Nanopore sequencing of the pharmacogene CYP2D6 allows simultaneous haplotyping and detection of duplications

Yusmiati Liau[1], Simran Maggo[1], Allison L. Miller[1], John F. Pearson[1], Martin A. Kennedy[1], Simone L. Cree[1]

[1] Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand

Abstract

Background: The accurate genotyping of *CYP2D6* is hindered by the very polymorphic nature of the gene, high homology with its pseudogene *CYP2D7*, and the occurrence of structural variations. Long read sequencing offers the promise of overcoming some of these challenges, along with the advantage of straightforward variant phasing. We have established methods for sequencing and analysis of DNA amplicons containing the whole *CYP2D6* gene, using the GridION nanopore sequencer. **Materials and methods:** Seven reference and 25 clinical samples covering various haplotypes including gene duplication were barcoded and sequenced over two sequencing runs. Sequenced raw reads were analyzed using a pipeline of bioinformatics tools including two mapping tools and two variant calling tools. **Results:** Using minimap2 and nanopolish (mapping and variant calling tools respectively) resulted in the most accurate variant detection. Haplotypes of all samples could be detected accurately to the subvariant level in most samples, while the remaining samples had either novel variants or variant patterns not matched to the current PharmVar *CYP2D6* haplotype database. Allele duplication could also be determined by analyzing the allelic balance between the sample haplotypes. **Conclusion:** Nanopore sequencing of *CYP2D6* offers a high throughput method for genotyping, accurate haplotyping, and detection of new variants and duplicated alleles.

Introduction

CYP2D6 is one of the most important and widely studied genes in pharmacogenetics. It codes for the cytochrome P450 2D6 enzyme which is responsible for the metabolism or bioactivation of ~25% of all available drugs, including antidepressants such as selective serotonin reuptake inhibitors (SSRIs), analgesics such as codeine, and anti-cancer drugs such as tamoxifen (1). *CYP2D6* is also one of the most polymorphic human genes, leading to considerable variability in enzyme activity between individuals with the need for accurate genotype information for clinical purposes (2).

CYP2D6 is a relatively small gene (4.4 kilobases from start to stop codon) located on the long arm of chromosome 22. However, the complexities of this gene make accurate genotyping challenging (3-5). The first challenge lies in the number of variants in the gene. PharmVar (https://www.pharmvar.org/gene/CYP2D6) has compiled more than 100 allelic variations and more than 200 genetic variants (6) in this gene. Moreover, the GnomAD database of more than 140K individuals suggests there may be more than 1000 variants in this gene, including 374 missense and 28 loss-of-function variants (7).

In addition to its polymorphic sequence, *CYP2D6* can be affected by structural variations including gene deletion, gene duplication, hybrid gene conversion or tandem arrangement with its upstream pseudogene, *CYP2D7*. Since *CYP2D6* and *CYP2D7* share more than 90% sequence similarity and occur in tandem on the chromosome, care must be taken to selectively target CYP2D6 to obtain accurate genotype information for clinical utility (8, 9).

Various methods have been developed to genotype *CYP2D6,* from the basic restriction fragment length polymorphism to the more sophisticated real time polymerase chain reaction (PCR), microarray, mass spectrometry, or next generation sequencing (10-12). Most methods are limited by the ability to detect only a subset of pre-selected variants. While next generation sequencing (NGS) methods have the ability to screen all *CYP2D6* variants, these methods are hampered by the risk of misalignment of *CYP2D6* and *CYP2D7* due to the short-reads generated. Most of these methods also do not provide information for variant phasing to confidently determine haplotypes. In practice, this is often carried out using allele specific primers to isolate one or both alleles, followed by Sanger sequencing of the amplification products (13). This is labour intensive, requires a greater investment of time, and can be technically demanding.

Long read sequencing is envisaged to resolve challenges faced by previous short read technology. For *CYP2D6* genotyping, long amplicon sequencing not only permits unequivocal variant detection without interference from the homologous pseudogene, it also means that variant phasing can be carried out in a reliable and straightforward fashion. A preliminary report

35 described *CYP2D6* sequencing of a single sample using the Oxford Nanopore Technologies (ONT)
36 MinION with an earlier, more error-prone version of the technology (14). In this paper, we
37 explore the application of current ONT long read sequencing chemistry to simultaneously analyze
38 *CYP2D6* from multiple reference and clinical samples. We performed the sequencing on the
39 GridION platform, a compact nanopore-based instrument from ONT (Oxford, UK) which enables
40 sequencing and data computation in the same module in a real time manner.

41

42 Materials and Methods

43 Samples

44 Thirty-two samples were genotyped over two sequencing runs, with 12 samples in the first and
45 24 samples in the second run. Four samples from the first run were repeated in the second run as
46 control samples. Seven *CYP2D6* genotyped samples (15) were obtained from Coriell Institute
47 (Camden, NJ, USA) and 25 unrelated clinical samples were collected from two adverse drug
48 reaction cohorts established in our laboratory, Understanding Adverse Drug Reactions Using
49 Genomic Sequencing (UDRUGS) and Genetics of Antidepressants study (GO-A) (16). These studies
50 received ethical approval respectively from Southern Health and Disability Ethics Committee
51 (URA/11/11/065) and Northern B Health and Disability Ethics Committee (18/NT/21), Ministry
52 of Health, New Zealand. The clinical samples consisted of 14 blood samples and 11 saliva samples.
53 DNA from blood samples was extracted using a modified salting-out method (17) followed by a
54 phenol-chloroform purification step (Supplementary File 1). Saliva samples were collected via an
55 Oragene DNA (OG-500) kit (DNA Genotek Inc., ON, Canada) and DNA was extracted according to
56 the manufacturer's protocol.

57 CYP2D6 Long PCR

58 The 6.6 kilobases (kb) long PCR amplifying the whole *CYP2D6* gene including upstream and
59 downstream non-coding regions was performed according to Gaedigk, 2007 (18). Specific
60 sequences from ONT were added to the 5' end of forward and reverse primers to enable sample
61 barcoding in subsequent steps. 250 ng of genomic DNA was amplified in the presence of 1x Kapa
62 LongRange Buffer, 0.3 mM dNTPs, 1.75 mM $MgCl_2$, 0.7 μM of each primer, 1 M of betaine, and 1.25
63 units of Kapa LongRange Taq Polymerase (Sigma-Aldrich) in a 50 μL reaction. PCR was run as
64 follows: 94°C for 3 minutes, 30 cycles of 94°C for 25 seconds, 68°C for 10 seconds, 68°C for 7
65 minutes, and a final extension at 68°C for 7 minutes.

66 Deletion and multiplication of the *CYP2D6* gene can be identified by amplifying a 3.5 kb fragment
67 with CYP-13 and CYP-24 primers (19) for deletion and Frag B primers (18) for duplication in a

68 duplex PCR together with the 6.6kb primers. This PCR was carried out as above using 0.4 μM of
69 the 6.6 kb primers and 0.3 μM of the deletion or duplication primers.

70 In samples where a gene duplication was detected, the duplicated allele was amplified with tailed
71 FragD primers (18) which produced 8.6 kb fragments, and these were subsequently barcoded
72 and sequenced on the GridION. This primer set recognizes the REP-DUP region, a unique
73 combination of *CYP2D6* and *CYP2D7* downstream regions commonly found in duplications of the
74 CYP2D6 gene. This PCR was similar to that for *CYP2D6* described above, except 0.4 μM primers
75 and extension for 10 minutes instead of 7 minutes were used. The sequences of all primers used
76 in this study are listed in Supplementary Table 1.

77 Barcoding PCR

78 Specific barcodes were assigned to the 6.6 kb PCR products of each sample and 8.6 kb PCR
79 products (where applicable) using the PCR Barcoding Expansion 1-96 kit (EXP-PBC096) (ONT).
80 Barcodes were added using a second-round PCR to 0.5 nM of CYP2D6 6.6 kb or 8.6 kb purified
81 PCR product in the presence of 1x LongAMP PCR buffer, 2 mM of $MgCl_2$, 0.3 mM of dNTPs, and 5
82 units of LongAmp Taq DNA polymerase in 50 μL reaction. PCR was performed as follows: 95°C
83 for 3 minutes, 15 cycles of 95°C for 15 secs, 62°C for 15 seconds, 65°C for 7 minutes (6.6 kb
84 products) or 10 minutes (8.6 kb products), with a final extension of 65°C for 7 minutes or 10
85 minutes.

