PACBIO®

Application Note
Targeted Sequencing

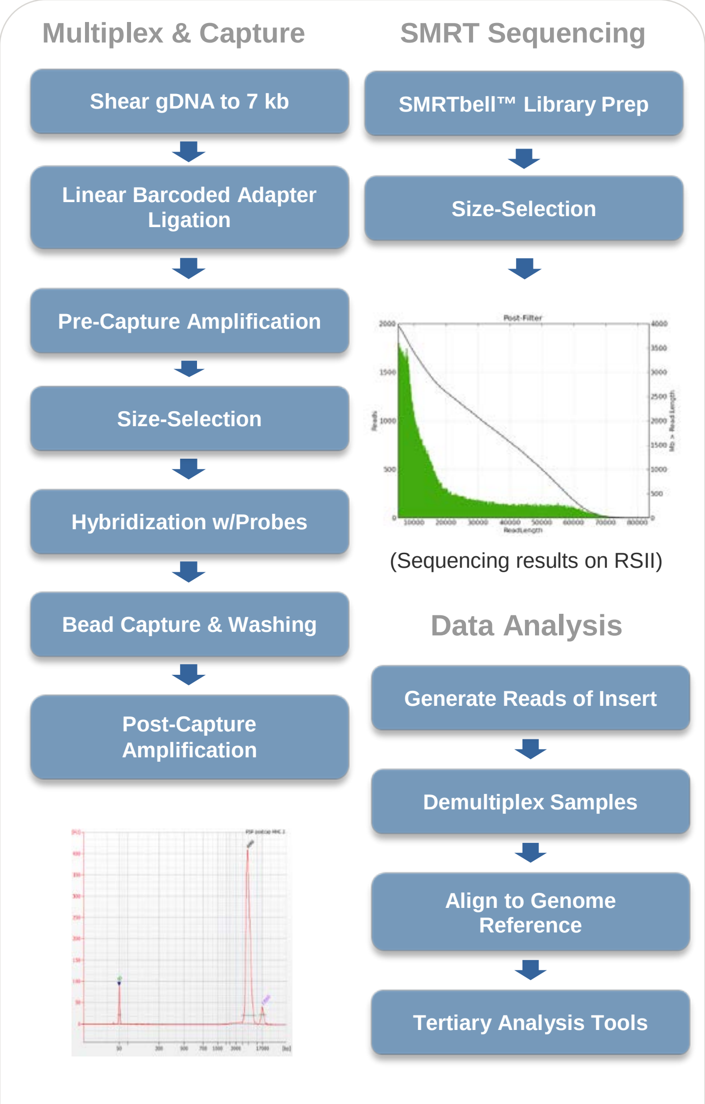
Multiplex Target Enrichment Using Barcoded Multi-Kilobase Fragments and Probe-Based Capture Technologies

Introduction

Target enrichment capture methods allow scientists to rapidly interrogate important genomic regions of interest for variant discovery, including SNPs, gene isoforms, and structural variation. Custom targeted sequencing panels are important for characterizing heterogeneous, complex diseases and uncovering the genetic basis of inherited traits with more uniform coverage when compared to PCR-based strategies. With the increasing availability of high-quality reference genomes, customized gene panels are readily designed with high specificity to capture genomic regions of interest, thus enabling scientists to expand their research scope from a single individual to larger cohort studies or population-wide investigations. Coupled with PacBio® long-read sequencing, these technologies can capture 5 kb fragments of genomic DNA (gDNA), which are useful for interrogating intronic, exonic, and regulatory regions, characterizing complex structural variations, distinguishing between gene duplications and pseudogenes, and interpreting variant haplotypes. In addition, SMRT® Sequencing offers the lowest GC-bias and can sequence through repetitive regions.

SMRT Sequencing-compatible, linear barcoded adapters are available as a cost-effective solution to multiplex up to 12 samples for targeted capture investigations. Incorporating these barcoded adapters into samples, prior to pooling and capture, allows investigators to analyze multiple samples simultaneously for a given panel on a single SMRT Cell run, or as appropriate for sufficient gene panel coverage. It is recommended to use these specific barcoded adapters, which have performed well with our modified library preparation workflow and sequencing platform. These adapters include universal primer sequences and have been tested for even amplification and unambiguous distinction between multiplexed samples. Use of these barcoded adapters is independent of the probe-based capture solution employed. Here are some typical results from probe-based target capture investigations coupled with SMRT Sequencing.

