The Pharmacogenomics Journal www.nature.com/tpj

ARTICLE OPEN

Application of long-read sequencing to elucidate complex pharmacogenomic regions: a proof of principle

Maaike van der Lee[1,2], William J. Rowell[3], Roberta Menafra[4], Henk-Jan Guchelaar[1,2], Jesse J. Swen[1,2]✉ and Seyed Yahya Anvar[1,2,4,5]✉

The use of pharmacogenomics in clinical practice is becoming standard of care. However, due to the complex genetic makeup of pharmacogenes, not all genetic variation is currently accounted for. Here, we show the utility of long-read sequencing to resolve complex pharmacogenes by analyzing a well-characterised sample. This data consists of long reads that were processed to resolve phased haploblocks. 73% of pharmacogenes were fully covered in one phased haploblock, including 9/15 genes that are 100% complex. Variant calling accuracy in the pharmacogenes was high, with 99.8% recall and 100% precision for SNVs and 98.7% precision and 98.0% recall for Indels. For the majority of gene-drug interactions in the DPWG and CPIC guidelines, the associated genes could be fully resolved (62% and 63% respectively). Together, these findings suggest that long-read sequencing data offers promising opportunities in elucidating complex pharmacogenes and haplotype phasing while maintaining accurate variant calling.

The Pharmacogenomics Journal (2022) 22:75–81; https://doi.org/10.1038/s41397-021-00259-z

INTRODUCTION

Pharmacogenomics (PGx) is crucial for individualizing drug dosages and thereby improving drug therapy outcomes [1, 2]. PGx relies on inferred phenotypes based on known variants in pharmacogenes. Nonetheless, not all genetic variability in drug response and enzyme activity can be explained by routine PGx genetic assays [3, 4], due to several factors. First, current genotyping assays are unable to fully resolve the genetic makeup of all genes involved in drug response [5–7]. Second, the mechanism of action of a drug and/or its metabolic pathway is not always fully understood [4, 8]. It is essential to be able to explain all genetic components driving variable drug response in order to assess what part of variability is genetic and what part can be explained by other factors. This is, however, challenged as the majority of pharmacogenes are at least in part located in complex genomic regions or contain variants like tandem-repeats and pseudogene hybrid conformations [9]. Currently applied genotyping technologies are based either on SNV (Single Nucleotide Variant) microarrays or short-read sequencing [10, 11]. Both approaches are limited in characterizing these complex regions [12–15], as they fail to adequately and reliably resolve highly homologous regions and identify PGx variants [7, 16, 17]. Moreover, with haplotype phasing it could be determined if variants are located on the same allele or if they are on different alleles, potentially leading to differences in phenotype assignment. Currently, PGx diplotypes are phased based on linkage disequilibrium. While this results in accurate haplotypes on a population scale it does not always result in accurate assumptions on an individual level. The impact of these challenges in clinical practice is high [5]. For example, the complex gene *CYP2D6*, is involved in the metabolism of 20–30% of commonly prescribed drugs [18] and cannot be fully characterized by short-read sequencing.

In recent years the long-read sequencing technologies from Oxford Nanopore and PacBio have shown to be capable of characterizing complex (pharmaco)genomic regions [19–21]. For these regions, long and high-quality reads significantly improve variant calling precision and allow for resolution of fully phased diplotypes.

The value of long-read sequencing for disease diagnostic purposes has previously been illustrated [7, 16, 22–26]. PacBio sequencing has been shown capable of characterising *CYP2D6*, by covering the entire gene locus in one high-quality long read [7, 16, 26–29]. More recently, long-read sequencing has also been applied for the HLA genes in relation to PGx [29, 30]. In addition, its application has been used in numerous challenging clinical diagnostic research assays such as long tandem repeat in *FMR1* gene linked to Fragile X syndrome [22] and in resolving the *PKD1* gene to detect mutations associated with polycystic kidney disease [23]. Finally, long-read sequencing facilitates haplotype phasing without the need for computational approaches and/or pedigree information. This can be of crucial importance in PGx leading to more accurate phenotype predictions [15]. The combination of PGx complexity and haplotype phasing indicates that long-read sequencing has the potential to substantially improve our ability to correctly predict drug metabolizer phenotypes. In this proof-of-concept paper, we assess the potential of long-read PacBio sequencing to resolve complex PGx regions by using available sequencing data of the well-characterised Genome in a Bottle (GIAB) reference sample HG002.

[1]Department of Clinical Pharmacy and Toxicology, Leiden University Medical Center, Leiden, the Netherlands. [2]Leiden Network of Personalized Therapeutics, Leiden, the Netherlands. [3]Pacific Biosciences, Menlo Park, CA, USA. [4]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [5]Present address: OKRA Technologies, Cambridge, UK. ✉email: j.j.swen@lumc.nl; yahya.anvar@okra.ai

Published online: 5 November 2021

RESULTS

Data description

Previously published sequencing data of the well-characterized HG002 GIAB sample were obtained [19]. This data consists of 6,728,123 reads with a median length of 13.4 kb, covering 97.5% of the genome (Fig. 1) with an average mapped coverage of 28-fold. Approximately 5 million genetic variants were detected using GATK (Genome Analysis Toolkit) HaplotypeCaller [31] and DeepVariant [32].

