Int J Legal Med (2018) 132:1007–1024
https://doi.org/10.1007/s00414-017-1709-0

ORIGINAL ARTICLE

Full-gene haplotypes refine CYP2D6 metabolizer phenotype inferences

Frank R. Wendt[1,2] · **Antti Sajantila**[3] · **Rodrigo S. Moura-Neto**[4] · **August E. Woerner**[1,2] · **Bruce Budowle**[1,2,5]

Received: 19 June 2017 / Accepted: 11 October 2017 / Published online: 26 October 2017

Abstract CYP2D6 is a critical pharmacogenetic target, and polymorphisms in the gene region are commonly used to infer enzyme activity score and predict resulting metabolizer phenotype: poor, intermediate, extensive/normal, or ultrarapid which can be useful in determining cause and/or manner of death in some autopsies. Current genotyping approaches are incapable of identifying novel and/or rare variants, so *CYP2D6* star allele definitions are limited to polymorphisms known a priori. While useful for most predictions, recent studies using massively parallel sequencing data have identified additional polymorphisms in *CYP2D6* that are predicted to alter enzyme function but are not considered in current star allele nomenclature. The 1000 Genomes Project data were used to produce full-gene haplotypes, describe their distribution in super-populations, and predict enzyme activity scores. Full-gene haplotypes generated lower activity scores than current approaches due to inclusion of additional damaging polymorphisms in the star allele. These findings are critical for clinical implementation of metabolizer phenotype prediction because a fraction of the population may be incorrectly considered normal metabolizers but actually may be poor or intermediate metabolizers.

Keywords *CYP2D6* · Full-gene haplotypes · Metabolizer phenotype · Massively parallel sequencing

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00414-017-1709-0) contains supplementary material, which is available to authorized users.

✉ Frank R. Wendt
Frank.Wendt@my.unthsc.edu

[1] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., CBH-250, Fort Worth, TX 76107, USA

[2] Graduate School of Biomedical Sciences, University of North Texas Health Science Center, 3500 Camp Bowie Blvd.,CBH-250, Fort Worth, TX 76107, USA

[3] Laboratory of Forensic Biology, Department of Forensic Medicine, University of Helsinki, P.O Box 40, 00014 Helsinki, Finland

[4] Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941, Brazil

[5] Center of Excellence in Genomic Medicine (CEGMR), King Abdulaziz University, Jeddah 22252, Saudi Arabia

Introduction

The cytochrome p450 family 2 subfamily D polypeptide 6 (CYP2D6) enzyme is part of a large family of CYPs responsible for a substantial portion of hepatic phase I metabolism of foreign compounds and endogenous toxins [1, 2]. The enzyme has been implicated in drug metabolism variation of clinical and medico-legal relevance [3, 4]. Current methods of pharmacogene analyses employ targeted approaches, such as genome wide association studies of candidate single nucleotide polymorphisms (SNPs) and SNP-targeted massively parallel sequencing (MPS) of known a priori SNPs and/or insertion/deletion polymorphisms (INDELs) [5]. The *CYP2D6* allele nomenclature, in the Human Cytochrome P450 Allele Nomenclature Database (HCYPAND) [6], identifies and defines polymorphisms that confer each *CYP2D6* star (*) allele (the collection of polymorphisms within the gene region). *CYP2D6* * allele genotypes (g), or diplotypes, commonly provide inferences of metabolizer phenotype (MP; poor (gPM), intermediate (gIM), normal/extensive (gNM/gEM), or ultrarapid (gUM) CYP2D6 activity) [1, 2, 6, 7]. It has been demonstrated both clinically and medico-legally that *CYP2D6* information can result in increased prescription

efficacy and inference of idiosyncratic effect during accident reconstruction [4, 8–11]. In fact, Koren et al. [4] and Koski et al. [8, 9] have used targeted genotyping of *CYP2D6* to make inferences regarding the cause and/or manner of death in a series of medico-legal cases. However, targeted genotyping approaches inherently are incapable of revealing novel polymorphisms that may further refine *CYP2D6* * alleles and the associated metabolic differences observed in clinical cases and medico-legal investigations. MPS of the full gene region has the potential to reveal additional polymorphisms and refine predictions of CYP2D6 activity.

Wendt et al. [12] characterized 418 polymorphisms in *CYP2D6* exons, introns, 3′ and 5′ untranslated regions, and promoter region in the 1000 Genomes Project dataset. Ninety-seven (23.2%) of the polymorphisms are currently used by HCYPAND to classify some of the < 150 *CYP2D6* * alleles observed to date. The remaining 321 polymorphisms have a wide range of allele frequencies in the African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) super-populations. Most notable are those polymorphisms predicted to damage, or most likely damage, CYP2D6 function. Ignoring these loci when determining *CYP2D6* * alleles may lead to inaccurate predictions of enzyme function and incorrect conclusions for drug therapy or potential cause and/or manner of death investigations. The HCYPAND database has established inclusion criteria for newly observed CYP alleles [13] which put more emphasis on enzyme-altering polymorphisms for the sake of providing minimalist nomenclature. However, in medico-legal autopsy cases where incident reconstruction and cause and/or manner of death may be inferred from toxicology and genetic data [4, 8, 9], a more comprehensive and inclusive nomenclature may be necessary. Herein, all typed loci from the 1000 Genomes Project data were used to generate and characterize full-gene *CYP2D6* haplotypes. These data indicate that the phenotypic impact of some previously defined *CYP2D6* * alleles may have been mischaracterized due to damaging polymorphisms occurring elsewhere in the gene relative to a HCYPAND causal SNP or INDEL. Mischaracterizations based on targeted genotyping and traditional * allele nomenclature may have clinical and medico-legal consequences, as a fraction of 1000 Genomes Project samples may be inaccurately placed into MP categories.

Materials and methods

Polymorphisms in the *CYP2D6* gene region (introns, exons, 5′ and 3′ untranslated regions (UTRs), and promoter) were downloaded from phase 3 of the 1000 Genomes Project [14, 15] and analyzed individually in 5 super- and 26 sub-populations (Table S1) according to Wendt et al. 2017 [12]. 1000 Genomes Project *CYP2D6* haplotypes containing reportedly phased polymorphisms [16] were aligned to the hg38 and hg19 reference genomes and the M33388 GenBank accession reference sequence (Table S2) for ease of community comparison [17] https://www.ncbi.nlm.nih.gov/nuccore/M33388?report=GenBank, https://www.ncbi.nlm.nih.gov/nucleotide/45024927 [18]. *CYP2D6* full-gene haplotypes were named relative to the reverse DNA strand of three reference sequences with the following nomenclature format: reference sequence (genome name or GenBank Accession Number)-HCYPAND *CYP2D6* * allele designation-polymorphism rs number, if known, followed by the base at that position. Note that if an rs number is not provided for a specific location, the nucleotide position and base change, relative to the indicated reference, are provided. For example, haplotype 1 is named M33388-CYP2D6: 5157insCCCACCCCTT, hg19-CYP2D6: rs28439297A; rs28680494G; rs1080983G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371699G; rs28371701C; rs28371702T; rs1058164G; rs16947C; rs28371730G; rs1135840G; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T, and hg38-CYP2D6: none, relative to the M33388, hg19, and hg38 reference genomes, respectively.

Full-gene *CYP2D6* haplotypes and diplotypes for 2504 individuals in five super- and 26 sub-populations were created using excel-based workbooks. Private mutations (those polymorphisms observed once in the 1000 Genomes dataset), except those considered clinically relevant by HCYPAND and those predicted by Wendt et al. 2017 [12] to damage, or most likely damage, the resulting protein, were removed from haplotype formation to simplify downstream analyses. 1000 Genomes Project haplotypes were named according to the same convention described above. Genetic Data Analysis [19] was used to determine haplotype and diplotype frequencies, observed (H_o) and expected (H_e) heterozygosities, pairwise genetic distances, and perform tests for detection of departures from Hardy-Weinberg Equilibrium (HWE). TreeView Version 1.6.6 Build 7601 [20, 21] and RStudio® [22] were used to create phylogenetic trees and multidimensional scaling (MDS) plots, respectively. Network analyses were performed using Population Analysis with Reticulate Trees (PopART) [23]. For comparison to HCYPAND *CYP2D6* * alleles, full-gene haplotypes lacking amino acid changes and damaging intronic sequences are considered derivatives of *CYP2D6**1 and full-gene haplotypes conferring R296C (or C296R for hg19-based haplotypes) and S486T (or T486S for hg19-based haplotypes), but no other amino acid changes or damaging intronic polymorphisms were considered derivatives of *CYP2D6**2.

Activity scores were assigned to each allele (i.e., 0, 0.5, 1, or 2) and individual (i.e., 0, 0.5, 1, 1.5, 2, or 3) in two ways: (1) Full-gene haplotypes were assigned a most similar

HCYPAND-recognized allele which was then assigned an activity score based on Gaedigk et al. [24] without considering the impact of additional polymorphisms not recognized by HCYPAND [6]. (2) Considering all polymorphisms, a best- and worst-case activity score were assigned to each full-gene haplotype. For example, haplotype 33 (M33388-*CYP2D6*: 310 T; 843G; 1067G; 5157insCCCACCCCTT) received an activity score of 1 based on the absence of HCYPAND causal polymorphisms only (resembling a normally active * allele). The presence of non-HCYPAND intronic polymorphism rs78854695, 1067G, damages splicing [12], but its specific impact on CYP2D6 has not been confirmed empirically. So haplotype 33 was assigned best- and worst-case activity scores of 0.5 and 0, respectively.

