Computational Tools to Assess the Functional Consequences of Rare and Noncoding Pharmacogenetic Variability

Yitian Zhou¹ and Volker M. Lauschke¹,*

Interindividual differences in drug response are a common concern in both drug development and across layers of care. While genetics clearly influences drug response and toxicity of many drugs, a substantial fraction of the heritable pharmacological and toxicological variability remains unexplained by known genetic polymorphisms. In recent years, population-scale sequencing projects have unveiled tens of thousands of coding and noncoding pharmacogenetic variants with unclear functional effects that might explain at least part of this missing heritability. However, translating these personalized variant signatures into drug response predictions and actionable advice remains challenging and constitutes one of the most important frontiers of contemporary pharmacogenomics.

Conventional prediction methods are primarily based on evolutionary conservation, which drastically reduces their predictive accuracy when applied to poorly conserved pharmacogenes. Here, we review the current state-of-the-art of computational variant effect predictors across variant classes and critically discuss their utility for pharmacogenomics. Besides missense variants, we discuss recent progress in the evaluation of synonymous, splice, and noncoding variations. Furthermore, we discuss emerging possibilities to assess haplotypes and structural variations. We advocate for the development of algorithms trained on pharmacogenomic instead of pathogenic data sets to improve the predictive accuracy in order to facilitate the utilization of next-generation sequencing data for personalized clinical decision support and precision pharmacogenomics.

Interindividual differences in drug response are common, resulting in insufficient drug efficacy or adverse drug reactions in 25–50% of patients.¹ Besides environmental, physiological, and demographic factors, genetic variation is estimated to explain 20–30% of this variability.² Combined, these estimates suggest that between 1 in 20 and 1 in 6 unintended pharmacological consequences could be predicted based on genetic factors. Most pharmacogenomic research has focused on the functional consequences of common genetic variants. Among the associations with highest clinical relevance are correlations between reduced function DPYD alleles and fluoropyrimide toxicity, thiopurine S-methyltransferase (TPMT) deficiency and myelosuppression due to thiopurines, and CYP2C19 genotypes with response to antidepressants.³ Furthermore, fueled by the rapid advances in next-generation sequencing (NGS) technologies and the increasing availability of genomic information on a population scale, it is becoming increasingly appreciated that rare genetic variants also can have important impacts on drug response.⁴ ⁵

Profiling of the genetic variability of 130,000 individuals across 208 pharmacogenes revealed almost 70,000 single nucleotide variants (SNVs), and each individual was found to carry ~144 missense, 151 synonymous, and 12 nonsense pharmacogenomic variants (Figure 1 and ref. 6). Notably, 98.5% of these variants are rare variants with minor allele frequency (MAF) less than 1%, and these rare variants are estimated to account for 30–40% of the genetically encoded functional variability in drug response.⁷ In addition to SNVs, 97% of pharmacogenes studied harbor copy number variations, and for 85 of these genes, copy number variations account for >5% of all loss-of-function alleles.⁸ Particularly, cytochrome P450s (CYPs) and multidrug resistance (MDR) transporters of the adenosine triphosphate (ATP)-binding cassette (ABC) superfamily are highly variable, resulting in substantial interindividual variability in gene activity.⁹ ¹¹

A multitude of commercial pharmacogenetic testing solutions have become available. The provided reports include functional predictions and inferred phenotypic consequences based on internal data and proprietary algorithms. Most suppliers focus on testing and interpretation of well-established variations in CYP2C19, CYP2D6, CYP2C9, VKORC1, and the HLA locus. However, more extensive sequencing-based analyses that include rare and novel variants are also available. Commercial pharmacogenetic testing plays an increasingly important role in clinical decision support by providing accessible user interfaces and consolidated recommendation summaries. Furthermore, such providers take care of data security and privacy. Notably however, the processes underlying the generation of the result report are nontransparent, reporting and testing standards are heterogeneous, and interpretations can at times be discrepant between providers.¹² ¹³

The prevalence of rare pharmacogenetic variants has by now been well established; however, their functional interpretation

¹Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden. *Correspondence: Volker M. Lauschke (volker.lauschke@ki.se)
Received April 1, 2021; accepted May 7, 2021. doi:10.1002/cpt.2289

626 CLINICAL PHARMACOLOGY & THERAPEUTICS | VOLUME 110 NUMBER 3 | September 2021
remains challenging. While common candidate variants can be interrogated experimentally, mostly using heterologous expression systems, these methods are not sufficiently scalable for the analysis of tens of thousands of variants. Similarly, the functional impact of common variants can be analyzed using epidemiological studies in which the presence of a variant is associated with a given pharmacological or toxicological end point, such as drug pharmacokinetics, incidence of adverse drug reactions, or patient outcomes. Yet, rare-variant association studies are challenging and often not feasible due to the vast number of individuals that would need to be analyzed to draw statistically meaningful conclusions. Thus, in the absence of feasible experimental or epidemiological alternatives, the use of computational tools constitutes the preferred strategy to rapidly interpret the functional impact of novel or rare pharmacogenetic variants.

