Davis, Carey P., Increased Resolution Screening of the Pharmacogenetic Gene *CYP2D6* with Microarray Technology. Doctor of Philosophy (Biomedical Sciences), July 2019, 178 pp., 76 tables, 22 figures, 144 references.

Abstract

Autopsy is a primary methodology used for assessing cause and/or manner of death in medicolegal investigations. Some autopsies, however, do not resolve the cause of death unambiguously or there is no evident pathology to determine the cause of death. In addition, in some cases toxicology screens are negative or difficult to interpret because it is challenging to determine if a high concentration of a drug in the body derived from one large dose or has built up over time. Determining the genetic constitution of victims at specific target genes may clarify the cause of some of these unexplained deaths or at least indicate susceptibility to triggering effects.

Cytochrome P450 (CYP450) is a super family of enzymes that detoxify foreign chemicals and are involved in the metabolism of drugs. One gene in this family that encodes CYP450 enzymes, *CYP2D6*, accounts for the metabolism of 25% of all drugs currently on the market. By examining the variability in the *CYP2D6* gene, a SNP panel was developed and used to aid in personalized medicine with a long-term outcome of reducing risk in patients who partake in drug therapy. However, paralogs of the *CYP2D6* gene can interfere with obtaining accurate typing results.

The hypothesis of this dissertation is that it is possible to develop a targeted panel for clinically relevant variants in the *CYP2D6* gene using array-based technology that can provide accurate and reliable genotyping results. The goal was first to demonstrate that high throughput sequencing, also known as next generation or massively parallel sequencing, could reliably

sequence a complex target using a model system, i.e. the hypervariable regions of the human mitochondrial genome. Then, using this advanced sequencing capability define the baseline genetic variation of the *CYP2D6* gene in a selected population and identify those genetic markers associated with metabolism capacity that would be verified against a database of actionable variants that cause reaction to drug exposure. The entire *CYP2D6* gene was sequenced to identify SNPs at a population level. These SNPs were compared against a known database of clinically relevant samples with known metabolic responses of the same ethnic background to verify actionable variants. Once these variants were identified, a PCR assay workflow leveraging microarray technology was developed to quickly and efficiently screen individuals of interest by overcoming paralog interference. This assay can be used prior to administering drugs or post mortem to gain information about potential adverse drug reactions.

KEYWORDS Pharmacogenomics, Cytochrome p450 family 2 subfamily D polypeptide 6 (*CYP2D6*), Massively Parallel Sequencing, Single nucleotide polymorphism (SNP), Microarray, pseudogenes

INCREASED RESOLUTION SCREENING OF THE PHARMACOGENETIC GENE *CYP2D6* WITH MICROARRAY TECHNOLOGY

Carey P. Davis, M.S.

APPROVED:

Major Professor

Committee Member

Committee Member

Committee Member

University Member

Chair, Department of Microbiology, Immunology, and Genetics

Dean, Graduate School of Biomedical Sciences

INCREASED RESOLUTION SCREENING OF THE PHARMACOGENETIC GENE *CYP2D6* WITH MICROARRAY TECHNOLOGY

A DISSERTATION

Presented to the Graduate Council of the

University of North Texas Health Science Center at Forth Worth in

Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

By:

Carey P Davis, M.S.

July 2019

ACKNOWLEDGEMENTS

When I first found out I was accepted into the PhD program at UNTHSC, I was both elated and terrified. Knowing that I would get to work with true forensic greats meant that I would truly be stepping into the big leagues. I knew that this would change the direction of my life, but had no idea what was truly in store for me. When I arrived in Texas, I was quickly humbled by the sheer knowledge around me and perhaps more shocking the approachability of everyone.

First, I must thank my major professor, Dr. Bruce Budowle. He challenged me as soon as I walked in the door and continues to challenge me to be the best scientist I can be. He taught me that we succeed better as a team, and through collaboration we can change the world. He extended his hard earned network to me that eventually led to opportunities I never would have imagined.

Drs. Arthur Eisenberg and Ranajit Chakraborty helped me gain the knowledge and confidence to pursue my dreams, and I will be forever in the debt of these great men.

I would like to thank my committee for their tenacity and support while I take an unorthodox route through this whole experience. Dr. Robert Luedkte kept me on my toes, and reminded me the clinical significance of what scientists can bring to the table. Dr. Lisa Hodge has supported me throughout this entire endeavor. I want to thank Drs. Nicole Phillips and August Woerner for agreeing to being on my committee and giving the support and feedback needed to get to this point. I greatly appreciate their willingness to help when it would have been easier to let me fade away.

I want to thank my colleagues both at UNTHSC and Illumina for believing in me even when I stopped believing in myself. The love and support I received from you helped more than you will ever know.

Lastly, I want to thank my family and friends for supporting me through this long journey. It's been a true adventure, and I'm so lucky to have such amazing people around me.

~ Carey

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 13. No shift in intensity of SNPs nor a theta shift to the left or right of the intended cluster location for SNPs not targeted by specific PCR as shown by SNPs outside of the amplified region

LIST OF ABBREVIATIONS

ABI	Applied Biosystems Incorporated
ADHD	Attention deficit hyperactivity disorder
ADME	Absorption, distribution, metabolism, excretion
ADR	Adverse drug reaction
ASPE	Allele specific primer extension
BAM	Binary alignment map
BCL	Base call and quality score file
BWA	Burrough-Wheeler Alignment
CCD	Charge-coupled device
CNV	Copy number variation
CPIC	Clinical Pharmacogenetics Implementation Consortium
CPF	Clusters passing filter
Ct	Threshold cycle
CYP2D6	Cytochrome p450, family 2, subfamily D, polypeptide 6
CYP2D7P	Cytochrome p450, family 2, subfamily D, polypeptide 7 pseudogene
CYP2D8P	Cytochrome p450, family 2, subfamily D, polypeptide 8 pseudogene
CYP450	Cytochrome p450, mono-oxygenase
dATP	Deoxyadenosine triphosphate
ddATP	Dideoxyadenosine triphosphate
ddCTP	Dideoxoycytosine triphosphate
ddGTP	Dideoxyguanosine triphosphate
ddTTP	Dideoxythymine triphosphate
ddNTP	Dideoxynucleotide triphosphate
DNA	Deoxyribonucleic acid

dNTP	Deoxynucleotide triphosphate
DME	Drug metabolizing enzyme
DPWG	Dutch Pharmacogenetics Working Group
EM	Extensive metabolizer
Fl-NTP	Fluorescently labeled nucleoside triphosphate
G6PD	Glucose-6-phosphate dehydrogenase
GA	Genome analyzer
GSA	Global screening array
GWAS	Genome wide association study
HWE	Hardy-Weinberg equilibrium
IM	Intermediate metabolizer
INS	Insertion
DEL	Deletion
INDEL	Insertion/deletion
LD	Linkage disequilibrium
MAF	Minor allele frequency
MPS	Massively parallel sequencing
MSM	Multi-sample amplification master mix
mtDNA	mitochondrial dexoyribonucleic acid
NAT	N-acetyltransferase
NGS	Next generation sequencing
PCR	Polymerase chain reaction
PM	Poor metabolizer
PTC	Phenylthiocarbamide
qPCR	Real-time polymerase chain reaction
RFLP	Restriction fragment length polymorphism
SAM	Sequence alignment/map format

SBS	Sequencing by synthesis
SIFT	Sorting intolerant from tolerant
SNP	Single nucleotide polymorphism
SSCP	Single-strand conformation polymorphism
STS	Sanger type sequencing
SWGDAM	Scientific Working Group on DNA Analysis Methods
TCA	tricyclic antidepressant
TER	Translation termination
TMPT	Thiopurine methyltransferase
UM	Ultra-rapid metabolizer
UNTCHI	University of North Texas Center for Human Identification
VCF	Variant call format
WGA	Whole genome amplification
WGG	Whole genome genotyping
WHO	World Health Organization

PART 1

INTRODUCTION

Over 100,000 deaths occur in the United States each year due to adverse drug reactions (ADRs). ADR is defined by the World Health Organization as "a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function." (WHO, 1972). Autopsy is a primary methodology used for assessing cause and/or manner of death in medicolegal investigations (Ludwig, 2002). The cause and manner of death needs to be determined for each case. Some autopsies, however, do not resolve the cause of death unambiguously or there is no evident pathology to determine the cause of death. In addition, in some cases toxicology screens are negative or difficult to interpret because it is challenging to determine if a high concentration of a drug in the body derived from one large dose or has built up over time (Pirmohamed *et al.*, 2003). Some drug-related deaths may be due to accidental lethal dosages as opposed to suicide. In addition, individual responses to drug exposure or other stresses such as exercise may vary based on an individual's genetic makeup. Determining the genetic constitution of victims at specific target genes may clarify the cause of some of these unexplained deaths or at least indicate susceptibility to triggering effects.

Cytochrome (CYP) P450 is a super family of enzymes that detoxify foreign chemicals and are involved in the metabolism of drugs (Ingelman-Sundberg, 2004). A specific gene encodes each CYP450 enzyme. This family of genes resides on many different chromosomes. The polymorphisms found in the genes encoding these enzymes result in four categories of general enzymatic activities affecting metabolism: poor (PM), intermediate (IM), extensive (EM), or ultrarapid (UM) metabolizers. These categories are the result of the number of copies of the encoded enzyme (i.e., alleles copy number) and activity of the respective enzymes per specific locus a person has, from two null alleles for PMs to multiple functional extra copies for UMs. The

normal phenotype (i.e., EM) is the result of two alleles of the CYP gene that encode a functional enzyme, and IMs have only one CYP allele encoding a functional enzyme at a specific locus or compromised function as a result of specific SNPs (Phillips, 2001). The field of pharmacogenetics studies the genetic bases for variable responses to various drug therapies. One gene, *CYP2D6*, accounts for the metabolism of 25% of all drugs currently on the market (Ingelman-Sundberg, 2004) and thus is of clinical importance. Current SNP arrays cannot reliably genotype known variants of the *CYP2D6* gene. Due to the high homology between *CYP2D6* and its pseudogenes *CYP2D7P* and *CYP2D8*, it is difficult to differentiate the variants from the active gene, *CYP2D6*, from its inactive pseudogenes. Cross-hybridization of these homologous regions across multiple assay platforms means there currently is no one method that reliably screens for clinically relevant variants within the *CYP2D6* gene. While, microarray technology can be a low-cost, high-scalability option to analyze known clinically relevant SNPs, a method needs to be developed to overcome the cross hybridization between the intended variant of *CYP2D6* and its pseudogenes.

The hypothesis of this dissertation is that it is possible to develop a targeted panel for clinically relevant variants in the *CYP2D6* gene using array-based technology that can provide accurate and reliable genotyping results. The goal was first to demonstrate that high throughput sequencing, also known as next generation or massively parallel sequencing, could reliably sequence a complex target using a model system, i.e. the hypervariable regions of the human mitochondrial genome. Then, using this advanced sequencing capability define the baseline genetic variation of the *CYP2D6* gene in a selected population and identify those genetic markers associated with metabolism capacity that would be verified against a database of actionable variants that cause reaction to drug exposure. The entire *CYP2D6* gene was sequenced to identify SNPs at a population level. These SNPs were compared against a known database of clinically

relevant samples with known metabolic responses of the same ethnic background to verify actionable variants. Once these variants were identified, a PCR assay workflow leveraging microarray technology was developed to quickly and efficiently screen individuals of interest by overcoming paralog interference. This assay can be used prior to administering drugs or post mortem to gain information about potential adverse drug reactions.

This dissertation addresses four major specific aims. Specific aim 1: Test using model system, the hypervariable regions, HV1 and HV2, of mitochondrial DNA, to demonstrate massively parallel sequencing can reliably sequence challenging targets. Specific aim 2: develop capture assays to enrich the full region of the gene *CYP2D6* for sequence analysis. Specific aim 3: completely sequence *CYP2D6* in a Caucasian sample population and describe observed genetic variability. These studies will determine a baseline of variation for the *CYP2D6* gene in a specific local Texas population. Specific aim 4: design a screening microarray assay to analyze the clinically relevant SNPs that overcomes paralog interference.

This dissertation is divided into three parts, introduction, DNA sequencing of target molecules, technology development, and conclusions with four separate chapters. The introduction and Chapter 1 discusses the field of pharmacogenetics and the gene of interest, *CYP2D6*, as it is necessary to understand how genetic contributions can affect drug metabolism to achieve the goal of a screening method of relevant variants, the gene *CYP2D6*, i.e. how it was discovered, difficulties and successes with characterizing this gene, and how it impacts medicine on an individual and at the population level. In addition, the first chapter discusses both DNA sequencing and microarray technologies to lay a foundation of capabilities and limitations using these high throughput systems. Part two addresses technology development and outcomes of my research. Chapter 2 discusses the method development of a mitochondrial DNA testing assay to

test a proof of concept that next generation sequencing can accurately analyze stretches of DNA that are both complex and highly repetitive before attempting to sequence the complex gene, *CYP2D6*. Chapter 3 builds upon the expertise gained with the development of a mitochondrial DNA sequencing assay to enable construction of a custom panel to sequence the gene *CYP2D6* on a population of Caucasian samples. The data generated from this assay were compared to determine locations within the gene that are more susceptible for genetic variation and compare the less common variants with a known database of variants. Chapter 4 discusses the workflow optimization of an on-market microarray to more efficiently and accurately call the SNPs of interest by overcoming pseudogene interference. Part three will close this dissertation with the conclusion and future directions in chapter 5.

CHAPTER1

Introduction and Foundational Background of Pharmacogenetics and gene CYP2D6

There are three responses to drug therapy, the drug works as intended, the drug has no effect, or the drug can cause adverse reactions, and these responses vary among people. The latter two responses can have serious consequences. Thus, there is emphasis on the inter-individual differences in drug response exerted by the variation in the capacity of drug metabolism. Since more than 2 million cases of adverse drug reactions (ADRs) result in 100,000 deaths annually in the United States, identifying these genetic variations that impact metabolic phenotype can have a major influence on treatment and reduce health risk (Shasty, 2006). Due to this potential impact, pharmacogenetic research has seen an explosion of activity by clinicians, geneticists, and pharmaceutical scientists. *Pharmacogenetics* is the branch of science that attempts to explain variability of one or more drug responses based on genetic variation (Kalow, 2001). Pharmacogenetics has the potential to identify the particular drug and the dose of drug that is most likely to be effective and safe for each patient, as well as those drugs and dosages that would be harmful to a patient. This field is rapidly developing and focuses on one particular gene instead of a whole host of genes (which is the domain of pharmacogenomics) (Ingelman-Sundberg, 2001). Metabolic differences in part are caused by genetic polymorphisms that encode proteins (enzymes) that inhibit or induct drug metabolism. The genetics of the drug metabolizing enzymes plays a critical role for understanding differences in drug response among individuals and may determine where ADRs can occur.

The first discoveries of inter-individual differences in response to a xenobiotic was perhaps described by Pythagoras in 510 BC when he noticed that some individuals develop hemolytic anemia in response to fava bean ingestion (Ingelman-Sundberg, 2004). Millennia later, Sir Archibald Garrod, studied patients at St. Bartholomew's Hospital in London who had alcaptonuria, and patients that had porphyria which was caused by sulphonal (a hypnotic) (Garrod, 1902).

Garrod observed that parental consanguinity was more common than usual among parents with children with alcaptonuria.

The first example of an actual pharmacogenetic study was an examination of taste blindness. This study focused on the variations in the ability to taste a foreign chemical; specifically, the ability to taste phenylthiocarbamide (PTC). While synthesizing chemicals in search for a sugar substitute, Fox (1932) observed some people could only detect a very slight taste while others claimed a bitter taste when PTC crystals were placed on the tongue.

The 1950s was the decade in which pharmacogenetics formalized and emerged as a distinct discipline. New techniques were developed to allow more accurate and precise measurements of enzyme activity, drug metabolites, and drug responses. Alf Alving observed in World War II that approximately 10% of African-American soldiers, but only a small number of Caucasian soldiers, developed acute hemolytic crises when given an average dose of primaquine or other chemically related antimalarial drugs (Clayman, 1952). This reaction was later shown to be caused by a deficiency of glucose-6-phosphate dehydrogenase (G6PD), which alters erythrocyte metabolism (Carson, 1956).

While these early landmarks of pharmacogenetic research concerned relatively common deficiencies of drug-metabolizing enzymes, the field was rediscovered with the detection of the debrisoquine/sparteine polymorphism of drug oxidation in the 1970s. In 1975, Robert Smith and his colleagues ingested 32 mg of debrisoquine, a sympathicolytic antihypertensive drug. He later recounted that "Within two hours severe orthostatic hypotension set in with blood pressure dropping to 70/50 mm Hg, hypotensive symptoms persisted for up to two days after the dose…" (Smith, 1986). His colleagues had no significant cardiovascular effects. At around the same time,

researchers studying sparteine, an anti-arrhythmic, showed that some individuals suffered from side effects including nausea, diplopia, blurred vision, and headaches after taking the drug (Eichelbaum *et al.*, 1979). In both of these studies, the researchers concluded that the reaction was caused by an impaired ability to oxidize the drug due to an autosomal recessive allele defect. The enzyme that metabolizes debrisoquine is encoded by the *CYP2D6* gene. Thus, these studies were the first to implicate genetic variants of the *CYP2D6* gene may alter metabolic function of this enzyme.

The drug metabolizing enzymes are part of a xenobiotic metabolism system occurring in virtually all eukaryotes and in many prokaryotes (Motulsky, 1957). This system protects the organism against the potential harmful effects of foreign compounds such as those obtained through nutrition. This drug metabolism can be divided into phase I (e.g. oxidation, reduction, hydrolysis, alkylation and dealkylation) and phase II (conjugation) reactions. In the phase I reactions the functional groups of the foreign compounds are modified, largely by the cytochrome P450 (CYP) family enzymes (such as *CYP2D6*, *CYP2C9*, *CYP2C19*) (Evans, *et al.*, 1999). The phase II enzymes conjugate (e.g. acetate, sulfate, glutathione, and glucuronic acid) the phase I enzyme products. Examples of Phase II enzymes are N-acetyltransferases (NAT1 and NAT2) and thiopurine methyltransferase (TMPT).

One of the highest impact groups of drug metabolizing enzymes is the Cytochrome P450 super family. Cytochrome P450 is a super family of enzymes, named because they are bound to membranes within a cell (cyto) and contain a heme pigment (chrome and P) that absorbs light at a wavelength of 450 nm when exposed to carbon monoxide (Garfinkel, 1958, Klingenberg, 1958). CYP450 enzymes are essential for the production of cholesterol, steroids, prostacyclins, and thromboxane A_2 and are involved in the detoxification of foreign chemicals and metabolism of

drugs. There are more than 50 CYP450 enzymes, but the *CYP1A2*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *CYP3A4*, and *CYP3A5* enzymes metabolize 90 percent of drugs (Wilkinson, 2005). These enzymes are expressed predominantly in the liver, but they also occur in the small intestine (reducing drug bioavailability), lungs, placenta, and kidneys (Slaughter *et al.*, 1995). Within this super family, Cytochrome P450 2D6 (*CYP2D6*) is a heme-containing enzyme that is responsible for the metabolism of at least 25% of known drugs varying across many different therapeutic classes from anti-attention deficit hyperactivity disorder (ADHD) drugs to vasodilators (Table 1). If each individual can be tested quickly and efficiently for relevant genes of the CYP450 family, with particular attention to the *CYP2D6* gene to effect better treatment, the number of deaths and severe reactions due to ADRs could be significantly reduced.

Table 1. Drugs of Different Therapeutic Classes Known to be Metabolized by *CYP2D6.* Table borrowed from Zanger *et al.*, 2004.

Therapeutic class	Drug	Pathway catalyzed by *CYP2D6*
Analgesica/antitussives	Codeine	O-demethylation
	Dextromethorphan	O-demethylation
	Dihydrocodeine	O-demethylation
	Ethylmorphine	O-deethylation
	Hydrocodone	N-demethylation
	Norcodeine	O-demethylation
	Oxycodone	O-demethylation
Anti-ADHD drug	Atomoxetine	Aromatic hydroxylation
Antiarrhythmics	Aprindine	Aromatic hydroxylation
	Encainide	O-demethylation
	Flecainide	O-dealkylation (?)
	Mexiletine	Aromatic hydroxylation
	N-propylajmaline	Benzylic hydroxylation
	Procainamide	Arylamine N-oxidation
	Propafenone	Aromatic hydroxylation
	Sparteine	Aliphatic hydroxylation

Therapeutic class	Drug	Pathway catalyzed by *CYP2D6*
Antidementia drugs	Galanthamine	O-demethylation
	Nicergoline	N-demethylation
Tricyclic antidepressants	Amitriptyline	Benzylic hydroxylation
	Clomipramine	Aromatic hydroxylation
	Desipramine	Aromatic hydroxylation
	Imipramine	Aromatic hydroxylation
	Nortriptyline	Benzylic hydroxylation
Other antidepressants	Citalopram	N-demethylation
	Desmethylcitalopram	N-demethylation
	Fluoxetine	N-demethylation
	Fluvoxamine	unclear
	Maprotiline	Unclear
	Mianserin	Aromatic hydroxylation
	Minaprine	Aromatic hydroxylation
	Mirtazapine	Aromatic hydroxylation
	Paroxetine	Demethylenation
	Venlafaxine	O-demethylation
Antidiabetic	Phenformine	Aromatic hydroxylation
Antiestrogen	Tamoxifen	Aromatic hydroxylation
Antihypertensives	Debrisoquine	Benzylic hydroxylation
	Guanoxan	Aromatic hydroxylation
	Indoramin	Aromatic hydroxylation
Antiemetics	Dolasetron	Aromatic hydroxylation
	Ondansetron	Aromatic hydroxylation
	Tropisetron	Aromatic hydroxylation
Antihistamines	Mequitazine	Aromatic Hydroxylation
	Promethazine	Aromatic hydroxylation
Antipsychotics	Haloperidol	N-dealkylation
	Perphenazine	N-dealkylation
	Risperidone	Aliphatic hydroxylation
	Thioridazine	Sulfoxidation
	Zuclopenthixol	N-dealkylation
Appetite suppressant	Dexfenfluramine	N-dealkylation

Therapeutic class	Drug	Pathway catalyzed by *CYP2D6*
Beta adrenergic blocking agents	Alprenolol	Aromatic hydroxylation
	Bufuralol	Benzylic hydroxylation
	Bunitrolol	Aromatic hydroxylation
	Bupranolol	Aromatic hydroxylation
	Carvedilol	Aromatic hydroxylation
	Metoprolol	Aliphatic hydroxylation
	Propranolol	Aromatic hydroxylation
	Timolol	O-dealkylation
Calcium antagonist	Perhexiline	Aliphatic hydroxylation
MAO-inhibitors	Amiflamine	N-demethylation
	Brofaromine	O-demethylation
Recreational drugs	Methoxyamphetamine	O-demethylation
	MDMA, MDME	Demethylenation
Vasodilatators	Cinnarizine	Aromatic hydroxylation
	Flunarizine	Aromatic hydroxylation

CYP2D6 Important Variants and Haplotypes:

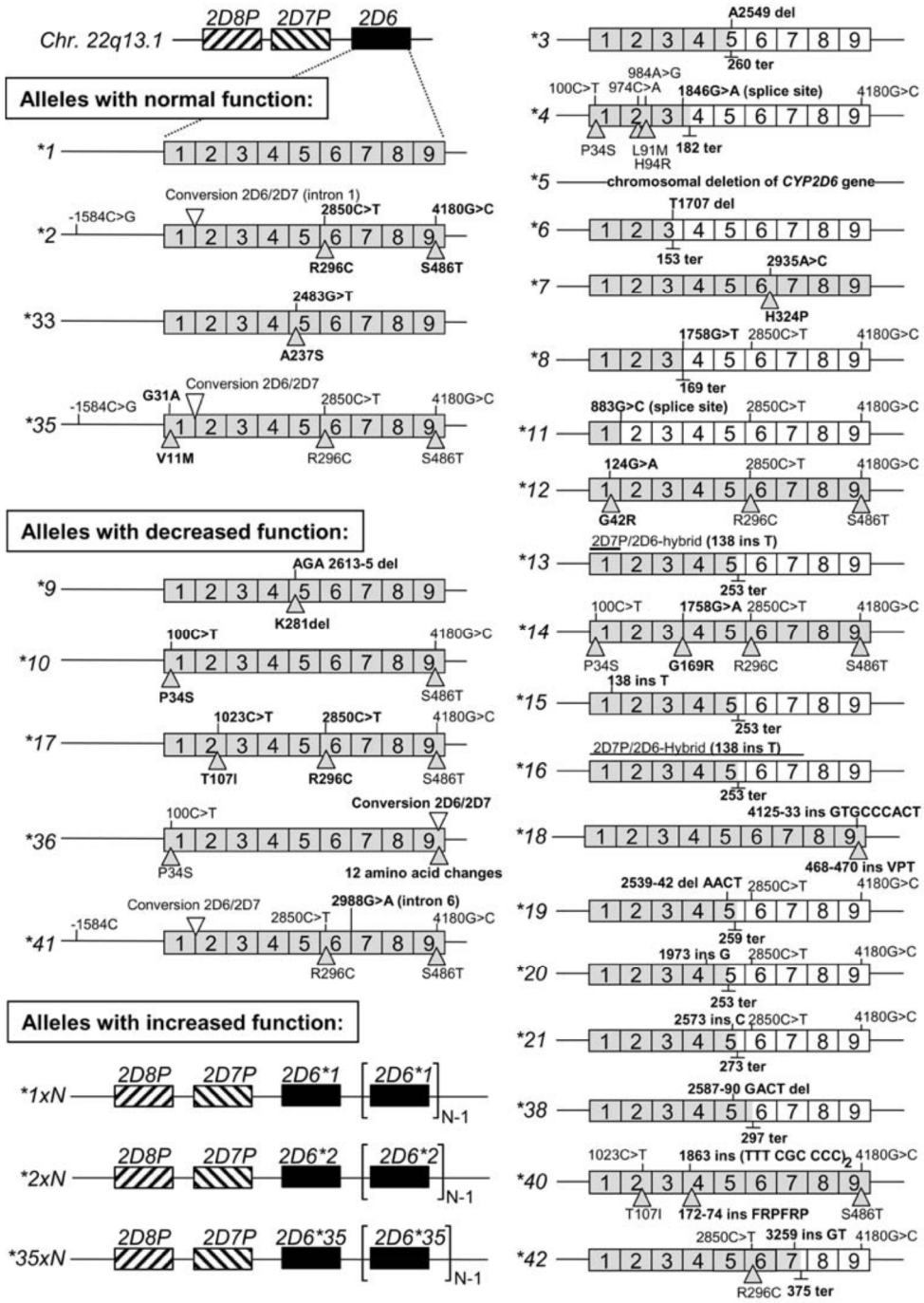
While there are currently more than 100 allelic variants and sub variants described for this gene (https://www.pharmvar.org), this review will focus on some of the most notable variants and haplotypes. The wild type allele for the *CYP2D6* gene was described in 1989 by Kimura *et al.* (1989). This allele became known as *CYP2D6**1. Since this first discovery, *CYP2D6* genotyping traditionally has been done according to an algorithm where several specific SNPs are tested, and if none are found, then the algorithm defaults to *CYP2D6* *1 (Gaedigk *et al.*, 1999). While there are a variety of SNPs, deletions, and duplications that occur in the *CYP2D6* gene, null alleles are the most studied. Null alleles do not encode a functional protein product. Only null alleles result in the PM phenotype if present in homozygous or compound heterozygous individuals. Homozygous individuals have the same null allele on both chromosomes (e.g. *3, *3), whereas

compound heterozygous individuals have two different null alleles on each chromosome (e.g., *3, *4). Different mechanisms can lead to total loss of function. First, several alleles have single base pair substitutions or small insertions/deletions that interrupt the reading frame or interfere with correct splicing and lead to premature termination of protein products. These alleles are 2D6*3, *4, *6, *8, *11, *15, *19, *20, *38, *40, *42, and *44 (Table 2) (Kagimoto *et al.*, 1990). For example, the 2D6*3 allele has an A deletion (A2637) in exon 5 causing a frameshift that results in a truncated, nonfunctional protein and is found only in Caucasian populations (Marez *et al.*, 1997). Only a few alleles encode for full-length but nonfunctional proteins (i.e., 2D6*7; Evert *et al.*, 1997; 2D6*12; Marez *et al.*, 1996; *14; Wang, 1992; *18; Yokoi *et al.*, 1996), and at least three alleles are the result of larger chromosomal deletions resulting in either removal of the entire *CYP2D6* gene (2D6*5, Gaedigk *et al.*, 1991) or in *CYP2D6*/2D7 hybrid genes with interrupted open reading frames (2D6*13 and 2D6*16; Daly *et al.*, 1996). The most frequent null allele in Caucasians is 2D6*4, which occurs with an allele frequency of about 20 to 25% and is responsible for 70-90% of all PMs. This allele has a variant that occurs due to a G to A transition in the first nucleotide of exon 4 (G1934A) and causes a shift of the consensus acceptor splice site of the third intron by one base, thereby having one additional base. This additional base causes an altered reading frame and a premature stop codon (Gaedigk *et al.*, 1999). In Asian populations the frequency of the 2D6*4 allele is around 1% or less (Wang *et al.*, 1993) and occurs in around 6 to 7% of Africans and African Americans (Leathart *et al.*, 1998).

*CYP2D6**9 is an allele associated with decreased enzyme activity. This allele was characterized initially in a family with IM phenotypes and also was found in a liver sample expressing a low amount of a variant *CYP2D6* protein (Broly *et al.*, 1993). The 2D6*9 allele was shown to lack codon 281 and was found to be enzymatically functional (Tyndale *et al.*, 1991). The

2D6*17 allele has three SNPs (T107I, R296C, and S486T) that have been shown to cause a decrease in activity in transfected COS-1 cells (Oscarson *et al.*, 1997), and the frequency of this allele is about 30% in African or African-American populations, explaining why Africans have higher median metabolic ratio (MR) values (Eichelbaum *et al.*, 1985). The allele is almost absent from European Caucasians (Griese *et al.*, 1998). *CYP2D6*4* currently has a total of 27 known variants (Gaedigk *et al.*, 2018). A list of all known variants for *CYP2D6* have been accumulated by the Pharmacogene Variation (PharmVar) Consortium at www.pharmvar.org.

Figure 1: Structure of functional and non-functional *CYP2D6* alleles. Only alleles with available phenotype information are shown. The 9 exons are indicated by numbered boxes with DNA polymorphisms indicated on top (del = deletion, ins = insertion). Predicted amino acid changes and translation termination (ter) codons are indicated below. Open reading frames are indicated by shaded boxes. Silent mutations and some promoter and intronic polymorphisms as well as alleles with uncertain function are not shown. Figure borrowed from (Zanger *et al.*, 2004).


CYP2D6 Functional Alleles
CYP2D6 Nonfunctional Alleles
Chr. 22q13.1
2D8P
2D7P
2D6
Alleles with normal function:
*1
Conversion 2D6/2D7 (intron 1)
-1584C>G
2850C>T
4180G>C
*2
R296C
S486T
2483G>T
*33
A237S
G31A
Conversion 2D6/2D7
*35
V11M
Alleles with decreased function:
AGA 2613-5 del
*9
K281del
100C>T
*10
P34S
1023C>T
*17
T107I
Conversion 2D6/2D7
*36
12 amino acid changes
2988G>A (intron 6)
-1584C
*41
Alleles with increased function:
*1xN
*2xN
*35xN
N-1
A2549 del
*3
260 ter
984A>G
100C>T
974C>A
1846G>A (splice site)
*4
P34S
L91M
H94R
182 ter
*5
chromosomal deletion of CYP2D6 gene
T1707 del
*6
153 ter
2935A>C
*7
H324P
1758G>T
*8
169 ter
883G>C (splice site)
*11
124G>A
*12
G42R
2D7P/2D6-hybrid (138 ins T)
*13
253 ter
*14
G169R
138 ins T
*15
2D7P/2D6-Hybrid (138 ins T)
*16
4125-33 ins GTGCCCACT
*18
468-470 ins VPT
2539-42 del AACT
*19
259 ter
1973 ins G
*20
2573 ins C
*21
273 ter
2587-90 GACT del
*38
297 ter
1863 ins (TTT CGC CCC)2
*40
172-74 ins FRPFRP
3259 ins GT
*42
375 ter

Individuals with extremely rapid metabolism for drugs (i.e., UMs) due to *CYP2D6* activity have multiple copies of a functional *CYP2D6* locus fused in a head-to-tail orientation as a result of unequal crossover events and other mechanisms (Bertilsson *et al.*, 1993). These multi-duplications or copy number variants (CNV) have been found with a frequency of 1 to 2% in northern European populations (Dahl *et al.*, 1995), but at much higher frequencies in Ethiopians (Aklillu *et al.*, 1996) and in Saudi Arabians (McLellan *et al.*, 1997). A table (Table 2) of selected * alleles shows how varied these alleles appear across the gene amongst Caucasians.

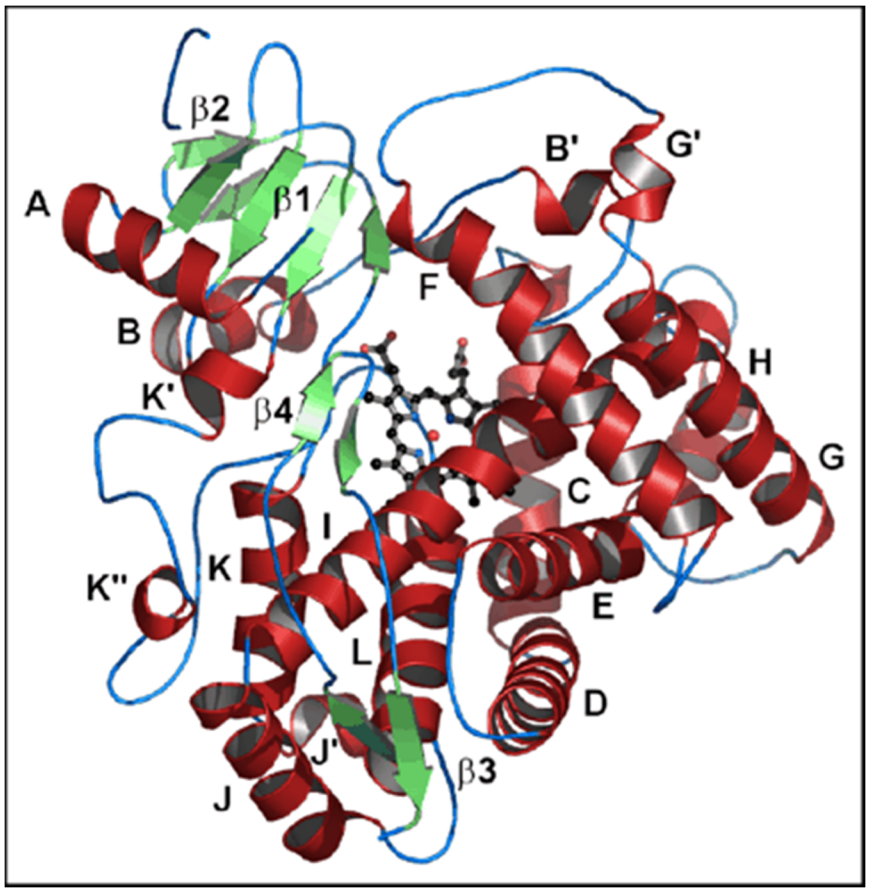
Table 2. Frequency of *CYP2D6* variants within Caucasian population. Table modified from Del Tredici *et al*., 2018.

