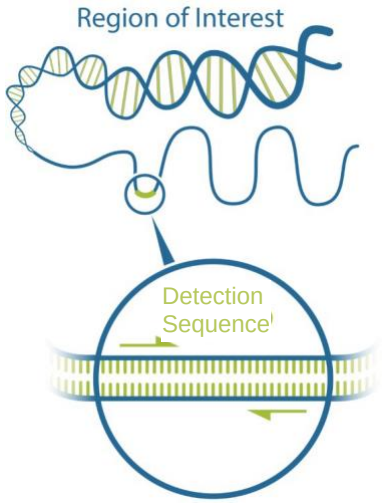
Phasing CYP2D6 with Xdrop™: distinguishing genes from pseudogenes

Background

CYP2D6 codes for cytochrome P450, an enzyme responsible for metabolizing or activating nearly 25% of all currently available drugs.[1] Its high polymorphism leads to considerable variability in enzyme activity among individuals[2] which influences their response to different drugs, including antipsychotics, beta-blockers, antidepressants, antihypertensives, antidiabetics and more. Genotyping this rather small gene (4.4 kb) has proven challenging based on short sequencing reads. The gene exhibits structural variations, like duplications and deletions, and hybrid gene conversions. It is also flanked by 2 genes/pseudogenes, CYP2D7 and CYP2D8, with over 90% sequence homology. [3,4] However, Xdrop™ enrichment followed by long read sequencing resolves the problem.

Xdrop™ enriches long (~100 kb) target DNA regions by amplifying a Detection Sequence corresponding to a small portion of the Region Of Interest (ROI) or in flanking regions. This amplicon is exclusively used to detect, select and enrich the full-length ROI, which is then captured and sequenced.



We evaluated Xdrop™ technology for the purpose of preparing samples for long-read sequencing to phase structurally complex genes. For that purpose, we used biobanked DNA (Coriell Institute, NA23348) previously genotyped as CYP2D6*7/CYP2D6*35A.[5]

The Xdrop™ Technology

The Xdrop™ technology combines high-resolution droplet PCR (dPCR) with droplet sorting and Multiple Displacement Amplification in droplets (dMDA).

Firstly, Xdrop™ partitions the DNA into millions of double emulsion droplets. Droplets containing the target DNA molecules are identified by a 120-160 bp targeted dPCR specific to a Detection Sequence within or adjacent to the region of interest.

The detection and sorting of droplets are performed using a standard cell sorter, which allows the PCR positive droplets containing the ROI to be collected. The sorted long DNA fragments are finally amplified in droplets (dMDA) to ensure unbiased DNA amplification.

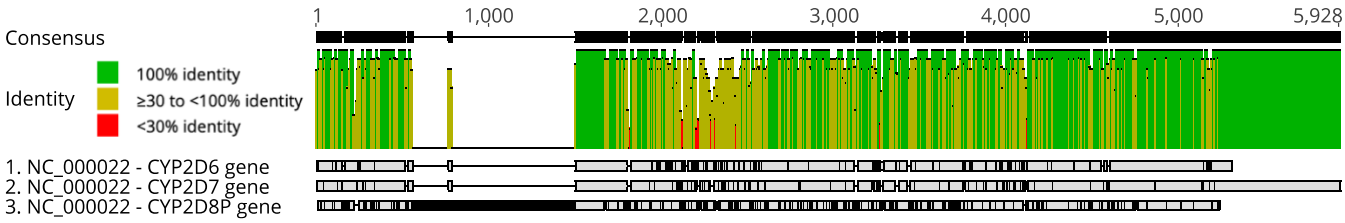
The Xdrop™ enrichment and amplification technology is compatible with both long- and short-read library preparation and sequencing.

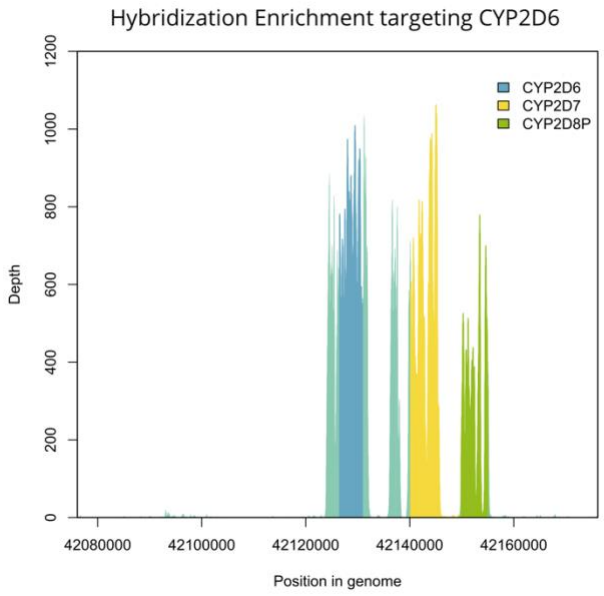
Main Applications of Xdrop™ targeted enrichment

- Structural Variations
- Tandem Repeats
- GC-rich Regions
- Gap-closing
- Integration of Transgenes & Viral DNA

Implications of pseudogene homology

The DNA used in this evaluation included the 3 genes/pseudogenes CYP2D6, CYP2D7 and CYP2D8P. The pairwise identity analysis below reveals the high homology among these 3 sequences (visualized with Geneious Prime). Standard hybridization-enrichment panels designed for CYP2D6, such as Agilent® SureSelect®, generally also enrich the highly similar CYP2D7 and CYP2D8 pseudogenes. As a result, correct mapping short reads generated from these enrichment products is impossible. Reads are randomly assigned to the 3 genes (see figure on next page).





