Research Article

Prediction of Deleterious Non-synonymous Single-Nucleotide Polymorphisms of Human Uridine Diphosphate Glucuronosyltransferase Genes

Yuan Ming Di,¹ Eli Chan,² Ming Qian Wei,³ Jun-Ping Liu,⁴ and Shu-Feng Zhou^{1,5}

Received 24 March 2009; accepted 15 June 2009; published online 2 July 2009

Abstract. UDP glucuronosyltransferases (UGTs) are an important class of Phase II enzymes involved in the metabolism and detoxification of numerous xenobiotics including therapeutic drugs and endogenous compounds (e.g. bilirubin). To date, there are 21 human *UGT* genes identified, and most of them contain single-nucleotide polymorphisms (SNPs). Non-synonymous SNPs (nsSNPs) of the human *UGT* genes may cause absent or reduced enzyme activity and polymorphisms of *UGT* have been found to be closely related to altered drug clearance and/or drug response, hyperbilirubinemia, Gilbert's syndrome, and Crigler-Najjar syndrome. However, it is unlikely to study the functional impact of all identified nsSNPs in humans using laboratory approach due to its giant number. We have investigated the potential for bioinformatics approach for the prediction of phenotype based on known nsSNPs. We have identified a total of 248 nsSNPs from human *UGT* genes. The two algorithms tools, sorting intolerant from tolerant (SIFT) and polymorphism phenotyping (PolyPhen), were used to predict the impact of these nsSNPs on protein function. SIFT classified 35.5% of the *UGT* nsSNPs as "deleterious"; while PolyPhen identified 46.0% of the *UGT* nsSNPs as "potentially damaging" and "damaging". The results from the two algorithms were highly associated. Among 63 functionally characterized nsSNPs in the *UGTs*, 24 showed altered enzyme expression/activities and 45 were associated with disease susceptibility. SIFT and Polyphen had a correct prediction rate of 57.1% and 66.7%, respectively. These findings demonstrate the potential use of bioinformatics techniques to predict genotype-phenotype relationships which may constitute the basis for future functional studies.

KEY WORDS: phenotype; PolyPhen; SIFT; SNP; UGT.

INTRODUCTION

The UDP glucuronosyltransferases (UGTs) represent a superfamily of endoplasmic-reticulum-bound enzymes that catalyze glucuronidation, a process that increases the polarity of xenobiotics and some endogenous compounds to facilitate their excretion via the bile or urine (1). Glucuronidation accounts for ~35% of all drugs metabolized by Phase II drug-metabolizing enzymes and therefore plays an important role in the detoxification and excretion of drugs and/or their

metabolites (1,2). Additionally, UGTs play a critical role in the disposition of several important endogenous substrates, including bilirubin, bile acids, steroids, thyroxine, biogenic amines, and fat-soluble vitamins (3,4).

Up to date, there are 21 human UGTs identified (<http://www.ugtalleles.ulaval.ca>, access date: 11 June 2009). Based on sequence identities, human UGTs comprise UGT1, 2, 3, and 8 families (3,5). The *UGT1* locus in humans has been mapped to chromosome 2q37 (3,6). All *UGT1* genes contain unique first exons, and they are subsequently spliced into common exons 2 through 5, leading to different *N*-terminal halves but identical *C*-terminal halves of the gene products (3). Unlike the *UGT1* family, the human *UGT2* mRNAs are transcribed from separate genes and are divided into UGT2A and 2B subfamilies (4,7). In humans, the genes encoding several UGT2 enzymes form a cluster on chromosome 4 (*UGT2B7-UGT2B4-UGT2B15*; 4). *UGT3* has two members, namely *UGT3A1* and *3A2*, and *UGT8* family has one member.

In the human genome, the most frequent type of DNA variation is single base change, namely single-nucleotide polymorphisms (SNPs) (8,9). Currently, there are a total of 14,708,752 SNPs in human genome identified and deposited to the NCBI dbSNP (<http://www.ncbi.nlm.nih.gov/sites/entrez>, Build 129). A non-synonymous (nsSNPs) or missense variant is a single base change in a coding region that leads to amino acid substitution in the corresponding protein (8,10,11). Many

Electronic supplementary material The online version of this article (doi:10.1208/s12248-009-9126-z) contains supplementary material, which is available to authorized users.

¹ Discipline of Chinese Medicine, School of Health Sciences, RMIT University, Bundoora, Melbourne, Victoria 3083, Australia.

² Department of Pharmacy, Faculty of Science, National University of Singapore, 18 Science Drive 4, Singapore, Singapore.

³ Division of Molecular and Gene Therapies, School of Medical Science, Griffith University, Gold Coast, Queensland 4222, Australia.

⁴ Department of Immunology, Central and Eastern Clinical School, Monash University, Prahran, Melbourne, Victoria 3181, Australia.

⁵ To whom correspondence should be addressed. (e-mail: shufeng.zhou@rmit.edu.au)

ABBREVIATIONS: UGTs, UDP glucuronosyltransferase; nsSNP, non-synonymous single-nucleotide polymorphism.

nsSNPs do not cause any change in the corresponding protein and are regarded as tolerated nsSNPs (9,12). On the other hand, when mutations occur and nsSNPs do cause clinical phenotypic consequences in individuals (e.g. disease or biochemical abnormality), they are termed deleterious nsSNPs (9,12).

In the case of UGTs, deleterious nsSNPs of UGT have shown to alter bilirubin metabolism, drug disposition, and predisposition to diseases in individuals carrying the polymorphism (13,14). Variation in *UGT1A* coding sequences may lead to deficiency of UGT1A1 activity, often resulting in high level of unconjugated bilirubin which is closely related to diseases such as Crigler-Najjar syndrome (CN; 15) and Gilbert's syndrome (16). Nonsense mutations in the *UGT1* gene resulting in a premature termination codon and absent UGT1A1 activity have been reported in CN patients (17,18). Furthermore, some genetic variations in the *UGT1A1* promoter (*28) is associated with increased drug toxicity and hyperbilirubinemia in cancer patients receiving irinotecan-based chemotherapy (19–22), and reports also show increased risk of tranilast-induced hyperbilirubinemia (23). *UGT1A7**3 leading to a low-enzyme activity has been linked to increased risk of the cancer of the colon (24), mouth (25,26), esophagus (26), stomach (26), and the liver (27). A number of *UGT2* genetic variants have been shown to relate to altered drug metabolism and disease risk (1,28,29). For example, the *UGT2B7* promoter variant is associated with significantly reduced glucuronidation of morphine in sickle cell disease and contributes to the variability in hepatic clearance of morphine in these patients (28). A *UGT2B15* variant has been shown to relate to prostate cancer (30).

The identification of deleterious nsSNPs in the human genome is important. This process relates human phenotypes to variation at the DNA level and links genetic and phenotypic differences in individuals; it is believed that nsSNPs and SNPs in regulatory regions together have the highest impact on phenotype (8). The identification and characterization of SNPs of major drug-metabolizing enzymes such as UGTs is crucial for understanding individual differences in drug metabolism, therapeutic efficacy, inherited diseases and also for the predisposition toward diseases such as cancer (13). However, humans have vast numbers of SNPs and presently it is unlikely to investigate their functional impact of these SNPs by experiments. Alternatively, bioinformatics approaches have gained increased use in the prediction of the functional effect of SNPs. Two algorithms sorting intolerant from tolerant (SIFT) and polymorphism phenotyping (PolyPhen) represent powerful bioinformatic tools, they enable high-throughput prediction of the potential impact of nsSNPs and large-scale polymorphism analyses. This process has the prospective to reduce the number of SNPs that may have clinical implications and needs to be evaluated in future clinical studies.

The elementary theory of SIFT have been developed and illustrated by Ng and Henikoff (9,31,32); it is used to predict the effect of an amino acid substitution on protein function according to sequence homology and the physical properties of amino acids. SIFT examines sequence similarity among related genes and does not require the knowledge of protein structural or functional information; therefore, it can be applied to a much larger number of proteins (9). The PolyPhen algorithm is a structure-sequence-based amino acid substitution prediction method; it utilizes the data available in Swiss-Prot (<http://au.expasy.org/sprot/userman.html>) and predicts the possible impact of amino acid substitutions

on the structure and function of a human protein (8). Mapping of an amino acid replacement to a known 3D structure reveals whether the replacement is likely to destroy the hydrophobic core of a protein, electrostatic interactions, interactions with ligands or other important features of a protein (8,33). Several programs including TMHMM algorithm, Coils2 program, SignalIP programs, and PHAT transmembrane-specific matrix scores are also used to predict the possible functional effect of nsSNPs (8,34–37). Multiple alignment-based profile scores provide the major contribution to the prediction therefore reliable prediction can be reasonably achieved even if the proteins has no known 3D structure (8). In this study, we have investigated the potential effect of known human *UGT* nsSNPs on protein function using both SIFT and PolyPhen algorithms.

