Single molecule long-read real-time amplicon-based sequencing of *CYP2D6*: a proof-of-concept with hybrid haplotypes

Rachael Dong[1*], Megana Thamilselvan[2*], Jerome Clifford Foo[3,4,5], Jacqueline DeJarnette[1], Xiuying Hu[3], Beatriz Carvalho Henriques[3], Yabing Wang[3], Keanna Wallace[3], Sudhakar Sivapalan[3], Avery Buchner[1,3], Vasyl Yavorskyy[2,3], Kristina Martens[6], Wolfgang Maier[7], Neven Henigsberg[8], Joanna Hauser[9], Annamaria Cattaneo[10,11], Ole Mors[12], Marcella Rietschel[4], Gerald Pfeffer[6,13], Katherine J. Aitchison[1,3,14,15**]

[1]Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Canada
[2]College of Natural and Applied Sciences, Faculty of Science, Department of Biological Sciences, University of Alberta, Edmonton, Canada
[3]College of Health Sciences, Department of Psychiatry, University of Alberta, Edmonton, Canada
[4]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany
[5]Institute of Psychopharmacology, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany
[6]Cumming School of Medicine, Hotchkiss Brain Institute, Department of Clinical Neurosciences, University of Calgary, Calgary, Canada
[7]Department of Psychiatry and Psychotherapy, University of Bonn, Bonn, Germany
[8]Centre for Excellence for Basic, Clinical and Translational Research, Croatian Institute for Brain Research, University of Zagreb School of Medicine, Zagreb, Croatia
[9]Department of Psychiatry, Poznan University of Medical Sciences, Poznan, Poland
[10]Biological Psychiatry Unit, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy
[11]Department of Pharmacological and Biomolecular Sciences, University of Milan, Milan, Italy
[12]Psychosis Research Unit, Aarhus University Hospital, Risskov, Denmark
[13]Cumming School of Medicine, Alberta Child Health Research Institute & Department of Medical Genetics, University of Calgary, Calgary, Canada
[14]College of Health Sciences, Department of Medical Genetics, University of Alberta, Edmonton, Canada
[15]Women and Children's Research Institute, University of Alberta, Edmonton, Canada

*These authors contributed equally to this work

**Corresponding author:

Dr. Katherine J. Aitchison, Neuroscience and Mental Health Institute; Women and Children's Research Institute; College of Health Sciences, Faculty of Medicine and Dentistry, Departments of Psychiatry and Medical Genetics, University of Alberta, Edmonton, AB T6G 2E1, Canada

Telephone: 1-780-492-4018; Email: kaitchis@ualberta.ca

Running Title: ***SMRT multiplexed amplicon sequencing of CYP2D6***

Abstract

CYP2D6 is a widely expressed human xenobiotic metabolizing enzyme, best known for its role in the hepatic phase I cytochrome P450 enzyme system, where it metabolizes ~20% of medications. It is also expressed in other organs including the brain, where its potential role in physiology and mental health traits and disorders is under further investigation. Owing to the presence of homologous pseudogenes in the *CYP2D* locus and transposable repeat elements in the intergenic regions, the gene encoding the CYP2D6 enzyme, *CYP2D6*, is one of the most hypervariable known human genes - with more than 165 core haplotypes. Haplotypes include structural variants, with a subtype of these known as hybrid haplotypes or fusion genes comprising part of *CYP2D6* and part of its adjacent pseudogene, *CYP2D7*. The fusion genes are particularly challenging to identify. High fidelity (HiFi) single molecule real-time (SMRT) long-read sequencing can cover whole *CYP2D6* haplotypes in a single continuous sequence, and is therefore ideal for structural variant detection. In addition, it is highly accurate and suitable for novel haplotype identification, which is necessary as new *CYP2D6* haplotypes are continuously being discovered, and many more likely remain to be identified in relatively understudied populations such as Indigenous Peoples. The aim of the present work was to develop an efficient and accurate HiFi SMRT amplicon-based method capable of detecting the full range of *CYP2D6* haplotypes including fusion genes. We report proof-of-concept for 24 amplicons including three positive controls, aligned to fusion gene haplotypes, with prior cross-validation data. Amplicons with *CYP2D7-D6* fusion genes, including positive controls, aligned to the **13* subhaplotypes predicted (**13F*, **13A2*) with 100% accuracy, with the exception of one that aligned at 99.9%. Alignment of the **68* was 100% and above 99.9% to the *CYP2D6*68* partial sequences EU5300606 and JF307779, respectively. The best alignments for the remaining *CYP2D6-2D7* fusion genes were ≥99.7% (to 3 significant figures). Lower percentage alignment for *CYP2D6-2D7* fusion genes may reflect imperfect PCR optimization and/or the possibility that we may have haplotypes not yet in public databases. Further work on these is in progress. Moreover, we have adapted this method for non-hybrid haplotypes. This technique could therefore suffice for the characterization of the full range of *CYP2D6* haplotypes. The method that we have developed could be extended to other complex loci and to other species in a multiplexed high throughput assay.

Introduction

CYP2D6 is a xenobiotic metabolizing enzyme widely expressed in multiple organs including the liver, intestine, brain, gonads, and thyroid gland[1-3]. While it is best known for its role in the hepatic phase I drug cytochrome P450 enzyme system, where it metabolizes an estimated 20% of medications[4-6], it also has physiological roles. In the brain, it is found in cerebellar Purkinje cells and cortical pyramidal neurones[3,7]. It is colocalized with the dopamine transporter[8], and dopamine transporter inhibitors such as cocaine also inhibit CYP2D6[9,10]. Enzyme activity appears to modulate resting brain perfusion, with suggestive involvement in regions associated with alertness or serotonergic function[11], and a rodent model developed to further explore this[12]. It is involved in steroid biosynthesis (conducting the 21-hydroxylation of progesterone and allopregnanolone[13], as well as in the synthesis of dopamine from *m*- and *p*-tyramine[14,15], and of serotonin from 5-methoxyndolethylamines[16] including 5-methoxytryptamine (a metabolite and precursor of melatonin[17]). The enzyme is induced in alcoholism[18] and in pregnancy[19]. There has also been suggestive evidence of an association between enzyme status and personality traits[20-22].

The gene encoding the CYP2D6 enzyme, *CYP2D6*, lies at chromosome 22q13.2 adjacent to two pseudogenes, *CYP2D7* and *CYP2D8*. The three genes are highly homologous[23-25], and this, together with transposable repeat elements in the intergenic regions[24], predisposes the locus to the generation of structural variants and novel mutations. Indeed, with more than 140 core haplotypes (or strings of genetic variants) identified to date[26,27], *CYP2D6* is one of the most hypervariable known human genes. Variants include single nucleotide variants (SNVs), small insertions and/or deletions (indels), and structural variants[28]. Structural variants include gene duplications and multiplications, complete deletions of the entire gene, and recombination events involving *CYP2D7*[28-32]. The recombination events involving *CYP2D7* result in the formation of hybrid or fusion genes[28-31,33-37] (Figure 1). *CYP2D7-2D6* fusions have a 5'-portion derived from *CYP2D7* and a 3'-portion derived from *CYP2D6*; these hybrids are non-functional[28]. *CYP2D6-2D7* fusions have a 5'-portion derived from *CYP2D6* and a 3'-portion derived from *CYP2D7*.

The Pharmacogene Variation Consortium assigns levels of function (no function, decreased, normal, or increased) to *CYP2D6* haplotypes that correspond to enzyme or phenotype predictions (poor metabolizers, intermediate, normal, or ultrarapid, respectively)[26,38-40]. As an example of haplotype to phenotype prediction, an individual with two no function (or null) haplotypes is a poor metabolizer, with no active enzyme. This has implications for the metabolism

of relevant medications. In this example, such individuals are unable to metabolize codeine from its inactive prodrug status to the metabolite with analgesic effect and hence do not experience any analgesic effect with this medication[41]. The *CYP2D6-2D7* fusion or hybrid haplotypes have zero function (e.g., **36*, **68*, and the ones with the 1847G>A SNV) or uncertain/unknown function(e.g., **61* and **63*)[26,28]. Hybrid haplotypes are found either as a single haplotype or in tandem with another *CYP2D6* haplotype[28-31,36,37].