	Allele	Caucasian			
		N	Obs freq %†	Exp freq (Avg)‡	Exp freq (Min–Max)‡
(A) ALLELES WITH A SINGLE *CYP2D6* GENE COPY					
NO FUNCTION	***3**	1,188	1.6	1.3	0.5–2.2
	***4**	14,062	18.7	19.4	17.5–21.0
	***4N**	31	0.04	–§	–§
	***6**	864	1.2	1.2	1–1.7
	***36**	12	0.02	0	0
	Total	16,157	21.6		
DECREASED FUNCTION	***9**	2,150	2.9	2.6	1.7–3.0
	***10**	1,034	1.4	2.9	1.0–8.0
	***17**	200	0.3	0.3	0.1–0.4
	***29**	71	0.1	0.1	0.02–0.1
	***41**	7,175	9.6	8.5	7.6–9.8
	Total	10,630	14.2		
NORMAL FUNCTION	***1**	28,036	37.3	37.9	33.5–41.9
	***2**	192	0.3#	25.4	15.9–33.7
	***2A**	11,717	15.6	–§	–§
	***35**	4,189	5.6	4.8	4.8

	Total	44,132	58.7		
(B) STRUCTURAL VARIANTS					
NO FUNCTION	***3xN**	**5**	**0.01**	–§	–§
	*4xN	622	0.8	0.3	0.1–0.4
	*5	2,332	3.1	2.5	0–3.8
	*6xN	3	0	0	0
	***36xN**	1	0	0	0
	Total	2,963	3.9		
DECREASED FUNCTION	***9xN**	**6**	**0.01**	–§	–§
	*10xN	11	0.01	0	0
	*17xN	0	0	0	0
	***29xN**	3	0	0	0
	***36-*10**	43	0.1	0	0
	***36-*10xN**	**1**	**0**	–§	–§
	***36xN-*10**	**2**	**0**	–§	–§
	*41xN	18	0.02	0.1	0–0.1
	Total	84	0.1		
NORMAL FUNCTION	*1xN	602	0.8	0.3	0–0.7
	*2xN	75	0.1#	0.6	0.5–0.8
	***2AxN**	**460**	**0.6**	–§	–§
	***35xN**	37	0.05	0.1	0–0.2
	Total	1,174	1.6		

The substrate of 2D6 typically contains a basic nitrogen and a planar aromatic ring with a well-defined active site cavity above the heme group. The crystal structure was solved in 2006 (Rowland *et al.*, 2006) (Figure 3).

Figure 3. Ribbon diagram of the *CYP2D6* gene. Image borrowed from Rowland *et al.*, 2006



Despite being a relatively small gene (~4400 nucleotides from starting ATG to stop codon) compared to other metabolizing enzyme encoding genes (e.g. ~90,300 nucleotides for *CYP2C19*), as more data were collected, it became apparent how the polymorphic nature of *CYP2D6,* as well as its surrounding locus add to the complexity of being able to accurately genotype it. The first challenge is the copy number variation (CNV) of this gene. Zero to 12 copies of *CYP2D6* have been described (Gaedigk *et al.*, 1991 and Johansson *et al.*, 1993). Studying a large set of de-identified clinical samples, it was discovered in the United States that 12.6 % of all patients had zero, one, or three or more copies of the *CYP2D6* gene (Beoris *et al.*, 2016). While this issue seems relatively straightforward there are two highly homologous pseudogenes, *CYP2D7P* and

CYP2D8P, to the functional gene *CYP2D6* (Figure 1) that complicate typing. The human CYP2D locus actually consists of three highly homologous genes, *CYP2D8P*, *CYP2D7P*, and *CYP2D6*, which are located within a contiguous region of about 45 kb on the q arm of chromosome 22 (22q13.1) and the gene *CYP2D6* contains 9 exons within 4,383 bp based on the NCBI 37 genome assembly (Kimura *et al.*, 1989).

The *CYP2D8P* locus is a true pseudogene with multiple deletions and insertions and no open reading frame. The *CYP2D7P* gene also is a pseudogene, only containing a single inactivating mutation in the first exon of the coding sequence. This variant causes a shift in the reading frame and premature translation termination (Kimura *et al.*, 1989). Figure 2 shows the placement of the pseudogenes relative to the active gene, *CYP2D6*, as well as where select variants are located within the active gene (Koch, 2004).

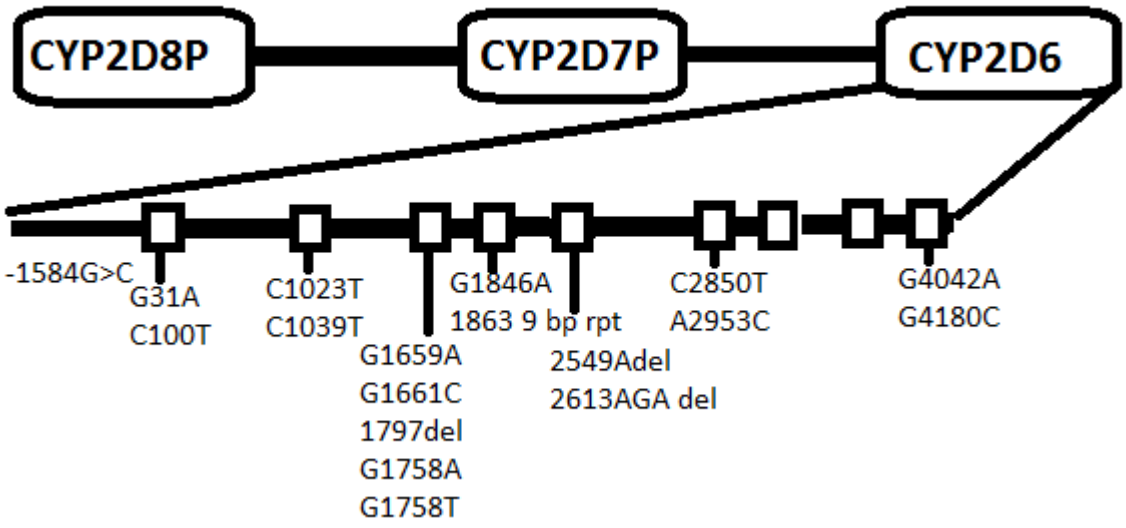
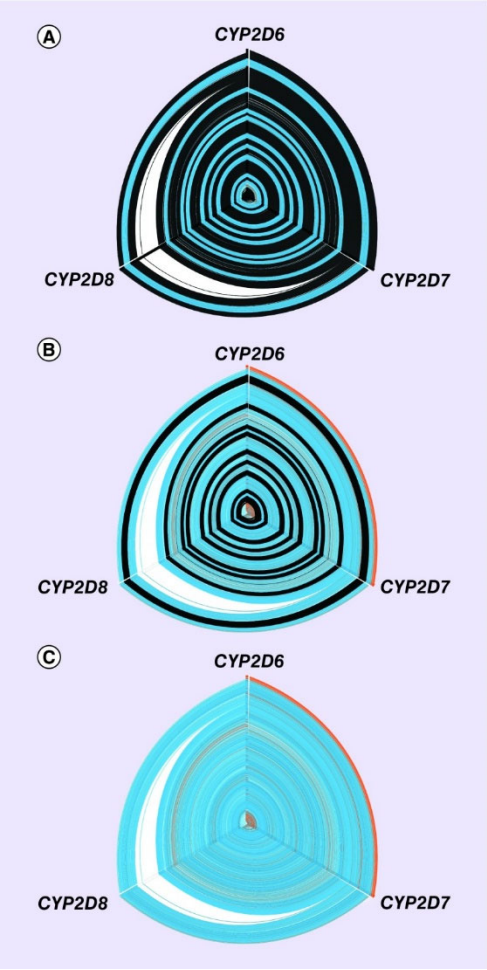


Figure 2: The location of the two pseudogenes, *CYP2D8P* and *CYP2D7P*, in relation to *CYP2D6* with select common polymorphisms for the active gene *CYP2D6* Image borrowed from Koch (2004).

Chromosome 22 was the first human chromosome to be completely sequenced in the course of the Human Genome Project (Dunham *et al.*, 1999). The *CYP2D6* gene (but not the CYP2D pseudogenes) was missing on the published sequence suggesting that the DNA sample may have been derived from an individual with a *CYP2D6* gene deletion (Idle *et al.*, 2000). Currently, different annotations are provided for the CYP2D locus depending on the genome viewer, some of which show an entry for *CYP2D6*, whereas others only show the pseudogenes. Upon inspection, of these publicly available sequences, it was revealed that the sequences provided are derived from the pseudogenes and that a functional *CYP2D6* gene sequence is still missing (Zanger *et al.*, 2004). The automatic procedures used to attach biological information to particular sequences, known as annotation, have some issues with this locus due to the pseudogenes. Computer algorithms do not discern well the pseudogenes from the active gene. Due to these limitations, the original sequence reported remains the accepted reference sequence (Kimura *et al.*, 1989).

Figure 4. Hive panel displaying multiple sequence alignment of *CYP2D6*, *CYP2D7P,* and *CYP2D8P*. The hive plot edges display sequence similarity between all three genes. Three principle axes (0, 120, and 240˚) of the hive plots represent the nucleotide composition of the multiple sequence alignment for the indicated gene: (A) exonic sequences (intronic sequences shown in black), (B) intronic sequences (exonic sequences shown in black), and (C) exonic and intronic sequences.

ClustalW was used to align the three genes (plus flanking 300 bp for each gene). Blue: aligned sequence is identical across the three genes; orange: aligned sequences are identical between the labeled genes; white: sequence gaps created by inserted nucleotides unique to the principle axis colored white (Image borrowed from Yang *et al.*, 2017).



Adding to the layer of the complexity of discerning the pseudogenes from their functional genes, there are a variety of chimeras and/or hybrids that exist between the pseudogene *CYP2D7P* and the functional gene *CYP2D6*. Two main hybrids are created based on their structure – either *CYP2D6/CYP2D7P* or *CYP2D7P/CYP2D6* hybrids, and they differ in the amount of sequence

derived from each gene. *CYP2D6/CYP2D7P* hybrids usually have a 5′-derived *CYP2D6* region and a 3′-derived *CYP2D7P* region while the *CYP2D7P/CYP2D6* hybrids have the reverse. There have been at least ten different *CYP2D7P/CYP2D6* hybrids reported thus far, each differing in the amount of each respective gene (Sim *et al.*, 2012). This complex molecular architecture and pseudogene homology result in technical challenges with targeted genotyping, full gene sequencing, and genotype interpretation. Targeted genotyping on a microarray can miss chimeras and hybrids as well as inaccurately call the number of CNVs (Pinto *et al.*, 2011). The pseudogenes play a major role in miscalls on genotyping assays. The high homology of the pseudogenes as well as potential different allele calls within separate copies of *CYP2D6* makes it incredibly challenging to assign the correct detected variant to a particular allele. Pseudogene interference can mask the true call of a SNP within the functional gene. Sequencing the full gene is difficult due to the length variation of the gene if hybrids are involved. The majority of sequencers on the market cannot sequence the entire gene in one read making read pairing extremely difficult in the case of hybrids (Treangen *et al.,* 2011 and Twyford *et al.* 2012). The homology between the pseudogenes and the functional gene adds another layer of complexity to the alignment of individual reads to the correct sequence (Meynert *et al.*, 2014). Without an accurate assessment of the gene in an individual, the ability to correlate drug metabolism to the gene becomes impossible.

ADRs in which *CYP2D6* had an impact have been implicated in a number of criminal investigations. The first noteworthy case to show that a genetically determined poor drug metabolism led to fatal drug intoxication involved a 9 year old boy who died of fluoxetine intoxication (Sallee *et al.*, 2000). The boy was treated for extreme behavioral problems with a combination of psychotherapeutic agents including fluoxetine. An extremely high concentration

of fluoxetine and its major active metabolite norfluoxetine found in several tissues in post-mortem toxicological analysis led to a pharmacogenetic analysis. This analysis revealed that the child had completely defective alleles at the *CYP2D6* gene. Another case involved a neonate who died at the age of 13 days from morphine intoxication (Koren *et al.*, 2006). The newborn received the morphine from the breast milk from his mother who had been prescribed codeine for pain management after an episiotomy. Upon pharmacogenetic investigation, the mother was found to carry a *CYP2D6* gene duplication associated with increased metabolism of codeine to morphine. The morphine levels in her breast milk were lethal to the neonate.

These cases and ones similar show the importance of *CYP2D6* in drug metabolism, and have prompted practical guidelines on the interpretation of *CYP2D6* for selected drugs by the Clinical Pharmacogenetics Implementation Consortium (CPIC) (Bell *et al*., 2017, Crews *et al.*, 2014, Hicks *et al.*, 2015, and Hicks *et al.*, 2017) and the Dutch Pharmacogenetics Working Group (DPWG) (Swen *et al.*, 2011).

These pseudogenes along with the polymorphic nature of the gene itself would prove to make sequencing and characterization of *CYP2D6* quite difficult. Understanding the work previously accomplished to characterize this gene will aid in discerning which polymorphisms are most noteworthy and crucial to developing a relevant screening assay. There are two technologies that have been used predominately to characterize genes –massively parallel sequencing and microarray. They have been beneficial but have limitations. A review of the technologies is provided to lay a foundation of how they work and why there is need to improve upon the methodologies that are developed with these platforms.

In 1977, Frederick Sanger published a seminal work describing a method in which a DNA polymerase incorporated modified nucleotides that block another nucleotide from attaching and thus ceasing strand elongation. Chain termination was achieved by exchanging the 3′ hydroxyl group with a hydrogen (Sanger *et al.*, 1977). Mixing proportions of the four native deoxynucleotide triphosphates (dNTP) with one of four of dideoxynucleotides (ddNTP) (the analogs with hydrogen) at a lower concentration, allows for a collection of nucleotide-specific terminated fragments for each of the four bases. The fragments were then size separated on a polyacrylamide gel with the A, C, G, and T reactions for the same template run in adjacent lanes. The fragment positions were determined using ^{32}P, the radioactive isotope of phosphorus and exposure to an x-ray film.

The next major leap in DNA sequencing came in 1986 when Applied Biosystems, Inc. (ABI) commercialized a fluorescent DNA sequencing instrument developed at the California Institute of Technology (Smith *et al.*, 1986). Instead of the use of radiolabeled deoxyadenosine triphosphate (dATP), the platform enabled replacement with fluorescently labeled primers (different fluor for each nucleotide reaction), thus removing the laborious processes of gel drying, X-ray film exposure and developing, reading autoradiographs, performing hand entry of the resulting sequences, and use of radiolabels. In this instrument, a scanning laser beam crossed the surface of the gel plates to excite the differentially labeled fluorescent primers whose emission was detected during electrophoretic separation of the fragments. There were two major types of fluorescently tagged nucleotides creating two types of dye-labeled sequencing. Dye-labeled primer sequencing in which the fluorescent dyes are attached to the 5′ end of the primer oligonucleotide (Smith *et al.*, 1985, Smith *et al.*, 1986, and Voss *et al.*, 1989), and dye-labeled

terminator sequencing, in which the dyes are attached to terminating dideoxynucleoside triphosphates (Prober *et al.*, 1987 and Bergot *et al.*, 1994). Originally, this process was performed using a slab gel, but in 1999 capillary electrophoresis sequencing instruments came on the market first with the MegaBACE™ sequencer from Molecular Dynamics and then the ABI PRISM® 3700 (Marsh *et al.*, 1997). Eventually the dye-labeled terminator sequencing became the dominant technology and was the technology was used to sequence the first human genome (Venter *et al.*, 2001 and International Human Genome Consortium, 2001). It is the technology that has been used in forensics for the last 20 years for mitochondrial DNA (mtDNA) sequencing.

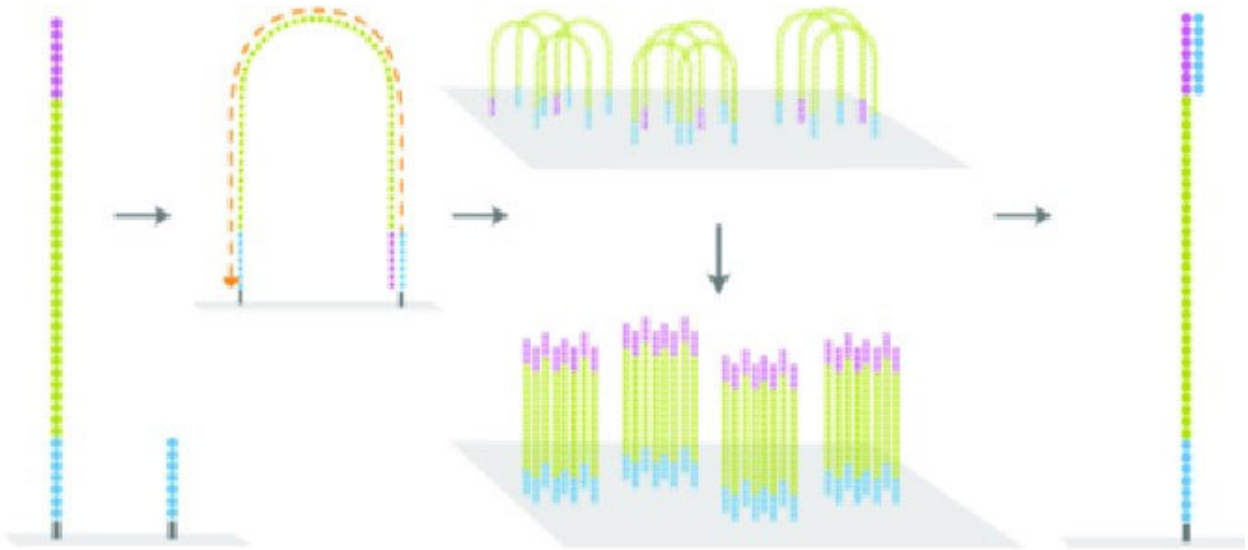
Capillary electrophoresis remained the mainstay for forensic work until recently with the advent of next generation sequencing (NGS). NGS technology vastly reduces preparation time and sequences on a massively parallel scale. DNA to be sequenced is used to construct a library of fragments with synthetic DNA (adapters) added covalently to each fragment end by use of a DNA ligase. These adapters are universal sequences that are specific to each sequencing platform that can be used to polymerase-amplify the library fragments during specific steps of the process. The library fragments are amplified (i.e., cloned) *in situ* on a solid surface, either a bead or a flat glass microfluidic channel by binding the library fragments to the surface with the complementary adapter sequence attached. The fragments are amplified many times to produce enough signal to be seen by the detection method in the instrument. Across all the currently available sequencing instruments, the process is a stepwise reaction series including: a nucleotide addition step, a detection step that determines the identity of the incorporated nucleotides on each fragment, and a wash step that removes nucleotides, fluorescent labels or blocking groups.

While there have been several NGS instruments on the market since 2005, the Illumina sequencing-by-synthesis (SBS) system is currently the most popular platform on the market and

the one used in my studies. For this purpose, only the Illumina sequencing technology will be described here.

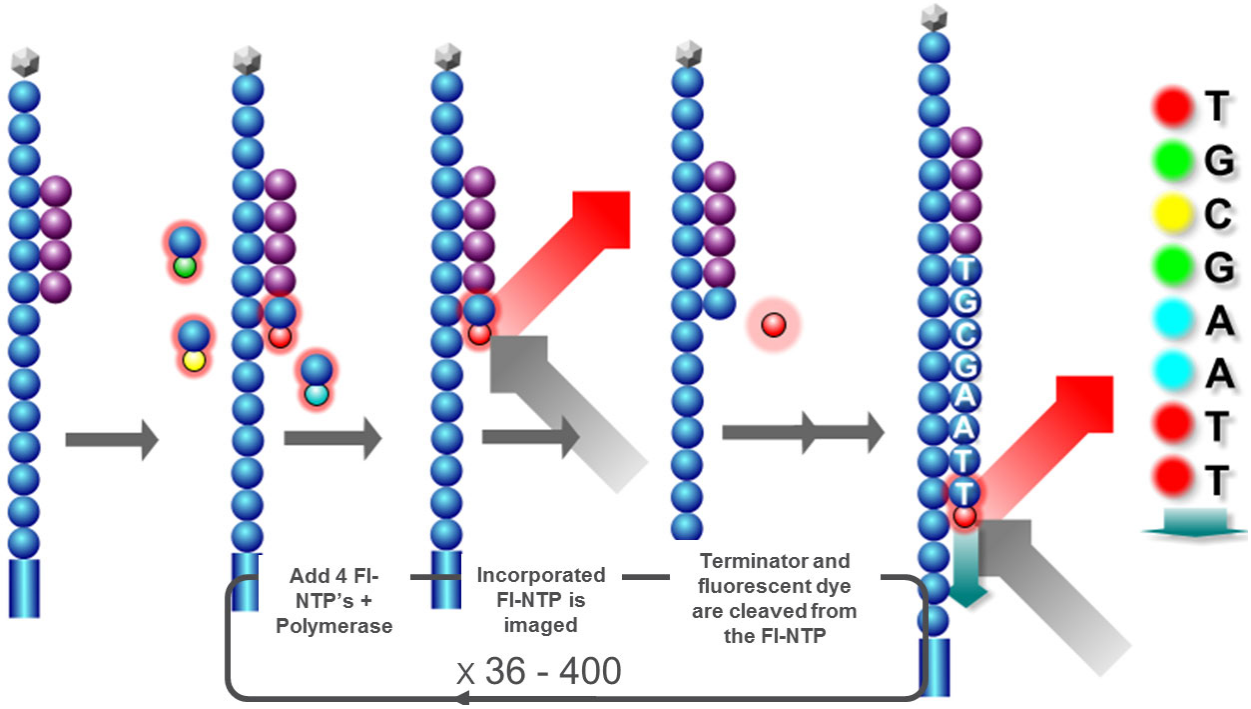
In 2006, Solexa released the Genome Analyzer (GA), and in 2007 the company was purchased by Illumina, Inc. The sequencer uses SBS technology. The library with fixed adapters is denatured to single strands and annealed to the flowcell, followed by bridge amplification to form clusters which contain clonal DNA fragments (Figure 5). Bridge amplification has the strand of DNA bend over and attach to a second oligo on the flow cell forming a bridge. A polymerase synthesizes the reverse strand. The two strands release and straighten. The process is repeated until a cluster of DNA forward and reverse strand clones exists. When the clusters are formed, the clusters of the strand not being sequenced (either forward or reverse) are washed away to allow a clone of just one strand type.

Figure 5. Bridge Amplification on Illumina flow cell (image borrowed from https://www.well.ox.ac.uk/ogc/wp-content/uploads/2017/09/Illumina_Sequencing_Overview_15045845_D.pdf)



Before sequencing, the library splices into single strands by a linearization enzyme, and then four kinds of nucleotides which contain different cleavable fluorescent dyes and a removable blocking group complement the template one base at a time (Figure 6). Once incorporated into the growing strand, the fluorescent signal is captured by a charge-coupled device (CCD) (Bentley *et al.*, 2008).

Figure 6: Incorporation of fluorescently labeled nucleoside triphosphates (Fl-NTP) by SBS method. Image borrowed from https://www.well.ox.ac.uk/ogc/wp-content/uploads/2017/09/Illumina_Sequencing_Overview_15045845_D.pdf.



If the blocking group is not removed in time for the next cycle, no nucleotide will be added to the sequence causing the sequence to be one nucleotide shorter. This is a phenomenon known as

phasing. If the nucleotide did not have the removable blocking group or it was removed too early, two or more nucleotides can be added at once making this strand longer. This is known as pre-phasing. Due to these misincorporations, the longer the read length the less accurate the sequence can become. (https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/sav/sequencing-analysis-viewer-user-guide-15020619-f.pdf). This phenomenon forces sequencing to be best with shorter fragments, usually 300 bp or less depending on the platform. The short read length capabilities of NGS instruments is overcome by tiling fragments across the region of interest, and combined bioinformatically. This limitation compared to Sanger-type sequencing which could routinely sequence much longer fragments makes it difficult to sequence areas of high repetition or complexity as the sequences cannot align properly if there is not an anchoring portion of the DNA sequence (one that aligns to only one place on the genome).

In the last few years, the increase in throughput and shortened turnaround time has been astounding. The new NovaSeq 6000 system offers output up to 6 Tb and 20 billion reads in less than two days (Table 1). This throughput equates to sequencing ~48 full human genomes with 30 x coverage, assuming 120 Gb of data per sample achieves 30x genome coverage (https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/novaseq-6000-system-specification-sheet-770-2016-025.pdf). When compared to the first human genome sequencing which took over a decade with a massive collaboration of 40 institutions and only yielded 10x coverage the technology has advanced at an impressive rate (Venter *et al.*, 2001 and International Human Genome Consortium, 2001). The NovaSeq 6000 System is built for flexibility being able to run four different types of flow cells (SP, S1, S2, and S4) and either one or two flow cells simultaneously. Depending on the

flow cell chosen and the read length, the output per flow cell can vary significantly (65 Gb up to 3000 Gb per flow cell). These choices affect both the read quality (the longer the read the lower the quality score) and run time. All specifications are outlined in Table 3.

Table 3: NovaSeq 6000 System flow cell specifications. Table borrowed from Illumina, Inc. (https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/novaseq-6000-system-specification-sheet-770-2016-025.pdf)

Flow cell type	SP	S1	S2	S4
Lanes per flow cell	2	2	2	4
Output per flow cell[a,b]				
2 × 50 bp	65-80 Gb	134-167 Gb	333-417 Gb	N/A
2 × 100 bp	N/A	266-333 Gb	667-833 Gb	1600-2000 Gb
2 × 150 bp	200-250 Gb	400-500 Gb	1000-1250 Gb	2400-3000 Gb
2 × 250 bp	325-400 Gb	N/A	N/A	N/A
Single reads CPF	0.65-0.8 B	1.3-1.6 B	3.3-4.1 B	8-10 B
Paired-end reads CPF	1.3-1.6 B	2.6-3.2 B	6.6-8.2 B	16-20 B
Quality scores[c]				
2 × 50 bp	≥ 85%			
2 × 100 bp	≥ 80%			
2 × 150 bp	≥ 75%			
2 × 250 bp	≥ 75%			
Run time[d]				
2 × 50 bp	~13 hr	~13 hr	~16 hr	N/A
2 × 100 bp	N/A	~19 hr	~25 hr	~36 hr
2 × 150 bp	~25 hr	~25 hr	~36 hr	~44 hr
2 × 250 bp	~38 hr	N/A	N/A	N/A

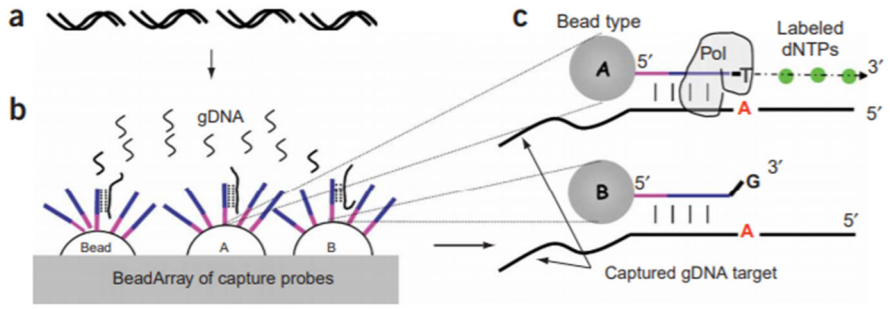
Legend for Table 3:

a) Output and read number specifications based on a single flow cell using Illumina PhiX control library
b) N/A = not applicable CPF = clusters passing filter
c) Quality scores are based on NovaSeq S2 and SP Reagent Kits run on the NovaSeq 6000 System using an Illumina PhiX control library
d) Run times are based on running two flow cells of the same type.

While sequencing was coming into its own microarray technology was advancing as well. The first known paper on microarray technology combined combinatorial, solid phase DNA synthetic chemistry with the benefits of photolithography to create the first *in situ* synthesized array (Fodor *et al.*, 1991). Here, the production of arrays of 10-amino acid peptides and, separately, arrays of dinucleotides was demonstrated. Three years later, with the formation of the company, Affymetrix, the technology improved to generate DNA arrays consisting of 256 different octa-nucleotides (Pease *et al.*, 1994). A major advantage of the Affymetrix technology at that time was that only a small set of reagents (the 4 modified nucleotides, plus a small handful of reagents necessary for the de-blocking and coupling steps) was needed to prepare a complex array. The major limitation with this technology was its lack of flexibility. Each model of array required the construction of a unique set of photolithographic masks in order to direct the light to the array at each step of the synthesis process. Eventually, this limitation was solved by using micro-mirrors to direct light at the pixels on the array (Nuwaysir *et al.*, 2002). Affymetrix has become one of dominant companies in the microarray field.

An alternative method, self-assembled arrays, involved synthesizing DNA on small polystyrene beads and depositing the beads on the end of a fiber optic array in which the ends of the fibers were etched to provide a well that is slightly larger than one bead. Different types of DNA could be synthesized on different beads and applying a mixture of beads to the fiber optic cable would result in a randomly assembled array. This technology was eventually licensed by Illumina, Inc. In earlier versions of this array, the beads were optically encoded with different fluorophore combinations allow determination of which oligos were in which position on the array (Ferguson *et al.*, 2000; Michael *et al.*, 1998; Steemers *et al*, 2000; Walt, 2000). This process limited the total number of unique beads that could be distinguished. The present methods for decoding the beads involve hybridizing and detecting a number of short, fluorescently labeled oligos in a sequential series of steps (Gunderson *et al.*, 2004). In 2005, a single base-resolution direct genomic assay without using PCR was achieved to effectively provide unlimited multiplexing (Gunderson *et al*., 2005). This process used whole-genome genotyping (WGG), a novel approach combining the processes shown in Figure 7. This approach is the basis of Illumina, Inc. products, and will be used in subsequent chapters.

Figure 7: WGG on DNA arrays (a) Whole genome amplification (WGA) to generate large amounts of amplified DNA (b) hybridization of the WGA genomic DNA (gDNA) to the oligonucleotide probe array (c) array-based allele-specific primer extension (ASPE) reaction with dNTPs by a polymerase (pol) using two separate bead types whose capture sequence is identical except for the 3′terminal base. Figure borrowed from Gunderson *et al*., 2005.



Due to the complexity of *CYP2D6* the first few assays developed to analyze this gene for variants focused on one aspect at a time, and thus exploited less complex methods than MPS and microarray. For instance, several methods started by examining the copy number variation of the gene. This variation is determined by the number of repeats of a particular gene and has been linked to differing levels of metabolic efficiency (Meijerman *et al.*, 2007). Southern blotting-restriction fragment length polymorphism (RFLP) was one of the first techniques applied to detect genetic polymorphisms. Cloned cDNA was used as a probe to detect *CYP2D6* restriction fragments after digestion of genomic DNA with the restriction enzyme *XbaI*. The most frequent fragments found with *XbaI* were 29, 44, and 13 kb in length (Gonzalez *et al.*, 1988). This method was found to have limited application because it was time consuming, required relatively large amounts of good quality DNA, and would not identify heterozygotes with the 2D6*5 allele (Steen *et al.*, 1995b).

Every general PCR is characterized by the logarithmic amplification of the target sequence. The exponential increase of amplified copy number is monitored with fluorescent probes in a technique known as real-time PCR (Wilhelm *et al.*, 2003). The introduction of real-time PCR (qPCR) techniques allows discrimination between genes having deletions, duplications, and multi-duplications, allowing for more accurate genotyping of the *CYP2D6* gene. Quantitative PCR compares threshold cycles (Ct) between the target gene and a reference sequence that does not

vary in copy content. A ΔCt value is calculated to determine CNV. This assay is based on the amplification efficiency of the two different assays that are competing in a single reaction. It has been shown that a 4% change in amplification efficiency could result in an error of up to 400% in ΔCt calculation (Guescini *et al*., 2008) and CNV results obtained by qPCR have been questioned (Armour *et al.*, 2009). In 2003, method was developed to reduce many of the steps associated with long-range PCR (amplifying large stretches of DNA in one continuous strand) was a real-time PCR method (Schaeffeler *et al.*, 2003). This method quantifies *CYP2D6* in relation to albumin as an internal reference gene. The assay was shown to correctly measure the overall number of *CYP2D6* gene copies in a variety of patient groups. It also demonstrated that the strategy of copy number determination in combination with SNP analysis is sufficient with certain empirical knowledge for correct phenotype prediction.

CYP2D6 genotyping strategies have two major limitations to overcome. They have interference with the highly homologous 2D7 and 2D8 pseudogenes and the frequent occurrence of alleles with large structural alterations. The former problem is usually overcome by using nested PCR strategies with a first PCR step designed to amplify a *CYP2D6*-specific region prior to a second larger region amplification (Heim *et al.*, 1990). Small nucleotide changes including SNPs and insertions/deletions of one or a few bases are then detected in a second PCR step either designed as a PCR/RFLP assay or as a shorter allele-specific PCR fragment used in a "multiplex" approach (Stiiven *et al.*, 1996). Another method used to determine polymorphisms is the single-strand conformation polymorphism (SSCP) strategy by which both known and unknown polymorphisms can be detected (Broly *et al.*, 1995). Newer technologies used for genotyping *CYP2D6* include real-time PCR methods (Hiratsuka *et al.*, 2000) as well as microarrays for DNA analysis (Murphy *et al.*, 2001).

In 2007, a new genotyping method to detect variant alleles of *CYP2D6* was described. This method identifies the SNP 4469 C>T with one fluorescent hybridization probe (SimpleProbes™) using the LightCycler™, (both products from Roche, Basel, Switzerland). This particular SNP is found in 21 of the 65 known variant alleles, comprising ~30% of the alleles observed in Caucasian populations. This SimpleProbe covers the SNP position encompassing alleles 2D6*2 and 2D6*41 with a T, alleles 2D6*1, *3, *4, *6, *9, *10, *15 with a C, and the deletion mutation allele 2D6*5. Using this new one-step SimpleProbe methodology, it is now possible to identify this one SNP within 2 hours which is observed in >90% in Caucasian populations (Nielson *et al.*, 2007). By examining this SNP position, the patient can then be placed in one of the four major metabolizing groups, aiding in the ability to administer the most appropriate dose of drug needed. This test could be implemented in hospitals and general practice to ensure the patient is receiving an appropriate dose of the drug, reducing accidental overdoses and other adverse reactions associated with improper drug concentrations.

