3-1-2016

Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6.

Wanqiong Qiao

Yao Yang

Robert Sebra

Geetu Mendiratta

Andrea Gaedigk
Children's Mercy Hospital

See next page for additional authors

Recommended Citation

Qiao, W., Yang, Y., Sebra, R., Mendiratta, G., Gaedigk, A., Desnick, R. J., Scott, S. A. Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6. *Human mutation* 37, 315-323 (2016).

Published in final edited form as:

Hum Mutat. 2016 March ; 37(3): 315–323. doi:10.1002/humu.22936.

Long-read single-molecule real-time (SMRT) full gene sequencing of cytochrome P450-2D6 (*CYP2D6*)

Wanqiong Qiao[1,*], **Yao Yang**[1,*], **Robert Sebra**[1,2], **Geetu Mendiratta**[1], **Andrea Gaedigk**[3,4], **Robert J. Desnick**[1], and **Stuart A. Scott**[1]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[2]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[3]Division of Clinical Pharmacology, Toxicology & Therapeutic Innovation, Children's Mercy Kansas City, Kansas City, MO 64108, USA

[4]School of Medicine, University of Missouri-Kansas City, Kansas City, MO 64108, USA

Abstract

The CYP2D6 enzyme metabolizes ~25% of common medications, yet homologous pseudogenes and copy-number variants (CNVs) make interrogating the polymorphic *CYP2D6* gene with short-read sequencing challenging. Therefore, we developed a novel long-read, full gene *CYP2D6* single-molecule real-time (SMRT) sequencing method using the Pacific Biosciences platform. Long-range PCR and *CYP2D6* SMRT sequencing of 10 previously genotyped controls identified expected star (*) alleles, but also enabled suballele resolution, diplotype refinement, and discovery of novel alleles. Coupled with an optimized variant calling pipeline, *CYP2D6* SMRT sequencing was highly reproducible as triplicate intra- and inter-run non-reference genotype results were completely concordant. Importantly, targeted SMRT sequencing of upstream and downstream *CYP2D6* gene copies characterized the duplicated allele in 15 control samples with *CYP2D6* CNVs. The utility of *CYP2D6* SMRT sequencing was further underscored by identifying the diplotypes of 14 samples with discordant or unclear *CYP2D6* configurations from previous targeted genotyping, which again included suballele resolution, duplicated allele characterization, and discovery of a novel allele and tandem arrangement (*CYP2D6*36+*41*). Taken together, long-read *CYP2D6* SMRT sequencing is an innovative, reproducible, and validated method for full-gene characterization, duplication allele-specific analysis and novel allele discovery, which will likely improve CYP2D6 metabolizer phenotype prediction for both research and clinical testing applications.

CORRESPONDENCE TO: Stuart A. Scott, PhD, Assistant Professor, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1497, New York, NY, 10029, Tel. 212-241-3780, Fax. 212-241-1464, stuart.scott@mssm.edu.

*These authors contributed equally to this manuscript and should be regarded as joint first authors.

Keywords

CYP2D6; single-molecule real-time (SMRT) sequencing; long-read sequencing; Pacific Biosciences; pharmacogenetics; pharmacogenomics; gene duplication

INTRODUCTION

One of the most influential discoveries in the field of human pharmacogenetics has been the identification of polymorphic debrisoquine metabolism in 1977 (Mahgoub, et al., 1977), and the subsequent realization that the 'poor metabolism' trait was inherited in an autosomal recessive fashion due to variant alleles encoding a hepatic cytochrome P450 oxidase (Meier, et al., 1983). The responsible enzyme, cytochrome P450 2D6 (CYP2D6; MIM# 124030), was purified and characterized (Gonzalez, et al., 1988; Gough, et al., 1990; Heim and Meyer, 1990; Kimura, et al., 1989), and is now believed to be directly involved in the metabolism of ~25% of all commonly used drugs (Owen, et al., 2009). The *CYP2D6* gene on chromosome 22q13.2 is highly polymorphic, with over 100 variant star (*) alleles catalogued by the Human Cytochrome P450 (CYP) Allele Nomenclature Committee (Sim and Ingelman-Sundberg, 2010), many of which are associated with reduced or no enzyme activity. Importantly, *CYP2D6* is also prone to copy number variation (CNV), including both gene duplication and deletion, and complex rearrangements with the *CYP2D7* pseudogene, which can significantly influence the interpretation of *CYP2D6* genotyping, sequencing, and phenotype prediction (Ramamoorthy and Skaar, 2011).

Clinical *CYP2D6* testing by targeted genotyping is widely available with result interpretation that typically categorizes individuals into one of four predicted CYP2D6 metabolism phenotypes based on genotype: ultrarapid (UM), extensive (EM), intermediate (IM), and poor (PM) (Gaedigk, et al., 2008; Owen, et al., 2009). The growing interest and potential utility of clinical *CYP2D6* testing is evidenced by recently published practice guidelines for *CYP2D6* genotype-directed codeine (Crews, et al., 2012; Crews, et al., 2014), tricyclic antidepressant (TCA) (Hicks, et al., 2013), and selective serotonin reuptake inhibitor (SSRI) (Hicks, et al., 2015) treatment by the Clinical Pharmacogenetics Implementation Consortium (CPIC) (Relling and Klein, 2011).

Interrogating the polymorphic *CYP2D6* gene is challenging due to the high sequence homology with its pseudogenes (Gaedigk, 2013). As such, many of the currently available *CYP2D6* genetic tests incorporate an initial long-range PCR to specifically amplify long fragments of the *CYP2D6* gene (~2–7 kb) prior to targeted genotyping or other mutation scanning technique (e.g., TaqMan, Sanger sequencing, etc.). The pseudogene homology also can interfere with common next-generation sequencing platforms as the capture of targeted *CYP2D6* regions and subsequent read alignment may erroneously be derived from or attributed to *CYP2D7*, respectively. Moreover, accurate prediction of *CYP2D6* metabolizer status necessitates direct analysis of the duplicated gene copy (or copies) when an increased copy number is detected, particularly when identified concurrently with normal activity and loss-of-function alleles in compound heterozygosity (e.g., **1/*4, DUP*) (Ramamoorthy and Skaar, 2011). Given the importance and polymorphic nature of *CYP2D6* and the paucity of

available *CYP2D6* next-generation sequencing assays, we developed a novel, third-generation single-molecule real-time (SMRT) sequencing assay using the Pacific Biosciences platform with long read lengths that span the entire *CYP2D6* gene, including targeted sequencing of duplicated *CYP2D6* copies when present.

MATERIALS AND METHODS

Samples and Subjects

Commercially available DNA samples with previously reported *CYP2D6* genotypes (Fang, et al., 2014; Pratt, et al., 2010) were acquired from the Coriell Biorepository (Camden, NJ, USA). In addition, peripheral blood samples from healthy adult donors who self-reported their racial and ethnic background [African-American (AA), Asian, Caucasian or Hispanic] and gave informed consent for the use of their DNA for research were obtained from the New York Blood Center (NY, USA) with Institutional Review Board approval as previously described (Martis, et al., 2013b). All personal identifiers were removed, and isolated DNA samples were tested anonymously. Genomic DNA was isolated using the Puregene® DNA Purification kit (Qiagen, Valencia, CA) according to the manufacturer's instructions.

