RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Performance of In Silico Tools for the Evaluation of *UGT1A1* Missense Variants

Carina Rodrigues,[1,2] Alice Santos-Silva,[1] Elísio Costa,[1] and Elsa Bronze-da-Rocha[1]*

[1]UCIBIO/REQUIMTE, Laboratório de Bioquímica, Departamento de Ciências Biológicas, Faculdade de Farmácia, Universidade do Porto, Porto, Portugal; [2]Escola Superior de Saúde, Instituto Politécnico de Bragança, Bragança, Portugal

Communicated by Sean V. Tavtigian

Received 3 November 2014; accepted revised manuscript 31 August 2015.
Published online 17 September 2015 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22903

ABSTRACT: Variations in the gene encoding uridine diphosphate glucuronosyltransferase 1A1 (*UGT1A1*) are particularly important because they have been associated with hyperbilirubinemia in Gilbert's and Crigler–Najjar syndromes as well as with changes in drug metabolism. Several variants associated with these phenotypes are nonsynonymous single-nucleotide polymorphisms (nsSNPs). Bioinformatics approaches have gained increasing importance in predicting the functional significance of these variants. This study was focused on the predictive ability of bioinformatics approaches to determine the pathogenicity of human *UGT1A1* nsSNPs, which were previously characterized at the protein level by in vivo and in vitro studies. Using 16 Web algorithms, we evaluated 48 nsSNPs described in the literature and databases. Eight of these algorithms reached or exceeded 90% sensitivity and six presented a Matthews correlation coefficient above 0.46. The best-performing method was MutPred, followed by Sorting Intolerant from Tolerant (SIFT). The prediction measures varied significantly when predictors such us SIFT, polyphen-2, and Prediction of Pathological Mutations on Proteins were run with their native alignment generated by the tool, or with an input alignment that was strictly built with *UGT1A1* orthologs and manually curated. Our results showed that the prediction performance of some methods based on sequence conservation analysis can be negatively affected when nsSNPs are positioned at the hypervariable or constant regions of *UGT1A1* ortholog sequences.

Hum Mutat 36:1215–1225, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: *UGT1A1*; nsSNPs; bioinformatics; genotype; phenotype; protein function

Introduction

Uridine diphosphate glucuronosyltransferase (UGT1A1) is an enzyme involved in the metabolism and detoxification of numerous xenobiotic and endogenous compounds, including bilirubin [Strassburg et al., 2008]. To date, more than 300 single-nucleotide polymorphisms (SNPs) within the *UGT1A1* gene (MIM #191740; HUGO: 12530) have been identified. Many of these variants have been associated with human diseases; however, in some cases, the absence of functional studies has led to their pathogenicity remaining unclear. As with other genes, *UGT1A1* presents many nonsynonymous SNPs (nsSNPs) in which genotype–phenotype correlations are not established. Some of them have been associated with Gilbert's syndrome (GS) [Huang et al., 2000; Kaniwa et al., 2005; Costa 2006; Farheen et al., 2006] and with Crigler–Najjar syndrome (CNS) types I and II [Labrune et al., 1994; Seppen et al., 1994; Ciotti et al., 1998; Labrune et al., 2002; Yusoff et al., 2006; Sneitz et al., 2010], as well as with changes in drug clearance and/or response [Sai et al., 2004]. GS is a benign and common condition characterized by a deficiency in bilirubin conjugation due to reduced activity of the enzyme glucuronyltransferase (approximately 30%), which causes moderate hyperbilirubinemia [Huang et al., 2000; Costa, 2006]. The most common variant associated with GS is the TA duplication at position c.-41_-40dupTA (variant *UGT1A1**28*) of the start codon in the promoter region of the *UGT1A1* gene [Bosma et al., 1995]. CNS types I and II are also associated with mutations in the *UGT1A1* gene, which lead to the absence or severe reduction of UGT1A1 enzyme activity [Deiss, 1999; Costa, 2006]. The clinical classification of the two CNS types is also based on the response to phenobarbital administration, which induces UGT1A1 enzyme activity, and by the presence of kernicterus [Costa, 2006].

As the functional evaluation of nsSNPs is time-consuming and expensive, bioinformatics tools have gained increasing importance and have been used to guide additional functional studies. These methods can be divided into two main categories: sequence-based approaches and sequence- and structure-based methods. The former uses multiple sequence alignments (MSA) and incorporates different methodologies to measure residue conservation, namely, Sorting Intolerant from Tolerant (SIFT) [Kumar et al., 2009], Prediction of Pathological Mutations on Proteins (PMUT) [Ferrer-Costa et al., 2005], and multivariate analysis of protein polymorphism (MAPP) [Stone and Sidow, 2005]. These tools assume that functional SNPs occur at evolutionarily conserved sites and that the majority of nsSNP are functionally neutral [Ng and Henikoff, 2006; Tavtigian et al., 2006]. Thus, the prediction of a SNP's functional effect is created based on conservation and variation in a specific position. The latter group uses sequence and structure features, which predict the possible impact of an amino acid substitution. Examples of this type of program include polymorphism phenotyping, version-2 (Polyphen-2) [Adzhubei et al., 2010] and molecular phenotyping of coding nsSNPs (SNPeffect) [Reumers et al., 2006]. Some structure-based methods are generally more reliable when the three-dimensional (3D) structure of the protein is considered and

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Elsa Bronze-da-Rocha, Departamento de Ciências Biológicas, Laboratório de Bioquímica, Faculdade de Farmácia da Universidade do Porto, Rua Jorge Viterbo Ferreira 228, Porto 4050-313, Portugal. E-mail: elsa.rocha@ff.up.pt

Contract grant sponsors: Fundação para a Ciência e Tecnologia (FCT) and Fundo Social Europeu (FSE) (grant SFRH/BD/42791/2007).

the nsSNPs are classified based on size, polarity, and protein stability changes. Nevertheless, the use of these methods might be limited due to the lack of protein structural information. However, these prediction methods can, to some extent, analyze one sequence and modulate substitutions in the structure of a homologous protein rather than in the exact protein structure of interest. The nsSNPs functional prediction methods can also incorporate annotations from the Swiss-Prot database or use information from other prediction programs to identify transmembrane regions or secondary structures [Wang and Moult, 2001; Ferrer-Costa et al., 2004]. Moreover, the consensus deleteriousness (CONDEL) tool combines different Web tool outputs into a unified classification [Gonzalez-Perez et al., 2011]. Most of the sequence-based amino acid substitutions that prediction methods accept as input include a protein sequence or protein identifier (ID). Afterwards, the search is performed against a system databases to find homologous sequences and produce an MSA [Karchin et al., 2005; Ng and Henikoff, 2006; Jordan et al., 2010; Thusberg et al., 2011]. To automatically generate MSAs, the homologous sequences are selected from several databanks such as the SWISS-PROT or NCBI's nonredundant protein databases. The obtained MSAs contain orthologs, paralogs, and multiple versions of the same sequence that can potentially alter conservation profiles. Between two paralogs, the average amino acid sequence identity is only 30% [Wong and Zang, 2014]. Ideally, only orthologous sequences should be used in an MSA, but paralogs have been widely used, probably due to the limited number of orthologous sequences. Some observations have noted that a MSA constructed only with orthologous sequences can lead to a more reliable evolutionary analysis, which improves the predictor's performance [Shu et al., 2003; Tavtigian et al., 2006]. The aim of the present study was to investigate the prediction ability of 16 Web available algorithms to assign biological or biochemical roles to a set of *UGT1A1* variants that have been phenotypically characterized by *in vivo* and *in vitro* studies.