So far, a multitude of computational algorithms have been developed that aim to distinguish deleterious from neutral variations. Yet, most prediction algorithms have been devised for the detection of disease-causing variants, and their application to pharmacogenetics is not straightforward. Specifically, most prediction methods use variants with known association to genetic disease as their “deleterious training set,” whereas their “neutral training sets” comprise variations that are common in the general population and are thus unlikely to be pathogenic. As a consequence of this input data selection, evolutionary conservation has emerged as the best metric for these methods to predict variant functionality due its strong association with variant pathogenicity. Importantly however, pharmacogenes that lack endogenous functions are rarely associated with disease and have overall low evolutionary constraints. Moreover, a multitude of pharmacogenetic variations with clear functional consequences, such as rs4244285 (CYP2C19*2), rs35742686 (CYP2D6*3), and rs3892097 (CYP2D6*4) are common in the general population, resulting in their inclusion in the neutral rather than deleterious training data sets.

These differences between disease-associated genes and pharmacogenes illustrated above can pose problems for the functional interpretation of pharmacogenetic variants using conventional tools. In this review, we provide an overview of available in silico prediction methods and critically discuss their suitability for pharmacogenetic applications. Furthermore, we highlight recent developments in extending functionality predictions from missense variants to also include synonymous variants, as well as noncoding variations in untranslated regions or regulatory elements. We conclude that the performance of computational prediction methods has experienced drastic improvements over the last years and recent data indicates that current algorithms achieve accuracies similar to conventional in vitro tests. These data incentivize the design of clinical trials in which the cost-effectiveness and patient benefits of NGS coupled with computer-assisted data interpretation can be tested in a prospective setting.

**FUNCTIONAL INTERPRETATION OF MISSENSE VARIANTS**

**General algorithms**

Missense variants, i.e., variants that cause amino acid substitutions, constitute the most common class of exonic variants. They can alter gene product functions through multiple mechanisms, including alteration of active sites, disruption of protein folding, and modification of posttranslational processes. So far, more than four million missense variants have been observed in the human genome, and each individual harbors around 144 missense variants in pharmacogenes alone.6,15 Due to their substantial number, missense variants have been the main focus of computational prediction tools. The majority of these tools estimate functional impact based on (i) features of the corresponding genetic or amino acid sequence using primarily evolutionary conservation, (ii) physicochemical properties of the amino acid exchange, or (iii) information related to the structure of the gene product. In this section, we provide a critical overview of the concept, principles, and recent progress of missense variant prediction tools and discuss their utility specifically for the assessment of pharmacogenetic variants.

As illustrated above, when the functional prediction of missense variants is based on sequence features, the assessment primarily relies on evolutionary conservation, a phylogenetic metric that evaluates the variability of the sequence in question across taxa (Table 1). The rationale behind this approach is that conserved sequences are under evolutionary constraint and, thus, mutations in such sequences are more likely to be deleterious. Evolutionary conservation is utilized by most prediction algorithms; some use alignments of the amino acid sequence, while others use genetic sequence alignments or a combination of both. The former is naturally more efficient but only applicable to missense variations, whereas the latter allows the interrogation of the entire human genome and can provide functional inference also for noncoding variations.

On the basis of evolutionary conservation, prediction tools were established either through methods that base their classifications on scores calculated with mathematical models defined a priori (sometimes called “direct models”), or through various machine-learning methods (or “artificial intelligence models”), most commonly random forests, hidden Markov models, support vector machines, or deep neural networks. With the thriving of computational science in recent years and rich genetic databases brought by NGS technologies, machine learning solutions have largely superseded direct models. In total, over 30 prediction tools focusing specifically on missense variants have been presented to date, of which SIFT,
Table 1 Selection of computational tools that predict functions of different variants types

| Variant type | Computational tools | Basis of decision | Sources of training/validation data | Availability | Ref. |
|--------------|---------------------|-------------------|------------------------------------|--------------|------|
| Missense     | SIFT                | Sequence conservation | Literature (LacI, HIV-1 Protease and Bacteriophage T4 Lysozyme) | Web server available | 17   |
|              | PolyPhen2           | Eight sequence-based features and three structure-based features | Uniprot database | Web server available | 18   |
|              | PROVEAN             | Similarity of homologous amino acid sequences | Uniprot/Swiss-Prot database and the literature | Web server available | 19   |
|              | VEST3               | Prioritization of variants that underlie Mendelian disease | Positive: HGMD; Neutral: ESP | Web server available | 100  |
|              | REVEL               | 18 pathogenicity scores from 13 prediction tools | Positive: HGMD; Neutral: ESP | Pre-predicted data set available | 21   |
| Synonymous   | SILVA               | Conservation, codon usage, sequence features, splicing-related motifs and pre-mRNA folding energy | Positive: Pathogenic synonymous variants from the literature; Neutral: 1KGP | Bioinformatic script available | 41   |
|              | DDIG-SN             | 54 features based on DNA, RNA, and protein level | Positive: HGMD; Neutral: 1KGP | Web server available | 42   |
|              | IDSV                | Splicing, conservation, codon usage, sequence features, translation efficiency, RNA folding and functional region annotation | Positive: ClinVar and pathogenic synonymous variants from the literature; Neutral: VariSNP | Bioinformatic script available | 43   |
| Splice, frameshift, and stop-gain | LOFTEE              | Preset rules for variant position and conservation | None | Bioinformatic script available | 50   |
| Frameshift and stop-gain | ENTPRISE-X           | Amino acid sequence features, predicted protein 3D structures and protein function | Positive: ClinVar; Neutral: ESP and 1KGP | Web server currently inaccessible | 52   |
| Frameshift   | DDIG-in             | Variant position, DNA conservation score, evolution feature, variant length, alignment score, disorder score, secondary structure, and accessible surface area | Positive: HGMD; Neutral: 1KGP | Web server available | 51   |
| Splice       | MMSplice            | Combination of features extracted from six modules | Positive: GENCODE; Neutral: random sequence for donor and acceptor modules, ClinVar and others for other modules | Bioinformatic script available | 58   |
| SpliceAl     | Blockwise features obtained from 40, 200, 1,000, and 5,000 nucleotides | GENCODE | Bioinformatic script available | 60   |
| Regulatory   | gkm-SVM             | Cell type relevance and features extracted from both positive and negative sequences | Positive: putative regulatory sequences sensitive to DNase I; Neutral: random genomic sequences | Bioinformatic script downloadable | 71   |
| GenoCanyon   | Conservation and biochemical signal | UCSC genome browser and NCBI GWAS catalog | Pre-predicted data set available | 70   |
| Haplotype    | Hubble.2D6          | Features based on two models to predict activity score and measured metabolic activity from simulated sequences and real-life data | PharmVar | Bioinformatic script available | 82   |
| VarCoPP      | 11 features at the variant, gene and gene-pair level | Positive: DIDA; Neutral: 1KGP | Web server available | 84   |
| Structural   | SVScore             | Gene, exon features and aggregation of per-base pathogenicity score | Finnish WGS data sets, ClinGen and 1KGP | Bioinformatic script available | 92   |
| SVFX         | Somatic and germline models using genomic and tissue-specific epigenomic features | Positive: PCAWG, ClinVar; Neutral: 1KGP, GnomAD and ClinVar | Bioinformatic script available | 93   |

