drugs. Among these, cytochrome P450 2D6 (CYP2D6) is considered one of the most important because it contributes to the metabolism of about 25% of drugs [**47**]. Drugs metabolized by CYP2D6 include opioids, chemotherapeutic agents, antidepressants, and antipsychotics, among others [**57**].

The activity of CYP2D6 varies considerably between individuals due to the high level of polymorphisms observed in the *CYP2D6* gene. There are more than 100 haplotypes defined for *CYP2D6* by the Pharmacogene Variation Consortium [**58**]. These are called star alleles (e.g., *CYP2D6*1*, **2*, etc.) and are characterized by single nucleotide variants (SNVs), insertion-deletion variants (indels), structural variants (SVs), or a combination of these. They include fully functional, decreased-function, and nonfunctional alleles, which provide a wide spectrum of CYP2D6 enzymatic activity ranging from ultrarapid to poor metabolism. Different ethnic groups have distinct frequencies of star alleles and metabolic phenotypes [**49**]; however, further studies are warranted for individuals of African or Asian ancestry because these populations are underrepresented in the estimation of *CYP2D6* genetic diversity [**59**].

Drug therapy without preemptive knowledge of a patient's CYP2D6 phenotype status can lead to severe adverse reactions or a loss of efficacy due to inappropriate drug choice and/or dosing. For example, codeine is one of the most common and widely used opioids whose analgesic effect is elicited by CYP2D6 through the formation of morphine. Patients who are CYP2D6 poor metabolizers exhibit very low plasma concentrations of morphine following codeine administration, which can compromise the efficacy of pain management because the affinity of morphine to the μ-opioid receptor is 200-fold stronger compared with that of codeine [**60**]. Conversely, a patient can experience life-threatening morphine intoxication after receiving

a small dose of codeine if they are a CYP2D6 ultrarapid metabolizer [**61**]. Anecdotally, a breastfed infant died from morphine poisoning because the mother was prescribed a normal dose of codeine for childbirth-related pain. In this case, the mother was a CYP2D6 ultrarapid metabolizer passing what was presumed to be a toxic amount of morphine to her newborn through her breast milk [**62**].

For these reasons, there is considerable interest in genotyping *CYP2D6*. However, genotype analysis of *CYP2D6* is complex because a large fraction of its existing variation cannot be accurately assessed with a single approach. SVs in *CYP2D6*, such as gene deletions, duplications, and hybrids, are particularly challenging to detect due to high sequence homology (>95%) with a nonfunctional paralog, *CYP2D7*, located upstream of *CYP2D6* [**50**]. Therefore, *CYP2D6* is prone to genotype misclassification and incorrect phenotype prediction [**63**]. In laboratory settings, several orthogonal genotyping methods, such as TaqMan assays, long-range polymerase chain reaction (PCR), quantitative multiplex PCR, high-resolution melting analysis, and Sanger sequencing, are employed to call star alleles. However, many of these methods are time consuming and heavily biased toward the detection of known variants. In clinical settings, due to practical limitations, only a handful of major star alleles, if any, are tested and rarely involve SVs. Hence, a new approach for genotyping *CYP2D6* is needed that is more robust and capable of higher throughput.

In this study, we developed Stargazer, a new bioinformatics tool for calling star alleles in *CYP2D6* from next-generation sequencing (NGS) data. NGS is a powerful platform for variant detection because of its high-throughput data generation, comprehensive genotyping capabilities, and ever-decreasing cost. Additionally, NGS does not require previous knowledge about the variants of interest, and can uncover novel functional variants, which is not possible for many of

the aforementioned genotyping methods. Furthermore, its cost-effectiveness can be increased for variant discovery by applying custom capture panels. To assess the accuracy of Stargazer, we applied it to the NGS data of 32 ethnically diverse HapMap trios. We report a correlation of 99.0% between *CYP2D6* genotype calls determined with Stargazer and by orthogonal methods. We are now extending Stargazer to call star alleles for other clinically important pharmacogenes. Accurate diplotype calls from NGS data using Stargazer provide a promising approach for precision medicine to maximize drug efficacy and minimize toxicity for individual patients. Stargazer is publicly accessible through https://stargazer.gs.washington.edu/stargazerweb/.

2.3 Materials and Methods

Samples

We built Stargazer using NGS data from 32 ethnically diverse trios. These trios were selected from the International HapMap Project, and they are comprised of 13 European, 5 Yoruban, 4 African American, 3 Han Chinese, 3 Mexican American, 2 Peruvian, and 2 Puerto Rican families (**Table 2.1**). These trios were originally sequenced to assess the performance of PGRNseq, a recently developed custom capture panel of key pharmacogenes including *CYP2D6* [**64**]. They were specifically chosen for this study because they are a genetically diverse set of samples in which we would likely encounter a wide range of *CYP2D6* variants, including SVs, to test Stargazer's genotyping abilities and limitations. In addition, these trios allow for the analysis of Mendelian inheritance patterns to further the validation of Stargazer's star allele calls. These trios were also previously genotyped for *CYP2D6* by a variety of orthogonal methods (see below), allowing us to assess the accuracy of Stargazer's diplotype calls.

Orthogonal genotyping methods

HapMap trios were genotyped for *CYP2D6* according to procedures described elsewhere [**51**,**65**,**66**,**67**]. Briefly, SNVs and indels were detected using TaqMan assays. Gene deletions, duplications, and multiplications were assessed by long-range PCR and quantitative multiplex PCR. *CYP2D6/CYP2D7* hybrids were identified using quantitative multiplex PCR, high-resolution melting analysis, and/or Sanger sequencing.

Custom capture panel and NGS

HapMap trios were sequenced twice—once with PGRNseq v1.1 and once with PGRNseq v2.0—to a mean coverage of ~400x and ~160x, respectively. Both sequencing runs were performed with Illumina HiSeq 2500 machines using 100-base pair (bp) paired-end reads. Three samples failed during one of the two sequencing runs: NA19835 in PGRNseq v1.1, and NA19686 and NA11834 in PGRNseq v2.0. Note that the probes designed to capture *CYP2D6* and *CYP2D7* were more specific and extensive in PGRNseq v2.0 compared with PGRNseq v1.1; however, both versions generated reads that mapped to all the exons, introns, untranslated regions, and promoters of *CYP2D6* and *CYP2D7* (**Supplementary Figure 2.1**). Two samples, NA12878 and NA19238, were also previously sequenced by genome sequencing (WGS) to a mean coverage of ~30x with Illumina HiSeq X instruments using 150-bp paired-end reads. These data were used to test Stargazer's generalizability to WGS data.

Input and output data of Stargazer

The Stargazer *CYP2D6* genotyping pipeline is outlined in **Figure 2.1**. The pipeline uses BAM files comprising sequence reads aligned with BWA-MEM [**68**] to human reference genome assembly GRCh37. BAM files are then used to generate a VCF file with GATK-HaplotypeCaller (v3.4) [**69**], from which Stargazer extracts all SNVs and indels located within 3 kilobases (kb)

from either end of *CYP2D6*. More specifically, Stargazer stores the genomic position of each variant, reference allele, alternate allele(s), genotype status (homozygous or heterozygous), and allelic depth for each sample. Stargazer uses the variant information from the VCF file to call star alleles based on SNVs and indels.

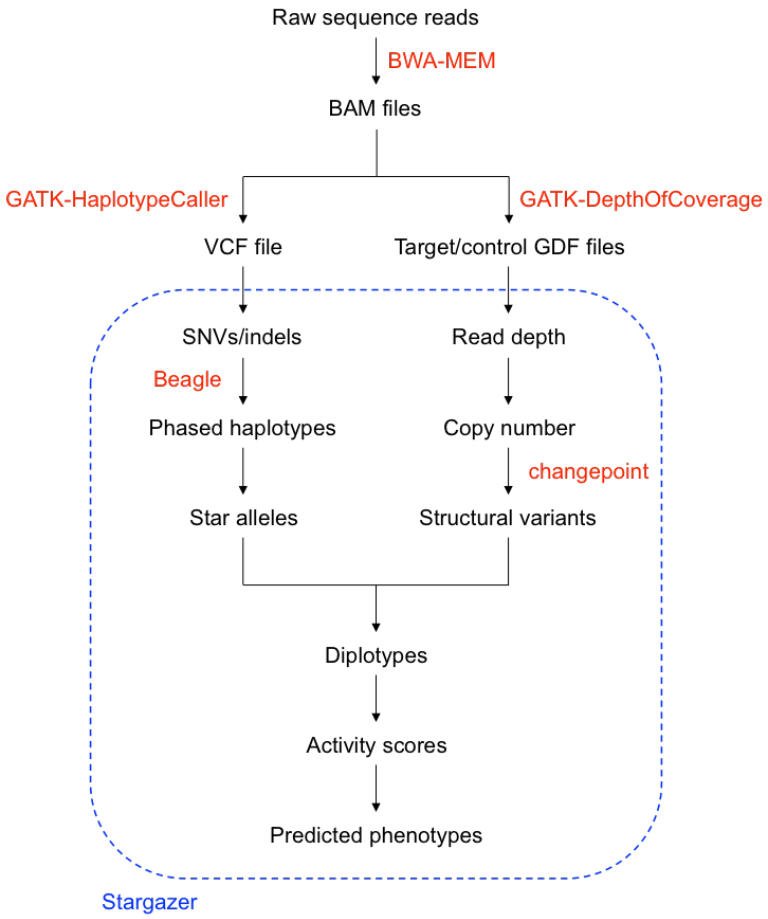


Figure 2.1 Schematic diagram of the Stargazer genotyping pipeline
Stargazer takes as input a VCF file, a target GDF file, and a control GDF file. Stargazer uses the variant information from the VCF file to call star alleles based on SNVs/indels. Using the target and control GDF files, Stargazer converts read depth to copy number for detection of structural variation. The output data of Stargazer include each sample's *CYP2D6* diplotype and plots to visually inspect copy number for *CYP2D6* and *CYP2D7*. Based on called *CYP2D6* diplotypes, the program outputs predicted phenotypes as well. Several external software tools, shown in red, are used within and outside of Stargazer.

BAM files are also used to calculate read depth for *CYP2D6* and *CYP2D7* with GATK-DepthOfCoverage (v3.4) [**69**]. For convenience, we will refer to this output as a target GDF (GATK-DepthOfCoverage format) file. Since the high homology between *CYP2D6* and *CYP2D7* can cause reads to align to erroneous or multiple locations, only uniquely mapping

reads with a mapping quality ≥20 are counted. Similarly, a control GDF file is produced from a user-chosen locus, which serves as a read depth normalization factor. Stargazer computes paralog-specific copy numbers using read depth from the target and control GDF files in order to detect SVs.

In the initial development of Stargazer, three genes—*VDR*, *RYR1*, and *EGFR*—were evaluated as control loci. These genes are covered by PGRNseq and are 63, 154, and 188 kb in size, respectively. They are also reported to exhibit low rates of whole gene deletion and/or duplication according to the Database of Genomic Variants [**70**]. All three genes produced the same copy number results for *CYP2D6* and *CYP2D7*. The analyses shown in the results section were all performed using *RYR1* as the control locus.

The output data of Stargazer include each sample's *CYP2D6* diplotype, predicted phenotype, and plots to visually inspect copy number for *CYP2D6* and *CYP2D7* (**Figure 2.2**). Note that when calling diplotypes, Stargazer only considers those variants that are currently used by the Pharmacogene Variation Consortium. Stargazer also returns all detected SNVs and indels, including those that are novel and those that are known but not currently used to define any star allele. As follow-up, these variants can be functionally annotated using variant annotation tools such as SeattleSeq Annotation (http://snp.gs.washington.edu/SeattleSeqAnnotation).

Prediction of star alleles

From a VCF file, Stargazer uses Beagle (v4.1) [**71**] to haplotype phase heterozygous variants for *CYP2D6* with over 2,500 reference samples from the 1000 Genomes Project. Stargazer then matches phased haplotypes to star alleles using a translation table built from publicly available data (https:// www.pharmvar.org). The table contains information on more than 90 star alleles

and 185 SNVs and indels, including variant positions and nucleotide changes in relation to the reference *CYP2D6*1* allele and human reference genome assembly GRCh37.

Detection of SVs

From a target GDF file, Stargazer converts read depths for *CYP2D6* and *CYP2D7* to copy numbers by performing intra- and intersample normalizations. Intrasample normalization accounts for individual variation in the sequencing efficiency using read depth from a control GDF file, while the intersample normalization considers the heterogeneity in coverage across all samples. Stargazer then automates the detection of SVs with changepoint (v2.2.2)—an R package that approximates one or more points at which the statistical properties of a sequence of observations change [**72**]. Here, the sequence is DNA, the observation is per-base copy number, and the statistical property is the mean copy number. If there is a significant shift in the mean copy number (e.g., from two to one), the algorithm returns the change point location and the two mean values (e.g., two and one).

Identification of diplotypes

For samples without SVs, Stargazer determines *CYP2D6* diplotypes by combining the star allele used to assign each phased haplotype. For samples with a whole gene deletion, the affected haplotype is assigned the *CYP2D6*5* deletion allele, which is then combined with the star allele assigned to the other haplotype to form a diplotype. For samples with a whole gene duplication, the affected haplotype is assigned "x2" (e.g., *CYP2D6*1x2*, **2x2*, etc.) because it has two gene copies of *CYP2D6*. For samples with more complex SVs, such as *CYP2D6/CYP2D7* hybrids, individual algorithms have been developed to determine diplotypes. The identification of diplotypes is discussed in more detail, in the context of HapMap trios, in the results section.

Assignment of predicted phenotypes

There are four CYP2D6 metabolizer classes: poor, intermediate, normal, and ultrarapid. To predict these phenotypes, Stargazer first translates *CYP2D6* diplotypes into a standard unit of enzyme activity known as an activity score [**48**]. The fully functional reference *CYP2D6*1* allele is assigned a value of 1, decreased-function alleles such as *CYP2D6*10* and **17* receive a value of 0.5, and nonfunctional alleles including *CYP2D6*4* and **5* have a value of 0. The sum of values assigned to both alleles constitutes the activity score of a diplotype. Consequently, subjects with *CYP2D6*1/*1*, **1/*4*, and **4/*5* diplotypes have an activity score of 2, 1, and 0, respectively. These activity scores are used to predict the four metabolizer classes as follows: poor, 0; intermediate, 0.5; normal, 1–2; and ultrarapid, >2.

2.4 Results

Identification of diplotypes by Stargazer for HapMap trios

We used Stargazer (v1.0.0) to call *CYP2D6* diplotypes for 32 HapMap trios sequenced with PGRNseq (**Table 2.1**). Data from PGRNseq v1.1 and PGRNseq v2.0 served as technical validation and produced the same diplotype calls for all samples. Moreover, all diplotype calls were inherited in a predictable manner, as exemplified in **Figure 2.3**. Diplotypes were identified using the following algorithms.

Table 2.1 Comparison of *CYP2D6* genotype calls for 32 HapMap trios between orthogonal methods and Stargazer using PGRNseq v1.1 or v2.0 data

Three samples failed during one of the two sequencing runs ('-'). When the *CYP2D6* haplotype calls of Stargazer were compared to those determined with the orthogonal methods, the concordance rate was 99.0% (190 out of 192 haplotypes). The two discordant haplotypes were found in samples NA19200 and NA19202 of the Y045 trio ('[]'). Predicted phenotypes were assigned based on the activity score and Clinical Pharmacogenetics Implementation Consortium guidelines.

No.	Sample	Ethnicity	Family	Relation	Orthogonal Methods	PGRNseq v1.1	PGRNseq v2.0	Activity Score	Phenotype
1	NA12801	European	1454	Father	**4/*6*	**4/*6*	**4/*6*	0	poor
2	NA12802	European	1454	Mother	**2/*41*	**2/*41*	**2/*41*	1.5	normal
3	NA12805	European	1454	Child	**2/*4*	**2/*4*	**2/*4*	1	normal
4	NA12891	European	1463	Father	**41/*68+*4*	**41/*68+*4*	**41/*68+*4*	0.5	intermediate
5	NA12892	European	1463	Mother	**2/*3*	**2/*3*	**2/*3*	1	normal
6	NA12878	European	1463	Child	**3/*68+*4*	**3/*68+*4*	**3/*68+*4*	0	poor
7	NA19834	African American	2424	Father	**2/*2*	**2/*2*	**2/*2*	2	normal
8	NA19835	African American	2424	Mother	**1/*2*	-	**1/*2*	2	normal

9	NA19836	African American	2424	Child	*1/*2	*1/*2	*1/*2	2	normal
10	NA19239	Yoruban	Y117	Father	*15/*17	*15/*17	*15/*17	0.5	intermediate
11	NA19238	Yoruban	Y117	Mother	*1/*17	*1/*17	*1/*17	1.5	normal
12	NA19240	Yoruban	Y117	Child	*15/*17	*15/*17	*15/*17	0.5	intermediate
13	NA12750	European	1444	Father	*2/*2	*2/*2	*2/*2	2	normal
14	NA12751	European	1444	Mother	*1/*2	*1/*2	*1/*2	2	normal
15	NA12740	European	1444	Child	*1/*2	*1/*2	*1/*2	2	normal
16	NA19685	Mexican American	M011	Father	*1/*2x2	*1/*2x2	*1/*2x2	3	ultrarapid
17	NA19684	Mexican American	M011	Mother	*1/*4	*1/*4	*1/*4	1	normal
18	NA19686	Mexican American	M011	Child	*1/*1	*1/*1	-	2	normal
19	HG00421	Han Chinese	SH007	Father	*2/*10x2	*2/*10x2	*2/*10x2	2	normal
20	HG00422	Han Chinese	SH007	Mother	*2/*10	*2/*10	*2/*10	1.5	normal
21	HG00423	Han Chinese	SH007	Child	*10/*10x2	*10/*10x2	*10/*10x2	1.5	normal
22	HG01979	Peruvian	PEL027	Father	*2/*68+*4	*2/*68+*4	*2/*68+*4	1	normal
23	HG01980	Peruvian	PEL027	Mother	*1/*2	*1/*2	*1/*2	2	normal
24	HG01981	Peruvian	PEL027	Child	*1/*2	*1/*2	*1/*2	2	normal
25	NA12003	European	1420	Father	*4/*35	*4/*35	*4/*35	1	normal
26	NA12004	European	1420	Mother	*2/*41	*2/*41	*2/*41	1.5	normal
27	NA10838	European	1420	Child	*2/*4	*2/*4	*2/*4	1	normal
28	NA12155	European	1408	Father	*1/*5	*1/*5	*1/*5	1	normal
29	NA12156	European	1408	Mother	*1/*4	*1/*4	*1/*4	1	normal
30	NA10831	European	1408	Child	*4/*5	*4/*5	*4/*5	0	poor
31	NA19128	Yoruban	Y077	Father	*17/*17	*17/*17	*17/*17	1	normal
32	NA19127	Yoruban	Y077	Mother	*2/*17	*2/*17	*2/*17	1.5	normal
33	NA19129	Yoruban	Y077	Child	*17/*17	*17/*17	*17/*17	1	normal
34	NA19700	African American	2367	Father	*4/*29	*4/*29	*4/*29	0.5	intermediate
35	NA19701	African American	2367	Mother	*1/*17	*1/*17	*1/*17	1.5	normal
36	NA19702	African American	2367	Child	*4/*17	*4/*17	*4/*17	0.5	intermediate
37	NA19771	Mexican American	M031	Father	*2/*4	*2/*4	*2/*4	1	normal
38	NA19770	Mexican American	M031	Mother	*1/*2	*1/*2	*1/*2	2	normal
39	NA19772	Mexican American	M031	Child	*2/*4	*2/*4	*2/*4	1	normal
40	HG01060	Puerto Rican	PR14	Father	*1/*41	*1/*41	*1/*41	1.5	normal
41	HG01061	Puerto Rican	PR14	Mother	*1/*4	*1/*4	*1/*4	1	normal
42	HG01062	Puerto Rican	PR14	Child	*1/*4	*1/*4	*1/*4	1	normal
43	NA10860	European	1362	Father	*1/*4N+4	*1/*4N+4	*1/*4N+4	1	normal
44	NA10861	European	1362	Mother	*4/*35	*4/*35	*4/*35	1	normal
45	NA11984	European	1362	Child	*1/*35	*1/*35	*1/*35	2	normal
46	HG02259	Peruvian	PEL042	Father	*1/*2	*1/*2	*1/*2	2	normal
47	HG02260	Peruvian	PEL042	Mother	*1/*1	*1/*1	*1/*1	2	normal
48	HG02261	Peruvian	PEL042	Child	*1/*2	*1/*2	*1/*2	2	normal
49	NA07357	European	1345	Father	*1/*6	*1/*6	*1/*6	1	normal
50	NA07345	European	1345	Mother	*1/*1	*1/*1	*1/*1	2	normal
51	NA07348	European	1345	Child	*1/*6	*1/*6	*1/*6	1	normal
52	NA06984	European	1328	Father	*4/*68+*4	*4/*68+*4	*4/*68+*4	0	poor
53	NA06989	European	1328	Mother	*9/*9	*9/*9	*9/*9	1	normal
54	NA12331	European	1328	Child	*4/*9	*4/*9	*4/*9	0.5	intermediate
55	NA18507	Yoruban	Y009	Father	*2/*4x2	*2/*4x2	*2/*4x2	1	normal
56	NA18508	Yoruban	Y009	Mother	*2/*5	*2/*5	*2/*5	1	normal
57	NA18506	Yoruban	Y009	Child	*2/*5	*2/*5	*2/*5	1	normal
58	NA19900	African American	2425	Father	*3/*29	*3/*29	*3/*29	0.5	intermediate
59	NA19901	African American	2425	Mother	*1/*1	*1/*1	*1/*1	2	normal
60	NA19902	African American	2425	Child	*1/*29	*1/*29	*1/*29	1.5	normal
61	NA19789	Mexican American	M037	Father	*1/*1	*1/*1	*1/*1	2	normal
62	NA19788	Mexican American	M037	Mother	*2/*78+*2	*2/*78+*2	*2/*78+*2	2	normal
63	NA19790	Mexican American	M037	Child	*1/*78+*2	*1/*78+*2	*1/*78+*2	2	normal
64	HG00463	Han Chinese	SH021	Father	*36+*10/*36+*10	*36+*10/*36+*10	*36+*10/*36+*10	1	normal
65	HG00464	Han Chinese	SH021	Mother	*1/*36+*10	*1/*36+*10	*1/*36+*10	1.5	normal
66	HG00465	Han Chinese	SH021	Child	*36+*10/*36+*10	*36+*10/*36+*10	*36+*10/*36+*10	1	normal
67	HG00592	Han Chinese	SH057	Father	*1/*10	*1/*10	*1/*10	1.5	normal
68	HG00593	Han Chinese	SH057	Mother	*2/*36+*10	*2/*36+*10	*2/*36+*10	1.5	normal
69	HG00594	Han Chinese	SH057	Child	*1/*2	*1/*2	*1/*2	2	normal
70	HG01190	Puerto Rican	PR40	Father	*5/*68+*4	*5/*68+*4	*5/*68+*4	0	poor
71	HG01191	Puerto Rican	PR40	Mother	*2/*41	*2/*41	*2/*41	1.5	normal
72	HG01192	Puerto Rican	PR40	Child	*5/*41	*5/*41	*5/*41	0.5	intermediate
73	NA12342	European	1330	Father	*4/*41	*4/*41	*4/*41	0.5	intermediate
74	NA12343	European	1330	Mother	*1/*5	*1/*5	*1/*5	1	normal
75	NA12336	European	1330	Child	*5/*41	*5/*41	*5/*41	0.5	intermediate
76	NA10853	European	1349	Father	*2/*41	*2/*41	*2/*41	1.5	normal
77	NA10854	European	1349	Mother	*1/*4	*1/*4	*1/*4	1	normal
78	NA11834	European	1349	Child	*2/*4	*2/*4	-	1	normal
79	NA19200	Yoruban	Y045	Father	*5/[*76+*1]	[*1]/*5	[*1]/*5	1	normal
80	NA19201	Yoruban	Y045	Mother	*1/*17	*1/*17	*1/*17	1.5	normal
81	NA19202	Yoruban	Y045	Child	*1/[*76+*1]	[*1]/*1	[*1]/*1	2	normal
82	NA18516	Yoruban	Y013	Father	*1/*17	*1/*17	*1/*17	1.5	normal
83	NA18517	Yoruban	Y013	Mother	*5/*10	*5/*10	*5/*10	0.5	intermediate
84	NA18515	Yoruban	Y013	Child	*1/*10	*1/*10	*1/*10	1.5	normal
85	NA19818	African American	2418	Father	*1/*17	*1/*17	*1/*17	1.5	normal
86	NA19819	African American	2418	Mother	*2/*4x2	*2/*4x2	*2/*4x2	1	normal
87	NA19828	African American	2418	Child	*2/*17	*2/*17	*2/*17	1.5	normal
88	NA12399	European	1354	Father	*1/*1	*1/*1	*1/*1	2	normal
89	NA12400	European	1354	Mother	*1/*68+*4	*1/*68+*4	*1/*68+*4	1	normal
90	NA12386	European	1354	Child	*1/*1	*1/*1	*1/*1	2	normal
91	NA11891	European	1377	Father	*1/*1	*1/*1	*1/*1	2	normal
92	NA11892	European	1377	Mother	*6/*41	*6/*41	*6/*41	0.5	intermediate
93	NA10865	European	1377	Child	*1/*41	*1/*41	*1/*41	1.5	normal
94	NA12272	European	1418	Father	*1/*1	*1/*1	*1/*1	2	normal
95	NA12273	European	1418	Mother	*1/*1	*1/*1	*1/*1	2	normal
96	NA10837	European	1418	Child	*1/*1	*1/*1	*1/*1	2	normal

For samples without SVs, diplotypes were determined by combining the star allele assigned to each of the two haplotypes. For example, phasing algorithms estimated from subject NA12805 two haplotypes—of which one matched *CYP2D6*2* and the other **4*—to form a *CYP2D6*2/*4* diplotype (**Figure 2.2A**).

For samples with a whole gene deletion, diplotypes were determined such that one haplotype contained the *CYP2D6*5* deletion allele while the other was assigned a star allele based on detected SNVs and indels. For example, subject NA18508 had only one *CYP2D6* gene copy, and all detected SNVs were hemizygous and matched *CYP2D6*2*. Stargazer called this sample as having a *CYP2D6*2/*5* diplotype (**Figure 2.2B**).

For samples with a whole gene duplication, Stargazer resolved the identity of the extra *CYP2D6* gene copy in the affected haplotypes. For example, Stargazer detected three gene copies in subject NA19685 with a *CYP2D6*1/*2* diplotype (**Figure 2.2C**). This sample could tentatively have a duplication on the *CYP2D6*1* or **2* allele, or in other words could have a *CYP2D6*1x2/*2* or **1/*2x2* diplotype. Stargazer used the allelic depth ratios of the SNVs defining the *CYP2D6*2* allele to determine which allele carried the duplication. If the *CYP2D6*2* allele carried the extra copy, the sample would have a read ratio of 2:1 reads for the respective SNVs. Indeed, most samples that were heterozygous for these SNVs had read ratios close to 1, whereas NA19685 was a significant outlier with an average ratio of 2.4 from the PGRNseq 1.1 data. Stargazer called the sample as having a *CYP2D6*1/*2x2*. The read ratio approach can also be used to distinguish between diplotypes having duplications on both alleles and those having multiple copies on one allele (e.g., *CYP2D6*1x2/*2x2* vs. **1/*2x3*).

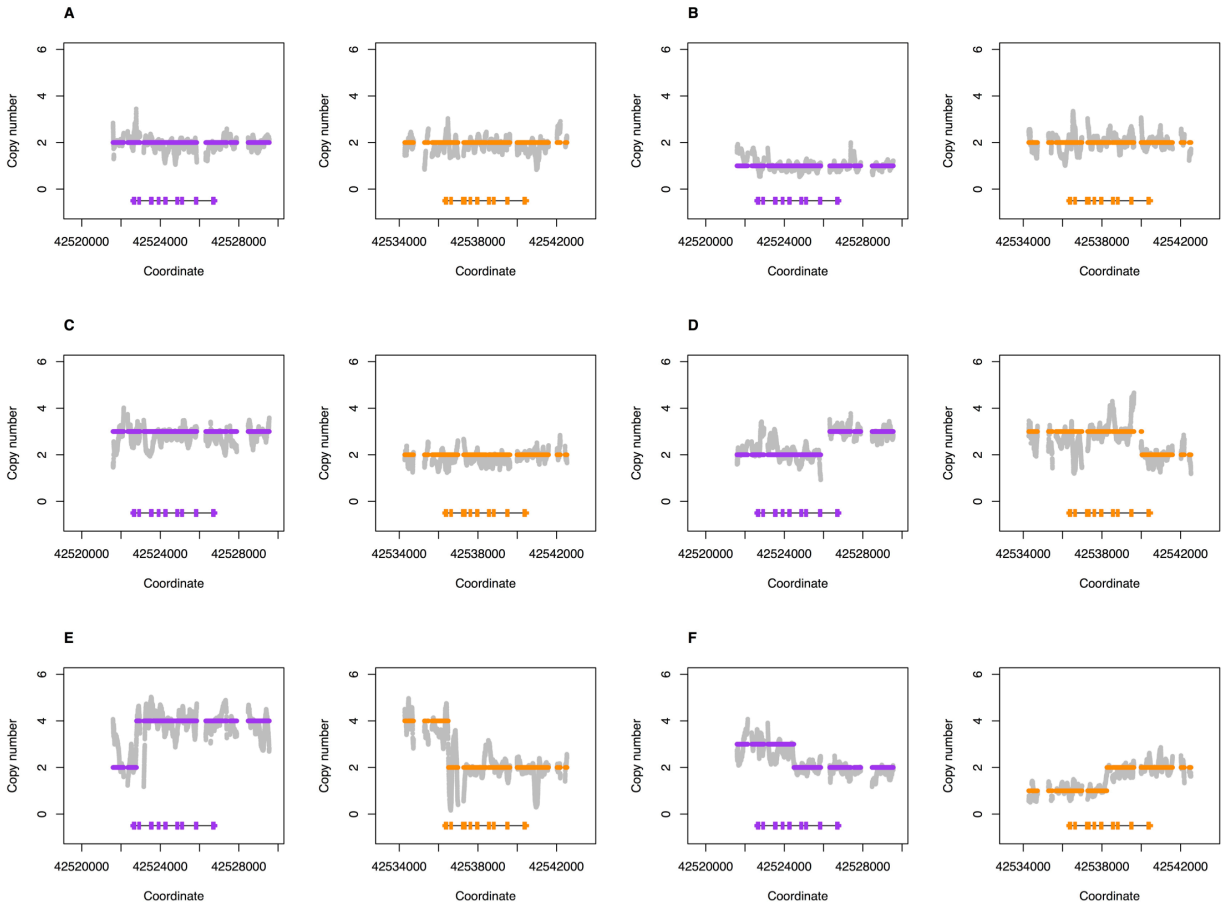


Figure 2.2 Examples of structural variation detected by Stargazer in HapMap trios
Grey dots are copy number calculated from read depth. Dots colored purple and orange are the mean copy number for *CYP2D6* and *CYP2D7*, respectively, determined by the changepoint algorithm. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. All panels were generated from PGRNseq v2.0 data. (A) shows European sample NA12805 that has a *CYP2D6*2/*4* diplotype without structural variation; this sample was included for comparison. (B) shows a gene deletion in Yoruban sample NA18508 with a *CYP2D6*2/*5* diplotype. (C) shows a gene duplication in Mexican American sample NA19685 with a *CYP2D6*1/*2x2* diplotype. (D) shows a complex structural variation involving a gene duplication and a gene hybrid in Peruvian sample HG01979 genotyped as *CYP2D6*2/*68+*4*. (E) shows a complex structural variation involving multiple gene duplications and gene hybrids in Han Chinese sample HG00465 genotyped as *CYP2D6*36+*10/*36+*10*. (F) shows a complex structural variation involving a gene hybrid in Mexican American sample NA19790 genotyped as *CYP2D6*1/*78+*2*.

For samples with complex structural variation, diplotypes were called using individual algorithms. For example, *CYP2D6*68+*4* is a tandem duplication where the *CYP2D6*4* gene copy is defined by only one SNV while *CYP2D6*68* is a hybrid gene featuring a *CYP2D7* sequence from intron 1 onward and four *CYP2D6* SNVs before the breakpoint. Stargazer called one of the two haplotypes of HG01979 as having this tandem structure because all five SNVs were detected, they were haplotype phased together, and the hybrid gene was also observed

(**Figure 2.2D**). The other haplotype was matched to *CYP2D6*2* and Stargazer called this sample as having a *CYP2D6*2/*68+*4* diplotype. Similar approaches were employed to determine diplotypes involving other tandem duplications, such as *CYP2D6*36+*10* (**Figure 2.2E**) and **78+*2* (**Figure 2.2F**), where the *CYP2D6*36* and **78* alleles each contain a gene hybrid to *CYP2D7*.

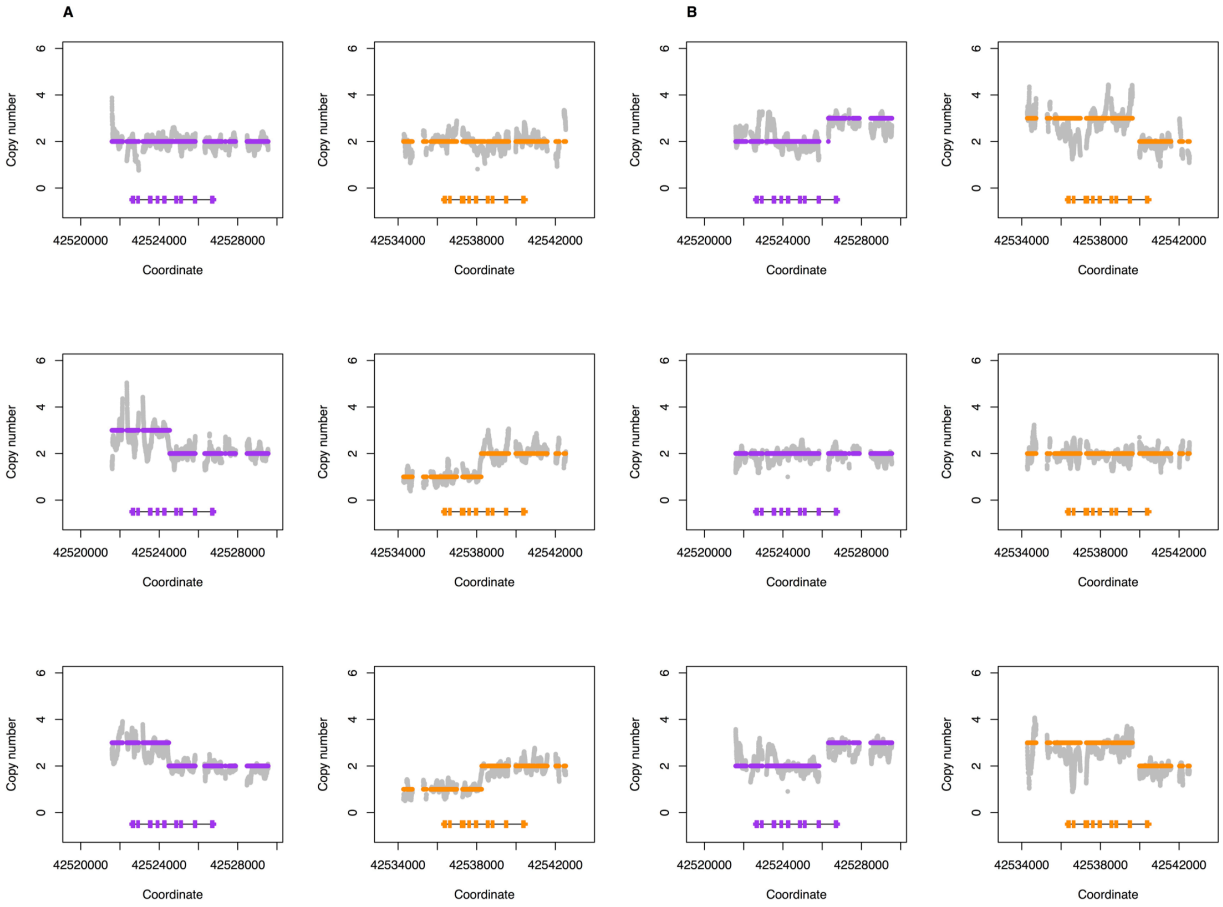


Figure 2.3 Segregation of complex structural variants detected in two HapMap trios
For each trio, data from the father is shown in the top panel, data from the mother is shown in the middle panel, and data from the child is shown in the bottom panel. Grey dots are copy number calculated from read depth. Dots colored purple and orange are the mean copy number for *CYP2D6* and *CYP2D7*, respectively, determined by the changepoint algorithm. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. All panels were generated from PGRNseq v2.0 data. (A) shows segregation of *CYP2D6*78+*2* in the Mexican American M037 family. (B) shows segregation of *CYP2D6*68+*4* in the European 1463 family.

For samples with more than one SV, Stargazer tested all possible pairwise combinations of SVs to determine diplotypes. More specifically, Stargazer first fit every combination of SVs against a sample's observed *CYP2D6* and *CYP2D7* copy number profiles and then selected the

combination that produced the least deviance. For example, HG00463 and HG00465 (family SH021) carry both a gene duplication and a gene hybrid on each of their chromosomes, and their profiles are best explained as having a *CYP2D6*36+*10* tandem arrangement on each chromosome (**Supplementary Figures 2.2A** and **2.2B**). There was one additional sample with multiple SVs—HG01190—whose profile was best explained by the combination of a *CYP2D6*5* deletion and *CYP2D6*68+*4* (**Supplementary Figure 2.2C**).

Summary of genotype and phenotype calls by Stargazer for HapMap trios

Stargazer called 20 unique haplotypes from the 32 HapMap trios. The frequencies of these haplotypes in the 64 unrelated parents are shown in **Supplementary Table 2.1**. As expected, fully functional *CYP2D6*1* and **2* alleles had the highest frequencies (31.3 and 18.8%, respectively) among the parents, followed by the nonfunctional *CYP2D6*4* allele (8.6%). Stargazer also detected *CYP2D6*17* (7.0%), which is commonly found in subjects of African ancestry including African Americans, and *CYP2D6*10* (2.3%), which occurs most frequently in East Asians. Both *CYP2D6*17* and **10* are decreased-function alleles. In addition, Stargazer identified many haplotypes with structural variation in the parents: 4.7% with a gene deletion (*CYP2D6*5*), 11.7% with a gene duplication (*CYP2D6*2x2*, **4x2*, **4N+*4*, **10x2*, **36+*10*, **68+*4*, and **78+*2*), and 8.6% with a gene hybrid (*CYP2D6*4N*, **36*, **68*, and **78*). This translates to 9.4, 21.9, and 15.6% of the parents having at least one gene deletion, duplication, or hybrid, respectively. Based on the diplotype calls, 4.7, 10.9, 82.8, and 1.6% of the parents were predicted to be poor, intermediate, normal, and ultrarapid metabolizers, respectively.

Comparison between Stargazer and orthogonal genotyping methods

When the *CYP2D6* haplotype calls of Stargazer were compared with those determined by orthogonal methods, the concordance rate was 99.0% (190 out of 192 haplotypes; **Table 2.1**). The two discordant haplotypes were found in NA19200 and NA19202 from the Y045 trio. For these samples, the orthogonal methods called *CYP2D6*5/*76+*1* and *CYP2D6*1/*76+*1* diplotypes, respectively, while Stargazer called *CYP2D6*1/*5* and *CYP2D6*1/*1* diplotypes. The nonfunctional *CYP2D6*76* allele—a *CYP2D6/CYP2D7* hybrid—was identified using long-range PCR and Sanger sequencing. The allele is essentially a *CYP2D7* gene that has a *CYP2D6* downstream sequence with a switch region to *CYP2D6* past exon 9, and lacks a *CYP2D7*-specific sequence also referred to as 'spacer' [**73**]. Since this allele has a *CYP2D6*-specific sequence, it may produce positive results with some long-range PCR reactions that are deemed diagnostic for the presence of a gene duplication. Nonetheless, Stargazer did not detect any significant change in copy number either at the switch region or in the spacer, so the program did not call *CYP2D6*76* (**Supplementary Figure 2.3**). Note that the phenotype prediction is the same whether the allele is detected or not.

