The Value of Long Read Amplicon Sequencing for Clinical Applications

Kornelia Neveling[1], Arjen Mensenkamp[1], Cheryl Heiner[3], Ian McLaughlin[3], John Harding[3], Lori Aro[3], Linske de Bruijn[1], Edris Askar[1], Simone van der Heuvel[1], Evelien Hoenselaar[1], Ronny Derks[1], Maartje van der Vorst[1], Marcel Nelen[1], Lisenka Vissers[1], Lonneke Haer-Wigman[1], Amber Den Ouden[1], Marjolijn Ligtenberg[1,2], Richarda de Voer[1]

[1]Department of Human Genetics, Radboud university medical center, Nijmegen, the Netherlands
[2]Department of Pathology, Radboud university medical center, Nijmegen, the Netherlands
[3]Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

NGS is commonly used for amplicon sequencing in clinical applications to study genetic disorders and detect disease-causing mutations. But this approach can be plagued by limited ability to phase sequence variants and makes interpretation of sequence data difficult when pseudogenes are present. Long-read high-fidelity amplicon sequencing can provide very accurate, efficient, high throughput (through multiplexing) sequences from single molecules, with read lengths largely limited by PCR. The data is easy to interpret; phased variants and breakpoints are present within high fidelity individual reads.

Here we demonstrate SMRT Sequencing of the ***PMS2*** and ***OPN1 (MW and LW)*** genes. In both cases, homologous regions make NGS and MLPA results very difficult to interpret. For *PMS2*, 3 amplicons ranging in size from 11.4 kb to 16.8 kb were designed using primers unique to this gene, covering 36 kb of sequence. SMRT sequencing produced HiFi reads with coverage ranged from 200-fold to 1500-fold; data clearly indicated 2 deletions >1000 kb with precise breakpoint mapping.

For *OPN1LW* and *OPN1MW*, a 14 kb and a nearly 16 kb amplicon can be generated covering the regions of interest. We show results of sequencing these genes from patient samples with different mutations, including sequencing the full-length amplicons. For all cases, PacBio sequencing was 100% concordant, finding all mutations originally identified by orthogonal technologies.

These results demonstrate several areas of added value from long amplicon sequencing:

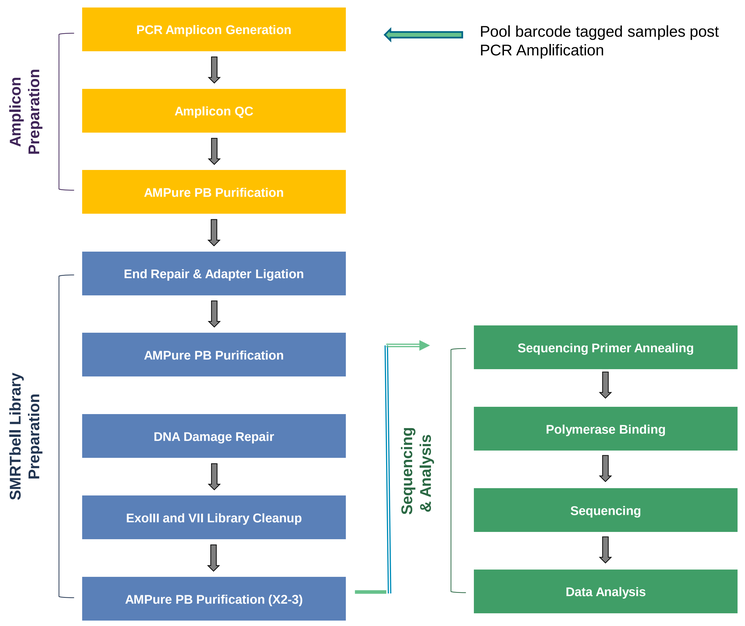
Efficiency

- Less PCR, no nesting
- Fewer added tests (i.e. MLPA)
- Multiplexing for high throughput

Improved results, easier data interpretation and analysis

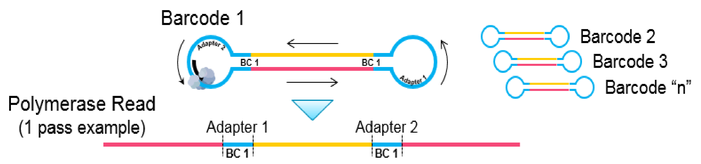
- Ability to distinguish between genes and pseudogenes
- Variant phasing within long reads
- Precise breakpoint detection

Long-Read SMRT Sequencing Workflow



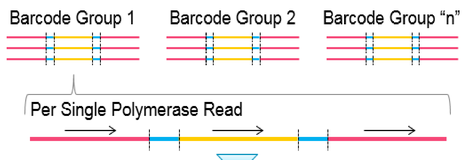
Sequencing Analysis Workflow

Circular Consensus Sequencing (CCS) Analysis

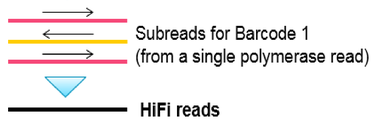


In SMRT Analysis:

1. Pre-Process Filtering (Analysis Parameters)
2. Demultiplex



3. Generate Circular Consensus
   - The CCS analysis method combines multiple passes from a single molecule resulting in high individual read accuracy (>99%)
   - CCS generate HiFi reads ready for further analysis (alignment, variant calling, etc.) with standard informatic tools



Methods and Results

PMS2

Figure 1: Design of the *PMS2* LR-PCR fragments.