86 Modified Beads Purification

87 Both initial and barcoding PCR products were purified using magnetic beads (HighPrep™ PCR,
88 MagBio, Gaithersburg, MA, USA). In the first sequencing run, we observed that short (~1 kb), off-
89 target DNA fragments were sequenced. Therefore, for the second run, to remove any DNA
90 fragments shorter than 3-4 kb, modified beads were made based on the protocol of Ramawater
91 and Schwessinger (20). In brief, Magbio beads (e.g. 5 mL) were pelleted and re-dissolved in buffer
92 containing 10 mM Tris-HCL, 1 mM EDTA pH 8, 1.6 M NaCl, 11% (w/v) PEG 8000, 0.20% (v/v)
93 Tween-20, and Millipore water to make 5 mL bead solution. For each sample, 0.8x volume of
94 beads was used.

95 Nanopore Sequencing Library Preparation

96 Each DNA library was prepared according to the 1D PCR barcoding (96) amplicons ONT protocol,
97 SQK-LSK109. Briefly, after pooling the barcoded amplicons, 200 fmoles of the DNA pool were
98 subjected to end-repair and ONT adapter ligation. After purification, approximately 350 ng (~80
99 fmoles) were introduced to the R9.4 flowcell to be sequenced for 48 hours on the GridION.

100 **Nanopore Sequencing Data analysis**

101 DNA sequences were extracted from fast5 files using Guppy v2.2.2 with flip-flop basecalling (ONT,
102 Oxford, UK). Only pass reads (qscore > 7) were used for subsequent genotyping analysis as
103 outlined in Figure 1. Guppy basecaller did not accommodate demultiplexing or length filtering
104 when we performed this analysis, and these were done subsequently using Porechop
105 (https://github.com/rrwick/Porechop) and NanoFilt (21). Filtered fastq files obtained from
106 NanoFilt were then mapped to reference sequence(NG_008376.3, 8593 bps) using minimap2
107 (22) or NGMLR (23), and variants were called from the mapped bam files to a vcf file using
108 nanopolish (24) or Clairvoyante (25). Threshold for calling variants was set at 0.2 for nanopolish
109 and 0.25 for Clairvoyante as recommended. A trained model for ONT was also used when calling
110 variants using Clairovyante. Variant phasing in samples containing more than one heterozygote
111 variant was carried out using WhatsHap (26) with maximum coverage of 25X, resulting in phased
112 VCF and phased BAM files. Quality control metrics were obtained using NanoOK (27) and
113 MinIONQC (28).

114 Variant nomenclature follows the numbering based on NG_008376.3 counting from the sequence
115 start. Haplotype information for each allele was determined by comparing the variants in each
116 allele to the PharmVar haplotype table. Duplicated alleles were identified by examining the ratio
117 of allele frequency between variants from the two alleles of each sample. The presence of
118 duplicated alleles was indicated by a skewed allele balance from 1:1 to 2:1, or more in the case of
119 multiple copies.

120 Genotype information for reference samples was available through the published Get-RM dataset
121 (15). Genotype information for the clinical samples was determined using Sanger Sequencing
122 based on key variants (Supplementary Table 2).

123 **Sanger sequencing**

124 Nested PCRs covering different regions (Supplementary Table 1) on the 1:1000 diluted 6.6 kb
125 *CYP2D6* PCR product were performed and used for Sanger sequencing on an AB3130XL Genetic
126 Analyzer (Thermo Fisher Scientific) (Supplementary Table 1). Nested PCRs were carried out in
127 20 μL reactions containing 1x PCR buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.2 μM of each primer,
128 0.5 unit of TAQ-TI Heat-Activated DNA Polymerase (Fisher Biotec, Wembley, WA, Australia) and
129 2 uL of diluted PCR product. DNA was amplified in a touchdown PCR consisting of 15 cycles of
130 94°C for 15 seconds, 70°C for 15 seconds, and 72°C for 1 minute with annealing temperature
131 decrease of 1°C/cycle followed by 20 cycles of 94°C for 15 seconds, 55°C for 15 seconds, and 72°C
132 for 1 minute and a final extension at 72°C for 5 minutes. Sanger sequencing was performed using
133 1 μL product of nested PCR, 1 μL of primer, 1x sequencing buffer, and 0.5 μl of Big Dye Terminator

134 (ThermoFisher Scientific, Carlsbad, USA) in 10 μL reaction. Dye incorporation was performed in
135 a thermal cycler with 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds, and 60°C for 4 minutes,
136 and subsequently purified using Sephadex G-50 (Sigma-Aldrich, Saint Louis, USA).

137 # Results

138 Both sequencing runs resulted in similar yield (~3.6 gigabases), but different numbers of reads.
139 The first sequencing run resulted in more than 90% reads having length less than 6 kb, possibly
140 due to unspecific amplification during barcode PCR. In the second run, short reads were
141 significantly reduced and target sequences of 6 kb were enriched (~55% of total reads) with the
142 use of modified beads during PCR product purification (Figure 2). Although the first run resulted
143 in more reads in total, there were more barcoded reads with with >6 kb length in the second run
144 (35% compared to 4%). Despite the differences in number of reads between runs and between
145 samples, all samples had adequate coverage for subsequent analysis and genotyping, with the
146 lowest sample coverage at 235x (Table 1).

147 ## Comparison of Mapping and Variant Calling Tools

148 Amplicon data analysis was performed by comparing two mapping (minimap2 (22) and NGMLR
149 (23) and two variant calling tools (nanopolish (24) and Clairvoyante (25)), resulting in a total of
150 four different analyses. These comparisons were carried out with seven reference samples from
151 the first sequencing run. False positive and negative variants were confirmed by visually
152 inspecting the Binary Alignment MAP (BAM) file on Integrative Genomics Viewer (IGV, Broad
153 Institute, Boston, MA) (29). The minimum negative predictive value was 99.96% (NGMLR –
154 nanopolish) and the minimum accuracy was 99.68% (NGMLR – Clairvoyante). The calculation for
155 these two measurements was dominated by the relatively high number of true negative variants
156 (6514 – 6561 variants). Clairvoyante generally generated more false positive variants compared
157 with nanopolish, resulting in higher False Discovery Rate values (Table 2). In addition, we
158 observed a misalignment event in NGMLR in the intron 1 CYP2D7 conversion region (seven
159 SNPSs, 4414-4445) (Supplementary Figure 1). Due to this misalignment, the combination of
160 NGMLR and nanopolish inaccurately called 4423C>G and 4427T>C as 4423insG and 4427delT
161 (Supplementary Figure 1). Based on our analyses, the combination of minimap2 and nanopolish
162 for mapping and variant calling resulted in the highest sensitivity and positive predictive value
163 and lowest false discovery rate compared to the other tool combinations, and hence was used to
164 analyze the entire sample cohort.

165 **CYP2D6 Genotyping Results**

166 Together, the reference and clinical samples cover a total of 16 different star (*) alleles (*1, *2, *3,
167 *4, *6, *7, *9, *10, *15, *17, *28, *29, *33, *35A, *41, and *71), as well as one sample with gene
168 deletion (*5) and four samples with gene duplication. The only sample with a gene deletion
169 (S022) was heterozygous, so only one allele was available to be sequenced. This allele contains a
170 hemizygous variant (6684G>T), which corresponds to *33. This variant was confirmed on Sanger
171 sequencing and the sample was assigned as *5/*33.

172 The genotyping results for all samples are shown in Table 3. Genotypes for all samples completely
173 matched the reference genotypes based on getRM data and Sanger Sequencing, and 21 of them
174 could be genotyped to the subvariant level, and matched to known subvariant catalogue entries
175 on PharmVar. Five samples matched to certain alleles but not to the sub alleles, while six samples
176 contained variants that did not match any known allele pattern on PharmVar. These findings are
177 described below.

