Received: 28 October 2019 | Revised: 15 February 2020 | Accepted: 3 March 2020

DOI: 10.1111/cbdd.13676

SPECIAL ISSUE - AI IN DRUG DISCOVERY

WILEY

Benchmarking analysis of deleterious SNP prediction tools on CYP2D6 enzyme

Karla Cristina do Vale Ferreira[1,2] | **Leonardo Ferreira Fialho**[1] | **Octávio Luiz Franco**[1,2,3] | **Sérgio Amorim de Alencar**[1] | **William Farias Porto**[3,4]

[1]Programa de Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília, Brazil

[2]Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília, Brazil

[3]S-Inova Biotech, Pós Graduação em Biotecnologia, Universidade Católica Dom Bosco, Campo Grande, Brazil

[4]Porto Reports, Brasília, Brazil

Correspondence
William Farias Porto, S-Inova Biotech, Pós Graduação em Biotecnologia, Universidade Católica Dom Bosco, Campo Grande, MS, Brazil.
Email: williamfp7@yahoo.com.br

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul; Fundação de Apoio a Pesquisa do Distrito Federal

Abstract

The cytochrome P450 family is composed of hemeproteins involved in the metabolic transformation of endogenous and exogenous substances. The CYP2D6 enzyme is responsible for the metabolism of ~25% of clinically used drugs and is mainly expressed in the liver. The *CYP2D6* gene is known to have a large number of single nucleotide polymorphisms (SNPs). Nevertheless, these variations could modify the CYP2D6 enzyme's function, resulting in poor metabolizing or ultra-extensive metabolizing phenotypes, when metabolism is slower or accelerated, respectively. Currently, there are several computational tools for predicting functional changes caused by genetic variations. Here, we evaluated the predictive power of 20 web servers using a data set of 37 *CYP2D6* missense SNPs (2 neutral and 35 deleterious) previously reported in literature with enzymatic assays with the purified protein. The results suggest that the most appropriate tools for CYP2D6 SNP prediction are SDM and PoPMuSiC, which could aid in the classification of novel missense SNPs in this gene, providing the identification of mutations potentially associated with drug metabolism and pointing new directions for precise medicine.

KEYWORDS

cytochrome P450, disease, drug metabolism, hemeproteins

1 | INTRODUCTION

Intrinsic genetic variations in enzymes and transporters that metabolize drugs can influence the efficacy and toxicity of various drugs (Ahmed, Zhou, Zhou, & Chen, 2016). Different patients may respond differently to the same drug and dose. As a key element in precision medicine, the study of individuals' responses to medication based on their genomic information allows the evaluation of some specific genetic variants responsible for an individual's specific response to the drug (Ahmed et al., 2016). Sometimes, the effective dose of medication for a particular patient may be fatal or result in therapeutic failure in others, leading to serious adverse effects or even no effect (Ahmed et al., 2016).

In terms of drug metabolism, the monooxygenases from the cytochrome P450 (CYP) hemeprotein family play a crucial role in metabolizing a wide variety of substances by means of oxidation process (Sridhar, Liu, Foroozesh, & Stevens, 2012). Among such enzymes, the CYP2D6 enzyme represents ~3% of all CYP enzymes in the liver; nevertheless, this enzyme is involved in the metabolism of endogenous compounds, including hormones and a large number of commonly prescribed exogenous molecules, comprising many anti-arrhythmic, β-blockers, neuroleptics, selective serotonin reuptake inhibitors, and

Porto Reports: http://www.portoreports.com

tricyclic antidepressants (Ereshefsky, Riesenman, & Lam, 1995; Wang, Li, Dong, & Yue, 2014; Zanger et al., 2001). In addition, the conversion of prodrugs into active compounds or the conversion of drugs into toxic metabolites could involve this enzyme. Thus, CYP2D6 is related to the metabolism of about 25% of currently used drugs, despite its low amount in the liver (Ingelman-Sundberg, 2005b; Ingelman-Sundberg, Sim, Gomez, & Rodriguez-Antona, 2007).

The *CYP2D6* gene is highly polymorphic and the single nucleotide polymorphisms (SNPs) are the most common variations (Ingelman-Sundberg et al., 2007). SNPs are natural gene variations that may have a significant influence on metabolism, clinical efficacy, and side effects in drug therapy, depending on whether the amino acid sequence of the encoded protein is modified and its physiological consequences (Landau, 2005).

Genetic variations are related to dysfunctions in the metabolic capacity of drugs, influencing pharmacokinetics, pharmacodynamics, or both, which confer intrinsic differences in drug response (Ahmed et al., 2016). Therefore, polymorphisms in the *CYP2D6* gene could generate four distinct phenotypes based on the metabolizing capacity, including poor, intermediate, extensive, and ultra-rapid metabolizers (Arneth, Shams, Hiemke, & Härtter, 2009; Byeon et al., 2018). Hence, CYP2D6 has been considered an important target for pharmacogenomics and pharmacogenetics studies (Hicks et al., 2013; Ingelman-Sundberg, 2005; Koski, Ojanperä, Sistonen, Vuori, & Sajantila, 2007).

Currently, the potential effect of missense SNPs has been substantially assessed by means of computational tools. Several approaches have been used for this, because testing these SNPs in the laboratory can be expensive and time-consuming. Therefore, computational tool analysis has become a more accessible approach for preliminary analysis (Shen, Prescott, & Zhao, 2006). Missense SNPs could be classified as deleterious or neutral by a series of in silico tools used to predict missense SNP effects on protein function. In general, depending on the strategy used to develop the algorithm, these tools can be classified into four different groups. The first group is the sequence homology-based method, which uses sequence conservation information to assess whether an amino acid substitution affects protein function by means of homologous sequence multiple alignments (Ramensky, 2002). The second group consists of supervised learning methods, where the algorithm "learns" the variants' conservation patterns and/or physicochemical properties and uses this information to differentiate the SNPs studied by the user as neutral or potentially deleterious (Rodriguez-Casado, 2012). The third group comprises the structure-based methods that evaluate amino acid positions, considering factors such as solvent accessibility and the free energy difference between the wild-type and the mutated amino acids, to analyze the impact of modifications on the protein structure (Gonzalez-Castejon et al., 2011). Finally, the fourth group covers the consensus-based methods, a combination of a variety of methods into a consensus classifier, which could result in significantly improved prediction performance (González-Pérez & López-Bigas, 2011).

