**Technical Note**
Targeted Sequencing

# Multiplexing Targeted Sequencing using Barcodes

## Experimental Design

### Introduction

Targeted sequencing using Pacific Biosciences' SMRT® Sequencing technology is an efficient way to examine subregions of complex genomes. Sample barcoding enhances throughput by enabling multiplexing of targets and samples for simultaneous sequencing. This strategy allows for significant cost savings by providing researchers the flexibility to sequence multiple targets across multiple biological samples in a simplified workflow. PacBio has developed and evaluated a set of barcodes specifically designed for Circular Consensus Sequencing ("CCS") on the PacBio® *RS.*

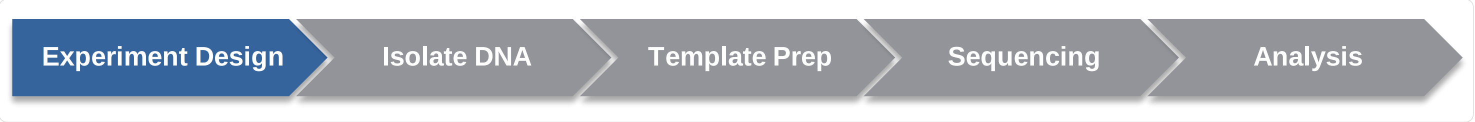
This note provides:

- An end-to-end demonstration of barcoding on the PacBio *RS* instrument.
- The methods used in designing PacBio barcodes.
- The workflow for barcode incorporation into PCR-amplified DNA molecules.
- An informatics workflow for PacBio barcode identification from sequencing reads.
- Results of PacBio barcode identification from sequencing runs.

### Experimental Design and Methods

The Fluidigm® EGFR-MET cancer gene panel was used as an example amplicon set to demonstrate barcode multiplexing using SMRT Sequencing. This panel is comprised of 49 amplicon regions, representing a two-gene panel. Specifically, it includes 28 amplicons from EGFR and 21 amplicons from MET, each with an average length of 500 bp. For more information about using the Fluidigm Target Enrichment System and preparing SMRTbell™ templates, see the *Pacific Biosciences® Technical Note – Targeted Sequencing on the PacBio RS using Fluidigm® Access™ Array System for Target Enrichment*.

To test the efficiency of barcodes, two experiments were carried out. A simple positive control experiment was run in which all 49 amplicons of the EGFR-MET panel were amplified for a single sample (Corriell NA17316) in two independent runs. In the first run, *only* barcode F0 was used to label the sample while *only* barcode F1 was used in the second run. Additionally, a full 48-multiplex experiment was performed where the same sample was amplified using 48 different barcode pairs. This resulted in 2352 distinct products to simulate maximum sample multiplexing by the Fluidigm array.



## Barcode Design

The best barcodes are those which are the most distinguishable from each other to ensure accurate mapping and identification during sequencing. A set of 96 barcodes, each 16 bp in length, was designed under several design constraints to ensure that barcodes were unlikely to be misread during sequencing or mis-mapped during alignment. Specifically, the PacBio barcodes were designed to have no homopolymer stretches, low pairwise sequence similarity (enforced by poor alignment scores), and to be simultaneously distinguishable in both forward and reverse-complement orientations.

Figure 1 shows a simple heatmap of alignment scores for all pairs of designed PacBio barcodes. There is no discernible structure, indicating that all barcodes are uniformly different from each other.

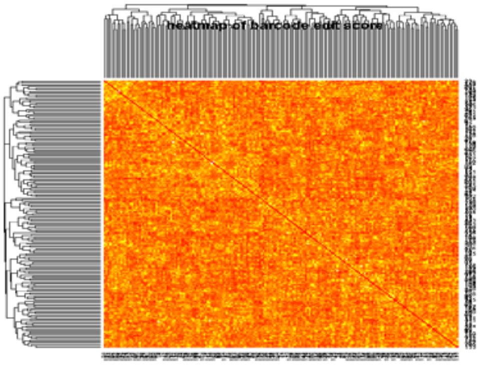


**Figure 1: Barcode Alignment Score HeatMap.**

## Barcode Incorporation Method

PacBio barcodes were incorporated into amplicons using the Fluidigm Access Array System. This microfluidic solution allowed us to amplify up to 48 barcoded samples to produce the targeted amplicons for the panel.