178 One sample (S016) contained only one variant (6504C>T, rs79738337) which is associated only
179 with *60, but lacks the key variant of *60 (6088insTA). As 6504C>T is an intronic variant, we
180 propose this sample to be a new subvariant of *1. Another sample (S007) had two variants
181 (4325G>A and 5695T>C), associated with *71 but lacked the 2617C>G variant and therefore
182 cannot be matched to either *71.001 or *71.002 as per the PharmVar catalogue. We propose this
183 sample to be a new subvariant of *71.

184 One sample (S003) had a 6750delA variant in one allele which was called as *3. In the other allele,
185 there are nine SNPs and one insertion variant, which taken together, cannot be matched to any
186 star (*) allele in PharmVar (4470C>T, 7436A>G, 7455T>C, 7466G>A, 7585A>C, 7679A>G,
187 7685G>A, 7746insT, 7753C>T, 7783A>G). From the 10 variants found in this allele, three of them
188 are in the exonic regions, one of which (7436A>G) is a missense variant, classified as moderate
189 by Variant Effect Predictor (VEP, Ensembl) (30) and is one of the key variants associated with
190 *108. However, the *108 allele has another key variant also in exon 7, 7427A>G (rs61736517),
191 which was not present in our sample. 7466G>A, also a missense variant and predicted to be
192 deleterious by VEP, has been associated with *4.021, but this sample lacks other key variants of
193 *4. We propose this sample is a new allele.

194 We found five variants not yet catalogued in PharmVar. Two of the Coriell samples (NA18518 and
195 NA23297) with *17 allele had an additional intronic variant (4853C>T, rs376217512). This is a
196 rare variant with allele frequency (AF) of ~1%. One sample with variants matching to *41 allele
197 (S013) contained two non-catalogued variants, 5578C>G (rs143170489) in intron 2 and
198 5814C>T (rs61736504) in exon 3. Both are rare with AF ~0.1%. The exonic variant is predicted

199 to be benign when annotated using VEP. Another sample with *4 allele (S023) also had a rare (AF:
200 0.001) intron 2 variant, 5604C>T, rs557722765.

201 The fifth variant not catalogued in PharmVar was an upstream variant (3836G>A, rs1080992)
202 which was found in four samples, and was the only variant in the particular allele of each sample.
203 This rare variant with an AF of less than 1% (7) has been reported in a non CEU population (31).
204 No functional information is available for this variant.

205 Detection of duplicated alleles

206 The presence of a duplicated allele in samples was determined by interrogating the ratio of
207 coverage between alleles. Samples without a duplication were expected to have a 1:1 ratio
208 between alleles, while samples with a duplication were expected have a skewed ratio of ~2:1
209 (Figure 3). All four samples which indicated gene duplication during the initial PCR screening
210 showed an uneven allele coverage ratio, and the duplicated alleles could be determined from the
211 over-represented allele. In addition, we sequenced the 8-kb fragments in three out of the four
212 samples with duplication. The 8-kb fragments originated only from the duplicated allele, and
213 aligned to the NG_008376.3 reference sequence from position 3860 to 8593. Sequencing of these
214 fragments revealed the same variants as the duplicated alleles identified from the 6 kb fragments
215 and thus confirmed the duplicated allele in all three samples (NA07439, NA23297, S018) (Table
216 4).

217 Validation of duplicated allele and variant phasing using the 8-kb fragment.

218 Interrogating the variants in the 8 kb fragments served two purposes. First, this information
219 identified the duplicated allele as explained in the previous section. Second, comparison of the
220 variants found in the 8-kb amplicon to those found in the 6.6kb amplicon allowed validation of
221 variant phasing performed in the 6.6 kb amplicon. All variants found in the 8-kb amplicons of the
222 three samples matched completely those found on the over-represented (duplicated) allele on
223 the 6.6 kb fragments. In addition to validation of the duplicated allele, this approach also validated
224 the accuracy of variant phasing of the 6.6kb amplicons. All phased variants in all samples could
225 be matched to a particular allele or subvariant, confirming the accuracy of variant phasing.

226 Identification and Filtering of False Positive Variants

227 Nanopore sequencing still fails to accurately basecall homopolymer strings. When aligned to the
228 reference, this usually leads to a drop in coverage and variant calling tools report this as a
229 deletion. Two to 13 false positive deletion/insertion events were found in each sample, and all
230 were in regions with three or more sequential identical bases. In addition to errors due to
231 homopolymers, false positive transition variants also occurred in two or three variants per

232 sample. These variants are non-random, always transitions (A↔G or C↔T) and called only in
233 specific position: 2381A>G, 2446T>C, 5207C>T, 6440T>C, 8046C>T, 8047C>T, and 8723G>A.
234 These false positive variants have lower quality in the nanopolish produced VCF files compared
235 to the true variants. We applied a previously reported strategy (32) by using quality/coverage
236 ratio to differentiate true and false variants. Most of the true variants have quality/coverage ratio
237 of >1, while most of the false variants have ratio < 0.5. For variants with ratio value between 0.5-
238 1, the trueness of the variant was determined visually in the BAM files using IGV.

239 Although all samples could be genotyped unequivocally, several variants failed to be detected and
240 were not reported in the VCF file. One variant that was present but not reported in many samples
241 was 5042T>G. This SNP sits between four Gs in the upstream and another G downstream, and
242 when the alternate G exists, the read depth dropped significantly, presumably due to the inability
243 of nanopore sequencing to detect homopolymers correctly. As a result, this SNP was often called
244 incorrectly as a T-deletion instead of a SNP. However this variant is quite common and is shared
245 by many alleles including the wildtype *1, *2, *4, *10, *41 and many others, and the absence did
246 not interfere with correct haplotyping of any sample.

247

248 # Discussion

249 We performed long amplicon nanopore sequencing of the whole *CYP2D6* gene (including
250 upstream and downstream regions) on the GridION nanopore sequencing platform. To our
251 knowledge, this is the first extensive analysis of *CYP2D6* on a nanopore sequencing platform
252 across multiple star (*) alleles, including assessment of structural variation. An earlier report of
253 *CYP2D6* nanopore sequencing only used one reference sample and was done on the earlier
254 version of ONT chemistry and flowcell. High error rates were associated with that version, and
255 unresolved ambiguity was reported in the sample diplotype, although variant phasing was
256 performed accurately (14). The initial relatively high error rate associated with nanopore
257 sequencing has improved as the technology continues to evolve. This is mainly attributed to
258 improvements in the pore, sequencing chemistries and bioinformatic algorithms available for
259 analysis. The accuracy of ONT nanopore sequencing has improved from less than 80% with the
260 R7.3 flowcell (33) used by Ammar et. al, to more than 90% with the current R9.4 version (34).

261 We used the R9.4 flowcell with 1D sequencing chemistry, and generated DNA sequences using
262 the recently updated guppy basecaller from ONT. Two different mapping and variant calling tools
263 were then used to analyze the reference samples to select the best bioinformatic workflow for
264 our dataset. We observed that a combination of minimap2 as a mapping tool and nanopolish as a
265 variant calling tool generated the least number of false positive and false negative variants

266 compared to other workflows. Using this combination of tools, the entire cohort of 32 samples in
267 our study was accurately assigned to their star (*) alleles, with the addition of subvariant and
268 duplicated allele information.

269 The advantages of long amplicon sequencing of *CYP2D6* have been recently reported by two
270 papers, using the PacBio single molecule real-time (SMRT) sequencing platform (35, 36).
271 Although currently the PacBio has lower error rates compared to current ONT nanopore
272 sequencing, its utility might be limited by the high capital cost of the equipment. In contrast, ONT
273 nanopore sequencing platforms have very low capital costs, and they also offer a unique
274 opportunity to deliver results in real-time settings such as in the field (24, 37) or at point of care.

275 One of the issues reported with nanopore sequencing is the limitation in reliable detection of
276 small insertion and deletion (indel) events (32, 38), and we encountered several false positive
277 indels in homopolymer regions. However, true indels outside the homopolymer regions could be
278 detected accurately, including key indel variants associated with *CYP2D6* star (*) alleles such as
279 6750delA (*3), 5098delT (*6), 6816delAAG (*9), and 4337insT (*15). All samples with these
280 haplotypes could be identified accurately. In addition, other indel variants were also detected
281 including 4933delC, 7746insA, and 8857delACA which are associated with several star (*) alleles
282 including *2, *41, and *35.