Such in silico prediction tools have become a target of a number of studies to define the predictive capacity of these tools, in order to indicate the best programs for the evaluation of certain proteins (Arooj et al., 2019; Grimm et al., 2015; Kerr et al., 2017). The predictive capacity of each algorithm is evaluated by statistical measures of performance, which refer to the probability of identifying true deleterious and true neutral mutations (Rodrigues, Santos-Silva, Costa, & Bronze-da-Rocha, 2015). However, the most suitable software to perform the functional prediction of the SNPs present in the *CYP2D6* gene is not yet known. Therefore, in this study we aim to analyze the predictive ability of 20 web-based algorithms of CYP2D6 variants that were phenotypically characterized by in vitro studies.

2 | METHODS AND MATERIALS

2.1 | Data sets and accession codes

The NCBI Variation Viewer browser (Sayers et al., 2011) was used to access the dbSNP database (build 153). The *CYP2D6* Isoform 1 sequence was retrieved with the RefSeq accession code NM_000106.6. The CYP2D6 protein sequence (RefSeq: NP_000097.3) was obtained from NCBI platform, and the protein structure was obtained from the Protein Data Bank (PDB ID: 3QM4; Wang, Savas, Hsu, Stout, & Johnson, 2012). From this, other databases were used to determine the phenotypes related to drug metabolism associated with the SNPs present in the CYP2D6 enzyme. The Online Mendelian Inheritance in Man (OMIM) database is the primary repository of comprehensive, curated information on genes and genetic phenotypes and the relationships between them (Amberger, Bocchini, Schiettecatte, Scott, & Hamosh, 2015). The Pharmacogene Consortium Variation (PharmVar) is a resource for pharmacogenetics and genomics communities, which has information on genes involved in drug metabolism that contribute to drug transport and response (Gaedigk et al., 2018). The SuperCYP database contains information on metabolism and effects on drug degradation of 57 CYPs, allowing the tolerance to drug cocktails to be verified and to find alternative combinations, to efficiently use the metabolic pathways (Preissner et al., 2010).

Finally, only those SNPs with experiments with the purified protein were selected for our data set. The prevalence of our data set was measured as follows:

$$\text{Prevalence} = \frac{\text{Deleterious SNPs}}{\text{Total SNPs}} \times 100$$

The prevalence indicates how often deleterious SNPs occur in the data set and could be used as an indicative of how statistical tests should be used.

2.2 | Frequency data

The SNP frequency data of *CYP2D6* gene were obtained from the 1,000 Genomes Project (phase 3) browser (Altshuler et al., 2012). The browser provides the frequencies of all SNPs identified in the genomes of 2,504 individuals from 26 populations obtained through a combination of low-coverage (2–6×) whole genome sequence data, targeted deep (50–100×) exome sequencing, and dense SNP genotype data. The 26 populations studied were grouped by the predominant component of ancestry into five super-populations: African (AFR; 661 individuals), Admixed American (AMR; 347 individuals), East Asian (EAS; 504 individuals), South Asian (489 individuals), and European (EUR; 503 individuals). The fixation index (F_{st}) was measured according the model developed by Wright (Wright, 1951).

2.3 | Analysis of the distance between the SNP site and the drug target

The distance of amino acid variants in relation to the CYP2D6 enzyme active site was determined by the distance between the iron of the heme group of the enzyme and the α-carbon from the respective amino acid. The amino acid was considered as far from the active site if the distance was higher than 12 Å (Sun, Hou, & Zhang, 2014; Sun, Ji, Fu, Wang, & Zhang, 2013).

2.4 | Conservation analysis

The ConSurf server is a computational tool for estimating the conservation of evolutionary amino acid positions in a protein molecule based on the phylogenetic relations between homologous sequences (Celniker et al., 2013). Using the CYP2D6 3D structure obtained from the Protein Data Bank (PDB ID: 3QM4; Wang et al., 2012), ConSurf (in ConSeq mode) carried out a search for close homologous sequences using CSI-BLAST (Angermüller, Biegert, & Söding, 2012; three iterations and 0.0001 e-value cutoff) against the UNIREF-90 (Suzek, Huang, McGarvey, Mazumder, & Wu, 2007) protein database. The sequences were then clustered and highly similar sequences removed using CD-HIT (cutoff 95%; Li & Godzik, 2006).

2.5 | In silico functional analysis of missense SNPs in the *CYP2D6* gene

In order to evaluate the functional effect of the SNPs present in the *CYP2D6* gene, we used 20 computational tools classified into four groups. The first group is represented by the sequence homology-based methods (Table S1), whose tools used were SIFT (Kumar, Henikoff, & Ng, 2009), PROVEAN (Choi, Sims, Murphy, Miller, & Chan, 2012), Panther (Thomas et al., 2003), FATHMM (Shihab et al., 2013), and Mutation Assessor (Reva, Antipin, & Sander, 2011). The second group represented the supervised learning methods (Table S2), whose tools used were Hansa (Acharya & Nagarajaram, 2012), SNAP (Bromberg, Yachdav, & Rost, 2008), Suspect (Yates, Filippis, Kelley, & Sternberg, 2014), MutPred (Li et al., 2009), and PolyPhen-2 (Adzhubei et al., 2010). The third group of structure-based methods (Table S3) used the following tools: SDM (Pandurangan, Ochoa-Montaño, Ascher, & Blundell, 2017), PoPMuSiC (Dehouck, Kwasigroch, Gilis, & Rooman, 2011), HoTMuSiC (Pucci, Bourgeas, & Rooman, 2016), FoldX (Schymkowitz et al., 2005), and SNPMuSiC (Ancien, Pucci, Godfroid, & Rooman, 2018). Finally in the fourth group, of the consensus-based methods (Table S4), we selected the following tools: PredictSNP (Bendl et al., 2014), Condel (González-Pérez & López-Bigas, 2011), Meta-SNP (Capriotti, Altman, & Bromberg, 2013), TransFIC (Gonzalez-Perez, Deu-Pons, & Lopez-Bigas, 2012), and SNPeffect (Reumers, 2004).

