SureSelect^{XT} Application Note

Use of Agilent SureSelect to perform targeted long-read nanopore sequencing

Authors

Isaac Lee, Rachael Workman, Winston Timp,
Department of Biomedical Engineering,
Johns Hopkins University

Josh Zhiyong Wang, Agilent Technologies

Introduction

Large-scale genomic anomalies – structural variations (SVs) – are pervasive in cancer. Due to the scale of the SVs and the repetitive nature of the sequences usually flanking them, the long-reads possible with nanopore sequencing provide an approach to advance the understanding of SVs. In this application note, SureSelect^{XT} is applied to nanopore long-read sequencing, enriching for CDKN2A and SMAD4 tumor suppressor genes, to improve the depth and variant calling accuracy of nanopore sequencing.

This application note focuses on optimizing the SureSelect^{XT} protocol to long-read sequencing and using open-source softwares nanopolish and sniffles to improve the base calling accuracy and detect single nucleotide variants (SNVs) and structural variants (SVs), demonstrating the utility of SureSelect system on third-generation long-read sequencing platforms.

Materials and Methods

The enriched DNA library used in nanopore sequencing library preparation was generated using the standard Agilent SureSelect^{XT} protocol with the following modifications. First, the DNA shearing was optimized for fragmentation centering at 2 kb with 3 – 4 ug purified genomic DNA. After end-repair and adaptor ligation, amplification was performed using PCR reagents and conditions titrated for long range PCR products (e.g. one minute elongation time for each 1 kb length). 750 ng amplified adaptor-ligated ~2kb DNA library underwent RNA probe hybridization and capture. And post-capture PCR was performed with conditions titreated for long-range PCR products. As a quality check, we profiled the size distribution and yield on the Bioanalyzer. Finally we used the enriched post-capture DNA for Oxford Nanopore Technologies (ONT) DNA sequencing library preparation by ligating it to ONT adaptors before getting it loaded and sequenced per ONT's protocol.

SureSelect^{XT} Target Capture for long-read nanopore sequencing

One of the target enrichment systems available from Agilent Technologies is the SureSelect^{XT} solution-phase hybridization-capture system. In addition to predetermined panels of capture probes, the user can utilize custom capture libraries with 120nt biotinylated RNA baits to enrich genomic regions ranging from less than 50 kb to over 100 Mb for deep sequencing of specific genomic regions. With probe designs generated by their design algorithms, which considers complicated factors such as sequence complexity and GC content, it is possible to perform DNA enrichment and sequencing in a highly efficient, cost-effective manner.

We have applied the Agilent SureSelect^{XT} protocol to nanopore long-read sequencing. Note that the vast majority of conventional DNA sequencing library preparation protocols are geared toward creating short, 200-300 bp DNA fragments, tailored to short-read second generation sequencing, e.g. Illumina, Ion Torrent. To apply the enrichment system to long-read sequencing, we adjusted the protocol, altering the shearing conditions to generate DNA fragments with a size distribution centered at 2kb and PCR conditions to allow for amplification of the long DNA strands. The probe design was optimized using Agilent's probe design algorithm and validated experimentally to increase the on-target percentage. Optimizations to the probe design include strategic placement of probes with appropriate, i.e. larger, probe spacing to enrich for larger regions, utilization of stringent probes to decrease non-specific binding, and increased number of probes around regions previously determined to contain SVs.

The modified SureSelect^{XT} protocol described in the following section was performed with 3-4ug of purified PDAC503 pancreatic cancer cell line gDNA with NA12878 lymphatic cell line gDNA as a control[1]. The resulting enriched library underwent library preparation and nanopore sequencing on an Oxford MinION. From an average of 200Mb (100k reads) total sequencing output, ~30% on-target percentage was achieved, yielding an average of >300-fold enrichment in the targeted region [Figure 1 and table 1]. These reads were bioinformatically processed to identify SV and even SNVs after read polishing.

