From gene to dose: long-read sequencing and -allele tools to refine phenotype predictions of CYP2C19

Graansma, L.J.; Zhai, Q.L.; Busscher, L.; Menafra, R.; Berg, R.R. van den; Kloet, S.L.; Lee, M. van der

From gene to dose: Long-read sequencing and *-allele tools to refine phenotype predictions of *CYP2C19*

Lonneke J. Graansma[1], Qinglian Zhai[1], Loes Busscher[2], Roberta Menafra[2], Redmar R. van den Berg[2], Susan L. Kloet[2] and Maaike van der Lee[1]*

[1]Department of Clinical Pharmacy and Toxicology, Leiden University Medical Center, Leiden, Netherlands, [2]Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, Netherlands

Background: Inter-individual differences in drug response based on genetic variations can lead to drug toxicity and treatment inefficacy. A large part of this variability is caused by genetic variants in pharmacogenes. Unfortunately, the Single Nucleotide Variant arrays currently used in clinical pharmacogenomic (PGx) testing are unable to detect all genetic variability in these genes. Long-read sequencing, on the other hand, has been shown to be able to resolve complex (pharmaco) genes. In this study we aimed to assess the value of long-read sequencing for research and clinical PGx focusing on the important and highly polymorphic *CYP2C19* gene.

Methods and Results: With a capture-based long-read sequencing panel we were able to characterize the entire region and assign variants to their allele of origin (phasing), resulting in the identification of 813 unique variants in 37 samples. To assess the clinical utility of this data we have compared the performance of three different *-allele tools (Aldy, PharmCat and PharmaKU) which are specifically designed to assign haplotypes to pharmacogenes based on all input variants.

Conclusion: We conclude that long-read sequencing can improve our ability to characterize the *CYP2C19* locus, help to identify novel haplotypes and that *-allele tools are a useful asset in phenotype prediction. Ultimately, this approach could help to better predict an individual's drug response and improve therapy outcomes. However, the added value in clinical PGx might currently be limited.

KEYWORDS

long-read sequencing, CYP2C19, personalized medicine, pharmacogenomics, bioinformatics

1 Introduction

One drug does not have the same effect for everyone; inter-individual differences in drug response can lead to toxicity and drug inefficacy (Weinshilboum and Wang, 2004). A large part of this variability is caused by genetic variants in genes, called 'pharmacogenes', which are involved in the pharmacokinetic and pharmacodynamic processes occurring as part of drug metabolism (Scharfe et al., 2017; Tafazoli et al., 2021). Variants in these pharmacogenes are associated with diverse drug responses. After the discovery that over 97% of the general
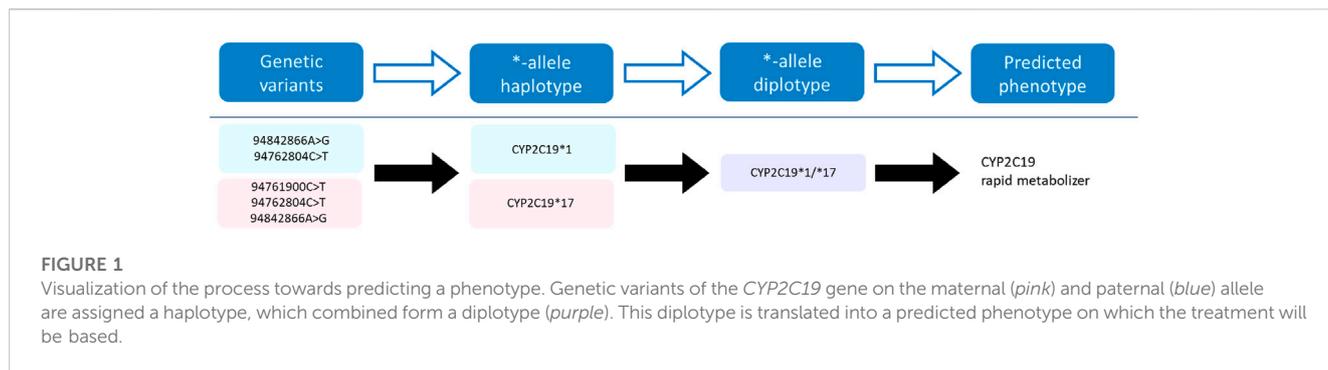
FIGURE 1
Visualization of the process towards predicting a phenotype. Genetic variants of the *CYP2C19* gene on the maternal (*pink*) and paternal (*blue*) allele are assigned a haplotype, which combined form a diplotype (*purple*). This diplotype is translated into a predicted phenotype on which the treatment will be based.

TABLE 1 Overview of *-allele tools and their specifications. For every *-allele tool selected, the supported input file format(s), the used reference genome, the prerequisite software and the reference database on which the *-allele tool analysis is based are depicted. When 'WGS' is stated, these tools assume input files with WGS data. 'Tweak' indicates that an adjustment was necessary before the *-allele tool supported this input file.

		Stargazer	Aldy[a]	PharmCat[a]	Cyrius	StellarPGX	PharmaKU[a]	PharmVIP[a]
Input format	VCF	X	X (tweak)	X			X	X
	BAM	X	X		WGS	WGS		X
	CRAM		X		WGS	WGS		
	SAM		X					
Reference Genome	GRCh37	X	X		X	X	X	
	GRCh38		X	X	X	X	X	X
Prerequisite Software	Windows	X						
	Linux	X	X			X		
	MacOS	X	X			X		
	Java			X				
	DRAGEN				X			
	Webbased						X	X
Reference database	PharmVar	X	X			X		
	PharmGKB		X	X			X	(PharmCat)[b]
	Own					X		
	Other		X				X (CPIC)[b]	

[a]*-allele tools assessed in this study.
[b]This is based on the tool/database specified.
BAM, binary alignment map; CRAM, Compressed Reference-oriented Alignment Map; SAM, sequence alignment map; VCF, variant call format; WGS, whole genome sequence.

population carries at least one pharmacogenomic (PGx) variant which can potentially affect drug response (Dunnenberger et al., 2015), personalized medicine is slowly becoming standard of care instead of a 'one size fits all' approach. This implementation of PGx in clinical practice to guide treatment decisions is crucial to maximize the effectiveness of a treatment and minimize harm.

The pharmacogene *CYP2C19* is a member of the cytochrome P450 (CYP) superfamily and is involved in the metabolism of many commonly prescribed drugs such as clopidogrel and proton-pump-inhibitors (harmGKB). Moreover, *CYP2C19* is highly polymorphic; 50%–65% of the population is characterized with an not 'normal' metabolic capacity based on the genetic make-up of their *CYP2C19* resulting in the need for dose adjustments (Hicks et al., 2017). This

combination of clinical relevance and high abundance of genetic variants makes it a highly important pharmacogene which is frequently tested in hospital laboratories to guide treatment decisions (Pratt et al., 2018). In order to use genetic information of *CYP2C19*, the star (*)-allele nomenclature and accompanying predicted phenotypes are used (Figure 1). Based on the combination of genetic variants identified, a haplotype or "*-allele" is assigned to the maternal and paternal allele according to the *-nomenclature system (Ingelman-Sundberg et al., 2000; Gaedigk et al., 2018). The two *-haplotypes are then combined into a diplotype. Based on the enzyme activity corresponding to the assigned diplotype, this diplotype is translated into a predicted phenotype (Figure 1). Different predicted phenotype categories are recognized which

are used in the dosing guidelines of the Clinical Pharmacogenetics Implementation Consortium (CPIC) (Caudle et al., 2017) and the Dutch Pharmacogenetics Working Group (DPWG) (Swen et al., 2008). For *CYP2C19,* CPIC defines five different metabolizer types: Poor Metabolizers (PM), Intermediate Metabolizers (IM), Normal Metabolizers (NM), Rapid Metabolizers (RM) and Ultra-rapid Metabolizers (UM). DPWG on the other hand does not use the RM predicted phenotype. To aid in the interpretation of the high amount of variants that can be detected with sequencing approaches, *-allele tools have been developed. There is a great variety in these bioinformatic tools; each tool is based on a different reference database and genome, runs on different software and supports different input file formats (Table 1). Since they all make different assumptions, the output of each tool may differ. A detailed analysis of the methods behind each tool, the performance of each tool and an analysis of their applicability using long-read sequencing is yet to be performed.