2.6 | Statistical analysis

The prediction results of the impact of missense *CYP2D6* SNPs using the tools described above were further evaluated using the following statistical measures of performance:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})}$$

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

$$\text{DOR} = \frac{(\text{TP}+0.5)\times(\text{TN}+0.5)}{(\text{FP}+0.5)\times(\text{FN}+0.5)}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Sensitivity refers to the probability of identifying true deleterious mutations, whereas specificity represents the probability

of identifying true neutral mutations (Hicks, Wheeler, Plon, & Kimmel, 2011). The accuracy is the proportion of correct classifications (Berrar, 2018). Finally, the diagnostic odds ratio (DOR) is a prevalence-independent indicator of performance, ranging from 0 to infinity, where values lower than 1 indicate an inverse prediction; values equals to 1, a random prediction; and values higher than 1, a correct prediction, with higher values indicating better discriminatory test performance (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003). Since some divisions by zero may occur, the approximated DOR is calculated by adding 0.5 to all counts (Glas et al., 2003).

The effects of conservation on the measures of performance were also evaluated. The SNP data set was divided in variable, intermediary, and conserved positions according the ConSurf results, and then, the measures of performance were calculated.

3 | RESULTS

3.1 | Missense SNP selection

From the 586 Missense SNPs available on the NCBI Variation Viewer database (Sayers et al., 2011), a filtering procedure was performed to select only those validated by literature, OMIM, PharmVar or SuperCYP, selecting those which were validated by experiments with the purified protein, using this information to identify neutral and deleterious SNPs (Figure 1). After that, 37 missense SNPs were selected (Table S5), of which 2 were neutral and 35 deleterious, reaching a prevalence of 94.59%.

3.2 | SNP frequency in population

CYP2D6 activity varies widely within a population and among ethnically distinct populations (Bernard, 2006). Thus, the 1,000 Genomes Project was used to determine the frequency of the 37 missense SNPs on the distribution of the five major population groups (Table S6).

The AMR population showed the highest prevalence among the SNPs evaluated, with nine deleterious SNPs observed (R26H, S34P, V136I, A237S, R296C, V338M, E418K, R440H, and T486S). The population of East Asia presented eight deleterious SNPs (S34P, G42E, G169R, E215K, A237S, R296C, E418K, and T486S) and one neutral (K147R). African (R26H, S34P, V136I, E215K, R296C, V338M, E418K, and T486S) and European populations (R26H, R28C, S34P, F219S, A237S, R296C, R440H, and T486S) presented eight deleterious SNPs. The lowest number of variations was observed in the South Asian population, where there were only six delterious SNPs (R26H, S34P, A237S, R296C, H324P, and T486S).

The frequency data for 21 SNPs (A5V, R25Q, G42R, A85V, A90V, V104A, F120L, L142S, E156A, I297L, D336N, D337G, V342M, R343G, R344Q, R388H, E410K, R440C, R441C, H463D, and R497C) were not available on 1,000 Genomes Project. Ten SNPs had frequency in more

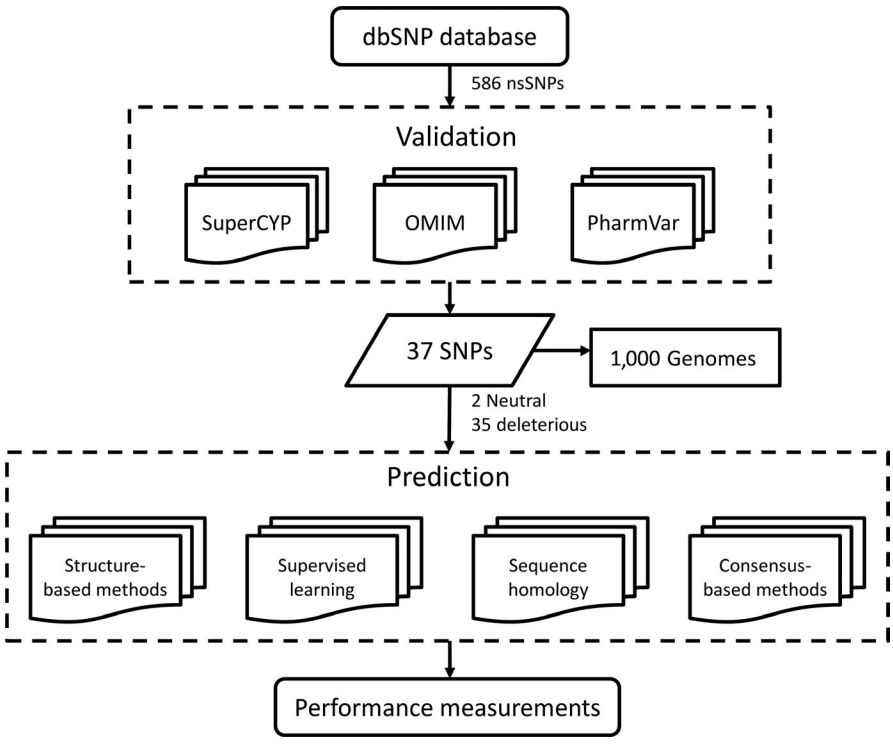

FIGURE 1 Flowchart used to evaluate the performance measures of missense SNP predictors present in the *CYP2D6* gene. SuperCYP, OMIM and PharmVar databases were used to determine phenotypes related to drug metabolism, associated with SNPs present in the CYP2D6 enzyme. Frequency information was retrieved from 1,000 Genomes Project. Analysis of the functional effect of the SNPs was carried out using 20 computational tools divided into four distinct groups that were later evaluated by the statistical measures of SEN, SPC, ACC and DOR

than one population (R26H, S34P, V136I, E215K, A237S, R296C, V338M, E418K, R440H, and T486S). However, according to the pairwise F_{ST}, only five of them presented moderate to great genetic differentiation in their respective populations: S34P (EAS); V136I and V338M (AFR); and R296C and T486S (EAS and AFR; Table 1).

