



Incorporating sequence capture into library preparation for MinION, GridION and PromethION

Protocols for PCR-based and PCR-free sequence capture allows users to select many loci of interest simultaneously prior to sequencing, making more efficient use of the sequencing run

Contact: publications@nanoporetech.com More information at: www.nanoporetech.com and publications.nanoporetech.com

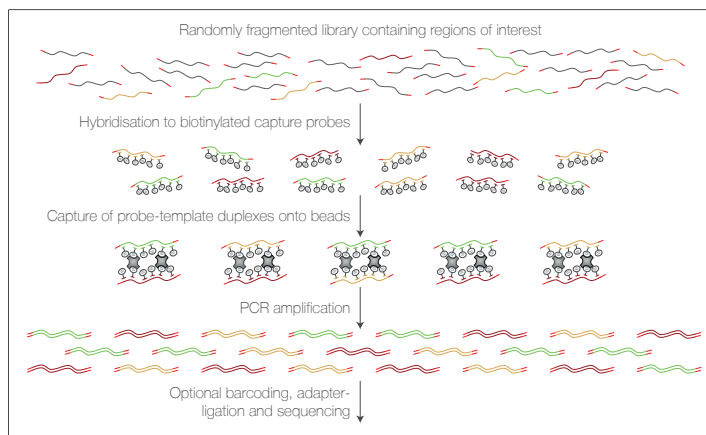


Fig. 1 Overview of the exon-capture workflow

Sequence capture uses complementary probes to enrich for regions of interest

Sequence capture is a technique which allows specific regions to be enriched from a genome. If the experimental goals can be reached by analysing data from a reduced set of loci, target enrichment can save time and money, both on data generation and analysis. Smaller regions of interest can be enriched by PCR, but as the number of target areas increases, PCR becomes less feasible. It is difficult to multiplex large numbers of PCR reactions, and the quantity of reagents and input DNA quickly becomes too large for many such PCR reactions to be performed individually. Sequence capture enriches many target loci in parallel, and is performed by hybridising the library fragments to probes which are specific to the regions of interest (Fig. 1).

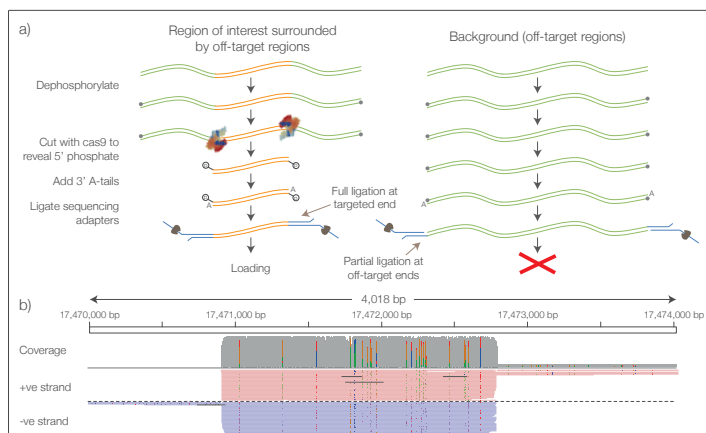


Fig. 3 CRISPR/Cas9 target enrichment a) overview of library prep b) example of mapped reads

90-minute Cas9 library preparation for PCR-free enrichment of target loci

Here, we introduce a PCR-free enrichment method for nanopore sequencing, using Cas9 (Fig. 3). In contrast to amplification-based methods, Cas9 allows native DNA strands to be sequenced, meaning that fragment length and modifications are preserved. In the method, all fragments in the sample DNA are initially dephosphorylated to prevent ligation later in the protocol. Cas9 is then used to cleave the DNA at predetermined sites, exposing ligatable ends at the target sites. All 3' ends are then dA-tailed, before ligation of sequencing adapters. The adapters only ligate to cleaved ends because of the newly revealed phosphates. The entire library is then added to the flow cell. The fraction of reads corresponding to the region of interest is enriched several thousand-fold.

