DMD Fast Forward. Published on February 9, 2009 as doi:10.1124/dmd.108.026047

TITLE PAGE:

A BIOINFORMATICS APPROACH FOR THE PHENOTYPE PREDICTION OF NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS IN HUMAN CYTOCHROME P450S

LIN-LIN WANG, YONG LI, SHU-FENG ZHOU

Department of Nutrition and Food Hygiene, School of Public Health, Peking University, Beijing 100191, P. R. China (LL Wang & Y Li)

Discipline of Chinese Medicine, School of Health Sciences, RMIT University, Bundoora, Victoria 3083, Australia (LL Wang & SF Zhou).

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

RUNNING TITLE PAGE:

a) Running title: Prediction of phenotype of human *CYPs*.

b) Author for correspondence:

A/Prof. Shu-Feng Zhou, MD, PhD

Discipline of Chinese Medicine, School of Health Sciences, RMIT University, WHO Collaborating Center for Traditional Medicine, Bundoora, Victoria 3083, Australia.

Tel: + 61 3 9925 7794; fax: +61 3 9925 7178.

Email: shufeng.zhou@rmit.edu.au

c) Number of text pages: 21

Number of tables: 10

Number of figures: 2

Number of references: 40

Number of words in Abstract: 249

Number of words in Introduction: 749

Number of words in Discussion: 1459

d) Non-standard abbreviations: CYP, cytochrome P450; nsSNP, non-synonymous single nucleotide polymorphism.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

ABSTRACT

Non-synonymous single nucleotide polymorphisms (nsSNPs) in coding regions that can lead to amino acid changes may cause alteration of protein function and account for susceptivity to disease. Identification of deleterious nsSNPs from tolerant nsSNPs is important for characterizing the genetic basis of human disease, assessing individual susceptibility to disease, understanding the pathogenesis of disease, identifying molecular targets for drug treatment and conducting individualized pharmacotherapy. Numerous nsSNPs have been found in genes coding for human cytochrome P450s (CYPs) but there is poor knowledge on the relationship between the genotype and phenotype of nsSNPs in *CYPs*. We have identified a total of 791 validated nsSNPs in 57 validated human *CYP* genes from the NCBI dbSNP and SWISS-Prot databases. Using the PolyPhen and SIFT algorithms, 39-43 % of nsSNPs in *CYP* genes were predicted to have functional impacts on protein function. There was a significant concordance between the predicted results using SIFT and PolyPhen. A prediction accuracy analysis found that about 70% of nsSNPs were predicted correctly as damaging. Of nsSNPs predicted as deleterious, the prediction scores by SIFT and PolyPhen were significantly associated with the numbers of nsSNPs with known phenotype confirmed by benchmarking studies including site-directed mutagenesis analysis and clinical association studies. These amino acid substitutions are supposed to be the pathogenetic basis for the alteration of CYP enzyme activity and the association with disease susceptivity. This prediction analysis of nsSNPs in human *CYPs* would be useful for further genotype-phenotype studies on individual differences in drug metabolism and clinical response.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

INTRODUCTION

The most common type of genetic variation in the human genome occurs as single nucleotide polymorphisms (SNPs) (Cargill et al., 1999; Nadeau, 2002). Up to 14 April 2008, a total of 16,673,796 SNPs for 44 organisms and 14,708,752 SNPs in humans have been identified and deposited to the NCBI dbSNP (http://www.ncbi.nlm.nih.gov/sites/entrez, dbSNP Build 129). A small number of the SNPs have been found to be associated with some rare human diseases. However, not all SNPs can cause amino acid changes and correlate with human diseases. Non-synonymous SNPs (nsSNPs) are SNPs that occur in a coding region and cause an amino acid change in the corresponding protein (Ramensky et al., 2002). According to the two online databases, Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/omim) (Hamosh et al., 2005) and Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk) (Stenson et al., 2008), nsSNP variants account for almost half of all genetic changes related to human diseases. Hence, these nsSNPs are considered to be deleterious nsSNPs since they lead to dramatic phenotypic consequences. On the other hand, there are some nsSNPs that do not alter protein function although the first-order structure of the protein has changed, these are called tolerant nsSNPs. It is important to differentiate deleterious nsSNPs from tolerant nsSNPs in order to characterize the genetic basis of human diseases, to assess individual susceptibility to these diseases, to understand the pathogenesis of diseases, to identify molecular and potentially therapeutic targets and to predict clinical phenotypes.

Approximately 67,000-200,000 common nsSNPs have been discovered in the human (Ramensky et al., 2002; Hinds et al., 2005). Using an experimental approach to explore the possible impact on protein function and the association between these nsSNPs and disease would be extremely time-consuming and almost unlikely and probably suffer from low statistical power to distinguish disease-causing nsSNPs from non-disease-causing nsSNPs (Zhernakova et al., 2009). However, to prioritize candidate nsSNPs for their likely impact on protein function and disease susceptibility by bioinformatics methods can overcome

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

this problem. Bioinformatics approaches that are based on the biochemical severity of the amino acid substitution, and the protein sequence and/or structural information, can offer a more feasible means for phenotype prediction. The algorithm "Sorting intolerant from tolerant" (SIFT, http://blocks.fhcrc.org/sift/SIFT.html) (Ng and Henikoff, 2003) is used to predict the functional effect of an amino acid substitution according to sequence homology and the physical properties of amino acids. This can be applied to naturally occurring nsSNPs and laboratory-induced missense mutations. Another important algorithm is Polymorphism Phenotyping (PolyPhen, http://genetics.bwh.harvard.edu/pph/) (Ramensky et al., 2002), which predicts the possible impact of an amino acid substitution on the structure and function of a human protein based on straightforward physical and comparative considerations. Additionally, other algorithms also employ sequence and/or structural information, such as SNPs3D (http://www.snps3d.org/), PMUT (http://mmb2.pcb.ub.es: 8080/PMut/), and topoSNP (http://gila.bioengr.uic.edu/snp/toposnp/). Generally, these computational methods provide a feasible, high-throughput way to determine the impact of large numbers of nsSNPs on protein function.

Polymorphisms have been found in genes coding for drug metabolizing enzymes, drug transporters and drug targets, all of which are important in determining clinical response to drug treatment (Ingelman-Sundberg et al., 2007; Tomalik-Scharte et al., 2008; Zhou et al., 2009a). A considerable body of research on the relationship between the genotype and phenotype of nsSNPs in human cytochrome P450 (CYP) genes have been undertaken, but this is still limited to a small fraction of nsSNP identified. CYPs represent the most important phase I drug metabolizing enzymes, oxidizing a number of endogenous substrates such as steroids and eicosanoids and xenobiotics including various carcinogens, toxins, and more than 90% of therapeutic drugs (Nebert and Russell, 2002). Based on amino acid sequence similarity, *CYP*s are grouped into different families, subfamilies, and specific enzymes, of which CYP1, CYP2 and CYP3 members are the main enzymes contributing to the oxidative metabolism of more than 95% of clinical drugs. The CYP1 family is known to metabolize a number of procarcinogens and toxicants, such

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

as polycyclic aromatic hydrocarbons and arylamines (Nebert and Russell, 2002). The CYP2 and CYP3 families play a major role in the metabolism of clinical drugs, environmental compounds, arachidonic acid, bile acids and some steroids, whereas the CYP4 enzymes mainly metabolize endogenous compounds such as fatty acids, arachidonic acid, leukotrienes, and prostaglandins (Nebert and Russell, 2002). A wide interindividual variation has been observed in hepatic CYP content and activity, which contributes to the *in vivo* differences in the response to drugs (Ingelman-Sundberg et al., 2007). Genetic mutations can lead to variation in the enzyme expression and activity of many CYPs, especially CYP2C9, 2C19, and 2D6 (Ingelman-Sundberg et al., 2007; Zhou et al., 2008), resulting in changes in the clearance of a number of drugs. Consequently, genetic polymorphisms in *CYPs* may eventually cause a large variability in drug response and in susceptibility to adverse drug reactions. Although deleterious nsSNPs of *CYP*s have received great interest from experimental scientists, the functional impact of most nsSNPs in human *CYPs* is still unknown. As such, this study has been undertaken to predict the phenotype of nsSNPs of human *CYP* genes using *in silico* approaches and the predicted results were compared with published phenotypic studies.

METHODS

Nomenclature and validation of human *CYP* genes

The human *CYP* genes investigated in this study were named in accordance with the Human CYP allele Nomenclature Committee (http://www.imm.ki.se/CYPalleles/criteria.htm). Genomic sequence numbers of these genes are available at the CYP-allele nomenclature website and GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). The data on human *CYP* genes were all collected from Entrez Gene on NCBI Website (http://www.ncbi.nlm.nih.gov/sites/entrez) and HUGO gene database (http://www.genenames.org/). With improved knowledge of molecular biology, some previously used gene names have expired and some genes have no formal names, with only number codes. Since such genes cannot be located in current databases, they were excluded from this study.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Data mining for nsSNPs of human *CYP* genes

For the creation of the nsSNP of human CYPs identified so far, information on all the nsSNPs was collected from the NCBI dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and SWISS-Prot databases (http://ca.expasy.org/sprot/). Fields including gene symbol, gene name, mRNA accession number, protein accession number, SNP ID, amino acid residue 1 (wild-type), amino acid position, and amino acid residue 2 (missense) were captured.

Data compiling for phenotype of SNPs of human *CYP* genes

For the creation of the phenotype datasets, information on all the phenotype of SNPs of human *CYP* genes was compiled from the PubMed (http://www.ncbi.nlm.nih.gov/PubMed/), OMIM (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) and UniProtKB/Swiss-Prot databases (http://ca.expasy.org/sprot/). The information on the effect of the nsSNP on enzyme activity and the correlation between the nsSNP and disease obtained from both *in vivo* and *in vitro* experiments (e.g. site-directed mutagenesis analysis and clinical association studies) was all collected.

Prediction of the phenotype of nsSNPs in human *CYP* genes

Protein structural attributes like solvent accessibility, secondary structure formation and resulting protein stability are essential parameters to understand the impact of point mutations (Terp et al., 2002). Residue changes that have an impact on the biophysical and structural properties of protein are known to be pathogenic or deleterious (Ferrer-Costa et al., 2002). Predicting the putative effects of nsSNPs on protein function was performed using SIFT (http://blocks.fhcrc.org/sift/SIFT.html) and PolyPhen (http://genetics.bwh.harvard.edu/pph/). These two bioinformatics tools enable high-throughput prediction of the potential impact of nsSNPs and large-scale polymorphism analyses. Both SIFT and PolyPhen

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

provide a series of prediction score categories based on the probability that an nsSNP will be tolerant or deleterious. SIFT predicts the functional importance of amino acid substitutions based on sequence homology and the physical properties of amino acids (Ng and Henikoff, 2003). SIFT can be applied not only to naturally occurring nsSNPs but also to artificial missense mutations. SIFT uses multiple alignment information to predict tolerated and deleterious substitutions for every position of the query sequence. Positions important for function should be conserved in an alignment of the protein family, while unimportant positions should appear diverse in an alignment. Using NCBI PSI-BLAST (www.ncbi.nlm.nih.gov/BLAST/), SIFT performs as the following: 1) searching for similar sequences; 2) choosing closely related sequences that may share similar function to the query sequence; 3) obtaining the alignment of these chosen sequences; and 4) calculating normalized probabilities for all possible substitutions from the alignment. SIFT scores are designated as tolerant (0.201-1.00), borderline (0.101-0.20), potentially intolerant (0.051-0.10), or intolerant (0.00-0.05) (Ng and Henikoff, 2003). Likewise, PolyPhen predicts the possible impact of an amino acid substitution on the structure and function of a human protein, based on the sequence, phylogenetic and structural information characterizing the substitution. PolyPhen performs the prediction through sequence-based characterisation of the substitution site, calculation of position specific independent counts (PSIC) profile scores for two amino acid variants, and calculation of structural parameters like effect on the hydrophobicity, electrostatic interactions, and so forth and contacts (Ramensky et al., 2002). PolyPhen scores are classified as probably benign (0.000-0.999), borderline (1.000-1.249), potentially damaging (1.250- 1.499), possibly damaging (1.500-1.999), or damaging (≥ 2.000) (Ramensky et al., 2002; Xi et al., 2004).

Validation of the prediction

nsSNPs with experimental evidence of changed enzyme activity or disease association were considered "really" deleterious. The phenotypic data are from both *in vivo* and *in vitro* studies, in which the site-directed mutagenesis analyses often provide direct evidence indicating the functional impact of nsSNPs.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Prediction accuracy was analysed according to these positive findings from these benchmarking experiments.

Statistical Analyses

Concordance analysis between the functional consequences of each nsSNP predicted by the two *in silico* methods was assessed using Spearman's rank correlation coefficient ρ. Correlated analysis between prediction score for deleterious nsSNPs and number of functional nsSNPs confirmed by *in vivo* and *in vitro* experiments was used by Pearson's χ^2 test. *P*-values below 0.05 were considered statistically significant.

RESULTS

Selection of human nsSNPs of *CYP* genes

As shown in Figure 1, five steps were conducted to select and validate human *CYP* genes and the nsSNPs. From step 1 to step 2, a total of 57 validated human *CYP* genes were compiled (Table 1), whereas 58 pseudogenes identified were excluded for further nsSNP search. From step 3 to step 4, a total of 791 nsSNPs were collected from 54 human *CYP* genes, with a mean value of 14.6 nsSNP per *CYP* gene (see Supplementary Table 1). In our data search, some previously reported SNPs in dbSNP have been identified as invalid by later studies due to wrong sequencing and alignment. These erroneous SNPs have expired or have merged with other SNPs. Some *CYP* genes have been renamed. We have carefully cross-examined the databases and removed those old CYPs and invalid SNPs.