High precision and recall in variant calling

For the 100 selected pharmacogenes, precision and recall compared to the benchmark truth set GIAB v3.3 was determined. For SNVs, GATK HaplotypeCaller and DeepVariant achieved similar precision and recall above 99.8% (Table 1). However, the DeepVariant caller achieved a much better performance in detecting indels (>98%) compared to GATK (precision: 94.5% and recall: 86.1%). When comparing to the genome wide results reported by Wenger et al, the precision and recall in detecting variants in the pharmacogenes are superior [19]. When stratifying results on complex regions (Table S2), accuracy remained high, with recall and precision >95% for all regions for both indels and SNVs. For the GATK caller, the accuracies were lower, (85-100% compared to 97-100% for DeepVariant caller). The drop in accuracy could be attributed to lower performance for tandem repeats and homopolymers (Table S2 and Fig. S1).

To assess the accuracy of SV calling in pharmacogenes, SV calls were compared with the SV benchmark set for all SVs over 50 bp. However, the high confidence GIAB regions did not cover all 100 genes. 46 genes were excluded, 12 genes were partially and 42 were fully overlapping with the GIAB curated data (Table S3). In total, 22 SVs (>50 bp) were identified in the 54 pharmacogenes compared to 23 catalogued in the benchmark set (Table S4). Two calls were regarded as false negative and one call as false positive. Together, assessing the performance of detecting SVs in PGx regions resulted in recall of 91.3% and precision of 94.5%. The high recall and precision in pharmacogenes suggest that there is no loss of accuracy with the use of long-read sequencing data compared to current benchmarks, whilst improving the detection of complex genetic variants.

Haplotype phasing and haploblocks

Using WhatsHap [33], reads were phased and resolved into haploblocks based on all identified variants. Each haploblock describes one stretch of fully phased sequence allowing for a complete characterisation of that region, representing a maternal or paternal allele. Notably, 71.2% of the genome could be phased into 16,193 haploblocks with a total haploblock length of 2.3 billion base pairs and a median haploblock size of 40,302 bp (range: 1–2.9 million bp). A clear distinction in haploblock size was observed between intergenic regions (median of 14,960 bp) and Gencode features (median of 56,743 bp), Fig. 2A. The vast majority of Gencode features was fully phased into haploblocks (Fig. S2 and S3). In particular, 71% of all protein coding features could be completely phased (≥90%) and an additional 22% were partially phased while 7% remained unresolved (≤10% phased). Similar patterns were observed for other Gencode features (Figs. S2 and S3). Read length does not seem to be the main limiting factor in resolving haplotypes as the percentage of a feature covered in haploblocks is independent of feature length (Fig. 2B). In addition, the majority of haploblocks (57.7%) exceed the median read length, indicating that not read length but the number of heterozygous variants and number of reads aligned to a given genomic region are the limiting factors in haploblock construction.

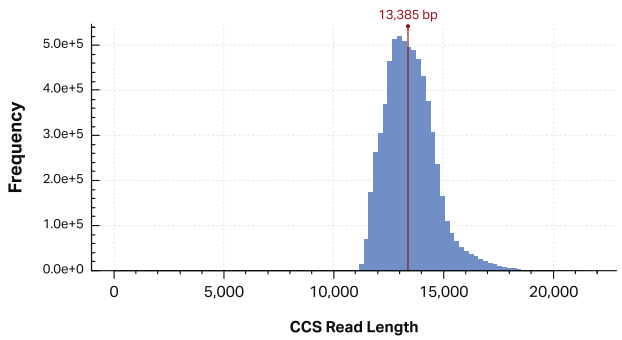

Fig. 1 Read length distribution. Distribution of read length of genome in a bottle sample HG002 after sequencing on Pacific Bioscience sequel platform and construction of circular consensus sequence.

Pharmacogenes

For each of the 100 selected pharmacogenes the portion of the genes located in a complex region was determined—with complex defined as genomic regions that overlap with segmental duplications (SD) or repeats. In total, 15 pharmacogenes were classified as 100% complex whereas eight pharmacogenes did not show any overlap with SDs or repeats (Fig. 3A).

For each of the 100 loci, almost all variants could be accurately called (precision and recall >99.8%). Subsequent phasing resulted in haploblocks with a median length of 140,473 bp, resulting in the majority (73/100) of the features being fully phased into haploblocks (Fig. 3A). Most significantly, of the 15 pharmacogenes classified as fully complex, 9 could be fully phased, 4 for at least 60% and the last two could not be phased. Of the notoriously complex HLA-genes, 35 out of 37 were fully resolved, the remaining two (*HLA-DRB5* and *HLA-DRB1*) were resolved for 6.4% and 67.1%, respectively.

Nonetheless, several important pharmacogenes could only be partially phased into haploblocks. For example, *G6PD*, *DPYD* and *CYP2C19* were resolved for 0%, 55% and 34%, respectively. As *G6PD* is located on the X chromosome and the individual sequenced is male, it is not possible to phase the locus into two alleles resulting in 0% of the locus being covered in phased haploblocks. For *DPYD* the cause lies in a combination of long gene length (~900,000 bp) and a low number of variants leading to large stretches without heterozygous variants resulting in broken haploblocks (Fig. S4). For *CYP2C19*, there is a large portion in the centre of the gene which is homozygous for all variants. More specifically, in the entire *CYP2C19* locus there are 52 variants of which 33 are homozygous, resulting in fragmented phased blocks (Fig. S4). Yet, as all regions have been sequenced, it is still possible to assign haplotypes and phenotypes using the current Dutch Pharmacogenetics Working Group (DPWG) and Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines and phasing assumptions.