Validation via comparison with short-read Sequencing

To validate the performance of the hybridization capture protocol, the control NA12878 DNA was sheared to ~200 bp fragments and selected using the original SureSelect protocol, and then sequenced via Illumina short-read sequencing on MiSeq sequencing platform (V3 chemistry 2x150bp). Coverage information and SNV candidates were extracted to compare the results with Nanopore sequencing. As shown in Figure 1, the alignment coverage over the targeted regions roughly match between the two sequencing platforms.

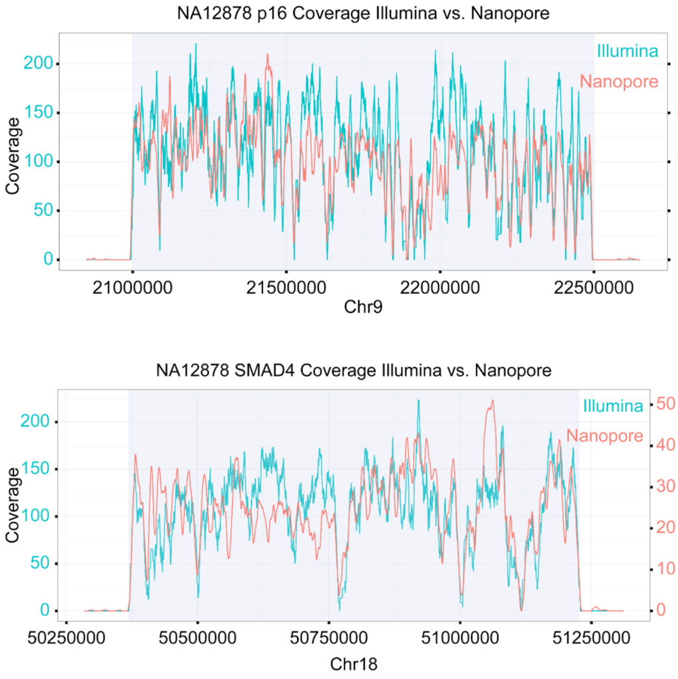

Fig 1. Nanopore and Illumina sequencing coverage of the capture region.

	Total yield (reads)	On-target	On-target percentage	Fold enrichment	Coverage
Illumina NA12878	4.4m	3.7m	85%	641X	113X
Nanopore NA12878	107k	32k	30%	353X	27X
Nanopore PDAC	56k	20k	26%	332X	20X

Table 1. Numbers from the capture alignment results

Detecting Structural Variants and SNPs with bwa-mem, nanopolish, and sniffles on nanopore sequencing data

The reference genome used for analysis of the data is human hg38 genome. The hdf5 output files were first exported as a single fasta file using poretools. Then the reads were aligned using bwa-mem and exported to a bam file:

```
bwa mem -t <$threads> -x ont2d <$reference> <$input_fasta> | samtools view -S -b - > <$output.bam>
```

The alignment file was then supplied to nanopolish along with the original sequence to perform error correction, the result of which was subsequently sorted and exported as a bam file:

```
path/to/nanopolish -t <$threads> -v --sam -r <$input_fasta> -b <$bam_path> -g <$reference> --models nanoplish_models.fofn | samtools view -b - | samtools sort - -o <$output.eventalign.sorted.bam>
```

Nanopolish corrects errors on the aligned sequences via a hidden Markov Model, wherein the observed output is the k-mer current signal, the states are the true nucleotide sequence, and the conditional probabilities are dependent on the previous state, k-mer current signal, as well as the sequence of the reference alignment[2]. All of the outputs together were used to either call SNPs or build the consensus sequence for a given window as follows:

```
path/to/nanopolish variants --snps <$output_file> -w <$chr>:<$start>-<$end> -r <$input_fasta> -b <$input_bam> -g <$reference> -e <$event-aligned_bam> -t <$thread> --models-fofn nanopolish_models.fofn
```

The "snps" option generated a vcf file of the detected SNVs. On the Illumina reads and the nanopore reads before nanopolish, samtools mpileup and bcftools commands were used to obtain the SNV candidates. SNV calling on the error-corrected sequences of the control NA12878 yielded 1,017 SNVs, of which 947 were in agreement with the SNVs for the same cell line published via Platinum Genomes Project [3]. When compared to the 4,138 SNVs called with raw nanopore data, only 2,485 of which were in agreement with published SNV data, we determined that the majority of the inaccurate SNVs are filtered out through error-correction [Figure 2 and table 2]. SNV analysis of illumina sequencing data resulted in 1,211 SNVs, of which 1,133 matched the published data.