Methods

Variants Search

To define a set of *UGT1A1* gene (MIM #191740; HUGO: 12530) SNPs previously characterized at the protein level, we performed a literature review and a search for variants using the databases: http://www.polydoms.cchmc.org/polydoms, http://www.ensembl.org, http://www.genecards.org, and http://www.ncbi.nlm.nih.gov/SNP. Information was obtained from the specific site for UGT nomenclature (http://www.pharmacogenomics.pha.ulaval.ca/cms/ugt_alleles/) to find *UGT1A1* variants related to unconjugated hyperbilirubinemia in GS and CNS. Despite the description of more than 300 SNPs in the database, this study only included 87 variants, which are described in the literature as being associated or not with hyperbilirubinemia. Among these, 38 were classified as pathogenic based on in vitro or in vivo studies that included site-directed mutagenesis, expression studies, assays of liver biopsy specimen, administration of phenobarbital, and the duodenal bile pattern; however, 10 SNPs were associated with normal bilirubin levels. These two groups of variants represent the gold standard for this study. The remaining 37 nsSNPs were also described as being associated with unconjugated hyperbilirubinemia, but the functional impact at the protein level had not yet been determined (Supp. Table S1).

Running Predictors

The sequence-based tools included: SIFT, MAPP; Protein Analysis Through Evolutionary Relationships (PANTHER), Predictor of Human Deleterious SNPs (PhD-SNP), Gene Ontology (GO) database in the form of a GO-based score (SNPs&GO), Mutation Assessor (Xvar), and Align Grantham Variance/Grantham Difference (A-GVGD). SIFT predicts whether an amino acid substitution or insertions/deletions affect protein function and is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences [Kumar et al., 2009; Sim et al., 2012]. MAPP predictions are established by assessing the physicochemical variation in each column of a sequence alignment, and require a MSA and a phylogenetic tree detailing the evolutionary distance between species [Stone and Sidow, 2005]. PANTHER is a system constructed with complete genomes (82) ordered into gene families and subfamilies, and their evolutionary relationships are collected in phylogenetic trees, MSAs, and statistical models (hidden Markov models or HMMs) [Thomas et al., 2003; Mi et al., 2013]. PhD-SNP is created on a support vector machine (SVM) adept at using different sequence and evolutionary information to predict variant pathogenicity. SNPs&GO collects a specific framework of information derived from the protein sequence, protein sequence profile, and protein function to predict human disease-related single point protein mutations [Calabrese et al., 2009]. Xvar predictions are based on the evolutionary conservation of the affected amino acid in protein homologs [Reva et al., 2011]. A-GVGD provides a class probability constructed based on the evolutionary conservation and chemical nature of the amino acid residues to predict whether a mutation is enriched as deleterious or enriched as neutral. A-GVGD achieves this by combining multiple protein sequence alignments along with information on the side chain composition, polarity and steric characteristics of the mutant amino acid to determine whether the mutation might lead to the modification of the protein structure [Tavtigian et al., 2006]. The sequence- and structure-based approaches comprised: Polymorphism Phenotyping version-2 (Polyphen-2), Pathogenic Mutation prediction (PMUT) on proteins; Screening for Nonacceptable Polymorphisms (SNAP); the SVM-based method Hansa; SNPeffect; Mutation Prediction (MutPred); Functional Analysis Through Hidden Markov models (FATHMM version 2.3); CONsensus DELeteriousness (CONDEL); and Meta-Predictor (Meta-SNP). PolyPhen-2 is based on the possible impact of an amino acid substitution on the structure and function of a human protein by analyzing several features comprising the sequence and eight sequence-based and three structure-based features, which were selected using machine learning (Bayesian classification) [Adzhubei et al., 2010]. PMUT uses different types of sequence information to label mutations, as well as neural networks to process this information for the prediction of pathological mutations [Ferrer-Costa et al., 2005]. SNAP is a neural-network-based method that uses in silico-derived protein information (e.g., secondary structure, conservation, solvent accessibility) to predict the functionality of mutated proteins [Bromberg and Rost, 2007; Bromberg et al., 2008]. Hansa is based on the SVM method, which uses a set of discriminatory features (10) to categorize missense mutations as neutral or deleterious [Hicks et al., 2013] and to map mutations onto the query protein as "disease" or "neutral" [Acharya and Nagarajaram, 2012]. SNPeffect version 2.0 provides a platform for predicting the effect of coding nsSNPs on the structure and function of the affected protein; the high resolution structural data of the SNPeffect server are used to unequivocally model the mutant structures so that changes in protein stability and binding can be evaluated [Reumers et al., 2006]. MutPred uses a random forest algorithm based on the probabilities of gain or loss of properties relating to features of protein structure and dynamics to predict the functional properties and amino acid sequence, provide evolutionary information, and calculate the molecular cause of disease-associated substitution [Li et al., 2009]. FATHMM version 2.3 is a species-independent method with optional species-specific weightings for the prediction of the functional effects of protein missense variants that allows the user to discriminate deleterious and neutral polymorphisms along with molecular and phenotypic consequences [Shihab et al., 2013]. Meta-SNP is a random forest-based binary classifier used to distinguish between disease-related and polymorphic nonsynonymous single-nucleotide variants (nsSNVs), which takes as input the output of the four predictors (PANTHER, PhD-SNP, SIFT, and SNAP) [Capriotti et al., 2013]. CONDEL and Meta-SNP were included as integrative tools. CONDEL establishes a weighted average of normalized scores of three individual methods, SIFT, Polyphen-2, and Xvar (CONDEL first version, available until April 2014; CONDEL v1); a recent version of this method integrates another two different tools (FATHMM and MAPP). The main characteristics of these methods are summarized in Table 1 according to the guidelines on this topic [Vihinen, 2012, 2013], and the most important steps used for this study are depicted in Figure 1.

Most of these tools only require as input the protein sequence or identifier and the amino acid substitution of interest. Usually, the majority of the outputs of these tools provide a binary classification of variants: “neutral/tolerant” or “deleterious/pathogenic.” When the output of a specific tool had more than two categories, we grouped them into a binary classification (neutral and deleterious). This is a strategy already used by other authors with the same tools [Tavtigian et al., 2006; Hicks et al., 2011]. In the Xvar method, the nsSNPs can be predicted as “neutral,” “low,” “medium,” and “high” [Hicks et al., 2011]; however, in this study, the variants classified as “neutral” and “low” were considered neutral, and the variants predicted as “medium” and “high” were clustered as deleterious variants. Using Polyphen-2, Adzhubei and coworkers (2010) considered that the variants predicted as “benign” were neutral and the other classes (“possibly damaging” or “probably damaging”) were deleterious variants; in the present study, we followed this classification for neutral and deleterious variants. The Align-GVGD tool classifies variants as “neutral,” “unclassified,” and “deleterious” [Tavtigian et al., 2006; Hicks et al., 2011]. For our study, the variants predicted as “neutral” and “unclassified” were categorized as neutral variants, as previously reported by others [Hicks et al., 2011]. The output for the MutPred tool contains a general score (*g*) inferring the probability that the amino acid substitution is deleterious/disease-associated, and a top five property score (*p*) for its impact on the structural and functional properties of the protein. According to the tool developer, a *P* value less than 0.05 is considered significant, even when the user applies their own limits of significance. In this study, we considered the cutoff of *g* 0.75 and $p < 0.05$ (referred to as confident hypotheses) to differentiate between benign and pathogenic mutations, as reported elsewhere [Li et al., 2009]. The tool Human Splicing Finder (HSF) (Version 2.4.1) was used to test whether any of the pathogenic missense substitutions interfere with splicing signals or motifs in some human sequences [Desmet et al., 2009]. All field tests using the described algorithms were run on a PC from January 7 to July 20, 2014 (Supp. Tables S2A, S2B, and S3), with the exception of Meta-SNP and HSF, which were run from January 15 to March 10, 2015.