MATERIALS AND METHODS

Gene Nomenclature and Dataset

The human *UGT* genes examined in this study were named in agreement with the UGT Nomenclature Committee (<http://www.ugtalleles.ulaval.ca/>, access date: 11 June 2009). The data on human UGT genes were collected from Entrez Gene on NCBI Website (<http://www.ncbi.nlm.nih.gov/sites/entrez>, access date: 11 June 2009). Expired and merged gene names were excluded from the study. The majority of the variants included in this analysis were identified during the screening of 21 human *UGT* genes from MutDB and PolyDoms. MutDB integrates genetic variants from Swiss-Prot and dbSNP and links to functional disruption prediction scores. PolyDoms incorporate the results of multiple algorithmic procedures and functional criteria applied to the entire Entrez dbSNP dataset. Information including gene symbol, gene name, mRNA accession number (NM), protein accession number (NP), SNP ID, amino acid residue 1 (wild-type), amino acid position, and amino acid residue 2 (missense) were collected. Most of the data and information describing the genes and variants, including Entrez Gene ID are available at <http://www.mutdb.org> and <http://www.polydoms.cchmc.org/polydoms>. Supplementary variants were identified from Entrez Gene on NCBI (<http://www.ncbi.nlm.nih.gov/sites/entrez>, access date: 11 June 2009) and added to the dataset after cross-examination. The information on the effect of the nsSNPs on enzyme activity and the correlation between the nsSNPs and disease were extracted from *in vivo* and *in vitro* experiments (e.g. site-directed mutagenesis analysis) according to literature.

Prediction of the Phenotype of nsSNPs in Human *UGT* Genes

Identified *UGT* nsSNPs and prediction effect of the variant amino acid substitution on protein function was performed using SIFT (<http://blocks.fhcrc.org/sift/SIFT.html>) and Polyphen (<http://genetics.bwh.harvard.edu/pph/>), listed in Supplementary Table 1.

In this study, SIFT version 3 was used. This algorithm uses a query sequence to search for similar sequences that may share similar function, generates the alignment of the chosen sequences, and predicts the probability score of the impact of an amino acid substitution on protein function

Table I. A Summary of Allelic Variants of Human *UGT* Genes

Gene symbol	No. nsSNPs	No. of nsSNPs scored by SIFT	No. of nsSNPs scored by PolyPhen	No. of nsSNPs scored by both algorithms
<i>UGT1A1</i>	51	51	50	50
<i>UGT1A3</i>	20	20	20	20
<i>UGT1A4</i>	12	12	12	12
<i>UGT1A5</i>	25	25	25	25
<i>UGT1A6</i>	9	9	9	9
<i>UGT1A7</i>	14	14	14	14
<i>UGT1A8</i>	29	29	29	29
<i>UGT1A9</i>	11	11	11	11
<i>UGT1A10</i>	25	25	25	25
<i>UGT2A1</i>	5	5	5	5
<i>UGT2A2</i>	3	3	3	3
<i>UGT2A3</i>	1	1	1	1
<i>UGT2B4</i>	9	9	9	9
<i>UGT2B7</i>	10	10	10	10
<i>UGT2B10</i>	6	6	6	6
<i>UGT2B11</i>	6	6	6	6
<i>UGT2B15</i>	2	2	2	2
<i>UGT2B17</i>	0	0	0	0
<i>UGT2B28</i>	5	5	5	5
<i>UGT3A1</i>	1	1	1	1
<i>UGT3A2</i>	2	2	2	2
<i>UGT8</i>	2	2	2	2
Total	248	248	247	247

effects. SIFT scores ranges from 0 to 1, outcome scores from 0.00–0.05 are elected as intolerant, 0.051–0.10 as potentially tolerant, 0.101–0.20 as borderline, and 0.201–1.00 as tolerant (31,38). Information such as dbSNP ID and GI number of an amino acid substitution can be employed to predict the effect on protein functions in SIFT. The algorithm and instructions for analysis of amino acid substitutions are available at <http://www.blocks.fhcrc.org/sift/SIFT.html>. The data employed in the SIFT analysis were the *UGT* gene sequences available in the NCBI non-redundant database (<http://www.ncbi.nlm.nih.gov>, access date: 11 June 2009), orthologous sequences were used as paralogous sequences confounds predictions (32).

PolyPhen uses empirically derived rules based on previous research in protein structure, interaction, and evolution that automatically predict whether a replacement is likely to be deleterious for the protein on the basis of three-dimensional structure and multiple alignments of homologous sequences (8,33). PolyPhen input is a protein amino acid sequence or the SWALL database (<http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+SWALL>) ID or accession number together with sequence position and two amino acid variants characterizing the polymorphism (8). PolyPhen scores range from $0 \leq 2 \leq X$, outcome scores of 0.00–0.99 are classified as benign, 1.00–1.24 as borderline, 1.25–1.19 as potentially damaging, 1.50–1.99 as possibly damaging, and ≥ 2 as damaging (38). Additional details of the algorithm and instructions for analysis of amino acid substitutions are available at <http://www.bork.embl-heidelberg.de/PolyPhen/>.

Validation of the Prediction Results

nsSNPs with experimental evidences of changing enzyme activity or disease association were regarded as deleterious. The

phenotypic data are from both *in vivo* and *in vitro* studies, in which analysis of site-directed mutagenesis or enzymatic changes often provide direct evidence indicating the functional impact of nsSNPs. Prediction accuracy was analyzed according to these positive findings from these experiments.

As a test for the ability of both SIFT and PolyPhen algorithms to identify substitutions impacting enzymatic activity of *UGT* proteins, scores were obtained and compared for the collected nsSNPs of human *UGT* genes related to loss of enzyme activity and disease based on experimental and clinical studies.

Statistical Analysis

Given that SIFT and PolyPhen employ different approaches and also different datasets as foundations for their analysis, it is important to find the concordance of the two prediction tools on functional consequences of each nsSNP prediction. Concordance analysis of each nsSNP predicted by SIFT and PolyPhen were assessed using Spearman's rank correlation coefficient ρ using SPSS 15. Prediction scores of each nsSNP were plotted on scatter graphs and analyzed using linear trend lines.

RESULTS

Validated nsSNPs of Human *UGT* Genes

Two hundred and forty eight amino acid substitution variants were identified in the systemic screening of 21 human *UGT* genes for the analysis of the potential impact of all nsSNPs in human *UGT* genes. With development and updated data in bioinformatics, some previously reported SNPs in dbSNP have been identified as invalid by later studies due to wrong sequencing and alignment. These incorrect SNPs have

either been terminated or merged into other SNPs. We have cross-examined the databases and removed those invalid SNPs. Number of variants from four *UGT* subfamilies (*UGT1*, *UGT2*, *UGT3*, and *UGT8*) are shown in Table I. Three SNPs were identified in the screening of *UGT2B17* but none were non-synonymous, therefore not included in this study.

Prediction of Functional Effect of nsSNPs of Human *UGT* Genes

SIFT scores for all 248 nsSNPs identified in this study was found, Polyphen classified *UGT1A1* (L15R) as benign but did not provide a prediction score, therefore 247 nsSNPs in this study were scored by both SIFT and PolyPhen, these nsSNPs were used in the statistical concordance test.

As shown in Table II, 88 of 248 or 35.5% of identified *UGT* nsSNPs exhibited SIFT scores of <0.05 and are classified as “Intolerant” variants by SIFT. One hundred fourteen of 248 or 46.0% of identified *UGT* nsSNPs had prediction scores of ≥ 1.5 and are classified as “probably damaging” by PolyPhen. SIFT predicted a few nsSNPs of *UGT* genes to have no deleterious nsSNPs causing functional effects on protein function of *UGT2A3*, of *2B15*, of *2B28*, *3A1*, and 8 (Table III), including Ala496Thr, Asp84Tyr, Lys522Thr, Leu364His, Asn441Ser, Cys120Gly, and Pro225Leu. PolyPhen also predicted some nsSNPs of *UGT* genes to have no deleterious nsSNPs, these included *UGT2A3*, *2B15*, and 8 (Table III). *UGT2A3*, *2B15*, and 8 did not have any deleterious nsSNPs predicted by both SIFT and PolyPhen (Table III), which are Ala496Thr, Asp84Tyr, Lys522Thr, and Ile367Met.