Structural variations were detected using a structural variance caller sniffles using the bwa-mem output bam file:

```
path/to/sniffles -m <$bam_file> \
    -s <$minimum number of reads agreeing the call> -c <$minimum CIGAR events> \
    -q <$minimum mapping quality> -v <$vcf_output_name>
```

The output of this command is a variant call format (vcf) file containing loci of SV candidates. From the control NA12878 data, 6 SVs were detected in the CDKN2A region and 3 in the SMAD4 region. These SVs were compared to a list of SVs detected via 100x depth whole-genome PacBio long-read sequencing, provided by the Genome In A Bottle consortium [figure 3]. From the PDAC503 data, one SV in the CDKN2A region and two in the SMAD4 region were detected in addition to the SV observed via the absence in coverage [figure 4]. The window in PDAC503 that showed no coverage (approximately chr9:21,950,000-22,450,000) was previously discovered as a site of homozygous deletion by Norris et. al, and the rest of the four SVs were novel SV candidates[1].

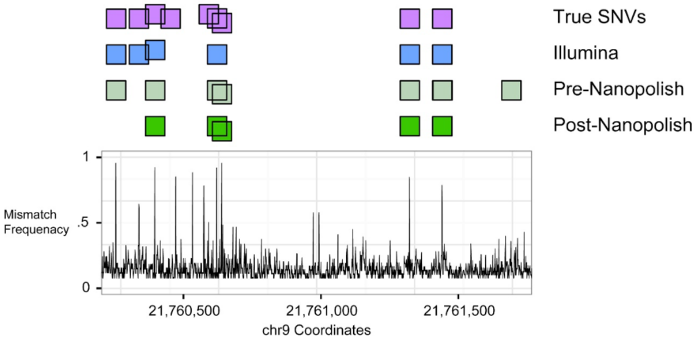

Fig 2. Comparison of detected SNVs and pileup of mismatch in the raw nanopore reads for a 2,000bp window

	SNV Comparisons		
	Illumina	Pre-polish	Post-polish
Avg. Coverage	113	27	27
Correct	1133	2485	947
Total	1211	4138	1017
Precision	**94%**	60%	**93%**
Sensitivity	**32%**	69%	**26%**

Number of True SNVs: 3587(Eberle,et al. bioRxiv, 2016)