(Continued)
PolyPhen-2, PROVEAN, VEST3, and REVEL are arguably the most extensively used (Table 1). Furthermore, missense variant effects are often analyzed using non–class-specific algorithms, such as CADD, Eigen, and FATHMM that allow functional predictions across variant categories.

Importantly, while there is an extensive repertoire of prediction methods and models to choose from, which tool achieves the best performance remains an area of ongoing debate. For instance, a comparison of 23 prediction tools using three independent pathogenic variant data sets indicated that REVEL and VEST3 achieved the highest predictive accuracy, outperforming the other tested methods, including SIFT, PolyPhen-2, and CADD. However, substantial variability in performance between benchmark sets was observed. Furthermore, while all predictors claimed high predictive accuracy with an area under receiver operating characteristic curve (AUC$_{ROC}$) pivoting around 0.9, drastic drops in performance when applied to independent variant cohorts are common. This phenomenon is explained, at least in part, by the use of training data that overlap across algorithms, resulting in overfitting and hindering the comparative evaluation of these tools.

### Gene-specific, class-specific, or feature-specific algorithms

To improve the overall utility of functional prediction tools, new algorithms have been developed that are specifically geared toward certain genes. For instance, BRCA-ML constitutes an ensemble machine-learning method that can predict the functional impact and implications for breast cancer risk for missense variants in BRCA1 and BRCA2. Similar gene-specific methods were developed for the assessment of variants in DYPD. In addition to gene-specific tools, algorithms were developed for entire gene families. Recently, machine-learning models based on sequence and structural features were developed to identify both loss-of-function (LOF) and gain-of-function (GOF) missense variants specifically in genes encoding voltage-gated sodium and calcium channels of the SCN$n$xA and CACNA1$x$ gene families that have been associated with multiple neurological and neurodevelopmental diseases.

As discussed above, training and development of these methods rely primarily on the assessment of evolutionary conservation and, accordingly, deleteriousness of a given variant, which most commonly coincides with reduced gene activity. Only a few tools have been developed to analyze whether a variant might result in increased function. While such hyperactivating variations are most relevant in cancer, few examples in pharmacogenes have been identified. Missense variants that result in increased activity, such as $CYP2B6^{*4}$ (p.K262R), $DPYD$ p.P1023T, and $CYP2D6^{*53}$ (p.F120I and p.A122S), can be analyzed using BSIFT, an extension of the conventional SIFT algorithm that can identify both deleterious and activating variant given a query mutation within a protein coding sequence. While this tool has not been tested on pharmacogenetic increased-function variants, it can be questioned whether it will be suitable for such variants, as this method also relies on evolutionary metrics. No dedicated tools have to our knowledge been presented for the analysis of activating mutations in noncoding regions, such as $CYP2C19^{*17}$ and $CYP2B6^{*22}$.

Besides algorithms for specific genes, gene families, or functional classes, methods have been developed that are specifically useful for variants with other features. For instance, by comparing observed allele frequency in a patient cohort with the expected allele frequency derived from disease prevalence and frequency in the general population, a binomial test was developed that can filter extremely rare benign variants to aid Mendelian disease analysis.

We recently developed an optimized prediction framework specifically developed for the functional analysis of pharmacogenetic variants in lowly conserved genes (APF). This framework was trained on 337 variants across 44 pharmacogenes (mostly $CYP$s) with high-quality, quantitative functional data. The resulting method provided accurate and, importantly, quantitative estimation of the impacts of variations on enzyme activity and achieved 93% sensitivity and specificity for LOF and functionally neutral
pharmacogenetic variants, respectively. Importantly, APF showed better performance in predicting rare (MAF < 1%) and rare (MAF < 0.1%) pharmacogenetic variants, thus facilitating the functional interpretation of the vast number of rare pharmacogenetic variants discovered in NGS projects. Furthermore, APF achieved similar predictive accuracy as in vitro assessments and outperformed other prediction tools for the evaluation of DPYD and TPMT variants.35

In conclusion, it is becoming increasingly clear that specialized computational methods improve the accuracy of pharmacogenetic variant effect predictions, particularly for rare variants in genes with low evolutionary constraint (Figure 2). We expect that developments in the near future will continue in this direction and further refine algorithms that focus on the evaluation of specific gene or variant properties rather than universal prediction methods.