Potential Effect of nsSNPs of *UGTs* on Amino Acid Change

Representative deleterious nsSNPs and the corresponding amino acid substitution of various *UGT* genes are listed in Table IV. As many as 31 *UGT1A1* nsSNPs of 80 *UGT* nsSNPs were predicted as deleterious by both algorithms. For *UGT1A3*, *1A4*, *1A5*, *1A6*, *1A7*, *1A8*, *1A9*, and *1A10*, deleterious nsSNPs predicted by both SIFT and PolyPhen includes Arg45Trp, Ile322Thr, Leu60His, Ser69Tyr, Ile318Thr, Thr202Ala, Val264Glu, and Ile211Thr. Thirteen *UGT* nsSNPs (Gly307Arg, Gly309Arg, Phe396Leu, Asp457Glu, Leu46Pro, Ser299Phe, Ala381Thr, Leu497Pro, Cys155Arg, Pro288Leu, and Lys367Ile) predicted deleterious by both SIFT and PolyPhen algorithms

were from the *UGT2* family, including *UGT2A1*, *2A2*, *2B4*, *2B7*, *2B10*, and *2B11*. For *UGT3A2*, Ala343Thr was predicted as deleterious, respectively, by both SIFT and PolyPhen algorithms.

Concordance Analysis of Predicted Results by SIFT and PolyPhen

Table V shows the concordance analysis between the functional consequences for 247 nsSNPs predicted by SIFT and PolyPhen. Raw scores rather than arbitrarily defined categories were used for the correlation analysis. The row percentage of PolyPhen scores that falls into a SIFT category was calculated. Remarkably, 85.7% of “benign” predictions from PolyPhen falls into SIFT “tolerant” predictions and 84.2% of “probably damaging” predictions from PolyPhen falls into SIFT “intolerant” predictions. Scatter graphs plotted using 247 prediction scores from SIFT and PolyPhen showed negative correlation (Fig. 1). Spearman’s rank correlation coefficient $\rho = -0.709$ ($P \leq 0.01$) illustrate the significant concordance between the prediction scores from SIFT and PolyPhen algorithms.

Validation of the Prediction Results

The confirmed phenotype of nsSNPs manifests as alteration of enzyme activity and susceptibility to diseases such as CN I, CN II and Gilbert’s syndrome. Up to date, a total of 63 nsSNPs of human *UGT* genes are identified to demonstrate a relation to decreased activity, loss of enzyme activity, or susceptibility to diseases based on experimental and clinical studies. Using positive findings from the experiments, if the variants were predicted to be deleterious, it is considered a correct prediction. An incorrect/error prediction was considered when such nsSNPs were predicted as tolerant.

The confirmed variants were collected from results derived from site-directed mutagenesis studies of the enzyme using biochemical characterization (18,39–47) or clinical data from family-based and association studies (48–56). The biochemical/*in vitro* and *in vivo* *UGT* variants and the predictions for their functional impact scores by SIFT and PolyPhen are displayed in Table VI. If a confirmed *UGT* nsSNP was predicted as “intolerant” by SIFT and “probably damaging” by PolyPhen, then it is termed “true positive finding”. If a confirmed *UGT* nsSNP was predicted as “tolerated” by SIFT and “benign” by PolyPhen then it is termed “false negative finding”. According to these criteria, approximately 57% and 67% of characterized *UGT* nsSNPs were correctly predicted as deleterious and are “true positive predictions” by SIFT and PolyPhen, respectively (Table VII). SIFT predicted 27 confirmed *UGT* alleles as “tolerated” and Polyphen predicted 21 confirmed *UGT* alleles as “benign”, based on this finding, false negative error for SIFT and PolyPhen was 33% and 43%, respectively (Table VII). Ninety-seven percent of confirmed allelic variants of human *UGTs* were distributed in the *UGT1A* family, with two additional variants from the *UGT2B* family. Even more interesting is *UGT1A1* alone contributes 74.6% of the confirmed allelic variants of human *UGTs*, this correlates with the fact that *UGT1A1* was the top *UGT* gene with most frequent deleterious nsSNPs predicted by both SIFT and PolyPhen.

Based on the results of *in vivo* and *in vitro* research, among 63 functionally characterized nsSNPs in the *UGTs*, there were 24 showing altered enzyme expression/activities

Table II. Prediction Results of nsSNPs of Human *UGT* Genes

Prediction result	SIFT ^a		Polyphen ^b	
	Number of nsSNPs	Percent	Number of nsSNPs	Percent
Tolerated	160	35.5	114	46
Deleterious	88	64.5	133	53.6
Not scored	0		1	0.4
Total	248	100	248	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 deleterious, those <1.5 are predicted to be tolerated

Table III. Distribution of Deleterious nsSNPs in Human *UGT* Genes Predicted by SIFT and PolyPhen

<i>UGT</i> gene	Number of deleterious nsSNPs predicted by SIFT	<i>UGT</i> gene	Number of deleterious nsSNPs predicted by PolyPhen	<i>UGT</i> gene	Number of deleterious nsSNPs predicted by SIFT or PolyPhen	Number of deleterious nsSNPs predicted by SIFT and PolyPhen
<i>UGT1A1</i>	32	<i>UGT1A1</i>	38	<i>UGT1A1</i>	38	32
<i>UGT1A3</i>	8	<i>UGT1A3</i>	10	<i>UGT1A3</i>	11	7
<i>UGT1A4</i>	3	<i>UGT1A4</i>	7	<i>UGT1A4</i>	7	3
<i>UGT1A5</i>	5	<i>UGT1A5</i>	6	<i>UGT1A5</i>	7	4
<i>UGT1A6</i>	3	<i>UGT1A6</i>	4	<i>UGT1A6</i>	5	4
<i>UGT1A7</i>	2	<i>UGT1A7</i>	3	<i>UGT1A7</i>	3	2
<i>UGT1A8</i>	8	<i>UGT1A8</i>	10	<i>UGT1A8</i>	10	8
<i>UGT1A9</i>	6	<i>UGT1A9</i>	7	<i>UGT1A9</i>	7	6
<i>UGT1A10</i>	7	<i>UGT1A10</i>	8	<i>UGT1A10</i>	9	6
<i>UGT2A1</i>	1	<i>UGT2A1</i>	1	<i>UGT2A1</i>	1	1
<i>UGT2A2</i>	1	<i>UGT2A2</i>	1	<i>UGT2A2</i>	1	1
<i>UGT2A3</i>	0	<i>UGT2A3</i>	0	<i>UGT2A3</i>	0	0
<i>UGT2B4</i>	4	<i>UGT2B4</i>	3	<i>UGT2B4</i>	4	3
<i>UGT2B7</i>	2	<i>UGT2B7</i>	3	<i>UGT2B7</i>	3	2
<i>UGT2B10</i>	2	<i>UGT2B10</i>	4	<i>UGT2B10</i>	4	2
<i>UGT2B11</i>	3	<i>UGT2B11</i>	3	<i>UGT2B11</i>	3	3
<i>UGT2B15</i>	0	<i>UGT2B15</i>	0	<i>UGT2B15</i>	0	0
<i>UGT2B28</i>	0	<i>UGT2B28</i>	2	<i>UGT2B28</i>	2	0
<i>UGT3A1</i>	0	<i>UGT3A1</i>	1	<i>UGT3A1</i>	1	0
<i>UGT3A2</i>	1	<i>UGT3A2</i>	2	<i>UGT3A2</i>	2	1
<i>UGT8</i>	0	<i>UGT8</i>	0	<i>UGT8</i>	0	0
Total	88		113		118	85

and 45 associated with disease susceptibility. In agreement with earlier studies, reduced levels of UGT enzyme activities were observed in CN II and Gilbert's syndrome, whilst absence of UGT enzyme activities were observed only in CN I patients. Additional deleterious nsSNPs of human *UGT* were predicted by SIFT and PolyPhen but the phenotypic prediction of these nsSNPs has not yet been confirmed using experiments.

More than two-thirds of confirmed variants were *UGT1A1* variants (Table VI). In the confirmed variants of *UGT1A1*, the associated phenotypes were CN I, CN II (15), and Gilbert's syndrome (16) which are all closely related to bilirubin levels. To date, 47 mutant *UGT1A1* alleles have been identified (Table VI; 17,18,39,42–51,53–59). Many of these SNPs were predicted to have phenotypical effects by the algorithms and the correct prediction rates were 68% and 81% by SIFT and PolyPhen, respectively.

A number of the confirmed allelic variants of *UGT1A1* associated with Gilbert's syndrome, including *UGT1A1**27 (Pro229Gln), *29 (Arg367Gly), *62 (Phe83Leu), *69 (Ile159Thr), *70 (Ala321Gly), *72 (Asp359Asn), and *73 (Pro364Leu). Both SIFT and PolyPhen correctly predicted the phenotype of Pro364Leu. In addition, *UGT1A1**7 (Tyr486Asp) and *9 (Gln331Arg) closely associated with Crigler-Najjar II syndrome (3,18,45,59), were predicted as probably damaging with high PolyPhen scores. Furthermore, another *UGT* allelic variant *UGT1A1**11 (Gly308Glu) identified in Crigler-Najjar I syndrome (3,43), was also correctly predicted by both SIFT and PolyPhen algorithms.