3.3 | Distance between the SNP and the drug target

Table 2 shows the distance of the 37 variants in relation to the active site of CYP2D6 demonstrated that 34 out of the 37 SNPs are located far from the enzyme active site (>12 Å), having negligible influence on the drug binding to the enzyme.

3.4 | Conservation analysis

The 37 SNPs were mapped by means of ConSurf to identify the native residue conservation degree (Figure 2). These analyses showed that among the missense SNPs, fourteen deleterious ones (A5V, R25Q, R28C, V104A, F120L, E215K, F219S, R296C, I297L, D337G, R344Q, R388H, E418K, and T486S) and one neutral (A85V) were located in regions considered as variable, one neutral (K147R) and eleven deleterious (R26H, G42E, G42R, V136I, G169R, A237S, H324P, D336N, V342M, E410K, and H463D) were located in intermediate regions, and ten deleterious (S34P, A90V, L142S, E156A, V338M, R343G, R440C, R440H, R441C, and R497C) are located in highly conserved regions (Figure 2).

3.5 | Performance measurements

The 20 prediction tools were grouped according to their algorithms. The tools from the same group presented a similar behavior when predicting the 37 missense SNPs (Figure 3). Table S7 presents a detailed description of each prediction. The homology-based group tends to have a balanced prediction with similar number of neutral and deleterious predictions; the consensus group tends to perform more neutral predictions, while the structure-based one tends to perform more deleterious predictions; and the machine learning group is very heterogeneous in terms of prediction outcome.

To evaluate the performance of all the tools in this study, we used a set of statistics, such as SEN, SPC, ACC, and DOR (Table 3). For this purpose, the tests were classified as true positive (TP) if the test result corresponded to the deleterious class (pathogenic or harmful) and as true negative (TN) if the result corresponded to the neutral class (neutral or benign). Consequently, a false positive (FP) is a negative test that is classified as positive, and a false negative (FN) is a positive test classified as negative (Grimm et al., 2015).

Due to the high prevalence of deleterious SNPs in our data set (~95%), SEN, SPC, and ACC are biased. As observed in Table 3, the ACC is virtually the same as SEN, with an indifferent contribution of SPC. Thus, as observed in Figure 3, the structure-based group has the more accurate tools, as they

TABLE 1 Wright's Fixation indices (F_{st}) obtained from pairwise population comparisons

Mutation	AMR-AFR	AMR-EUR	AMR-EAS	AMR-SAS	AFR-EUR	AFR-EAS	AFR-SAS	EUR-EAS	EUR-SAS	EAS-SAS
R26H	<0.01	<0.01	<0.01	<0.01	0.01	0.01	<0.01	<0.01	<0.01	0.01
R28C	–	<0.01	–	–	<0.01	–	–	<0.01	<0.01	–
S34P	<0.01	<0.01	0.18	<0.01	0.02	0.24	0.01	0.14	<0.01	0.18
G42E	–	–	<0.01	–	–	<0.01	–	<0.01	–	<0.01
V136I	0.04	<0.01	<0.01	<0.01	0.05	0.05	0.05	–	–	–
K147R	–	–	<0.01	–	–	<0.01	–	<0.01	–	<0.01
G169R	–	–	<0.01	–	–	0.01	–	0.01	–	<0.01
E215K	<0.01	–	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	–	<0.01
F219S	–	<0.01	–	–	<0.01	–	–	<0.01	<0.01	–
A237S	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
R296C	0.05	<0.01	0.05	<0.01	0.04	0.18	0.04	0.06	<0.01	0.07
V338M	0.04	<0.01	<0.01	<0.01	0.05	0.05	0.05	–	–	–
H324P	–	–	–	<0.01	–	–	0.01	–	<0.01	<0.01
E418K	<0.01	<0.01	<0.01	<0.01	0.01	<0.01	0.01	<0.01	–	<0.01
R440H	<0.01	<0.01	<0.01	<0.01	<0.01	–	–	<0.01	<0.01	–
T486S	0.04	<0.01	0.05	<0.01	0.02	<0.01	0.02	0.03	<0.01	0.03

Abbreviations: AFR, African; AMR, Admixed American; EAS, East Asian; EUR, European; SAS, South Asian.

TABLE 2 Information of the distance between the active site and CYP2D6 enzyme variants

rsID	Mutation	Distance (Å)
rs773790593	A5V	ND
rs138417770	R25Q	ND
rs28371696	R26H	ND
rs138100349	R28C	ND
rs1065852	S34P	29.0
rs118203758	G42E	35.0
rs5030862	G42R	35.0
rs267608310	A85V	21.6
rs267608309	A90V	18.9
rs76187628	V104A	18.4
rs61736507	F120L	8.6
rs61736512	V136I	12.0
rs375135093	L142S	18.2
rs569229126	K147R	23.2
rs267608302	E156A	20.9
rs5030865	G169R	34.9
rs567606867	E215K	17.5
rs371793722	F219S	23.7
rs28371717	A237S	27.3
rs16947	R296C	18.0
rs949717872	I297L	15.3
rs5030867	H324P	27.2
rs745746329	D336N	31.4
rs748712690	D337G	32.3
rs59421388	V338M	30.2
rs750996195	V342M	32.3
rs267608295	R343G	30.1
rs76088846	R344Q	26.4
rs77312092	R388H	28.6
rs769157652	E410K	26.6
rs28371733	E418K	25.4
rs777560972	R440C	12.7
rs267608319	R440H	12.7
rs730882251	R441C	9.5
rs749471493	H463D	28.4
rs1135840	T486S	15.1
rs370580423	R497C	33.7

Abbreviation: ND, not determined.

performed more deleterious predictions. In order to overcome this bias, the DOR parameter was measured. The DOR indicated that some predictors have an inverse prediction, cases of SNAP2, Condel, and SNP Effect (Table 3). In contrast, the DOR indicated that the best programs were SDM and PoPMuSiC (both from structure-based methods), with the highest DOR values.

