A *De Novo* Algorithm for Allele Reconstruction from Oxford Nanopore Amplicon Reads, with Application to *CYP2D6*

Scott D. Brown[1], Lisa Dreolini[1], Agata Minor[2,3], Michelle Mozel[4], Nancy Wong[2], Sharon Mar[2], Amanda Lieu[2], Maimun Khan[2], Amanda Carlson[2], Monica Hrynchak[2,3], *Robert A. Holt[1,5,6], *Perseus I. Missirlis[4,7]

[1] Department of Basic and Translational Research; BC Cancer Research Institute; 675 W 10th Ave, Vancouver, BC, V5Z 1L3, Canada

[2] Department of Pathology and Laboratory Medicine; Molecular Cytogenetics Laboratory; Royal Columbian Hospital; 330 E Columbia St, New Westminster, BC, V3L 3W7, Canada

[3] Department of Pathology and Laboratory Medicine; University of British Columbia; G227-2211 Wesbrook Mall, Vancouver, BC, V6T 2B5, Canada

[4] Department of Anesthesiology and Perioperative Medicine; Royal Columbian Hospital; 330 E Columbia St, New Westminster, BC, V3L 0A2, Canada

[5] Department of Molecular Biology and Biochemistry; Simon Fraser University; SSB8166 – 8888 University Drive, Burnaby, BC, V5A 1S6, Canada

[6] Department of Medical Genetics; University of British Columbia; C201 – 4500 Oak Street, Vancouver, BC, V6H 3N1, Canada

[7] Department of Anesthesiology, Pharmacology & Therapeutics; University of British Columbia; 2775 Laurel Street, Vancouver, BC, V5Z 1M9, Canada

[*] Co-senior Authors - Robert A. Holt and Perseus I. Missirlis

Correspondence: sbrown@bcgsc.ca, rholt@bcgsc.ca or perseus.missirlis@fraserhealth.ca

Abstract

The Oxford Nanopore Technologies' sequencing platform offers a path towards bedside genomics, producing long reads that can completely cover a gene of interest, and thus detect any known or novel variant the gene contains. However, the analysis of these long reads to identify actionable genotypes remains challenging and typically requires customization depending on the target gene. Here, we describe a generic algorithm to accurately reconstruct allele sequences derived from long-reads of genomic-amplicon origin. Rather than calling variants directly from these long-reads, our method takes a "sequence-first" approach,

performing an unbiased reconstruction of the underlying amplicon sequences to generate high-confidence reconstructed allele sequences. This is done without user input of the expected target gene, allowing for any source amplicon to be reconstructed. These high-confidence reconstructed allele sequences are then compared to the genomic reference sequence of the gene to infer the specific diplotype present in the sample. This approach is agnostic towards the number of genes and alleles present and readily detects novel variants. We demonstrate our approach using three independent data sets for *CYP2D6*, a diverse and complex gene with over 175 known alleles of clinical significance affecting drug dosing. We show how our approach can accurately recover validated *CYP2D6* diplotypes from 20 Coriell samples sequenced using different primer sets, on different Oxford Nanopore Technologies flow cell versions, and to different depths. This includes inferring occurrences of copy number variation from relative abundances of each allele, a critical factor for ascribing functional effects to a diplotype.

Keywords: MinION, pharmacogenomics, CYP2D6, long-read sequencing, allele reconstruction, haplotyping

---

Background

Gene diversity can significantly impact the therapeutic effect of drugs in patients. One such example is the cytochrome P450 (CYP) superfamily of enzymes that play a key role in drug metabolism and can result in either supratherapeutic or subtherapeutic serum drug concentrations (McDonnell and Dang, 2013). Genotyping of genes prior to treatment can inform drug selection, dosing decisions, and avoidance of adverse side effects (Chang *et al.*, 2015). *CYP2D6* is one of the most clinically significant genes and *CYP2D6* genotype-guided dosing exists for a number of indications, including anti-fungal treatment in hematopoietic stem cell transplantation, breast cancer chemotherapeutics, tricyclic antidepressants, and opioids for pain management (Teusink *et al.*, 2016; Kiyotani *et al.*, 2008; Hicks *et al.*, 2017; Somogyi *et al.*, 2015). As such, a method for obtaining a fast and accurate readout of a patient's *CYP2D6* genotype is of great importance. *CYP2D6* is highly polymorphic; currently there are over 175 known functional alleles (each designated as a "star-allele", with *CYP2D6*1* being the reference), many of which differ from another allele by only a single base substitution (Whirl-Carrillo *et al.*, 2021, 2012), and new alleles are constantly being added as they are discovered.

Traditional methods to interrogate genotypes are generally slow and expensive; typically relying on custom sets of PCR reactions to interrogate specific sites along the length of each gene of interest. More contemporaneous approaches have leveraged high-throughput sequencing data, either whole genome (Chen *et al.*, 2021; Deserranno *et al.*, 2023), or amplicon-based (Hari *et al.*, 2023; Liau, Cree, *et al.*, 2019; Liau, Maggo, *et al.*, 2019; Qiao *et al.*, 2016; Ammar *et al.*, 2015; Charnaud *et al.*, 2022). These methods often make use of Pacific Biosciences or Oxford Nanopore Technologies (ONT) long reads, allowing the entire gene to be sequenced in a single read, and thus, facilitating accurate DNA sequencing and haplotyping with a single assay. Regardless of

sequencing strategy, these methods take a "variant-first" analysis approach, relying on variant calling from the sequence reads first, followed by phasing of the variants into haplotypes.

We sought to develop a method of diplotyping genes that uses a "sequence-first" rather than "variant-first" approach. We aimed to perform allele sequence reconstruction from long read sequences by generating high-confidence consensus sequences of each allele present in the original sample, then genotyping each resulting high-confidence consensus sequence to assign a haplotype. Further, we ensured our method was unbiased and generalizable, not being restricted to a specific gene. As such, our only input requirements are that the dataset is generated from genomic amplicon sequencing using ONT long reads and the dataset contains reads which completely cover the genomic regions of interest. No information is needed *a priori* regarding the specific genomic target, nor the number of targets. We demonstrate the utility of our approach using ONT sequencing data from *CYP2D6* amplicons and scripts that automate *CYP2D6* haplotyping of reconstructed allele sequences to call the most likely star-allele and identify any discrepant, potentially novel, variants.

Results

Selection of* CYP2D6 *amplicon.

The *CYP2D6* gene locus is 4.3 kb long and is located on the negative strand of chromosome 22 (chr22:42,126,499-42,130,810; hg38). Variants affecting CYP2D6 enzyme function may fall anywhere within this region, with the range of currently known variants (insertions, deletions, and single nucleotide variants) spanning chr22:42,126,578-42,130,783 (Whirl-Carrillo *et al.*, 2021, 2012). We used primers previously developed by Ammar et al. (Ammar *et al.*, 2015) to selectively amplify a 5.1 kb region of *CYP2D6* (chr22:42,126,072-42,131,142), and not the nearly identical *CYP2D7* pseudogene (**Figure 1**).

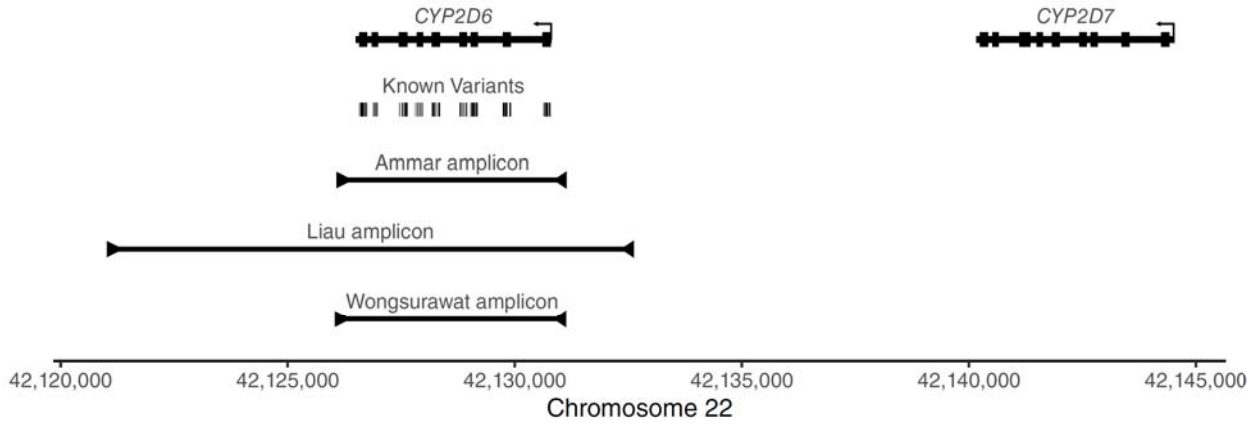


Figure 1. Graphical summary of the genomic region encompassing *CYP2D6*. The *CYP2D6* gene is shown, with larger boxes for the exons. The *CYP2D7* pseudogene is shown upstream of the *CYP2D6* transcriptional start site. All known variants of genetic relevance are shown as ticks. The three amplicon regions assessed in our work are shown as individual tracks.

Oxford Nanopore Technologies sequence generation.

Genomic DNA from three Coriell samples with known *CYP2D6* haplotypes (NA02016; **2xN/*17*, NA10005; **17/*29*, NA17300; **1/*6*) were used. Sequencing libraries were constructed using the ONT Ligation Sequencing Kit V14. We loaded each amplicon library onto a separate R10.4.1 Flongle flow cell, each being run for 24 h on a MinION Mk1b USB sequencer.