To assess the clinical utility, diplotypes and phenotypes were assigned based on the variant panel from the Ubiquitous pharmacogenomics (U-PGx) consortium and a previously developed pipeline [15]. A total of 1,418 variants were identified in 10 key pharmacogenes included in the panel, from which, 38 variants were considered in the phenotyping panel (Table S5). Clinically relevant variants in the *CYP3A5*, *CYP2D6* and *VKORC1* genes were identified. For CYP3A5, the rs776746 (g.99672916 C > T) variant was found on both alleles resulting in a *CYP3A5**3/*3 genotype and a Poor Metabolizer phenotype. For *CYP3A5* a PM status is regarded as not actionable due to this being the most common phenotype in Caucasians. For *CYP2D6* and *VKORC1* the inferred phenotype was divergent from the wildtype. In the *CYP2D6* locus, both the rs3892097 (g.42128945 C > T) and the rs1065852 (g.42130692 G > A) variant were found to be heterozygous. With phasing, it was determined that the variants were located on the same allele resulting in a *CYP2D6**1/*4 diplotype and inferred

Table 1. Variant calling performance for pharmacogenes.

Variant caller	SNVs			Indels		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
GATK haplotype caller	99.88	99.96	99.92	94.47	86.12	90.10
DeepVariant (CCS model)	99.84	100.0	99.92	98.74	98.00	98.37

Measured against the Genome in a Bottle benchmark v.3.3.2. using both GATK variant caller and DeepVariant. *SNV* single nucleotide variant, *Indels* insertions and deletions, *GATK* genomic analysis toolkit, *CCS* circular consensus sequence.

CYP2D6 intermediate metabolizer (IM) phenotype. Moreover, given the presence of the non-functional *CYP2D7* pseudogene which shares >90% of its sequence with *CYP2D6*, it is of importance to exclude any interference of *CYP2D7* reads to accurately determine *CYP2D6* haplotypes [5]. The reads were sufficiently long to allow for a clear distinction between *CYP2D6* and *CYP2D7* without any ambiguously mapped reads (Fig. S4). The same was observed for *CYP2B6* and its pseudogene *CYP2B7P* and for the *CYP3A* locus of which all genes share high sequence homology (Fig. S4). For *VKORC1*, a homozygous variant (NC_000016.10:g.31093557 G > A) was identified, leading to the 1173TT genotype resulting in a decreased activity (Fig. S4). Overall, these results indicate that, according to publicly available consensus guidelines, this individual would require dose adjustments for drugs that are a substrate to CYP2D6 and VKORC1.

Clinical relevance

In total, 15 genes included in this study are represented in the CPIC and/or DPWG guidelines, resulting in a total of 56 and 67 gene-drug interactions for the DPWG and CPIC guidelines respectively (Fig. 3, Table S6). Of these genes 10 (66.7%) were completely resolved in phased haploblocks. The genes which were fully resolved are involved in 35 of the gene-drug interactions in DPWG and for 35 gene-drug interactions in CPIC. For the remaining genes, variants could still be accurately identified, allowing for haplotype assignments according to current clinical practice which uses non-phased genetic data.

DISCUSSION

In this proof-of-concept study, we have shown that long-read sequencing yields high quality variant calls in all selected pharmacogenes. Compared to the genome-wide analysis [19], results for PGx genes are superior with regards to variant calling accuracy and resolution of larger phased haploblocks. In addition, the majority of the selected pharmacogenes could be fully resolved in phased haploblocks.

Based on variant calling alone, long-read whole genome data can be used for routine PGx similar to the way NGS is used [15, 34, 35]. Moreover, long-read sequencing offers the benefit of resolving paternal and maternal alleles. Given the polymorphic nature of pharmacogenes the likelihood of one individual carrying multiple variants in one pharmacogene is extremely high [19, 36], increasing the importance of haplotype phasing. Additionally, this high abundance of variants resulted in significantly larger haploblocks for the pharmacogenes compared to Gencode features.

Long read sequencing is comparable to short-read sequencing in regards to SNV detection and performs better in regards to haplotype phasing and complex SVs [19]. Haplotype phasing can potentially make the difference between an inferred intermediate metabolizer phenotype (two truncating variants on the same allele) and a poor metabolizer phenotype (two truncating variants on different alleles). Current PGx haplotyping strategies utilize computational phasing, leading to accurate phasing on a population scale but not always on an individual level. As drug adjustments are made on an individual level, accuracy in regards to phasing for one individual is crucial [37]. Here we have shown that long-read sequencing enables the majority of pharmacogenes to be fully phased into haploblocks without the need for pedigree data or for computational phasing.