Only two variant alleles of the human *UGT2* mRNAs have been confirmed in the *UGT2* gene family, which are Asp67Tyr and Asp85Tyr (60,61). Hyperbilirubinemia can be caused by a number of reasons including liver disease and hemolysis (62). Rate of production of bilirubin also needs to be

considered when bilirubin level is abnormal, e.g. in inherited hemolytic syndromes such as glucose 6-phosphate dehydrogenase deficiency and sickle cell anemia. *UGT2B7* promoter variant 840G>A is associated with significantly reduced glucuronidation of morphine in sickle cell disease and contributes to the variability in hepatic clearance of morphine in these patients (28). *UGT2B15* variant (Asp85Tyr) have been shown to relate to prostate cancer risk (30,63).

DISCUSSION

Humans have vast numbers of SNPs and presently there are more than a million SNPs in dbSNP that can be screened for association with diseases. Prediction tool SIFT was tested against unbiased experimental datasets in which mutagenesis was performed throughout the entire protein, and both wild-type and negative phenotypes were assayed, only three datasets fit the criteria and the scarcity of unbiased data sets indicates how difficult characterization of mutant proteins on a large scale can be (31). Therefore, determining disease causing SNPs via site-directed mutagenesis experiments and gene knockout/knockin experiments is complex, lengthy, and unrealistic taking into consideration the mass amount of SNPs (64). Using bioinformatic tools to predict nsSNPs most likely to be damaging, this process acts as a filter and reduces the number of SNPs required to be screened for association with disease to those that most likely alter gene function (8,9). By predicting nsSNPs of *UGT*, we are able to distinguish mutations that are unlikely to affect protein function (9), reducing the number of nsSNPs for experimentation thus saving time and resources. Based on prediction results of deleterious nsSNPs, on a much smaller scale we then via experiments and clinical studies are able to screen for polymorphisms of *UGT* that may potentially cause disease and increase drug toxicity. Moreover, the

Table IV. Deleterious nsSNPs Predicted by Both SIFT and PolyPhen

Gene symbol	SNP ID	Amino acid change	Gene symbol	SNP ID	Amino acid change
<i>UGT1A1</i>	VAR_026134	Pro34Gln	<i>UGT1A5</i>	rs17851756	Ile322Thr
<i>UGT1A1</i>	VAR_026135	His39Asp	<i>UGT1A5</i>	rs28934877	Asn400Asp
<i>UGT1A1</i>	–	Gln185Pro	<i>UGT1A6</i>	rs1042708	Ser69Tyr
<i>UGT1A1</i>	VAR_007698	Arg209Trp	<i>UGT1A6</i>	rs17851756	Ile320Thr
<i>UGT1A1</i>	VAR_007699	Gly276Arg	<i>UGT1A6</i>	rs28934877	Asn398Asp
<i>UGT1A1</i>	VAR_026138	Glu291Val	<i>UGT1A6</i>	rs34993780	Tyr485Asp
<i>UGT1A1</i>	VAR_026139	Ile294Thr	<i>UGT1A6</i>	rs1042709	Ala509Pro
<i>UGT1A1</i>	VAR_007701	Gly308Glu	<i>UGT1A7</i>	rs17851756	Ile318Thr
<i>UGT1A1</i>	rs17851756	Ile322Thr	<i>UGT1A7</i>	rs28934877	Asn396Asp
<i>UGT1A1</i>	VAR_007702	Gln331Arg	<i>UGT1A8</i>	rs45504099	His53Asn
<i>UGT1A1</i>	VAR_026140	Arg336Leu	<i>UGT1A8</i>	VAR_015546	Thr202Ala
<i>UGT1A1</i>	VAR_026142	Arg336Trp	<i>UGT1A8</i>	rs17863762	Cys276Tyr
<i>UGT1A1</i>	VAR_026143	Trp354Arg	<i>UGT1A8</i>	rs17851756	Ile318Thr
<i>UGT1A1</i>	VAR_007703	Gln357Arg	<i>UGT1A8</i>	rs34946978	Pro361Leu
<i>UGT1A1</i>	–	Pro364Leu	<i>UGT1A8</i>	rs55750087	Arg364Gly
<i>UGT1A1</i>	VAR_012283	Arg367Gly	<i>UGT1A8</i>	rs28934877	Asn396Asp
<i>UGT1A1</i>	VAR_007704	Ala368Thr	<i>UGT1A8</i>	rs34993780	Tyr483Asp
<i>UGT1A1</i>	VAR_007705	Ser375Phe	<i>UGT1A9</i>	rs4663870	Val264Glu
<i>UGT1A1</i>	VAR_026144	His376Arg	<i>UGT1A9</i>	rs17851756	Ile318Thr
<i>UGT1A1</i>	VAR_026145	Gly377Val	<i>UGT1A9</i>	rs34946978	Pro361Leu
<i>UGT1A1</i>	VAR_007706	Ser381Arg	<i>UGT1A9</i>	rs55750087	Arg364Gly
<i>UGT1A1</i>	VAR_026146	Pro387Ser	<i>UGT1A9</i>	rs28934877	Asn396Asp
<i>UGT1A1</i>	VAR_026147	Gly395Val	<i>UGT1A9</i>	rs34993780	Tyr483Asp
<i>UGT1A1</i>	VAR_019412	Asn400Asp	<i>UGT1A10</i>	–	Ile211Thr
<i>UGT1A1</i>	VAR_007707	Ala401Pro	<i>UGT1A10</i>	rs17851756	Ile318Thr
<i>UGT1A1</i>	VAR_026148	Arg403Cys	<i>UGT1A10</i>	rs34946978	Pro361Leu
<i>UGT1A1</i>	VAR_026149	Trp461Arg	<i>UGT1A10</i>	rs55750087	Arg364Gly
<i>UGT1A1</i>	–	Glu463Ala	<i>UGT1A10</i>	rs28934877	Asn396Asp
<i>UGT1A1</i>	VAR_026150	Ala478Asp	<i>UGT1A10</i>	rs34993780	Tyr483Asp
<i>UGT1A1</i>	VAR_007709	Tyr486Asp	<i>UGT2A1</i>	rs4148301	Gly307Arg
<i>UGT1A1</i>	–	Leu496Asn	<i>UGT2A2</i>	rs4148301	Gly309Arg
<i>UGT1A3</i>	rs45625338	Arg45Trp	<i>UGT2B4</i>	VAR_011329	Phe396Leu
<i>UGT1A3</i>	rs45595237	Arg49Trp	<i>UGT2B4</i>	rs13119049	Asp457Glu
<i>UGT1A3</i>	rs17851756	Ile322Thr	<i>UGT2B4</i>	rs41298245	Cys511Arg
<i>UGT1A3</i>	rs34946978	Pro365Leu	<i>UGT2B7</i>	rs61361928	Leu46Pro
<i>UGT1A3</i>	rs55750087	Arg368Gly	<i>UGT2B7</i>	rs34620993	Ser299Phe
<i>UGT1A3</i>	rs28934877	Asn400Asp	<i>UGT2B10</i>	rs4095564	Ala381Thr
<i>UGT1A3</i>	rs34993780	Tyr487Asp	<i>UGT2B10</i>	rs3208727	Leu497Pro
<i>UGT1A4</i>	rs17851756	Ile322Thr	<i>UGT2B11</i>	rs7697037	Cys155Arg
<i>UGT1A4</i>	rs28934877	Asn400Asp	<i>UGT2B11</i>	rs3890590	Pro288Leu
<i>UGT1A4</i>	VAR_009507	Tyr487Asp	<i>UGT2B11</i>	rs11936018	Lys367Ile
<i>UGT1A5</i>	rs28898602	Leu60His	<i>UGT3A2</i>	rs2591714	Ala343Thr
<i>UGT1A5</i>	rs41270755	Arg49Trp			

detection of individual single-nucleotide polymorphisms that alter enzyme function can be a useful tool for the identification of disease risk or to personalize drug regimens (65,66).

A total number of 248 nsSNPs were identified through the screening of 21 human *UGT* genes from NCBI dbSNP, Mutdb, and Polydoms. However, according to published results from *in vivo* and *in vitro* studies, only about a quarter of nsSNPs in the dataset of validated human *UGT* genes were found to attribute to alteration of enzyme activity and correlation with diseases. These confirmed phenotypes of nsSNPs were related to reduced or absence of enzyme activity, and correlated to susceptibility to diseases such as CN, Gilbert's syndrome and hyperbilirubinemia related to certain therapeutic drugs.