INTERPRETING SYNONYMOUS VARIANTS

Synonymous variants, i.e., variants in the coding region that do not affect the amino acid sequence, can alter the functionality of gene products by interrupting transcriptional processes, modulating splicing, altering messenger RNA structure stability, and affecting the rate of translation.36–38 Furthermore, synonymous variants are as likely to be pathogenic as nonsynonymous variants.39 Notably however, the majority of prediction tools commonly used for coding variants are not applicable to the interpretation of synonymous variations.

Synonymous variants are most commonly predicted using algorithms for the genome-wide prediction of pathogenic variation, such as CADD,22 FAT H M M ,24,25 and TRAP.40 Furthermore, algorithms are available that are focused specifically on the prediction of synonymous variants, including SILVA,41 DDIG-SN,42 and IDSV43 (Table 1). However, as for missense variants, these tools are trained on pathogenic data, which were limited in size and not sufficiently curated.44 For instance, SILVA was trained on only 33 experimentally determined deleterious variants and 785 putatively neutral variants from the 1000 Genome Project, which is too few for accurate model training. In contrast, DDIG-SN used all pathogenic synonymous variants from the Human Gene Mutation Database (HGMD); however, the reliability of annotations in this database remains questionable.45 To address these issues, a curated synonymous variant data set was developed by including only recent pathogenic HGMD entries and matching them with proximal putatively benign synonymous variants from the VariSNP database.46 The authors found that an ensemble score of TRAP, SILVA, and FATHMM, achieved the highest performance in this balanced data set.

So far, only a few synonymous variants were identified with functional impact on drug response and toxicity. The most extensively studied example is a haplotype consisting of two synonymous variants (rs1128503 and rs1045642 corresponding to p.G412G and p.I1145I) in ABCBI (encoding Pgp/MDR1) that disrupts cotranslational folding of the transporter.47 The variants are further in high linkage disequilibrium with the putatively neutral triallelic missense variant rs2032582 (p.S893A/T), and the haplotype has been associated with toxicity and response to taxanes in breast cancer patients.48 Furthermore, synonymous haplotypes in COMT affect translation,49 thereby potentially altering the metabolism of COMT substrates, such as levodopa and methyldopa. Due to the limited number of synonymous variants with confirmed effects on gene activity, benchmarking of the available methods for pharmacogenetic applications remains difficult. Thus, further expansion of the available database, e.g., by comprehensive experimental characterizations, might provide an important step forward to refine synonymous variant predictions.

NONSENSE, FRAMESHIFT, AND SPLICE VARIANTS

Nonsense variants are defined as variants that result in the premature termination of the polypeptide and comprise frameshift, splice, and stop-gain variants. While it is intuitive that such nonsense variants should be classified as functionally deleterious, not all nonsense variants abolish gene function. The LOFTEE algorithm implements a set of simple rules to filter functionally neutral frameshift, splice, and stop-gain variants.50 For instance, these filters flag variants as low confidence LOF if they locate near the end of the transcript, where the ancestral allele presumptively abolishes gene function or where splicing only affects untranslated regions. Furthermore, a multitude of algorithms is available for the functional assessment of individual variant classes. For frameshift variants, variant pathogenicity can be predicted by DDIG-in,51 ENTPRISE-X,52 SIFT-indel53 and VEST-indel54 (Table 1). As for missense variants, these methods are trained on pathogenic variant sets from HGMD or ClinVar, using common variations from population-scale sequencing project as benign reference data.

To predict the effect of variants on splicing, specific algorithms are available for the evaluation of variants inside or outside of splice sites. To evaluate the effect of variants within splice regions,
MaxEntScan\textsuperscript{55} and SpliceScan\textsuperscript{56} constitute the most commonly used tools, whereas CryptSplice\textsuperscript{57} and MMSplice\textsuperscript{58} can evaluate effects on the formation of cryptic splice sites, exon skipping, splice site selection, and splicing efficiency (Table 1). For methodological details of these approaches, we refer readers to an excellent recent review on this topic.\textsuperscript{59} Recently, SpliceAI was developed as a method to predict the effect of genomic variants on splicing events using data from patients with rare diseases, and the predicted splice sites were found to be strongly deleterious across human populations.\textsuperscript{60} Application of SpliceAI to common pharmacogenomic splice variants resulted in the successful identification of \textit{CYP2C19}\textsuperscript{7,2}, \textit{CYP2D6}\textsuperscript{4}, and \textit{DPYD}\textsuperscript{2A}. In contrast, \textit{CYP3A5}\textsuperscript{3} was not flagged by either SpliceAI or MaxEntScan, suggesting that prediction of deep intronic variants remains challenging.

