

Targeted Sequencing & Phasing on the PacBio® RS II

Using the Roche NimbleGen SeqCap EZ System

Introduction

As a cost-effective alternative to whole-genome human sequencing, targeted sequencing of specific regions, such as exomes or panels of relevant genes, has become increasingly common. These methods typically include direct PCR amplification of the genomic DNA (“gDNA”) of interest, or the capture of these targets via probe-based hybridization. Commonly, these approaches are designed to amplify or capture exonic regions and thereby result in amplicons or fragments that are a few hundred base pairs in length, a length that is well-addressed with short-read sequencing technologies. These approaches typically provide very good coverage and can identify SNPs in the targeted exonic regions, but do not readily identify mutations in introns that may lead to splice variants. These short read approaches are also unable to haplotype these variants.

Here we describe a targeted sequencing workflow that combines Roche NimbleGen’s SeqCap EZ enrichment technology with Pacific Biosciences’ SMRT® Sequencing to provide a more comprehensive view of variants and haplotype information over multi-kilobase, contiguous regions. While the SeqCap EZ technology is typically used to capture 200 bp fragments, we demonstrate that 6,000 bp fragments can also be utilized to enrich for long fragments that extend beyond the targeted capture site and well into (and often across) the adjacent intronic regions. When combined with SMRT Sequencing, multi-kilobase genomic regions can be phased and variants, including complex structural variants (Reference 1), can be detected in exons, introns and intergenic regions.

Materials and Methods

Using Covaris® g-TUBE® devices, we generated 10 kb fragments by shearing 2 µg of genomic DNA according to the manufacturer’s recommended protocols. The sheared gDNA was next size-selected using Sage Science’s BluePippin™ system to select fragments between 5-9 kb, the SeqCap EZ adapters were added and the resulting library amplified by priming off of these