Long-read sequencing also offers a full characterization of every variant in the selected PGx loci, including structural and rare variants, as indicated by the high precision and recall for SNVs, Indels and SVs. For example, the median read length (13.4kbp) is approximately three times larger than the size of the *CYP2D6* locus (4.4kbp), which allows for full characterization of the locus and potential CNVs. The large difference between DeepVariant and GATK for Indels can be explained by the use of long-read PacBio CCS data for the training of the DeepVariant caller. GATK was designed with the error mode of short read sequencing as a basis, with ~100 times more substitutions then indels. DeepVariant on the other hand has learned the error mode from the PacBio HiFi training data, which has a ratio of 30 times more indels compared to substitutions [19, 32]. Specifically, Indels and tandem repeat identification is significantly improved with the use of long reads and DeepVariant [19, 25]. This difference highlights once more the added benefit of long reads over short read sequencing in regard to the identification of complex variants.

For the studied individual, 1418 SNVs were identified in the selected clinical PGx loci (10 genes) of which 94% were fully phased, indicating a high abundance of variants in the pharmacogenes. Moreover, the phased nature of this data can help improve our understanding of haplotypes and variant combinations. Thus, long-read sequencing technologies have the potential of transforming our knowledge of genetic factors that play a role in variable drug response.

Prior to implementation of long-read sequencing into clinical practice, tools to assist the interpretation are needed. Several groups have made efforts to develop such translational tools for PGx [38–40]. However, there are still limitations to these tools. First, they often cover the entire range of known variants and their associated haplotypes. However, not for every *-haplotype the clinical impact is known, therefore this will occasionally result in haplotype of which the effect is unknown making it difficult to implement in clinical practice [34]. Secondly, the tools do not always provide the same result for the same individual [34], indicating that the assumptions on which these tools are based are not comparable. To only include clinically relevant *-haplotypes in our analysis, we have limited our analysis of the clinical utility to the panel of variants defined by the U-PGx consortium. It should be noted, however, that this does lead to the exclusion of the majority of variants in all PGx loci, due to the fact that there is not yet sufficient knowledge about the function of these variants.

To illustrate the impact of long reads on clinical PGx we have assessed the sequencing results in the context of the DPWG and CPIC guidelines. Based on the genetic variants observed in the studied individual, the guidelines recommended drug or dose adjustment for 22 drugs. Of all gene-drug interactions in the guidelines (53 for DPWG and 54 for CPIC), the vast majority (35 for both) was associated with a (partial) complex gene which could be fully resolved in a haploblock. As we have shown in this study, long-read sequencing is capable of resolving these complexities

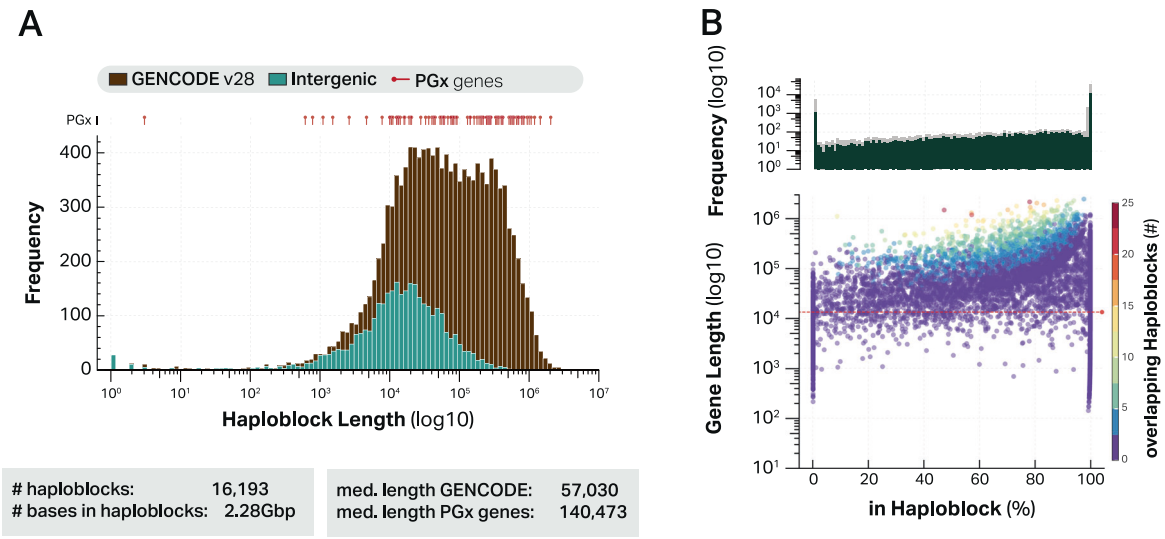

Fig. 2 Haploblock resolution of GENCODE features. A haploblock length distribution stratified by Gencode features and intergenic regions, overlap with pharmacogenes is highlighted in red. **B** For each protein coding feature the percentage that were resolved into haploblocks compared to the feature length. The red line reflects the mean read length. The majority of haploblocks are larger than the mean read length, indicating that not read length but the number of heterozygous variants is decisive for the length of a haploblock.

and constructing large haploblocks, allowing for more accurate haplotype calling. SNV panels and short-read sequencing, on the other hand, are capable of accurate variant identification but are limited in their ability to solve all complexities and in regard to haplotype phasing.