The nsSNPs mainly occurred in the following human *CYP* genes: *CYP21A2* (68 nsSNPs), *2D6* (52 nsSNPs), *2A6* (37 nsSNPs), *2B6* (32 nsSNPs), *3A4* (32 nsSNPs), *1A2* (31 nsSNPs), *2C19* (31 nsSNPs) , *17A1* (31 nsSNPs), *1B1* (30 nsSNPs), *2C9* (28 nsSNPs), *11B1* (26 nsSNPs), *1A1* (25 nsSNPs) , *5A1* (23

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

nsSNPs) , *27B1* (22 nsSNPs), and *11B2* (20 nsSNPs). This accounted for 8.60, 6.57, 4.68, 4.05, 4.05, 3.92, 3.92, 3.92, 3.79, 3.54, 3.29, 3.16, 2.91, 2.78, and 2.53 %, respectively. *CYP3A5*, *4A22*, *4V2*, *27A1*, *2C8*, *8A1*, *19A1*, *2A13*, *4F12*, *4F2*, *2J2*, and *11A1* contained 10-20 nsSNPs. *CYP2R1*, *4X1*, *7B1* and *27C1* had 1 nsSNP only. However, none of nsSNPs was found in *CYP2U1*, *4Z1*, and *46A1*.

Prediction of functional effect of nsSNPs of human *CYP* genes

Among the 791 nsSNPs of human *CYP* genes, 308 (38.94%) and 338 (42.73%) of which were predicted to be deleterious by SIFT and PolyPhen, respectively, whereas 460 (58.15%) and 430 (54.36%) were predicted as tolerated (Table 2). Thus, a slightly higher number of deleterious nsSNPs was obtained when the PolyPhen algorithm was used compared to the SIFT algorithm.

There was a significant similarity in the distribution of top 10 *CYP* genes with most frequent deleterious nsSNPs predicted by different algorithms, but some differences were noted (Table 3). When SIFT was applied for prediction, the 10 *CYP* genes with most frequent deleterious nsSNPs were *CYP21A2* (number of deleterious nsSNPs: 40), *17A1* (24), *1A2* (18), *1B1* (16), *2C9* (15), *2D6* (15), *27B1* (14), *1A1* (13), *2C19* (13), and *3A4* (12), accounting for 58.44% (180/308) of predicted deleterious nsSNPs. The top *CYP* genes with most frequent deleterious nsSNPs were *CYP21A2* (40), *17A1* (25), *1A2* (17), *1B1* (17), *2C9* (16), *2D6* (16), *27B1* (16), *1A1* (15), *3A4* (14), and *11B1* (14), accounting for 56.21% (190/338) of deleterious nsSNPs predicted using the PolyPhen algorithm. The 10 *CYP* genes containing most deleterious nsSNPs predicted by either SIFT or PolyPhen were *CYP21A2* (49), *17A1* (26), *1A2* (24), *2D6* (24), *2C9* (19), *1B1* (18), *3A4* (17), *27B1* (16), *2C19* (16), *1A1* (15), and *2B6* (15) (Table 3).

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Effect of predicted deleterious nsSNPs on amino acid changes

Representative deleterious nsSNPs predicted by both SIFT and PolyPhen algorithms and the corresponding amino acid substitutions of various *CYP* genes are listed in Table 4. For *CYP1A1*, both algorithms predicted 13 nsSNPs including Met66Val, Ile78Thr, Arg135Trp, Arg279Trp, Ile286Thr, Ile448Asn, Arg464Cys, Arg464Ser, Phe470Val, Arg477Trp, Pro492Arg, and Arg511Leu as damaging or potentially damaging. For *CYP2B6*, the predicted deleterious snSNPs by both SIFT and PolyPhen algorithms included Arg22Cys, Gly99Glu, Lys139Glu, Pro167Ala, and Ile328Thr. For *CYP2D6*, seven nsSNPs were predicted as deleterious by both SIFT and PolyPhen algorithms, including Gly42Arg, Gly169Cys, Ser311Leu, His324Pro, Arg343Gly, Trr355Cys, and Arg365His. For *CYP11B1*, Pro42Ser, Pro94Leu, Leu293Val and Ala368Asp were predicted to be intolerant by both SIFT and PolyPhen algorithms. For *CYP11B2*, *19A1*, *21A2*, and *27B1*, Val403Glu, Pro207Ser, Pro30Gln, and Val374Ala were predicted as deleterious, respectively, by both SIFT and PolyPhen algorithms.

Table 5 shows the common amino acid change of deleterious nsSNPs in human *CYP* genes predicted by the SIFT and PolyPhen algorithms. The most common amino acid of contig reference (wild-type) was Arg (n = 79), followed by Pro (n = 23), Gly (n = 20), Ile (n = 17), Leu (n = 15), Phe (n = 12), and Thr (n = 12), while Cys was the most common amino acid of missense (n = 30), followed by Leu (n = 16), His (n = 15), Ser (n = 15), Trp (n = 15), Arg (n = 13) and Gln (n = 12). Arg→Cys was the most frequent substitution (n = 22) due to nsSNPs in human *CYP* genes, followed by Arg→His (n = 13), Arg→Trp (n = 12), Arg→Gln (n = 10) and Pro→Leu (n = 10).

Potential effects of some selected amino acid substitution due to nsSNPs in human *CYP* genes according to the PolyPhen algorithm are shown in Table 6. These included disruption of annotated functional site (e.g. Cys437Tyr in *CYP19A1*), disruption of ligand binding site (e.g. Arg101Gln in *CYP2A13*,

Arg128Leu2A6 in *CYP2A6*, Arg433Trp and Trp120Arg in *CYP2C19*, and Arg130Gln in *CYP3A4*), and hydrophobicity change at buried site (e.g. Arg311Cys in *CYP2A6* and Gln214Leu in *CYP2C9*). Disruption of annotated functional site in CYP19A1 (aromatase) would lead to a complete loss of the enzyme activity and cause aromatase deficiency. Disruption of ligand binding site in human CYPs would alter ligand-enzyme interactions.

Concordance analysis of predicted results by SIFT and PolyPhen

To compare the prediction capacity, we conducted a concordance analysis between the functional consequences for 766 nsSNPs predicted by SIFT and PolyPhen (Table 7). Raw scores rather than the arbitrarily defined categories were used for the correlation analysis. There was a significant concordance between the predictions using both SIFT and PolyPhen algorithms (Spearman's $\rho = -0.640$, $P \leq 0.001$).

Validation of the prediction of the functional impact of nsSNPs

Overall prediction accuracy

When one nsSNP, found experimentally to be associated with a remarkable phenotype such as altered enzyme activity or disease, was predicted as deleterious, it was considered that the prediction on this nsSNPs was correct. The prediction was defined as an error if such a deleterious nsSNP was predicted as tolerant.

Based on the results of published *in vitro* and *in vivo* studies, 259 nsSNPs of human *CYP* genes in the databases and literatures have been reported to alter enzyme activity and correlate with disease. For each *CYP* gene, the amino acid substitutions and consequent phenotypic implications have been compiled (see Table 8 & Supplementary Table 1). These confirmed phenotypes of nsSNPs were related to alteration of enzyme activity, and then connected to susceptibility to disease such as congenital adrenal hyperplasia,

cerebrotendinous xanthomatosis, corticosterone methyloxidase deficiency, primary congenital glaucoma type 3A and vitamin D-dependent rickets, and poor metabolism of drugs.

Most amino acid changes of human CYPs cause decreased enzyme activity but with some exceptions such as Lys262Arg (785A>G) in *CYP2B6* (Kirchheiner et al., 2003), Ile269Phe (805A>T) in *CYP2C8* (Dai et al., 2001), and Ile331Val (991A>G) in *CYP2C19* (Rudberg et al., 2008) which cause increased enzyme activity and accelerated drug clearance. The phenotype of nsSNPs in human *CYP* genes also present in different metabolism status such as Leu160His (479T>A; *CYP2A6*2*) in *CYP2A6* accounting for poor metabolism of nicotine (Oscarson et al., 1999); Ser224Pro (670T>C; *CYP2A6*11*) in *CYP2A6* accounting for poor metabolism of tegafur in a Japanese gastric cancer patient (Daigo et al., 2002); Met1Val (1A>G) in *CYP2C19* for poor *S*-mephenytoin metabolism (Ferguson et al., 1998); Ile359Leu (1075A>C) in *CYP2C9* for poor tolbutamide and warfarin metabolism (King et al., 2004); Pro34Ser (100C>T), Gly42Arg (124G>A), Gly169Arg (505G>T), and Thr107Ile (320C>T) in *CYP2D6* for poor debrisoquine and sparteine metabolism (Marez et al., 1997); and Thr143Ala (427A>G), Arg158Cys (472C>T), Ile192Asn (575T>A), and Asn404Tyr (1210A>T) in *CYP2J2* for poor metabolism of arachidonic acid and linoleic acid (King et al., 2002).

In 259 confirmed phenotypes of human *CYP*s, about half of allelic variants were present in *CYP21A2* (61/259, 23.55%), *17A1* (29/259, 11.20%), *27B1* (18/259, 6.95%), and *2A6* (14/259, 5.40%) (Table 8 & Supplementary Table 2). Intriguingly, the frequency of deleterious nsSNPs found in *CYP21A2* gene, which contained 61 nsSNPs, was more than twice that of the second most frequent gene *CYP17A1*. The mutation of *CYP21A2* gene is related to congenital adrenal hyperplasia and hyperandrogenism (Tajima et al., 1993), while *CYP17A1* mutations are associated with adrenal hyperplasia type 5 (Monno et al., 1993).

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Figure 2 displays the prediction for the functional impact of the 259 nsSNPs in human *CYP* genes by SIFT and PolyPhen programs. The two algorithms had similar prediction accuracy. According to the above criteria, approximately 68.57% and 69.80% of the 259 nsSNPs were correctly predicted as deleterious using SIFT and PolyPhen, respectively; while the error prediction was 31.43% and 30.20%, respectively (Table 9). Based on the data from site-directed mutagenesis assay, 66.87% and 68.71% of the prediction was correct using SIFT and PolyPhen, respectively. The error prediction was 33.13% and 31.29%, respectively. In nsSNPs predicted as deleterious, the prediction scores by SIFT and PolyPhen were correlated with the numbers of functional nsSNPs confirmed by available phenotype data, whereas only PolyPhen scores were correlated with numbers of functional nsSNPs when confirmed by site-directed mutagenesis (Table 10).

In addition to the above results, additional deleterious nsSNPs were predicted by both SIFT and PolyPhen algorithms. These deleterious nsSNPs comprised 81 RefSNPs from NCBI dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and 13 SNPs from Swiss-Prot database (http://ca.expasy.org/sprot/). dbSNP of NCBI maps each submitted SNP assay (ss) to the genome and assigns a RefSNP accession ID (rs number) to each submitted SNP assay. However, the phenotypic prediction of these nsSNPs as deleterious has not been confirmed by data from functional studies (Table 4).

Prediction for Individual CYPs

CYP1A2 is a major hepatic CYP enzyme that metabolizes a number of drugs including phenacetin, caffeine, theophylline, tacrine, flutamide, thalidomide, clozapine, lidocaine, propranolol, 5,6-dimethylxanthenone-4 acetic acid and tizanidine (Zhou et al., 2000; Zhou et al., 2009a). CYP1A2 is one of the major enzymes that bioactivate a variety of procarcinogens and mutagens and thus induction of

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

CYP1A2 may increase the carcinogenicity of these compounds. This enzyme also metabolizes several important endogenous substances including steroids, retinols, melatonin, uroporphyrinogen and arachidonic acid (Zhou et al., 2009a). There are wide inter-individual differences (40- to 130-fold) in CYP1A2 expression and activity, and approximately 15- and 40-fold interindividual variations in CYP1A2 mRNA and protein expression levels have been observed in human livers (Zhou et al., 2009b). To date, more than 15 variant alleles of human *CYP1A2* gene have been identified (http://www.imm.ki.se/CYPalleles), and 142 SNPs have been found in the *CYP1A2* upstream sequence, introns and exons in NCBI dbSNP (http://www.ncbi.nlm.nih.gov/). Among the SNPs located in seven exons, many of them are nsSNPs including Leu15Phe, Ser18Cys, Phe21Leu, Phe41Arg, Pro42Arg, Glu44Lys, Gly73Arg, Thr83Met, Phe125Ile, Glu168Gln, Met180Val, Phe186Leu, Phe205Val, Ser211Cys, Arg281Trp, Ser298Arg, Gly299Ser, Ile314Val, Asp348Asn, Arg377Gln, Ile386Phe, Cys406Tyr, Arg431Trp, Thr438Ile, Arg456His, Arg457Trp and Gln478His. By using SIFT or PolyPhen algorithm, the nsSNPs Glu44Glys, Gly73Arg, Phe125Ile, Glu168Gln, Met180Val, Phe186Leu, Phe205Val, Ser211Cys, Arg281Trp, Ser298Arg, Gly299Ser, Ile314Val, Asp348Asn, Arg377Gln, Ile386Phe, Arg431Trp, Thr438Ile, Arg456His, Arg457Trp and Gln478His have been predicted as damaging (see Supplementary Table 1). nsSNPs such as Leu15Phe, Ser18Cys, Pro42Arg, Glu44Lys, Phe186Leu, Asp348Asn, Ile386Phe, Arg431Trp, and Arg456His have been found to cause a decreased enzyme activity in *in vitro* and *in vivo* studies (Zhou et al., 2009b), which indicates that our prediction is correct.