Tools Depending on MSAs

Some of the evaluated tools generate their own alignments and do not allow users to create and submit their own MSA. In these cases, the protein sequence was submitted in FASTA format or protein ID using the selected methods (Xvar; PANTHER; PhD-SNP; Hansa; SNPs&GO; and MutPred). Other tools require a MSA built by the user as input, such as the A-GVGD and MAPP tools. Moreover, MAPP needs a phylogenetic unrooted tree. SIFT and PMUT generate an alignment internally but they also permit user-generated alignments. The Web server version of Polyphen-2 has its own alignment pipeline, but user-generated alignments can be submitted to the stand-alone software version that can be downloaded onto a local computer. Using these last three methods, we were able to test whether their performance would be affected by the alignment employed. For this purpose, and also to run the A-GVGD and MAPP algorithms, we performed a search of *UGT1A1* orthologs in databases and built a MSA with 27 sequences. Most of the sequences were retrieved from the inparanoid (http://inparanoid.sbc.su.se/cgi-bin/index.cgi) and Ensembl databases (http://www.ensembl.org/info/docs/compara/). Orthologs of the human UGT1A1 protein were aligned by a multiple alignment program for amino acid or nucleotide sequences using Multiple Alignment from Fast Fourier Transform (MAFFT, version 7) [Katoh et al., 2013]. The identified orthologs were selected and those that produced large gaps were removed from the MSA. Functional predictions from SIFT were performed by two input options: in an automated manner, where the SIFT search for protein sequence homologies to the query protein; and based on the sequence that allows the calculation of the probability for each possible amino acid change. Users can select UniRef90, SeqSWISS-PROT, SWISS-PROT/TrEMBL, or NCBI’s nonredundant protein databases to search for homologies by using the PSI-BLAST method with default settings or by changing the median conservation of the sequences. There are two other possibilities, the use of multiple related sequences or MSA. We ran these methods with the 27 orthologs of MSA to test tool performance.

Statistical Analysis

The predictive capacity of the functional significance of *UGT1A1* nsSNPs of the studied algorithms was evaluated by statistical measures of performance, such as sensitivity (SEN), specificity (SPC), Matthew correlation coefficient (MCC) [Matthews, 1975], and accuracy (ACC). Sensitivity refers to the probability of identifying true deleterious mutations, whereas specificity represents the probability of identifying true neutral mutations [Hicks et al., 2011]. As reported in previous studies, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were used to compute sensitivity (SEN) or a true positive rate, specificity (SPC), or a true negative rate, ACC and MCC [Hicks et al., 2011], as follows:

$$\mathrm{SEN} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \quad \mathrm{SPC} = \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}}$$

$$\mathrm{ACC} = \frac{(\mathrm{TP}+\mathrm{TN})}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$$

The MCC scores range from +1 (a perfect prediction) to −1 (an inverse prediction) where 0 represents an average random prediction [Baldi et al., 2000]. This measurement has been favored over “accuracy,” as the last is less sensitive to the different numbers of pathogenic and nonpathogenic variant classes. Usually, the value for MCC increases in a slower manner and reaches a maximum of 0.5 when 75% of cases were correctly predicted. Random results (50% of both negative and positive correctly predicted) give a value of 0. The McNemar’s $\chi 2$ test ($\alpha = 0.05$) was used to assess the

Table 1. Selected Tools for the Analysis of Functionally Characterized *UGT1A1* nsSNPs

Computational methods	Type of analysis	Description	URL	Input	Output	References
SIFT (sorting intolerant from tolerant)	Sequence/evolutionary conservation	Conservation among protein homologs (sequence based). Uses multiple sequence alignments (MSAs) from precomputed BLAST searches from the NCBI.	http://blocks.fhcrc.org/sift/SIFT.html	Protein sequences, database SNP (dbSNP) identification (ID), alignment are optional and there are other possibilities. Allows the user to provide other alignments.	SIFT scores ranges from 0 to 1. The amino acid substitution is predicted as damaging, if the score is ≤0.05, and as tolerated if the score is >0.05. It also gives the number of sequences and the median conservation at a particular position.	Kumar et al. (2009)
MAPP (multivariate analysis of protein polymorphism)	Sequence/ evolutionary conservation	Sequence-based SNP prediction tool that considers the physicochemical variation present in a column of a MSA of homologous proteins.	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Protein sequence alignment in FASTA format and evolutionary tree in parenthesis representation with branch lengths.	MAPP produces an output table in which each row corresponds to a position (column) in the alignment. The tool displays a list of amino acid predicted that do not impair the protein function at this position (good amino acids) and a list of amino acids predicted as deleterious at this position (bad amino acids). The MAPP score is a continuous variable for all 20 amino acids at each position.	Stone and Sidow (2005)
A-GVGD (Align Grantham Variance/Grantham Difference)	Biophysical characteristics of amino acid/evolutionary conservation	Combines the biophysical characteristics of amino acids and protein MSAs to predict where missense substitutions in genes of interest fall in a spectrum from enriched deleterious to enriched neutral.	http://agvgd.iarc.fr/about.php	MSAs (in FASTA format) or select from their small library of alignments.	The biochemical variation at each alignment position is given by the Grantham Variation (GV) score for positions in a protein MSA (PMSA) and Grantham Deviation (GD) that measures the biochemical difference between the reference amino acid encoded by the variant. The difference between these properties and those of the amino acid variant assessed is calculated and GD is generated, in different classes, ranging from C0 to C65 (C0, C15, C25, C35, C45, C55, C65); C65 is most likely to interfere with function and C0 is less likely to compromise function.	Tavtigian et al. (2006)
PANTHER (Protein Analysis Through Evolutionary Relationships)	Sequence/evolutionary conservation	This tool uses evolutionary information encoded in the protein sequence profile for SNP prediction. It uses the hidden Markov model (HMM)-based statistical methods and MSAs to perform evolutionary analysis of coding SNPs.	http://www.pantherdb.org/tools/csnpScore.do	Protein sequence and position of the SNP.	The probability of a variant being pathogenic is calculated from the variation over each alignment column by a substitution position-specific evolutionary conservation (subPSEC) score. This score is a negative logarithm of the probability ratio of the wild type and mutant amino acids at a certain position: it ranges between 0 (neutral) and approximately 10 (most likely to be deleterious).	Thomas et al. (2003)
SNPs&GO Gene Ontology (GO) database, in the form of a GO-based score	Sequence/evolutionary conservation (gene ontology)	The tool collects information derived from the protein sequence, the local sequence environment of the SNP, the protein sequence profile, features derived from sequence alignment, as well as protein function, adding a degree of complexity to the functional analysis of SNPs.	http://snps-and-go.biocomp.unibo.it	Protein sequence and information about the substitutions.	The output provides the links to the sequence or structure given in input; and in the second part of the output, the protein sequence is visualized in a column and includes all the mutated residues. The table includes the mutated residue, the prediction (either "disease" or "neutral"), the reliability index (RI), the probability associated to the disease-related class, and the information about the prediction method. If the probability corresponding to disease-related is larger than 0.5, the variation is predicted as disease-related, and if lower than 0.5, is predicted as neutral. The RI scoring is evaluated from 0 (unreliable) to 10 (reliable).	Calabrese et al. (2009)
PhD-SNP (Predictor of Human Deleterious SNP) is a SVM-based classifier	Sequence/evolutionary conservation	Is based on a support vector machine (SVM-based classifier). The predictor based on a single SVM trained and tested on protein sequence and profile information.	http://snps.biofold.org/phd-snp/phd-snp.html	Protein sequence and position of the SNP.	The output consists of a table with the information for the wild-type residue, the new residue; and the related mutation is predicted as disease-related ("disease") or as neutral polymorphism ("neutral"). The reliability index (RI) value is evaluated from the output of the SVMSVM (O) as RI = 20*abs (O-0.5).	Capriotti et al. (2006)

(*Continued*)