Base calling of the raw .fast5 files was performed using the GPU-enabled version of Guppy (a base caller provided by ONT). No quality filter was used at this stage, as quality filtering was performed at the per-base level later in our analysis pipeline. Omitting the quality filter allows for high-quality bases in lower-quality reads to still contribute information, maximizing the depth of coverage at all positions. Porechop was used to remove any adapter sequences and split chimeric reads. We obtained 87,906, 56,948, and 50,389 total reads from NA17300, NA10005, and NA02016 respectively. The average read length was 1,736 bases (± 2,078.35) (**Figure 2**).

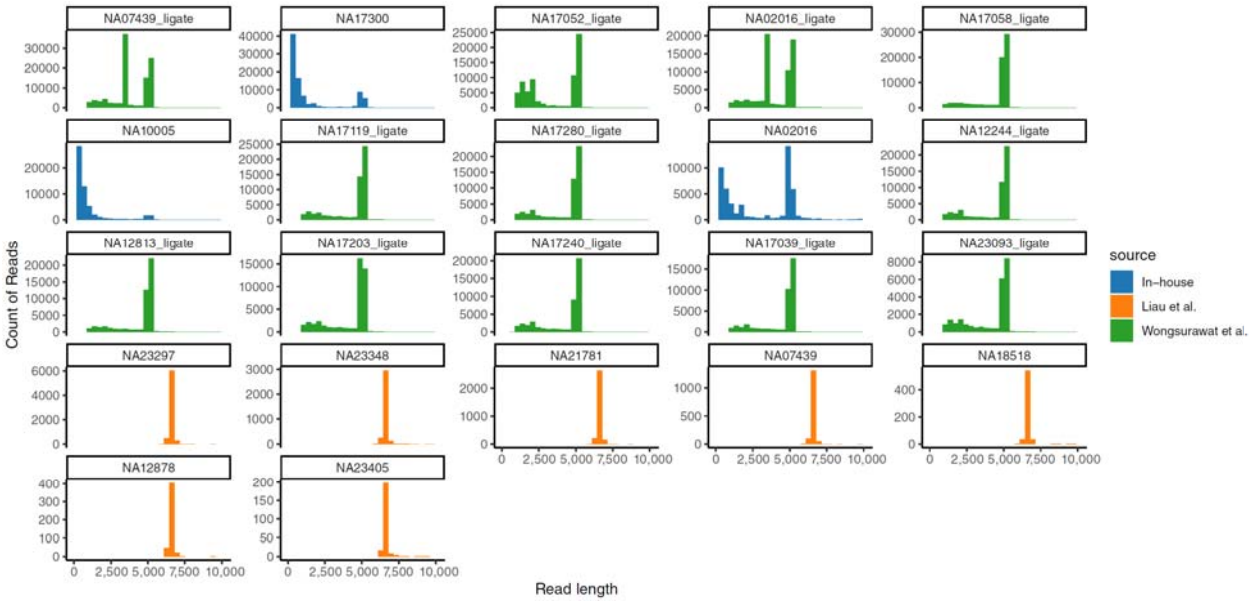


Figure 2. Distribution of read lengths for all samples. Samples ordered by decreasing numbers of total reads. In all cases, a population of reads at the expected 5.1 – 6.6 kb range is seen. In-house generated samples (blue) show higher than expected numbers of short reads, suggestive of degraded DNA amplicons during library preparation. The x-axis was truncated at 10,000 bp, removing an average of 0.37 % of reads per sample from display.

Inclusion of existing data sets to test different sequence qualities and amplicons.

To validate our approach on different ONT flow cell versions, *CYP2D6* amplicons, sequencing depths, and haplotypes, we obtained data from Liau et al. (Liau, Maggo, *et al.*, 2019) for seven additional Coriell samples. Liau et al. generated a 6.6 kb

amplicon for *CYP2D6* **(Figure 1)**, which covered more of the upstream and downstream regions of *CYP2D6*, and sequenced these on older MinION R9.4 flow cells, yielding an average of 2,292 reads per sample. This larger 6.6 kb amplicon should have a greater ability to detect gene duplication events, being more likely to span the junction site of a duplicated copy, though the exact location of this junction is unknown. Of note, Liau et al. also sequenced an even larger 8.6 kb amplicon for samples with confirmed duplications, covering the repetitive region downstream of *CYP2D6* (Gaedigk *et al.*, 2019), but this sequence data was unavailable.

Data for twelve additional Coriell samples were obtained from Wongsurawat et al. (Wongsurawat *et al.*, 2025), with an average of 55,068 reads per sample. Wongsurawat et al. generated a 5.1 kb amplicon using primers from Puaprasert et al. (Puaprasert *et al.*, 2018), covering a region similar to Ammar et al., and were multiplexed on MinION R9.4.1 flow cells. Wongsurawat et al. included two additional primer sets, one designed to identify gene deletions by producing a 3.5 kb amplicon only upon gene deletion (otherwise priming at sites 15 kb apart; too large to be efficiently amplified), and another designed to amplify a 3.5 kb region between duplicated copies of *CYP2D6* (with priming sites on either side of *CYP2D6* facing "outwards"). This data was generated using two different library construction kits; the Rapid Barcoding Kit and the Ligation Sequencing Kit. We only used Ligation Sequencing Kit data for our analysis. Combined with our in-house generated data, we have compiled 22 ONT sequence datasets covering 20 distinct Coriell samples.

Assigning reads to inferred amplicons based on whole genome alignment.

We designed our method to perform an unbiased reconstruction of allele sequences. We assumed that every amplicon present in the sequence library will have a subset of reads supporting that amplicon. As such, rather than supplying the algorithm with the anticipated amplicon, we used the sequence reads to infer the genomic loci that were amplified. To achieve this, we aligned all reads to the human reference genome (hg38) using minimap2, allowing for only the single best mapping position of each read, and assessed the depth of coverage across the whole genome. The regions with high coverage were taken to be representative of amplicons present in the source sequencing library (**Figure 3A**). Importantly, this approach does not assume any predefined number of amplicons, allowing this approach to work, theoretically, for multiplexed samples of any number of amplicons. All of our samples yielded an inferred amplicon region at the expected *CYP2D6* locus, and four samples yielded a second distinct inferred amplicon region (**Supplementary Table S1**). Two of these non-*CYP2D6* inferred amplicons (NA02016_ligate, NA07439_ligate) were from the region between *CYP2D6* and *CYP2D7*, corresponding to the duplication-detecting amplicon from Wongsurawat et al. The remaining two non-*CYP2D6* inferred amplicons were unexpected, one (NA17203_ligate) was from the *CYP2D7* locus and one (NA17052_ligate) was from the *SORCS2* locus on chromosome 4. Both of these cases were from the Wongsurawat et al. dataset, potentially suggesting that the primers or PCR conditions used for this dataset were more susceptible to off-target amplification. Indeed, careful inspection of the identified *SORCS2* locus sequence shows it contains a string of 15 contiguous bases at the start (GCCCTGGGAGGTAGG), and 13 contiguous bases at the end (GAAGCCTGCAAAG), which

exactly match sequence within the *CYP2D6* locus, and are present in the primers used by Wongsurawat et al. Also possible, for the *CYP2D7*-mapping amplicon from NA17203_ligate, is that these reads were derived from a *CYP2D6-CYP2D7* hybrid allele, where the majority of the amplicon sequence was derived from *CYP2D7*. Regardless, to remain unbiased, all inferred amplicons are carried forward in the analysis as possible loci of interest.

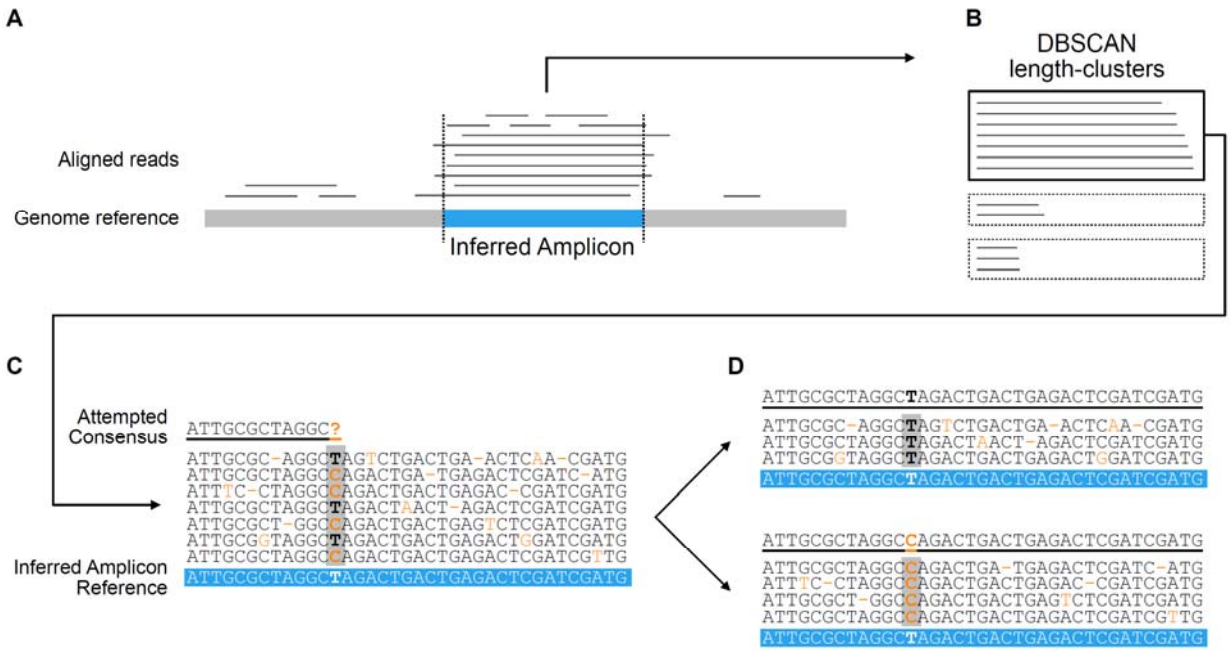


Figure 3. Schematic overview of algorithm. (A) Sequence reads are aligned to the genome, and regions of high depth are extracted and inferred as the source amplicon (blue). (B) These reads are clustered by length using the DBSCAN algorithm, and the most abundant cluster(s) and are selected for allele reconstruction (solid outlined box). (C) Selected reads are aligned to the inferred amplicon reference sequence (extracted from the genomic reference; blue), and each position of the alignment is assessed sequentially, building a consensus sequence base by base (top underlined sequence). Variants as compared to the reference are coloured orange. If a position is encountered as appearing heterozygous with an unknown consensus base, the reads are partitioned into two subsets (D). The reads in each subset are used, independently, to re-attempt allele reconstruction. In the example shown, this yields two consensus sequences (top underlined sequence), each with homozygous support for each base.