In the *-allele nomenclature, 71 *CYP2C19* variants are currently taken into account, of which only three are recommended for standard clinical testing by the association for molecular pathology (Pratt et al., 2021). In routine PGx for both clinical practice and research purposes, the variants of interest are generally genotyped with SNV (Single Nucleotide Variant) arrays (Mukerjee et al., 2018). These arrays directly interrogate the genomic positions of known variants. However, in Gnomad a total of 975 variants have been reported in *CYP2C19* so far. The fact that routine PGx uses a limited number of variants means that they are unable to identify all variants (including rare and novel variants). Moreover, the variants that are detected cannot be phased to their allele of origin. Phasing determines whether variants are located on the same allele or on opposing alleles (e.g., *CYP2C19* *2/*3 (PM) or *CYP2C19* *1/*2+*3 (IM)) which can be of major importance for enzyme activity. Hence, phasing is expected to improve haplotype assignment and therefore phenotype prediction (van der Lee et al., 2022). These two limitations, (novel) variant detecting and haplotype phasing, are hindering the ability of SNV arrays to fully characterize pharmacogenes.

To resolve these limitations, improve phenotype prediction and get a better understanding of genetic make-up, more advanced technologies are needed. While Next-Generation short read Sequencing (NGS) can help to identify all variants within the genes of interest (van der Lee et al., 2020a), it remains difficult to resolve complex regions and to perform direct (read based) haplotype phasing with short reads (100–200 bp). However, long-read sequencing has the ability to overcome the limitations mentioned above, since it is able to resolve larger regions and enables us to look into phased haplotypes. Due to the length of the reads (~5.000 base pairs (bp) for capture based approaches), variants can be assigned to their allele of origin (Midha et al., 2019). While single pass accuracy is still limited with Pacific Bioscience (PacBio, 2023) HiFi sequencing, the circular consensus reads - which are obtained by combining multiple sequencing passes of the same molecule—are highly accurate (Wenger et al., 2019). To date, long-read sequencing for PGx has only been applied in a single gene setting (Qiao et al., 2016; Borràs et al., 2017; van der Lee, 2021) or with publicly available data of one individual (van der Lee et al., 2022). Recently, a long-read sequencing panel using PacBio

technology, which consists of a selection of clinically relevant pharmacogenes, has been developed at the Leiden University Medical Center (LUMC).

While the application of long-read sequencing for clinical PGx might still be too costly and time intensive, due to the rapid decline in costs it can be expected that long-read sequencing originating from research applications becomes more abundantly available. This offers the opportunity to repurpose this data to extract a PGx profile using all *-allele variants and the phasing information. Hence, it is of importance to assess the performance of long-read sequencing for clinical PGx as well as for research. Therefore the aim of this study is to investigate the benefit of long-read sequencing for both of these applications, by resolving the *CYP2C19* locus, identifying (novel) variants, and assessing the performance of different *-allele tools. The knowledge gained from this exploratory study might give a focus and direction for further PGx research.

2 Methods

2.1 Long-read sequencing panel

48 samples originating from old PGx studies performed at the clinical pharmacy and toxicology department of the Leiden University Medical Center (LUMC), were available and anonymized. These samples were sequenced using the PacBio long-read sequencing panel developed by the department of clinical pharmacy and toxicology and the department of human genetics of the LUMC. This panel includes core PGx genes as well as genes that were of interest for specific projects at the departments. For every gene, ~10.000 bp upstream and downstream of the transcription start and end site, respectively, were included in the panel. In short, DNA concentration and quality were checked by using the Qubit Fluorometer and the Qubit dsDNA Broad Range Assay kit (Invitrogen, Carlsbad, CA) and verified on the Femto Pulse system (Agilent Technologies, CA, United States). Subsequently, the DNA was sheared to an average size of ~8 kb using the Diagenode Megaruptor three and purified using washed AMPure XP beads (Beckman-Coulter Woerden, Netherlands). End repair, A-tailing and adapter ligation was performed using 500 ng of sheared DNA product and the Twist Library Preparation Kit 1 (Twist Bioscience, CA, United States) and the PacBio annealed barcoded adapter (10 μM, desalt-purified (Integrated DNA Technologies (IDT), Coralville, IA)). After further purification and size selection (3.7x diluted washed AMPure XP beads), the DNA was amplified using the Takara LA Taq HotStart kit (TaKaRa Bio United States, Inc.). The reaction was performed in two reaction volumes (100 μL) containing: 50–100 ng of DNA, 0.5 μM PacBio universal primer (/5Phos/gcagtcgaacatgtagctgactcaggtcac (IDT, Coralville, IA)), 0.1 mM of each dNTP, 1x LA PCR buffer, and 0.03 U Takara LA Taq. The PCR parameters were 2 min at 95˚C, followed by six cycles of 20 s at 95˚C, 15 s at 64˚C and 10 min at 68˚C, and a final extension of 5 min at 68˚C. After amplification the two reaction volumes were pooled and the product was checked for concentration and quality with the Qubit Fluorometer and the Femto Pulse system. The product of eight samples was equimolarly pooled. Next, the capture was performed using the Twist Hybridization and Wash Kit (Twist Bioscience, CA, United States) and the Twist Probe

Custom panels (Twist Bioscience, CA, United States). The pools were then amplified using the Takara LA Taq HotStart kit in two reaction volumes (100 μL) each containing: 50 μL captured sample pool, 0.5 μM PacBio universal primer, 0.2 mM of each dNTP, 1x LA PCR buffer, and 0.03 U Takara LA Taq. After another step of quality control with the Qubit Fluorometer and the Femto Pulse system, sequence libraries were prepared using 500 ng of the captured pooled samples. The library was sequenced on the Sequel® II (Pacific Biosciences, CA, United States) on a 8 M SMRT cell at an on-plate concentration of 80 p.m. with the following specifications: sequencing primer V4, Sequencing kit 2.0 and binding kit 2.0 and a 30 h movie time. HiFi CCS (Circular concensus reads) were obtained for further processing (Wenger et al., 2019; PacBio, 2023).

2.2 Data preprocessing

The data was preprocessed using the LUMC developed variant calling pipeline specifically for the PacBio PGx sequencing project (Redmar van den Berg, 2021). The CCS subreads were demultiplexed using LIMA (Lima, 2023). Duplicate reads were marked using pbmarkdup (Pdmarkdup, 2023). Next, demultiplexed CCS bam files were mapped to the reference genome (GRCh38) using pbmm2 (Pbmm2, 2023). Thereafter, the variant calling was performed with GATK4 (GATK, 2023) and phased using WhatsHap (Martin et al., 2016). Finally, the results were aggregated and reported with MultiQC (Multiqc, 2023). The output was reported in both BAM (Binary Alignment Map) and VCF (Variant Call Format) files for each sample. Samples with less than 10% of the target bases reaching at least 30X coverage were excluded from the analysis.