Interestingly, the amino acid position conservation shed some light on the prediction results. In variable positions (Table S8), only structure-based methods were able to correctly discriminate the deleterious SNPs; in intermediate conservation positions (Table S9), the other programs showed some increase in correct prediction rates; and in conserved predictions (Table S10), almost all programs performed accurate predictions.

4 | DISCUSSION

The family of cytochrome P450 enzymes has been the focus of pharmaceutical research for decades, as evidenced by over 100,000 articles in PubMed (Preissner et al., 2010), including determination of P450 2D6 structure by X-ray crystallography (Wang et al., 2012) adverse drug effects caused by inhibition of cytochrome P450 enzymes (Wu et al., 2019) and computational study of genetic variants to predict disease-associated amino acid mutations (Arooj et al., 2019). Despite the vast amount of information on CYP, understanding how the metabolism of this family of enzymes works is still complicated.

The *CYP2D6* gene is highly polymorphic and may include several SNPs at the same time, rather than a single-site mutation (Crews et al., 2012). Although this gene is known to have cumulative SNPs, it is possible to determine what a single polymorphism can cause in enzymatic activity. In this study, we selected 37 SNPs previously validated by in vitro studies (Table S5), of which 35 single polymorphisms indicated specific alterations where there was an increase, decrease, or absence of CYP2D6 enzyme metabolism. Despite most variants result in changes in metabolic function, more than 90% of disease-inducing mutations are located away from drug-binding sites (>12 Å), occurring at the protein interaction interfaces, not in the ligand-binding cavity, thus having negligible influence on the drug binding to enzyme (Sun et al., 2014; Sun et al., 2013). From 37 SNPs evaluated here, 34 out of 37 are located far from the active site (Table 2), with a distance >12 Å indicating that non-synonymous SNPs have little influence on the target binding site. These results are consistent with recent findings that most disease-inducing mutations occur at the protein interaction interfaces rather than the ligand-binding cavity (Monteiro, Franco, Alencar, & Porto, 2019; Sun et al., 2013; 2014).

ConSurf analysis demonstrated that most SNPs in the *CYP2D6* gene are present in regions considered to be variable or intermediate, which may have contributed to the decrease in accuracy of the tools of the sequence homology, machine learning and consensus groups (Tables S8-10).

FIGURE 2 Evolutionary conservation of CYP2D6 amino acid residues obtained from multiple sequence alignment by ConSurf. Color intensity increases with degree of conservation. The amino acids are colored based on their conservation grades and conservation levels. A grade of 1 indicates rapidly evolving (variable) sites, which are color-coded in turquoise; 5 indicates sites that are evolving at an average rate, which are colored white; and 9 indicates slowly evolving (evolutionarily conserved) sites, which are color-coded in purple. The deleterious predicted SNPs are marked below the amino acid sequence as red (one variant) or black (two variants) arrows, and the neutral SNPs are marked as blue arrows. e - an exposed residue according to the neural-network algorithm; b – a buried residue according to the neural-network algorithm; f – a predicted functional residue (highly conserved and exposed); s - a predicted structural residue (highly conserved and buried)

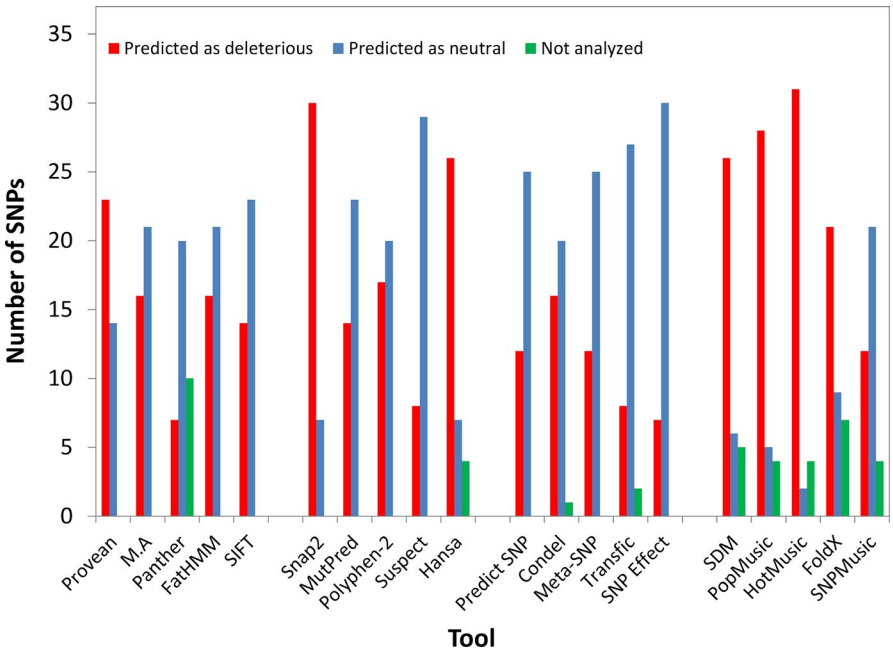

FIGURE 3 The deleterious and neutral predictions of CYP2D6 SNP data set (2 neutral and 35 deleterious). The prediction programs are divided into four groups according their algorithms (for a detailed description of each tool, see Tables S1-S4). The red bar represents the predicted as deleterious, the blue one, the predicted as neutral and the green one, those that were not analyzed. A detailed description of each prediction is available on Table S7