Each target was amplified by two target-specific primers that were each tailed with "constant region" sequences (CS1 on the Forward primer and CS2 on the Reverse primer). These constant regions enabled barcoding since they could be used to anchor the barcode sequences during amplification within the Fluidigm Access Array System. This four-primer strategy was implemented with a two barcode scheme, where each end of the insert was tagged by distinct barcodes (See Figure 2).

**Barcode Left** + **CS1** + **Amplicon** + **CS2** + **Barcode Right**

Note that these barcodes can be used as tails of PCR primers and are therefore amenable to other multiplex targeted enrichment strategies (in addition to the Fluidigm strategy detailed here).

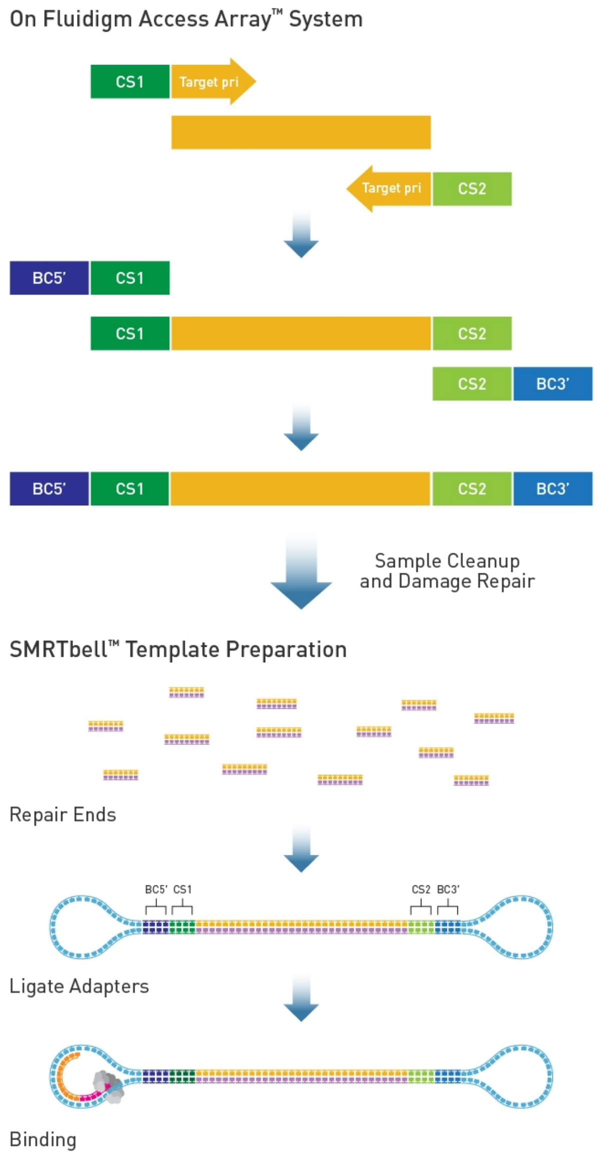


**Figure 2: Template Preparation Workflow.**

In the Fluidigm Access Array System barcoding design, the barcodes, constant regions, and SMRTbell adapter

In the Fluidigm Access Array System barcoding design, the barcodes, constant regions, and SMRTbell adapter together comprised 129 bases of non-template sequence that was read by the enzyme, in addition to the target sequence itself (over 500-bases). However, the entire structure is easily covered due to the long readlength of SMRT Sequencing.

### *Template Preparation and Sequencing*

Barcoded amplicons generated using the Fluidigm Access Array System were converted into SMRTbell libraries using standard SMRTbell library preparation protocols.[1] See the *Pacific Biosciences Technical Note - Targeted Sequencing on the PacBio RS using Fluidigm Access Array System for Target Enrichment* for information on the upfront template preparation process using the Access Array System.

Barcoded SMRTbell libraries were sequenced using the standard 2X45 minute sequencing protocol with an early version of the PacBio C2 chemistry to generate CCS reads. Two SMRT Cells were run on the Corriell NA17316 single-barcode pair x 49-amplicon control sample and seven SMRT Cells were run for the 49-amplicon x 48-plex barcode pooled sample.