CYP2D6 hybrid haplotypes are common in the general population, with a frequency estimated as at least 6.7%[42]. In our sample, the Genome-based therapeutic drugs for depression (GENDEP) study[43], out of 853 patients with depression, the frequency is 2.6% (22/853). Owing to the range of enzyme phenotype corresponding to *CYP2D6* hybrids, their accurate detection is important for predicting prescribing implications[42], as well as potentially for neuroscience and physiology more generally.

The detection of *CYP2D6* hybrid haplotypes is however, challenging for many genomic technologies, with incorrect and incomplete characterization being commonplace (Carvalho Henriques et al., 2021b). For example, the AmpliChip CYP450 Test did not cover hybrid haplotypes, and hence none of the 19 patients subsequently identified as having hybrid haplotypes had previously had these found, with 2 having been genotyped as *CYP2D6*1/*1* (wild-type) and 4 as 'no call'[43]. We have previously reported the use of methods including Sanger sequencing to characterize hybrid haplotypes[43-45]. However, Sanger sequencing poses limitations for haplotype phasing (determining which combination of variants lies on which allele) and discriminating whether sequence is derived from *CYP2D6* or *CYP2D7*[46]. As a relatively labour-intensive and time-consuming technique, Sanger sequencing is best suited to low throughput.

Short-read next generation sequencing (NGS) is useful for the detection of SNVs and small indels, but less useful for structural variant detection with haplotype phasing, which requires information across longer sequence or read lengths[47]. Single molecule real-time (SMRT) long-read sequencing provided by Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) can achieve structural variant detection with haplotype phasing. However, until relatively recently, these SMRT technologies had a lower accuracy than short-read NGS. In 2019, circular consensus sequencing (CCS) was optimized to generate highly accurate (99.8%) long high fidelity (HiFi) reads[47], with a median length of 13.4 kb[48]. At this accuracy level, SNVs and short indels may be identified as well as structural variants. With *CYP2D6* haplotypes being under 10 kb and including

all types of variation, this technology is eminently suited for the identification of the full range of haplotypes including novel, unidentified haplotypes. The latter is important as to date, there are populations (e.g., Indigenous peoples) in which this gene is relatively less studied than in others.

Some *CYP2D6* sequencing using HiFi SMRT has already been conducted in previous research. For 25 individuals with prior AmpliChip CYP450 genotype, including four with "*XN*" representing more than one copy of specific haplotypes (e.g., *CYP2D6*1/*2XN*), all genotypes were concordant with the genotype resulting from the SMRT data other than one[49]. In this case, the prior genotype was *CYP2D6*4/*4* and the new genotype was *CYP2D6*4/*5,* with the *CYP2D6*5* representing a complete deletion of the *CYP2D6* gene that had been missed by the AmpliChip (the design of which is now recognized as being able to detect only a subset of *CYP2D6*5* haplotypes[43]). In addition, one novel trinucleotide deletion and one novel SNV were detected in this group of samples and confirmed by Sanger sequencing. SMRT data for *CYP2D6* has also been compared to data from targeted Illumina NGS in 17 individuals[50]. These 17 included one hybrid haplotype (*CYP2D6*36*), including a duplication thereof and its occurrence together with *CYP2D6*10* in a hybrid tandem (*CYP2D6*36+*10*). A recent study has applied the HiFi SMRT technology to 561 patients treated with tamoxifen, and to separate cohorts treated with tamoxifen and venlafaxine[40]. In the tamoxifen-treated dataset, only four individuals with hybrids were identified by the SMRT, and the hybrid haplotype was not specified (Supplementary Table S2[40]). The latest relevant SMRT paper used a long-range polymerase chain reaction (L-PCR) technique that amplified a 6.1 kb region spanning from 712 bp upstream to 1176 bp downstream of the NB_008376.4 *CYP2D6* RefSeq coding sequence, and, in addition (by design), amplified the corresponding region from *CYP2D7*, generating a 7.6 kb amplicon in an analysis of 377 Solomon Islanders[51]. From the SMRT data, 27/365 (7.6%) samples appeared to have a *CYP2D6-2D7* fusion haplotype with breakpoints in exon 8 (consistent with a *CYP2D6*63*), and 7/365 (2%) samples appeared to have *CYP2D7-2D6* fusions (*CYP2D6*13*). However, there was a degree of discrepancy between the above and TaqMan CNV intron 9 and exon 9 data, with not all of the samples with a *CYP2D6-2D7* predicted fusion haplotype having a higher intron 2 than exon 9 CNV count, and 1/7 of the predicted *CYP2D7-2D6* fusions not having a high exon 9 count. In addition, the upstream region covered was insufficient for submission of novel haplotypes to PharmVar, which requires at least 1600 bp upstream of the ATG start sequence (to cover the -1584C/G SNP).

The research gap that we address herein is therefore: creating an efficient and accurate HiFi SMRT amplicon-based method capable of detecting the full range of *CYP2D6-2D7* and *CYP2D7-2D6* fusion genes, which could be extended to other *CYP2D6* haplotypes and adapted for multigene panels.

Methods

Samples

Used herein was the subset of 95 DNA samples from patients with depression in the Genome-based therapeutic drugs for depression (GENDEP) pharmacogenomics clinical trial, specifically, the 19 samples with *CYP2D6* hybrid haplotypes previously described by Carvalho Henriques et al.[43], plus one additional putative hybrid identified by TaqMan copy number variant (CNV) screening using the methodology described[43]. GENDEP was designed to identify pharmacogenomic predictors of response to two antidepressants, nortriptyline and escitalopram, the metabolism of which both involve CYP2D6[52]. DNA was extracted from venous blood. The 19 samples had prior data from multiple technologies (the AmpliChip CYP450 test, TaqMan SNV and CNV data, the Ion AmpliSeq Pharmacogenomics Panel, PharmacoScan, and Sanger sequencing) resulting in consensus genotype calls, while the one additional sample had only prior AmpliChip CYP450 (*CYP2D6*1/*2*) and TaqMan CNV data that were unequal across different regions of *CYP2D6* (intron 2, intron 6, and exon 9 calls of 3, 3, and 2, respectively), indicating the presence of a *CYP2D6-2D7* fusion gene. In total, there were 20 samples, two of which had consensus genotypes consistent with two hybrid haplotypes. In addition, we used two samples from the Genetic Testing Reference Material Program (GeT-RM) collection as positive controls for *CYP2D6-2D7* (NA18545, *CYP2D6*5/*36x2+*10x2*) and *CYP2D7-2D6 (*NA19785, *CYP2D6*1/*13+*2)* hybrid haplotypes, respectively[53].

L-PCRs using barcoded universal primers for multiplexing amplicons

We adapted protocols for amplifying hybrid-specific amplicons E, G, and H as described[29-31,43] for the PacBio barcoded universal primers for multiplexing amplicons method[54]. The E amplicon is for *CYP2D6-2D7* fusion genes, while the G and H cover the *CYP2D7-2D6* fusion genes. A 5' universal sequence and a 5' amino modifier C6 (5AmMC6, to prevent unbarcoded amplicons from

being sequenced) were added to each primer (Supplementary Table S1, Thermo Fisher Scientific) for the first-round L-PCR.

First-Round L-PCRs were performed using KAPA HiFi HotStart ReadyMix (Roche Molecular Systems). Reactions (50 µl, in duplicate) contained primers at 0.3 µM each, dNTPs at 0.6 µM, $MgCl_2$ at 2.75 µM, 2% DMSO, and 3 µl of template at 25-202 ng/ µl (more template was used for samples previously showing a relatively low amplicon generating efficiency). For the E amplicon, DNA was amplified for 30 cycles with denaturation at 98°C for 30 s, cycling at 98°C for 10 s, annealing at 85°C for 30 s, and extension at 73°C for 7.5 min (latterly adjusted to 8 min), followed by a terminal extension step of 73°C for 10 min. For the G amplicon, conditions were: denaturation at 98°C for 30 s; 30 cycles of 98°C for 10 s, and annealing and extension at 72°C for 6 min; followed by a terminal extension step of 72°C for 7 min. For the H amplicon, conditions were: denaturation at 95°C for 30 s; 35 cycles of 98°C for 20 s, annealing at 74°C for 30 s and extension at 73°C for 5 min; followed by a terminal extension step of 72°C for 10 min.