Materials and Methods



(Sequencing results on RSII)

Figure 1 – Workflow Overview

This multiplexed target-capture workflow (Figure 1) incorporates the linear barcoded adapters immediately after gDNA shearing, allowing pooling of samples prior to probe hybridization and SMRTbell™ library preparation.

Note that all enrichment and library preparation steps were carried out according to the protocols posted on our website (**Supported Protocol**). For more information, and the most up-to-date copy of this Note, see our PacBio Literature resources on the **PacBio Targeted Sequencing Applications** page. The oligo sequences used for generating the recommended 12 linear barcoded adapters (Table 1) can be ordered directly from a preferred oligo-synthesis house. The sequences and protocol for annealing these oligos to form double-stranded, linear barcoded adapters can be found with the **Oligo Order Sheet** and **Supported Protocol**).

Barcode ID		Linear Barcoded Adapter Sequences
1	lbc0001	gcagtcgaacatgtagctgactcaggtcacTCAGACGATGCGTCATggtagT
2	lbc0009	gcagtcgaacatgtagctgactcaggtcacCTGCGTGCTCTACGACggtagT
3	lbc0017	gcagtcgaacatgtagctgactcaggtcacCATAGCGACTATCGTGggtagT
4	lbc0026	gcagtcgaacatgtagctgactcaggtcacCGAGCACGCGCGTGTGggtagT
5	lbc0038	gcagtcgaacatgtagctgactcaggtcacTGCTCGCAGTATCACAggtagT
6	lbc0040	gcagtcgaacatgtagctgactcaggtcacCAGTGAGAGCGCGATAggtagT
7	lbc0048	gcagtcgaacatgtagctgactcaggtcacCACACTCTAGAGCGAggtagT
8	lbc0052	gcagtcgaacatgtagctgactcaggtcacCAGACTCTCACACGCggtagT
9	lbc0058	gcagtcgaacatgtagctgactcaggtcacAGATATCATCAGCGAGggtagT
10	lbc0059	gcagtcgaacatgtagctgactcaggtcacTGCAGTGATCGATGAggtagT
11	lbc0062	gcagtcgaacatgtagctgactcaggtcacGACAGCATCTGCGCTCggtagT
12	lbc0070	gcagtcgaacatgtagctgactcaggtcacCTGCGCAGTACGTGCAggtagT
Primer/ Blocker		**Sequence**
Universal Sequence		gcagtcgaacatgtagctgactcaggtcac

Table 1 – Linear Barcoded Adapters and Primer for Multiplexed Targeted Capture.Universal Primer Sequences are used for both amplification steps of multiplexed libraries and as blockers during probe-hybridization. These universal sequence are integrated in the linear barcoded adapters.

Seven (7) kb fragments were generated using 5 µg of gDNA using Covaris® g-TUBE® devices according to the manufacturer's recommended protocols (Figure 2). This was followed by a DNA repair step and ligation of the linear barcoded adapters. Individual barcoded samples were then amplified by priming off the integrated universal sequence before pooling the samples for probe hybridization. After capture and washing, the samples were once-again amplified. The sample in Figure 3 exhibited a well-defined peak at ~7,000 bp after the amplification step and was ready for SMRTbell library preparation. We then size-selected (5 kb - 50 kb) the SMRTbell library using Sage Science's BluePippin™ system (Figure 4) prior to sequencing and analysis. After sequencing, the mapped reads of insert plot showed a dominant primary peak around 8 kb (Figure 5).

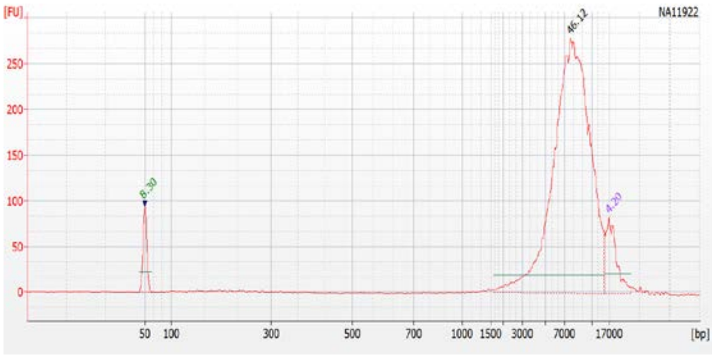


Figure 2 – Bioanalyzer plot of genomic DNA sheared to 7 Kb.