CYP2A6 plays an important role in the metabolism of many therapeutic drugs, environmental toxicants, as well as metabolic activation of procarcinogens such as nicotine, nitrosamines and aflatoxin B_1. CYP2A6 metabolizes about 1% of clinical drugs, including halothane, tegafur, cisapride, chlormethiazole, losigamone, letrozole, fadrozole, coumarin, pilocarpine, cyclophosphamide and ifosfamide (Zhou et al., 2008). Both *in vitro* and *in vivo* studies have demonstrated a wide (20- to >100-fold) inter-individual variation in CYP2A6 expression and activity, which is due primarily to genetic polymorphisms in the

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

CYP2A6 gene. The *CYP2A6* gene spans a region of approximately 6 kb pairs consisting of 9 exons and has been mapped to the long arm of chromosome 19 (between 19q12 and 19q13.2). It is located within a 350-kb pair gene cluster together with the *CYP2A7* and *2A13* genes, two *CYP2A7* pseudogenes, as well as genes in the *CYP2B* and *2F* subfamilies. To date, more than 33 variant alleles (**1B* to **34*) of the *CYP2A6* gene have been identified (http://www.imm.ki.se/CYPalleles). There are more than 28 non-synonymous SNPs in exons 1-9 of *CYP2A6*. These include 13G>A (Gly5Arg), 86G>A (Ser29Asn), 352T>C (Phe118Leu), 361G>C (Gly121Arg), 383G>A (Arg128Gln), 383G>A (Arg128Leu), 391T>G (Ser131Ala), 451G>A (Glu151Lys), 457T>C (Ala153Pro), 457T>C (Ala153Ser), 474C>G (Asp158Glu), 478C>A (Leu160Ile), 479T>A (Leu160His), 607C>A (Arg203Ser), 773C>A (Thr258Lys), 835G>C (Glu279Gln), 874G>A (Val292Met), 881C>G (Thr294Ser), 902G>C (Gly301Ala), 931C>T (Arg311Cys), 1093G>A (Val365Met), 1175T>A (Phe392Tyr), 1226A>G (Gln409Arg), 1252A>G (Asn418Asp), 1257G>C (Glu419Asp), 1412T>C (Ile471Thr), 1427A>G (Lys476Arg), 1436G>T (Gly479Val), and 1454G>T (Arg485Leu). The SIFT or PolyPhen program predicted the following nsSNPs as damaging: Gly5Arg, Phe118Leu, Gly121Arg, Arg128Gln, Arg128Leu, Ala153Pro, Ala153Ser, Leu160His, Arg257Cys, Thr258Lys, Arg311Cys, Asn418Ser, Ile471Thr, Gly479Val, and Arg485Leu. Functional assays have demonstrated that Val110Leu, Phe118Leu, Arg128Gln, Arg203Cys, Ser224Pro, Val365Met, Tyr392Phe, Ile471Thr, and Lys476Arg caused a marked decrease in enzyme activity or abolished the activity. However, Arg485Leu did not alter the enzyme activity.

The *CYP2A13* gene is located in the *CYP2* gene cluster on chromosome 19 and the nucleotide and protein sequences of CYP2A13 are highly similar to CYP2A6 with 95.3% and 93.5% identity, respectively. In the olfactory mucosa and respiratory tract, CYP2A13 is highly expressed as a functional protein. CYP2A13 has similar substrate specificity to 2A6 with some marked differences. CYP2A13 is active in the metabolism of a number of procarcinogens. CYP2A13 is the most efficient enzyme in the metabolic activation of the tobacco-specific procarcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone, a

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

tobacco-specific lung carcinogen (Brown et al., 2007). CYP2A13, but not CYP2A6, is also highly efficient in metabolizing the mycotoxins aflatoxin B_1 to its carcinogenic metabolites 8,9-epoxide and 1-8,9-epoxide. Although CYP2A13 is less active for coumarin 7-hydroxylation than CYP2A6, it is much more active for nicotine, cotinine, and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. Interestingly, CYP2A13 also metabolizes phenacetin and theophylline, two typical substrates of CYP1A2. To date, eight allelic variants of *CYP2A13* have been described (http://www.imm.ki.se/CYPalleles). There are several nsSNPs in its exons 1, 5, 6, and 8, include Arg25Gln, Arg101Gln, Asp158Glu, Arg257Cys, Pro321Leu, Val323L, and Phe392Tyr, Phe453Tyr, and Arg494Cys. Arg101Gln, Arg257Cys, Pro321Leu, Val323Leu, Phe453Tyr, and Arg494Cys were predicted to be deleterious by either SIFT or PolyPhen algorithm. Several alleles of *CYP2A13* found in Caucasian, African and Asian populations have functional impact on substrate metabolism and cancer risk (D'Agostino et al., 2008). A 30-42% decrease in coumarin 7-hydroxylation was observed for *CYP2A13*2* (Arg25GGln plus Arg257Cys) and **8* (1706C>G leading to Asp158Glu) (Schlicht et al., 2007). The Arg257Cys variant was 37 to 56% less active than the wild-type protein toward substrates such as hexamethylphosphoramide, 2'-methoxyacetophenone, *N*,*N*-dimethylaniline, and *N*-nitrosomethylphenylamine and it displayed a >2-fold decrease in catalytic efficiency toward 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (Zhang et al., 2002). The Arg at the 257 position is conserved in the CYP2As and seems to be located near the carboxyl end of the G-helix according to alignments based on sequence conservation. The residue is expected to be located on the surface of the protein, away from any of the proposed substrate access channels. However, conformational changes that occur with substrate binding may require the various helices involved in substrate binding to rotate and move, as found in the structure of P450 BM3 in complex with a substrate analog palmitoleic acid. It is likely that a mutation in the loop region impedes such changes and thus alters substrate binding or product release in a substrate-independent fashion. An Arg253Ala mutation close to the end of the G helix interfered with the interaction of rabbit CYP2B4 with the P450 reductase, leading to a ~50%

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

decrease in catalytic activity (Lehnerer et al., 2000). It is unknown if the Arg257Cys mutation affects CYP2A13- P450 reductase interactions.

The *CYP2B6* gene has been mapped to chromosome 19 between 19q12 and 19q13.2 which consists of nine exons encoding 491 amino acids. CYP2B6 can metabolise ~8% of all pharmaceutical drugs to some extent. These include cyclophosphamide, ifosfamide, tamoxifen, ketamine, artemisinin, nevirapine, efavirenz, bupropion, sibutramine, propofol, *S*-mephenytoin, selegiline, *S*-mephobarbital, thioTEPA, valproic acid, pethidine, perhexiline, and diazepam. CYP2B6 can metabolize procarcinogens including aflatoxin B_1, 6-aminochrysene, and 7,12-dimethylbenz[*a*]anthracene. There are large inter-individual differences in hepatic CYP2B6 protein and mRNA levels ranging from 20- to 278-fold. To date, at least 28 allelic variants of *CYP2B6* (**1B* to **29*) have been described (http://www.imm.ki.se/CYPalleles). There are many nsSNPs in exons 1-9 of *CYP2B6*, including 62A>T (Gln21Leu), 64C>T (Arg22Cys), 76A>T (Thr26Ser), 83A>G (Asp28Gly), 85C>A (Arg29Ser), 86G>C (Arg29Pro), 136A>G (Mer46Val), 296G>A (Gly99Glu), 415A>G (Lys139Gln), 419G>A (Arg140Gln), 499C>G (Pro167Ala), 503C>T (Thr168Ile), 516G>T (Gln172His), 546C>G (Ile182Met), 547G>A (Val183Ile), 593T>C (Met198Thr), 769G>A (Asp257Asn), 777C>A (Ser259Arg), 785A>G (Lys262Arg), 867C>G (Asn289Lys), 917C>G (Thr306Ser), 893T>C (Ile328Thr), 1006C>T (Arg336Cys), 1172T>A (Ile391Asn), 1190A>G (His397Arg), 1268C>A (Thr423Asn), 1282C>A (Pro428Thr), 1375A>G (Met459Val), and 1459C>T (Arg487Cys). By using the SIFT or PolyPhen algorithm, we predicted that Gln21Leu, Arg22Cys, Asp28Gly, Arg29Ser, Arg29Pro, Gly99Glu, Lys139Glu, Glu148AAsp, Pro167Ala, Met198Thr, Ser259Arg, Thr306Ser, Ile328Thr, Arg336Cys, and Pro428Thr were deleterious. *In vitro* and *in vivo* studies have demonstrated that Met46Val, Gly99Glu, Lys139Glu, Arg140Gln, TThr168Ile, Gln172His, Met198Thr, Lys262Arg, Ile328Thr, Arg336Cys, Ile391Asn, Pro428Thr caused a marked decrease in enzyme activity.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

CYP2C8 accounts for ~7% of total hepatic CYP contents and metabolizes ~5% of drugs cleared by Phase I reaction. The *CYP2C8* gene was cloned in 1999, which was found to span 31 kb and contain 9 exons. The prototypical substrate for CYP2C8 is the potent antimicrotubule drug paclitaxel, and its 6α-hydroxylation has been widely used in *in vitro* reaction phenotyping (Cresteil et al., 2002). CYP2C8 contributes substantially to the biotransformation of a variety of clinical drugs, including antimalarial agents (e.g. amodiaquine and chloroquine, thiazolidinedione antidiabetic drugs (e.g. troglitazone, rosiglitazone, pioglitazone (also minor contribution from CYP2C9 and 3A4)), statins (e.g. cerivastatin and fluvastatin, atorvastatin, and simvastatin, also contribution from CYP3A4 and 2C9), opioids (e.g. morphine, methadone, buprenorphine, and loperamide), repaglinide (a hypoglycaemic drug that stimulates insulin secretion), and *R*-ibuprofen. CYP2C8 is also involved in the oxidation of verapamil, gallopamil (a methoxy derivative of verapamil and a calcium antagonist used for the treatment of angina pectoris), dapsone, amiodarone, diclofenac, perphenezine, amitriptyline, carbamazepine, bortezomib, cisapride, and omeprazole, but other CYPs play a more important role in the metabolism of these drugs. Endogenous retinoids and arachidonic acid are also metabolized by human CYP2C8. To date, at least 13 variants of *CYP2C8* have been identified and designated *CYP2C8*1B* to **14* (http://www.imm.ki.se/CYPalleles). There are about ten nsSNPs found in its exons 2, 3, 4, 5, 7, 8, and 9, including 244G>T (Ala82Ser), 416A>G (Arg139Lys), 556C>G (Arg186Gly), 541G>A (Val181Ile), 730A>G (Ile244Val), 792C>G (Ile264Met), 805A>T (Ile269Phe), 1081C>T (Leu361Phe), and 1196A>G (Lys399Arg). The nsSNPs Arg186Gly, Ile269Phe, and His411Leu were predicted to be deleterious by the PolyPhen algorithm. Functional studies have shown that Arg139Lys, Arg186Gly, Ile269Phe, and Lys399Arg caused decreased enzyme activity. However, both Arg139Lys and Lys399Arg were predicted as tolerant by both SIFT and PolyPhen algorithms.

The *CYP2C9* gene has been mapped to the long arm of chromosome 10, located in a densely packed region also containing genes encoding CYP2C8, 2C18 and 2C19. *CYP2C9* encodes a protein of 490

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

amino acids, with a molecular weight of 55.6 kDa. CYP2C9 metabolizes approximately 15% clinical drugs, including non-steroid anti-inflammatory drugs (e.g. diclofenac, ibuprofen, ketoprofen, suprofen, naproxen, flurbiprofen, indomethacin, meloxicam, piroxicam, tenoxicam, and lornoxicam), sulfonylurea hypoglycemics (e.g. tolbutamide, glyburide, glimepiride, gliclazide and glipizide), cyclooxygenase-2 inhibitors (e.g. celecoxib, etoricoxib, and valdecoxib), antiepileptics (e.g. phenytoin and phenobarbital), angiotensin II receptor inhibitors (e.g. losartan, irbesartan, and candesartan), anticancer drugs (e.g. cyclophosphamide and tamoxifen), and anticoagulants (e.g. *S*-acenocumarol, phenprocoumon and *S*-warfarin) (Miners and Birkett, 1998). About 30 allelic variants have been detected within *CYP2C9* (http://www.imm.ki.se/CYPalleles) and some of them caused decreased enzyme activity and poor drug metabolism. nsSNPs in *CYP2C9* include 269T>C (Leu90Pro), 334A>C (Ile112Leu), 371G>A (Arg124Gln), 374G>A (Arg125His), 389C>G (Thr130Arg), 395G>A (Arg132Gln), 430C>T (Arg144Cys), 448C>T (Arg150Cys), 449G>A (Arg150His), 641A>T (Gln214Leu), 752A>G (His251Arg), 815A>G (Glu272Gly), 895A>G (Thr299Ala), 980T>C (Ile327Thr), 1003C>T (Arg335Trp), 1010C>T (Pro337Arg), 1073A>G (Tyr358Cys), 1075A>C (Ile359Leu), 1076T>C (Ile359Thr), 1080C>G (Asp360Glu), 1238T>C (Leu413Pro), 1341A>C (Leu447Phe), 1429G>A (Ala477Thr), and 1465C>T (Pro489Ser). Using SIFT or PolyPhen programs, Arg124Gln, Arg125His, Thr130Arg, Arg144Cys, Arg144His, Gln214Leu, His251Arg, Glu272Gly, Thr299Ala, Ile327Thr, Arg335Trp, Pro337Arg, Ile359Leu, Tyr358Cys, Asp360Glu, Asp397Ala, Leu413Pro, Gly417Asp, Leu447Phe, and Pro489Ser. Benchmarking functional assays have found that Leu90Pro, Arg125His, Thr130Arg, Arg132Gln, Arg144Cys, Arg150His, Gln214Leu, Thr299Ala, Arg335Trp, Ile359Leu, Asp360Glu, Ala477Thr, and Pro489Ser caused decreased enzyme activity. However, both SIFT and PolyPhen failed to predict the phenotype of Leu90Pro, Arg150His and Ile359Leu.