There are 49 non-synonymous SNPs in exons 1–5 of *UGT1A1* (Supplementary Table 1). To date, 47 mutant *UGT1A1* alleles have been identified (17,18,39,42–51,53–59). Deleterious nsSNPs of *UGT1A1* such as Ser375Phe, Gly308-

Glu, Ala291Val, and His39Asp which are phenotypically presented as CN I were all correctly predicted by both SIFT and PolyPhen. The Tyr486Asp (*UGT1A1**7) mutation in exon 1 and is the most abundant mutation in CN II in Japanese patients (67), was correctly predicted as deleterious nsSNP by both SIFT and PolyPhen. Variant *UGT1A1**27 (Pro229Gln) identified in CN II with reduced UGT enzyme activities *in vitro* and *in vivo* was not predicted to be deleterious by SIFT and PolyPhen. In seven confirmed alleles associated with Gilbert's syndrome (*UGT1A1**27, *29, *62, *69, *70, *72, and *73), only Arg367Gly and Pro364Leu were correctly predicted as deleterious nsSNPs of *UGT1A1*. Other examples of incorrect predictions includes *UGT1A1**6 (Gly71Arg) and *UGT1A1**15 (Cys177Arg). Overall, SIFT and PolyPhen predicted *UGT1A1* SNPs to have phenotypical effects with correct prediction rates of 68% and 81%, respectively.

Table V. Concordance Analysis Between the Functional Consequences of each nsSNP Predicted by SIFT and Polyphen

		SIFT prediction				
		Tolerant	Borderline	Potentially intolerant	Intolerant	Total
		1.000~0.201	0.200~0.101	0.100~0.050	0.049~0.000	
Polyphen prediction						
Benign	0.00~0.99	84 (85.7%)	5 (5.1%)	6 (6.1%)	3 (3.1%)	98
Borderline	1.00~1.24	17 (81%)	2 (9.5%)	2 (9.5%)	0	21
Potentially damaging	1.25~1.49	7 (53.8%)	3 (23.1%)	2 (15.4%)	1 (7.7%)	13
Possibly damaging	1.50~1.99	8 (24.2%)	8 (24.2%)	2 (6.1%)	15 (45.5%)	33
Probably damaging	≥2.00	6 (7.3%)	5 (6.1%)	2 (2.4%)	69 (84.2%)	82
Total		122	23	14	88	247

SIFT and PolyPhen are *in silico* algorithm tools that use protein sequence alignment, physicochemical differences, mapping to know protein 3-D structures to predict the functional impact of nsSNPs on protein structure and activities (8,9,31,33). Even though SIFT and PolyPhen employ different approaches and types of reference data for their predictions and different scales for scoring, there was significant concordance observed on functional consequences of each nsSNP prediction on human *UGTs* (Spearman's $\rho =$

-0.709 , $P \leq 0.01$). SIFT and PolyPhen can discriminate diseased variants from neutral variants, results from this study are consistent with reports by Ng and Henikoff (9) who predicted 757 of 3,084 or 25% nsSNPs from dbSNP to be damaging and having protein activity impacts. Ramensky *et al.* (8) correctly predicted 27.6% of nsSNPs to affect protein function in the human genome variation database. Xi *et al.* (38) reported 30–50% of over 500 amino acid substitution variants identified in DNA repair genes were predicted to

Fig. 1. Scatter graph displaying the negative correlation between prediction scores for nsSNPs in human *UGT* genes using SIFT and PolyPhen programs **a** and **b**

Table VI. Confirmed Allelic Variants of Human *UGT* Genes using *In Vivo* and *In Vitro* Studies

Allele naming	Residue 1	Position	Residue 2	SNP Id	Phenotype	Enzyme activity		Site directed mutagenesis	SIFT prediction score	PolyPhen prediction score	Reference
						<i>In vitro</i>	<i>In vivo</i>				
<i>UGT1A1</i> *3	Ser	375	Phe	VAR_007705	CN1	Inactive	Inactive	Yes	0.02	1.949	(18,43)
<i>UGT1A1</i> *6	Gly	71	Arg	rs4148323		Reduced	Reduced	Yes	0.1	1.285	(45,57)
<i>UGT1A1</i> *7	Tyr	486	Asp	VAR_007709	CN2	Reduced	Reduced	Yes	0	3.39	(45,57,59)
<i>UGT1A1</i> *8	Arg	209	Trp	VAR_007698	CN2	4.40%	Reduced	Yes	0	3.408	(58,70)
<i>UGT1A1</i> *9	Gln	331	Arg	VAR_007702	CN2	Reduced	Reduced	Yes	0	2.814	(17)
<i>UGT1A1</i> *11	Gly	308	Glu	VAR_007701	CN1	Inactive	Absent	Yes	0	2.946	(43)
<i>UGT1A1</i> *12	Leu	175	Gln	VAR_019411	CN2	38.40%	Reduced	Yes	0.1	1.547	(46)
<i>UGT1A1</i> *14	Gly	276	Arg	VAR_007699	CN1	Inactive	Inactive	Yes	0	2.6	(46)
<i>UGT1A1</i> *15	Cys	177	Arg	VAR_007697	CN1	Inactive	Inactive	Yes	0.14	2.739	(46)
<i>UGT1A1</i> *16	Gln	357	Arg	VAR_007703	CN1	Absent	Absent	No	0.01	2.105	(44)
<i>UGT1A1</i> *17	Ser	381	Arg	VAR_007706	CN1	Absent	Absent	No	0.01	1.507	(44)
<i>UGT1A1</i> *18	Ala	401	Pro	VAR_007707	CN1	Absent	Absent	No	0.02	1.92	(44)
<i>UGT1A1</i> *20	Ala	368	Thr	VAR_007704	CN1	Absent	Absent	No	0.01	1.957	(44)
<i>UGT1A1</i> *22	Ala	291	Val	VAR_026138	CN1	Absent	Absent	No	0	2.766	(44)
<i>UGT1A1</i> *23	Lys	428	Glu	VAR_007708	CN1	Absent	Absent	No	0.09	1.124	(44)
<i>UGT1A1</i> *27	Pro	229	Gln	VAR_009505	Gilbert	Reduced	Reduced	No	0.3	0.759	(39)
<i>UGT1A1</i> *29	Arg	367	Gly	VAR_012283	Gilbert	Reduced	Reduced	No	0	2.94	(39)
<i>UGT1A1</i> *30	Leu	15	Arg	VAR_019410	CN2	Reduced	Reduced	Yes	0.16	No score	(47)
<i>UGT1A1</i> *31	Pro	387	Arg	VAR_026146	CN1	Absent	Absent	Yes	0	3.158	(41)
<i>UGT1A1</i> *32	Arg	336	Trp	VAR_026142	CN1	0–10%	Absent	Yes	0	3.235	(40)
<i>UGT1A1</i> *33	Ile	294	Thr	VAR_026139	CN2	40–55%		Yes	0	1.839	(40)
<i>UGT1A1</i> *34	Met	310	Val	–	CN2	26–51%		Yes	0.52	1.956	(42)
<i>UGT1A1</i> *351	Ile	431	Thr	–	CN2	61–81%		Yes	0.26	1.42	(42)
<i>UGT1A1</i> *38	Asn	400	Asp	VAR_019412	CN2			No	0.01	2.623	(52)
<i>UGT1A1</i> *42	Glu	463	Ala	–	CN2			No	0	2.702	(48)
<i>UGT1A1</i> *44	His	39	Asp	VAR_026135	CN1			Yes	0.01	3.673	(51)
<i>UGT1A1</i> *51	His	376	Arg	VAR_026144	CN2			Yes	0.01	2.669	(51)
<i>UGT1A1</i> *52	Gly	377	Val	VAR_026145	CN2			Yes	0	2.847	(51)
<i>UGT1A1</i> *55	Leu	496	Asn	–	CN1			Yes	0.01	2.375	(51)
<i>UGT1A1</i> *62	Phe	83	Leu	VAR_026136	Gilbert			Yes	0.58	0.863	(56)
<i>UGT1A1</i> *63	Pro	364	Leu	–	–	Reduced	–	No	0.01	3.219	(50)
<i>UGT1A1</i> *69	Ile	159	Thr	–	Gilbert	–	–	Yes	0.2	1.46	(49)
<i>UGT1A1</i> *70	Ala	321	Gly	–	Gilbert	–	–	Yes	0.37	1.051	(49)
<i>UGT1A1</i> *72	Asp	359	Asn	–	Gilbert	–	–	Yes	0.05	2.45	(49)
<i>UGT1A1</i> *73	Pro	364	Leu	–	Gilbert	–	–	Yes	0.01	3.219	(49)
<i>UGT1A1</i> *94	Trp	461	Arg	VAR_026149	CN1	–	No detectable activity	Yes	0	4.657	(53)
<i>UGT1A1</i> *95	Arg	336	Gln	VAR_026141	CN1	–	–	Yes	0.11	1.445	(55)
<i>UGT1A1</i> *96	Arg	336	Leu	VAR_026140	CN1	–	–	Yes	0	3.03	(55)
<i>UGT1A1</i> *97	Gly	395	Val	VAR_026147	CN1	–	–	Yes	0.01	2.292	(55)
<i>UGT1A1</i> *98	Pro	387	Ser	VAR_026146	CN1	–	–	Yes	0	3.158	(55)
<i>UGT1A1</i> *101	Trp	354	Arg	VAR_026143	CN2	–	–	Yes	0	4.753	(55)
<i>UGT1A1</i> *102	Pro	34	Gln	VAR_026134	CN2	–	–	Yes	0	2.803	(55)
<i>UGT1A1</i> *103	Arg	403	Cys	VAR_026148	CN2	–	–	Yes	0	2.92	(55)
<i>UGT1A1</i> *106	Ala	478	Asp	VAR_026150	CN2	–	–	Yes	0	2.149	(55)
<i>UGT1A1</i> *107	Trp	40	Arg	–	CN2	–	–	Yes	0.14	4.05	(54)
<i>UGT1A1</i> *109	Gln	18	Pro	–	CN2	–	–	Yes	0.03	1.535	(54)

