An end-to-end workflow for pharmacogenomic target enrichment, star allele annotation and *CYP2D6* resolution using nanopore sequencing

Carly Tyer[1], Lynn Ly[1], Sergey Nurk[2], Nikolai Karpov[3], Rocio Esteban[2], Hayley Greenfield[2], Ashwini Singh[2], Scott Hickey[1], Philipp Rescheneder[2], Sissel Juul[3]

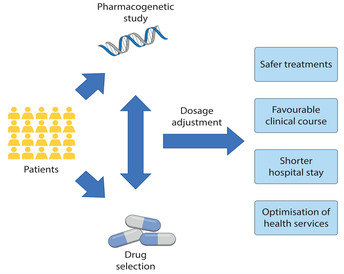
[1]Oxford Nanopore Technologies, Inc. San Francisco, CA
[2]Oxford Nanopore Technologies, Plc. Oxford, UK
[3]Oxford Nanopore Technologies, Inc. New York City, NY

Oxford Nanopore Technologies APPLICATIONS

Abstract

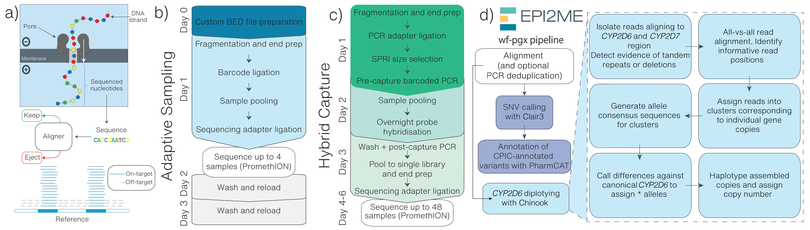
We present a targeted end-to-end pharmacogenomics workflow encompassing enrichment, sequencing, and star allele annotation of 18 genes with CPIC guidelines, including *CYP2D6*. Two scalable targeting approaches were used: 1) adaptive sampling (AS), an amplification-free software-based enrichment approach that preserves native DNA methylation and yields longer on-target reads, and 2) hybrid capture. We sequenced 94 patient-derived cell lines using these methods to provide sufficient coverage for accurate haplotyped variant calling and resolution of complex allelic structures. To resolve the *CYP2D6* locus, the wf-pgx EPI2ME pipeline incorporates Chinook for network clustering and reference-based assembly to reconstruct individual gene copies. We compare wf-pgx results to Genetic Testing Reference Material Coordination Program (GeT-RM) genotypes, demonstrating concordance with or improvement over 99% of PharmCAT calls and 94% of *CYP2D6* genotypes, including improvements to reference calls in 12 cases. Additionally, AS methylation analysis shows variable methylation at key CpGs that affect drug metabolism independently from genomic sequence. Both approaches offer scalable solutions for high-throughput PGx analysis that may inform future reference standards and phenotypic research.

1. Introduction



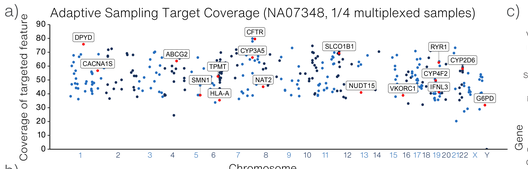
Pharmacogenetic (PGx) variation affects individual metabolism by changing the efficacy of uptake, signal transduction, and breakdown of pharmaceutical compounds. Such variation informs clinical guidance on prescriptions and dosage based on genetic makeup and it is critical to identify variants that impact enzymatic activity in metabolizing a drug. *CYP2D6*, which affects the metabolism of 20-25% of commonly prescribed drugs, is challenging to resolve due to its high homology to neighboring pseudogenes, high polymorphism, and structural complexity. Long reads are required to discriminate between genes and pseudogenes of high similarity, or those found in challenging regions of the genome. This study aims to address these challenges using Oxford Nanopore sequencing. We sequenced 94 patient-derived cell lines, including those with previously characterized allelic variants, to evaluate the efficacy of this approach in resolving complex *CYP2D6* alleles and other PGx genes.

2. Materials and Methods



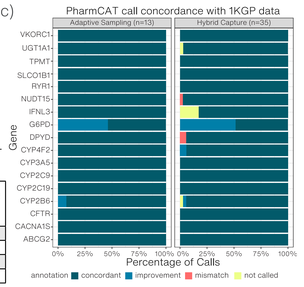
Adaptive sampling (AS) is a rapid and adaptable software-based approach to enrich regions of interest by depleting off-target regions during sequencing itself, with no upfront sample manipulation (Fig. a). During sequencing, strands are basecalled and mapped to a reference genome in real time. Strands that align outside of the target regions within the first 400 base pairs are ejected, while strands that are on target are allowed to sequence to completion. The adaptive sampling workflow, which targets 375+ genes specified in an easily configurable BED file, allows for sequencing of up to four samples on one PromethION™ Flow Cell (Fig. b). The second method is a hybrid capture workflow using the Twist Alliance Long-Read PGx panel targets, which allows scalable high-coverage sequencing of 49 clinically implicated genes for up to 48 samples on one PromethION Flow Cell (Fig. c). Starting with high accuracy (HAC) FASTQ or BAM data, we used the EPI2ME wf-pgx pipeline to call and annotate SNVs with PharmCAT[1] and to diplotype *CYP2D6* with Chinook (Fig. d). The pipeline is compatible with both hybrid capture and adaptive sampling data.