CYP2D6 Variant Nomenclature and Genotyping

The *CYP2D6* allele designations refer to those defined by the Cytochrome P450 Allele Nomenclature Committee (http://www.cypalleles.ki.se/cyp2d6.htm) (Sim and Ingelman-Sundberg, 2010), which uses the M33388.1 GenBank reference sequence (with minor corrections) for *CYP2D6* variant coordinates (http://www.ncbi.nlm.nih.gov/nuccore/M33388) (Kimura, et al., 1989), nucleotide numbering, and star (*) allele definitions. All variant nucleotide positions are numbered according to the historical M33388.1 reference sequence with the 'A' of the ATG start codon as nucleotide 1; however, Supp. Table S1 summarizes all relevant *CYP2D6* variants with both M33388.1 coordinates and current Human Genome Variation Society (HGVS) nomenclature using the NM_000106.5 reference sequence and the 'A' of the ATG start codon as nucleotide 1. In addition, all protein-level variant nomenclature noted below has been confirmed using Mutalyzer (https://www.mutalyzer.nl) and NM_000106.5 with the 'A' of the ATG start codon as nucleotide 1.

Genotyping of 15 variant *CYP2D6* alleles (**2 − *11*, **15*, **17*, **29*, **35*, and **41*) and the gene duplication, was performed using the xTAG CYP2D6 Kit v3 (Luminex Corporation, TX, USA) according to the manufacturer's instructions. Briefly, the regions surrounding the *CYP2D6* variants were long-range PCR-amplified, subjected to allele-specific primer extension, hybridized to specific Luminex microspheres, and sorted on a Luminex 100 xMAP™ platform. Genotypes for each sample were determined using TDAS CYP2D6 version 1.01 software (Luminex Corporation), and the wild-type (*CYP2D6*1*) allele was assigned in the absence of other detectable variant alleles. Targeted genotype results were known prior to SMRT sequencing analysis.

Copy Number Analysis

CYP2D6 copy number was interrogated using commercially available TaqMan® real-time qPCR Copy Number Assays (Applied Biosystems, Carlsbad, CA) as per the manufacturer's

instructions and described previously (Fang, et al., 2014; Martis, et al., 2013a). In brief, FAM™-labeled TaqMan® minor groove binding probe and unlabeled PCR primer assays [Hs04083572_cn (intron 2), Hs00010001_cn (exon 9)] were individually run in a duplex qPCR with a VIC™-labeled RNase P TaqMan® Copy Number Reference Assay (catalogue number: 4403326; Applied Biosystems). Quadruplicate experiments were each performed in 10 μl reactions containing ~10 ng of DNA, 1X TaqMan® Genotyping Master mix, 0.5 μl each of TaqMan® Copy Number and Reference Assays in 384 well plates. Covered plates were run in a 7900HT Fast Real-Time PCR System (Applied Biosystems) and the amplification consisted of a denaturation step at 95°C for 10 min followed by 40 amplification cycles (95°C for 15 sec and 60°C for 60 sec). Data was captured using absolute quantitation with a manual C_T threshold and autobaseline, followed by analysis using CopyCaller™ v1.0 Software (Applied Biosystems) where the number of copies of target sequence was determined by relative quantitation with the comparative C_T ($\Delta\Delta C_T$) method. This method measures the C_T difference (ΔC_T) between target and reference sequences, and then compares the ΔC_T values of test samples to a calibrator sample known to have two copies of the target sequence. The copy number of the target was calculated to be two times the relative quantity.

Full Gene *CYP2D6* Sample Preparation and Barcoding

For sequencing analysis, the entire *CYP2D6* gene ('downstream' copy) was initially amplified as an 8.1 kb PCR using previously reported primers (Gaedigk, et al., 2007). Long-range PCR reactions were performed in 20 μl containing ~40 ng of DNA, 1X SequalPrep™ Reaction buffer (Invitrogen), 0.5 μM of forward and reverse primers (Table 1), and 1.8 units of SequalPrep™ Polymerase. Amplification consisted of an initial denaturation step at 94°C for 2 min followed by 10 amplification cycles (94°C for 10 sec, 63°C for 30 sec, and 68°C for 13 min), another 20 amplification cycles (94°C for 10 sec, 63°C for 30 sec, and 68°C for 13 min + 20 sec/cycle), and a final extension at 72°C for 5 min. These products were used as templates for nested PCR and barcoding prior to SMRT sequencing.

Nested PCR amplified a 5.0 kb fragment and incorporated universal oligonucleotide tags (Table 1) to allow for subsequent barcoding and multiplexed SMRT sequencing. Amplification was as above but with both annealing and extension at 68°C, and an extension time of 6 min given the shorter amplicon length. Nested PCR products were used as template for a final amplification that incorporated unique forward and reverse barcodes for each sample (Table 1). The PCR conditions for this reaction were identical to the second round nested PCR.

CYP2D6 Duplication Analysis

Duplicated *CYP2D6* gene copies were interrogated directly by a unique long-range PCR amplicon that specifically amplified 'upstream' (duplicated) copies of the gene. As previously described (Gaedigk, et al., 2007), these primers specifically amplify an 8.6 kb fragment that encompasses the entire upstream *CYP2D6* copy (or 10.2 kb if the duplicated gene copy carries *CYP2D7*-derived sequences as in *CYP2D6*36* (Gaedigk, et al., 2006)), allowing for star (*) allele interrogation of the duplicated gene copy. The PCR conditions for the upstream long-range PCR were identical to the reaction carried out to amplify

downstream *CYP2D6* gene copies. These products were also used as templates for nested PCR and barcoding, as detailed above, prior to SMRT sequencing.

Single-Molecule Real-Time (SMRT) Sequencing

Multiplex SMRT sequencing was performed as previously described (Yang, et al., 2015). In brief, all PCR amplicons were purified by Agencourt® AMPure® XP beads and quantified by Nanodrop 1000. After purification, equal molecule quantities of PCR amplicons were pooled, with the required volume of each amplicon calculated by the following formula:

$$V(i)=\frac{M}{n \times C(i)}$$

Where *M* is the total mass of pooled PCR amplicons, *n* is the total number of samples, *V(i)* is the volume of each PCR amplicon, and *C(i)* is the concentration of each amplicon. A total of 500 ng of pooled PCR amplicons were submitted for SMRT sequencing.

SMRT sequencing was performed according to the P5-C3 Pacific Biosciences protocol with a movie collection time of 180 minutes. In brief, pooled PCR amplicons were quantified using Qubit fluorometric analysis (Life Technologies) and a Bioanalysis 12000 chip (Agilent Technologies) to assess PCR amplicon quality, size, and quantity. Additionally, barcoded and pooled amplicons were purified using Ampure XP Solid Phase Reversible Immobilization (Beckman Coulter) at 0.8-fold volume. SMRTbell libraries were constructed using end-repair, ligation, and exonuclease purification strategies detailed in the Pacific Biosciences P5-C3 Template Preparation Kit protocols. SMRTbell templates were then bound to polymerase molecules for 4 hours at 25°C using 3 nM of the amplicon SMRTbell library and excess P5 DNA polymerase at a concentration of 9 nM as previously described (Rasko, et al., 2011). The polymerase-template complexes were immobilized at 250 pM for 30 min on nanofabricated SMRTcells containing an array of zero-mode waveguides (ZMWs), and ZMWs were analyzed for sequencing to generate reads using a 1×180-minute collection protocol. Circular consensus sequencing (CCS) was then employed using multiple passes on each SMRTbell to generate CCS reads with higher accuracy for data analysis and the Reads of Insert pipeline was utilized with a filter of 85% accuracy and 1-pass, prior to variant analyses.