### *Barcode Identification*

Barcode identification was done using standard sequence alignment tools[2] on PacBio CCS reads. Specifically, the HMMer package was used to identify the barcodes using Hidden Markov Models ("HMM") as the core alignment method[3, 4]. HMMs were used because they are fully probabilistic, and allow the proper building of constraints into the model while maintaining highly sensitive and specific-search performance.

A *start* barcode model was pinned to the beginning of the read and a *finish* barcode model was pinned to the end of the CCS read (see Figure 3). They were then independently assessed. HMM pinning was done by assigning probabilities that geometrically decrease as the hit begins further away from the ends of the read (Figure 3). The two ends were then combined by adding the log-likelihood-odds scores of the start and finish HMM hits in pairs that corresponded to the different barcode hypotheses. For example, since there were 48 different barcodes in the design, there were 48 different hypotheses. The optimal barcode call was given by the maximum log-likelihood-odds hypothesis. This process was performed for both the forward and reverse complement of the reads.

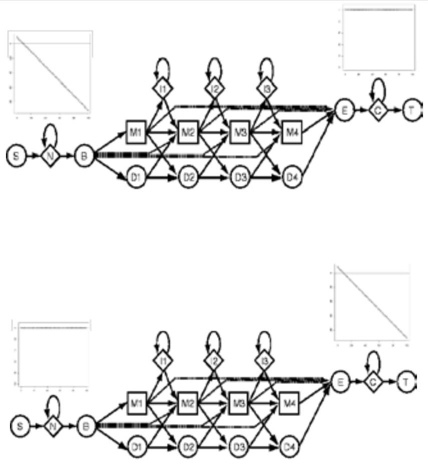


**Figure 3: Start and Finish Barcode HMMs. The HMMs are probabilistically pinned to the beginning (top of figure) and end (bottom of figure) of the read where the barcodes should occur.**

## Results and Discussion

### *Barcoding Yields*

For the 48-multiplex experiment, our yield was 139,397 CCS reads from 7 SMRT Cells. On average, 15,907 barcode-identified CCS reads were detected per SMRT Cell, or 79.9% of all CCS reads. Additionally requiring a read to contain not only a barcode but also a good amplicon alignment gave a yield of 75.6%. Incomplete amplicon products are believed to be the main cause of the yield loss.

For the simple positive control experiment, the yields were similar. On average, over the two independent runs, 78.6% of all CCS reads contained the correct barcode pair and 76.0% contained not only the barcode but also a good amplicon alignment.

### *Barcode Calling Errors*

In the simple positive control experiment, we barcoded a single sample with two different PacBio barcode pairs and ran the two barcode pairs separately in two independent runs. Barcode F0 was used in the first run and barcode F1 was used in the second run. False barcodes were estimated because only a single barcode pair was used per run and, therefore, any other barcode calls were false. In the first run we saw 3 false calls (0.035%) and in the second run we saw 3 false calls (0.027%). Note that some of these false call hits were *perfect* matches to an alternate barcode possibly indicating a small level of contamination in the primer mixes and not false barcode calls.

### *Barcode Score and Amplicon Alignment*

The majority of CCS reads (75.6%) contained exactly the identifying barcode pairs and the targeted amplicons as expected. Some reads showed shorter amplicon hits or poor barcode hits (see Figure 4)

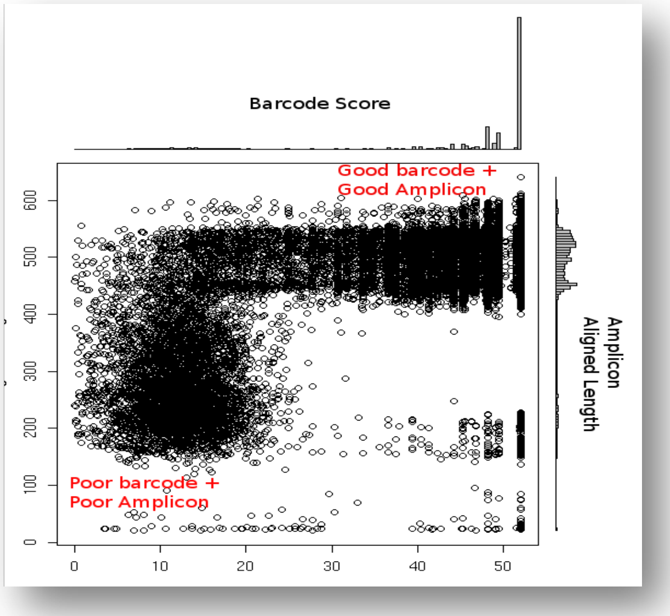


**Figure 4: Barcode Score Versus Amplicon Aligned Length. This plot shows that a majority of reads have proper amplicon lengths and high-scoring barcode calls (upper-right corner).**

This may have been due to improper amplification during PCR or other artifacts.