Figure 2: Run metrics of a 16 kb amplicon run: (I) the majority of mapped CCS reads (HiFi reads) represent the 11.4, 13.6 and 16.8 kb *PMS2* fragments; (II) the N50 polymerase read length is >10 kb; and (III) the insert read length density plot shows the three LR_PCR fragments.

Figure 3: Complete coverage of *PMS2* by long-read sequencing (upper panel). Informative SNPs in *PMS2* intron 14 and *PMS2CL* intron 5 reduce false-positive variant calls due to removal of *PMS2CL*-specific reads (lower panel).

Methods and Results Continued

PMS2 Continued

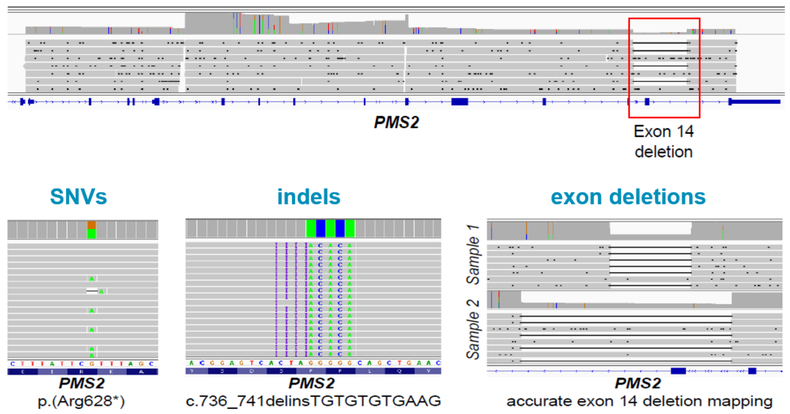


Figure 4: Long-read sequencing of *PMS2* is able to detect exon deletions >1kb in size (upper panel), SNVs (lower panel; left), small indels (lower panel; middle) and accurate breakpoint mapping of most exon deletions (lower panel; right).

OPN1

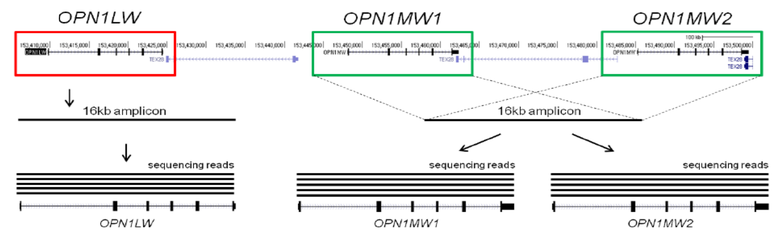


Figure 5: Representation of 16 kb LR-PCRs for *OPN1LW* and *OPN1MW.*

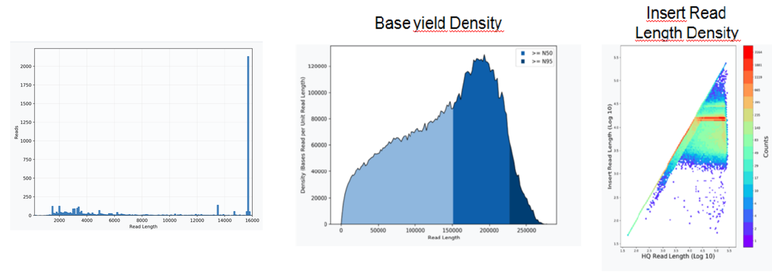


Figure 6: Run metrics of a 16kb amplicon run. The N50 polymerase read length is 151kb. The insert read length density plot shows the ~16kb amplicons. The run output was 18.5 Gb.

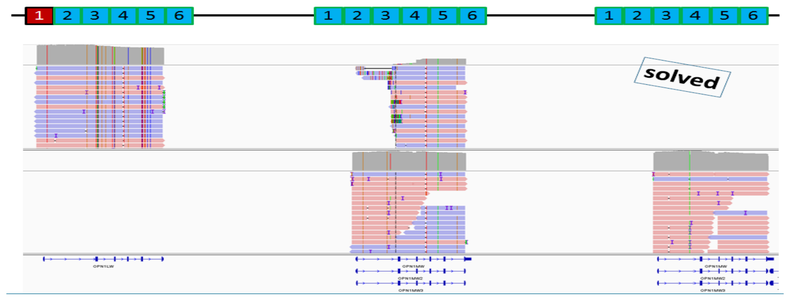


Figure 7: Protanopia patient, with two PCR products (1xLW and 1x MW) . Following sequencing, three MW copies were detected, one has an exon 1 that belongs to LW. All three copies map to different locations in the genome. The data confirm patients' phenotype.

Conclusions

Targeted long-read sequencing with PacBio is highly accurate (>99.99%) and detects all types of variants, sequencing through various contexts. Long amplicon sequencing reduces the laboratory complexity by eliminating additional testing and can be multiplexed for high throughput.