Correct SNP variant calling

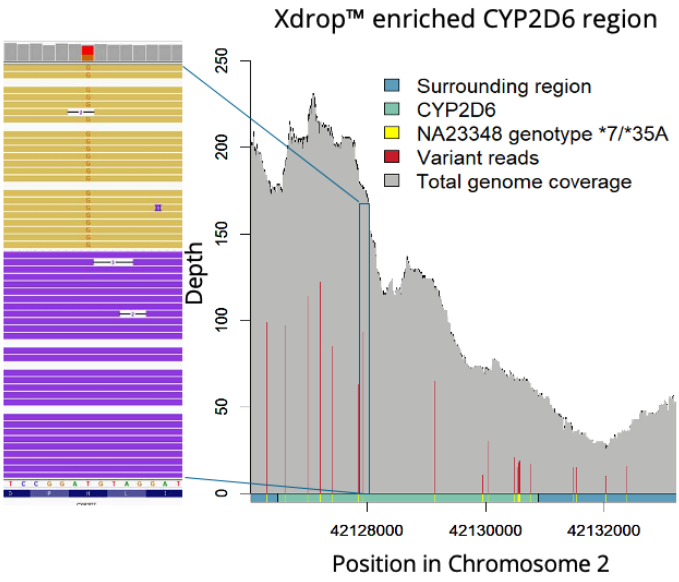
We enriched DNA fragments from 10 ng input material using Xdrop™ and a single primer set. We then sequenced the libraries prepared directly from the enriched samples on Oxford Nanopore's MinION (ONT), which resulted in 162,383 total reads.

The CYP2D6*7 allele contains 1 single nucleotide polymorphism (SNP), while CYP2D6*25A contains 22 SNPs. We were able to call all 23 variants correctly, including phasing (see figure on the right).

De novo assembly to call structural variants

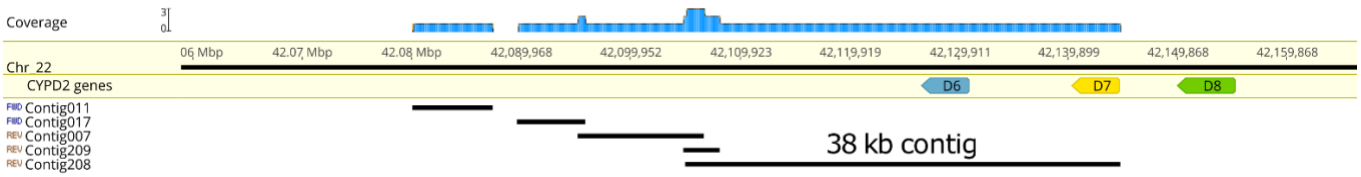
We merged the sequencing reads of 3 replicates, each from 10 ng DNA input, to de novo assemble a contig spanning the CYP2D6 and CYP2D7 gene regions, using roughly 1.6 Gb of raw data.

We extracted the reads from the 100 kb region surrounding the Detection Sequence, used Pacasus[6] to split chimeric reads (inverted repeats of the same read) introduced during dMDA, and then de novo assembled the resulting reads using Canu v.1.9[7], minimap2 and Racon[8], and finally visualized the results in Geneious Prime (see bottom figure).



Conclusions

Xdrop™ empowers applications with a simple design and the Indirect Sequence Capture of long DNA fragments (~100 kb). Coupled with long-read sequencing, Xdrop™ overcomes difficulties in phasing haplotypes in single, genetically heterogeneous samples. The high fidelity of the enrichment enables detecting one or more SNPs in an allele and enables de novo assembly to resolve structural variation.



Aknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 848497.

References

1. Ingelman-Sundberg, M., et al. 2007. Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacol. Ther. 116: 496.
2. Gaedigk, A., et al. 2018. Ten Years' Experience with the CYP2D6 Activity Score: A Perspective on Future Investigations to Improve Clinical Predictions for Precision Therapeutics. J. Pers. Med. 8: 15.
3. Nofziger, C. and Paulmichl. M. 2018. Accurately genotyping CYP2D6: Not for the faint of heart. Pharmacogenomics 19: 999.
4. Yang, Y., et al. 2017. Sequencing the CYP2D6 gene: From variant allele discovery to clinical pharmacogenetic testing. Pharmacogenomics 18: 673.
5. Liau, Y., et al. 2020. Nanopore sequencing of the pharmacogene CYP2D6 allows simultaneous haplotyping an detection of duplications. bioRxiv. Doi: 10.1101/576280.
6. Warris, S., et al. 2018. Correcting palindromes in long reads after whole-genome amplification. BMC Genomics 19: 798.
7. Koren, S., et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722.
8. Vaser R., et al. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737.