Nonetheless, it should be mentioned that not all pharmacogenes could be fully resolved. The key reason for this was the absence of heterozygous variants to allow for haploblock construction. This, in turn, leads to broken haploblocks and pharmacogenes which cannot be fully resolved. For the individual we studied this effect was apparent for *CYP2C19* and *DPYD* in particular. However, variant identification was still possible in the entire gene locus allowing for non-phased haplotype assignments. For these genes which could not be fully resolved, conventional haplotype approaches based on non-phased sequencing data can still be applied resulting in haplotype and phenotype predictions in line with current clinical practice. Moreover, for *DPYD* three of the four clinically relevant variants were still phased, two of which in the same haploblock. Indicating that a lack of complete phasing does not mean that none of the relevant variants can be phased. As the coverage was sufficient in all pharmacogenes, this lack of phasing is caused by the individuals genetic make-up, being a lack of heterozygous variants in this region, and not by the sequencing in itself, this is not easily resolved. For another individual the same problem of broken haploblocks might be observed in other genes depending on their genetics. While long-read sequencing for clinical pharmacogenomics seems promising, the costs and turn-around time associated with it are currently too high for potential high throughput PGx diagnostics [41]. Currently, this makes long-read sequencing not compatible with the quick SNV-arrays used in clinical PGx. However, sequencing costs are quickly decreasing. Moreover, pre-emptive genotyping becomes more popular which makes the longer turn-around time no longer an issue.

In this study, genetic data from a high-quality DNA sample was used. In clinical practice, high quality might not always be guaranteed. Nonetheless, previous applications of long-read sequencing in a clinical setting or with the use of clinically obtained DNA have resulted in good quality results [22–26]. Moreover, since 2020 a PacBio ultra-low DNA input workflow requiring only 5 ng of DNA has been available [42]. It is therefore expected that high quality sequencing results can be obtained with routinely collected clinical samples.

The accuracy and value of long-read sequencing has previously been investigated in whole genome data, which might make a targeted approach as we have presented here seem unnecessary [19]. However, it is well-established that the complexity of pharmacogenomic regions of the genome compromises the current assays in resolving their genetic makeup and thereby limiting the reliability and completeness of the phenotyping assays. The difference in genetic makeup of pharmacogenes compared to the general protein-coding genes makes the direct extrapolation from whole genome results unreliable. Most importantly, they contain more variants that together influence the drug response [19, 36, 43]. This high number of polymorphisms leads to the hypothesis that pharmacogenes can more easily be phased due to the higher abundance of heterozygous variants, as was confirmed in our study. Indeed, accuracy in the pharmacogenes was higher than that in other genes whereas short reads have a much lower accuracy in detecting genetic variants in these complex regions. The ability of long-read sequencing to resolve pharmacogenes was shown previously in targeted sequencing studies [7, 16, 26–30]. However, this study aimed at providing a comprehensive overview of the utility of long-read sequencing in resolving complex pharmacogenes and to inform on regions that remain challenging.

This study was limited to high quality data from a single subject and serves as a proof-of-concept for the application of long-read sequencing in PGx. Despite this limitation we feel that this is sufficient to serve as a proof-of-concept study investigating the potential of long-read sequencing for PGx. Based on these data regarding the variant calling accuracy and ability to resolve complex pharmacogene into phased haploblocks, we conclude that long-read sequencing data offers great opportunities to elucidate complex PGx loci and haplotype phasing while maintaining accurate variant calling in the selected pharmacogenes.

METHODS

Data description

Publicly available long-read sequencing results of GIAB sample HG002 was sequenced with PacBio sequencing and analysed with the use of CCS (Circular Consensus Sequencing) reads, were obtained [19]. A GIAB sample was selected as these are extremely well characterised with benchmark results available [44]. CCS reads were generated using CCS software v.3.0.0