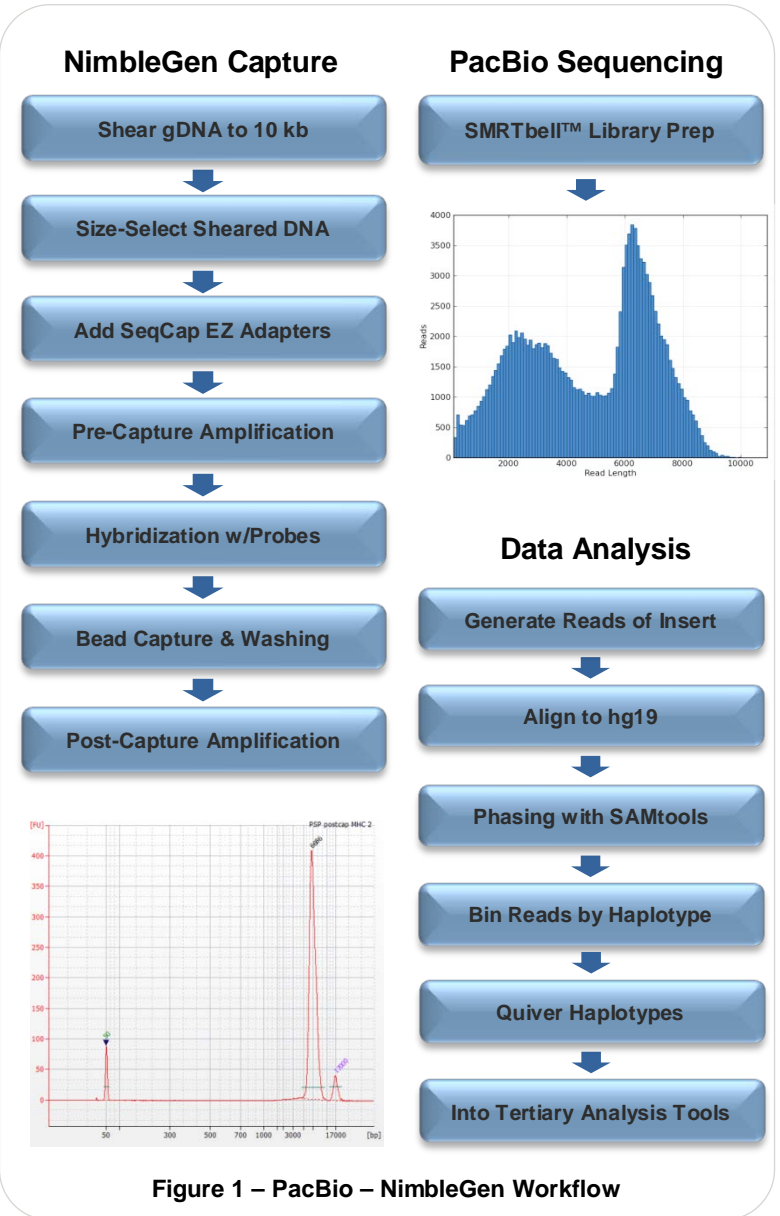


Figure 1 – PacBio – NimbleGen Workflow

adapters. The target specific capture probes were then hybridized to the library and the captured material separated from the non-hybridized material with washing. Then, the captured and washed library underwent a final amplification. After the post-capture amplification step, the sample in Figure 1 exhibited a well-defined peak at 6,700 bp. The sample was now ready for SMRTbell library preparation and sequencing.

All enrichment and library preparation steps were carried out according to the protocols posted on PacBio SampleNet (see Reference 2). The templates were then sequenced; the mapped reads of insert plot from Sample 1 shows a pronounced primary peak around 6 kb.

RS_ReadsOfInsert_Mapping was used to generate one read per molecule using zero minimum passes and a minimum predicted accuracy of 75%, and then these reads were aligned against the Human Genome Reference hg19 using Basic Local Alignment and Successive Refinement (“BLASR”) software. For each targeted region, SAMtools was used to phase and bin reads by haplotype, and then Quiver was applied to polish each haplotype to high consensus accuracy. This entire workflow is summarized on GitHub (see Reference 3).

Results and Discussion

Table 1 describes the human samples that were enriched with the Roche Nimblegen SeqCap EZ Human Major Histocompatibility Complex (MHC) and/or Comprehensive Cancer kits and then sequenced on PacBio and Illumina systems for comparison. For the three samples sequenced on the PacBio RS II, an average of 61% of the reads were on target, representing an average enrichment factor of 1,300X. These values were comparable to the Illumina results.

Targeted Sequencing of the MHC Region

Sample 1 was enriched for the largely contiguous 5 Mb MHC region on Chromosome 6. The PacBio data produced about 50% of the reads on target, representing

an enrichment factor of approximately 600X. Figure 2 shows an example of the typical coverage for one full-length HLA gene, HLA-DQA1. PacBio sequencing reads of insert show even coverage across the entire 6,500 bp gene (Figure 2, top).

These reads were then phased using SAMtools and separated and grouped by haplotype (Figure 2, bottom; blue for one haplotype, pink for the other).