All variants will be described according to Human Genome Variation Society (HGVS) nomenclature for GRCh38 (den Dunnen et al., 2016), using genomic positions on chromosome 10 NC_000010.11 (location of *CYP2C19*). If more applicable, dbSNP Reference SNP (RefSNP or rs) numbers are used (Sherry et al., 2001), as they are widely known and recognized in the PGx field.

2.3 Variant characterization

For variant characterizations, all VCF files were cross-referenced with a bed file containing the genomic coordinates of the start and end positions of the included genes to obtain a VCF file specific to the PGx genes. Moreover, clinically relevant variants were flagged based on their presence in PharmVar (PharmVar, 2022) or, if the gene was not available in PharmVar, their presence in the Ubiquitous pharmacogenomics consortium's (U-PGx) variant panel (upgx, 2020; van der Wouden et al., 2019). Thereafter, the number of known variants (variants present in Pharmvar and/or the U-PGx panel) and novel variants (variants not used in the clinic) identified per gene were calculated using Excel. In this study, we refer to variants which are not in current PGx nomenclature as novel variants. For *CYP2C19* all clinically relevant variants were obtained from PharmVar.

The predicted impact of *CYP2C19* variants was assessed using the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) including SIFT (Kumar et al., 2009) and PolyPhen (Adzhubei et al.,
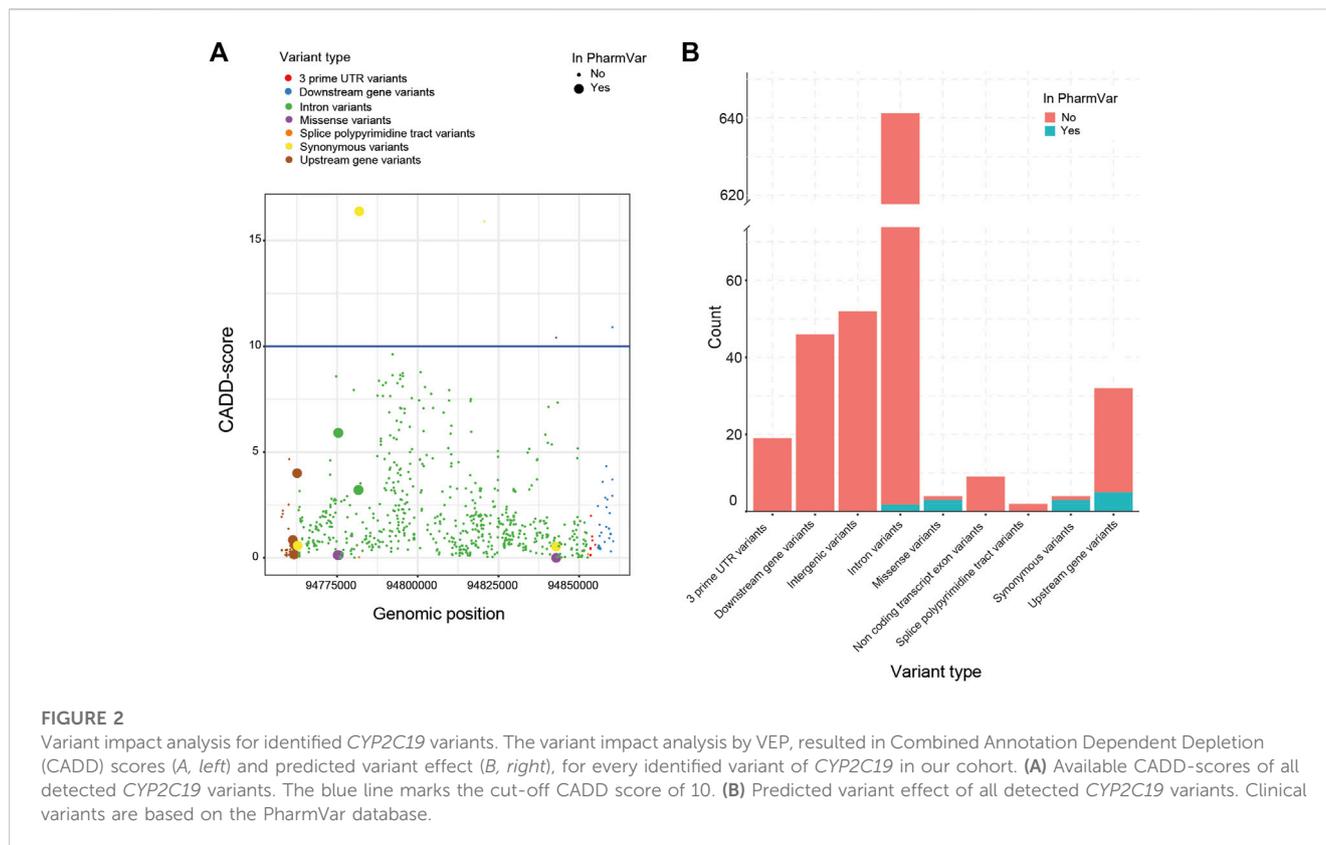
2010) for missense variants and the Combined Annotation Dependent Depletion (CADD) score (Rentzsch et al., 2021) was used to predict the magnitude of the impact of non-synonymous variants, selecting on *CYP2C19* (transcript: ENST00000371321.9). The ~10.000 bp upstream and downstream regions were also taken into account during the impact analysis. While *in silico* algorithms have a limited accuracy in predicting enzymatic function (80% accuracy) (Han et al., 2017), we use them here as one of the tools to explore the potential impact of (novel) variants as no PGx specific *in silico* tools are available. The CADD score was selected as prediction tool to assess what part of the identified variants is classified as high impact (CADD score >10) and is a candidate for further functional assessment. It has previously been shown that, of the non-class specific tools, the CADD score performs best on pharmacogenes (Han et al., 2017).

2.4 *-Allele calling and phenotype prediction

For every sample the predicted metabolizer type was assigned based on the identified genetic variants in *CYP2C19*. To investigate the performance of *-allele tools on long-read sequencing data and to study potential differences in underlying assumptions of these tools, a comparison was made between manual *-allele assignments and assignments from four different *-allele tools. Manual curation was used as the ground truth as we could check phasing and variants manually for all samples, thereby omitting the risk of inaccurate phasing assumptions. Moreover, the results were used to determine the predicted phenotypes of the samples according to the CPIC guidelines (Caudle et al., 2017). To ensure a broad selection, these four tools were selected based on three criteria: 1) every tool uses reference genome GRCh38, 2) all tools support VCF as input file and 3) they all require different software to run. For the automatic processing, four *-allele tools were selected: PharmVIP (Piriyapongsa et al., 2021), Aldy v3.0 (Numanagić et al., 2018), PharmaKU (John et al., 2021) and PharmCat 1.8.0 (Sangkuhl et al., 2020). PharmVIP and PharmaKU were used as described in the literature (Piriyapongsa et al., 2021) (John et al., 2021). Since they are both web-based, no installations were necessary. In order to run Aldy, Ubuntu 20.04.4 was installed inside VirtualBox (6.1), Aldy was run using the VCF option by changing the default from hg19 to hg38. PharmCat was run on Java 16.0.2. Since these tools use the CPIC definition of the metabolizer types, the CPIC guidelines were also used for the manual assignment. The manual assignment was performed using the phased VCF files and the PharmVar database of *CYP2C19* (June 2022) (PharmVar, 2022).