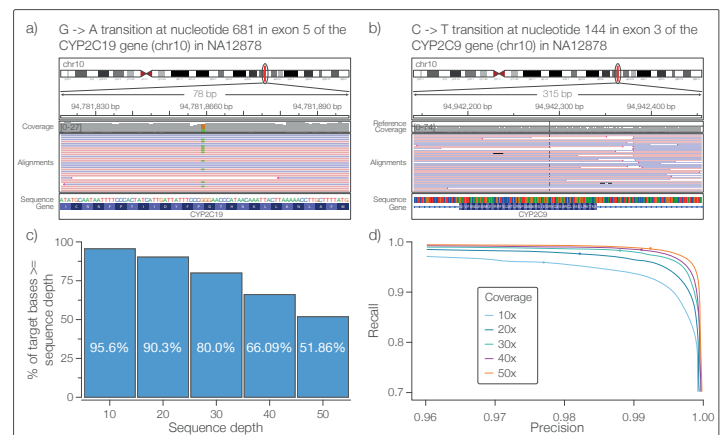


Fig. 2 SNP-calling of exome data a) CYP2C19 b) CYP2C9 c) uniformity d) precision and recall

Single-nucleotide polymorphism (SNP) detection in exome-capture data

SNPs occur approximately once every 1,000 nucleotides in the human genome. The errors in nanopore data are essentially random, making it possible to derive a consensus sequence with extremely high accuracy. Unlike errors, true variants are consistent across several reads. Figs. 2a) and b) show examples of polymorphisms in chromosome 10 of NA12878, which have been verified with other technologies, and which can be clearly seen in IGV plots of the nanopore data using 'quick consensus mode' with the 'hide indels' option. We typically obtain >80% of target bases covered at 30x or greater from approximately 5 Gb of aligned bases (Fig. 2c), with a precision of > 0.988 and recall of > 0.983 (values given for 30x).

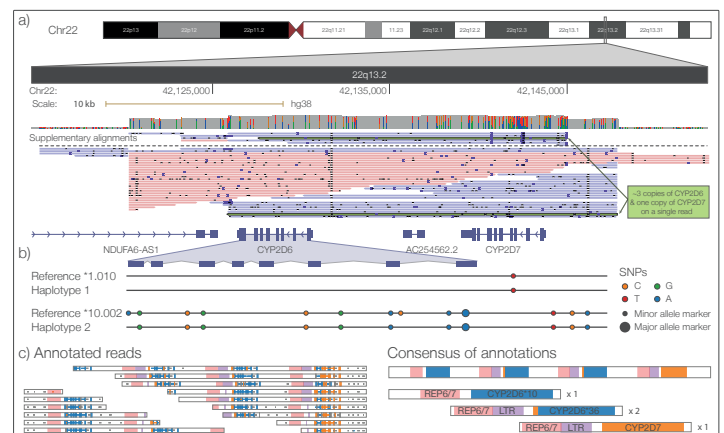


Fig. 4 a) CYP2D6 locus, aligned reads b) NA18256 and reference haplotypes c) gene conversion

Enriching and haplotyping the polymorphic and repetitive CYP2D6 gene

The enzyme CYP2D6 metabolises ~25% of common drugs and its activity between individuals varies widely, due to high levels of polymorphism. To determine an individual's optimal drug dosage, we need to identify all variants that influence the enzyme's activity. Long reads help to distinguish between CYP2D6 and its close paralogue, CYP2D7 (Fig. 4a) and between their gene conversion products. Using Cas9 we enriched the region from NA18256, previously shown to have one copy of allele *10 and three copies of allele *36. Previous tests did not distinguish between the *10 and *36 variants, the latter being a gene conversion variant. We identified the expected alleles and obtained reads that captured three copies of CYP2D6 (Fig. 4b). Analysis indicates that the haplotype contains one copy of *10 allele and two copies of *36 (Fig. 4c).