SMRT Data Analysis and Variant Calling

SMRT sequencing analysis included demultiplexing, alignment, sequencing quality score recalibration, and *CYP2D6* variant calling (Figure 1). Raw sequencing data in FASTQ format were demultiplexed into unique samples according to barcode sequences at both ends of sequencing reads using the NGSutils next-generation sequencing data analysis software kit (Breese and Liu, 2013). The 12 bp at the proximal ends of the barcodes were used for demultiplexing, which allowed one mismatch (including insertions/deletions). Sequencing reads were aligned to the targeted *CYP2D6* gene region (chr22:42,522,044–42,527,019; hg19) using BWA-MEM version 0.7.12 with dedicated parameter settings for Pacific Biosciences SMRT sequencing (Li, 2013). A python script subsequently was developed and used to correct random sequencing errors using alignment information from the SAM files.

The Amplicon Long-read Error Correction (ALEC) script with additional details on its functionality is accessible at Github (https://github.com/scottlab/ALEC.git) and a manuscript detailing and evaluating its functionality is currently in preparation. Base quality score recalibration and variant calling were performed using the best practices of the Genome Analysis Tool Kit (GATK) (DePristo, et al., 2011; McKenna, et al., 2010) with the target *CYP2D6* gene region and dbSNP138 as reference. Identified genotypes were translated to common star (*) allele nomenclature using the M33388.1 *CYP2D6* reference sequence (Kimura, et al., 1989) with modifications and haplotype definitions according to the Human Cytochrome P450 (CYP) Allele Nomenclature Committee (http://www.cypalleles.ki.se/cyp2d6.htm) (Sim and Ingelman-Sundberg, 2010). For HGVS nomenclature of identified *CYP2D6* variants, see Supp. Table S1.

Allele-specific PCR and Variant Phasing

Allele-specific PCR (AS-PCR) was performed to phase selected missense variants identified by SMRT sequencing. The novel 3226A>G and 3235A>G variants identified in NA17222 were phased in relation to the neighboring 2850C>T (*2) allele, and the 2615–2617delAAG (*9) variant identified in CAUC073 was phased in relation to the neighboring 3183G>A allele (Supp. Figure S1). To increase the specificity of the AS-PCR primers, artificial nucleotide mismatches were introduced into the third position from the 3'end of each allele-specific primer (Hirotsu, et al., 2010; Liu, et al., 2012) (Table 1). AS-PCR reactions were performed in 20 μl using the long-range PCR products as templates (see *Full Gene* CYP2D6 *Sample Preparation and Barcoding*), 1X Takara PCR buffer, 0.2 mM dNTPs, 0.2 μM of forward and reverse primers (Table 1), and 2 units of Takara Taq DNA polymerase. Amplification consisted of an initial denaturation step at 95°C for 3 min followed by 25 amplification cycles (95°C for 45 sec, 66°C for 30 sec, and 72°C for 1 min), and a final extension at 72°C for 5 min. All amplicons were subjected to Sanger sequencing to confirm the phase of sequenced variants in relation to the anchored allele-specific primer. Primer specificities were confirmed and validated by parallel reactions using templates that did not harbor the targeted phasing alleles.

In addition, the *CYP2D6* full gene haplotypes were phased in both the NA17222 and CAUC073 samples by allele-specific long-range PCR (ASXL-PCR) and Sanger sequencing as previously described (Gaedigk, et al., 2015) using primers anchored to *CYP2D6* −1584C (NA17222) and −2523G (CAUC073).

RESULTS

SMRT Sequencing Read Length and Quality Metrics

127,116 reads were generated for all 72 amplicons described below by sequencing five SMRTcells, where 81,649 (73.2%) had barcodes successfully distinguished and demultiplexed. The insert subread lengths had an average of 1648 bp and a median of 896 bp; however, only those above 1000 bp (45.82%) were used for downstream variant calling analyses to increase the coverage of full length subreads (Figure 2A). The sequencing depth after size selection is shown in Figure 2B. The most common apparent pre-assembly sequencing errors were insertions (1.82%), followed by deletions (1.19%) and single

nucleotide miscalls (0.59%). The identified sequencing errors were nonsystematic and randomly distributed across reads, as opposed to being concentrated at more distal positions. Although alignment of SMRT sequencing data with BWA-MEM takes into account some of the characteristics of SMRT sequencing (Li, 2013), a correction was performed to remove randomly distributed artificial sequencing errors according to the SAM file error feature (see *Materials and Methods*).

Full Gene *CYP2D6* SMRT Sequencing Validation

Ten DNA samples with previously determined and/or reported (Pratt, et al., 2010) *CYP2D6* diplotypes were used to develop and validate *CYP2D6* SMRT sequencing, including seven commercially available and three internal samples (Table 2). All samples were genotyped for 15 *CYP2D6* variant alleles (**2 − *11*, **15*, **17*, **29*, **35*, and **41*) and the gene duplication, as well as undergoing copy number assessment with qPCR probes specific to intron 2 and exon 9. These validation samples together harbored ten different variant *CYP2D6* alleles (**2 − *4*, **6*, **9*, **10*, **17*, **29*, **35*, and **41*). Although NA16688 did not have a gene duplication by Luminex genotyping (**2/*10*), qPCR analysis identified three copies of intron 2 and two copies of exon 9, suggesting the presence of a **36+*10* tandem allele due to the *CYP2D7*-derived exon 9 and downstream sequences characteristic of **36* (Gaedigk, et al., 2006). All 10 samples subsequently were amplified by long-range PCR using both *CYP2D6* 'upstream' and 'downstream' gene copy primers, and all first round amplicons were used as templates for nested PCR to add universal flanking primers and SMRT sequencing barcodes (Figure 3A; see *Materials and Methods*). Only downstream copies of *CYP2D6* could be amplified by long-range PCR for those samples with two gene copies by qPCR; however, NA16688 amplified both upstream and downstream products (Figure 3B) consistent with the qPCR results. The resulting 5.0 kb barcoded amplicons were purified, quantitated, pooled, and subjected to SMRT sequencing.

SMRT sequencing identified the expected genotypes of all the validation samples, but also resulted in additional variant information for most samples (Table 2 and Supp. Table S2). The previously reported NA16688 *CYP2D6* diplotype was refined from **2/*10* to **2M/*36+*10B*, NA17280 was refined from **2/*3* to **3A/*59*, and the diplotypes of other samples were refined to suballeles where possible (e.g., **1A*, **2M*, etc.). In addition, NA17222 was genotyped by Luminex as **1/*2* due to heterozygosity at −1584C>G, 1661G>C, 2850C>T and 4180G>C; however, *CYP2D6* SMRT sequencing detected two additional coding variants in this sample [3226A>G (rs61736517; p.H352R) and 3235A>G (rs202102799; p.Y355C)] (Table 2 and Supp. Table S2). AS-PCR anchored at 2850C>T followed by Sanger sequencing confirmed that these two novel variants were on the 2850C (**1*) haplotype (Supp. Figure S1) and the full gene haplotype was elucidated by ASXL-PCR (Supp. Table S3). This novel allele has been named *CYP2D6*108* by the Cytochrome P450 Allele Nomenclature Committee. Similarly, an internal sample (ASIAN048) genotyped by Luminex as *CYP2D6*1/*29* due to heterozygosity at 1659G>A (rs61736512; p.V136M), was also submitted to the Cytochrome P450 Allele Nomenclature Committee as *CYP2D6* SMRT sequencing confirmed the heterozygous 1659G>A variant and no other coding variants in this sample (Table 2 and Supp. Table S2). Although related to both **29* and **70*, this novel allele has been named *CYP2D6*107* by the Nomenclature Committee, as

1659G>A (p.V136M) has previously been shown to result in reduced enzyme activity when expressed independently of other *CYP2D6* coding variants (Wennerholm, et al., 2001).