Stargazer calls using WGS data

To assess the performance of Stargazer on WGS data, we evaluated two samples (NA19238 and NA12878) that were sequenced with PGRNseq v1.1, PGRNseq v2.0, and WGS. Although NA12878 carried a *CYP2D6*68+*4* tandem duplication, Stargazer called the same diplotypes regardless of the sequencing platform (**Figure 2.4**).

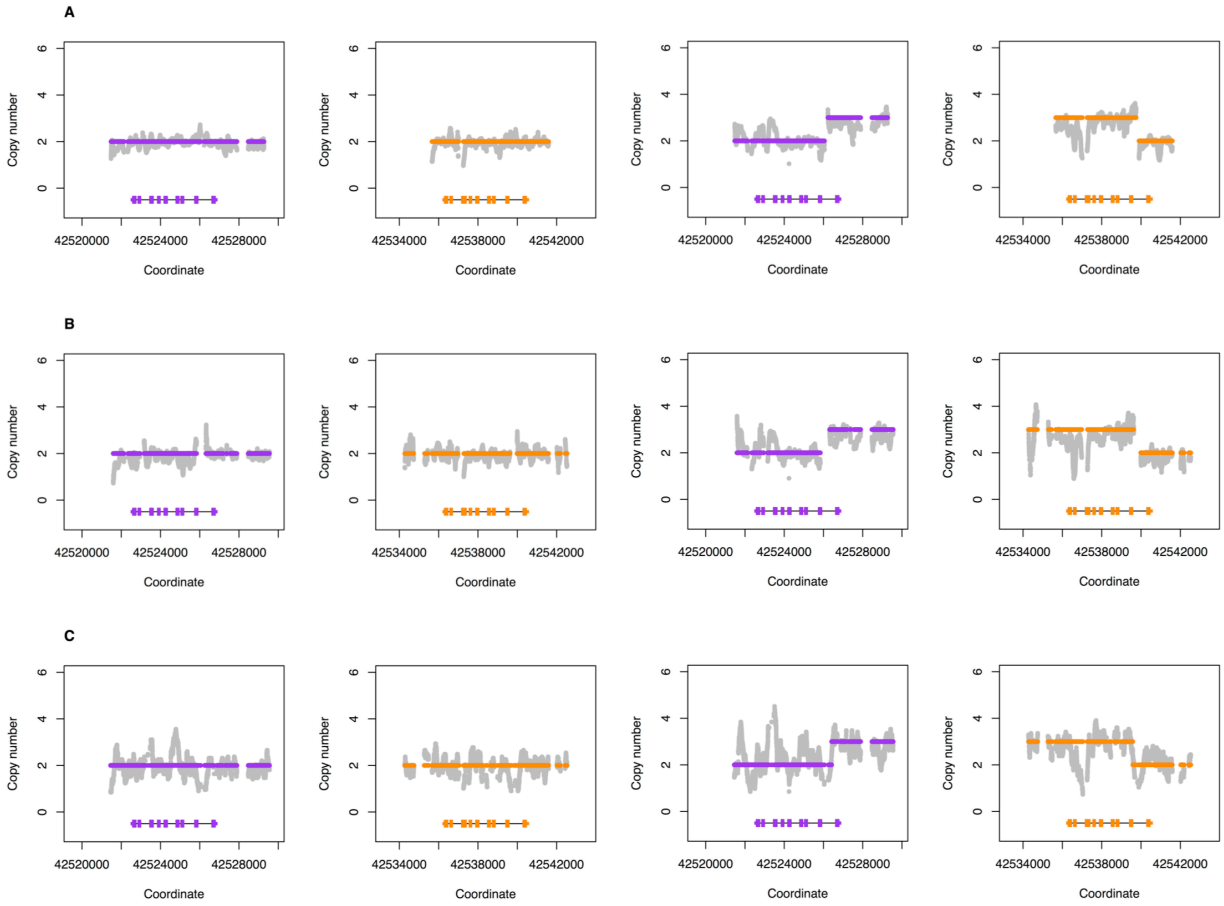


Figure 2.4 Comparison of custom capture and whole genome sequencing for Stargazer's detection of *CYP2D6* structural variation
Grey dots are copy number calculated from read depth. Dots colored purple and orange are the mean copy number for *CYP2D6* and *CYP2D7*, respectively, determined by the changepoint algorithm. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. Two subjects NA19238 and NA12878 were sequenced with (A) PGRNseq v1.1 at ~400X coverage with 100bp paired-end reads, (B) PGRNseq v2.0 at ~160X coverage with 100bp paired-end reads, and (C) whole genome sequencing at ~30X coverage with 150bp paired-end reads. In all three cases, Stargazer called the correct diplotypes *CYP2D6*1/*17* and **3/*68+*4*, respectively.

Testing Stargazer at various sequencing coverages

We applied Stargazer to simulated datasets generated by randomly downsampling sequence reads from PGRNseq v2.0 data (**Supplementary Table 2.2**). When 15% of reads were used (corresponding to 23.7x coverage), 186 out of 188 haplotypes were correctly called; the two misclassified haplotypes carried an SV. However, the misclassifications did not affect the phenotype prediction. Based on these results, our recommendation is to use Stargazer for datasets with a mean read coverage greater than 20x.

SNVs and indels detected by NGS

From the PGRNseq v1.1 and PGRNseq v2.0 data, 142 SNVs and indels were detected at 138 loci within 3 kb from either end of *CYP2D6*, 86 of which are not currently used to define star alleles (**Supplementary Table 2.3**). Among these—according to SeattleSeq Annotation—five are missense pathogenic variants while the remaining ones are either synonymous or within the 3' or 5' untranslated region, downstream or upstream of the gene, or within an intron. We did not find any novel variants that are obviously detrimental to CYP2D6 function, such as nonsense, frameshift, or splice site pathogenic variants.

2.5 Discussion

We developed Stargazer—a new software tool for calling star alleles in various polymorphic pharmacogenes from NGS data. When building Stargazer, we used *CYP2D6* as a model for the detection and interpretation of SVs in the context of other observed SNVs and indels. We purposefully chose *CYP2D6* as a starting point because it is one of the most complex genetic loci to genotype in the human genome. Two other programs—Cypiripi [**74**] and Astrolabe [**75**]—have been published to genotype *CYP2D6* from NGS data. Unlike Stargazer, which uses statistical phasing with population haplotype frequencies to call diplotypes, both Cypiripi and Astrolabe rely on probabilistic scoring system for matching detected variants to the most likely diplotypes comprised of pre-defined haplotypes. Although both tools can reliably call simple diplotypes, they have difficulties with the detection of complex SVs, such as *CYP2D6/CYP2D7* hybrids. We show that Stargazer can reliably detect those hybrids from targeted or WGS data.

More specifically, we show that Stargazer correctly genotyped *CYP2D6* for 32 ethnically diverse HapMap trios. These trios were previously validated by a variety of orthogonal methods,

and comparisons show that Stargazer is 99.0% concordant with these methods. In the future, we will test additional verified samples in order to further validate Stargazer's performance. All diplotype calls by Stargazer were inherited according to expectations including population-specific star alleles such as *CYP2D6*10* and **17*. Stargazer also produced the same diplotype calls for all samples from the two independent PGRNseq v1.1 and PGRNseq v2.0 datasets.

We plan to extend Stargazer to *CYP2A6*—another highly polymorphic pharmacogene displaying many SNVs and indels as well as SVs. CYP2A6 metabolizes nicotine, and sequence variation in *CYP2A6* has been linked to nicotine dependence and withdrawal symptoms upon smoking cessation [**76**]. Similar to *CYP2D6*, *CYP2A6* has several star alleles with a gene hybrid to its nearby paralog *CYP2A7* [**77**]. We also plan to develop Stargazer for other cytochrome P450 genes.

As larger genomic datasets become available, several aspects of Stargazer will improve. These include the statistical estimation of phased haplotypes, primarily haplotypes based on rare variants. In the current version of Stargazer, we incorporated a large panel of reference samples from the 1000 Genomes Project. This approach performed well for our dataset, but we are aware that in further applications, rare variants may have frequencies that are too low to be phased reliably. To ameliorate this issue, we plan to merge multiple large reference panels to obtain additional haplotype information. Novel variants will require physical phasing backed by sequence reads. When short reads cannot provide adequate phasing, long-read sequencing from Oxford Nanopore Technologies or Pacific Biosciences can be used to generate reference haplotype information. Recently, both technologies have been successfully applied to sequence *CYP2D6* [**78**,**79**].

Certain features of Stargazer are specific for targeted sequencing such as PGRNseq. For example, for the purpose of normalization, Stargazer requires multiple samples to be analyzed at a time. This is because sequencing with custom capture typically yields uneven coverage across the genes of interest, and Stargazer's copy number estimation is based on population statistics. If the sample size is too small or a large fraction of samples share the same type of structural variation, population statistics can be shifted dramatically, generating biased copy number data. However, this problem can be addressed by including one or more reference samples with known copy numbers. For WGS data, where coverage is usually distributed more evenly, the intersample normalization may be skipped, allowing Stargazer to analyze a single sample.

We reported five missense variants that are not currently used to define any star allele. However, interpretation of these variants is difficult without functional characterization. In fact, the same is true for many variants in existing star alleles (e.g., *CYP2D6*22* is defined by a nonsynonymous SNV in exon 9 with unknown effect). Therefore, there is clearly a need to more rigorously characterize the function of the rapidly increasing number of haplotypes to facilitate phenotype prediction. In the future, it is possible that data from deep mutational scanning for the CYP2D6 enzyme could be incorporated into Stargazer to aid the characterization of the functional consequences of all possible single pathogenic variants of this protein [**80**].

The HapMap trios used in this study consist of seven distinct ethnic groups and therefore represent a sampling of the global distribution of *CYP2D6* genotypes. Characterized by multiple genotyping platforms including NGS, these trios can serve as a reference resource for other *CYP2D6* genotyping projects.

There is growing awareness of individual variation in drug response. For example, in March 2013, the Food and Drug Administration cautioned against the use of codeine in children

of any age to treat pain after surgery to remove the tonsils or adenoids [**81**]. Shortly after, a prospective study showed that children who were CYP2D6 ultrarapid metabolizers and taking codeine after those surgeries were at a higher risk for toxicity and death [**82**]. In April 2017, the Food and Drug Administration issued the agency's strongest warning against codeine, alerting that the medication should not be used to treat pain or cough in children younger than 12 years. While limiting the therapeutic use of codeine addresses the concern for patient safety, tailoring codeine or other drug treatments based on an individual's *CYP2D6* genotype could achieve the same goal. There may also be therapeutic settings where alternative treatments are not fully interchangeable and health outcomes could suffer from restrictions in drug choice. For example, national guidelines recommend codeine as a front-line drug for the treatment of pain in patients with sickle cell disease, and many hematologists prefer codeine to other analgesics that have comparable efficacy but higher potential for abuse and physical dependence [**83**]. With additional validation, Stargazer may offer an alternative approach for optimizing treatment response in all patients.

2.6 Acknowledgements

The authors acknowledge the Pharmacogenomics Research Network for supporting the development of PGRNseq. This work was supported by NIH grants HL069757, GM092676, GM116691, GM115318, GM115277, and S10OD021553, and the University of Washington's Graduate School Fund for Excellence and Innovation.

Chapter 3 Interrogation of Structural Variants by Stargazer Improves the Association Between *CYP2D6* Genotype and CYP2D6-mediated Metabolic Activity

This chapter has been submitted for publication: Dalton, R.,* Lee, S.B.,* Claw, K., Prasad, B., Phillips, B.R., Shen, D.D., Wong, L.H., Fade, M., McDonald, M.G., Dunham, M.J., Fowler, D.M., Rettie, A.E., Schuetz, E., Gaedigk, A., Thornton, T.A., Nickerson, D.A., Thummel, K.E. & Woodahl, E.L. (2019). *These authors contributed equally to this work.

3.1 Summary

The *CYP2D6* gene locus is challenging to accurately genotype due to a large number of single nucleotide variants (SNVs) and complex structural variation. Our goal was to determine whether the *CYP2D6* genotype-phenotype association is increased when diplotype assignments incorporate structural variant alleles identified by Stargazer, a novel allele-calling software tool applied to next-generation sequencing data. We compared *CYP2D6* diplotypes and activity scores in human liver samples made using PGRNseq SNV data alone and using Stargazer that included structural variation data. CYP2D6 activity was measured with dextromethorphan and metoprolol as probe substrates. Without incorporating structural variation data, the diplotypes of 70 samples were incorrectly assigned (22%) leading to 25 incorrect activity scores (8%). We also identified novel rare coding variants, one of which, A449D, exhibited decreased (44%) catalytic activity. The accuracy of CYP2D6 phenotype prediction is improved when the *CYP2D6* gene locus is interrogated using next-generation sequencing approaches coupled with Stargazer.

3.2 Introduction

The combination of a wide range of clinically relevant substrates and a well-studied relationship between genetic variation and enzymatic activity makes cytochrome P450 2D6 (CYP2D6) a logical candidate for clinical pharmacogenetic testing [**84**,**85**]. To guide the implementation of pharmacogenomic testing, the Clinical Pharmacogenetics Implementation Consortium (CPIC)

has published five guidelines for drugs metabolized by CYP2D6, i.e. codeine, tamoxifen, selective serotonin reuptake inhibitors, tricyclic antidepressants, and antiemetics [**57**,**86**,**87**,**88**,**89**,**90**,**91**]. A high degree of variability in CYP2D6 activity has been observed among individuals as well as populations, which can largely be attributed to variation in the *CYP2D6* gene [**50**,**49**]. The *CYP2D6* gene is notoriously difficult to interrogate, however, due to challenges imposed by the complexity of the *CYP2D* gene locus, which includes a highly homologous pseudogene, *CYP2D7*, and a high degree of *CYP2D6* variation [**50**,**92**,**93**]. Due to such variation, targeted genotyping panels may miss rare clinically relevant variants that are present in an individual but not interrogated, leading to phenotype misclassification [**50**,**93**]. This is particularly problematic in diverse populations, where there are limited or no data regarding genetic variation [**49**,**93**,**94**].

In addition to single nucleotide variants and indels (referred to collectively as SNVs for brevity), it is also important to consider structural variation when assigning *CYP2D6* diplotypes. Gene deletions, duplications, and multiplications, as well as gene rearrangements between *CYP2D7* and *CYP2D6*, known as hybrid genes, have all been observed. Of note, *CYP2D6/CYP2D7* hybrid genes occur in a number of configurations and often result in loss of enzyme function [**50**,**93**,**73**,**66**,**95**]. Such structural variation is not uncommon: the *CYP2D6*5* gene deletion occurs at a frequency of 2-6% worldwide, while duplications/multiplications of functional gene units, such as *CYP2D6*1xN* or **2xN*, occur at frequencies up to 12% [**49**,**93**,**96**].

Until recently, identification of *CYP2D6* alleles from genotyping or sequencing data has involved "manual" assignment of haplotypes and diplotypes by comparing the variants found in a sample to those listed in *CYP2D6* haplotype definition tables [**96**,**97**,**58**]. The large amount of data obtained via *CYP2D6* sequencing makes the manual method cumbersome and prone to

error, highlighting the need to turn the mental algorithm of manual allele identification into a bioinformatics tool. Stargazer, a software tool for assigning star alleles from next-generation sequencing data, meets these needs [**56**]. Stargazer combines SNV calls with structural variation data to improve the accuracy of diplotype calls and greatly reduces the labor involved in the process.

Variation in the *CYP2D6* gene is not the sole contributor to inter-individual variability in CYP2D6 activity. CYP2D6 protein level has recently been described as being more predictive of metabolic drug clearance than CYP2D6 activity score (AS) [**98**], implying that the causal pathway between genetics and clearance is modulated by other proteins. Aldo-keto reductase 1D1 (AKR1D1), a protein that may contribute to variation in CYP activity, is involved in bile acid homeostasis and has been predicted to be a "master regulator" of the expression of several CYPs [**99**]. Another protein that is essential for CYP2D6-dependent drug clearance is cytochrome P450 oxidoreductase (POR), which mediates the transfer of electrons from NADPH to CYP enzymes, a rate-limiting step in all CYP-mediated metabolism. *POR* is a polymorphic gene and has been associated with variation in CYP2D6 activity [**100**,**101**].

The major goal of this study was to compare the predictive power of *CYP2D6* AS assignments based on diplotypes determined using SNV data alone to diplotypes that also incorporated structural variation data. To that end, we assessed CYP2D6 activity in a panel of 314 human liver tissue samples using two CYP2D6 probe substrates, dextromethorphan and metoprolol, and performed next-generation-based sequencing using PGRNseq, a targeted gene panel to assign *CYP2D6* diplotypes, coupled with Stargazer. We also explored the contribution of POR and AKR1D1 to the variability in measured CYP2D6 activity and characterized the catalytic function of selected rare variants identified in the liver samples.

3.3 Results

Liver donor demographics

Donor demographics (N=314) were presented previously (**Supplementary Table 3.1**) [**102**,**103**]. Data on liver disease and concomitant medications were missing or incomplete for many donors and are thus not reported.

Effects of *CYP2D6* mRNA and protein content on activity

Formation rates of dextrorphan from dextromethorphan and α-hydroxymetoprolol from metoprolol were used as a measure of CYP2D6 activity and were well correlated with one another (R^2=0.90). Both dextrorphan and α-hydroxymetoprolol metabolic formation rates were significantly correlated with CYP2D6 mRNA content (**Figures 3.1A** and **3.1B**) and CYP2D6 protein content (**Figures 3.1C** and **3.1D**). The associations between dextrorphan and α-hydroxymetoprolol metabolite formation rates and *CYP2D6* mRNA content were moderate: R^2 values of 0.24 and 0.21, respectively. CYP2D6 protein content showed a stronger association with CYP2D6 activity: R^2 values of 0.51 and 0.46, respectively.

CYP2D6 allele and diplotype frequencies and activity score reassignment following structural variation data analysis

We identified 25 different major *CYP2D6* haplotypes (star alleles) in the liver bank samples using the Stargazer algorithm (**Table 3.1**), comprising 73 unique diplotypes (**Supplementary Table 3.2**). Of these, only 17 unique alleles were identified from SNV data alone. Representative examples of structural variants detected by Stargazer are displayed in **Figure 3.2**. Allele frequencies reflect the predominantly European descent of the samples; the three most common

alleles were *CYP2D6*1* (32.0%), **2* (14.3%), and **4* (13.5%), which is consistent with published reports for this population [**49**].

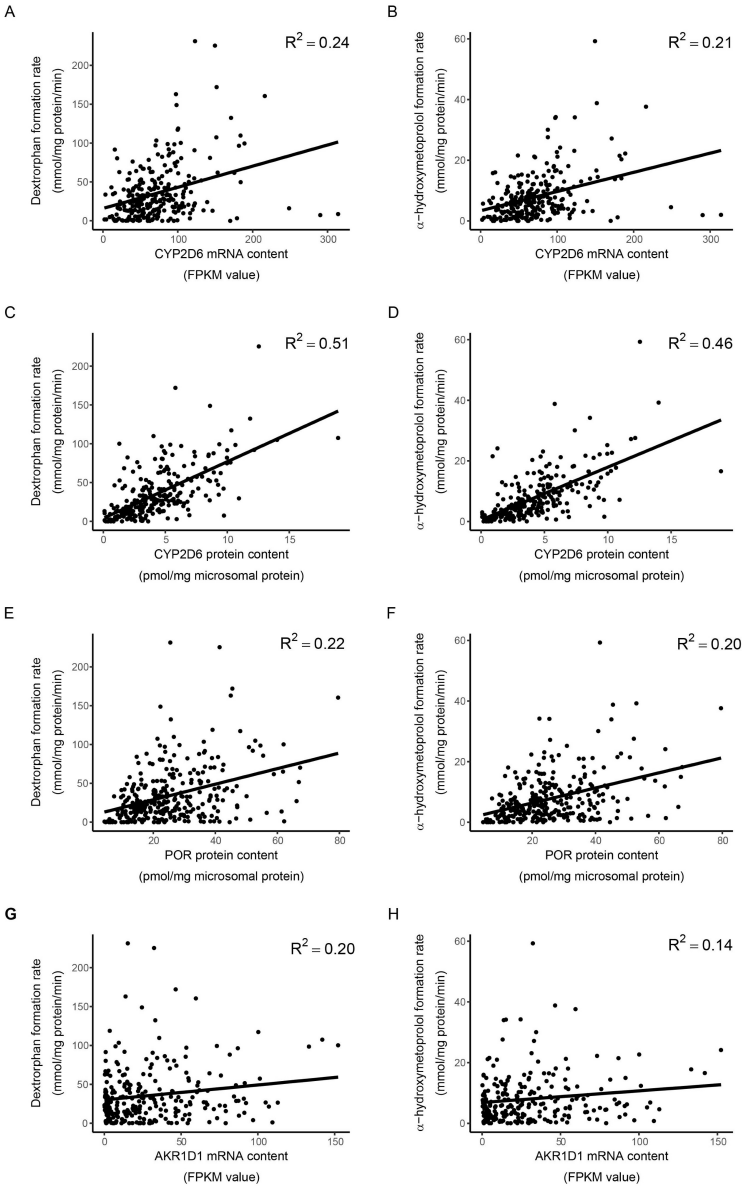


Figure 3.1 Association between CYP2D6 metabolite formation rate and *CYP2D6* mRNA and protein content, POR protein, and *AKR1D1* mRNA content

CYP2D6 metabolite formation rate was associated with CYP2D6 mRNA content quantitated by RNAseq: dextrorphan formation rate (panel A) and alpha-hydroxymetoprolol formation rate (panel B); with CYP2D6 protein content quantitated by LC-MS/MS: dextrorphan formation rate (panel C) and alpha-hydroxymetoprolol formation rate (panel D); with POR protein content quantitated by LC-MS/MS: dextrorphan formation rate (panel E) and alpha-hydroxymetoprolol formation rate (panel F); and with *AKR1D1* mRNA content quantitated by RNAseq: dextrorphanol formation rate (panel G) and alpha-hydroxymetoprolol formation rate (panel H).

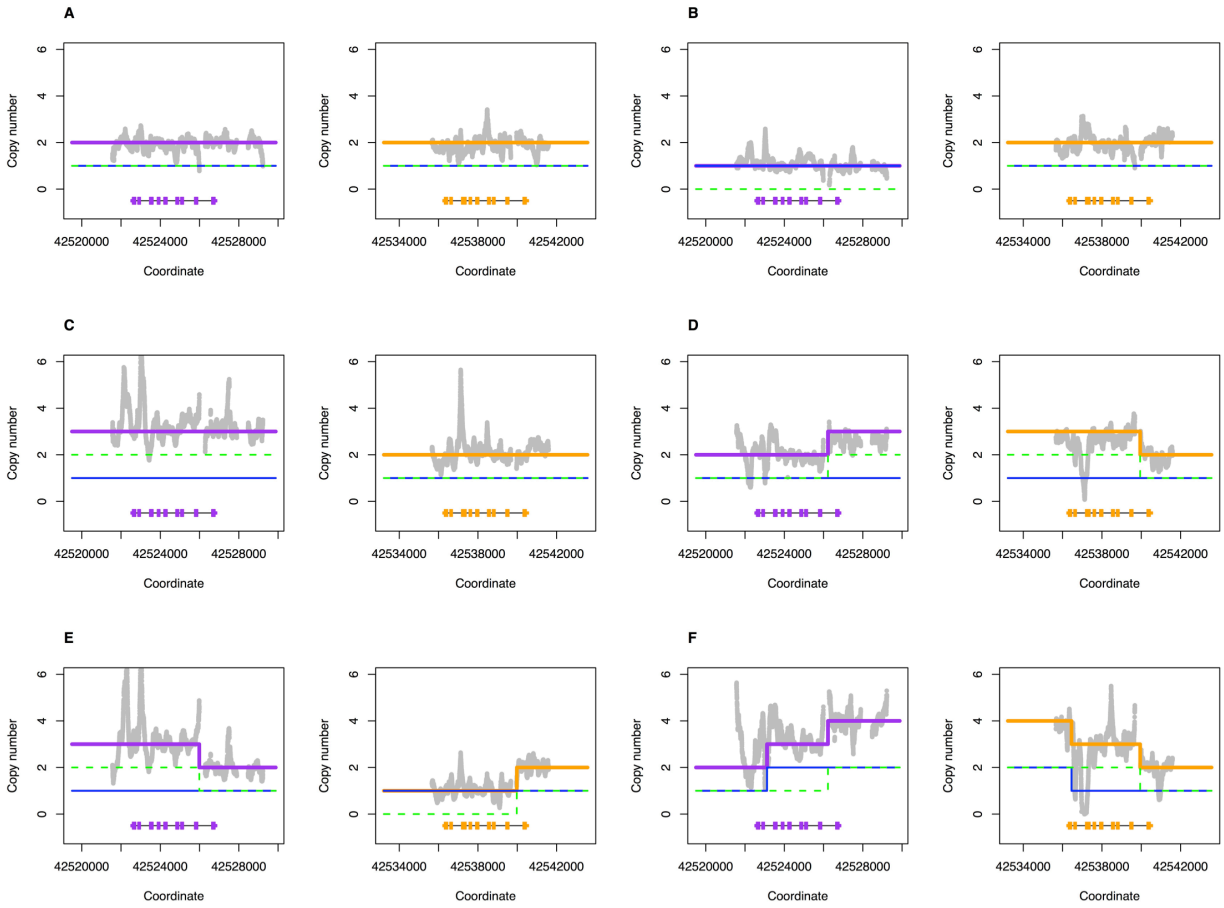


Figure 3.2 Examples of structural variation detected by Stargazer in Liver Bank samples
Grey dots denote copy number calculated from read depth. Blue solid and green dashed lines represent two sequence structures that are being tested against the observed copy number of the sample. Purple and orange lines represent theoretical copy number profiles for *CYP2D6* and *CYP2D7*, respectively, formed by combining the two sequence structures in question. Each panel contains scaled CYP2D6 and CYP2D7 gene models, in which the exons and introns are depicted with boxes and lines, respectively: (A) *CYP2D6*2/*4*: no structural variation; (B) *CYP2D6*2/*5*: gene deletion; (C) *CYP2D6*35/*2x2*: gene duplication; (D) *CYP2D6*9/*68+*4*: complex structural variation involving a gene duplication and a gene hybrid; (E) *CYP2D6 *41/*13+*2*: complex structural variation involving a gene hybrid; (F) *CYP2D6*68+*4/*36+*10*: complex structural variation involving multiple gene duplications and gene hybrids.

Of the 25 allelic variants identified (**Table 3.1**), nine are known to confer normal enzyme function and were assigned a value of 1 to calculate the AS. Normal function alleles had an aggregate frequency of 54.1% and included *CYP2D6*1*, **2*, **33*, **34*, **35*, **39*, **43*, **41x2*, and **13+*2* (previously known as **77+*2*) [**66**]. Seven decreased function alleles, *CYP2D6*9*, **10*, **17*, **29*, **41*, **59*, and **36+*10*, were found at an aggregate frequency of 18.9% and received a value of 0.5. Seven no function alleles, *CYP2D6*3*, **4*, **5*, **6*, **20*, **4N+*4*, and **68+*4*,

comprised 26.1% of the total alleles and received a value of 0. We identified two increased function alleles, *CYP2D6*1x2* (0.5%) and *CYP2D6*2x2* (0.5%), which received values of 2.

Table 3.1 Distribution of *CYP2D6* haplotypes identified in 314 Liver Bank samples

Type	Allele	N (%)	Activity
Reference	**1*	201 (32.0%)	normal
SNV	**2*	90 (14.3%)	normal
	**33*	5 (0.8%)	normal
	**34*	1 (0.2%)	normal
	**35*	37 (5.9%)	normal
	**39*	1 (0.2%)	normal
	**43*	2 (0.3%)	normal
	**9*	22 (3.5%)	decreased
	**10*	15 (2.4%)	decreased
	**17*	4 (0.6%)	decreased
	**29*	1 (0.2%)	decreased
	**41*	70 (11.1%)	decreased
	**59*	5 (0.8%)	decreased
	**3*	7 (1.1%)	none
	**4*	85 (13.5%)	none
	**6*	6 (1.0%)	none
	**20*	3 (0.5%)	none
SV	**1x2*	3 (0.5%)	increased
	**2x2*	3 (0.5%)	increased
	**41x2*	1 (0.2%)	normal
	**13+*2*	1 (0.2%)	normal
	**36+*10*	2 (0.3%)	decreased
	**4N+*4*	4 (0.6%)	none
	**68+*4*	35 (5.6%)	none
	**5*	24 (3.8%)	none

AS were calculated for diplotypes identified from SNV data alone (17 alleles) and for diplotypes identified with the inclusion of structural variation data (25 alleles). Of the samples analyzed, 70 (22.3%) had incorrectly assigned diplotypes based on SNV data alone. Of importance, the inclusion of structural variation changed the AS for 25 of those samples, representing approximately 8% of the investigated liver tissue samples. **Figure 3.3** describes the diplotypes of the 70 samples with structural variation alleles that were identified with Stargazer as well as the changes in AS assignments. The column on the left contains diplotypes assigned based on SNV data alone, the corresponding AS, and the number of samples with that diplotype. The column on the right contains diplotypes assigned by Stargazer using SNV and structural variation data, the corresponding AS, and the number of samples with that diplotype. For example, using SNV data alone, 39 samples were assigned a diplotype of *CYP2D6*1/*1*, which has an AS of 2. When diplotypes were reassigned using Stargazer, 8 of the samples were **1/*5*, resulting in a decrease in AS to 1. The *CYP2D6*5* gene deletion was identified in 25 samples,

occurring at an allele frequency of 3.8%. The importance of including *CYP2D6*5* is evident as the corrected diplotype calls decreased the AS of 18 samples (5.7%) in this sample set. The remaining 7 samples with *CYP2D6*5* had another no function allele; thus, AS did not change. A total of 7 duplicated alleles were identified, i.e. *CYP2D6*1x2*, **2x2*, and **41x2*; these are duplications of functional and decreased function alleles, increasing the AS for samples in which they were identified. We also detected a number of tandem structures: *CYP2D6*4N+*4*, **68+*4*, **36+*10*, and **13*2*. Since *CYP2D6*4N*, **68*, **36*, and **13* are nonfunctional, their inclusion did not change the AS value assigned to these alleles [**95**].

SNV data alone			SNV + structural variation data			
Diplotype	**AS**	**n**	**Diplotype**		**AS**	**n**
Deletion						
**1/*1*	2	39	**1/*5*	↓	1	8
**2/*2*	2	9	**2/*5*	↓	1	4
**35/*35*	2	2	**5/*35*	↓	1	1
**41/*41*	1	5	**5/*41*	↓	0.5	4
**4/*10*	0.5	3	**4/*5*	↓	0	1
**3/*3*	0	1	**3/*5*	→	0	1
**4/*4*	0	17	**4/*5*	→	0	4
Duplication						
**1/*1*	2	39	**1/*1x2*	↑	3	1
**1/*2*	2	36	**1x2/*2*	↑	3	1
**2/*35*	2	3	**2x2/*35*	↑	3	1
**2/*41*	1.5	16	**2x2/*41*	↑	2.5	1
**1/*4*	1	41	**1x2/*4*	↑	2	1
**2/*4*	1	19	**2x2/*4*	↑	2	1
**10/*41*	1	1	**10/*41x2*	↑	1.5	1
Hybrid tandem						
**2/*41*	1.5	16	**13+*2/*41*	→	1.5	1
**1/*4*	1	41	**1/*68+*4*	→	1	12
**1/*4*	1	41	**1/*4N+*4*	→	1	2
**2/*4*	1	19	**2/*68+*4*	→	1	5
**4/*35*	1	5	**35/*68+*4*	→	1	2
**4/*33*	1	2	**4N+4/*33*	→	1	1
**10/*10*	1	1	**10/*36+*10*	→	1	1
**4/*9*	0.5	4	**9/*68+*4*	→	0.5	1
**4/*41*	0.5	14	**41/*68+*4*	→	0.5	3
**4/*41*	0.5	14	**4N+*4/41*	→	0.5	1
**3/*4*	0	2	**3/*68+*4*	→	0	2
**4/*4*	0	17	**4/*68+*4*	→	0	5
Hybrid tandem + hybrid tandem						
**4/*10*	0.5	3	**36+*10/*68+*4*	→	0.5	1
**4/*4*	0	17	**68+*4/*68+*4*	→	0	1
Deletion + hybrid tandem						
**4/*4*	0	17	**5/*68+*4*	→	0	2

Figure 3.3 Diplotypes and activity scores assigned with SNV data alone and with Stargazer structural variation data
Columns on the left show diplotypes and activity scores (AS) assigned using allele calls from SNV data alone. Corrected diplotypes and AS, based on Stargazer allele assignments, are displayed in the columns on the right. AS are color-coded as follows: 3 (dark blue); 2.5 (medium blue); 2 (light blue); 1.5 (light yellow); 1 (dark yellow); 0.5 (orange); and 0 (red). Arrows

indicate the direction of the change in AS assignments with the incorporation of structural data: decrease (↓), increase (↑), and no change (→).

The inclusion of structural data for *CYP2D6* diplotype calling significantly improved the predictive value of the AS assignments for both probe substrates, dextromethorphan and metoprolol (**Figure 3.4**; **Supplementary Table 3.3**). Using SNV data alone, AS predicted 36% and 31% of the variability in CYP2D6 activity when using dextromethorphan and metoprolol as probes, respectively. With structural variation data from Stargazer included, these coefficients increased to 41% and 36%, respectively.

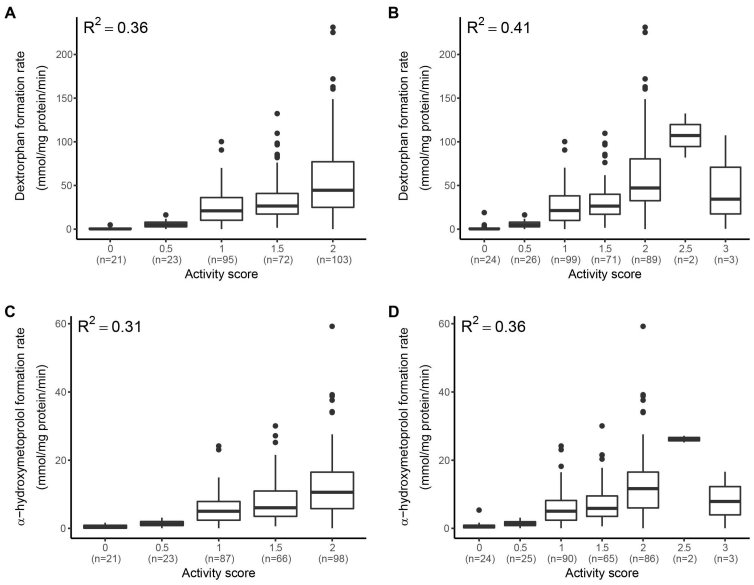


Figure 3.4 Association between CYP2D6 metabolite formation rate and activity score
Dextrorphan formation rate by activity score (AS) assigned with SNV data alone (panel A) and with Stargazer (panel B). Alpha-hydroxymetoprolol formation rate by activity score assigned with SNV data alone (panel C) and with Stargazer (panel D). Boxes represent interquartile range with interior line representing the median. Error bars represent 1.5x the interquartile range. Number of samples in each AS category is given in parentheses.

Other factors contributing to variability in CYP2D6 activity

Mean CYP2D6 activity varied significantly between liver collection sites (University of Washington vs. St. Jude Children's Research Hospital), consistent with our collaborators' reports from these samples [**102**,**103**]; therefore, collection site was included as a covariate. POR protein content, but not mRNA content (data not shown), was significantly associated with metabolite formation rates for both dextrorphan (R^2=0.22, $p<0.001$) and α-hydroxymetoprolol (R^2=0.20,

p<0.001) (**Figures 3.1E** and **3.1F**). *AKR1D1* mRNA content (protein content data not available) was also associated with CYP2D6 activity for both dextrorphan (R^2=0.20, p<0.001) and α-hydroxymetoprolol (R^2=0.14, p<0.001) (**Figures 3.1G** and **3.1H**).

Multiple regression analysis of CYP2D6 activity

Multiple linear regression analysis using robust standard errors was performed to determine which variables, in addition to AS, explain the variability in CYP2D6 activity (**Table 3.2**). Donor age and sex were not included due to a lack of significance in a primary analysis, which is consistent with the literature [**104**]. Although race has been described to contribute to variation in CYP2D6 activity, we did not include it as a covariate because almost all liver donors were of European descent (95.5%). We also excluded liver disease and donor treatment with CYP2D6 inhibitors because the data were incomplete for the majority of donors.

The final model included CYP2D6 AS, microsomal POR protein concentration, hepatic *AKR1D1* mRNA content, and liver collection site. The R^2 for dextrorphan formation rate was 0.51, an improvement over the value of 0.41 obtained when AS was the sole predictor. The R^2 for α-hydroxymetoprolol formation rate was 0.47, also an improvement over the value of 0.36 obtained when only AS was used. The covariates that contributed to these increases in R^2 were POR protein content and *AKR1D1* mRNA content as collection site was included in both models.

Table 3.2 Multiple linear regression: association between CYP2D6 activity and activity score, POR protein content, and *AKR1D1* mRNA content

	Dextromethorphan			Metoprolol		
	β	*SE*	*p-value*	*β*	*SE*	*p-value*
(Intercept)[a]	-27.2	6.87	9.78×10^{-5}	-5.84	1.59	3.02×10^{-4}
AS=0.5	6.17	8.24	0.45	1.01	1.91	0.6
AS=1	27.45	6.6	4.40×10^{-5}	5.37	1.53	5.34×10^{-4}
AS=1.5	30.13	6.72	1.14×10^{-5}	6.65	1.56	2.81×10^{-5}
AS=2	57.6	6.63	5.75×10^{-16}	12.19	1.53	6.99×10^{-14}
AS=2.5	108	27.01	8.48×10^{-5}	22.83	6.19	2.80×10^{-4}
AS=3	60.56	19.97	2.68×10^{-3}	10.12	4.58	0.03
POR protein	0.67	0.14	1.65×10^{-6}	0.18	0.03	8.09×10^{-8}
AKR1D1 mRNA	0.2	0.06	1.07×10^{-3}	0.04	0.01	0.02
Site: UW	27.58	4.25	4.80×10^{-10}	4.33	0.98	1.68×10^{-5}
DF	243			227		
R^2/adj. R^2	0.51/0.49			0.47/0.45		

AS, activity score; POR, cytochrome P450 oxidoreductase; AKR1D1, aldo-keto reductase 1D1; UW, University of Washington; β, β-coefficient; SE, standard error; DF, degrees of freedom.

[a]AS=0 is incorporated into the intercept.