PCR products were visualized on a 1% agarose gel, quantified using the Qubit dsDNA HS assay using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific), and visualized using an Agilent DNA 12000 kit on an Agilent 2100 Bioanalyzer (Agilent Technologies) (Figure 2). As some nonspecific products and primer oligomers were seen on the Agilent output for amplicons E and H, respectively, size selection using AMPure PB Beads (PacBio) was used. Purified sample and peak concentrations were measured with the Qubit dsDNA HS and Agilent DNA 12000 assays, respectively.

In the second-round L-PCR, barcoded universal primers (barcoded universal F/R primers plate 96 v2, PacBio) were attached to the forward and reverse universal sequences in the amplicons from the first-round L-PCR. Reactions (25 µl, in duplicate or more, using KAPA HiFi HotStart ReadyMix) contained 2.5 µl barcoded primers, dNTPs at 0.6 µM, $MgCl_2$ at 2.75 µM, 2% DMSO, and 1.5-3 ng template. For the E amplicons, cycling conditions were the same as those used in the first-round L-PCR, with reduction in the number of cycles to 20. For the G amplicons, conditions were: denaturation at 98°C for 30 s; followed by 30 cycles of 98°C for 10 s, annealing at 70°C for 15 s, and extension 72°C for 6 min; then a terminal extension step at 72°C for 7 min. Conditions for the H amplicon were: denaturation at 95°C for 30 s; 35 cycles of 98°C for 20s, annealing at 74°C for 30 s and extension at 73°C for 5 min; followed by a terminal extension step at 72°C for

10 min. After quantification of amplicons using the same methods as for the first-round, size selection and removal of excess primers was conducted using AMPure PB Beads (PacBio).

Samples were pooled in equimolar amounts (~23 fmol, in duplicate), calculated using approximations of estimated amplicon lengths and the NEBioCalculator (New England Bio Labs), to generate a pool mass of 0.3-3 μg, with size selection using AMPure PB Beads of the pool, and visualization using the Agilent DNA 12000 kit (Figure 3).

Sequencing

SMRTbell library construction using appropriate pool input mass following visualization using an Agilent TapeStation, size selection using AMPure PB Beads, and sequencing with a 30 hour movie time on an 8M cell using the Sequel IIe platform were conducted at The Centre for Applied Genomics (TCAG), the Hospital for Sick Children, Toronto, Canada. The above procedures were also conducted at the Omics and Precision Agriculture Laboratory, Global Institute for Food Security, University of Saskatchewan, where size selection using a Blue Pippin was also conducted. Data demultiplexing was conducted by the bioinformatics teams at these facilities.

Data Analysis and Sequence Alignment

After filtering CCS reads based on amplicon length, alignment versus reference sequences was conducted using SMRT Link Software version 10.2. Reference sequences (from the National Center for Biotechnology Information (NCBI) or PharmVar) for alignment were deduced from the consensus genotypes previously generated (Supplementary Table S1).

Results

Amplicons from twenty-one hybrid haplotypes were submitted (for one sample with two types of hybrids, only one was submitted) for sequencing, plus three positive controls. Data from twenty-four hybrid haplotypes were analyzed including positive controls (13 E, 9 G, and two H) (Tables 1-3). For the G amplicons, the alignment versus reference *CYP2D6*13* sequences was 100% (Table 1). The H amplicons are presented in Table 2. The alignment versus reference *CYP2D6*13* sequence for the positive control (NA19785) H amplicon was also 100%. The sample amplified using the H primers that previously had a consensus genotype with a degree of uncertainty aligned best to *CYP2D6*13F* at 99.94%. Technical replicates of this sample all aligned to *CYP2D6*13F* with a percentage of at least 99.5%. Further examination of the data

pertaining to this sample enabled resolution of genotype to *CYP2D6*13/*4.013*. The E amplicons aligned to **36* and **68* hybrid haplotypes as well as to sequences of the various hybrid haplotypes that have the 1847G>A (splice defect) SNV that is the defining SNV for the **4* haplotypes (Table 3. Alignment of the **68* was 100% and above 99.9% to the *CYP2D6*68* partial sequences EU5300606 and JF307779, respectively. The best alignments for the remaining *CYP2D6-2D7* fusion genes were ≥99.7% (to 3 significant figures). The available reference sequences for the **4* hybrid haplotypes include EU530605 (**4-like*[29]), EU530604 (**4N*[29]), and PV00250 (PharmVar **4.013*). Alignments against all three of these are presented (Table 3) for the relevant amplicons apart from one, for which alignment against PV00250 was not possible (sample 5).

Discussion

In summary, we were able to develop and optimize an amplicon-based method of detecting a range of *CYP2D6-2D7* and *CYP2D7-2D6* fusion genes using PacBio barcoded universal primers (BUP) for multiplexing amplicons. This is the most challenging type of *CYP2D6* variant to detect and characterize. Data analysis was highly efficient. Data were cross-validated versus previous data from multiple technologies[43], and all of the resulting best percentage alignments were ≥99.7%. This method would therefore appear to be more accurate and efficient than any of the other SMRT HiFi methodologies for *CYP2D6* haplotype detection reported to date[40,50,51] for the detection and characterization of *CYP2D6* fusion genes. In addition, we have since used the forward primer for the E amplicon and the reverse primer for the G amplicon to generate an L-PCR product for non-hybrid *CYP2D6* haplotypes (data not shown). Therefore, the combination of four primer pairs (E forward, E reverse, G forward, G reverse) is sufficient for an amplicon-based method of *CYP2D6* characterization apart from a minority of *CYP2D7-2D6* hybrids for which the H amplicon appears to be required. Moreover, the E forward primer is 1909 base pairs (bp) upstream from the ATG start site, and the G reverse primer 619 bp downstream from the TAG stop site. The amplicon generated therefore covers the upstream (to -1600 bp) and downstream (to 265 bp) regions required for novel haplotype submission to PharmVar.

For G amplicons, the alignment was excellent at 100%. There may be at least two potential reasons for this. Firstly, in the first-round L-PCRs for the Gs, we were able to generate amplicons with minimal non-specificity (Figure 2). Secondly, a **13* haplotype was the first type of *CYP2D6*

fusion gene to be identified[34], and hence has been relatively well studied since (with there now being 10 publicly available sequences) in comparison to the other hybrid haplotypes. The percentage alignment for the H positive control was 100%, and the best alignment for the H sample was above 99.9%.

Within the E amplicon group, best alignments were at 100% versus partial reference sequences, with the remaining being at ≥99.7%. The greater degree of variability in the alignment statistics for hybrid haplotypes in the **4* hybrid haplotypes as compared to other *CYP2D6-2D7* hybrid haplotypes may reflect the fact that (like the various **13* hybrid haplotypes[28]) this is a family of hybrid haplotypes containing the 1847G>A SNV. For the samples with a **36* haplotype, the alignment statistics for the **36* core haplotype and various subhaplotypes were comparable. Sample NA18545 has not previously been sequenced, with the genotype being deduced from CNV testing using L-PCR and quantitative CNV analysis (Gaedigk, personal communication). The alignment to the **36* partial sequence was 100% for this sample, and to the PharmVar core **36* haplotype 99.84%. As this sample has a **36* duplication, it may have more than one **36* subhaplotype. Further optimization of the E amplicon procedure as below may contribute to resolving this. Of note, our E amplicons were longer (by ~1.7 kb) than previously described[29,31]. Initial setting of the extension time in the cycling parameters reflected the shorter predicted length, and while we subsequently set extension to 8 min, this could likely be further optimized to 8.5 min. Accurate amplicon sizing is important not only for L-PCR optimization but also for molar calculations in amplicon pooling.

Owing to the presence of some non-specific products at less than the correct length resulting from the E L-PCRs, size selection to remove products less than ~3 kb was conducted using AMPure PB beads. Whilst the final pool profile indicates the persistence of such products, HiFi sequencing was nonetheless able to produce alignments. This may be at least partly attributable to the relatively small number of multiplexed amplicons, resulting in a high degree of redundancy and read depth.