Table 1. Continued

Computational methods	Type of analysis	Description	URL	Input	Output	References
Xvar (Mutation Assessor)	Sequence evolutionary conservation	Predictions are based on evolutionary conservation of the affected amino acid in protein homologs.	http://mutationassessor.org/	Protein sequence and variant information (chromosome position; reference allele, substituted allele).	A functional impact score is determined by a combination of a conservation score and a specificity score. Variants classified as "mild" or "low" are expected not to affect protein function, whereas variants classified as "medium" or "high" are estimated to result in altered function.	Reva et al. (2011)
PMUT (Prediction of Pathological Mutations on proteins)	Sequence and structure based approach	This tool is based in sequence alignment with structural factors to characterize missense substitutions using a feed-forward neural network.	http://mmb2.pcb.ub.es:8080/Pmut	Protein sequence and amino acid substitution of interest	Scoring mutation matrices are based on volume, solvent accessibility, hydrophobicity, and secondary structure characteristics, which are used to label mutations "neural network" and predict variant(s) as "neutral" or "pathological." Index scoring varies from 0 to 1.0 (high) and the reliability index (RI) from 0 to 9.	Ferrer-Costa et al. (2005)
Polyphen-2 (Polymorphism Phenotyping vs-2)	Protein and structure based/decision tree	Is based on eight sequenced-based and three structure-based predictive features The algorithm calculates a Bayes posterior probability that a given mutation is deleterious.	http://genetics.bwh.harvard.edu/pph2/	Protein sequence and amino acid substitution of interest.	Polyphen-2 gives a qualitative prediction. The probabilistic score HumDiv/HumVar varies from 0 to 1: a mutation is classified as "probably damaging" if score is above 0.85–1, "possibly damaging" if score is above 0.15–0.84 and the remaining mutations are classified as benign (0.00–0.14). Sensitivity and specificity varies from 0 to 1.	Adzhubei et al. (2010)
SNAP (screening for nonaceptable polymorphisms)	Sequence- and structure-based approach	Sequence, functional, and structural (secondary structure, conservation, solvent accessibility, etc.) annotations, and biophysical and evolutionary information to make predictions regarding the functionality of mutated proteins.	http://www.rostlab.org/services/SNAP	Protein sequence and amino acid substitution of interest.	The prediction of "neutral" or "nonneutral" is based on protein structural features. The output also consists in a reliability index and expected accuracy. SNAP will report only predictions with reliability over a set threshold (default = 0) and minimum expected accuracy (the tool will only report predictions with accuracy over a set threshold: default = 50%).	Bromberg and Rost (2007)
Hansa, prediction of neutral mutations	Sequence- and structure-based approach	Position-specific probabilities, local protein structural status, and the intrinsic properties of the wild-type and mutated residues. Hansa combines 10 different properties of these substitutions to partition disease and neutral mutations, and also data from PhD-SNP and Parepro algorithms.	http://hansa.cdfd.org.in:8080/	The user provides a database ID, GenBank, RefSeq, SWISSPROT, or PDB identifier for the query protein and amino acid substitution of interest.	This tool predicts the deleterious effects of a mutation using 10 neutral disease missense mutation discriminatory (NDMSMD) features, and classifies the mutation either as "disease" or "neutral."	Acharya and Nagarajara (2012)
MutPred, classify an amino acid substitution	Sequence- and structure-based approach	Classify amino acid substitutions (AAS) as disease associated or neutral in humans. In addition, it predicts molecular cause of disease/deleterious AAS. It indirectly exploits the structural and functional data available for functional prediction.	http://mutpred.mutdb.or/	Protein sequence and amino acid substitution of interest.	The output of MutPred contains a general score (g) that is the probability that the amino acid substitution is deleterious/disease associated, and top five property scores (p), where p predicts the impact of certain structural and functional properties. Scores with $g > 0.5$ and $p < 0.05$ are referred as "actionable hypotheses"; scores with $g > 0.75$ and $p < 0.05$ are referred as "confident hypotheses"; and scores with $g > 0.75$ and $p < 0.01$ are referred as "very confident hypotheses." Certain combinations of high values of general scores and low values of property scores are referred to as hypotheses.	Li et al. (2009)
FATHMM (v2.3) (Functional Analysis Through Hidden Markov Models)	Sequence- and structure-based approach	Predicts the functional effects of protein missense mutations by combining sequence conservation within HMMs, indicating the alignment of homologous sequences and conserved protein domains, with "pathogenicity weights" representing the overall tolerance of the protein/domain to mutations, which control the phenotype ontology of this method.	http://fathmm.biocompute.org.uk	Protein sequence and amino acid substitution of interest.	The molecular consequences of mutations are statistically inferred by mapping SUPERFAMILY domains onto the Gene Ontology, the Human Phenotype Ontology and the Mammalian Phenotype Ontology. Mutations are predicted as "damaging" and "tolerated." Scores range from less than zero ("preliminary") to greater than zero ("tolerated"); scores equal to zero indicate no significant changes in the underlying amino acid probabilities.	Shihab et al. (2013)

(*Continued*)

Table 1. Continued

Computational methods	Type of analysis	Description	URL	Input	Output	References
SNPeffect, molecular phenotyping of coding nsSNP	Sequence- and structure-based approach	Analyses of the consequences of SNPs on several functional properties: structural and thermodynamic properties affecting protein dynamics and stability, aggregation-prone regions, amylogenic regions, integrity of functional-binding sites, cellular processing, posttranslational modification, domain annotation, and cellular localization of proteins.	http://snpeffect.vib.be	Protein sequence and amino acid substitution of interest.	Variant analysis is performed using four tools: TANGO, WALTZ, LIMBO, and FoldX. SNPeffect output annotates the variants and calculates the effects produced by variants on known genes (e.g., amino acid changes).	Reumers et al. (2006); De Baets et al. (2012)
CONDEL (consensus deleteriousness [versions 1 and 2])	Sequence- and structure-based approach. Integrative tool	CONDEL assesses the outcome of nsSNVs using a CONsensus DELeteriounes score that combine several methods. The first version integrated the output of computational tools: Xvar, SIFT, and Polyphen. The new version also integrates the output of MAPP and FATHMM.	http://bg.upf.edu/condele	Protein sequence and amino acid substitution of interest.	CONDEL integrates the output of several predictive tools to calculate a weighted average of the scores (WAS) of these tools. In the second version, it also includes the weighted average of the normalized scores (WAS). CONDEL integrates the outputs of several tools to calculate their respective weighted average of the scores (WAS). The output score classifies the missense SNVs as "deleterious" or "neutral". Optionally, CONDEL can integrate a third tool, Mutation Assessor.	Gonzalez-Perez and Lopez-Bigs (2011); Hiltemann et al. (2014)
Meta-SNP, meta-predictor of disease-causing variants	Sequence- and structure-based approach Integrative tool	Meta-SNP is a random forest-based binary classifier to discriminate between disease-related and polymorphic nsSNVs integrates four existing methods: PANTHER, PhD-SNP, SIFT, and SNAP.	http://snps.biofold.org/meta-snp/index.html	Protein sequence and amino acid substitution of interest.	Consists of a table with the predictions of the four existing methods: PANTHER, PhD-SNP, SIFT, and SNAPs. The variant is classified as disease related ("disease") or polymorphic ("neutral"). Under the prediction of each method, the values returned by them are reported. The reliability index (RI) value is calculated from the Meta-SNP output as: RI = 20*abs (O-0.5).	Capriotti et al. (2013)

differences between the proportions of correct predictions obtained by the three tools executed with their native alignment [SIFT_Self; Polyphen-2_Self; PMUT_Self] and with the alignment of 27 orthologs (SIFT_Orth[27]; Polyphen-2_Orth[27]; PMUT_Orth[27]) for the same variants.

Results

This study compared the predictive ability of 16 Web-available tools (Table 1) to infer the functional effects of 48 *UGT1A1* nsSNPs already characterized at the protein level. To measure their performance (Table 2), we used four statistical measures, SEN, SPC, MCC, and ACC.