Results

Histograms of SIFT [25–30], PolyPhen-v2 [25, 31, 32], and PROVEAN [33–35] scores from Wendt et al. [12] for HCYPAND-recognized and non-HCYPAND polymorphisms (Fig. 1) indicate that 29, 20, and 30 polymorphisms, respectively, from the 1000 Genomes Project dataset for *CYP2D6* are predicted to negatively impact protein function. After removal of 138 private mutations, except those considered causal by HCYPAND and those considered damaging or most likely damaging based on Wendt et al. [12], 446 unique full-gene haplotype string sequences were observed (Table S2). Full-gene *CYP2D6* haplotypes 11, 3, and 1 are identical to reference sequences M33388, hg19, and hg38, respectively. A majority of haplotypes were observed once in the global population, so the average global frequency of full-gene *CYP2D6* haplotypes was quite low (0.00224 ± 0.0115 with a range of 0.165 (haplotype 1) to 0.000200 (haplotypes 205 through 446)). Haplotypes 1 through 18, had global frequencies ≥ 1%, with an average frequency of 0.0394 ± 0.0438 (Table 1 and Fig. 2). The average super-population frequencies for haplotypes 1 through 18 were 0.0348 ± 0.0382 in AFR, 0.0508 ± 0.0832 in AMR, 0.0110 ± 0.0169 in EAS, 0.0539 ± 0.0559 in EUR, and 0.0638 ± 0.102 in SAS.

One haplotype observed in this study was consistent with HCYPAND *CYP2D6* * allele reference list (*CYP2D6**1A [6], M33388-*CYP2D6*: None, hg19-*CYP2D6*: rs28439297A; rs28680494G; rs1080983G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs10808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371699G; rs28371701C; rs28371702T; rs1058164G; rs16947C; rs28371730G; rs1135840G; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T; rs536156813delCCCACCCCTT, hg38-*CYP2D6*: rs536156813delCCCACCCCTT). Though not specifically reported in the HCYPAND *CYP2D6* * allele table, 407 (91.3%) haplotypes could be associated with at least one * allele based on the presence of defining amino acid changes and causal polymorphisms; however, 38 (i.e., 8.52%) of them could not be associated. These 38 haplotypes were observed 125/5008 times (2.50%) in the 1000 Genomes Project and contain combinations of functionally relevant polymorphisms [6]. Figure 3 and Figure S1 represent the variant composition of haplotypes 1 through 18 and all 446 haplotypes, respectively. The average number of polymorphisms per haplotype was 15 ± 10, 20 ± 10, and 14 ± 10 as designated by comparison with the M33388, hg19, and hg38 reference sequences, respectively. The majority of each haplotype is functionally irrelevant polymorphic sites; however, 326, 328, and 326 haplotypes harbor a damaging, or most likely damaging, variant relative to the M33388, hg19, and hg38 reference genomes, respectively.

Network analysis was performed to determine the relatedness of two sets of haplotypes: (1) haplotypes having ≥ 1% global haplotype frequency (haplotypes 1–18; Fig. 4), and (2) haplotypes observed more than once in the entire 1000 Genomes Project dataset (haplotypes 1–204; Figure S2). In Fig. 4 and Figure S2, the haplotypes with relatively low global frequencies appear to be derived from the major haplotypes (1, 2, and 3). All major haplotypes except 15 and 17 were observed in the AFR super-population, though the frequency is relatively low due to the AFR population having the widest haplotype spread (Figure S2). Haplotype 9 is exclusive, and haplotypes 5, 6, 9, 13, and 18 are almost exclusive, to the AFR super-population. The clustering of rare haplotypes suggests that these may be specific to one super-population, such as AFR (minor haplotypes stemming from haplotypes 5, 6, 8, and 9) or EAS (minor haplotypes stemming from haplotypes 2, 15, and 17).

There were 961 unique diplotypes observed across 2504 individuals. The average global diplotype frequency was 0.00104 ± 0.00336. Ten diplotypes had global frequencies ≥ 1%, with an average global frequency of 0.0275 ± 0.0177 (Table 1). The average diplotype frequency in the AFR, AMR, EAS, EUR, and SAS super-populations was 0.00253 ± 0.00277, 0.00587 ± 0.0133, 0.00658 ± 0.0227, 0.00466 ± 0.00688, and 0.00568 ± 0.0142, respectively (Fig. 5). The average global H_o and H_e of *CYP2D6* were 0.820 ± 0.0898 and 0.861 ± 0.0923, respectively. Before Bonferroni correction ($p < 0.05$), *CYP2D6* deviated from HWE expectations in three super-populations (AFR, AMR, and EUR) and nine sub-populations (ASW, GWD, LWK, MSL, YRI, CLM, CDX, CHS, and GBR) (Fig. 6). After Bonferroni correction ($p < 0.00161$), *CYP2D6* deviated from HWE expectations in the AFR super-population and the ASW and LWK sub-populations.

Activity scores were assigned to each individual using the functional consequence of the most closely related HCYPAND-recognized * allele to each full-gene haplotype. Two hundred eighty-two individuals harbor full-gene

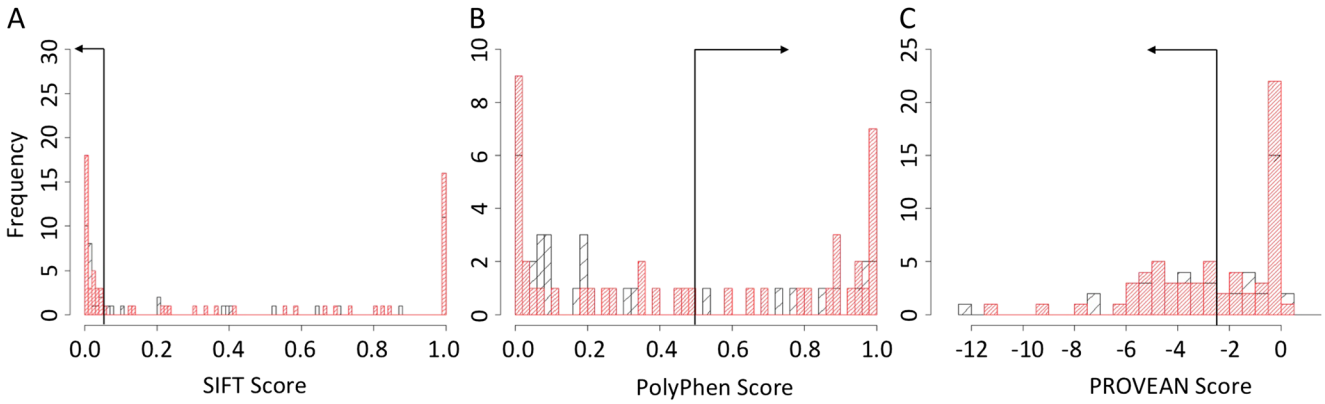


Fig. 1 Distribution of variant effect prediction for Clinical Pharmacogenetics Implementation Consortium polymorphisms and those not used for CYP2D6 nomenclature. Distribution of Sort Intolerant From Tolerant (SIFT) (**a**) [25–30], Polymorphism Phenotyping v2 (PolyPhen-v2) (**b**) [25, 31, 32], and Protein Variant Effect Analyzer (PROVEAN) (**c**) [33–35] scores from Wendt et al. [12] for The Human CYP Allele Nomenclature Database recognized polymorphisms (black bars) and polymorphisms not considered by the database (red bars). Arrows indicate the threshold applied to scores from each algorithm to predict the damaging or benign effect a polymorphism has on protein function; scores < 0.05, > 0.5, < − 2.5 for SIFT, PolyPhen-v2, and PROVEAN, respectively, may impact protein function

haplotypes that, by HCYPAND nomenclature, should confer normal activity enzymes (activity score = 1), but full-gene data reveal additional causal polymorphisms that were predicted to damage enzyme function (Fig. 7). Twenty-two of these individuals (21 AFR and 1 AMR) have diplotypes where two full-gene haplotypes produce conflicting activity scores relative to the HCYPAND approach. The average individual activity scores were 1.40 ± 0.644, 1.36 ± 0.638, and 1.30 ± 0.675 based on HCYPAND guidelines, full-gene best- (Student's t test, $p < 0.05$ relative to the HCYPAND approach), and full-gene worst-case (Student's t test, $p < 0.001$ relative to the HCYPAND approach) haplotype activity score designation. Haplotype 8 (M33388-CYP2D6: -2060G; -2053T; -1594C; -1418A; -1408A; -1235G; -740T; 176A; 310T; 744delC; 746G; 843G; 1659A; 1661C; 2123T; 2215G; 2292A; 2850T; 3183A; 4180C; 4719G; 4864C; 5157insCCCACCCCTT) is responsible for the majority of conflicting MP predictions (111 AFR and 2 AMR individuals). By HCYPAND nomenclature, this haplotype would be assigned an activity score of 0.5 based on similarity to *CYP2D6**29 [40]; however, full-gene haplotype data reveal the presence of non-HCYPAND INDEL rs267608275delC (744delC). Due to lack of empirical observation of the influence rs267608275delC has on CYP2D6 function, worst- and best-case activity scores of 0 and 0.5, respectively, were assigned for full-gene haplotypes, indicating that rs267608275delC may or may not further decrease CYP2D6 function. This significant decrease in average activity score suggests that by current *CYP2D6* * allele and activity score designation, ~ 11% of the individuals in this dataset may be wrongly considered gNM-F, gNM-S, or gIM when they likely belong in the gNM-S, gIM, and gPM categories, respectively.