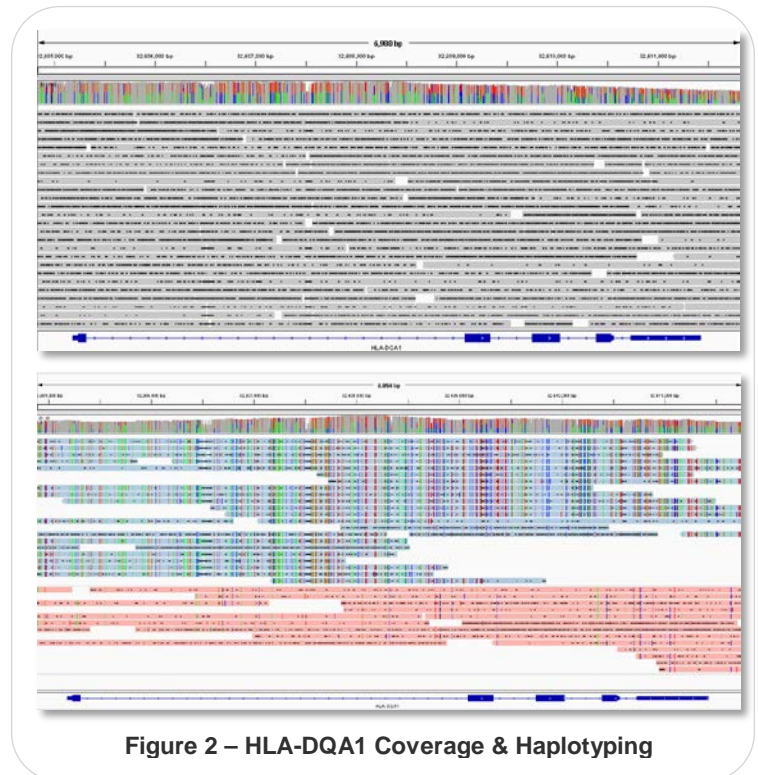


Figure 2 – HLA-DQA1 Coverage & Haplotyping

Sample #	Sample Description	SeqCap EZ Kit	Sequencing Platform	Average Fragment Size (bp)	# SMRT Cells or Runs	Mapped Reads of Insert Mean (bp)	% Reads on Target	Enrichment Factor
1	PacBio Reference	Human MHC (5 Mb region)	PacBio® RS II	6,000	4	4,788	48.4	600x
2	PacBio Reference	Comprehensive Cancer (4 Mb; 578 genes)	PacBio RS II	6,000	4	4,658	65.5	1500x
3	NA12762	Comprehensive Cancer	PacBio RS II	6,000	3	4,352	68.9	1800x
4	NA12762	Comprehensive Cancer	Illumina HiSeq 2500	200	Single, 2 x 100 bp - paired end run	N/A	63.9	1400x

Table 1 – Enrichment & Sequencing Statistics From Human Samples

Sample	NimbleGen – PacBio Type	GenDx – PacBio Type	Sanger-Based Assembly Type
Sample 1	DQA1*02:01 DQA1*01:05	DQA1*02:01 DQA1*01:05	DQA1*02:01 DQA1*01:05

Table 2 – Typing of the HLA-DQA1 Locus

Samples 2 and 3 were enriched with the NimbleGen Comprehensive Cancer design as described above. This panel covers a 4 Mb region that targets the exons for 578 oncology-related genes.

Similar to the results with the MHC design, this method provides broad sequencing coverage across many of the genes. Figure 3 shows an example of a 35 kb gene (KMT2D) from Sample 2 which is covered in its entirety. The heterozygous SNPs distributed across the gene also allow for haplotyping.

Figure 4 illustrates the sequencing results from Samples 3 and 4 comparing PacBio to Illumina data. Across this 10 kb region of BRCA1, Illumina sequencing detected 8 SNPs. PacBio sequencing was able to detect these same SNPs (black boxes) and an additional 7 SNPs that were either not detected, or detected with insufficient coverage, in Illumina data (red boxes). Using longer fragments, PacBio sequencing can capture variants that extend much further from the location of the target capture probes.

For clearer visualization of the variants, reads of insert with a predicted accuracy of >97% were used. Quiver was then used to generate a consensus sequence for each haplotype. Sample 1 was independently typed using sequence data from a Sanger-based assembly and amplicons generated using NGS-go® HLA primers from GenDx. The methods were 100% concordant (Table 2).