MutPred showed the best performance in almost all measures in the following order, SEN (97%), SPC (70%), MCC (0.73), and ACC (0.92). Higher SEN values were also obtained with PANTHER (96.7%), PMUT_Orth (94.7%), Hansa (92.1%), SNPs&GO (91.9%), SIFT_Orth (91.9%), and Meta-SNAP (92%). However, the ACC obtained for PANTHER (0.77), PMUT_Orth (0.81), Hansa (0.81), SNPs&GO (0.83), SIFT_Orth (0.89), and Meta-SNP (0.83) was lower than for MutPrep. Concerning specificity, higher values were obtained for SIFT_Orth (80%), followed by MutPrep SNAP and PMUT_self (70%), whereas the other methods showed lower values (Table 2). All of the applied predictors showed higher SEN (median value: 82.9%) than SPC (median value: 45.7%).

Considering MCC, which has been described as the best parameter to measure a predictor's performance [Johnson et al., 2005], the MutPred method presented the best value (0.73) and was followed by SIFT_Orth(27) (0.69), PhD-SNP (0.62), A-GVGD_Orth(9) (0.48), whereas SNPs&GO, SNAP, and Meta-SNP showed MCC values of 0.46. The tool with the worst performance was FATHMM, which had the lowest value for MCC (0.05). CONDEL (version 2), which integrates the output of Xvar, SIFT, Polyphen-2, MAPP, and FATHMM, achieved a MCC value of 0.05. Tools using worker's alignments as the input (SIFT, Polyphen-2, and PMUT) allow the use of another MSA for performance evaluation (Supp. Text S1). The predictive ability of these tools increased with the use of orthologous MSA, particularly PMUT that was the most positively affected. Significant differences between the proportions of correct predictions were obtained when Polyphen-2 and PMUT were run with orthologous MSA versus when they were executed with the automatically generated alignment (p = 0.021 and p = 0.007, respectively). SIFT was less influenced by the MSA employed (p = 0.063); however, all the metrics were improved when using the orthologous alignment (Table 2). SIFT has multiple options in its automated manner (MSA built by the tool); the best result, achieved by the "SIFT Sequence" option, used the UniRef90 dataset that produces the MSA and eliminates sequences with more than 90% identity to the query. The protein sequence alignment of 27 *UGT1A1* orthologs used as the input for SIFT, Polyphen-2, and PMUT in FASTA format is presented in Supp. Text S1. For the A-GVGD method, another alignment was required with few orthologous sequences. The best results were obtained using a MSA built with nine orthologs (including human *UGT1A1*), available as online supporting information (Supp. Text S2). The alignment of 27 orthologs is shown (Supp. Fig. S1) and corresponds to a section of the protein sequence acceptor-binding region (Supp. Fig. S1A), as well as to a fragment of the donor-binding site (Supp. Fig. S1B).

Phenotype analysis through SNPeffect did not give reliable structural information for the UGT1A1 protein to proceed with a FoldX stability analysis, which estimates the importance of the interactions that contribute to the stability of proteins and protein complexes.

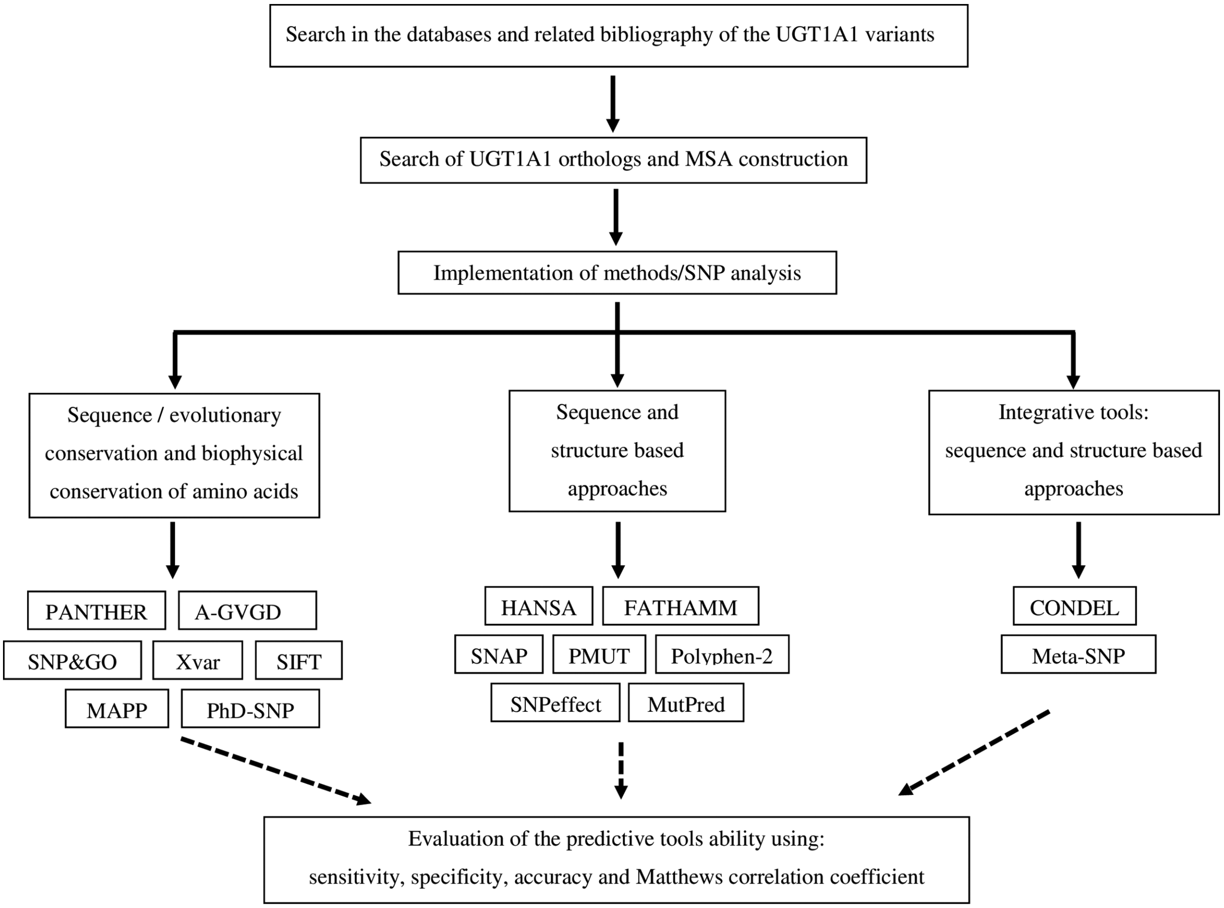
Search in the databases and related bibliography of the UGT1A1 variants
Search of UGT1A1 orthologs and MSA construction
Implementation of methods/SNP analysis
Sequence / evolutionary conservation and biophysical conservation of amino acids
Sequence and structure based approaches
Integrative tools: sequence and structure based approaches
PANTHER
A-GVGD
SNP&GO
Xvar
SIFT
MAPP
PhD-SNP
HANSA
FATHAMM
SNAP
PMUT
Polyphen-2
SNPeffect
MutPred
CONDEL
Meta-SNP
Evaluation of the predictive tools ability using: sensitivity, specificity, accuracy and Matthews correlation coefficient

Figure 1. Flow chart used to test 48 UGT1A1 nsSNPs by in silico analysis using 16 algorithms.

The SNPeffect tool allows the user to reduce the homology threshold, but this implies a loss in model accuracy. When the homology was set to 50%, we still did not retrieve further results. This tool was therefore excluded from the performing analysis because it failed to classify most of the UGT1A1 variants. Polyphen-2 and SIFT unsuccessfully generated predictions for two variants in our data set. The PANTHER Web tool failed to generate predictions for nine variants (out of 48). This may occur when the sequence alignment generated by the tool is poor or when the variant is located at a residue absent in the majority of the MSA column and, consequently, is unable to be modeled by the Human Markov Model [Thomas et al., 2003]. However, PANTHER was reliable when predictions were obtained. Using more evolutionarily distant orthologs, we found that the majority of variants were classified as neutral.