Frameshift, splice, and stop-gain variants account for a considerable fraction of the genetically encoded functional variability in pharmacogenes. As splice prediction methods are not based on evolutionary conservation but rather on biochemical principles and consensus motifs, it is not surprising that nonsense prediction seems to perform equally well on pharmacogenetic variants compared with pathogenic data. In contrast, all current frameshift prediction methods are trained on disease data, calling their transferability to poorly conserved pharmacogenes into question. Of note, our analysis included only four commonly studied variants and more extensive research on larger variant sets, including parameter optimizations, are necessary to draw more firm conclusions.

**NONCODING VARIANTS**

Genomic regions that do not code for proteins account for >99% of the human genome, and noncoding variants are expected to contribute substantially to the interindividual variability expression levels, as well as spatial and temporal gene regulation.\textsuperscript{51,62}

Despite their drastically higher abundance, noncoding variants are substantially understudied compared with coding variations, primarily because of their context dependency. In contrast to missense variants, effects of noncoding variants commonly only manifest in specific cell types, developmental stages, or diseases, which renders their systematic study and experimental validations difficult. Noncoding variation often interacts with transcription factors or other regulatory factors, thus only causing effects in the presence of the respective binding partner. Moreover, noncoding variants can have indirect effects; for instance, they frequently reside in long noncoding RNAs, which in turn regulate the expression of coding transcripts.\textsuperscript{63}

As for other variant types, the first approaches to flag noncoding variants of functional relevance were based on evolutionary conservation scores, such as GERP++\textsuperscript{64}, PhastCons\textsuperscript{65}, and PhyloP.\textsuperscript{66} Importantly however, while transcription factor binding might be conserved across taxa, the nucleotide sequence of the corresponding binding site is often highly variable and, as a consequence, evolutionary conservation can only serve as a weak proxy for the functional importance of variants in regulatory elements.\textsuperscript{67,68}

To improve the predictive power, the next generation of noncoding prediction algorithms included functional genomic data, including chromatin accessibility, histone modifications, and transcription factor binding data. Widely used examples include CADD,\textsuperscript{52} FATHMM,\textsuperscript{24,25} GWAVA,\textsuperscript{69} and GenoCanyon\textsuperscript{70} (Table 1). Importantly however, these tools are based on extensive population-scale data sets, which fail to capture context specificity.

To overcome these limitations, algorithms were developed that predict regulatory elements and their strength of interaction with transcription factors based only on DNA sequence as input. The first method for this strategy, gkm-SVM, was trained on cell type-specific enhancer sequences with length-matched, genomic DNA base composition content–matched, and repeat-matched random sequences as negative training data sets.\textsuperscript{71} Notably, variant-induced changes in gkm-SVM score accurately predicted functional regulatory variants, indicating that modulation of local transcription factor–DNA interactions underlies the pathogenicity of noncoding variations.\textsuperscript{72} Other notable approaches that predict transcription factor binding based on DNA sequence include DeepSEA\textsuperscript{73} DeepBind,\textsuperscript{74} Sasquatch,\textsuperscript{75} Basenji,\textsuperscript{76} ExPecto,\textsuperscript{6} and DeFine.\textsuperscript{77}

Inversely, colocalization of expression quantitative trait loci and data from genome-wide association studies allows for identifying causal variants, which provides an interesting orthogonal approach to highlight variants with tissue-specific impacts on gene expression.\textsuperscript{78} These tools allow researchers for the first time to predict gene expression changes from hundreds of human cell types based on primary sequence information alone, thus opening up promising possibilities to identify causal variants underlying interindividual differences in expression levels.

Current approaches for the functional prediction of noncoding variation are mechanism-based and, as a result, it can be expected that their performance will not decrease substantially when applied to pharmacogenomic variation. However, of note, \textit{CYP2C19}\textsuperscript{17}, arguably the most clinically relevant noncoding pharmacogenetic variant, was not identified as functionally relevant by any of the aforementioned algorithms for noncoding variant prediction, providing anecdotal evidence that the computational assessment of such variants remains challenging. Further data in this regard are limited and thus incentivize the systematic analysis of algorithm performance on noncoding pharmacogenomic variation. In an interesting case study, Xie and colleagues analyzed associations of noncoding variation with the sensitivity of lymphoblastoid cells to anthracyclines and identified candidate transcription factors as putative mediators of drug response.\textsuperscript{79} While more systematic evaluations remain to be conducted, this work indicates the potential utility of such a strategy to associate noncoding variants in regulatory domains with pharmacological and toxicological variation.

**INTERPRETING VARIANT HAPLOTYPES**

All algorithms discussed above interpret variant functionality in isolation, i.e., without considering combinatorial effects either in cis (variants within the same locus) or in trans (variants in distant loci). However, many pharmacogenetic alleles are characterized by a combination of variants. For example, more than 140 haplotypes have so far been described for \textit{CYP2D6} and, given the large number of rare SNVs identified in population-scale data sets, it is likely that many more remain to be described. Similarly, haplotypes across the \textit{CYP2C18/CYP2C19} locus have recently been associated with rapid metabolism of escitalopram.\textsuperscript{80} While the functional consequences of some haplotypes have been experimentally
determined, the functionality of the majority remains as of yet unknown. Most commonly, haplotypes are computationally interpreted using a "dominant model," i.e., a haplotype is considered to have decreased activity if at least one of the variants is predicted to result in decreased activity, irrespective of whether the other variants in the haplotype also alter gene function or are functionally neutral. However, as shown for DPYD haplotypes, this strategy might not always correctly reflect clinical observations. Furthermore, combinations of variants with opposite functional tendency, i.e., haplotypes consisting of increased and decreased function variants, cannot be predicted.