Functional characterization of rare *CYP2D6* variants

A total of 180 SNVs were identified by PGRNseq analysis in the 314 livers tested (**Supplementary Table 3.4**). Of these 113 are not currently used to define a star allele. To explore the potential contribution of rare coding SNVs, we selected 7 nonsynonymous SNVs with CADD scores of 0-32 for functional characterization in a yeast expression system (**Supplementary Table 3.5**). Five exhibited metabolic activity comparable to the reference enzyme, while one (rs79392742; A449D) displayed decreased activity (44.4±6.9%) comparable to the decreased activity P34S control (35.4±4.7%) (**Figure 3.5**). This finding was confirmed using Vivid^{TM} 7-ethoxymethoxy-3-cyanocoumarin, a standard fluorogenic CYP2D6 substrate (**Supplementary Figure 3.1**) and consistent with the SWISS-MODEL [**105**], which predicted that A449D (a small non-polar to bulky polar residue change; CADD score of 32) could interfere with heme binding (**Supplementary Figure 3.2**). The liver carrying A449D was genotyped as *CYP2D6*1/*4*, although it is not known on which allele, and exhibited CYP2D6 activity within the range of livers with an AS=1. Another variant, rs141756339 (R474Q; found in a sample genotyped as *CYP2D6*2/*41*) showed increased activity (127±8.1%; **Figure 3.5**), which was also confirmed using Vivid^{TM} 7-ethoxymethoxy-3-cyanocoumarin (**Supplementary Figure 3.1**).

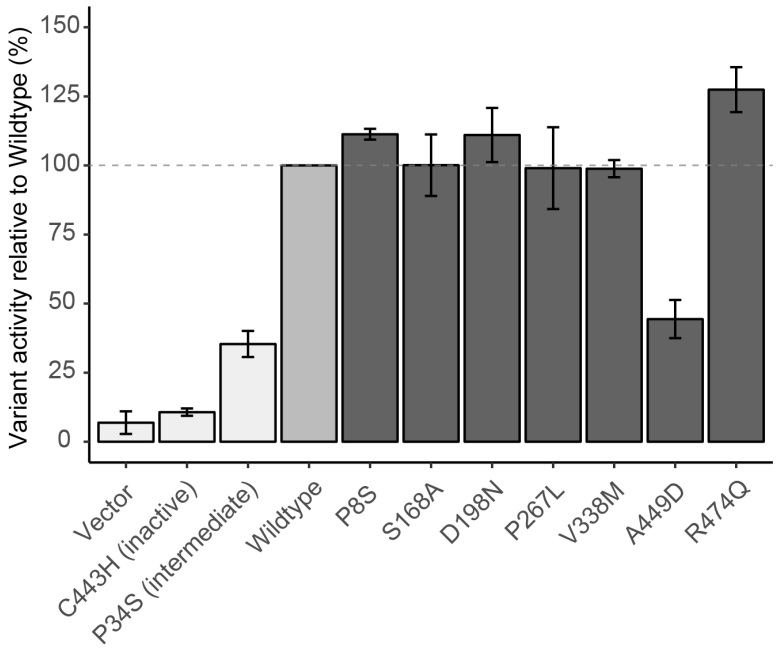


Figure 3.5 Functional characterization of rare *CYP2D6* coding variants
Each *CYP2D6* variant was induced in an isogenic yeast strain and function was characterized with a tic-ABP1P CYP2D6 probe. Fluorescence was normalized to that of CYP2D6 wildtype (medium grey; horizontal dashed line represents 100% activity). Control strains (light grey): empty vector, inactive variant (C443H), and decreased function variant (P34S). Novel rare variant strains (dark gray): P8S, S168A, D198N, P267L, V338M, A449D, and R474Q. Error bars indicate standard errors from at least 2 independent replicates.

3.4 Discussion

The inclusion of structural variation information is essential to accurately assign *CYP2D6* diplotypes and AS and leads to a better prediction of CYP2D6 activity than when AS is assigned based on SNV data alone. It follows that the identification of gene deletions, gene duplications/multiplications, and hybrid alleles will maximize the clinical utility of *CYP2D6* pharmacogenetic tests. An efficient way to achieve that requirement is through gene sequencing, as demonstrated with the use of PGRNseq, and the inclusion of structural variation for activity phenotype prediction using the Stargazer interpretive platform. Our data demonstrate the value of this approach. Compared to diplotypes assigned using SNV data alone, diplotypes of 70 samples were incorrectly assigned due to the presence of structural variation, representing over 20% of the samples tested. Importantly, the AS changed for 25 samples, representing approximately 8% of the samples tested.

When diplotypes were assigned based on SNV data alone, AS explained about 36% and 31% of the variation in CYP2D6 activity for dextromethorphan and metoprolol, respectively. When diplotypes were reassigned based on the Stargazer algorithm that utilizes both SNV and structural variation data, these numbers increased to approximately 41% and 36% for dextromethorphan and metoprolol, respectively. Samples with incorrect AS assignments contained alleles that contribute to "extreme" phenotypes: the *CYP2D6*5* allele harboring a deletion of the entire gene or alleles carrying two or more functional gene copies including *CYP2D6*1xN* and *2xN*. The *CYP2D6*4N*, **36*, **68*, and **13* hybrid alleles are nonfunctional and thus did not affect AS assignments when in tandem arrangements with alleles identified via SNV only analysis. A *CYP2D6*68+*4* allele for example, receives a value of 0 regardless whether it is called as *CYP2D6*68+*4* or as *CYP2D6*4*. Identifying these tandem structures accurately is still important, however, since not all hybrid alleles are nonfunctional. Stargazer is also informative when resolving potentially ambiguous diplotypes. For example, many commercially available tests may detect a duplication in an individual with a *CYP2D6*1/*4* diplotype, but would be unable to determine if the additional gene copy is on the **1* or the **4* allele. Consequently, it cannot be determined whether the patient has a functional or nonfunctional gene duplication.

Only one diplotype was irreconcilable between the two assignment methods. Using SNV data alone, the sample appeared to be homozygous for 100C>T and heterozygous for 1846G>A (positions according to M3338819) and was called *CYP2D6*4/*10*. Stargazer analysis, however, determined that this sample had a copy number of 1 and called *CYP2D6*4/*5* based on the presence of 1846G>A. The heterozygous variant call, which showed a significant allele imbalance, was most likely due to a mapping error and interpreted as hemizygosity by Stargazer.

The sample's CYP2D6 activity (<0.5 mmol/mg protein/min for both probe drugs) supports a *CYP2D6*4/*5* call.

The metabolite formation rates for samples with an AS=3 were lower than expected and the small number of samples (N=3) with functional gene duplications makes interpretation difficult. One sample in particular, genotyped as a *CYP2D6*1/*1x2*, had very low CYP2D6 activity for both probe drugs. It is possible that this was due to poor liver tissue and/or human liver microsome (HLM) quality, a hypothesis supported by the observation that *CYP3A4* activity, determined using the same batch of HLMs (unpublished data), was also much lower than expected. In general, variation not accounted for in this study may be from phenoconversion caused by liver donors' concomitant medications, genetic variation in genes not covered by PGRNseq, or distant regulatory SNVs such as rs5758550, which has recently been described to impact CYP2D6 expression levels [**106**,**107**,**108**].

Of the 113 SNVs detected in the liver samples that are not listed in haplotype translation tables [**96**,**97**], 73 are within the boundaries of *CYP2D6* haplotype definitions, 17 have not yet been assigned a rs number, and 10 were missense SNVs, thus are high priorities for further characterization. We evaluated the catalytic activity of 7 rare nonsynonymous variants and showed that A449D causes decreased activity similar to P34S (100C>T), a diagnostic SNV for the decreased function *CYP2D6*10* allele. In contrast, R474Q exhibited increased activity in comparison to the reference enzyme. These nonsynonymous variants were evaluated for function individually in the reference background, and not within the context of their haplotype, which remains unknown. Nevertheless, these findings highlight the value of gene resequencing when predicting phenotype. At the population level, the consideration of rare variation may enhance clinical outcomes for precision pharmacotherapy.

We found that POR protein content, required for CYP enzyme function, was also correlated with CYP2D6 activity. This is consistent with our collaborators' reports that POR is correlated with CYP2C19 and CYP2A6 activity in these samples [**102**,**103**]. The inclusion of POR pharmacogenetic data in the clinical setting may improve CYP2D6 phenotype prediction compared to *CYP2D6* genotyping alone. Indeed, a study examining the effects of *POR* variants on CYP2D6 activity found that the common A503V amino acid change with allele frequencies ranging from approximately 19% in African Americans to 37% in Chinese Americans, is associated with a decrease in CYP2D6 activity of 40-50% when using dextromethorphan and bufuralol as probes [**101**].

AKR1D1 mRNA content was also significantly correlated with CYP2D6 activity. AKR1D1 is involved in bile acid homeostasis and is hypothesized to regulate CYPs through activation of nuclear receptors that transcriptionally regulate CYP expression. Bile acids are ligands of nuclear receptors PXR and CAR, which induce the transcription of *CYP3A4*, *CYP2C9*, and *CYP2C19*. CYP2D6 is generally considered to be non-inducible, so this mechanism of CYP regulation should theoretically not affect CYP2D6 activity. Perhaps some other, yet unknown, mechanism applies.

A multiple linear regression model was used to assess the combined contributions of CYP2D6 AS, POR protein content, *AKR1D1* mRNA content, and liver procurement site to variability in CYP2D6 activity. Together, these factors explained 51% and 47% of the variability in CYP2D6 activity when dextromethorphan and metoprolol were used as probes, respectively. Due to the lack of racial diversity in the liver bank donors, we were unable to assess for differences in activity possibly explained by race, although such associations are generally attributed to differences in the frequency of altered function alleles [**49**,**65**].

In summary, for a gene locus as complex as *CYP2D6*, analysis of SNVs alone is clearly not sufficient for making accurate diplotype calls and AS predictions. Most currently available *CYP2D6* pharmacogenetic tests are designed to detect the most common alleles and interrogate the presence of select structural variants (e.g. deletions/duplications). The combination of Stargazer and targeted next-generation sequencing offers a solution to many of the limitations of existing tests. First, it can detect rare alleles typically not included on targeted genotyping panels. Second, the approach identifies the nature of the duplicated gene copy rather than detecting a nonspecific, and potentially ambiguous, duplication signal. Third, Stargazer can identify hybrid gene structures, including those in tandem arrangements with non-hybrid alleles, solving the problem of incorrect duplication calls or other miscalls due to the interference of these complex structures. If we are to strive for precision in the therapeutic treatment of disease, a feasible first step, as described in this study, is to consider the full complement of structural variants, SNVs, and indels for *CYP2D6* diplotype assignment. As techniques such as multi-omics data integration mature, phenotype predictions can be improved with additional data on variation in proteins related to CYP2D6 function, such as POR and AKR1D1.

3.5 Materials and Methods

Chemicals and reagents

Dextromethorphan, metoprolol tartrate, NADPH, potassium phosphate, EDTA, and acetonitrile were purchased from Sigma-Aldrich (St Louis, MO, USA). Dextrorphan-d3 and Carvedilol-d3 were purchased from Cerilliant (Round Rock, TX, USA) and C/D/N Isotopes (Pointe-Claire, QC, CA), respectively.

Human liver samples

Human liver tissue samples (N=314) were obtained from the University of Washington Human Liver Bank (N=48) and the St. Jude Liver Resource at St. Jude Children's Research Hospital (N=266). Institutional Review Boards at the University of Washington and St. Jude Children's Research Hospital approved the collection and use of these samples for research. All links between archived tissues and donors were destroyed, further details described previously [**102**]. Preparation and total protein quantitation of HLMs were previously described [**102**].

Protein and mRNA quantitation

Microsomal CYP2D6 and POR protein content was quantitated using a surrogate peptide-based LC-MS/MS method [**102**,**109**]. Methods for mRNA quantitation were previously described [**103**].

DNA isolation, *CYP2D6* sequencing, RNAseq analysis, and *CYP2D6* allele and activity score assignment

Genomic DNA was isolated from human livers as described [**102**]. The PGRNseq platform (v.1.1) was used to identify *CYP2D6* genomic variation. PGRNseq captures all *CYP2D6* exons as well as 2 kb of upstream and 1 kb of downstream sequence. It also captures the *CYP2D7* pseudogene [**64**]. Allele and subsequent star-allele diplotype assignments were first made manually using SNV data only. Allele definitions are according to the Pharmacogene Variation Consortium (PharmVar) [**58**,**110**]. Manual diplotype assignment was performed by cross-referencing SNVs found in each individual sample to the *CYP2D6* allele definitions on PharmVar, taking a general approach described previously [**85**]. The Stargazer algorithm was then used to detect *CYP2D6* structural variation in PGRNseq data and the star-allele diplotype

assignments determined manually were corrected accordingly [**56**]. AS were assigned to each allele based on CPIC criteria [**57**,**86**,**87**,**88**,**89**,**90**,**91**].

CYP2D6 metabolite formation rate in HLMs

HLM incubations were conducted using a metabolite formation approach to estimate intrinsic clearance, defined as maximum velocity (V_{max}) divided by Michaelis-Menten constant (K_m) [**111**,**112**,**113**]. Optimization experiments confirmed that substrate concentrations were below K_m, incubation times and total HLM protein content were within linear conditions of metabolite formation, and <10% substrate was depleted over the course of the incubation (data not shown). Incubations were performed in triplicate using 20 μg of total HLM protein diluted in 100 mM potassium phosphate, pH 7.4 and 1 mM EDTA buffer. HLMs were pre-incubated with substrate (1.5 μM dextromethorphan or 4 μM metoprolol) for 5 min at 37°C. NADPH (final concentration 1 mM) was added to start the reaction and samples were incubated for 20 min (metoprolol) and 30 min (dextromethorphan) at 37°C, respectively. Reactions were terminated by the addition of ice-cold acetonitrile.

Metabolite and parent drug quantitation

Quantitation of dextromethorphan, dextrorphan, metoprolol, and α-hydroxymetoprolol was performed with modifications to published methods [**114**,**115**,**116**] on an Agilent Technologies G1956B mass spectrometer coupled to an Agilent 1200 series HPLC using a Zorbax SB-C18 2.1 mm x 150 mm x 5 μm column (Agilent Technologies, Santa Clara, CA). The column was maintained at 30°C with a flow rate of 0.3 ml/min for dextrorphan, and 35°C with a flow rate of 0.25 ml/min for α-hydroxymetoprolol. Quantitation methods are detailed in **Supplementary Methods**.

Functional characterization of rare *CYP2D6* variants

Rare variants were characterized *in vitro* using a cytochrome P450 *S. cerevisiae* model; detailed in **Supplementary Methods** [**117**,**118**,**119**,**120**,**121**,**122**]. Methods include construction of yeast expression plasmids, induction of three control and seven novel CYP2D6 variant genes, and functional characterization of variants using both a click chemistry-compatible ticlopidine 5'-carboxypropargyl amide probe (tic-ABP1P) that has specificity for CYP2D6 activity with minimum reactivity towards other yeast proteins and a fluorogenic CYP2D6 substrate, Vivid™ 7-ethoxymethoxy-3-cyanocoumarin.

3.6 Acknowledgements

This work was supported by the Northwest Alaska-Pharmacogenomics Research Network (NWA-PGRN) (U01GM092676 and P01GM116691); the Center for Exposures, Diseases, Genomics, and Environment (EDGE) (P30ES007033); and NIGMS R24GM115277. M.J.D. is a Senior Fellow in the Genetic Networks program at the Canadian Institute for Advanced Research and is supported in part by a Faculty Scholars grant from the Howard Hughes Medical Institute. D.M.F. is a CIFAR Azrieli Global Scholar. The authors thank Linda Risler at the Clinical Pharmacokinetics Laboratory at the University of Washington for training in HLM incubations and Rachel Tyndale at the University of Toronto for discussions of AKR1D1 associations with P450 activities.

Chapter 4 Calling Star Alleles with Stargazer in 28 Pharmacogenes with Whole Genome Sequences

This chapter has been submitted for publication: Lee, S.B., Wheeler, M.M., Thummel, K.E. & Nickerson, D.A. (2019).

4.1 Summary

Variation in the enzymatic activity of pharmacogenes is defined by star alleles (haplotypes) comprised of single nucleotide variants, small insertion-deletions, and large structural variants. We recently developed Stargazer, a next-generation sequencing-based tool to call star alleles for the clinically important *CYP2D6* gene. Here, we present the utility of extending Stargazer to call star alleles for 28 pharmacogenes using whole genome sequencing (WGS) data. We applied Stargazer to WGS data from 70 ethnically diverse samples from the Genetic Testing Reference Materials Coordination Program (GeT-RM). These reference samples were extensively characterized by GeT-RM using multiple pharmacogenetic testing assays. In all 28 genes, Stargazer recalled 100% of star alleles (N=92) present in GeT-RM's consensus genotypes (N=1559). Stargazer also detected star alleles not previously reported by GeT-RM including complex structural variants. Our results demonstrate that combining WGS data and Stargazer enables automated, accurate, and comprehensive genotyping of pharmacogenes in the human genome.

4.2 Introduction

Genetic variation contributes significantly to the wide inter-individual variability in pharmacological responses and gives rise to differences in systemic drug exposure, safety, and efficacy [**3**]. Not accounting for this genetic variation can lead to severe adverse reactions or a loss of efficacy, due to inappropriate drug choice and/or dosing [**62**,**123**]. For example, multiple

loss-of-function variants in the *CYP2C9* gene can greatly diminish drug metabolism by blocking enzyme synthesis or reducing its catalytic function [**124**,**125**]. Individuals who are homozygous for these variants are called CYP2C9 poor metabolizers and are at risk of abnormal bleeding if prescribed the average dose of anticoagulant, warfarin [**126**].

Pharmacogenetic (PGx) testing offers the potential for precision drug therapy, through the combination of genetic information and corresponding drug response phenotypes. Optimal pharmacotherapy can be determined by PGx testing to increase the overall efficacy and prevent adverse drug reactions [**30**]. The Food and Drug Administration provides additional guidance by requiring applicable PGx test information be included in the drug labeling [**43**]. However, to date, broad implementation of PGx testing has met several challenges and only a few PGx tests are currently routinely used in the clinic [**44**].

A major barrier to broad implementation has been the complexity of many pharmacogenes. Several genes require that PGx testing include a large number of genetic variants to provide accurate predictions of enzymatic activity [**46**]. For example, the clinically important *CYP2D6* gene has more than 100 star alleles (haplotypes) defined by single nucleotide variants (SNVs), small insertion-deletions (indels), and/or large structural variants (SVs) [**47**]. These *CYP2D6* alleles encode enzymes with normal, decreased, increased, or no function, which translate to inferred clinical phenotypes that range from ultrarapid to poor metabolism [**48**]. Importantly, the frequency of star alleles and phenotypes can vary across different populations [**49**], highlighting the need for comprehensive variant testing.

Another major challenge has been that a large fraction of existing star alleles cannot be accurately assessed with a single methodology. As stated above, *CYP2D6* alleles include SVs such as deletions, duplications, and complex gene hybrids. Many of these SVs are difficult to

detect due to high sequence homology (>95%) with a nearby non-functional paralog [**50**]. Thus, several orthogonal genotyping methods including TaqMan assays, long-range PCR, quantitative multiplex PCR, High Resolution Melt analysis, and Sanger sequencing are required to accurately call all SVs in *CYP2D6* [**51**]. These methods do reliably detect the star alleles needed for clinical application but can be time consuming and biased towards the detection of known SVs.

The Centers for Disease Control and Prevention-based Genetic Testing Reference Materials Coordination Program (GeT-RM) has established genomic DNA reference materials to help the genetic testing community obtain characterized reference materials [**127**]. A GeT-RM collaborative project recently published genotyping results for 137 ethnically diverse Coriell DNA samples and 28 pharmacogenes [**128**,**129**]. These samples were genotyped using several commercial and laboratory-developed PGx testing assays [**128**]. More recently, GeT-RM has made whole genome sequencing (WGS) data for 70 of the 137 reference samples publicly available [**129**].

In this study, we utilized the genotyping results from GeT-RM and the available WGS data to continue the development of a new next-generation sequencing-based tool. We extended the SV-aware algorithm of Stargazer [**56**] to assess star alleles in 28 pharmaogenes (**Table 4.1**). Among these genes, *CYP2A6*, *GSTM1*, *GSTT1*, and *UGT2B17* are known to display extensive gene deletion polymorphisms [**130**]. Additionally, *CYP2A6*, *CYP2B6*, and *CYP2D6* have been shown to frequently exhibit complex SVs, which include gene hybrids with their paralogs *CYP2A7*, *CYP2B7*, and *CYP2D7*, respectively [**131**,**132**,**95**]. To evaluate the accuracy of this algorithm, we compared star alleles detected by Stargazer to those previously reported by GeT-RM. In addition, we provide an in-depth characterization of the WGS data of these 70 reference samples, which includes the identification of star alleles not tested in previous genotyping efforts.

In order to verify Stargazer's SV calls, we also explored the Database of Genomic Variants (DGV) [**70**] for variant reports submitted by various studies including the 1000 Genomes Project (1KGP) [**133**].

4.3 Results

Evaluating Stargazer's genotyping accuracy

We applied Stargazer to assess 1960 genotypes in 28 pharmacogenes in 70 WGS samples from GeT-RM. To estimate the accuracy of Stargazer, we compared these genotypes with those previously published by GeT-RM [**128**]. For these samples, GeT-RM reported a total of 1559 consensus genotypes comprised of 92 star alleles (**Table 4.1**). These consensus genotypes were verified by two or more PGx testing assays [**128**]. In all 28 genes, Stargazer recalled 100% of star alleles present in GeT-RM's consensus genotypes (**Supplementary Table 4.1**).

Table 4.1 Star alleles previously reported by GeT-RM and assessed by Stargazer's analysis of whole genome sequencing

Gene	Reference Allele	Star Alleles Found in Consensus GeT-RM Genotypes (N=92)	Star Alleles Only Found in Non-consensus GeT-RM Genotypes (N=31)
CYP1A1	**1*	**2, *4, *5*	*none*
CYP1A2	**1A*	**1C, *1F, *1L*	*none*
CYP2A6	**1*	**2, *4 (del), *9, *17, *20*	**8*
CYP2B6	**1*	**2, *6, *7, *18*	**4, *5, *15, *20, *22, *27*
CYP2C8	**1*	**2, *3, *4*	*none*
CYP2C9	**1*	**2, *3, *5, *6, *8, *9, *11*	**18*
CYP2C19	**1*	**2, *3, *4, *8, *13, *15, *17*	**6, *27*
CYP2D6	**1*	**2, *2x2 (dup), *4, *5 (del), *6, *9, *10, *14, *15, *17, *29, *35, *41, *xN (dup)*	**21, *36+*10 (hyb), *40*
CYP2E1	**1*	**7*	**4, *5*
CYP3A4	**1*	**1B, *2, *3, *22*	**15, *16*
CYP3A5	**1*	**3, *6, *7*	*none*
CYP4F2	**1*	**2, *3*	*none*
DPYD	**1*	**9*	**4*
GSTM1	**A*	**B, *0 (del)*	*none*
GSTP1	**A*	**B, *C, *D*	*none*
GSTT1	**A*	**0 (del)*	**AxN (dup), *B*
NAT1	**4*	**11, *14, *17*	*none*
NAT2	**4*	**5, *6, *7, *14*	**12, *13*
SLC15A2	**1*	**2*	*none*
SLC22A2	**1*	**3, *6, *7*	**2, *K432Q*
SLCO1B1	**1A*	**1B, *5, *14, *15, *17, *21*	*none*
SLCO2B1	**1*	*none*	**S464F*
TPMT	**1*	**3C, *8*	*none*
UGT1A1	**1*	**6, *28, *60*	**7, *27, *36, *37*
UGT2B7	**1*	**2*	**3*
UGT2B15	**1*	**2, *5*	**4*
UGT2B17	**1*	**2 (del)*	*none*

VKORC1	**1*	**2, *3, *4*	*none*

Structural variant-defined alleles are indicated by "del" (deletion), "dup" (duplication), and "hyb" (hybrid).

WGS confirmation of star alleles present in 'non-consensus' GeT-RM genotypes

A subset of GeT-RM genotypes (N=401) could not be verified by multiple PGx testing assays previously (**Supplementary Table 4.1**). This is either because certain star alleles were tested by a single method or multiple test results disagreed [**128**]. A total of 31 star alleles were only found in these 'non-consensus' genotypes (**Table 4.1**). Stargazer's output confirmed the presence of most of these star alleles with the exception of four alleles: *CYP2A6*8*, *CYP2B6*27*, *CYP2C9*18*, and *GSTT1*AxN*. These four alleles were not present in the WGS data (**Supplementary Table 4.2**). For example, GeT-RM predicted eight samples to contain a *GSTT1* duplication (*GSTT1*AxN*) using the Agena Bioscience iPLEX ADME PGx Pro Panel (Agena Bioscience, San Diego, CA). Nonetheless, Stargazer's output showed that 3 of these samples had a normal copy number of two and the remaining 5 samples had a deletion (*GSTT1*0*) instead (**Supplementary Figure 4.1**). To provide validation for these deletion events, we searched DGV and found that four of these five samples were also previously shown by 1KGP to have copy number loss in *GSTT1* (**Supplementary Table 4.3**).

SV-defined alleles previously undercalled by GeT-RM

Stargazer's output showed that three gene deletions (*CYP2A6*4*, *GSTM1*0*, and *GSTT1*0*) and one *CYP2D6/CYP2D7* hybrid (*CYP2D6*36+*10*) were previously under-reported by GeT-RM (**Figure 4.1**; **Table 4.2**). For example, GeT-RM tested gene deletions in *GSTM1* using the Affymetrix DMET Plus Array (Affymetrix, Santa Clara, CA, USA) and the Agena Bioscience iPLEX ADME PGx Pro Panel. Both assays identified 32 samples with homozygous deletions but found zero samples with heterozygous deletions. In contrast, Stargazer detected both heterozygous and homozygous deletions in 21 and 32 samples, respectively. By cross-

referencing to DGV reports from 1KGP, we validated copy number loss in *GSTM1* for 13 of the 21 samples with heterozygous deletions (**Supplementary Table 4.3**). Similarly, we used 1KGP's DGV reports to verify Stargazer's genotype calls for samples with *CYP2A6*4* (N=2) and *GSTT1*0* (N=13) (**Supplementary Table 4.3**).

Table 4.2 Star alleles with structural variation previously undercalled by GeT-RM

Star Allele	Assays[a]	N of Heterozygotes From GeT-RM	N of Homozygotes From GeT-RM	N of Heterozygotes From Stargazer	N of Homozygotes From Stargazer
*CYP2A6*4*	[1, 2]	4	2	7	2
*CYP2D6*36+*10*	[2, 3]	7	0	4	4
*GSTM1*0*	[1, 2]	0	32	21	32
*GSTT1*0*	[1, 2]	16	18	36	18

[a][1] Affymetrix DMET Plus Array (Affymetrix, Santa Clara, CA); [2] Agena Bioscience iPLEX ADME PGx Pro Panel (Agena Bioscience, San Diego, CA); [3] Agena Bioscience iPLEX ADME CYP2D6 Panel (Agena Bioscience, San Diego, CA).

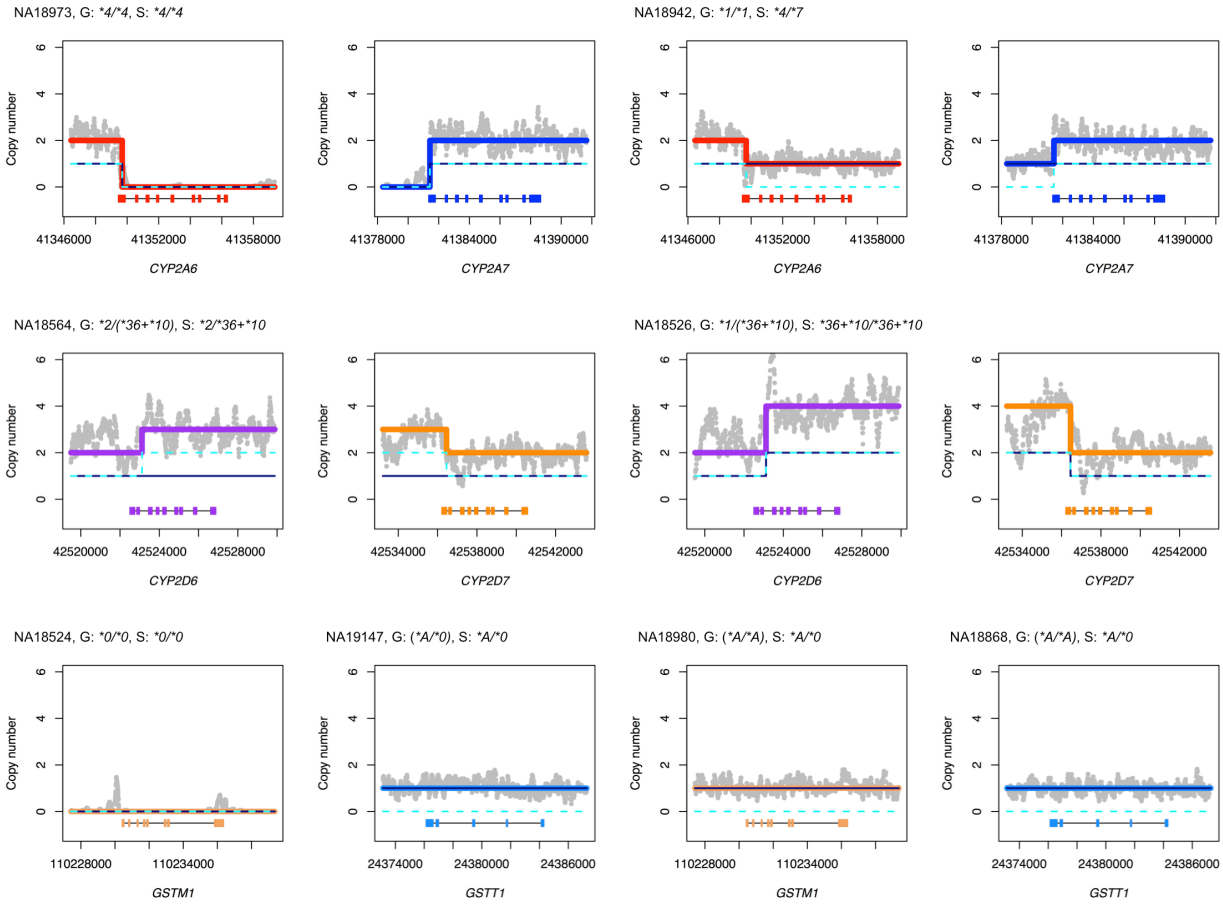


Figure 4.1 Examples of star alleles with structural variation undercalled by GeT-RM
Panels display Stargazer's result for copy number analysis for individual samples (N=8). Genotypes from GeT-RM and Stargazer (abbreviated as "G" and "S" for brevity) are also shown, with "()" indicating non-consensus genotypes. The left and right panels exhibit samples whose structural variant calls are matched and not matched, respectively. Grey dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined.

Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. Reports in the Database of Genomic Variants supported Stargazer's gene deletion calls in NA18942, NA18980, and NA18868.

WGS identification of additional star alleles not previously reported by GeT-RM

Using the WGS data, Stargazer detected 38 additional star alleles not previously reported by GeT-RM (**Table 3**). These alleles were found in 127 of 1960 genotypes assessed (**Supplementary Table 4.1**). Seven of these alleles contained SVs and were comprised of five gene duplications, one *CYP2B6/CYP2B7* hybrid, and one *CYP2D6/CYP2D7* hybrid (**Figure 4.2**). Of the five duplications, only *CYP2A6*1x2*, *CYP2D6*2x2*, and *CYP2D6*4x2* are currently listed in existing PGx databases, suggesting that the remaining two, *CYP2D6*34x2* and *CYP2E1*7x2*, may be novel. *CYP2D6*34x2* was identified from a single African sample (NA19207) and *CYP2E1*7x2* was identified from three African samples (NA19095, NA19226, and NA19908); note that NA19908 was homozygous for *CYP2E1*7x2* (**Figure 4.2**). DGV reports support the presence of *CYP2A6*1x2* in NA18861 (DGV gold standard; copy number gain observed by multiple studies) and *CYP2E1*7x2* in NA19908 (copy number gain observed by 1KGP) (**Supplementary Table 4.3**).

Table 4.3 Star alleles identified by Stargazer's analysis of whole genome sequencing and not previously reported by GeT-RM

Gene	Star Alleles (N=38)
CYP1A1	**2A, *2B, *13*
CYP2A6	**1x2 (dup), *7, *15, *18, *19, *21, *22, *23, *24, *25, *35*
CYP2B6	**17, *23, *29 (hyb)*
CYP2C19	**35*
CYP2D6	**4x2 (dup), *34, *34x2 (dup), *39, *46, *68+*4 (hyb)*
CYP2E1	**7x2 (dup)*
DPYD	**5, *6*
GSTM1	**Ax2 (dup)*
NAT1	**3, *10, *26*
SLC22A2	**4*
SLCO1B1	**24, *27, *30, *31, *35*
TPMT	**16*

Structural variant-defined alleles are indicated by "dup" (duplication) and "hyb" (hybrid).

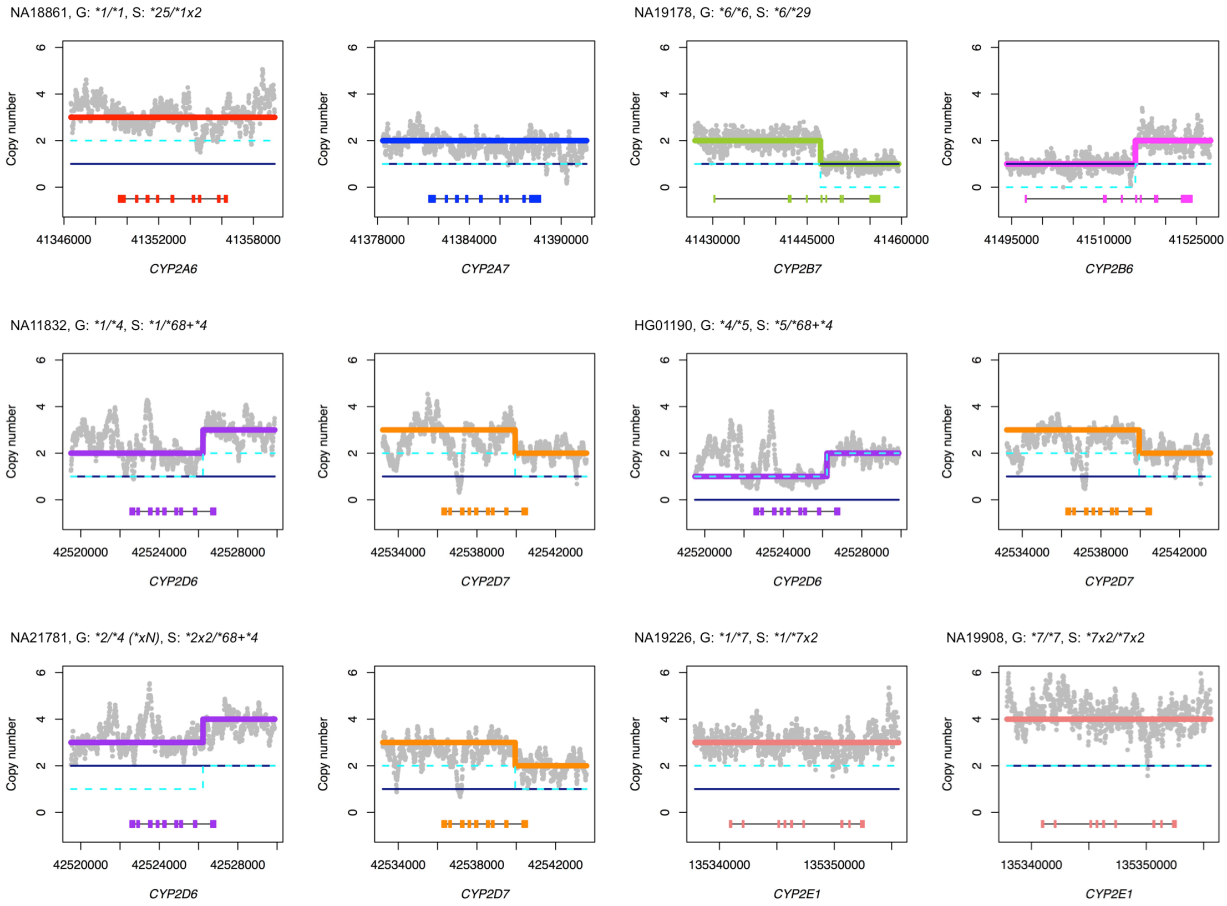


Figure 4.2 Examples of star alleles with structural variation not reported by GeT-RM
Panels display Stargazer's results for copy number analysis for individual samples (N=7). Genotypes from GeT-RM and Stargazer (abbreviated as "G" and "S" for brevity) are also shown, with "()" indicating non-consensus genotypes. Grey dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined. Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. Stargazer's genotype call for HG01190 involving two different SVs (a *CYP2D6* deletion and a *CYP2D6/CYP2D7* hybrid) was independently verified previously.[19] Reports in the Database of Genomic Variants supported Stargazer's gene duplication calls in NA18861 and NA11908.

Surprisingly, Stargazer detected three gene copies for the *CYP3A4*, *CYP3A5*, *UGT2B7*, and *UGT2B15* genes in a single sample (NA18540). *CYP3A4* and *CYP3A5* are 77 kbp apart on chromosome 7 while *UGT2B7* and *UGT2B15* are 426 kbp apart on chromosome 4. GeT-RM did not report any SVs in these genes for this sample or other samples tested [**128**]. Copy number analyses using Stargazer showed no breakpoints in flanking genomic regions (**Supplementary Figure 4.2**), indicating that this sample likely has chromosomal trisomy for chromosomes 4 and 7 (i.e., *CYP3A4*1/*1/*1*, *CYP3A5*1/*1/*3*, *UGT2B7*1/*1/*2*, and *UGT2B15*2/*2/*4*). This

result has been independently confirmed through karyotyping by Redon et al., 2006, which have additionally revealed trisomy in chromosomes 9, 14, and 21 [**134**]. This aberrant karyotype most likely arose during cell immortalization.

Statistical phasing of SNVs/indels for star alleles

Using statistical phasing [**71**] with the 1KGP haplotype reference panel [**135**], Stargazer revised a total of 64 GeT-RM genotypes (**Supplementary Table 4.1**). For instance, both GeT-RM and Stargazer found four heterozygous SNVs in 14 samples that were indicative of the *CYP1A2*1A/*1L* or **1C/*1F* genotype. GeT-RM reported both genotypes as equally likely, while the phasing algorithm indicated the *CYP1A2*1A/*1L* genotype to be more likely. In addition, Stargazer revised GeT-RM genotypes in two related samples, NA12156 (mother) and NA10831 (child), to follow expected inheritance patterns. As an example, GeT-RM and Stargazer genotyped the mother as *UGT1A1*28/*60* and **1/*28, *60*, respectively. The mother's correct genotype should be the latter because the child was genotyped as *UGT1A1*28, *60/*28, *60* by both GeT-RM and Stargazer.

Resolving ambiguous *CYP2D6* duplications using WGS allelic depth

Both GeT-RM and Stargazer found seven samples with a gene duplication in the *CYP2D6* gene. For four of these samples, GeT-RM reported genotypes containing gene duplications of an unspecified star allele (*CYP2D6*xN*), while Stargazer resolved these ambigous duplications using allelic depth of WGS reads (**Supplementary Table 4.1**). For instance, GeT-RM genotyped the sample NA19819 as *CYP2D6*2/*4/*xN* because the sample contained three gene copies as well as the *CYP2D6*2* (normal function) and **4* (no function) alleles. In contrast, Stargazer

called the sample as *CYP2D6*2/*4x2*, a genotype that was previously independently verified for this sample [**56**].