CYP2C19 is involved in the metabolism of a number of drugs (~10%), including proton pump inhibitors (e.g. omeprazole, lansoprazole, pantoprazole, and rabeprazole), tricyclic antidepressants (e.g. imipramine,

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

amitriptyline and nortriptyline), selective serotonin reuptake inhibitors (e.g. citalopram, fluoxetine, and sertraline), benzodiazepines (e.g. diazepam, flunitrazepam, quazepam, and clobazam), barbiturates (e.g. hexobarbital, mephobarbital, and phenobarbital), phenytoin, *S*-mephenytoin, bortezomib, voriconazole, selegiline, nelfinavir and proguanil (Desta et al., 2002). There are over 20 known allelic variants (http://www.imm.ki.se/CYPalleles) in *CYP2C19*. Among all identified SNPs of *CYP2C19*, there are about 30 nsSNPs found in exons 3, 5, 7, 8, and 9. These include 1A>G (Met1Val), 50T>C (Leu17Pro), 55A>C (Ile19Leu), 221T>C (Met74Thr), 276G>C (Glu92Asp), 358T>C (Trp120Arg), 365A>C (Glu122Ala), 431G>A (Arg144His), 449G>A (Arg150His), 502T>C (Phe168Leu), 518C>T (Ala173Val), 527A>G (Asn176Ser), 680C>T (Pro227Leu), 836A>C (Gln279Pro), 839C>A (Ser280Tyr), 905C>G (Thr302Arg), 985C>T (Arg329Cys), 991G>A (Val331Ile), 1030C>T (His344Tyr), 1180G>A (Val394Met), 1228C>T (Arg410Cys), 1297C>T (Arg433Trp), and 1390C>A (Pro464Thr). Functional studies have demonstrated that several nsSNPs, including Met1Val, Trp120Arg, Arg132Gln, Arg144His, Pro227Leu, Arg433Trp and Arg442Cys, led to an abolished or decreased enzyme activity and accounted for a poor drug metabolism phenotype (Zhou et al., 2008). All of these nsSNPs were predicted to affect protein function by SIFT or PolyPhen. Additional nsSNPs were predicted deleterious, including Leu17Pro, Ala161Pro, Ala173Val, Asn176Ser, Trp212Cys, Thr302Arg, Arg329Cys, Val394Met, and Arg410Cys, and their functional impact should be investigated. Two variants on special positions, Trp120Arg and Arg433Trp, were predicted as probably damaging with high PolyPhen scores, because they can cause disruption of ligand binding site.

The *CYP2D6* gene is mapped to chromosome 22q13.1 and consists of nine exons with an open reading frame of 1,491 base pairs coding for 497 amino acids. CYP2D6 is responsible for the metabolism of up to 25% of the commonly used drugs. Drugs that are extensively metabolized by CYP2D6 include tricyclic antidepressants (e.g. clomipramine, imipramine, doxepin, desipramine, and nortriptyline), SSRIs (fluoxetine, fluvoxamine, and paroxetine), other non-tricyclic antidepressants (atomoxetine, maprotiline,

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

mianserin, and venlafaxine), neuroleptics (e.g. chlorpromazine, perphenazine, thioridazine, zotepine, zuclopenthixol, mianserin, olanzapine, risperidone, sertindole, and haloperidol), and β-blockers (e.g. atenolol, bufuralol, carvedilol, metoprolol, bisoprolol, propranolol, bunitrolol, bupranolol, timolol and alprenol) (Ingelman-Sundberg et al., 2007; Tomalik-Scharte et al., 2008; Zhou et al., 2009a). CYP2D6 also extensively metabolizes opioids (e.g. codeine, dihydrocodeine and tramadol), antiemetics (tropisetron, ondansetron, dolasetron, and metoclopramid), antihistamines (e.g. terfenadine, oxatomide, loratadine, astemizole, epinastine, promethazine, mequitazine, azelastine, diphenhydramine and chlorpheniramine), and antiarrhythmics (e.g. sparteine, propafenone, encainide, flecainide, cibenzoline, aprindine, lidocaine, procainamide and mexiletine). There is a large inter-individual variation in the enzyme activity of CYP2D6. Unlike other CYPs, CYP2D6 is not inducible, and thus genetic mutations are largely responsible for the interindividual variation in enzyme expression and activity. To date, over 90 allelic variants of *CYP2D6* have been reported (http://www.imm.ki.se/CYPalleles). There are about 30 nsSNPs in *CYP2D6* reported. These include 31G>A (Val11Met), 77G>A (Arg26His), 100C>T (Pro34Ser), 124G>A (Gly42Arg), 271C>A (Leu91Met), 281A>G (His94Arg), 320C>T (Thr107Ile), 358T>A (Phe120Ile), 364G>T (Gly122Ser), 463G>A (Glu155Lys), 496A>G (Asn166Asp), 501C>A (His167Gln), 502T>G (Ser168Ala), 505G>T (Gly169Cys), 635G>A (Gly212Glu), 692T>C (Leu231Pro), 709G>T (Ala237Ser), 886A>G (Asn285Ser), 886T>C (Cys296Arg), 899C>G (Ala300Gly), 932C>T (Ser311Lys), 971A>C (His324Pro), 986G>A (Gly329Val), 1012G>A (Val338Met), 1094G>A (Arg365His), 1117G>A (Gly373Ser), 1405C>G (Pro469Ala), 1408A>G (Thr470Ala), 1432C>T (His478Tyr), 1435G>C (Gly479Arg), 1441T>G (Phe481Val), and 1457C>G (Thr486Ser). Many of these SNPs were predicted to have phenotypical effect by SIFT or PolyPhen programs, including Arg28Cys, Pro34Ser, Gly42Arg, Ala85Val, Leu91Met, Trp152Gly, Trp152Arg, Gly169Cys, Gly169Arg, Leu213Pro, Met279Lys, Ser311Leu, His324Pro, Arg329Leu, Val338Met, Arg343Gly, Tyr355Cys, Arg365His, Ile369Thr, Val374Met, Arh380His, Glu418Lys, Pro430Leu, and Pro469Ala. Although most of allele variants are usually found in terms of haplotypes, it has been identified that the phenotype of nsSNPs in

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

CYP2D6 is associated with the alteration of drug metabolism status such as Pro34Ser (*CYP2D6*10* and **14*), Thr107Ile (**17*), Gly42Arg (**12*) for impaired sparteine metabolism, Gly169Arg for poor debrisoquine metabolism, and Arg441Cys for a loss of enzyme activity (Zhou et al., 2008).

CYP3A4 has the highest abundance in the human liver (~40%) and metabolizes more than 50% of clinical drugs (Zhou, 2008). The substrate specificity of the CYP3A4 enzymes is very broad, with an extremely large number of structurally divergent chemicals being metabolized often in a regio- and stereo-selective fashion. *CYP3A4* gene is located on chromosome 7q22.1 and is about 27 kb long consisting of 13 exons and 12 introns. More than 19 *CYP3A4* variants *(*1B* through to **20*) have been identified to date (http://www.imm.ki.se/CYPalleles). Among the SNPs in *CYP3A4*, there are 26 nsSNPs found in its exons 1, and 3-13. These include 44T>C (Lys15Pro), 167G>A (Gly56Asp), 203A>G (Tyr68Cys), 286A>G (Lys96Glu), 352A>G (Ile118Val), 484C>T (Arg162Trp), 485G>A (Arg162Gln), 520G>A (Glu174Lys), 554C>G (Thr185Ser), 559A>T (Thr187Ser), 566T>C (Phe189Ser), 577A>G (Ile193Val), 653C>G (Pro218Arg), 664T>C (Ser222Pro), 754T>G (Ser252Ala), 878T>C (Lys293Pro), 1046C>A (Thr349Asn), 1117C>T (Lys373Phe), 1247C>T (Pro416Lys), 1292T>C (Ile431Thr), 1334T>C (Met445Thr), 1399C>T (Pro467Ser), and 1432A>T (Ser478Cys). We have predicted that Leu15Pro, Gly56Asp, Tyr68Cys, Lys96Glu, Arg30Gln, Arg162Trp, Thr185Ser, Phe189Ser, Pro218Arg, Ser222Pro, Thr363Met, Lys373Phe, Pro416Leu, Met445Thr, Met445Arg, Met445Lys, and Pro467Ser were deleterious using either SIFT or PolyPhen. *In vitro* and *in vivo* studies have shown that Arg130Gln, Thr185Ser, Phe189Ser, Lys293Pro, Thr363Met, Lys373Phe, and Pro416Lys have a functional impact on enzyme activity. However, the nsSNP Lys293Pro was predicted as tolerant by both algorithms

Steroid 11β-hydroxylase deficiency was caused by nonsense mutations in *CYP11B1* including Pro42Ser, Asn133His, Thr318Met, Thr319Met, Arg374Gln, and Arg448His. Amino acid alterations in *CYP11B2*,

such as Val385Ala, Arg181Trp, Thr185Ile, Glu198Asp, Leu461Pro, and Thr498Ala, reduced 18-hydroxylase and abolished 18-oxidase activities, and then elevated ratio of 18-hydroxycorticosterone to aldosterone in serum which is the characteristics of corticosterone methyloxidase II deficiency (Joehrer et al., 1997). By using SIFT or PolyPhen, a number of nsSNPs were predicted to be deleterious. These include Cys10Tyr, Phe42Ser, Pro94Leu, Asn133His, Met160Ile, Leu293Val, Thr318Met, Thr319Met, Ala368Asp, Arg374Gln, Glu383Val, Pro414Ala, Arg448His, and Cys494Phe.

High frequencies of deleterious nsSNPs were shown in *CYP17A1* with 21 missense mutations accounting for adrenal hyperplasia type 5. The degree of loss of enzyme activity varied from 62% to 100% due to different amino acid changes in CYP17A1, for instances, the Pro35Leu mutant retained 38% 17α-hydroxylase activity and 33% 17,20-lyase activity compared to the wild-type (Biason-Lauber et al., 2000). The Arg496His mutant showed 30% 17α-hydroxylase activity and 29% 17,20-lyase activity of the wild-type, while Arg96Trp presented 25% of both enzyme activities. Asn177Asp exhibited 10% of the two enzyme activities; Arg347His and Arg358Gln selectively ablated 17,20-lyase activity, while preserving most 17α-hydroxylase activity (Biason-Lauber et al., 2000; Gupta et al., 2001). Phe417Cys ablated both 17,20-lyase and 17α-hydroxylase activities due to loss of heme-binding and phosphorylation (Biason-Lauber et al., 2000; Gupta et al., 2001). These nsSNPs were predicted to affect protein function by SIFT and PolyPhen except for Arg358Gln.

Mutations in *CYP19A1* (i.e. aromatase) are related to aromatase deficiency with various mutations producing different degrees of enzyme activity loss (Ma et al., 2005). Aromatase is a critical enzyme for estrogen biosynthesis, and aromatase inhibitors are of increasing importance in the management of breast cancer. Four nsSNPs including Trp39Arg, Thr201Met, Arg264Cys, and Met364 have been found in *CYP19A1* (Ma et al., 2005). The Cys264, Thr364, and double variant Arg39Cys264 allozymes

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

demonstrated a significant decrease in the level of enzyme activity and expression after transient expression in COS-1 cells (Ma et al., 2005). A slight decrease in protein level was also found for the Arg39 allozyme, while Met201 showed no significant changes in either activity or protein level when compared with the wild-type enzyme. There was also a 4-fold increase in K_m value for Thr364. The nsSNPs Trp39Arg, Pro207Ser, Glu210Lys, Arg264Cys, Met364Thr, Arg374Cys, Arg434Cys, and Cys436Tyr were predicted to affect protein function by SIFT and/or PolyPhen. Thr201Met was predicted to be tolerant by both programs, which is in agreement with the result from functional assay.

There are at least 118 alleles of *CYP21A2* (http://www.imm.ki.se/CYPalleles) and many of them have functional impact. Of the described missense mutations in *CYP21A2* gene in humans (see Supplementary Table 2), severities of enzymatic activity loss are exhibited in the following order: Gly64Glu (191G>A; *CYP21A2*47*), Ala362Val (1185C>T; *CYP21A2*49*), Gly375Ser (1123G>A; *CYP21A2*72*), and Arg408Cys (1222C>T; *CYP21A2*73*) with complete loss of enzymatic activity (Lajic et al., 2002). Arg356Pro (1067G>C; *CYP21A2*33*), Arg356Gln (1067G>A; *CYP21A2*34*) and Gly291Ser (871G>A; *CYP21A2*23*) showed 0.15, 0.65, and <1% activity of the wild-type, respectively. Arg483Pro (1448G>C; *CYP21A2*28*) and Ile173Asn (515T>A) exhibited 1-2% of activity, whereas Pro30Leu (89C>T; *CYP21A2*8*), Val304Met (910G>A; *CYP21A2*71*), Arg339His (1016G>A; *CYP21A2*24*) and Pro482Ser (1444C>T; *CYP21A2*61*) had 46, 50, 50 and 70% of activity of the wild-type, respectively (Helmberg et al., 1992; Lajic et al., 2002). These reduced enzyme activities obviously correspond to the degrees of disease manifestation in the patients. Most of these nsSNPs (e.g. Gly64Glu, Ile173Asn, and Gly291Ser) were predicted to be damaging to the mutated protein.