Table 2. SNV detection results

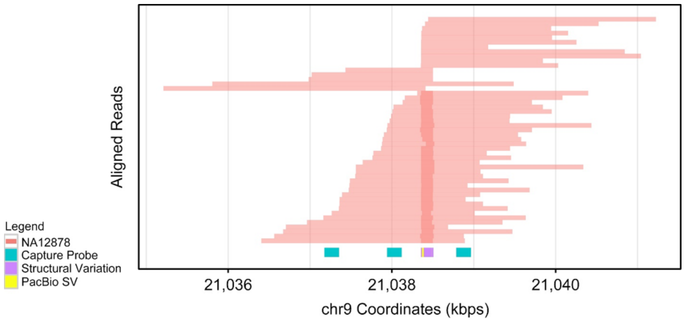

Fig 3. A structural variation detected in NA12878

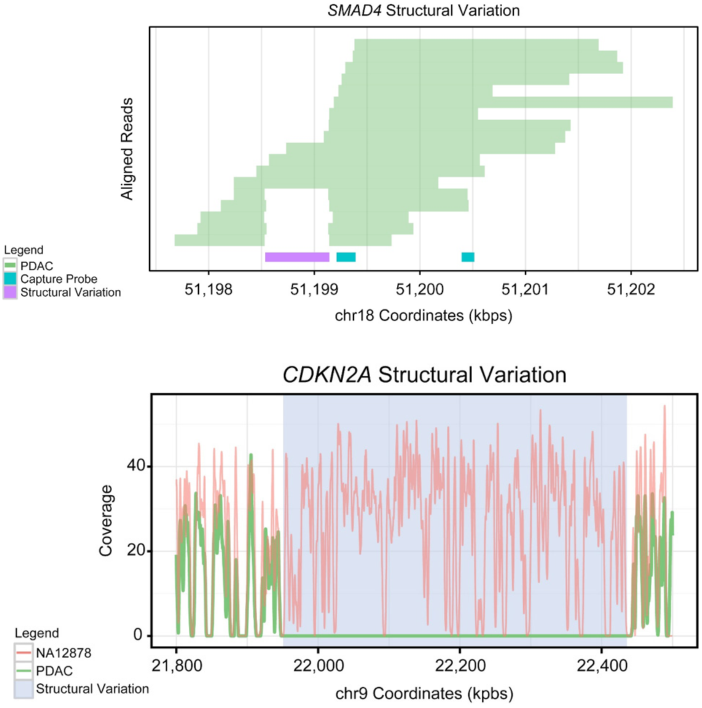

Fig 4. Structural variations detected in PDAC503

Conclusion

Agilent's SureSelect^{XT} target enrichment platform offers a highly reliable and efficient method to study structural variants and single nucleotide variants in specific regions of interest via long-read sequencing. When targeting a 2.4 Mbp region (1.5 Mbp for CDKN2A gene and 850 kbp for SMAD4), which is < 0.5% of the genome, ~30 % on-target was achieved reliably from nanopore sequencing. When performing SNV and SV detection using nanopolish and sniffles algorithms, one can detect previously annotated and novel SNVs and SVs from the SureSelect^{XT} nanopore sequencing data. As expected, SV detection, which is challenging even with high depth sequencing and intense computational processing from short-read data, is efficient and cost-effective using long-read sequencing coupled with the target enrichment. And we envision that both the SNV and SV detection using this technique can be improved with further optimization of the protocol, e.g. enrichment of even longer DNA fragments (>5kb). Furthermore, the workflow described in this application note can be easily adopted to other long-read sequencer platform such as PacBio sequencers by using SureSelect enriched 2kb (or 2+ kb) DNA as input DNA for PacBio sequencing library preparation.

1. Norris, Alexis L., Hirohiko Kamiyama, Alvin Makohon-Moore, Aparna Pallavajjala, Laura A. Morsberger, Kurt Lee, Denise Batista, et al. 2015. "Transflip Mutations Produce Deletions in Pancreatic Cancer." Genes, Chromosomes & Cancer, May. doi:10.1002/gcc.22258.
2. Loman, Nicholas James, Joshua Quick, and Jared T. Simpson. 2015. "A Complete Bacterial Genome Assembled de Novo Using Only Nanopore Sequencing Data." bioRxiv. doi:10.1101/015552.
3. Eberle, Michael A., Epameinondas Fritzilas, Peter Krusche, Morten Kallberg, Benjamin L. Moore, Mitchell A. Bekritsky, Zamin Iqbal, et al. 2016. "A Reference Dataset of 5.4 Million Phased Human Variants Validated by Genetic Inheritance from Sequencing a Three-Generation 17-Member Pedigree." bioRxiv. doi:10.1101/055541.

LEARN MORE OR BUY ONLINE:
www.agilent.com/genomics/sureselectxt

Trusted Answers. Together.

Find an Agilent customer center in your country:
www.agilent.com/genomics/contactus
U.S. and Canada
1-800-227-9770
agilent_inquiries@agilent.com
Europe
info_agilent@agilent.com
Asia Pacific
info_agilent@agilent.com

For Research Use Only.
Not for use in diagnostic procedures.

PR7000-0669

Printed in USA, April 21, 2017
5991-8056EN

Agilent Technologies