We also applied 16 tools to predict the phenotype of the 37 nsSNPs that were not characterized at a protein level. These were referred to as ND (Supp. Table S2A) but were associated with hyperbilirubinemia. Most tools classified some of the variants as neutral: six and seven for PhD-SNP; three for SIFT_Orth; seven for A-GVGD-7; and three for Hansa (Supp. Table S3).

Results from the tool Human Splicing Finder (3.0) for the 48 *UGT1A1*-characterized nsSNPs (Supp. Table S2A) and for the 37 nsSNPs not characterized are also presented (Supp. Table S2B). Most of the predicted variants were classified as "potential alteration for splicing," which involves the creation or alteration of an exonic splicing enhancer site, activation of an exonic cryptic donor site, or the creation of an exonic splicing silencer site, compared with those where no significant splicing motif alteration was detected.

Discussion

Determining the pathogenicity of an nsSNP could be an important step in the establishment of the genetic basis of its pathology, assessment of individual susceptibility to disease, understanding its pathogenesis, identification of molecular targets for drug treatment, and improvement in individual therapy. An increasing number of algorithms have been developed to predict the impact of missense mutations on protein function. In this context, we evaluated the performance of 16 available tools to predict the functionality of our selected *UGT1A1* nsSNP variants.

Comparison of Tool Performance Obtained in Other Studies

The majority of bioinformatics tools described in the literature are sequence- and structure-based approaches. MutPred, a sequence- and structure-based approach, was found to be the best performance tool to predict the phenotype of *UGT1A1* nsSNPs [Li et al., 2009]. This tool foresees the disruption of molecular function and works specifically for well-studied proteins when the homology and solved structure are available [De Baets et al., 2012]. It was expected that the lack of a 3D solved structure of the UGT1A1 protein could be a limitation for this analysis, but the establishment of MutPred's $g > 0.75$ cutoff to differentiate between benign and disrupted/pathogenic variants was found to accurately classify neutral variants from our data set, showing the higher specificity of this tool (70%). The other tools with a good performance were SIFT (Orth) and PhD-SNP.

Table 2. Sensitivity (SEN), Specificity (SPC), Matthew's Correlation Coefficients (MCC), and Accuracy (ACC) of the Selected SNP-Based Pathogenicity Detection Tools that Calculate the 48 *UGT1A1* snSNPs (*n* = 10 Neutral; *n* = 38 Deleterious) Whose Functional Impact on UGT1A1 Protein Was Already Evaluated by *In Vitro* and *In Vivo* Studies

Tools	SEN (%)	SPC (%)	MCC	ACC
PolyPhen 2_ Self	81.6	30.0	0.12	0.71
PolyPhen 2_ Orth (27)	88.9	40.0	0.24	0.75
SIFT_Self	86.5	44.4	0.31	0.78
SIFT_Orth (27)	91.9	80.0	0.69	0.80
A-GVGD_Orth (27)	65.0	67.0	0.25	0.65
A-GVGD_Orth (9)	88.0	60.0	0.48	0.82
PMUT_Self	47.4	70.0	0.14	0.52
PMUT_Orth (27)	94.7	30.0	0.33	0.81
Xvar	81.6	30.0	0.12	0.73
SNPs&GO	92.1	50.0	0.46	0.83
PhD-SNP	92.1	70.0	0.62	0.88
MAPP	79.0	50.0	0.26	0.73
CONDEL (v.1)	86.5	50.0	0.35	0.80
CONDEL (v.2)	84.0	20.0	0.05	0.71
Hansa	92.1	40.0	0.37	0.81
SNAP	81.6	70.0	0.46	0.79
MutPred	97.4	70.0	0.73	0.92
PANTHER	96.7	11.1	0.15	0.77
FATHMM	66.0	40.0	0.05	0.60
Meta-SNP	92.0	50	0.46	0.83

UGT1A1 gene: MIM #191740; HUGO: 12530.
SEN, sensitivity; SPC, specificity; MCC, Matthew's correlation coefficients; ACC, accuracy; AUC, area under the curve; _Self, run with native alignment (default settings); _Ortho (27), run with the orthologs alignment (MSA of 27 orthologs). _Ortho (9) runs with special alignment built for A-GVGD tool (MSA of nine orthologs).

These results are in accordance with those obtained in other studies, in which MutPred showed a good performance [Li et al., 2009; Thusberg et al., 2011]. Thusberg et al. (2011) evaluated eight tools that were also included in our study in a total dataset of 40,000 variants and concluded that no single method could be rated as the best for all parameters. Their study showed that the SNPs&GO and MutPred approaches reached the best performance for ACC and MCC (for SNPs&GO: MCC = 0.65; ACC = 0.82; and for MutPred: MCC = 0.63; ACC = 0.81). Two other studies evaluated predictor's performance using a set of *UGT1A1* nsSNPs and reached a lower performance [Di et al., 2009; Galehdari et al., 2013]. The study that analyzed SIFT and Polyphen performance (first version) encompassed other UGT1A1 variants but did not include neutral variants [Di et al., 2009]. For Polyphen, the correct prediction rate was 66.7% and for SIFT it was 57.1%; the prediction rate was significantly lower compared with that found in the present study, which was 88.9% for SIFT and 81.6% for Polyphen-2, respectively (data not shown). In another study, the performance of six SNP-based pathogenicity tools (SIFT, Polyphen-2, MutPred, PhD-SNP, Provean, and FATHMM) were tested with 59 *UGT1A1* nsSNPs associated with CNS [Galehdari et al., 2013]. The results showed the highest ACC values for SIFT (0.63), followed by MutPrep and Polyphen-2, both with 0.62, whereas for MCC the highest value was obtained with SIFT (0.34) followed by MutPrep and Polyphen-2 (0.30 for both). These are statistically lower values when compared with those obtained in the present study for MutPrep (MCC: 0.73; ACC: 0.92) and SIFT_Orth (MCC: 0.69; ACC: 0.89). The FATHMM method showed the lowest performance in our work as well as in Galehdari et al. (2013) study. These discrepancies could be related to the use of different criterion for variant selection. In our study, the selection of *UGT1A1* variants was based on a literature search. The selected variants were pathogenic and had already been validated by functional studies, and the neutral variants were described as being associated with normal bilirubin levels.

Performance of Algorithms That Run with User MSA

There are tools that generate MSAs internally and allow the user option to create and submit their own MSAs. Some SNPs predictors do not always perform optimally with their own program-generated MSA, and more accurate results could be achieved with gene-specific MSAs optimized by the user [Hicks et al., 2011]. Considering that orthologous sequences are more reliable in providing phylogenetic information, several groups have improved predictions by restricting the analysis to orthologs rather than paralogs [Shu et al., 2003; Tavtigian et al., 2006]. Orthologs are corresponding genes in different lineages and are a consequence of speciation, whereas paralogs result from gene duplication [Lynch et al., 2004]. A conserved position within a MSA may be due to evolutionary selection pressure that preserves protein functions, but it can also occur by chance. Increasing the orthologs added to a MSA leads to greater power in discriminating local substitutions as pathogenic and nonpathogenic [Wong et al., 2014]. So, the quality of MSA is a critical step because it is used to infer how an amino acid substitution is tolerated at a given position. The three tools, SIFT_Orth, Polyphen-2, and PMUT, allow great control of user-defined sequences in the alignment and flexibility when adding or removing sequences in the MSA; however, additional work is needed to obtain an alignment with the relevant protein sequences. There is also the potential to skew results by variations in the number and types of species included in the MSA. We verified that Polyphen-2 and PMUT were the most affected by the MSA employed; contrarily, the performance of SIFT was not significantly changed when it was run with the orthologous MSA.