Three of the validation samples were used for reproducibility analyses, including triplicate intra-run sequencing of NA17247 and inter-run sequencing of NA12244 and NA17280. SMRT sequencing identified expected star (*) allele diplotypes across all replicates. In addition, when considering all identified variants in each sample (including intronic), triplicate SMRT sequencing of NA17247 and NA17280 had average non-reference genotype concordances of 1.0. This high reproducibility was enabled by a novel, long-read error correction script that was developed and employed during the variant calling analysis pipeline (see *Materials and Methods*). Notably, the uncorrected NA17247 and NA17280 average non-reference genotype concordances were 0.978 and 0.875, respectively. Non-reference genotype concordances of NA12244 were not informative due to only a small number of variants called versus the hg19 reference sequence, which contains the *CYP2D6* intron 1 conversion, 1661G>C, 2850C>T, 4180G>C, and other common *CYP2D6*2* single nucleotide variants.

CYP2D6 Duplication Allele-Specific SMRT Sequencing

Ten commercially available and five internal DNA samples with previously determined and/or reported (Fang, et al., 2014; Pratt, et al., 2010) *CYP2D6* CNV alleles were used as controls to assess if the upstream/downstream long-range PCR strategy could enable the successful characterization of duplicated *CYP2D6* copies by SMRT sequencing (Table 3; Supp. Table S4). As expected, all samples with qPCR results indicating three or four total copies successfully amplified a long-range upstream PCR product in addition to the downstream copies (Figure 2B).

SMRT sequencing of upstream and downstream *CYP2D6* gene copies identified the haplotypes of the duplicated alleles in all samples (Table 3), which were consistent with the previously reported commercially available CNV controls (Fang, et al., 2014; Pratt, et al., 2010) with copy number gains by qPCR. Moreover, when coupled with the qPCR results, duplication SMRT sequencing could refine the *CYP2D6* diplotypes with suballele resolution and exact number of allele copies (e.g., **1Ax2/*2M*, **1A/*4x3*). The NA19152 sample was genotyped as *CYP2D6*1/*29, DUP* by Luminex but was revised to **29/*43x2* with SMRT sequencing, consistent with the recently reported genotyping of this cell line (Fang, et al., 2014).

CYP2D6 SMRT Sequencing and Diplotype Clarification

A previously reported pharmacogenetic reference material study identified 12 Coriell samples that could not be confidently assigned consensus *CYP2D6* diplotypes due to discrepant results from five different targeted genotyping platforms (Pratt, et al., 2010). These 12 samples and an additional two internal samples with unclear star (*) allele configurations based on targeted genotyping were subjected to *CYP2D6* SMRT sequencing in an effort to resolve their diplotypes. As detailed in Table 4 and Supp. Table S5, qPCR and SMRT sequencing identified the *CYP2D6* copy number and diplotypes in all 14 samples. Some of the previously reported discrepancies were due to different variants being

interrogated by the commercial platforms used in the reference material study. In addition to confirming consensus diplotypes, *CYP2D6* SMRT sequencing enabled suballele resolution, genotype refinement, duplicated allele characterization, and discovery of a novel tandem arrangement (*CYP2D6*36+*41*). Specifically, *CYP2D6* SMRT sequencing refined NA17244 from **2/*4,DUP* to **2Mx2/*4x2*, NA17084 from **1/*10* to **1/*36+*10*, and NA17287 from **1/*1* to **1A/*83*.

Interestingly, Luminex genotyping could not infer a *CYP2D6* star (*) allele diplotype for sample CAUC073 but called heterozygous variants at 100C>T, 1661G>C, 2613delAGA, 3183G>A and 4180G>C, which were confirmed by SMRT sequencing and suggested a **9/*10B* diplotype (but with 3183G>A) (Table 4 and Supp. Table S5). However, phasing by AS-PCR indicated that the 3183G>A (rs59421388; p.V338M) variant was on the same haplotype as 2615_2617delAAG (**9*; rs5030656; p.K281del) (Supp. Figure S1), and the full gene haplotype was further elucidated by ASXL-PCR (Supp. Table S3). This novel allele has been named *CYP2D6*109* by the Nomenclature Committee. NA17243 was genotyped as *CYP2D6*4/*35* by Luminex, which was confirmed in the downstream copy by *CYP2D6* SMRT sequencing; however, sequencing also revealed complex upstream copies that involve a significant conversion with *CYP2D7* and, therefore, are not likely to encode a functional enzyme (Supp. Table S5).

DISCUSSION

We developed a novel, third-generation SMRT sequencing assay capable of long read lengths that span the entire ~5.0 kb of the *CYP2D6* gene. *CYP2D6* SMRT sequencing was validated against controls with previously reported consensus diplotypes and CNV alleles, which resulted in expected star (*) alleles, but with additional genotype refinement, diplotype reclassification, duplication allele-specific characterization, and novel allele discovery (*CYP2D6*107−*109*). The application of *CYP2D6* SMRT sequencing was highlighted by characterizing samples with discrepant or unclear *CYP2D6* configurations from previous targeted genotyping, suggesting that this technique should have significant utility for both research and clinical testing.

Although targeted genotyping of *CYP2D6* is widely available using several commercial platforms, sequencing the polymorphic *CYP2D6* gene is challenging due to homologous pseudogenes and relatively common structural rearrangements (Gaedigk, 2013). As such, short-read sequencing is not an effective approach for *CYP2D6* full gene characterization. Although a Pacific Biosciences *CYP2D6* SMRT sequencing assay has not previously been reported, long-read MinION nanopore sequencing from Oxford Nanopore Technologies Inc. was very recently tested using *CYP2D6* long-range PCR amplicons from the NA12878 CEPH HapMap cell line (Ammar, et al., 2015). However, despite using an alignment algorithm developed for long error-prone sequencing reads (BLASR) (Chaisson and Tesler, 2012), the MinION nanopore results suggested an ambiguous *CYP2D6* diplotype due to the presence of three distinct haplotypes (**2*, **3*, and **4*), which were hypothesized by the authors to be due to either PCR template switching or sample contamination (Ammar, et al., 2015). Although the potential utility of long-read sequencing was clearly highlighted by this small study, which also included sequencing of the *HLA-A* and *HLA-B* regions (Ammar, et

al., 2015), it is possible that their 30% variant calling threshold may have resulted in false positive or negative variants due to the polymorphic nature and complexity of the *CYP2D6* locus and the ~25–30% error rate characteristic of MinION long-read sequencing (Ashton, et al., 2015).

During our validation of *CYP2D6* SMRT sequencing, the random error rate characteristic of the Pacific Biosciences platform resulted in a number of likely false positive variant calls, which often involved the important *CYP2D6*4* loss-of-function variant (1846G>A; rs3892097) due to sequence alignment issues from its neighboring poly-G tract. This prompted our development of an amplicon long-read error correction script, which was applied to the sequencing data prior to a re-alignment to reduce the number of low frequency and poor quality variants, typically single base insertion/deletions. This proved to be a critical component of *CYP2D6* SMRT sequencing data analysis as evidenced by the increased reproducibility observed following correction compared to the uncorrected genotype concordances.

Although there are over 100 variant *CYP2D6* star (*) alleles catalogued by the Nomenclature Committee, most targeted genotyping platforms only interrogate a small subset with established functional effect. Consequently, the ‘normal’ wild-type designation (i.e., *CYP2D6*1/*1*) is often a default diplotype that is assigned in the absence of any other detected variant alleles included in a test panel. This commonly used system results in some alleles being incorrectly classified as *CYP2D6*1* when they actually carry a less common functional variant allele that was not directly genotyped. Although our method development study was not designed to determine the prevalence of *CYP2D6* star (*) alleles that are missed or incorrectly classified by common targeted genotyping assays (particularly as some samples in our study were specifically chosen due to ambiguous genotyping results), it is notable that ~20% of the samples in our study tested by SMRT sequencing were revised to either a non-genotyped or novel star (*) allele.