DISCUSSION

Predicting the phenotypic consequences of nsSNPs using algorithms *in silico* may provide a greater understanding of genetic differences in susceptibility to disease and drug response. Numerous experiments on the function of nsSNPs have found that genetic mutations in the *CYP* gene family are responsible for interindividual variation in enzyme activity, contribute to metabolic dysfunction and are associated with several important clinically relevant diseases. With regard to lots of nsSNPs having both phenotype information with experimental evidence and prediction results using computational approach, the relationship between prediction consequences of nsSNPs and real phenotype confirmed by experiments is also studied in this study.

A total of 791 validated nsSNPs were obtained from 57 validated *CYP* genes from the NCBI dbSNP and SWISS-Prot databases. Each *CYP* gene had an average of 14.6 nsSNPs. However, only 33% (259/791) nsSNPs in the dataset of validated nsSNPs in *CYP* genes were found to attribute to alteration of enzyme activity and correlate with disease according to published *in viv*o and *in vitro* studies. These confirmed phenotypes of nsSNPs related to alteration of enzyme activity, and then connected to susceptibility to disease such as adrenal hyperplasia, cerebrotendinous xanthomatosis, corticosterone methyloxidase deficiency, primary congenital glaucoma type 3A and vitamin D-dependent rickets, and poor metabolism of drug. In 259 confirmed allelic variants in *CYP*s, about half of allelic variants were distributed in *CYP21A2* (24%), *17A1* (11%), *27B1* (7%), and *2A6* (5%).

A number of haplotypes (combination of SNPs) exist in human CYP family. Haplotypes are considered as better predictors for phenotype than individual SNPs. However, it is too complicated to design a computational method to predict the genotype of haplotype currently, although a few algorithms have

been developed to analyse haplotype frequencies and predict haplotype phases based on individual genetic information (Xu et al., 2002).

Although a number of sophisticated *in silico* approaches have been used to predict the function of nsSNPs on protein structure and activity, the underpinnings for these algorisms are protein sequence alignment (Wang and Moult, 2001), physiochemical differences (Henikoff and Henikoff, 1992), mapping to known protein three-dimensional structures (Sunyaev et al., 2001; Stitziel et al., 2003), and combinations thereof (Liu et al., 2007). Different *in silico* algorithms focus on different aspects of this information, among which the SIFT and PolyPhen algorithms are the main representatives in this field. Significant concordance was observed between the functional consequences of nsSNP predicted by the SIFT and PolyPhen algorithms (Spearman's $\rho = -0.640$; $P \leq 0.001$). Defining that the variants whose positions with normalized probabilities < 0.05 in SIFT and < 1.5 are predicted to be deleterious in PolyPhen are predicted to be deleterious, 38.94 % and 42.73 % of the amino acid substitutions are predicted by SIFT and PolyPhen algorithms respectively to have functional effects on enzymatic activity. They are consistent with results from Xi et al (2004) who reported that 20-50% of the large number of amino acid substitutions observed in DNA repair genes were supposed to impact function. But the ratio of deleterious nsSNPs in certain range of genes is slightly higher than that in normal human genome, which possibly is a characteristic of certain genes. Ramensky et al (2002) reported that 28% of validated nsSNPs in the human genome variation database predicted to affect protein function. Ng and Henikoff (2002) reported that 25% of 3084 nsSNPs from dbSNP would impact protein activity using SIFT. According to statistics of PolyPhen website, there are 33.2% of 76,434 entries in total predicted to be possibly and probably damaging based on dbSNP build 126.

The prediction accuracy for these *in silico* algorithms is also investigated. A number of *in vivo* and *in vitro* experiments have provided directly or indirectly evidence for nsSNPs functional effect on alteration of enzyme activity, metabolic dysfunction or correlation with diseases. Prediction accuracy is analysed based on these evidences. It has been estimated that 63%~75% of amino acid substitutions were predicted correctly by the SIFT and PolyPhen algorithms (Chasman and Adams, 2001; Sunyaev et al., 2001; Ng and Henikoff, 2002). Using the SIFT algorithm, Ng and Henikoff (2002) identified 69% of deleterious nsSNPs (3626/5218) from substitutions annotated to be involved in disease from databases and 63% of deleterious nsSNPs from substitutions in LacI that affect function. Based on probabilistic models developed by the authors themselves, 75% of nsSNPs in a total of 733 amino acid substitutions in LacI that affect function were predicted correctly by Chasman and Adams (2001), whereas 69% of nsSNPs in a total of 1551 allele variants involved in disease from databases that affect function were predicted correctly by Sunyaev et al. (2001). In this study, the consistent results with previous studies were found that both SIFT and PolyPhen were shown to successfully predict the effect of about 70% of allele variants in *CYP* gene family. But more higher prediction accuracy (96%) was evaluated of the algorithms by Xi et al (2004) using the PolyPhen and SIFT programs on the phenotype of APEX1 variants. Although the data for evaluation were obtained from benchmarking studies, there maybe exist certain bias because of small samples of only 26 substitutions and it should be careful to extrapolate this data to other gene family or whole genome.

In addition, Pearson χ^2 test shows that, of nsSNPs predicted as deleterious, the prediction scores by SIFT and PolyPhen algorithms were significantly correlated with the numbers of nsSNPs with known phenotype. This result supports that nsSNPs with less scores from SIFT and more scores from PolyPhen have more probability to have phenotypical effects. Using oligonucleotides for *in vitro* synthesis of mutant DNA, site-directed mutagenesis, providing direct evidence for SNP functions, has been widely used in the study of protein structure-function relationships, gene expression and vector modification

(Domanski and Halpert, 2001). Furthermore, site-directed mutagenesis has widely been employed to explore genotype-phenotype relationship in human CYP superfamily (Domanski and Halpert, 2001). However, the prediction accuracy based on site-directed mutagenesis was approximately 70%, similar to that based on the all benchmark results.

Some amino acids are critical for the action of human CYP enzymes and their changes will lead to functional consequence. For example, *CYP2A6*6* contains a 383G>A mutation leading to an Arg128Gln substitution (Kitagawa et al., 2001). By using PolyPhen algorithm, a change of position 128 would result in disruption of ligand binding site (Table 6), which is consistent with available functional findings by Kitagawa et al. (2001). When expressed in insect cells using a baculovirus system, coumarin 7-hydroxylation was significantly reduced (1/8 of normal) in cell lysate from *CYP2A6*6*-transfected *Sf*9 cells compared with that lysate from *CYP2A6*1*-transfected cells. Although CYP2A6.6 retained about one-half the heme content of CYP2A6.1, the reduced carbon oxide-bound Soret peak was completely lost (Kitagawa et al., 2001), suggesting that the inactivation of CYP2A6.6 is mainly due to disordering of the holoprotein structure rather than a failure of heme incorporation.

The residue 214 is important for the catalytic activity of CYP2C9. There is a natural SNP 641A>T (Gln214Leu; *CYP2C9*28*) occurring in Japanese with a very low frequency (0.002) (Maekawa et al., 2006). Functional characterization of novel *CYP2C9* alleles using a mammalian cell expression system *in vitro* revealed that CYP2C9.28 had 2-fold higher K_m values and 3-fold lower V_{max} values than the wild-type, suggesting an important role of Gln214 for substrate recognition. The PolyPhen algorithm predicted that Gln214Leu in CYP2C9 would lead to a hydrophobicity change at buried site (Table 6). In the structures of CYP2C9 without ligand bound or with bound *S*-warfarin, residues 212–222 in the F–G loop form helices F' and G' while residues 101-106 in the B-C loop form helix B' (Williams et al., 2003). The

F-G loop was not observed in rabbit CYP2C5 (Williams et al., 2000) and bacterial CYPs (Ravichandran et al., 1993). It is likely that the residue change at 214 may lead to altered ligand recognition. Residues Phe69, Phe100, Leu102, Leu208, Leu362, Leu366 and Phe476 form a hydrophobic patch in the active site while Arg105 and Arg108, previously implicated in the formation of the putative anionic-binding site, both point away from the cavity (Williams et al., 2003). In contrast to the putative basic residues, two acidic residues are present in the active site of 2C9: Asp 293 and Glu300. Asp 293 is close to Phe110 and Phe114, and forms a hydrogen bond to the backbone nitrogen of Ile112 and consequently is well ordered, whereas Glu300 points into the active site but exhibits some flexibility in the ligand-free structure of CYP2C9 (PDB:1OG1). Both Gln214 and Asn217 are near Phe476 and may provide potential hydrogen-bonding interactions with ligands.

Both residues 120 and 433 are predicted to be important for ligand binding in CYP2C19 using the PolyPhen algorithm. Changes in these two residues constitute *CYP2C19*5A* (1297C>T; Arg433Trp) and **8* (358T>C; Trp120Arg), respectively. Functional studies have demonstrated that both mutations affect the structure and stability of the protein (Ibeanu et al., 1998a; Ibeanu et al., 1999). Arg433Trp causes greatly reduced enzyme activity and *CYP2C19*8* with the 358T>C SNP encodes a protein with decreased enzyme activity (Ibeanu et al., 1998b). Because the variant alleles (*CYP2C19*5* and**8*) only account for a minor percentage of *CYP2C19* defective alleles, it is unlikely that these alleles will result in clinically significant consequences.

The SNP 389G>A (Arg130Gln) in *CYP3A4* is predicted to disrupt ligand binding site by the PolyPhen algorithm (Table 6). This mutation constitutes a natural occurring allele (*CYP3A4*8*) in humans (Eiselt et al., 2001). The mutated protein CYP3A4.8 showed decreased enzyme activity as shown in *in vitro* functional assays. The reported CYP3A4 structures contained a closed, buried active site connected to

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

bulk solvent via several tunnels. Yano et al. (2004) reported an active site volume of 1386 $Å^3$ while Williams et al. (2004) found a small volume – 520 $Å^3$. The residue 130 of CYP3A4 appears to play a role in ligand binding.

There are several aspects affecting the prediction accuracy. Firstly, part of allele variants may be actually neutral but incorrectly annotated as causing disease. It is easy to get wrong conclusion because these mutations were observed in vitro or in patients or are in linkage disequilibrium with another substitution that can result in alteration of protein function and disease phenotype (Ng and Henikoff, 2002). Secondly, there exist 19% false positive error in SIFT and 9% false positive error in PolyPhen (Ng and Henikoff, 2006). If all of the nsSNPs from dbSNP were functionally neutral, there are 19% or 9% predicted as deleterious nsSNPs. Additionally, damaging mutations in redundant motifs partially accounts for erroneous prediction (Ng and Henikoff, 2002). Programs of the SIFT and PolyPhen that identify SNPs by aligning expressed sequence tags or genomic sequences possibly detect base differences between the functional gene and a pseudogene or another gene of the genome with redundant motifs that actually have lost their function. Thus, programs would erroneously report these differences as SNPs in functional gene. Finally, both SIFT and PolyPhen require homologous sequences and disregard the information on SNP haplotype and non-coding SNPs. The SNP haplotype includes SNP co-occurrences, complex haplotype or other relationships among the SNPs. Some non-coding SNPs occurring in promoter or enhancer regions or splicing junctions can also change protein function. All these above limitations may affect the validity and accuracy of the two algorithms.

In conclusion, the present study has identified 39-43 % of nsSNPs of human *CYP* genes to be deleterious using *in silico* methods. A prediction accuracy analysis found that 70% of nsSNPs were predicted correctly as damaging. Of nsSNPs predicted as deleterious, the prediction scores by SIFT and PolyPhen

were significantly associated with the numbers of nsSNPs with known phenotype confirmed by benchmarking studies including site-directed mutagenesis analysis. These amino acid substitutions are supposed to be the pathogenetic basis of increased susceptivity to certain diseases and altered drug metabolism. The prediction of nsSNPs in human *CYPs* would be useful for further genotype-phenotype studies on the individual difference in drug metabolism and clinical response.

REFERENCES

Biason-Lauber A, Kempken B, Werder E, Forest MG, Einaudi S, Ranke MB, Matsuo N, Brunelli V, Schonle EJ and Zachmann M (2000) 17alpha-hydroxylase/17,20-lyase deficiency as a model to study enzymatic activity regulation: role of phosphorylation. *J Clin Endocrinol Metab* **85:**1226-1231.

Brown PJ, Bedard LL, Reid KR, Petsikas D and Massey TE (2007) Analysis of CYP2A contributions to metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone in human peripheral lung microsomes. *Drug Metab Dispos* **35:**2086-2094.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ and Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22:**231-238.

Chasman D and Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* **307:**683-706.

Cresteil T, Monsarrat B, Dubois J, Sonnier M, Alvinerie P and Gueritte F (2002) Regioselective metabolism of taxoids by human CYP3A4 and 2C8: structure-activity relationship. *Drug Metab Dispos* **30:**438-445.