Reads clustered into groups based on length.

In our in-house generated data, we observed a higher than expected frequency of short reads (**Figure 2).** On average, 60.8 % of reads were under 1 kb, compared to 0 % for data from Liau et al. (only aligned reads were provided) and 0.08 % from Wongsurawat et al. We aligned our short reads to the *CYP2D6* reference sequence, demonstrating that they mapped to random positions spread across the amplicon (**Supplementary Figure S1**). This suggested the short reads were derived from fragmented amplicon during library construction, and were not representative of *bona fide* short amplicons within the sample. We assert that

full-length (non-fragmented) reads which were derived from the same amplicon would all have the same length (or highly similar lengths when allowing for sequence errors). To select the subset of reads for each inferred amplicon most likely to be informative, we performed length-clustering on the reads from each inferred amplicon using the DBSCAN algorithm (Schubert *et al.*, 2017) (**Figure 3B**). After clustering, we first selected the cluster that had reads with a mean length that was closest to the source inferred amplicon, representing reads that fully covered the amplicon. Secondly, we also assessed any highly abundant clusters, based on the premise that there may be other smaller or larger amplicons that would otherwise go undetected (**Supplementary Table S1**). All selected clusters were used for subsequent allele sequence reconstruction.

Allele sequence reconstruction performed through unbiased, recursive consensus sequence generation.

At this stage, we had obtained subsets of length-matched reads which were likely to have been derived from the same locus. These subsets, however, contained a mix of reads supporting each of the underlying haplotypes present in the original sample. If all sequence reads were error-free, this would have been a trivial exercise to partition reads based on their exact sequence. Due to sequencing errors, reads derived from the same haplotype did not share the exact same sequence. In order to reconstruct the underlying allele sequences, we took an empirical approach for determining which allele sequences the reads supported. This was achieved by finding partitions of the reads such that all reads in a partition supported the same sequence (while allowing for some noise due to sequence errors). First, we attempted to build a consensus sequence from all of the reads in the set, using similar logic to our previous work (Brown *et al.*, 2023). Briefly, reads were aligned to the inferred amplicon reference sequence, and each position of the alignment was assessed, sequentially, to call a consensus base at that position. If a position was reached that did not have sufficient evidence of a homozygous base (defined as the most-observed base being at least two times as frequent as the second-most-observed base; see Materials and Methods) (**Figure 3C**), the reads were partitioned into subsets based on the observed bases at this heterozygous position (**Figure 3D**). The consensus sequence generation was restarted from the beginning, recursively, on each of these resulting subsets. This process continued until a homozygous consensus sequence was successfully generated from a subset of reads, or the size of a subset fell below the lower limit threshold for reporting (5% of the total starting read set of length-clustered reads, selected empirically).

Our approach considered base quality in two ways. In order for a base to contribute to the consensus sequence, the base must have had a minimum quality of 10 (selected empirically). If this criterion was not met, a secondary criterion was tested, requiring a quality score above the 33rd percentile of observed qualities at that position (selected empirically). This ensured that positions with lower-quality bases across all reads were not completely lost from the consensus due to lack of coverage, while still ensuring that the consensus sequence was built using the best quality bases available at each position. Strand-specific errors are also common in ONT data (Krishnakumar *et al.*, 2018; Branton *et al.*, 2008), resulting in a specific sequencing error occurring primarily on one strand. When assessing all reads without considering the strand, such errors could appear as a heterozygous position. To avoid erroneously calling a position heterozygous, we assess, separately, the positive and negative

strand reads at each position, requiring support of heterozygosity from both strands independently to call a single consensus base. All reconstructed consensus sequences are available in **Supplementary File S1**.

Focusing on the results for read sets derived from the correct genomic locus (*CYP2D6*) and of the correct mean length (5.1 - 6.6 kb) for the 22 samples, all yielded at least two distinct consensus sequences (14 samples had exactly two, five samples had three, and three samples had four distinct consensus sequences). While having two sequences is expected for a diploid genome, having greater than two consensus sequences could support gene duplication events where the copies have sequence differences. This could also, however, be from a common sequencing error that is abundant enough to result in a consensus sequence. Relative read depths may help distinguish these two cases, as described later.

Two samples had length-clusters that were shorter than the expected amplicon size, but were abundant enough to have allele sequence reconstruction attempted (**Supplementary Table S1**). One yielded two consensus sequences mapping to exons 8-9 of *CYP2D6* (NA10005), and one yielded a consensus sequence mapping to exons 4-5 of *CYP2D6* (NA17300). This supports the notion that these short reads were derived from degraded DNA template occurring during library construction. Indeed, when inspecting these short consensus sequences, they match the respective regions of the long consensus sequences from the same sample (**Supplementary Figure S2**). Two samples had allele sequence reconstruction attempted for reads that did not originate from the *CYP2D6* locus or supplementary amplicons. As described above, these are most likely derived from off-target amplification from the genome during PCR, but could also be the result of *CYP2D6-CYP2D7* hybrid alleles (in the case of the *CYP2D7*-derived consensus sequence in NA17203_ligate). The *SORCS2*-derived consensus sequences were run through BLASTn (Sayers *et al.*, 2025), and both mapped back to the *SORCS2* locus, supporting that these reads were from off-target amplification.

Automated reporting of most likely* CYP2D6 *star alleles and putative novel variants.

To identify the specific *CYP2D6* alleles that each resulting consensus sequence supported, we first aligned each consensus sequence to the *CYP2D6*1* reference sequence. All variants (indels and mismatches) were enumerated, and those occurring within coding regions or splice sites were used for *CYP2D6* allele determination. We automated this process by using the *CYP2D6* Allele Definition Table from PharmGKB (Whirl-Carrillo *et al.*, 2021, 2012). Briefly, for each known star allele, all canonical variants were parsed from the PharmGKB definition table. For alleles with variants that have multiple possible bases allowed, some of which may include the reference wildtype base, all possible allowed bases are stored. Then, the set of identified variants for the consensus sequence was compared to the set of all required or possible variants from all known alleles, and each known star allele was given a score by summing the number of disagreements in the identified variants from the consensus sequence. Thus, a consensus sequence that had exactly the same variants as a defined star-allele would receive a score of zero for that allele. The allele with the lowest score was returned as the best candidate allele assignment for the consensus sequence. In the event of an imperfect match, missing and/or extra variants in the consensus sequence compared to the candidate star allele

reference were also returned, allowing for manual review of the candidate allele assignment (**Table 1**). In this way, novel variants and alleles can also be readily identified, rather than simply the best match to known alleles.

Inferring copy number from reconstructed alleles.

Copy number of genes can be an important factor for determining functional effects. In our results, the relative copy number of the reconstructed alleles could be inferred based on the number of reads that contributed to each allele. In most cases, we obtained a single consensus sequence per *CYP2D6* allele. In these cases, we simply compared the number of reads for each allele. In cases where multiple distinct consensus sequences were obtained for the same *CYP2D6* allele, we pooled the read counts for all sequences of the same allele. The presence of multiple distinct consensus sequences for the same allele could be used as evidence of duplication events; however, the sequences should be manually inspected to determine if the differences are believed to be from true underlying sequence differences or from regions of poor sequence quality such as those near the ends of the amplicon. Importantly, the lack of multiple distinct sequences does not preclude the presence of a duplication, as each duplicated copy may indeed have an identical underlying sequence. For these reasons, we propose the read depths as a more reliable estimate of relative copy number. To interpret the allele read depth ratios, we tested for ratios > 2:1. In our results, the exact number of duplication events could not be reliably determined, likely due to biases in different allele sequences and sequencing errors, especially in samples with lower read counts. Therefore, the alleles with a greater than 2:1 ratio were assigned an *xN* duplication.

Known diplotypes are recapitulated from Coriell samples.