The AmpliChip CYP450 test provided by Roche identifies 33 *CYP2D6* alleles, including variants associated with impaired enzyme activity and seven gene duplications, as well as two *CYP2C19* variants. This microarray was the first test approved by the FDA for *CYP2D6* analysis. This array has been proven to be an effective method to predict the PM phenotype, and relatively useful to determine the EM and IM phenotypes. A low sensitivity of UM prediction was observed, making a fast efficient method to identify the UM phenotype still in need (Rebsamen *et al.*, 2009).

The difficulty of identification of structural variants is most noted in the detection of deleted and duplicated alleles. Long-range PCR can be used to identify the variant alleles. 2D6*5 and 2D6*16 are two of the most common deletion alleles that can be detected using the same amplification reaction which results in products of different length if one of these deletion alleles

is present (Steen *et al.*, 1995a). Duplication or multi-duplication of either functional 2D6*1, 2D6*2, or 2D6*35 alleles or the nonfunctional 2D6*4 allele can either be detected by appropriate RFLP assays or by using one of several available long-distance PCR assays (Johansson *et al.*, 1996; Sachse *et al.*, 1997).

Since these research advancements, multiple companies have realized the great potential of pharmacogenetic testing, and myriad products are now on the market (Table 4). Unfortunately, there is no one kit that accurately captures all clinically relevant *CYP2D6* variants speaking volumes in regard to the complexity of this gene.

Table 4. Commercially available *CYP2D6* interrogating sequencing and microarray kits. Modified and expanded from Yang *et al.*, 2017.

Assay	**Star (*) allele haplotypes interrogated**	**Company**
xTAG *CYP2D6* Kit v3	**2, *3, *4, *5, *6, *7, *8, *9, *10, *11, *15, *17, *29, *35, *41, *xN*	Luminex[†]
Ion AmpliSeq Pharmacogenomics Research Panel	**2, *2A, *3, *4, *5, *6, *7, *8, *9, *10, *11, *12, *14, *15, *17, *20, *29, *35, *41, *29/*70, *xN*	ThermoFisher/ Ion Torrent[‡]
DMET Plus	**2, *3, *4, *5, *6, *7, *8, *9, *10, *11, *12, 14A, *14B, *15, *17, *18, *19, *20, *21, *29, *38, *40, *41, *42, *44, *56A, *56B, *64*	ThermoFisher/ Affymetrix
PharmacoScan	**2, *3, *4, *5, *6, *7, *8, *9, *10, *11, *12, *14A, *14B, *15, *17, *18, *19, *20, *21, *29, *38, *40, *41, *42, *44, *56A, *56B, *64*, plus copy number variation	ThermoFisher/ Affymetrix

iPLEX *CYP2D6* Panel	**1*, (**2;*28;*32;*55;*59*), (**2A;*31;*51*), **2D*, (**2L;*45B;*46*), **2M, *3, *4, *4B, *4J, *4K, *4M, *4N;P, *53,*6, *6C, *7, *8, *9*, (**10A;*37;*54*), (**10B;*47;*49;*52;*72*), **11, *12, *14A, *14B, *15, *17, *18, *19, *20, *21[a], *21B, *27, *29, *30, *34, *35, *36, *38, *39, *40, *41, *42, *44, *45A, *56A, *56B, *57, *58, *63, *64, *65, *68, *69, *70, *71, *82, *83, *84*	Agena Bioscience
iPLEX PGx Pro Panel	**1A*, (**2A;*31;*51*), (**2L;*35;*71*), **3, *4, *4M, *6, *7, *8, *9*, (**10;*36;*37;*47;*49;*52;*54;*57;*65;*72*), **11, *12, *14A, *14B, *15, *17, *18, *19, *20, *21A, *21B, *30, *40, *41, *42, *44, *56A, *56B, *58, *64, *69*	Agena Bioscience
GenoChip Tamox	**3, *4, *5, *6, *7, *8, *9, *10, *11, *17, *29, *41, *xN*	Akabiotech
INFINITI CYP450 2D6I	**2, *2A *3, *4, *5, *6, *7, *8, *9, *10, *12, *14, *17, *29, *41, *xN*	AutoGenomics
VeraCode ADME Core Panel	**2A, *3, *4, *5, *6, *7, *8, *9, *10, *11, *12, *14, *15, *17, *18, *19, *20, *21, *38, *41, *42, *44, *56*	Illumina
GenoChip *CYP2D6*	**3, *4, *5, *6, *7, *8, *9, *10, *11, *17, *29, *41, *xN*	PharmGenomics

GSA v2	*3A, *3B, *6, *7, *9, *22, *23, *25, *33, *38, *44, *48, *106	Illumina

References:

Akilillu E, Persson I, Bertilsson L, Johansson I, Rodrigues F, Ingelman-Sundberg M. "Frequent Distribution of Ultrarapid Metabolizers of Debrisoquine in an Ethiopian Population Carrying Duplicated and Multiduplicated Functional *CYP2D6* Alleles." *J Pharmacol Exp Ther*. 1996; 278:441-446.

Armour JA, Palla R, Zeeuwen PL, den Heijer M, Schalkwijk J, *et al.* "Accurate, high-throughput typing of copy number variation using paralog ratios from dispersed repeats. *Nucl Acid Res.* 2007;35:e19.

Bell GC, Caudle KE, Whirl-Carrilo M, Gordon RJ, Hikino K, *et al.* "Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for *CYP2D6* genotype and use of ondansetron and tropistron." *Clin. Pharmacol. Ther.* 2017;102(2):213-218.

Bentley D, Balasubramanian S, Swerdlow H, Smith G, Milton J, Brown C, et al. "Accurate whole genome sequencing using reversible terminator chemistry." *Nature* 2008; 456:53-59.

Beoris M, Amos Wilson J, Garces JA, Lukowiak AA. "*CYP2D6* copy number distribution in the US population." *Pharmacogenet Genomics.* 2016;26(2):96-9.

Bergot BJ, Chakerian V, Connell CR, Eadie JS, Fung S, et al. "Spectrally resolvable rhodamine dyes for nucleic acid sequence determination." U.S. Patent 5,366,860, 1994.

Bertilsson L, Dahl M, Sjoqvist F, Aberg-Wistedt A, Humble M, Johansson I, Lundqvist E, Ingelman-Sundberg M. "Molecular Basis for Rational Megaprescribing in Ultrarapid Hydroxylators of Debrisoquine." *Lancet*. 1993; 341:63.

Broly F, Meyer U. "Debrisoquine Oxidation Polymorphism: Phenotypic Consequences of a 3-base-pair Deletion in Exon 5 of the *CYP2D6* Gene." *Pharmacogenetics*. 1993; 3:123-130.

Broly F, Marez D, Sabbagh N, Legrand M, Millecamps S, Lo Guidice J, Boone P, Meyer U. "An Efficient Strategy for Detection of Known and New Mutations of the *CYP2D6* Gene Using Single Strand Conformation Polymorphism Analysis." *Pharmacogenomics*. 1995; 5:373-384.

Carson P, Flanagan C, Ickes C, Alving A. "Enzymatic Deficiency in Primaquine-Sensitive Erythrocytes." *Science*. 1956; 124:484-485.

Clayman C, Arnold J, Hockwald R, Yount E, Edgcomb J, Alving A. "Toxicity of Primaquine in Caucasians." *JAMA*. 1952; 149:1563-1568.

Crews KR, Gaedigk A, Dunnenberger HM, Leeder JS, Klein TE, *et al.* "Clinical Pharmacogenetics Implementation Consortium guidelines for Cytochrome P450 2D6 genotype and codeine therapy: 2014 Update." *Clin. Pharmacol. Ther.*2014;95(4):376-382.

Dahl M, Johansson I, Bertilsson L, Ingelman-Sundberg M, Sjoqvist F. "Ultrarapid Hydroxylation of Debrisoquine in a Swedish Population. Analysis of the Molecular Genetic Basis." *J Pharmacol Exp Ther*. 1995; 274:516-520.

Daly A, Fairbrother K, Andreassen O, London S, Idle J, Steen V. "Characterization and PCR-based Detection of Two Different Hybrid CYP2D7P/*CYP2D6* Alleles Associated with the Poor Metabolizer Phenotype." *Pharmacogenetics*. 1996; 6:319-328.

Del Tredici A, Malhotra A, Dedek M, Espin F, Roach D *et al.* "Frequency of *CYP2D6* alleles including structural variants in the United States." *Front Pharmacol.* 2018;9:305.

Dunham I, Shimizu N, Roe B, Chissoe S, Hunt A, Collins A, *et al*. "The DNA Sequence of Human Chromosome 22." *Nature*. 1999; 22:489-495.

Eichelbaum M, Spannbrucker N, Steincke B, and Dengler HJ. "Defective N-oxidation of sparteine in man: a new pharmacologic defect." *European Journal of Clinical Pharmacology.* 1979;16:183-187.

Eichelbaum M, Woolhouse N. "Inter-ethnic Difference in Sparteine Oxidation Among Ghanians and Germans." *Eur J Clin Pharmacol*. 1985; 28:79-83.

Evans WE and Relling MV. "Pharmacogenomics: translating functional genomics into rational therapeutics." *Science*. 1999; 286:487-491.

Evert B, Eichelbaum M, Haubruck H, Zanger U. "Functional Properties of *CYP2D6* 1 (wild-type) and *CYP2D6* 7 ($His_{324}Pro$) Expressed by Recombinant Baculovirus in Insect Cells." *Naunyn-Schmied Arch Pharmacol*. 1997; 355:309-318.

Ferguson JA, Steemers FJ, Walt DR. "High-density fiber-optic DNA random microsphere array." *Anal Chem*. 2000;72(22):5618-24.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. "Light-directed, spatially addressable parallel chemical synthesis." *Science*. 1991;251(4995):767-73.

Fox A. "The Relationship Between Chemical Constitution and Taste." *Proc Natl Acad Sci*. 1932; 18:115-120.Gaedigk A, Blum M, Gaedigk R, Eichelbaum M, Meyer U. "Deletion of the Entire Cytochrome P450 *CYP2D6* Gene as a Cause of Impaired Drug Metabolism in Poor Metabolizers of the Debrisoquine/Sparteine Polymorphism." *Am J Hum Genet*. 1991; 48:943-950.

Gaedigk A, Gotschall R, Forbes N, Simon S, Kearns G, Leeder J. "Optimization of Cytochrome P4502D6 (*CYP2D6*) Phenotype Assignment Using a Genotyping Algorithm Based on Allele Frequency Data." *Pharmacogenetics*. 1999; 9:669-682.

Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, *et al.* "The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database." *Clin Pharmacol Ther*. 2018;103:399-401.

Garfinkel D. "Studies on pig liver microsomes. I. enzymic and pigment composition of different microsomal fractions." *Archives of Biochemistry and Biophysics*. 1958;77:493-509.

Garrod A. “The Incidence of Alkaptonuria. A Study in Chemical Individuality.” *Lancet*. 1902; 1616-1620.

Gonzalez F, Skoda R, Kimura S, Umeno M, Zanger U, Nebert D, Gelboin H, Hardwick J, Meyer U. “Characterization of the Common Genetic Defect in Humans Deficient in Debrisoquine Metabolism.” *Nature*. 1988;331:442-446.

Griese E, Zanger U, Brudermanns U, Gaedigk A, Mikus G, Morike K, Stuven T, Eichelbaum M. “Assessment of the Predictive Power of Genotypes for the in-vivo Catalytic Function of *CYP2D6* in a German Population.” *Pharmacogenetics*. 1998; 8:15-26.

Guescini M, Sisti D, Rocchi MB, Stocchi L, Stocchi v. “A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition.” *BMC Bioinformatics.* 2008;30:326-338.

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, *et al.* “Decoding randomly ordered DNA arrays.” *Genome Research.* 2004;14:870-877.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, and Chee MS. “A genome-wide scalable SNP genotyping assay using microarray technology.” *Nature Genetics*. 2005;37(5):549-554.

Heim M, Meyer U. “Genotyping of Poor Metabolizers of Debrisoquine by Allele-Specific PCR Amplification.” *Lancet*. 1990; 336:529-532.

Hicks JK, Bishop JR, Sangkuhl K, Müller D, Ji Y, *et al.* “Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for *CYP2D6* and *CYP2C19* genotypes and dosing of selective serotonin reuptake inhibitors.” *Clin. Pharmacol. Ther.* 2015;98(2):127-134.

Hicks JK, Sangkuhl K, Swen JJ, Ellingrod V, Müller D, *et al.* “Clinical Pharmacogenetics Implementation Guideline (CPIC) for *CYP2D6* and *CYP2C19* genotypes and dosing of tricyclic anti-depressants: 2016 update.” *Clin. Pharmacol. Ther.* 2017;102(1):37-44.

Hiratsuka M, Agatsuma Y, Omori R, Narahara K, Inoue T, Kishikawa Y, Mizugaki M. “High Throughput Detection of Drug-Metabolizing Enzyme Polymorphisms by Allele-Specific Fluorogenic 5’ Nuclease Chain Reaction Assay.” *Biol Pharm Bull*. 2000; 23:1131-1135.

Idle J, Corchero J, Gonzalez F. “Medical Implications of HGP’s Sequence of Chromosome 22.” *Lancet*. 2000; 355:319.

Ingelman-Sundberg M. “Pharmacogenomics: an opportunity for a safer and more efficient pharmacotherapy. *J. Intern Med.* 2001; 250:186-200.

Ingelman-Sundberg M. “Pharmacogenetics of Cytochrome P450 and It’s Applications in Drug Therapy: the Past, Present, and Future.” *TRENDS in Pharmacol Sciences*. 2004; 25:193-200.

International Human Genome Sequencing Consortium. “Initial sequencing and analysis of the human genome.” *Nature* 2001;409:860-921.

Johansson I, Lundqvist E, Bertilsson L, Dahl ML, Sjoqvist *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.* 1993;90(24), 11825-11829.

Johansson I, Lundqvist E, Dahl M, Ingelman-Sundberg M. "PCR-based Genotyping for Duplicated and Deleted *CYP2D6* Genes." *Pharmacogenetics*. 1996; 6:351-355.

Kagimoto M, Heim M, Kagimoto K, Zeugin T, Meyer U. "Multiple Mutations of the Human Cytochrome P450IID6 gene (*CYP2D6*) in Poor Metabolizers of Debrisoquine: Study of the Functional Significance of Individual Mutations by Expression of Chimeric Genes." *J Biol Chem*. 1990; 265:17209-17214.

Kimura S, Umeno M, Skoda R, Meyer U, Gonzalez F. "The Human Debrisoquine 4-hydroxylase (CYP2D) locus: Sequence and Identification of the Polymorphic *CYP2D6* gene, a Related Gene, and a Pseudogene." *Am J Hum Genet*. 1989; 45:889-904.

Klingenberg M. "Pigments of rat liver microsomes." *Archives of Biochemistry and Biophysics*. 1958;75:376-386.

Koch W. "Technology platforms for pharmacogenomics diagnostic assays." *Nature Reviews Drug Discovery*. 2004;3:749-761.

Koren G, Cairns J, Chitayat D, Gaedigk A, and Leeder SJ. "Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother." *Lancet* 2006;368:704.

Leathart J, London S, Steward A, Adams J, Idle J, Daly A. "*CYP2D6* Phenotype-Genotype Relationships in African-Americans and Caucasians in Los Angeles." *Pharmacogenetics*. 1998; 8:529-541.

Mahgoub A, Idle J, Dring L, Lancaster R, Smith R. "Polymorphic hydroxylation of debrisoquine in man." *Lancet.* 1977; 2:584-586.

Marez D, Legrand M, Sabbagh N, Lo-Guidice J, Boone P, and Broly F. "An additional allelic variant of the *CYP2D6* gene causing impaired metabolism of sparteine." *Hum Genet*. 1996;97(5):668-670.

Marez D, Legrand M, Sabbagh N, Guidice J, Spire C, Lafitte J, Meyer U, Broly F. "Polymorphism of the Cytochrome P450 *CYP2D6* Gene in a European Population: Characterization of 48 Mutations and 53 Alleles, Their Frequencies and Evolution." *Pharmacogenetics*. 1997; 7:193-202.

Marsh M, Tu O, Dolnik V, Roach D, Solomon N, et al. "High-throughput DNA sequencing on a capillary array electrophoresis system." *J. Capill Electrophor.* 1997; 4:83-89.

McLellan R, Oscarson M, Seidegard J, Evans D, Ingelman-Sundberg M. "Frequent Occurrence of *CYP2D6* Gene Duplication in Saudi Arabians." *Pharmacogenetics*. 1997; 7:187-191.

Meijerman I, Sanderson L, Smits P, Beijnen J, Schellens J. "Pharmacogenetic Screening of the Gene Deletion and Duplications of *CYP2D6*." *Drug Metab Rev*. 2007;39(1):45-60.

Meynert A, Ansari M, FitzPatrick D, Taylor M. "Variant detection sensitivity and biases in whole genome and exome sequencing." *BMC Bioinformatics.* 2014;15:247-258.

Michael KL, Taylor LC, Schultz SL, Walt DR. "Randomly ordered addressable high-density optical sensor arrays." *Anal Chem.* 1998;70(7):1242-8.

Motulsky AG. "Drug reactions enzymes, and biochemical genetics." *J. Am. Med. Assoc.,*1957; 165:835-837.

Murphy G, Pollock B, Kirshner M, Pascoe N, Cheuk W, Mulsant B, Reynolds C. "*CYP2D6* Genotyping with Oligonucleotide Microarrays and Nortriptyline Concentrations in Geriatric Depression." *Neuropsychopharmacology*. 2001; 25:737-743.

Nielson K, Lyders Hansen E, Gille. "Genotyping of the Cytochrome P450 2D6 4469 C>T Polymorphism using SimpleProbes™." *Scand J Clin Lab Invest*. 2007; 67:280-290.

Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, *et al*. "Gene expression analysis using oligonucleotide arrays produced by maskless photolithography." *Genome Research.* 2002;12:1749-1755.

Oscarson M, Hidestrand M, Johansson I, Ingelman-Sundberg M. "A Combination of Mutations in the *CYP2D6**17 (*CYP2D6*Z) Allele Causes Alterations in Enzyme Function." *Mol Pharmacol*. 1997; 52:1034-1040.

Pease AC, Solas D, Sullican EJ, Cronin MT, Holmes CP, Fodor SPA. "Light generated oligonucleotide arrays for rapid DNA sequence analysis." *Proc. Natl. Acad. Sci.* 1994;91:5022-5026.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, *et al.* "Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants." *Nature Biotechnology.* 2011;29:512-520.

Pirmohamed M, Park BK. "Adverse drug reactions: back to the future." *British J of Clinical Pharmacology.* 2003; 55(5):486-492.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, et al. "A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides." *Science*. 1987;283(4825):336-41.

Rebsamen M, Desmeules J, Daali Y, Chiappe A, Diemand A, Rey C, Chabert J, Dayer P, Hochstrasser D, Rossier M. "The AmpliChip CYP450 test: Cytochrome P450 2D6 Genotype Assessment and Phenotype Prediction." *Pharmacogenomics*. 2009; 9:34-41.

Rowland P, Blaney F, Smyth M, Jones J, Vaughn L, Oxbrow A, *et al.* "Crystal Structure of Human Cytochrome P450 2D6." *J. Biol. Chem.* 2006; 11:7614-7622.

Sallee FR, DeVane CL, and Ferrell RE. "Fluoxetine-related death in a child with cytochrome P-450 2D6 genetic deficiency." *J. Child Adolesc Psychopharmacol*. 2000;10:27-34.

Sanger F, Nicklen S, Coulson AR. "DNA sequencing with chain-terminating inhibitors." *Proc. Natl. Acad. Sci.* 1997; 74: 5463-67.

Sanger M, Anandatheerthavarada H, Martin M, Guengerich F, Avadhani. "Identification of Genetic Variants of Human Cytochrome P450 2D6 with Impaired Mitochondrial Targeting." *Molec Gen Metabolism*. 2010;99(1):90-7.

Sasche C, Brockmoller J, Bauer S, Roots I. "Cytochrome P450 2D6 Variants in a Caucasian Population: Allele Frequencies and Phenotypic Consequences." *Am J Hum Genet*. 1997; 60:284-295.

Shasty B. "Pharmacogenetics and the Concept of Individualized Medicine." *Pharmacogenomics*. 2006; 6:16-21.

Sim SC, Daly AK, Gaedigk A. "*CYP2D6* update: revised nomenclature for *CYP2D7/2D6* hybrid genes." *Pharmacogenet. Genomics*. 2012;22(9):692-694.

Slaughter R, Edwards D. "Recent advances: The Cytochrome P450 enzymes." *Ann Pharmacother*. 1995; 29:619-624.

Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. "The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis." *Nucleic Acids Res*. 1985;11;13(7):2399-412.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, et al. "Fluorescence detection in automated DNA sequence analysis." *Nature* 1986; 321:674-79.

Smith R. "Introduction: Human Genetic Variations in Oxidative Drug Metabolism." *Xenobiotics*. 1986; 16:361-365.

Steen V, Molven A, Aarskog N, Gulbrandsen A. "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*. 1995a; 4:2251-2257.

Steemers FJ, Ferguson JA, Walt DR. "Screening unlabeled DNA targets with randomly ordered fiber-optic gene arrays." *Nature Biotechnology.* 2000;18:91-94.

Steen V, Andreassen O, Daly A, Tefre T, Borresen A, Idle J, Gulbrandsen A. "Detection of the Poor Metabolizer-associated *CYP2D6*(D) Gene Deletion Allele by Long-PCR Technology." *Pharmacogenetics*. 1995b;5:215-223.

Stiiven T, Griese E, Kroemer H, Eichelbaum M, Zanger U. "Rapid Detection of *CYP2D6* Null Alleles by Long Distance- and Multiplex-Polymerase Chain Reaction." *Pharmacogenomics*. 1996; 6:417-421.

Swen JJ, Nijenhuis M, DeBoer A, Grandia L, Maitland-van der Zee AH, *et al.* "Pharmacogenetics: from bench to byte—an update of guidelines." *Clin. Pharmacol. Ther.* 2011;89(5):662-73.

Treangen T and Salzberg S. "Repetitive DNA and next-generation sequencing: computational challenges and solutions." *Nat Rev Genet.* 2011;13(1) 36-46.

Twyford AD and Ennos RA. “Next-generation hybridization and introgression.” *Heredity*. 2012;108(3):179-189.

Tyndale R, Aoyama T, Broly F, Matsunaga T, Inaba T, Kalow W, Gelboin H, Meyer U, Gonzalez F. “Identification of a New Variant *CYP2D6* Allele Lacking the Codon Encoding Lys-281: Possible Association with the Poor Metabolizer Phenotype.” *Pharmacogenetics*. 1991; 1:26-32.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al. “The Sequence of the Human Genome.” *Science*. 2001;291(5507):1304-51.

Voss H, Schwager C, Wirkner U, Sproat B, Zimmermann J et al. “Direct genomic fluorescent on-line sequencing and analysis using in vitro amplification of DNA.” *Nucleic Acids Res.* 1989;11:17(7):2517-27.

Walt DR. “Techview: molecular biology. Bead-based fiber-optic arrays.” *Science*. 2000;287(5452):451-2.

Wang SL “Phenotypes and Genotypes of Debrisoquine hydroxylation polymorphisms in Chinese.” *M.Sc. thesis*. National Cheng Kung University, Tainan, Taiwan. 1992.

Wang S, Huang J, Lai M, Liu B, Lai M. “Molecular Basis of Genetic Variation in Debrisoquin Hydroxylation in Chinese Subjects: Polymorphism in RFLP and DNA Sequence of *CYP2D6*.” *Clin Pharmacol Ther*. 1993; 53:410-418.

WHO. International drug monitoring: the role of national centres. *Tech Rep Ser WHO*. 1972, no 498.

Wilheim J, Pingoud A. “Real-time Polymerase Chain Reaction.” *ChemBioChem*. 2003;4:1120-1128.

Wilkinson G. “Drug Metabolism and Variability Among Patients in Drug Response.” *N Engl J Med*. 2005; 352:2211-2221.

Yang Y, Botton M, Scott E, and Scott S. “Sequencing the *CYP2D6* gene: the variant allele discovery to clinical pharmacogenetic testing.” *Pharmacogenomics*. 2017;18(7):673-685.

Yokoi T, Kosaka Y, Chida M, Chiba K, Nakamura H, Ishizaki T, Kinoshita M, Sato K, and Kamataki T. “A new *CYP2D6* allele with a nine base insertion in exon 9 in a Japanese population associated with poor metabolizer phenotype.” *Pharmacogenetics*. 1996;6(5):395-401.

Zanger U, Raimundo S, Eichelbaum M. “Cytochrome P450 2D6: Overview and Update on Pharmacology, Genetics, Biochemistry.” *Naunyn-Schmied Arch Pharmacol.* 2004; 369:23-37.

PART 2

TECHNOLOGY DEVELOPMENT

CHAPTER 2

Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Enhanced data acquisition for DNA samples encountered in forensic testing

MPS chemistry has had difficulties with sequencing through highly repetitive regions as well as high GC or high AT regions. Before moving forward with a region as complex as *CYP2D6*, at the time it was important to optimize and verify that MPS technology could sequence through challenging regions. The mitochondrial DNA hypervariable region was chosen as a target model to determine the capabilities of MPS due to its commonalities in complexity with *CYP2D6* on a smaller scale. The hypervariable region is GC rich, highly polymorphic and contains heteroplasmies that behave similarly to pseudogenes in complex deconvolution. The hypervariable region has both long cytosine nucleotide stretches (homopolymer c stretches) and point and length heteroplasmies. Heteroplasmy is the presence of more than one type of mtDNA within a cell, tissue or individual. Point heteroplasmy manifests as two nucleotides being at the same nucleotide location (i.e. two species of the mitochondrial genome that differ at a nucleotide position). Length heteroplasmy is differences in the mitochondrial genome due insertions and/or deletions (typically observed at homopolymeric stretches). Additionally, at the time there was no streamlined approach to analyze the vast amounts of data that are generated with next generation sequencing. Testing a model first builds the knowledge and technical skills needed to tackle a gene as complex as *CYP2D6*. Thus, at the time my study was performed the non-coding region was considered to be a good model to test that next generation sequencing can be effective in a targeted assay setting. A protocol was developed to sequence the hypervariable region of the human mitochondrial DNA genome. This dissertation is a culmination of work for the past ten years, and while novel when published in 2015 the author acknowledges that this work has become routine and almost trivial by today's standards.

(in print: **Davis C**, Peters D, Warshauer D, King J, and Budowle B. "Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Improved methods for DNA samples encountered in forensic testing" (2015) *Legal Medicine* 17(2):123-7.)

Key Words: mitochondrial DNA, HVI, HVII, massively parallel sequencing, sanger type sequencing, heteroplasmy

1 Abstract

Mitochondrial DNA testing is a useful tool in the analysis of forensic biological evidence. In cases where nuclear DNA is damaged or limited in quantity, the higher copy number of mitochondrial genomes available in a sample can provide information about the source of a sample. Currently, Sanger-type sequencing (STS) is the primary method to develop mitochondrial DNA profiles. This method is laborious and time consuming. Massively parallel sequencing (MPS) can increase the amount of information obtained from mitochondrial DNA samples while improving turnaround time by decreasing the numbers of manipulations and more so by exploiting high throughput analyses to obtain interpretable results. In this study 18 buccal swabs, three different tissue samples from five individuals, and four bones samples from casework were sequenced at hypervariable regions I and II using STS and MPS. Sample enrichment for STS and MPS was PCR-based. Library preparation for MPS was performed using Nextera® XT DNA Sample Preparation Kit and sequencing was performed on the MiSeq™ (Illumina, Inc.). MPS

yielded full concordance of base calls with STS results, and the newer methodology was able to resolve length heteroplasmy in homopolymeric regions. This study demonstrates short amplicon MPS of mitochondrial DNA is feasible, can provide information not possible with STS, and lays the groundwork for development of a whole genome sequencing strategy for degraded samples.

2. Introduction

Mitochondrial DNA (mtDNA) analysis has been applied in a number of fields [1-4]. Since mtDNA is found in higher copy number per cell than nuclear DNA, it has been an invaluable genetic marker in forensics with samples that are often limited in quantity and quality [5-7]. Sequencing a portion of the human mitochondrial genome (mtGenome), typically hypervariable regions I and II (HVI and HVII, respectively), using Sanger type sequencing (STS) and capillary electrophoresis has been validated [6, 7] and prescribed under SWGDAM guidelines for forensic purposes [8]. While mtDNA sequencing can be very informative, it has not reached maximum use in forensic science due to the limitations of STS being time consuming, labor intensive, and relatively expensive. In addition, STS of mtDNA cannot readily sequence regions of amplicons downstream of length heteroplasmy homopolymers. The downstream sequence is uninterpretable due to multiple length molecules being sequenced simultaneously [9].

Massively parallel sequencing (MPS) is a high throughput technology that can rapidly generate high quality sequence from targeted areas of the human genome. The technology has reached a level of robustness such that it can be considered a viable approach to analyze challenging forensic samples. King et al. [10] have shown that a MPS system employing Nextera® XT DNA Sample Preparation Kit and the MiSeq™ platform (Illumina, Inc., San Diego, CA, USA) provides a practical protocol with long PCR and whole genome mtDNA sequencing of reference samples. To

make the transition from a reference sample capacity to that of an evidence sample sequencing capacity will require that amplicons be relatively short in length. For reference samples, the initial PCR enrichment employed generated amplicons of approximately 8 kb in length. Obviously, such length amplicons are not routinely practical for the types of samples currently analyzed by mtDNA sequencing, e.g., hair shaft, bones, and teeth. Therefore, the study described herein attempted to analyze by MPS short amplicons of the well-defined regions HVI and HVII. Concentrating only on a short portion of the mtDNA genome does not exploit the full throughput of MPS since whole genome sequencing has been demonstrated. However, this study lays the foundation that 1) MPS can provide concordant results with STS for mtDNA analyses of the type of samples that may be encountered in a forensic laboratory; 2) demonstrates amplicons containing length heteroplasmy can be sequenced without additional laboratory effort; and 3) length heteroplasmy variants can be resolved. Based on these results, efforts are underway to develop multiplex assays containing short amplicons that span additional portions or the entire mtDNA genome.

3. Methods and Materials

3.1 Samples and DNA extraction

Buccal swabs were collected from eighteen individuals. Additionally, blood was drawn by a trained phlebotomist into a lavender-top Vacutainer® tubes (Becton, Dickson and Company; Franklin, NJ, USA) blood collection tube containing EDTA. Blood, buccal cells, and forcibly removed hair were collected from an additional five individuals. All samples were collected with informed consent. These samples were extracted using the Qiagen DNA Blood Mini kit (Qiagen Inc., Valencia, CA) following manufacturer's recommendations for each tissue type. Four femur bone samples, for which STS data were available, were selected from the University of North Texas Center for Human Identification (UNTCHI) missing person identification laboratory. These

bone samples were selected because they displayed various issues with analysis by STS. The outer surface of each bone fragment was immersed in 50% bleach (3% NaOCl) for 15 min in a 50 mL conical tube. The bone fragments were washed five times with nuclease-free water, and then immersed briefly in 100% ethanol. Next, the fragments were crushed to powder using a 6750 Freezer/Mill (SPEX SamplePrep® L.L.C., Metuchen, NJ, USA), filled with liquid nitrogen, using a 10 minute chill before a 5 minute grind time at 15 impacts per second. Each aliquot was weighed and separated into 0.5 g aliquots. The bone powder was extracted using the Qiagen DNA Mini kit following manufacturer's recommendations for bones. The quantity of DNA was determined using the Quantifiler® Human DNA Quantification Kit on the ABI 7500 Real-Time PCR System (Life Technologies, Carlsbad, CA) following manufacturer's recommendations. Three of the four bones were sequenced in duplicate by MPS and one bone sample (i.e., bone 7000) yielded DNA (a total yield of 1 ng) sufficient for only one analysis.

3.2 Sanger Type Sequencing

Initially, PCR amplification for STS was performed using primers R1 (forward primer 5′-CACCAGTCTTGTAAACCGGAGA-3′) and R2 (reverse primer 5′-CTTTGGGGTTTGGTTGGTTC-3′) to amplify positions 15910-564 [11] at 0.3 µL each, 0.6 µL of AmpliTaq® Gold Polymerase (Life Technologies), 0.6 µL of dNTP mix (10 mM), 0.9 µL of $MgCl_2$ (25 mM), 1.5 µL of BSA, 1.5 µL of 10X PCR Buffer II (Life Technologies), 8.3 µL of nuclease-free water, and 1 µL of DNA (1 ng). The samples were amplified using the following parameters: 95ºC hold for 11 min, 28 cycles of 95ºC for 10 sec, 60ºC for 45 sec, and 72ºC for one minute, a hold for 10 min at 15ºC, and a final hold at 4ºC. Amplified products were purified by adding 2 µL of ExoSAP-IT® (Affymetrix, Santa Clara, CA) to each well containing 15 µL of amplified product, and placed in a thermal cycler with cycling parameters of 37ºC for 15 min, 80ºC

for 15 min, and a hold at 4ºC. The samples were cycle sequenced using the BigDye® Terminator™ v3.1 cycle sequencing kit (Life Technologies) at 1 µL, 5 µL of BetterBuffer BigDye™ dilution buffer (Gel Company Inc., San Francisco, CA), 6 µL of nuclease-free water, 1.5 µL of the forward or reverse primer, and 1 µL of PCR product accounting for 8 wells per sample. The samples were placed in the thermal cycler and amplified with the following parameters: hold at 96ºC for 3 min, 25 cycles of 96ºC for 15 sec, 50ºC for 10 sec, and 60ºC for 3 min, and a final hold at 4ºC. Samples were purified using the BigDye® XTerminator™ kit. Here 27.5 µL of nuclease-free water, 22.5 µL of SAM™ solution, and 5 µL of BigDye® XTerminator™ were added to the amplified products and then placed in a plate vortex for 30 min at a speed setting of 7.5.

Samples were subjected immediately to electrophoresis on the Applied Biosystems 3130*xl* Genetic Analyzer using Performance Optimized Polymer (POP-4™ polymer) on a 36-cm 16 capillary array of which 8 capillaries were dedicated to sample. The sequences were analyzed using the DNA Sequencing Analysis software v5.2 (Life Technologies) and Sequencher 5.1 (Gene Codes Corporation, Ann Arbor, MI).

3.4 MPS

PCR amplification for MPS was performed in four separate reactions per sample according to the Human mtDNA D-Loop Hypervariable Region Guide (Illumina http://support.illumina.com/downloads/human_mtdna_d_loop_hypervariable_region_guide_15034858.ilmn) to create four amplicons covering the following position ranges: 29-285, 172-408, 15997-16236, and 16159-16401. The quantity and quality of the PCR amplicons were determined using the Agilent High Sensitivity DNA chip on the Agilent® 2100 Bioanalyzer System (Agilent Technologies Inc., Santa Clara, CA). DNA amplicons were normalized to 0.2 ng/µL, and

combined at a 1:1:1:1 ratio for a total of 20 µL (5 µL each) to allow for multiple preparations, if desired.