The main limitation of the work reported herein is the variable percentage alignments, some of which may reflect factors such as PCR optimization achieved to date. Additionally, we recognize that the sample size used here is limited and further work in larger samples is needed for verification of the present methods and results. The alignment previously achieved by Sanger sequencing (after manual curation of the data) was comparable to that achieved by SMRT (e.g.,

sample 33, alignment 100%; sample 5, alignment 99.75% to **4-like*). The present SMRT alignment process was largely automated, and the entire SMRT process reported herein was much more efficient than Sanger sequencing. Further work is in progress, including using more of our own data as reference sequences.

This is the first report of an amplicon-based method of *CYP2D6* SMRT sequencing on a range of fusion genes with haplotypes previously characterized by multiple technologies. Although some versions of *CYP2D6-2D7* fusion genes (e.g., *CYP2D6*61*) were not in our sample set, we are not aware of any reason for our method not working for any *CYP2D6-2D7* fusion gene.

Although CYP2D6 plays a key role in the metabolism of ~20-25% of clinically used drugs[4,5], and is included in drug labelling by regulatory bodies (the U.S. Food and Drug Administration, the European Medicines Agency, the Pharmaceuticals and Medical Devices Agency) and prescribing recommendations by consortia (Clinical Pharmacogenetics Implementation Consortium, the Dutch Pharmacogenetics Working Group), identification of *CYP2D6* variants is not yet routine in clinical practice. This is despite the fact that dispensing data indicate that many patients are being prescribed medications for which the identification of *CYP2D6* variants prior to these medications being dispensed could be helpful[55]. One of the reasons for this is the complexity of the locus, and in particular the fact that the fusion genes are challenging to identify and accurately characterize. Another reason is that clinical implementation requires an efficient, high throughput method that requires relatively little personnel time. The HiFi sequencing method reported herein is suitable for high throughput, efficient, with accurate characterization of the full range of *CYP2D6* haplotypes and can be completed in a week or less. Moreover, the method that we have developed could be extended to other loci, and might be particularly useful for other complex loci with structural variants, such as *CYP2A6*. In this manner a group of genes may be efficiently characterized in a multiplexed high throughput assay.

Acknowledgements

We thank the University of Alberta Faculty of Medicine and Dentistry Department of Experimental Oncology and High Content Analysis Core for access to instrumentation and/or services provided, including Michael Wong for his contribution to our Agilent data, and the University of Alberta Department of Biological Sciences Molecular Biology Service Unit (MBSU) for Sanger sequencing (contribution to data in Supplementary Table S1). We thank Andrea

Gaedigk for personal communications regarding the details of GeT-RM samples NA19785 and NA18545 used in this report. We thank Matthew Seetin of Pacific Biosciences for bioinformatics training and support. The work described in this paper was funded by: a Canada Foundation for Innovation (CFI), John R. Evans Leaders Fund (JELF) grant (32147—Pharmacogenetic translational biomarker discovery, to KJA), an Alberta Innovates Strategic Research Project (SRP51_PRIME - Pharmacogenomics for the Prevention of Adverse Drug Reactions in mental health; G2018000868 to KJA and Chad Bousman), an Alberta Centennial Addiction and Mental Health Research Chair (to KJA), an Alberta Innovation and Advanced Education Small Equipment Grant (to KJA), the Department of Psychiatry, and the Faculty of Medicine and Dentistry at the University of Alberta. The infrastructure from GP's lab for the running of the Ion AmpliSeq Pharmacogenomics Panel was supported by a Hotchkiss Brain Institute Dementia Equipment Fund grant (to CB and GP for the Ion) and a Canada Foundation for Innovation John R. Evans Leaders Fund Grant (CFI-JELF) (36624 - Neuromuscular genetics program, to GP). GENDEP was funded by a European Commission Framework 6 grant, LSHB-CT-2003-503428. Roche Molecular Systems supplied the AmpliChip CYP450 Test arrays and some associated support. GlaxoSmithKline and the Medical Research Council (UK) contributed by funding add-on projects in the London centre. This paper is an original work and the views expressed here can only be attributed to the authors, not necessarily reflecting those of the NHS, the NIHR or the UK Department of Health and Social Care.

References

1. Miksys S.L., Cheung C., Gonzalez F.J., Tyndale R.F. Human CYP2D6 and mouse CYP2Ds: organ distribution in a humanized mouse model. *Drug Metab Dispos*, **33**, 1495-1502 (2005), 10.1124/dmd.105.005488
2. Aitchison K., Datla K., Rooprai H., Fernando J., Dexter D. Regional distribution of clomipramine and desmethylclomipramine in rat brain and peripheral organs on chronic clomipramine administration. *J Psychopharmacol*, **24**, 1261-1268 (2010), 10.1177/0269881109105789

3. GTEx Portal. CYP2D6 Gene Page. Accessed 2022 Sep 7 https://www.gtexportal.org/home/gene/ENSG00000100197
4. Ingelman-Sundberg M. Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J*, **5**, 6-13 (2005), 10.1038/sj.tpj.6500285
5. Zanger U.M., Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther*, **138**, 103-141 (2013), 10.1016/j.pharmthera.2012.12.007
6. Saravanakumar A., Sadighi A., Ryu R., Akhlaghi F. Physicochemical Properties, Biotransformation, and Transport Pathways of Established and Newly Approved Medications: A Systematic Review of the Top 200 Most Prescribed Drugs vs. the FDA-Approved Drugs Between 2005 and 2016. *Clin Pharmacokinet*, **58**, 1281-1294 (2019), 10.1007/s40262-019-00750-8
7. Siegle I., Fritz P., Eckhardt K., Zanger U.M., Eichelbaum M. Cellular localization and regional distribution of CYP2D6 mRNA and protein expression in human brain. *Pharmacogenetics*, **11**, 237-245 (2001), 10.1097/00008571-200104000-00007
8. Niznik H.B., Tyndale R.F., Sallee F.R., Gonzalez F.J., Hardwick J.P., Inaba T., et al. The dopamine transporter and cytochrome P45OIID1 (debrisoquine 4-hydroxylase) in brain: resolution and identification of two distinct [3H]GBR-12935 binding proteins. *Arch Biochem Biophys*, **276**, 424-432 (1990), 10.1016/0003-9861(90)90741-g
9. Shen H., He M.M., Liu H., Wrighton S.A., Wang L., Guo B., et al. Comparative metabolic capabilities and inhibitory profiles of CYP2D6.1, CYP2D6.10, and CYP2D6.17. *Drug Metab Dispos*, **35**, 1292-1300 (2007), 10.1124/dmd.107.015354
10. Han D.D., Gu H.H. Comparison of the monoamine transporters from human and mouse in their sensitivities to psychostimulant drugs. *BMC Pharmacol*, **6**, 6 (2006), 10.1186/1471-2210-6-6
11. Kirchheiner J., Seeringer A., Godoy A.L., Ohmle B., Maier C., Beschoner P., et al. CYP2D6 in the brain: genotype effects on resting brain perfusion. *Mol Psychiatry*, **16**, 237, 333-241 (2011), 10.1038/mp.2010.42