Predicting *CYP2D6* metabolizer phenotype status from patient diplotypes is based on the available genotype/phenotype data, but ultimately is challenging and an imperfect inference (Hertz, et al., 2015). A related system based on a continuum of activity scores for different allele activities has also been proposed for phenotype prediction (Hicks, et al., 2014); however, any *CYP2D6* phenotype prediction classification system is ultimately based on the genotype data that is available. Full gene resolution techniques such as *CYP2D6* SMRT sequencing will result in the identification of more precise diplotypes, and ultimately more refined phenotype prediction, but the increased identification of rare and novel star (*) alleles indicates that functional studies are increasingly going to be needed to determine the effect of these low frequency sequence variants on enzyme activity. This incomplete interrogation of *CYP2D6* by previously reported targeted genotyping studies has likely contributed to conflicting results between research groups (Kiyotani, et al., 2013; Province, et al., 2014), suggesting that *CYP2D6* SMRT sequencing may be a very useful method for clinical research.

Although sequencing the entire *CYP2D6* gene can identify variants of uncertain significance, this approach has the clear advantage of providing a more comprehensive

landscape of the *CYP2D6* gene for allele discovery. However, in addition to research applications, clinical *CYP2D6* testing is increasingly accessible and being adopted by physicians to inform pharmacotherapy, which is further supported by recent *CYP2D6* genotype-directed practice guidelines (Crews, et al., 2012; Crews, et al., 2014; Hicks, et al., 2015; Hicks, et al., 2013). As such, full gene *CYP2D6* SMRT sequencing may also be useful for routine clinical testing. Of note, the ability to barcode and multiplex samples in SMRT sequencing runs makes the per sample cost of *CYP2D6* SMRT sequencing comparable to many currently available commercial *CYP2D6* genotyping assays.

In conclusion, *CYP2D6* SMRT sequencing is a validated, third-generation long-read sequencing method, which is highly reproducible when coupled with our optimized variant calling pipeline. The capacity to interrogate the entire *CYP2D6* gene in a single sequencing read as well as specifically characterize duplicated alleles when present facilitates full gene resolution and improved *CYP2D6* metabolizer phenotype prediction for both research and clinical testing applications. In addition, our long-range amplicon SMRT sequencing strategy could easily be expanded beyond *CYP2D6* as a multiplexed pharmacogenetic or Mendelian disease gene panel, and/or for interrogating other structurally challenging regions of the human genome.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) through grant K23 GM104401 (S.A.S.), and through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. The authors thank Dr. Suparna Martis for technical assistance during the outset of this study.

References

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. [PubMed: 25901276]

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol. 2015; 33(3):296–300. [PubMed: 25485618]

Breese MR, Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics. 2013; 29(4):494–6. [PubMed: 23314324]

Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012; 13:238. [PubMed: 22988817]

Crews KR, Gaedigk A, Dunnenberger HM, Klein TE, Shen DD, Callaghan JT, Kharasch ED, Skaar TC. Clinical Pharmacogenetics Implementation C. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype. Clin Pharmacol Ther. 2012; 91(2):321–6. [PubMed: 22205192]

Crews KR, Gaedigk A, Dunnenberger HM, Leeder JS, Klein TE, Caudle KE, Haidar CE, Shen DD, Callaghan JT, Sadhasivam S, Prows CA, Kharasch ED, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. Clin Pharmacol Ther. 2014; 95(4):376–82. [PubMed: 24458010]

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43(5):491–8. [PubMed: 21478889]

Fang H, Liu X, Ramirez J, Choudhury N, Kubo M, Im HK, Konkashbaev A, Cox NJ, Ratain MJ, Nakamura Y, O'Donnell PH. Establishment of CYP2D6 reference samples by multiple validated genotyping platforms. Pharmacogenomics J. 2014

Gaedigk A. Complexities of CYP2D6 gene analysis and interpretation. Int Rev Psychiatry. 2013; 25(5):534–53. [PubMed: 24151800]

Gaedigk A, Bradford LD, Alander SW, Leeder JS. CYP2D6*36 gene arrangements within the cyp2d6 locus: association of CYP2D6*36 with poor metabolizer status. Drug Metab Dispos. 2006; 34(4): 563–9. [PubMed: 16415111]

Gaedigk A, Ndjountche L, Divakaran K, Dianne Bradford L, Zineh I, Oberlander TF, Brousseau DC, McCarver DG, Johnson JA, Alander SW, Wayne Riggs K, Steven Leeder J. Cytochrome P4502D6 (CYP2D6) gene locus heterogeneity: characterization of gene duplication events. Clin Pharmacol Ther. 2007; 81(2):242–51. [PubMed: 17259947]

Gaedigk A, Riffel AK, Leeder JS. CYP2D6 Haplotype Determination Using Long Range Allele-Specific Amplification: Resolution of a Complex Genotype and a Discordant Genotype Involving the CYP2D6*59 Allele. J Mol Diagn. 2015

Gaedigk A, Simon SD, Pearce RE, Bradford LD, Kennedy MJ, Leeder JS. The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. Clin Pharmacol Ther. 2008; 83(2):234–42. [PubMed: 17971818]

Gonzalez FJ, Skoda RC, Kimura S, Umeno M, Zanger UM, Nebert DW, Gelboin HV, Hardwick JP, Meyer UA. Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. Nature. 1988; 331(6155):442–6. [PubMed: 3123997]

Gough AC, Miles JS, Spurr NK, Moss JE, Gaedigk A, Eichelbaum M, Wolf CR. Identification of the primary gene defect at the cytochrome P450 CYP2D locus. Nature. 1990; 347(6295):773–6. [PubMed: 1978251]

Heim M, Meyer UA. Genotyping of poor metabolisers of debrisoquine by allele-specific PCR amplification. Lancet. 1990; 336(8714):529–32. [PubMed: 1975039]

Hertz DL, Snavely AC, McLeod HL, Walko CM, Ibrahim JG, Anderson S, Weck KE, Magrinat G, Olajide O, Moore S, Raab R, Carrizosa DR, et al. In vivo assessment of the metabolic activity of CYP2D6 diplotypes and alleles. Br J Clin Pharmacol. 2015

Hicks JK, Bishop JR, Sangkuhl K, Muller DJ, Ji Y, Leckband SG, Leeder JS, Graham RL, Chiulli DL, ALL, Skaar TC, Scott SA, et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and CYP2C19 Genotypes and Dosing of Selective Serotonin Reuptake Inhibitors. Clin Pharmacol Ther. 2015

Hicks JK, Swen JJ, Gaedigk A. Challenges in CYP2D6 phenotype assignment from genotype data: a critical assessment and call for standardization. Curr Drug Metab. 2014; 15(2):218–32. [PubMed: 24524666]

Hicks JK, Swen JJ, Thorn CF, Sangkuhl K, Kharasch ED, Ellingrod VL, Skaar TC, Muller DJ, Gaedigk A, Stingl JC. Clinical Pharmacogenetics Implementation C. Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants. Clin Pharmacol Ther. 2013; 93(5):402–8. [PubMed: 23486447]

Hirotsu N, Murakami N, Kashiwagi T, Ujiie K, Ishimaru K. Protocol: a simple gel-free method for SNP genotyping using allele-specific primers in rice and other plant species. Plant Methods. 2010; 6:12. [PubMed: 20409329]

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 Hum Genet. 1989; 45(6):889–904. [PubMed: 2574001]

Kiyotani K, Mushiroda T, Zembutsu H, Nakamura Y. Important and critical scientific aspects in pharmacogenomics analysis: lessons from controversial results of tamoxifen and CYP2D6 studies. J Hum Genet. 2013; 58(6):327–33. [PubMed: 23657426]

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013:1–3. arXiv 1303.3997v2.