New star alleles defined with WGS findings

Stargazer identified SNVs, indels, and SVs not present in existing haplotype translation tables. These variants represent nine new star alleles and were found in a total of 20 Stargazer genotypes in a population-specific manner (**Table 4**; **Supplementary Table 4.1**). More specifically, functional annotation of detected SNVs/indels added four new star alleles defined by a frameshift variant, a splice site variant, and two nonsense variants (**Table 4.4**). Copy number analyses with Stargazer also identified five SVs (**Figure 4.3**). To enable automated detection of these SVs, we defined and included five new star alleles as part of the Stargazer algorithm (**Table 4.4**). We also found DGV reports supporting *CYP2E1*S1* in NA19143 (copy number gain observed by [**136**]) and *SLC22A2*S2* in NA19226 and NA19819 (DGV gold standard; copy number loss observed by multiple studies) (**Supplementary Table 4.3**).

Table 4.4 New star alleles discovered by Stargazer's whole genome analysis.

Star Allele	Description[a]	N of African Samples	N of East Asian Samples	N of European Samples
*CYP2A6*1+*S6*	duplication of *CYP2A7* (chr19:41358125-41389907)	0	1	0
*CYP2E1*S1*	duplication in exons 7-9 (chr10:135350465-135439323)	4	0	0
*SLC22A2*S1*	deletion in intron 9 (chr6:160649735-160654861)	3	0	0
*SLC22A2*S2*	deletion affecting 3'-UTR (chr6:160627751-160638068)	2	0	0
*UGT2B15*S1*	deletion affecting exons 4-6 (chr4:69510975-69528283)	0	0	1
*CYP2C9*S1*	nonsense (rs756250160A>T)	0	1	0
*SLCO1B1*S1*	nonsense (rs183501729C>T)	0	1	0
*SLCO1B1*S2*	splice site (rs77271279G>T)	2	0	0
*SLCO2B1*S1*	frameshift (rs72408262GCACAGAAAA>G)	0	1	4

[a]Genomic coordinates and nucleotide changes are according to Human Genome version 19.

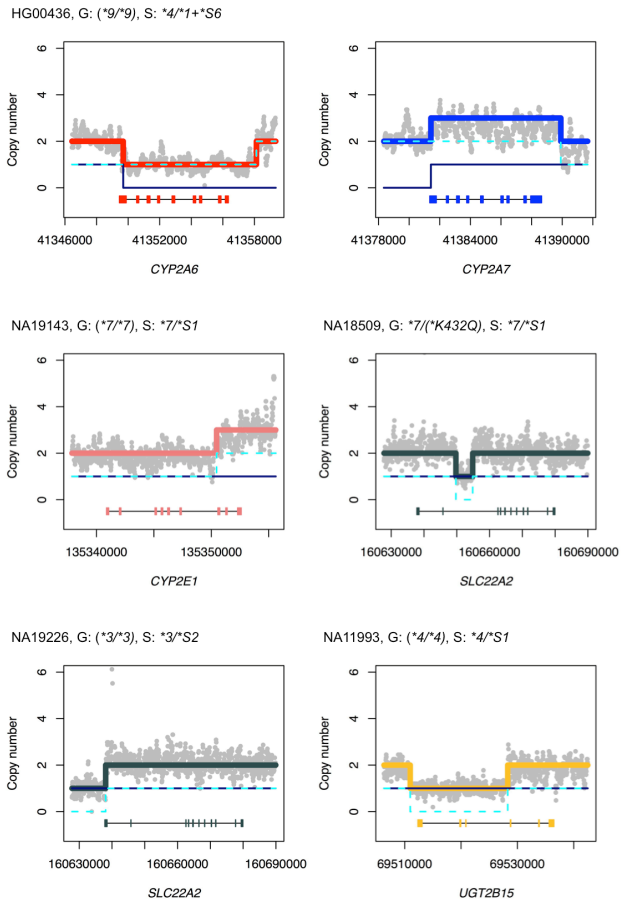


Figure 4.3 Examples of new star alleles with structural variation
Panels display Stargazer's results for copy number analysis for individual samples (N=5). Genotypes from GeT-RM and Stargazer (abbreviated as "G" and "S" for brevity) are also shown, with "()" indicating non-consensus genotypes. Grey dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined. Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. *CYP2A6*1+*S6* was identified in only one sample (HG00436) exhibiting both a *CYP2A6* deletion (*CYP2A6*4*) and a duplication in the *CYP2A7* paralog (*CYP2A6*1+*S6*). Reports in the Database of Genomic Variants supported Stargazer's partial duplication call in NA19143 and partial deletion call in NA19226.

4.4 Discussion

Here, we present an extension of the Stargazer algorithm to call star alleles in 28 pharmacogenes from next-generation sequencing data. Stargazer is one of the first bioinformatics tools that enable systematic identification of star alleles (e.g., Cypiripi [**74**], Astrolabe [**75**], PharmCAT [**137**], Aldy [**138**]). However, Stargazer is the only tool that uses statistical haplotype phasing [**71**], which is informed by population haplotype frequencies to call star alleles more accurately. In addition, other tools tend to have difficulties with the detection of complex SVs, such as

CYP2D6/CYP2D7 hybrids [**74**,**75**,**137**]. Lastly, to our knowledge, Stargazer is the most comprehensive tool available and assesses star alleles in more genes than has previously been reported [**137**,**138**].

To evaluate the performance of Stargazer, we utilized public WGS data from 70 genotyping reference samples from GeT-RM. These samples were extensively characterized using multiple standard methods (e.g., allele-specific PCR, molecular inversion probes, hybridization-based arrays, and TaqMan assays) [**128**]. To verify Stargazer's SV calls, we explored DGV reports from various sources including 1KGP. In all 28 genes, Stargazer recalled 100% of star alleles present in GeT-RM's consensus genotypes. Stargazer also identified additional star alleles not previously reported by GeT-RM, including both known and novel SVs, and correctly found trisomies of the chromosomes 4 and 7 in the sample NA18540. Altogether, these results demonstrate that Stargazer has high sensitivity for the detection of SVs and can accurately assess star alleles in these 28 genes.

With statistical phasing, Stargazer revised star alleles in reference samples that previously had ambiguous or incorrect GeT-RM genotypes. In the current version of Stargazer, we incorporated the 1KGP haplotype reference panel to increase sample size and improve phase accuracy [**139**]. This approach performed well for our dataset, but we are aware that further applications may be challenged by low-frequency variants and limited by the magnitude and extent of linkage disequilibrium [**140**]. To ameliorate this issue, we plan to merge multiple large, high-quality reference panels to obtain additional haplotype information.

CYP2D6 duplications have been reported for normal function, decreased function, and non-functional alleles [**141**,**142**,**143**]. GeT-RM previously reported ambiguous *CYP2D6* genotypes involving duplication of an unspecified allele (*CYP2D6*xN*). For instance, the sample

NA19819 was genotyped as *CYP2D6*2/*4/*xN* by GeT-RM where *CYP2D6*2* and **4* are normal function and non-functional alleles, respectively. Therefore, the sample could tentatively be *CYP2D6*2x2/*4* or **2/*4x2*, which would predict two completely different phenotypes—normal metabolizer and intermediate metabolizer, respectively. By using allelic depth of WGS reads, Stargazer correctly called the *CYP2D6*2/*4x2* genotype for the sample [**56**] and resolved other genotypes with ambiguous *CYP2D6* duplications. Furthermore, both GeT-RM and Stargazer predicted the samples HG00436, NA19109, and NA19226 to be ultrarapid metabolizers (*CYP2D6*1/*2x2*, **29/*2x2*, and **2/*2x2*, respectively), which is a major phenotypic consequence of carrying *CYP2D6* gene duplications. Collectively, these results highlight that allelic decomposition performed by Stargazer enables accurate phenotype prediction for samples with gene duplications.

We report nine new star alleles defined by variants discovered in WGS data. Although the enzymatic activity of these alleles remains to be functionally characterized, seven alleles likely have an impact on enzyme activity. *UGT2B15*S1*, for instance, is likely non-functional because it includes deletion of the last three exons of the *UGT2B15* gene. Another new allele, *CYP2A6*1+*S6*, contains a gene duplication in the paralog *CYP2A7*. The duplication does not directly affect the *CYP2A6* sequence, but it could still change CYP2A6 activity because CYP2A7 transcript level has been shown to alter CYP2A6 expression via competition for miRNA binding [**144**]. Conversely, the partial gene deletions in the *SLC22A2*S1* and **S2* alleles do not affect the translated region of the *SLC22A2* gene and thus are unlikely to have a functional consequence.

As of April 2019, there are 359 gene/drug pairs (e.g., *CYP2D6*/codeine) described by the Clinical Pharmacogenetics Implementation Consortium with accompanying levels of

recommendation for changing drug choice and dosing decisions [**8**]. The 28 pharmacogenes currently targeted by Stargazer include 132 of these gene/drug pairs. We plan to further extend Stargazer to additional pharmacogenes, including 26 PGx loci whose consensus genotypes could not be determined in Pratt et al., 2016 because they were only characterized by one laboratory.

In summary, by leveraging WGS data, we confirmed the consensus results reported by GeT-RM and expanded the current PGx variation catalogs for the 70 important reference samples. Therefore, our WGS characterization can be added to this public reference resource for other PGx genotyping projects. As sequencing costs continue to decline and as the clinical value of whole genome (or targeted panel) sequencing continues to emerge, there will be an increasing need for systematic and highly efficient analysis algorithms. Our results show that WGS data combined with Stargazer offers a feasible path for accurate PGx testing and the optimization of individual drug treatment responses.

4.5 Materials and Methods

PGx genotypes and WGS data from GeT-RM

We accessed PGx genotypes and WGS data for 70 ethnically diverse Coriell DNA samples from GeT-RM (**Supplementary Table 4.1**; **Supplementary Table 4.4**). Both PGx genotypes and WGS data are publically available through the GeT-RM website [**129**]. GeT-RM generated consensus and non-consensus genotypes for 28 pharmacogenes using a variety of testing platforms, as detailed in Pratt et al., 2016. WGS was performed to a depth of >30X using paired-end 150 bp sequence reads on the Illumina HiSeq X (Illumina, San Diego, CA, USA). WGS data were downloaded in the BAM file format, which contains sequence reads aligned to Human Genome version 19 with the program ISAAC [**145**].

Extension of Stargazer to 28 pharmacogenes

The first version of Stargazer (v1.0.0) included a haplotype translation table for >100 star alleles in the *CYP2D6* gene [**56**]. Extension of Stargazer (v1.0.4) involved construction of 27 additional haplotype translation tables for >500 star alleles. Star allele information was compiled from several public PGx databases: the Pharmacogene Variation Consortium [**58**], the Pharmacogenomics Knowledgebase [**146**], the *UGT* database [**147**], the *NAT* database [**148**], and the *TPMT* database [**149**]. Generation of haplotype translation tables involved lifting cDNA coordinates of PGx variants to genomic coordinates from Human Genome version 19. The new version of Stargazer and all haplotype translation tables are available for download: https://stargazer.gs.washington.edu/stargazerweb/.

Description of Stargazer's algorithm

Stargazer's algorithm has been described previously [**56**]. Briefly, for this extended version, SNVs/indels in each gene were assessed from a VCF file generated from BAM files using GATK-HaplotypeCaller [**69**]. The VCF file was phased using the program Beagle [**71**] with the 1KGP haplotype reference panel [**135**]. Phased SNVs/indels were then matched to star alleles in each gene's haplotype translation table. BAM files were also used to calculate read depth using GATK-DepthOfCoverage [**69**]. Read depth was converted to copy number by intra-sample normalization. Following normalization, SVs were detected by testing pairwise combinations of expected haplotype copy number profiles against the sample's observed copy number profile for both haplotypes. SV results were incorporated to inform the final star allele assignment. Output data of Stargazer included individual genotypes and copy number plots to visually inspect SVs calls (see **Figure 4.1** for examples of these plots).

Assessment of differences in genotype calls between GeT-RM and Stargazer

Differences in genotype calls between GeT-RM and Stargazer were carefully evaluated by considering the consistency of star allele assignment across individual PGx testing assays as well as assessment of WGS reads. WGS reads were assessed through visual inspection using Integrative Genomics Viewer [**150**], as exemplified in **Supplementary Table 4.2**. For validation of Stargazer's results involving SVs, we used existing reports available in DGV. A total of 33 Stargazer genotypes were verified this way (26 samples overlapped with DGV) (**Supplementary Table 4.3**).

Assessment of novel variation by Stargazer

Stargazer's output includes a VCF file of detected SNVs/indels not present in existing PGx databases. This VCF file is functionally annotated using SeattleSeq Annotation [**151**]. Functional annotation enables prediction of nonsynonymous variants, which may impact enzyme activity. For samples with SVs that do not match expected copy number profiles, Stargazer performs change point analysis [**72**]. This analysis identifies approximate breakpoints, which are then used to identify subsequent SVs with similar breakpoints in copy number profiles.

4.6 Acknowledgements

The authors acknowledge the Genetic Testing Reference Materials Coordination Program (GeT-RM) for their generous contribution of pharmacogenetic genotypes and whole genome sequencing data. This work was supported by the National Institute of General Medical Sciences (NIGMS) (R24GM115277, P50GM115318 and P01GM116691). S.B.L. is a recipient of Macrogen PhD Fellowship.

Chapter 5 Summary and Future Directions

5.1 Research Summary

The fundamental requirement of precision medicine is the ability to correctly genotype all medically important genes on a clinically relevant timescale. Here, genotyping means the comprehensive detection of all types of genetic variation—single nucleotide variants (SNVs), small insertion-deletions (indels), and large structural variants (SVs)—and careful interpretation of detected variants in the context of haplotype assembly (e.g., star alleles). This dissertation highlights that many of the key pharmacogenes that determine drug response phenotypes are genetically complex and highly polymorphic. Thus, they require several orthogonal methods for accurate genotyping (e.g., TaqMan assays, long-range PCR, quantitative multiplex PCR, High Resolution Melt analysis, and Sanger sequencing) [**51**,**52**]. However, these methods and the interpretation of results are expensive and laborious, which has hindered the development of any scalable and systematic pharmacogenetic (PGx) analysis framework.

Next-generation sequencing (NGS) technologies allow massively parallel sequencing of thousands of genes with high-throughput data generation, comprehensive genotyping capabilities, and ever-decreasing cost [**55**]. Therefore, NGS can serve as a powerful PGx analysis platform if there is an accompanying algorithm that researchers and physicians can apply to interpret the massive amounts of data generated by NGS. The preceding chapters exemplify how an SV-aware calling algorithm I have developed, Stargazer, can accurately genotype various polymorphic pharmacogenes using NGS data. I validated Stargazer's genotyping accuracy for 28 pharmacogenes using characterized reference materials, and will continue to extend the algorithm to additional genes. The work presented in this dissertation also shows that the combination between NGS and Stargazer allows the identification and quantification of rare and

novel PGx variation, including complex SVs. This represents a tremendous advantage over traditional genotyping techniques mentioned above, which tend to be biased toward the detection of known PGx variation. Taken together, NGS and Stargazer offer an optimal path for accurate and systematic PGx genotyping and prediction of individual drug responses. This approach will be increasingly useful in clinical practice, particularly as whole genome sequencing (WGS) and targeted panel sequencing become more widely available.

Despite its high genotyping accuracy, the current version of Stargazer is limited in the detection of certain classes of variation that may be relevant to improved PGx interpretation. First, since Stargazer relies on read-depth signal to detect SVs, it cannot detect SV events that do not involve change in read depth such as inversions and translocations. Although these events have not been reported yet in the 28 genes targeted by Stargazer, in the future we will update Stargazer's SV detection algorithm to incorporate additional types of NGS signals, including split-read and paired-end, which are produced by programs like LUMPY [**152**]. Second, at present Stargazer uses small-size (< 10 bp) indels called by GATK-HaplotypeCaller [**69**], but accurate calling of intermediate-size (10-50 bp) indels is much more challenging due to the short NGS reads. Therefore, we plan to employ programs like IMSindel [**153**] that performs *de novo* assembly and gapped global-local alignment with split-read analysis to identify intermediate-size indels more reliably than GATK-HaplotypeCaller alone. Third, the current Stargazer genotyping pipeline could miss variants resulting from somatic mosaicism unless sequencing was performed at ultra-deep coverage for the accurate identification and quantification of mosaicism that may have PGx implications [**154**].

5.2 Exploiting Large and Diverse Genomic Projects for Pharmacogenomic Discovery

Although substantial progress has been made in the identification and implementation of clinically actionable PGx variants, the majority of their drug phenotype associations were discovered in cohorts of European individuals. For example, warfarin is an anticoagulant drug that has a narrow therapeutic window with risks of bleeding or thrombosis and its dosing requirements are heavily influenced by genetic polymorphisms in pharmacogenes such as *CYP2C9* and *VKORC1* [**39**]. A study recently showed that of 12 prospective randomized trials investigating genotype-guided warfarin dosing, individuals of European ancestry accounted for 79.9% of all study populations combined, whereas Latinos accounted for 2%, African Americans for 9.5%, and individuals of Asian ancestry for 6.9% [**155**]. The lack of exhaustive studies in non-European populations raised the concern about whether algorithms developed in European populations can be generalized to more-diverse populations. This concern became reality when numerous reports came out showing the poor performance of European-derived warfarin dosing algorithms in African Americans, Brazilians, and Caribbean Hispanics [**156**,**157**]. There could be multiple reasons, but the most likely explanation is that European-derived algorithms missed important variants in other ethnic groups (e.g., *CYP2C9*8*) (**Table 5.1**). In addition, PGx variants may have differing effects depending on the ethnic group, resulting in the use of incorrect effect sizes and incorrect dose estimates.

Table 5.1 Genetic variants that contribute to warfarin dose requirements and their frequencies in diverse ethnic groups
This table is reproduced from Table 1 in [**155**].

Allele	rs ID	Warfarin Dose	European	Latino	African	Asian	American Indian/Alaska Natives
*CYP2C9*2*	rs1799853	Decreased	10	7	2	<1	5
*CYP2C9*3*	rs1057910	Decreased	7	4	<1	3	3
*CYP2C9*5*	rs28371686	Decreased	<1	<1	2	<1	-
*CYP2C9*6*	rs9332131	Decreased	<1	<1	1	<1	-
*CYP2C9*8*	rs7900194	Decreased	<1	<1	5	<1	<1
*CYP2C9*11*	rs28371685	Decreased	<1	<1	2	<1	<1
CYP2C9 18786T	rs7089580	Increased	22	12	21	1	-
VKORC1 -1639G>A	rs9923231	Decreased	39	41	5	88	60
VKORC1 -8191A>G	rs61162043	Increased	61	57	46	12	-

Enhanced discovery of PGx variants in under-studied populations could incentivize the development of dosing algorithms that benefit diverse ethnic groups. To this end, large-scale and

multi-ethnic genomic projects can serve as valuable resources for closing the gap in PGx data (**Table 5.2**). For example, the Trans-Omics for Precision Medicine (TOPMed) program is comprised of 144 thousand participants, from >80 different studies, consisting of approximately 60% with substantial non-European ancestry [**158**]. To demonstrate the importance of PGx discovery in large and diverse sample sets, I recently used Stargazer to perform a *CYP2D6* genotype analysis of African American individuals (N=3,418) from the Jackson Heart Study that were whole genome sequenced by the TOPMed program [**159**]. From the WGS data, I found five novel star alleles comprised of gene duplications (*CYP2D6*29x3*, **34x2*, and **42x2*) and multiplications (*CYP2D6*1x3* and **2x3*). All of these alleles were never reported before and could result in increased enzymatic activity. This analysis is currently being expanded to other ethnic cohorts in the TOPMed program.

Table 5.2 Large-scale and multi-ethnic genomic projects

Project	Country	Size	Website
International HapMap Project	International	1,301	http://www.hapmap.org
1000 Genomes Project (1KGP)	International	2,504	http://www.internationalgenome.org
Exome Sequencing Project (ESP)	USA	6,503	https://esp.gs.washington.edu
Exome Aggregation Consortium (ExAc)	USA	60,706	http://exac.broadinstitute.org
100,000 Genomes Project	UK	100,000	https://www.genomicsengland.co.uk
Trans-Omics for Precision Medicine (TOPMed)	USA	144,000	https://www.nhlbiwgs.org
UK Biobank	UK	500,000	https://www.ukbiobank.ac.uk
All of Us (AoU)	USA	1,000,000	https://allofus.nih.gov

5.3 High-throughput Functional Characterization of Pharmacogenetic Variation

As NGS expands to millions of diverse individuals, we need new scalable and systematic techniques for assaying the effects of genetic variation on protein function, as the interpretation of these variants is difficult without functional characterization. For example, there are currently 61 star alleles defined for the *CYP2C9* gene, but only 14 of them are functionally annotated (e.g., decreased activity), and the rest have either uncertain or unknown function [**160**]. Even for those alleles with functional annotation, because star alleles are defined by haplotypes and, thus, can have more than one variant, it is often difficult to know which of the variants in a given star

allele is responsible for the allele's altered activity—plus, it could very well be a real combinatorial effect of two or more variants. Unfortunately, the trend described above is true for most of the pharmacogenes that have star alleles. Further, the algorithms designed to provide clinical decision support, those that interpret and report PGx test results, must be able to do so for all potential variants, even those that have yet to be observed.

To this end, several tools have been developed to summarize large amounts of functional genomic data (e.g., evolutionary conservation, gene model, histone or transcription factor ChIP-seq signals) into scores that can be used to predict consequences of coding and noncoding variants, such as CADD score [**161**]. Although these scores can be informative, they are limited in their accuracy [**162**]. Therefore, there is clearly a need to experimentally test the function of the rapidly increasing number of PGx variants to facilitate more comprehensive phenotype prediction. Recent advances in protein engineering and methodology may enable the high-throughput functional analysis necessary for the development of such reporting systems.

One technique, deep mutational scanning (DMS), allows massively parallel measurement of the functional consequences of all possible single-amino-acid substitutions for a given enzyme. This is accomplished through the coupling of rounds of selection on a library of protein variants with NGS of individual DNA barcodes [**80**]. The change in the frequency of each variant from input library to post-selection serves as a measure of its function. DMS can also be extended to create different combinations of multiple mutations for testing the enzymatic activity of various haplotypes. Another technique, massively parallel reporter assay (MPRA), is an efficient way of assessing the functional effects of thousands of SNVs and single nucleotide deletions simultaneously for a single regulatory element (i.e., the enhancer or promoter of a gene) [**163**]. Like DMS, MPRA also involves saturation mutagenesis and construction of

libraries; however, the main difference is that MPRA uses the expression of a reporter gene (e.g., luciferase) as the functional output. Although both DMS and MPRA are limited in their naturalness, well-characterized PGx variants (e.g., the reduced function *CYP2C9*2* allele with the missense mutation R144C for DMS and the -1639G>A promoter variant in *VKORC1* [**164**,**165**] for MPRA) could serve as positive controls and be used to construct a model to predict the actual effect of any allele.

5.4 Extending Stargazer for Long-read Sequencing Data

Thus far, this dissertation has focused on Illumina's short-read sequencing data (50 to 300 base pairs). However, there exist two other sequencing platforms, known as third-generation sequencing, that can produce much longer reads (>10,000 base pairs) and are becoming more and more mainstream: Oxford Nanopore Technologies and Pacific Biosciences. The former applies an ionic current through nanopores and measures the changes in current as the DNA molecule passes through the nanopore [**166**]. The latter captures light pulses emitted during the replication process of the DNA molecule in a small sequencing chamber [**167**].

I am planning to extend Stargazer to support long-read sequencing data from these two technologies. One major advantage of long-read sequencing is the generation of fully phased alleles, which would remove Stargazer's need for statistical haplotype phasing when calling star alleles. For example, both technologies have recently been successfully applied to sequence the entire *CYP2D6* gene at a single pass (**Figure 5.1**) [**78**,**79**]. This approach could also be used to generate more accurate haplotype information of rare variants for statistical phasing. Another important advantage of long-read sequencing is the inclusion of large repeat elements, such as the Alu-rich regions downstream of *CYP2D6* and *CYP2D7* that are thought to be responsible for generating a wide variety of common SVs in *CYP2D6* [**73**,**168**]. Despite their significance, these

repeated regions cannot be properly assembled by short-read-based methods. In contrast, long-read sequencing will allow the investigation of inter-individual variation within these regions, which could boost our understanding of the evolution of the *CYP2D* locus and result in better drug phenotype prediction because some of the repeat elements have been shown to regulate gene expression [**169**]. Finally, long-read sequencing allows direct detection of SVs, whereas short-read sequencing must rely on some sort of inference (**Figure 5.2**) [**170**]. For example, currently Stargazer converts read depth signal to copy number to detect gene deletions, duplications, and hybrids; however, long-read sequencing can directly capture these SV events in individual reads. This will dramatically increase the confidence and precision of SV calling with more accurate breakpoints. Potential drawbacks of long-read sequencing include its higher base-calling error rate and far lower throughput when compared to Illumina's short-read sequencing [**171**]. However, note that Pacific Biosciences' latest circular consensus sequencing has recently been shown to generate high-fidelity (HiFi) reads that match short reads in terms of accuracy for small variant detection, while enabling SV detection and *de novo* assembly at similar contiguity and markedly higher concordance than noisy long reads [**172**].

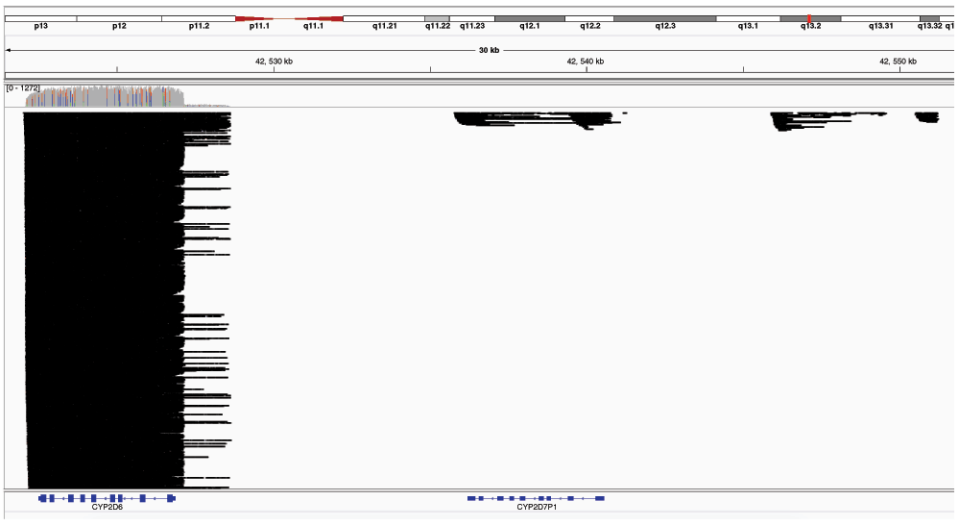


Figure 5.1 Oxford Nanopore Technologies sequence reads aligned to the *CYP2D6*, *CYP2D7*, and *CYP2D8* genes
The majority of reads aligned across the entire length of *CYP2D6* as was expected by selective PCR amplification. Downstream, an insignificant number of read fragments aligned to *CYP2D7* and *CYP2D8* (*CYP2D8* is located from 42,545,874 to 42,551,097;

exon-intron diagram not shown in gene annotation track). Due to the extremely high coverage at *CYP2D6*, not all reads are shown in this pileup diagram. This figure is reproduced from Figure 1 in [**78**].

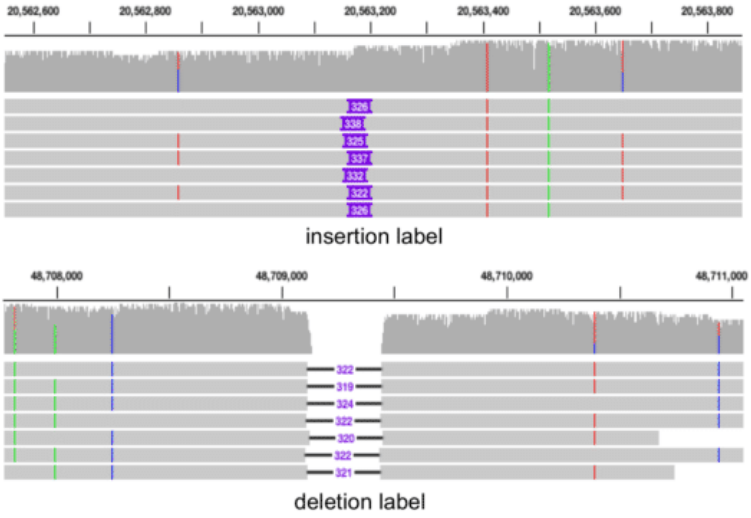


Figure 5.2 Direct visualization of structural variation detected by Pacific Biosciences sequencing technology
(Top) An insertion larger than the defined threshold is indicated by a purple box. The width of the box is proportional to the size of the insertion, and the base pair size is written on the box if it fits. (Bottom) A deletion is indicated by a black line. The base pair size of the deletion is written on a white box at the center of the line. Examples are from HG002 sequenced by Genome in a Bottle. This figure is reproduced from Figure 2 in [**170**].

5.5 The $1,000 Genome, the $1,000,000 Interpretation

Although clinical integration of NGS technology is closer to reality than ever before, the lack of algorithms that can interpret the massive amounts of data produced by NGS still remains as a major challenge for precision medicine. Bruce R. Korf, former president of the American College of Medical Genetics, puts it this way: "We are close to having a $1,000 genome sequence, but this may be accompanied by a $1,000,000 interpretation" [**173**]. This could not be more true for pharmacogenetics, which has relied on the manual assignment of haplotypes from NGS data to identify star alleles and predict drug responses. However, the volume and complexity of NGS data make the manual method cumbersome and prone to error, demanding the need to transform the "mental algorithm" into a bioinformatics tool. My program Stargazer meets this need and enables fast, automated, and accurate genotyping of pharmacogenes from NGS data. Furthermore, PGx genotyping with NGS and Stargazer to predict individual drug responsiveness is only one aspect of precision medicine, and this approach can be expanded to

other biomedical fields that require systematic and accurate genotype analysis (e.g., blood typing, HLA typing). Hence, pharmacogenetics serves a test system for issues that will be faced across many aspects of precision medicine. As algorithms like Stargazer continue to adapt to the challenges presented by NGS, the lessons learned in overcoming these barriers will provide a valuable roadmap as precision medicine expands into all aspects of preventive and diagnostic care.

Bibliography

[1] RM Turner, BK Park, and M Pirmohamed, "Parsing interindividual drug variability: an emerging role for systems pharmacology," *Wiley Interdiscip Rev Syst Biol Med*, vol. 7, no. 4, pp. 221-41, Jul-Aug 2015.

[2] IMS Institute for Healthcare Informatics. (2015, November) Global Medicines Use in 2020. [Online]. https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/global-medicines-use-in-2020

[3] WE Evans and MV Relling, "Moving towards individualized medicine with pharmacogenomics," *Nature*, vol. 429, no. 6990, pp. 464-8, May 2004.

[4] BB Spear, M Heath-Chiozzi, and J Huff, "Clinical application of pharmacogenetics," *Trends Mol Med*, vol. 7, no. 5, pp. 201-4, May 2001.

[5] N Shehab et al., "US Emergency Department Visits for Outpatient Adverse Drug Events, 2013-2014," *JAMA*, vol. 316, no. 20, pp. 2115-2125, Nov 2016.

[6] J Sultana, P Cutroneo, and G Trifirò, "Clinical and economic burden of adverse drug reactions," *J Pharmacol Pharmacother*, vol. 4, no. Suppl 1, pp. S73-7, Dec 2013.

[7] GD Sweeney, "Variability in the human drug response," *Thromb Res Suppl*, vol. 4, pp. 3-15, 1983.

[8] Clinical Pharmacogenetics Implementation Consortium. (2019, March) Genes-Drugs. [Online]. https://cpicpgx.org/genes-drugs/

[9] EV Koonin and YI Wolf, "Constraints and plasticity in genome and molecular-phenome evolution," *Nat Rev Genet*, vol. 11, no. 7, pp. 487-98, Jul 2010.

[10] UM Zanger and M Schwab, "Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation," *Pharmacol Ther*, vol. 138, no. 1, pp. 103-41, Apr 2013.

[11] DW Nebert, K Wikvall, and WL Miller, "Human cytochromes P450 in health and disease," *Philos Trans R Soc Lond B Biol Sci*, vol. 368, no. 1612, p. 20120431, Jan 2013.

[12] DR Nelson, JV Goldstone, and JJ Stegeman, "The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s," *Philos Trans R Soc Lond B Biol Sci*, vol. 368, no. 1612, p. 20120474, Jan 2013.

[13] H Wang, KM Donley, DS Keeney, and SM Hoffman, "Organization and evolution of the Cyp2 gene cluster on mouse chromosome 7, and comparison with the syntenic human cluster," *Environ Health Perspect*, vol. 111, no. 15, pp. 1835-42, Nov 2003.

[14] R Feyereisen, "Arthropod CYPomes illustrate the tempo and mode in P450 evolution," *Biochim Biophys Acta*, vol. 1814, no. 1, pp. 19-28, Jan 2011.

[15] WL Miller and RJ Auchus, "The molecular biology, biochemistry, and physiology of human steroidogenesis and its disorders," *Endocr Rev*, vol. 32, no. 1, pp. 81-151, Feb 2011.

[16] H Sezutsu, G Le Goff, and R Feyereisen, "Origins of P450 diversity," *Philos Trans R Soc Lond B Biol Sci*, vol. 368, no. 1612, p. 20120428, Jan 2013.

[17] NL Kirischian and JY Wilson, "Phylogenetic and functional analyses of the cytochrome P450 family 4," *Mol Phylogenet Evol*, vol. 62, no. 1, pp. 458-71, Jan 2012.

[18] DR Nelson et al., "Comparison of cytochrome P450 (CYP) genes from the mouse and

human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants," *Pharmacogenetics*, vol. 14, no. 1, pp. 1-18, Jan 2004.

[19] JV Goldstone et al., "Cytochrome P450 1 genes in early deuterostomes (tunicates and sea urchins) and vertebrates (chicken and frog): origin and diversification of the CYP1 gene family," *Mol Biol Evol*, vol. 24, no. 12, pp. 2619-31, Dec 2007.

[20] Y Zhou, M Ingelman-Sundberg, and VM Lauschke, "Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects," *Clin Pharmacol Ther*, vol. 102, no. 4, pp. 688-700, Oct 2017.

[21] RE Janha et al., "Inactive alleles of cytochrome P450 2C19 may be positively selected in human evolution," *BMC Evol Biol*, vol. 14, p. 71, Apr 2014.

[22] M Ingelman-Sundberg, "Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity," *Pharmacogenomics J*, vol. 5, no. 1, pp. 6-13, 2005.

[23] EE Thompson et al., "CYP3A variation and the evolution of salt-sensitivity variants," *Am J Hum Genet*, vol. 75, no. 6, pp. 1059-69, Dec 2004.

[24] M Schirmer et al., "Genetic signature consistent with selection against the CYP3A4*1B allele in non-African populations," *Pharmacogenet Genomics*, vol. 16, no. 1, pp. 59-71, Jan 2006.

[25] Dobon B, C Rossell, S Walsh, and J Bertranpetit, "Is there adaptation in the human genome for taste perception and phase I biotransformation?," *BMC Evol Biol*, vol. 19, no. 1, p. 39, Jan 2019.

[26] DW Nebert, "Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist?," *Clin Genet*, vol. 56, no. 4, pp. 247-58, Oct 1999.

[27] PE Carson, CL Flanagan, CE Ickes, and AS Alving, "Enzymatic deficiency in primaquine-sensitive erythrocytes," *Science*, vol. 124, no. 3220, pp. 484-5, Sep 1956.

[28] AG Motulsky, "Drug reactions, enzymes, and biochemical genetics," *J Am Med Assoc*, vol. 165, no. 7, pp. 835-7, Oct 1957.

[29] F Vogel, "Moderne problem der humangenetik," *Ergeb. Inn. Med. U. Kinderheik.*, vol. 12, pp. 52-125, 1959.

[30] MV Relling and WE Evans, "Pharmacogenomics in the clinic," *Nature*, vol. 526, no. 7573, pp. 343-50, Oct 2015.

[31] ES Vesell and JG Page, "Genetic control of drug levels in man: antipyrine," *Science*, vol. 161, no. 3836, pp. 72-3, Jul 1968.

[32] FJ Gonzalez et al., "Characterization of the common genetic defect in humans deficient in debrisoquine metabolism," *Nature*, vol. 331, no. 6155, pp. 442-6, Feb 1988.

[33] M Ingelman-Sundberg, "Pharmacogenomic biomarkers for prediction of severe adverse drug reactions," *N Engl J Med*, vol. 358, no. 6, pp. 637-9, Feb 2008.

[34] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860-921, Feb 2001.

[35] DF Carr, A Alfirevic, and M Pirmohamed, "Pharmacogenomics: Current State-of-the-Art," *Genes (Basel)*, vol. 5, no. 2, pp. 430-43, May 2014.

[36] R Weinshilboum, "Inheritance and drug response," *N Engl J Med*, vol. 348, no. 6, pp. 529-

37, Feb 2003.

[37] M Pirmohamed, "Pharmacogenetics and pharmacogenomics," *Br J Clin Pharmacol*, vol. 52, no. 4, pp. 345-347, Oct 2001.

[38] SL Van Driest et al., "Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing," *Clin Pharmacol Ther*, vol. 95, no. 4, pp. 423-31, Apr 2014.

[39] J Moaddeb and SB Haga, "Pharmacogenetic testing: Current Evidence of Clinical Utility," *Ther Adv Drug Saf*, vol. 4, no. 4, pp. 155-169, Aug 2013.

[40] J Taube, D Halsall, and T Baglin, "Influence of cytochrome P-450 CYP2C9 polymorphisms on warfarin sensitivity and risk of over-anticoagulation in patients on long-term treatment," *Blood*, vol. 96, no. 5, pp. 1816-9, Sep 2000.

[41] S Rost et al., "Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2," *Nature*, vol. 427, no. 6974, pp. 537-41, Feb 2004.

[42] RS Epstein et al., "Warfarin genotyping reduces hospitalization rates results from the MM-WES (Medco-Mayo Warfarin Effectiveness study)," *J Am Coll Cardiol*, vol. 55, no. 25, pp. 2804-12, Jun 2010.

[43] Food and Drug Administration. (2018, August) Table of Pharmacogenomic Biomarkers in Drug Labeling. [Online]. https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm

[44] AK Daly and I Cascorbi, "Opportunities and limitations: the value of pharmacogenetics in clinical practice," *Br J Clin Pharmacol*, vol. 77, no. 4, pp. 583-6, Apr 2014.

[45] JD Robarge, L Li, Z Desta, A Nguyen, and DA Flockhart, "The star-allele nomenclature: retooling for translational genomics," *Clin Pharmacol Ther*, vol. 82, no. 3, pp. 244-8, Sep 2007.

[46] JJ Swen et al., "Pharmacogenetics: from bench to byte--an update of guidelines," *Clin Pharmacol Ther*, vol. 89, no. 5, pp. 662-73, May 2011.

[47] SF Zhou, "Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I," *Clin Pharmacokinet*, vol. 48, no. 11, pp. 689-723, 2009.

[48] A Gaedigk et al., "The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype," *Clin Pharmacol Ther*, vol. 83, no. 2, pp. 234-42, Feb 2008.

[49] A Gaedigk, K Sangkuhl, M Whirl-Carrillo, T Klein, and JS Leeder, "Prediction of CYP2D6 phenotype from genotype across world populations," *Genet Med*, vol. 19, no. 1, pp. 69-76, Jan 2017.

[50] A Gaedigk, "Complexities of CYP2D6 gene analysis and interpretation," *Int Rev Psychiatry*, vol. 25, no. 5, pp. 534-53, Oct 2013.

[51] A Gaedigk et al., "Cytochrome P4502D6 (CYP2D6) gene locus heterogeneity: characterization of gene duplication events," *Clin Pharmacol Ther*, vol. 81, no. 2, pp. 242-51, Feb 2007.