283 The challenge that homopolymer regions present to nanopore sequencing also means that this
284 technology remains limited for investigating the two long homoplymer strings of A's in the
285 upstream region of the *CYP2D6* gene (2943-2964 and 3098-3107). The regions in between these
286 two strings are notoriously difficult to sequence, resulting in a gap in most PharmVar catalouged
287 alleles or subvariants. Variation in the number of A's within 2943-2964 has been reported
288 previously (36), however this region has not been included as part of the haplotype definition
289 (6). Although nanopore sequencing was not able to reliably detect variation in the poly A string,
290 the regions adjacent to the strings were sequenced accurately in this study, including the
291 detection of variants such as 2966A>G, which are still ambiguous in many star (*) alleles.

292 Sample barcoding permits efficient use of the flowcell and reduction in per sample assay costs.
293 For high throughput laboratories or for research studies on larger cohorts, sample number could
294 potentially be multiplexed up to 96 using the ONT barcode kit. Our 24-plex sequencing run
295 resulted in the minimum read depth of ~500x, which is more than adequate for accurate variant
296 calling and phasing, while some samples had tens of thousands of reads. The lowest read depth
297 in our sample cohort was ~200x, and we observed no difference in accuracy compared to samples
298 with higher read depth. Using nanopore sequencing to sequence a long amplicon from the *GBA*
299 gene, Leija-Salazar et al reported that coverage >100x is sufficient for accurate variant calling
300 (32).

301 The main advantage of long read sequencing is enabling straightforward variant phasing. In this
302 study, all variants could be phased accurately (provided more than one heterozygote variant is
303 present), resulting in highly confident haplotyping to the subvariant level. In addition, duplicated
304 alleles when present, were also identified. Using the 8-kb PCR amplicon derived from duplicated
305 alleles as a confirmation sequence, we showed that duplicated alleles could be detected from the
306 6-kb *CYP2D6* amplicon by the skewed allele balance of reads. The allele ratio could be observed
307 from the VCF file or by visualizing the phased BAM file on IGV. All our duplication samples have
308 an allele ratio of ~2:1, indicating two copies of the duplicated gene. Due to limitation of the
309 samples used in this cohort, it remains inconclusive whether this method could determine the
310 exact number of gene copies accurately in cases with more than two copies.

311 Nanopore based long read *CYP2D6* amplicon sequencing is an efficient tool for detecting key
312 alleles and haplotypes and challenges that may arise owing to gene deletion and conversion are
313 yet to be addressed using this method. The whole gene deletion allele (*5) still requires detection
314 by a separate duplex PCR, and despite the accurate determination of haplotypes and duplication,
315 the methods described here have not been extended to the detection of other CYP2D6 structural
316 variations such as the CYP2D6-CYP2D7 hybrids or tandem rearrangement.

317

318 Conclusion

319 In summary, nanopore sequencing of the *CYP2D6* gene allows accurate detection and phasing of
320 variants and determination of sample haplotypes to the subvariant level. New variants not yet
321 catalogued on the PharmVar CYP2D6 page (rs1080992, rs376217512, rs143170489,
322 rs61736504, rs557722765) were also reported in several samples, and key indel variants could
323 be reliably detected. The ability of nanopore sequencing to detect duplicated alleles when present
324 provides better prediction of CYP2D6 phenotype. In addition, sample barcoding permits
325 sequencing and analysis of multiple samples simultaneously, improving cost and time efficiency.

References

1. Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacology and Therapeutics. 2007;116(3):496-526.
2. Gaedigk A, Dinh JC, Jeong H, Prasad B, Leeder JS. Ten years' experience with the CYP2D6 activity score: A perspective on future investigations to improve clinical predictions for precision therapeutics. Journal of Personalized Medicine. 2018;8(2):1-15.
3. Gaedigk A. Complexities of CYP2D6 gene analysis and interpretation. International Review of Psychiatry. 2013;25(5):534-53.
4. Yang Y, Botton MR, Scott ER, Scott SA. Sequencing the CYP2D6 gene: From variant allele discovery to clinical pharmacogenetic testing. Pharmacogenomics. 2017;18(7):673-85.
5. Nofziger C, Paulmichl M. Accurately genotyping CYP2D6 : not for the faint of heart. Pharmacogenomics. 2018;19(13):999-1002.
6. Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, Klein TE. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. Clinical Pharmacology and Therapeutics. 2018;103(3):399-401.
7. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285-91.
8. Kimura S, Umeno M, Skodaj RC, Meyert UA, Gonzalez FJ. Sequence and Identification of the Polymorphic CYP2D6 Gene , a Related Gene , and a Pseudogene. Am J Hum Genet. 1989;45:889-904.
9. Riffel AK, Dehghani M, Hartshorne T, Floyd KC, Leeder JS, Rosenblatt KP, et al. CYP2D7 sequence variation interferes with TaqMan CYP2D6*15 and *35 genotyping. Frontiers in Pharmacology. 2016;6(JAN):1-10.
10. Sachse C, Brockmöller J, Bauer S, Roots I. Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. American journal of human genetics. 1997;60(2):284-95.
11. Rebsamen MC, Desmeules J, Daali Y, Chiappe A, Diemand A, Rey C, et al. The AmpliChip CYP450 test: Cytochrome P450 2D6 genotype assessment and phenotype prediction. Pharmacogenomics Journal. 2009;9(1):34-41.
12. Chua EW, Cree SL, Ton KNT, Lehnert K, Shepherd P, Helsby N, et al. Cross-comparison of exome analysis, next-generation sequencing of amplicons, and the iPLEX® ADME PGx panel for pharmacogenomic profiling. Frontiers in Pharmacology. 2016;7(JAN):1.
13. 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. Journal of Molecular Diagnostics. 2015;17(6):740-8.
14. Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Research. 2015;4(0):17.
15. Pratt VM, Everts RE, Aggarwal P, Beyer BN, Broeckel U, Epstein-Baak R, et al. Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. Journal of Molecular Diagnostics. 2016;18(1):109-23.
16. Maggo SDS, Chua EW, Chin P, Cree S, Pearson J, Doogue M, et al. A New Zealand platform to enable genetic investigation of adverse drug reactions. New Zealand Medical Journal. 2017;130(1466):62-9.
17. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Research. 1988;16(3):1215-.
18. Gaedigk A, Ndjountché L, Divakaran K, Dianne Bradford L, Zineh I, Oberlander TF, et al. Cytochrome P4502D6 (CYP2D6) gene locus heterogeneity: Characterization of gene duplication events. Clinical Pharmacology and Therapeutics. 2007;81(2):242-51.
19. Steen V, Andreassen O, Daly A, Tefre T, Borresen A, Idl eJ, et al. Detection of the poor metabolizer-associated CYP2D6(D) gene deletion allele by long-PCR technology. Pharmacogenetics 1995;5(4):214-23.