3. Pharmacogene enrichment and star allele concordance



b)

Method	Samples Multiplexed	Genes Targeted	Flow Cell	Average Coverage	Average Number of Reads	Average N50	Average GB Yield
Adaptive Sampling	4	375	1	73x	463,838	10,519	3.22
Adaptive Sampling	4	375	2	68x	539,943	6,933	3.02
Hybrid Capture	48	50	1	294x	188,744	4,040	0.71
Hybrid Capture	48	50	2	361x	245,170	4,090	0.93



When sequencing up to four samples, the adaptive sampling workflow results in high, on-target capture efficiencies across all targets, averaging 71x per sample (Fig. a). When sequencing up to 48 samples, the hybrid capture workflow results in even higher coverage, averaging 328x per sample. Long reads are necessary to distinguish between genes and pseudogenes of high similarity and other challenging regions of pharmacogenes. The adaptive sampling workflow results in N50s of ~8 kb, and the hybrid capture workflow results in read N50s over 4 kb, exceeding typical hybrid capture approaches (Fig. b). We performed star (*) allele calling with PharmCAT on Nanopore VCFs from wf-pgx and Illumina VCFs from the 1000 Genomes Project[3] (1KGP). We compared these calls to each other and, when available, GeT-RM calls. 99% of star allele calls agreed with 1KGP's calls or improved them due to better phasing (Fig. c). Some genes were not called due to biotin probe dropout. *CYP4F2* had two disagreements where Nanopore and GeT-RM both called *1/*1, but 1KGP falsely called *1/*17 (Fig. 5b). We also re-evaluated short-read genotypes using stricter filtering, but that introduced additional short-read false negatives like in *CYP2B6* (Fig. 5a).

4. Chinook, a tool for *CYP2D6* locus reconstruction

a)

Hybrid Capture - Agree					
Sample	Chinook	Sample	Chinook	Sample	Chinook
HG00156	*5/*5	NA17039	*2/*17	NA18980	*2/*36+*10
HG00276	*4/*5	NA17102	*1/*40	NA18992	*1/*5
HG00373	*2/*2	NA17113	*17x2/*45	NA19109	*2x2/*28
HG00421	*2/*10x2	NA17114	*1/*5	NA19122	*2x2/*17
HG00423	*10/*10x2	NA17115	*1/*2	NA19143	*10/*45
HG00436	*2x2/*71	NA17204	*1/*35	NA19174	*4/*40
HG00589	*1/*21	NA17227	*1/*9	NA19178	*1/*1
HG01086	*1/*31	NA17230	*4/*41	NA19207	*2x2/*10
HG01190	*68+*4/*5	NA17234	*1/*41	NA19226	*2/*2x2
HG02373	*36+*10/*14	NA17237	*1/*4	NA19239	*15/*17
HG03225	*5/*56	NA17240	*1/*15	NA19785	*1/*15+*2
NA02016	*2x2/*17	NA17244	*2x2/*4x2	NA19819	*2/*4x2
NA06993	*4/*35	NA17247	*1/*2	NA19920	*1/*4x2
NA07348	*1/*6	NA17269	*2/*41	NA21781	*2x2/*68+*4
NA07439	*4x2/*41	NA17287	*1/*83	NA23093	*1/*36+*10
NA10005	*17/*29	NA17288	*1/*1	NA23296	*2x2/*4
NA10847	*1/*41	NA17290	*1/*41	NA23297	*10x2/*17
NA10860	*1/*4N+*4	NA17454	*1x2/*2x2	NA24027	*2x2/*6
NA11832	*1/*68+*4	NA18518	*17/*29	NA24217	*2/*4x3
NA12154	*68+*4/*33	NA18540	*36+*10/*41		
NA12244	*35/*41	NA18552	*1/*14		

b)

Adaptive Sampling - Agree	
HG00276	*4/*5
HG01190	*68+*4/*5
NA07348	*1/*6
NA11832	*1/*68+*4
NA18518	*17/*29
NA19174	*4/*40
NA19207	*2,*2,*10
NA19226	*2/*2x2

Allele calls match GeT-RM
Allele calls improve upon GeT-RM
Allele calls disagree with GeT-RM

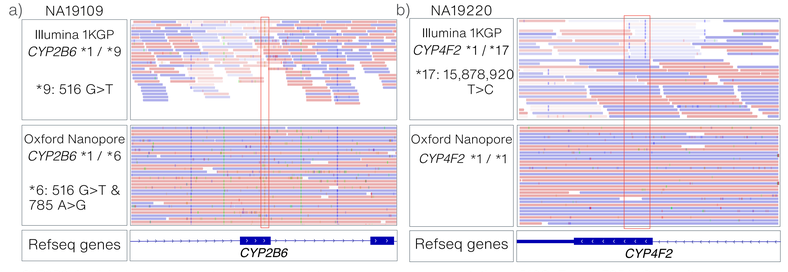
c)