CYP2D6 haplotypes can be functionally interpreted using Hubble.2D6, a tool that uses transfer learning and training on a set of expert-annotated CYP2D6 star alleles from PharmVar. Hubble.2D6 achieved 88% accuracy in an independent validation set; however, when transfer learning or functional annotations were excluded, test accuracy dropped to 44% and 40%, respectively, demonstrating that both model components synergize to yield a model with high accuracy. As such, this approach constitutes a promising step toward the computational prediction of pharmacogenetic haplotypes, particularly if training data are sparse, and it will be interesting to see whether similar methods can improve analyses of other pharmacogenes with complex linkage disequilibrium, such as CYP2A6, CYP2B6, DYPD, or the NAT loci.

Single nucleotide variants in distant genetic loci can exert synergistic effects by affecting the function of different genes that synergistically impact a phenotype of interest. This phenomenon is well established in more than 44 diseases in which the disease is caused by digenic combinations of mutations, such as long QT syndrome, genetic deafness, and Usher syndrome. These pathogenic variant combinations served as training data for a computational method called VarCoPP that can predict such variant interactions in trans in a biological context.

Digenic effects have been described for absorption, distribution, metabolism, and excretion, and their functional interpretation is mainly based on their involvement in the same metabolic pathways. Examples of such quantitative interactions are between variants in VKORC1 and CYP2C9 that jointly modify warfarin dose requirements, combinations of TPMT and NUDT15 variations as determinants of severe thiopurine hematotoxicity, and interactions of CYP2C19 and CYP2D6 variants in determining clinical outcome in tamoxifen-treated patients with breast cancer.

INTERPRETING STRUCTURAL VARIANTS

Structural variants (SVs) are defined as deletions, duplications, insertions, inversions, or other complex genomic rearrangements that cause changes of more than 50 base pairs on DNA. SVs are generally more difficult to analyze than SNVs or indels, but, during the past few years, advancements in sequencing technologies have substantially improved the ability to detect SVs in human genomes. A recent study analyzed 14,891 genomes across five different human populations and identified 433,371 SVs, which substantially surpassed previous estimates, suggesting that structural variation constitutes a common phenomenon with important impacts on a plethora of traits. Structural variation is highly relevant in multiple pharmacogenes, such as CYP2B6 and CYP2D6, where deletions, duplications, and other complex rearrangements can result in altered gene copy numbers, which in turn can be functional or inactive.

Notably, however, the functional interpretation of SVs remains challenging, at least in part due to the inaccurate detection of breakpoints. Most commonly the functional effects of SVs are interpreted using simplified rules, i.e., all SVs that overlap with any exon are considered LOF, as are duplications that only cover parts of a gene’s coding region. In contrast, duplications that span the entire gene are considered gain-of-function (GOF). Using these simplified rules, CYP2D6*1xN and CYP2D6*2xN would be correctly identified as GOF as the entire coding region is duplicated. Sequence differences in the downstream repetitive sequence would not affect this conclusion. Similarly, duplications of GSTM1 and SULT1A1 would be correctly flagged as GOF. The simplified rules would also correctly classify the deletion allele CYP2D6*5 as LOF. However, no clear predictions could be made regarding the duplication of reduced-function alleles, such as CYP2D6*9xN or CYP2D6*10xN. If found de novo, complex structural rearrangements, such as the CYP2D6-CYP2D7 hybrid alleles CYP2D6*13, CYP2D6*61, or CYP2D6*63, would likely be classified as LOF alleles, depending on the genomic annotation.

So far, only a few computational methods have been developed that utilize more sophisticated rules. SVScore utilizes pathogenicity assessments for individual variants affected by the SV, provided by tools such as CADD, and integrates these per-base scores into an aggregate metric across the corresponding genomic interval. More recently, a machine learning–based framework was presented that assigns SV pathogenicity by comparing genomic and tissue-specific epigenomic features of SVs in diseased and healthy individuals. This method achieved high predictive accuracy in both somatic and germline SV sets across various diseases, including cancer, cardiovascular, and inflammatory bowel disease. However, whether those algorithms are useful for the prediction of SVs in poorly conserved pharmacogenes remains to be determined.

CONCLUSIONS

Computational interpretation of pharmacogenomic variability is considered an important pillar for the clinical implementation of NGS-guided therapy. Importantly, associations between genetic factors and pharmacological or toxicological phenotypes are complex, involving coding and noncoding variants, as well as complex structural variations, which complicates genotype to phenotype translations. It is becoming increasingly clear that the most commonly used methods that are developed for the genome-wide detection of pathogenic variants perform relatively poorly when applied to specific data sets, such as pharmacogenetic variations. As a consequence, a considerable fraction of genetic interindividual variability remains unexplained, which complicates the clinical implementation of pharmacogenetic data. Current trends go toward the development of refined algorithms trained for specific variant classes, genes, or applications, which promises to improve predictive accuracy. Importantly though, the repertoire of methods that have been developed for or, at least, have been tested on poorly conserved pharmacogenes remains limited, and, while methodological improvements for pathogenic assessment algorithms are actively...
discussed in the literature,96,97 these developments are only starting to catch on in pharmacogenomics. Furthermore, algorithms are commonly not accessible via user-friendly interfaces, which hampers dissemination to users without bioinformatic expertise and reduces the adoption of these methods in clinical practice.