20. Nagar R, Schewessinger B. DNA size selection (>3-4kb) and purification of DNA using an improved homemade SPRI beads solution.2018.
21. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: Visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666-9.
22. Li H. Minimap2: pairwise alignment for nucleotide sequences. 2017;34(May):3094-100.
23. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods. 2018;15(6):461-8.
24. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228-32.
25. Luo R, Sedlazeck FJ, Lam T-W, Schatz M. Clairvoyante: a multi-task convolutional deep neural network for variant calling in Single Molecule Sequencing. bioRxiv. 2018;310458.
26. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Gunnar W, et al. WhatsHap : fast and accurate read-based phasing. bioRxiv 2016;085050:1-18.
27. Leggett RM, Heavens D, Caccamo M, Clark MD, Davey RP. NanoOK: Multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. Bioinformatics. 2015;32(1):142-4.
28. Lanfear R, Schalamun M, Kainer D, Wang W, Schwessinger B. MinIONQC: fast and simple quality control for MinION sequencing data. Bioinformatics. 2018(July):1-3.
29. Robinson JT, Thorvaldsdoottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol. 2011;29(1):24-6.
30. McLaren W, Gil L, Hunt S, Riat H, Ritchie G, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17(1):122.
31. He X, He N, Ren L, Ouyang Y, Zhang N, Ma Y, et al. Genetic polymorphisms analysis of CYP2D6 in the Uygur population. BMC Genomics. 2016;17(1):1-9.
32. Leija-Salazar M, Sedlazeck FJ, Mokretar K, Mullin S, Toffoli M, Athanasopoulou M, et al. Detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. bioRxiv. 2018(November 2018):288068-.
33. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nature Methods. 2015;12(4):351-6.
34. Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, et al. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Scientific Reports. 2018;8(1):1-12.
35. Qiao W, Yang Y, Sebra R, Mendiratta G, Gaedigk A, Desnick RJ, et al. Long-read single-molecule real-time (SMRT) full gene sequencing of cytochrome P450-2D6 (CYP2D6) HHS Public Access. Hum Mutat. 2016;37(3):315-23.
36. Buermans HPJ, Vossen RHAM, Anvar SY, Allard WG, Guchelaar HJ, White SJ, et al. Flexible and Scalable Full-Length CYP2D6 Long Amplicon PacBio Sequencing. Human Mutation. 2017;38(3):310-6.
37. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protocols. 2017;12:1261-76.
38. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Briefings in Bioinformatics. 2017(June).
39. Wright GEB, Niehaus DJH, Drögemöller BI, Koen L, Gaedigk A, Warnich L. Elucidation of CYP2D6 genetic diversity in a unique African population: Implications for the future application of pharmacogenetics in the xhosa population. Annals of Human Genetics. 2010;74(4):340-50.

Table 1. Metrics of first and second CYP2D6 sequencing runs

No.	Parameters	1st run	2nd run
1.	Total yield	3.24M reads (3.63 Gbases)	748K reads (3.62 Gbases)
2.	Read length	N50: 2955 bps Median: 307 bps	N50: 6741 bps Median: 6437 bps
3.	Number of barcodes	14	26
4.	Number of barcoded reads	690K reads (47.6% of total pass reads)	242K reads (46% of total pass reads)
6.	Number of barcoded reads with length > 6kb	57K reads (3.96% of total pass reads)	186K reads (35.4% of total pass reads)
7.	Number of reads per barcode (length > 6 kb and qscore > 7)	235-11832	555-21853
8.	Mean qscore[a]	6.1	8.1
9.	Error rates (subs – del – ins)[b]	6.53% - 4.56% - 5.13%	4.99% - 3.97% - 3.78%

a Mean qscores were obtained from MinION QC
b Error rates were obtained from NanoOK

Table 2. Comparison between mapping and variant calling tools

Mapping tool	Variant Calling Tool	TP Variants (median[range])	FN Variants (median[range])	FP Variants (median[range])	PPV (%)	FDR (%)	Sensitivity (%)
minimap2	nanopolish	27 [1-32]	1 [0-4]	6 [5-7]	79.12	20.87	96.43
minimap2	Clairvoyante	23 [1-33]	1 [0-5]	17 [10-19]	59.18	40.82	95.83
NGMLR	nanopolish	25 [1-30]	3 [0-6]	11 [6-13]	66.52	33.48	89.29
NGMLR	Clairvoyante	26 [1-30]	2 [0-4]	19 [12-21]	54.55	45.45	92.86

Abbreviation: TP: True Positive, FN: False Negative, FP: False Postive, PPV: Positive Predictive Value, FDR: False Discovery Rate. PPV, FDR, and Sensitivity were calculated based on median values.

Table 3 Genotyping Results

No.	Sample	Sequencing run	Reference Genotype (Get-RM or Sanger)	Allele	Genotype (GridION)	No. of Variants		
						subs	del	ins
Reference Samples								
1	NA07439	run 1	*4XN/*41	hap1	*4.004XN	15	0	0
				hap2	*41.003	23[a]	1	0
		run 2	*4XN/*41	hap1	*4.004XN	14[b]	0	0
				hap2	*41.003	22[a,b]	1	0
2	NA23348	run 1	*7/*35	hap1	*35.001	23[b]	1	0
				hap2	*7.001	1	0	0
		run 2	*7/*35	hap1	*7.001	1	0	0
				hap2	*35.001	23[b]	1	0
3	NA18518	run 1	*17/*29	hap1	*17.001	20[b]	0	0
				hap2	*29.001	17[c]	1	0
4	NA21781	run 1	*2/*4(XN)	hap1	*2.001XN	22[b]	1	0
				hap2	*4.001	14[b,d]	0	0
5	NA12878	run 1	*3/*4	hap1	*3.001	0	1	0
				hap2	*4.001	17[b]	0	0
6	NA23405	run 1	*1/*7	hap1	*7.001	1	0	0
				hap2	*1.001	0	0	0
7	NA23297	run 1	*10(XN) / *17	hap1	*17.001	20[b]	0	0
				hap2	*10.002XN	13[b]	0	0
Clinical Samples								
8	S001	run 1	*1/*1	hap1	*1.001	0	0	0
				hap2	*1	1	0	0
		run 2	*1/*1	hap 1	*1	1	0	0
				hap 2	*1.001	0	0	0
9	S002	run 1	*6/*4	hap1	*6.005	2	1	0
				hap2	*4.001	17[b]	0	0
		run 2	*6/*4	hap 1	*6.005	2	1	0
				hap 2	*4.001	17[b]	0	0
10	S003	run 1	*1/*3	hap1	new allele	9	0	1
				hap2	*3.001	0	1	0
11	S004	run 1	*4/*9	hap1	*4.001	17[b]	0	0
				hap2	*9.001	0	1	0
12	S005	run 1	*2/*10	hap1	*2.001	22[b]	1	0
				hap2	*10.002	13[b]	0	0
13	S006	run 2	*2/*9	hap 1	*2.001	22[b]	1	0
				hap 2	*9.001	0	1	0
14	S007	run 2	*71/*41	hap 1	*71	2	0	0
				hap 2	*41.001	22[b]	1	0

15	S008	run 2	*15/*2	hap 1	*15.001	1	0	1
				hap 2	*2.001	22[b]	1	0
16	S009	run 2	*1/*6	hap 1	*6.005	2	1	0
				hap 2	*1	1[b]	0	0
17	S010	run 2	*4/*10	hap 1	*10.002	13[b]	0	0
				hap 2	*4.015	18[b]	0	0
18	S011	run 2	*2/*2	hap 1	*2.001	22[b]	1	0
				hap 2	*2.001	22[b]	1	0
19	S012	run 2	*4/*4	hap 1	*4.001	17[b]	0	0
				hap 2	*4.001	17[b]	0	0
20	S013	run 2	*2/*41	hap 1	*41.001	23	1	0
				hap 2	*2.001	23	1	0
21	S014	run 2	*4/*28	hap 1	*28.001	25	1	0
				hap 2	*4.015	19	0	0
22	S015	run 2	*4/*35	hap 1	*4.001	18	0	0
				hap 2	*35.001	24	1	0
23	S016	run 2	*10/*1	hap1	*1	1	0	0
				hap2	*10.002	14	0	0
24	S017	run 2	*35/*41	hap1	*41.001	22[b]	1	0
				hap2	*35.001	23[b]	1	0
25	S018	run 2	*1/*1	hap1	*1.016XN	1[b]	0	0
				hap2	*1.012	2	0	0
26	S019	run 2	*1/*1	hap1	*1.012	1	0	0
				hap2	*1.001	0	0	0
27	S020	run 2	*1/*1	hap1	*1	1	0	0
				hap2	*1.001	0	0	0
28	S021	run 2	*1/*1	hap1	*1	1	0	0
				hap2	*1.001	0	0	0
29	S022	run 2	*5/*33	hap1	*5.001			
				hap2	*33.001	1	0	0
30	S023	run 2	*4/*4	hap1	*4.001	17[b]	0	0
				hap2	*4	18[b]	0	0
31	S024	run 2	*3/*4	hap1	*3.001	0	1	0
				hap2	*4.001	17[b]	0	0
32	S025	run 2	*4/*41	hap1	*41.001	22[b]	1	0
				hap2	*4.001	16[e]	0	0

a. Variant 7189G>A was not called in the vcf file but visually confirmed on Bam file.
b. Variant 5042T>G was incorrectly called as delT
c. Variant 6416A>G was not called in the vcf file but visually confirmed on Bam file.
d. Variants 2775C>T, 3201G>A and 4300C>T were not called in the vcf file but visually confirmed on Bam file.
e. Variant 2775C>T was not called in the vcf file but visually confirmed on Bam file.