Due to limitations in the processing of samples in the PharmVIP tool, a complete analysis of all samples could unfortunately not be performed. PharmVIP limits the storage of the results to 10 days and ten samples at any one time, making it not applicable for clinical practice and not suitable for this study. Hence, the remaining three *-allele tools are assessed in more detail and no PharmVIP analyses are included in the results.

The results of the tool-based analyses were combined into one file, after which the manual *-allele assignments were added. Based on this, the accuracy (compared to the manual curation) and ease of use of the different *-allele tools and added value of phasing were assessed.

FIGURE 2
Variant impact analysis for identified *CYP2C19* variants. The variant impact analysis by VEP, resulted in Combined Annotation Dependent Depletion (CADD) scores (*A, left*) and predicted variant effect (*B, right*), for every identified variant of *CYP2C19* in our cohort. **(A)** Available CADD-scores of all detected *CYP2C19* variants. The blue line marks the cut-off CADD score of 10. **(B)** Predicted variant effect of all detected *CYP2C19* variants. Clinical variants are based on the PharmVar database.

2.5 Statistical analysis

If not specifically stated, analyses were performed using the software as described in the tool documentation. Data was processed with Excel, R version 1.4.1717 and Python 3.9.12. Visualizations were performed with R version 1.4.1717 and Adobe Illustrator 25.2.3.

3 Results

3.1 Variant identification

To assess the ability of long-read sequencing to detect (novel) genetic variants, the number of variants identified per gene was analyzed. Out of the 48 samples, 11 samples are excluded from the study due to a lack of coverage (<10% Target Bases 30X) (Supplementary Figure S1). For the remaining 37 samples, an average of 84.1% of the target bases reached 30X coverage (range: 33%–96%). Moreover, the average read length was 5,418 bp (range: 3,980–6,277 bp) and the average haploblock size 7,507 bp (range 4,863–9,806 bp). After this exclusion, 813 unique *CYP2C19* variants can be identified in the entire *CYP2C19* region (including upstream and downstream regions) of which 13 are present in the current PGx nomenclature. 303 of the 813 variants are singletons; they are only detected once in our cohort. The same trend is observed for the other genes in the panel, where many more variants are identified compared to the number currently used in the guidelines (Supplementary Table S1). This discrepancy remains

when only looking at the core gene (without taking the flanking regions into account). For *CYP2C19,* 683 variants are identified in the core gene (start to end position of the gene) of which 8 variants are present in PharmVar. Besides identifying the *CYP2C19* variants in the cohort, the variants were also phased. In total, an average of 76% (11%–100%) of the variants in the *CYP2C19* locus could be phased to their allele of origin in relation to at least one other variant.

Next, we investigated the predicted impact of all 813 variants using the Combined Annotation Dependent Depletion (CADD)-score generated by the Ensembl Variant Effect Predictor (VEP). Based on these CADD-scores, no relationship between the presence of a variant in the current nomenclature and its CADD-score could be established (Figure 2A). This lack of association between CADD-scores was visible for all VEP categories (e.g., missense, splice site, upstream). In total, only four variants had a CADD score above the cut-off of 10. The most common clinical deleterious variant in our cohort (g.94781859G>A, *2) was assigned the highest score (16.4). The other three variants are not part of *-allele nomenclature (one synonymous, one missense and one downstream).

All variants found in the cohort that are present in PharmVar are either upstream ($n = 5$), intronic ($n = 2$), synonymous ($n = 3$) or missense variants ($n = 3$) (Figure 2B). For the other variants identified in the cohort, these numbers were: upstream ($n = 27$), intronic ($n = 643$), synonymous ($n = 1$) and missense variants ($n = 1$), the remainder of the variant types is shown in Figure 2B. This enrichment of variants in the missense category in the clinical group (3 missense variants of 13 known variants) compared to the novel group (1 missense variant out of 675 novel variants) can be expected as the non-coding variants are often not regarded as impactful and
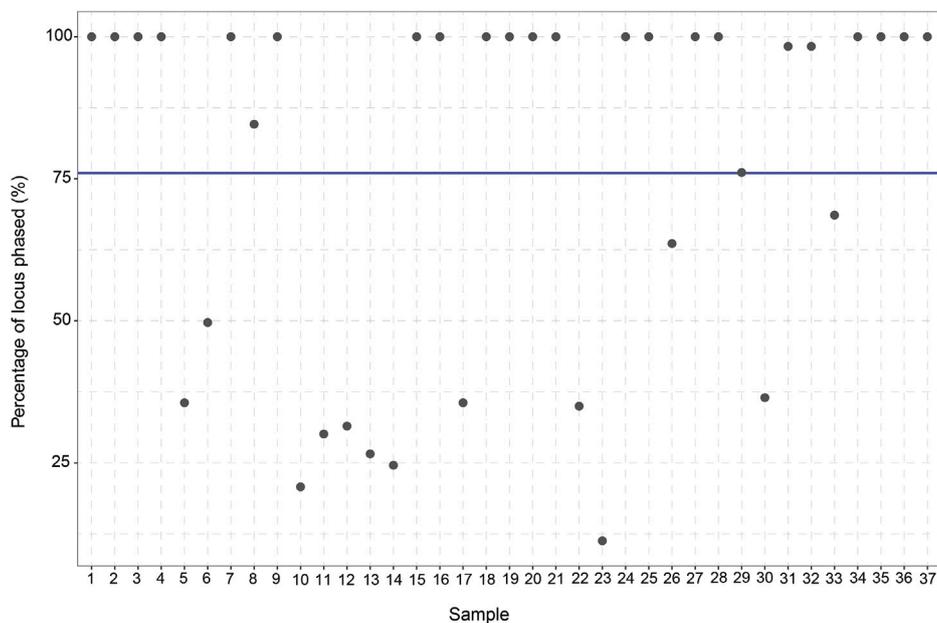
FIGURE 3
Percentage of *CYP2C19* variants phased to their allele of origin per sample. For every sample (*n* = 37) the percentage (%) of the variants which could be phased to their allele of origin was calculated. The blue line displays the average percentage of phased variants for the 37 samples (76%).

may not be included in the guidelines and nomenclature as rapidly as variants in coding regions are. The four missense variants were predicted to be not deleterious by SIFT (tolerated) and PolyPhen (benign). Three of these missense variants are part of the *-alleles (PharmVar, 2022). The first (g.94775507G>A>) is characteristic for *11 which is associated with a normal function allele. The second (g.94842866A>G) is a core variant of multiple *-alleles including *1 and is not expected to have an impact on enzyme function. The third missense variant (g.94775165G>C) is not a core variant and only associated with three minoralleles of *2 (*2.002, *2.010 and *2.012).

Out of the four missense variants identified in the entire cohort, three are in the PharmVar database. The fourth missense variant is not present in PharmVar even though literature confirmed the potential deleterious effect of this variant (g.94842860C>T) (Medicine Nlo. rs59734894, 2022; PharmGKB, 2022) and the CADD score is 10.4. The variant is reported to be associated with a decreased expression of *CYP2C19* compared to the reference allele, which would lead to a decreased metabolic capacity of CYP2C19 and subsequent dose adjustment. The individual carrying this variant was characterized as a *1/*17, unfortunately due to a decrease in coverage in intron five for this individual we are unable to determine if the novel variant (exon 7) is on the same allele as the *17 variant (upstream) or not.

3.2 Phenotype prediction

The data of the identified *CYP2C19* variants was used for a manual and *-allele tool based *-allele prediction.