MDS plots and unrooted phylogenetic trees (Fig. 8) of super-populations show distant separation of the AFR and EAS populations while EUR, AMR, and SAS cluster relatively close together. Considering sub-populations, those belonging to the AFR and EAS super-populations cluster together according to super-population affiliation. Sub-populations belonging to the AMR, EUR, and SAS super-populations have considerable overlap with one another.

Discussion

Full-gene *CYP2D6* haplotypes may be able to refine MP predictions, ultimately identifying more gIM and gPM individuals than the HCYPAND haplotype nomenclature. The analyses presented here may be limited by the relative low depth of sequence coverage per sample within the 1000 Genomes Project database, small sample size for each sub population, and the precision of variant effect prediction algorithms. These factors impact detection of rare variants that may contribute to CYP2D6 function. In fact, the splice defects defining three * alleles (*CYP2D6*4*, *CYP2D6*11*, and *CYP2D6*41*) were not identified by the software analyses performed in Wendt et al., and the 843G SNP was incorrectly identified as damaging [6], emphasizing the importance of using variant effect predictors with caution and employing multiple prediction algorithms to identify single nucleotide variants of interest [41]. Additional rare variants will continue to be discovered for PM, IM, and/or UM individuals and/or isolated populations such as Finns or Ashkenazi Jews. While potentially contributing significantly to phenotype, variant effect prediction is not possible; however, the analysis herein may allow for some prediction of phenotype using a number of previously unexploited polymorphisms [42–45]. Additional research is needed (e.g., functional enzyme studies, targeted mutagenesis, and/or quantitative trait analyses) to empirically characterize the phenotype generated from rare polymorphisms, combinations of rare/deleterious polymorphisms in the same haplotype,

Table 1 Full-gene *CYP2D6* haplotypes with ≥ 1% global frequency. Eighteen cytochrome p450 family 2, subfamily D, polypeptide 6 (*CYP2D6*) full-gene haplotypes with global allele frequencies ≥ 1% in the 1000 Genomes Project dataset. Italicized nucleotide positions and rs numbers indicate polymorphisms that change the amino acid sequence relative to the indicated reference genome; underlined nucleotide positions and rs numbers alter CYP2D6 function [6, 24, 36–38]

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
1	1	Normal	1	1	1	M33388-*CYP2D6*: 5157insCCCACCCCTT	–	hg19-*CYP2D6*: rs28439297A; rs28680494G; rs1080983G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371699G; rs28371701C; rs28371702T; rs1058164G; *rs16947C*; rs28371730G; *rs1135840G*; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T	Splice defect; C296R; T486S	hg38-*CYP2D6*: none	–
2	4	None	0	0	0	M33388-*CYP2D6*: -2060G; -2053T; -1426T; -1235G; -1000A; *100T*; 310T; 843G; 1039T; 1661C; 2097G; 3582G; *4180C*; 4401T; 4719G; 4808G; 4864C; 5157insCCCACCCCTT	P34S; splice defect; S486T	hg19-*CYP2D6*: rs1080983G; rs1080985C; rs28588594T; rs1080989A; rs28624811C; rs28633410G; *rs1065852T*; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371701C; rs1081003T; rs58440431G; *rs16947C*; rs2004511G; rs28371730G; rs28371738T; rs116390392G; rs71184866insACA; rs4078247G	P34S; C296R	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs28588594T; rs28735595G; rs1080989A; *rs1065852T*; rs28371699T; rs28371702G; rs1081003T; rs1058164C; rs58440431G; rs2004511G; *rs1135840C*; rs28371738T; rs35028622G; rs4078247G; rs34386013C	P34S; splice defect; S486T
3	2	Normal	1	1	1	M33388-*CYP2D6*: -2060G; -2053T; -1770A; -1584G; -1235G; -740T; -678A; 214C; 221A; 223G; 227C; 232C; 233C; 245G; 310T; 746G; 843G; 1661C; *2850T*; 3584A; *4180C*; 4481A; 4655delACA;	Splice defect; R296C; S486T	hg19-*CYP2D6*: None	–	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs1080983A; rs1080985G; rs28735595G; rs28624811T; rs28633410A; rs1808995C; rs1080996A; rs74644586G;	Splice defect; R296C; S486T

Table 1 (continued)

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
						4719G; 4864C; 5157insCCCACCCCTT				rs76312385C; rs75276289C; rs28695233C; rs1081000G; rs28371699T; rs28371701G; rs28371702G; rs1058164C; *rs16947T*; rs28371730A; *rs1135840C*; rs116390392A; rs71184866delACA; rs35028622G; rs34386013C	
4	2	Normal	1	1	1	M33388-*CYP2D6*: -2060G; -2053T; -1770A; -1235G; -740T; -678A; 214C; 221A; 223G; 227C; 232C; 233C; 245G; 310T; 746G; 843G; 1661C; *2850T*; 2988A; 3584A; *4180C*; 4481A; 4655delACA; 4719G; 4864C; 5157insCCCACCCCTT	Splice defect; R296C; Splice defect; S486T	hg19-*CYP2D6*: rs1080985C; rs28371725A	Splice defect	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs1080983A; rs28735595G; rs28624811T; rs28633410TA; rs1808995C; rs1080996A; rs74644586G; rs76312385C; rs75276289C; rs28695233C; rs1081000G; rs28371699T; rs28371701G; rs28371702G; rs1058164C; *rs16947T*; rs28371725A; rs28371730A; *rs1135840C*; rs116390392A; rs71184866delACA; rs35028622G; rs34386013C	Splice defect; R296C; splice defect; S486T
5	17	Decreased	0.5	0.5	0.5	M33388-*CYP2D6*: -2060G; -2053T; -1235G; -740T; 214C; 221A; 223G; 227C; 232C; 233C;	Splice Defect; T107I; R296C; S486T	hg19-*CYP2D6*: rs1080983G; rs1080985C; rs28633410G; rs28371701C; *rs28371706T*; rs116390392G;	T107I	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs28735595G; rs28624811T;	Splice defect; T107I; R296C; S486T

Table 1 (continued)

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
						245G; 310T; 843G; *1023A*; 1661C; *2850T*; 3584A; *4180C*; 4709A; 4719G; 4766A; 4864C; 5157insCCCACCCCTT		rs71184866insACA; rs4078249A; rs4078248A		rs1808995C; rs1080996A; rs74644586G; rs76312385C; rs75276289C; rs28695233C; rs1081000G; rs28371699T; rs28371702G; *rs28371706T*; rs1058164C; *rs16947T*; rs28371730A; *rs1135840C*; rs4078249A; rs35028622G; rs4078248A; rs34386013C	
6	1	Normal	1	1	1	M33388-*CYP2D6*: -43insG; 4928A; 5157insCCCACCCCTT	–	hg19-*CYP2D6*: rs28439297A; rs28680494G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs75085559insG; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371699G; rs28371701C; rs28371702T; rs1058164G; *rs16947C*; rs28371730G; *rs1135840G*; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T; rs112866416A	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs75085559insG; rs112866416A	–
7	4	None	0	0	0	M33388-*CYP2D6*: -2060G; -2053T; -1426T; -1235G; -1000A; *100T*; 310T; 746G; 843G; *974A*; *984G*; 997G; 1661C; 1846A; 2097G; 3582G; *4180C*; 4401T; 4719G; 4808G; 4864C;	P34S; L91 M; H94R; Splice Defect; Splice Defect; S486T	hg19-*CYP2D6*: rs1080983G; rs1080985C; rs28588594T; rs1080989A; rs28624811C; rs28633410G; *rs1065852T*; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; *rs28371703A*; *rs28371704G*; rs28371705G; rs3892097A; rs58440431G; *rs16947C*; rs2004511G;	P34S; L91M; H94R; splice defect; C296R	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs28588594T; rs28735595G; rs1080989A; *rs1065852T*; rs28371699T; rs28371701G; rs28371702G; *rs28371703A*; *rs28371704G*;	P34S; L91M; H94R; splice defect; splice defect; S486T

Table 1 (continued)

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
						5157insCCCACCCCTT		rs28371730G; rs28371738T; rs116390392G; rs71184866insACA; rs4078247G		rs28371705G; rs1058164C; rs3892097A; rs58440431G; rs2004511G; *rs1135840C*; rs28371738T; rs35028622G; rs4078247G; rs34386013C	
8	29	Decreased	0.5	0.5	0	M33388-*CYP2D6*: -2060G; -2053T; -1594C; -1418A; -1408A; -1235G; -740T; -176A; 310T; 744delC; 746G; 843G; *1659A*; 1661C; 2123T; 2215G; 2292A; *2850T*; *3183A*; *4180C*; 4719G; 4864C; 5157insCCCACCCCTT	Splice defect; V136I; R296C; V338M; S486T	hg19-*CYP2D6*: rs1080983G; rs1080984C; rs1080985C; rs1080986A; rs1080987A; rs28633410G; rs1080993A; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs267608275delC; *rs61736512A*; rs76327133T; rs80262685G; rs75203276A; *rs59421388A*; rs28371730G; rs116390392G; rs71184866insACA	Splice defect; V136I; V338M	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs1080984C; rs1080986A; rs1080987A; rs28735595G; rs28624811T; rs1080993A; rs28371699T; rs267608275delC; rs28371701G; rs28371702G; *rs61736512A*; rs1058164C; rs76327133T; rs80262685G; rs75203276A; *rs16947T*; *rs59421388A*; *rs1135840C*; rs35028622G; rs34386013C	Splice defect; V136I; R296C; V338M; S486T
9	17	Decreased	0.5	0.5	0.5	M33388-*CYP2D6*: -2060G; -2053T; -1235G; -740T; 214C; 221A; 223G; 227C; 232C; 233C; 245G; 310T; 654T; 843G; *1023A*; 1661C; *2850T*; 3584A; *4180C*; 4709A; 4719G; 4766A; 4864C;	Splice defect; T107I; R296C; S486T	hg19-*CYP2D6*: rs1080983G; rs1080985C; rs28633410G; rs376217512T; rs28371701C; *rs28371706T*; rs116390392G; rs71184866insACA; rs4078249A; rs4078248A	T107I	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs28735595G; rs28624811T; rs1808995C; rs1080996A; rs74644586G; rs76312385C; rs75276289C; rs28695233C; rs1081000G;	Splice defect; T107I; R296C; S486T