12. Cheng J., Zhen Y., Miksys S., Beyoglu D., Krausz K.W., Tyndale R.F., et al. Potential role of CYP2D6 in the central nervous system. *Xenobiotica*, **43**, 973-984 (2013), 10.3109/00498254.2013.791410
13. Niwa T., Okada K., Hiroi T., Imaoka S., Narimatsu S., Funae Y. Effect of psychotropic drugs on the 21-hydroxylation of neurosteroids, progesterone and allopregnanolone, catalyzed by rat CYP2D4 and human CYP2D6 in the brain. *Biol Pharm Bull*, **31**, 348-351 (2008), 10.1248/bpb.31.348
14. Hiroi T., Imaoka S., Funae Y. Dopamine Formation from Tyramine by CYP2D6. *Biochemical and Biophysical Research Communications*, **249**, 838-843 (1998). http://www.sciencedirect.com/science/article/pii/S0006291X98992324.
15. Funae Y., Kishimoto W., Cho T., Niwa T., Hiroi T. CYP2D in the brain. *Drug Metab Pharmacokinet*, **18**, 337-349 (2003), 10.2133/dmpk.18.337
16. Yu A.M., Idle J.R., Herraiz T., Küpfer A., Gonzalez F.J. Screening for endogenous substrates reveals that CYP2D6 is a 5-methoxyindolethylamine O-demethylase. *Pharmacogenetics*, **13**, 307-319 (2003), 10.1097/01.fpc.0000054094.48725.b7
17. Yu A.M., Idle J.R., Byrd L.G., Krausz K.W., Kupfer A., Gonzalez F.J. Regeneration of serotonin from 5-methoxytryptamine by polymorphic human CYP2D6. *Pharmacogenetics*, **13**, 173-181 (2003), 10.1097/01.fpc.0000054066.98065.7b
18. Miksys S., Rao Y., Hoffmann E., Mash D.C., Tyndale R.F. Regional and cellular expression of CYP2D6 in human brain: higher levels in alcoholics. *J Neurochem*, **82**, 1376-1387 (2002), 10.1046/j.1471-4159.2002.01069.x
19. Pan X., Ning M., Jeong H. Transcriptional Regulation of CYP2D6 Expression. *Drug Metab Dispos*, **45**, 42-48 (2017), 10.1124/dmd.116.072249
20. Bertilsson L., Alm C., De Las Carreras C., Widen J., Edman G., Schalling D. Debrisoquine hydroxylation polymorphism and personality. *Lancet*, **1**, 555 (1989), 10.1016/s0140-6736(89)90094-9
21. Gan S.H., Ismail R., Wan Adnan W.A., Zulmi W., Kumaraswamy N., Larmie E.T. Relationship between Type A and B personality and debrisoquine hydroxylation capacity. *Br J Clin Pharmacol*, **57**, 785-789 (2004), 10.1111/j.1365-2125.2004.02076.x

22. Gonzalez F.J. Regulation of hepatocyte nuclear factor 4 alpha-mediated transcription. *Drug Metab Pharmacokinet,* **23**, 2-7 (2008), 10.2133/dmpk.23.2

23. Kimura S., Umeno M., Skoda R.C., Meyer U.A., Gonzalez F.J. The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic *CYP2D6* gene, a related gene, and a pseudogene. *American Journal of Human Genetics,* **45**, 889-904 (1989). http://www.ncbi.nlm.nih.gov/pubmed/2574001.

24. Yasukochi Y., Satta Y. Evolution of the *CYP2D* gene cluster in humans and four non-human primates. *Genes Genet Syst,* **86**, 109-116 (2011), 10.1266/ggs.86.109

25. Yang Y., Botton M.R., Scott E.R., Scott S.A. Sequencing the *CYP2D6* gene: from variant allele discovery to clinical pharmacogenetic testing. *Pharmacogenomics,* **18**, 673-685 (2017), 10.2217/pgs-2017-0033

26. PharmVar. CYP2D6. Accessed 2023 Aug 11 https://www.pharmvar.org/gene/CYP2D6

27. Gaedigk A., Ingelman-Sundberg M., Miller N.A., Leeder J.S., Whirl-Carrillo M., Klein T.E., et al. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clinical Pharmacology and Therapeutics,* **103**, 399-401 (2018), 10.1002/cpt.910

28. PharmVar. *CYP2D6* Structural Variation Document. Accessed 2023 May 24 https://a.storyblok.com/f/70677/x/723d6993d7/cyp2d6_structural-variation_v2-5.pdf

29. Kramer W.E., Walker D.L., O'Kane D.J., Mrazek D.A., Fisher P.K., Dukek B.A., et al. *CYP2D6*: novel genomic structures and alleles. *Pharmacogenetics and Genomics,* **19**, 813-822 (2009), 10.1097/FPC.0b013e3283317b95

30. Gaedigk A., Jaime L.K., Bertino J.S., Jr., Berard A., Pratt V.M., Bradfordand L.D., et al. Identification of Novel *CYP2D7-2D6* Hybrids: Non-Functional and Functional Variants. *Front Pharmacol,* **1**, 121 (2010), 10.3389/fphar.2010.00121

31. Black J.L., 3rd, Walker D.L., O'Kane D.J., Harmandayan M. Frequency of undetected *CYP2D6* hybrid genes in clinical samples: impact on phenotype prediction. *Drug Metabolism and Disposition,* **40**, 111-119 (2012), 10.1124/dmd.111.040832

32. Gaedigk A. Complexities of CYP2D6 gene analysis and interpretation. *Int Rev Psychiatry,* **25**, 534-553 (2013), 10.3109/09540261.2013.825581

33. Panserat S., Mura C., Gérard N., Vincent-Viry M., Galteau M.M., Jacoz-Aigrain E., et al. An unequal cross-over event within the CYP2D gene cluster generates a chimeric CYP2D7/CYP2D6 gene which is associated with the poor metabolizer phenotype. *Br J Clin Pharmacol,* **40**, 361-367 (1995), 10.1111/j.1365-2125.1995.tb04558.x

34. Daly A.K., Fairbrother K.S., Andreassen O.A., London S.J., Idle J.R., Steen V.M. Characterization and PCR-based detection of two different hybrid *CYP2D7P*/*CYP2D6* alleles associated with the poor metabolizer phenotype. *Pharmacogenetics,* **6**, 319-328 (1996), 10.1097/00008571-199608000-00005

35. Gaedigk A., Coetsee C. The *CYP2D6* gene locus in South African Coloureds: unique allele distributions, novel alleles and gene arrangements. *European Journal of Clinical Pharmacology,* **64**, 465-475 (2008), 10.1007/s00228-007-0445-7

36. Gaedigk A., Fuhr U., Johnson C., Bérard L.A., Bradford D., Leeder J.S. *CYP2D7-2D6* hybrid tandems: identification of novel CYP2D6 duplication arrangements and implications for phenotype prediction. *Pharmacogenomics,* **11**, 43-53 (2010), 10.2217/pgs.09.133

37. Gaedigk A., Riffel A.K., Berrocal B.G., Solaesa V.G., Dávila I., Isidoro-García M. Characterization of a complex CYP2D6 genotype that caused an AmpliChip CYP450 Test no-call in the clinical setting. *Clin Chem Lab Med,* **52**, 799-807 (2014), 10.1515/cclm-2013-0943

38. Caudle K.E., Sangkuhl K., Whirl-Carrillo M., Swen J.J., Haidar C.E., Klein T.E., et al. Standardizing *CYP2D6* Genotype to Phenotype Translation: Consensus Recommendations from the Clinical Pharmacogenetics Implementation Consortium and Dutch Pharmacogenetics Working Group. *Clin Transl Sci,* **13**, 116-124 (2020), 10.1111/cts.12692

39. Jukić M.M., Smith R.L., Molden E., Ingelman-Sundberg M. Evaluation of the CYP2D6 Haplotype Activity Scores Based on Metabolic Ratios of 4,700 Patients Treated With Three Different CYP2D6 Substrates. *Clin Pharmacol Ther,* **110**, 750-758 (2021), 10.1002/cpt.2246

40. van der Lee M., Allard W.G., Vossen R., Baak-Pablo R.F., Menafra R., Deiman B., et al. Toward predicting CYP2D6-mediated variable drug response from CYP2D6 gene sequencing data. *Sci Transl Med*, **13**, (2021), 10.1126/scitranslmed.abf3637
41. Crews K.R., Monte A.A., Huddart R., Caudle K.E., Kharasch E.D., Gaedigk A., et al. Clinical Pharmacogenetics Implementation Consortium Guideline for *CYP2D6*, *OPRM1*, and *COMT* Genotypes and Select Opioid Therapy. *Clin Pharmacol Ther*, **110**, 888-896 (2021), 10.1002/cpt.2149
42. Dalton R., Lee S.B., Claw K.G., Prasad B., Phillips B.R., Shen D.D., et al. Interrogation of CYP2D6 Structural Variant Alleles Improves the Correlation Between CYP2D6 Genotype and CYP2D6-Mediated Metabolic Activity. *Clin Transl Sci*, **13**, 147-156 (2020), 10.1111/cts.12695
43. Carvalho Henriques B., Buchner A., Hu X., Wang Y., Yavorskyy V., Wallace K., et al. Methodology for clinical genotyping of *CYP2D6* and *CYP2C19*. *Translational Psychiatry*, **11**, 596 (2021), 10.1038/s41398-021-01717-9
44. Carvalho Henriques B., Buchner A., Hu X., Wang Y., Yavorskyy V., Rachael D., et al. Validation of methodology for efficient genotyping of *CYP2D6* and *CYP2C19* *Research Square*, (2021). https://doi.org/10.21203/rs.3.rs-123677/v1.
45. Carvalho Henriques B., Buchner A., Hu X., Yavorskyy V., Wang Y., Martens K., et al. Cross-validation of technologies for genotyping CYP2D6 and CYP2C19. *bioRxiv*, 2019.2012.2024.870295 (2020), 10.1101/2019.12.24.870295
46. Ardui S., Ameur A., Vermeesch J.R., Hestand M.S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*, **46**, 2159-2168 (2018), 10.1093/nar/gky066
47. Wenger A.M., Peluso P., Rowell W.J., Chang P.C., Hall R.J., Concepcion G.T., et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, **37**, 1155-1162 (2019), 10.1038/s41587-019-0217-9
48. van der Lee M., Rowell W.J., Menafra R., Guchelaar H.-J., Swen J.J., Anvar S.Y. Application of long-read sequencing to elucidate complex pharmacogenomic regions: a proof of principle. *Pharmacogenomics Journal*, **22**, 75-81 (2022), 10.1038/s41397-021-00259-z