Across the 22 distinct Coriell samples analyzed, our algorithm successfully reconstructed *CYP2D6* diplotypes that matched the reported GeT-RM consensus in 18 cases (81.8 %; **Table 1**) (Pratt *et al.*, 2010, 2016; Gaedigk *et al.*, 2019). This high level of concordance included 12 samples (54.5 %) containing either synonymous substitutions (C6015G, C6057T) or known homopolymer-induced false-positive deletions (7598delC, 7599delC; **Figure 4C**), which are disregarded during interpretation. This also included six samples (27.2 %) where allele copy number carried inherent uncertainty in our calls or the GeT-RM reference (NA02016, NA02016_ligate, NA07439, NA07439_ligate, NA23297, NA23348). These cases with copy number variation included the two Coriell samples from Wongsurawat et al., where the consensus sequence we generated from their duplication-detecting amplicon supports the presence of a duplicated allele (**Supplementary Figure S3**). Another case (NA10005) presented with a putative uncharacterized variant within a **29* allele; however, this likely represents mispartitioned **17* reads rather than a true variant, given the low coverage of the consensus sequence and absence in the paired **29* haplotype (**Figure 4A-B**). In one sample (NA17300) our reconstruction identified an additional uncharacterized variant (G6996A) in a **6* allele **(Figure 4D**). This variant occurs downstream of the premature stop codon caused by the 6727delT frameshift mutation, and thus would not have further functional effect. One more case (NA17203_ligate) showed a third low-abundance consensus with imperfect match to **4*, which was interpreted as **4*. This case also yielded a consensus from *CYP2D7*-derived reads.

Recently, NA17203 has been reported as harboring a *CYP2D6-CYP2D7* hybrid allele (**68*) (Wongsurawat *et al.*, 2025; Charnaud *et al.*, 2022), though this has not yet been reported by GeT-RM. Manually assessing the alignment of the NA17203_ligate *CYP2D7*-derived consensus sequence to the *CYP2D6* and *CYP2D7* reference sequences supports that this sequence may be representative of the **68 CYP2D6-CYP2D7* hybrid allele, with a higher number of variants detected from exon 2 – 9 of *CYP2D6* compared to *CYP2D7* (**Supplementary Figure S4**). Indeed, comparing our consensus sequence to the **68* sequence obtained by Rubben et al. (Rubben *et al.*, 2022) from NA12878, we see only a single indel difference, occurring between exons 3 and 4 (**Supplementary Figure S5**), and thus our consensus sequence supports the same functional hybrid **68* allele.

Of the four samples classified as inconsistent with GeT-RM diplotypes, two (NA17058_ligate, NA17280_ligate) yielded reconstructions strongly suggestive of an error in the reported GeT-RM calls, which were originally based on targeted genotyping panels rather than full-length sequencing, and are known to be imperfect (van der Maas *et al.*, 2025) (**Figure 5**). These inconsistent calls are also supported by recent reports (Charnaud *et al.*, 2022; Wongsurawat *et al.*, 2025). The remaining two cases (NA12878 and NA21781) involved the **68 CYP2D6-CYP2D7* hybrid allele, which was not detected by the Liau et al. primer design. While no hybrid allele reads were captured in this dataset, the underlying constituent haplotypes were reconstructed correctly, with only the **68* classification absent. Taken together, these results demonstrate that the algorithm is robust and largely concordant with reference diplotypes, with nearly all discrepancies attributable to synonymous changes, sequencing artefacts, or structure variants beyond the scope of the amplicon design.

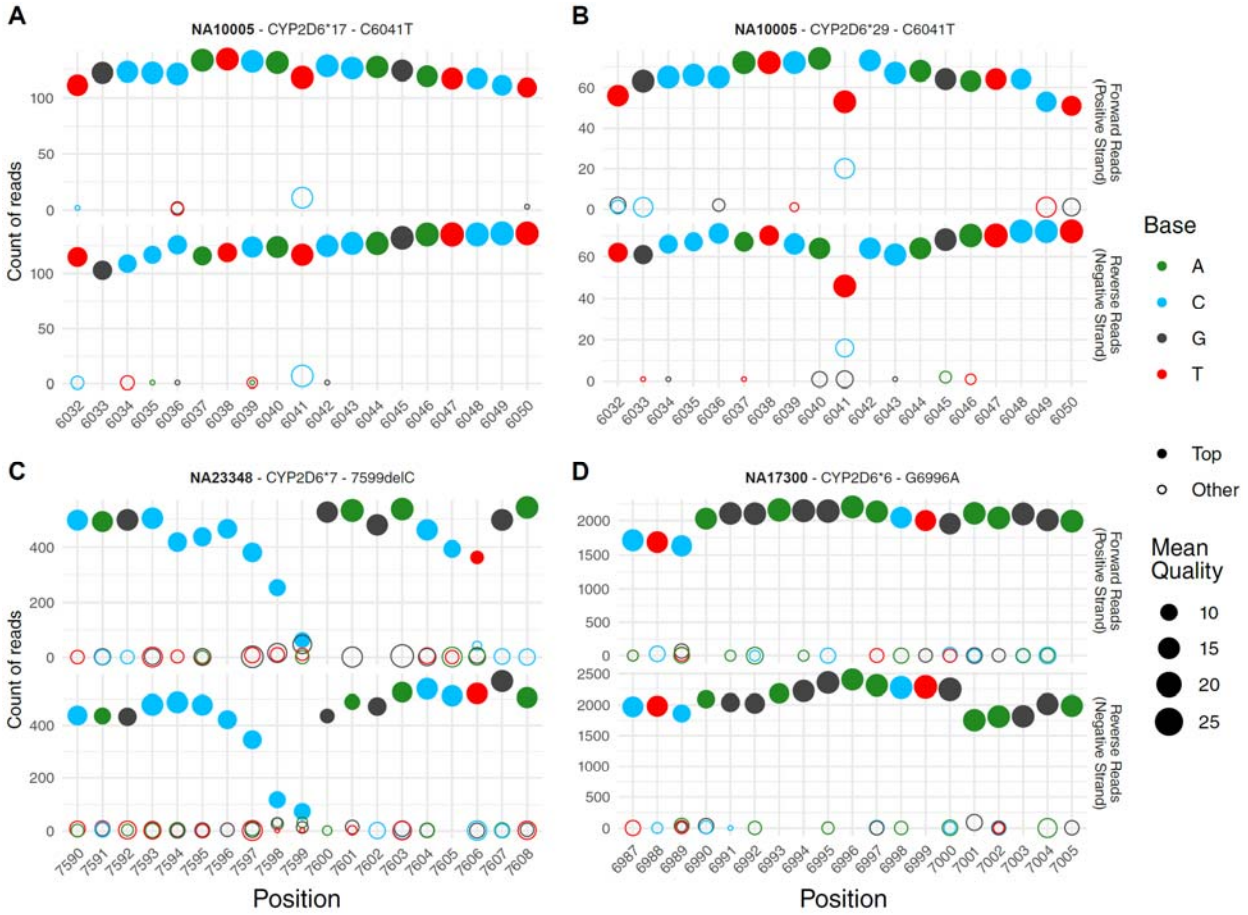


Figure 4. Example signal-to-noise plots for different identified variants. In each panel, the top-observed base at each position is displayed as a solid point, and all other detected bases are hollow points. The size of the point denotes the average quality of the base from all reads used to generate the consensus. The top graph shows results from forward reads, and the bottom graph from reverse reads. (A) The expected C6041T variant from NA10005 *CYP2D6*17* is shown, and compared to the potentially erroneous call from the *CYP2D6*29* allele (B). The poorer signal-to-noise ratio in *CYP2D6*29* suggests that at least some of these reads are likely derived from *CYP2D6*17*, but were incorrectly partitioned into this read group. (C) An example of the apparent decrease in coverage in homopolymer regions, shown here with a deletion 7599delC being called in NA23348. (D) The novel G6996A variant from NA17300 *CYP2D6*6* allele is shown, with no evidence of this being an erroneous call due to poor signal-to-noise or poor quality. This variant occurs downstream of the premature stop codon due to the 6727delT mutation in *CYP2D6*6,* and thus will not have any effect on the resulting protein.

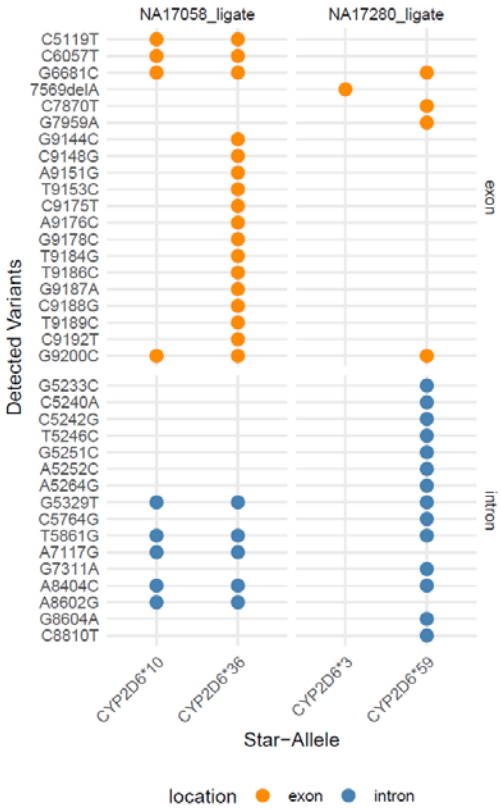


Figure 5. Landscape of variants identified in reconstructed alleles for NA17058_ligate and NA17280_ligate. In each sample, two alleles were reconstructed. Exonic variants are displayed as orange bubbles, intronic variants as blue bubbles. In both samples, the supported alleles differ by a large number of variants, and do not have significant overlap with each other.

Table 1. Summary of *CYP2D6* allele reconstruction and diplotyping.