3.2.1 Manual

The manual *-allele assignment resulted in an overview of the haplotype predictions for every sample. The *1/*17 (RM) diplotype was observed to be the most frequent diplotype in our cohort (35%) and only eight samples (21.6%) did not have any actionable PGx variants. Six of the 37 patients (16.2%) were carriers of *CYP2C19*38.003 + *g*.94781616A>G which is not yet part of the *-allele nomenclature. All six samples showed 100% phasing (Figure 3), affirming that g.94781616A>G is indeed present on the *CYP2C19*38 allele. According to PharmVar, the intronic g.94781616A>G variant is currently only associated with *CYP2C19*3.002 and not present in any other (minor)*-alleles. This makes it a potential novel (minor)*-allele with a relatively high frequency.

3.2.2 *-Allele tools

Based on the criteria put (GRCh38 as reference, support of VCF files and different platforms as their basis), three *-allele tools were selected; Aldy, PharmaKU and PharmCat. Advantages and disadvantages of these tools are described in Supplementary Table S2.

Assignments from Aldy were in agreement with the manually assigned diplotypes based on major (e.g., *2) and minor alleles (e.g., *2.002) for 56.8% of the assignments (*n* = 21) (Table 2). Due to this tool's transparency, we could observe that the Aldy resource file was not up to date with PharmVar as several *-alleles were not included, explaining the differences between the manual prediction and the assessment of Aldy. Moreover, Aldy is the only tool which also returns minor alleles (PharmaKU and PharmCat only return major alleles). When only looking at the major alleles, as they are most commonly used in clinical practice, the prediction by Aldy

TABLE 2 Overview of haplotype prediction by different *-allele tools. The 37 samples were assessed by different *-allele tools (Aldy, PharmaKU and PharmCat) and manually. The predicted haplotypes are depicted for each sample.

Samples	Manual, *-alleles (predicted phenotype)	Aldy, *-alleles (predicted phenotype)	PharmaKU, *-alleles (predicted phenotype)	PharmCAT[b]
1	*1.002/[a]38.003 + 94781616A>G (NM)	*1.001 + rs17885098/*1.005 (NM)	*1/*1 (NM)	*1/*5
2	*1.002/*38.003 + 94781616A>G 6 (NM)	*1.001 + rs17885098/*1.005 (NM)	*1/*1 (NM)	*1/*5
3	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
4	*1.002/[b]17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
5	*1.002 +94842865C>T/ *2.002–94842865C>T (IM)	*1.002/*2.002 (IM)	*1/*2 (IM)	*1/*2
6	*1.002/*1.002 (NM)	*1.002/*1.002 (NM)	*1/*1 (NM)	*1/*1
7	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
8	*1.002/[b]17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
9	*1.002/[b]17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
10	*1.002[b]*1.002 (NM)	*1.002/*1.002 (NM)	*1/*1 (NM)	*1/*1
11	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
12	*1.002[b]*1.002 (NM)	*1.002/*1.002 (NM)	*1/*1 (NM)	*1/*1
13	*1.002/*1.002 (NM)	*1.002/*1.002 (NM)	*1/*1 (NM)	*1/*1
14	*1.002/*1.006 (NM)	*1.002/*1.006 (NM)	*1/*1 (NM)	*1/*1
15	*17.001/*38.003 + 94781616A>G (RM)	*1.001/*17.001 (RM)	*1/*17 (RM)	*4/*38 (*17/ *38)[a]
16	*1.002[b]*2.011 + 94775507G>A (IM)	*2.001/*11.001 + rs4986894 (IM)	*1/*2 (IM)	*1/*2
17	*1.002/*2.012 (IM)	*1.002/*2.002 (IM)	*1/*2 (IM)	*1/*2
18	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
19	*1.002[b]*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
20	*1.002[b]*38.003 + 94781616A>G (NM)	*1.001 + rs17885098/*1.005 (NM)	*1/*1 (NM)	*1/*5
21	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
22	*1.002[b]*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
23	*1.002[b]*1.002 (NM)	*1.002/*1.002 (NM)	*1/*1 (NM)	*1/*1
24	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
25	*1.002[b]*2.011 (IM)	*1.002 + rs4986894/*2.001 (IM)	*1/*2 (IM)	*1/*2
26	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
27	*1.002[b]*2.011 (IM)	*1.002 + rs4986894/*2.001 (IM)	*1/*2 (IM)	*1/*2
28	*1.002/*17.001 (RM)	*1.002/*17.001 (RM)	*1/*17 (RM)	*1/*4 (*1/*17)[a]
29	*1.002[b]*1.002 (NM)	*1.002/*1.002 (NM)	*1/*1 (NM)	*1/*1
30	*1.002/*2.011 (IM)	*1.002 + rs4986894/*2.001 (IM)	*1/*2 (IM)	*1/*2
31	*1.002/*2.011 (IM)	*1.002 + rs4986894/*2.001 (IM)	*1/*2 (IM)	*1/*2
32	*1.002/*2.011 (IM)	*1.002 + rs4986894/*2.001 (IM)	*1/*2 (IM)	*1/*2
33	*1.002/*1.002 (NM)	*1.002/*1.002 (NM)	*1/*1 (NM)	*1/*1
34	*1.002/*2.011 (IM)	*1.002 + rs4986894/*2.001 (IM)	*1/*2 (IM)	*1/*2
35	*1.002/*38.003 + 94781616A>G (NM)	*1.001 + rs17885098/*1.005 (NM)	*1/*1 (NM)	*1/*5
36	*2.011/*17.001 (IM)	*2.001/*17.001 + rs4986894 (IM)	*2/*17 (IM)	*2/*4 (*2/*17)[a]

TABLE 2 (*Continued*) Overview of haplotype prediction by different *-allele tools. The 37 samples were assessed by different *-allele tools (Aldy, PharmaKU and PharmCat) and manually. The predicted haplotypes are depicted for each sample.

Samples	Manual, *-alleles (predicted phenotype)	Aldy, *-alleles (predicted phenotype)	PharmaKU, *-alleles (predicted phenotype)	PharmCAT[b]
37	*1.002[b]*38.003 + 94781616A>G (NM)	*1.001 + rs17885098/*1.005 (NM)	*1/*1 (NM)	*1/*5

[a]The haploype predictions depicted are the first predictions in the output list. The haplotype prediction between brackets is added, when the prediction corresponding to the manual assessment, was present in the list but not as first output.
[b]No phenoypes are assigned for PharmCat due to ambiguity in haplotype assignments.
NM, normal metabolizer; IM, intermediate metabolizer; RM, rapid metabolizer.

overlapped for 81.1% with our manual assignment (*n* = 30). Unfortunately, the Aldy output does not include predicted phenotypes or drug guidelines.

PharmaKU does include phenotype predictions and guidelines and predicted 83.8% of the major diplotype calls accurately compared to the manual analysis (*n* = 31) (Table 2). The simple upload screen makes PharmaKU an easy and clear tool to use. Unfortunately, the program behind PharmaKU is a black box compared to the transparent script of Aldy as the tool is fully webbased. After uploading the VCF file of interest, the output file starts downloading automatically. Moreover, the analysis only includes major *-alleles; minor alleles are not taken into account.

The final *-allele tool tested was PharmCat. Since this tool reports all possible haplotypes for a sample, with only 34 variants as reference database, the prediction accuracy was rather poor compared to the manual prediction (45.9% overlap, *n* = 17) when selecting the first reported diplotype (Table 2). This is due to the software's algorithm which assumes any genomic positions not present in the VCF file to be missing and not wildtype. With background knowledge or a prior manual assessment, the accurate prediction was frequently present in the output list (only not as the first *-allele assignment) increasing the accuracy to 86.5% compared to the manual assignment. Similar to PharmaKU, the output only includes core *-alleles, but it does include drug guidelines and hyperlinks to literature with more information.