D'Agostino J, Zhang X, Wu H, Ling G, Wang S, Zhang QY, Liu F and Ding X (2008) Characterization of CYP2A13*2, a variant cytochrome P450 allele previously found to be associated with decreased incidences of lung adenocarcinoma in smokers. *Drug Metab Dispos* **36:**2316-2323.

Dai D, Zeldin DC, Blaisdell JA, Chanas B, Coulter SJ, Ghanayem BI and Goldstein JA (2001) Polymorphisms in human CYP2C8 decrease metabolism of the anticancer drug paclitaxel and arachidonic acid. *Pharmacogenetics* **11:**597-607.

Daigo S, Takahashi Y, Fujieda M, Ariyoshi N, Yamazaki H, Koizumi W, Tanabe S, Saigenji K, Nagayama S, Ikeda K, Nishioka Y and Kamataki T (2002) A novel mutant allele of the CYP2A6 gene (CYP2A6*11) found in a cancer patient who showed poor metabolic phenotype towards tegafur. *Pharmacogenetics* **12:**299-306.

Desta Z, Zhao X, Shin JG and Flockhart DA (2002) Clinical significance of the cytochrome P450 2C19 genetic polymorphism. *Clin Pharmacokinet* **41:**913-958.

Domanski TL and Halpert JR (2001) Analysis of mammalian cytochrome P450 structure and function by site-directed mutagenesis. *Curr Drug Metab* **2:**117-137.

Eiselt R, Domanski TL, Zibat A, Mueller R, Presecan-Siedel E, Hustert E, Zanger UM, Brockmoller J, Klenk HP, Meyer UA, Khan KK, He YA, Halpert JR and Wojnowski L (2001) Identification and functional characterization of eight CYP3A4 protein variants. *Pharmacogenetics* **11:**447-458.

Ferguson RJ, De Morais SM, Benhamou S, Bouchardy C, Blaisdell J, Ibeanu G, Wilkinson GR, Sarich TC, Wright JM, Dayer P and Goldstein JA (1998) A new genetic defect in human CYP2C19: mutation of the initiation codon is responsible for poor metabolism of S-mephenytoin. *J Pharmacol Exp Ther* **284:**356-361.

Ferrer-Costa C, Orozco M and de la Cruz X (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol* **315:**771-786.

Gupta MK, Geller DH and Auchus RJ (2001) Pitfalls in characterizing P450c17 mutations associated with isolated 17,20-lyase deficiency. *J Clin Endocrinol Metab* **86:**4416-4423.

Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:**D514-517.

Helmberg A, Tusie-Luna MT, Tabarelli M, Kofler R and White PC (1992) R339H and P453S: CYP21 mutations associated with nonclassic steroid 21-hydroxylase deficiency that are not apparent gene conversions. *Mol Endocrinol* **6:**1318-1322.

Henikoff S and Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89:**10915-10919.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA and Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* **307:**1072-1079.

Ibeanu GC, Blaisdell J, Ferguson RJ, Ghanayem BI, Brosen K, Benhamou S, Bouchardy C, Wilkinson GR, Dayer P and Goldstein JA (1999) A novel transversion in the intron 5 donor splice junction of CYP2C19 and a sequence polymorphism in exon 3 contribute to the poor metabolizer phenotype for the anticonvulsant drug S-mephenytoin. *J Pharmacol Exp Ther* **290:**635-640.

Ibeanu GC, Blaisdell J, Ghanayem BI, Beyeler C, Benhamou S, Bouchardy C, Wilkinson GR, Dayer P, Daly AK and Goldstein JA (1998a) An additional defective allele, CYP2C19*5, contributes to the S-mephenytoin poor metabolizer phenotype in Caucasians. *Pharmacogenetics* **8:**129-135.

Ibeanu GC, Goldstein JA, Meyer U, Benhamou S, Bouchardy C, Dayer P, Ghanayem BI and Blaisdell J (1998b) Identification of new human CYP2C19 alleles (CYP2C19*6 and CYP2C19*2B) in a Caucasian poor metabolizer of mephenytoin. *J Pharmacol Exp Ther* **286:**1490-1495.

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

Joehrer K, Geley S, Strasser-Wozak EM, Azziz R, Wollmann HA, Schmitt K, Kofler R and White PC (1997) CYP11B1 mutations causing non-classic adrenal hyperplasia due to 11 beta-hydroxylase deficiency. *Hum Mol Genet* **6:**1829-1834.

King BP, Khan TI, Aithal GP, Kamali F and Daly AK (2004) Upstream and coding region CYP2C9 polymorphisms: correlation with warfarin dose and metabolism. *Pharmacogenetics* **14:**813-822.

King LM, Ma J, Srettabunjong S, Graves J, Bradbury JA, Li L, Spiecker M, Liao JK, Mohrenweiser H and Zeldin DC (2002) Cloning of CYP2J2 gene and identification of functional polymorphisms. *Mol Pharmacol* **61:**840-852.

Kirchheiner J, Klein C, Meineke I, Sasse J, Zanger UM, Murdter TE, Roots I and Brockmoller J (2003) Bupropion and 4-OH-bupropion pharmacokinetics in relation to genetic polymorphisms in CYP2B6. *Pharmacogenetics* **13:**619-626.

Kitagawa K, Kunugita N, Kitagawa M and Kawamoto T (2001) CYP2A6*6, a novel polymorphism in cytochrome p450 2A6, has a single amino acid substitution (R128Q) that inactivates enzymatic activity. *J Biol Chem* **276:**17830-17835.

Lajic S, Clauin S, Robins T, Vexiau P, Blanche H, Bellanne-Chantelot C and Wedell A (2002) Novel mutations in CYP21 detected in individuals with hyperandrogenism. *J Clin Endocrinol Metab* **87:**2824-2829.

Lehnerer M, Schulze J, Achterhold K, Lewis DF and Hlavica P (2000) Identification of key residues in rabbit liver microsomal cytochrome P450 2B4: importance in interactions with NADPH-cytochrome P450 reductase. *J Biochem* **127:**163-169.

Liu J, Lewohl JM, Harris RA, Dodd PR and Mayfield RD (2007) Altered gene expression profiles in the frontal cortex of cirrhotic alcoholics. *Alcohol Clin Exp Res* **31:**1460-1466.

Ma CX, Adjei AA, Salavaggione OE, Coronel J, Pelleymounter L, Wang L, Eckloff BW, Schaid D, Wieben ED, Adjei AA and Weinshilboum RM (2005) Human aromatase: gene resequencing and functional genomics. *Cancer Res* **65:**11071-11082.

Maekawa K, Fukushima-Uesaka H, Tohkin M, Hasegawa R, Kajio H, Kuzuya N, Yasuda K, Kawamoto M, Kamatani N, Suzuki K, Yanagawa T, Saito Y and Sawada J (2006) Four novel defective alleles

and comprehensive haplotype analysis of CYP2C9 in Japanese. *Pharmacogenet Genomics* **16:**497-514.

Marez D, Legrand M, Sabbagh N, Guidice JM, Spire C, Lafitte JJ, Meyer UA and 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:**193-202.

Miners JO and Birkett DJ (1998) Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *Br J Clin Pharmacol* **45:**525-538.

Monno S, Ogawa H, Date T, Fujioka M, Miller WL and Kobayashi M (1993) Mutation of histidine 373 to leucine in cytochrome P450c17 causes 17 alpha-hydroxylase deficiency. *J Biol Chem* **268:**25811-25817.

Nadeau JH (2002) Single nucleotide polymorphisms: tackling complexity. *Nature* **420:**517-518.

Nebert DW and Russell DW (2002) Clinical importance of the cytochromes P450. *Lancet* **360:**1155-1162.

Ng PC and Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res* **12:**436-446.

Ng PC and Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:**3812-3814.

Ng PC and Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7:**61-80.

Oscarson M, McLellan RA, Gullsten H, Agundez JA, Benitez J, Rautio A, Raunio H, Pelkonen O and Ingelman-Sundberg M (1999) Identification and characterisation of novel polymorphisms in the CYP2A locus: implications for nicotine metabolism. *FEBS Lett* **460:**321-327.

Ramensky V, Bork P and Sunyaev S (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30:**3894-3900.

Ravichandran KG, Boddupalli SS, Hasermann CA, Peterson JA and Deisenhofer J (1993) Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's. *Science* **261:**731-736.

Rudberg I, Mohebi B, Hermann M, Refsum H and Molden E (2008) Impact of the ultrarapid CYP2C19*17 allele on serum concentration of escitalopram in psychiatric patients. *Clin Pharmacol Ther* **83:**322-327.

Schlicht KE, Michno N, Smith BD, Scott EE and Murphy SE (2007) Functional characterization of CYP2A13 polymorphisms. *Xenobiotica* **37:**1439-1449.

Stenson PD, Ball E, Howells K, Phillips A, Mort M and Cooper DN (2008) Human Gene Mutation Database: towards a comprehensive central mutation database. *J Med Genet* **45:**124-126.

Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S and Liang J (2003) Structural location of disease-associated single-nucleotide polymorphisms. *J Mol Biol* **327:**1021-1030.

Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS and Bork P (2001) Prediction of deleterious human alleles. *Hum Mol Genet* **10:**591-597.

Tajima T, Fujieda K, Nakayama K and Fujii-Kuriyama Y (1993) Molecular analysis of patient and carrier genes with congenital steroid 21-hydroxylase deficiency by using polymerase chain reaction and single strand conformation polymorphism. *J Clin Invest* **92:**2182-2190.

Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N and Krawczak M (2002) Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Hum Mutat* **20:**98-109.

Tomalik-Scharte D, Lazar A, Fuhr U and Kirchheiner J (2008) The clinical role of genetic polymorphisms in drug-metabolizing enzymes. *Pharmacogenomics J* **8:**4-15.

Wang Z and Moult J (2001) SNPs, protein structure, and disease. *Hum Mutat* **17:**263-270.

Williams PA, Cosme J, Sridhar V, Johnson EF and McRee DE (2000) Mammalian microsomal cytochrome P450 monooxygenase: structural adaptations for membrane binding and functional diversity. *Mol Cell* **5:**121-131.

Williams PA, Cosme J, Vinkovic DM, Ward A, Angove HC, Day PJ, Vonrhein C, Tickle IJ and Jhoti H (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **305:**683-686.

Williams PA, Cosme J, Ward A, Angove HC, Matak Vinkovic D and Jhoti H (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* **424:**464-468.

Xi T, Jones IM and Mohrenweiser HW (2004) Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics* **83:**970-979.

Xu CF, Lewis K, Cantone KL, Khan P, Donnelly C, White N, Crocker N, Boyd PR, Zaykin DV and Purvis IJ (2002) Effectiveness of computational methods in haplotype prediction. *Hum Genet* **110:**148-156.

Yano JK, Wester MR, Schoch GA, Griffin KJ, Stout CD and Johnson EF (2004) The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-A resolution. *J Biol Chem* **279:**38091-38094.

Zhang X, Su T, Zhang QY, Gu J, Caggana M, Li H and Ding X (2002) Genetic polymorphisms of the human CYP2A13 gene: identification of single-nucleotide polymorphisms and functional characterization of an Arg257Cys variant. *J Pharmacol Exp Ther* **302:**416-423.

Zhernakova A, van Diemen CC and Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat Rev Genet* **10:**43-55.

Zhou S, Paxton JW, Tingle MD and Kestell P (2000) Identification of the human liver cytochrome P450 isoenzyme responsible for the 6-methylhydroxylation of the novel anticancer drug 5,6-dimethylxanthenone-4-acetic acid. *Drug Metab Dispos* **28:**1449-1456.

Zhou SF (2008) Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. *Curr Drug Metab* **9:**310-322.

Zhou SF, Chan E, Zhou ZW, Xue CC, Lai X and Duan W (2009a) Insights into the structure, function, and regulation of human cytochrome P450 1A2. *Curr Drug Metab* **(in press)**.

Zhou SF, Di YM, Chan E, Du YM, Chow VD, Xue CC, Lai X, Wang JC, Li CG, Tian M and Duan W (2008) Clinical pharmacogenetics and potential application in personalized medicine. *Curr Drug Metab* **9:**738-784.

Zhou SF, Liu YH, Mo SL and Chan E (2009b) Insights into the substrate specificity, inhibitors, and polymorphism and the clinical impact of human cytochrome P450 1A2. *J Clin Pharmacol* **(in press)**.

FOOTNOTES:

a) Acknowledgements

This work was supported by grants from the National Natural Sciences Foundations of People's Republic of China (No. 30671760) and the State Scholarship Fund of China (Dr LL Wang was its holder). We would like to thank Mr. Brian May (RMIT University, Melbourne, Australia) for his assistance in the preparation of this paper.

b) Person to receive reprint request

A/Prof. Shu-Feng Zhou, MD, PhD

Discipline of Chinese Medicine, School of Health Sciences, RMIT University, Bundoora, Victoria 3083, Australia.

Tel: + 61 3 9925 7794; fax: +61 3 9925 7178.

Email: shufeng.zhou@rmit.edu.au.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

LEGENDS FOR FIGURES

Figure 1. Flow chart of the processes for collecting and filtering nsSNPs in human *CYP* genes.

Figure 2. Pie diagrams displaying the prediction accuracy for nsSNPs in human *CYP* genes using (A) SIFT and (B) PolyPhen programs. The phenotypic data are from both *in vivo* and *in vitro* studies.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

TABLES:

Table 1. List of human *CYP* genes and their nsSNPs.