Table 4. Variants in duplicated allele in 6.6kb and 8kb fragments.

No.	Sample	Variants found in duplicated allele in 6.6 kb amplicon (from position 3860 – 8593)	Variants found in 8.6 kb amplicon (from position 3860 – 8593)
1.	NA07439 (*4XN) (run 1)	4300C>T, 4510G>T, 5042T>G, 5238C>T, 5862G>C, 6047G>A, 6298A>G, 7585A>C, 7783A>G, 8381G>C, 8602C>T, 8923T>G	4300C>T, 4510G>T, 5238C>T, 5862G>C, 6047G>A, 6298A>G, 7585A>C, 7783A>G, 8381G>C, 8602C>T, 8923T>G *
2.	NA23297 (*10XN)	4300C>T, 4510G>T, 5238C>T, 5862G>C, 6298A>G, 7585A>C, 7783A>G, 8381G>C, 8602C>T, 8923T>G*	4300C>T, 4510G>T, 5238C>T, 5862G>C, 6298A>G, 7585A>C, 7783A>G, 8381G>C, 8602C>T, 8923T>G*
3.	NA07439 (*4XN) (run 2)	4300C>T, 4510G>T, 5238C>T, 5862G>C, 6047G>A, 6298A>G, 7585A>C, 7783A>G, 8381G>C, 8602C>T, 8923T>G *	4300C>T, 4510G>T, 5238C>T, 5862G>C, 6047G>A, 6298A>G, 7585A>C, 7783A>G, 8381G>C, 8602C>T, 8923T>G
4.	S018 (*1XN)	5369G>A*	5369G>A*