Liu J, Huang S, Sun M, Liu S, Liu Y, Wang W, Zhang X, Wang H, Hua W. An improved allele-specific PCR primer design method for SNP marker analysis and its application. Plant Methods. 2012; 8(1):34. [PubMed: 22920499]

Mahgoub A, Idle JR, Dring LG, Lancaster R, Smith RL. Polymorphic hydroxylation of Debrisoquine in man. Lancet. 1977; 2(8038):584–6. [PubMed: 71400]

Martis S, Mei H, Vijzelaar R, Edelmann L, Desnick RJ, Scott SA. Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. Pharmacogenomics J. 2013a; 13(6):558–66. [PubMed: 23164804]

Martis S, Peter I, Hulot JS, Kornreich R, Desnick RJ, Scott SA. Multi-ethnic distribution of clinically relevant CYP2C genotypes and haplotypes. Pharmacogenomics J. 2013b; 13(4):369–77. [PubMed: 22491019]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. [PubMed: 20644199]

Meier PJ, Mueller HK, Dick B, Meyer UA. Hepatic monooxygenase activities in subjects with a genetic defect in drug oxidation. Gastroenterology. 1983; 85(3):682–92. [PubMed: 6603386]

Owen RP, Sangkuhl K, Klein TE, Altman RB. Cytochrome P450 2D6. Pharmacogenet Genomics. 2009; 19(7):559–62. [PubMed: 19512959]

Pratt VM, Zehnbauer B, Wilson JA, Baak R, Babic N, Bettinotti M, Buller A, Butz K, Campbell M, Civalier C, El-Badry A, Farkas DH, et al. Characterization of 107 genomic DNA reference materials for CYP2D6, CYP2C19, CYP2C9, VKORC1, and UGT1A1: a GeT-RM and Association for Molecular Pathology collaborative project. J Mol Diagn. 2010; 12(6):835–46. [PubMed: 20889555]

Province MA, Goetz MP, Brauch H, Flockhart DA, Hebert JM, Whaley R, Suman VJ, Schroth W, Winter S, Zembutsu H, Mushiroda T, Newman WG, et al. CYP2D6 genotype and adjuvant tamoxifen: meta-analysis of heterogeneous study populations. Clin Pharmacol Ther. 2014; 95(2): 216–27. [PubMed: 24060820]

Ramamoorthy A, Skaar TC. Gene copy number variations: it is important to determine which allele is affected. Pharmacogenomics. 2011; 12(3):299–301. [PubMed: 21449666]

Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, et al. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 2011; 365(8):709–17. [PubMed: 21793740]

Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clin Pharmacol Ther. 2011; 89(3):464–7. [PubMed: 21270786]

Sim SC, Ingelman-Sundberg M. The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. Hum Genomics. 2010; 4(4): 278–81. [PubMed: 20511141]

Wennerholm A, Johansson I, Hidestrand M, Bertilsson L, Gustafsson LL, Ingelman-Sundberg M. Characterization of the CYP2D6*29 allele commonly present in a black Tanzanian population causing reduced catalytic activity. Pharmacogenetics. 2001; 11(5):417–27. [PubMed: 11470994]

Yang Y, Sebra R, Pullman BS, Qiao W, Peter I, Desnick RJ, Geyer CR, DeCoteau JF, Scott SA. Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS). BMC Genomics. 2015; 16:350. [PubMed: 25943404]

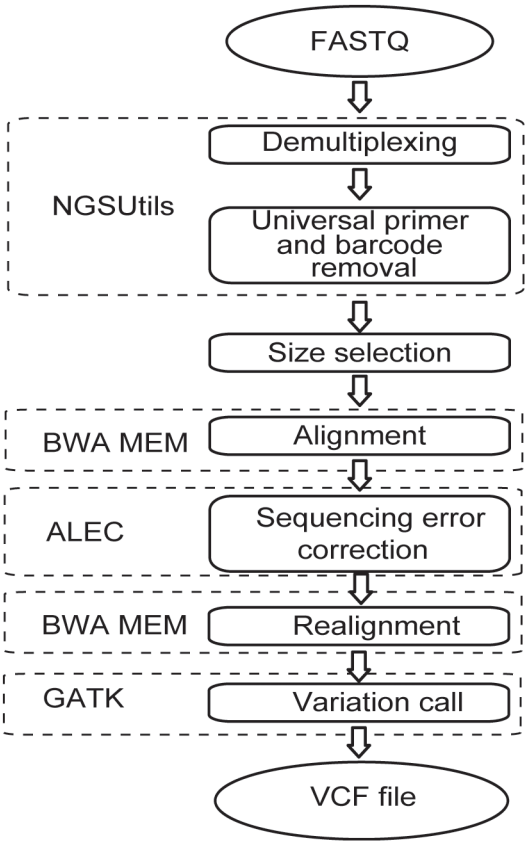

FIGURE 1. Illustration of the *CYP2D6* SMRT sequencing analysis pipeline
For a detailed description see *Materials and Methods*.

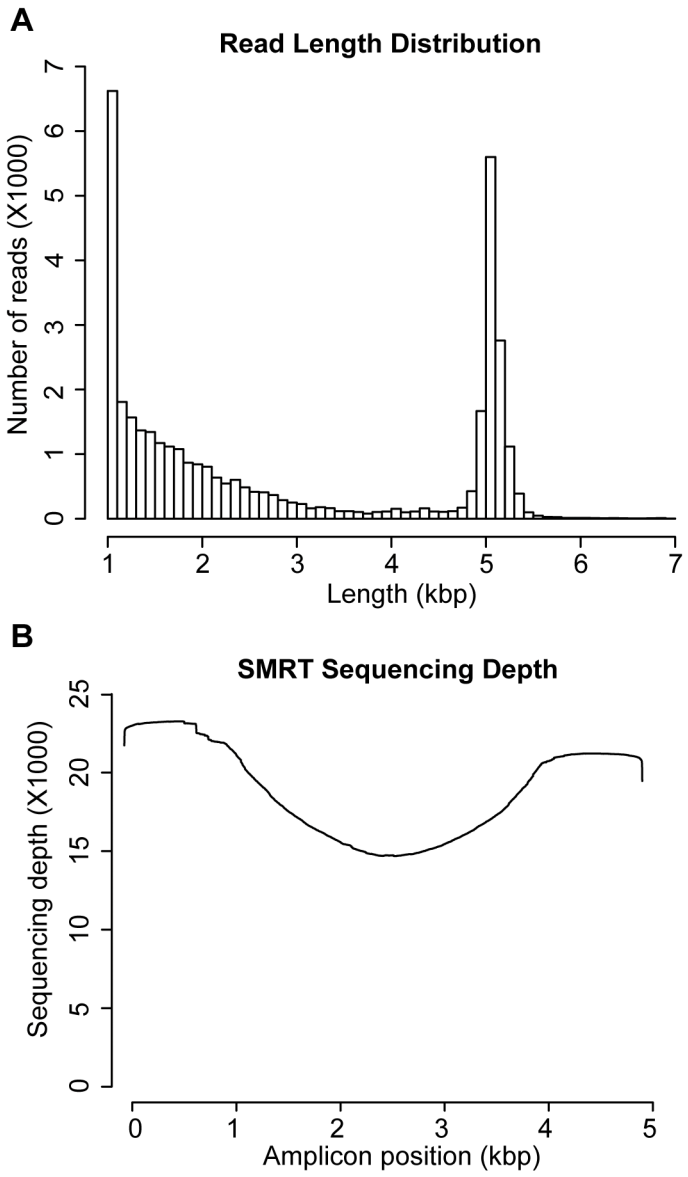

FIGURE 2. *CYP2D6* SMRT sequencing metrics

(A) The distribution of *CYP2D6* SMRT sequencing read lengths across all upstream and downstream amplicons that were subjected to variant calling. **(B)** The read depth for all analyzed nucleotides across the 5.0 kb *CYP2D6* region.