Sample ID	Source	Flow cell	Total Reads	Exon and Splice Variants in Consensus Compared to *CYP2D6*1*	Star Allele Candidate	Amplicon Depth	Allele Depth	Copy Number Ratio	Genotype Call	GeT-RM Genotype
NA02016	In-house	Flongle R10.4.1	50,389	G6681C, C7870T, G9200C	**2*	3,793	8,466	> 3:1	**2xN/*17*	**2xN/*17* (Pratt *et al.*, 2010)
				G6681C, C7870T, G9200C	**2*	2,873				
				G6681C, C7870T, G9200C	**2*	1,800				
				C6041T, G6681C, C7870T, G9200C	**17*	2,322	2,322			
NA02016_ligate	Wongsurawat et al. 2025	MinION R9.4.1	66,015	G6681C, C7870T, G9200C	**2*	8,697	17,576	> 4:1	**2xN/*17*	**2xN/*17* (Pratt *et al.*, 2010)
				G6681C, C7870T, G9200C	**2*	5,551				
				G6681C, C7870T, G9200C	**2*	3,328				
				C6041T ,G6681C, C7870T, G9200C	**17*	3,792	3,792			
NA07439	Liau et al. 2019	MinION R9.4	1,467	C5119T, C6057T[‡], G6681C, G6866A, 7599delC[†], G9200C	**4*	642	642	> 3:1	**4xN/*41*	**4x2/*41* (Gaedigk *et al.*, 2019)
				G6681C, 7598delC[†], 7599delC[†], C7870T, G8008A, G9200C	**41*	175	175			
NA07439_ligate	Wongsurawat et al. 2025	MinION R9.4.1	102,224	C5119T, C6057T[‡], G6681C, G6866A, G9200C	**4*	23,238	23,238	> 3:1	**4xN/*41*	**4x2/*41* (Gaedigk *et al.*, 2019)
				G6681C, C7870T, G8008A, G9200C	**41*	6,284	6,284			
NA10005	In-house	Flongle R10.4.1	56,948	C6041T, G6681C, C7870T, G9200C	**17*	288	288	~ 1:1	**17/*29*	**17/*29* (Pratt *et al.*, 2010)
				G6679A, G6681C, C7870T, G8203A, G9200C	**29*	277	429			
				C6041T, G6679A, G6681C, C7870T, G8203A, G9200C	**29* **(+C6041T)**	152				
NA12244_ligate	Wongsurawat et al. 2025	MinION R9.4.1	49,461	G5050A, G6681C, C7870T, G9200C	**35*	16,047	16,047	~ 1:1	**35/*41*	**35/*41* (Pratt *et al.*, 2010)
				G6681C, C7870T, G8008A, G9200C	**41*	12,450	12,450			
NA12813_ligate	Wongsurawat et al. 2025	MinION R9.4.1	40,070	C5119T, C5992A, A6002G, C6015G[‡], G6681C, G6866A, G9200C	**4*	7,972	7,972	~ 1:1	**2/*4*	**2/*4* (Pratt *et al.*, 2016; Gaedigk *et al.*, 2019)
				G6681C, C7870T, G9200C	**2*	8,926	8,926			
NA12878	Liau et al. 2019	MinION R9.4	486	7569delA, 7599delC[†]	**3*	168	168	~ 1:1	**3/***4***	**3/(*68)+*4* (Gaedigk *et al.*, 2019)
				C5119T, C5992A, A6002G, C6015G[‡], G6681C, G6866A, 7599delC[†], G9200C	**4*	120	120			
NA17039_ligate	Wongsurawat et al. 2025	MinION R9.4.1	38,634	G6681C, C7870T, G9200C	**2*	5,962	9,262	~ 1:1	**2/*17*	**2/*17* (Pratt *et al.*, 2010)
				G6681C, C7870T, G9200C	**2*	3,300				
				C6041T, G6681C, C7870T, G9200C	**17*	3,574	10,002			
				C6041T, G6681C, C7870T, G9200C	**17*	6,428				
NA17052_ligate	Wongsurawat et al. 2025	MinION R9.4.1	70,348	None.	**1*	15,461	29,915	~ 1:1	**1/*1*	**1/*1* (Pratt *et al.*, 2010)
				None.	**1*	14,454				
NA17058_ligate	Wongsurawat et al. 2025	MinION R9.4.1	66,012	C5119T, C6057T[‡], G6681C, G9200C	**10*	18,988	18,988	~ 1:1	**10/***36***	**10/*10* (Pratt *et al.*, 2010)
				C5119T, C6057T[‡], G6681C, G9144C, C9148G, A9151G, T9153C, C9175T, A9176C, G9178C, T9184G, T9186C, G9187A, C9188G, T9189C, C9192T, G9200C	**36*	18,465	18,465			
NA17119_ligate	Wongsurawat et al. 2025	MinION R9.4.1	55,696	None.	**1*	9,120	9,120	~ 1:1	**1/*2*	**1/*2* (Pratt *et al.*, 2010)
				G6681C, C7870T, G9200C	**2*	6,958	9,105			
				C7870T, G9200C	**2*	2,147				

Sample ID	Source	Flow cell	Total Reads	Exon and Splice Variants in Consensus Compared to *CYP2D6*1*	Star Allele Candidate	Amplicon Depth	Allele Depth	Copy Number Ratio	Genotype Call	GeT-RM Genotype
NA17203_ligate	Wongsurawat et al. 2025	MinION R9.4.1	45,220	C5119T, C5992A, A6002G, C6015G‡, G6681C, G6866A, G9200C	**4*	5,209	6,802	~ 1:1:1	**4/*35/***68***	**4/*35* (Pratt *et al.*, 2010)
				C5119T, C5992A, A6002G, C6015G‡, G6681C, C7870T, G9200C	**4* **(-G6866A)** *or *160* **(+C5119T)**	1,593				
				G5050A, G6681C, C7870T, G9200C	**35*	5,466	5,466			
				CYP2D6-CYP2D7 hybrid	**68*	3,551	3,551			
NA17240_ligate	Wongsurawat et al. 2025	MinION R9.4.1	44,930	None.	**1*	6,507	6,507	~ 1:1	**1/*10*	**1/*10* (Pratt *et al.*, 2010)
				C5119T, C6057T‡, G6681C, G9200C	**10*	5,383	5,383			
NA17280_ligate	Wongsurawat et al. 2025	MinION R9.4.1	51,996	7569delA	**3*	14,469	14,469	> 2:1	****3xN/*59***	**2/*3* (Pratt *et al.*, 2010)
				G6681C, C7870T, G7959A, G9200C	**59*	6,453	6,453			
NA17300	In-house	Flongle R10.4.1	87,906	None.	**1*	4,916	4,916	~ 1:1	**1/*6*	**1/*6* (Pratt *et al.*, 2010)
				6727delT, G6996A	**6* **(+G6996A)**	4,608	4,608			
NA18518	Liau et al. 2019	MinION R9.4	620	C6041T, G6681C, 7599delC†, C7870T, G9200C	**17*	164	164	~ 1:1	**17/*29*	**17/*29* (Pratt *et al.*, 2016; Gaedigk *et al.*, 2019)
				T5902C‡, G6679A, G6681C, 7599delC†, C7870T, G8203A, G9200C	**29*	65	65			
NA21781	Liau et al. 2019	MinION R9.4	3,018	G6681C, 7599delC†, C7870T, G9200C	**2*	569	1,078	> 3:1	**2xN/***4***	**2x2/(*68)+*4* (Gaedigk *et al.*, 2019)
				T5902C‡, G6681C, 7599delC†, C7870T, G9200C	**2*	509				
				C5119T, C5992A, A6002G, C6015G‡, G6681C, G6866A, 7599delC, G9200C	**4*	283	283			
NA23093_ligate	Wongsurawat et al. 2025	MinION R9.4.1	23,207	C5119T, C6057T‡, G6681C, G9144C, C9148G, A9151G, T9153C, C9175T, A9176C, G9178C, T9184G, T9186C, G9187A, C9188G, T9189C, C9192T, G9200C	**36*	3,187	3,187	~ 1:1:1	**1/*10/*36*	**1/*36+*10* (Gaedigk *et al.*, 2019)
				C5119T, C6057T‡, G6681C, G9200C	**10*	2,048	2,048			
				None.	**1*	2,103	2,103			
NA23297	Liau et al. 2019	MinION R9.4	6,833	C5119T, C6057T‡, G6681C, 7599delC†, G9200C	**10*	2,960	2,960	> 3:1	**10xN/*17*	**10x2/*17* (Gaedigk *et al.*, 2019)
				C6041T, G6681C, 7599delC†, C7870T, G9200C	**17*	858	858			
NA23348	Liau et al. 2019	MinION R9.4	3,384	7599delC†, A7955C	**7*	1,251	1,251	> 5:1	**7xN/*35*	**7/*35* (Pratt *et al.*, 2016; Gaedigk *et al.*, 2019)
				G5050A ,G6681C, 7599delC†, C7870T, G9200C	**35*	212	212			
NA23405	Liau et al. 2019	MinION R9.4	235	7599delC†	**1*	80	80	~ 1:1	**1/*7*	**1/*7* (Pratt *et al.*, 2016; Gaedigk *et al.*, 2019)
				7599delC†, A7955C	**7*	49	49			

†Deletion occurring in a homopolymer base stretch, a known common sequencing error, which is ignored. ‡Synonymous variant. Discrepancies bolded and underlined.