49. Buermans H.P.J., Vossen R.H.A.M., Anvar S.Y., Allard W.G., Guchelaar H.-J., White S.J., et al. Flexible and Scalable Full-Length *CYP2D6* Long Amplicon PacBio Sequencing. *Human Mutation*, **38**, 310-316 (2017), 10.1002/humu.23166

50. 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*, **66**, 139-149 (2021), 10.1038/s10038-020-0815-x

51. Charnaud S., Munro J.E., Semenec L., Mazhari R., Brewster J., Bourke C., et al. PacBio long-read amplicon sequencing enables scalable high-resolution population allele typing of the complex CYP2D6 locus. *Communications Biology*, **5**, 168 (2022), 10.1038/s42003-022-03102-8

52. Carvalho Henriques B., Yang E.H., Lapetina D., Carr M.S., Yavorskyy V., Hague J., et al. How Can Drug Metabolism and Transporter Genetics Inform Psychotropic Prescribing? *Front Genet*, **11**, 491895 (2020), 10.3389/fgene.2020.491895

53. Gaedigk A., Turner A., Everts R.E., Scott S.A., Aggarwal P., Broeckel U., et al. Characterization of Reference Materials for Genetic Testing of *CYP2D6* Alleles: A GeT-RM Collaborative Project. *The Journal of Molecular Diagnostics*, **21**, 1034-1052 (2019), 10.1016/j.jmoldx.2019.06.007

54. Procedure & Checklist - Preparing SMRTbell Libraries using PacBio Barcoded Universal Primers for Multiplexing Amplicons. Accessed 2022 Aug 23 https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-SMRTbell-Libraries-using-PacBio-Barcoded-Universal-Primers-for-Multiplexing-Amplicons.pdf

55. Fan M., Yarema M.C., Box A., Hume S., Aitchison K.J., Bousman C.A. Identification of high-impact gene-drug pairs for pharmacogenetic testing in Alberta, Canada. *Pharmacogenet Genomics*, **31**, 29-39 (2021), 10.1097/FPC.0000000000000418

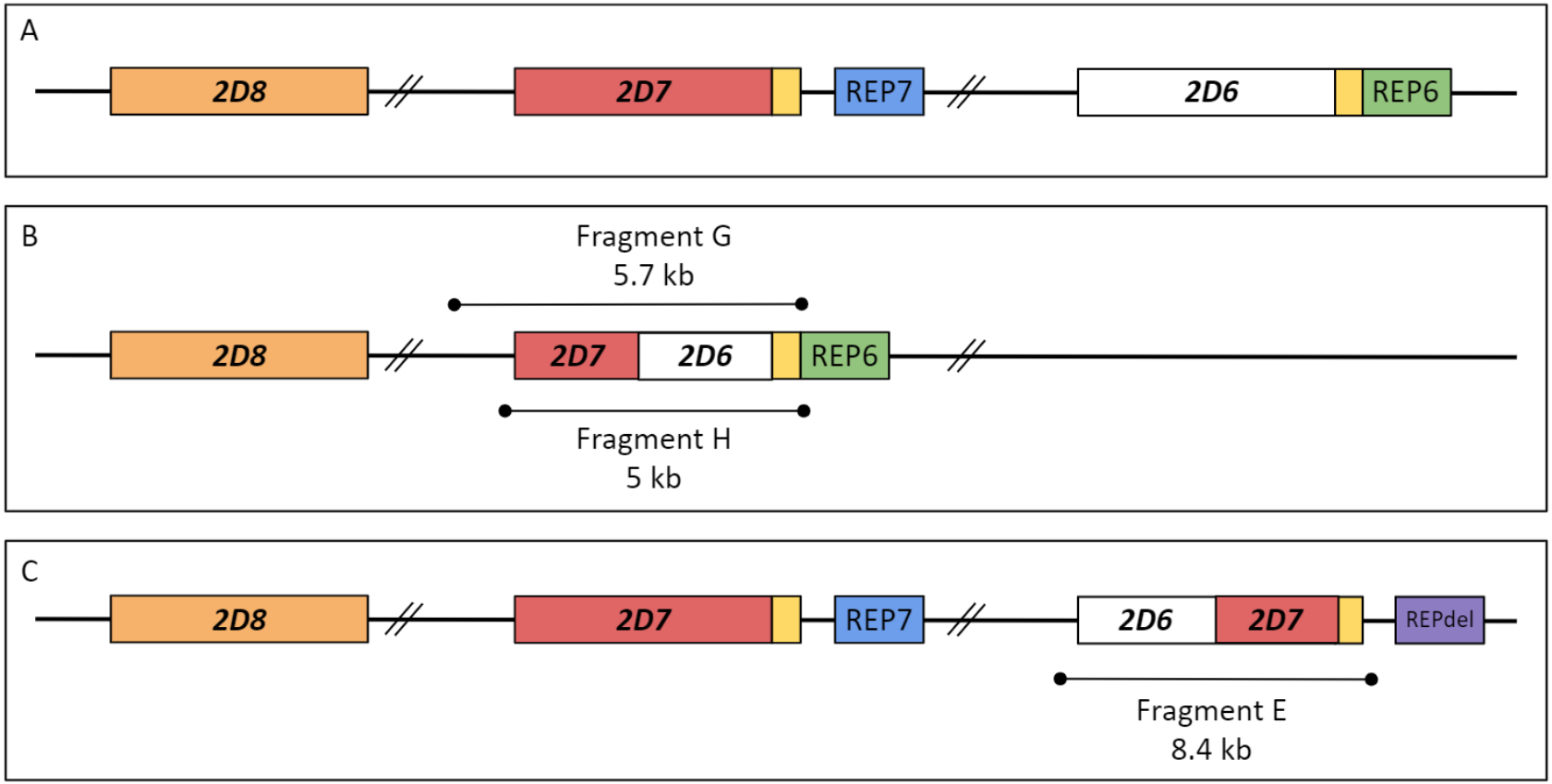


Figure 1. *CYP2D* wild-type locus and fusion genes, with amplicons diagrammed. (**A**) *CYP2D* wild-type locus. (**B**) Example *CYP2D7-2D6* fusion gene. The G and H amplicons are predicted to be 5.7 and 5 kb, respectively. (**C**) Example *CYP2D6-2D7* fusion gene. The E amplicon is estimated as ~8–8.4 kb.