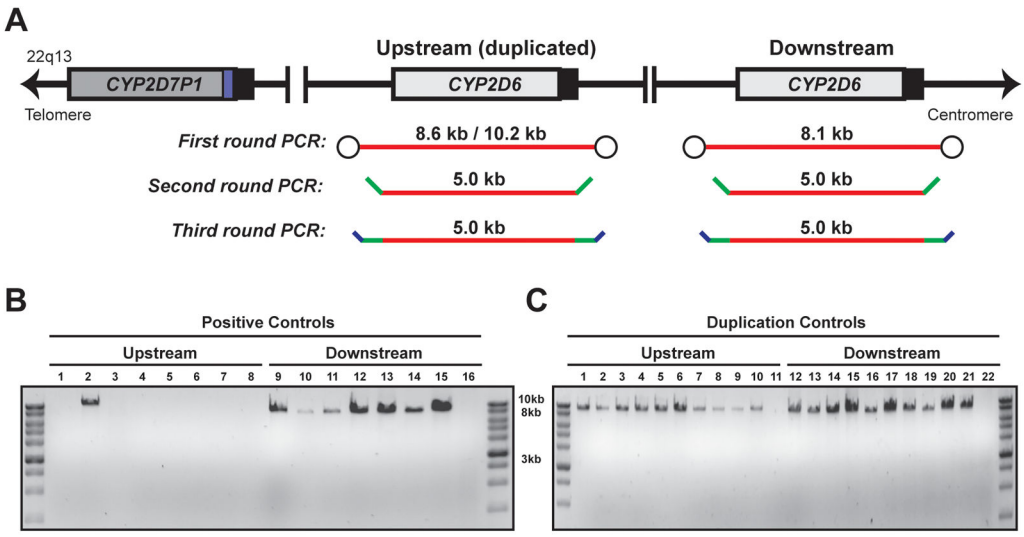

FIGURE 3. The *CYP2D6* gene and overview of amplicon preparation for SMRT sequencing

(A) Schematic illustration of the *CYP2D6* gene locus, including both upstream (duplicated) and downstream copies, and location of long-range PCR amplicons (red lines). Green and blue dashes represent universal oligonucleotide tags and barcodes, respectively. **(B)** Representative amplicons from the positive control and duplication control samples. First round 8.6 or 10.2 kb PCR products are on the top image and third round 5.0 kb PCR products are on the bottom image. Lanes 8 and 16 of the positive controls and lanes 11 and 22 of the duplication controls represent no template controls. Note the one positive control sample in lane 2 (NA166688) with an unexpected 10.2 kb upstream amplicon.

TABLE 1

Oligonucleotide primers for long range PCR amplification

Primer Sequence	T_a	Product Length
Upstream PCR		
5′-CCAGAAGGCTTTGCAGGCTTCAG-3′ [a]	63°C	8.6 kb or 10.2 kb [b]
5′-CGGCAGTGGTCAGCTAATGAC-3′ [a]		
Downstream PCR		
5′–CCAGAAGGCTTTGCAGGCTTCAG-3′ [a]	63°C	8.1 kb
5′-CAGGCATGAGCTAAGGCACCCAGA-3′ [a]		
Second Round Nested PCR [c]		
5′-[fwd tag]-GAACCTCTGGAGCAGCCCATACCC-3′	68°C	5.0 kb
5′-[rev tag]-ACTGAGCCCTGGGAGGTAGGTAG-3′ [a]		
Third Round Barcode PCR [d]		
5′-[barcode]-ATGGGTTCCAGAGTCAATCT-3′	68°C	5.0 kb
5′-[barcode]-GAAAGGTCTGGAGTCTTGAT-3′		
2850C>T AS-PCR		
FWD-C: 5′-AGCAGCTTCAATGATGAGAACC**G**GC-3′[e]	66°C	513 bp
FWD-T: 5′-AGCAGCTTCAATGATGAGAACC**G**GT-3′[e]		
REV: 5′-GATGCGGAAGCCCTGTACTT-3′		
3183G>A AS-PCR		
FWD: 5′-GGCAAGGTCCTACGCTTCCA-3′	66°C	717 bp
REV-G: 5′-CGCCGCACCTGCCCTAT**A**AC-3′[e]		
REV-A: 5′-CGCCGCACCTGCCCTAT**A**AT-3′[e]		

AS-PCR: allele-specific PCR; T_a: Annealing temperature

[a] Primer sequences from Gaedigk A, et al. 2007 (Gaedigk, et al., 2007).

[b] Upstream amplicons were 8.6 kb in length; however, *CYP2D6*36* upstream alleles amplified a 10.2 kb product due to exon 9 and downstream sequence conversion to *CYP2D7*.

[c] Forward universal tag sequence: ATGGGTTCCAGAGTCAATCT; reverse universal tag sequence: GAAAGGTCTGGAGTCTTGAT.

[d] All unique barcodes were 18 nucleotides in length.

[e] All 3′ allele-specific nucleotides are underlined, and artificially mutated nucleotides are underlined with bold font.

TABLE 2

Positive control samples tested with *CYP2D6* targeted genotyping, copy number analysis, and SMRT sequencing

	***CYP2D6* Diplotype**		**TaqMan Copy Number**		***CYP2D6* SMRT Sequencing**	
Samples	**Reported [a]**	**Luminex v3**	**Intron 2**	**Exon 9**	**Upstream/Downstream Copy**	**Diplotype**
NA12244	**35/*41*	**35/*41*	2	2	Downstream	**35/*41* [b]
NA16688	**2/*10*	**2/*10*	3	2	Upstream + Downstream	**2M/*36+*10B*
NA17222	**1/*2*	**1/*2*	2	2	Downstream	**2M/*108* [c]
NA17246	**4/*35*	**4/*35*	2	2	Downstream	**4/*35* [b,d]
NA17247	**1/*2*	**1/*2*	2	2	Downstream	**1A/*2M*
NA17280	**2/*3*	**2/*3*	2	2	Downstream	**3A/*59* [d,e],e
NA17296	**1/*9*	**1/*9*	2	2	Downstream	**1A/*9*
ASIAN048	-	**1/*29* [f]	2	2	Downstream	**1A/*107* [g]
HISP291	-	**1/*17*	2	2	Downstream	**1A/*17*
CAUC053	-	**1/*6*	2	2	Downstream	**1A/*6A*

Please note that all variant coordinates are summarized in Supp. Table S1 using both the M33388.1 reference sequence and current HGVS nomenclature with NM_000106.5.

[a] Based on consensus *CYP2D6* genotypes reported in Pratt VM, et al. 2010 (Pratt, et al., 2010).

[b] All identified **17*, **35* and **41* alleles have the *CYP2D7* gene conversion in intron 1.

[c] SMRT sequencing confirmed the common **2* alleles identified by Luminex genotyping (−1584C>G, 1661G>C, 2850C>T, and 4180G>C), but also detected the novel 3226A>G (p.H352R) and 3235A>G (p.Y355C) coding variants. Phasing by AS-PCR and ASXL-PCR indicated that both 3226A>G and 3235A>G were on the same haplotype as the 2850C allele, which has been named**108* by the Cytochrome P450 Allele Nomenclature Committee (Sim and Ingelman-Sundberg, 2010).