[52] WE Kramer et al., "CYP2D6: novel genomic structures and alleles," *Pharmacogenet Genomics*, vol. 19, no. 10, pp. 813-22, Oct 2009.

[53] J Zhang, R Chiodini, A Badr, and G Zhang, "The impact of next-generation sequencing on genomics," *J Genet Genomics*, vol. 38, no. 3, pp. 95-109, Mar 2011.

[54] S Goodwin, JD McPherson, and WR McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat Rev Genet*, vol. 17, no. 6, pp. 333-51, May 2016.

[55] R Nielsen, JS Paul, A Albrechtsen, and YS Song, "Genotype and SNP calling from next-generation sequencing data," *Nat Rev Genet*, vol. 12, no. 6, pp. 443-51, Jun 2011.

[56] SB Lee et al., "Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model," *Genet Med*, vol. 21, pp. 361-372, 2019.

[57] Clinical Pharmacogenetics Implementation Consortium, "Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update," *Clin Pharmacol Ther*, vol. 95, no. 4, pp. 376-82, Apr 2014.

[58] PharmVar Steering Committee, "The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database," *Clin Pharmacol Ther*, vol. 103, no. 3, pp. 399-401, Mar 2018.

[59] A LLerena et al., "Interethnic variability of CYP2D6 alleles and of predicted and measured metabolic phenotypes across world populations," *Expert Opin Drug Metab Toxicol*, vol. 10, no. 11, pp. 1569-83, Nov 2014.

[60] J Kirchheiner et al., "Pharmacokinetics of codeine and its metabolite morphine in ultra-rapid metabolizers due to CYP2D6 duplication," *Pharmacogenomics J*, vol. 7, no. 4, pp. 257-65, Aug 2007.

[61] Y Gasche et al., "Codeine intoxication associated with ultrarapid CYP2D6 metabolism," *N Engl J Med*, vol. 351, no. 27, pp. 2827-31, Dec 2004.

[62] G Koren, J Cairns, D Chitayat, A Gaedigk, and SJ Leeder, "Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother," *Lancet*, vol. 368, no. 9536, p. 704, Aug 2006.

[63] ME Naranjo et al., "High frequency of CYP2D6 ultrarapid metabolizers in Spain: controversy about their misclassification in worldwide population studies," *Pharmacogenomics J*, vol. 16, no. 5, pp. 485-90, Oct 2016.

[64] AS Gordon et al., "PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation," *Pharmacogenet Genomics*, vol. 26, no. 4, pp. 161-168, Apr 2016.

[65] A Gaedigk, LD Bradford, SW Alander, and JS Leeder, "CYP2D6*36 gene arrangements within the cyp2d6 locus: association of CYP2D6*36 with poor metabolizer status," *Drug Metab Dispos*, vol. 34, no. 4, pp. 563-9, Apr 2006.

[66] A Gaedigk et al., "CYP2D7-2D6 hybrid tandems: identification of novel CYP2D6 duplication arrangements and implications for phenotype prediction," *Pharmacogenomics*, vol. 11, no. 1, pp. 43-53, Jan 2010.

[67] A Gaedigk, GP Twist, and JS Leeder, "CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR," *Pharmacogenomics*, vol. 13, no. 1, pp. 91-111, Jan 2012.

[68] H Li and R Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754-60, Jul 2009.

[69] A McKenna et al., "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res*, vol. 20, no. 9, pp. 1297-303, Sep 2010.

[70] JR MacDonald, R Ziman, RK Yuen, L Feuk, and SW Scherer, "The Database of Genomic Variants: a curated collection of structural variation in the human genome," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D986-92, Jan 2014.

[71] SR Browning and BL Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering," *Am J Hum Genet*, vol. 81, no. 5, pp. 1084-97, Nov 2007.

[72] R Killick and IA Eckley, "changepoint: an R package for changepoint analysis," *J Stat Softw*, vol. 58, no. 3, pp. 1-19, June 2014.

[73] A Gaedigk et al., "Identification of Novel CYP2D7-2D6 Hybrids: Non-Functional and Functional Variants," *Front Pharmacol*, vol. 1, p. 121, Oct 2010.

[74] I Numanagić et al., "Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 12, pp. i27-34, Jun 2015.

[75] GP Twist et al., "Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences," *NPJ Genom Med*, vol. 1, p. 15007, Jan 2016.

[76] T Kubota et al., "CYP2A6 polymorphisms are associated with nicotine dependence and influence withdrawal symptoms in smoking cessation," *Pharmacogenomics J*, vol. 6, no. 2, pp. 115-9, Mar-Apr 2006.

[77] T Fukami, M Nakajima, H Sakai, HL McLeod, and T Yokoi, "CYP2A7 polymorphic alleles confound the genotyping of CYP2A6*4A allele," *Pharmacogenomics J*, vol. 6, no. 6, pp. 401-12, Nov-Dec 2006.

[78] R Ammar, TA Paton, D Torti, A Shlien, and GD Bader, "Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes," *F1000Res*, vol. 4, p. 17, Jan 2015.

[79] W Qiao et al., "Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6," *Hum Mutat*, vol. 37, no. 3, pp. 315-23, Mar 2016.

[80] DM Fowler and S Fields, "Deep mutational scanning: a new style of protein science," *Nat Methods*, vol. 11, no. 8, pp. 801-7, Aug 2014.

[81] BM Kuehn, "FDA: No codeine after tonsillectomy for children," *JAMA*, vol. 309, no. 11, p. 1100, Mar 2013.

[82] CA Prows et al., "Codeine-related adverse drug reactions in children following tonsillectomy: a prospective study," *Laryngoscope*, vol. 124, no. 5, pp. 1242-50, May 2014.

[83] RS Gammal et al., "Pharmacogenetics for Safe Codeine Use in Sickle Cell Disease," *Pediatrics*, vol. 138, no. 1, p. e20153479, Jul 2016.

[84] SJ Gardiner and EJ Begg, "Pharmacogenetics, drug-metabolizing enzymes, and clinical practice," *Pharmacol Rev*, vol. 58, no. 3, pp. 521-90, Sep 2006.

[85] VM Lauschke and M Ingelman-Sundberg, "Prediction of drug response and adverse drug reactions: From twin studies to Next Generation Sequencing," *Eur J Pharm Sci*, vol. 130, pp. 65-77, Mar 2019.

[86] 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*, vol. 98, no. 2,

pp. 127-34, Aug 2015.

[87] JK Hicks et al., "Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update," *Clin Pharmacol Ther*, vol. 102, no. 1, pp. 37-44, Jul 2017.

[88] GC Bell et al., "Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 genotype and use of ondansetron and tropisetron," *Clin Pharmacol Ther*, vol. 102, no. 2, pp. 213-218, Aug 2017.

[89] MP Goetz et al., "Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and Tamoxifen Therapy," *Clin Pharmacol Ther*, vol. 103, no. 5, pp. 770-777, May 2018.

[90] Clinical Pharmacogenetics Implementation Consortium, "Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants," *Clin Pharmacol Ther*, vol. 93, no. 5, pp. 402-8, May 2013.

[91] Clinical Pharmacogenetics Implementation Consortium, "Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype," *Clin Pharmacol Ther*, vol. 91, no. 2, pp. 321-6, Feb 2012.

[92] JK Hicks, JJ Swen, and A Gaedigk, "Challenges in CYP2D6 phenotype assignment from genotype data: a critical assessment and call for standardization," *Curr Drug Metab*, vol. 15, no. 2, pp. 218-32, Feb 2014.

[93] Y Yang, MR Botton, ER Scott, and SA Scott, "Sequencing the CYP2D6 gene: from variant allele discovery to clinical pharmacogenetic testing," *Pharmacogenomics*, vol. 18, no. 7, pp. 673-685, May 2017.

[94] A Fohner et al., "Pharmacogenetics in American Indian populations: analysis of CYP2D6, CYP3A4, CYP3A5, and CYP2C9 in the Confederated Salish and Kootenai Tribes," *Pharmacogenet Genomics*, vol. 23, no. 8, pp. 403-14, Aug 2013.

[95] JL 3rd Black, DL Walker, DJ O'Kane, and M Harmandayan, "Frequency of undetected CYP2D6 hybrid genes in clinical samples: impact on phenotype prediction," *Drug Metab Dispos*, vol. 40, no. 1, pp. 111-9, Jan 2012.

[96] PharmGKB. (2019, Jan.) Gene-specific Information Tables for CYP2D6.

[97] PharmVar. (2019, Jan.) CYP2D6. [Online]. https://www.pharmvar.org/gene/CYP2D6

[98] M Ning, JD Duarte, LH Rubin, and H Jeong, "CYP2D6 Protein Level Is the Major Contributor to Interindividual Variability in CYP2D6-Mediated Drug Metabolism in Healthy Human Liver Tissue," *Clin Pharmacol Ther*, vol. 104, no. 5, pp. 974-982, Nov 2018.

[99] X Yang et al., "Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver," *Genome Res*, vol. 20, no. 8, pp. 1020-36, Aug 2010.

[100] SN Hart et al., "Genetic polymorphisms in cytochrome P450 oxidoreductase influence microsomal P450-catalyzed drug metabolism," *Pharmacogenet Genomics*, vol. 18, no. 1, pp. 11-24, Jan 2008.

[101] D Sandee et al., "Effects of genetic variants of human P450 oxidoreductase on catalysis by CYP2D6 in vitro," *Pharmacogenet Genomics*, vol. 20, no. 11, pp. 677-86, Nov 2010.

[102] Y Shirasaka et al., "Interindividual variability of CYP2C19-catalyzed drug metabolism due

to differences in gene diplotypes and cytochrome P450 oxidoreductase content," *Pharmacogenomics J*, vol. 16, no. 4, pp. 375-87, Aug 2016.

[103] JA Tanner et al., "Predictors of Variation in CYP2A6 mRNA, Protein, and Enzyme Activity in a Human Liver Bank: Influence of Genetic and Nongenetic Factors," *J Pharmacol Exp Ther*, vol. 360, no. 1, pp. 129-139, Jan 2017.

[104] Z Bebia et al., "Bioequivalence revisited: influence of age and sex on CYP enzymes," *Clin Pharmacol Ther*, vol. 76, no. 6, pp. 618-27, Dec 2004.

[105] T Schwede, J Kopp, N Guex, and MC Peitsch, "SWISS-MODEL: An automated protein homology-modeling server," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3381-5, Jul 2003.

[106] RR Shah and RL Smith, "Addressing phenoconversion: the Achilles' heel of personalized medicine," *Br J Clin Pharmacol*, vol. 79, no. 2, pp. 222-40, Feb 2015.

[107] B Ray, E Ozcagli, W Sadee, and D Wang, "CYP2D6 haplotypes with enhancer single-nucleotide polymorphism rs5758550 and rs16947 (*2 allele): implications for CYP2D6 genotyping panels," *Pharmacogenet Genomics*, vol. 29, no. 2, pp. 39-47, Feb 2019.

[108] D Wang, AC Papp, and X Sun, "Functional characterization of CYP2D6 enhancer polymorphisms," *Hum Mol Genet*, vol. 24, no. 6, pp. 1556-62, Mar 2015.

[109] B Prasad and JD Unadkat, "Optimized approaches for quantification of drug transporters in tissues and cells by MRM proteomics," *AAPS J*, vol. 16, no. 4, pp. 634-48, Jul 2014.

[110] PharmVar Steering Committee, "The Evolution of PharmVar," *Clin Pharmacol Ther*, vol. 105, no. 1, pp. 29-32, Jan 2019.

[111] JB Houston, KE Kenworth, and A Galetin, "Typical and Atypical Enzyme Kinetics," in *Drug Metabolizing Enzymes: Cytochrome P450 and Other Enzymes in Drug Discovery and Development*, J Lee, RS Obach, and MB Fisher, Eds.: CRC Press, 2003, pp. 211-254.

[112] LC Wienkers and JC Stevens, "Cytochrome P450 Reaction Phenotyping," in *Drug Metabolizing Enzymes: Cytochrome P450 and Other Enzymes in Drug Discovery and Development*, J Lee, RS Obach, and MB Fisher, Eds.: CRC Press, 2003, pp. 255-310.

[113] K Venkatakrishnan, LL von Moltke, RS Obach, and DJ Greenblatt, "Drug metabolism and drug interactions: application and clinical value of in vitro models," *Curr Drug Metab*, vol. 4, no. 5, pp. 423-59, Oct 2003.

[114] B Zhu et al., "Assessment of cytochrome P450 activity by a five-drug cocktail approach," *Clin Pharmacol Ther*, vol. 70, no. 5, pp. 455-61, Nov 2001.

[115] RJ Ryu et al., "Pharmacokinetics of metoprolol during pregnancy and lactation," *J Clin Pharmacol*, vol. 56, no. 5, pp. 581-9, May 2016.

[116] LA McConnachie, B Phillips, M Bajpai, DD Shen, and RJ Ho, "Only truncated, not complete cytochrome p450 2D6 RNA transcript and no detectable enzyme activity are expressed in human lymphocytes," *Drug Metab Dispos*, vol. 31, no. 9, pp. 1103-7, Sep 2003.

[117] MG McDonald et al., "Expression and Functional Characterization of Breast Cancer-Associated Cytochrome P450 4Z1 in Saccharomyces cerevisiae," *Drug Metab Dispos*, vol. 45, no. 12, pp. 1364-1371, Dec 2017.

[118] RD Gietz and RH Schiestl, "High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method," *Nat Protoc*, vol. 2, no. 1, pp. 31-4, 2007.

[119] AT Wright and BF Cravatt, "Chemical proteomic probes for profiling cytochrome p450

activities and drug interactions in vivo," *Chem Biol*, vol. 14, no. 9, pp. 1043-51, Sep 2007.

[120] KP Vatsis, HM Peng, and MJ Coon, "Abolition of oxygenase function, retention of NADPH oxidase activity, and emergence of peroxidase activity upon replacement of the axial cysteine-436 ligand by histidine in cytochrome P450 2B4," *Arch Biochem Biophys*, vol. 434, no. 1, pp. 128-38, Feb 2005.

[121] S Yoshioka, S Takahashi, H Hori, K Ishimori, and I Morishima, "Proximal cysteine residue is essential for the enzymatic activities of cytochrome P450cam," *Eur J Biochem*, vol. 268, no. 2, pp. 252-9, Jan 2001.

[122] K Auclair, P Moënne-Loccoz, and PR Ortiz de Montellano, "Roles of the proximal heme thiolate ligand in cytochrome p450(cam)," *J Am Chem Soc*, vol. 123, no. 21, pp. 4877-85, May 2001.

[123] M Pirmohamed, F Kamali, AK Daly, and M Wadelius, "Oral anticoagulation: a critique of recent advances and controversies," *Trends Pharmacol Sci*, vol. 36, no. 3, pp. 153-63, Mar 2015.

[124] DP Dai et al., "CYP2C9 polymorphism analysis in Han Chinese populations: building the largest allele frequency database," *Pharmacogenomics J*, vol. 14, no. 1, pp. 85-92, Feb 2014.

[125] RL Haining, AP Hunter, ME Veronese, WF Trager, and AE Rettie, "Allelic variants of human cytochrome P450 2C9: baculovirus-mediated expression, purification, structural characterization, substrate stereoselectivity, and prochiral selectivity of the wild-type and I359L mutant forms," *Arch Biochem Biophys*, vol. 333, no. 2, pp. 447-58, Sep 1996.

[126] S Sanderson, J Emery, and J Higgins, "CYP2C9 gene variants, drug dose, and bleeding risk in warfarin-treated patients: a HuGEnet systematic review and meta-analysis," *Genet Med*, vol. 7, no. 2, pp. 97-104, Feb 2005.

[127] LV Kalman et al., "Development and Characterization of Reference Materials for Genetic Testing: Focus on Public Partnerships," *Ann Lab Med*, vol. 36, no. 6, pp. 513-20, Nov 2016.

[128] VM Pratt et al., "Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project," *J Mol Diagn*, vol. 18, no. 1, pp. 109-23, Jan 2016.

[129] Centers for Disease Control and Prevention. (2019, Feb.) RM Materials - Material Availability. [Online]. https://wwwn.cdc.gov/clia/Resources/GetRM/MaterialsAvailability.aspx

[130] International HapMap Consortium, "Common deletion polymorphisms in the human genome," *Nat Genet*, vol. 38, no. 1, pp. 86-92, Jan 2006.

[131] M Oscarson et al., "Characterization of a novel CYP2A7/CYP2A6 hybrid allele (CYP2A6*12) that causes reduced CYP2A6 activity," *Hum Mutat*, vol. 20, no. 4, pp. 275-83, Oct 2002.

[132] Swiss HIV Cohort Study, "Partial deletion of CYP2B6 owing to unequal crossover with CYP2B7," *Pharmacogenet Genomics*, vol. 17, no. 10, pp. 885-90, Oct 2007.

[133] 1000 Genomes Project Consortium, "Mapping copy number variation by population-scale genome sequencing," *Nature*, vol. 470, no. 7332, pp. 59-65, Feb 2011.

[134] R Redon et al., "Global variation in copy number in the human genome," *Nature*, vol. 444,

no. 7118, pp. 444-54, Nov 2006.

[135] 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68-74, Oct 2015.

[136] K Wang et al., "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome Res*, vol. 17, no. 11, pp. 1665-74, Nov 2007.

[137] TE Klein and MD Ritchie, "PharmCAT: A Pharmacogenomics Clinical Annotation Tool," *Clin Pharmacol Ther*, vol. 104, no. 1, pp. 19-22, Jul 2018.

[138] I Numanagić et al., "Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes," *Nat Commun*, vol. 9, no. 1, p. 828, Feb 2018.

[139] SR Browning and BL Browning, "Haplotype phasing: existing methods and new developments," *Nat Rev Genet*, vol. 12, no. 10, pp. 703-14, Sep 2011.

[140] MW Snyder, A Adey, JO Kitzman, and J Shendure, "Haplotype-resolved genome sequencing: experimental methods and applications," *Nat Rev Genet*, vol. 16, no. 6, pp. 344-58, Jun 2015.

[141] I Johansson et al., "Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine," *Proc Natl Acad Sci U S A*, vol. 90, no. 24, pp. 11825-9, Dec 1993.

[142] ML Dahl, I Johansson, L Bertilsson, M Ingelman-Sundberg, and F Sjöqvist, "Ultrarapid hydroxylation of debrisoquine in a Swedish population. Analysis of the molecular genetic basis," *J Pharmacol Exp Ther*, vol. 274, no. 1, pp. 516-20, Jul 1995.

[143] M Chida et al., "New allelic arrangement CYP2D6*36 x 2 found in a Japanese poor metabolizer of debrisoquine," *Pharmacogenetics*, vol. 12, no. 8, pp. 659-62, Nov 2002.

[144] M Nakano et al., "CYP2A7 pseudogene transcript affects CYP2A6 expression in human liver by acting as a decoy for miR-126," *Drug Metab Dispos*, vol. 43, no. 5, pp. 703-12, May 2015.

[145] C Raczy et al., "Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms," *Bioinformatics*, vol. 29, no. 16, pp. 2041-3, Aug 2013.

[146] M Whirl-Carrillo et al., "Pharmacogenomics knowledge for personalized medicine," *Clin Pharmacol Ther*, vol. 92, no. 4, pp. 414-7, Oct 2012.

[147] UGT Nomenclature Committee. (2005, June) UGT Alleles Nomenclature Home Page. [Online]. https://www.pharmacogenomics.pha.ulaval.ca/ugt-alleles-nomenclature/

[148] NAT Nomenclature Committee. NAT Alleles Nomenclature Home Page. [Online]. http://nat.mbg.duth.gr

[149] TPMT Nomenclature Committee. (2017, May) TPMT Alleles Nomenclature Home Page. [Online]. https://www.imh.liu.se/tpmtalleles/tabell-over-tpmt-alleler?l=en

[150] JT Robinson et al., "Integrative genomics viewer," *Nat Biotechnol*, vol. 29, no. 1, pp. 24-6, Jan 2011.

[151] SB Ng et al., "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, no. 7261, pp. 272-6, Sep 2009.

[152] RM Layer, C Chiang, AR Quinlan, and IM Hall, "LUMPY: a probabilistic framework for structural variant discovery," *Genome Biol*, vol. 15, no. 6, p. R84, Jun 2014.

[153] D Shigemizu et al., "IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis," *Sci Rep*, vol. 8, no. 1, p. 5608, Apr 2018.

[154] KM Walsh, MB Bracken, WK Murk, J Hoh, and AT Dewan, "Association between reduced copy-number at T-cell receptor gamma (TCRgamma) and childhood allergic asthma: A possible role for somatic mosaicism," *Mutat Res*, vol. 690, no. 1-2, pp. 89-94, Aug 2010.

[155] JB Kaye et al., "Warfarin Pharmacogenomics in Diverse Populations," *Pharmacotherapy*, vol. 37, no. 9, pp. 1150-1163, Sep 2017.

[156] G Suarez-Kurtz and MR Botton, "Pharmacogenetics of coumarin anticoagulants in Brazilians," *Expert Opin Drug Metab Toxicol*, vol. 11, no. 1, pp. 67-79, Jan 2015.

[157] J Duconge et al., "A Novel Admixture-Based Pharmacogenetic Approach to Refine Warfarin Dosing in Caribbean Hispanics," *PLoS One*, vol. 11, no. 1, p. e0145480, Jan 2016.

[158] Trans-Omics for Precision Medicine Program. About TOPMed. [Online]. https://www.nhlbiwgs.org

[159] Trans-Omics for Precision Medicine Program, "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program," *bioRxiv*, p. 563866, 2019, doi: https://doi.org/10.1101/563866.

[160] PharmVar. (2019, January) CYP2C9. [Online]. https://www.pharmvar.org/gene/CYP2C9

[161] M Kircher et al., "A general framework for estimating the relative pathogenicity of human genetic variants," *Nat Genet*, vol. 46, no. 3, pp. 310-5, Mar 2014.

[162] F Gnad, A Baucom, K Mukhyala, G Manning, and Z Zhang, "Assessment of computational methods for predicting the effects of missense mutations in human cancers," *BMC Genomics*, vol. 14, no. Suppl 3, p. S7, 2013.

[163] M Kircher et al., "Saturation mutagenesis of disease-associated regulatory elements," *bioRxiv*, p. 505362, 2018, doi: https://doi.org/10.1101/505362.

[164] BF Gage and LJ Lesko, "Pharmacogenetics of warfarin: regulatory, scientific, and clinical issues," *J Thromb Thrombolysis*, vol. 25, no. 1, pp. 45-51, Feb 2008.

[165] Y Zhu et al., "Estimation of warfarin maintenance dose based on VKORC1 (-1639 G>A) and CYP2C9 genotypes," *Clin Chem*, vol. 53, no. 7, pp. 1199-205, Jul 2007.

[166] AD Tyler et al., "Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications," *Sci Rep*, vol. 8, no. 1, p. 10931, Jul 2018.

[167] A Rhoads and KF Au, "PacBio Sequencing and Its Applications," *Genomics Proteomics Bioinformatics*, vol. 13, no. 5, pp. 278-89, Oct 2015.

[168] VM Steen, A Molven, NK Aarskog, and AK Gulbrandsen, "Homologous unequal cross-over involving a 2.8 kb direct repeat as a mechanism for the generation of allelic variants of human cytochrome P450 CYP2D6 gene," *Hum Mol Genet*, vol. 4, no. 12, pp. 2251-7, Dec 1995.

[169] J Häsler and K Strub, "Alu elements as regulators of gene expression," *Nucleic Acids Res*, vol. 34, no. 19, pp. 5491-7, 2006.

[170] Pacific Biosciences. (2017, March) IGV 3 Improves Support for PacBio Long Reads.

[Online]. https://www.pacb.com/blog/igv-3-improves-support-pacbio-long-reads/

[171] SE Levy and RM Myers, "Advancements in Next-Generation Sequencing," *Annu Rev Genomics Hum Genet*, vol. 17, pp. 95-115, Aug 2016.

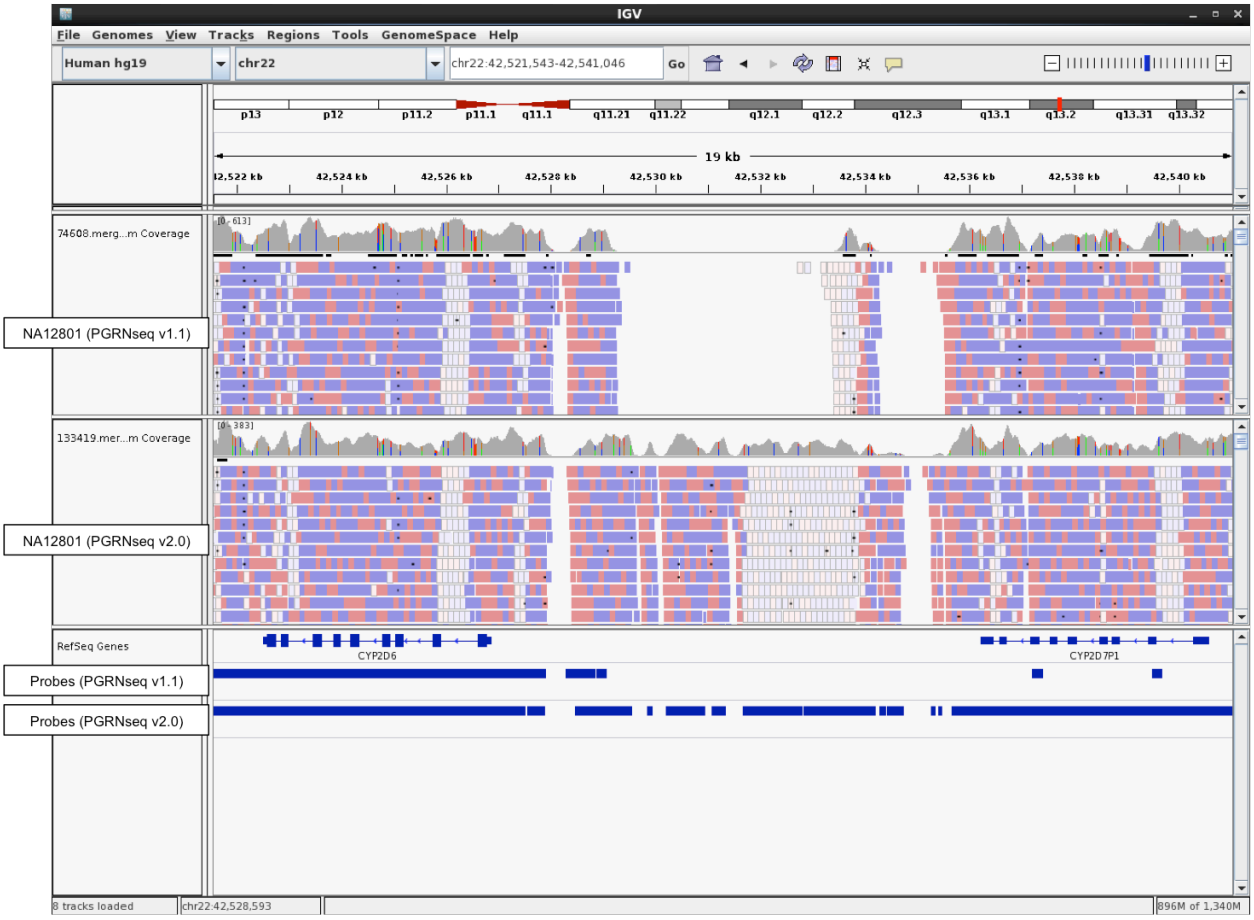
[172] AM Wenger et al., "Highly-accurate long-read sequencing improves variant detection and assembly of a human genome," *bioRxiv*, p. 519025, 2019, doi: https://doi.org/10.1101/519025.

[173] Bio-IT World. (2010, September) The Road to the $1,000 Genome. [Online]. http://www.bio-itworld.com/2010/09/28/1Kgenome.html

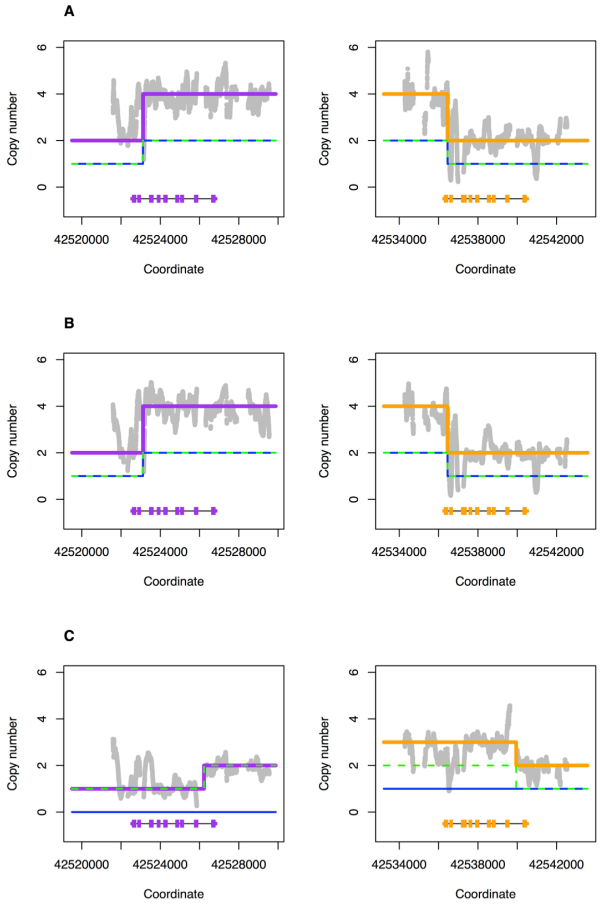
[174] RR Shah and A Gaedigk, "Precision medicine: does ethnicity information complement genotype-based prescribing decisions?," *Ther Adv Drug Saf*, vol. 9, no. 1, pp. 45-62, Jan 2018.

Appendix A: Supplementary Materials for Chapter 2

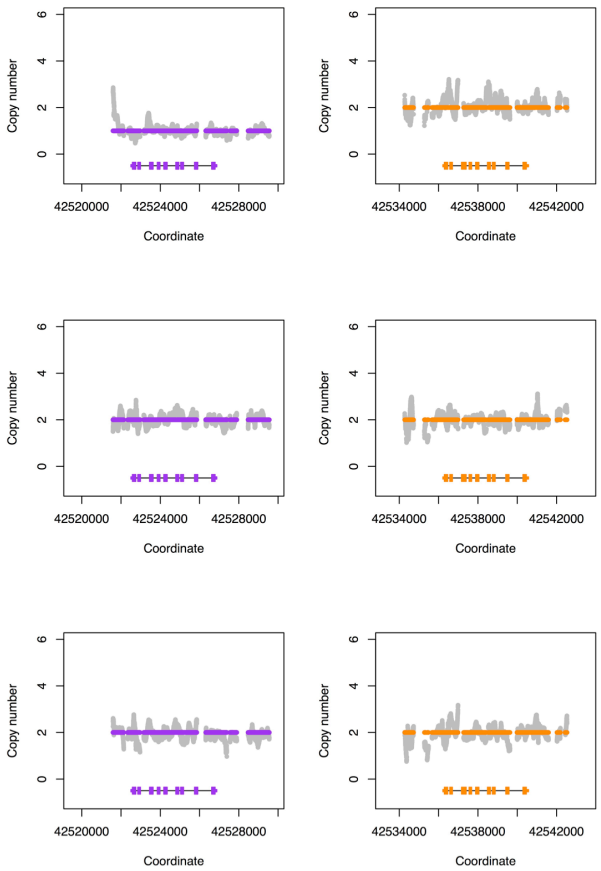
Supplementary Figures and Tables



Supplementary Figure 2.1 Next-generation sequencing of *CYP2D6* and *CYP2D7* with PGRNseq versions v1.1 and v2.0
A screenshot from Integrative Genomics Viewer[1] is showing sequence reads and coverage for subject NA12801. In PGRNseq v1.1, note that *CYP2D7* has been captured with the probes of *CYP2D6* owing to their high sequence homology.



Supplementary Figure 2.2 Testing all possible pairwise combinations of structural variants
In order to call diplotypes for samples that carry more than one structural variation, Stargazer first fits every pairwise combination of structural variants to the *CYP2D6* and *CYP2D7* copy number profiles of such sample before finding the combination that produces the least deviance. This approach was used to genotype (A) Han Chinese HG00463 with a *CYP2D6*36+*10/*36+*10* diplotype, (B) Han Chinese HG00465 with a *CYP2D6*36+*10/*36+*10* diplotype, and (C) Puerto Rican HG01190 with a *CYP2D6*5/*68+*4* diplotype. Grey dots denote copy number calculated from read depth. The blue solid and green dashed lines represent two structural variants that are being tested against the observed copy number of the sample. The purple and orange lines represent theoretical copy number profiles for *CYP2D6* and *CYP2D7*, respectively, formed by combining the two structural variants in question. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. All panels were generated from PGRNseq v2.0 data.



Supplementary Figure 2.3 Discrepant diplotype calls between orthogonal methods and Stargazer in the Y045 family
For this trio, the orthogonal methods called the father (NA19200, the top panel), the mother (NA19201, the middle panel), and the child (NA19202, the bottom panel) as having *CYP2D6*5/*76+*1*, *CYP2D6*1/*17*, and *CYP2D6*1/*76+*1* diplotypes, respectively, while Stargazer called *CYP2D6*1/*5*, *CYP2D6*1/*17*, and *CYP2D6*1/*1* diplotypes. The *CYP2D6*76* allele is essentially a *CYP2D7* gene that has a *CYP2D6* downstream sequence with its switch region past exon 9, and lacks a *CYP2D7*-specific sequence also referred to as 'spacer.' Because Stargazer did not detect any significant change in copy number at the switch region or in the spacer, the program did not call the *CYP2D6*76* allele. Grey dots are copy number calculated from read depth. Dots colored purple and orange are the mean copy number for *CYP2D6* and *CYP2D7*, respectively, determined by the changepoint algorithm. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. All panels were generated from PGRNseq v2.0 data.

Supplementary Table 2.1 *CYP2D6* haplotypes identified in 32 HapMap trios called by Stargazer and the frequencies of these haplotypes in the 64 unrelated parents

No.	Allele	Count	%	Activity Score	Structural Variation
1	**1*	40	31.3%	1	reference
2	**2*	24	18.8%	1	none
3	**3*	2	1.6%	0	none
4	**4*	11	8.6%	0	none
5	**5*	6	4.7%	0	deletion
6	**6*	3	2.3%	0	none
7	**9*	2	1.6%	0.5	none
8	**10*	3	2.3%	0.5	none
9	**15*	1	0.8%	0	none
10	**17*	9	7.0%	0.5	none
11	**29*	2	1.6%	0.5	none
12	**35*	2	1.6%	1	none
13	**41*	8	6.3%	0.5	none
14	**2x2*	1	0.8%	2	duplication
15	**4x2*	2	1.6%	0	duplication
16	**4N+*4*	1	0.8%	0	hybrid
17	**10x2*	1	0.8%	1	duplication
18	**36+*10*	4	3.1%	0.5	hybrid
19	**68+*4*	5	3.9%	0	hybrid
20	**78+*2*	1	0.8%	1	hybrid

Supplementary Table 2.2 Testing Stargazer on data generated from a range of read coverage
Sequence reads from PGRNseq v2.0 data were randomly downsampled to represent read coverages ranging from 7.9x to 158.1x. Stargazer was then applied to each simulated dataset and resulting diplotype and phenotype calls were recorded.

Fraction	Coverage	Correct Haplotype	Incorrect Haplotype	Uncalled Haplotype	Correct Phenotype	Incorrect Phenotype	Uncalled Phenotype
0.05	7.9	154	30	4	83	8	3
0.10	15.8	181	3	4	91	1	2
0.15	23.7	186	2	0	94	0	0
0.20	31.6	184	0	4	92	0	2
0.30	47.4	188	0	0	94	0	0
0.40	63.2	186	0	2	93	0	1
0.50	79.0	187	1	0	93	1	0
0.60	94.9	186	0	2	93	0	1
0.70	110.7	186	0	2	93	0	1
0.80	126.5	188	0	0	94	0	0
0.90	142.3	188	0	0	94	0	0
1.00	158.1	188	0	0	94	0	0

Supplementary Table 2.3 SNVs/indels detected from the PGRNseq data that are within 3kb from either end of *CYP2D6*
Many of these variants are not currently used to define any star allele. An asterisk (*) in the alternative allele column indicates an overlapping deletion that spans a position of interest.