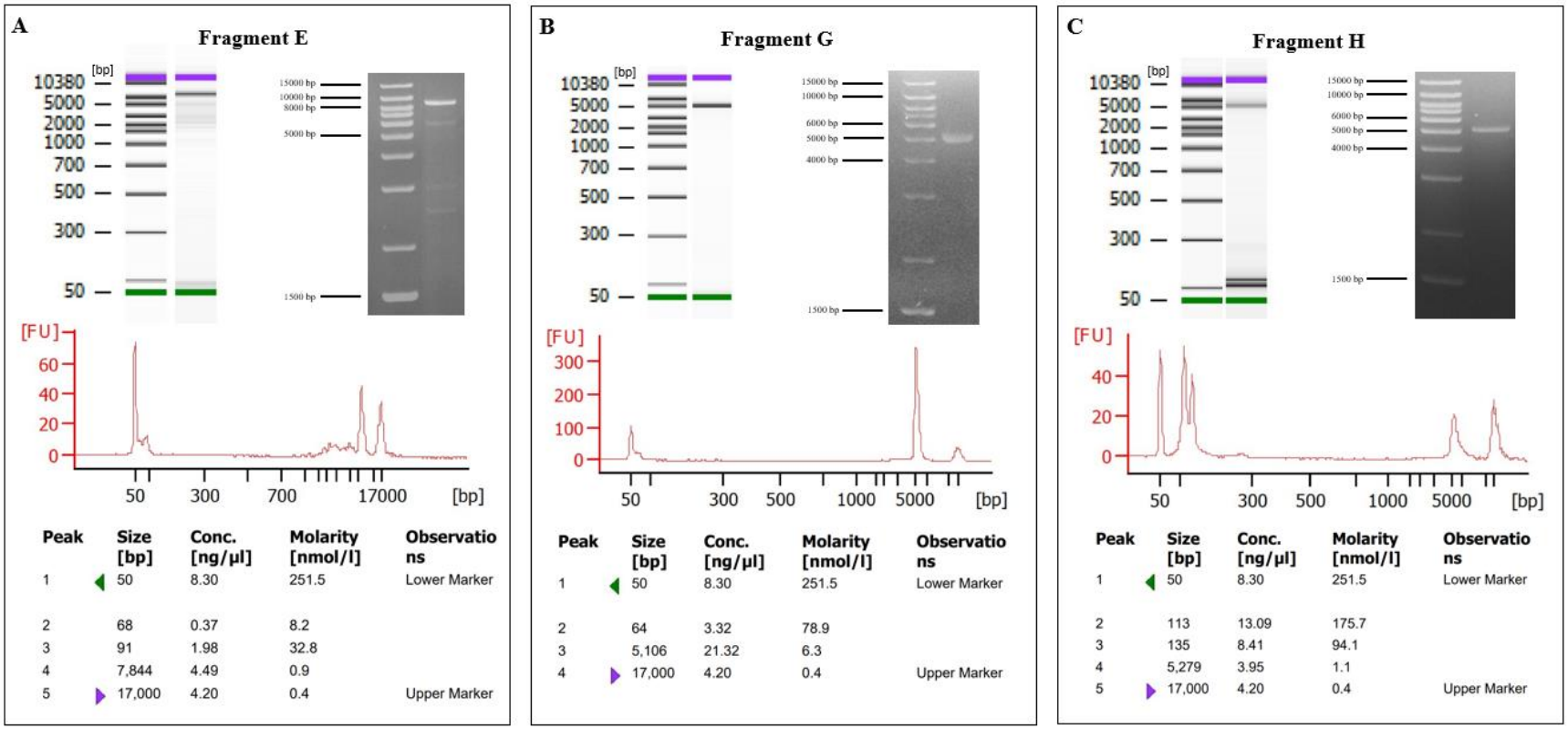
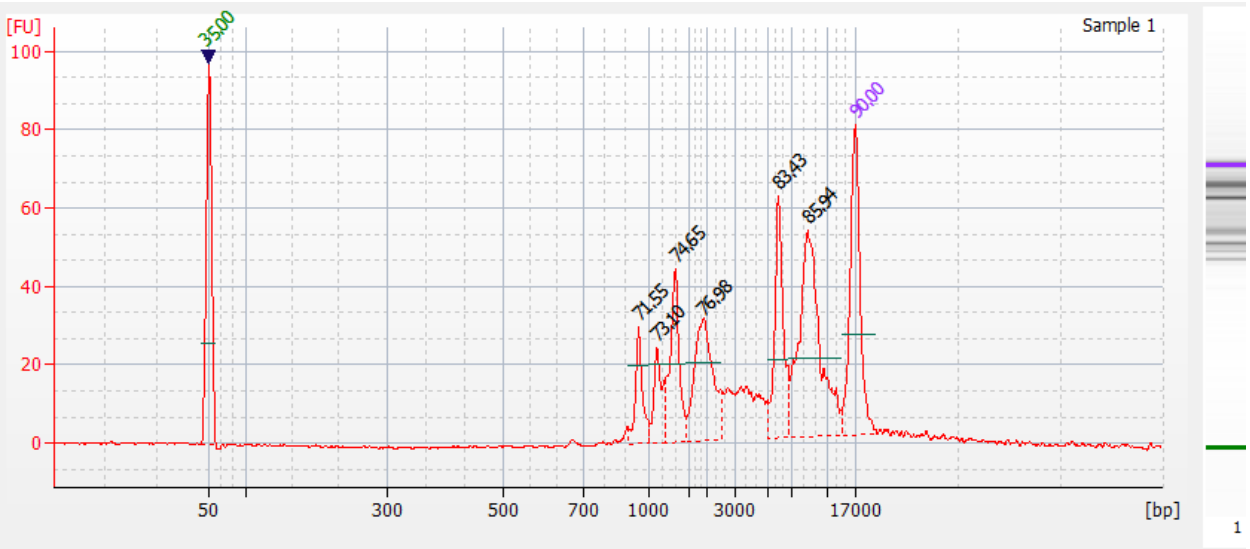


Figure 2. Quality control of first-round L-PCR by agarose gel and Agilent 2100 Bioanalyzer (DNA 12000 assay) electrophoresis. (**A**) Representative E amplicon. (**B**) Representative G amplicon. (**C**) Amplicon H. For the Agilent output, the x-axis indicates length in base pairs (bp) and the y-axis fluorescence intensity in fluorescence units [FU]. The lower marker (50 bp) and upper marker (17000 bp) are the first and last peaks, respectively. Electropherogram plots are transformed into automated gel electrophoresis images (top left), where the bottom marker (green), top marker (purple), and PCR products are visualized alongside the ladder (bp). The 1% agarose gel with a 1 kb Plus DNA Ladder (Thermo Fisher Scientific) is shown in the top right.



		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2		953	1.39	2.2	
3		1,100	1.28	1.8	
4		1,328	2.53	2.9	
5		1,876	3.54	2.9	
6		5,876	3.02	0.8	
7		8,448	6.16	1.1	
8	▶	17,000	4.20	0.4	Upper Marker

Figure 3. Sample pool as visualized on the Agilent 1200 assay. The ~5.9 kb peak represents the G amplicons and the ~8.4 kb peak the E amplicons. The H amplicon are hidden in the region between the G peak and the nonspecific lower peaks.

Table 1. Alignment of G amplicons to reference sequences.

Sample #	Consensus Genotype	Haplotype	Accession number	Filter	Aligned Length	I	D	X	Percent Alignment
22	**13/*4.013*	**13F*	EU093102	4000 – 6000	5002	0	0	0	100
23	**13/*1*	**13F*	EU093102	4000 – 6000	5002	0	0	0	100
24	**13/*1*	**13F*	EU093102	4000 – 6000	5002	0	0	0	100
25	**13+*2/*1*	**13A2*	GQ162807	4000 – 6000	5039	0	0	0	100
28	**13+*10/*36*	**13A2*	GQ162807	4000 – 6000	5039	0	0	0	100
30	**13/*4.013*	**13F*	EU093102	4000 – 6000	5002	0	0	0	100
32	**13+*2/*1*	**13A2*	GQ162807	4000 – 6000	5039	0	0	0	100
33	**13+*2/*1*	**13A2*	GQ162807	4000 – 6000	5039	0	0	0	100
35	**13+*2/*41*	**13A2*	GQ162807	4000 – 6000	5039	0	0	0	100
37 (NA19785)	**13+*2/*1*	**13A2*	GQ162807	4000 – 6000	5039	0	0	0	100

Note: samples 22 and 30 are technical replicates from the same participant. I, D, and X represent the numbers of insertions, deletions, and mismatches, respectively, as compared with the reference sequence.