The high degree of phasing in these samples was shown to be of importance in assigning the *CYP2C19* haplotypes for multiple samples (e.g., sample 25 and sample 27). These samples were manually assigned with a *CYP2C19**1.002/*2.011 diplotype. However, when using *-allele tool Aldy, these samples were predicted to be *CYP2C19**1.002 + rs4986894/*2.001 (Supplementary Figure S2). Rs4986894 corresponds to g.94762608T>C, which is the variant responsible for the distinction between *CYP2C19**2.001 and *CYP2C19**2.011; *CYP2C19**2.001 + rs4986894 is *CYP2C19**2.011. Due to the 100% phasing in those samples (Figure 3), it can be concluded with certainty that the rs4986894 is present on the *CYP2C19**2 allele. Hence, the correct haplotype would be *CYP2C19**1.002/*2.011. Other tools (PharmCat and PharmaKU) both assigned *1/*2 as predicted haplotype to these samples, as these tools do not use minor alleles the discrepancy was not seen there).

Overall, 35% of the individuals were regarded as normal metabolizers based on the manual assignments. A further 27% was IM and 38% was classified as RM. No individuals were classified as PM or UM. PharmCat phenotypes were not assigned due to the ambiguity in the genotyping as described above. For the

remaining tools there was no discrepancy on a phenotype level (Table 2).

4 Discussion

In current routine PGx, phenotype predictions are most commonly based on SNV array data. Despite the advantages of this method, such as rapid turnaround times and a straightforward interpretation, it also has limitations; it is not possible to detect all variants and direct (read based) phasing cannot be performed. In the present study, we assessed the performance of long-read sequencing and *-allele tools for the characterization of *CYP2C19*. We showed that with long-read sequencing, novel PGx variants can be identified and that the majority (on average 76% for the *CYP2C19* locus) of these variants can be phased to their allele of origin. Moreover, we have been able to identify a potentially impactful variant currently not used in PGx nomenclature as well as a novel minor allele of *CYP2C19**38. Finally, the majority of the assessed *-allele tools result in accurate predicted phenotype assignments while the diplotypes did differ between the different tools. These findings show the benefits and potential of applying long-read sequencing in PGx for *CYP2C19* in a research setting and provide a glimpse into the future of clinical PGx.

Based on the long-read sequencing data, a missense variant (g.94842860C>T, *T* = 0.00013 in Caucasians (Medicine Nlo. rs59734894, 2022) in *CYP2C19* could be identified which, according to literature, decreases *CYP2C19* expression but is not (yet) included in the current PGx databases. Since the assigned diplotype of the patient carrying g.94842860C>T is *1/*17 (RM), which is associated with a higher activity of CYP2C19, carrying this missense mutation might have major consequences. It is possible that carrying this mutation reverses the higher enzyme activity due to *CYP2C19**17, making this patient a normal or even intermediate metabolizer to whom different drug dosages apply than the recommended dose for a rapid metabolizer. Hence, this variant warrants further study and should, if previous findings are confirmed, be added to the PGx nomenclature. The same holds true for other variants identified. For example, splice site variants can have a high impact since they might result in aberrant proteins (Riolo et al., 2021) and a non-functional CYP2C19 enzyme. Intriguingly, none of the splice site variants detected in *CYP2C19* (*n* = 2) are present in the current PGx databases. One of these variants (g.94781806A>G) is located in intron four on the junction with exon five and the other (94852716A>G) in intron eight on the junction with exon 9 (Kent et al., 2002). No literature describing the potential impact of the variants could be found and they were

assigned CADD scores below 10. Moreover, only a few specific intron variants are recorded in PharmVar. The emphasis lies on variants affecting the coding sequence, even though it has been proven that intronic variants can potentially create or disrupt a splice site, affecting the enzyme activity (Ingelman-Sundberg and Sim, 2010).

Using the CADD scores to identify potentially deleterious variants resulted in four variants with a CADD score of 10 or higher. One of these variants was a well-known clinical variant (the g.94781859G>A, *2), another was confirmed by literature to be potentially deleterious. The remaining two (one synonymous and one downstream) have an unknown impact. It should be kept in mind that many pharmacovariants have small effects (decreased function) which do not result in completely inactive or absent protein. These smaller effects do add up in a clinical setting but are difficult to predict with available *in silico* tools. Hence, high CADD scores might be useful to identify potential high impact variants but on the other hand a low CADD score does not mean that the variant has no impact on enzyme function.

Besides applying long-read sequencing for variant identification, it also enabled us to phase the identified variants back to their allele of origin using read backed phasing with WhatsHap. The variability in the phasing percentages can have two major causes. First, the read length might not have been sufficient due to DNA fragmentation, resulting in limited coverage. As we selected samples based on overall quality this is not likely to be the major cause. On the other hand, this fragmentation and lack of coverage can be the cause of the 11 samples that did not meet our criteria. The second reason for low phasing is a possible lack of heterozygous variants; to phase two reads together, forming a phasing haploblock, at least one heterozygous variants is needed.

Moreover, it is important to keep in mind that read backed phasing differs from statistical phasing, resulting in possible discrepancies between the *-allele assignments. This was seen in our study for multiple samples, where the manual assignment based on read backed phased VCF and BAM files was *CYP2C19* *1.002/*2.011 while *-allele tool Aldy, using statistical phasing, assigned *CYP2C19* *1.002 + rs4986894/*2.001. Conventionally, haplotypes are assigned using statistical phasing based on population statistics which might not be accurate for the individual. For those haplotype assignments, linkage disequilibrium (LD) is used. However, it is known that some variants can occur separately despite their strong linkage disequilibrium (van der Lee et al., 2020b) and the statistical phasing depends on the LD-threshold that is set (e.g., $r^2 > 0.8$ or $r^2 > 0.85$), which results in differences between the predictions. Read backed phasing, as with long-reads, is more accurate for individual patients.

Finally, the long-read sequencing data was used to predict the phenotype of every sample and to explore the performance of different *-allele tools. For six out of the 37 patients, the manual *-allele assignment identified a *CYP2C19**38.003 haplotype. However, in all six cases, the clinical variant g.94781616A>G (rs7088784) was found on the same allele as *CYP2C19**38.003, confirmed by closer analysis of the BAM files. This variant is currently only associated with *CYP2C19**3.002 (PharmVar, 2022). In combination with the high frequency of this occurring in our cohort (16%), this might point towards a novel suballele of *CYP2C19**38. This new minor allele would include all the

variants of *CYP2C19**38.003 plus g.94781616A>G, the shared variant with *CYP2C19**3.002. It should be noted that g.94781616A>G is not a core SNV of *CYP2C19**3, nor is it a variant with a high impact; according to PharmVar the variant impact is 'No Change'. This would point towards a new minor allele without a clinical impact.