*5042 missed by the variant calling tool

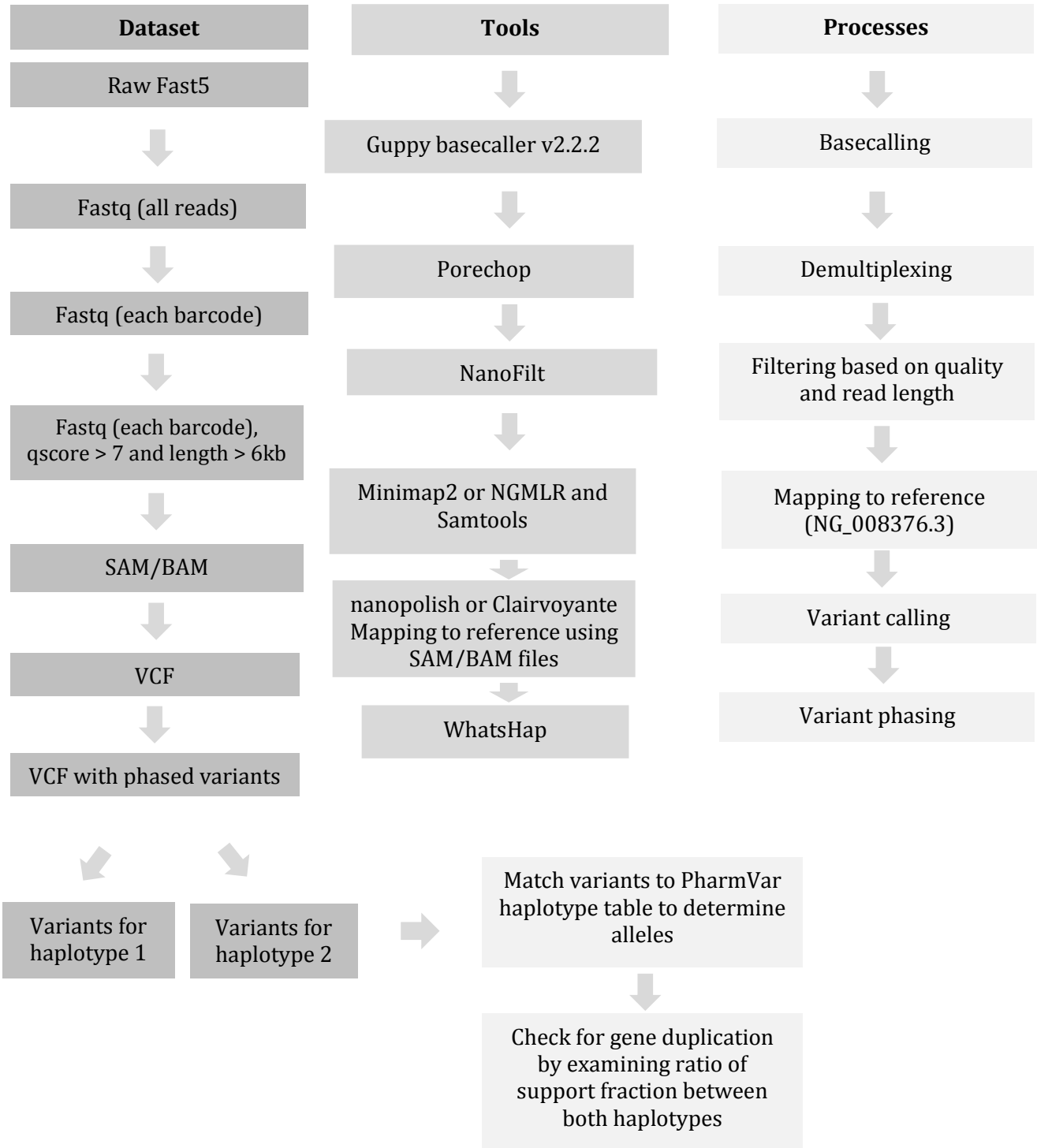


Figure 1. Bioinformatics workflow for CYP2D6 nanopore sequencing

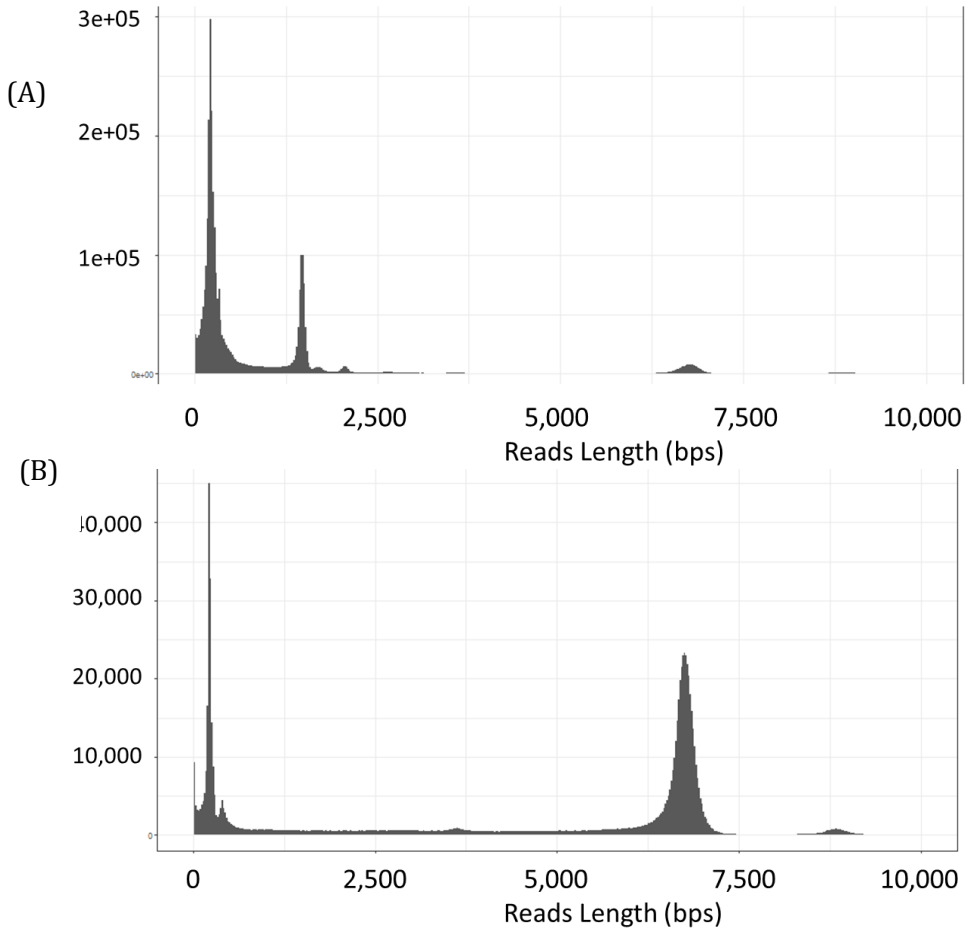
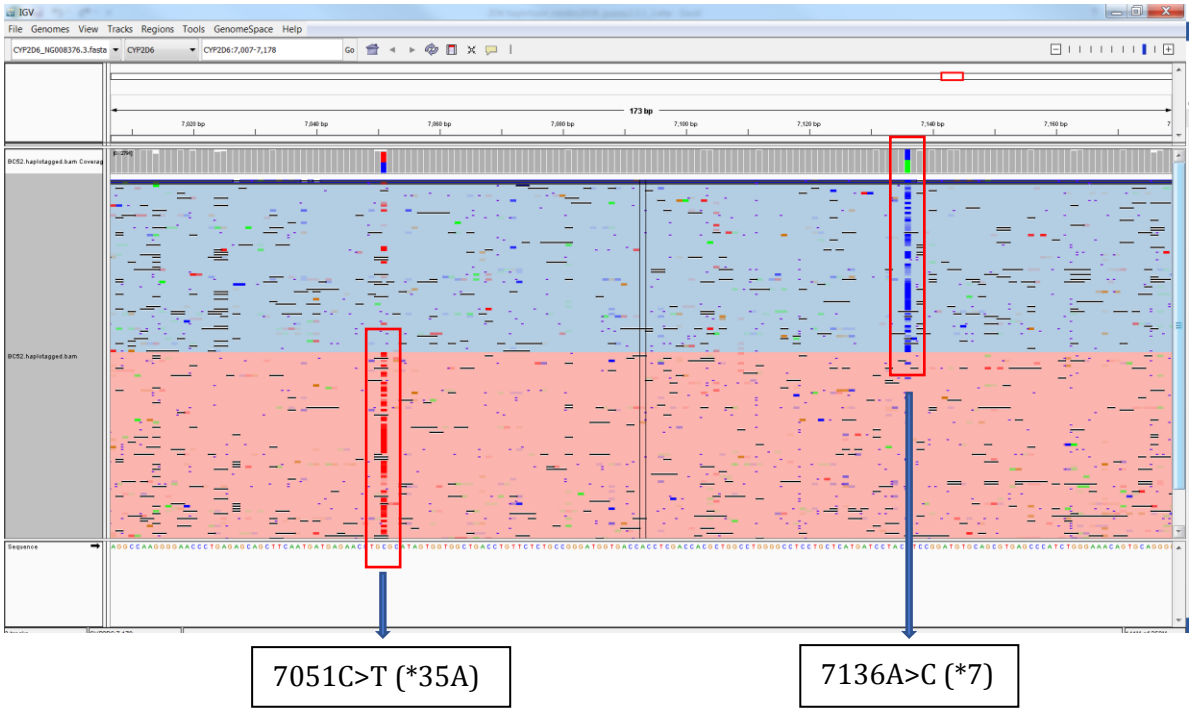


Figure 2. Read Length distribution of (A) first run, and (B) second run of CYP2D6 sequencing

(A)



(B)

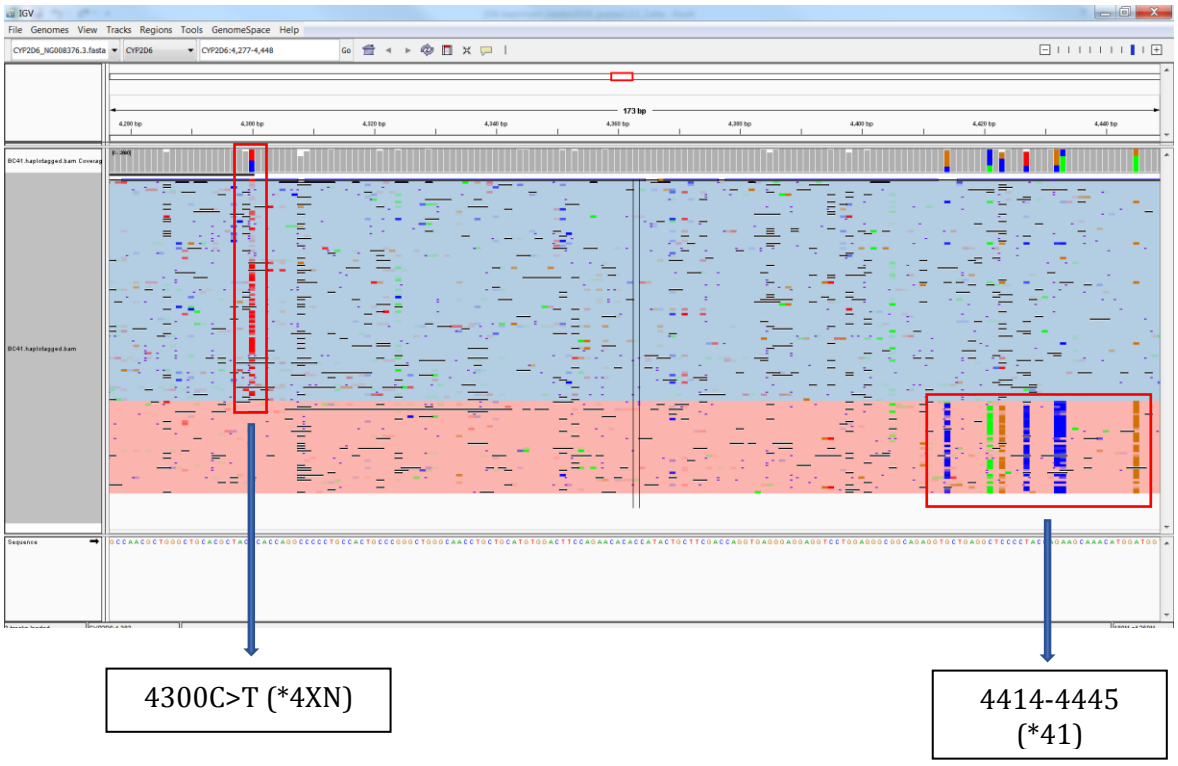
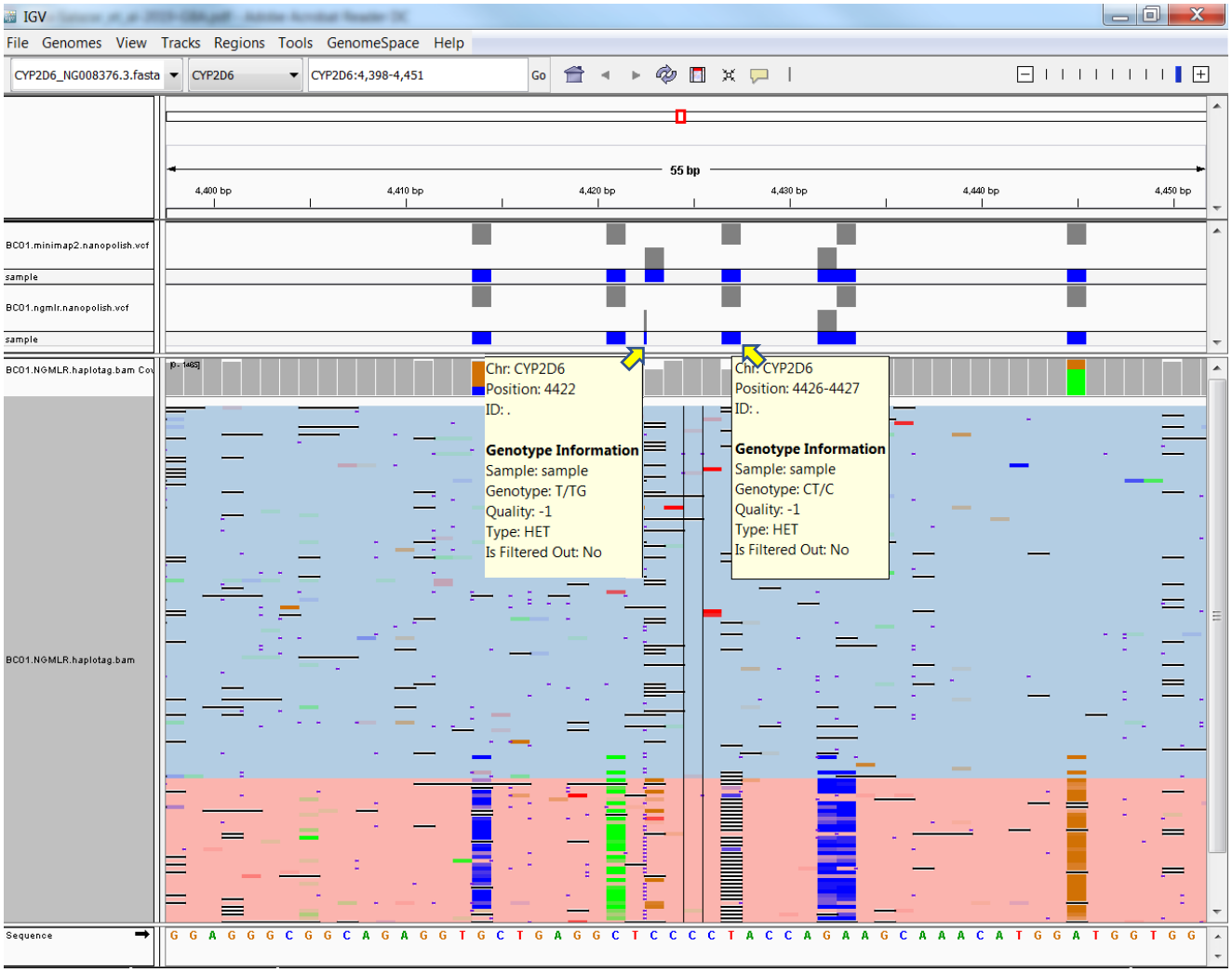