Table 1 (continued)

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
						5157insCCCACCCCTT				rs28371699T; rs376217512T; rs28371702G; *rs283717067T*; rs1058164C; *rs16947T*; rs28371730A; *rs1135840C*; rs4078249A; rs35028622G; rs4078248A; rs34386013C	
10	4	None	0	0	0	M33388-*CYP2D6*: -2060G; -2053T; -1426T; -1235G; -1000A; *100T*; 310T; 746G; 843G; *974A*; *984G*; 997G; 1661C; 1846A; 2097G; 3582G; *4180C*; 4401T; 4719G; 4808G; 4864C	P34S; L91M; H94R; splice defect; splice defect; S486T	hg19-*CYP2D6*: rs1080983G; rs1080985C; rs28588594T; rs1080989A; rs28624811C; rs28633410G; *rs1065852T*; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; *rs28371703A*; *rs28371704G*; rs28371705G; rs3892097A; rs58440431G; *rs16947C*; rs2004511G; rs28371730G; rs28371738T; rs116390392G; rs71184866insACA; rs4078247G; rs536156813delCCCACCCCTT	P34S; L91M; H94R; splice defect; C296R	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs28588594T; rs28735595G; rs1080989A; *rs1065852T*; rs28371699T; rs28371701G; rs28371702G; *rs28371703A*; *rs28371704G*; rs28371705G; rs1058164C; rs3892097A; rs58440431G; rs2004511G; *rs1135840C*; rs28371738T; rs35028622G; rs4078247G; rs34386013C; rs536156813delCCCACCCTT	P34S; splice defect; dplice defect; L91 M; H94R; splice defect; S486T
11	1	Normal	1	1	1	M33388-*CYP2D6*: none	–	hg19-*CYP2D6*: rs28439297A; rs28680494G; rs1080983G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371699G; rs28371701C; rs28371702T; rs1058164G; *rs16947C*;	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs536156813delCCCACCCTT	–

Table 1 (continued)

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
								rs28371730G; *rs1135840G*; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T; rs536156813delCCCACCCCTT			
12	1	Normal	1	1	1	M33388-*CYP2D6*: 1170A; 5157insCCCACCCCTT	–	hg19-*CYP2D6*: rs28439297A; rs28680494G; rs1080983G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371699G; rs28371701C; rs28371702T; rs1081004A; rs1058164G; *rs16947C*; rs28371730G; *rs1135840G*; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs1081004A	–
13	1	Normal	1	1	1	M33388-*CYP2D6*: 310T; 843G; 1067G	Splice defect; splice defect	hg19-*CYP2D6*: rs28439297A; rs28680494G; rs1080983G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371701C; rs78854695G; rs1058164G; *rs16947C*; rs28371730G; *rs1135840G*; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T; rs536156813delCCCACCCCTT	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs28371699T; rs28371702G; rs78854695G; rs536156813delCCCAC-CCCTT	Splice defect; splice defect
14	35	Normal	1	1	1	M33388-*CYP2D6*: -2060G; -2053T; -1770A; -1584G; -1235G; -740T; -678A; *31A*; 214C; 221A; 223G; 227C; 232C; 233C; 245G; 310T; 746G; 843G; 1661C; *2850T*;	V11M; splice defect; S486T	hg19-*CYP2D6*: *rs769258A*	V11M	hg38-*CYP2D6*: rs28439297G; rs28680494T; rs1080983A; rs1080985G; rs28735595G; rs28624811T; rs28633410A; *rs769258A*; rs1808995C;	V11M; splice sefect; R296C; S486T

Table 1 (continued)

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
						3584A; *4180C*; 4481A; 4655delACA; 4719G; 4864C; 5157insCCCACCCCTT				rs1080996A; rs74644586G; rs76312385C; rs75276289C; rs28695233C; rs1081000G; rs28371699T; rs28371701G; rs28371702G; rs1058164C; *rs16947T*; rs28371730A; *rs1135840C*; rs116390392A; rs71184866delACA; rs35028622G; rs34386013C	
15	1	Normal	1	1	1	M33388-*CYP2D6*: -498A; 5154T; 5157insCCCACCCCTT	–	hg19-*CYP2D6*: rs28439297A; rs28680494G; rs1080983G; rs1080985C; rs28735595A; rs28624811C; rs28633410G; rs74966855A; rs1808995G; rs1080996C; rs74644586C; rs76312385T; rs75276289G; rs28695233A; rs1081000A; rs28371699G; rs28371701C; rs28371702T; rs1058164G; *rs16947C*; rs28371730G; *rs1135840G*; rs116390392G; rs71184866insACA; rs35028622T; rs34386013T; rs374153932T	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs74966855A; rs374153932T	–
16	1	Normal	1	1	1	M33388-*CYP2D6*: 270A; 5157insAAGGGGT-GGG	–	hg19-*CYP2D6*: rs28439297T; rs28680494C; rs1080983C; rs1080985G; rs28735595T; rs28624811G; rs28633410C; rs1808995C; rs1080996G; rs74644586G; rs76312385A; rs75276289C; rs28695233T; rs1081000T; rs29001678A; rs28371699C; rs28371701G; rs28371702A; rs1058164C; *rs16947G*; rs28371730C; *rs1135840C*; rs116390392C;	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs29001678A	–

Table 1 (continued)

Haplotype number	Most similar CYP2D6 * allele	Reported activity	Gaedigk et al. activity score	Full-gene best-case activity score	Full-gene worst-case activity score	M33388 nomenclature	M33388 amino acid changes	Hg19 nomenclature	Hg19 amino acid changes	Hg38 nomenclature	Hg38 amino acid changes
								rs71184866insTGT; rs35028622A; rs34386013A			
17	1	Normal	1	1	1	M33388-*CYP2D6*: 270A; 5157insAAGGGGT-GGG	–	hg19-*CYP2D6*: rs28439297T; rs28680494C; rs1080983C; rs1080985G; rs28735595T; rs28624811G; rs28633410C; rs1808995C; rs1080996G; rs74644586G; rs76312385A; rs75276289C; rs28695233T; rs1081000T; rs29001678A; rs28371699C; rs28371701G; rs28371702A; rs1058164C; *rs16947G*; rs28371730C; *rs1135840C*; rs116390392C; rs71184866insTGT; rs35028622A; rs34386013A	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs29001678A	–
18	1	Normal	1	1	1	M33388-*CYP2D6*: 5154A; 5157insAAGGGGT-GGG	–	hg19-*CYP2D6*: rs28439297T; rs28680494C; rs1080983C; rs1080985G; rs28735595T; rs28624811G; rs28633410C; rs1808995C; rs1080996G; rs74644586G; rs76312385A; rs75276289C; rs28695233T; rs1081000T; rs28371699C; rs28371701G; rs28371702A; rs1058164C; *rs16947G*; rs28371730C; *rs1135840C*; rs116390392C; rs71184866insTGT; rs35028622A; rs34386013A; rs374153932A	Splice defect; C296R; T486S	hg38-*CYP2D6*: rs374153932A	–

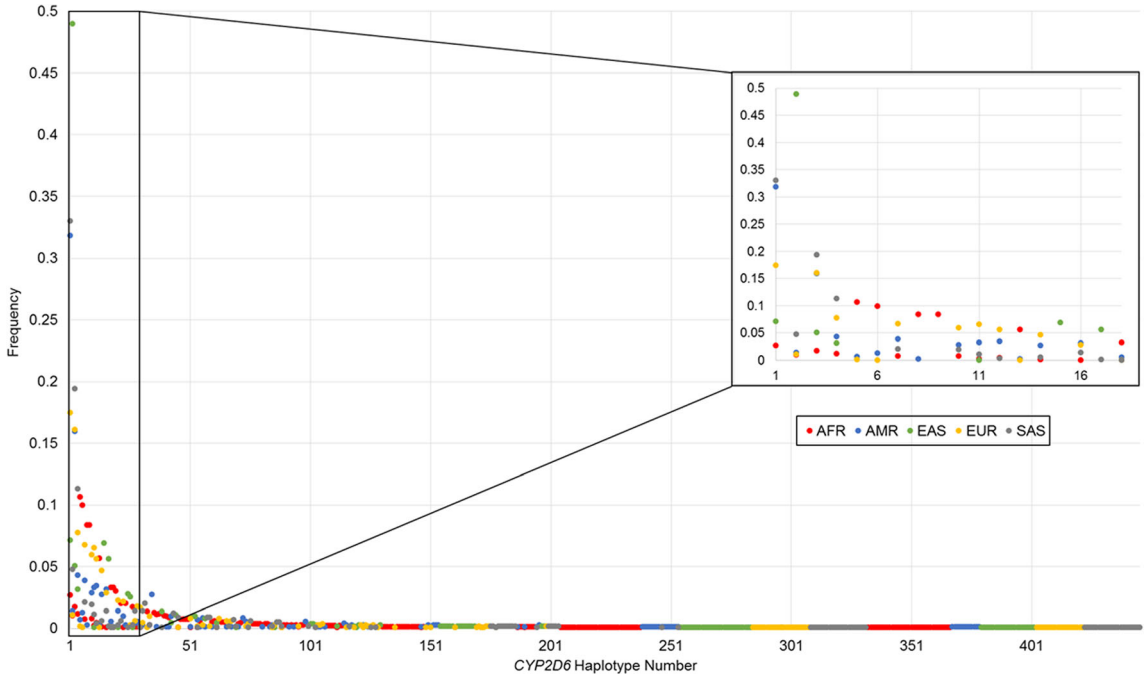


Fig. 2 *CYP2D6* haplotype frequencies. Observed frequencies of 446 full-gene *CYP2D6* haplotypes in five major super-populations (African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS)). The inset graph shows the frequency of those haplotypes with greater than 30 observations globally