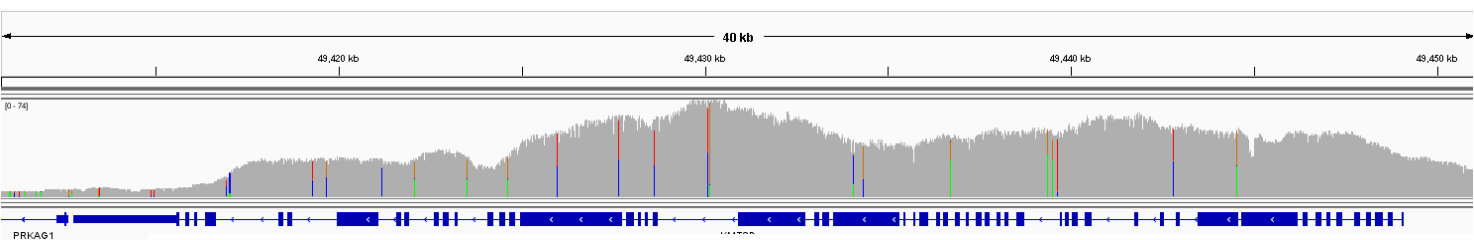


Figure 3 – Sequencing Across An Entire 35 kb Gene

Targeted Sequencing of Cancer Genes

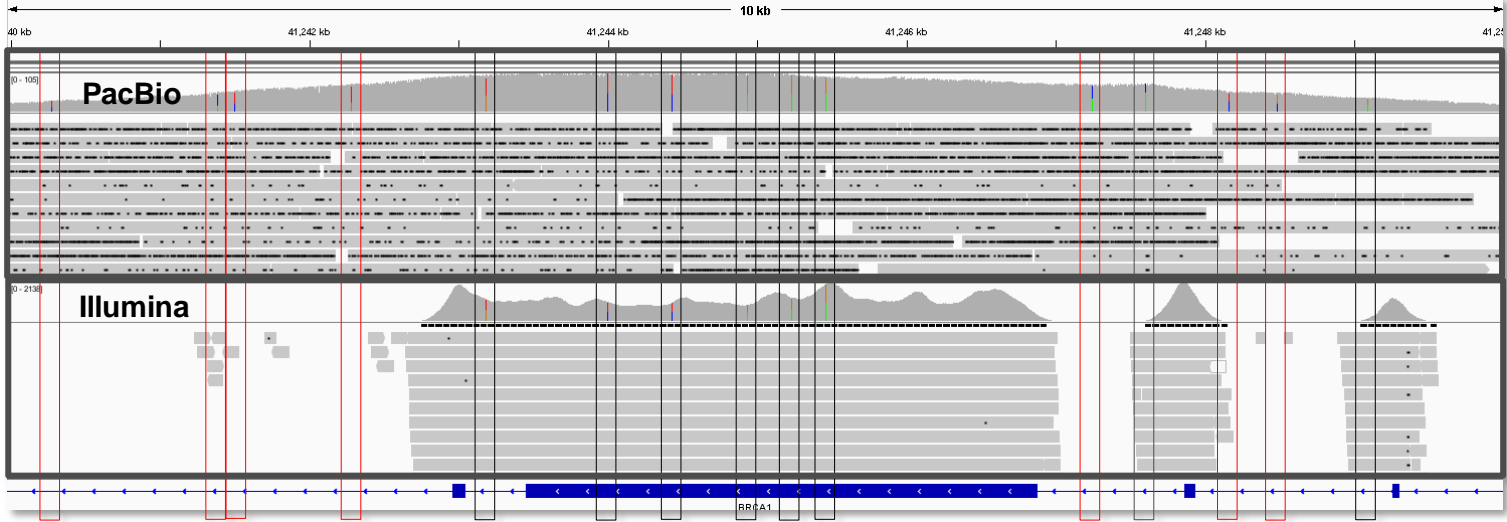


Figure 4 – PacBio Sequencing Detects Additional Variants in the BRCA1 Gene

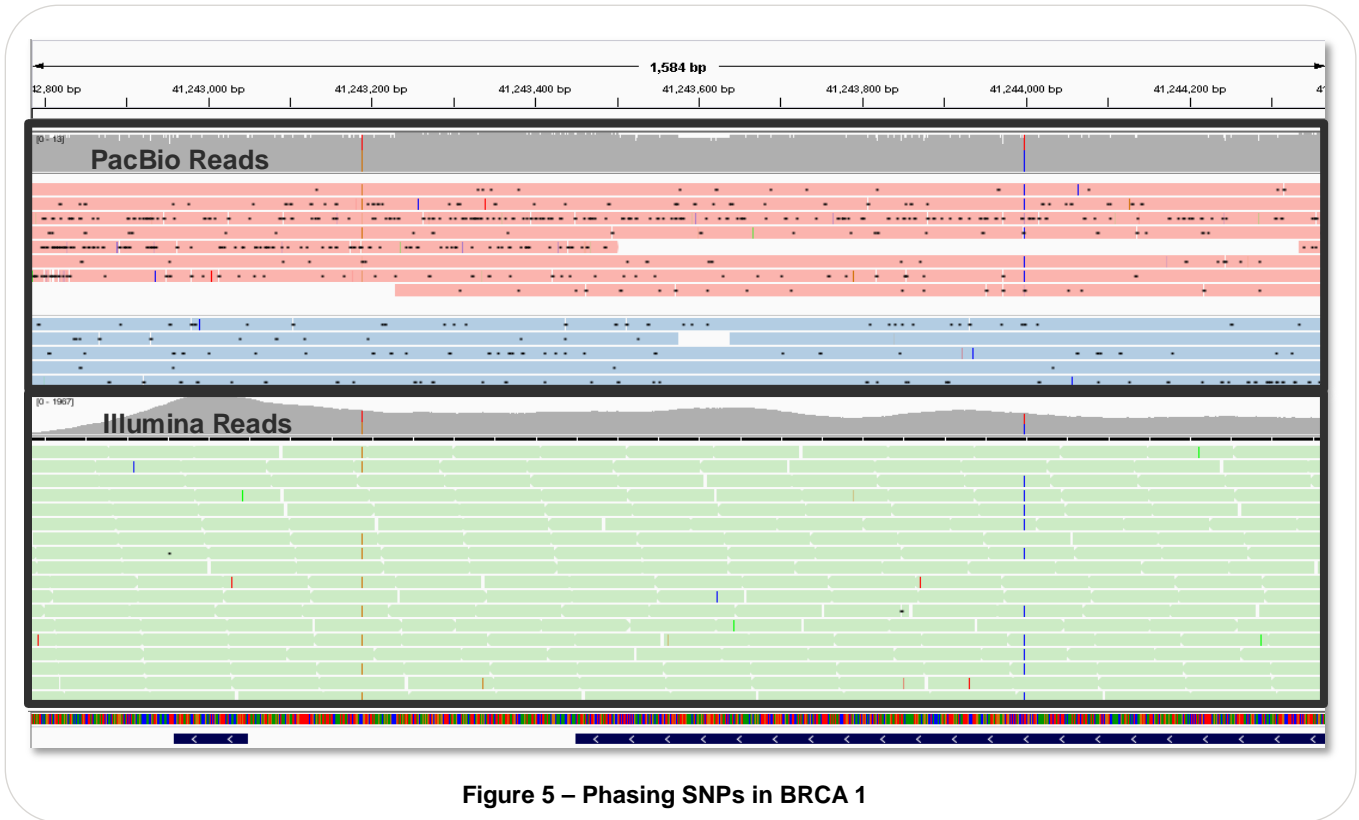


Figure 5 – Phasing SNPs in BRCA 1

Figure 5 provides a comparison of phased PacBio sequencing reads from Sample 3 (top panel) to unphased Illumina reads (lower panel) from Sample 4. The PacBio reads were phased using SAMtools and then separated and grouped by haplotype (blue for one haplotype, pink for the other). In this example, two heterozygous SNPs (red boxes) that are 800 bp apart can easily be phased by looking at the arrangement of base calls of the two SNPs across individual PacBio reads, which in this sample are greater than 4 kb. For clearer visualization of the variants, reads of insert with a predicted accuracy of >97% are shown. By contrast, phasing is not possible in this example using 2 x 100 bp paired-end Illumina reads.

Conclusion

Targeted sequencing of 6 kb fragments using Roche NimbleGen's SeqCap EZ enrichment combined with SMRT Sequencing provides even coverage over multi-kilobase regions of the genome. With PacBio long reads, heterozygous SNPs can be used to phase the reads and generate accurate haplotypes.

Compared with short-read sequencing technologies that provide little to no coverage in the intronic regions, this method provides a more comprehensive view of the targeted regions of interest. Off-the-shelf and custom-designed capture probe sets can be ordered from the NimbleGen website (see Reference 5).

References

1. Wang et al. "PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations" *BMC Genomics* (2015) 16:214
2. NimbleGen Enrichment and SMRTbell Template Prep Protocol on PacBio SampleNet
3. Targeted Phasing Consensus Scripts on GitHub
4. PacBio Targeted Sequencing Webpage
5. Roche NimbleGen SeqCap EZ System Webpage

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2015, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacificbiosciences.com/licenses.html>. PacBio, SMRT, SMRTbell and Iso-Seq are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science, Inc. NGS-go and NGSengine are trademarks of GenDx. All other trademarks are the sole property of their respective owners. 100-483-900-01