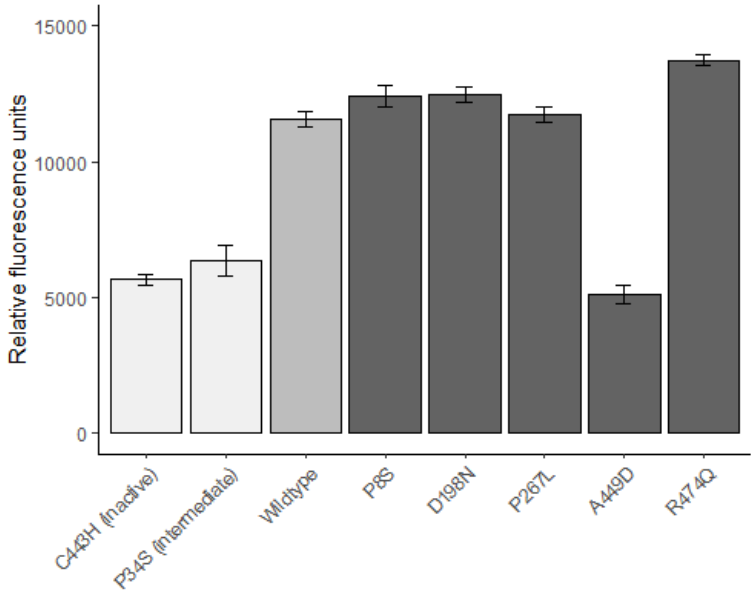
No.	Position[a]	rs ID	Reference Allele	Alternative Allele	Function[b]	Table[c]
1	42526969	rs1080993	C	T	upstream	yes
2	42527471	rs28633410	T	C	upstream	yes
3	42527484	rs75324300	CCACA	C	upstream	yes
4	42527533	rs28624811	A	G	upstream	yes
5	42527793	rs1080989	C	T	upstream	yes
6	42527886	rs375413467	A	AT	upstream	yes
7	42528341	rs76210340	C	T	upstream	yes
8	42528382	rs1080985	C	G	upstream	yes
9	42528538	rs58188898	G	A	upstream	yes
10	42528568	rs1080983	T	C	upstream	yes
11	42522965	rs28371732	C	T	synonymous	yes
12	42523539	rs28371726	A	G	synonymous	yes
13	42524218	rs28371718	G	T	synonymous	yes
14	42524323	rs17002852	A	G	synonymous	yes
15	42524924	rs111606937	A	G	synonymous	yes
16	42525132	rs1058164	G	C	synonymous	yes
17	42525756	rs1081003	G	A	synonymous	yes
18	42525798	rs28371705	G	C	synonymous	yes
19	42524947	rs3892097	C	T	splice	yes
20	42522613	rs1135840	G	C	missense	yes
21	42523505	rs150552908	C	T	missense	yes
22	42523558	rs202102799	T	C	missense	yes
23	42523610	rs59421388	C	T	missense	yes
24	42523943	rs16947	A	G	missense	yes
25	42524817	rs5030866	C	T	missense	yes
26	42525077	rs28371710	C	T	missense	yes
27	42525134	rs61736512	C	T	missense	yes
28	42525772	rs28371706	G	A	missense	yes
29	42525811	rs28371704	T	C	missense	yes
30	42525821	rs28371703	G	T	missense	yes
31	42526694	rs1065852	G	A	missense	yes
32	42526763	rs769258	C	T	missense	yes
33	42523003	rs116917064	A	G	intron	yes
34	42523209	rs28371730	T	C	intron	yes
35	42523211	rs2004511	T	C	intron	yes
36	42523358	rs28371729	G	T	intron	yes
37	42523409	rs1985842	G	T	intron	yes
38	42523805	rs28371725	C	T	intron	yes
39	42524132	rs76015180	C	T	intron	yes
40	42524696	rs58440431	T	C	intron	yes
41	42525952	rs71328650	C	A	intron	yes
42	42526049	rs147296446	C	G	intron	yes
43	42526484	rs28371699	A	C	intron	yes
44	42526549	rs56011157	C	T	intron	yes
45	42526561	rs28695233	G	T	intron	yes
46	42526562	rs75276289	G	C	intron	yes
47	42526567	rs1080998	G	A	intron	yes
48	42526571	rs1080997	C	G	intron	yes
49	42526573	rs1080996	T	G	intron	yes
50	42526580	rs1080995	G	C	intron	yes
51	42524243	rs35742686	CT	C	frameshift	yes
52	42525085	rs5030655	CA	C	frameshift	yes
53	42526656	rs72549357	C	CA	frameshift	yes
54	42522312	rs116390392	T	C	downstream	yes
55	42522392	rs28371738	G	A	downstream	yes
56	42524175	rs5030656	CCTT	C	coding	yes
57	42526995	rs372204775	C	T	upstream	no
58	42527116	rs572914357	A	G	upstream	no
59	42527191	rs374672076	C	G	upstream	no
60	42527291	rs74966855	G	T	upstream	no
61	42527392	rs573767354	C	T	upstream	no
62	42527578	rs542624837	C	A	upstream	no
63	42527627	rs190388483	G	A	upstream	no
64	42528392	rs1080984	A	G	upstream	no
65	42528477	rs775251692	C	T	upstream	no
66	42528851	rs28680494	A	C	upstream	no
67	42528858	rs28439297	C	T	upstream	no
68	42528976	rs28360521	C	T	upstream	no
69	42529156	rs73887946	C	T	upstream	no
70	42529176	rs138419008	T	C	upstream	no
71	42529219	rs28568508	A	G	upstream	no
72	42529295	rs537833339	T	C	upstream	no
73	42529321	rs28566059	C	T	upstream	no
74	42529330	rs76404195	C	T	upstream	no
75	42529366	rs118017929	C	G	upstream	no
76	42529407	rs28542726	G	T	upstream	no
77	42529540	rs28369142	C	A	upstream	no
78	42524191	rs28371719	C	A	synonymous	no
79	42524795	rs28371713	A	G	synonymous	no
80	42525039	rs1135825	G	A	synonymous	no
81	42523636	rs3915951	C	A	missense	no

82	42523483	rs61731586	C	T	missense	no
83	42523514	rs61745683	C	T	missense	no
84	42523528	rs1058172	C	T	missense	no
85	42524826	.	G	A	missense	no
86	42522765	rs548264542	G	T	intron	no
87	42522774	rs772203297	G	A	intron	no
88	42523247	.	C	CT	intron	no
89	42523309	rs867985262	C	T	intron	no
90	42523315	rs79596243	T	C	intron	no
91	42524090	rs74516776	G	A	intron	no
92	42524150	rs377220693	C	T	intron	no
93	42524435	rs1807313	T	A	intron	no
94	42524501	rs75203276	C	T	intron	no
95	42524578	rs80262685	T	C	intron	no
96	42524596	rs544825251	C	T	intron	no
97	42524670	rs76327133	G	A	intron	no
98	42524708	rs111564371	T	C	intron	no
99	42524713	rs112568578	C	G	intron	no
100	42524743	rs113889384	G	A	intron	no
101	42524963	rs554875652	C	G	intron	no
102	42524975	rs200720666	C	T	intron	no
103	42524982	rs113678157	C	T	intron	no
104	42525003	rs76326664	T	C	intron	no
105	42525247	rs546937511	C	T	intron	no
106	42525305	rs142302759	T	G	intron	no
107	42525390	rs557722765	G	A	intron	no
108	42525438	rs184086520	A	G	intron	no
109	42525547	rs267608277	G	A	intron	no
110	42525625	rs1081004	C	T	intron	no
111	42525728	rs78854695	A	C	intron	no
112	42526044	.	T	C	intron	no
113	42526050	.	T	G	intron	no
114	42526050	rs267608275	TG	T	intron	no
115	42526055	.	C	G	intron	no
116	42526141	rs376217512	G	A	intron	no
117	42526524	rs29001678	G	A	intron	no
118	42521489	rs145458535	C	T	downstream	no
119	42521560	.	G	A	downstream	no
120	42521577	rs201551607	G	A	downstream	no
121	42521625	rs72395204	G(GGGTGGGGAA)x2	G(GGGTGGGGAA)x0	downstream	no
122	42521625	rs72395204	G(GGGTGGGGAA)x2	G(GGGTGGGGAA)x1	downstream	no
123	42521625	rs72395204	G(GGGTGGGGAA)x2	G(GGGTGGGGAA)x3	downstream	no
124	42521639	rs374153932	T	A	downstream	no
125	42521639	rs374153932	T	*	downstream	no
126	42521715	.	GCT	G	downstream	no
127	42521807	rs561101513	G	T	downstream	no
128	42521865	rs112866416	C	T	downstream	no
129	42521918	rs532182046	T	C	downstream	no
130	42521919	rs550576546	G	A	downstream	no
131	42521920	rs562602203	T	G	downstream	no
132	42521929	rs34385013	G	A	downstream	no
133	42521985	rs4078247	T	C	downstream	no
134	42522027	rs4078248	C	T	downstream	no
135	42522074	rs35028622	C	A	downstream	no
136	42522084	rs4078249	C	T	downstream	no
137	42522137	rs71184866	C	CTGT	downstream	no
138	42522349	rs148648640	T	C	downstream	no
139	42522464	rs77845838	G	A	downstream	no
140	42522498	rs555084863	G	A	downstream	no
141	42526836	rs75085559	A	AC	5' UTR	no
142	42522550	rs201759814	G	A	3' UTR	no

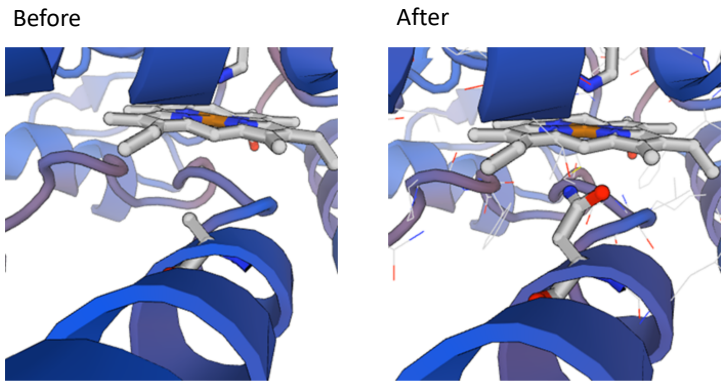
[a]Human Genome version 19.
[b]SeattleSeq Annotation was used to provide functional annotation for the variants.
[c]Variant is present in the translation table.

Appendix B: Supplementary Materials for Chapter 3

Supplementary Figures and Tables



Supplementary Figure 3.1 Characterization of rare CYP2D6 variants using Vivid™ 7-ethoxymethoxy-3-cyanocoumarin
CYP2D6 variants were induced in an isogenic yeast strain and function measured with a standard fluorogenic CYP2D6 substrate on a plate reader. Controls (light gray): inactive variant (C443H) and decreased function variant (P34S). Wildtype: medium gray. Novel rare variant strains (dark gray): P8S, D198N, P267L, A449D, and R474Q. Error bars indicate standard deviations from 4 independent replicates.



Supplementary Figure 3.2 Protein modeling of CYP2D6 with and without the rare A449D variant
The SWISS-MODEL was used to render CYP2D6 structure, with and without the A449D variant. The rendered visuals suggest that A449D affects CYP2D6 activity by interfering with heme binding. The A449D amino acid change also has a CADD score of 32, which is considerably higher than the score of 15 that is used as cut-off for determining detrimental mutations.

Supplementary Table 3.1 Liver donor demographics (N=314)

Age (years)[a]	
mean ± SD	39.7 ± 22.4
range	0-87
Sex (N, %)	
male	183 (58.2%)
female	126 (40.1%)
unknown	5 (1.6%)
Population (N, %)	
European	300 (95.5%)
African	9 (2.9%)
Asian	1 (0.3%)
Hispanic	1 (0.3%)
unknown	3 (1.0%)

[a]Age was recorded in years for all subjects and an age of zero was given to donors under one year old; more granular data (e.g. weeks, months) were not available.

Supplementary Table 3.2 *CYP2D6* diplotypes detected in the liver samples

Diplotype	N	AS	Phenotype
**1/*1x2*	1	3	UM
**2/*1x2*	1	3	UM
**35/*2x2*	1	3	UM
**41/*2x2*	1	2.5	UM
**1/*1*	30	2	NM
**1/*2*	35	2	NM
**1/*33*	2	2	NM
**1/*35*	12	2	NM
**2/*2*	4	2	NM
**2/*33*	1	2	NM
**2/*34*	1	2	NM
**2/*35*	2	2	NM
**35/*35*	1	2	NM
**4/*1x2*	1	2	NM
**4/*2x2*	1	2	NM
**1/*10*	8	1.5	NM
**1/*17*	3	1.5	NM
**1/*41*	18	1.5	NM
**1/*59*	1	1.5	NM
**1/*9*	9	1.5	NM
**10/*35*	4	1.5	NM
**10/*41x2*	1	1.5	NM
**2/*41*	14	1.5	NM
**2/*59*	3	1.5	NM
**2/*9*	1	1.5	NM
**29/*39*	1	1.5	NM
**35/*41*	8	1.5	NM
**41/*77+*2*	1	1.5	NM
**1/*20*	1	1	NM
**1/*3*	2	1	NM
**1/*4*	25	1	NM
**1/*4N+*4*	2	1	NM
**1/*5*	8	1	NM
**1/*6*	2	1	NM
**1/*68+*4*	12	1	NM
**10/*36+*10*	1	1	NM
**2/*3*	1	1	NM
**2/*4*	13	1	NM
**2/*5*	4	1	NM
**2/*6*	1	1	NM
**2/*68+*4*	5	1	NM
**3/*43*	1	1	NM
**33/*4N+*4*	1	1	NM
**35/*20*	1	1	NM
**35/*68+*4*	2	1	NM
**4/*33*	1	1	NM
**4/*35*	3	1	NM
**4/*43*	1	1	NM
**41/*41*	1	1	NM
**41/*59*	1	1	NM
**5/*35*	1	1	NM
**6/*35*	1	1	NM
**9/*41*	6	1	NM
**9/*9*	1	1	NM
**4/*10*	1	0.5	IM
**4/*17*	1	0.5	IM
**4/*41*	10	0.5	IM
**4/*9*	3	0.5	IM
**41/*4N+*4*	1	0.5	IM
**41/*68+*4*	3	0.5	IM
**5/*41*	4	0.5	IM
**6/*41*	1	0.5	IM
**68+*4/*36+*10*	1	0.5	IM
**9/*68+*4*	1	0.5	IM
**3/*5*	1	0	PM
**3/*68+*4*	2	0	PM
**4/*20*	1	0	PM
**4/*4*	7	0	PM
**4/*5*	4	0	PM
**4/*6*	1	0	PM
**4/*68+*4*	5	0	PM
**5/*68+*4*	2	0	PM
**68+*4/*68+*4*	1	0	PM

AS, activity score; UM, ultrarapid metabolizer; NM, normal metabolizer; IM, intermediate metabolizer; PM, poor metabolizer.

Supplementary Table 3.3 Multiple linear regression: comparison of association between CYP2D6 activity and activity scores assigned using SNV data only or including structural variation data from Stargazer

Prediction of CYP2D6 activity using activity scores assigned using SNV data only

	Dextromethorphan			Metoprolol		
	β	***SE***	***p-value***	***β***	***SE***	***p-value***
(Intercept)[a]	-3.71	6.23	0.55	-0.15	1.47	0.92
AS=0.5	6.51	8.59	0.45	1.13	2.03	0.58
AS=1	22.82	6.87	**9.94×10^{-4}**	4.61	1.64	**0.01**
AS=1.5	32.21	7.05	**7.17×10^{-6}**	7.26	1.68	**2.25×10^{-5}**
AS=2	52.99	6.81	**1.13×10^{-13}**	11.59	1.62	**6.59×10^{-12}**
Site: UW	30.63	4.16	**1.62×10^{-12}**	5.17	0.99	**3.51×10^{-7}**
DF	308			289		
R^2/adj. R^2	0.36/0.35			0.31/0.30		

Prediction of CYP2D6 activity using activity scores assigned using Stargazer, based on both SNV and structural variation data

	Dextromethorphan			Metoprolol		
	β	***SE***	***p-value***	***β***	***SE***	***p-value***
(Intercept)[a]	-2.11	5.6	0.71	0.19	1.33	0.89
AS=0.5	3.16	7.74	0.68	0.43	1.85	0.82
AS=1	23.29	6.23	**2.19×10^{-4}**	4.79	1.49	**1.46×10^{-3}**
AS=1.5	28.98	6.46	**1.02×10^{-5}**	6.45	1.55	**4.13×10^{-5}**
AS=2	55.22	6.3	**$<2.00\times10^{-16}$**	12.05	1.5	**2.41×10^{-14}**
AS=2.5	94.87	20.18	**3.90×10^{-6}**	23.63	4.78	**1.31×10^{-6}**
AS=3	49.49	16.75	**3.37×10^{-3}**	8	3.97	**0.04**
Site: UW	28.79	4	**4.90×10^{-12}**	4.73	0.96	**1.26×10^{-6}**
DF	306			287		
R^2/adj. R^2	0.41/0.40			0.36/0.35		

AS, activity score; UW, University of Washington; β, β-coefficient; SE, standard error; DF, degrees of freedom.
[a]AS=0 is incorporated into the intercept.

Supplementary Table 3.4 SNVs/indels detected in the Liver Bank by PGRNseq data that were within 3 kb from either end of *CYP2D6*

No.	Table[a]	Region[b]	Position[c]	rs ID	Reference Allele	Alternative Allele	Function[d]	AA	Residue	CADD	AC
1	Yes	Yes	42522312	rs116390392	T	C	downstream	-	-	2.511	413
2	Yes	Yes	42522392	rs28371738	G	A	downstream	-	-	1.136	145
3	Yes	Yes	42522613	rs1135840	G	C	missense	THR, SER	486	3.127	261
4	Yes	Yes	42522940	rs769157652	C	T	missense	GLU, LYS	410	15.05	1
5	Yes	Yes	42522965	rs28371732	C	T	synonymous	SER	401	0.891	2
6	Yes	Yes	42523003	rs116917064	A	G	Intron	-	-	3.767	368
7	Yes	Yes	42523209	rs28371730	T	C	Intron	-	-	1.232	408
8	Yes	Yes	42523211	rs2004511	T	C	Intron	-	-	0.848	146
9	Yes	Yes	42523358	rs28371729	G	T	Intron	-	-	5.316	17
10	Yes	Yes	42523409	rs1985842	G	T	Intron	-	-	0.698	201
11	Yes	Yes	42523505	rs150552908	C	T	missense	GLY, SER	373	11.82	68
12	Yes	Yes	42523539	rs28371726	A	G	synonymous	HIS	361	7.593	36
13	Yes	Yes	42523558	rs202102799	T	C	missense	TYR, CYS	355	15.61	19
14	Yes	Yes	42523567	rs61736517	T	C	missense	HIS, ARG	352	0.008	1
15	Yes	Yes	42523610	rs59421388	C	T	missense	VAL, MET	338	22.9	1
16	Yes	Yes	42523805	rs28371725	C	T	Intron	-	-	6.324	75
17	Yes	Yes	42523854	rs79292917	C	T	synonymous	PRO	325	4.125	5
18	Yes	Yes	42523943	rs16947	A	G	missense	CYS, ARG	296	0.545	406
19	Yes	Yes	42524132	rs76015180	C	T	Intron	-	-	6.039	2
20	Yes	Yes	42524175	rs5030656	CCTT	C	coding	-	-	-	22
21	Yes	Yes	42524218	rs28371718	G	T	synonymous	PRO	267	4.991	3
22	Yes	Yes	42524243	rs35742686	CT	C	frameshift	-	-	-	8
23	Yes	Yes	42524310	rs28371717	C	A	missense	ALA, SER	237	11.43	5
24	Yes	Yes	42524323	rs17002852	A	G	synonymous	HIS	232	4.784	4
25	Yes	Yes	42524502	rs267608300	C	T	Intron	-	-	1.687	5
26	Yes	Yes	42524696	rs58440431	T	C	Intron	-	-	5.504	142
27	Yes	Yes	42524814	rs199535154	A	G	missense	LEU, PRO	213	14.3	3
28	Yes	Yes	42524815	rs150163869	G	A	synonymous	LEU	213	3.7	4
29	Yes	Yes	42524817	rs5030866	C	T	missense	GLY, GLU	212	3.567	6
30	Yes	Yes	42524820	rs3831704	T	TC	frameshift	-	-	-	3
31	Yes	Yes	42524924	rs111606937	A	G	synonymous	GLY	176	6.286	126
32	Yes	Yes	42524947	rs3892097	C	T	splice	-	-	14.92	126
33	Yes	Yes	42525044	rs1135824	T	C	missense	ASN, ASP	166	0.003	1
34	Yes	Yes	42525085	rs5030655	CA	C	frameshift	-	-	-	6
35	Yes	Yes	42525089	rs78482768	G	C	missense	GLN, GLU	151	0.009	1
36	Yes	Yes	42525132	rs1058164	G	C	synonymous	VAL	136	0.087	261
37	Yes	Yes	42525134	rs61736512	C	T	missense	VAL, ILE	136	7.841	1
38	Yes	Yes	42525733	rs267608289	T	C	Intron	-	-	3.956	1
39	Yes	Yes	42525756	rs1081003	G	A	synonymous	PHE	112	3.186	25
40	Yes	Yes	42525772	rs28371706	G	A	missense	THR, ILE	107	5.817	4
41	Yes	Yes	42525798	rs28371705	G	C	synonymous	THR	98	3.782	119
42	Yes	Yes	42525811	rs28371704	T	C	missense	HIS, ARG	94	0.424	115
43	Yes	Yes	42525821	rs28371703	G	T	missense	LEU, MET	91	14.36	115
44	Yes	Yes	42525952	rs71328650	C	A	Intron	-	-	2.593	259
45	Yes	Yes	42526049	rs147296446	C	G	Intron	-	-	3.125	275
46	Yes	Yes	42526484	rs28371699	A	C	Intron	-	-	6.725	258
47	Yes	Yes	42526549	rs56011157	C	T	Intron	-	-	6.694	408
48	Yes	Yes	42526561	rs28695233	G	T	Intron	-	-	2.422	407
49	Yes	Yes	42526562	rs75276289	G	C	Intron	-	-	1.702	407
50	Yes	Yes	42526567	rs1080998	G	A	Intron	-	-	0.469	407
51	Yes	Yes	42526571	rs1080997	C	G	Intron	-	-	1.522	407
52	Yes	Yes	42526573	rs1080996	T	G	Intron	-	-	2.388	407
53	Yes	Yes	42526580	rs1080995	G	C	Intron	-	-	3.908	407
54	Yes	Yes	42526694	rs1065852	G	A	missense	PRO, SER	34	18.34	145
55	Yes	Yes	42526717	rs28371696	C	T	missense	ARG, HIS	26	14.41	2
56	Yes	Yes	42526763	rs769258	C	T	missense	VAL, MET	11	3.418	39
57	Yes	Yes	42526775	rs72549358	C	T	missense	VAL, MET	7	1.817	1
58	Yes	Yes	42526969	rs1080993	C	T	upstream	-	-	4.618	1
59	Yes	Yes	42527224	rs566383351	G	A	upstream	-	-	1.184	1
60	Yes	Yes	42527471	rs28633410	T	C	upstream	-	-	0.843	401
61	Yes	Yes	42527533	rs28624811	A	G	upstream	-	-	0.79	405
62	Yes	Yes	42527793	rs1080989	C	T	upstream	-	-	0.265	145
63	Yes	Yes	42527886	rs375413467	A	AT	upstream	-	-	-	62
64	Yes	Yes	42528028	rs28735595	C	T	upstream	-	-	0.256	164
65	Yes	Yes	42528096	rs59099247	C	T	upstream	-	-	0.489	1
66	Yes	Yes	42528224	rs28588594	G	A	upstream	-	-	0.835	41
67	Yes	Yes	42528382	rs1080985	C	G	upstream	-	-	0.174	486
68	No	No	42521442	rs535293691	G	A	downstream	-	-	4.086	2
69	No	No	42521489	rs145458535	C	T	downstream	-	-	1.976	79
70	No	No	42521577	rs201551607	G	A	downstream	-	-	0.172	2
71	No	No	42521625	rs72395204	G(GGGTGGGGAA)x2	G(GGGTGGGGAA)x0	downstream	-	-	-	52
72	No	No	42521625	rs72395204	G(GGGTGGGGAA)x2	G(GGGTGGGGAA)x1	downstream	-	-	-	98
73	No	No	42521625	rs72395204	G(GGGTGGGGAA)x2	G(GGGTGGGGAA)x3	downstream	-	-	-	7
74	No	No	42521676	-	G	A	downstream	-	-	0.257	1
75	No	No	42521799	-	GGCCTGTGGGCCCT	G	downstream	-	-	-	1
76	No	No	42521807	rs561101513	G	T	downstream	-	-	2.048	1
77	No	No	42521811	-	C	G	downstream	-	-	4.42	1
78	No	No	42521865	rs112866416	C	T	downstream	-	-	2.139	4
79	No	No	42521918	rs532182046	T	C	downstream	-	-	3.416	6
80	No	No	42521919	rs550576546	G	A	downstream	-	-	7.514	6
81	No	No	42521920	rs562602203	T	G	downstream	-	-	5.554	5
82	No	No	42521929	rs34385013	G	A	downstream	-	-	2.417	261

83	No	No	42521969	-	G	A	downstream	-	-	3.234	1
84	No	No	42521985	rs4078247	T	C	downstream	-	-	0.999	145
85	No	No	42522027	rs4078248	C	T	downstream	-	-	8.703	4
86	No	No	42522074	rs35028622	C	A	downstream	-	-	1.735	259
87	No	No	42522079	rs77827855	G	A	downstream	-	-	0.286	2
88	No	No	42522084	rs4078249	C	T	downstream	-	-	3.795	4
89	No	No	42522101	rs866408541	C	T	downstream	-	-	5.332	1
90	No	No	42522137	rs71184866	C	CTGT	downstream	-	-	-	405
91	No	No	42522141	-	G	T	downstream	-	-	3.167	2
92	No	No	42522142	-	A	G	downstream	-	-	3.045	2
93	No	Yes	42522349	rs148648640	T	C	downstream	-	-	0.894	1
94	No	Yes	42522464	rs77845838	G	A	downstream	-	-	5.327	7
95	No	Yes	42522550	rs201759814	G	A	3' UTR	-	-	4.413	41
96	No	Yes	42522649	rs141756339	C	T	missense	ARG, GLN	474	11.57	1
97	No	Yes	42522724	rs79392742	G	T	missense	ALA, ASP	449	32	1
98	No	Yes	42522765	rs548264542	G	T	Intron	-	-	6.939	1
99	No	Yes	42522774	rs772203297	G	A	Intron	-	-	1.665	3
100	No	Yes	42522846	-	C	T	Intron	-	-	5.595	1
101	No	Yes	42523070	rs1009883497	GAC	G	Intron	-	-	-	1
102	No	Yes	42523228	-	C	T	Intron	-	-	1.957	1
103	No	Yes	42523247	-	C	CT	Intron	-	-	-	1
104	No	Yes	42523309	rs867985262	C	T	Intron	-	-	3.3	120
105	No	Yes	42523315	rs79596243	T	C	Intron	-	-	2.375	126
106	No	Yes	42523400	rs28578778	A	G	Intron	-	-	0.507	7
107	No	Yes	42523514	rs61745683	C	T	missense	VAL, ILE	370	6.769	65
108	No	Yes	42523528	rs1058172	C	T	missense	ARG, HIS	365	34	124
109	No	Yes	42523609	rs771811053	ACGT	A	coding	-	-	-	1
110	No	Yes	42523636	rs3915951	C	A	missense	ARG, LEU	329	14.89	63
111	No	Yes	42523650	rs370010370	C	T	Intron	-	-	4.449	1
112	No	Yes	42523695	rs768806497	CCTT	C	Intron	-	-	-	1
113	No	Yes	42523721	rs867265808	G	A	Intron	-	-	7.023	1
114	No	Yes	42523813	rs143276168	G	A	Intron	-	-	1.046	2
115	No	Yes	42523820	rs772490933	G	A	Intron	-	-	5.597	1
116	No	Yes	42524073	rs187203531	C	G	Intron	-	-	3.596	7
117	No	Yes	42524130	rs28371722	C	T	Intron	-	-	2.929	1
118	No	Yes	42524191	rs28371719	C	A	synonymous	LEU	276	5.956	1
119	No	Yes	42524219	rs148769737	G	A	missense	PRO, LEU	267	15.39	1
120	No	Yes	42524501	rs75203276	C	T	Intron	-	-	1.877	1
121	No	Yes	42524578	rs80262685	T	C	Intron	-	-	1.351	1
122	No	Yes	42524670	rs76327133	G	A	Intron	-	-	2.853	1
123	No	Yes	42524708	rs111564371	T	C	Intron	-	-	1.56	208
124	No	Yes	42524713	rs112568578	C	A	Intron	-	-	0.492	1
125	No	Yes	42524713	rs112568578	C	G	Intron	-	-	0.492	207
126	No	Yes	42524733	-	G	A	Intron	-	-	2.875	2
127	No	Yes	42524743	rs113889384	G	A	Intron	-	-	1.308	199
128	No	Yes	42524759	-	G	C	Intron	-	-	0.675	1
129	No	Yes	42524776	rs368129875	C	T	Intron	-	-	1.347	1
130	No	Yes	42524795	rs28371713	A	G	synonymous	PHE	219	4.015	210
131	No	Yes	42524860	rs763284150	C	T	missense	ASP, ASN	198	8.097	1
132	No	Yes	42524975	rs200720666	C	T	Intron	-	-	6.896	2
133	No	Yes	42525038	rs1135826	A	C	missense	SER, ALA	168	0.028	1
134	No	Yes	42525039	rs1135825	G	T	missense	HIS, GLN	167	0.021	1
135	No	Yes	42525045	rs768668762	G	A	synonymous	ALA	165	0.032	1
136	No	Yes	42525180	rs61736507	G	A	synonymous	PHE	120	4.546	3
137	No	Yes	42525305	rs142302759	T	G	Intron	-	-	2.797	1
138	No	Yes	42525390	rs557722765	G	A	Intron	-	-	6.221	2
139	No	Yes	42525416	rs143170489	G	C	Intron	-	-	0.508	3
140	No	Yes	42525438	rs184086520	A	G	Intron	-	-	4.17	1
141	No	Yes	42525463	rs1046791114	G	A	Intron	-	-	1.17	3
142	No	Yes	42525500	rs189736703	C	T	Intron	-	-	1.057	5
143	No	Yes	42525547	rs267608277	G	A	Intron	-	-	4.796	5
144	No	Yes	42525625	rs1081004	C	T	Intron	-	-	4.952	34
145	No	Yes	42525628	-	T	A	Intron	-	-	4.379	1
146	No	Yes	42525728	rs78854695	A	C	Intron	-	-	5.434	1
147	No	Yes	42526421	rs926658668	A	G	Intron	-	-	3.423	1
148	No	Yes	42526477	-	T	C	Intron	-	-	4.764	1
149	No	Yes	42526524	rs76527171	G	A	Intron	-	-	6.971	24
150	No	Yes	42526772	rs1346961531	G	A	missense	PRO, SER	8	1.759	1
151	No	Yes	42526811	rs376446555	AT	A	5' UTR	-	-	-	3
152	No	Yes	42526816	rs756687222	C	G	5' UTR	-	-	5.87	3
153	No	Yes	42526836	rs75085559	A	AC	5' UTR	-	-	-	5
154	No	Yes	42527109	-	C	T	upstream	-	-	3.693	1
155	No	Yes	42527116	rs572914357	A	G	upstream	-	-	2.938	2
156	No	Yes	42527158	rs1080992	C	T	upstream	-	-	2.402	4
157	No	Yes	42527191	rs374672076	C	G	upstream	-	-	1.49	116
158	No	Yes	42527381	-	C	T	upstream	-	-	0.566	1
159	No	Yes	42527530	-	T	C	upstream	-	-	0.539	1
160	No	Yes	42527580	rs949806652	C	T	upstream	-	-	4.046	1
161	No	Yes	42527684	rs114155186	C	A	upstream	-	-	0.902	1
162	No	Yes	42527741	rs200937699	G	T	upstream	-	-	0.61	1
163	No	Yes	42527929	rs528127834	C	A	upstream	-	-	0.791	2
164	No	Yes	42527989	rs757622584	TC	T	upstream	-	-	-	1
165	No	Yes	42528392	rs1080984	A	G	upstream	-	-	1.42	1
166	No	No	42528477	rs775251692	C	T	upstream	-	-	5.077	3
167	No	No	42528538	rs58188898	G	A	upstream	-	-	10.95	3
168	No	No	42528568	rs1080983	T	C	upstream	-	-	0.709	414
169	No	No	42528682	rs551293469	G	A	upstream	-	-	4.149	4
170	No	No	42528851	rs28680494	A	C	upstream	-	-	1.804	260

171	No	No	42528858	rs28439297	C	T	upstream	-	-	3.928	260
172	No	No	42528976	rs28360521	C	T	upstream	-	-	2.547	146
173	No	No	42529156	rs73887946	C	T	upstream	-	-	14.44	1
174	No	No	42529176	rs138419008	T	C	upstream	-	-	12.69	2
175	No	No	42529219	rs28568508	A	G	upstream	-	-	4.072	148
176	No	No	42529253	-	T	A	upstream	-	-	4.734	1
177	No	No	42529321	rs28566059	C	T	upstream	-	-	6.32	104
178	No	No	42529366	rs118017929	C	G	upstream	-	-	6.933	5
179	No	No	42529407	rs28542726	G	T	upstream	-	-	7.75	35
180	No	No	42529540	rs28369142	C	A	upstream	-	-	0.099	8

AA, amino acid changes; AC, allele count.

[a]Variant is present in the translation table.

[b]Variant is located within the haplotype definition region.

[c]Human Genome version 19.

[d]SeattleSeq Annotation was used to provide functional annotation for the variants.

Supplementary Table 3.5 Rare nonsynonymous SNVs detected in the Liver Bank for functional characterization and the diplotypes in which SNVs were identified

Variant	Position[a]	rs ID	Reference Allele	Alternative Allele	Diplotype
P8S	42526772	rs1346961531	G	A	**1/*68+*4*
S168A	42525038	rs1135826	A	C	**1/*5*
D198N	42524860	rs763284150	C	T	**1/*1*
P267L	42524219	rs148769737	G	A	**1/*2*
V338M	42523610	rs59421388	C	T	**29/*39*
A449D	42522724	rs79392742	G	T	**1/*4*
R474Q	42522649	rs141756339	C	T	**2/*41*

[a]Human Genome version 19.

Supplementary Materials and Methods

Metabolite and parent drug quantitation

For dextromethorphan and dextrorphan, calibration standards were prepared over the range of 0.004 to 1 μM; dextrorphan-d3 (0.5 ng/μl) was used as an internal standard. Ions m/z 258.1 and 261.1 were monitored for dextrorphan and dextrorphan-d3, respectively. Mobile phase was 0.1% formic acid in water (A) and acetonitrile (B) using the following gradient: 0 min 20% B, 1 min 20% B, 5 min 40% B, 6 min 20% B, and 9 min 20% B.

For metoprolol and α-hydroxymetoprolol, calibration standards were prepared over the range of 0.003 to 0.176 μM; carvedilol-d4 (1 ng/μl) was used as an internal standard. Ions m/z 284.1 and 411.1 were monitored for α-hydroxyl-metoprolol and carvedilol-d4, respectively. Mobile phase was 10 mM ammonium formate, pH 4, in water (A) and acetonitrile (B) using the following gradient: 0 min 15% B, 3.5 min 60% B, 4.5 min 90% B, 5 min 90% B, 5.1 min 15% B, and 9.5 min 15% B.

Synthesis of CYP2D6 probe, ticlopidine 5'-carboxypropargyl amide (tic-ABP1P)

Bulk solvents were obtained from Fisher Scientific (Fair Lawn, NJ). All other chemicals and anhydrous solvents were purchased from Sigma-Aldrich (St. Louis, MO) unless otherwise indicated. Flash chromatography was performed with a Combiflash® Rf purification system (Teledyne Isco, Lincoln, NE). Proton NMR spectra were taken on an Agilent DD2 500 MHz spectrometer at 25°C. Coupling constants (*J*) are reported in Hertz (Hz) and peak multiplicities are denoted by the following abbreviations: s=singlet, d=doublet, t=triplet, dd=doublet of doublets, m=multiplet, b=broad. Chemical shift values (δ) are reported in ppm and are referenced to the residual solvent signals (δ=7.25 for chloroform and δ=3.30 for methanol). High resolution

mass spectroscopy (HRMS) was performed on a Thermo Fisher LTQ Orbitrap equipped with an ESI probe.

Ticlopidine 5'-Methyl Carboxylate

Yield=81% after purification by flash chromatography (0-10% ethyl acetate in hexanes). ^{1}H NMR (500 MHz, $CDCl_3$): δ 8.21 (bs, 1H), 7.88 (bd, *J*=8.32 Hz, 1H), 7.45 (d, *J*=8.32 Hz, 1H), 7.09 (d, *J*=5.12 Hz, 1H), 6.72 (d, *J*=5.12 Hz, 1H), 3.91 (s, 3H), 3.86 (bs, 2H), 3.67 (bs, 2H), 3.0-2.8 (m, 4H). HRMS ESI$^+$ calculated for $C_{16}H_{17}ClNSO_2$ (M+H): 322.0669, found: 322.0673.

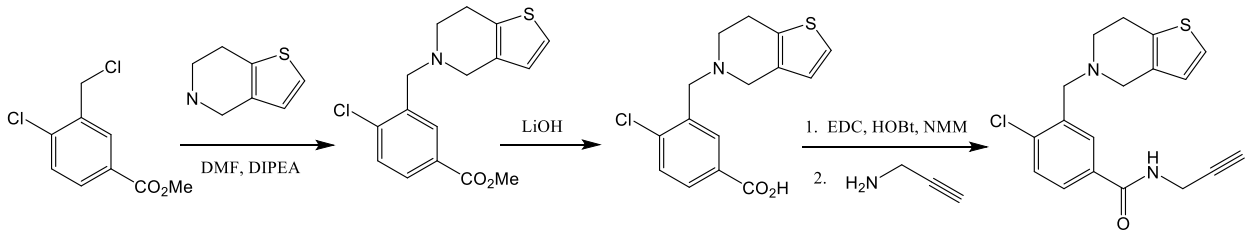
Ticlopidine 5'-Carboxylic Acid

The product acid was obtained as its hydrochloride salt in quantitative yield and was used in the next reaction step without further purification. ^{1}H NMR (500 MHz, CD_3OD): δ 8.34 (d, *J*=2.09 Hz, 1H), 8.09 (dd, *J*=2.09, 8.41 Hz, 1H), 7.69 (d, *J*=8.41 Hz, 1H), 7.35 (d, *J*=5.23 Hz, 1H), 6.86 (d, *J*=5.23 Hz, 1H), 4.60 (s, 2H), 4.31 (bs, 2H), 3.63 (t, *J*=6.15 Hz, 2H), 3.19 (t, *J*=6.15 Hz, 2H). HRMS ESI$^+$ calculated for $C_{16}H1_7ClNSO_2$ (M+H): 308.0512, found: 308.0517.

Ticlopidine 5'-Carboxypropargyl Amide (tic-ABP1P)

Yield=70% after purification by flash chromatography (0-20% ethyl acetate in hexanes). ^{1}H NMR (500 MHz, CD_3OD): δ 8.01 (d, *J*=2.26 Hz, 1H), 7.70 (dd, *J*=2.26, 8.33 Hz, 1H), 7.49 (d, *J*=8.33 Hz, 1H), 7.12 (d, *J*=5.14 Hz, 1H), 6.70 (d, *J*=5.14 Hz, 1H), 4.13 (d, *J*=2.53, 2H), 3.85 (s, 2H), 3.61 (s, 2H), 2.90-2.83 (m, 4H), 2.59 (t, *J*=2.53, 1H). HRMS ESI$^+$ calculated for $C_{16}H_{17}ClNSO_2$ (M+H): 345.0828, found: 345.0831.

Reaction scheme



Construction of yeast expression plasmid: p41KGAL1-CYP2D6-HA

S. cerevisiae codon-optimized CYP2D6 sequence (Uniprot: P10635) was synthesized by a commercial supplier (Integrated DNA Technologies, Coralville, IA) and cloned into a low-copy p41K*GAL1* plasmid.[1] To create CYP2D6 variants, point mutations were engineered with KAPA HiFi DNA Polymerase (Kapa Biosystems, Wilmington, MA) using a standard site-directed mutagenesis method. The final *CYP2D6* plasmid constructs and protein expression levels were verified by Sanger sequencing and Western blot, respectively.

CYP2D6 induction in yeast

CYP2D6 variant proteins were induced in a humanized cytochrome P450 *S. cerevisiae* strain as described previously [**117**] with the following modifications. Yeast strain S288C derivative (*MATα HAP1+ ura3Δ0*::p*GPD-MYC*-b5-URA3 *HO*::p41K*GAL1-POR*-FLAG-*TRP1 leu2Δ1 his3Δ1 trp1Δ63 pep4Δ0 prb1Δ0*) was transformed with p41K*GAL1-CYP2D6*-HA using a standard lithium acetate procedure [**118**] and was propagated on liquid media supplemented with 200 μg/ml G418 to maintain the plasmid. At an OD_{600} of 0.025, cells were pelleted, and resuspended in fresh media containing 2% (w/v) galactose to induce the over-expression of CYP2D6-HA. Cells were grown at 30°C and collected after 7 doublings for western blotting and functional assay.