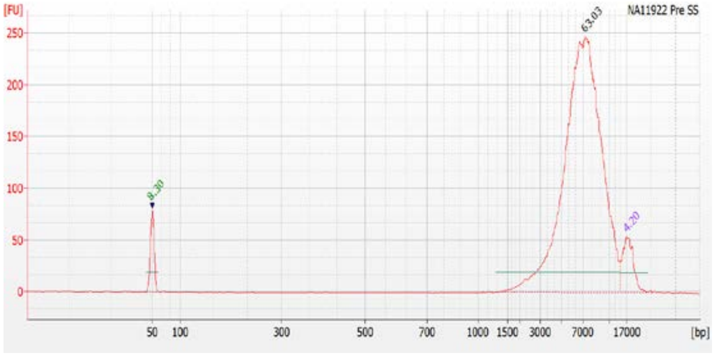


Figure 3 – Bioanalyzer plot of gDNA samples after pre-capture amplification before size selection.

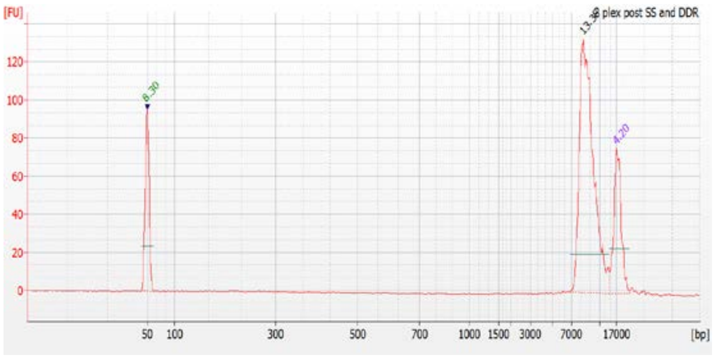


Figure 4 – Bioanalyzer plot of final size-selected SMRTbell library with DNA damage repair.

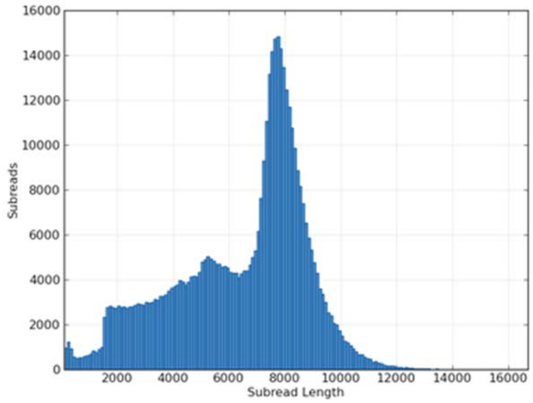


Figure 5 – Read length histogram of enriched reads mapped to reference genome.

For enrichment-fold analysis, the RS_Resequencing_Barcode.1 protocol was used to align subreads to the reference genome. The default parameters were used to filter out any low-quality, short reads. Subsequent to alignment, **BEDtools** was used to estimate coverage statistics and calculate On/Off-Target rates. The On-Target rate was used to calculate fold enrichment (Enrichment Factor) by taking into account the reference genome size and the probe capture target region size. Typical representation of On/Off-target rates observed are shown in Table 2. The entire workflow is summarized on the **PacBio Multiplexed Targeted Capture GitHub** page.

Representative Target Enrichment Rate for 12-Plex Capture Using 1.6Mb Neurology Panel on 1 SMRT Cell (Average of 3 Cells)	
On Target Reads	46,129
Total Reads	67,792
Reads on Target	68%
Human Genome Size (Mb)	3,200
Total Target Size (Mb)	1.6
Enrichment Factor	1,372 x

Table 2 – Representative On/Off-target rates and fold enrichment observed using NimbleGen Neurology gene panel.

For phasing analysis, the RS_*ReadsOfInsert_Mapping* module in the **SMRT Analysis Software** was used to generate one read per molecule using zero minimum passes and a minimum predicted accuracy of 75%.These reads were then 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 (SMRT Analysis Software) was applied to polish each haplotype to high consensus accuracy. This entire workflow is summarized on **PacBio Targeted Phasing Consensus GitHub** page.