<i>UGT1A1*113</i>	Gly	493	Arg	-	CN2	-	-	No	0.23	2.086	(71)
<i>UGT1A3*4a</i>	Arg	45	Trp	-		-	Low	Yes	0.01	2.078	(72)
<i>UGT1A3*7a</i>	Phe	110	Ile	-		-	Intermediate	Yes	0.39	0.804	(73)
<i>UGT1A3*8a</i>	Ala	158	Val	-		-	Low	Yes	0.54	1.223	(73)
<i>UGT1A3*10a</i>	Val	47	Ala	rs6431625		-	Intermediate	Yes	0.1	0.726	(73)
<i>UGT1A4*2</i>	Pro	24	Thr	-		-	Substrate dependent	No	0.58	2.3	(74)
<i>UGT1A4*3b</i>	Leu	48	Val	-		-	Low activity	Yes	1	0.072	(75)
<i>UGT1A7*4</i>	Trp	208	Arg	rs11692021		-	Low	Yes	0.39	1.199	(76)
<i>UGT1A7*5</i>	Gly	115	Ser	-		-	Low	Yes	0.74	0.262	(77)
<i>UGT1A7*6</i>	Glu	139	Asp	-		-	High	Yes	0.59	0.318	(77)
<i>UGT1A8*3</i>	Cys	277	Tyr	rs17863762		-	Severely reduced	Yes	0	3.076	(78)
<i>UGT1A9*3a</i>	Met	33	Thr	-		-	Substrate-dependent impact	Yes	1	1.074	(77)
<i>UGT1A9*5</i>	Asp	256	Asn	-		-	Decreased activity	Yes	0.61	0.994	(79)
<i>UGT1A10*6</i>	Thr	202	Ile	VAR_018355		-	Reduced by 50%	Yes	0.03	1.377	(79,80)
<i>UGT1A10*7</i>	Ile	211	Thr	-		-	No activity	Yes	0.02	2.003	(81)
<i>UGT2B10*2</i>	Asp	67	Tyr	-		-	Associated with lower level of nicotine and cotinine glucuronidation in HLM	No	0.08	0.524	(60)
<i>UGT2B15*2</i>	Asp	85	Tyr	rs1902023		-	Increased activity on androgens	No	0.08	0.299	(61)

Table VII. Evaluation of Predicting Effect of SIFT and PolyPhen on Confirmed Human *UGT* nsSNPs

	Number	%
SIFT prediction		
True positive predictions	36	57
False negative predictions	27	43
Total	63	
PolyPhen prediction		
True positive predictions	42	67
False negative predictions	21	33
Total	63	

exhibit reduced activity. Zhang *et al.* (68) studied variants in H⁺/peptide cotransporter (PEPT1) involved in drug transportation, SIFT correctly predicted the SNP that reduced transport capacity to “affect protein function”, confirming the accuracy of SIFT on prediction of individual protein predictions. Our recent study on phenotype prediction of 791 validated nsSNPs in human cytochrome P450s using SIFT and PolyPhen, found that 70% of nsSNPs were correctly predicted as damaging (69).

Furthermore, the scatter graph plotted using functional consequence predictions from SIFT and PolyPhen showed negative correlation between the two sets of scores. Low scores from SIFT indicate an “intolerant” prediction, correlated with high scores from PolyPhen which indicates a “probably damaging” prediction. High scores from SIFT indicate a “tolerant” prediction, correlated with low scores from PolyPhen indicating a “benign” prediction. This correlation was further established by row percentage calculated between scores for each one of 247 nsSNPs predicted by SIFT and PolyPhen. The two algorithms show agreements at two extremes of the scores, a high percentage (85.7%) of “benign” predictions from PolyPhen was predicted “tolerant” by SIFT and a high percentage (84.2%) of “probably damaging” predictions from PolyPhen was predicted “intolerant” by SIFT. There were slight correlations for scores in between, showing a lack of agreement crossing different prediction categories of the two algorithms, demonstrating a strong correlation in their predictions on potential effect of known human *UGT* nsSNPs on protein function. The concordance between SIFT and PolyPhen prediction scores suggest they can be used in combination in the future to improve accuracy of prediction studies.

Although bioinformatics tools show their potential in reducing the number of nsSNPs for disease association studies by filtering nsSNPs that are most likely to be disease related, error predictions do occur. In this study, based on collected data from 63 *in vitro* and *in vivo* studies, the false negative predictions for SIFT and PolyPhen is 43% and 33%, respectively. There are several aspects affecting the prediction accuracy for prediction tools like SIFT and PolyPhen algorithms. Firstly, SIFT and PolyPhen rely on several different databases for SNP information, polluted databases with erroneous SNP reports and bias of the data towards disease-related allelic variants are likely to lead to an over prediction of the number of deleterious nsSNPs (8). For example, Ramensky *et al.* (8) compared a fraction of nsSNPs predicted to be damaging for HGV-base entries and found that the overall prediction rate for the category “Suspected” nsSNPs was 31.4%, for the category “Proven” nsSNPs was 28.9% and for “Proven” nsSNPs from systematic studies on healthy

individual was 27.6%. Furthermore, programs identifying SNPs may detect base differences between the functional gene and a pseudogene and erroneously report these differences as SNPs in the function protein. Including nsSNPs erroneously mapped from pseudogenes in the SNP database will affect prediction accuracy on prediction tools using SNP information from these databases (9). The growth of public SNP data and improvements of SNP database data will assist in acquiring correct information on SNPs to assist in improving prediction accuracy of bioinformatics tools. Secondly, SIFT has a weighted false positive error of 19% and PolyPhen 9%, indicating that if all of the nsSNPs from dbSNP were functionally neutral, 19% or 9% would have been predicted as damaging by SIFT and PolyPhen, respectively (9). Additionally, there are factors overlooked when the predictions occurs, SIFT predicts on amino acid substitution in the protein product and does not take into account mutations that affect transcription, translation, splicing, and other possible pretranslational alterations (9), erroneous predictions may then arise. Furthermore, the prediction might appear incorrect based on the lack of association with obvious altered phenotypes. SIFT and PolyPhen may be sensitive to a mutation and predict it to be damaging to protein function, but if the phenotype is undiagnosed, or has not yet been assayed for therefore condemned an "error prediction". On the other hand, SIFT and PolyPhen may predict a deleterious mutation to be "tolerated" and because it is not an obvious phenotype, we believe that it is a correct prediction whereas in actual fact it is a deleterious mutation that is recessive or undiagnosed. Therefore it is important to identify the association of SNPs with various phenotype/diseases. Identifying deleterious nsSNPs using bioinformatics tools like SIFT and PolyPhen may lay the foundation and initiate the process of identifying SNPs that has clinical implications. There lies the ability to reduce the number and/or the field of possible deleterious nsSNPs that needs to be evaluated in clinical studies.

In summary, *in silico* analysis predicted that 35.3–46% of over 200 amino acid substitution variants currently identified in the human UGT genes might cause reduced enzyme activity and/or elevated bilirubin levels. The identification of these variants using bioinformatic tools such as SIFT and PolyPhen narrows down study fields and is the first step of the progress to evaluate their possible phenotypic importance in costly clinical studies.