Libraries were prepared from the normalized PCR amplification products (1 ng total input) using the Nextera® XT DNA Sample Preparation kit according to the Human mtDNA D-Loop Hypervariable Region Guide. Size, quantity, and quality of the libraries were determined with the Agilent DNA 1000 kit on the Agilent 2100 Bioanalyzer system. During library preparation, unique indexes were added to the DNA fragments of each sample to allow for pooling and demultiplexing of the data at the analysis stage. The libraries were normalized to 2 nM and pooled for sequencing.

Three separate runs on the MiSeq™ were performed to obtain data from the 18 buccal swabs (a total of 18 barcoded samples), 15 tissue samples (a total of 15 barcoded samples), and four bone samples (a total of 7 barcoded samples). These mtDNA samples were combined with 25% PhiX control and sequenced at a concentration of 10 pM according to manufacturer's specifications with 2 x 151 cycles.

MPS data were analyzed using the mtDNA MiSeq Reporter (MSR) plug-in, and interpreted using the mtDNA Variant Analyzer software (Illumina, San Diego, CA). The software allows for adjustable thresholds; a default detection threshold of 0.10 (10% of reads per nucleotide position), and a default analysis threshold of 0.25 (25% of reads per nucleotide position) were used for this study. The sequence of each sample was compared individually to the revised Cambridge Reference Sequence (rCRS) [12], and sequences of samples from different tissues within a single individual were compared as well.

4. Results

Although MPS allows for high throughput beyond what was tested here, only a maximum of 18 samples were sequenced simultaneously so that exceedingly high coverage could be obtained. The samples were grouped based on sample type and placed in three separate sequence analyses to generate high average coverage. The clustering and sequencing for each run were accomplished in 24 hours with an average of 1149K/mm^2 (±232) clusters and an average coverage of 230,003X (±174,091). Phred scores of 35 were used as a threshold of good quality data, and PhiX was used as a sequencing control.

The MPS generated mtDNA sequences from three different tissue types produced the same nucleotide calls within each of the five source individuals and also were concordant with base calls from STS sequence data (Table 1). However, MPS was able to sequence and provide interpretable data in a homopolymer C stretch in HVI of sample 400. This homopolymer stretch was unresolvable by STS due to length heteroplasmy. With STS all molecules essentially are sequenced together in one reaction. Length variants are superimposed over each other and downstream sequence is out of register and ambiguous (Figure 1A, B). With MPS, each cluster is sequenced independently; therefore, all length variants (above threshold) can be determined. There were four length variants detected, three contained a stretch of 10, 11, or 12 Cs and one contained 4 As and 11Cs. The high average coverage afforded by MPS (19,559X in this sample) ensured sampling of length variants sufficient to identify four different length homopolymers in this region (Table 2). At this time proportions of the length heteroplasmy cannot be provided. While the variants can be seen using Integrative Genomics Viewer (IGV [13]), the reads are substantially down sampled. Moreover, each time the data are repopulated in IGV, a different count is obtained. Software or scripts will be needed to quantify length variants. Although it would

be more cost effective to barcode many different samples and sequence them simultaneously, there may be situations where a higher depth of coverage may be desired. In such cases, length variants may be resolved and facilitate tissue-to-tissue or sample-to-sample comparisons. More research is needed to determine the degree of detectable variation among heteroplasmic samples from single individuals.

The eighteen buccal samples from various individuals yielded full target sequence for the amplified mtDNA regions (positions 16000-400) with Sanger and MPS. Complete concordance was observed between the mtDNA types produced with STS and MPS (Table 3). Four instances of length heteroplasmy, five instances of point heteroplasmy, as well as a double deletion in one sample were observed among the 18 samples. MPS correctly identified all calls in the initial analysis with the software with no manual intervention. While the point heteroplasmy was observed with STS, the length heteroplasmy in the homopolymer C stretch was unresolvable with STS. For all four instances of length heteroplasmy, MPS was able to resolve the individual sequences.

Three of the four bone samples yielded concordant results between MPS and STS (Table 4). In addition, the replicates performed from the point of extraction showed no considerable difference in coverage by MPS. One bone (sample 5612, Table 4) initially indicated a potential difference at position 16093. STS showed a point heteroplasmy (i.e., Y) whereas MPS listed it only as a C (Figure 2A). Clearly, the lower contributing T is less than 25% of the signal at position 16093. Therefore, this "apparent" difference is due to the MPS threshold setting and not the ability to detect heteroplasmy at this position. The analysis threshold initially was set at 25%, and the

MPS reads at 16093 were 80.9% C and 19.1% T (Figure 2B). These values were indicative of point heteroplasmy and thus all bone sample sequences were concordant between MPS and STS. It should be noted that the heteroplasmy contribution in STS data is not quantifiable. In contrast, MPS data are quantifiable. At this time, there is no recommended heteroplasmy detection threshold; a full validation study would be required to set the value. In the interim, the same general heteroplasmy interpretation guidelines for STS were applied with MPS data [8].

Bone sample number 7000 was difficult to sequence using STS for the HVII region due to the presence of length heteroplasmy in the homopolymer stretch and different primers (C1-5′-CTCACGGGAGCTCTCCATGC-3′ and D1- 5′-CTGTTAAAAGTGCATACCGCCA -3′ covering positions 29-429) were used to generate just the HVII amplicon. Sequencing by STS then was successful (data not shown). Only 0.25 ng of DNA was available for each PCR that would be sequenced with MPS. The sample yielded full results with the first round of testing with coverage ranging from 274 to 650,866 X and an average of 258,601 X (+/- 211,477). The base call sequence results were concordant with those generated by STS, and the length heteroplasmy was correctly called on the first analysis. These results suggested that challenged samples can be analyzed with MPS. Given the depth of coverage, the data indicate that smaller quantities of template DNA may yield typeable results. Future studies will define the minimal amount of template DNA that can yield interpretable sequence data.

5. Discussion

MPS provided concordant mtDNA sequence results from short amplicons. In addition, this technology increased the amount of information obtained from mtDNA samples, particularly regarding resolving and quantifying length and point heteroplasmy. In this study, 18 buccal swabs,

three tissue types from five individuals, and four bones from casework type samples were sequenced with STS and MPS using the Nextera® XT kit for library preparation and the MiSeq™ platform. These data support that MPS can be an attractive alternative for mtDNA sequencing, especially regarding resolution of heteroplasmic length variants, and has the potential of being an efficient and sensitive method for forensic analyses. Future work will be developing short amplicon multiplexes that span portions or the entire mtDNA genome so that greater discrimination power can be attained from degraded and low quantity samples.

6. Acknowledgements

The authors would like to thank Cydne Holt, Tom Richardson, Tamsen Dunn, Kathryn Stephens, Paulina Walichiewicz, and John Walsh of Illumina, Inc. for their expertise and advice on analysis of these samples.

7. References

[1] S. Bannwarth, V. Procaccio, A.S. Lebre, C. Jardel, A. Chaussenot, C. Hoarau, et al. Prevalence of rare mitochondrial DNA mutations in mitochondrial disorders. J Med Genet. 50 (2013) 704-14.

[2] J. Nunnari, A. Suomalainen. Mitochondria: in sickness and in health. Cell. 148 (2012) 1145-59.

[3] T. Kivisild, M. Reidla, E. Metspalu, A. Rosa, A. Brehm, E. Pennarun, et al. Ethiopian Mitochondrial DNA Heritage: Tracking Gene Flow Across and Around the Gate of Tears. The American Journal of Human Genetics. 75 (2004) 752-70.

[4] M. Richards, V. Macaulay, A. Torroni, H.-J. Bandelt. In search of geographical patterns in European mitochondrial DNA. The American Journal of Human Genetics. 71 (2002) 1168-74.

[5] P. Gill, P.L. Ivanov, C. Kimpton, R. Piercy, N. Benson, G. Tully, et al. Identification of the remains of the Romanov family by DNA analysis. Nat Genet. 6 (1994) 130-5.

[6] M.M. Holland, T.J. Parsons. Mitochondrial DNA sequence analysis-validation and use for forensic casework. Forensic Sci Rev. 11 (1999) 21-50.

[7] M.R. Wilson, J.A. DiZinno, D. Polanskey, J. Replogle, B. Budowle. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Leg Med. 108 (1995) 68-74.

[8] S.W.G.D.A.M. (SWGDAM). Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories. (2013).

[9] W Bär, B Brinkmann, B Budowle, A Carracedo, P Gill, M Holland, P Lincoln, W Mayr, N Morling, B Olaisen, P Schneider, G Tully, and M Wilson. DNA Commission of the International Society for Forensic Genetics: Guidelines for mitochondrial DNA typing. Int J Leg Med. 113 (2000) 193-6.

[10] J. L. King, B. L. LaRue, N. Novroski, M. Stoljarova, S. B. Seo, X. Zeng, et al., High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq, Forensic Sci. Int. Genet. (2014) In Press. http://dx.doi.org/10.1016/j.fsigen.2014.06.001

[11] PC Curtis, JL Thomas, NR Phillips, RK Roby. Optimization of primer-specific filter metrics for the assessment of mitochondrial DNA sequence data. Mitochondrial DNA 21(6) (2010) 191-7.

[12] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 23 (1999) 147.

[13] Integrative Genomics Viewer**,** http://www.broadinstitute.org/igv/.

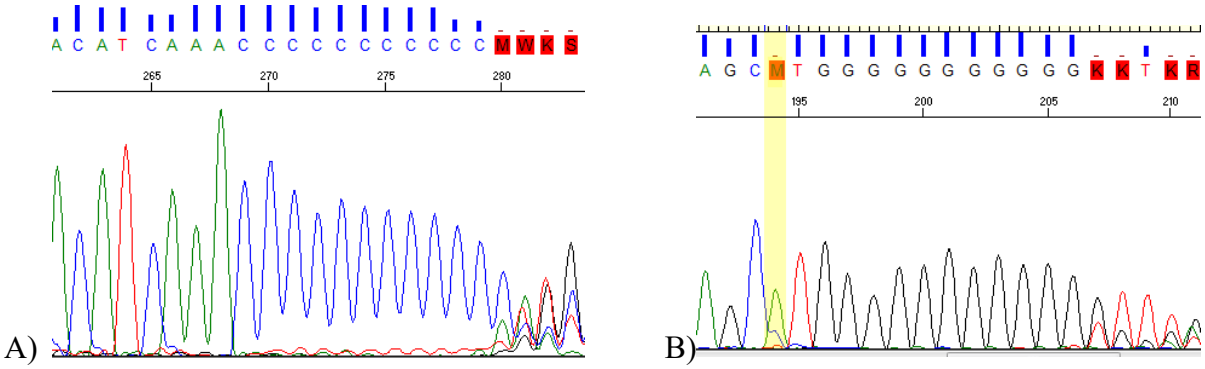
8. Tables

Table 1: Haplotypes for STS and MPS for tissue type samples[a]

Sample	Tissue type	Method	Variants											
		STS	73G	189R	199C	203A	204C	250C	263G	315.1C	16129A	16172C	16223T	16311C
	blood	MPS	73G	189R	199C	203A	204C	250C	263G	315.1C	16129A	16172C	16223T	16311C
	buccal	MPS	73G	189R	199C	203A	204C	250C	263G	315.1C	16129A	16172C	16223T	16311C
100	hair	MPS	73G	189R	199C	203A	204C	250C	263G	315.1C	16129A	16172C	16223T	16311C
		STS	73G	150T	263G	315.1C	16192T	16311C						
	blood	MPS	73G	150T	263G	315.1C	16192T	16311C						
	buccal	MPS	73G	150T	263G	315.1C	16192T	16311C						
200	hair	MPS	73G	150T	263G	315.1C	16192T	16311C						
		STS	73G	183G	263G	315.1C	16192T	16256T	16270T	16286T	16320T			
	blood	MPS	73G	183G	263G	315.1C	16192T	16256T	16270T	16286T	16320T			
	buccal	MPS	73G	183G	263G	315.1C	16192T	16256T	16270T	16286T	16320T			
300	hair	MPS	73G	183G	263G	315.1C	16192T	16256T	16270T	16286T	16320T			
		STS	73G	152C	217C	263G	309.1C	315.1C	16051G	16092C	16129C	16183C	16189C	16362C
	blood	MPS	73G	152C	217C	263G	309.1C	315.1C	16051G	16092C	16129C	16183C	16189C	16362C
	buccal	MPS	73G	152C	217C	263G	309.1C	315.1C	16051G	16092C	16129C	16183C	16189C	16362C
400	hair	MPS	73G	152C	217C	263G	309.1C	315.1C	16051G	16092C	16129C	16183C	16189C	16362C
		STS	73G	263G	315.1C	16129A	16223T	16294T	16362C					
	blood	MPS	73G	263G	315.1C	16129A	16223T	16294T	16362C					
	buccal	MPS	73G	263G	315.1C	16129A	16223T	16294T	16362C					
500	hair	MPS	73G	263G	315.1C	16129A	16223T	16294T	16362C					

a. The haplotype data exclude the length heteroplasmy of sample 400

Figure 1. STS electropherograms of homopolymer C stretch for sample 400. The left panel (A) shows the forward strand sequence. The right panel (B) shows the reverse strand sequence.



A) B)

Table 2. Length heteroplasmy sequences detected using MPS for sample 400.

rCRS	A	A	A	A	C	C	C	C	C	T	C	C	C	C	
type I	A	A	A	**C**	C	C	C	C	C	**C**	C	C	C		
type II	A	A	A	**C**	C	C	C	C	C	**C**	C	C	C	C	
type III	A	A	A	A	C	C	C	C	C	**C**	C	C	C	C	**C**
type IV	A	A	A	**C**	C	C	C	C	C	**C**	C	C	C	C	**C**

Table 3: Haplotypes for STS and MPS for 18 buccal samples

Sample	Sequencing method		Variants													
	STS	309.1C	315.1C	16218Y	16320T											
1	MPS	309.1C	315.1C	16218Y	16320T											
	STS	73G	263G	315.1C	16236T	16270T	16399G									
2	MPS	73G	263G	315.1C	16236T	16270T	16399G									
	STS	73G	150T	263G	315.1C	16192T	16311C									
3	MPS	73G	150T	263G	315.1C	16192T	16311C									
	STS	195C	214R	249-	263G	290-	291-	309.1C	315.1C	16092C	16176T	16218T	16223T	16298C	16325C	16327T
4	MPS	195C	214R	249-	263G	290-	291-	309.1C	315.1C	16092C	16176T	16218T	16223T	16298C	16325C	16327T
	STS	73G	195C	263G	315.1C	16126C	16294T	16296T	16304C							
5	MPS	73G	195C	263G	315.1C	16126C	16294T	16296T	16304C							
	STS	73G	150T	263G	309.1C	309.2C	315.1C	16172C	16223T	16257A	16261T					
6*	MPS	73G	150T	263G	309.1C	309.2C	315.1C	16172C	16223T	16257A	16261T					
	STS	263G	309.1C	309.2C	315.1C	16129A	16183C	16189C	16355T	16356C	16362C					
7*	MPS	263G	309.1C	309.2C	315.1C	16129A	16183C	16189C	16355T	16356C	16362C					
	STS	257G	263G	309.1C	315.1C	16183C	16189C	16362C								
8*	MPS	257G	263G	309.1C	315.1C	16183C	16189C	16362C								
	STS	263G	309.1C	309.2C	315.1C	16242T	16304C									
9*	MPS	263G	309.1C	309.2C	315.1C	16242T	16304C									
	STS	263G	309.1C	315.1C	16311C											
10	MPS	263G	309.1C	315.1C	16311C											
	STS	73G	263G	309.1C	315.1C	16256T	16270T	16399G								
11	MPS	73G	263G	309.1C	315.1C	16256T	16270T	16399G								
	STS	73G	263G	309.1C	315.1C	16126C	16294T	16296T	16304C							
12	MPS	73G	263G	309.1C	315.1C	16126C	16294T	16296T	16304C							
	STS	263G	309.1C	315.1C	16295T	16304C										
13	MPS	263G	309.1C	315.1C	16295T	16304C										
	STS	73G	146C	150T	263G	309.1C	315.1C	16126C	16292T	16294T	16296T					
14	MPS	73G	146C	150T	263G	309.1C	315.1C	16126C	16292T	16294T	16296T					
	STS	73G	152C	194T	263G	315.1C	16093Y	16223T	16362C	16390A						
15	MPS	73G	152C	194T	263G	315.1C	16093Y	16223T	16362C	16390A						
	STS	73G	146C	152C	263G	315.1C	374R	16256T	16270T	16278T	16318M	16399G				
16	MPS	73G	146C	152C	263G	315.1C	374R	16256T	16270T	16278T	16318M	16399G				
	STS	73G	152C	199C	204C	207A	250C	263G	309.1C	315.1C	16129A	16223T	16391A			
17	MPS	73G	152C	199C	204C	207A	250C	263G	309.1C	315.1C	16129A	16223T	16391A			
	STS	73G	146C	154C	200G	215G	263G	310C	318C	326G	16223T					
18	MPS	73G	146C	154C	200G	215G	263G	310C	318C	326G	16223T					

*Haplotype data excludes the length heteroplasmy of this sample

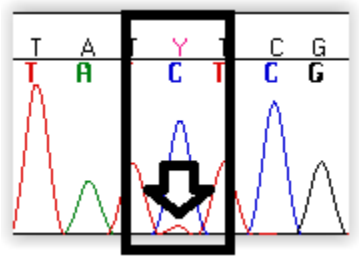
Table 4: Haplotypes for STS and MPS for bone samples

Sample	Sequencing method	Variants														
	STS	73G	263G	315.1C	**16093Y***	16223T	16239T	16260T	16274A	16325C	16362C					
5612	MPS	73G	263G	315.1C	**16093C**	16223T	16239T	16260T	16274A	16325C	16362C					
	STS	263G	315.1C	16192T	16260T											
6997	MPS	263G	315.1C	16192T	16260T											
	STS	64T	73G	146C	153G	155C	203A	222T	235G	263G	315.1C	16111T	16184T	16223T	16290T	16319A	16362C
H2	MPS	64T	73G	146C	153G	155C	203A	222T	235G	263G	315.1C	16111T	16184T	16223T	16290T	16319A	16362C
	STS	55C	56G	64T	73G	263G	279C	309.1C	315.1C	16126C	16223T	16325C	16362C				
7000	MPS	55C	56G	64T	73G	263G	279C	309.1C	315.1C	16126C	16223T	16325C	16362C				

*All samples had threshold set at 25%, position 16093 for sample 5612 detected heteroplasmy at 19.1% showing all sites concordant

Figure 2. Point heteroplasmy determination. The top panel (A) is a STS electropherogram. The bottom panel (B) is MPS output from the MRS plug-in. Both sequencing methods detected the point heteroplasmy at position 16093.

A)



B)

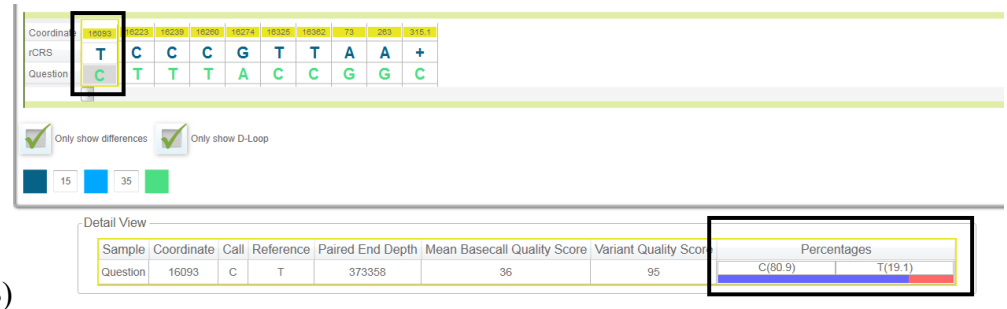


Figure 2: Position 16093 Y for STS and MPS in respective software

The hypervariable region of the mtDNA yields multiple noted difficult sequence challenges. This work demonstrated complex sequences can be typed with MPS. The ability to accurately sequence homopolymer C stretches, with length heteroplasmies, without complications as well as accurately calling point heteroplasmies indicated that potential sequence variation with the challenges of *CYP2D6* could be undertaken. This lower cost proof of concept protocol supported pursuing next generation sequencing of *CYP2D6* discussed in the next chapter.

CHAPTER 3

Characterizing the CYP2D6 gene with complete sequence data

Introduction

The previous section demonstrated that complex target molecules could be sequenced reliably, allowing knowledge to be gained to design a panel, sequence a complex gene, and provided the expertise to interpret large, complex data. Genetic studies on *CYP2D6* to determine gain or loss of function has been targeted using a variety of SNP-based capture assays usually performed with enrichment by PCR (Friedrich *et al.,* 2014; Levo *et al*., 2003; Regan *et al.*, 2012). While this approach can capture the known variation of *CYP2D6*, such analyses are biased and can miss yet to be defined variation. A full unbiased analysis has shown to be more beneficial when sequencing the gene in its entirety (Wendt *et al.*, 2018a). The results can help clarify the gene diversity of the locus as well as show whether a SNP-based assay is sufficient for diagnostic purposes. Therefore, MPS was used to more completely determine genetic variation at the *CYP2D6* locus. This study mines all variants from the 1000 genomes project related to *CYP2D6*, as well as a full gene sequence of 65 Caucasian individuals to gain a more complete analysis of variants found at the individual and population level (for Caucasians only). This study demonstrates the utility of MPS to quickly and accurately obtain comprehensive sequence data.

The human genome is highly organized, and recombination is not distributed evenly across the genome (McVean *et al.*, 2004). There are recombination hotspots and coldspots resulting in block-like patterns in the genome, yielding information where one can infer how vital to the organism that section of DNA is (Mackiewicz D *et al.*, 2013). Those areas that have many variants within the same "block," or hotspots, are more likely to tolerate mutations, and tend to be 1-2 kb in length. The hotspot distribution is usually positively correlated with GC content and repetitive element distribution (Myers *et al.,* 2005). Those areas that have very few variants within the block, or a coldspot, may indicate conserved regions where mutations may be detrimental to the

organism, i.e. changes may not be evolutionarily preferred (Patil *et al*., 2001). These variants are of particular interest in this study to determine what variants may be inferred to cause metabolic changes.

Methods and Materials

Custom Probe Design

Probes were designed for the entire gene *CYP2D6* using Illumina's design software DesignStudio (Illumina, 2014). The probes were designed to cover the entire gene (introns and exons) as well as 5000 bp on either side of the gene. In order to ensure all areas were covered with particular attention to exons targeting 1000 X coverage designing probes in a tiling method to properly cover exons. Introns were targeted at 50 X coverage.

Samples and DNA Extraction:

Whole blood samples were collected by venipuncture with lavender-top Vacutainer® tubes (Becton, Dickson and Company; Franklin, NJ, USA) from 65 anonymized and apparently unrelated Caucasian individuals according to protocols approved by the University of North Texas Health Science Center's Institutional Review Board. DNA was extracted using the QIAamp® DNA Blood Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's recommendations.

DNA Library Preparation:

The quantity of DNA was determined using the Quantifiler® Human DNA Quantification Kit on the ABI 7500 Real-Time PCR System (Life Technologies, Carlsbad, CA) following the

manufacturer's recommendations. DNA was normalized to 20 ng/µL in DNA free water, then 52.5 µL of extract were added to a microTUBE and sheared in a Covaris™ S2 (Woburn, MA) for 120 seconds to target 500 bp long fragments. Fifty µL (1 µg) were removed from the microTUBE to be used for library preparation.

Libraries were prepared using the TruSeq™ Custom Enrichment DNA Sample Preparation Kit (Illumina, Inc.) using the custom probes, designed previously, following the manufacturer's guidelines. Twelve libraries were pooled to run per lane.

Sequencing:

Libraries were sequenced using the GAIIx™ (Illumina, Inc.) using paired end cluster generation kit v4 and TruSeq SBS sequencing kit v5 using 2 x 146 bp read length with 6 bp single read indexes according to the manufacturer's protocol.

Data Analysis:

The sequence data obtained from the Genome Analyzer IIx were converted from base call and quality score file (BCL) files to FASTQ with demultiplexing of the samples. The FASTQ files were aligned with Burroughs-Wheeler Alignment (BWA) (Li *et al.*, 2009a) to obtain binary alignment/map (BAM) files (Li H *et al*. 2009b). The BAM files were converted to sequence alignment/map format (SAM) files (Li H *et al*. 2009b), and then a variant caller was run to obtain variant call format files (VCFs) (Danecek *et al.*, 2011).

The SNPs were analyzed to determine what effect a SNP caused, i.e. if an amino acid change occurred. The data were evaluated by both PolyPhen (Adzhubei *et al.*, 2010) and SIFT (Kumar *et al.*, 2009) to determine protein functionality with each variant.

Results

The data from the Texas population were evaluated. The entire gene plus surrounding areas were sequenced with no gaps at an average read depth of 288 reads with a minimum read depth of 21 reads and a maximum read depth of 2317. This difference in read count was expected as more probes were designed to allow for tiling 1000 X coverage in exonic regions whereas intronic regions were targeted at 50 X coverage. No read below the Phred quality score of Q30 was used to ensure the best quality and reads below 30 X coverage were considered low quality and analyzed with caution.

275 distinct SNPs were found within the Texas Caucasian population sample.

Of those Texas SNPs, only eight were found to be potentially damaging by both PolyPhen and SIFT (Table 1, supplementary tables 1-65 for individual SNP profile). Interestingly, the same location had two separate SNPs twice making it only six separate locations where SNPs occurred that could be potentially damaging. Both of the locations sporting two separate nucleotide changes were also observed in 1000 genomes. One location yielded two known SNPs rs72552269 and rs1065852, and one location yielded the known SNP rs28371703 that accounts for the G>T/T and G>G/T forms. These SNPs appear to be in a relatively high frequency, but probably because they cover two separate SNPs at the same location. SNP rs28371703 has a relatively low MAF of 0.095 according to Pharmgkb (https://www.pharmgkb.org/), and a comparable MAF of .0962 in the Texas population. The second SNP, rs1065852, had a MAF of 0.23 according to dbSNP, and a similar MAF of 0.2578 with the Texas population sample. This SNP is correlated with *CYP2D6*4*, a known variant that causes poor metabolism if found in both copies. This variant has been reported to be present at a frequency of 20% in the Caucasian population (Bradford, 2002). One

novel potentially damaging SNP (T>T/G) was discovered in the Texas population samples and was not in the 1000 genomes data. It resides at location 42525793 and causes a missense variation. This SNP has not been described in the published literature or existing on-line databases providing evidence that sequencing the entire gene and surrounding areas can aid in better identifying potentially harmful variants in *CYP2D6* as this SNP would not have been found by testing for only known variants. Table 2 includes all variants with their predicted protein functionality found in the Texas population for reference. All the SNPs were tested against a known database of clinical annotations on PharmGKB and found that only the SNP at position 42525793 was unknown. All others correlated with this information.

Linkage disequilibrium (LD) potentially can be a useful tool to determine if variants can deviate from HWE expectations (Wendt *et al.*, 2018b). Future work could perform this test to determine if any SNPs presented here might have disequilibrium, but more testing would need to be performed to create a larger data set in order to truly address LD.

Conclusion

Traditionally, large genome wide association studies (GWAS) were performed to look for potentially damaging variants. By sequencing the entire gene instead of focusing on common variant SNPs, a more comprehensive view of SNP patterns was developed. This comprehensive view helped determine potential hot spots of polymorphic activity as well as find rare variants that may have a larger impact than previously known. Sequencing also allows for the analysis of potential SNPs in the intronic region. This study alone observed 56 unique SNPs in the intronic region. This study adds to the baseline information of *CYP2D6* variation and demonstrates that complete gene sequencing can lead to a more informed analysis of potential variants than targeted

SNP assays by allowing the interrogation of areas of the genome that are not already well characterized. Discovering a novel SNP that could have a negative effect on protein functionality shows the importance of sequencing the entire region to gain a full picture of variants within *CYP2D6.*

References:

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P*, et al.* "A method and server for predicting damaging missense mutations." *Nat Methods.* 2010;7(4):248-249.

Bernard S, Neville KA, Nguyen AT, Flockhart DA. "Interethnic differences in genetic polymorphisms of *CYP2D6* in the U.S. population: clinical implications." *Oncologist* 2006;11:126-135.

Bradford LD. "*CYP2D6* allele frequency in European Caucasians, Asians, Africans and their descendents." *Pharmacogenomics* 2002;3:229-243.

Danecek P, Auton A, Abecasis G, Albers C, Banks E, DePristo M, *et al*.(2011) "The variant call format and VCFtools." *Bioinformatics.* 2011;27(15): 2156-58.

Eichelbaum M, Ingelman-Sundberg M, Evans WE. "Pharmacogenomics and individualized drug therapy." *Annu Rev Med.* 2006;57:119-137.

Friedrich DC, Genro JP, Sortica VA, Suarez-Kurtz G, Elizabete de Moraes M, Pena SDJ, *et al*. "Distribution of *CYP2D6* Alleles and Phenotypes in the Brazilian Population." *PLoS ONE*. 2014;9(10):e110691.

Illumina, Inc. "Illumina designstudio: Web-based design tool for custom targeting next-generation sequencing projects." http://designstudio.illumina.com/ 2014.

Ingelman-Sundberg M, Rodriguez-Antona C. "Pharmacogenetics of drug metabolizing enzymes: implications for a safer and more effective drug therapy." *Philos Trans R Soc Lond B Biol Sci.* 2005;360:1563-1570.

Kimura S, Umeno M, Skoda RC, Meyer UA, and Gonzalez FJ. "The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic *CYP2D6* gene, a related gene, and a pseudogene." *Am J Hum Genet.* 1989;45(6):889-904.

Kirchheiner J, Nickchen K, Bauer M, Wong ML, Licinio J, *et al.* "Pharmacogenetics of antidepressants and antipsychotics: the contribution of allelic variations to the phenotype of drug response." *Mol Psychiatry*. 2004;9:442-473.

Kumar P, Henikoff S, Ng PC. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm." *Nat Protoc*. 2009;4(7):1073-81.

Levo A, Koski A, Ojanperä, Vuori E, and Sajantila A. "Post-mortem SNP analysis of *CYP2D6* gene reveals correlation between genotype and opioid drug (tramadol) metabolite ratios in blood." *Foren Sci Int*. 2003;135(1):9-15.

Li H and Durbin R. "Fast and accurate short read alignment with Burrows-Wheeler Transform." *Bioinformatics*. 2009a;25:1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* "The Sequence Alignment/Map format and SAMtools." *Bioinformatics*. 2009b;25(16):2078-9.

Mackiewicz D, de Oliveira, Moss de Oliveira S, Cebrat S. "Distribution of Recombination Hotspots in the Human Genome – A Comparison of Computer Simulations with Real Data." *PLoS ONE*. 2013;8(6):e65272.

McVean GA, Myers S, Hunt S, Deloukas P, Bentley D, Donnelly P. "The fine-scale structure of recombination rate variation in the human genome." *Science*. 2004;304(5670):581-584.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. "A fine-scale map of recombination rates and hotspots across the human genome." *Science.* 2005;310:321-324.

Patil N, Berno A, Hinds D, Barrett W, Doshi J, Hacker *et al.* "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." *Science*. 2001;294(5547):1719-1723.

Regan M, Leyland-Jones B, Bouzyk M, Pagani O, Tang W, Kammler R, *et al*. "*CYP2D6* genotype and Tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: The Breast International Group 1-98 Trial." *J Natl Cancer Inst* 2012;104 (6):441-451.

Sistonen J, Sajantila A, Lao O, Corander J, Barbujani G, *et al*. "*CYP2D6* worldwide genetic variation shows high frequency of altered activity variants and no continental structure." *Pharmacogenet Genomics* 2007;17: 93-101.

Teh LK and Bertilsson L. "Pharmacogenomics of *CYP2D6*: molecular genetics, interethnic differences and clinical importance." *Drug Metab Pharmacokinet.* 2012;27:55-67.

Wendt FR, Sajantila A, Moura-Neto RS, Woerner AE, Budowle, B. "Full-gene haplotypes refine *CYP2D6* metabolizer phenotype inferences. *Int J Legal Med.* 2018a;132(4):1007-1024.

Wendt FR, Pathak G, Sajantila A, Chakraborty R, Budowle B. "Global genetic variation of select opiate metabolism genes in self-reported healthy individuals." *Pharmacogenomics J.* 2018b;18(2):281-294.

Zhou SF. "Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I." *Clin Pharmacokinet*. 2009a;48:689-723.

Zhou SF "Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part II." *Clin Pharmacokinet.* 2009b;48:761-804.