Discussion

We have presented our solution to reconstruct allele sequences from ONT-based amplicon sequencing, demonstrating its effectiveness on three distinct datasets encompassing different amplicons and flow cells. Our approach is unbiased, only requiring that a subset of reads exist which fully cover the amplicon of interest, but not requiring *a priori* knowledge of what the specific amplicon is. The appeal of this approach in diverse genes such as *CYP2D6* is that potential duplications and structural variants may be reconstructed without pre-existing knowledge of the sequence. For our application of this approach to the *CYP2D6* locus, we provide post-hoc analysis scripts to automate the variant detection and star-allele assignment, facilitating rapid interpretation of results. Our approach makes no assumptions regarding the number of gene amplicons (or alleles thereof) present in the data, only requiring sufficient coverage of each amplicon in order to be reconstructed and reported. Since the primary output of our approach is the reconstructed allele sequence, novel variants and alleles are readily detectable. In addition to the reconstructed allele sequence and assigned star-allele, our pipeline also outputs a file with per-position base frequencies and average base qualities, allowing interrogation of specific positions to assess the confidence of the resulting base. This gives the user more information, allowing manual interpretation of putative novel variants (**Figure 4**), much like is possible with traditional Sanger sequencing. Within the context of tamoxifen metabolism by CYP2D6, a recent study demonstrated that a neural-network model given full-length gene sequences was better able to predict metabolism rates compared to rates predicted from the assigned star-alleles alone (Van der Lee *et al.*, 2021). This demonstrates the increased power of allele sequence reconstruction, and suggests that such reconstructions may be even more important in the future as more machine learning models are trained across a wider range of genes and pathways.

We observed a higher than expected number of short reads in our data. In assessing these reads after mapping to the genome, they appear to be spread across the whole expected *CYP2D6* amplicon region, and are not coming from any single contaminant (**Supplementary Figure S1**). Such a signal is consistent with the concept of DNA amplicon fragmentation during sequence library preparation, resulting in these DNA fragments having adapters ligated and subsequently sequenced. We note that Liau et al. (Liau, Maggo, *et al.*, 2019) also observed a large fraction of short reads in their data (> 90 % of their reads were shorter than the expected ~6 kb amplicon in their first sequencing run, ~45 % in their second run), so this may be an expected artefact of sequence library preparation which would require handling in the downstream analysis pipeline. We opted to only consider short reads if they were evidenced to have been derived from a single conserved, short amplicon.

Existing methods for diplotyping *CYP2D6* from amplicon or whole genome sequence data take a “variant-first” approach (Hari *et al.*, 2023; Chen *et al.*, 2021; Deserranno *et al.*, 2023; Liau, Maggo, *et al.*, 2019; Ammar *et al.*, 2015), using algorithms such as Clair3 (Zheng *et al.*, 2022) to call variants, and then these variants are subsequently phased into haplotypes using tools like WhatsHap (Patterson *et al.*, 2015). Our method differs from existing methods in that it is not reliant on variant

calling directly from sequence reads. Rather, we distil the set of sequence reads into high-confidence, reconstructed allele sequences, which can be directly compared to representative reference allele sequences for haplotype determination. This is especially helpful when analyzing ONT reads, which have higher rates of sequencing errors than NGS approaches. Our allele sequence reconstruction approach is not reliant on a specific number or identity of amplicons within the data, theoretically handling even extreme cases of copy number variation or gene duplication. Our method uses commonly available Linux tools and open-source software, defined in Anaconda environments and should be readily deployable on any Linux system.

Our approach requires amplicon data, and specifically requires reads to be uniform in their coverage of the amplicon. As such, the ONT Rapid (transposase-based) library construction kits are incompatible, as are "shotgun" based approaches, as they result in read sets with variable start and end positions. Further, with data from a single amplicon, there is no way to unambiguously assess copy number. Based on the number of reads contributing to each reconstructed allele, we can approximate the copy number ratio if a sample has two distinct allele sequences. We note that this is a relative abundance measure rather than an absolute measure of copy number. Despite this, relative abundance is likely sufficient to ascertain clinical effect, as all *CYP2D6* alleles capable of ultrarapid metabolism do so by duplicating at least once (Hicks *et al.*, 2017). Thus, determining the relative abundance of normal functioning alleles may be sufficient to designate an ultrarapid metabolizer. Further copy number increases will elevate the "Activity Score" of the diplotype, but will not change the metabolizer status, nor drug prescription recommendations. On the other hand, since there is no genomic reference to accurately estimate relative copy number, if certain diplotypes consist of duplicated homozygous normal function alleles (such as **1/*1x2*, **2/*2x2*, **33/*33x2*, etc.), it is impossible to determine the copy number and thus the metabolizer phenotype. Multiplexing of multiple genes may help, assuming that one locus is copy number invariant. Alternatively, a control region of the genome could be included that is known to be copy number invariant, as is done for data used by Aldy 4 (Hari *et al.*, 2023). The data from Wongsurawat et al. (Wongsurawat *et al.*, 2025) demonstrated that our analysis is not hindered by the presence of multiple amplicons, supporting this as a reasonable approach in the future.

For the three datasets examined in this manuscript, none contained any detectable population of reads that contained full-length gene duplication events, likely due to the standard *CYP2D6* primers resulting in too large of an amplicon to be efficiently amplified in the event of a gene duplication. Wongsurawat et al. included specific primers targeting the region between duplicated gene copies, which produced a readily identifiable 3.5 kb amplicon in their two samples with known gene duplications. Two samples from Liau et al. with known *CYP2D6*68* alleles (a gene hybrid with *CYP2D7*) were both called incorrectly, with no reads showing the expected fusion event. Alternative primer sets would be worth exploring in the future to increase the chances of capturing these structural changes.

We have demonstrated that *CYP2D6* allele sequence reconstruction and diplotyping can be achieved from modest long-read sequencing (ONT Flongles), requiring limited specialized equipment, and having low cost. It is appealing that a singular

sequencing reaction coupled with this algorithm may be able to ascertain *CYP2D6* diplotypes with accurate phasing of known or novel variants, as well as detect gene duplications and structural variants without the need for multiple experiments. This information can then be used to designate metabolizer status in order to facilitate rational drug prescription as per CPIC guidelines (Hicks *et al.*, 2017). Currently, to achieve this clinically, a laborious series of assays are conducted, such as Taqman real-time PCR reactions for common SNPs and indels, copy number verification by qPCR and finally long-range PCR is used for confirmation of duplications or deletions (Pratt *et al.*, 2021). These assays must have regulatory approval, being performed in a qualified lab using approved medical devices. Regulatory approval of our approach is possible, particularly by transitioning to the GMP-ready Q-Line of ONT devices. This would necessitate, at the expense of more complexity, barcoding and multiplexing of samples on larger MinION flow cells to increase throughput.

Materials and Methods

CYP2D6 *gene amplification.* Genomic DNA was quantified using the Nanodrop, and diluted to 200 ng of DNA in 33.5 uL of nuclease-free water. Primer sequences were obtained from Ammar et al. (Ammar *et al.*, 2015), using CYP2D6-2F (TAGCTCCCTGACGCCATGATTTGTCTT) and CYP2D6-2R (CCTGGTTATCCCAGAAGGCTTTGCAG). We used the standard protocol for Phusion High-Fidelity DNA Polymerase (NEB M0530): 5× Phusion HF Buffer, 10 mM dNTPs, 10 uM forward primer, 10 uM reverse primer, 200 ng template genomic DNA, and 0.5 uL Phusion DNA polymerase in a 50 uL reaction. PCR cycling conditions were: 98°C for 30 sec., followed by 35 cycles of 98°C for 10 sec. and 72°C for 120 sec., followed by a final step of 72°C for 10 min. and hold at 4°C. PCR products were cleaned up using the Qiagen QIAquick PCR Purification Kit following manufacturer's instructions.

Sequence library construction. Libraries were prepared using the ONT Ligation Sequencing Kit V14 (SQK-LSK114) following the provided protocol (LSK114_ACDE_9163_v114_revN_29Jun2022). Samples were quantified with Qubit dsDNA HS Assay kit (Invitrogen Q32851). One hundred fmols of amplicon DNA was repaired using NEBNext Ultra II End Prep Enzyme Mix provided in the NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing kit (NEB E7180S). Samples were purified with 30 uL of Ampure XP beads (Beckman, A63880), and eluted with 31 µL of nuclease-free water. Adapter ligation was performed with the NEBNext Quick T4 DNA Ligase provided in the NEBNext Companion Module, ONT Ligation Buffer (LNB) and ONT adapter mix (LA). Samples were purified with 20 uL Ampure XP beads, washed with ONT Long Fragment Buffer, and eluted with 7 µL ONT Elution Buffer (included in SQK-LSK114 kit).

Sequencing using ONT MinION. The Flongle R10.4.1 (FLO-FLG114) flow cell was placed into the Flongle Adapter provided with the Flongle Starter Pack (FLGIntSP). Prior to sequencing samples, flow cell check was performed following the 'flow-cellpcheck-PQE_1004-V1' protocol to ensure a minimum of 50 active pores were present. The flow cell was prepared and run according to ONT instructions: the recommended 5-10 fmol of prepared library was loaded per flow cell, and the run time was set to 24 hours. The Flongle Adapter was USB connected to an M1 Macbook Pro with 16 GB of RAM. On-device base calling was disabled, instead performing base calling in a later step on our higher-powered computational resources.