Interestingly, the *-allele assignment by *-allele tools did not show g.94781616A>G in their output at all. As it is only a minor mutation of *CYP2C19**3.002, it is only reported if its parent allele is present. The absence of g.94781616A>G in the *-allele tool output illustrates how the *-allele tools work; they return the diplotype prediction in which most variants are accounted for and generally do not mention all additional variants observed, making the identification of possible new *-alleles by using *-allele tools challenging. Moreover, *-allele tools do not always agree as 45% of the calls in this study were the same in all tools based on the major alleles and the first diplotype presented by PharmCat. After manually checking the PharmCat diplotypes this agreement increased to 81%. This is largely caused by different assumptions made by the tools and the various (updates of) reference databases they use. For example, for one individual sample PharmCat predicted *17/*38, while *17/*1.001 was predicted by Aldy. This discrepancy can be explained by the fact that *38 is not included in the Aldy3 database. After adding *38 to Aldy's database, all *38 alleles were detected and predicted accurately. On a phenotype level, there were no discrepancies between manual assignments, Aldy and PharmaKU indicating that the difference between the tools for our cohort were minor and not of clinical influence. However, care should still be taken as the difference could be of clinical relevance when they concern rare variants which are part of *-alleles nomenclature but maybe not yet present in all the tools algorithms. Furthermore, the most applicable tool will depend on the skills and interests of the user. For a physician, a straightforward tool with an accurate prediction and clear output would be ideal. Hence, physicians might favor PharmaKU. However, for the researcher with a grasp of bioinformatics and the urge to understand every detail, the more transparent and adaptable *-allele tool Aldy would be the first choice. Lastly, it is important to mention that most *-allele tools are dynamic tools and, as a result, are subject to change and updates.

While we were able to identify interesting variants and haplotypes, the sample size of this cohort is small. An analysis of more samples would yield more information and increase the power of our study. However, despite the small cohort of this study, we already identified novel variants and a novel minor allele. The fact that this is possible with only a limited number of samples strengthens our analysis; repeating this study for a larger cohort will presumably result in even more discoveries. Hence, it is anticipated that, by using long-read sequencing, new clinically relevant PGx variants and haplotypes will be discovered in the near future.

It is important to note the difference in requirements between clinical PGx and PGx in a research setting. In a clinical setting, quick results of well-known variants are needed to guide drug treatment. Although long-read sequencing is decreasing rapidly in costs and in turn-around time, it is not yet close to the SNP assays used in clinical practice. Therefore, the role of long-read sequencing in current PGx will be limited to those cases where conventional PGx cannot

provide an answer or to cases where sequencing data is available and PGx can be extracted as a bonus. Nonetheless, as pre-emptive genotyping is slowly being adapted, the longer turn-around time of sequencing based PGx will become less of an issue and the added data that can be obtained will be a major benefit. Therefore, we envision that in the near future long-read sequencing based clinical PGx will become more common.

Non-etheless, one major limitation does remain: the impact of novel variants and haplotypes. While more and more variants are identified in sequencing based studies, the clinical impact of these variants is still unknown. Meanwhile, most clinical studies focus only on well-known and established variants with the use of SNP panels. The clinical data collected in these studies is of crucial importance to assign functional effects to novel variants and haplotypes. Efforts should be made to integrate these type of studies by adding sequencing to clinical PGx studies and clinical data to sequencing studies when possible. The same holds true for a clinical setting, when sequencing data is used for clinical PGx the outcomes of the patients treatment can be used to inform researchers and clinicians on potential effects of novel variants. Vice versa, sequencing data can help to identify novel and potentially deleterious variants which might cause an unexpected drug response.

The establishment of more cohorts with clinical data and advanced genetic data also offers the opportunity to develop better phenotype prediction tools. *-Allele tools offer the opportunity to assign haplotypes known within the current nomenclature. However, the current categorical system is unable to account for small individual variant effects and relies on classifying an individual into limited phenotype categories. Previous studies have shown that using more advanced phenotyping methods which predict drug response on a continuous scale substantially improves the explained variability of the drug metabolizing enzyme CYP2D6 (McInnes et al., 2020; van der Lee et al., 2021). For CYP2C19, similar models could be developed based on (long-read) sequencing data and clinical outcomes. However, datasets which have both advanced genetic data and clinical outcomes are limited.

In conclusion, our study highlights the value of long-read sequencing for PGx in regard to accurate phenotype prediction. It shines a light on the possible role that long-read sequencing can play, together with *-allele tools, in future clinical PGx and in research. This study only shows the tip of the iceberg and highlights that a new focus on computational tools and big data is required to ultimately improve our ability to predict drug metabolism and thereby drug outcomes for the individual patient.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: European Genome-Phenome Archive [https://ega-archive.org/], EGAS00001006929.

Author contributions

Designed the research: MV, SK. Performed the analyses: LG, QZ, RM, RV. Processed the samples: LB, MV. Wrote the manuscript: LG. Reviewed and edited the manuscript: MV, SK, QZ, RM, RV, LB.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1076574/full#supplementary-material

References

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

Borràs, D. M., Vossen, R., Liem, M., Buermans, H. P. J., Dauwerse, H., van Heusden, D., et al. (2017). Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing. *Hum. Mutat.* 38 (7), 870–879. doi:10.1002/humu.23223

Caudle, K. E., Dunnenberger, H. M., Freimuth, R. R., Peterson, J. F., Burlison, J. D., Whirl-Carrillo, M., et al. (2017). Standardizing terms for clinical pharmacogenetic test results: Consensus terms from the clinical pharmacogenetics implementation consortium (CPIC). *Genet. Med.* 19 (2), 215–223. doi:10.1038/gim.2016.87

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., et al. (2016). HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mutat.* 37 (6), 564–569. doi:10.1002/humu.22981

Dunnenberger, H. M., Crews, K. R., Hoffman, J. M., Caudle, K. E., Broeckel, U., Howard, S. C., et al. (2015). Preemptive clinical pharmacogenetics implementation: Current programs in five US medical centers. *Annu. Rev. Pharmacol. Toxicol.* 55 (1), 89–106. doi:10.1146/annurev-pharmtox-010814-124835

Gaedigk, A., Ingelman-Sundberg, M., Miller, N. A., Leeder, J. S., Whirl-Carrillo, M., Klein, T. E., et al. (2018). The pharmacogene variation (PharmVar) consortium: Incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clin. Pharmacol. Ther.* 103 (3), 399–401. doi:10.1002/cpt.910

GATK (2023). Gatk: Broad Institute. Available at: https://gatk.broadinstitute.org/hc/en-us.

Han, S. M., Park, J., Lee, J. H., Lee, S. S., Kim, H., Han, H., et al. (2017). Targeted next-generation sequencing for comprehensive genetic profiling of pharmacogenes. *Clin. Pharmacol. Ther.* 101 (3), 396–405. doi:10.1002/cpt.532

PharmGKB (2023). CYP2C19 clinical annotations: PharmGKB. Available at: https://www.pharmgkb.org/gene/PA124/clinicalAnnotation.

Hicks, J. K., Sangkuhl, K., Swen, J. J., Ellingrod, V. L., Müller, D. J., Shimoda, K., et al. (2017). Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update. *Clin. Pharmacol. Ther.* 102 (1), 37–44. doi:10.1002/cpt.597

Ingelman-Sundberg, M., Daly, A. K., Oscarson, M., and Nebert, D. W. (2000). Human cytochrome P450 (CYP) genes: Recommendations for the nomenclature of alleles. *Pharmacogenetics* 10 (1), 91–93. doi:10.1097/00008571-200002000-00012

Ingelman-Sundberg, M., and Sim, S. C. (2010). Intronic polymorphisms of cytochromes P450. *Hum. Genomics* 4 (6), 402–405. doi:10.1186/1479-7364-4-6-402

John, S. E., Channanath, A. M., Hebbar, P., Nizam, R., Thanaraj, T. A., and Al-Mulla, F. (2021). PharmaKU: A web-based tool aimed at improving outreach and clinical utility of pharmacogenomics. *J. Personalized Med.* 11 (3), 210. doi:10.3390/jpm11030210

Kent, W. J. S. C., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). USCS genome browser 2002. Available at: https://genome.ucsc.edu/.

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

Lima (2023). Lima. Available at: https://lima.how/.

Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G. W., et al. (2016). WhatsHap: Fast and accurate read-based phasing. *bioRxiv* 2016, 085050. doi:10.1101/085050

McInnes, G., Dalton, R., Sangkuhl, K., Whirl-Carrillo, M., Lee, S. B., Tsao, P. S., et al. (2020). Transfer learning enables prediction of CYP2D6 haplotype function. *PLoS Comput. Biol.* 16 (11), e1008399. doi:10.1371/journal.pcbi.1008399

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The Ensembl variant effect predictor. *Genome Biol.* 17 (1), 122. doi:10.1186/s13059-016-0974-4

Midha, M. K., Wu, M., and Chiu, K-P. (2019). Long-read sequencing in deciphering human genetics to a greater depth. *Hum. Genet.* 138 (11), 1201–1215. doi:10.1007/s00439-019-02064-y

Mukerjee, G., Huston, A., Kabakchiev, B., Piquette-Miller, M., van Schaik, R., and Dorfman, R. (2018). User considerations in assessing pharmacogenomic tests and their clinical support tools. *NPJ Genom Med.* 3, 26. doi:10.1038/s41525-018-0065-4

Multiqc (2023). Multiqc. Available at: https://multiqc.info/.

Medicine Nlo. rs59734894 (2022). Reference SNP (rs) Report. Available at: https://www.ncbi.nlm.nih.gov/snp/rs59734894.

Numanagić, I., Malikić, S., Ford, M., Qin, X., Toji, L., Radovich, M., et al. (2018). Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nat. Commun.* 9 (1), 828. doi:10.1038/s41467-018-03273-1

PacBio (2023). HiFi sequencing. Available at: https://www.pacb.com/technology/hifi-sequencing/.

Pbmm2 (2023). Pbmm2. Available at: https://github.com/PacificBiosciences/pbmm2.

Pdmarkdup (2023). Pdmarkdup. Available at: https://github.com/PacificBiosciences/pbmarkdup/.

PharmVar (2022). Pharmacogene Variation Consortium [Web Page]. Available at: https://www.pharmvar.org/(Accessed 31 05, 2022).

PharmGKB (2022). Annotation of rs59734894: Stanford university. Available at: https://www.pharmgkb.org/variantAnnotation/1451143360.

Piriyapongsa, J., Sukritha, C., Kaewprommal, P., Intarat, C., Triparn, K., Phornsiricharoenphant, K., et al. (2021). PharmVIP: A web-based tool for pharmacogenomic variant analysis and interpretation. *J. Pers. Med.* 11 (11), 1230. doi:10.3390/jpm11111230

Pratt, V. M., Del Tredici, A. L., Hachad, H., Ji, Y., Kalman, L. V., Scott, S. A., et al. (2018). Recommendations for clinical CYP2C19 genotyping allele selection: A report of the association for molecular pathology. *J. Mol. Diagnostics* 20 (3), 269–276. doi:10.1016/j.jmoldx.2018.01.011

Pratt, V. M., Turner, A., Broeckel, U., Dawson, D. B., Gaedigk, A., Lynnes, T. C., et al. (2021). Characterization of reference materials with an association for molecular pathology pharmacogenetics working group tier 2 status: CYP2C9, CYP2C19, VKORC1, CYP2C cluster variant, and ggcx: A GeT-RM collaborative project. *J. Mol. Diagn* 23 (8), 952–958. doi:10.1016/j.jmoldx.2021.04.012

Qiao, W., Yang, Y., Sebra, R., Mendiratta, G., Gaedigk, A., Desnick, R. J., et al. (2016). Long-read single molecule real-time full gene sequencing of cytochrome P450-2D6. *Hum. Mutat.* 37 (3), 315–323. doi:10.1002/humu.22936

Redmar van den Berg, J. (2021). *LUMC/PacBio-variantcalling: v3.0.0.* Zenodo.

Rentzsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-splice—Improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* 13 (1), 31. doi:10.1186/s13073-021-00835-9

Riolo, G., Cantara, S., and Ricci, C. (2021). What's wrong in a jump? Prediction and validation of splice site variants. *Methods Protoc.* 4 (3), 62. doi:10.3390/mps4030062

Sangkuhl, K., Whirl-Carrillo, M., Whaley, R. M., Woon, M., Lavertu, A., Altman, R. B., et al. (2020). Pharmacogenomics clinical annotation tool (PharmCAT). *Clin. Pharmacol. Ther.* 107 (1), 203–210. doi:10.1002/cpt.1568

Scharfe, C. P. I., Tremmel, R., Schwab, M., Kohlbacher, O., and Marks, D. S. (2017). Genetic variation in human drug-related genes. *Genome Med.* 9 (1), 117. doi:10.1186/s13073-017-0502-5

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29 (1), 308–311. doi:10.1093/nar/29.1.308

Swen, J. J., Wilting, I., de Goede, A. L., Grandia, L., Mulder, H., Touw, D. J., et al. (2008). Pharmacogenetics: From bench to byte. *Clin. Pharmacol. Ther.* 83 (5), 781–787. doi:10.1038/sj.clpt.6100507

Tafazoli, A., Guchelaar, H. J., Miltyk, W., Kretowski, A. J., and Swen, J. J. (2021). Applying next-generation sequencing platforms for pharmacogenomic testing in clinical practice. *Front. Pharmacol.* 12, 693453. doi:10.3389/fphar.2021.693453

Upgx (2020). Ubiquitous pharmacogenomics. Available at: https://upgx.eu/.

van der Lee, M., Allard, W. G., Bollen, S., Santen, G. W. E., Ruivenkamp, C. A. L., Hoffer, M. J. V., et al. (2020). Repurposing of diagnostic Whole exome sequencing data of 1,583 individuals for clinical pharmacogenetics. *Clin. Pharmacol. Ther.* 107 (3), 617–627. doi:10.1002/cpt.1665

van der Lee, M., Allard, W. G., Vossen, R., Baak-Pablo, R. F., Menafra, R., Deiman, B., et al. (2021). Toward predicting CYP2D6-mediated variable drug response from CYP2D6 gene sequencing data. *Sci. Transl. Med.* 13 (603), eabf3637. doi:10.1126/scitranslmed.abf3637

van der Lee, M., Kriek, M., Guchelaar, H. J., and Swen, J. J. (2020). Technologies for pharmacogenomics: A review. *Genes (Basel).* 11 (12), 1456. doi:10.3390/genes11121456

van der Lee, M., Rowell, W. J., Menafra, R., Guchelaar, H. J., Swen, J. J., and Anvar, S. Y. (2022). Application of long-read sequencing to elucidate complex pharmacogenomic regions: A proof of principle. *Pharmacogenomics J.* 22 (1), 75–81. doi:10.1038/s41397-021-00259-z

van der Lee, M. (2021). *Towards solving the missing heritability in pharmacogenomics [Doctoral dissertation].* Leiden, The Netherlands: Leiden University Scholarly Publications.

van der Wouden, C. H., van Rhenen, M. H., Jama, W. O. M., Ingelman-Sundberg, M., Lauschke, V. M., Konta, L., et al. (2019). Development of the PGx-passport: A panel of actionable germline genetic variants for pre-emptive pharmacogenetic testing. *Clin. Pharmacol. Ther.* 106 (4), 866–873. doi:10.1002/cpt.1489

Weinshilboum, R., and Wang, L. (2004). Pharmacogenomics: Bench to bedside. *Nat. Rev. Drug Discov.* 3 (9), 739–748. doi:10.1038/nrd1497

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P-C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37 (10), 1155–1162. doi:10.1038/s41587-019-0217-9