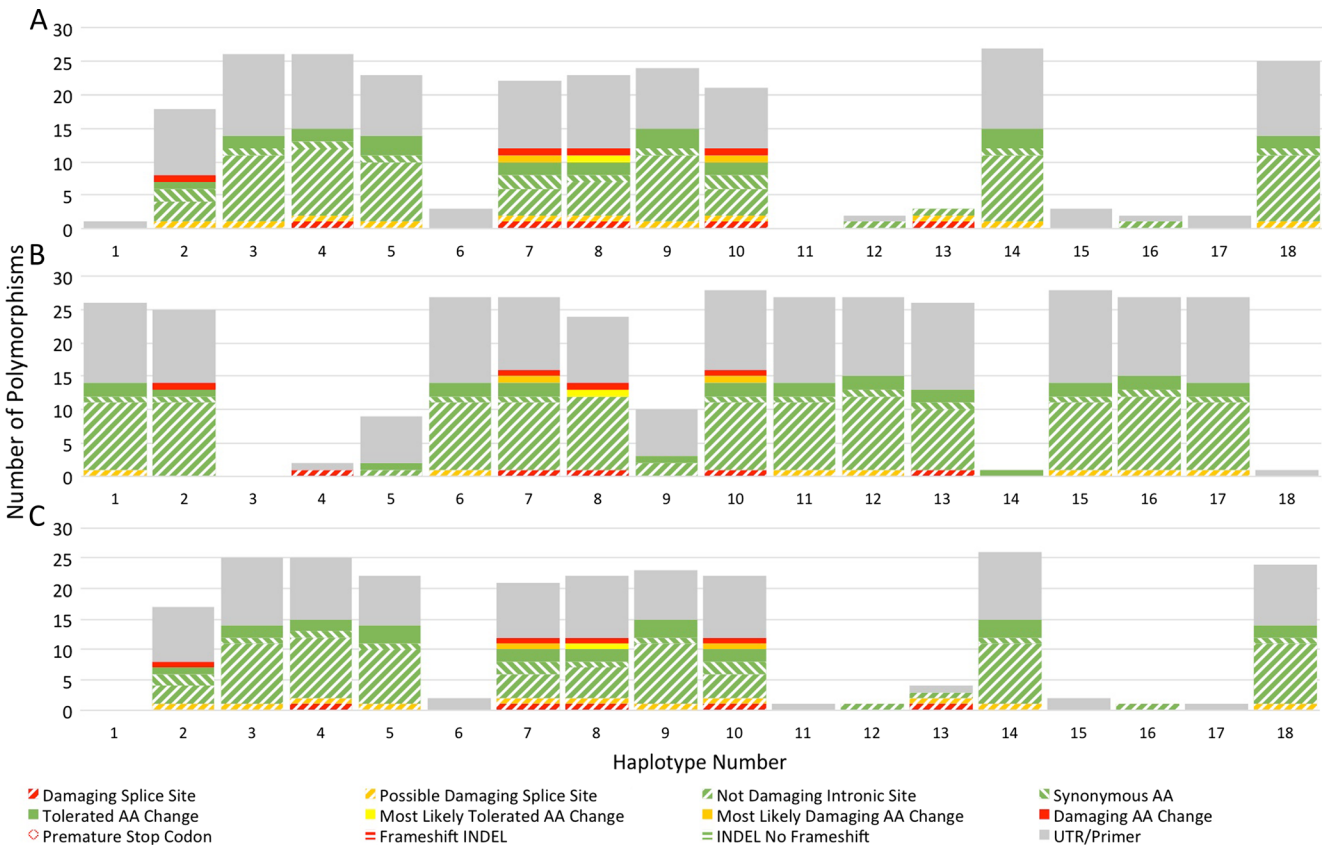


Fig. 3 *CYP2D6* haplotype composition**.** Haplotype composition of 18 full-gene *CYP2D6* star alleles, with global frequencies ≥ 1%, aligned to GenBank accession M33388 (**a**), hg19 (**b**), and hg38 (**c**). Variant effect predictions performed by Wendt et al. [13] using Sort Intolerant From Tolerant (SIFT) [25–30], Polymorphism Phenotyping v2 (PolyPhen-v2) [25, 31, 32], Protein Variant Effect Analyzer (PROVEAN) [33–35], and Human Splicing Finder [39]

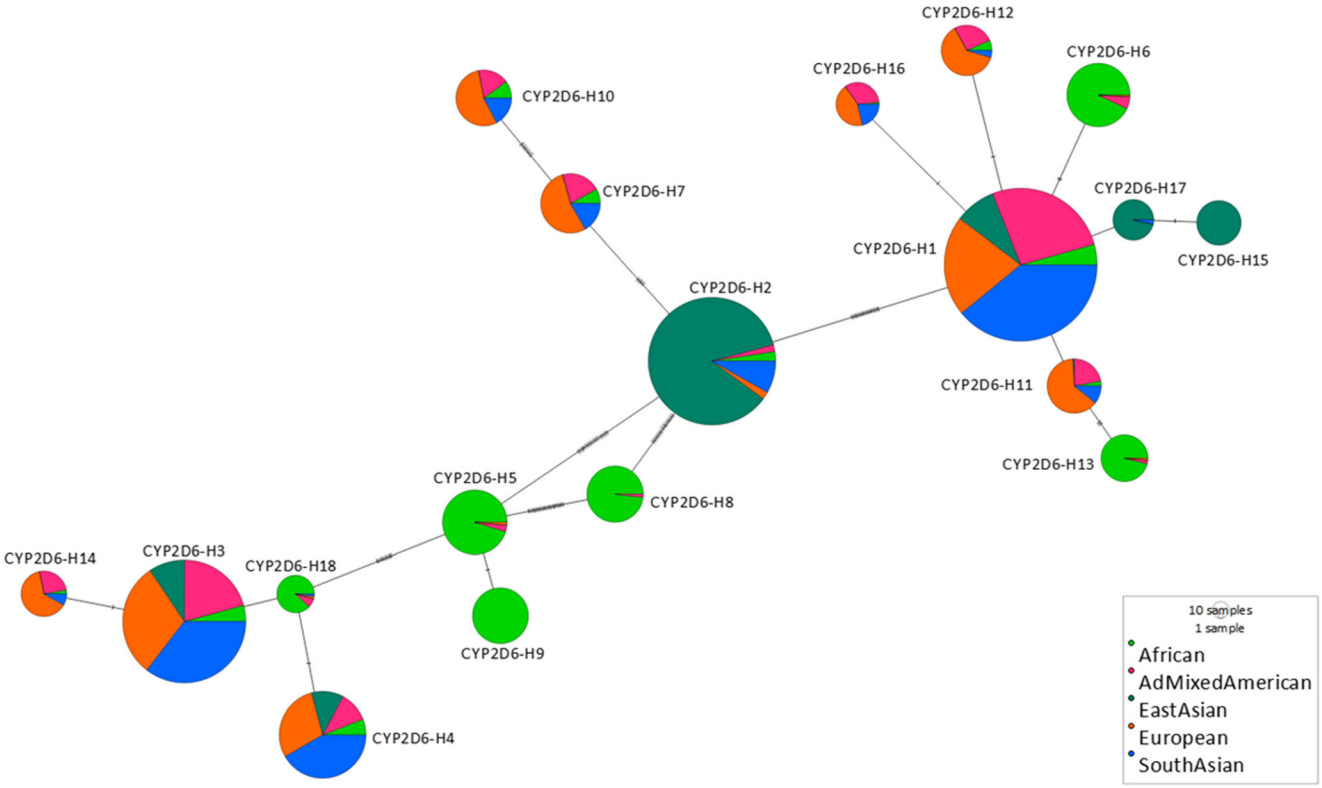


Fig. 4 Network analysis of *CYP2D6* haplotypes. Network analysis of *CYP2D6* full-gene haplotypes 1 through 18. The size of each circle is proportional to the global frequency of each haplotype, segments within each circle are proportional to the super-population haplotype frequency, and lines connecting circles are dashed with the number of mutations separating two haplotypes

combinations of deleterious polymorphisms in different genes that may influence one another, and their distributions in under-represented populations.