TABLE 3 Sensitivity (SEN), specificity (SPC), accuracy (ACC) and approximated diagnostic odds ratio (DOR) of the twenty predictive tools of SNPs

Tools	SEN (%)	SPC (%)	ACC (%)	DOR
SIFT	40	100	43	3.37
PROVEAN	63	50	62	1.67
Mutation assessor	46	100	49	4.23
Panther	28	100	33	2.03
FatHMM	46	100	49	4.23
Hansa	81	50	79	3.92
SNAP2	80	0	76	0.76
Suspect	23	100	23	1.55
MutPred	40	100	43	3.37
PolyPhen-2	49	100	51	4.73
PredictSNP	34	100	38	2.66
Condel	44	50	44	0.79
Meta-SNP	34	100	38	2.66
TransFIC	24	100	29	1.67
SNP Effect	17	50	19	0.22
SDM	83	50	81	4.64
PoPMuSiC	87	50	85	6.11
HoTMuSiC	94	0	88	2.36
FoldX	71	50	70	2.41
SNPMuSiC	39	100	42	3.21

The results indicate that the tools from the structure-based group were the most suitable for CYP2D6. However, we must take into account some biases from our SNP data set. Firstly, the high prevalence value (~95%) hinders the correct assessment of neutral SNPs; and secondly, some SNPs affect only specific drugs; thus, the two neutral SNPs might be deleterious, but were not covered by literature. In fact, in a general manner, the predictors are constructed to identify changes in several genes; and the low accuracy is likely related to the fact that these tools were not specifically constructed for the *CYP2D6* gene. Based on this, a tool was built specifically for the CYP, the web server called SCYPPred (Li, Wei, Wang, & Chou, 2012), which was developed for the prediction of human cytochrome P450 SNPs based on the SVM flanking sequence method, but, unfortunately, this server was unavailable during the manuscript preparation. Despite these limitations, the information here reported could be useful to select the most suitable combination of algorithms for cytochrome P450 family, for reducing the amount of variants to be analyzed. Arooj et al. (2019) used a similar approach to analyze harmful CYP11B2 variants, applying a combination of computational tools to reduce the amount of variants to perform in silico structural analysis (Arooj et al., 2019).

5 | CONCLUSION

Once CYP2D6 is a key in drug metabolism, the correct prediction of its variants effect could be valuable to develop more precise treatments according the populations. From the frequency information obtained from 1,000 Genomes project, we observed that some variations have some degree of differentiation in specific populations (Table 1), which could point different strategies for them. However, only five variations have frequencies >1% in some populations, and thus, the other variants could be considered

as rare mutations; and those could be important for developing personalized medicine. To this end, instead of the traditional "one-size-fits-all" approach to drug therapy, the application of targeted drug therapy could allow a more personalized approach, thus minimizing the occurrence of therapeutic failures or adverse effects (Ahasic & Christiani, 2015; Hertz & McLeod, 2016).

There is a vast range of tools for predicting the functional effect of missense SNPs based on different methodologies. Identifying the best prediction tools is of great importance for pharmacogenomics to determine responses to drug action in relation to the genetic variations of individuals. The best tools for CYP2D6 were SDM and PoPMuSiC independent of the amino acid position conservation. With the application of such tools, predictions with greater confidence could be reached for this gene. This could aid in the identification of deleterious SNPs potentially associated with deficiencies in the metabolism of the CYP2D6 enzyme present in the human population.

ACKNOWLEDGEMENTS

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Apoio a Pesquisa do Distrito Federal (FAPDF), and Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul (FUNDECT).

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supplementary material of this article.

ORCID

Karla Cristina do Vale Ferreira https://orcid.org/0000-0001-7461-3860
Octávio Luiz Franco https://orcid.org/0000-0001-9546-0525
Sérgio Amorim de Alencar https://orcid.org/0000-0001-8236-6468
William Farias Porto https://orcid.org/0000-0002-4739-8250

REFERENCES

Acharya, V., & Nagarajaram, H. A. (2012). Hansa: An automated method for discriminating disease and neutral human nsSNPs. *Human Mutation*, *33*(2), 332–337. https://doi.org/10.1002/humu.21642

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Ahasic, A. M., & Christiani, D. C. (2015). Personalized critical care medicine: How far away are we? *Seminars in Respiratory and Critical Care Medicine*, *36*(6), 809–822. https://doi.org/10.1055/s-0035-1564852

Ahmed, S., Zhou, Z., Zhou, J., & Chen, S. Q. (2016). Pharmacogenomics of drug metabolizing enzymes and transporters: Relevance to precision medicine. *Genomics, Proteomics and Bioinformatics*, *14*(5), 298–313. https://doi.org/10.1016/j.gpb.2016.03.008

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Lacroute, P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*(D1), 789–798. https://doi.org/10.1093/nar/gku1205

Ancien, F., Pucci, F., Godfroid, M., & Rooman, M. (2018). Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Scientific Reports*, *8*(1), 4480. https://doi.org/10.1038/s41598-018-22531-2

Angermüller, C., Biegert, A., & Söding, J. (2012). Discriminative modelling of context-specific amino acid substitution probabilities. *Bioinformatics*, *28*(24), 3240–3247. https://doi.org/10.1093/bioinformatics/bts622

Arneth, B., Shams, M., Hiemke, C., & Härtter, S. (2009). Rapid and reliable genotyping procedure for detection of alleles with mutations, deletion, or/and duplication of the CYP2D6 gene. *Clinical Biochemistry*, *42*(12), 1282–1290. https://doi.org/10.1016/j.clinbiochem.2009.04.009

Arooj, A., Pervez, M. T., Gillani, Z., Chohan, T. A., Tayyab Chaudhry, M., Babar, M. E., & Shah, A. T. (2019). In-Silico analysis of ns-SNPs associated with CYP11B2 Gene. *BioRxiv*, 602–615. https://doi.org/10.1101/602615