Table 1. List of potentially damaging variants found in Texas population

Variant	rsnumber	Coordinate	Type	Exonic	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen
C>C/T	rs1058172	42523528	snv	yes	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)
T>T/G	rs5030867	42523858	snv	yes	missense_variant	1061	971	324	H/P	cAt/cCt	deleterious(0.04)	probably_damaging(0.938)
T>T/G		42525793	snv	yes	missense_variant	389	299	100	D/A	gAc/gCc	deleterious(0)	probably_damaging(1)
G>G/T	rs28371703	42525821	snv	yes	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)
G>T/T	rs28371703	42525821	snv	yes	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)
C>C/T	rs118203758	42526669	snv	yes	missense_variant	215	125	42	G/E	gGg/gAg	deleterious(0.02)	probably_damaging(0.963)
G>G/A	rs72552269	42526694	snv	yes	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)
G>A/A	rs1065852	42526694	snv	yes	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)

Table 2: Allele Variants with Predicted Protein Functionality for Texas Caucasians

Variant	Coordinate	Type	Genotype	Exonic	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Number of Instances
G>G/A	42522550	snv	het	yes	3_prime_UTR_variant	1610	0	0					13
G>G/C	42522613	snv	het	yes	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	26
G>C/C	42522613	snv	hom	yes	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	10
A>A/C	42522660	snv	het	yes	synonymous_variant	1500	1410	470	T	acT/acG			1
A>A/G	42522768	snv	het	yes	intron_variant	0	0	0					2
G>G/A	42522774	snv	het	yes	intron_variant	0	0	0					1
A>G/G	42523003	snv	hom	yes	intron_variant	0	0	0					22
A>A/G	42523003	snv	het	yes	intron_variant	0	0	0					30
T>C/C	42523209	snv	hom	no	intron_variant	0	0	0					28
T>T/C	42523209	snv	het	no	intron_variant	0	0	0					24
T>T/C	42523211	snv	het	no	intron_variant	0	0	0					19
T>C/C	42523211	snv	hom	no	intron_variant	0	0	0					7
T>T/C	42523315	snv	het	no	intron_variant	0	0	0					1
G>G/T	42523358	snv	het	no	intron_variant	0	0	0					2
A>A/G	42523400	snv	het	no	intron_variant	0	0	0					3

G>G/T	42523409	snv	het	no	intron_variant	0	0	0					36
C>C/T	42523528	snv	het	yes	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	25
A>A/G	42523539	snv	het	yes	synonymous_variant	1173	1083	361	H	caT/caC			1
C>C/T	42523613	snv	het	yes	missense_variant	1099	1009	337	D/N	Gac/Aac	tolerated(0.21)	benign(0.003)	2
C>C/A	42523636	snv	het	yes	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	22
C>C/T	42523805	snv	het	no	intron_variant	0	0	0					19
C>T/T	42523805	snv	hom	no	intron_variant	0	0	0					1
G>G/A	42523813	snv	het	no	intron_variant	0	0	0					1
T>T/G	42523858	snv	het	yes	missense_variant	1061	971	324	H/P	cAt/cCt	deleterious(0.04)	probably_damaging(0.938)	1
A>G/G	42523943	snv	hom	yes	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	26
A>A/G	42523943	snv	het	yes	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	26
T>T/A	42524054	snv	het	no	intron_variant	0	0	0					1
C>C/A	42524057	snv	het	no	intron_variant	0	0	0					1
A>A/T	42524058	snv	het	no	intron_variant	0	0	0					1
G>G/A	42524090	snv	het	no	intron_variant	0	0	0					1

C>C/T	42524138	snv	het	no	intron_variant	0	0	0					1
CCTT>CCTT/C	42524175	deletion	het	yes	inframe_deletion, splice_region_variant			281	K/-	AAG/-			3
G>G/T	42524218	snv	het	yes	synonymous_variant	891	801	267	P	ccC/ccA			3
CT>CT/C	42524243	deletion	het	yes	frameshift_variant, feature_truncation	865	775	259					2
C>C/A	42524310	snv	het	yes	missense_variant	799	709	237	A/S	Gct/Tct	tolerated(0.34)	benign(0.11)	2
A>A/G	42524323	snv	het	yes	synonymous_variant	786	696	232	H	caT/caC			1
A>A/C	42524408	snv	het	no	intron_variant	0	0	0					1
T>T/A	42524435	snv	het	no	intron_variant	0	0	0					1
A>A/C	42524471	snv	het	no	intron_variant	0	0	0					1
T>T/C	42524696	snv	het	no	intron_variant	0	0	0					19
T>C/C	42524696	snv	hom	no	intron_variant	0	0	0					7
T>T/C	42524708	snv	het	no	intron_variant	0	0	0					36
C>C/G	42524713	snv	het	no	intron_variant	0	0	0					36
G>G/A	42524743	snv	het	no	intron_variant	0	0	0					35
A>A/G	42524795	snv	het	yes	synonymous_variant	747	657	219	F	ttT/ttC			36

A>A/G	42524924	snv	het	yes	synonymous_variant	618	528	176	G	ggT/ggC			34
C>C/T	42524947	snv	het	yes	splice_acceptor_variant	0	0	0					18
C>T/T	42524947	snv	hom	yes	splice_acceptor_variant	0	0	0					7
C>C/T	42524975	snv	het	no	intron_variant	0	0	0					7
C>C/T	42524982	snv	het	no	intron_variant	0	0	0					3
T>T/G	42525010	snv	het	no	intron_variant	0	0	0					1
G>G/C	42525132	snv	het	yes	synonymous_variant	498	408	136	V	gtC/gtG			27
G>C/C	42525132	snv	hom	yes	synonymous_variant	498	408	136	V	gtC/gtG			9
G>G/A	42525180	snv	het	yes	synonymous_variant	450	360	120	F	ttC/ttT			2
G>G/C	42525416	snv	het	no	intron_variant	0	0	0					2
C>C/T	42525625	snv	het	no	intron_variant	0	0	0					8
A>A/C	42525728	snv	het	yes	intron_variant	0	0	0					2
G>G/A	42525756	snv	het	yes	synonymous_variant	426	336	112	F	ttC/ttT			4
T>T/G	42525793	snv	het	yes	missense_variant	389	299	100	D/A	gAc/gCc	deleterious(0)	probably_damaging(1)	4
G>G/C	42525798	snv	het	yes	synonymous_variant	384	294	98	T	acC/acG			16
G>C/C	42525798	snv	hom	yes	synonymous_variant	384	294	98	T	acC/acG			6

T>T/C	42525811	snv	het	yes	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	17
T>C/C	42525811	snv	hom	yes	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	6
G>G/T	42525821	snv	het	yes	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	15
G>T/T	42525821	snv	hom	yes	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	5
C>C/T	42525840	snv	het	yes	synonymous_variant	342	252	84	L	ctG/ctA			1
C>C/A	42525952	snv	het	no	intron_variant	0	0	0					27
C>A/A	42525952	snv	hom	no	intron_variant	0	0	0					8
C>C/G	42526049	snv	het	no	intron_variant	0	0	0					24
C>G/G	42526049	snv	hom	no	intron_variant	0	0	0					13
C>C/G	42526055	snv	het	no	intron_variant	0	0	0					1
A>A/C	42526484	snv	het	no	intron_variant	0	0	0					24
A>C/C	42526484	snv	hom	no	intron_variant	0	0	0					11
G>G/A	42526524	snv	het	no	intron_variant	0	0	0					2
C>T/T	42526549	snv	hom	no	intron_variant	0	0	0					27
C>C/T	42526549	snv	het	no	intron_variant	0	0	0					25
G>T/T	42526561	snv	hom	no	intron_variant	0	0	0					26
G>G/T	42526561	snv	het	no	intron_variant	0	0	0					26

G>C/C	42526562	snv	hom	no	intron_variant	0	0	0					26
G>G/C	42526562	snv	het	no	intron_variant	0	0	0					26
G>A/A	42526567	snv	hom	no	intron_variant	0	0	0					27
G>G/A	42526567	snv	het	no	intron_variant	0	0	0					25
C>G/G	42526571	snv	hom	no	intron_variant	0	0	0					27
C>C/G	42526571	snv	het	no	intron_variant	0	0	0					25
T>G/G	42526573	snv	hom	no	intron_variant	0	0	0					27
T>T/G	42526573	snv	het	no	intron_variant	0	0	0					25
G>C/C	42526580	snv	hom	no	intron_variant	0	0	0					27
G>G/C	42526580	snv	het	no	intron_variant	0	0	0					25
C>C/T	42526669	snv	het	yes	missense_variant	215	125	42	G/E	gGg/gAg	deleterious(0.02)	probably_damaging(0.963)	1
G>G/A	42526694	snv	het	yes	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20
G>A/A	42526694	snv	hom	yes	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	7
C>C/T	42526763	snv	het	yes	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	8
A>A/AC	42526836	insertion	het	yes	5_prime_UTR_variant, feature_elongation		0	0					1

Supplemental Tables:

Supplementary Table 1: SNPs within sequencing range of *CYP2D6* for Sample 0001

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	48.8	2331	11,931,136	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	99.5	1736	91,725	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.8	2060	42,049	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	59.7	2048	8,241,222	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	30.9	1822	1,259,562	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	20.2	1979	1,577,398	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	11.8	1398	1,231,165	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	99.8	2552	42,545	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs28371717	C>C/A	42524310	snv	het	yes	48.3	1883	972,909	missense_variant	799	709	237	A/S	Gct/Tct	tolerated(0.34)	benign(0.11)	1
rs58440431	T>T/C	42524696	snv	het	no	38.8	2605	15,941,009	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	17.2	2597	2,143,446	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	17.9	2584	2,117,462	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	23.5	2581	1,972,605	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	25.4	2706	2,016,685	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	14.3	2113	1,807,301	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	41.4	1974	1,155,815	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	50.2	1788	889,896	synonymous_variant	498	408	136	V	gtC/gtG			44
rs28371705	G>G/C	42525798	snv	het	yes	33.5	1099	730,367	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	33.6	1071	702,356	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	31.6	1064	720,332	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs71328650	C>C/A	42525952	snv	het	no	50.1	1752	872,875	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	53	1672	784,885	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	48	1777	923,853	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	98	1214	241,189	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	96.3	996	37,958	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	96	976	39,937	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	95.8	959	40,916	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	95.4	942	43,898	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	95.9	918	38,879	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	96.1	955	37,915	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	48.8	1720	879,838	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 2: SNPs within sequencing range of *CYP2D6* for Sample 0004

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	48.7	2008	1,029,976	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	37.6	1743	1,084,654	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	47.3	1804	950,851	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	39.6	1690	1,019,668	intron_variant	0	0	0					33
rs28371725	C>C/T	42523805	snv	het	no	51.1	1632	797,833	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	49.8	2127	10,641,055	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs5030656	CCTT>CCTT/C	42524175	deletion	het	yes	40.7	1566	942,646	inframe_deletion, splice_region_variant			281	K/-	AAG/-			2
rs111564371	T>T/C	42524708	snv	het	no	18.6	2404	1,953,447	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	19	2395	1,938,454	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	23.8	2454	1,868,584	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	24.5	2470	1,861,605	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	14.3	1927	1,645,274	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	47.3	1485	782,702	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	47.7	1952	1,020,930	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	48.5	1870	961,906	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	45.7	1708	927,781	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	37.9	1491	925,564	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	34.3	1349	886,462	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	34	1335	880,453	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	34	1308	863,444	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	33	1281	854,421	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	33.9	1276	844,432	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	36.9	1273	802,468	intron_variant	0	0	0					0

Supplementary Table 3: SNPs within sequencing range of *CYP2D6* for Sample 0005

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	50.1	1211	604,607	intron_variant	0	0	0					9
rs150518553	A>A/G	42524323	snv	het	yes	47.1	1186	626,557	synonymous_variant	786	696	232	H	caT/caC			0

Supplementary Table 4: SNPs within sequencing range of *CYP2D6* for Sample 0047

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	19.4	928	748,180	3_prime_UTR_variant	1610	0	0					0
rs1135840	G>G/C	42522613	snv	het	yes	45.4	895	488,405	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	99.1	959	9,946	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.6	1039	41,031	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	68.1	1037	330,704	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	22.5	1093	846,246	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	16.1	1200	1,004,192	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs16947	A>G/G	42523943	snv	hom	yes	99.5	1390	71,381	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs74516776	G>G/A	42524090	snv	het	no	34.5	1156	755,398	intron_variant	0	0	0					0
rs58440431	T>T/C	42524696	snv	het	no	57	1478	634,841	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	13.9	1455	1,252,202	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	14.2	1457	1,247,207	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	18.1	1418	1,161,257	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	17.5	1501	1,238,263	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	11	1110	981,121	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	59.5	1069	433,635	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	31.8	1025	698,326	synonymous_variant	498	408	136	V	gtC/gtG			44
rs28371705	G>G/C	42525798	snv	het	yes	57.2	519	221,295	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	55.8	505	222,280	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	53	481	223,251	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs71328650	C>C/A	42525952	snv	het	no	35.9	1004	642,359	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	40.8	880	520,359	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	33.7	1039	688,349	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	97.5	764	19,745	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	95.8	642	27,614	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	95.7	635	27,608	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	95.6	662	29,632	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	95.7	654	28,626	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	95.8	638	27,611	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	95.9	662	27,632	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	69.3	1103	338,764	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 5: SNPs within sequencing range of *CYP2D6* for Sample 0298

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs116917064	A>A/G	42523003	snv	het	yes	52.4	307	146,161	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	60.4	281	111,169	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	59.2	282	115,167	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	22.1	336	261,740	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	11.1	244	217,270	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>A/G	42523943	snv	het	yes	50.7	375	185,190	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
	C>C/T	42524138	snv	het	no	49.5	221	111,109	intron_variant	0	0	0					0
rs58440431	T>T/C	42524696	snv	het	no	51.8	303	146,157	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	46.1	229	123,105	splice_acceptor_variant	0	0	0					19
rs28371705	G>G/C	42525798	snv	het	yes	40	120	72,480	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	42.3	149	86,630	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	41	136	79,550	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>C/T	42526549	snv	het	no	35.3	255	165,900	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	33	233	156,770	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	32.3	229	155,740	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	33	231	154,760	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	32.9	226	151,740	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	33	218	146,720	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	34.4	209	137,720	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	50	270	135,135	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 6: SNPs within sequencing range of *CYP2D6* for Sample 0412

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	52.3	560	267,293	intron_variant	0	0	0					9

Supplementary Table 7: SNPs within sequencing range of *CYP2D6* for Sample 0899

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	47.9	2600	13,531,243	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	42.5	2325	1,334,988	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	46.4	2117	1,132,981	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	38.9	1742	1,063,676	intron_variant	0	0	0					33
rs28371725	C>C/T	42523805	snv	het	no	48.2	2085	10,791,004	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	50.2	2736	13,601,371	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	16.5	2858	2,381,470	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	17.7	2880	2,367,509	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	22.8	2848	2,198,649	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	25.3	2863	2,136,725	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	13.2	2195	1,899,289	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	49.5	2023	10,221,000	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	47.7	2042	1,065,970	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	48.5	1884	970,912	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	46.3	1977	1,060,914	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	39.6	1751	1,057,694	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	35.3	1584	1,024,559	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	35.3	1565	1,013,552	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	35.9	1531	980,550	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	34.5	1472	964,507	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	35.7	1462	940,521	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	37.7	1469	914,553	intron_variant	0	0	0					0

Supplementary Table 8: SNPs within sequencing range of *CYP2D6* for Sample 2931

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	47	268	142,126	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	36.1	227	145,820	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	49.8	239	120,119	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	37.3	213	133,790	intron_variant	0	0	0					33
rs28371725	C>C/T	42523805	snv	het	no	45.3	279	152,126	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	51	300	147,153	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
	T>T/A	42524054	snv	het	no	9.3	268	243,250	intron_variant	0	0	0					0
	C>C/A	42524057	snv	het	no	9.5	265	239,250	intron_variant	0	0	0					0
	A>A/T	42524058	snv	het	no	9.5	264	239,250	intron_variant	0	0	0					0
rs111564371	T>T/C	42524708	snv	het	no	23.9	277	210,660	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	23.5	285	218,670	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	28.3	280	200,790	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	33.4	312	207,104	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	16.7	201	165,330	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	48.6	175	90,850	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	46.2	197	106,910	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	42.9	196	112,840	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	51.2	215	105,110	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	41.8	194	113,810	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	40.2	174	104,700	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	39.9	173	104,690	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	41.4	174	102,720	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	36.9	157	99,580	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	38.7	155	95,600	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	40.4	151	90,610	intron_variant	0	0	0					0

Supplementary Table 9: SNPs within sequencing range of *CYP2D6* for Sample 3250

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	97.8	555	12,543	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	97.9	343	7,334	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99	575	6,569	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	63.2	653	240,412	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	98.3	639	11,626	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
	A>A/C	42524471	snv	het	no	45	591	325,266	intron_variant	0	0	0					0
rs111564371	T>T/C	42524708	snv	het	no	33	888	593,292	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	33.8	911	602,308	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	42.5	938	538,397	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	44.2	967	538,427	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	27.7	645	465,178	synonymous_variant	618	528	176	G	ggT/ggC			0
rs200720666	C>C/T	42524975	snv	het	no	19.2	559	451,107	intron_variant	0	0	0					0
rs113678157	C>C/T	42524982	snv	het	no	17.9	549	449,980	intron_variant	0	0	0					0
rs1058164	G>C/C	42525132	snv	hom	yes	98.1	536	10,525	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	50.6	411	203,208	intron_variant	0	0	0					6
rs71328650	C>A/A	42525952	snv	hom	no	99	588	6,580	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	99.3	608	4,604	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	98.6	441	6,435	intron_variant	0	0	0					45
rs29001678	G>G/A	42526524	snv	het	no	45.7	396	214,180	intron_variant	0	0	0					2
rs56011157	C>T/T	42526549	snv	hom	no	97.1	339	10,329	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	95.5	291	13,277	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	95	283	14,268	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	94.8	308	16,292	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	95	282	14,267	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	94.7	286	15,270	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	95.5	313	14,299	intron_variant	0	0	0					0

Supplementary Table 10: SNPs within sequencing range of *CYP2D6* for Sample 4275

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	49.7	2505	12,601,243	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	30.7	1962	1,356,602	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	45.7	1930	1,043,877	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	40.5	1178	700,477	intron_variant	0	0	0					33
rs16947	A>A/G	42523943	snv	het	yes	49.4	2724	13,761,344	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs28371717	C>C/A	42524310	snv	het	yes	49.1	1880	955,920	missense_variant	799	709	237	A/S	Gct/Tct	tolerated(0.34)	benign(0.11)	1
rs111564371	T>T/C	42524708	snv	het	no	19.2	2675	2,161,512	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	20	2654	2,121,530	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	24.8	2653	1,993,658	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	24.7	2846	2,139,703	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	16.1	1992	1,665,320	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	48.5	1716	883,830	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	45.9	1669	899,764	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	48.3	1654	853,797	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	46.7	1594	850,744	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	38.2	1421	878,542	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	35.7	1290	830,460	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	35	1276	827,446	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	36.1	1274	813,459	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	34.9	1222	793,426	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	36.2	1204	767,436	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	38.4	1186	729,454	intron_variant	0	0	0					0

Supplementary Table 11: SNPs within sequencing range of *CYP2D6* for Sample 8868

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	25.7	305	226,780	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>A/G	42523003	snv	het	yes	54.3	397	180,214	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	70.4	325	95,226	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	69.2	322	99,222	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	23.1	374	287,860	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	12.4	242	212,300	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>A/G	42523943	snv	het	yes	68.4	364	114,247	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	61.3	313	120,190	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	67.3	246	80,165	splice_acceptor_variant	0	0	0					19
rs28371705	G>G/C	42525798	snv	het	yes	50.4	141	70,710	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	54.9	177	79,960	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	50.5	189	93,950	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>C/T	42526549	snv	het	no	65.3	291	101,190	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	64.2	274	98,176	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	63.8	273	98,173	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	65.5	275	95,180	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	66	268	91,177	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	58.9	203	83,119	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	58.7	206	85,121	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	69	314	97,216	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 12: SNPs within sequencing range of *CYP2D6* for Sample 12349

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	32.4	136	92,440	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	33.3	93	62,310	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	48.1	104	54,500	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	35.2	122	79,430	intron_variant	0	0	0					33
rs16947	A>A/G	42523943	snv	het	yes	50.4	139	69,700	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs5030656	CCTT>CCTT/C	42524175	deletion	het	yes	57.9	94	40,550	inframe_deletion, splice_region_variant			281	K/-	AAG/-			2
rs111564371	T>T/C	42524708	snv	het	no	16.6	145	121,240	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	16.4	146	122,240	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	19	137	111,260	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	18.1	149	122,270	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	18.9	97	77,180	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	47.8	90	47,430	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	52.7	110	52,580	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	50	121	60,600	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	53.3	107	50,570	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	32.9	82	55,270	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	34.9	83	54,290	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	34.1	82	54,280	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	25	97	72,240	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	23.1	91	70,210	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	23.1	91	70,210	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	19.5	87	70,170	intron_variant	0	0	0					0
rs769258	C>C/T	42526763	snv	het	yes	53.8	107	49,570	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 13: SNPs within sequencing range of *CYP2D6* for Sample 12812

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	47.4	213	112,101	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	37.6	166	103,620	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	45.9	187	98,830	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	39.9	178	107,710	intron_variant	0	0	0					33
rs28371725	C>C/T	42523805	snv	het	no	47.4	212	111,100	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	45.9	279	151,128	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	21.6	269	210,580	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	21.4	277	217,590	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	25.5	278	207,710	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	27	244	178,660	synonymous_variant	747	657	219	F	ttT/ttC			0
rs1058164	G>G/C	42525132	snv	het	yes	45.9	157	85,720	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	43.6	94	53,410	intron_variant	0	0	0					6
rs71328650	C>C/A	42525952	snv	het	no	44.4	171	95,760	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	51.5	165	80,850	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	49.1	171	87,840	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	39.6	154	93,610	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	37.2	145	91,540	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	36.1	144	92,520	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	35	140	91,490	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	28.6	133	95,380	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	30.6	124	86,380	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	33.6	126	83,420	intron_variant	0	0	0					0

Supplementary Table 14: SNPs within sequencing range of *CYP2D6* for Sample 13242

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	55.9	171	75,950	intron_variant	0	0	0					9

Supplementary Table 15: SNPs within sequencing range of *CYP2D6* for Sample 13510

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs116917064	A>A/G	42523003	snv	het	yes	45.2	1660	910,750	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	56.1	1685	738,943	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	56.6	1676	725,947	intron_variant	0	0	0					20
rs28371729	G>G/T	42523358	snv	het	no	34.1	1269	835,433	intron_variant	0	0	0					0
rs1058172	C>C/T	42523528	snv	het	yes	23.6	1548	1,180,365	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs28371725	C>C/T	42523805	snv	het	no	48	1300	673,621	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	50.2	1882	936,942	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	48.1	1589	825,764	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	46.3	1291	692,596	splice_acceptor_variant	0	0	0					19
rs28371705	G>G/C	42525798	snv	het	yes	33.4	913	606,304	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	34	953	620,320	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	32.6	938	623,302	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>C/T	42526549	snv	het	no	39.1	1126	686,440	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	36.1	1045	668,377	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	35.9	1039	665,372	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	36.9	1023	644,377	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	37.8	1014	631,383	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	37.3	975	611,364	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	38.3	963	594,368	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	50.6	1174	579,593	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 16: SNPs within sequencing range of *CYP2D6* for Sample 15230

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	49.6	231	116,114	intron_variant	0	0	0					9
rs769258	C>C/T	42526763	snv	het	yes	44.7	188	104,840	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 17: SNPs within sequencing range of *CYP2D6* for Sample 15449

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs116917064	A>A/G	42523003	snv	het	yes	52.7	184	86,960	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	41.1	191	112,780	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	40.8	191	113,780	intron_variant	0	0	0					20
rs28371729	G>G/T	42523358	snv	het	no	33.6	131	87,440	intron_variant	0	0	0					0
rs1058172	C>C/T	42523528	snv	het	yes	24.9	229	172,570	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	22	129	99,280	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>A/G	42523943	snv	het	yes	59.5	259	105,154	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	43.8	169	95,740	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	40.5	149	88,600	splice_acceptor_variant	0	0	0					19
rs200002499	T>T/G	42525010	snv	het	no	39.5	129	78,510	intron_variant	0	0	0					0
rs28371705	G>G/C	42525798	snv	het	yes	27.6	77	55,210	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	29.9	98	68,290	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs56011157	C>C/T	42526549	snv	het	no	40.6	129	76,520	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	40.2	122	73,490	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	39.8	123	74,490	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	41.3	126	74,520	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	41.9	124	72,520	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	42.3	123	71,520	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	42.9	119	68,510	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	49	157	80,770	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20
rs769258	C>C/T	42526763	snv	het	yes	53.4	178	83,950	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 18: SNPs within sequencing range of *CYP2D6* for Sample 15637

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	51.4	74	36,380	intron_variant	0	0	0					9

Supplementary Table 19: SNPs within sequencing range of *CYP2D6* for Sample 18139

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	92.5	80	6,740	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	94.1	34	2,320	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	98.4	61	1,600	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	56.2	73	32,410	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	98.7	76	1,750	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	44.6	101	56,450	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	43.3	106	59,450	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	44.6	102	56,450	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	37.8	82	51,310	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	24.1	55	41,130	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>C/C	42525132	snv	hom	yes	95.5	44	2,420	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	60.5	38	15,230	intron_variant	0	0	0					6
rs71328650	C>A/A	42525952	snv	hom	no	91.7	73	6,660	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	95	80	4,760	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	100	71	1,710	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	96.2	52	2,500	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	93.9	49	3,460	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	93.8	48	3,450	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	93	43	3,400	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	92.5	40	3,370	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	92.7	41	3,380	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	90	40	4,360	intron_variant	0	0	0					0
rs118203758	C>C/T	42526669	snv	het	yes	51	51	25,260	missense_variant	215	125	42	G/E	gGg/gAg	deleterious(0.02)	probably_damaging(0.963)	0

Supplementary Table 20: SNPs within sequencing range of *CYP2D6* for Sample 19055

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	37.1	70	44260	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>G/G	42523003	snv	hom	yes	97.7	44	1430	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	97.6	41	1400	intron_variant	0	0	0					65
rs2004511	T>C/C	42523211	snv	hom	no	97.6	41	1400	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	70	40	12280	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	30	40	28120	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>A/G	42523943	snv	het	yes	92	26	2230	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>C/C	42524696	snv	hom	no	100	12	1110	intron_variant	0	0	0					20
rs3892097	C>T/T	42524947	snv	hom	yes	37.3	59	37220	splice_acceptor_variant	0	0	0					19
rs28371705	G>C/C	42525798	snv	hom	yes	100	9	1900	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>C/C	42525811	snv	hom	yes	100	11	1110	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>T/T	42525821	snv	hom	yes	100	15	1150	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>T/T	42526549	snv	hom	no	97.6	42	1410	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	97.3	37	1360	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	97.3	37	1360	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	97.3	37	1360	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	97.1	34	1330	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	97.1	34	1330	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	96.9	32	1310	intron_variant	0	0	0					0
rs1065852	G>A/A	42526694	snv	hom	yes	96.7	61	2590	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 21: SNPs within sequencing range of *CYP2D6* for Sample 19256

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	21.8	1942	1,517,424	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>G/G	42523003	snv	hom	yes	99.5	1709	81,689	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.7	1890	61,875	intron_variant	0	0	0					65
rs2004511	T>C/C	42523211	snv	hom	no	99.8	1883	31,880	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	37.3	2232	1,396,831	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	18.8	1197	970,225	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	99.7	1886	61,875	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>C/C	42524696	snv	hom	no	99.3	1628	121,615	intron_variant	0	0	0					20
rs3892097	C>T/T	42524947	snv	hom	yes	88.4	1227	1,421,082	splice_acceptor_variant	0	0	0					19
	T>T/G	42525793	snv	het	yes	22.7	573	440,129	missense_variant	389	299	100	D/A	gAc/gCc	deleterious(0)	probably_damaging(1)	0
rs28371705	G>C/C	42525798	snv	hom	yes	99.2	603	5,597	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>C/C	42525811	snv	hom	yes	99.7	732	2,726	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>T/T	42525821	snv	hom	yes	98.8	688	8,663	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>T/T	42526549	snv	hom	no	97.2	929	26,903	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	95.4	829	38,791	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	95.2	820	39,781	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	95.3	826	39,786	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	95	806	40,765	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	94.8	733	38,693	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	95.2	771	37,729	intron_variant	0	0	0					0
rs1065852	G>A/A	42526694	snv	hom	yes	98.7	1119	141,102	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 22: SNPs within sequencing range of *CYP2D6* for Sample 19872

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	16.9	1873	1,555,316	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>A/G	42523003	snv	het	yes	56.8	2172	9,341,228	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	64.4	2280	8,091,464	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	64.2	2252	8,051,445	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	34.5	2016	1,318,693	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	18.4	1187	968,219	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>A/G	42523943	snv	het	yes	49.5	1848	931,913	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	47.8	1646	859,786	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	42.4	1414	813,598	splice_acceptor_variant	0	0	0					19
rs28371705	G>G/C	42525798	snv	het	yes	32.4	902	610,292	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	33.3	1009	670,334	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	32.2	1008	679,323	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>C/T	42526549	snv	het	no	55	1542	694,848	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	52	1407	675,731	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	51.9	1395	670,724	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	52.1	1338	641,696	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	53.1	1351	633,716	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	52.3	1292	615,675	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	53.2	1292	604,686	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	65.5	1786	6,141,167	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20
rs769258	C>C/T	42526763	snv	het	yes	33.1	1979	1,322,655	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 23: SNPs within sequencing range of *CYP2D6* for Sample 20394

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
	A>A/G	42522768	snv	het	yes	4.5	66	63,300	intron_variant	0	0	0					0
	T>T/C	42523315	snv	het	no	23.3	43	33,100	intron_variant	0	0	0					0
rs28371725	C>T/T	42523805	snv	hom	no	91.9	62	5,570	intron_variant	0	0	0					9
rs111564371	T>T/C	42524708	snv	het	no	36.8	89	55,320	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	42.1	96	55,400	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	36.4	99	63,360	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	27.8	72	52,200	synonymous_variant	618	528	176	G	ggT/ggC			0

Supplementary Table 24: SNPs within sequencing range of *CYP2D6* for Sample 20540

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	47.4	441	231,208	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	99	297	3,291	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.5	388	2,386	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	58.4	389	162,227	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	28.4	381	272,108	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	17.2	440	361,750	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	12.9	294	256,380	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	98.7	388	5,382	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs35742686	CT>CT/C	42524243	deletion	het	yes	44.5	368	206,165	frameshift_variant, feature_truncation	865	775	259					2
rs58440431	T>T/C	42524696	snv	het	no	41.4	440	258,182	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	17.3	440	364,760	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	19.2	444	357,850	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	23.3	433	332,101	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	21.8	395	308,860	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	13.6	274	235,370	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	47.6	292	152,138	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	43.4	282	159,122	synonymous_variant	498	408	136	V	gtC/gtG			44
rs28371705	G>G/C	42525798	snv	het	yes	37.8	148	92,560	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	39.6	153	90,590	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	34	157	101,520	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs71328650	C>C/A	42525952	snv	het	no	45.7	304	165,139	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	47.8	312	163,149	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	52.2	323	154,168	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	98.8	244	3,240	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	97.2	213	6,206	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	97.1	207	6,200	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	97.2	212	6,206	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	97.1	208	6,201	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	97.6	207	5,202	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	97.8	225	5,220	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	49.6	343	173,170	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 25: SNPs within sequencing range of *CYP2D6* for Sample 22179

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	36.2	265	169,960	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>A/G	42523003	snv	het	yes	64.6	207	73,133	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	64.2	228	81,145	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	63.5	222	81,141	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	38.8	209	128,810	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	25	112	84,280	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>A/G	42523943	snv	het	yes	48.9	190	97,930	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	27.9	179	129,500	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	34.7	122	79,420	splice_acceptor_variant	0	0	0					19
rs28371704	T>T/C	42525811	snv	het	yes	34	98	62,320	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs56011157	C>C/T	42526549	snv	het	no	56.6	145	63,820	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	50.4	127	63,640	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	50.4	125	62,630	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	52.7	131	62,690	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	53.1	130	61,690	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	52.5	122	58,640	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	54.8	126	57,690	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	72.1	190	53,137	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 26: SNPs within sequencing range of *CYP2D6* for Sample 22445

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	93.5	108	7,101	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	97.6	85	280	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	97.7	88	285	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	60	90	3654	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	14.1	85	7312	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs16947	A>G/G	42523943	snv	hom	yes	96.8	93	390	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
	A>A/C	42524408	snv	het	no	16.7	54	459	intron_variant	0	0	0					0
rs1807313	T>T/A	42524435	snv	het	no	10.5	57	516	intron_variant	0	0	0					0
rs111564371	T>T/C	42524708	snv	het	no	34.6	105	6836	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	35.9	103	6637	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	44.8	120	6452	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	34.8	135	8847	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	23	87	6720	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	88.9	45	540	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>A/A	42525952	snv	hom	no	87.8	98	1286	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	98.9	90	189	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	96.9	64	262	intron_variant	0	0	0					45
rs29001678	G>G/A	42526524	snv	het	no	54.8	42	1923	intron_variant	0	0	0					2
rs56011157	C>T/T	42526549	snv	hom	no	93.9	33	231	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	93.1	29	227	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	92.6	27	225	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	95.7	23	122	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	94.4	18	117	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	100	18	18	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	100	19	19	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	12.3	57	507	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 27: SNPs within sequencing range of *CYP2D6* for Sample 22668

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	28.7	195	13956	3_prime_UTR_variant	1610	0	0					0
rs1135840	G>G/C	42522613	snv	het	yes	35.5	183	11865	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	97.2	176	5,171	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	100	208	207	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	71.5	207	59,148	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	23.6	242	18557	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	36.8	261	16596	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	26.5	151	11140	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	98.9	189	2,187	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	43.9	199	11187	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	20.5	190	15139	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	21.6	185	14540	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	27.3	198	14454	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	27	176	12747	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	11.1	189	16821	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	38.8	179	10969	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	48.3	120	6258	synonymous_variant	498	408	136	V	gtC/gtG			44
rs28371705	G>G/C	42525798	snv	het	yes	36.6	71	4526	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	32.9	85	5728	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	31	87	6027	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs71328650	C>C/A	42525952	snv	het	no	44.8	203	11291	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	37.1	205	12976	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	25	268	20167	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	98.2	219	4,215	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	97.8	178	4,174	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	97.1	171	5,166	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	95.4	130	6,124	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	96.2	130	5,125	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	96.8	126	4,122	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	96.6	119	4,115	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	57.6	236	100,136	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 28: SNPs within sequencing range of *CYP2D6* for Sample 22680

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	49.4	1354	684,667	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	35	1078	699,376	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	47.3	1252	657,590	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	39.5	1033	624,407	intron_variant	0	0	0					33
rs16947	A>A/G	42523943	snv	het	yes	48.5	1446	744,700	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	19	1673	1,353,318	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	19.4	1672	1,346,324	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	24.1	1611	1,223,388	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	26.1	1672	1,235,436	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	13.4	1213	1,044,162	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	43.9	1081	605,474	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	49.9	1001	501,499	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	48.9	1053	538,514	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	46.5	976	521,453	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	39.8	890	536,354	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	36.5	798	507,291	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	36.2	793	506,287	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	36.6	780	494,285	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	33.7	727	482,245	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	35.8	717	460,257	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	37.4	732	456,273	intron_variant	0	0	0					0
rs769258	C>C/T	42526763	snv	het	yes	47.3	1056	556,500	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 29: SNPs within sequencing range of *CYP2D6* for Sample 22842

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1058172	C>C/T	42523528	snv	het	yes	19.3	171	13833	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14

Supplementary Table 30: SNPs within sequencing range of *CYP2D6* for Sample 23033

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	37.9	240	14991	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs1135836	A>A/C	42522660	snv	het	yes	17.4	249	20443	synonymous_variant	1500	1410	470	T	acT/acG			0
	G>G/A	42522774	snv	het	yes	17.9	196	16035	intron_variant	0	0	0					0
rs116917064	A>G/G	42523003	snv	hom	yes	96.9	193	6,185	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	98.2	173	3,167	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	59.4	187	76,111	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	25.7	218	16256	intron_variant	0	0	0					33
rs3915951	C>C/A	42523636	snv	het	yes	15	134	11320	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	96.8	186	6,179	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	38.6	207	12780	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	19.9	201	16140	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	20.9	196	15541	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	24.6	187	14146	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	24	204	15549	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	21	159	12433	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	49.6	115	5857	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	52.6	95	4550	intron_variant	0	0	0					6
rs1081003	G>G/A	42525756	snv	het	yes	44.6	121	6754	synonymous_variant	426	336	112	F	ttC/ttT			2
rs71328650	C>C/A	42525952	snv	het	no	47.6	164	8678	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	78.1	194	42,150	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	50.3	144	7172	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	95.6	114	5,108	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	92.6	94	787	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	91.1	91	882	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	90.1	101	1091	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	91.9	99	891	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	92.7	96	789	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	92	100	892	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	48.3	147	7671	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 31: SNPs within sequencing range of *CYP2D6* for Sample 23619