Copy number variation (CNV) of some *CYP2D6* * alleles and *CYP2D7* pseudogene conversion do occur in some individuals, namely UMs, and may influence the HWE and LD results [46]. It is likely that some 1000 Genomes Project individuals from the AFR super-population carry CNVs based on deviations from HWE expectations [46], but the project does not explore CNV in detail due to limitations of short-read sequencing [47, 48]. The data presented herein have been analyzed as though only two copies of *CYP2D6* are present in each individual, so unless an individual contains the SNP rs1135823G>T (1617G>T), the UM phenotype was not identified. A number of unique haplotypes have been identified that may be true haplotype observations but may also be attributed to duplication of two common haplotypes and/or *CYP2D7* pseudogene conversions [49, 50]. This phenomenon is particularly true for African populations which exhibit relatively frequent gene duplications (up to 30%) [51, 52]. The development of continuous read single-molecule DNA sequencing strategies, such as nanopore technology [53], may help reduce ambiguity in sequencing regions with a high degree of structural variation.

Unique haplotypes have been identified from the full-gene region of *CYP2D6*, including introns, exons, 5′ and 3′ UTRs, and the promoter region. While comprehensively assessing the gene itself, there are a number of distant regulatory elements that may impact enzyme function. Wang et al. [54] identified long-range haplotypes that include polymorphisms within enhancer regions that may refine these haplotype definitions. However, inclusion of such regulatory elements in *CYP2D6*-based MP predictions may need to be explored further due to potential regulation of other enzymes, potentially confounding MP prediction. Additionally, private mutations not predicted to damage the resulting protein have been removed (note that those considered for HCYPAND-recognized * alleles have been included) from this analysis; however, if incorporated, may produce finer granularity of haplotype definition but likely would not alter activity score unless empirically shown to alter enzyme function.

Full-gene *CYP2D6* data have provided additional resolution to the MP compared to predictions used to date, possibly resolving some medico-legal autopsy negative cases. Although empirical data are required to confirm their enzyme activity, approximately 11% of the healthy individuals in this study may be wrongly identified as NMs according to traditional *CYP2D6* genotyping and activity score predictions of MP. Clinically, these individuals likely would be classified as IM or PM and be treated accordingly. Enhanced predictive capabilities of MP may be made with comprehensive *CYP2D6* diplotype information and/

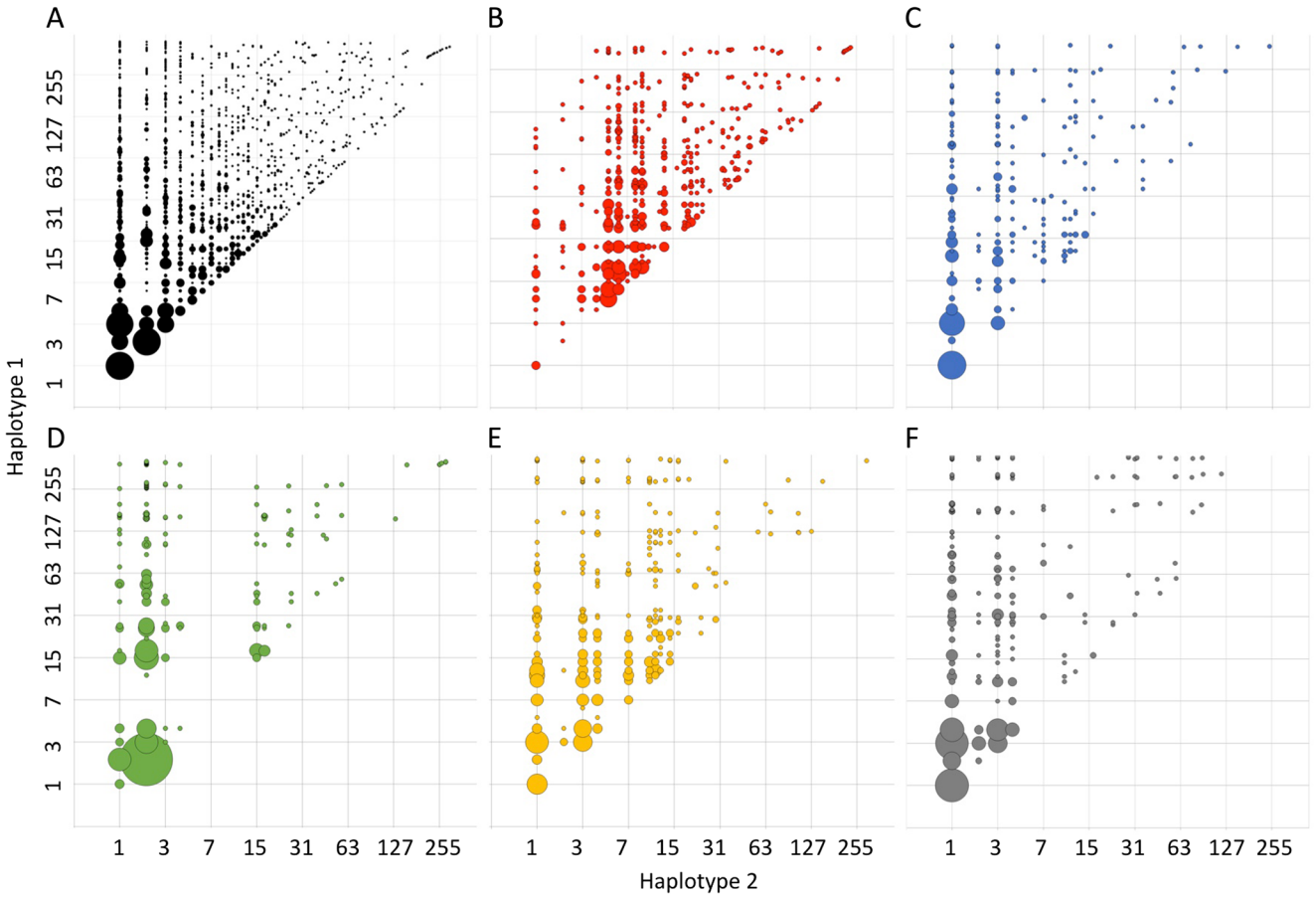


Fig. 5 CYP2D6 diplotype frequencies. Relative *CYP2D6* diplotype frequencies in the global population (**a**), African (**b**), Admixed American (**c**), East Asian (**d**), European (**e**), and South Asian (**f**) super-populations. The *x*- and *y*-axes indicate the first and second haplotype, respectively, of an individual diplotype; the size of each solid circle is proportional to the frequency of that diplotype

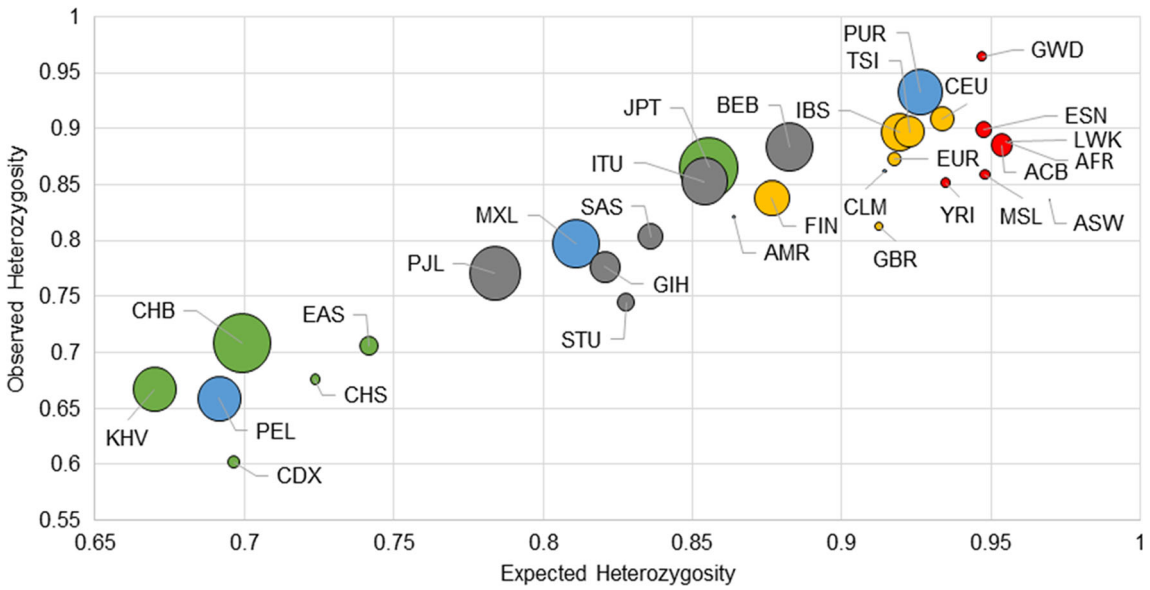


Fig. 6 Heterozygosity and Hardy-Weinberg Equilibrium summary. Observed and expected heterozygosity of *CYP2D6* in five major super- and 26 sub-populations. Super-populations and their associated sub-populations are color-coded (red for African (AFR), blue for Admixed American (AMR), green for East Asian (EAS), yellow for European (EUR), and gray for South Asian (SAS)); the size of each data point corresponds to the Hardy-Weinberg Equilibrium (HWE) *p* value for the locus within that population. The AFR super-population and the Luhya in Webuye, Kenya (LWK), and American of African Ancestry in Southwest United States (ASW) sub-populations deviated significantly from HWE expectations after Bonferroni correction ($p < 0.00161$)