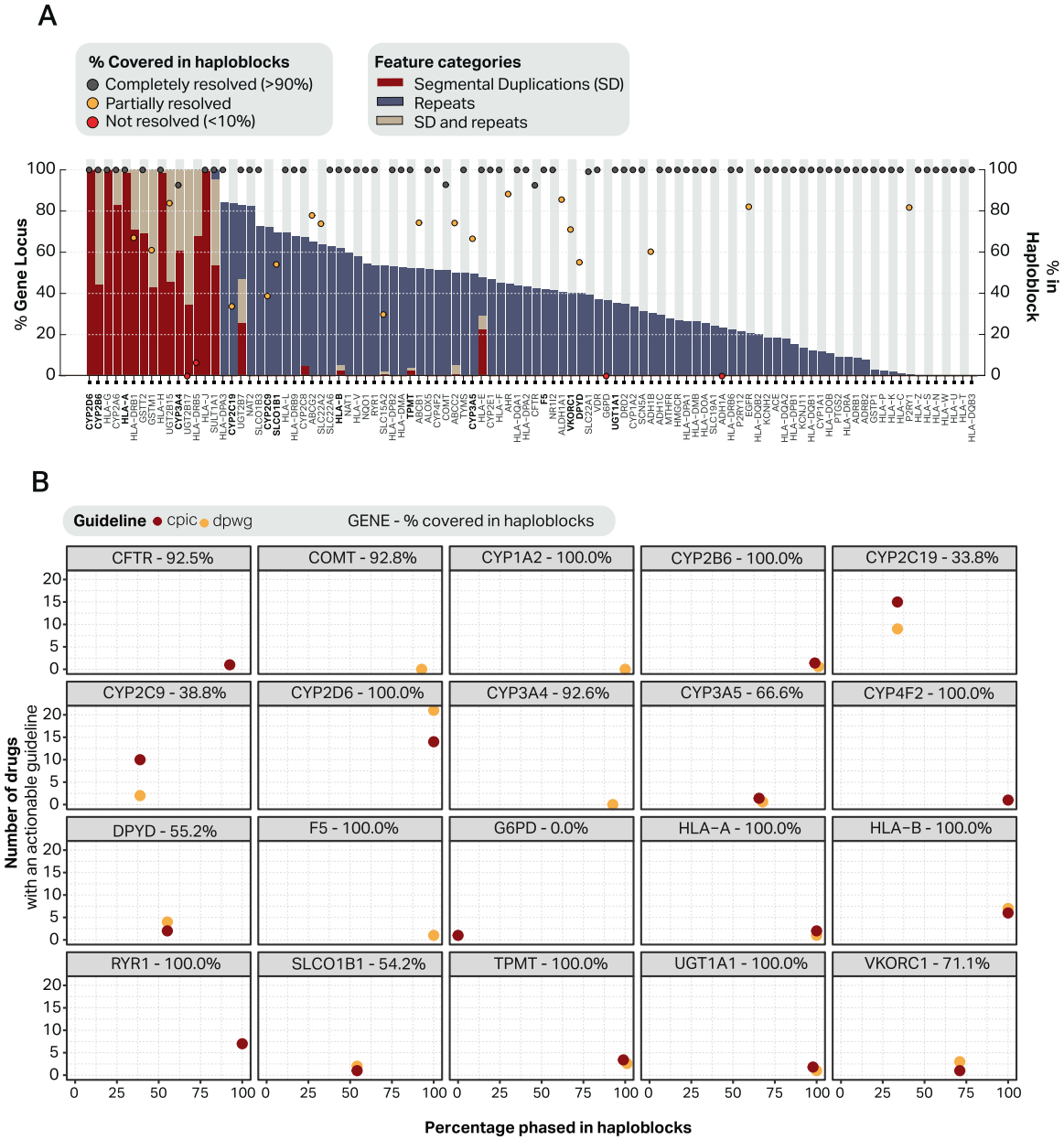

Fig. 3 Complexity of pharmacogenes and proportion solved in haploblocks. In (**A**), the pharmacogenes and their complexity related to the percentage covered in haploblocks. In bold genes included in the Ubiquitous pharmacogenomics (U-PGx) passport. **B** for genes included in the CPIC of DPWG guidelines the number of available actionable guidelines is mapped to the percentage of each gene which is phased into haploblocks. Actionable is defined as guidelines which recommends a dose change or drug switch. For each gene the percentage resolved in haploblocks is included in the panel headers. CPIC Clinical Pharmacogenetics Implementation Consortium, DPWG Dutch Pharmacogenetics Working group.

[19]. The obtained HiFi reads were aligned to GRCh38 reference genome using NGMLR aligner v0.2.7. Genetic variants were identified using GATK HaplotypeCaller (v.4.0.6.0) and DeepVariant (v.0.7.1). A set of 64 pharmacogenes that were previously described by Lauschke et al. [9] along with notoriously complex HLA-genes were selected for the PGx analysis (Table S1).

Haploblock constructing

Variants called by GATK were phased using WhatsHap [45] to obtain phased SNV and Indel variants. From the phased reads, haploblocks were constructed and stored in GTF and BED files. Each haploblock was constructed by matching phased reads based on the variants they contain in order to increase the length of the sequence which can be resolved. One haploblock represents one stretch of unbroken sequence based on overlapping phased reads and stops when a region in the genome is covered only by reads without any variants, there is no longer a difference in variants between the two alleles or if the region lacks coverage.

Subsequently, all loci were categorised into one of three features: Gencode features (v28), PGx genes and intergenic features. Where a feature is defined as an annotated genomic region such as protein-coding genes, segmental duplicated regions, pseudogenes, etc. Gencode reference annotation for genetic features in the human genome (release 28) was used to investigate haploblock resolution of important loci such as protein coding genes. The Gencode project aims to classify and identify all gene features in human genomes including all annotations [46]. For each autosomal Gencode and PGx feature, the percentage of the feature that is covered in a haploblock is calculated (number of basepairs in haploblocks/total feature length). Regions with ≥90% haploblock coverage are classified as fully phased, whereas regions with no overlapping haploblocks remain unphased. All other regions are marked as partially phased.

Segmental duplications (SD) and repeat tracks are obtained from UCSC (University of California Santa Cruz) Genome Browser. Bedtools was used to identify overlapping regions between all tracks and annotations files discussed. For each locus, the percentage of segments overlapping with SDs or repeats is defined as 'complex'.

Clinical relevance

A previously developed pipeline was employed to assign haplotypes and phenotypes to clinically relevant pharmacogenes based on the DPWG guidelines [15]. The selected genes were based on the U-PGx consortium's panel and consisted of 10 key pharmacogenes and 38 variants. The pipeline, which was originally designed for NGS data, did not include the *UGT1A1* and *HLA-B* genes which are present in the U-PGx consortium panel due to their complexities [15]. All genotypes are assessed based on their presence in the guidelines and on the number of drugs with an actionable advice, where actionable is defined as "a gene-drug interaction requiring a drug switch, dose adjustment or intensive monitoring". For all pharmacogenes mentioned in the CPIC and DPWG guidelines, the number of gene-drug interactions are calculated.