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
	A>A/G	42522768	snv	het	yes	9.3	161	14615	intron_variant	0	0	0					0
rs28371725	C>C/T	42523805	snv	het	no	44.2	120	6753	intron_variant	0	0	0					9

Supplementary Table 32: SNPs within sequencing range of *CYP2D6* for Sample 25425

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	50.2	1075	535,540	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	32.4	914	616,295	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	46.6	906	481,420	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	38.5	843	517,324	intron_variant	0	0	0					33
rs78209835	C>C/T	42523613	snv	het	yes	49	774	395,379	missense_variant	1099	1009	337	D/N	Gac/Aac	tolerated(0.21)	benign(0.003)	0.4
rs16947	A>A/G	42523943	snv	het	yes	49.4	1165	589,575	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	20.4	1278	1,017,260	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	21.6	1282	1,004,276	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	26.5	1272	935,337	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	25.2	1352	1,010,341	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	14.7	1027	873,150	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	50.1	912	455,457	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	50.5	952	468,478	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	50	945	472,472	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	46	895	483,412	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	35.5	737	475,262	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	31.5	665	454,209	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	30.5	660	458,201	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	32.7	657	442,215	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	30.3	621	432,188	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	31.4	618	424,194	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	32.5	615	415,200	intron_variant	0	0	0					0

Supplementary Table 33: SNPs within sequencing range of *CYP2D6* for Sample 25802

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	27.4	169	122,460	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>G/G	42523003	snv	hom	yes	99.2	129	1,128	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	100	189	1,186	intron_variant	0	0	0					65
rs2004511	T>C/C	42523211	snv	hom	no	99.5	188	1,187	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	51.8	168	81,870	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	24.8	121	91,300	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	100	127	1,127	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>C/C	42524696	snv	hom	no	100	96	1,960	intron_variant	0	0	0					20
rs3892097	C>T/T	42524947	snv	hom	yes	84.5	84	13,710	splice_acceptor_variant	0	0	0					19
rs28371705	G>C/C	42525798	snv	hom	yes	100	27	1,270	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>C/C	42525811	snv	hom	yes	100	29	1,290	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>T/T	42525821	snv	hom	yes	100	30	1,290	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>T/T	42526549	snv	hom	no	94.2	139	8,131	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	90.8	119	11,108	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	90.3	113	11,102	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	89.2	102	11,910	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	90	100	10,900	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	89.7	98	10,870	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	90.2	103	10,920	intron_variant	0	0	0					0
rs1065852	G>A/A	42526694	snv	hom	yes	99.3	150	1,149	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 34: SNPs within sequencing range of *CYP2D6* for Sample 25861

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	100	52	52	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	72.2	37	1026	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	97.2	36	135	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	66.7	45	1530	intron_variant	0	0	0					33
rs16947	A>A/G	42523943	snv	het	yes	86	50	743	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	29.7	64	4519	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	29.5	61	4318	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	31.5	73	5023	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	30.1	83	5825	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	16.2	76	6212	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	90.2	41	437	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	66.1	59	2039	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	73.3	75	2055	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	95.7	47	245	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	97.4	38	137	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	94.1	34	232	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	93.9	33	231	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	88.2	34	430	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	90	30	327	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	93.3	30	228	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	93.5	31	229	intron_variant	0	0	0					0

Supplementary Table 35: SNPs within sequencing range of *CYP2D6* for Sample 31450

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	17	338	279,570	3_prime_UTR_variant	1610	0	0					0
rs1135840	G>G/C	42522613	snv	het	yes	44.9	327	179,146	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	97.2	250	7,241	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	98.7	309	4,304	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	56.7	308	133,174	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	29.1	347	246,101	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	32.1	365	247,117	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	15.2	250	212,380	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	98.4	385	6,376	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	38.4	331	204,127	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	21.6	333	261,720	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	21.6	335	262,720	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	28.1	335	241,940	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	25.8	353	262,910	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	17.3	249	205,430	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	47	248	131,116	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	46.1	182	97,830	synonymous_variant	498	408	136	V	gtC/gtG			44
rs28371705	G>G/C	42525798	snv	het	yes	39.3	122	74,480	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	43.4	129	73,560	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	40.9	134	78,540	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs71328650	C>C/A	42525952	snv	het	no	49.5	285	144,141	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	46.8	284	151,133	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	47.3	260	137,123	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	94.8	191	10,181	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	91.2	170	15,155	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	91.1	169	15,153	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	91.1	168	15,153	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	90.7	161	15,146	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	90.6	159	15,144	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	91.6	154	13,141	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	51.5	234	113,120	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 36: SNPs within sequencing range of *CYP2D6* for Sample 33439

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	48.1	136	7065	intron_variant	0	0	0					9
rs61736507	G>G/A	42525180	snv	het	yes	51.3	80	3941	synonymous_variant	450	360	120	F	ttC/ttT			0.26
rs143170489	G>G/C	42525416	snv	het	no	52.9	52	2427	intron_variant	0	0	0					0.26

Supplementary Table 37: SNPs within sequencing range of *CYP2D6* for Sample 33556

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	46.2	286	154,132	intron_variant	0	0	0					9
rs61736507	G>G/A	42525180	snv	het	yes	55.3	161	72,890	synonymous_variant	450	360	120	F	ttC/ttT			0.26
rs143170489	G>G/C	42525416	snv	het	no	38.1	113	70,430	intron_variant	0	0	0					0.26

Supplementary Table 38: SNPs within sequencing range of *CYP2D6* for Sample 35150

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	47.8	1490	777,712	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	36.3	1226	780,445	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	46.9	1257	663,586	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	38.8	1014	621,393	intron_variant	0	0	0					33
rs28371725	C>C/T	42523805	snv	het	no	48.7	1193	612,580	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	49.1	1641	834,804	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	17.2	1576	1,303,271	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	18.5	1613	1,310,298	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	23	1633	1,252,375	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	23.7	1725	1,311,407	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	13.3	1325	1,145,176	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	44.6	1183	655,527	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	51	723	354,369	intron_variant	0	0	0					6
rs71328650	C>C/A	42525952	snv	het	no	47.3	1208	636,571	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	46.8	1207	642,564	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	45.8	1082	586,496	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	36.3	994	633,360	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	34.1	912	601,311	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	33.7	906	600,305	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	34.4	914	599,314	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	34	874	577,297	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	34.1	880	577,299	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	35.7	883	568,315	intron_variant	0	0	0					0

Supplementary Table 39: SNPs within sequencing range of *CYP2D6* for Sample 44235

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs116917064	A>G/G	42523003	snv	hom	yes	98.6	290	4,283	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99	288	3,285	intron_variant	0	0	0					65
rs2004511	T>C/C	42523211	snv	hom	no	98.9	285	3,282	intron_variant	0	0	0					20
rs28578778	A>A/G	42523400	snv	het	no	21	286	22660	intron_variant	0	0	0					1
rs1058172	C>C/T	42523528	snv	het	yes	14.7	369	31454	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	19.8	212	17042	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	98.4	312	5,307	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>C/C	42524696	snv	hom	no	98.6	211	3,208	intron_variant	0	0	0					20
rs3892097	C>T/T	42524947	snv	hom	yes	87.2	211	27,184	splice_acceptor_variant	0	0	0					19
rs1081003	G>G/A	42525756	snv	het	yes	37.6	141	8853	synonymous_variant	426	336	112	F	ttC/ttT			2
rs28371705	G>G/C	42525798	snv	het	yes	37.3	102	6438	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	41	134	7955	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	39.8	131	7751	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs147296446	C>C/G	42526049	snv	het	no	21.2	203	16043	intron_variant	0	0	0					0
rs56011157	C>T/T	42526549	snv	hom	no	94.4	142	8,134	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	94.5	128	7,121	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	94.5	127	7,120	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	94.7	131	7,124	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	95.4	130	6,124	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	94.8	116	6,110	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	95.9	124	5,118	intron_variant	0	0	0					0
rs1065852	G>A/A	42526694	snv	hom	yes	97.7	222	5,216	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 40: SNPs within sequencing range of *CYP2D6* for Sample 45389

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	96.2	105	4,101	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	92.9	84	6,780	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	98.9	88	1,860	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	51	98	48,500	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	99.2	126	1,125	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	39.6	112	67,440	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	41.2	114	67,470	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	45.5	123	67,560	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	35.1	95	61,330	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	31	85	58,260	synonymous_variant	618	528	176	G	ggT/ggC			0
rs200720666	C>C/T	42524975	snv	het	no	27.8	79	57,220	intron_variant	0	0	0					0
rs113678157	C>C/T	42524982	snv	het	no	20.8	72	57,150	intron_variant	0	0	0					0
rs1058164	G>C/C	42525132	snv	hom	yes	100	51	1,510	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	63	46	17,290	intron_variant	0	0	0					6
rs71328650	C>C/A	42525952	snv	het	no	91	67	6,610	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	96.6	89	3,860	intron_variant	0	0	0					0
	C>C/G	42526055	snv	het	no	5.5	91	86,500	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	95.5	67	3,640	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	91.3	46	4,420	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	87.2	39	5,340	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	86.8	38	5,330	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	88.9	36	4,320	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	91.2	34	3,310	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	91.2	34	3,310	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	91.4	35	3,320	intron_variant	0	0	0					0

Supplementary Table 41: SNPs within sequencing range of *CYP2D6* for Sample 45556

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	31.3	112	77,350	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	96.9	96	3,930	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	100	124	1122	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	43.9	123	69,540	intron_variant	0	0	0					20
rs28578778	A>A/G	42523400	snv	het	no	22.3	94	73,210	intron_variant	0	0	0					1
rs1985842	G>G/T	42523409	snv	het	no	26.5	98	72,260	intron_variant	0	0	0					33
rs3915951	C>C/A	42523636	snv	het	yes	19.2	74	59,140	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs143276168	G>G/A	42523813	snv	het	no	44	92	51,400	intron_variant	0	0	0					0.4
rs16947	A>G/G	42523943	snv	hom	yes	99.1	116	1,115	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	40.3	120	71,480	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	24.6	119	89,290	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	24.6	126	95,310	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	27	127	92,340	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	28.1	139	100,390	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	20.4	96	74,190	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	31.5	89	61,280	splice_acceptor_variant	0	0	0					19
rs200720666	C>C/T	42524975	snv	het	no	13.6	66	57,900	intron_variant	0	0	0					0
rs1058164	G>G/C	42525132	snv	het	yes	59.3	59	24,350	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	42.5	40	23,170	intron_variant	0	0	0					6
rs1081003	G>G/A	42525756	snv	het	yes	52.1	73	35,380	synonymous_variant	426	336	112	F	ttC/ttT			2
rs71328650	C>C/A	42525952	snv	het	no	49.1	116	59,570	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	74.5	98	25,730	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	50.8	63	31,320	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	96.7	60	2,580	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	98.1	54	1,520	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	96.3	54	2,520	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	96	50	2,480	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	96	50	2,480	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	96	50	2,480	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	96.1	51	2,490	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	51.4	105	51,540	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 42: SNPs within sequencing range of *CYP2D6* for Sample 47298

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	99	202	2,199	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	99	100	1,980	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	98.6	147	2,144	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	63.6	164	59,103	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	98.5	199	3,196	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	30	210	147,630	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	32.7	212	142,690	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	39	218	133,850	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	42.2	206	119,870	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	23	140	107,320	synonymous_variant	618	528	176	G	ggT/ggC			0
rs200720666	C>C/T	42524975	snv	het	no	17.4	109	90,190	intron_variant	0	0	0					0
rs1058164	G>C/C	42525132	snv	hom	yes	99.2	124	1,122	synonymous_variant	498	408	136	V	gtC/gtG			44
rs1081004	C>C/T	42525625	snv	het	no	51.4	70	34,360	intron_variant	0	0	0					6
rs71328650	C>A/A	42525952	snv	hom	no	98.6	145	2,143	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	98.3	116	2,114	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	98.4	123	2,121	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	97.3	110	3,107	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	98	101	2,990	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	98	101	2,990	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	97	99	3,960	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	96.8	93	3,900	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	97.9	94	2,920	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	97.9	95	2,930	intron_variant	0	0	0					0

Supplementary Table 43: SNPs within sequencing range of *CYP2D6* for Sample 47812

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	97.6	165	4,161	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	94.4	71	4,670	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	100	106	1,106	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	67.4	132	43,890	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	99.5	210	1,209	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	24.8	202	152,500	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	25.5	208	155,530	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	34.2	193	127,660	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	33.3	172	114,570	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	31.4	169	116,530	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>C/C	42525132	snv	hom	yes	99.1	107	1,106	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>A/A	42525952	snv	hom	no	98.1	107	2,105	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	98.3	121	2,119	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	100	90	1,900	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	98.7	79	1,780	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	97.4	78	2,760	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	97.4	77	2,750	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	97.5	80	2,780	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	97.4	76	2,740	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	97.4	77	2,750	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	96.2	79	3,760	intron_variant	0	0	0					0

Supplementary Table 44: SNPs within sequencing range of *CYP2D6* for Sample 50261

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	32	372	253,119	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>G/G	42523003	snv	hom	yes	97.8	278	6,272	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.2	362	3,358	intron_variant	0	0	0					65
rs2004511	T>C/C	42523211	snv	hom	no	98.6	364	5,359	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	34.5	400	262,138	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	19.4	289	233,560	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	97.3	258	7,251	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>C/C	42524696	snv	hom	no	96.8	314	10,304	intron_variant	0	0	0					20
rs3892097	C>T/T	42524947	snv	hom	yes	87.1	234	30,203	splice_acceptor_variant	0	0	0					19
	T>T/G	42525793	snv	het	yes	24	101	76,240	missense_variant	389	299	100	D/A	gAc/gCc	deleterious(0)	probably_damaging(1)	0
rs28371705	G>C/C	42525798	snv	hom	yes	98	98	2,960	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>C/C	42525811	snv	hom	yes	97.5	121	3,118	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>T/T	42525821	snv	hom	yes	98.1	106	2,101	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>T/T	42526549	snv	hom	no	96.4	192	7,185	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	93.4	166	11,155	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	93.8	164	10,152	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	94.1	170	10,159	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	94.1	169	10,159	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	93.8	163	10,152	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	94.7	169	9,160	intron_variant	0	0	0					0
rs1065852	G>A/A	42526694	snv	hom	yes	97.9	283	6,277	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 45: SNPs within sequencing range of *CYP2D6* for Sample 51179

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs28371725	C>C/T	42523805	snv	het	no	32.4	1722	1,163,558	intron_variant	0	0	0					9

Supplementary Table 46: SNPs within sequencing range of *CYP2D6* for Sample 53551

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	95.9	169	7,162	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	95.8	120	5,115	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.3	153	1,151	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	56.8	155	67,880	intron_variant	0	0	0					33
rs5030867	T>T/G	42523858	snv	het	yes	44.1	177	99,780	missense_variant	1061	971	324	H/P	cAt/cCt	deleterious(0.04)	probably_damaging(0.938)	0
rs16947	A>G/G	42523943	snv	hom	yes	97.8	180	4,176	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	34	262	173,890	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	33.8	266	176,900	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	37	265	167,980	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	43.6	241	136,105	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	38.4	166	101,630	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>C/C	42525132	snv	hom	yes	97.9	145	3,142	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>A/A	42525952	snv	hom	no	98.1	162	3,159	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	98.6	148	2,145	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	94.9	137	7,130	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	92.2	115	9,106	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	90.1	101	10,910	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	89.8	99	10,880	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	88.7	98	11,860	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	88.8	80	9,710	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	91	78	7,710	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	91.3	80	7,730	intron_variant	0	0	0					0

Supplementary Table 47: SNPs within sequencing range of *CYP2D6* for Sample 53841

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	61.1	108	42,660	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	35.4	132	84,460	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	51.4	105	51,540	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	35.2	122	79,430	intron_variant	0	0	0					33
rs28371726	A>A/G	42523539	snv	het	yes	21.4	131	103,280	synonymous_variant	1173	1083	361	H	caT/caC			0.26
rs16947	A>A/G	42523943	snv	het	yes	63.2	155	56,960	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs1058164	G>G/C	42525132	snv	het	yes	42.6	68	39,290	synonymous_variant	498	408	136	V	gtC/gtG			44
rs78854695	A>A/C	42525728	snv	het	yes	42.9	79	44,330	intron_variant	0	0	0					0
rs147296446	C>C/G	42526049	snv	het	no	28.6	119	85,340	intron_variant	0	0	0					0
rs56011157	C>C/T	42526549	snv	het	no	41.7	96	56,400	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	37.6	93	58,350	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	37.6	93	58,350	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	36.2	94	60,340	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	38.4	87	53,330	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	39.5	83	49,320	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	35.8	81	52,290	intron_variant	0	0	0					0

Supplementary Table 48: SNPs within sequencing range of *CYP2D6* for Sample 56195

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	24.1	307	233,740	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>G/G	42523003	snv	hom	yes	98.3	300	5,289	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.7	345	1,340	intron_variant	0	0	0					65
rs2004511	T>C/C	42523211	snv	hom	no	99.4	340	2,337	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	38.7	391	238,150	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	25.8	236	175,610	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	97.6	295	7,288	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>C/C	42524696	snv	hom	no	97.8	229	5,224	intron_variant	0	0	0					20
rs3892097	C>T/T	42524947	snv	hom	yes	68	278	89,189	splice_acceptor_variant	0	0	0					19
	T>T/G	42525793	snv	het	yes	27.2	82	59,220	missense_variant	389	299	100	D/A	gAc/gCc	deleterious(0)	probably_damaging(1)	0
rs28371705	G>C/C	42525798	snv	hom	yes	95.8	95	4,910	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>C/C	42525811	snv	hom	yes	97.7	130	3,127	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	98.3	119	2,113	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>T/T	42526549	snv	hom	no	93.5	185	12,173	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	92.8	168	12,155	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	92.1	164	13,151	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	91.8	161	13,146	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	91.1	157	14,143	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	90.8	154	14,139	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	91.8	160	13,146	intron_variant	0	0	0					0
rs1065852	G>A/A	42526694	snv	hom	yes	99.7	291	1,290	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 49: SNPs within sequencing range of *CYP2D6* for Sample 60533

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	51.7	58	28,300	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	92.6	54	4,500	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	96.6	59	2,570	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	71.9	57	16,410	intron_variant	0	0	0					20
rs28578778	A>A/G	42523400	snv	het	no	25.7	35	26,900	intron_variant	0	0	0					1
rs1985842	G>G/T	42523409	snv	het	no	32.4	34	23,110	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	98.4	64	1,630	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs28371718	G>G/T	42524218	snv	het	yes	35.6	59	38,210	synonymous_variant	891	801	267	P	ccC/ccA			0
rs58440431	T>T/C	42524696	snv	het	no	25.9	85	63,220	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	30.2	86	60,260	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	32.9	85	57,280	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	40.5	74	44,300	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	20	50	40,100	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	14.1	79	67,110	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	23.8	80	61,190	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	38.7	31	19,120	synonymous_variant	498	408	136	V	gtC/gtG			44
rs78854695	A>A/C	42525728	snv	het	yes	17.2	29	24,500	intron_variant	0	0	0					0
rs1081003	G>G/A	42525756	snv	het	yes	57.7	26	11,150	synonymous_variant	426	336	112	F	ttC/ttT			2
rs71328650	C>C/A	42525952	snv	het	no	64.2	67	24,430	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	80	45	9,360	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	33.3	54	36,180	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	94.4	36	2,340	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	90.3	31	3,280	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	90	30	3,270	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	89.7	29	3,260	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	88.9	27	3,240	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	89.3	28	3,250	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	90.6	32	3,290	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	39	41	25,160	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 50: SNPs within sequencing range of *CYP2D6* for Sample 62592

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	52.6	156	74,820	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	35.1	95	61,330	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	45.1	102	56,460	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	40.2	97	58,390	intron_variant	0	0	0					33
rs28371725	C>C/T	42523805	snv	het	no	54.1	111	51,600	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	57	128	55,730	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	18.7	150	122,280	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	18.3	153	125,280	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	20.3	148	118,300	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	31.8	148	101,470	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	21.8	101	79,220	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	48.4	93	48,450	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	52.3	111	53,580	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	62	100	38,620	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	50.9	108	53,550	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	40.2	87	52,350	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	42.1	76	44,320	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	42.1	76	44,320	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	39.7	73	44,290	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	39.1	69	42,270	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	41.5	65	38,270	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	44.9	69	38,310	intron_variant	0	0	0					0

Supplementary Table 51: SNPs within sequencing range of *CYP2D6* for Sample 63712

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	42.3	421	243,178	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	98.5	348	5,338	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.4	313	2,309	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	52.9	315	148,166	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	26.1	371	274,970	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	22.3	458	356,102	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	16.8	286	238,480	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	98.8	429	5,423	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	42.8	465	266,199	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	19.7	473	379,930	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	19.7	493	395,970	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	25.1	482	361,121	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	24.6	443	334,109	synonymous_variant	747	657	219	F	ttT/ttC			0
rs3892097	C>C/T	42524947	snv	het	yes	42.9	316	180,135	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	52.1	287	137,149	synonymous_variant	498	408	136	V	gtC/gtG			44
rs28371705	G>G/C	42525798	snv	het	yes	40.3	159	95,640	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	42.9	202	113,850	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	39.3	210	125,810	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs71328650	C>C/A	42525952	snv	het	no	56.1	337	147,188	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	52.2	312	149,163	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	44.3	319	177,141	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	95.7	279	12,267	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	95.2	229	11,217	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	95	223	11,211	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	95.2	229	11,218	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	94.8	211	11,200	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	94.7	207	11,196	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	95.1	205	10,195	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	52.7	355	168,187	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 52: SNPs within sequencing range of *CYP2D6* for Sample 65602

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	26.2	830	611,217	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>G/G	42523003	snv	hom	yes	98.6	739	10,725	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.7	760	2,751	intron_variant	0	0	0					65
rs2004511	T>C/C	42523211	snv	hom	no	99.5	757	4,753	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	38.3	944	581,361	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	21.3	597	468,127	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	99.3	733	5,724	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>C/C	42524696	snv	hom	no	99.8	616	1,612	intron_variant	0	0	0					20
rs3892097	C>T/T	42524947	snv	hom	yes	81.5	607	112,495	splice_acceptor_variant	0	0	0					19
	T>T/G	42525793	snv	het	yes	23.9	202	153,480	missense_variant	389	299	100	D/A	gAc/gCc	deleterious(0)	probably_damaging(1)	0
rs28371705	G>C/C	42525798	snv	hom	yes	98	201	4,197	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>C/C	42525811	snv	hom	yes	98.1	274	5,265	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>T/T	42525821	snv	hom	yes	97.4	278	7,265	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>T/T	42526549	snv	hom	no	93.2	483	33,449	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	93.6	419	27,392	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	93.6	410	26,380	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	92.8	390	28,361	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	93	384	27,357	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	92	352	28,323	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	92.2	361	28,332	intron_variant	0	0	0					0
rs1065852	G>A/A	42526694	snv	hom	yes	97.7	619	14,604	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 53: SNPs within sequencing range of *CYP2D6* for Sample 66900

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs201759814	G>G/A	42522550	snv	het	yes	22.8	308	237,700	3_prime_UTR_variant	1610	0	0					0
rs116917064	A>A/G	42523003	snv	het	yes	54.7	246	111,134	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	72.3	238	66,172	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	71.8	241	68,173	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	44.9	296	163,133	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	19.8	167	134,330	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs28371725	C>C/T	42523805	snv	het	no	56.8	155	67,880	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	48.2	191	99,920	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	48.6	222	114,108	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	43.3	210	119,910	splice_acceptor_variant	0	0	0					19
rs28371705	G>G/C	42525798	snv	het	yes	27.5	102	74,280	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	20.8	98	76,200	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	21.7	93	72,200	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>C/T	42526549	snv	het	no	52.5	224	106,117	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	47.3	201	106,950	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	46.2	197	106,910	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	46.4	183	98,850	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	47.8	185	96,880	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	45.3	182	99,820	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	47.2	176	93,830	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	60.6	254	100,154	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 54: SNPs within sequencing range of *CYP2D6* for Sample 70613

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs116917064	A>A/G	42523003	snv	het	yes	51.4	148	72,760	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	51.5	171	83,880	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	50.3	171	85,860	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	18.7	171	139,320	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	19	126	102,240	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs28371725	C>C/T	42523805	snv	het	no	46.8	156	83,730	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	52.4	168	80,880	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	50	108	54,540	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	38.8	130	79,500	splice_acceptor_variant	0	0	0					19
rs28371705	G>G/C	42525798	snv	het	yes	27.6	58	42,160	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	31.1	75	51,230	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs56011157	C>C/T	42526549	snv	het	no	43.5	115	65,500	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	39.8	108	65,430	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	39	105	64,410	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	38.1	97	60,370	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	37.9	96	59,360	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	35.2	91	59,320	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	38.9	90	55,350	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	59.7	159	64,950	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 55: SNPs within sequencing range of *CYP2D6* for Sample 72638

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	96.2	157	6,150	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	99.3	137	1,135	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.3	141	1,140	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	53.4	131	61,700	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	98.1	161	3,157	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs5030656	CCTT>CCTT/C	42524175	deletion	het	yes	42.5	116	69,510	inframe_deletion, splice_region_variant			281	K/-	AAG/-			2
rs111564371	T>T/C	42524708	snv	het	no	38.5	195	120,750	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	39.5	207	124,810	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	44.7	219	121,980	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	41.6	226	132,940	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	20	142	112,280	synonymous_variant	618	528	176	G	ggT/ggC			0
rs200720666	C>C/T	42524975	snv	het	no	14.8	136	115,200	intron_variant	0	0	0					0
rs1058164	G>C/C	42525132	snv	hom	yes	99.1	117	1,116	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>A/A	42525952	snv	hom	no	98	147	3,144	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	97.5	119	3,116	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	99.2	118	1,117	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	93.4	91	6,850	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	94.7	75	4,710	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	94.6	74	4,700	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	94.7	75	4,710	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	94.4	72	4,680	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	94.5	73	4,690	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	94.7	76	4,720	intron_variant	0	0	0					0

Supplementary Table 56: SNPs within sequencing range of *CYP2D6* for Sample 79391

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs116917064	A>A/G	42523003	snv	het	yes	47.1	584	309,275	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	56	503	221,281	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	54.5	493	224,268	intron_variant	0	0	0					20
rs1058172	C>C/T	42523528	snv	het	yes	20.4	505	402,103	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs16947	A>A/G	42523943	snv	het	yes	50.1	666	332,333	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	52.8	513	242,271	intron_variant	0	0	0					20
rs3892097	C>C/T	42524947	snv	het	yes	51.9	421	202,218	splice_acceptor_variant	0	0	0					19
rs28371705	G>G/C	42525798	snv	het	yes	33.6	212	140,710	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	39.9	252	149,990	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	36.4	254	159,910	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs56011157	C>C/T	42526549	snv	het	no	35.6	324	208,115	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	34.7	297	194,103	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	34.7	294	192,102	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	35.1	285	185,100	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	35.6	278	179,990	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	33.2	256	171,850	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	36.2	271	173,980	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	50	402	201,201	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 57: SNPs within sequencing range of *CYP2D6* for Sample 81119

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	47.3	2049	1,080,968	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	99.4	1393	91,381	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.6	1771	71,761	intron_variant	0	0	0					65
rs2004511	T>T/C	42523211	snv	het	no	58.3	1771	7,391,032	intron_variant	0	0	0					20
rs1985842	G>G/T	42523409	snv	het	no	30.8	1451	1,002,446	intron_variant	0	0	0					33
rs1058172	C>C/T	42523528	snv	het	yes	18.9	1598	1,292,302	missense_variant	1184	1094	365	R/H	cGc/cAc	deleterious(0)	probably_damaging(0.999)	14
rs3915951	C>C/A	42523636	snv	het	yes	11.9	1093	960,130	missense_variant, splice_region_variant	1076	986	329	R/L	cGc/cTc	tolerated(0.07)	benign(0.039)	0
rs16947	A>G/G	42523943	snv	hom	yes	99.8	2225	42,218	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs58440431	T>T/C	42524696	snv	het	no	39.7	2306	1,388,915	intron_variant	0	0	0					20
rs111564371	T>T/C	42524708	snv	het	no	18.2	2286	1,867,416	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	18.8	2254	1,827,424	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	23.5	2217	1,696,521	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	24.5	2361	1,781,577	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	17.1	1666	1,377,285	synonymous_variant	618	528	176	G	ggT/ggC			0
rs3892097	C>C/T	42524947	snv	het	yes	43.3	1600	904,691	splice_acceptor_variant	0	0	0					19
rs1058164	G>G/C	42525132	snv	het	yes	49.3	1447	732,713	synonymous_variant	498	408	136	V	gtC/gtG			44
rs28371705	G>G/C	42525798	snv	het	yes	35.6	970	625,345	synonymous_variant	384	294	98	T	acC/acG			14
rs28371704	T>T/C	42525811	snv	het	yes	35.1	994	636,344	missense_variant	371	281	94	H/R	cAc/cGc	tolerated(0.74)	benign(0.004)	14
rs28371703	G>G/T	42525821	snv	het	yes	32.6	989	657,318	missense_variant	361	271	91	L/M	Ctg/Atg	deleterious(0.03)	possibly_damaging(0.877)	14
rs71328650	C>C/A	42525952	snv	het	no	49.6	1474	742,730	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	52.3	1379	657,720	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	51.3	1312	638,673	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	98.5	925	14,911	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	97	796	24,772	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	97	774	23,750	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	96.8	791	25,765	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	96.9	781	24,757	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	96.9	773	24,749	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	96.9	782	24,757	intron_variant	0	0	0					0
rs1065852	G>G/A	42526694	snv	het	yes	49.9	1405	704,700	missense_variant	190	100	34	P/S	Cca/Tca	deleterious(0.03)	possibly_damaging(0.581)	20

Supplementary Table 58: SNPs within sequencing range of *CYP2D6* for Sample 86636

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	47.4	969	508,458	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	36.8	834	526,306	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	49.8	892	445,442	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	42.4	716	411,302	intron_variant	0	0	0					33
rs16947	A>A/G	42523943	snv	het	yes	52.3	1156	550,604	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs28371718	G>G/T	42524218	snv	het	yes	47.2	837	442,395	synonymous_variant	891	801	267	P	ccC/ccA			0
rs111564371	T>T/C	42524708	snv	het	no	19.8	1149	922,227	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	20.3	1151	916,233	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	25.3	1191	887,301	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	24.5	1307	985,319	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	15.1	993	840,149	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	46.6	801	428,373	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	49.7	918	460,454	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	48	813	423,390	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	45.3	846	463,383	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	39.4	763	462,300	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	35.1	659	427,231	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	34.8	651	424,226	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	33.7	631	418,212	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	34.6	613	400,212	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	35.5	611	394,217	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	37.1	608	381,225	intron_variant	0	0	0					0
rs769258	C>C/T	42526763	snv	het	yes	49	937	477,459	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 59: SNPs within sequencing range of *CYP2D6* for Sample 86710

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	89.6	48	5,430	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	88.6	35	4,310	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	94.1	34	2,320	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	48.6	35	18,170	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	95.2	62	3,590	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	26.8	56	41,150	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	25.9	58	43,150	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	32.7	49	33,160	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	29.8	57	40,170	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	44.4	45	25,200	synonymous_variant	618	528	176	G	ggT/ggC			0
rs200720666	C>C/T	42524975	snv	het	no	14.6	41	35,600	intron_variant	0	0	0					0
rs1058164	G>C/C	42525132	snv	hom	yes	90.9	33	3,300	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	95.2	42	2,400	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	97.1	36	1,340	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	94.1	34	2,320	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	84	25	4,210	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	65.4	26	9,170	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	65.4	26	9,170	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	65.4	26	9,170	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	59.1	22	9,130	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	60.9	23	9,140	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	60.9	23	9,140	intron_variant	0	0	0					0

Supplementary Table 60: SNPs within sequencing range of *CYP2D6* for Sample 88291

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Ac	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	40.4	114	68,460	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	28.3	92	66,260	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	32.8	119	80,390	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	30.2	96	67,290	intron_variant	0	0	0					33
rs16947	A>A/G	42523943	snv	het	yes	38.9	132	80,510	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs28371718	G>G/T	42524218	snv	het	yes	33.1	125	83,410	synonymous_variant	891	801	267	P	ccC/ccA			0
rs111564371	T>T/C	42524708	snv	het	no	13.9	137	118,190	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	12.3	146	128,180	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	21.9	138	107,300	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	30	131	91,390	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	13.6	133	114,180	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	26.3	76	56,200	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	37.9	117	72,440	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	43.7	126	71,550	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	33.8	142	94,480	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	26.9	120	87,320	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	19.3	114	92,220	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	18	112	91,200	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	22.1	106	81,230	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	21	100	79,210	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	21.2	100	78,210	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	21.8	88	68,190	intron_variant	0	0	0					0
rs28371695, rs75085559	A>A/AC	42526836	insertion	het	yes	32	137	104,490	5_prime_UTR_variant, feature_elongation	0	0	0					0.26

Supplementary Table 61: SNPs within sequencing range of *CYP2D6* for Sample 88321

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	71	32	9,220	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	37.8	37	23,140	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	62.5	32	12,200	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	41.4	29	17,120	intron_variant	0	0	0					33
rs16947	A>A/G	42523943	snv	het	yes	56.7	30	13,170	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	40.7	28	16,110	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	40	30	18,120	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	42.4	33	19,140	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	36.4	33	21,120	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	10.8	37	33,400	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	65.2	23	8,150	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	54.3	36	16,190	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	36.8	19	12,700	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	47.6	21	11,100	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	22.7	22	17,500	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	25	20	15,500	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	25	20	15,500	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	33.3	21	14,700	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	33.3	24	16,800	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	36	25	16,900	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	46.2	26	14,120	intron_variant	0	0	0					0
rs769258	C>C/T	42526763	snv	het	yes	41.7	24	14,100	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 62: SNPs within sequencing range of *CYP2D6* for Sample 89899

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs78209835	C>C/T	42523613	snv	het	yes	39.5	114	69,450	missense_variant	1099	1009	337	D/N	Gac/Aac	tolerated(0.21)	benign(0.003)	0.4
rs769258	C>C/T	42526763	snv	het	yes	43.7	135	76,590	missense_variant	121	31	11	V/M	Gtg/Atg	tolerated(0.16)	unknown(0)	5

Supplementary Table 63: SNPs within sequencing range of *CYP2D6* for Sample 90174