We verified that most of the variants are classified as neutral by the A-GVGD tool when the alignments contain a large number of sequences. The presence of gaps in the vicinity of the alignment column leads to predictions toward neutral, as verified by Hicks et al. (2011). To improve the tool's performance, another alignment was required that contained few orthologous sequences (nine orthologs) and has a higher identity. Using this alignment, the results again classified most of the variants as neutral. Manual inspection of the alignment is recommended to ensure that predictions are as appropriate and accurate as possible.

The prediction Ability Can Be Affected by the Presence of Hypervariable or Constant Regions of UGT1A1 Orthologs Sequences

A reliable MSA of UGT1A1 orthologs is difficult to obtain due to the presence of hypervariable and constant regions observed in the UGT1A1 orthologs sequences. The *UGT1A1* gene is part of a complex locus that encodes several UDP-glucuronosyltransferases [Mackenzie, 1986]. The locus includes 13 unique alternate first exons, followed by four common exons. Each of the remaining nine 5' exons may be spliced to the four common exons, resulting in nine proteins with different N-terminals and identical C-terminals. Each first exon encodes the substrate binding site, which is regulated by its own promoter (acceptor-binding region; residues 26–291), and the other four exons (conserved region; residues: 292 and 490) encode donor-binding regions that contain the sugar-binding site [Mackenzie, 1986]. According to Li and Wu (2007), the donor-binding region is highly conserved, especially in the donor-interacting residues (the residues interacting with UDP glucuronic acid) and in the acceptor-binding region that presents four hypervariable regions among vertebrate UGTs. Classification of variants based on phylogenetic information assumes that pathogenic sites remain conserved and that nonpathogenic sites exhibit increased diversity. Some of the validated pathogenic variants located at the hypervariable region of the UGT1A1 protein are frequently classified as neutral due to the low conservation observed at those positions by sequence-based tools. Examples of these variants are p.G71R [Aono et al., 1995], p.F83L [Sutomo et al., 2002], and p.V225G [Costa et al., 2006], which are associated with GS and CNS. In contrast, neutral mutations are classified as deleterious when they are located at the constant region.

In the prediction analyses of the other 37 variants described in the literature as being associated with hyperbilirubinemia but that lacked functional studies (Table 2), we observed that some predictors with the best performance classified these variants as neutral. This can happen for several reasons: (1) all of these methods have an error associated with their prediction and none of these tools reach 100% accuracy; (2) most of these variants that were classified as neutral are located in the hypervariable region (p.M1V, p.Q6H, pH39D, p.W40R, p.V169E, p.Q185P) as observed for the first data set analyzed in this study; and (3) the variant could be classified as neutral and its association with hyperbilirubinemia may have resulted from an incorrect assumption. The establishment of a genotype–phenotype correlation for SNPs of the *UGT1A1* gene may be challenging due to the presence of other frequent mutations at the promoter region that could also be associated with elevated bilirubin levels. In fact, in a previous work [Rodrigues et al., 2012], we identified nine heterozygous SNPs by sequencing analysis of the coding regions of the *UGT1A1* gene. Three of these new variants were detected in GS patients (p.E180Q; p.M404T; p.R475C), four were detected in controls (p.I215V; p.M272V; p.V386I; p.I492T), and two were already described in the literature. Data showed that three of these new variants had been previously classified as benign, p.I215V, p.M272V, and p.V386I, and individuals presented total bilirubin concentrations of 5.1, 10.1, and 4.3 μmol/L, respectively. Three other variants were expected to have an effect on protein function, p.M404T, p.475C, and p.I492T, and individuals showed total bilirubin levels of 36.6, 51.3, and 5.8 μmol/L, respectively. However, in the presence of the two promoter polymorphisms, c.–41_–40dupTA and c.–3279T>G, higher total bilirubin levels were observed in GS patients and controls, both in heterozygotes and homozygotes [Rodrigues et al., 2012].

Possible Consequences on the Pre-mRNA Maturation and Processing of Some SNPs

In recent years, several bioinformatics tools have been developed to identify diseases that cause missense variants. Most of them only consider their impact at the protein functional level and do not take into account their effect on the pre-mRNA maturation and processing, something that is performed by specific predictive splicing tools. Several studies revealed that disease-causing missense mutations, which disrupt splicing, could be relatively higher than reported [Lopez-Bigas et al., 2005; Sterne-Weiler et al., 2011]. This could be due to the lack of a specific RNA to perform molecular and in vivo approaches. The 87 UGT1A1 variants analyzed in the present study by Human Splice Finder were classified as potential splicing disruption variants. Some of these variants were already characterized at the molecular level and have an impact on enzyme activity. This large number of missense variants, classified as putative splice variants, might be related to the regulation of gene expression of the UGT locus. The presence of new UGT1A1 proteins generated by alternative splicing of a further exon in the UGT1A1 locus was confirmed by immunofluorescence and coimmunoprecipitation assays using a specific anti-UGT1A1 antibody and generated a structural diversity of UGT1A1 proteins that are differentially expressed in several tissues [Lévesque et al., 2007]. These data suggest that the expression of UGT1A1 members is dependent on transcriptional regulatory mechanisms beyond those associated with tissue-specific expression [Lévesque et al., 2007]. Alternative splicing of exons may affect the fine balance of isoforms and, therefore, contribute to disease or to genetic modification of the disease phenotype and thus interfere with potential therapy [Garcia-Blanco et al., 2004]. However, these data will require further analysis.

Conclusion

The information provided by the in silico methods has the advantage of directing and complementing functional assays. The best performing methods obtained in this study for *UGT1A1* variants were MutPred and SIFT-Orth. The capacity of the SNP prediction methods can vary according to the available structural information and the MSA employed. We verified that methods primarily based on protein structural information such as SNPeffect failed to give reliable information for the *UGT1A1* variants. There are numerous available evolutionary tools that do not allow the user to select the alignments. One of the concepts that our data suggest is that the selection of user inputs improves the performance of these methods. Our data suggest that whenever possible, the user should consider optimizing the sequence alignment employed. The performance study of SNP predictors using a set of functionally well-characterized variants is essential to help redirect the in silico analysis of a particular gene.

Acknowledgments

The authors thank Dr. Adam Hospital, from the University of Barcelona (PMUT), for his help in analyzing our data and Dr. Ivan Adzhubey (Polyphen-2) from Harvard Medical School for providing the analysis with Standalone software on a local computer and guidance to analyze the results. Finally, we thank Professor Pedro Oliveira from the Biomedical Sciences Institute Abel Salazar, University of Porto, for help with the statistical analysis.

Disclosure statement: The authors declare no conflicts of interest.

References

Acharya V, Nagarajaram HÁ. 2012. Hansa: an automated method for discriminating disease and neutral human nsSNPs. Hum Mutat 33:332–337.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Aono S, Adachi Y, Uyama E, Yamada Y, Keino H, Nanno T, Koiwai O, Sato H. 1995. Analysis of genes for bilirubin UDP-glucuronosyltransferase in Gilbert's syndrome. Lancet 345:958–959.

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424.

Bosma PJ, Chowdhury JR, Bakker C, Gantla S, de Boer A, Oostra BA, Lindhout D, Tytgat GN, Jansen PL, Oude Elferink RP, Chowdhury NR. 1995. The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. N Engl J Med 333:1171–1175.

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acid Res 35:3823–3835.

Bromberg Y, Yachdav G, Rost B. 2008. SNAP predicts effect of mutations on protein function. Bioinformatics 24:2397–2398.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30:1237–1244.

Capriotti E, Altman RB, Bromberg Y. 2013. Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 14 Suppl 3:S2.

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734.

Ciotti M, Chen F, Rubaltelli FF, Owens IS. 1998. Coding defect and a TATA box mutation at the bilirubin UDP-glucuronosyltransferase gene cause Crigler–Najjar type I disease. Biochim Biophys Acta 1407:40–50.