Functional activity of CYP2D6 variant proteins in yeast

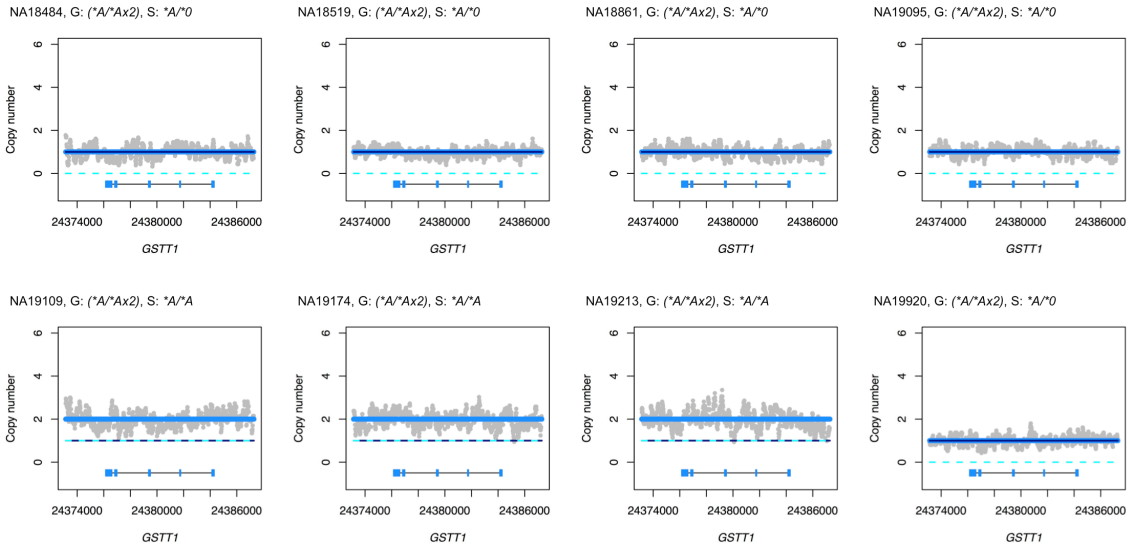
Isogenic yeast strains were generated for the induction of three control (wildtype, C443H, and P34S) and seven novel (P8S, S168A, D198N, P267L, V338M, A449D, and R474Q) *CYP2D6* variant genes. Each expression construct contained only a single variant. P34S confers decreased function of CYP2D6 [**58**]. C443H is a synthetic nonfunctional allele resulting in a catalytically inactive control [**120**,**121**,**122**]. CYP2D6 enzymatic activity was probed using a click chemistry compatible ticlopidine 5'-carboxypropargyl amide activity-based protein profiling probe (tic-ABP1P) that have specificity for CYP2D6 activity with minimum reactivity towards other yeast proteins. Yeast cells induced with CYP2D6 variants were collected at OD_{600} of 1, pelleted and saponin-permeabilized cells were pre-incubated with 2 mM NADPH (Sigma-Aldrich) for 20 min at 30°C. Next, cells were exposed to tic-ABP1P (20 μM) for 17 h in a rotator in the dark at 37°C to form CYP2D6 activity-dependent P450-probe adducts. To append a fluorophore reporter (Alexa FluorTM 488 picolyl azide) via a copper-catalyzed azide-alkyne cycloaddition reaction [**119**], the Click-iTTM kit (Sigma-Aldrich) was used. PBS washed cells were incubated with a Click-iTTM Plus reaction cocktail according to the manufacturer's instructions. Cells were washed in PBS 5 times prior to flow cytometry analysis. For each strain, a total of 20,000 events were collected using the flow cytometer BDTM LSRII with the FITC channel (BD Bioscience, San Jose, CA). A standard yeast gate was applied to all cytometry data, and fluorescence data was analyzed using FlowJo (Ashland, OR). The geometric mean fluorescence (GMF) of each strain was normalized to the GMF of the wildtype CYP2D6 control strain, and ratios of normalized GMF of variants to normalized GMF of wildtype CYP2D6 were calculated and converted to percentage.

CYP2D6 activity was also confirmed using a standard fluorogenic CYP2D6 substrate, Vivid^{TM} 7-ethoxymethoxy-3-cyanocoumarin. Yeast cells induced with CYP2D6 variants were

collected at OD_{600} of 2 and potassium phosphate buffer resuspended cells were pre-incubated with 2 mM NADPH (Sigma-Aldrich) for 20 min at 30°C. Next, 100 µl of cells containing 20 µM of Vivid^{TM} substrate were dispensed into each well of a black 96-well plate and incubated for an additional 20 min at 37°C. Fluorescence unit readings were collected over ~2 hours in a plate shaker using 408 nm and 450 nm excitation and emission wavelengths, respectively. The relative fluorescence unit of each strain was normalized to that of control wells containing buffer and substrate alone.

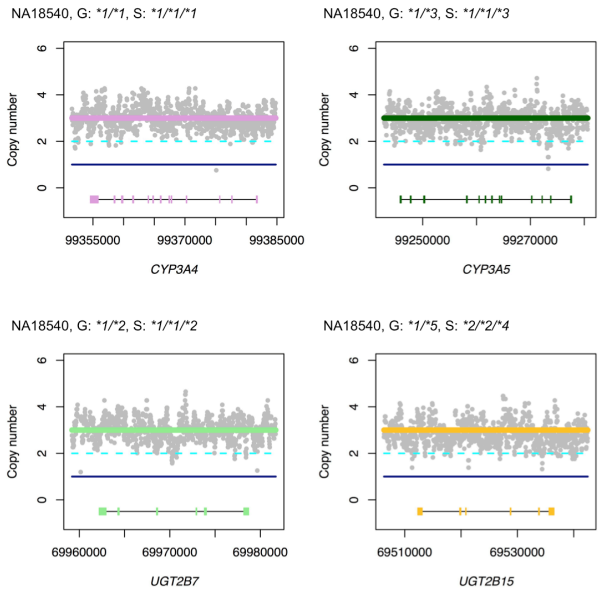
Appendix C: Supplementary Materials for Chapter 4

Supplementary Figures and Tables



Supplementary Figure 1 Whole genome sequencing data for samples previously reported by GeT-RM to have more than two gene copies of *GSTT1* (*GSTT1*AxN)*

Panels display Stargazer's result for copy number analysis for individual samples (N=8). Genotypes from GeT-RM and Stargazer (abbreviated as "G" and "S" for brevity) are also shown, with "()" indicating non-consensus genotypes. Grey dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined. Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. Reports in the Database of Genomic Variants supported Stargazer's gene deletion calls in NA18519, NA18861, NA19095, and NA19920.



Supplementary Figure 2 Four gene duplications found in a single sample (NA18540) by Stargazer
Panels display Stargazer's result for copy number analysis for the sample NA18540. Genotypes from GeT-RM and Stargazer (abbreviated as "G" and "S" for brevity) are also shown. Grey dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined. Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. This result is indicative of trisomy in chromosomes 4 and 7, and has been independently confirmed by [**134**].

Supplementary Table 4.1 Genotypes for 70 reference samples and 28 pharmacogenes identified by Stargazer's analysis of whole genome sequencing data

For comparison, GeT-RM genotypes previously published by [**128**] are also shown.

	CYP1A1		*CYP1A2*		*CYP2A6*		*CYP2B6*	
Sample ID	GeT-RM	Stargazer	GeT-RM	Stargazer	GeT-RM	Stargazer	GeT-RM	Stargazer
HG00276	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/*1	*1/*1	*2/(*4)	*2/*4
HG00436	*1/*1	*1/*13 [A]	*1A/*1F	*1A/*1F	*9/*9	*4/*1+*S6 [N]	*1/*6	*1/*6
HG00589	*2/*2	*2B/*2B [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*1	*1/*1	*1/*1	*1/*1
HG01190	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1 (*5)/*1 (*27)	*1/*5
NA06991	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*1/*1	*1/*1	*1/*6	*1/*6
NA07000	(*1/*1)	*1/*1	(*1F/*1F)	*1F/*1F	(*1/*1)	*1/*1	*1/*1	*1/*1
NA07019	*1/*4	*1/*4	*1A/*1F	*1A/*1F	*1/*1	*1/*1	*1 (*5)/*1 (*22)	*5/*22
NA07029	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/*1	*1/*1	*6/(*27)	*1/*6
NA07055	*1/*5	*1/*5	*1F/*1F	*1F/*1F	*1/*1	*1/*22 [A]	*6/(*27)	*1/*6
NA07056	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*6/(*22)	*6/*22
NA07348	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*1/*1	*1/*1	*1/*1	*1/*1
NA07357	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*1/*1	*1/*1	*1/*1	*1/*1
NA10831	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/*2	*1/*2	*1/*1	*1/*1
NA10847	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*1/*1	*1/*1	*1/*6	*1/*6
NA10851	*1/*1	*1/*2A [A]	*1F/*1F	*1F/*1F	*1/*2	*1/*2	*1/*1 (*27)	*1/*1
NA10854	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1
NA11832	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*1/*2	*1/*2	*1/*1	*1/*1
NA11839	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/*9	*1/*9	*1/*1 (*15)	*1/*15
NA11993	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*9/*17	*9/*17	*1/*1	*1/*1
NA12003	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*1/(*8)	*1/*1	*6/*6	*6/*6
NA12006	*1/*4	*1/*4	*1F/*1F	*1F/*1F	*1/*1	*1/*1	*1/*1 (*5)	*1/*5
NA12145	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1
NA12156	*1/*2	*1/*2B [A]	*1A/*1F	*1A/*1F	*1/*1	*1/*1	*1/*1	*1/*1
NA12717	*1/*1	*1/*1	*1F/*1F	*1F/*1F	*1/*1	*1/*1	*1/*7	*5/*6 [P]
NA12813	*1/*2	*2A/*2B [A]	*1F/*1L	*1F/*1L	*1/*1	*1/*1	*1/*2	*1/*2
NA12873	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/*9	*1/*9	*1/*6	*1/*6
NA18484	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*9	*1/*9	*1/*18	*1/*18
NA18509	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/*17	*1/*17	*1/*6	*1/*6
NA18518	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*17	*1/*17	*1/*6	*1/*6
NA18519	*1/*1	*1/*1	(*1A/*1L, *1C/*1F)	*1A/*1L [P]	*1/*1	*1/*1	*1/*6	*6/*17 [A]
NA18524	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/*1	*1/*1	*1/*1 (*27)	*1/*1
NA18526	*1/*1	*1/*2A [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*9	*7/*9 [A]	*1/*1	*1/*1
NA18540	*1/*2	*13/*2B [A]	*1L/*1L	*1L/*1L	*1/*1	*1/*1	*1/*6	*1/*6
NA18544	*1/*2	*2A/*2B [A]	*1F/*1L	*1F/*1L	*1/*1	*7/*18 [A]	*1/*1	*1/*1
NA18552	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*9/*9	*9/*9	*1/*6	*1/*6
NA18564	*1/*1	*1/*13 [A]	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1
NA18565	*2/*2	*2B/*2B [A]	*1F/*1F	*1F/*1F	*9/*9	*4/*15 [A]	*1/*1	*1/*1
NA18617	(*1/*1)	*1/*1	(*1A/*1A)	*1A/*1A	(*1/*4)	*1/*4	*1/*1	*1/*4
NA18855	*1/*1	*1/*1	(*1A/*1L, *1C/*1F)	*1A/*1L [P]	*1/(*8)	*1/*1	*6/*6	*6/*6
NA18861	*1/*1	*1/*1	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*1	*25/*1x2 [A]	*1/*18	*1/*18
NA18868	*1/*1	*1/*2A [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*17	*1/*17	*1/*6	*1/*6
NA18942	*1/*1	*13/*13 [A]	*1A/*1F	*1A/*1F	*1/*1	*4/*7 [A]	*1/*2	*1/*2
NA18952	*1/*2	*13/*2B [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*4/*4	*4/*4	*1/*1	*1/*1
NA18959	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/(*4)	*1/*4	*1/*1 (*27)	*1/*1
NA18966	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*1/(*4)	*1/*4	*1/*6	*1/*6
NA18973	*1/*1	*1/*1	*1F/*1L	*1F/*1L	*4/*4	*4/*4	*1/*6	*1/*6
NA18980	*1/*2	*13/*2B [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*9/*9	*9/*9	*6/*6	*6/*6
NA18992	*1/*1	*1/*13 [A]	*1A/*1A	*1A/*1A	*1/*9	*9/*19 [A]	*1/*6	*1/*6
NA19003	*1/*1	*1/*13 [A]	*1A/*1F	*1A/*1F	*1/*1	*1/*1	*1/*6	*1/*6
NA19007	*1/*2	*13/*2B [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/(*4)	*1/*4	*1/*1	*1/*23 [A]
NA19095	*1/*1	*1/*1	(*1A/*1L, *1C/*1F)	*1A/*1L [P]	*9/*9	*9/*9	*18/*18	*18/*18
NA19109	*1/*1	*1/*1	*1A/*1F	*1A/*1F	*17/*20	*17/*20	*1/*6	*1/*6
NA19122	*1/*1	*1/*1	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*1	*1/*35 [A]	*6/*6	*6/*6
NA19143	*1/*1	*1/*2A [A]	*1A/*1F	*1A/*1F	*1/*1	*1/*35 [A]	*6/*6	*6/*6
NA19147	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*9	*9/*23 [A]	*1/*18	*1/*18
NA19174	*1/*1	*1/*2A [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*9	*1/*9	*6/*18	*6/*18
NA19176	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/(*8)	*1/*1	*6/*6	*6/*6
NA19178	*1/*1	*1/*2A [A]	*1L/*1L	*1L/*1L	*1/*20	*1/*20	*6/*6	*6/*29 [A]
NA19207	*1/*1	*1/*2A [A]	*1A/*1F	*1A/*1F	*1/*17	*1/*17	*1/*6	*1/*6
NA19213	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*17	*17/*24 [A]	*6/*6	*6/*6
NA19226	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*17	*1/*17	*18/(*20)	*18/*20
NA19239	*1/*1	*1/*2A [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*17	*1/*17	*1/*6	*1/*6
NA19789	*1/*2	*1/*2B [A]	*1A/*1L, *1C/*1F	*1A/*1L [P]	*1/*1	*1/*1	*1/*1 (*27)	*1/*1
NA19819	*1/*1	*1/*2A [A]	*1F/*1L	*1F/*1L	*1/*1	*1/*1	*1/*1	*1/*1
NA19908	*1/*2	*2A/*2B [A]	*1L/*1L	*1L/*1L	*1/*17	*1/*17	*1/*6	*1/*6
NA19917	(*1/*1)	*1/*1	(*1A/*1F)	*1A/*1F	(*1/*1)	*1/*1	*6/*6	*6/*6
NA19920	*1/*1	*1/*1	*1A/*1A	*1A/*1A	*1/*1	*1/*35 [A]	*6/*6	*6/*6
NA20296	*1/*1	*1/*2A [A]	*1L/*1L	*1L/*1L	*1/*1	*1/*1	*1/*2	*1/*2
NA20509	*1/*1	*1/*2A [A]	*1A/*1F	*1A/*1F	*1/*1	*1/*1	*1/*7	*5/*6 [P]
NA21781	(*1/*1)	*1/*1	(*1F/*1F)	*1F/*1F	(*1/*1)	*1/*21 [A]	*1/*1	*1/*4

	CYP2C8		*CYP2C9*		*CYP2C19*		*CYP2D6*	
Sample ID	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**
HG00276	*1/*3	*1/*3	*1/*2	*1/*2	*1/*1	*1/*1	*4/*5	*4/*5
HG00436	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2x2	*1/*2x2
HG00589	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2 (*21)	*2/*21 [P]
HG01190	*1/*3	*1/*3	*1/*2	*1/*2	*1/*2	*1/*2	*4/*5	*5/*68+*4 [A]
NA06991	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*4	*1/*4
NA07000	*1/*1	*1/*1	*1/*1	*1/*1	*1/*17	*1/*17	*9/*2 (*35)	*9/*35
NA07019	*1/*1	*1/*1	*1/*1	*1/*1	*1/*17	*1/*17	*1/*4	*1/*4
NA07029	*1/*3	*1/*3	*1/*2	*1/*2	*8/*17	*8/*17	*1/*35	*1/*35
NA07055	*1/*1	*1/*1	*1/*1	*1/*1	*1/*17	*1/*17	*4/*4	*4/*4
NA07056	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*2/*4	*2/*4
NA07348	*1/*4	*1/*4	*1/*1	*1/*1	*2/*17	*2/*17	*1/*6	*1/*6
NA07357	*1/*4	*1/*4	*1/*1	*1/*1	*2/*17	*2/*17	*1/*6	*1/*6
NA10831	*1/*1	*1/*1	*1/*2	*1/*2	*1/*17	*1/*17	*4/*5	*4/*5
NA10847	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*41	*1/*41
NA10851	*1/*1	*1/*1	*1/*1	*1/*1	*1/*17	*1/*17	*1/*4	*1/*4
NA10854	*3/*3	*3/*3	*2/*2	*2/*2	*1/*1	*1/*1	*1/*4	*1/*4
NA11832	*1/*1	*1/*1	*1/*3 (*18)	*1/*3	*1/*2	*1/*2	*1/*4	*1/*68+*4 [A]
NA11839	*1/*3	*1/*3	*2/*3 (*18)	*2/*3	*1/*1	*1/*1	*1/*2	*1/*2
NA11993	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*9	*1/*9
NA12003	*1/*3	*1/*3	*1/*2	*1/*2	*1/*1	*1/*1	*4/*35	*4/*35
NA12006	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*4/*41	*4/*41
NA12145	*1/*4	*1/*4	*1/*1	*1/*1	*2/*17	*2/*17	*1/*4	*1/*4
NA12156	*1/*1	*1/*1	*1/*2	*1/*2	*1/*1	*1/*1	*1/*4	*1/*4
NA12717	*1/*1	*1/*1	*1/*1	*1/*1	*2/*2	*2/*2	*1/*1	*1/*1
NA12813	*1/*1	*1/*1	*1/*3 (*18)	*1/*3	*1/*17	*1/*17	*2/*4	*2/*4
NA12873	*1/*1	*1/*1	*1/*1	*1/*1	*1/*17	*1/*17	*1/*5	*1/*5
NA18484	*1/*4	*1/*4	*1/*9	*1/*9	*1 (*27)/*2	*2/*27	*1/*17	*1/*17
NA18509	*1/*1	*1/*1	*1/*1	*1/*1	*2/*2	*2/*2	*2/*17	*2/*17
NA18518	*1/*2	*1/*2	*1/*1	*1/*1	*2/*17	*2/*17	*17/*29	*17/*29
NA18519	*1/*2	*1/*2	*1/*5	*1/*5	*1/*17	*1/*17	*1/*29	*1/*29
NA18524	*1/*1	*1/*1	*1/*3	*1/*3	*1/*2	*1/*2	*1/(*36+*10)	*36+*10/*36+*10
NA18526	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/(*36+*10)	*36+*10/*36+*10
NA18540	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*2	*10/*41	*36+*10/*36+*10
NA18544	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*2	*10/*41	*10/*41
NA18552	*1/*1	*1/*1	*1/*1	*1/*1	*1/*4	*1/*4	*1/*14	*1/*14
NA18564	*1/*1	*1/*1	*1/*1	*1/*1	*2/*3	*2/*3	*2/(*36+*10)	*2/*36+*10
NA18565	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*10/(*36+*10)	*10/*36+*10
NA18617	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*2	*10/(*36+*10)	*36+*10/*36+*10
NA18855	*1/*1	*1/*1	*1/*9	*1/*9	*1 (*27)/*2	*2/*27	*1/*5	*1/*5
NA18861	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*5/*29	*5/*29
NA18868	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*2	*2/*5	*2/*5
NA18942	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*2/*2	*2/*2
NA18952	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*2/*2	*2/*2
NA18959	*1/*1	*1/*1	*1/*3 (*18)	*1/*3	*1/*1	*1/*1	*2/(*36+*10)	*2/*36+*10
NA18966	*1/*1	*1/*1	*1/*1	*1/*S1 [N]	*1/*1	*1/*1	*1/*2	*1/*2
NA18973	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2 (*21)	*2/*21 [P]
NA18980	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*2/(*36+*10)	*2/*36+*10
NA18992	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*5	*1/*5
NA19003	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*2	*1/*1	*1/*1
NA19007	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA19095	*1/*2	*1/*2	*1/*1	*1/*1	*1/*1	*1/*1	*1/*29	*1/*29
NA19109	*2/*2	*2/*2	*1/*1	*1/*1	*17/*17	*17/*17	*29/*2x2	*29/*2x2
NA19122	*1/*1	*1/*1	*1/*11	*1/*11	*1 (*15)/*2	*2/*35 [A]	*2/*17	*2/*17
NA19143	*1/*1	*1/*1	*1/*6	*1/*6	*1/*15	*1/*15	*2/*10	*2/*10
NA19147	*1/*1	*1/*1	*1/*1	*1/*1	*1/*17	*1/*17	*17/*29	*17/*39 [P] [A]
NA19174	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*2	(*4/*40)	*4/*40
NA19176	*2/*2	*2/*2	*1/*1	*1/*1	*2/*17	*2/*17	*1/*2	*1/*2
NA19178	*1/*1	*1/*1	*5/*9	*5/*9	*1 (*27)/*6	*6/*27	*1/*1	*1/*1
NA19207	*1/*2	*1/*2	*1/*1	*1/*1	*2/*17	*2/*17	*2/*10/*xN	*10/*34x2 [A]
NA19213	*1/*1	*1/*1	*1/*6	*1/*6	*1/*15	*1/*15	*1/*1	*1/*1
NA19226	*1/*1	*1/*1	*1/*8	*1/*8	*1/*2	*1/*2	*2/*2x2	*2/*2x2
NA19239	*1/*2	*1/*2	*1/*1	*1/*1	*13/*17	*13/*17	*15/*17	*15/*17
NA19789	*1/*3	*1/*3	*1/*2	*1/*2	*1/*1	*1/*1	*1/*1	*1/*1
NA19819	*1/*2	*1/*2	*1/*1	*1/*1	*1/*17	*1/*17	*2/*4/*xN	*2/*4x2 [A]
NA19908	*1/*1	*1/*1	*1/*5	*1/*5	*1/*17	*1/*17	*1/*2	*1/*46 [A]
NA19917	*1/*1	*1/*1	*1/*1 (*18)	*1/*1	*1 (*15)/*2	*2/*15	*1/*17 (*40)	*1/*40
NA19920	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*4/*xN	*1/*4x2 [A]
NA20296	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*34 [A]
NA20509	*1/*4	*1/*4	*1/*1	*1/*1	*2/*2	*2/*2	*4/*35	*4/*35
NA21781	*1/*1	*1/*1	*1/*1	*1/*1	*1/*2	*1/*2	*2/*4 (*xN)	*2x2/*68+*4 [A]

	CYP2E1		*CYP3A4*		*CYP3A5*		*CYP4F2*	
Sample ID	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**
HG00276	*1/*1	*1/*1	*1/*2	*1/*2	*3/*3	*3/*3	*1/*1	*1/*1
HG00436	(*7/*7)	*7/*7	*1/*1	*1/*1	*3/*3	*3/*3	(*2)/*3	*1/*3, *2 [P]
HG00589	(*5)/*7	*1/*7, *5 [P]	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
HG01190	*1/*7	*1/*7	*1/*1B	*1/*1B	*1/*1	*1/*1	*1/*3	*1/*3
NA06991	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA07000	(*1/*1)	*1/*1	*1/*1	*1/*1	*1/*3	*1/*3	(*3/*3)	*3/*3
NA07019	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*3	*1/*3
NA07029	*1/*1	*1/*1	*1/*1	*1/*1	*1/*3	*1/*3	*3/*3	*3/*3
NA07055	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA07056	*1/*1	*1/*1	*1/*22	*1/*22	*3/*3	*3/*3	*3/*3	*3/*3
NA07348	*1/*7	*1/*7	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA07357	(*7)/*7	*7/*7	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA10831	*1/*7	*1/*7	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA10847	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA10851	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA10854	*1/*1	*1/*1	*1/*1B	*1/*1B	*1/*3	*1/*3	*1/*3	*1/*3
NA11832	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA11839	(*5)/*7	*1/*7, *5 [P]	*1/*1B	*1/*1B	*1/*3	*1/*3	*1/*3	*1/*3
NA11993	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*3	*1/*3
NA12003	*1/*1	*1/*1	*1/*1B	*1/*1B	*1/*3	*1/*3	*1/*1	*1/*1
NA12006	*1/*1	*1/*1	*1/*3	*1/*3	*3/*3	*3/*3	(*2)/*3	*1/*3
NA12145	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA12156	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	(*2)/*3	*1/*3, *2 [P]
NA12717	*1/*1	*1/*1	*1B/*22	*1B/*22	*1/*3	*1/*3	*1/*3	*1/*3
NA12813	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*3	*1/*3
NA12873	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA18484	*1/*7	*1/*7	*1B/*1B	*1B/*1B	*1/*7	*1/*7	*1/*2	*1/*2
NA18509	(*7/*7)	*7/*S1 [N]	*1/*1B	*1/*1B	*1/*7	*1/*7	(*2)/*3	*1/*3, *2 [P]
NA18518	(*4)/*7	*1/*7, *4 [P]	*1B/*1B	*1B/*1B	*1/*6	*1/*6	*1/*2	*1/*2
NA18519	*1/*1	*1/*1	*1B/*1B	*1B/*1B	*1/*6	*1/*6	(*2)/*3	*1/*3, *2 [P]
NA18524	(*7)/*7	*7/*7	*1/*1	*1/*1	*1/*3	*1/*3	*1/*3	*1/*3
NA18526	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*3	*1/*3
NA18540	(*5)/*7	*1/*7, *5 [P]	*1/*1	*1/*1/*1 [T]	*1/*3	*1/*1/*3 [T]	*1/*3	*1/*3
NA18544	*7/*7	*7/*7	*1/*1	*1/*1	*1/*3	*1/*3	*1/*1	*1/*1
NA18552	(*5)/*7	*5/*7	*1/*1	*1/*1	*3/*3	*3/*3	*1/*3	*1/*3
NA18564	(*5)/*7	*1/*7, *5 [P]	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3
NA18565	*1/*1	*1/*1	*1/*1	*1/*1	*1/*3	*1/*3	*1/*1	*1/*1
NA18617	(*7/*7)	*7/*7	*1/*1	*1/*1	*3/*3	*3/*3	(*1/*1)	*1/*1
NA18855	(*7)/*7	*7/*7	*1/*1B	*1/*1B	*3/*6	*3/*6	*1/*1	*1/*1
NA18861	(*7)/*7	*7/*7	*1B/*1B	*1B/*1B	*1/*1	*1/*1	*1/*1	*1/*1
NA18868	(*7)/*7	*7/*7	*1B/*1B	*1B/*1B	*1/*3	*1/*3	*1/(*2)	*1/*1
NA18942	(*5)/*7	*1/*7, *5 [P]	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA18952	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA18959	*1/*1	*1/*1	*1/*1	*1/*1	*1/*3	*1/*3	*1/*1	*1/*1
NA18966	*1/*1	*1/*1	*1/*1 (*16)	*1/*16	*1/*3	*1/*3	*1/*3	*1/*3
NA18973	*1/*7	*1/*7	*1/*1	*1/*1	*1/*3	*1/*3	*1/*1	*1/*1
NA18980	(*7/*7)	*7/*7	*1/*1	*1/*1	*1/*3	*1/*3	*1/*1	*1/*1
NA18992	(*7)/*7	*7/*7	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA19003	*1/*7	*1/*7	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA19007	(*5)/*7	*1/*7, *5 [P]	*1/*1	*1/*1	*3/*3	*3/*3	*1/*3	*1/*3
NA19095	*1/*7	*1/*7x2 [A]	*1/*1B	*1/*1B	*1/*3	*1/*3	*1/*3	*1/*3
NA19109	*1/*1	*1/*1	*1B/*1B (*15)	*1B/*15	*1/*3	*1/*3	*1/*1	*1/*1
NA19122	*7/*7	*7/*7	*1B/*1B	*1B/*1B	*1/*1	*1/*1	*1/(*2)	*1/*2
NA19143	(*7/*7)	*7/*S1 [N]	*1B/*1B	*1B/*1B	*6/*7	*6/*7	*1/*1	*1/*1
NA19147	(*4)/*7	*1/*S1, *7, *4 [P] [N]	*1/*1B	*1/*1B	*1/*3	*1/*3	*1/*1	*1/*1
NA19174	(*7/*7)	*7/*7	*1B/*1B	*1B/*1B	*1/*6	*1/*6	*1/*1	*1/*1
NA19176	(*7)/*7	*7/*7	*1B/*1B	*1B/*1B	*1/*3	*1/*3	*1/*1	*1/*1
NA19178	(*7)/*7	*7/*7	*1B/*1B	*1B/*1B	*1/*1	*1/*1	*1/*1	*1/*1
NA19207	*7/*7	*7/*7	*1B/*1B	*1B/*1B	*3/*7	*3/*7	*1/*1	*1/*1
NA19213	(*7)/*7	*7/*7	*1B/*1B	*1B/*1B	*1/*6	*1/*6	*1/*1	*1/*1
NA19226	*1/*7	*1/*7x2 [A]	*1B/*1B (*15)	*1B/*15	*1/*6	*1/*6	*1/(*2)	*1/*2
NA19239	(*7)/*7	*7/*7	*1/*1B	*1/*1B	*1/*1	*1/*1	*1/*1	*1/*1
NA19789	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	(*2)/*3	*1/*3, *2 [P]
NA19819	*1/*7	*1/*7	*1/*1B	*1/*1B	*3/*6	*3/*6	(*2)/*3	*1/*3, *2 [P]
NA19908	*7/*7	*7x2/*7x2 [A]	*1B/*1B (*15)	*1B/*15	*1/*3	*1/*3	(*2)/*3	*2/*3
NA19917	(*1/*7)	*1/*7	*1/*1	*1/*1B	*1/*7	*1/*7	(*1/*1)	*1/*2
NA19920	(*4)/*7	*1/*S1, *7, *4 [P] [N]	*1B/*1B	*1B/*1B	*7/*7	*7/*7	*1/*1	*1/*1
NA20296	*1/*1	*1/*1	*1B/*1B	*1B/*1B	*1/*6	*1/*6	*1/*1	*1/*1
NA20509	*1/*1	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	*1/*1	*1/*1
NA21781	(*1/*1)	*1/*1	*1/*1	*1/*1	*3/*3	*3/*3	(*1/*1)	*1/*1

	DPYD		*GSTM1*		*GSTP1*		*GSTT1*	
Sample ID	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**
HG00276	*1/*1	*1/*1	*0/*0	*0/*0	*A/*B	*A/*B	(*A/*A)	*A/*0
HG00436	*1/*1	*1/*1	*0/*0	*0/*0	*A/*B	*A/*B	*0/*0	*0/*0
HG00589	*1/*1	*1/*5 [A]	*0/*0	*0/*0	*A/*A	*A/*A	*0/*0	*0/*0
HG01190	*1/*9	*1/*9	*0/*0	*0/*0	*A/*B	*A/*B	*0/*0	*0/*0
NA06991	*1/(*4)	*1/*4	*A/*B	*A/*B	*B/*B	*B/*B	(*A/*0)	*A/*0
NA07000	*1/*1	*1/*1	(*A/*A)	*A/*A	(*A/*A)	*A/*A	(*A/*0)	*A/*0
NA07019	*1/*1	*1/*1	*0/*0	*0/*0	*B/*B	*B/*B	(*A/*0)	*A/*0
NA07029	*1/*1	*1/*1	*A/*A	*A/*0	*A/*A	*A/*A	(*A/*0)	*A/*0
NA07055	*1/*1	*1/*5 [A]	*A/*A	*A/*0	*A/*A	*A/*A	(*A/*A)	*A/*A
NA07056	*1/*1	*1/*1	*0/*0	*0/*0	*A/*B	*A/*B	*0/*0	*0/*0
NA07348	*1/*9	*1/*6 [A]	*A/*A	*A/*0	*A/*A	*A/*A	(*A/*A)	*A/*0
NA07357	*1/*9	*5/*6 [A]	*0/*0	*0/*0	*A/*A	*A/*A	*0/*0	*0/*0
NA10831	*1/*9	*5/*9 [A]	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*A)	*A/*A
NA10847	*1/*9	*1/*9	*A/*A	*A/*0	*A/*B	*A/*B	(*B/*0)	*B/*0
NA10851	*1/*9	*1/*9	*0/*0	*0/*0	*A/*C, *B/*D	*A/*C [P]	*0/*0	*0/*0
NA10854	*1/*1	*1/*1	*0/*0	*0/*0	*A/*C, *B/*D	*A/*C [P]	(*A/*A)	*A/*A
NA11832	*1/*1	*1/*5 [A]	*0/*0	*0/*0	*A/*B	*A/*B	*0/*0	*0/*0
NA11839	*1/*1	*1/*5 [A]	*0/*0	*0/*0	*A/*B	*A/*B	(*A/*0)	*A/*0
NA11993	*1/*1	*1/*5 [A]	*0/*0	*0/*0	*A/*B	*A/*B	(*A/*0)	*A/*0
NA12003	*1/*1	*1/*1	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*A)	*A/*0
NA12006	*1/*1	*1/*1	*A/*A	*A/*0	*A/*A	*A/*A	(*A/*A)	*A/*A
NA12145	*1/*9	*1/*9	*0/*0	*0/*0	*A/*C, *B/*D	*A/*C [P]	(*A/*0)	*A/*0
NA12156	*1/*1	*1/*5 [A]	*0/*0	*0/*0	*A/*B	*A/*B	(*A/*A)	*A/*A
NA12717	*1/*1	*1/*1	*A/*A	*A/*0	*A/*B	*A/*B	(*A/*A)	*A/*A
NA12813	*1/(*4)	*1/*4	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*A)	*A/*0
NA12873	*1/*1	*1/*1	*0/*0	*0/*0	*A/*C, *B/*D	*A/*C [P]	(*A/*0)	*A/*0
NA18484	*1/*9	*1/*9	*A/*A	*A/*0	*B/*B	*B/*B	(*A/*AxN)	*A/*0
NA18509	*9/*9	*9/*9	*0/*0	*0/*0	*B/*B	*B/*B	(*A/*A)	*A/*0
NA18518	*1/*9	*5/*9 [A]	*A/*A	*A/*0	*A/*A	*A/*A	(*A/*A)	*A/*0
NA18519	*9/*9	*9/*9	*A/*A	*A/*A	*A/*A	*A/*A	(*A/*AxN)	*A/*0
NA18524	*1/*1	*1/*5 [A]	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*A)	*A/*A
NA18526	*1/*1	*1/*1	*0/*0	*0/*0	*A/*B	*A/*B	*0/*0	*0/*0
NA18540	*1/*9	*1/*9	*B/*B	*B/*B	*A/*A	*A/*A	*0/*0	*0/*0
NA18544	*1/*1	*1/*5 [A]	*B/*B	*B/*0	*B/*B	*B/*B	(*A/*0)	*A/*0
NA18552	*1/*1	*1/*5 [A]	*B/*B	*B/*0	*A/*A	*A/*A	(*A/*A)	*A/*0
NA18564	*1/*9	*1/*9	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*0)	*A/*0
NA18565	*1/*1	*1/*1	*0/*0	*0/*0	*A/*B	*A/*B	(*A/*0)	*A/*0
NA18617	*1/*1	*1/*1	(*0/*0)	*0/*0	(*A/*A)	*A/*A	(*A/*A)	*A/*0
NA18855	*9/*9	*9/*9	*A/*A	*A/*0	*A/*B	*A/*B	(*A/*A)	*A/*A
NA18861	*1/*9	*1/*9	*A/*B	*A/*B	*A/*B	*A/*B	(*A/*AxN)	*A/*0
NA18868	*1/*9	*1/*9	*A/*A	*A/*A	*A/*A	*A/*A	(*A/*A)	*A/*0
NA18942	*1/*1	*1/*1	*B/*B	*B/*0	*A/*A	*A/*A	*0/*0	*0/*0
NA18952	*1/*1	*5/*5 [A]	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*A)	*A/*A
NA18959	*1/*1	*1/*1	*B/*B	*B/*B	*A/*A	*A/*A	(*A/*A)	*A/*0
NA18966	*1/*1	*1/*1	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*A)	*A/*A
NA18973	*1/*9	*5/*9 [A]	*A/*B	*A/*B	*A/*A	*A/*A	(*A/*A)	*A/*0
NA18980	*1/*1	*1/*5 [A]	*A/*A	*A/*0	*A/*A	*A/*A	(*A/*A)	*A/*0
NA18992	*1/*1	*1/*1	*B/*B	*B/*0	*A/*A	*A/*A	(*A/*0)	*A/*0
NA19003	*1/*1	*1/*1	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*0)	*A/*0
NA19007	*1/*1	*1/*1	*0/*0	*0/*0	*A/*A	*A/*A	*0/*0	*0/*0
NA19095	*9/*9	*9/*9	*A/*A	*A/*A	*A/*B	*A/*B	(*A/*AxN)	*A/*0
NA19109	*1/*9	*1/*9	*A/*A	*A/*A	*A/*B	*A/*B	(*A/*AxN)	*A/*A
NA19122	*1/*1	*1/*1	*A/*A	*A/*0	*A/*B	*A/*B	(*A/*A)	*A/*0
NA19143	*1/*1	*1/*1	*A/*A	*A/*0	*A/*B	*A/*B	*0/*0	*0/*0
NA19147	*1/*9	*1/*9	*A/*B	*A/*B	*A/*B	*A/*B	(*A/*0)	*A/*0
NA19174	*1/*1	*1/*1	*0/*0	*0/*0	*A/*B	*A/*B	(*A/*AxN)	*A/*A
NA19176	*1/*1	*1/*1	*A/*A	*A/*0	*A/*B	*A/*B	*0/*0	*0/*0
NA19178	*1/*9	*1/*9	*A/*A	*A/*A	*A/*B	*A/*B	*0/*0	*0/*0
NA19207	*1/*9	*1/*9	*0/*0	*0/*0	*A/*B	*A/*B	*0/*0	*0/*0
NA19213	*1/*1	*1/*1	*A/*A	*A/*0	*A/*B	*A/*B	(*A/*AxN)	*A/*A
NA19226	*1/*9	*1/*9	*A/*A	*A/*A	*A/*A	*A/*A	*0/*0	*0/*0
NA19239	*1/*9	*1/*9	*A/*A	*A/*A	*A/*A	*A/*A	*0/*0	*0/*0
NA19789	*1/*1	*1/*5 [A]	*B/*B	*B/*0	*A/*A	*A/*A	(*A/*0)	*A/*0
NA19819	*1/*9	*1/*9	*A/*B	*A/*B	*A/*A	*A/*A	(*A/*A)	*A/*A
NA19908	*1/*9	*1/*9	*A/*A	*A/*Ax2 [A]	*B/*C	*B/*C	(*A/*A)	*A/*0
NA19917	*9/*9	*9/*9	(*A/*A)	*A/*A	(*A/*A)	*A/*A	(*A/*A)	*A/*0
NA19920	*9/*9	*9/*9	*A/*A	*A/*0	*A/*B	*A/*B	(*A/*AxN)	*A/*0
NA20296	*1/*9	*1/*9	*A/*A	*A/*0	*B/*B	*B/*B	(*A/*A)	*A/*A
NA20509	*1/*9	*1/*9	*0/*0	*0/*0	*A/*A	*A/*A	(*A/*A)	*A/*A
NA21781	*1/*1	*1/*5 [A]	(*0/*0)	*0/*0	(*A/*A)	*A/*A	(*0/*0)	*0/*0