Recall and precision

To assess the accuracy of detecting different types of genetic variants in PGx genes, variant calling results were compared to the benchmark results from GIAB v.3.3.2 HG002 using the hap.py pipeline [47]. For SNVs and Indels, the benchmark v3.3. sequence is based on short-read sequencing [48]. Both the GATK variant caller (v.4.0.6.0) and DeepVariant (v.0.7.1) with the PacBio model were used to identify genetic variants. To assess recall and precision in complex regions, results were stratified using the stratifications from GIAB (https://github.com/genome-in-a-bottle/genome-stratifications). In addition, benchmarking results in GC-rich regions, homopolymers, tandem repeats, segmental duplications and UCSC repeat tracks were included in the analysis.

To assess the accuracy of SV calling in PGx genes, publicly available SV calls obtained with pbsv were downloaded from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/PacBio_pbsv_05212019/ and compared to the GIAB benchmark using truvari as previously described (https://github.com/PacificBiosciences/sv-benchmark) [19]. GIAB high confidence regions and SV callset were obtained from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/ NIST_SVs_Integration_v0.6/. Since GIAB curation for SV is based on hg19, the genes were converted to the hg19 genome using the liftOver tool from UCSC. Bedtools was used to overlap pharmacogenes to high-confident regions from GIAB. The SV benchmark set only includes SV with a size larger than 50 bp, therefore the SV analysis is limited to SVs >50 bp.

CODE AVAILABILITY

The code developed to generate the results in this study is available upon request.

REFERENCES

1. Pirmohamed M. Personalized pharmacogenomics: predicting efficacy and adverse drug reactions. Annu Rev Genomics Hum Genet. 2014;15:349–70.
2. Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, Peterson JF, et al. Pharmacogenomics. Lancet. 2019;394:521–32.
3. Matthaei J, Brockmoller J, Tzvetkov MV, Sehrt D, Sachse-Seeboth C, Hjelmborg JB, et al. Heritability of metoprolol and torsemide pharmacokinetics. Clin Pharmacol Ther. 2015;98:611–21.
4. Klein K, Zanger UM. Pharmacogenomics of cytochrome P450 3A4: recent progress toward the "Missing Heritability" problem. Front Genet. 2013;4:12.
5. Gaedigk A. Complexities of CYP2D6 gene analysis and interpretation. Int Rev Psychiatry. 2013;25:534–53.
6. Ingelman-Sundberg M, Sim SC. Pharmacogenetic biomarkers as tools for improved drug therapy; emphasis on the cytochrome P450 system. Biochem Biophys Res Commun. 2010;396:90–4.
7. Buermans HP, Vossen RH, Anvar SY, Allard WG, Guchelaar HJ, White SJ, et al. Flexible and scalable full-length CYP2D6 long amplicon PacBio sequencing. Hum Mutat. 2017;38:310–6.
8. Lauschke VM, Ingelman-Sundberg M. Prediction of drug response and adverse drug reactions: from twin studies to next generation sequencing. Eur J Pharm Sci. 2019;130:65–77.
9. Lauschke VM, Milani L, Ingelman-Sundberg M.Pharmacogenomic biomarkers for improved drug therapy-recent progress and future developments.AAPS J. 2017;20:4
10. van der Wouden CH, Cambon-Thomsen A, Cecchin E, Cheung KC, Davila-Fajardo CL, Deneer VH, et al. Implementing pharmacogenomics in Europe: design and implementation strategy of the ubiquitous pharmacogenomics consortium. Clin Pharmacol Ther. 2017;101:341–58.
11. Gordon AS, Fulton RS, Qin X, Mardis ER, Nickerson DA, Scherer S PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. Pharmacogenet Genomics. 2016.
12. Kimura S, Umeno M, Skoda RC, Meyer UA, Gonzalez FJ. The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic CYP2D6 gene, a related gene, and a pseudogene. Am J Human Genet. 1989;45:889–904.
13. Gellner K, Eiselt R, Hustert E, Arnold H, Koch I, Haberl M, et al. Genomic organization of the human CYP3A locus: identification of a new, inducible CYP3A gene. Pharmacogenetics. 2001;11:111–21.
14. Yang W, Wu G, Broeckel U, Smith CA, Turner V, Haidar CE, et al. Comparison of genome sequencing and clinical genotyping for pharmacogenes. Clin Pharmacol Ther. 2016;100:380–8.
15. van der Lee M, Allard WG, Bollen S, Santen GWE, Ruivenkamp CAL, Hoffer MJV, et al. Repurposing of diagnostic whole exome sequencing data of 1,583 individuals for clinical pharmacogenetics. Clin Pharmacol Ther. 2020;107:617–27.
16. Qiao W, Yang Y, Sebra R, Mendiratta G, Gaedigk A, Desnick RJ, et al. Long-read single molecule real-time full gene sequencing of cytochrome P450-2D6. Human Mutat. 2016;37:315–23.
17. Buermans HP, Vossen RH, Anvar SY, Allard WG, Guchelaar HJ, White SJ, et al. Flexible and scalable full-length CYP2D6 long amplicon PacBio sequencing. Human Mutat. 2017;38:310–6.
18. Ingelman-Sundberg M. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. Trends Pharmacol Sci. 2004;25:193–200.
19. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.
20. Midha MK, Wu M, Chiu K-P. Long-read sequencing in deciphering human genetics to a greater depth. Hum Genet. 2019;138:1201–15.
21. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, et al. Sequencing of human genomes with nanopore technology. Nat Commun. 2019;10:1869.
22. Ardui S, Race V, Zablotskaya A, Hestand MS, Van Esch H, Devriendt K, et al. Detecting AGG interruptions in male and female FMR1 premutation carriers by single-molecule sequencing. Hum Mutat. 2017;38:324–31.
23. Borràs DM, Vossen R, Liem M, Buermans HPJ, Dauwerse H, van Heusden D, et al. Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing. Human Mutat. 2017;38:870–9.
24. Schüle B, McFarland KN, Lee K, Tsai YC, Nguyen KD, Sun C, et al. Parkinson's disease associated with pure ATXN10 repeat expansion. NPJ Parkinson's Dis. 2017;3:27.
25. Ameur A, Kloosterman WP, Hestand MS Single-molecule sequencing: towards clinical applications. Trends Biotechnol. 2019;37:72–85.
26. van der Lee M, Allard WG, Vossen R, Baak-Pablo RF, Menafra R, Deiman B, et al. Toward predicting CYP2D6-mediated variable drug response from CYP2D6 gene sequencing data. Sci Transl Med. 2021;13:eabf3637.
27. Fukunaga K, Hishinuma E, Hiratsuka M, Kato K, Okusaka T, Saito T, et al. Determination of novel CYP2D6 haplotype using the targeted sequencing followed by the long-read sequencing and the functional characterization in the Japanese population. J Hum Genet. 2021;66:139–49.
28. Liau Y, Maggo S, Miller AL, Pearson JF, Kennedy MA, Cree SL. Nanopore sequencing of the pharmacogene CYP2D6 allows simultaneous haplotyping and detection of duplications. Pharmacogenomics. 2019;20:1033–47.
29. Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Res. 2015;4:17.
30. Lacaze P, Ronaldson KJ, Zhang EJ, Alfirevic A, Shah H, Newman L, et al. Genetic associations with clozapine-induced myocarditis in patients with schizophrenia. Transl Psychiatry. 2020;10:37.
31. Broad Institute. Genome Analysis Toolkot (GATK) 2019 [Available from: https://software.broadinstitute.org/gatk/.
32. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnol. 2018;36:983–7.
33. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. J Comput Biol. 2015;22:498–509.
34. Caspar SM, Schneider T, Meienberg J, Matyas G Added value of clinical sequencing: WGS-based profiling of pharmacogenes. Int J Mol Sci. 2020;21:2308.
35. Bush WS, Crosslin DR, Owusu-Obeng A, Wallace J, Almoguera B, Basford MA, et al. Genetic variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network. Clin Pharmacol Ther. 2016;100:160–9.
36. Jin Y, Wang J, Bachtiar M, Chong SS, Lee CGL. Architecture of polymorphisms in the human genome reveals functionally important and positively selected