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., … Damborsky, J. (2014). PredictSNP: Robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Computational Biology*, *10*(1), e1003440. https://doi.org/10.1371/journal.pcbi.1003440

Bernard, S., Neville, K. A., Nguyen, A. T., & Flockhart, D. A. (2006). Interethnic differences in genetic polymorphisms of CYP2D6 in the U.S. population: Clinical implications. *The Oncologist*, *11*(2), 126–135. https://doi.org/10.1634/theoncologist.11-2-126

Berrar, D. (2018). Performance measures for binary classification. *Encyclopedia of Bioinformatics and Computational Biology*, *1*, 546–560. https://doi.org/10.1016/b978-0-12-809633-8.20351-8

Bromberg, Y., Yachdav, G., & Rost, B. (2008). SNAP predicts effect of mutations on protein function. *Bioinformatics*, *24*(20), 2397–2398. https://doi.org/10.1093/bioinformatics/btn435

Byeon, J.-Y., Kim, Y.-H., Lee, C.-M., Kim, S.-H., Chae, W.-K., Jung, E.-H., … Lee, Y. J. (2018). CYP2D6 allele frequencies in Korean population, comparison with East Asian, Caucasian and African populations, and the comparison of metabolic activity of CYP2D6 genotypes. *Archives of Pharmacal Research*, *41*(9), 921–930. https://doi.org/10.1007/s12272-018-1075-6

Capriotti, E., Altman, R. B., & Bromberg, Y. (2013). Collective judgment predicts disease-associated single

nucleotide variants. *BMC Genomics*, *14*(Suppl 3), S2. https://doi.org/10.1186/1471-2164-14-S3-S2

Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., … Ben-Tal, N. (2013). ConSurf: Using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry*, *53*(3–4), 199–206. https://doi.org/10.1002/ijch.201200096

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, *7*(10), e46688. https://doi.org/10.1371/journal.pone.0046688

Crews, K. R., Gaedigk, A., Dunnenberger, H. M., Klein, T. E., Shen, D. D., Callaghan, J. T., … Skaar, T. C. (2012). Clinical pharmacogenetics implementation consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype. *Clinical Pharmacology and Therapeutics*, *91*(2), 321–326. https://doi.org/10.1038/clpt.2011.287

Dehouck, Y., Kwasigroch, J. M., Gilis, D., & Rooman, M. (2011). PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, *12*(1), 151. https://doi.org/10.1186/1471-2105-12-151

Ereshefsky, L., Riesenman, C., & Lam, Y. W. F. (1995). Antidepressant Drug Interactions and the Cytochrome P450 System. *Clinical Pharmacokinetics*, *29*(Supplement 1), 10–19. https://doi.org/10.2165/00003088-199500291-00004

Gaedigk, A., Ingelman-Sundberg, M., Miller, N. A., Leeder, J. S., Whirl-Carrillo, M., & Klein, T. E. (2018). The Pharmacogene variation (PharmVar) Consortium: Incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clinical Pharmacology and Therapeutics*, *103*(3), 399–401. https://doi.org/10.1002/cpt.910

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, *56*(11), 1129–1135. https://doi.org/10.1016/S0895-4356(03)00177-X

Gonzalez-Castejon, M., Marin, F., Soler-Rivas, C., Reglero, G., Visioli, F., & Rodriguez-Casado, A. (2011). Functional non-synonymous polymorphisms prediction methods: Current approaches and future developments. *Current Medicinal Chemistry*, *18*(33), 5095–5103. https://doi.org/10.2174/092986711797636081

Gonzalez-Perez, A., Deu-Pons, J., & Lopez-Bigas, N. (2012). Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Medicine*, *4*(11), 89. https://doi.org/10.1186/gm390

González-Pérez, A., & López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics*, *88*(4), 440–449. https://doi.org/10.1016/j.ajhg.2011.03.004

Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., … Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, *36*(5), 513–523. https://doi.org/10.1002/humu.22768

Hertz, D. L., & McLeod, H. L. (2016). Integrated patient and tumor genetic testing for individualized cancer therapy. *Clinical Pharmacology and Therapeutics*, *99*(2), 143–146. https://doi.org/10.1002/cpt.294

Hicks, J. K., Swen, J. J., Thorn, C. F., Sangkuhl, K., Kharasch, E. D., Ellingrod, V. L., … Stingl, J. C. (2013). Clinical pharmacogenetics implementation consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants. *Clinical Pharmacology and Therapeutics*, *93*(5), 402–408. https://doi.org/10.1038/clpt.2013.2

Hicks, S., Wheeler, D. A., Plon, S. E., & Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human Mutation*, *32*(6), 661–668. https://doi.org/10.1002/humu.21490

Ingelman-Sundberg, M. (2005a). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): Clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics Journal*, *5*(1), 6–13. https://doi.org/10.1038/sj.tpj.6500285

Ingelman-Sundberg, M. (2005b). The human genome project and novel aspects of cytochrome P450 research. *Toxicology and Applied Pharmacology*, *207*(2 SUPPL.), 52–56. https://doi.org/10.1016/j.taap.2005.01.030

Ingelman-Sundberg, M., Sim, S. C., Gomez, A., & Rodriguez-Antona, C. (2007). Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacology and Therapeutics*, *116*(3), 496–526. https://doi.org/10.1016/j.pharmthera.2007.09.004

Kerr, I. D., Cox, H. C., Moyes, K., Evans, B., Burdett, B. C., van Kan, A., … Eggington, J. M. (2017). Assessment of in silico protein sequence analysis in the clinical classification of variants in cancer risk genes. *Journal of Community Genetics*, *8*(2), 87–95. https://doi.org/10.1007/s12687-016-0289-x

Koski, A., Ojanperä, I., Sistonen, J., Vuori, E., & Sajantila, A. (2007). A fatal doxepin poisoning associated with a defective CYP2D6 genotype. *American Journal of Forensic Medicine and Pathology*, *28*(3), 259–261. https://doi.org/10.1097/PAF.0b013e3180326701