Results

The following results are representative data from 6 kb gDNA fragment captures of multiple individual samples, each ligated with a different PacBio linear-barcoded adapter. One benefit of capturing long fragments is the ability to attain coverage across the entire gene. Since each read is multiple kilobases in length, each one will often contain multiple heterozygous SNPs, which can be used to assign each read to the appropriate haplotype (Figure 6).

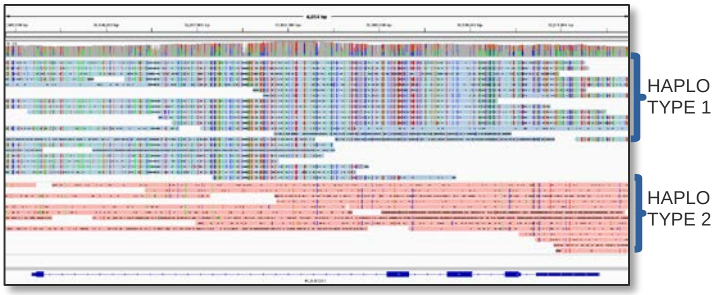


Figure 6 – Representative coverage of the 6.5kb HLA-DQA1 gene showing the reads can be sorted into two haplotypes.

Another advantage of long fragments is the improved mapping accuracy to the genome, which proved useful for differentiating genes from their pseudogenes or copy number variations, and other regions with high homology (Figure 7).

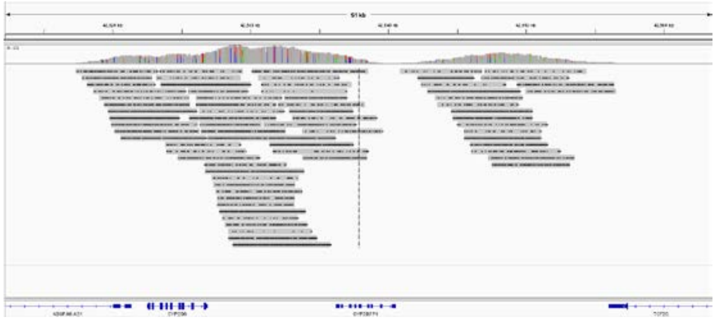


Figure 7 – Captured reads aligning to the CYP2D6 region. Capture probes were only designed for the exons of CYP2D6, yet they are able to capture long, 5 kb, homologous fragments that unambiguously map to CYP2D6, CYP2D7 and CYP2D8 revealing the structure in that region.

Similarly, capturing long fragments is valuable for detailed characterization of large repetitive regions where short-read technologies will have difficulty sequencing and mapping.

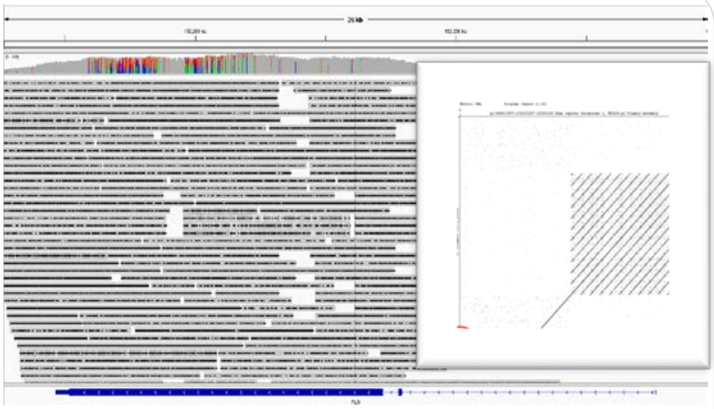


Figure 8 – Exon 3 of the Filaggrin (FLG) gene is approximately 12 kb long and made up of multiple 1 kb repeats. Capturing 5 kb fragments allows for unambiguous mapping to resolve the total number of repeats. Aligning the consensus capture sequence to the reference illustrates this sample has twelve 1 kb repeats (inset).

Samples were multiplexed prior to probe-capture for a cost-effective workflow and obtain even representation with high fold enrichment for SMRT Sequencing (Figures 9 & 10). This allows researchers to multiplex two to twelve samples in a single capture reaction, significantly reducing the overall cost of capture reagents, library preparation, and the number of SMRT Cells required (Table 3). On a per reaction basis, capture reagents typically cost from $100 to more than $1000, depending largely upon the number of capture probes in the design and the platform used. The ability to barcode multiple samples and pool them prior to performing a single capture reaction can significantly reduce the total cost per sample. Furthermore, the pooled, captured samples can then undergo a single PacBio template prep reaction, further saving time and reducing costs.