Hybrid Capture - Improve GeT-RM		
Sample	Chinook	GeT-RM
NA02016	*2x2/*17	*2XN/*17
NA16688	*2/*36+*10	*2/*10
NA17019	*1/*36+*10	*1/*10
NA17058	*10/*36+*10	*10/*10
NA17084	*1/*36+*10	*1/*10
NA17203	*68+*4/*35	*4/*35
NA17226	*4/*4N+*4	*4/*4
NA17245	*68+*4/*35	*2/*4
NA17264	*1/*68+*4	*1/*4
NA23873	*1/*68+*4	*1/*4
NA23874	*4/*68+*4	*4/*4
NA24008	*1/*68+*4	*1/*4

Hybrid Capture - Mismatch GeT-RM		
Sample	Chinook	GeT-RM
HG00463	*10/*36+*10	*36+*10/*36+*10
NA17232	*2/*35	*2/*2XN
NA18526	*1/*36+*10	*1/*36x2+*10
NA23246	*10x2/*36x3	*10x2/*36+*10

CYP2D6 is one of the most technically challenging pharmacogenes to resolve and has been implicated as a key factor in the metabolism of numerous drugs. *CYP2D6* is highly polymorphic, subject to copy number variation (including deletion of the entire gene), and its proximity to the *CYP2D7* pseudogene can give rise to hybrid copies. Accurate genotyping of the locus with conventional methods has been problematic. We compare Chinook v0.7.2 calls on 77 samples with available GeT-RM genotypes. Chinook was accurate in 94% of cases, with miscalls of the copy number in four cases – a limitation that will be addressed in future versions. Importantly, following manual investigation of read alignments, we believe that our calls improve upon GeT-RM genotypes for 12 samples, with most cases involving hybrid alleles or CNVs such as *68, *36 and *4N.

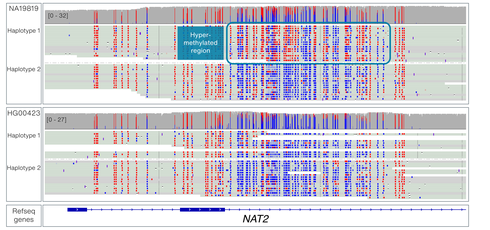
5. Long reads resolve short read mismapping errors



The *CYP2B6**6 allele is defined by the presence of two non-synonymous mutations, 516G>T and 785A>G, resulting in decreased function. In short read data, 785A>G falls in a low-mappability region due to the similarity between *CYP2B6* and the pseudogene *CYP2B7*. Alignments in this area have decreased MAPQ and some mismapping/clipping; the variant does not pass 1KGP's filters, leading to a miscall of *9. In contrast, Nanopore reads were confidently mapped and phased, matching GeT-RM's call of *6 (Fig. a). The definitive SNP for the *CYP4F2**17 allele (15,878,920 T>C) is located in a low-mappability region. The short read mappings revealed a set of reads that seemingly have the T>C variant, but those reads have a low MAPQ and may be mis-mapped from a neighboring region (MAF = 19%). In contrast, long read data, which is confidently mapped and phased, supports the reference T allele (MAF = 1%) (Fig. b).

6. Methylation calling provides additional insight for PGx

Because adaptive sampling enriches targets without amplification, reads retain per-base methylation information, enabling parallel analysis of sequence and epigenetic variation. Previous research has shown that variability in susceptibility to anti-tuberculosis drug-induced liver injury (ATDILI) is only partly explained by *NAT2* single-nucleotide polymorphisms; Jittikoon et al. found that hypermethylation at the *NAT2* promoter was associated with increased ATDILI occurrence in tuberculosis patients[4]. We performed haplotype-specific methylation analysis with modkit to find differentially methylated regions (DMR). One sample exhibited *NAT2* promoter hypermethylation, a risk factor for liver damage. Access to methylation data may help uncover more interactions between epigenetics and treatment outcomes.



Conclusion

We demonstrate that Oxford Nanopore sequencing enables accurate genotyping of pharmacogenetic variants. Targeted nanopore sequencing successfully enriches pharmacogenes, resulting in high coverage for haplotyped variant calling, including structural variations and resolution of complex allelic structures. The EPI2ME pipeline results in accurate star (*) allele calling with PharmCAT. For the challenging *CYP2D6* region, our custom Chinook tool results in accurate star (*) allele calling, providing improvements to the GeT-RM calls where hybrid alleles go undetected.

References

[1]Sangkuhl et al. Pharmacogenomics Clinical Annotation Tool (PharmCAT). *Clin Pharmacol Ther.* 2020 Jan;107(1):203-210.
[2]Pratt et al. Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. *J Mol Diagn.* 2016 Jan;18(1):109-23.
[3]Byrska-Bishop et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell.* 2022 Sep 1;185(18):3426-3440.e19.
[4]Jittikoon et al. Association of NAT2 promoter hypermethylation with susceptibility to hepatotoxicity due to antituberculosis drugs and biomarker potential. Sci Rep. 2025 Mar 25;15(1):10197. PMCID: PMC11937569.