	NAT1		*NAT2*		*SLC15A2*		*SLC22A2*	
Sample ID	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**
HG00276	**4/*4*	**4/*10 [A]*	**5/*6*	**5/*6*	**1/*1*	**1/*1*	**1/(*2)*	**2/*3*
HG00436	**4/*4*	**4/*10 [A]*	**7/*7*	**7/*7*	**2/*2*	**2/*2*	**1/*1*	**1/*1*
HG00589	**4/*4*	**10/*26 [A]*	**4/(*13)*	**4/*13*	**2/*2*	**2/*2*	**1/*1*	**1/*1*
HG01190	**4/*4*	**4/*10 [A]*	**4/*4*	**4/*4*	**1/*1*	**1/*1*	**1/*3*	**1/*3*
NA06991	**4/*11*	**4/*11*	**5/*6*	**5/*6*	**1/*1*	**1/*1*	**1/*3*	**2/*3*
NA07000	*(*4/*11)*	**4/*11*	*(*5/*5)*	**5/*5*	*(*2/*2)*	**2/*2*	*(*1/*1)*	**1/*2*
NA07019	**4/*4*	**4/*10 [A]*	**6/*6*	**6/*6*	**1/*1*	**1/*1*	**1/*3*	**2/*3*
NA07029	**4/*4*	**4/*4*	**5/*6*	**5/*6*	**1/*2*	**1/*2*	**1/(*2)*	**1/*2*
NA07055	**4/*17*	**10/*17 [A]*	**5/*5*	**5/*5*	**1/*2*	**1/*2*	*(*3/*3)*	**3/*3*
NA07056	**4/*4*	**4/*4*	**6/*6*	**6/*6*	**1/*2*	**1/*2*	*(*2/*2)*	**2/*2*
NA07348	**4/*4*	**4/*4*	**5/*5*	**5/*5*	**1/*2*	**1/*2*	**1/(*2)*	**2/*3*
NA07357	**4/*4*	**4/*4*	**5/*6*	**5/*6*	**1/*1*	**1/*1*	**1/(*2)*	**1/*2*
NA10831	**4/*4*	**4/*4*	**4/*5*	**4/*5*	**1/*2*	**1/*2*	*(*2/*2)*	**2/*2*
NA10847	**4/*4*	**4/*4*	**5/*5*	**5/*5*	**1/*2*	**1/*2*	**1/(*2)*	**1/*2*
NA10851	**4/*4*	**10/*10 [A]*	**4/*5*	**4/*5*	**1/*2*	**1/*2*	**1/(*2)*	**1/*2*
NA10854	**4/*17*	**4/*17*	**4/*6*	**4/*6*	**2/*2*	**2/*2*	**1/(*2)*	**1/*2*
NA11832	**4/*4*	**4/*4*	**5/(*12)*	**5/*12*	**1/*2*	**1/*2*	**3/(*3)*	**3/*3*
NA11839	**4/*4*	**4/*4*	**5/*6*	**5/*6*	**1/*2*	**1/*2*	*(*2/*2)*	**2/*2*
NA11993	**4/*4*	**4/*4*	**5/*5*	**5/*5*	**2/*2*	**2/*2*	**1/(*2)*	**1/*2*
NA12003	**4/*4*	**4/*4*	**5/*5*	**5/*5*	**1/*1*	**1/*1*	**1/(*2)*	**1/*2*
NA12006	**4/*11*	**10/*11 [A]*	**6/*6*	**6/*6*	**1/*2*	**1/*2*	*(*2/*2)*	**2/*2*
NA12145	**4/*14*	**4/*14*	**4/*5*	**4/*5*	**1/*2*	**1/*2*	**1/(*2)*	**2/*3*
NA12156	**4/*4*	**4/*10 [A]*	**5/*6*	**5/*6*	**2/*2*	**2/*2*	**1/(*2)*	**2/*3*
NA12717	**4/*4*	**3/*10 [A]*	**4/*5*	**4/*5*	**2/*2*	**2/*2*	**1/*3*	**2/*3*
NA12813	**4/*4*	**4/*4*	**5/*6*	**5/*6*	**1/*2*	**1/*2*	*(*2/*2)*	**2/*2*
NA12873	**4/*4*	**4/*4*	**5/*5*	**5/*5*	**1/*2*	**1/*2*	**1/*1*	**1/*1*
NA18484	**4/*4*	**10/*10 [A]*	**4/*14*	**4/*14*	**1/*2*	**1/*2*	*(*2)/*3*	**2/*3*
NA18509	**4/*4*	**4/*4*	*(*12/*13)*	**12/*13*	**1/*2*	**1/*2*	**7/(K432Q)*	**7/*S1 [N]*
NA18518	**4/*4*	**4/*10 [A]*	*(*4)/*14*	**4/*14*	**1/*2*	**1/*2*	**1/*7*	**1/*7*
NA18519	**4/*4*	**4/*10 [A]*	**4/(*13)*	**4/*13*	**1/*2*	**1/*2*	**1/*3*	**1/*3*
NA18524	**4/*4*	**4/*10 [A]*	**4/*4*	**4/*4*	**2/*2*	**2/*2*	**1/(*3)*	**1/*3*
NA18526	**4/*4*	**4/*4*	**4/*4*	**4/*4*	**1/*2*	**1/*2*	**1/*1*	**1/*4 [A]*
NA18540	**4/*4*	**4/*4*	**5/*7*	**5/*7*	**2/*2*	**2/*2*	**1/(*3)*	**1/*3*
NA18544	**4/*4*	**4/*10 [A]*	**4/*6*	**4/*6*	**2/*2*	**2/*2*	**1/*1*	**1/*1*
NA18552	**4/*4*	**10/*10 [A]*	**4/*4*	**4/*4*	**2/*2*	**2/*2*	**1/*3*	**2/*3*
NA18564	**4/*4*	**4/*4*	**4/*7*	**4/*7*	**2/*2*	**2/*2*	**1/*3*	**2/*3*
NA18565	**4/*4*	**4/*10 [A]*	**4/*7*	**4/*7*	**1/*2*	**1/*2*	**3/*3*	**3/*3*
NA18617	*(*4/*4)*	**10/*10 [A]*	*(*4/*6)*	**4/*6*	*(*2/*2)*	**2/*2*	*(*1/*1)*	**1/*4 [A]*
NA18855	**4/*11*	**10/*11 [A]*	**6/(*13)*	**6/*13*	**1/*1*	**1/*1*	**1/(K432Q)*	**2/*S1 [N]*
NA18861	**4/*4*	**10/*10 [A]*	**5/*5*	**5/*5*	**1/*1*	**1/*1*	*(*2)/*3*	**2/*3*
NA18868	**4/*4*	**4/*10 [A]*	**5/(*12)*	**5, *12/*13 [P]*	**1/*2*	**1/*2*	**1/*7*	**1/*7*
NA18942	**4/*4*	**10/*10 [A]*	**4/*6*	**4/*6*	**2/*2*	**2/*2*	**1/*1*	**1/*1*
NA18952	**4/*4*	**10/*10 [A]*	**4/*4*	**4/*4*	**2/*2*	**2/*2*	**1/*1*	**1/*1*
NA18959	**4/*4*	**10/*10 [A]*	**6/*7*	**6/*7*	**1/*2*	**1/*2*	**1/*1*	**1/*4 [A]*
NA18966	**4/*4*	**4/*10 [A]*	**4/*4*	**4/*4*	**2/*2*	**2/*2*	**1/(*2)*	**1/*2*
NA18973	**4/*4*	**4/*4*	**4/*6*	**4/*6*	**1/*2*	**1/*2*	**1/*1*	**1/*1*
NA18980	**4/*4*	**4/*10 [A]*	**4/*7*	**4/*7*	**2/*2*	**2/*2*	**1/(*2)*	**1/*2*
NA18992	**4/*4*	**4/*10 [A]*	**4/*4*	**4/*4*	**1/*2*	**1/*2*	**3/(*3)*	**3/*3*
NA19003	**4/*4*	**4/*10 [A]*	**4/*4*	**4/*4*	**2/*2*	**2/*2*	**1/(*2)*	**1/*2*
NA19007	**4/*4*	**10/*10 [A]*	**4/*7*	**4/*7*	**1/*2*	**1/*2*	**1/(*2)*	**1/*2*
NA19095	**4/*4*	**4/*4*	**6/*6*	**6/*6*	**2/*2*	**2/*2*	**1/(*2)*	**2/*3*
NA19109	**4/*4*	**4/*10 [A]*	**4/*6*	**4/*6*	**1/*2*	**1/*2*	**1/(*2)*	**2/*3*
NA19122	**4/*4*	**4/*10 [A]*	*(*12/*12)*	**12/*12*	**1/*2*	**1/*2*	**3/(*3)*	**3/*3*
NA19143	**4/*4*	**4/*10 [A]*	*(*13)/*14*	**13/*14*	**1/*2*	**1/*2*	**3/(*3)*	**3/*3*
NA19147	**4/*4*	**4/*4*	**5/(*13)*	**5/*13*	**1/*2*	**1/*2*	**1/*3*	**1/*3*
NA19174	**4/*4*	**4/*10 [A]*	**5/*6*	**5/*6*	**1/*1*	**1/*1*	*(*2/*2)*	**2/*2*
NA19176	**4/*4*	**10/*10 [A]*	**5/*6*	**5/*6*	**1/*2*	**1/*2*	**1/(K432Q)*	**1/*S1 [N]*
NA19178	**4/*4*	**4/*4*	**5/*6*	**5/*6*	**2/*2*	**2/*2*	**1/(*2)*	**1/*2*
NA19207	**4/*4*	**4/*4*	*(*4)/*14*	**12/*14*	**1/*1*	**1/*1*	*(*2/*2)*	**2/*2*
NA19213	**4/*4*	**10/*10 [A]*	*(*4)/*14*	**12/*14*	**1/*2*	**1/*2*	**1/(*3)*	**1/*3*
NA19226	**4/*4*	**4/*4*	**4/*5*	**4/*5*	**1/*2*	**1/*2*	*(*3/*3)*	**3/*S2 [N]*
NA19239	**4/*4*	**4/*10 [A]*	*(*4)/*14*	**12/*14*	**1/*2*	**1/*2*	**1/(*2)*	**1/*2*
NA19789	**4/*4*	**10/*10 [A]*	**4/*4*	**4/*4*	**1/*1*	**1/*1*	**1/(*3)*	**1/*3*
NA19819	**4/*4*	**4/*10 [A]*	**5/*6*	**5/*6*	**1/*2*	**1/*2*	**1/(*3)*	**3/*S2 [N]*
NA19908	**4/*4*	**4/*4*	**5/*5*	**5/*5*	**1/*2*	**1/*2*	**3/*6*	**3/*6*
NA19917	*(*4/*4)*	**4/*10 [A]*	*(*4/*6)*	**4/*6*	*(*2/*2)*	**2/*2*	*(*1/*1)*	**1/*3*
NA19920	**4/*4*	**4/*10 [A]*	**6/*6*	**6/*6*	**2/*2*	**2/*2*	*(*2/*3)*	**2/*3*
NA20296	**4/*4*	**4/*10 [A]*	**6/*6*	**6/*6*	**2/*2*	**2/*2*	**3/*3*	**3/*3*
NA20509	**4/*4*	**4/*4*	**5/*6*	**5/*6*	**1/*2*	**1/*2*	**3/(*3)*	**3/*3*
NA21781	*(*4/*4)*	**4/*4*	-	**5/*6*	*(*1/*2)*	**1/*2*	*(*1/*1)*	**1/*2*

	SLCO1B1		*SLCO2B1*		*TPMT*		*UGT1A1*	
Sample ID	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**	**GeT-RM**	**Stargazer**
HG00276	*1A/*15	*1A/*15	*1/*1	*1/*1	*1/*1	*1/*16 [A]	*28/*60	*1/*28, *60 [P]
HG00436	*1B/*1B	*1B/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*27/(*28, *60)	*1/*27, *28, *60 [P]
HG00589	*1B/*1B	*1B/*S1 [N]	*1/*1	*1/*S1 [N]	*1/*3C	*1/*3C	*1/(*7)	*1/*7
HG01190	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1	(*37)/*60	*37/*60
NA06991	*15/*15	*15/*15	*1/*1	*1/*1	*1/*1	*1/*1	*60/*60	*60/*60
NA07000	*1A/*15	*1A/*15	(*1/*1)	*1/*1	*1/*1	*1/*1	(*1/*1)	*1/*1
NA07019	*1A/*14	*1A/*14	*1/*1	*1/*1	*1/*1	*1/*1	*1/*60	*1/*60
NA07029	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA07055	*1A/*14	*1A/*14	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA07056	*1A/*14	*1A/*14	*1/*1	*1/*1	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA07348	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1	*60/*60	*60/*60
NA07357	*1A/*15	*1A/*15	*1/*1	*1/*1	*1/*1	*1/*1	*60/(*28, *60)	*28, *60/*60
NA10831	*1A/*14	*1A/*14	*1/*1	*1/*1	*1/*1	*1/*1	(*28, *60)/(*28, *60)	*28, *60/*28, *60
NA10847	*5/*15	*5/*15	*1/*1	*1/*1	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA10851	*1A/*14	*1A/*14	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*60
NA10854	*1A/*1B	*1A/*1B	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA11832	*1B/*14	*14/*35 [A]	*1/*1	*1/*1	*1/*1	*1/*1	(*28, *60)/*60	*28, *60/*60
NA11839	*1A/*1B	*1A/*1B	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA11993	*1A/*17	*1A/*17	*1/*1	*1/*1	*1/*1	*1/*1	(*28, *60)/*60	*28, *60/*60
NA12003	*1A/*15	*1A/*15	*1/*1	*1/*1	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA12006	*1B/*21	*1B/*21	*1/*1	*1/*1	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA12145	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA12156	*1A/*21	*1A/*21	*1/*1	*1/*1	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA12717	*1A/*1A	*1A/*1A	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA12813	*1A/*21	*1A/*21	*1/(*S464F)	*S1/*S464F [N]	*1/*1	*1/*1	(*28, *60)/(*28, *60)	*28, *60/*28, *60
NA12873	*1A/*1A	*1A/*1A	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA18484	*1A/*1B	*1B/*S2 [N]	*1/*1	*1/*1	*1/*1	*1/*1	(*28, *60)/*60	*28, *60/*60
NA18509	*1A/*1B	*1A/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA18518	*1A/*1B	*1A/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*60/*60	*60/*60
NA18519	*1B/*1B	*1B/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*60/(*28, *60)	*28, *60/*60
NA18524	*1B/*21	*1B/*21	*1/*1	*1/*S1 [N]	*1/*1	*1/*1	*1/*1	*1/*1
NA18526	*1A/*15	*1A/*15	*1/*1	*1/*1	*1/*1	*1/*1	*1/*60	*1/*60
NA18540	*1B/*17	*17/*1B	(*S464F/*S464F)	*S464F/*S464F	*1/*1	*1/*1	*1/*28	*1/*28
NA18544	*1B/*17	*17/*1B	*1/*1	*1/*1	*1/*1	*1/*1	*28/(*60)	*1/*28, *60 [P]
NA18552	*1B/*15	*1B/*15	*1/*1	*1/*1	*1/*1	*1/*1	*1/*60	*1/*60
NA18564	*1A/*1B	*1A/*1B	*1/*1	*1/*1	*1/*1	*1/*1	*1/*60	*1/*60
NA18565	*1A/*1A	*1A/*1A	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*1/*60	*1/*60
NA18617	(*1B/*1B)	*1B/*1B	(*1/*S464F)	*1/*S464F	*1/*1	*1/*1	(*1/*6)	*1/*6
NA18855	*1B/*1B	*1B/*1B	*1/*1	*1/*1	*1/*3C	*1/*3C	(*28, *60)/(*28, *60)	*28, *60/*28, *60
NA18861	*1A/*14	*1A/*14	*1/*1	*1/*1	*1/*1	*1/*1	*60/*60	*60/*60
NA18868	*1B/*1B	*1B/*24 [A]	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	(*28, *60)/(*28, *60)	*28, *60/*28, *60
NA18942	*1B/*1B	*1B/*1B	(*S464F/*S464F)	*S464F/*S464F	*1/*1	*1/*1	*28/*60	*1/*28, *60 [P]
NA18952	*1A/*1A	*1A/*1A	*1/*1	*1/*S1 [N]	*1/*1	*1/*1	*1/*60	*1/*60
NA18959	*1B/*1B	*1B/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*1/*60	*1/*60
NA18966	*1A/*1A	*1A/*1A	(*S464F/*S464F)	*S464F/*S464F	*1/*3C	*1/*3C	*1/*6	*1/*6
NA18973	*1A/*1B	*1A/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*1/*6	*1/*6
NA18980	*1A/*1B	*1A/*1B	*1/(*S464F)	*S1/*S464F [N]	*1/*1	*1/*1	*6/*6	*6/*6
NA18992	*1A/*17	*1A/*17	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA19003	*1B/*1B	*1B/*1B	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1
NA19007	*1B/*1B	*1B/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*1/*6	*1/*6
NA19095	*1B/*1B	*1B/*35 [A]	*1/*1	*1/*1	*1/*1	*1/*1	(*28, *60)/(*36, *60)	*28, *60/*36, *60
NA19109	*1B/*15	*1B/*15	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*1/*60	*1/*60
NA19122	*1B/*1B	*1B/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*28/(*60)	*1/*28, *60 [P]
NA19143	*1B/*1B	*1B/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	(*28, *60)/*60	*28, *60/*60
NA19147	*1B/*1B	*30/*35 [A]	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	(*28, *60)/(*28, *60)	*28, *60/*28, *60
NA19174	*1A/*1B	*1A/*27 [A]	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	*60/(*28, *60)	*28, *60/*60
NA19176	*1A/*1B	*1A/*1B	*1/*1	*1/*1	*1/*8	*1/*8	*28/(*60)	*28, *60/*60
NA19178	*1A/*1B	*1A/*1B	*1/*1	*1/*1	*1/*1	*1/*1	(*28, *60)/(*36, *60)	*28, *60/*36, *60
NA19207	*1A/*1B	*1B/*S2 [N]	*1/*1	*1/*1	*1/*1	*1/*1	(*36, *60)/(*28, *60)	*28, *60/*36, *60
NA19213	*1B/*14	*1B/*14	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	(*36, *60)/*60	*36, *60/*60
NA19226	(*1A)/*1B	*1B/*31 [A]	*1/*1	*1/*1	*1/*1	*1/*1	*1/*60	*1/*60
NA19239	*1B/*1B	*1B/*1B	*1/(*S464F)	*1/*S464F	*1/*1	*1/*1	(*28, *60)/(*37, *60)	*28, *60/*37, *60
NA19789	*1A/*1B	*1A/*1B	*1/*1	*1/*1	*1/*1	*1/*1	*60/*60	*60/*60
NA19819	*1B/*1B	*1B/*1B	*1/*1	*1/*1	*1/*1	*1/*1	*28/(*60)	*1/*28, *60 [P]
NA19908	*1B/*1B	*1B/*1B	(*S464F/*S464F)	*S464F/*S464F	*1/*1	*1/*1	(*28, *60)/*60	*28, *60/*60
NA19917	(*1A/*1B)	*1A/*1B	(*1/*1)	*1/*1	*1/*1	*1/*1	*28/(*60)	*1/*28, *60 [P]
NA19920	*1A/*1B	*1A/*1B	*1/(*S464F)	*1/*S464F	*1/*3C	*1/*3C	(*37)/*60	*1/*37, *60 [P]
NA20296	*1B/*1B	*1B/*27 [A]	*1/*1	*1/*1	*1/*3C	*1/*3C	*1/*60	*1/*60
NA20509	*1A/*15	*1A/*15	*1/*1	*1/*1	*1/*1	*1/*1	(*28)/(*28, *60)	*1/*28, *60
NA21781	*5/(*15)	*5/*15	(*1/*1)	*1/*1	*1/*1	*1/*1	*28/(*60)	*1/*28, *60 [P]

Sample ID	*UGT2B7* GeT-RM	Stargazer	*UGT2B15* GeT-RM	Stargazer	*UGT2B17* GeT-RM	Stargazer	*VKORC1* GeT-RM	Stargazer
HG00276	*1/*1	*1/*1	*1/*5	*1/*5	*2/*2	*2/*2	*2/*3	*2/*3
HG00436	*1/*2	*1/*2	*2/*5	*2/*5	(*1/*2)	*1/*2	*2/*3	*2/*3
HG00589	*1/*2	*1/*2	*1/*2	*1/*2	*2/*2	*2/*2	*2/*2	*2/*2
HG01190	*1/*2	*1/*2	(*4)/*5	*4/*5	(*1/*1)	*1/*1	(*3/*3)	*3/*3
NA06991	*1/*2	*1/*2	(*4)/*5	*4/*5	(*1/*1)	*1/*1	*2/*2	*2/*2
NA07000	(*1/*2)	*1/*2	(*1/*2)	*2/*4	(*1/*1)	*1/*1	*2/*3	*2/*3
NA07019	*2/*2	*2/*2	(*4)/*5	*4/*5	(*1/*2)	*1/*2	(*3/*3)	*3/*3
NA07029	*2/*2	*2/*2	*1/*5	*2/*4 [P]	(*1/*2)	*1/*2	*2/*3	*2/*3
NA07055	*1/*1	*1/*1	*1/*2	*1/*2	(*1/*2)	*1/*2	*2/*2	*2/*2
NA07056	*2/*2	*2/*2	(*4)/*5	*4/*5	(*1/*2)	*1/*2	*2/*3	*2/*3
NA07348	*1/*2	*1/*2	*2/*5	*2/*5	(*1/*2)	*1/*2	(*3/*3)	*3/*3
NA07357	*1/*2	*1/*2	*1/*5	*2/*4 [P]	(*1/*1)	*1/*1	(*3/*4)	*3/*4
NA10831	*1/*2	*1/*2	*5/*5	*5/*5	*2/*2	*2/*2	*2/*2	*2/*2
NA10847	*1/*2	*1/*2	*2/*2	*2/*2	(*1/*1)	*1/*1	(*3/*4)	*3/*4
NA10851	*2/*2	*2/*2	*5/*5	*5/*5	*2/*2	*2/*2	*2/*4	*2/*4
NA10854	*1/*1	*1/*1	*2/*5	*2/*5	(*1/*1)	*1/*1	*2/*3	*2/*3
NA11832	*1/*2	*1/*2	*1/*5	*2/*4 [P]	(*1/*2)	*1/*2	(*3/*4)	*3/*4
NA11839	*1/*2	*1/*2	*1/*5	*2/*4 [P]	(*1/*2)	*1/*2	*2/*3	*2/*3
NA11993	*1/*2	*1/*2	(*4/*4)	*4/*S1 [N]	-	*1/*1	*2/*3	*2/*3
NA12003	*1/*2	*1/*2	*2/*5	*2/*5	(*1/*1)	*1/*1	*2/*2	*2/*2
NA12006	*1/*1	*1/*1	*5/*5	*5/*5	(*1/*1)	*1/*1	*2/*4	*2/*4
NA12145	*1/*2	*1/*2	*1/(*4)	*1/*4	(*1/*2)	*1/*2	*2/*3	*2/*3
NA12156	*1/*2	*1/*2	*5/*5	*5/*5	-	*1/*2	*2/*3	*2/*3
NA12717	*1/*2	*1/*2	(*4/*4)	*4/*4	(*1/*2)	*1/*2	(*3/*4)	*3/*4
NA12813	*1/*2	*1/*2	*2/*5	*2/*5	*2/*2	*2/*2	(*3/*4)	*3/*4
NA12873	*1/*2	*1/*2	*1/*5	*2/*4 [P]	(*1/*2)	*1/*2	(*3/*4)	*3/*4
NA18484	*1/*1	*1/*1	*1/*2	*1/*2	-	*1/*1	*1/*3	*1/*3
NA18509	*1/*2	*1/*2	*1/*2	*1/*2	(*1/*1)	*1/*1	*1/*1	*1/*1
NA18518	*1/*2	*1/*2	*1/*2	*1/*2	(*1/*1)	*1/*1	(*3/*3)	*3/*3
NA18519	*1/*2	*1/*2	*1/*1	*1/*1	(*1/*1)	*1/*1	*1/(*3)	*1/*3
NA18524	*2/*2	*2/*2	*1/*5	*2/*4 [P]	(*1/*2)	*1/*2	*2/*2	*2/*2
NA18526	*1/*2	*1/*2	*1/*1	*1/*1	*2/*2	*2/*2	*2/*2	*2/*2
NA18540	*1/*2	*1/*1/*2 [T]	*1/*5	*2/*2/*4 [P] [T]	*2/*2	*2/*2/*2 [T]	*2/*2	*2/*2
NA18544	*1/*1	*1/*1	*1/*2	*1/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA18552	*1/*1	*1/*1	*1/*2	*1/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA18564	*1/*2	*1/*2	*2/*2	*2/*2	(*1/*2)	*1/*2	*2/*2	*2/*2
NA18565	*2/*2	*2/*2	*2/*2	*2/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA18617	(*1/*2)	*1/*2	(*1/*1)	*1/*1	(*2/*2)	*2/*2	*2/*2	*2/*2
NA18855	*1/*2	*1/*2	*1/*2	*1/*2	(*1/*2)	*1/*2	*1/*1	*1/*1
NA18861	*1/*2	*1/*2	*1/*2	*1/*2	(*1/*1)	*1/*1	*1/(*3)	*1/*3
NA18868	*1/*2	*1/*2	*1/*1	*1/*1	(*1/*2)	*1/*2	*1/(*3)	*1/*3
NA18942	*1/*1	*1/*1	*1/(*4)	*1/*4	*2/*2	*2/*2	*2/*2	*2/*2
NA18952	*1/(*3)	*1/*3	*1/*5	*1/*5	*2/*2	*2/*2	*2/*3	*2/*3
NA18959	*1/*2	*1/*2	*1/*2	*1/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA18966	*1/*2	*1/*2	*1/*2	*1/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA18973	*1/(*3)	*1/*3	*1/*2	*1/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA18980	*2/(*3)	*2/*3	*1/*2	*1/*2	*2/*2	*2/*2	*2/*3	*2/*3
NA18992	*2/(*3)	*2/*3	*2/*2	*2/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA19003	*1/*2	*1/*2	*2/*2	*2/*2	(*1/*2)	*1/*2	*2/*2	*2/*2
NA19007	*1/(*3)	*1/*3	*1/*2	*1/*2	*2/*2	*2/*2	*2/*2	*2/*2
NA19095	*1/*2	*1/*2	*1/*1	*1/*1	(*1/*2)	*1/*2	*1/*3	*1/*3
NA19109	*1/*1	*1/*1	*1/*2	*1/*2	(*1/*2)	*1/*2	*1/*3	*1/*3
NA19122	*1/*1	*1/*1	*2/*2	*2/*2	(*1/*1)	*1/*1	*1/*1	*1/*1
NA19143	*1/*2	*1/*2	*1/*1	*1/*1	(*1/*2)	*1/*2	(*3/*3)	*3/*3
NA19147	*1/*1	*1/*1	*2/*2	*2/*2	(*1/*2)	*1/*2	*2/(*3)	*2/*3
NA19174	*1/*1	*1/*1	*1/*2	*1/*2	(*1/*1)	*1/*1	*1/(*3)	*1/*3
NA19176	*1/*1	*1/*1	*1/*5	*2/*4 [P]	(*1/*1)	*1/*1	(*3/*3)	*3/*3
NA19178	*1/*1	*1/*1	*2/*2	*2/*2	(*1/*1)	*1/*1	(*3/*3)	*3/*3
NA19207	*1/*2	*1/*2	*1/*2	*1/*2	(*1/*1)	*1/*1	*1/*1	*1/*1
NA19213	*1/*2	*1/*2	*1/*2	*1/*2	(*1/*2)	*1/*2	*1/*3	*1/*3
NA19226	*1/*2	*1/*2	*1/*1	*1/*1	(*1/*1)	*1/*1	(*3/*3)	*3/*3
NA19239	*1/*1	*1/*1	*1/*1	*1/*1	(*1/*2)	*1/*2	(*3/*3)	*3/*3
NA19789	*1/*1	*1/*1	*1/*1	*1/*1	(*1/*1)	*1/*1	(*3/*3)	*3/*3
NA19819	*1/*1	*1/*1	*1/(*4)	*1/*4	(*1/*1)	*1/*1	*2/*3	*2/*3
NA19908	*1/*1	*1/*1	*1/(*4)	*1/*4	*2/*2	*2/*2	(*3/*3)	*3/*3
NA19917	(*1/*1)	*1/*1	(*1/*1)	*1/*1	(*1/*2)	*1/*2	*1/(*3)	*1/*3
NA19920	*1/*2	*1/*2	*1/*2	*1/*2	(*1/*1)	*1/*1	*1/(*4)	*1/*4
NA20296	*1/*1	*1/*1	*1/*5	*2/*4 [P]	*2/*2	*2/*2	*1/*2	*1/*2
NA20509	*1/*2	*1/*2	*2/*5	*2/*5	(*1/*2)	*1/*2	*2/*3	*2/*3
NA21781	(*1/*2)	*1/*2	(*1/*1)	*1/*4	(*1/*1)	*1/*1	*2/*4	*2/*4

Consensus GeT-RM genotypes are shaded in green (N=1559). "()" indicates non-consensus GeT-RM genotypes (N=401). "[A]" indicates Stargazer's identification of 38 additional star alleles not previously reported by GeT-RM (N=127). "[P]" indicates revision is made by Stargazer's statistical phasing of star alleles using a haplotype reference panel (N=64). "[N]" indicates Stargazer's calling of nine new star alleles defined in this study (N=20). "[T]" indicates three gene copies present in the sample NA18540 caused by trisomy in chromosomes 4 and 7 (N=5).

Supplementary Table 4.2 Star alleles previously reported by GeT-RM and not identified by Stargazer's analysis of whole genome sequencing data

Star Allele	Assays[a]	Defining Variant[b]	Sample ID	GeT-RM[c]	Stargazer	WGS Evidence[d]
*CYP2A6*8*	[1, 2]	rs28399468 (41349732C>A)	NA12003	**1/(*8)*	**1/*1*	AD: C=49, A=0
			NA18855	**1/(*8)*	**1/*1*	AD: C=45, A=0
			NA19176	**1/(*8)*	**1/*1*	AD: C=27, A=0
*CYP2B6*27*	[1]	rs36079186 (41512918T>C)	HG01190	**1 (*5)/*1 (*27)*	**1/*5*	AD: T=40, C=0
			NA07029	**6/(*27)*	**1/*6*	AD: T=31, C=0
			NA07055	**6/(*27)*	**1/*6*	AD: T=35, C=0
			NA10851	**1/*1 (*27)*	**1/*1*	AD: T=46, C=0
			NA18524	**1/*1 (*27)*	**1/*1*	AD: T=44, C=0
			NA18959	**1/*1 (*27)*	**1/*1*	AD: T=29, C=0
			NA19789	**1/*1 (*27)*	**1/*1*	AD: T=36, C=0
*CYP2C9*18*	[2, 3]	rs72558193 (96745830A>C)	NA11832	**1/*3 (*18)*	**1/*3*	AD: A=45, C=0
			NA11839	**2/*3 (*18)*	**2/*3*	AD: A=59, C=0
			NA12813	**1/*3 (*18)*	**1/*3*	AD: A=46, C=0
			NA18959	**1/*3 (*18)*	**1/*3*	AD: A=38, C=0
			NA19917	**1/*1 (*18)*	**1/*1*	AD: A=49, C=2
*GSTT1*AxN*	[2]	CN=3	NA18484	*(*A/*AxN)*	**A/*0*	CN=1
			NA18519	*(*A/*AxN)*	**A/*0*	CN=1
			NA18861	*(*A/*AxN)*	**A/*0*	CN=1
			NA19095	*(*A/*AxN)*	**A/*0*	CN=1
			NA19109	*(*A/*AxN)*	**A/*A*	CN=2
			NA19174	*(*A/*AxN)*	**A/*A*	CN=2
			NA19213	*(*A/*AxN)*	**A/*A*	CN=2
			NA19920	*(*A/*AxN)*	**A/*0*	CN=1

WGS, whole genome sequencing; AD, allelic depth; CN, copy number.

[a][1] Affymetrix DMET Plus Array (Affymetrix, Santa Clara, CA); [2] Agena Bioscience iPLEX ADME PGx Pro Panel (Agena Bioscience, San Diego, CA); [3] Agena Bioscience iPLEX ADME CYP2C9 Panel (Agena Bioscience, San Diego, CA).

[b]Genomic coordinates and nucleotide changes are according to Human Genome version 19.

[c]"()" indicates non-consensus GeT-RM genotypes.

[d]AD was assessed through visual inspection of WGS reads using Integrative Genomics Viewer.

Supplementary Table 4.3 Structural variant reports in the Database of Genomic Variants supporting genotype calls from Stargazer over those from GeT-RM

Star Allele	Sample ID	GeT-RM[a]	Stargazer	Report ID	Type	Source
*CYP2A6*1x2*	NA18861	**1/*1*	**25/*1x2*	gssvG19016	CNV gain	DGV gold standard (multiple studies)
*CYP2A6*4*	NA18565	**9/*9*	**4/*15*	esv3644361	CNV loss	1KGP (Mills et al., 2011)
	NA18942	**1/*1*	**4/*7*	esv3644361	CNV loss	1KGP (Mills et al., 2011)
*CYP2E1*7x2*	NA19908	**7/*7*	**7x2/*7x2*	dgv172e214	CNV gain	1KGP (Mills et al., 2011)
*CYP2E1*S1*	NA19143	*(*7/*7)*	**7/*S1*	dgv8n64	CNV gain	Wang et al, 2007
*GSTM1*0*	NA10847	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA12006	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA12717	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA18544	**B/*B*	**B/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA18552	**B/*B*	**B/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA18942	**B/*B*	**B/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA18980	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA18992	**B/*B*	**B/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA19143	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA19213	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA19789	**B/*B*	**B/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA19920	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
	NA20296	**A/*A*	**A/*0*	esv3587155	CNV loss	1KGP (Mills et al., 2011)
*GSTT1*0*	HG00276	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA12003	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA12813	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA18519	*(*A/*AxN)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA18552	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA18861	*(*A/*AxN)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA18868	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA18959	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA18973	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA18980	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA19095	*(*A/*AxN)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA19908	*(*A/*A)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
	NA19920	*(*A/*AxN)*	**A/*0*	esv3647425	CNV loss	1KGP (Mills et al., 2011)
*SLC22A2*S2*	NA19226	*(*3/*3)*	**3/*S2*	gssvL111331	CNV loss	DGV gold standard (multiple studies)
	NA19819	**1/(*3)*	**3/*S2*	gssvL111331	CNV loss	DGV gold standard (multiple studies)

DGV, Database of Genomic Variants; CNV, copy number variation; 1KGP, 1000 Genomes Project.
[a]"()" indicates non-consensus GeT-RM genotypes.

Supplementary Table 4.4 Demographic and sequencing information for 70 reference samples

Sample ID	Population	Sequencing ID
HG00276	European (Finnish)	20b87673c1224e9db8bdbbe82899309c
HG00436	East Asian (Southern Han Chinese)	54db734bc1ec46b29fc6c5c6df35ca65
HG00589	East Asian (Southern Han Chinese)	316ab006177d41b484982d7fa4d851ad
HG01190	American (Puerto Rican)	9e01734a352a41f89266c2ae8c9c13de
NA06991	European (Utah/Mormon)	82b808e2886a42f986da3ba811dbeaa4
NA07000	European (Utah/Mormon)	5d81e03c86324f209c69093ddf77bb62
NA07019	European (Utah/Mormon)	dd864a425f814bca87a971eaf94cfa17
NA07029	European (Utah/Mormon)	28997710cbae49f5996f6075fcbf74bd
NA07055	European (Utah/Mormon)	a9963d642c584dfab81f5ae694208390
NA07056	European (Utah/Mormon)	3a7ec2f78f3c40df98248d3ba1354a20
NA07348	European (Utah/Mormon)	e92d97eb8a3f4c4cb7db5faf9882b167
NA07357	European (Utah/Mormon)	dd64f80e456a46e49555c0c7c30372b0
NA10831	European (Utah/Mormon)	021ab129bb594be5804b02e08e14d93d
NA10847	European (Utah/Mormon)	e030f757080d4e5e841d8e7feef7a665
NA10851	European (Utah/Mormon)	cb484635e004493fb395ac764578797b
NA10854	European (Utah/Mormon)	7866e1bc7fdc4304863db4a25d1a42e4
NA11832	European (Utah/Mormon)	22d9ea3d16804243afbfea7e776c5237
NA11839	European (Utah/Mormon)	c392e9300f6b490aa1c43f7ca7d7afd4
NA11993	European (Utah/Mormon)	20be9cf6bed64502b85c999ec59c784b
NA12003	European (Utah/Mormon)	141390df39414eaaaa9725a8349d9c45
NA12006	European (Utah/Mormon)	5f006ec8ba3c41a18d9ff92c0a62955a
NA12145	European (Utah/Mormon)	45e24be4dc7d4cb3a2742ce7c05730e8
NA12156	European (Utah/Mormon)	69d7b11affce444694e9955b90848028
NA12717	European (Utah/Mormon)	3959a73552a04f969d37a04dc869c7a9
NA12813	European (Utah/Mormon)	2c9f234af49b4f6a970d8ddef07358e5
NA12873	European (Utah/Mormon)	def4a73760dd42d38b17c33c3deb654b
NA18484	African (Yoruba)	3ddccd4bcda14140b3e89559d2cf3186
NA18509	African (Yoruba)	d84c8e4613064299b2a16cfa39d819b5
NA18518	African (Yoruba)	e1ddd983797640bf81b66f9a77b37439
NA18519	African (Yoruba)	20aaa40fcdee4290bfada8dbba5e5232
NA18524	East Asian (Han Chinese)	ba64fb6dbb0a4b36a5cbe53bd8706ca7
NA18526	East Asian (Han Chinese)	7d077b89f2514fb0b8e002f8d9a10189
NA18540	East Asian (Han Chinese)	f00e1071f840476c9872de73f0ea8a02
NA18544	East Asian (Han Chinese)	ecf7d003ffef4d429351037cb97bac09
NA18552	East Asian (Han Chinese)	5bb329dcb3654e8890985547e82135f6
NA18564	East Asian (Han Chinese)	4323a15d7b5d4bf2b204e0c0088ba923
NA18565	East Asian (Han Chinese)	b859eb68840b42d9a376ce22cff1cd08
NA18617	East Asian (Han Chinese)	0b7cc95044c54d86a81151d856d0c5b2
NA18855	African (Yoruba)	03bc76a2c27140bc8143c56767ca6877
NA18861	African (Yoruba)	543558ae08cd44b3850fc7b835484037
NA18868	African (Yoruba)	2b0b4c79e8104ed98ef5cc82d9ca8bd2
NA18942	East Asian (Japanese)	bda28ac619e64916bb0a28d1f2698e2e
NA18952	East Asian (Japanese)	92ac6fc0f69345aabb9e7bd47452ed70
NA18959	East Asian (Japanese)	7fdcb4bdabe743feb4abc155bc580b82
NA18966	East Asian (Japanese)	f52c133442ca4f93a47b689fc385b1f5
NA18973	East Asian (Japanese)	11a1c7b37a63449f9bc65799567b5710
NA18980	East Asian (Japanese)	7a29e8a53b6844ddbb42aa165b34fba3
NA18992	East Asian (Japanese)	b7988ee8179d4678921602d34400a63a
NA19003	East Asian (Japanese)	e968235bb8e14f02af53e2f17cee324f
NA19007	East Asian (Japanese)	6320750d96f747b695f78964df1fad17
NA19095	African (Yoruba)	75331ab394f24d56ac73cee5d41fa15b
NA19109	African (Yoruba)	8f90214c429f4afc9e0da555cb77f89a
NA19122	African (Yoruba)	9e3dcdd083ad40dd82fab180a916fb77
NA19143	African (Yoruba)	dfe4ba40717d4891a12ec2c856de671f
NA19147	African (Yoruba)	6b9f68c07c3e4e3e81669fe390ba1027
NA19174	African (Yoruba)	b462b7941e0642309d5d44e7aae6d42f
NA19176	African (Yoruba)	d7b5cf7015d44c23a949dc117c149c80
NA19178	African (Yoruba)	3fcca708192c4ffe8e57318c7d64e480
NA19207	African (Yoruba)	7a94709935ea412e8b7349828beaf338
NA19213	African (Yoruba)	a284eb8f85e54b0ba3c45e30e91a33a1
NA19226	African (Yoruba)	db923feb258241b8b0a3e3aaada9359b
NA19239	African (Yoruba)	69335e244b5f46e5b56e5360f8f826b7
NA19789	American (Mexican-American)	932a3f4c888844c39399f1cfdf2bc593
NA19819	African (African-American)	a04f109738f34a358850f5f69d7d8814
NA19908	African (African-American)	0d73bafef55a4a718489f3fdca91fd55
NA19917	African (African-American)	2176d8cef950450bbb29052359a2d2d1
NA19920	African (African-American)	228fa4f74dd4431f8eb3526f37e355ed
NA20296	African (African-American)	34b739132d0e403d9aa323b1c3edf12a
NA20509	European (Toscani in Italy)	b51b8299a62f4a3eab95d0467e950b7e
NA21781	unknown (Caucasian)	e246a44270e34d2aa228844732995abe