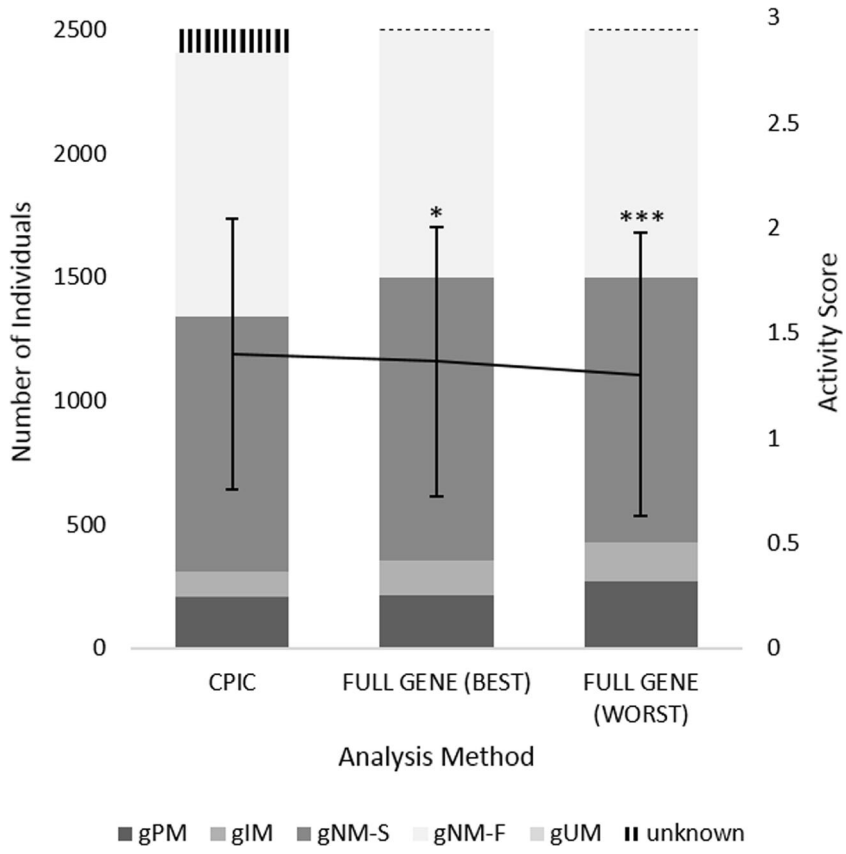
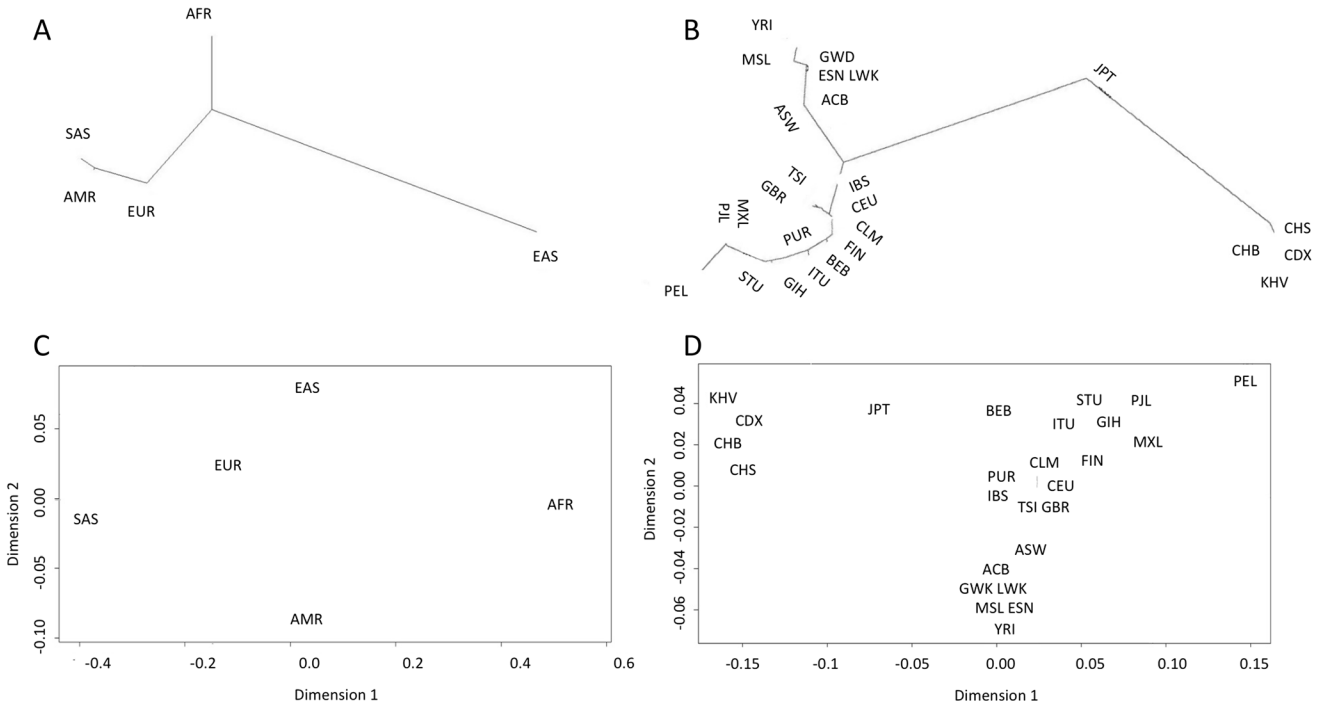


Fig. 7 Activity score predictions using Clinical Pharmacogenetics Implementation Consortium star allele nomenclature and full-gene *CYP2D6* haplotypes. Genotype-inferred (g) CYP2D6 MP (gPM, poor; gIM, intermediate; gNM-S, normal/extensive-slow; gNM-F, normal/extensive-fast; gUM, ultrarapid) distribution, and average activity scores using The Human CYP Allele Nomenclature Database (HCYPAND) [6] recognized *CYP2D6* * alleles and full-gene haplotypes for 2504 1000 Genomes Project self-reported healthy individuals. Asterisks indicate significant difference in average activity score (Student's *t* test; $*p < 0.05$; $***p < 0.001$) compared with HCYPAND nomenclature. vertical black bars indicate the average activity score ± one standard deviation. Individuals with combinations of causal HCYPAND polymorphisms were not assigned to a MP using HCYPAND nomenclature and were assigned to the "unknown" MP category



AFR: African; AMR: Ad Mixed American; EAS: East Asian; EUR: European; SAS: South Asian; ACB: African Caribbean in Barbados; ASW: American of African Ancestry in Southwest USA; BEB: Bengali from Bangladesh; CDX: Chinese Dai in Xishuangbanna, China; CEU: Utah Residence with Northern and Western Ancestry; CHB: Han Chinese in Beijing; CHS: Southern Han Chinese; CLM: Colombians from Medellin, Colombia; ESN: Esan in Nigeria; FIN: Finnish in Finland; GBR: British in England and Scotland; GIH: Gujarati Indian from Houston, Texas; GWD: Gambian in Western Divisions in Gambia; IBS: Iberian Population in Spain; ITU: Indian Telugu from the United Kingdom; JPT: Japanese in Tokyo, Japan; KHV: Kinh in Ho Chi Minh City, Vietnam; LWK: Luhya in Webuye, Kenya; MSL: Mende in Sierra Leone; MXL: Mexican Ancestry from Los Angeles, USA; PEL: Peruvians from Lima, Peru; PJL: Punjabi from Lahore, Pakistan; PUR: Puerto Ricans from Puerto Rico; STU: Sri Lankan Tamil from the United Kingdom; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigeria

Fig. 8 Population comparisons using full-gene *CYP2D6* haplotypes. Neighbor-joining trees (**a** and **b**) and multidimensional scaling plots (**c** and **d**) for five super- (**a** and **c**) and 26 sub-populations (**b** and **d**) in the 1000 Genomes Project using pairwise genetic distances based on full-gene *CYP2D6* haplotype assignment

or incorporation of a longer read sequencing platform into the *CYP2D6* interrogation workflow.

The case described by Koren et al. [4] of an application of *CYP2D6* genotyping to assist in a medico-legal investigation is a classic example of ultrarapid metabolism of codeine to morphine. In such cases, the medicaments were delivered as a pro-drug (inactive) that must be metabolized in order to deliver the intended effect (e.g., codeine or tamoxifen) [1, 2]. Given the prevalence of UM in various populations and the current use of these drugs, it is anticipated that similar cases will occur and molecular autopsies may shed light on the cause and/or manner of death. At the opposite end of the metabolizer phenotype spectrum are PMs and IMs. If considering codeine to morphine metabolism, *CYP2D6* genotyping post-mortem may not be informative. However, many antidepressants (e.g., nortriptyline) are active upon administration and depend on CYP2D6 to deactivate the drug. PMs and IMs, based on their *CYP2D6* variants, may experience adverse reactions to antidepressants which include, but are not limited to, delayed propagation of myocardium depolarization leading to cardiac arrhythmia, myocardial infarction, and death [10]. In the context of the data presented in this study, ~ 11% of individuals may be incorrectly classified as NMs or IMs as their full-gene haplotype data indicate one category lower (i.e., NM by targeted approach is an IM by full-gene approach; IM by targeted approach is a PM by full-gene approach). As such, understanding *CYP2D6* haplotype information of all metabolizer phenotypes can be quite relevant to the medico-legal community.