Table 2. Alignment of H amplicons to reference sequences

Sample #	Consensus Genotype	Pool #	Barcode	Accession number	Filter	Aligned Length	I	D	X	Percent Alignment
47 (NA19785)	**13+*2/*1*	2 & 3	bc1034	EU093102	4000 – 6000	5039	0	0	0	100
38 (NA19785)	**13+*2/*1*	1	bc1039	EU093102	4000 – 6000	5039	0	0	0	100
61	**13/*4*	3	bc1013	EU093102 (*13F)	4000 – 6000	5002	0	1	10	99.78
62	**13/*4*	3	bc1014	EU093102 (*13F)	4000 – 6000	5002	3	1	13	99.66
63	**13/*4*	3	bc1016	EU093102 (*13F)	4000 – 6000	5002	10	1	13	99.52
64	**13/*4*	3	bc1023	EU093102 (*13F)	4000 – 6000	5002	10	1	13	99.52
46	**13/*4*	2 & 3	bc1029	EU093102 (*13F)	4000 – 6000	5002	10	1	13	99.52
65	**13/*4*	3	bc1032	EU093102 (*13F)	4000 – 6000	5002	0	0	3	99.94
36	**13/*4*	1	bc1038	EU093102 (*13F)	4000 – 6000	5002	10	1	14	99.50
29	**13/*4*	1	bc1032	2325bc1013	4000 – 6000	5060				99.86
29	**13/*4*	1	bc1085	EU093102 (*13F)	4000 – 6000	5002	1	6	0	99.86
29	**13/*4*	1	bc1086	EU093102 (*13F)	4000 – 6000	5002	1	6	0	99.64
29	**13/*4*	1	bc1085	EU093102 (*13F)	4000 – 6000	5002	1	6	0	99.86

I, D, and X represent the numbers of insertions, deletions, and mismatches, respectively, as compared with the reference sequence.

Table 3. Alignment of E amplicons to reference sequences. For the hybrid haplotypes containing the 1847G>A defining SNV for *4, where possible, alignments to all three reference sequences are provided. For the rest of the samples aligned to more than one reference sequence, unless the percentage alignments were comparable, the best alignment of the longest length is reported. For technical replicates, the best of the replicates is reported.

Sample #	Consensus Genotype	Haplotype	Accession number	Filter	Aligned Length	I	D	X	Percentage Alignment
2	**68+*2/*1*[a]	**68*	EU530606	7680 - 9680	4550	0	0	0	100
2	**68+*2/*1*[a]	**68*	JF307779	7680 - 9680	4738	0	0	2	99.96
2	**68+*2/*1*[a]	**68*	233bc1065	5000-9000	5750	0	10	3	99.77
3	**68+*2/*1*[a]	**68*	EU530606	7680 - 9680	4550	0	0	0	100
3	**68+*2/*1*[a]	**68*	JF307779	7680 - 9680	4738	0	0	2	99.96
3	**68+*2/*1*[a]	**68*	233bc1065	5000-9000					
4	**13/*4.013*	**4-like*	EU530605	7680 – 9680	6400	0	10	10	99.76
4	**13/*4.013*	**4N*	EU530604	7680 – 9680	7000	1	3	16	99.51
4	**13/*4.013*	**4.013*	PharmVar **4.013*	7680 – 9680	6739	1	10	9	99.50
5	**4.013 +*4/*1*	**4-like*	EU530605	7680 - 9680	6400	2	3	2	99.89
5	**4.013 +*4/*1*	**4N*	EU530604	7680 - 9680	7000	2	3	6	99.84
6	**4.013+*4/*4*	**4-like*	EU530605	7680 – 9680	6400	1	3	12	99.75
6	**4.013+*4/*4*	**4N*	EU530604	7680 – 9680	7000	1	3	16	99.71
6	**4.013+*4/*4*	**4.013*	PV00250 (PharmVar **4.013)*	7680 – 9680	6739	0	10	10	99.70
7	**36 + *10/*35*	**36*	DQ211355 *(*36* partial sequence)	7680 -12000	3004	0	0	0	100
7	**36 + *10/*35*	**36*	PV00460 (PharmVar **36)*	7680 -12000	6746	6	8	17	99.48
7	**36 + *10/*35*	**36*	PV00705 (PharmVar **36.002*)	7680 – 9680	6744	0	3	7	99.75
7	**36 + *10/*35*	**36*	PV01364 (PharmVar **36.003*)	7680 – 9680	6739	4	0	10	99.67
7	**36 + *10/*35*	**36*	PV01720	7680 – 9680	6737	0	3	10	99.70

			(PharmVar *36.004*)						
8	*68+*4/*10	*68	EU530606	7680 -12000	4550	0	0	0	100
8	*68+*4/*10	*68	JF307779	7680 -12000	4738	0	0	2	99.96
9	*4.013+*4/*1	*4-like	EU530605	7680 – 9680	6400	5	4	4	99.78
9	*4.013+*4/*1	*4N	EU530604	7680 – 9680	7000	1	3	16	99.71
9	*4.013+*4/*1	*4.013	PV00250 (PharmVar *4.013)*	7680 – 9680	6739	0	10	10	99.70
10	*4.013+*4/*1	*4-like	EU530605	7680 – 9680	6400	1	1	5	99.89
10	*4.013+*4/*1	*4N	EU530604	7680 – 9680	7000	5	4	1	99.8
10	*4.013+*4/*1	*4.013	PV00250 (PharmVar *4.013)*	7680 – 9680	6739	0	4	10	99.67
11 (NA18545)	*5/*36x2+*10x 2	*36	DQ211355	7680 – 12000	3004	0	0	0	100
11 (NA18545)	*5/*36x2+*10x 2	*36	PV00460 (PharmVar *36)*	7680 – 12000	6448	7	2	1	99.84
11 (NA18545)	*5/*36x2+*10x 2	*36	PV00222 (PharmVar *36.001*)	7680 – 12000	6739	0	10	8	99.73
11 (NA18545)	*5/*36x2+*10x 2	*36	PV00705 (PharmVar *36.002*)	7680 – 9680	6737	0	8	7	99.77
11 (NA18545)	*5/*36x2+*10x 2	*36	PV00705 (PharmVar *36.003*)	7680 – 9680	6739	0	10	9	99.71
11 (NA18545)	*5/*36x2+*10x 2	*36	PV00705 (PharmVar *36.004*)	7680 – 9680	6737	0	8	10	99.73
12	*4.013+*4/*35	*4-like	EU530605	7680 – 9680	6400	5	2	6	99.77
12	*4.013+*4/*35	*4N	EU530604	7680 – 9680	7000	1	3	16	99.71
12	*4.013+*4/*35	*4.013	PV00250 (PharmVar *4.013)*	7680 – 9680	6739	0	10	10	99.70
14	*4.013+*4/*1	*4-like	EU530605	7680 – 9680	6400	1	3	12	99.75
14	*4.013+*4/*1	*4N	EU530604	7680 – 9680	7000	1	3	16	99.71
14	*4.013+*4/*1	*4.013	PV00250 (PharmVar *4.013)*	7680 – 9680	6739	0	10	10	99.70
15	*4.013 +*4/*4.002	*4-like	EU530605	7680 – 9680	6400	3	4	3	99.83

15	**4.013 +*4/*4.002*	**4N*	EU530604	7680 – 9680	7000	1	2	14	99.71
15	**4.013 +*4/*4.002*	**4.013*	PV00250 (PharmVar *4.013)*	7680 – 9680	6739	0	10	11	99.69
16	**4.013 +*4/*4*	**4-like*	EU530605	7680 – 9680	6400	1	3	12	99.75
16	**4.013 +*4/*4*	**4N*	EU530604	7680 – 9680	7000	1	3	16	99.71
16	**4.013 +*4/*4*	**4.013*	PV00250 (PharmVar *4.013)*	7680 – 9680	6739	0	10	10	99.70
18	**4.013+*4/*2*	**4-like*	EU530605	7680 – 9680	6400	1	1	3	99.91
18	**4.013+*4/*2*	**4N*	EU530604	7680 – 9680	7000	0	10	10	99.86
18	**4.013+*4/*2*	**4.013*	PV00250 (PharmVar *4.013)*	7680 – 9680	6739	2	1	7	99.70

[a]Consensus deduced from a combination of AmpliChip CYP450 and TaqMan CNV data in addition to the SMRT data. I, D, and X represent the numbers of insertions, deletions, and mismatches, respectively, as compared with the reference sequence.