Same as reference

Supplementary Table 64: SNPs within sequencing range of *CYP2D6* for Sample 94560

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>C/C	42522613	snv	hom	yes	98.5	1306	201,285	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>G/G	42523003	snv	hom	yes	98.7	688	9,677	intron_variant	0	0	0					0
rs28371730	T>C/C	42523209	snv	hom	no	99.6	1142	41,132	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	63.8	1150	415,732	intron_variant	0	0	0					33
rs16947	A>G/G	42523943	snv	hom	yes	99.5	1500	71,489	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs111564371	T>T/C	42524708	snv	het	no	30.9	1893	1,309,584	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	32.1	1927	1,307,618	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	38.7	2032	1,242,785	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	39.5	2100	1,270,828	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	26.8	1300	947,346	synonymous_variant	618	528	176	G	ggT/ggC			0
rs200720666	C>C/T	42524975	snv	het	no	17.4	1138	939,198	intron_variant	0	0	0					0
rs113678157	C>C/T	42524982	snv	het	no	15.2	1116	946,170	intron_variant	0	0	0					0
rs1058164	G>C/C	42525132	snv	hom	yes	99.1	1015	91,005	synonymous_variant	498	408	136	V	gtC/gtG			44
	C>C/T	42525840	snv	het	yes	47.2	974	512,457	synonymous_variant	342	252	84	L	ctG/ctA			0
rs71328650	C>A/A	42525952	snv	hom	no	99	1094	111,079	intron_variant	0	0	0					45
rs147296446	C>G/G	42526049	snv	hom	no	99.4	1125	71,117	intron_variant	0	0	0					0
rs28371699	A>C/C	42526484	snv	hom	no	99.6	905	4,901	intron_variant	0	0	0					45
rs56011157	C>T/T	42526549	snv	hom	no	98.1	685	13,672	intron_variant	0	0	0					0
rs28695233	G>T/T	42526561	snv	hom	no	96.7	601	20,581	intron_variant	0	0	0					0
rs75276289	G>C/C	42526562	snv	hom	no	96.6	591	20,571	intron_variant	0	0	0					0
rs76312385	G>A/A	42526567	snv	hom	no	96.5	599	21,577	intron_variant	0	0	0					0
rs74644586	C>G/G	42526571	snv	hom	no	96.2	574	22,552	intron_variant	0	0	0					0
rs1080996	T>G/G	42526573	snv	hom	no	96.6	584	20,564	intron_variant	0	0	0					0
rs1080995	G>C/C	42526580	snv	hom	no	96.7	611	20,589	intron_variant	0	0	0					0

Supplementary Table 65: SNPs within sequencing range of *CYP2D6* for Sample 97924

dbSNP ID	Variant	Coordinate	Type	Genotype	Exonic	Alt Variant Freq	Read Depth	Allelic Depths	Consequence	cDNA Position	CDS Position	Protein Position	Amino Acids	Codons	Sift	PolyPhen	Allele Freq Eur
rs1135840	G>G/C	42522613	snv	het	yes	48.3	1036	535,500	missense_variant	1547	1457	486	T/S	aCc/aGc	tolerated(0.99)	benign(0.029)	45
rs116917064	A>A/G	42523003	snv	het	yes	35.4	832	536,294	intron_variant	0	0	0					0
rs28371730	T>T/C	42523209	snv	het	no	50	900	445,445	intron_variant	0	0	0					65
rs1985842	G>G/T	42523409	snv	het	no	36.5	773	490,282	intron_variant	0	0	0					33
rs28371725	C>C/T	42523805	snv	het	no	49.7	794	398,394	intron_variant	0	0	0					9
rs16947	A>A/G	42523943	snv	het	yes	48	1091	566,523	missense_variant	976	886	296	C/R	Tgc/Cgc	tolerated(0.35)	benign(0.001)	66
rs35742686	CT>CT/C	42524243	deletion	het	yes	43	864	504,380	frameshift_variant, feature_truncation	865	775	259					2
rs111564371	T>T/C	42524708	snv	het	no	20.6	1253	990,257	intron_variant	0	0	0					0
rs112568578	C>C/G	42524713	snv	het	no	21.5	1249	980,268	intron_variant	0	0	0					0
rs113889384	G>G/A	42524743	snv	het	no	27.5	1249	904,343	intron_variant	0	0	0					0
rs28371713	A>A/G	42524795	snv	het	yes	25.3	1298	966,328	synonymous_variant	747	657	219	F	ttT/ttC			0
rs111606937	A>A/G	42524924	snv	het	yes	14.6	934	794,136	synonymous_variant	618	528	176	G	ggT/ggC			0
rs1058164	G>G/C	42525132	snv	het	yes	48.6	850	436,412	synonymous_variant	498	408	136	V	gtC/gtG			44
rs71328650	C>C/A	42525952	snv	het	no	52.9	828	388,435	intron_variant	0	0	0					45
rs147296446	C>C/G	42526049	snv	het	no	51	829	406,423	intron_variant	0	0	0					0
rs28371699	A>A/C	42526484	snv	het	no	49.4	858	434,424	intron_variant	0	0	0					45
rs56011157	C>C/T	42526549	snv	het	no	41	721	424,295	intron_variant	0	0	0					0
rs28695233	G>G/T	42526561	snv	het	no	36.8	665	420,245	intron_variant	0	0	0					0
rs75276289	G>G/C	42526562	snv	het	no	36.6	656	414,239	intron_variant	0	0	0					0
rs76312385	G>G/A	42526567	snv	het	no	37	654	412,242	intron_variant	0	0	0					0
rs74644586	C>C/G	42526571	snv	het	no	34.7	624	407,216	intron_variant	0	0	0					0
rs1080996	T>T/G	42526573	snv	het	no	36.2	616	393,223	intron_variant	0	0	0					0
rs1080995	G>G/C	42526580	snv	het	no	37.3	609	381,227	intron_variant	0	0	0					0

CHAPTER 4

Accurate Cost Effective Method of Screening CYP2D6 with Microarray Technology

As demonstrated in the chapter 3, sequencing the entire gene can be beneficial to finding novel SNPs. However, diagnostic laboratories in places such as hospitals that may want to screen patients genetically before administering drugs or aid in determining cause of death may find next generation sequencing techniques too demanding, time consuming, and cost prohibitive to employ. Therefore, it is anticipated that more laboratories will run microarrays than implement sequencing centers for routine testing. This option of using microarrays is sensible based on application and cost bases. One of the most popular microarray chips in this demographic is the Global Screening Array (GSA). This beadchip combines multi-ethnic genome-wide content, curated clinical research variants, and quality control markers for precision medicine research targeting ~640,000 markers (https://www.illumina.com/products/by-type/microarray-kits/infinium-global-screening.html). This beadchip has 321 probes targeting *CYP2D6* alone including 232 of those targeting CNV and the *CYP2D6* variants *3A, *3B,*6, *7, *9, *22, *23, *25, *33, *38, *44, *48, *106. These variants cover simple SNP substitutions such as *3, *6, *38, and *44, full length but non-functional copies such as *7, lacking a codon such as *9, and a few that target the metabolism of historically important drugs such as primaquine (*22) and tamoxifen (*33). This coverage gives a broad overview of the capabilities one would want to target for a proof of concept test.

This particular beadchip, however, falls victim to pseudogene interference making it a prime candidate to determine if pseudogene interference can be overcome. A methodology was developed to overcome this pseudogene interference, and this test can be accurately and reliably run now at a reasonable cost.

(submitted for publication to *BioTechniques*)

Abstract:

Aims:

The cytochrome P450 2D6 (*CYP2D6*) enzyme metabolizes 25% of known drugs currently on the market and is diverse amongst individuals. Currently, comprehensive screening is difficult due to the highly polymorphic nature of the *CYP2D6* gene and its two homologous pseudogenes. This study aims to develop a low cost method to screen *CYP2D6* using a microarray.

Methods and Materials:

A protocol was developed to integrate an optimized PCR into the microarray workflow to enrich for *CYP2D6*.

Results and Conclusions:

A long-range PCR assay was proven to reduce the pseudogene interference to characterize better the true *CYP2D6* SNP call, allowing for a more accurate and robust screening process at a relatively low cost for laboratories that are already running microarray processes.

Key words: DNA, microarray, *CYP2D6*, pharmacogenomics, pseudogene interference,

Introduction

The CYP450 super-family is comprised of over 50 enzymes involved in metabolism of foreign and endogenous compounds. The *CYP2D6* enzyme is of particular interest in that it is responsible for approximately 30% of Phase I metabolism of these compounds [1-8]. Moreover, the enzyme is involved in the phase I metabolism of approximately 25% of marketed drugs, converting pro-drugs to active metabolites, active drugs to active metabolites, or active drugs to inactive metabolites [7-9]. Adverse drug reactions have been associated with genetic variants of the gene *CYP2D6* which impact the functionality of the enzyme. The *CYP2D6* gene is located on chromosome 22q13.1 and contains 9 exons within 4,383 base pairs (bp) based on the NCBI 37 genome assembly [10]. Currently, there are more than 100 allelic variants and sub variants described for this gene [11].

This variation leads to four *CYP2D6* phenotypes: poor metabolizers (PM; two inactive alleles), intermediate metabolizers (IM; one reduced activity allele and one inactive allele or two reduced activity alleles), extensive metabolizers (EM; at least one functional allele), and ultrarapid metabolizers (UM; three or more functional copies) [12-14]. Genetic typing of *CYP2D6* can be useful for predicting response to drug exposure which is important to a number of applications, including healthcare and molecular autopsy. However, current diagnostic methods have complications that can make genetic typing of *CYP2D6* problematic.

While massively parallel sequencing (MPS) of *CYP2D6* is desirable as it can yield the full sequence of the gene [15-16], more cost effective per sample methods are sought. Microarrays containing known SNPs that define *CYP2D6* star alleles are one of these formats for genetic typing

[17]. However, typing of this pharmacogenetic gene on a microarray has been notoriously difficult due to SNPs that reside within homologs of *CYP2D6*, particularly *CYP2D7P* and *CYP2D8P* that contribute background noise. We propose that a higher concentration of a region of interest could out compete the original amplification during whole genome amplification to ensure the SNP call was indeed the region of interest in such highly complex locations. The study herein describes a methodology that preferentially selects for alleles of *CYP2D6*, thus reducing the noise from its closely related paralogs.

Methods and Materials

A first attempt to bypass the pseudogene interference issue was to add amplicons directly to a beadchip, but this approach was found to be cost prohibitive doubling the cost to run a sample as well as increased incurred costs of running an entirely separate beadchip. Due to this finding, it was decided that overcoming pseudogene interference within the sample workflow was a more efficient option.

In order to test the hypothesis that a higher concentration of a region of interest would be represented on the Global Screening Array (GSA) v2 beadchip, a proof of concept experiment was performed using two known commercially available kits (Figure 2). The ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA) targets 60 forensically relevant STRs and 170 SNPs that are used for human identity, ancestry, and phenotype determination. There are 149 SNPs in this assay that overlap with the SNPs targeted on the Global Screening Array (GSA) v2 beadchip (Illumina, Inc., San Diego, CA); see appendix for a complete list of overlapping SNPs. Twenty known purified DNA samples with known concentrations were purchased from Biochain Institute Inc. (Biochain Institute Inc, Newark, CA) and used for the study in accordance to the University of

North Texas Health Science Center's Institutional Review Board. Throughout the entire study, each experiment contained replicates of 30.

Since the amplified material would be assayed on a microarray instead of being analyzed by MPS, a modified amplification protocol was used to obtain the appropriate amplicons. The first amplification set up was performed following the manufacturer's recommendations except that an input amount of 25 ng was used. The amplification parameters for the first amplification remained the same except the second part of the amplification employed 25 cycles instead of the recommended ten cycles. The amplification parameters were: 98˚C for 3 minutes, 8 cycles of 96˚C for 45 seconds, 80˚C for 30 seconds, 54˚C for 2 minutes, 68˚C for 2 minutes, 25 cycles of 96˚C for 30 seconds, 68˚C for 3 minutes, and 68˚C for 10 minutes, and a final hold at 10˚C.

The amplification product was purified with a SPRI clean up using a 1:1 volume ratio of product to bead mixture, and the product was verified on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) using the DNA 1000 chip kit. Two hundred nanograms of product then were added into the microarray workflow at the precipitation step. All subsequent steps of the microarray workflow were processed as the manufacturer recommended which includes the final prepared DNA being visualized on the GSA v2 beadchip using an iScan (Illumina).

To determine if the targeted ForenSeq amplicons can overcome the background noise created by whole genome amplification (WGA), two separate individuals were used by adding the ForenSeq amplicons into the WGA of the same sample as well as to the other sample.

Results of Concept:

The addition of ForenSeq amplicons at the precipitation step positively correlated with an increase in intensity when the same sample was used for both the ForenSeq and WGA step (Figure

3). The addition of another sample of ForenSeq amplicons into the WGA of a different sample showed a shift in calls to the contributor of the ForenSeq amplicons (Figure 3). This shift occurred for 147 of the 149 SNPs. The remaining two SNPs yielded a reduction in intensity as well as five SNPs not targeted with ForenSeq had a reduction in intensity. In an effort to reduce negative effects further optimization was required.

First, the introduction of the ForenSeq amplicons was added into three separate locations in the microarray workflow (Figure 4). They were introduced at the first step in WGA, at fragmentation, and at the original step of precipitation. It was determined that the negative effects of reduced intensities were minimized if the ForenSeq amplicons were introduced at the first step of the microarray workflow, i.e. WGA. A titration experiment adding 15 ng to 300 ng of amplified material into the WGA was performed to determine optimal concentration. The DNA for the WGA remained at a constant 200 ng for these experiments. Figure 5 shows that in samples where both individuals had the same call at the targeted SNP an increase in intensity was observed and a slight theta shift occurred. This effect could be remedied with a new manifest where the samples are trained based on the new workflow. In samples that had differing SNPs, the titration showed that 15 ng were sufficient to turn an AA call to AB, or BB to AB, but were not capable to fully change the SNP to AA or BB. Input DNA at 150 ng and 300 ng were sufficient for the ForenSeq amplicons to be observed over the original WGA amplicon (Figure 6).

While the data supported that amplified material can be spiked into the WGA with no negative effect, this process creates a less than optimal workflow as one would have to spike amplified material from a post amplification area into a plate containing samples that had not been amplified (i.e. in a pre-amplification work area). To reduce the chance of contamination and potential sample switches a staging experiment was performed with 50 ng of input DNA. The

WGA and targeted PCR plates were created at the same time, and the WGA plate was staged at three temperatures, 4˚C, room temperature, and -20˚C and maintained for eight hours to allow for long amplification of targeted regions. There is a potential for primer dimerization, so the WGA plates were staged with and without the reagent multi-sample amplification mix (MSM) that presumably contains the polymerase. The 50 ng of amplified material with and without SPRI clean up were added to the WGA plate and the process allowed to continue as previously described (Figure 7). The workflow proved to be robust with no major effect observed in any of the variables or with the spiking of PCR products (Figures 8, 9, and 10). The SNPs were converted as observed earlier with any of the tested conditions. An example (Figure 11) is shown where the plate was staged at room temperature with MSM and no clean up performed on the ForenSeq amplified material.

CYP2D6 workflow method

Based on the above concept work, the next step was to determine if the same call conversion could occur with the *CYP2D6* target, given the paralog complication. The protocol to amplify one large fragment of *CYP2D6* developed by KAPA Biosystems in their application note (https://www.n-genetics.com/products/1104/1033/12516.pdf) was followed to amplify fragment A. Fragment A was chosen as it amplifies the entire *CYP2D6* gene and excludes the homologs *CYP2D7P* and *CYP2D8P*, which is impossible to achieve with the standard WGA protocol. Two microliters of this amplified material were added into the WGA protocol following the manufacturer's recommendations and visualized on the GSA v2 beadchip on the iScan.

Results

A shift at SNPs within the fragment A of *CYP2D6* was observed demonstrating that the amplified material can outcompete the background WGA for that location (Figure 12). Importantly, the SNPs not specifically targeted were not affected by the spiking of amplified material. An example (Figure 13) is shown for SNPs rs950114 and rs9501628.

Discussion

A low cost screening method of *CYP2D6* on the Illumina microarray GSA v2 beadchip was demonstrated to be effective in capturing the true SNP calls for *CYP2D6* by outperforming the background of the homologs *CYP2D7P* and *CYP2D8P* to generate a more accurate call, i.e., reducing background noise. This process could be implemented in laboratories already running microarrays with a reasonable increase in turnaround time as only one additional PCR is needed to generate more accurate calls. This PCR would add about $1.25 per sample assuming a 96 well plate format in addition to the normal processing of the microarray chip. This method could be used to assay *CYP2D6* in situations where genetic analyses could assist in drug prescription or postmortem in situations where drug exposure may be a triggering event in the cause of death.

Tables and Figures:

Figure 1: Pseudogene interference causes miscalls and skewed SNPgrams. Figure 1A shows a whole genome sample SNP with optimal clustering with well separated and tightly grouped clusters. Figure 1B shows a *CYP2D6* SNP with a skewed SNPgram and miscalls of AB when all sample truth calls should be AA at this SNP.

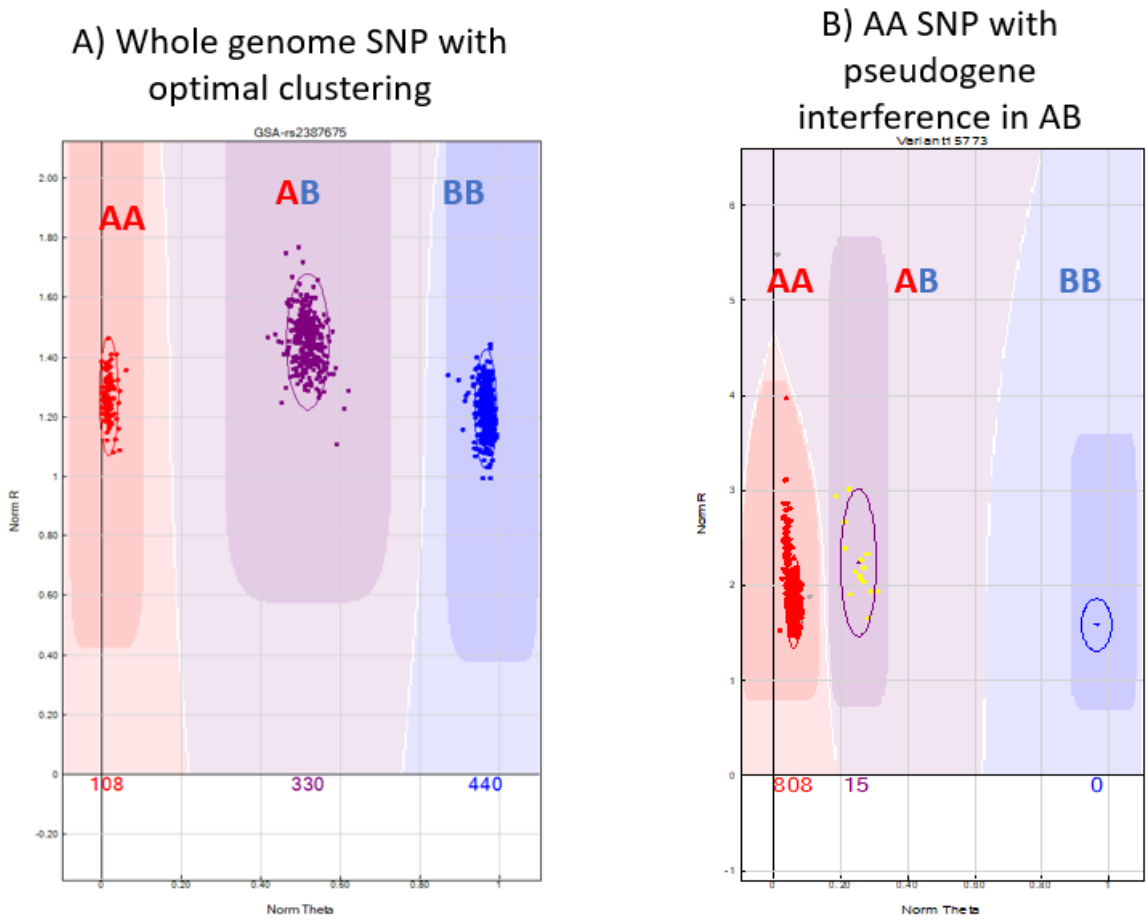


Figure 2: Workflow of Proof of Concept ForenSeq amplicons into Microarray workflow adding the amplified material with the microarray workflow at the precipitation step.

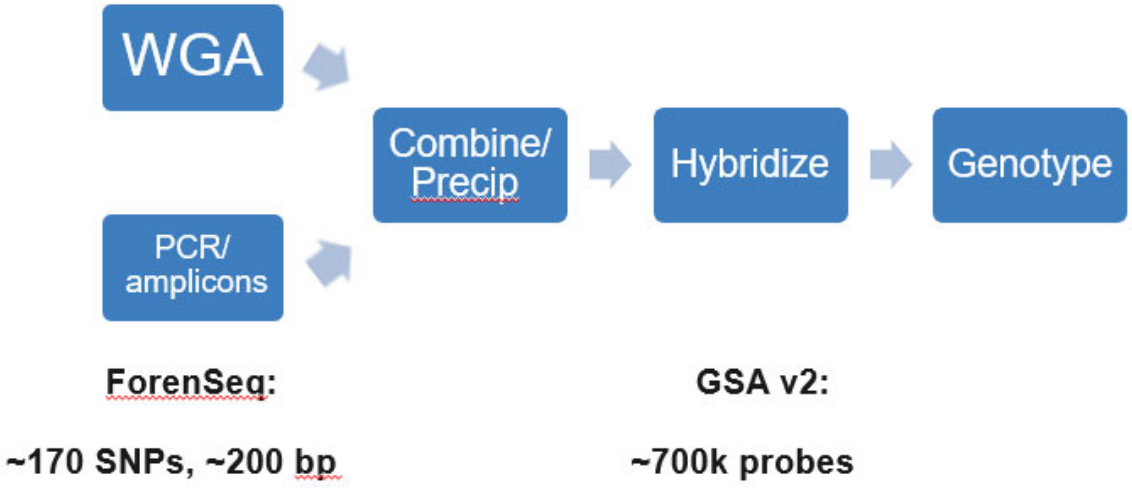


Figure 3: SNP plot for SNP rs4833103 showing addition of ForenSeq amplicons into same sample increases intensity, and adding ForenSeq amplicon into whole genome amplification (WGA) of another sample shows shift to the ForenSeq call [first letter is sex of donor of the WGA sample (F or M), and second letter is sex of the donor of ForenSeq amplified material].

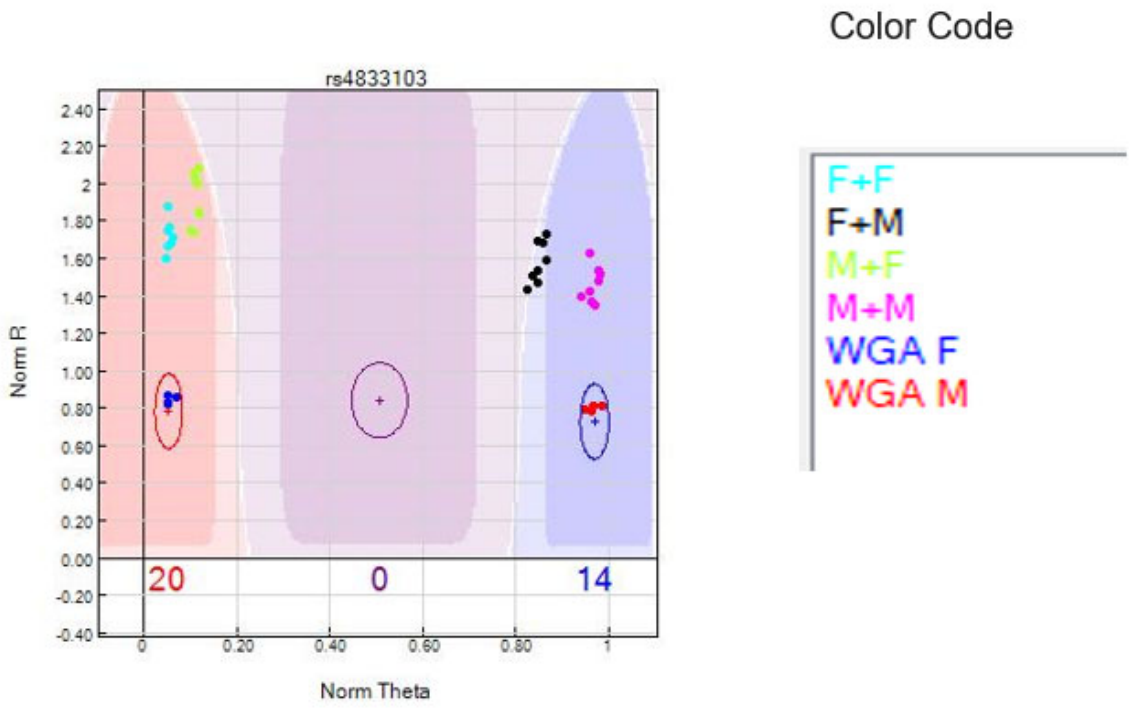


Figure 4: Three steps in the Infinium HTS microarray workflow where PCR amplicons were added and subsequently compared.

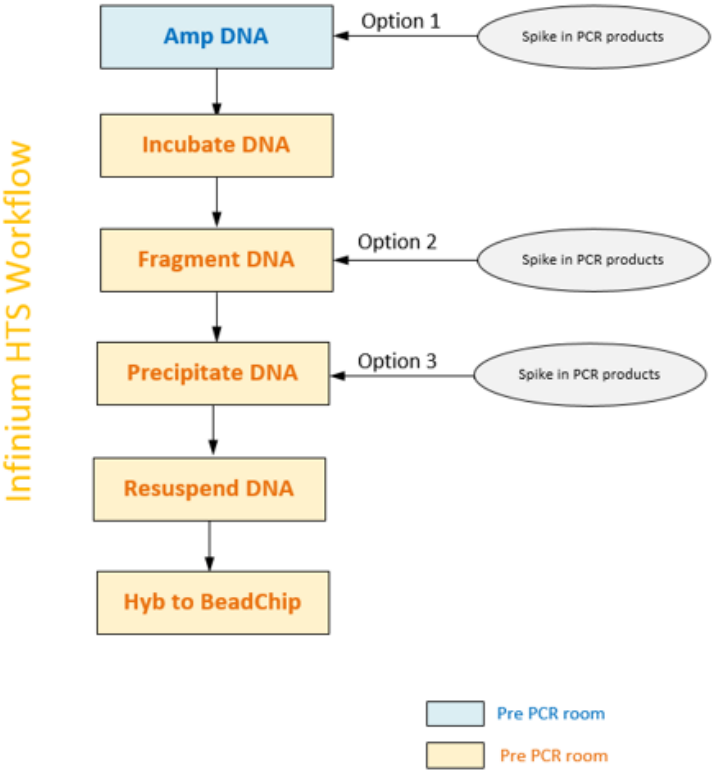


Figure 5: Titration of amplicons at the WGA step with two samples with the same call at SNP rs4959270 showing an increase in signal intensity and a slight theta shift. With a new cluster file generated, the intensity and theta shift would be normalized to account for the increase in amplicon at this location. Here, the samples were titrated in replicates of 30 from 15 ng to 300 ng where F is female and male is M, and WGA is the standard control workflow. The first letter of each pair denotes the standard control workflow sample and the sex associated with that sample, the second letter of the pair denotes the sex of the amplified material spiked into the WGA workflow.

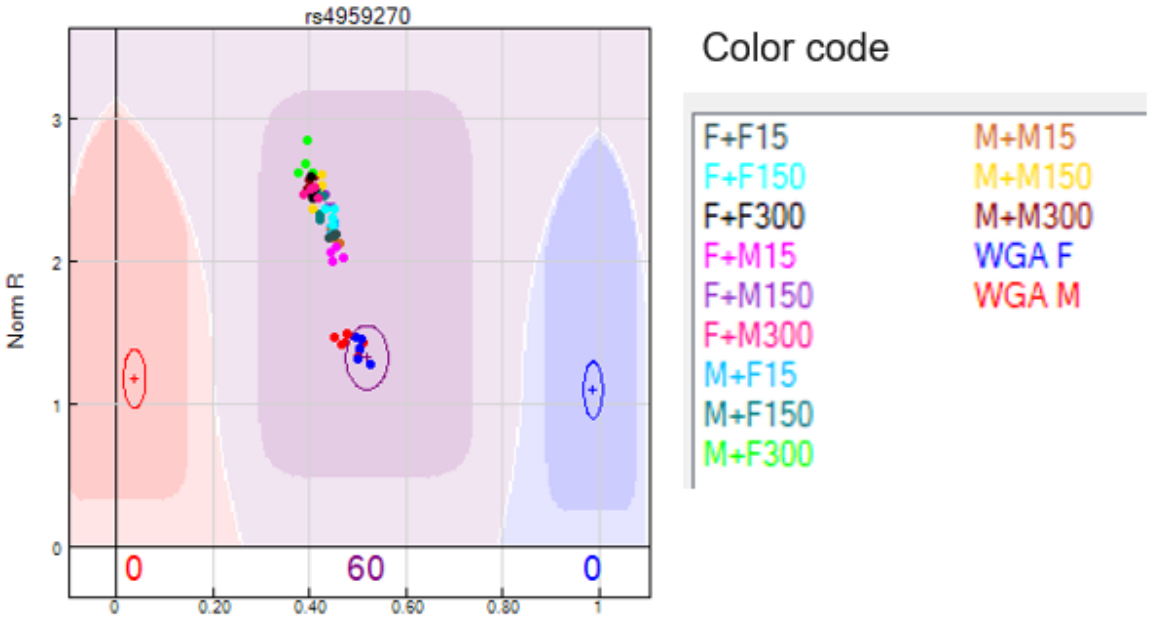


Figure 6: A titration of 300 ng to 15 ng showing that adding more amplified DNA was able to shift the call to the truth call of the ForenSeq amplified contributor in SNP rs7041158. F is female and M is male and WGA is the standard control workflow. The first letter of each pair denotes the standard control workflow sample and the sex associated with that sample, the second letter of the pair denotes the sex of the amplified material spiked into the WGA workflow.

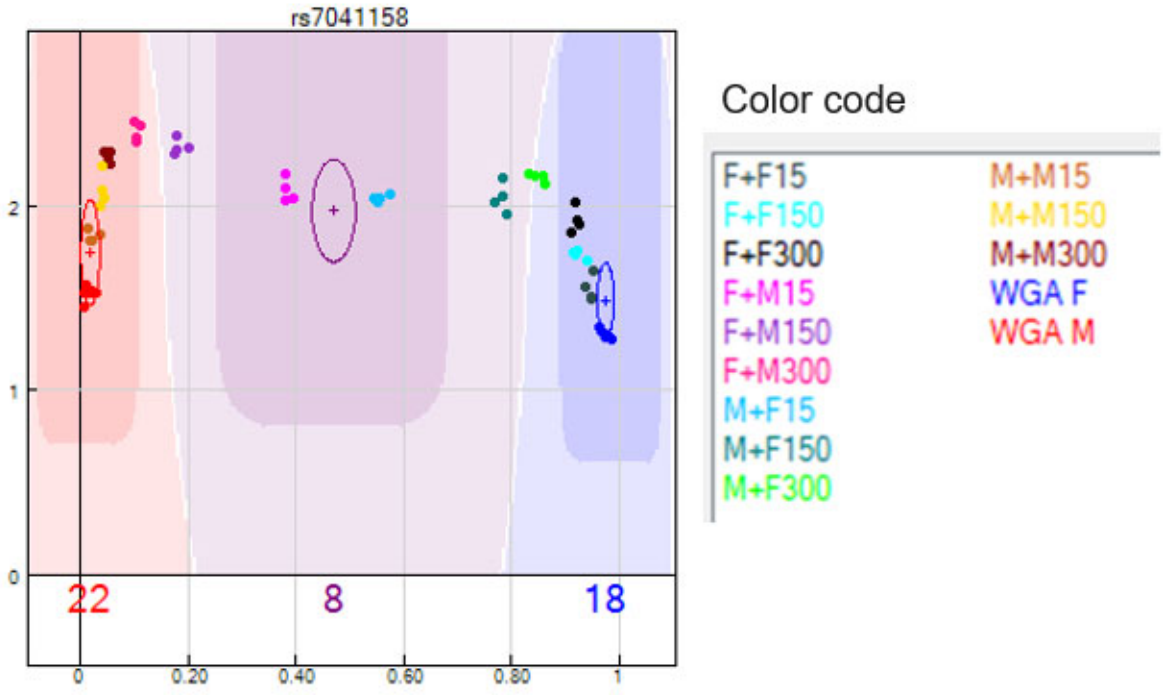


Figure 7: Staging experiment layout where WGA plate was staged for eight hours at different temperatures with and without the multi-sample amplification mix (MSM), and then amplified material was added with or without a 1:1 SPRI clean up first.

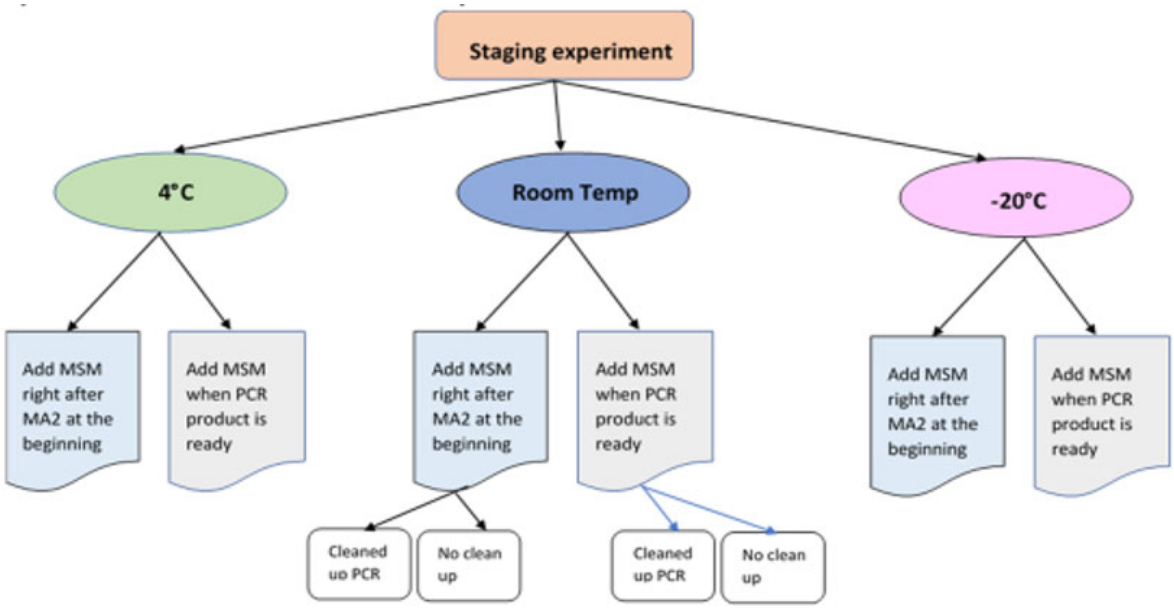


Figure 8: Mean call rate for WGA only staging experiment (no spiked amplicon) showing the WGA plate can be staged at varying temperatures with no effect to the overall sample nor by sex (F is Female and M is male).