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Landau, R. (2005). Pharmacogenetics: Implications for obstetric anesthesia. *International Journal of Obstetric Anesthesia*, *14*(4), 316–323. https://doi.org/10.1016/j.ijoa.2005.03.005

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., … Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, *25*(21), 2744–2750. https://doi.org/10.1093/bioinformatics/btp528

Li, L., Wei, D.-Q., Wang, J.-F., & Chou, K.-C. (2012). SCYPPred: A web-based predictor of SNPs for human cytochrome P450. *Protein and Peptide Letters*, *19*(1), 57–61. https://doi.org/10.2174/092986612798472785

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Monteiro, L. L. S., Franco, O. L., Alencar, S. A., & Porto, W. F. (2019). Deciphering the structural basis for glucocorticoid resistance caused by missense mutations in the ligand binding domain of glucocorticoid receptor. *Journal of Molecular Graphics and Modelling*, https://doi.org/10.1016/j.jmgm.2019.07.020

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., & Blundell, T. L. (2017). SDM: A server for predicting effects of mutations on protein stability. *Nucleic Acids Research*, *45*(W1), W229–W235. https://doi.org/10.1093/nar/gkx439

Preissner, S., Kroll, K., Dunkel, M., Senger, C., Goldsobel, G., Kuzman, D., … Preissner, R. (2010). SuperCYP: A comprehensive database on Cytochrome P450 enzymes including a tool for analysis

of CYP-drug interactions. *Nucleic Acids Research*, *38*(suppl_1), D237–D243. https://doi.org/10.1093/nar/gkp970

Pucci, F., Bourgeas, R., & Rooman, M. (2016). Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Scientific Reports*, *6*(1), 23257. https://doi.org/10.1038/srep23257

Ramensky, V. (2002). Human non-synonymous SNPs: Server and survey. *Nucleic Acids Research*, *30*(17), 3894–3900. https://doi.org/10.1093/nar/gkf493

Reumers, J. (2004). SNPeffect: A database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Research, 33(Database*, *issue)*, D527–D532. https://doi.org/10.1093/nar/gki086

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17), e118–e118. https://doi.org/10.1093/nar/gkr407

Rodrigues, C., Santos-Silva, A., Costa, E., & Bronze-da-Rocha, E. (2015). Performance of in silico tools for the evaluation of UGT1A1 missense variants. *Human Mutation*, *36*(12), 1215–1225. https://doi.org/10.1002/humu.22903

Rodriguez-Casado, A. (2012). In silico investigation of functional ns-SNPs – an approach to rational drug design. *Research and Reports in Medicinal Chemistry*, *2*, 31–42. https://doi.org/10.2147/RRMC.S28211

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., … Ye, J. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *39*(Database), D38–D51. https://doi.org/10.1093/nar/gkq1172

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research*, *33*(Web Server), W382–W388. https://doi.org/10.1093/nar/gki387

Shen, J., Prescott, D., & Zhao, H. (2006). Applications of computational algorithm tools to identify functional SNPs in cytokine genes. *Cytokine*, *35*(1–2), 62–66.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., … Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*, *34*(1), 57–65. https://doi.org/10.1002/humu.22225

Sridhar, J., Liu, J., Foroozesh, M., & Stevens, C. L. K. (2012). Insights on cytochrome P450 enzymes and inhibitors obtained through QSAR studies. *Molecules*, *17*(8), 9283–9305. https://doi.org/10.3390/molecules17089283

Sun, H.-Y., Hou, T.-J., & Zhang, H.-Y. (2014). Finding chemical drugs for genetic diseases. *Drug Discovery Today*, *19*(12), 1836–1840. https://doi.org/10.1016/j.drudis.2014.09.013

Sun, H. Y., Ji, F. Q., Fu, L. Y., Wang, Z. Y., & Zhang, H. Y. (2013). Structural and energetic analyses of SNPs in drug targets and implications for drug therapy. *Journal of Chemical Information and Modeling*, *53*(12), 3343–3351. https://doi.org/10.1021/ci400457v

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), 1282–1288. https://doi.org/10.1093/bioinformatics/btm098

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., … Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141. https://doi.org/10.1101/gr.772403

Wang, X., Li, J., Dong, G., & Yue, J. (2014). The endogenous substrates of brain CYP2D. *European Journal of Pharmacology*, *724*, 211–218. https://doi.org/10.1016/j.ejphar.2013.12.025

Wang, A., Savas, U., Hsu, M.-H., Stout, C. D., & Johnson, E. F. (2012). Crystal structure of human cytochrome P450 2D6 with Prinomastat bound. *Journal of Biological Chemistry*, *287*(14), 10834–10843. https://doi.org/10.1074/jbc.M111.307918

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.

Wu, Z., Lei, T., Shen, C., Wang, Z., Cao, D., & Hou, T. (2019). ADMET evaluation in drug discovery. 19. Reliable Prediction of human cytochrome P450 inhibition using artificial intelligence approaches. *Journal of Chemical Information and Modeling*, *59*(11), 4587–4601. https://doi.org/10.1021/acs.jcim.9b00801

Yates, C. M., Filippis, I., Kelley, L. A., & Sternberg, M. J. E. (2014). SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of Molecular Biology*, *426*(14), 2692–2701. https://doi.org/10.1016/j.jmb.2014.04.026

Zanger, U. M., Fischer, J., Raimundo, S., Stüven, T., Evert, B. O., Schwab, M., & Eichelbaum, M. (2001). Comprehensive analysis of the genetic factors determining expression and function of hepatic CYP2D6. *Pharmacogenetics*, *11*(7), 573–585. https://doi.org/10.1097/00008571-200110000-00004

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Ferreira KCDV, Fialho LF, Franco OL, de Alencar SA, Porto WF. Benchmarking analysis of deleterious SNP prediction tools on CYP2D6 enzyme. *Chem Biol Drug Des*. 2020;96:984–994. https://doi.org/10.1111/cbdd.13676