ACKNOWLEDGEMENTS

The authors appreciate the technical assistance of Dr Lin-Lin Wang of Institute of Reproductive and Child Health, Peking University, Beijing, China. Ms. Yuan Ming Di is a holder of RMIT University PhD Scholarship. The authors appreciate the support of RMIT Health Innovations Research Institute, RMIT University, Bundoora, Victoria, Australia. We also would like to thank Associate Professor Clifford Da Costa (Department of Mathematics and Statistics, RMIT University, Melbourne, Australia) for his assistance in the statistical analysis of the data in this paper.

REFERENCES

- Guillemette C. Pharmacogenomics of human UDP-glucuronosyltransferase enzymes. *Pharmacogenomics J.* 2003;3:136–58.
- Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science.* 1999;286:487–91.
- Mackenzie PI, Owens IS, Burchell B, et al. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics.* 1997;7:255–69.
- King CD, Rios GR, Green MD, Tephly TR. UDP-glucuronosyltransferases. *Curr Drug Metab.* 2000;1:143–61.
- Mackenzie PI, Bock KW, Burchell B, et al. Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenet Genomics.* 2005;15:677–85.
- Harding D, Jeremiah SJ, Povey S, Burchell B. Chromosomal mapping of a human phenol UDP-glucuronosyltransferase, GNT1. *Ann Hum Genet.* 1990;54:17–21.
- Turgeon D, Carrier JS, Levesque E, Hum DW, Belanger A. Relative enzymatic activity, protein stability, and tissue distribution of human steroid-metabolizing UGT2B subfamily members. *Endocrinology.* 2001;142:778–87.
- Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 2002;30:3894–900.
- Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 2002;12:436–46.
- Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet.* 2006;7:61–80.
- Herrgard S, Cammer SA, Hoffman BT, et al. Prediction of deleterious functional effects of amino acid mutations using a library of structure-based function descriptors. *Proteins.* 2003;53:806–16.
- Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics.* 2006;7:217.
- Ehmer U, Vogel A, Schutte JK, Krone B, Manns MP, Strassburg CP. Variation of hepatic glucuronidation: novel functional polymorphisms of the UDP-glucuronosyltransferase UGT1A4. *Hepatology.* 2004;39:970–7.
- Guillemette C, Millikan RC, Newman B, Housman DE. Genetic polymorphisms in uridine diphospho-glucuronosyltransferase 1A1 and association with breast cancer among African Americans. *Cancer Res.* 2000;60:950–6.
- Crigler JF Jr, Najjar VA. Congenital familial nonhemolytic jaundice with kernicterus. *Pediatrics.* 1952;10:169–80.
- Black M, Billing BH. Hepatic bilirubin udp-glucuronyl transferase activity in liver disease and Gilbert's syndrome. *N Engl J Med.* 1969;280:1266–71.
- Moghrabi N, Clarke DJ, Boxer M, Burchell B. Identification of an A-to-G missense mutation in exon 2 of the UGT1 gene complex that causes Crigler-Najjar syndrome type 2. *Genomics.* 1993;18:171–3.
- Bosma PJ, Chowdhury JR, Huang TJ, et al. Mechanisms of inherited deficiencies of multiple UDP-glucuronosyltransferase isoforms in two patients with Crigler-Najjar syndrome, type I. *Faseb J.* 1992;6:2859–63.
- Iyer L, Das S, Janisch L, et al. UGT1A1*28 polymorphism as a determinant of irinotecan disposition and toxicity. *Pharmacogenomics J.* 2002;2:43–7.
- Ando Y, Saka H, Ando M, et al. Polymorphisms of UDP-glucuronosyltransferase gene and irinotecan toxicity: a pharmacogenetic analysis. *Cancer Res.* 2000;60:6921–6.
- Wasserman E, Myara A, Lokiec F, et al. Severe CPT-11 toxicity in patients with Gilbert's syndrome: two case reports. *Ann Oncol.* 1997;8:1049–51.
- Kaniwa N, Kurose K, Jinno H, et al. Racial variability in haplotype frequencies of UGT1A1 and glucuronidation activity of a novel single nucleotide polymorphism 686C> T (P229L) found in an African-American. *Drug Metab Dispos.* 2005;33:458–65.
- Danoff TM, Campbell DA, McCarthy LC, et al. A Gilbert's syndrome UGT1A1 variant confers susceptibility to tranilast-induced hyperbilirubinemia. *Pharmacogenomics J.* 2004;4:49–53.
- Strassburg CP, Vogel A, Kneip S, Tukey RH, Manns MP. Polymorphisms of the human UDP-glucuronosyltransferase (UGT) 1A7 gene in colorectal cancer. *Gut.* 2002;50:851–6.
- Zheng Z, Park JY, Guillemette C, Schantz SP, Lazarus P. Tobacco carcinogen-detoxifying enzyme UGT1A7 and its association with orolaryngeal cancer risk. *J Natl Cancer Inst.* 2001;93:1411–8.
- Vogel A, Ockenga J, Ehmer U, et al. Polymorphisms of the carcinogen detoxifying UDP-glucuronosyltransferase UGT1A7 in proximal digestive tract cancer. *Z Gastroenterol.* 2002;40:497–502.