Base calling raw sequence data with Guppy. We assembled all steps of our pipeline into a Snakemake v6.3.0 workflow. We performed initial base calling of the raw .fast5 data using the GPU version of Guppy v6.5.7, selecting the "dna_r10.4.1_e8.2_400bps_sup.cfg" base calling configuration file. We ran base calling on a GPU cluster comprised of eight NVIDIA GeForce RTX 3090 GPUs, 256 GB RAM, and two Intel Xeon Silver 4212 12 core CPUs. All .fast5 files generated by MinKNOW for each sample were submitted to the cluster in its own job, requesting 20 GB RAM, 1 GPU, and 4 CPUs. To obtain the most useable data possible, we retain the failed reads, as reads with poor quality at critical positions will be filtered out later in our pipeline. We ran Porechop v0.2.4 (https://github.com/rrwick/Porechop) to remove any remaining adapter sequences and separate chimeric reads.

Independent control data. We obtained existing base called sequence data of *CYP2D6* from other Coriell samples from Liau et al. (Liau, Maggo, *et al.*, 2019) and Wongsurawat et al. (Wongsurawat *et al.*, 2025). These datasets were generated in different labs, using different primers, and sequenced on different versions of the ONT flow cells. These datasets are used to help us validate that our process is not dependent on a specific amplicon or data source.

Inference of genomic amplicon. All of these raw reads are aligned to the human genome (hg38) using minimap2 v2.17-r941 (Li, 2018) with the following parameters: "-x map-ont --secondary=no –t 4". To infer the reference amplicon sequence in an unbiased manner, we use samtools v1.15 (Danecek *et al.*, 2021) to get the depth of read coverage at every position of the genome, and extract out the regions that have depth above the 50th percentile. Any adjacent regions are merged using bedtools v2.26.0 (Quinlan and Hall, 2010) with the following parameters: "merge -d 100".

Candidate read extraction. We extract all the reads that had a primary or supplemental alignment to the identified genomic region(s). We observed a high number of short reads in our in-house-generated data. To further group reads into those completely covering the same amplicon sequence, we perform length clustering on the candidate reads. Clustering is done using the DBSCAN algorithm (Schubert *et al.*, 2017) from the scikit-learn library (v0.24.2), run on the read lengths, with the allowed neighbour length variability set to 10 and the minimum core point neighbours set to 0.1% of the total number of candidate reads. Clusters are selected for consideration for downstream allele reconstruction if the number of reads in the cluster is at least 25 % of the size of the largest cluster, selected empirically. Additionally, to ensure we do not miss the cluster with reads the full-length of our inferred amplicon, we select the cluster with the length closest to the inferred amplicon region if it is not already included.

Allele sequence reconstruction. We align each group of candidate reads to their inferred amplicon reference sequence using minimap2 v2.17-r941. The set of all alignments are parsed, position by position, to assess the number of reads supporting each base at each position. If a position has evidence of heterozygosity (here, defined as the top-observed base having less than double the depth of the second most-observed base, threshold determined empirically), the reads are partitioned into separate groups based on support of the heterozygous bases (**Figure 3C-D**), and consensus sequence construction tries again on each of the new read groups independently. This repeats, recursively, until pure (no evidence of heterozygosity) consensus sequence(s) are derived, or the read groups are too small to continue. Due to the phenomenon of strand-specific errors in ONT data (Krishnakumar *et al.*, 2018; Branton *et al.*, 2008), this test of purity is also performed on the forward and reverse reads independently.

Our allele reconstruction method is built as a Python (v3.6.13) script, and has a number of parameters that can be adjusted to fine-tune the behavior of the allele reconstruction. Based on all of the control data, we used the following settings, determined empirically, and explained

subsequently: --min_depth_factor 0.25, --global_min_depth_thresh 0.05, --min_base_qual_global 10, --min_base_qual_percentile 33, --sig_to_noise_thresh 2. The "--min_depth_factor" controls the minimum number of reads needed to support a base at any position, and represents the fraction of the maximum observed depth (here, 25 %). The "--global_min_depth_thresh" controls the minimum number of reads supporting a reconstructed allele sequence for it to be output, here set to 5 % of the number of input reads. The "--min_base_qual_global" is the minimum base quality score needed to be used for allele sequence reconstruction. If this threshold is not met, then the "--min_base_qual_percentile" is used, being a percentile threshold, per position, that is the lower base quality limit. This allows for data to be obtained from positions with lower quality overall. Finally, the "--signal_to_noise_thresh" is the value of the minimum ratio of the most frequent base to the second most frequent base at any position needed to be considered homozygous. This needs to be met when considering all reads, and is adjusted by a factor of 1.5 when also testing reads from each strand separately. Based on quality thresholds, a read may be "ambiguous" for a given position (supporting a base, but due to low quality the base is not used). Ambiguous reads do not contribute to depth calculations at that position. If the position is heterozygous, ambiguous reads for that position are discarded. The average quality score for each consensus base across all reads used for the consensus is recorded and output in the detailed results file, along with base frequencies, allowing manual inspection of the reconstructed allele sequences.

Automatic genotyping of consensus sequence. Each resulting reconstructed allele sequence is aligned to the *CYP2D6* reference sequence using Biopython v1.79 pairwise2.align.localxs, and the alignment is parsed to enumerate all variants. All known *CYP2D6* variants are catalogued from the *CYP2D6* Allele Definition Table from PharmGKB (Whirl-Carrillo *et al.*, 2021, 2012), downloaded from https://www.pharmgkb.org/page/cyp2d6RefMaterials on January 21, 2025. This file was manually curated to ensure consistency in how each variant type was reported, allowing for straightforward parsing. For variants that can have multiple possible bases allowed for a given allele, all possible allowed values are stored. The set of identified variants for each reconstructed allele sequence is compared to those from each known allele, and a score is calculated for each known allele based on the number of disagreements in identified variants within the exon and splice regions of the gene. The allele with the lowest score is returned as the best candidate allele assignment for the reconstructed allele sequence, along with any missing or extra variants present in the consensus sequence compared to the star allele reference should they exist. In this way, manual review of the candidate alleles and identified variants allows a user to confirm the allele assignment, or characterize a novel allele.

List of Abbreviations

CPU	Central Processing Unit
CYP	Cytochrome P450
DNA	Deoxyribonucleic Acid
GPU	Graphics Processing Unit
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PCR	Polymerase Chain Reaction

Declarations

Ethics approval and consent to participate

This study was approved by the Fraser Health Research Ethics Board and BC Cancer Agency Research Ethics Board (H21-01422).

Consent for publication

Not applicable.

Availability of data and materials

Custom code is available at the following GitHub repository: https://github.com/scottdbrown/allele-reconstruction-long-read-amplicon-data. Raw .fastq sequence data for our three sequencing runs is available at the SRA under Bioproject PRJNA1357883 (https://www.ncbi.nlm.nih.gov/bioproject/1357883).

Competing interests

The authors declare that they have no competing interests.

Funding

This project was supported by a Fraser Health Authority Strategic Priorities Grant and a directed donation funded by Royal Columbian Hospital Foundation.

Authors' Contributions

RAH, PIM, and SDB designed the study. LD performed library preparation and sequencing optimization. SM, AL, MK, AC, and MM performed library preparation and sequencing. AM, NW, and MH coordinated and oversaw the technical work conducted in the Molecular Cytogenetics Laboratory at Royal Columbian Hospital. SDB performed the sequence analysis and algorithm development. SDB and PIM wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank Dr. Martin Kennedy for sharing additional data from Liau et al. (Liau, Maggo, *et al.*, 2019).

References

Ammar,R. *et al.* (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research*, **4**, 1–17.

Branton,D. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol. 2008 2610*, **26**, 1146–1153.

Brown,S.D. *et al.* (2023) Complete sequence verification of plasmid DNA using the Oxford Nanopore Technologies' MinION device. *BMC Bioinformatics*, **24**, 1–16.

Chang,K.L. *et al*. (2015) Pharmacogenetics: Using genetic information to guide drug therapy. *Am. Fam. Physician*, **92**, 588–594.

Charnaud,S. *et al.* (2022) PacBio long-read amplicon sequencing enables scalable high-resolution population allele typing of the complex CYP2D6 locus. *Commun. Biol.*, **5**, 1–10.

Chen,X. *et al.* (2021) Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. *Pharmacogenomics J.*, **21**, 251–261.

Danecek,P. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, 1–4.

Deserranno,K. *et al.* (2023) Targeted haplotyping in pharmacogenomics using Oxford Nanopore Technologies' adaptive sampling. *Front. Pharmacol.*, **14**, 1–11.

Gaedigk,A. *et al.* (2019) Characterization of Reference Materials for Genetic Testing of CYP2D6 Alleles: A GeT-RM Collaborative Project. *J. Mol. Diagnostics*, **21**, 1034–1052.

Hari,A. *et al.* (2023) An efficient genotyper and star-allele caller for pharmacogenomics. *Genome Res.*, **33**, 61–70.

Hicks,J.K. *et al.* (2017) Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update. *Clin. Pharmacol. Ther.*, **102**, 37–44.

Kiyotani,K. *et al.* (2008) Impact of CYP2D6*10 on recurrence-free survival in breast cancer patients receiving adjuvant tamoxifen therapy. *Cancer Sci.*, **99**, 995–999.

Krishnakumar,R. *et al.* (2018) Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Sci. Reports 2018 81*, **8**, 1–13.

Van der Lee,M. *et al.* (2021) Toward predicting CYP2D6-mediated variable drug response from CYP2D6 gene sequencing data. *Sci. Transl. Med.*, **13**.

Li,H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Liau,Y., Cree,S.L., *et al.* (2019) A multiplex pharmacogenetics assay using the MinION nanopore sequencing device. *Pharmacogenet. Genomics*, **29**, 207–215.