variants in immune response and drug transporter genes. Hum Genomics. 2018;12:43.
37. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12:703–14.
38. Lee SB, Wheeler MM, Thummel KE, Nickerson DA Calling star alleles With Stargazer in 28 pharmacogenes with whole genome sequences. Clin Pharmacol Ther. 2019;106:1328–37.
39. Twist GP, Gaedigk A, Miller NA, Farrow EG, Willig LK, Dinwiddie DL, et al. Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. NPJ Genomic Med. 2016;1:15007.
40. Numanagić I, Malikić S, Ford M, Qin X, Toji L, Radovich M, et al. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. Nat Commun. 2018;9:828.
41. National Human Genome Research Institute. DNA sequencing costs: data [cited 2020. Available from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.
42. Pacific Biosciences PacBio https://www.pacb.com2019
43. Zhou Y, Ingelman-Sundberg M, Lauschke VM. Worldwide distribution of cytochrome P450 alleles: A meta-analysis of population-scale sequencing projects. Clin Pharmacol Ther. 2017;102:688–700.
44. National Institute of Standards and Technology. Genome in a bottle 2020 Available from: https://www.nist.gov/programs-projects/genome-bottle.
45. Martin M, Patterson M, Garg S, O Fischer S, Pisanti N, Klau GW, et al. WhatsHap: fast and accurate read-based phasing. bioRxiv 085050; https://doi.org/10.1101/085050
46. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–d73.
47. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37:555–60.
48. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol. 2019;37:561–6.

AUTHOR CONTRIBUTIONS

Designed the research: SYA. Performed the research: SYA, ML, WJR, RM. Wrote the manuscript: ML, SYA, HJG, JJS.

FUNDING INFORMATION

The research leading to these results has received funding from the European Community's Horizon 2020 Programme under grant agreement No. 668353 (U-PGx).

COMPETING INTERESTS

We would like to declare that William J. Rowell is a full-time employee of Pacific Biosciences.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41397-021-00259-z.

Correspondence and requests for materials should be addressed to Jesse J. Swen or Seyed Yahya Anvar.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.