Gene	Chromosomal location	Substrates	Number of amino acids	Number of exons	Number of nsSNPs
Family 1					
CYP1A1	15q22-q24	Xenobiotics	512	7	25
CYP1A2	15q24	Xenobiotics	516	7	31
CYP1B1	2p21	Xenobiotics, sterols	543	3	30
Family 2					
CYP2A6	19q13.2	Xenobiotics	494	9	37
CYP2A7	19q13.2	Unknown	494	9	18
CYP2A13	19q13.2	Xenobiotics	494	9	11
CYP2B6	19q13.2	Xenobiotics	491	9	32
CYP2C8	10q23.33	Xenobiotics	490	9	14
CYP2C9	10q24	Xenobiotics	490	9	28
CYP2C18	10q24	Xenobiotics	490	9	9
CYP2C19	10q24.1-q24.3	Xenobiotics	490	9	31
CYP2D6	22q13.1	Xenobiotics	497	9	52
CYP2E1	10q24.3-qter	Xenobiotics	493	9	19
CYP2F1	19q13.2	Xenobiotics	491	10	7
CYP2J2	1p31.3-p31.2	Fatty acids	502	9	10
CYP2R1	11p15.2	Vitamins	501	2	1
CYP2S1	19q13.1	Unknown	504	9	5
CYP2W1	7p22.3	Unknown	490	9	2
CYP2U1	4q25	Unknown	544	5	0
Family 3					
CYP3A4	7q21.1	Xenobiotics	503	13	32
CYP3A5	7q21.1	Xenobiotics	502	13	15
CYP3A7	7q21-q22.1	Xenobiotics	503	13	5
CYP3A43	7q21.1	Unknown	503	13	5
Family 4					
CYP4A11	1p33	Fatty acids	519	12	7
CYP4A22	1p33	Unknown	519	12	15
CYP4B1	1p34-p12	Fatty acids	511	12	18
CYP4F11	19p13.1	Unknown	524	12	5
CYP4F12	19p13.1	Fatty acids	524	13	11
CYP4F2	19pter-p13.11	Eicosanoids	520	13	11
CYP4F22	19p13.12	Unknown	531	14	2
CYP4F3	19p13.2	Eicosanoids	520	13	6
CYP4F8	19p13.1	Eicosanoids	520	13	3
CYP4V2	4q35.2	Unknown	525	11	15
CYP4X1	1p33	Unknown	509	12	1
CYP4Z1	1p33	Unknown	505	12	0
Family 5					
CYP5A1	7q34-q35	Eicosanoids	534	13	23
Family 7					
CYP7A1	8q11-q12	Sterols	504	6	2
CYP7B1	8q21.3	Sterols	506	6	1
Family 8					
CYP8A1	20q13.13	Eicosanoids	500	10	14

CYP8B1	3p22-p21.3	Sterols	501	1	5
Family 11					
CYP11A1	15q23-q24	Sterols	521	6	10
CYP11B1	8q21	Sterols	503	9	26
CYP11B2	8q21-q22	Sterols	503	9	20
Family 17					
CYP17A1	10q24.3	Sterols	508	8	31
Family 19					
CYP19A1	15q21.1	Sterols	503	10	13
Family 20					
CYP20A1	2q33.2	Unknown	462	13	4
Family 21					
CYP21A2	6p21.3	Sterols	495	10	68
Family 24					
CYP24A1	20q13	Vitamins	514	12	4
Family 26					
CYP26A1	10q23-q24	Vitamins	497	7	3
CYP26B1	2p13.3	Vitamins	512	6	3
CYP26C1	10q23.33	Vitamins	522	4	3
Family 27					
CYP27A1	2q33-qter	Sterols	531	8	15
CYP27B1	12q13.1-q13.3	Vitamins	508	9	22
CYP27C1	2q14.3	Unknown	372	8	1
Family 39					
CYP39A1	6p21.1-p11.2	Sterols	469	12	7
Family 46					
CYP46A1	14q32.1	Sterols	500	15	0
Family 51					
CYP51A1	7q21.2-q21.3	Sterols	509	10	3

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 2. Prediction results of nsSNPs of human *CYP* genes.

Prediction result	SIFT[a]		PolyPhen[b]	
	Number of nsSNPs	%	Number of nsSNPs	%
Deleterious	308	38.94	338	42.73
Tolerated	460	58.15	430	54.36
Not scored	23	2.91	23	2.91
Total	791	100	791	100

[a]See website: SIFT (http://blocks.fhcrc.org/sift/SIFT.html); Positions with normalized probabilities <0.05 are predicted to be deleterious, those ≥0.05 are predicted to be tolerated.
[b]See website: PolyPhen (http://genetics.bwh.harvard.edu/pph/) Positions with normalized probabilities <1.5 are predicted to be tolerated, those ≥1.5 are predicted to be deleterious.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 3. Distribution of deleterious nsSNPs in *CYP* genes predicted by SIFT and/or PolyPhen algorithms.

***CYP* gene**	**Number of deleterious nsSNPs predicted by SIFT**	***CYP* gene**	**Number of deleterious nsSNPs predicted by PolyPhen**	***CYP* gene**	**Number of deleterious nsSNPs predicted by either SIFT or PolyPhen**
CYP21A2	40	*CYP21A2*	40	*CYP21A2*	49
CYP17A1	24	*CYP17A1*	25	*CYP17A1*	26
CYP1A2	18	*CYP1A2*	17	*CYP1A2*	24
CYP1B1	16	*CYP1B1*	17	*CYP2D6*	24
CYP2C9	15	*CYP2C9*	16	*CYP2C9*	19
CYP2D6	15	*CYP2D6*	16	*CYP1B1*	18
CYP27B1	14	*CYP27B1*	16	*CYP3A4*	17
CYP1A1	13	*CYP1A1*	15	*CYP27B1*	16
CYP2C19	13	*CYP3A4*	14	*CYP2C19*	16
CYP3A4	12	*CYP11B1*	14	*CYP1A1*	15
CYP27A1	11	*CYP27A1*	13	*CYP2B6*	15
CYP2A6	11	*CYP2C19*	11	*CYP11B1*	14
CYP11B1	9	*CYP2B6*	11	*CYP2A6*	14
CYP2B6	9	*CYP2A6*	9	*CYP27A1*	13
CYP19A1	7	*CYP19A1*	9	*CYP3A5*	10
CYP3A5	7	*CYP3A5*	8	*CYP19A1*	9
CYP11B2	6	*CYP5A1*	7	*CYP4B1*	9
CYP2A13	6	*CYP11B2*	6	*CYP11B2*	8
CYP4B1	6	*CYP2A13*	6	*CYP5A1*	7
CYP2E1	5	*CYP4B1*	6	*CYP2E1*	7
CYP4F2	5	*CYP4V2*	6	*CYP4F2*	7
CYP2J2	4	*CYP24A1*	6	*CYP2A13*	6
CYP4V2	4	*CYP2E1*	4	*CYP4V2*	6
CYP11A1	3	*CYP2A7*	4	*CYP24A1*	6
CYP2A7	3	*CYP2F1*	4	*CYP2A7*	6
CYP2C18	3	*CYP8A1*	4	*CYP2J2*	5
CYP2C8	3	*CYP4F2*	3	*CYP2F1*	4
CYP2F1	3	*CYP2J2*	3	*CYP8A1*	4
CYP2R1	3	*CYP11A1*	3	*CYP11A1*	3
CYP24A1	2	*CYP2C8*	3	*CYP2C8*	3
CYP3A43	2	*CYP4F12*	3	*CYP4F12*	3
CYP4F12	2	*CYP2C18*	2	*CYP2C18*	3
CYP4F3	2	*CYP2R1*	2	*CYP2R1*	3
CYP4F8	2	*CYP3A43*	2	*CYP4A11*	3
CYP5A1	2	*CYP4A11*	2	*CYP4A22*	3
CYP8A1	2	*CYP4A22*	2	*CYP3A43*	2
CYP20A1	1	*CYP8A1*	2	*CYP8A1*	2
CYP3A7	1	*CYP8B1*	2	*CYP8B1*	2
CYP4A11	1	*CYP4F3*	1	*CYP4F3*	2
CYP4A22	1	*CYP4F8*	1	*CYP4F8*	2
CYP4F11	1	*CYP20A1*	1	*CYP20A1*	2
CYP8A1	1	*CYP39A1*	1	*CYP39A1*	1
		CYP7A1	1	*CYP7A1*	1
				CYP3A7	1
				CYP4F11	1
Total	308		338		411

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 4. Deleterious nsSNPs predicted by both SIFT and PolyPhen algorithms.

Gene Symbol	SNP ID	Amino acid change
CYP1A1	rs35035798	Met66Val
	rs17861094	Ile78Thr
	rs2229150	Arg93Trp
	rs45442501	Arg135Trp
	rs34260157	Arg279Trp
	rs4987133	Ile286Thr
	VAR_016938	Ile448Asn
	rs41279188	Arg464Ser
	VAR_016939	Arg464Cys
	rs36121583	Phe470Val
	rs56240201	Arg477Trp
	rs28399430	Pro492Arg
	rs56343424	Arg511Leu
CYP1A2	rs45565238	Gly73Arg
	rs55802037	Phe125Ile
	rs45540640	Phe205Val
	VAR_020851	Ser211Cys
	rs45468096	Arg281Trp
	rs28399418	Ile314Val
	VAR_025188	Arg377Gln
	VAR_020794	Ile385Phe
	rs28399424	Arg431Trp
	VAR_025189	Arg455His
	rs34151816	Arg457Trp
CYP1B1	-	Met1Thr
	VAR_008350	Trp57Cys
	rs28936700	Gly61Glu
	rs9282671	Tyr81Asn
	rs56339482	Leu107Val
	rs9341248	Ser206Asn
	rs55771538	Gly365Trp
	rs28936414	Arg368His
	rs28936413	Asp374Asn
	rs56305281	Pro379Leu
	rs55989760	Glu387Lys
	rs56010818	Arg390His
	-	Asn423Tyr
	rs56175199	Pro437Leu
	rs28936701	Arg469Trp
CYP2A6	rs61562160	Gly121Arg
	-	Arg128Leu
	rs4986891	Arg128Gln

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

	rs1801272	Leu160His
	rs11575924	Arg257Cys
	rs5031016	Ile471Thr
CYP2A7	rs3869579	Arg311Cys
CYP2A13	VAR_018335	Arg101Gln
	rs8192789	Arg257Cys
	rs3885816	Pro321Leu
	VAR_018356	Val323Leu
	VAR_018338	Phe453Tyr
	VAR_018339	Arg494Cys
CYP2B6	rs8192709	Arg22Cys
	rs36060847	Gly99Glu
	rs12721655	Lys139Glu
	rs3826711	Pro167Ala
	rs28399499	Ile328Thr
CYP2C8	-	Arg186Gly
	rs11572103	Ile269Phe
	VAR_001253	His411Leu
CYP2C9	rs12414460	Arg124Gln
	rs1799853	Arg144Cys
	rs2256871	His251Arg
	rs57505750	Ile327Thr
	rs58368927	Pro337Leu
	rs1057909	Tyr358Cys
	rs28371686	Asp360Glu
	-	Asp397Ala
	rs28371687	Leu413Pro
	VAR_008346	Gly417Asp
	rs59485260	Leu447Phe
	rs9332239	Pro489Ser
CYP2C18	rs59636573	Val330Leu
	rs2281891	Thr385Met
CYP2C19	rs28399504	Met1Val
	rs41291556	Trp120Arg
	VAR_008358	Arg132Gln
	rs57700608	Asn176Ser
	rs6413438	Pro227Leu
	rs58259047	Thr302Arg
	rs56337013	Arg433Trp
	VAR_021275	Arg442Cys

CYP2D6	rs5030862	Gly42Arg
	rs5030865	Gly169Cys
	rs1135830	Ser311Leu
	rs5030867	His324Pro
	VAR_008372	Arg343Gly
	rs28510588	Tyr355Cys
	rs1058172	Arg365His
CYP2E1	rs56864127	Arg126Gln
	rs60719153	Arg126Trp
CYP2F1	rs57670668	Arg98Pro
	rs2287942	Val175Gly
	rs7246981	Pro490Leu
CYP2J2	VAR_014319	Ile192Asn
	rs1056596	Leu378Gln
CYP2S1	rs8192795	Leu230Pro
CYP3A4	rs12721634	Leu15Pro
	rs59418896	Tyr68Cys
	rs3091339	Lys96Glu
	VAR_011600	Arg129Gln
	rs57409622	Arg162Trp
	rs12721627	Thr185Ser
	rs4987161	Phe189Ser
	VAR_011606	Thr362Met
	rs4986909	Pro416Leu
CYP3A5	rs55817950	Arg28Cys
	rs56244447	Leu82Arg
	rs28365083	Thr398Asn
	rs13220949	Arg439Lys
	rs41279854	Phe446Ser
	rs13233803	Arg495Thr
CYP3A43	rs45450092	Met145Ile
	rs45621431	Met275Ile
CYP4B1	rs4646491	Arg340Cys
	rs59694031	Cys369Ser
	rs2297809	Arg375Cys
CYP4F2	rs3093104	Ser7Tyr
CYP4F3	rs28371479	Ile271Thr

CYP4F8	rs2072600	Tyr125Phe
CYP4F12	rs17853419	Phe461Ser
	rs10421387	Gly470Trp
CYP5A1	rs6140	Ile332Thr
	rs13306050	Pro512Leu
CYP8A1	rs5584	Pro500Ser
	VAR_010915	Pro38Leu
	rs11699426	Val69Gly
CYP11A1	rs11544450	Gly15Cys
	rs1130843	Phe274Leu
CYP11B1	VAR_001260	Pro42Ser
CYP11B1	-	Pro94Leu
CYP11B1	rs5292	Leu293Val
CYP11B1	-	Ala368Asp
CYP11B2	rs5315	Val403Glu
CYP19A1	rs17853490	Pro207Ser
CYP21A2	VAR_026060	Pro30Gln
CYP24A1	rs6022990	Met374Thr
	rs6068812	Leu409Ser
CYP27A1	rs2229381	Thr175Met
	rs41272687	Pro384Leu
CYP27B1	rs2229103	Val374Ala

.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 5. Common amino acid change of deleterious nsSNPs in human *CYP* genes predicted by SIFT and PolyPhen algorithms.