*Figure 3. Phased variants for (A) sample with no duplicated allele (NA23348, *7/*35A) and (B) sample with duplicated allele (NA07439, *4XN/*41). Colours represent each of the phased alleles of the samples.*



Supplementary Figure 1. Misdetection of two out of the seven variants of intron 1 CYP2D7 conversion due to misalignment by NGMLR. The two variants were called incorrectly as insertion and deletion (4423insG and 4427delT) instead of the two variants of 4423C>G and 4427T>C.

Supplementary Table 1 Primer sequences

No.	Primer name	Sequence (5'-3')	Reference
Primers for long PCR			
1.	6.6kbF	tttctgttggtgctgatattgcATG GCA GCT GCC ATA CAA TCC ACC TG	(18)
2.	6.6kbR	acttgcctgtcgctctatcttcCGA CTG AGC CCT GGG AGG TAG GTA G	(18)
3.	FragD_8.6kbF	tttctgttggtgctgatattgcCCAGAAGGCTTTGCAGGCTTCAG	(18)
4.	FragD_8.6kbR	acttgcctgtcgctctatcttcCGGCAGTGGTCAGCTAATGAC	(18)
5.	CYP-13	ACC GGG CAC CTG TAC TCC TCA	(19)
6.	CYP-24	GCA TGA GCT AAG GCA CCC AGA C	(19)
7.	FragB_Dup_F	CCA TGG AAG CCC AGG ACT GAG C	(18)
8.	FragB_Dup_R	CGG CAG TGG TCA GCT AAT GAC	(18)
Primers for Sanger Sequencing			
1.	2D6_5UTRF	GCAGCTGCCATACAATCCAC	
2.	2D6_5UTRF	TGCTTTCTGGCCTCCATGTT	
3.	2D6Prom_R	TCCTCCATAACGTTCCCACCAGAT	(39)
4.	2D6Prom_F	CCATACCTGCCTCACTACCAAATG	(39)
5.	2D6Ex1F	TCTGGAGCAGCCCATACCCG	(39)
6.	2D6Ex1SR	CCCCAGACTACAGGTCCTAGTCCTATTTG	(12)
7.	2D6Ex2F	TCCTCCTTCCACCTGCTCAC	(39)
8.	2D6Ex2SR	CTTTGCCCCACCTCGTCTCT	(39)
9.	2D6Ex34F	AGCTGGAATCCGGTGTCGAA	(39)
10.	2D6Ex34SR	AGCCATCTCCAGGTAGACCCAG	(39)
11.	2D6Ex56F	ACAGGCAGGCCCTGGGTCTA	(12)
12.	2D6Ex56SR	CCTGGTCACCCATCTCTGGTC	(39)
13.	2D6Ex7F	CCAACATAGGAGGCAAGAAG	(39)
14.	2D6Ex7R	ACTGGACTCTAGGATGCTGG	(39)
15.	2D6Ex8F	CAGAATGTTGGAGGACCCAA	(39)
16.	2D6Ex8R	AGGAAAGCAAAGACACCATG	(39)
17.	2D6Ex9F	TGAAGGATGAGGCCGTCTGG	(39)
18.	2D6-R	ACTGAGCCCTGGGAGGTAGGTAG	(39)

Supplementary Table 2. Sequence variants to detect star (*) alleles in Sanger Sequencing

Star (*) Alleles	**Variants**	
	RSID	Position based on NG_008376.3 (1=sequence start)
***2**	rs16947	7051C>T
***3**	rs35742686	6750delA
***4**	rs1065852, rs3892097	4300C>T, 6047G>A,
***6**	rs5030655	5098delT
***7**	rs5030867	7136A>C
***9**	rs5030656	6816delAAG
***10**	rs1065852	4300C>T
***15**	rs774671100, rs1081004	4337insT, 5369G>A
***17**	rs28371706	5222C>A/T
***28**	rs72549358, rs78482768	4219G>A, 5905C>G
***29**	rs61736512, rs59421388	5860G>A, 7384G>A
***33**	rs28371717	6684G>T
***35**	rs769258	4231G>A
***41**	rs28371725	7189G>A
***71**	rs118203758	4325G>A

Supplementary File 1

Protocol for DNA extraction from blood samples

For each patient sample, 4 mL of peripheral blood will be mixed with 6 mL of red blood cell lysis buffer (1 mM EDTA) and spun at 2000 rcf for 10 minutes at room temperature. The supernatant will be carefully discarded and the pellet resuspended and the initial step repeated until a layer of white buffy cells is clearly visible. The supernatant can then be discarded and 1 mL of 1 M NaCl solution added. This solution is then repeatedly vortexed to eliminate any clumping. Once the solution is clear of clumps, 6 mL of white cell lysis solution (10 mM Tris HCl, 26 mM EDTA, 17.3 mM (0.5%) SDS) and 50 μL of RNase A solution is added and incubated at 37°C for at least 1 hour.

After incubation 1mL of 3M sodium acetate solution is added and mixed. The resulting solution is then added to a 15 mL tube containing 2.5 gr of high vacuum grease (Dow Corning Corporation, USA) and then 1 ml of phenol:chloroform:isoanyl alcohol 25:24:1 was added. The tubes are then capped and thoroughly mixed before centrifuging for 5 minutes at 1500 rcf. Five mL of the resultant supernatant is then transferred to a new tube containing 5mL of 100% isopropanol. The tubes are then gently inverted until DNA precipitates. The tubes containing DNA are then centrifuged at 2000 rcf for 5 minutes to form a DNA pellet at the bottom of the tube. The isopropanol is then carefully discarded and then washed with 10 mL of 70% ethanol (10 mL) by centrifuging at 2000 rcf for 3 minutes. After discarding the ethanol, the DNA pellet is dissolved in 500 mL of TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) at 37°C for at least 8 hours. The DNA solution were quantified using a spectrophotometer at 260 nm and diluted to 50 ng/μL.

EDTA: Ethylenediaminetetraacetic acid

TRIS HCL: tris(hydroxymethyl)aminomethane hydrochloric acid

SDS: Sodium dodecyl sulfate

TE buffer: Tris-EDTA buffer

NaCl: Sodium chloride