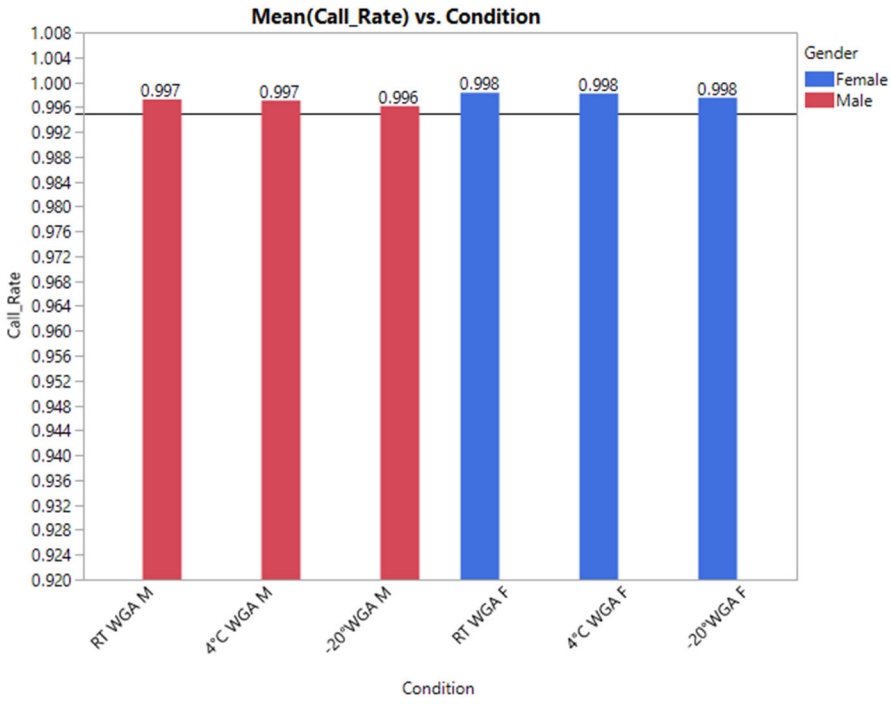


Figure 9: Comparison of (WGA Male +PCR Male) or (WGA Female + PCR Female) showing no difference found between temperature of staging when additional amplified material was added to the WGA. M is male and F is female to differentiate sex of samples.

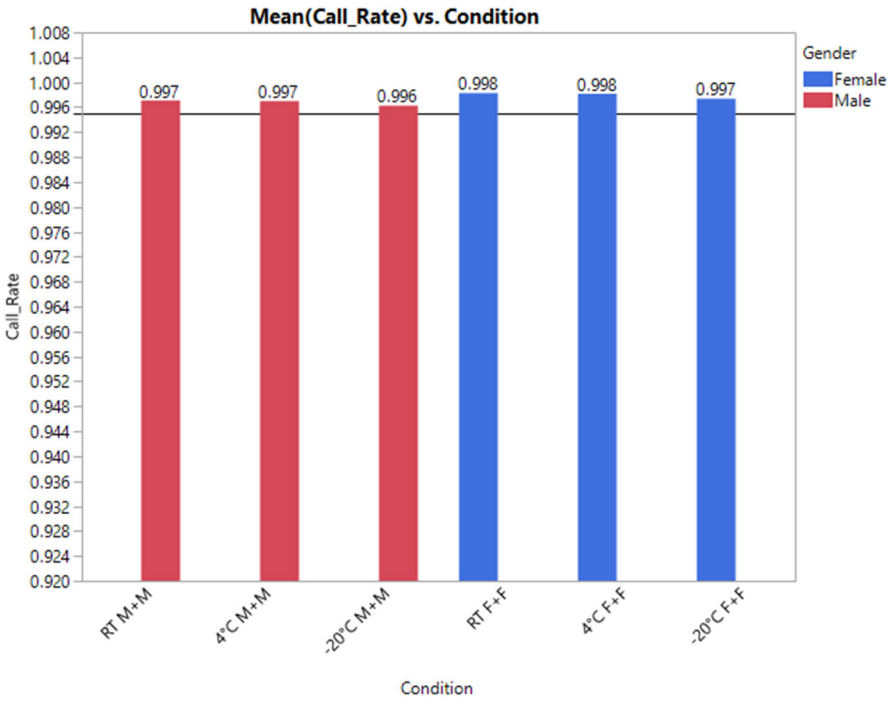


Figure 10: No differences were found between staging of plates at room temperature, 4˚C, or -20˚C, as well as whether the reagent MSM which houses the polymerase was staged with the plate or added directly prior to PCR. Here 30 replicates of each condition were performed (n = 30).

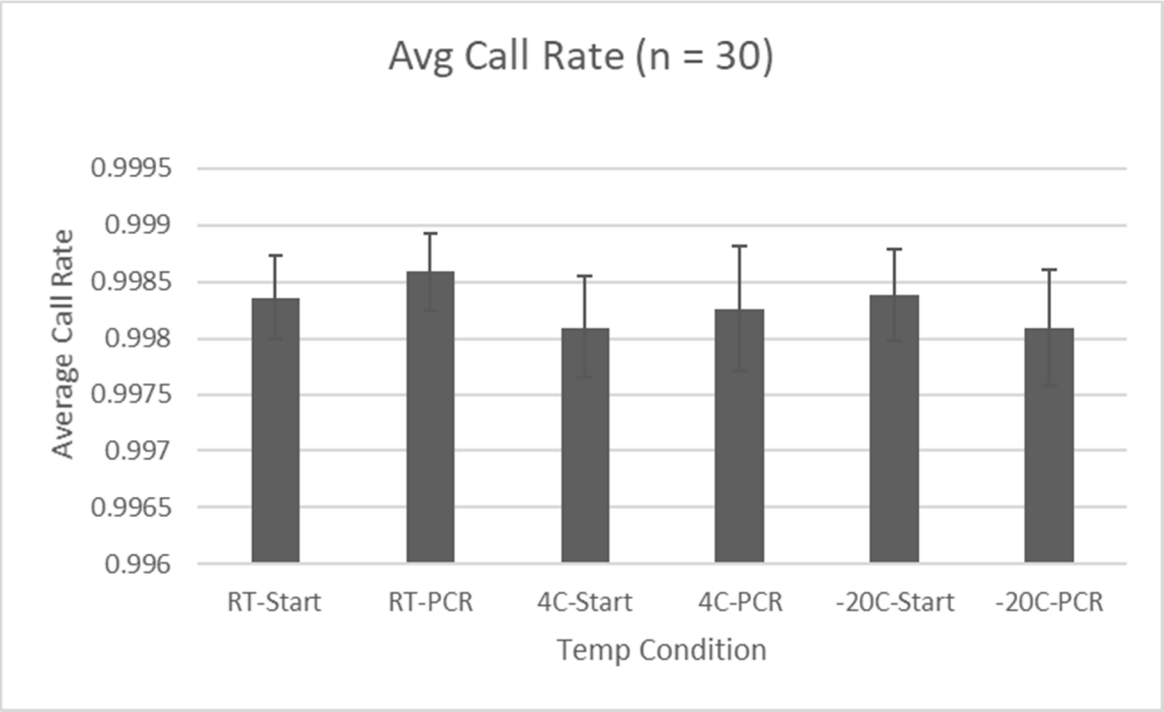


Figure 11: SNP rs1357617 displayed SNP conversion when WGA plate was staged with MSM for eight hours and 50 ng of ForenSeq amplified material were added without a clean up step. The male WGA shifted from AA to BB when a female with an BB call of amplified material was added. The female WGA shifted from BB to AA when the male amplified material was added. F is female and M is male to differentiate samples in this experiment.

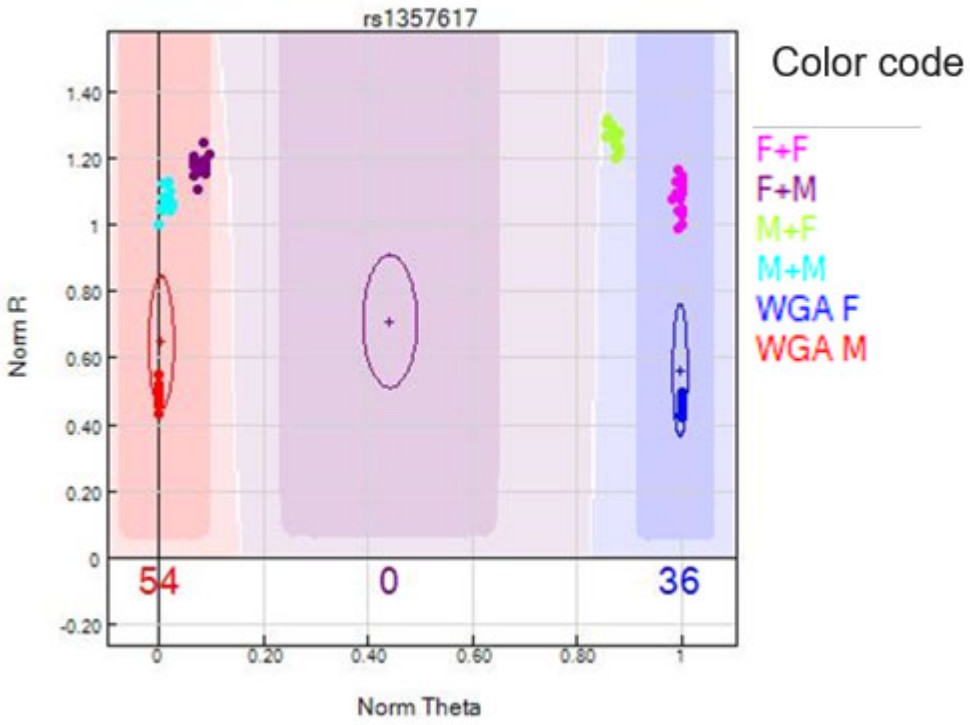


Figure 12: SNP shifts at SNP rs35742686 observed when fragment A amplicon of *CYP2D6* was spiked directly into WGA showing a WGA male at AA shifting to AB the female call when female amplified material was added. The WGA female shifts from AB to when male amplified material was added. F is female and M is male to differentiate samples in this experiment.

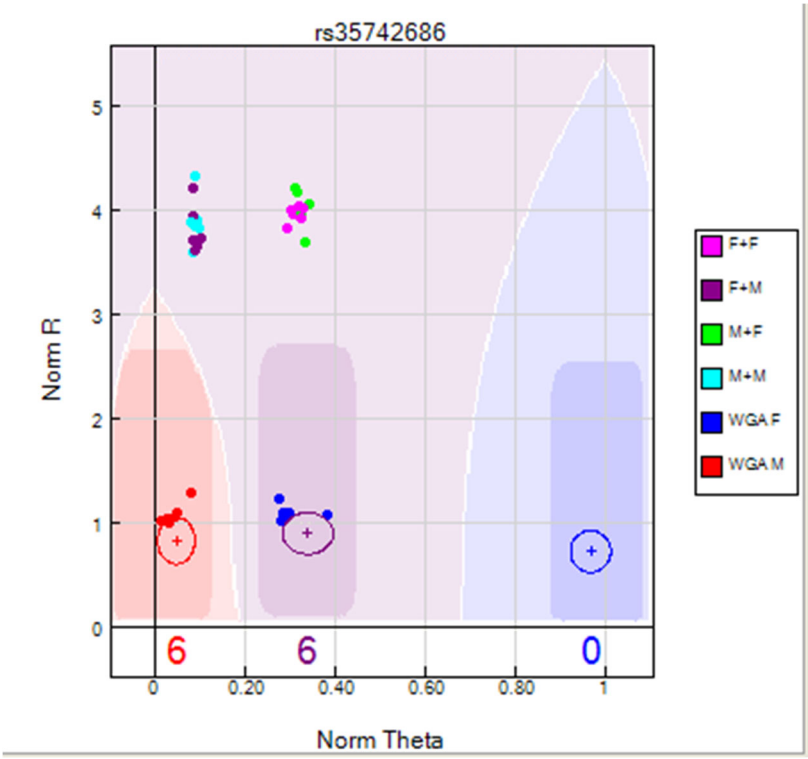
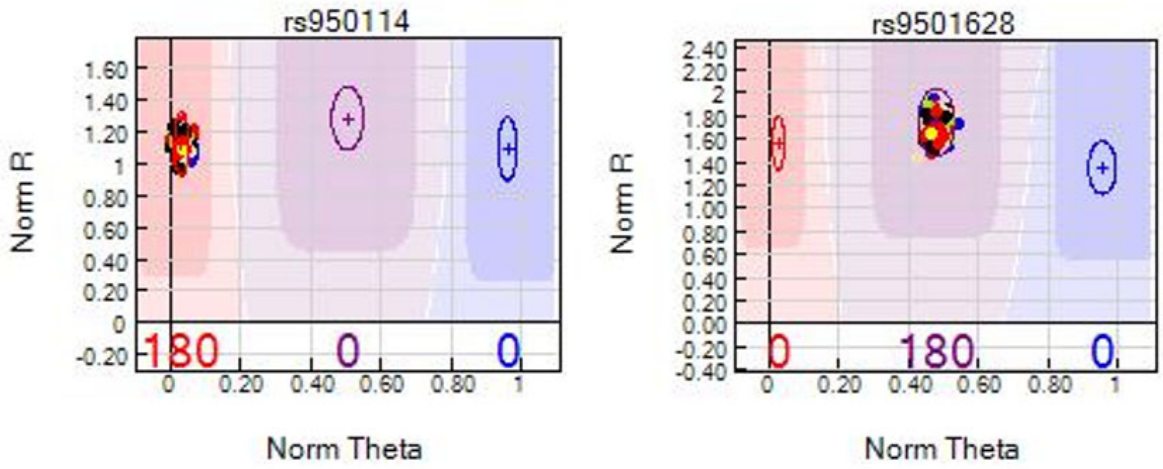


Figure 13: No shift in intensity of SNPs nor a theta shift to the left or right of the intended cluster location for SNPs not targeted by specific PCR as shown by SNPs outside of the amplified region.



Appendix:

Supplementary Table 1. SNPs in both GSA v2 and ForenSeq Universal Assay

Overlapping SNPs in ForenSeq and GSA v2

rs1005533	rs1528460	rs2593595	rs722098
rs1015250	rs1572018	rs260690	rs7226659
rs1024116	rs159606	rs279844	rs7251928
rs1028528	rs16891982	rs2814778	rs727811
rs1042602	rs1736442	rs2830795	rs729172
rs10488710	rs174570	rs2831700	rs733164
rs10497191	rs17642714	rs28777	rs735480
rs1058083	rs1800407	rs2920816	rs737681
rs10773760	rs1800414	rs310644	rs740598
rs10776839	rs1805005	rs321198	rs740910
rs1079597	rs1805006	rs3737576	rs7554936
rs1109037	rs1805008	rs3780962	rs763869
rs1110400	rs1805009	rs3811801	rs7657799
rs11547464	rs1834619	rs3814134	rs798443
rs11652805	rs1871534	rs3823159	rs7997709
rs12203592	rs1876482	rs3827760	rs8037429
rs1229984	rs1886510	rs430046	rs870347
rs12498138	rs192655	rs4364205	rs873196
rs12821256	rs1979255	rs4411548	rs876724

rs12896399	rs200354	rs445251	rs891700
rs12913832	rs2024566	rs4471745	rs901398
rs1294331	rs2040411	rs4530059	rs907100
rs12997453	rs2042762	rs459920	rs914165
rs13182883	rs2046361	rs4606077	rs917115
rs13218440	rs2056277	rs4833103	rs917118
rs1335873	rs2076848	rs4891825	rs938283
rs1336071	rs2107612	rs4918664	rs9522149
rs1355366	rs2111980	rs4959270	rs964681
rs1357617	rs214955	rs560681	rs987640
rs1382387	rs2166624	rs576261	rs9905977
rs1393350	rs2196051	rs6444724	rs993934
rs1413212	rs221956	rs671	rs9951171
rs1426654	rs2228479	rs6754311	
rs1454361	rs2238151	rs6811238	
rs1462906	rs2342747	rs683	
rs1463729	rs2378249	rs6955448	
rs1490413	rs2399332	rs7041158	
rs1493232	rs2402130	rs717302	
rs1523537	rs251934	rs719366	

References:

[1] Ingelman-Sundberg M, Rodriguez-Antona C. Pharmacogenetics of drug metabolizing enzymes: implications for a safer and more effective drug therapy. *Philos Trans R Soc Lond B Biol Sci* 360:1563-1570 [2005].

[2] Bradford LD. *CYP2D6* allele frequency in European Caucasians, Asians, Africans and their descendents. *Pharmacogenomics* 3: 229-243 [2002].

[3] Kirchheiner J, Nickchen K, *et al.* Pharmacogenetics of antidepressants and antipsychotics: the contribution of allelic variations to the phenotype of drug response. *Mol Psychiatry* 9: 442-473 [2004].

[4] Eichelbaum M, Ingelman-Sundberg M, Evans WE. Pharmacogenomics and individualized drug therapy. *Annu Rev Med* 57: 119-137 [2006].

[5] Zhou SF. Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. *Clin Pharmacokinet* 48: 689-723 [2009].

[6] Zhou SF. Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part II. *Clin Pharmacokinet* 48: 761-804 [2009].

[7] Ingelman-Sundberg M and Rodriguez-Antona C. Pharmacogenetics of drug-metabolizing enzymes: implications for safer and more effective drug therapy. *Philos Trans Soc Lond B Biol Sci* 360(1460):1563-70 [2005].

[8] Leppert W. *CYP2D6* in the metabolism of opioids for mild to moderate pain. *Pharmacology* 87(5-6):274-85 [2011].

[9] Ingelman-Sundberg M, Sim SC, Gomez A, and Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies; pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol Ther* 116(3):496-526 [2007].

[10] Kimura S, Umeno M, Skoda RC, Meyer UA, and Gonzalez FJ. The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic *CYP2D6* gene, a related gene, and a pseudogene. *Am J Hum Genet* 45(6): 889-904 [1989].

[11] The Human Cytochrome P450 (CYP) Allele Nomenclature Database. Available: http://www.cypalleles.ki.se/. Accessed 15 Jan 2015.

[12] Bernard S, Neville KA, Nguyen AT, and Flockhart DA. (2006) Interethnic differences in genetic polymorphisms of *CYP2D6* in the U.S. population: clinical implications. *Oncologist* 11:126-135 [2006].

[13] Teh LK and Bertilsson L. Pharmacogenomics of *CYP2D6*: molecular genetics, interethnic differences and clinical importance. *Drug Metab Pharmacokinet* 27:55-67 [2012].

[14] Sistonen J, Sajantila A, Lao O, *et al*. *CYP2D6* worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenet Genomics* 17: 93-101 [2007].

[15] Wendt FR, Sajantila A, Moura-Neto RS, Woerner AE, and Budowle B. Full-gene haplotypes refine *CYP2D6* metabolizer phenotype inferences. *Int J Legal Med* 132(4):1007-1024 [2018].

[16] Wendt FR, Sajantila A, and Budowle B. Predicted activity of UGT2B7, ABCB1, OPRM1, and COMT using full-gene haplotypes and their association with the *CYP2D6*-inferred metabolizer phenotype. *Forensic Sci Int Genetics* 33:48-58 [2018].

[17] Yuanyuan Dong, Huasheng Xiao, Qi Wang, *et al*. Analysis of genetic variations in CYP2C9, CYP2C19, *CYP2D6* and CYP3A5 genes using oligonucleotide microarray. *Int J Clin Exp Med* 8(10): 18917-18926 [2015].

PART 3

CONCLUSION

CHAPTER 5

Concluding Remarks on Future Directions

Summary

This dissertation describes development of a targeted panel for clinically relevant variants in the *CYP2D6* gene using array-based technology. The goal of this dissertation was to assess and develop capability to analyze gene data by MPS. Armed with this capability, I sequenced and analyzed *CYP2D6* and showed that variants may have metabolic consequences that could be detected. However, it is recognized that MPS will not be readily implemented into diagnostic laboratories and therefore microarray technology was further investigated, particularly pseudogene (i.e. paralog) assay interference of the *CYP2D6* gene was explored. A PCR assay workflow leveraging microarray technology was developed to quickly and efficiently screen individuals of interest by overcoming paralog interference. This assay can be used prior to administering drugs or post mortem to gain information about potential adverse drug reactions.

Specific aim 1 tested a model using the hypervariable regions, HV1 and HV2, of the mitochondrial genome to demonstrate massively parallel sequencing can reliably sequence challenging targets. This mtDNA protocol demonstrated the robustness of MPS by accurately sequencing homopolymer C stretches, discerning point heteroplasmies, and deciphering length heteroplasmies. The knowledge gained through that protocol optimization was used to develop a custom capture assay to enrich the full region of the gene *CYP2D6* for sequence analysis in specific aim 2. The custom capture assay was designed to capture not just the gene in question, but 5000 bp on either side to account for any variants in the promoter region or outside the gene itself that could change the protein function. This custom capture probe assay was used in specific aim 3 to completely sequence a Caucasian sample population and analyze the genetic variability against a database of known variants. This effort provided increased knowledge on how to analyze large complex data sets. These studies determined a baseline of variation for the *CYP2D6* gene in a

specific local Texas population. A majority of these variants were rare in this population. Larger data sets likely will find even more variants. These SNPs were compared against a known database of clinically relevant samples with known metabolic responses of the same ethnic background to verify actionable variants, where it became apparent that a novel SNP that could be potentially damaging was discovered in specific aim 3. While finding demonstrates the power of MPS, the translation and implementation of MPS into standard high-throughput laboratories that would test these types of variants will be difficult. Microarray technology is simpler to implement, thus moving platforms to microarray for further testing was prudent. Even though technology is easier to implement, it has a major challenge in that paralog interference makes it difficult to accurately type genes like *CYP2D6*. Thus, a screening microarray workflow was designed to analyze the clinically relevant SNPs that overcome paralog interference in specific aim 4. This workflow added an upfront PCR where that PCR product was seamlessly added into the standard WGA workflow with little added time and cost and overcame paralog interference to accurately call previously ambiguous variants.

With 1.6 million ADRs occurring each year in the United States it is imperative that solutions to reduce ADRs must occur (Agency, 2014). The safety of patients is paramount, and effort to strive toward the best health care possible should be pursued. While there are many factors that affect ADRs, the genetic component is one that may be addressed analytically. While the majority of drugs work as intended for a majority of the population, there are a notable portion of the population in which there is either no effect or they experience a negative effect. The current method of drug administration is by empirical trial and error. If a screening assay can help identify the right drug at the right dose without waiting to observe the effect, a safer outcome can occur. A quick screening array such as the one described in this dissertation can be performed on a patient

during routine examination or when admitted to the hospital arming physicians with another piece of information to make a better decision for drug therapy. This could reduce the number of ADRs that occur prior to or during hospital stays.

For example, the drug Rezulin (generic: Troglitazone) was released on the market in 1997 only to be quickly pulled from the market in 2000. Rezulin was developed as the first anti-diabetic drug that enhanced insulin sensitivity. It was pushed to the market quicker than normal because it was the only drug potentially available for treatment of diabetes that at the time could not be controlled by medication. This drug was found to cause severe hepatotoxicity and was pulled from the United States market in 2000, leaving patients with diabetes without a way to control their disease. A Food and Drug Administration epidemiologist determined that patient monitoring was not effective in protecting against liver failure, estimating that the drug could be linked to over 430 liver failures and that patients incurred 1,2000 times greater risk of liver failure while taking Rezulin (Willman, 2000). Upon further investigation it was determined that the metabolism of Rezulin eventually formed quinone and o-quinone methide (Dixit *et al*., 2011). Quinone is known to be metabolized by the CYP450 super family, and the cause of multiple ADRs across drugs (Hughes *et al*., 2017). Genetic testing of patients before taking a drug that may metabolize to something with known ADRs like quinone could reduce the number of people with a negative effect while still allowing a drug that worked well for many to benefit from it. An example of this same drug is that it was recently determined that a derivative of Troglitazone can help reduce proliferation in breast cancer cells (Salamone *et al.*, 2012). This drug has the potential to have far reaching benefits in the battle against breast cancer, but with the outcome of Troglitazone previously, it may never come to fruition without some preliminary testing to determine if a patient may react poorly to quinone or quinone derivatives. An expanded screening assay similar to the

one described in this dissertation could aid in this quest to safely allow drugs to market that would otherwise be lost to all. This test could be performed in a standard procedure for patients admitted to a hospital or prior to prescribing the drug to determine the best course of action for the patient.

Additionally, screening processes such as the one described here could be used during clinical trials to aid in pharmaceutical companies' development of drugs. When the absorption, distribution, metabolism, and excretion (ADME) of a drug is well established, then a genetic screening test of multiple genes could be developed covering the entire ADME pathway as multiple genes are likely to have an effect. If this screening process can be developed in tandem with clinical trials, then more drugs that are effective for a majority of the population can make it to the market with significantly less risk by coupling a screening test before a prescription can be filled to ensure the drug in question can be effective with the patient's genetic makeup. This can reduce research and development costs for pharmaceutical companies, potentially reducing the price of the drug, and allowing a path forward for drugs that are beneficial for a majority of the population. Both microarray and MPS technologies are capable of testing multiple genes simultaneously, and this dissertation provides a proof of concept that with further development either could be viable for such an assay.

Limitations

The presented dissertation is limited by two factors. First, is the selection of samples for a better understanding of *CYP2D6* SNPs. The population of local Texas Caucasians had self-reported ethnicity which may lead to a bias. There may need to be more samples sequenced to confidently identify all common variants in a population. The other limitation with the population

chosen is all were self-reported healthy individuals. While it is entirely possible those with a metabolic change are in the population tested, it may be beneficial to test a larger data set of samples with known metabolic issues to gain a more complete data set of damaging variants.

While the seminal work in this dissertation was able to overcome pseudogene interference to more accurately call clinically relevant *CYP2D6* SNPs, there are still limitations associated with this work. This work did not address CNV calling in either the sequencing or microarray portion. Caucasians with 0, 1, or more than 3 copies (all other than the standard 2 copies) accounts for 12.2% of the population, thus determining CNVs in an important diagnostic factor for this population. Accurate CNV calling is challenging with both technologies and will need significant further work to accomplish accurate typing. It should be a future work as knowing the clinically relevant variant exists is only a portion of the story needed to understand how an individual may metabolize drugs. Also, this study did not address all known variants of *CYP2D6*. This will take significant further work to interrogate all known variants on this gene but is certainly warranted as many companies are currently moving toward this.

Topic-specific future work

For future work, a larger population of healthy Caucasians as well as Caucasians with known metabolic changes should be sequenced to ensure all variants have been characterized. Next, this method should be extended to other populations. *CYP2D6* is known to have population specific variants, so to be beneficial across all population groups it would be helpful to have all major population groups included. Collaborations with population sequencing groups such as the 100,000 genomes project in the UK or Yale's new clinical medicine sequencing initiative could allow for data to be mined quicker and more efficiently. A specific probe design could be

accomplished and added to the current beadchip to interrogate a more complete list of *CYP2D6* variants. A multiplex targeted PCR to address the pseudogene interference while being cognizant of the CNV probes to still glean CNV probative value could be performed. Future studies could work on longer indels to be interrogated by microarrays. In theory, longer stretches of insertions and deletions can be accomplished, and with such complex genes such as *CYP2D6*, this could help obtain all the data of known variants that sequencing can accomplish for a much lower price. Once these are resolved, they can be expanded to the other major pharmacogenetic genes such as *CYP1A2*, *CYP2C9*, *CYP2C19*, *CYP2D6*, *CYP3A4*, and *CYP3A5* to cover the metabolism of 90% of the drugs on the market. It could be beneficial to collaborate with a pharmaceutical company during clinical trials to have controlled known test groups with metabolic information to further elucidate the interactions of *CYP2D6* on drug pathways. This could have far reaching effects because if it is known what portion of the population reacts negatively to a particular drug, a screening assay could be applied to ensure the right patient is receiving this particular medicine.

Conclusions

In this dissertation, the importance of sequencing the entire region of a gene and its surrounding areas was demonstrated with the number of variants found outside the exome, and discovery of a novel variant that has potentially damaging effects on protein functionality. Also, a novel low-cost screening method of *CYP2D6* on the Illumina microarray GSA v2 beadchip was demonstrated to be effective in capturing true SNP calls for *CYP2D6* by outperforming the background of the homologs *CYP2D7P* and *CYP2D8P* to generate a more accurate call, i.e., reducing background noise. This process could be implemented in laboratories already running microarrays with a reasonable increase in turnaround time and could be used to assay *CYP2D6* in

situations where genetic analyses could assist in drug prescription or postmortem in situations where drug exposure may be a triggering event in the cause of death.

References:

Agency for Healthcare Research and Quality "Saving lives and saving money: hospital-acquired conditions update: interim data from national efforts to make care safer, 2010-2014." Agency for Healthcare Research and Quality. https://www.ahrq.gov/sites/default/files/publications/files/interimhacrate2014_2.pdf

Dixit V, Bharatam P. "Toxic Metabolite Formation from Troglitazone (TGZ): New Insights from a DFT Study." *Chemical Research in Toxicology.* 2011;24(7):1113-1122.

Hughes T, Swamidass S. "Deep Learning to predict the formation of quinone species in drug metabolism." *Chem Res Toxicol.* 2017;30(2):642-656.

Salamone, S, Colin C, Grillier-Vuissoz I, Kuntz S, Mazerbourgh S, *et al.* "Synthesis of new troglitazone derivatives: Anti-proliferative activity in breast cancer cell lines and preliminary toxicological study." *European Journal of Medicinal Chem*. 2012;51:206-215.

Willman D. "The Rise and Fall of the Killer Drug Rezulin." *Los Angeles Times*. 2000;June4.

LIST OF PUBLICATIONS

(Asterisks indicate dissertation chapters)

***Davis CP** and Budowle B. "Cost Effective Method of Screening *CYP2D6* with Microarray Technology." *Biotechniques.* In Review.

Jaeger AC, Alvarez ML, **Davis CP**, Guzman E, Han Y, Way L, Walichiewicz P, Silva D, Pham N, Caves G, Bruand J, Schlesinger F, Pond SJ, Varlaro J, Stephens KM, and Holt CL. "Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories." (2017) *Forensic Sci Int Genet.* May; 28: 52-70.

Warshauer DH, **Davis CP**, Holt C, Han Y, Walichiewicz P, Richardson T, Stephens K, Jager A, King J, and Budowle B. "Massively parallel sequencing of forensically relevant single nucleotide polymorphisms using TruSeq™ forensic amplicon." (2015) *International Journal of Legal Medicine* 129(1): 31-6.

***Davis C**, Peters D, Warshauer D, King J, and Budowle B. "Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Improved methods for DNA samples encountered in forensic testing" (2015) *Legal Medicine* 17(2):123-7.

King J, LaRue B, Novroski N, Stoljarova M, Seo S, Warshauer D, **Davis C**, Zeng X, Parson W, Sajantila A, Budowle B. "High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq" (2014) *Forensic Science International: Genetics* 12:128-35.

Purps J, Siegert S, Willuweit S, Nagy M, Alves C, Salazar,…**Davis C**,…Roewer L. "A global analysis of Y-chromosomal haplotype diversity for 23 STR loci" (2014) *Forensic Science International: Genetics* 12:12-23.

Budowle B, Warshauer DH, Seo SB, King JL, **Davis C**, LaRue B. "Deep Sequencing Provides Comprehensive Multiplex Capabilities" (2013) *Forensic Science International: Genetics Supplement Series* 4:e334-e335.

Seo SB, King JL, Warshauer DH, **Davis CP**, Ge J, and Budowle B. "Single nucleotide polymorphism typing with massively parallel sequencing for human identification." (2013) *International Journal of Legal Medicine* DOI 10.1007/s00414-013-0879-7

Warshauer DH, Lin D, Hari K, Jain R, **Davis C**, LaRue B, King JL, and Budowle B. "STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data." (2013) *Forensic Science International: Genetics* 7:409-417.

Davis CP, Ge J, Sprecher C, Chidambaram A, Thompson J, Ewing M, Fulmer P, Rabbach D, Storts D, and Budowle B. "Prototype PowerPlex® Y23 System: A Concordance Study" (2013) *Forensic Science International: Genetics* 7(1):204-208.

Davis CP, Illescas M, Tirado C, Lopez R, Budowle B, and Dawson Cruz T. "A Case of Amelogenin Y-null: A simple primer binding site mutation or unusual genetic anomaly?" (2012) *Legal Medicine* 14(6):320-323.

Li L, Ge J, Zhang S, Guo J, Zhao S, Li C, Tang H, **Davis C**, Budowle B, Hou Y, and Liu Y. "Maternity exclusion with a very high autosomal STRs kinship index." (2012) *International Journal of Legal Medicine* 126(4):645-648. DOI: 10.1007/s00414-012-0668-8

Davis CP, Ge J, King J, Malik N, Weirich V, Eisenberg AJ, and Budowle B. "Variants observed for STR locus SE33: A concordance study." (2012) *Forensic Science International: Genetics* 6(4):494-497. doi:10.1016/j.fsigen.2011.12.002

Davis CP, King JL, Budowle B, Eisenberg AJ, and Turnbough MA. "Extraction platform evaluations: A comparison of AutoMate Express™, EZ1® Advanced XL, and Maxwell® 16 Bench-top DNA extraction systems." (2012) *Legal Medicine*. 14(1):36-39. doi:10.1016/j.legalmed.2011.09.005

Davis CP, Chelland LA, Pavlova VR, Illescas MJ, Brown KL, Dawson Cruz T. "Traditional Multiplex STR Amplification of Low Template DNA Samples with the Addition of Proofreading Enzymes." (2011) *Journal of Forensic Sciences*. 56(3):726-32 DOI: 10.1111/j.1556-4029.2011.01727.x

Davis CP, Ge J, Chidambaram A, King J, Turnbough M, Collins M, Dym O, Chakraborty R, Eisenberg AJ, and Budowle B. "Y-STR Loci Diversity in Native Alaskan Populations" (2011) *International Journal of Legal Medicine.* 125(4):559-63 DOI 10.1007/s00414-011-0568-3.

Sajantila A, Palo JU, Ojanperä I, **Davis CP**, Budowle B. "Pharmacogenetics in Medico-Legal Context." (2010) *Forensic Science International.* 203:44-52. DOI: 10.1016/j.forsciint.2010.09.011. (Review article)