Contig Reference	Number	Missense	Number	Common Amino Acid Change	Number
Arg	79	Ala	30	Arg→ Cys	22
Pro	23	Leu	16	Arg→ His	13
Gly	20	His	15	Arg→ Trp	12
Ile	17	Ser	15	Arg→ Gln	10
Leu	15	Trp	15	Pro→ Leu	10
Phe	12	Arg	13	Ile→ Thr	7
Thr	12	Gln	12	Thr→Met	5
Val	9	Pro	11	Leu→ Pro	5
Tyr	7	Glu	10	Gly→ Glu	5
Met	7	Thr	10	Arg→ Trp	4
Ser	6	Asn	7		
His	6	Val	7		
Asn	5	Leu	6		
Trp	4	Tyr	6		
Asp	3	Gly	6		
Glu	3	Met	5		
Cys	3	Ser	5		
Lys	2	Thr	4		
Ala	2	Val	4		

Table 6. Potential effect of amino acid substitution for nsSNPs in human *CYP* genes predicted by the PolyPhen algorithm.

Gene symbol	Protein Access	SNP ID	Allelic variants	PolyPhen predict	PolyPhen score	Substitution effect (according to PolyPhen)	Phenotype	Reference	Experiment	SDM[a]
CYP19A1	NP_000094	VAR_016965	Cys437Tyr	Probably damaging	3.84	Disruption of annotated functional site	Aromatase deficiency; complete loss of activity	(Ito et al., 1993)	*in vivo/in vitro*	yes
CYP2A13	NP_000757	VAR_018335	Arg101Gln	Probably damaging	2.752	Disruption of ligand binding site				
CYP2A6	NP_000753	-	Arg128Leu	Probably damaging	2.493	Disruption of ligand binding site	Decreased activity	(Mwenifumbo et al., 2008)	*In vivo/in vitro*	no
CYP2A6	NP_000753	rs58571639	Arg311Cys	Probably damaging	0.556	Hydrophobicity change at buried site				
CYP2C9	NP_000762	-	Gln214Leu	Probably damaging	1.661	Hydrophobicity change at buried site	Decreased activity	(Maekawa et al., 2006)	*in vivo/in vitro*	yes
CYP2C19	NP_000760	rs56337013	Arg433Trp	Probably damaging	3.627	Disruption of ligand binding site	Loss of activity	(Ibeanu et al., 1998a)	*in vivo*	no
CYP2C19	NP_000760	rs41291556	Trp120Arg	Probably damaging	4.57	Disruption of ligand binding site	Loss of activity	(Ibeanu et al., 1999)	*in vivo/in vitro*	yes
CYP2E1	NP_000764	rs56040284	Ala175Thr	Probably damaging	0.741	Disruption of ligand binding site				
CYP2E1	NP_000764	rs56864127	Arg126Gln	Probably damaging	2.043	Disruption of ligand binding site				
CYP2E1	NP_000764	rs60719153	Arg126Trp	Probably damaging	3.536	Disruption of ligand binding site				
CYP2F1	NP_000765	rs57670668	Arg98Pro	Probably damaging	2.434	Disruption of ligand binding site				
CYP3A4	NP_059488	VAR_011600	Arg130Gln	Probably damaging	2.789	Disruption of ligand binding site	Decreased activity	(Eiselt et al., 2001)	*In vitro*	yes

[a]SDM = Site-directed mutagenesis.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 7. Concordance analysis between the functional consequences of nsSNPs in human *CYP* genes predicted by SIFT and PolyPhen algorithms.

PolyPhen prediction	**Score**	**SIFT prediction**				
		Tolerant	**Borderline**	**Potentially intolerant**	**Intolerant**	**Total**
		1.000~0.201	**0.200~0.101**	**0.100~0.050**	**0.049~0.000**	
Benign	0.000~0.999	218	41	30	33	322
Borderline	1.000~1.249	27	13	10	13	63
Potentially damaging	1.250~1.499	11	7	7	23	48
Possibly damaging	1.500~1.999	27	14	8	51	100
Probably damaging	≥2.000	24	18	12	179	233
Total		307	93	67	299	766
						Spearman's ρ = -0.640; $P \leq 0.001$

Data are analysed by Spearman's rank correlation test.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 8. A summary of nsSNPs of human *CYP* genes with known phenotypes.

Gene Symbol	nsSNP number	Phenotype
CYP21A2	61	CAH; hyperandrogenism
CYP17A1	29	AH5; loss of 17alpha-hydroxylase activity and 17,20-lyase activity
CYP27B1	18	VDDR I; loss of activity
CYP2A6	14	Poor metabolism; decreased activity
CYP2C9	14	Poor tolbutamide metabolizer; decreased activity; increases the K_m value for substrates tested
CYP2B6	13	Decreased expression/activity; increased activity.
CYP1B1	11	GLC3A; peters anomaly
CYP11B1	10	Steroid 11β-hydroxylase deficiency
CYP27A1	9	CTX
CYP2D6	9	Poor metabolism; decreased activity
CYP11B2	8	CMO- I and CMO-II deficiency; loss of activity
CYP2C19	8	Poor metabolism; decreased activity
CYP3A4	8	Decreased activity; lower turnover for testosterone, chlorpyrifos & nifedipine
CYP3A5	8	Decreased activity
CYP4V2	8	BCD
CYP1A2	7	Decreased activity and expression
CYP19A1	6	Aromatase deficiency
CYP2C8	4	Decreased paclitaxel turnover; increased K_m for paclitaxel 6α-hydroxylation
CYP2J2	4	Reduced metabolism of arachidonic acid or linoleic acid
CYP11A1	3	Congenital lipoid adrenal hyperplasia; loss of activity
CYP26A1	2	CAH
CYP2E1	1	Reduced activity
CYP2R1	1	25-hydroxyvitamin D_3 deficiency; complete loss of activity
CYP3A7	1	Increased activity.
CYP4A11	1	Hypertension; decreased activity
Total	259	

Also see Supplementary Table 1.

Abbreviations: AH5, adrenal hyperplasia type 5; BCD, Bietti crystalline corneoretinal dystrophy; CTX, cerebrotendinous xanthomatosis; CAH, congenital adrenal hyperplasia; CMO, corticosterone methyloxidase; GLC3A, primary congenital glaucoma type 3A; VDDR-I, vitamin D-dependent rickets type I.

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 9. Evaluation of predicting accuracy of the SIFT and PolyPhen algorithms on human *CYP* nsSNPs based on *in vivo* and site-directed mutagenesis studies.

Criteria	**SIFT prediction**		**PolyPhen prediction**	
	Number	**%**	**Number**	**%**
Based on *in vivo/in vitro* studies				
Correct prediction	168	68.57	171	69.80
Error prediction	77	31.43	74	30.20
Total	245	100	245	100
Based on site-directed mutagenesis assays				
Correct prediction	109	66.87	112	68.71
Error prediction	54	33.13	51	31.29
Total	163	100	163	100

DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Table 10. Correlation analysis between prediction score for deleterious nsSNPs and number of functional nsSNPs by either SIFT or PolyPhen algorithm confirmed by *in vivo*/*in vitro* experiments.

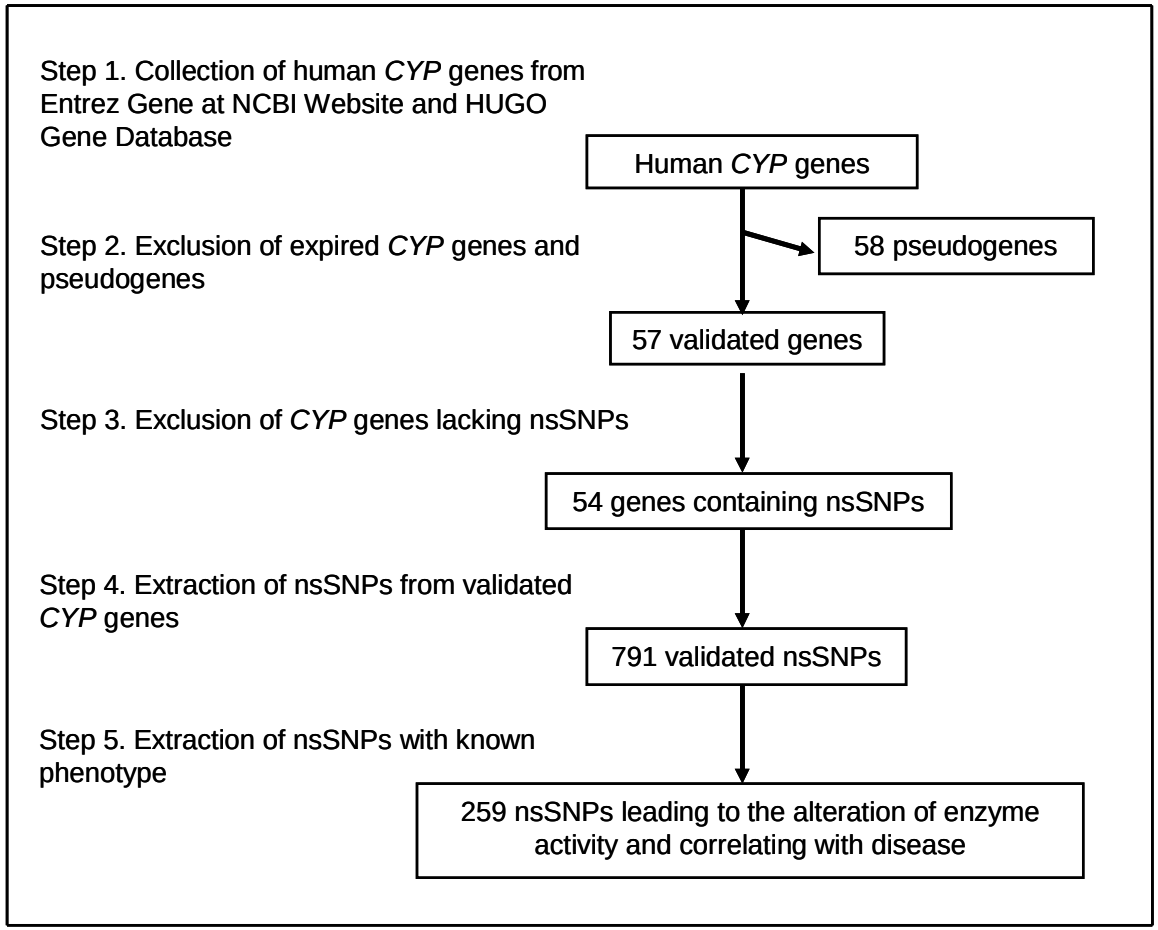
Algorithm	Category	nsSNPs predicted as deleterious	nsSNPs with phenotypical effect conformed by in vivo/in vitro studies		nsSNPs with phenotypical effect conformed by site-directed mutagenesis	
		Number	Number	%[a]	Number	%[b]
SIFT prediction	0	185	116	62.70	74	40.00
	0.01~0.05	123	54	43.90	35	28.46
	Sum	308	170	55.20	109	35.39
	P value		0.001		0.038	
PolyPhen prediction	Probably damaging	238	135	56.72	88	36.97
	Possibly damaging	100	36	36.00	24	24.00
	Sum	338	171	50.59	112	33.14
	P value		0.001		0.000	

[a]%=100 × (number of nsSNPs with phenotypical effect conformed by in vivo/in vitro studies/ number of nsSNPs predicted as deleterious).

[b]%=100 × (number of nsSNPs with phenotypical effect conformed by site-directed mutagenesis / number of nsSNPs predicted as deleterious).

Data are analysed by Pearson's χ^2 test.

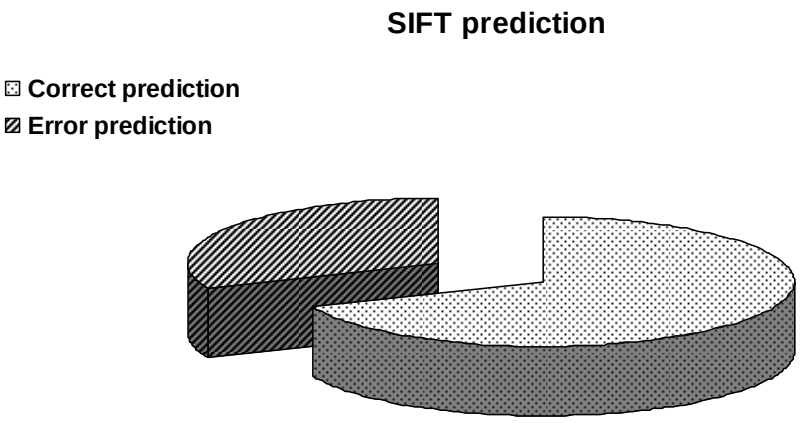
Figure 1



DMD Fast Forward. Published on February 9, 2009 as DOI: 10.1124/dmd.108.026047
This article has not been copyedited and formatted. The final version may differ from this version.

Figure 2

A



B