Liau,Y., Maggo,S., *et al*. (2019) Nanopore sequencing of the pharmacogene CYP2D6 allows simultaneous haplotyping and detection of duplications. *Pharmacogenomics*, **20**, 1033–1047.

van der Maas,S. *et al*. (2025) Dynamic star allele definitions in Pharmacogenomics: impact on diplotype calls, Phenotype predictions and statin therapy recommendations. *Front. Pharmacol.*, **16**, 1–8.

McDonnell,A.M. and Dang,C.H. (2013) Basic Review of the Cytochrome P450 System. *J. Adv. Pract. Oncol.*, **4**, 263–268.

Patterson,M.D. *et al.* (2015) WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.*, **22**, 498–509.

Pratt,V.M. *et al.* (2010) 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. Diagnostics*, **12**, 835–846.

Pratt,V.M. *et al.* (2016) Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative

Project. *J. Mol. Diagnostics*, **18**, 109–123.

Pratt,V.M. *et al.* (2021) Recommendations for Clinical CYP2D6 Genotyping Allele Selection: A Joint Consensus Recommendation of the Association for Molecular Pathology, College of American Pathologists, Dutch Pharmacogenetics Working Group of the Royal Dutch Pharmacists Association. *J. Mol. Diagnostics*, **23**, 1047–1064.

Puaprasert,K. *et al.* (2018) Real time PCR detection of common CYP2D6 genetic variants and its application in a Karen population study. *Malar. J.*, **17**, 8–11.

Qiao,W. *et al.* (2016) Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6. *Hum. Mutat.*, **37**, 315–323.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rubben,K. *et al.* (2022) Cas9 targeted nanopore sequencing with enhanced variant calling improves CYP2D6-CYP2D7 hybrid allele genotyping. *PLoS Genet.*, **18**, 1–21.

Sayers,E.W. *et al.* (2025) Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.*, **53**, D20–D29.

Schubert,E. *et al.* (2017) DBSCAN Revisited, Revisited. *ACM Trans. Database Syst.*, **42**.

Somogyi,A.A. *et al.* (2015) Pharmacogenetics of opioid response. *Clin. Pharmacol. Ther.*, **97**, 125–127.

Teusink,A. *et al.* (2016) Genotype-directed dosing leads to optimized voriconazole levels in pediatric patients receiving hematopoietic stem cell transplantation. *Biol. Blood Marrow Transplant.*, **22**, 482–486.

Whirl-Carrillo,M. *et al.* (2021) An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin. Pharmacol. Ther.*, **110**, 563–572.

Whirl-Carrillo,M. *et al.* (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, **92**, 414–417.

Wongsurawat,T. *et al.* (2025) Enhancing CYP2D6 genotyping with nanopore sequencing to address allele diversity P . vivax malaria elimination. *Sci. Rep.*, **15**, 33842.

Zheng,Z. *et al.* (2022) Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.*, **2**, 797–803.

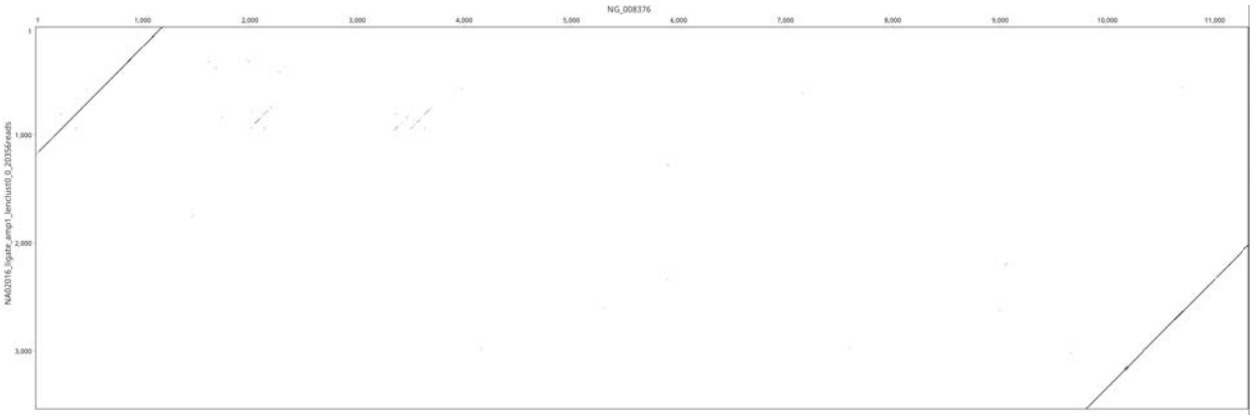
Supplemental

Supplementary Table S1. Summary table of all inferred amplicons, length-clusters, and allele reconstruction attempts with read counts at each stage.

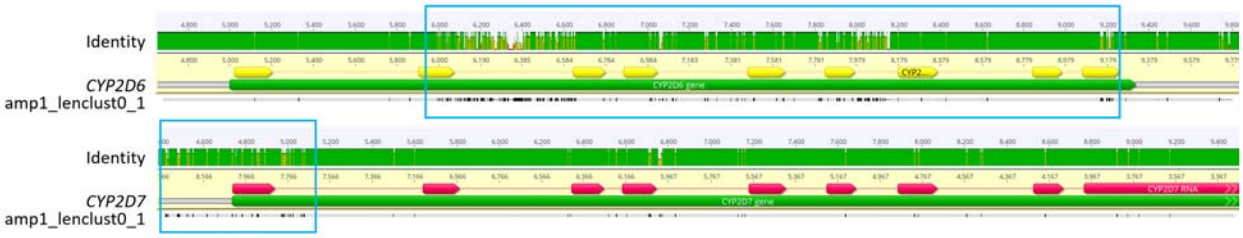
Supplementary File S1. All successfully constructed consensus sequences for all samples.

Supplementary Figure S1. *CYP2D6* read coverage from short reads. Screenshot from Geneious Prime. The length-cluster of short reads from NA17300 (40,717 reads with mean length 445 bases) was aligned to the *CYP2D6* reference sequence. Coverage is shown in blue, spanning the entire expected amplicon. Exons are marked with grey segments under the green *CYP2D6* gene. Coverage is variable, and the region of highest coverage (exon 4-5) yielded a single reconstructed consensus sequence.

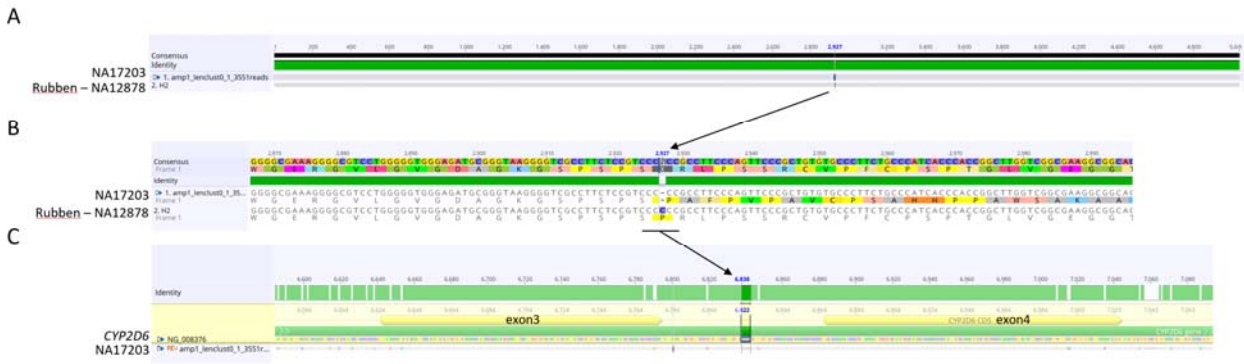
Supplementary Figure S2. Alignment of long and short consensus sequences to *CYP2D6*. Screenshots from Geneious Prime. For NA10005 (top) and NA17300 (bottom), short consensus sequences were obtained from the respective length-clusters of short reads. The sequences of these short consensus sequences match at least one of the long consensus sequences in each sample, with the exception of mismatches at the beginning and end of the sequences.



Supplementary Figure S3. Dot plot of duplication-detecting consensus sequence and *CYP2D6*. Screenshot from Geneious Prime. The consensus sequence from NA02016_ligate from the 3.5 kb duplication-detecting amplicon (y-axis) is compared to the *CYP2D6* reference sequence (x-axis), showing the consensus sequence supporting the region upstream and downstream of *CYP2D6* as would be expected in the event of an allele duplication.



Supplementary Figure S4. Suspected *CYP2D6-CYP2D7* hybrid allele sequence aligned to *CYP2D6* and *CYP2D7*. Screenshots from Geneious Prime. The *CYP2D7*-locus-dervied consensus sequence from NA17203_ligate is aligned to the *CYP2D6* reference (top) and *CYP2D7* reference (bottom), with the sequence identity as the green histogram above the sequences. Regions of lower sequence identity to either reference are boxed in light blue. This consensus sequence appears to best support a *CYP2D6-CYP2D7* hybrid allele, with exon 1 from *CYP2D6*, and the remainder of the sequence from *CYP2D7*.



Supplementary Figure S5. Alignment of NA17203_ligate *CYP2D6-CYP2D7* hybrid allele sequence to **68* hybrid allele sequence. Screenshots from Geneious Prime. (A) Our NA17203_ligate **68* allele sequence shows a perfect match to the **68* sequence identified from NA12878 identified by Rubben et al, (B) with the exception of a C deletion observed in our sequence. (C) When our **68* hybrid allele sequence is aligned to the *CYP2D6* reference sequence, this deletion falls between exons 3 and 4, and thus would not be expected to change the function of the resulting protein.