We thus advocate for a stream of dedicated research for the training and further development of computational methods for the functional interpretation of pharmacogenetic variability across variant classes. As discussed before,20 systematic experimental methods, such as deep mutational scanning, could provide a useful means to increase the amount of available high-quality pharmacogenomic training data and such a strategy already has promising results for the improvement of genome-wide prediction models.98 Furthermore, there are at least 41 active national biobank projects of which the majority offer data sharing with plans on linking genomic data to longitudinal medical records.99 These efforts are likely to provide important resources for the identification of naturally occurring functional variants, thereby further increasing the extent of available training data, particularly for transfer learning. We conclude that the predictive power of computational methods has overall drastically improved over the last several years. However, algorithms that specifically account for the genetic peculiarities of pharmacogenes have so far only been established for the interpretation of missense variants, and further developments for other variant classes, particularly noncoding variations, are needed to narrow the gap between variant identification and the translation of this information into clinical advice.

FUNDING
V.M.L. receives support from the Swedish Research Council (grant agreement numbers: 2016-01153, 2016-01154, and 2019-01837), by the EU/EFPIA/OICR/McGill/KTH/Diamond Innovative Medicines Initiative 2 Joint Undertaking (EUbOPEN grant number 875510), by the

Figure 3
Quantitative evaluation of computational predictions for the assessment of common pharmacogenomic alleles. The function of frequent pharmacogenomic alleles was predicted using 15 commonly used variant effect predictors and compared with activity data from the literature. Reduced-activity variants are shown in red; neutral variants are shown in blue. Variants whose function could not be predicted with the algorithm in question are shown in gray. Quantitative correlations (Pearson’s r) of prediction scores and functional activity are shown for each algorithm. Note that APF and MutationAssessor are decent quantitative predictors of pharmacogenomic allele activity with r > 0.5, whereas evolutionary conservation scores (GERP++, SiPhy, and PhastCons) and FATHMM could not provide meaningful quantitative predictions (r < 0.2).
Swedish Strategic Research Programmes in Diabetes (SFO Diabetes), and Stem Cells and Regenerative Medicine (SFO StratRegen). Furthermore, the lab acknowledges support from Merck KGaA and Eli Lilly and Company.

CONFLICT OF INTEREST
Y.Z. is cofounder and shareholder of PersoMedix AB. V.M.L. is cofounder and shareholder of PersoMedix AB, CEO and shareholder of HepaPredict AB, and conducted consultancy work for Enginzyme AB.

© The Authors. Clinical Pharmacology & Therapeutics published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

1. Lauschke, V.M. & Ingelman-Sundberg, M. Prediction of drug response and adverse drug reactions: from twin studies to next generation sequencing. *Eur. J. Pharm. Sci.* **130**, 65–77 (2019).

2. Sim, S.C., Kacesvka, M. & Ingelman-Sundberg, M. Pharmacogenomics of drug metabolizing enzymes: a recent update on clinical implications and endogenous effects. *Pharmacogenomics J.* **13**, 1–11 (2013).

3. Lauschke, V.M., Milani, L. & Ingelman-Sundberg, M. Pharmacogenetic biomarkers for improved drug therapy-recurrent progress and future developments. *AAPS J.* **20**, 4 (2017).

4. Drögemöller, B.I., Wright, G.E.B. & Warnich, L. Considerations for rare variants in drug metabolism genes and the clinical implications. *Expert Opin. Drug Metab. Toxicol.* **10**, 873–884 (2014).

5. Lauschke, V.M. & Ingelman-Sundberg, M. Precision medicine and rare genetic variants. *Trends Pharmacol. Sci.* **37**, 85–86 (2016).

6. Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y. & Lauschke, V.M. Integrating rare genetic variants into pharmacogenomic drug response predictions. *Human Genomics* **12**, 26 (2018).

7. Kozyra, M., Ingelman-Sundberg, M. & Lauschke, V.M. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. *Genet. Med.* **19**, 20–29 (2017).

8. Santos, M. et al. Novel copy-number variations in pharmacogenes contribute to interindividual differences in drug pharmacokinetics. *Genet. Med.* **20**, 622–629 (2018).

9. Gordon, A.S. et al. Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. *Hum. Mol. Genet.* **23**, 1957–1963 (2014).

10. Fujikura, K., Ingelman-Sundberg, M. & Lauschke, V.M. Genetic variation in the human cytochrome P450 supergene family. *Pharmacogenet. Genomics* **25**, 584–594 (2015).

11. Xiao, Q., Zhou, Y. & Lauschke, V.M. Ethnogeographic and inter-individual variability of human ABC transporters. *Hum. Genet.* **139**, 623–646 (2020).

12. Kalman, L.V. et al. Pharmacogenetic allele nomenclature: International workshop recommendations for test result reporting. *Clin. Pharmacol. Ther.* **99**, 172–185 (2016).

13. Luvsantseren, S. et al. Variant interpretation in current pharmacogenetic testing. *J. Pers. Med.* **10**, 204 (2020).

14. Zhou, Y., Ingelman-Sundberg, M. & Lauschke, V.M. Worldwide distribution of cytochrome P450 alleles: a meta-analysis of population-scale sequencing projects. *Clin. Pharmacol. Ther.* **102**, 688–700 (2017).