### *Barcode and Amplicon Coverage*

As can be seen in Figure 5, barcode coverage was fairly uniform. For the 48-multiplex barcode experiment, 47 of the 48 PacBio barcodes had coverage that was within 24% of the mean. Barcode F23 had lower coverage. Previous experiments (not detailed here) showed that there does not appear to be any strong systematic bias as this barcode had average coverage in those independent experiments. Therefore, we hypothesize that this was most likely due to an error in the pooling. With this outlier removed (965-fold coverage), the maximum to minimum coverage ratio is 1.59 (2710/1706).

Additionally, fully stratifying the reads into the 48 samples by 49 amplicons (2352 distinct products) revealed varying coverage (Figure 6). On average, each barcode and amplicon pair was covered 44.82 times. At the lowest end of coverage, there was a single barcode+amplicon pair that was not observed in the sequencing data. Overall, there does not appear to be much bias in the coverage over the normal distribution of counts.

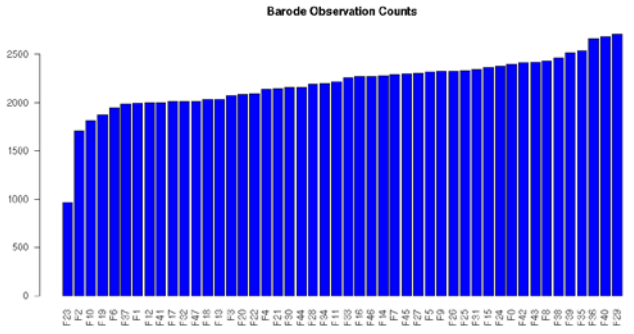


**Figure 5: Barcode Uniformity. Counts of the number of times each barcode is identified in the 48-barcode experiment.**

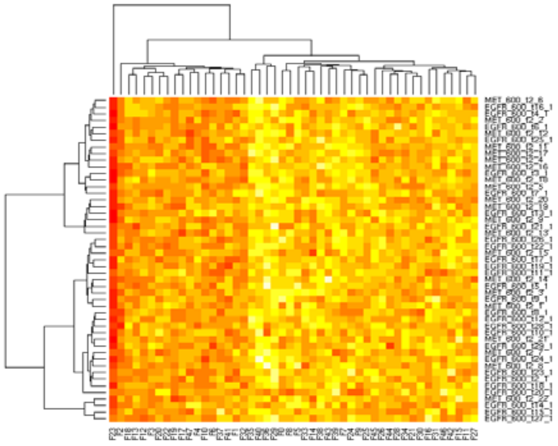


**Figure 6: Barcode and Amplicon Coverage Heatmap. Coverage counts after fully stratifying the reads into the 48 samples (columns) by 49 amplicons (rows). There are 2352 distinct sequence products. Dark red indicates lower coverage while light yellow corresponds to higher coverage.**

## Conclusion

Barcoding samples prior to SMRT Sequencing allows multiplexed sample preparation and sequencing. For target amplicons, with average lengths of 500-bases, we observed uniform barcode coverage with almost no false barcode calls.

The list of PacBio barcode sequences, and software for barcode identification in CCS reads is available from the PacBio DevNet website[2]. In addition, software for barcode identification from continuous long reads (without CCS) is easily producible. The current work used the Fluidigm Access Array System as a platform for multiplex target amplification; however, these barcodes can be easily incorporated using alternate target enrichment strategies.

## References

1. Travers, K., et al. Flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Research (2010), e159.

2. http://www.pacbiodevnet.com

3. Eddy SR. “Profile hidden Markov models”. Bioinformatics (1998) 14 (9): 755-763. doi:10.1093/bioinformatics/14.9.755, (1998).

4. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. "Hidden Markov models in computational biology. Applications to protein modeling". J. Mol. Biol. 1501-1531 doi:10.1006/jmbi.1994.1104. PMID 8107089, (February 1994) 235(5).

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