27. Vogel A, Kneip S, Barut A, et al. Genetic link of hepatocellular carcinoma with polymorphisms of the UDP-glucuronosyltransferase UGT1A7 gene. *Gastroenterology*. 2001;121:1136–44.
28. Darbari DS, van Schaik RH, Capparelli EV, Rana S, McCarter R, van den Anker J. UGT2B7 promoter variant -840G>A contributes to the variability in hepatic clearance of morphine in patients with sickle cell disease. *Am J Hematol*. 2008;83:200–2.
29. Holthe M, Rakvag TN, Klepstad P, et al. Sequence variations in the UDP-glucuronosyltransferase 2B7 (UGT2B7) gene: identification of 10 novel single nucleotide polymorphisms (SNPs) and analysis of their relevance to morphine glucuronidation in cancer patients. *Pharmacogenomics J*. 2003;3:17–26.
30. MacLeod SL, Nowell S, Plaxco J, Lang NP. An allele-specific polymerase chain reaction method for the determination of the D85Y polymorphism in the human UDP-glucuronosyltransferase 2B15 gene in a case-control study of prostate cancer. *Ann Surg Oncol*. 2000;7:777–82.
31. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res*. 2001;11:863–74.
32. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31:3812–4.
33. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet*. 2001;10:591–7.
34. Ng PC, Henikoff JG, Henikoff S. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*. 2000;16:760–6.
35. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001;305:567–80.
36. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. *Science*. 1991;252:1162–4.
37. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*. 1997;10:1–6.
38. Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics*. 2004;83:970–9.
39. Aono S, Adachi Y, Uyama E, et al. Analysis of genes for bilirubin UDP-glucuronosyltransferase in Gilbert's syndrome. *Lancet*. 1995;345:958–9.
40. Ciotti M, Chen F, Rubaltelli FF, Owens IS. Coding defect and a TATA box mutation at the bilirubin UDP-glucuronosyltransferase gene cause Crigler-Najjar type I disease. *Biochim Biophys Acta*. 1998;1407:40–50.
41. Ciotti M, Obaray R, Martin MG, Owens IS. Genetic defects at the UGT1 locus associated with Crigler-Najjar type I disease, including a prenatal diagnosis. *Am J Med Genet*. 1997;68:173–8.
42. Ciotti M, Werlin SL, Owens IS. Delayed response to phenobarbital treatment of a Crigler-Najjar type II patient with partially inactivating missense mutations in the bilirubin UDP-glucuronosyltransferase gene. *J Pediatr Gastroenterol Nutr*. 1999;28:210–3.
43. Erps LT, Ritter JK, Hersh JH, Blossom D, Martin NC, Owens IS. Identification of two single base substitutions in the UGT1 gene locus which abolish bilirubin uridine diphosphate glucuronosyltransferase activity *in vitro*. *J Clin Invest*. 1994;93:564–70.
44. Labrune P, Myara A, Hadchouel M, et al. Genetic heterogeneity of Crigler-Najjar syndrome type I: a study of 14 cases. *Hum Genet*. 1994;94:693–7.
45. Maruo Y, Nishizawa K, Sato H, Sawa H, Shimada M. Prolonged unconjugated hyperbilirubinemia associated with breast milk and mutations of the bilirubin uridine diphosphate-glucuronosyltransferase gene. *Pediatrics*. 2000;106:E59.
46. Seppen J, Bosma PJ, Goldhoorn BG, et al. Discrimination between Crigler-Najjar type I and II by expression of mutant bilirubin uridine diphosphate-glucuronosyltransferase. *J Clin Invest*. 1994;94:2385–91.
47. Seppen J, Steenken E, Lindhout D, Bosma PJ, Elferink RP. A mutation which disrupts the hydrophobic core of the signal peptide of bilirubin UDP-glucuronosyltransferase, an endoplasmic reticulum membrane protein, causes Crigler-Najjar type II. *FEBS Lett*. 1996;390:294–8.
48. Chalasani N, Chowdhury NR, Chowdhury JR, Boyer TD. Kernicterus in an adult who is heterozygous for Crigler-Najjar syndrome and homozygous for Gilbert-type genetic defect. *Gastroenterology*. 1997;112:2099–103.
49. Farheen S, Sengupta S, Santra A, et al. Gilbert's syndrome: high frequency of the (TA)₇ TAA allele in India and its interaction with a novel CAT insertion in promoter of the gene for bilirubin UDP-glucuronosyltransferase 1 gene. *World J Gastroenterol*. 2006;12:2269–75.
50. Huang CS, Luo GA, Huang ML, Yu SC, Yang SS. Variations of the bilirubin uridine-diphosphoglucuronosyl transferase 1A1 gene in healthy Taiwanese. *Pharmacogenetics*. 2000;10:539–44.
51. Kadakol A, Ghosh SS, Sappal BS, Sharma G, Chowdhury JR, Chowdhury NR. Genetic lesions of bilirubin uridine-diphosphoglucuronate glucuronosyltransferase (UGT1A1) causing Crigler-Najjar and Gilbert syndromes: correlation of genotype to phenotype. *Hum Mutat*. 2000;16:297–306.
52. Labrune P, Myara A, Chalas J, Le Bihan B, Capel L, Francoual J. Association of a homozygous (TA)₈ promoter polymorphism and a N400D mutation of UGT1A1 in a child with Crigler-Najjar type II syndrome. *Hum Mutat*. 2002;20:399–401.
53. Maruo Y, Serdaroglu E, Iwai M, et al. A novel missense mutation of the bilirubin UDP-glucuronosyltransferase gene in a Turkish patient with Crigler-Najjar syndrome type I. *J Pediatr Gastroenterol Nutr*. 2003;37:627–30.
54. Petit F, Gajdos V, Capel L, et al. Crigler-Najjar type II syndrome may result from several types and combinations of mutations in the UGT1A1 gene. *Clin Genet*. 2006;69:525–7.
55. Servedio V, d'apolito M, Maiorano N, et al. Spectrum of UGT1A1 mutations in Crigler-Najjar (CN) syndrome patients: identification of twelve novel alleles and genotype-phenotype correlation. *Hum Mutat*. 2005;25:325.
56. Sutomo R, Laosombat V, Sadewa AH, et al. Novel missense mutation of the UGT1A1 gene in Thai siblings with Gilbert's syndrome. *Pediatr Int*. 2002;44:427–32.
57. Aono S, Yamada Y, Keino H, et al. Identification of defect in the genes for bilirubin UDP-glucuronosyl-transferase in a patient with Crigler-Najjar syndrome type II. *Biochem Biophys Res Commun*. 1993;197:1239–44.
58. Huang CS, Luo GA, Huang MJ, Chen ES, Young TH, Chao YC. A novel compound heterozygous variation of the uridine-diphosphoglucuronosyl transferase 1A1 gene that causes Crigler-Najjar syndrome type II. *Pharmacogenetics*. 2001;11:639–42.
59. Udomuksorn W, Elliot DJ, Lewis BC, Mackenzie PI, Yoovathaworn K, Miners JO. Influence of mutations associated with Gilbert and Crigler-Najjar type II syndromes on the glucuronidation kinetics of bilirubin and other UDP-glucuronosyltransferase 1A substrates. *Pharmacogenet Genomics*. 2007;17:1017–29.
60. Chen G, Blevins-Primeau AS, Dellinger RW, Muscat JE, Lazarus P. Glucuronidation of nicotine and cotinine by UGT2B10: loss of function by the UGT2B10 Codon 67 (Asp>Tyr) polymorphism. *Cancer Res*. 2007;67:9024–9.
61. Levesque E, Beaulieu M, Green MD, Tephly TR, Belanger A, Hum DW. Isolation and characterization of UGT2B15(Y85): a UDP-glucuronosyltransferase encoded by a polymorphic gene. *Pharmacogenetics*. 1997;7:317–25.
62. Tukey RH, Strassburg CP. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu Rev Pharmacol Toxicol*. 2000;40:581–616.
63. Gsur A, Preyer M, Haidinger G, et al. A polymorphism in the UDP-Glucuronosyltransferase 2B15 gene (D85Y) is not associated with prostate cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2002;11:497–8.
64. Li Y, Wang Y, Li Y, Yang L. Prediction of the deleterious nsSNPs in ABCB transporters. *FEBS Lett*. 2006;580:6800–6.
65. Evans WE, Johnson JA. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu Rev Genomics Hum Genet*. 2001;2:9–39.
66. Ingelman-Sundberg M, Oscarson M, McLellan RA. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. *Trends Pharmacol Sci*. 1999;20:342–9.
67. Yamamoto K, Soeda Y, Kamisako T, et al. Analysis of bilirubin uridine 5'-diphosphate (UDP)-glucuronosyltransferase gene mutations in seven patients with Crigler-Najjar syndrome type II. *J Hum Genet*. 1998;43:111–4.
68. Zhang EY, Fu DJ, Pak YA, et al. Genetic polymorphisms in human proton-dependent dipeptide transporter PEPT1: implica-

- tions for the functional role of Pro586. *J Pharmacol Exp Ther*. 2004;310:437–45.
69. Wang L. A bioinformatics approach for the phenotype prediction of non-synonymous single nucleotide polymorphisms in human cytochrome P450s. *Drug Metab Dispos*. 2009;37:977–91.
70. Bosma PJ, Goldhoorn B, Oude Elferink RP, Sinaasappel M, Oostra BA, Jansen PL. A mutation in bilirubin uridine 5'-diphosphate-glucuronosyltransferase isoform 1 causing Crigler-Najjar syndrome type II. *Gastroenterology*. 1993;105:216–20.
71. Yusoff S, Van Rostenberghe H, Yusoff NM, et al. Frequencies of A(TA)7TAA, G71R, and G493R mutations of the UGT1A1 gene in the Malaysian population. *Biol Neonate*. 2006;89:171–6.
72. Iwai M, Maruo Y, Ito M, Yamamoto K, Sato H, Takeuchi Y. Six novel UDP-glucuronosyltransferase (UGT1A3) polymorphisms with varying activity. *J Hum Genet*. 2004;49:123–8.
73. Caillier B, Lepine J, Tojcic J, et al. A pharmacogenomics study of the human estrogen glucuronosyltransferase UGT1A3. *Pharmacogenet Genomics*. 2007;17:481–95.
74. Wiener D, Fang JL, Dossett N, Lazarus P. Correlation between UDP-glucuronosyltransferase genotypes and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone glucuronidation phenotype in human liver microsomes. *Cancer Res*. 2004;64:1190–6.
75. Saeki M, Saito Y, Jinno H, et al. Genetic variations and haplotypes of UGT1A4 in a Japanese population. *Drug Metab Pharmacokinet*. 2005;20:144–51.
76. Guillemette C, Ritter JK, Auyeung DJ, Kessler FK, Housman DE. Structural heterogeneity at the UDP-glucuronosyltransferase 1 locus: functional consequences of three novel missense mutations in the human UGT1A7 gene. *Pharmacogenetics*. 2000;10:629–44.
77. Villeneuve L, Girard H, Fortier LC, Gagne JF, Guillemette C. Novel functional polymorphisms in the UGT1A7 and UGT1A9 glucuronidating enzymes in Caucasian and African-American subjects and their impact on the metabolism of 7-ethyl-10-hydroxycamptothecin and flavopiridol anticancer drugs. *J Pharmacol Exp Ther*. 2003;307:117–28.
78. Huang YH, Galijatovic A, Nguyen N, et al. Identification and functional characterization of UDP-glucuronosyltransferases UGT1A8*1, UGT1A8*2 and UGT1A8*3. *Pharmacogenetics*. 2002;12:287–97.
79. Jinno H, Saeki M, Saito Y, et al. Functional characterization of human UDP-glucuronosyltransferase 1A9 variant, D256N, found in Japanese cancer patients. *J Pharmacol Exp Ther*. 2003;306:688–93.
80. Saeki M, Ozawa S, Saito Y, et al. Three novel single nucleotide polymorphisms in UGT1A10. *Drug Metab Pharmacokinet*. 2002;17:488–90.
81. Martineau I, Tchernof A, Belanger A. Amino acid residue ILE211 is essential for the enzymatic activity of human UDP-glucuronosyltransferase 1A10 (UGT1A10). *Drug Metab Dispos*. 2004;32:455–9.