[d] The identified **4* alleles were not revised to suballeles due to the extent of benign suballele variants detected across **4* samples.

[e] SMRT sequencing identified heterozygous 1661G>C, 2291G>A, 2850C>T, 2939G>A and 4180G>C (in addition to the **3A* variants), which revised the **2/*3* diplotype to **3A/*59*.

[f] Luminex genotyping called a **1/*29* genotype due to heterozygous 1659G>A but without 1661G>C, 2850C>T, 3183G>A, or 4180G>C.

[g] SMRT sequencing identified heterozygous 1659G>A (p.V136M) and no other coding variants, which has been named *CYP2D6*107* by the Cytochrome P450 Allele Nomenclature Committee (Sim and Ingelman-Sundberg, 2010).

TABLE 3

Duplication samples tested with *CYP2D6* targeted genotyping, copy number analysis, and SMRT sequencing

	***CYP2D6* Diplotype**		**TaqMan Copy Number**		***CYP2D6* SMRT Sequencing**	
Samples	**Reported** [a]	**Luminex v3**	**Intron 2**	**Exon 9**	**Upstream/Downstream Copy**	**Diplotype**
NA17221	**1xN/*2*	**1/*2, DUP*	3	3	Upstream + Downstream	**1Ax2/*2M*
NA02016	**2xN/*17*	**2/*17, DUP*	3	3	Upstream + Downstream	**2Mx2/*17*
NA17298	**1/*1xN*	**1/*1, DUP*	3	3	Upstream + Downstream	**1A/*1Ax2*
NA07439	**4xN/*41*	**4/*41, DUP*	3	3	Upstream + Downstream	**4x2/*41* [b]
NA17232	**2/*2xN*	**2/*35, DUP*	3	3	Upstream + Downstream	**2Mx2/*35*
NA19137	**2x2/*17*	**2/*17, DUP*	3	3	Upstream + Downstream	**2Mx2/*17*
NA18924	**2/*4x2*	**2/*4, DUP*	3	3	Upstream + Downstream	**2M/*4x2* [b]
NA19152	**29/*43x2*	**1/*29, DUP*	3	3	Upstream + Downstream	**29/*43x2* [c]
NA19171	**2x2/*41*	**2/*41, DUP*	3	3	Upstream + Downstream	**2Mx2/*41* [d]
NA19175	**1/*4x3*	**1/*4, DUP*	4	4	Upstream + Downstream	**1A/*4x3* [b]
AA457	-	**1/*4, DUP*	4	4	Upstream + Downstream	**1A/*4x3* [b]
ASIAN089	-	**1/*2, DUP*	3	3	Upstream + Downstream	**1A/*2x2* [e]
ASIAN141	-	**2/*41, DUP*	3	3	Upstream + Downstream	**2Mx2/*41*
HISP432	-	**1/*2, DUP*	4	4	Upstream + Downstream	**1/*2Mx3*

Please note that all variant coordinates are summarized in Supp. Table S1 using both the M33388.1 reference sequence and current HGVS nomenclature with NM_000106.5.

[a] Based on consensus *CYP2D6* genotypes reported in Fang H, et al. 2014 (Fang, et al., 2014).

[b] The identified **4* alleles were not revised to suballeles due to the extent of benign suballele variants detected across **4* samples.

[c] SMRT sequencing confirmed the expected **29* allele and duplication, and also identified the **43* (77G>A) allele (not interrogated by Luminex) on both upstream and downstream copies.

[d] SMRT sequencing confirmed the expected **2M* and **41* alleles on the downstream copies but also identified the synonymous 958G>A variant in heterozygosity.

[e] Some identified **2* alleles were similar to **2M* but did not harbor 4481G>A.

TABLE 4

Previously identified discrepant and/or unclear *CYP2D6* samples tested with copy number analysis and SMRT sequencing

	***CYP2D6* Diplotype**		**TaqMan Copy Number**		***CYP2D6* SMRT Sequencing**	
Samples	**Reported [a]**	**Luminex v3**	**Intron 2**	**Exon 9**	**Upstream/Downstream Copy**	**Diplotype**
NA17289	**2/*4*	**2/*4*	2	2	Downstream	**2M/*4* [b]
NA17084	**1/*10*	**1/*10*	3	2	Upstream + Downstream	**1A/*36+*10B* [c]
NA17252	**4/*5*	**4/*5*	1	1	Downstream	**4/*5* [b]
NA17244	**2A/*4, DUP*	**2/*4, DUP*	4	4	Upstream + Downstream	**2Mx2/*4x2* [b]
NA17287	**1/*1(*36/?)*	**1/*1*	2	1	Downstream	**1A/*83* [d]
NA09301	*DUP*	**1/*2, DUP*	3	3	Upstream + Downstream	**1A/*2x2* [e]
NA17218	**2/*2(*35)*	**2/*35*	2	2	Downstream	**2M/*35*
NA17213	**1/*2(*35)*	**1/*35*	2	2	Downstream	**1A/*35*
NA17256	**2(*35)/*2(*35)*	**35/*35*	2	2	Downstream	**35/*35*
NA17243	**2(*35)/*4*	**4/*35*	2	2	Upstream + Downstream	**4/*35* [b,f]
NA17261	**2(*35)/*4*	**4/*35*	2	2	Downstream	**4/*35* [b]
NA17119	**1/*2*	**1/*2*	2	2	Downstream	**1A/*2M*
CAUC073	-	*?*	2	2	Downstream	**10B/*109* [g]
HISP418	-	*?, DEL*	2	1	Upstream + Downstream	**5/*36+*41* [h]

Please note that all variant coordinates are summarized in Supp. Table S1 using both the M33388.1 reference sequence and current HGVS nomenclature with NM_000106.5.

[a] Based on discrepant *CYP2D6* genotyping from different targeted platforms reported in Pratt VM, et al. 2010 (Pratt, et al., 2010).

[b] The identified **4* alleles were not revised to suballeles due to the extent of benign suballele variants detected across **4* samples.

[c] SMRT sequencing detected **1A/*10B* on the downstream copy but also identified the common and benign 2850C>T variant in heterozygosity.

[d] SMRT sequencing identified heterozygous 843T>G, exon 9 *CYP2D7* gene conversion and 4180G>C, which revised the **1/*1* diplotype to **1A/*83*.

[e] SMRT sequencing confirmed the **2* duplication, and also identified the synonymous 2452T>C variant on the upstream **2* copy.

[f] SMRT sequencing confirmed the **4/*35* diplotype; however, the upstream copy-specific primers amplified a product despite only two copies being detected by qPCR. The upstream copies had significant conversion to *CYP2D7* and, therefore, are not likely to encode a functional enzyme.

[g] Luminex genotyping did not infer a star (*) allele diplotype but called heterozygous variants at 100C>T, 1661G>C, 2613delAGA, 3183G>A and 4180G>C, which were confirmed by SMRT sequencing and suggested a **9/*10B* diplotype (but with 3183G>A). Phasing by AS-PCR and XLAS-PCR indicated that the 3183G>A (p.V338M) variant was on the same haplotype as the 2613delAGA (**9*) allele, which has been named **109* by the Cytochrome P450 Allele Nomenclature Committee (Sim and Ingelman-Sundberg, 2010).

[h] Luminex genotyping did not infer a star (*) allele genotype but called heterozygous variants at 100C>T and 1846G>A, mutant variants at 1661G>C, 2850C>T, 2988G>A and 4180G>C, and the *5 gene deletion allele. SMRT sequencing revised the diplotype to **5/*36+*41* but with the synonymous 4193T>C variant on the downstream **41* tandem allele.