15. Karczewski, K.J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

16. Zhou, Y., Fujikura, K., Mkrtchian, S. & Lauschke, V.M. Computational methods for the pharmacogenetic interpretation of next generation sequencing data. *Front. Pharmacol.* **9**, 1437 (2018).

17. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).

18. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

19. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PloS One* **7**, e46688 (2012).

20. Lauschke, V.M. & Ingelman-Sundberg, M. Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. *NPJ Genom. Med.* **5**, 9 (2020).

21. Ioannidis, N.M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

22. Kircher, M., Witten, D.M., Jain, P., O’Roak, B.J., Cooper, G.M. & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

23. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genetic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).

24. Shihab, H.A. et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).

25. Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R. & Campbell, C. PATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).

26. Li, J. et al. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.* **46**, 7793–7804 (2018).

27. Mahmood, K. et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum. Genomics* **11**, 10 (2017).

28. Grimm, D.G. et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).

29. Hart, S.N., Polley, E.C., Shemis, H., Yadav, S. & Couch, F.J. Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML. *NPJ Breast Cancer* **6**, 13 (2020).

30. Shrestha, S. et al. Gene-Specific Variant Classifier (DPYD-Varifier) to identify deleterious allelics of dihydropyrimidine dehydrogenase. *Clin. Pharmacol. Ther.* **104**, 709–718 (2018).

31. Heyne, H.O. et al. Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci. Transl. Med.* **12**, eaay6848 (2020).

32. Lee, W., Zhang, Y., Mukhyala, K., Lazarus, R.A. & Zhang, Z. Bidirectional SIFT predicts a subset of activating mutations. *PLoS One* **4**, e8311 (2009).

33. Wang, J. et al. A novel statistical method for interpreting the pathogenicity of rare variants. *Genet. Med.* **23**, 59–68 (2021).

34. Zhou, Y., Mkrtchian, S., Kumondai, M., Hiratsuka, M. & Lauschke, V.M. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics J.* **19**, 115–126 (2019).

35. Zhou, Y., Bagli Hernandez, C. & Lauschke, V.M. Population-scale predictions of DPD and TPMT phenotypes using a quantitative pharmacogene-specific ensemble classifier. *Br. J. Cancer* **123**, 1782–1789 (2020).

36. Shabalina, S.A., Spiridonov, N.A. & Kashina, A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res.* **41**, 2073–2094 (2013).

37. Hunt, R.C., Simhardi, V.L., Iandoli, M., Sauna, Z.E. & Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **30**, 308–321 (2014).

38. Presnyak, V. et al. Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124 (2015).

39. Chen, R., Davydov, E.V., Shiroti, M. & Burt, A.J. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One* **5**, e13574 (2010).
85. Renaux, A. et al. ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res.* 47, W93–W98 (2019).

86. Schalekamp, T. et al. VKORC1 and CYP2C9 genotypes and acenocoumarol anticoagulation status: interaction between both genotypes affects overanticoagulation. *Clin. Pharmacol. Ther.* 89, 13–22 (2006).

87. Schaeffeler, E. et al. Impact of NUDT15 genetics on severe thiopurine-related hematotoxicity in patients with European ancestry. *Genet. Med.* 21, 2145–2150 (2019).

88. Sim, S., Lövrot, J., Lindh, J.D., Bergh, J. & Xie, H. Effect of CYP2C19 and CYP2D6 genotype on tamoxifen treatment outcome indicates endogenous and exogenous interplay. *Pharmacogenomics* 19, 1027–1037 (2018).

89. Collins, R.L. et al. A structural variation reference for medical and population genetics. *Nature* 581, 444–451 (2020).

90. Desta, Z. et al. PharmVar GeneFocus: CYP2B6. *Clin. Pharmacol. Ther.* 110, 82–97 (2021).

91. Nofziger, C. et al. PharmVar GeneFocus: CYP2D6. *Clin. Pharmacol. Ther.* 107, 154–170 (2020).

92. Ganel, L. & Abel, H.J., FinMetSeq Consortium & Hall, I.M. SVScore: an impact prediction tool for structural variation. *Bioinformatics* 33, 1083–1085 (2017).

93. Kumar, S., Harmanci, A.,Vytheeswaran, J. & Gerstein, M.B. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol.* 21, 274 (2020).

94. Lauschke, V.M. & Ingelman-Sundberg, M. How to consider rare genetic variants in personalized drug therapy. *Clin. Pharmacol. Ther.* 103, 745–748 (2018).

95. Russell, L.E., Zhou, Y., Almousa, A.A., Sodhi, J.K., Nwabufo, C.K. & Lauschke, V.M. Pharmacogenomics in the era of next generation sequencing – from byte to bedside. *Drug Metab. Rev.* 53, 253–278 (2021).

96. Ghosh, R., Oak, N. & Plon, S.E. Evaluation of *in silico* algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol.* 18, 225 (2017).

97. Ellard, S., Colclough, K., Patel, K.A. & Hattersley, A.T. Prediction algorithms: pitfalls in interpreting genetic variants of autosomal dominant monogenic diabetes. *J. Clin. Invest.* 130, 14–16 (2020).

98. Gray, V.E., Hause, R.J., Luebeck, J., Shendure, J. & Fowler, D.M. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst.* 6, 116–124.e3 (2018).

99. Kovanda, A., Zimani, A.N. & Peterlin, B. How to design a national genomic project—a systematic review of active projects. *Hum. Genomics* 15, 20 (2021).

100. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 14(suppl. 3), S3 (2013).