Costa E. 2006. Hematologically important mutations: bilirubin UDP-glucuronosyltransferase gene mutations in Gilbert and Crigler–Najjar syndromes. Blood Cells Mol Dis 36:77–80.

Costa E, Vieira E, Martins M, Saraiva J, Cancela E, Costa M, Bauerle R, Freitas T, Carvalho JR, Santos-Silva A, Barbot J, Dos Santos R. 2006. Analysis of the UDP-glucuronosyltransferase gene in Portuguese patients with a clinical diagnosis of Gilbert and Crigler–Najjar syndromes. Blood Cells Mol Dis 36:91–97.

De Baets G, Van Durme J, Reumers J, Maurer-Stroh S, Vanhee P, Dopazo J, Schymkowitz J, Rousseau F. 2012. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res 40:D935–D939.

Deiss A. 1999. Destruction of erythrocytes. In: *Wintrobe's clinical haematology*. Vol. 10. Baltimore , MD: Lippincott Williams & Wilkins. p 267–299.

Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. 2009. Human splicing finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

Di YM, Chan E, Wei MQ, Liu JP, Zhou SF. 2009. Prediction of deleterious non-synonymous single-nucleotide polymorphisms of human uridine diphosphate glucuronosyltransferase genes. AAPS J 11:469–480.

Farheen S, Sengupta S, Santra A, Pal S, Dhali GK, Chakravorty M, Majumder PP, Chowdhury A. 2006. 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 12:2269–2275.

Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21:3176–3178.

Ferrer-Costa C, Orozco M, de la Cruz X. 2004. Sequence-based prediction of pathological mutations. Proteins 57:811–819.

Galehdari H, Saki N, Mohammadi-Asl J, Rahim F. 2013. Meta-analysis diagnostic accuracy of SNP-based pathogenicity detection tools: a case of UTG1A1 gene mutations. Int J Mol Epidemiol Genet 4:774–785.

Garcia-Blanco MA, Baraniak AP, Lasda EL. 2004. Alternative splicing in disease and therapy. Nat Biotechnol 22:535–546.

Gonzalez-Perez A, Lopez-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88:440–449.

Hicks S, Plon SE, Kimmel M. 2013. Statistical analysis of missense mutation classifiers. Hum Mutat 34:405–406.

Hicks S, Wheeler DA, Plon SE, Kimmel M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32:661–668.

Hiltemann S, Mei H, deHollander M., Palli I, Spek P, Guido J, Andrew S. 2014. CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy. Gigascience 3:1–6.

Huang CS, Luo GA, Huang ML, Yu SC, Yang SS. 2000. Variations of the bilirubin uridine-diphosphoglucuronosyl transferase 1A1 gene in healthy Taiwanese. Pharmacogenetics 10:539–544.

Johnson MM, Houck J, Chen C. 2005. Screening for Deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response. Cancer Epidemiol Biomarkers Prev 14:1326–1329.

Jordan DM, Ramensky VE, Sunyaev SR. 2010. Human allelic variation: perspective from protein function, structure, and evolution. Curr Opin Struct Biol 20:342–350.

Kaniwa N, Kurose K, Jinno H, Tanaka-Kagawa T, Saito Y, Saeki M, Sawada J, Tohkin M, Hasegawa R. 2005. 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 33:458–465.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. 2005. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21:2814–2820.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. Mol Biol Evol 30:772–780.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081.

Labrune P, Myara A, Chalas J, Le Bihan B, Capel L, Francoual J. 2002. 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 20:399–401.

Labrune P, Myara A, Hadchouel M, Ronchi F, Bernard O, Trivin F, Chowdhury NR, Chowdhury JR, Munnich A, Odievre M. 1994. Genetic heterogeneity of Crigler–Najjar syndrome type I: a study of 14 cases. Hum Genet 94:693–697.

Lévesque E, Girard H, Journault K, Lépine J, Guillemette C. 2007. Regulation of the UGT1A1 bilirubin-conjugating pathway: role of a new splicing event at the UGT1A locus. Hepatology 45:128–138.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750.

Li C, Wu Q. 2007. Adaptive evolution of multiple-variable exons and structural diversity of drug-metabolizing enzymes. BMC Evol Biol 7:69–88.

Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. 2005. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett 579:1900–1903.

Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. Trends Genet 20:544–549.

Mackenzie PI. 1986. Rat liver UDP-glucuronosyltransferase. Sequence and expression of a cDNA encoding a phenobarbital-inducible form. J Biol Chem 261:6119–6125.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc 8:1551–1566.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80.

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. 2006. SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics 22:2183–2185.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39:e118.

Rodrigues C, Vieira E, Santos R, de Carvalho J, Santos-Silva A, Costa E, Bronze-da-Rocha E. 2012. Impact of UGT1A1 gene variants on total bilirubin levels in Gilbert syndrome patients and in healthy subjects. Blood Cells Mol Dis 48:166–172.

Sai K, Saeki M, Saito Y, Ozawa S, Katori N, Jinno H, Hasegawa R, Kaniwa N, Sawada J, Komamura K, Ueno K, Kamakura S, et al. 2004. UGT1A1 haplotypes associated with reduced glucuronidation and increased serum bilirubin in irinotecan-administered Japanese patients with cancer. Clin Pharmacol Ther 75:501–515.

Seppen J, Bosma PJ, Goldhoorn BG, Bakker CT, Chowdhury JR, Chowdhury NR, Jansen PL, Oude Elferink RP. 1994. Discrimination between Crigler–Najjar type I and II by expression of mutant bilirubin uridine diphosphate-glucuronosyltransferase. J Clin Invest 94:2385–2391.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 34:57–65.

Shu Y, Leabman MK, Feng B, Mangravite LM, Huang CC, Stryke D, Kawamoto M, Johns SJ, DeYoung J, Carlson E, Ferrin TE, Herskowitz I, Giacomini KM; Pharmacogenetics Of Membrane Transporters Investigators. 2003. Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1. Proc Natl Acad Sci USA 100:5902–5907.

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40 (Web Server issue):W452–W457.

Sneitz N, Bakker CT, de Knegt RJ, Halley DJ, Finel M, Bosma PJ. 2010. Crigler–Najjar syndrome in The Netherlands: identification of four novel UGT1A1 alleles, genotype–phenotype correlation, and functional analysis of 10 missense mutants. Hum Mutat 31:52–59.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR: 2011. Loss of exon identity is a common mechanism of human inherited disease. Genome Res 21:1563–1571.

Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15:978–986.

Strassburg CP, Lankisch TO, Manns MP, Ehmer U. 2008. Family 1 uridine-5'-diphosphate glucuronosyltransferases (UGT1A): from Gilbert's syndrome to genetic organization and variability. Arch Toxicol 82:415–433.

Sutomo R, Laosombat V, Sadewa AH, Yokoyama N, Nakamura H, Matsuo M, Nishio H. 2002. Novel missense mutation of the UGT1A1 gene in Thai siblings with Gilbert's syndrome. Pediatr Int 44:427–432.

Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43:295–305.

Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S, Vandergriff JA, Doremieux O. 2003. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res 31:334–341.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32:358–368.

Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics 13 Suppl 4:S2.

Vihinen M. 2013. Guidelines for reporting and using prediction tools for genetic variation analysis. Hum Mutat 34:275–282.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat 17:263–270.

Wong KC, Zhang Z. 2014. SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences. Bioinformatics 30: 112–119.

Yusoff S, Van Rostenberghe H, Yusoff NM, Talib NA, Ramli N, Ismail NZ, Ismail WP, Matsuo M, Nishio H. 2006. Frequencies of A(TA)7TAA, G71R, and G493R mutations of the UGT1A1 gene in the Malaysian population. Biol Neonate 89:171–176.