References

1. Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C (2007) Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacol Ther 116(3):496–526 **Review**
2. Ingelman-Sundberg M (2005) Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics J. 5(1):6–13 **Review**
3. Smith HS (2009) Opioid metabolism. Mayo Clin Proc 84(7):613–624. https://doi.org/10.1016/S0025-6196(11)60750-7 Review
4. Koren G, Cairns J, Chitayat D, Gaedigk A, Leeder SJ (2006) Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother. Lancet 368(9536):704
5. Koch WH (2004) Technology platforms for pharmacogenomic diagnostic assays. Nat Rev Drug Discov 3(9):749–761 **Review**
6. http://www.cypalleles.ki.se/cyp2d6.htm. Accessed 01JUN2016.
7. Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Klein T, Leeder JS (2017) Prediction of CYP2D6 phenotype from genotype across world populations. Genet Med 19(1):69–76. https://doi.org/10.1038/gim.2016.80
8. Koski A, Ojanperä I, Sistonen J, Vuori E, Sajantila AA (2007) Fatal doxepin poisoning associated with a defective CYP2D6 genotype. Am J Forensic Med Pathol 28(3):259–261
9. Koski A, Sistonen J, Ojanperä I, Gergov M, Vuori E, Sajantila A (2006) CYP2D6 and CYP2C19 genotypes and amitriptyline metabolite ratios in a series of medicolegal autopsies. Forensic Sci Int 158(2–3):177–183
10. Kerr GW, McGuffie AC, Wilkie S (2001) Tricyclic antidepressant overdose: a review. Emerg Med J 18(4):236–241 **Review**
11. Karlsson L, Zackrisson AL, Josefsson M, Carlsson B, Green H, Kugelberg FC (2015) Influence of CYP2D6 and CYP2C19 genotypes on venlafaxine metabolic ratios and stereoselective metabolism in forensic autopsy cases. Pharmacogenomics J 15(2):165–171. https://doi.org/10.1038/tpj.2014.50
12. Wendt FR, Pathak G, Sajantila A, Chakraborty R, Budowle B (2017) Global genetic variation of select opiate metabolism genes in self-reported healthy individuals. Pharmacogenomics J. https://doi.org/10.1038/tpj.2017.13
13. http://www.cypalleles.ki.se/criteria.htm. Accessed 14JUN2017.
14. Karolchik D, Hinrichs AS, Kent WJ (2012;Chapter 1:Unit1.4) The UCSC genome browser. Curr Protoc Bioinformatics. https://doi.org/10.1002/0471250953.bi0104s40
15. Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526(7571):68–74. https://doi.org/10.1038/nature15393
16. IGSR and the 1000 Genomes Project frequently asked questions web page. http://www.internationalgenome.org/faq/are-1000-genomes-variant-calls-phased/. Accessed 14JUN2017
17. Zhou SF, Di YM, Chan E, Du YM, Chow VD, Xue CC, Lai X, Wang JC, Li CG, Tian M, Duan W (2008) Clinical pharmacogenetics and potential application in personalized medicine. Curr Drug Metab 9(8):738–784 **Review**
18. Gaedigk A, Bhathena A, Ndjountché L, Pearce RE, Abdel-Rahman SM, Alander SW, Bradford LD, Rogan PK, Leeder JS (2005) Identification and characterization of novel sequence variations in the cytochrome P4502D6 (CYP2D6) gene in African Americans. Pharmacogenomics J 5(3):173–182 Erratum in: Pharmacogenomics J. 2005;5(4):276. Rogan, PK [added]
19. Genetic Data Analysis Software. Lewis and Zaykin. 1999.
20. Page RDM (1996) TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357–358
21. TreeView Version 1.6.6 Build 7601. http://taxonomy.zoology.gla.ac.uk/rod/treeview/treeview_manual.html#citing_treeview
22. RStudio Team (2015) RStudio: integrated development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/
23. Population Analysis with Reticulate Trees (PopART). Developed by Jessica Leigh, David Bryant, and Mike Steel. http://popart.otago.ac.nz
24. Gaedigk A, Simon SD, Pearce RE, Bradford LD, Kennedy MJ, Leeder JS (2008) The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. Clin Pharmacol Ther 83(2):234–242
25. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F (2016) The Ensembl Variant Effect Predictor. Genome Biol 17(1):122. https://doi.org/10.1186/s13059-016-0974-4
26. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073–1081. https://doi.org/10.1038/nprot.2009.86
27. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80 **Review**
28. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814

29. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12(3):436–446
30. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11(5):863–874
31. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249. https://doi.org/10.1038/nmeth0410-248
32. Adzhubei I, Jordan DM, Sunyaev SR (2013;Chapter 7:Unit7.20) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. https://doi.org/10.1002/0471142905.hg0720s76
33. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7(10):e46688. https://doi.org/10.1371/journal.pone.0046688
34. Choi Y, Chan AP (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31(16):2745–2747. https://doi.org/10.1093/bioinformatics/btv195
35. Choi Y (2012) A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. In Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine (BCB '12). ACM, New York, 414–417
36. Sakuyama K, Sasaki T, Ujiie S, Obata K, Mizugaki M, Ishikawa M, Hiratsuka M (2008) Functional characterization of 17 CYP2D6 allelic variants (CYP2D6.2, 10, 14A-B, 18, 27, 36, 39, 47-51, 53-55, and 57). Drug Metab Dispos 36(12):2460–2467. https://doi.org/10.1124/dmd.108.023242
37. Marez D, Legrand M, Sabbagh N, Lo Guidice JM, Spire C, Lafitte JJ, Meyer UA, Broly F (1997) Polymorphism of the cytochrome P450 CYP2D6 gene in a European population: characterization of 48 mutations and 53 alleles, their frequencies and evolution. Pharmacogenetics 7(3):193–202
38. Sachse C, Brockmöller J, Bauer S, Roots I (1997) Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. Am J Hum Genet 60(2):284–295
39. Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37(9):e67. https://doi.org/10.1093/nar/gkp215
40. Wennerholm A, Johansson I, Hidestrand M, Bertilsson L, Gustafsson LL, Ingelman-Sundberg M (2001) Characterization of the CYP2D6*29 allele commonly present in a black Tanzanian population causing reduced catalytic activity. Pharmacogenetics 11(5):417–427 Erratum in: Pharmacogenetics 2001 Nov;11(8):743
41. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 24(8):2125–2137. https://doi.org/10.1093/hmg/ddu733.
42. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, Esko T, Mägi R, Inouye M, Lappalainen T et al (2014) Sequencing Initiative Suomi (SISu) Project. Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet 10(7):e1004494. https://doi.org/10.1371/journal.pgen.1004494
43. Peltonen L, Jalanko A, Varilo T (1999) Molecular genetics of the Finnish disease heritage. Hum Mol Genet 8(10):1913–1923 **Review**
44. Kere J (2001) Human population genetics: lessons from Finland. Annu Rev Genomics Hum Genet 2:103–128 **Review**
45. Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A (2009) Genetic markers and population history: Finland revisited. Eur J Hum Genet 17(10):1336–1346. https://doi.org/10.1038/ejhg.2009.53
46. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM (2006) International HapMap Consortium. Common deletion polymorphisms in the human genome. Nat Genet 38(1):86–92
47. Drögemöller BI, Wright GE, Niehaus DJ, Emsley R, Warnich L (2013) Next-generation sequencing of pharmacogenes: a critical analysis focusing on schizophrenia treatment. Pharmacogenet Genomics 23(12):666–674. https://doi.org/10.1097/FPC.0000000000000006
48. Wright GE, Carleton B, Hayden MR, Ross CJ (2016) The global spectrum of protein-coding pharmacogenomic diversity. Pharmacogenomics J. https://doi.org/10.1038/tpj.2016.77
49. Contreras AV, Monge-Cazares T, Alfaro-Ruiz L, Hernandez-Morales S, Miranda-Ortiz H, Carrillo-Sanchez K, Jimenez-Sanchez G, Silva-Zolezzi I (2011) Resequencing, haplotype construction and identification of novel variants of CYP2D6 in Mexican Mestizos. Pharmacogenomics 12(5):745–756. https://doi.org/10.2217/pgs.11.8
50. Beoris M, Amos Wilson J, Garces JA, Lukowiak AA (2016) CYP2D6 copy number distribution in the US population. Pharmacogenet Genomics 26(2):96–99. https://doi.org/10.1097/FPC.0000000000000188
51. Aklillu E, Persson I, Bertilsson L, Johansson I, Rodrigues F, Ingelman-Sundberg M (1996) Frequent distribution of ultrarapid metabolizers of debrisoquine in an ethiopian population carrying duplicated and multiduplicated functional CYP2D6 alleles. J Pharmacol Exp Ther 278(1):441–446
52. Bertilsson L, Dahl ML, Dalén P, Al-Shurbaji A (2002) Molecular genetics of CYP2D6: clinical relevance with focus on psychotropic drugs. Br J Clin Pharmacol 53(2):111–122 **Review**
53. Feng Y, Zhang Y, Ying C, Wang D, Nanopore-based DC (2015) Fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics 13(1):4–16. https://doi.org/10.1016/j.gpb.2015.01.009 Review. Erratum in: Genomics Proteomics Bioinformatics. 2015 Dec;13(6):383. Genomics Proteomics Bioinformatics. 2015 Jun;13(3):200-201
54. Wang D, Papp AC, Sun X (2015) Functional characterization of CYP2D6 enhancer polymorphisms. Hum Mol Genet 24(6):1556–1562. https://doi.org/10.1093/hmg/ddu566