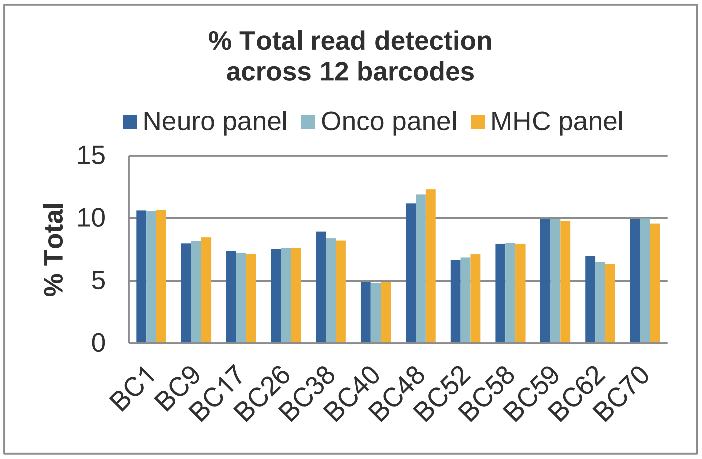


Figure 9 – Representative barcode coverage across 12 samples based on mapped subreads. Oncology, Neurology, and MHC panel results shown.

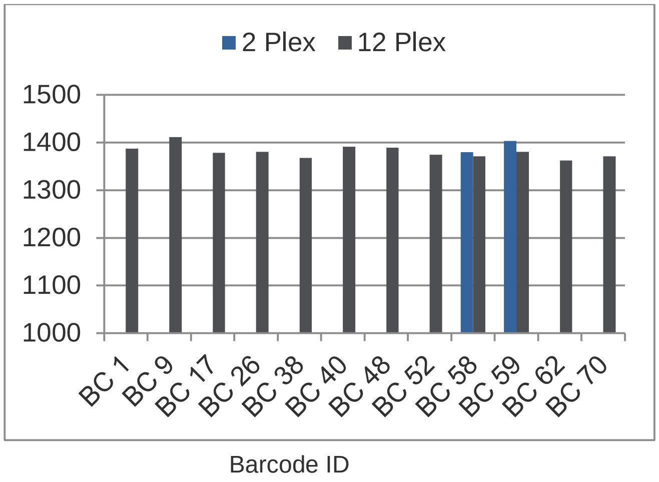


Figure 10 – Barcode coverage representation and fold-enrichment across a 2- and 12-plex target capture sample runs based on mapped subreads.

Cost Advantages of Barcoding Samples Prior to Capture (Targeting 1.6 Mb of Sequence per Sample Across 12 Samples)		
	Without Barcodes	With Barcodes
# Capture Rxns	12	1
# Library Prep Rxns	12	1
# SMRT Cells	12	3

Table 3 – Cost benefit of multiplexing target capture for SMRT Sequencing.

Conclusion

This workflow for multiplexed target capture of 5 kb genomic DNA fragments with SMRT Sequencing-compatible, linear barcoded adapters allows investigators to pool up to twelve samples prior to capture. This significantly reduces the cost of the capture and template preparation reagents, thereby reducing the overall cost per sample. Our typical results reveal even coverage over multi-kilobase regions of the genome and across the multiplexed samples. PacBio long reads offer the opportunity to phase SNPs, interrogate intronic regions, characterize complex structural variation, and distinguish duplication events and pseudogenes. This multiplexed workflow solution has been tested to be compatible with several probe-based capture technologies.

References

1. IDT barcodes for multiplex targeted sequencing order sheet
2. Shared Protocol: Multiplex targeted Sequence capture using Roche NimbleGen SeqCap EZ Library
3. Github: Analyzing your multiplexed targeted capture data

Other Targeted Sequencing & Barcode Solutions

4. Website: PacBio Targeted Sequencing Applications
5. Product Note: Multiplexing Amplicons up to 10 Kb
6. Shared Protocol: Guidelines for Using PacBio Barcodes for SMRT Sequencing
7. Procedure & Checklist – Preparing SMRTbell™ Libraries using PacBio Barcoded Universal Primers for Multiplex SMRT Sequencing