Targeted SMRT® Sequencing and Phasing Using Roche NimbleGen's SeqCap EZ Enrichment

Steve Kujawa[1], Lawrence Hon[1], Paul Peluso[1], Kristi Spittle[1], John Harting[1], Denise Raterman[2], Todd Richmond[2], Dan Burgess[2]
[1]Pacific Biosciences, Menlo Park, CA [2]Roche NimbleGen, Madison, WI

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 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 region, but are 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 regions. While the SeqCap EZ technology is typically used to capture 200 bp fragments, we demonstrate that 6 kb fragments can also be utilized to enrich for long fragments that extend beyond the targeted capture site and well into (and often across) the flanking intronic regions. When combined with the long reads of SMRT Sequencing, multi-kilobase regions of the human genome can be phased and variants detected in exons, introns and intergenic regions.

PacBio - NimbleGen Workflow

NimbleGen Capture Workflow

Shear gDNA to 6 kb

↓

Size-Select Sheared DNA

↓

Add SeqCap EZ Adapters

↓

Pre-Capture Amplification

↓

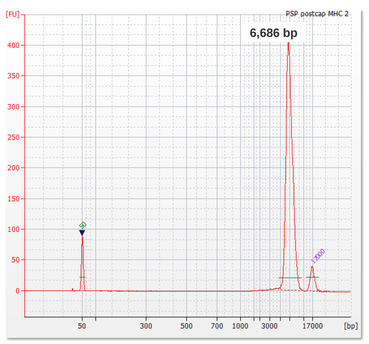
Hybridize Library w/ SeqCap EZ Probes

↓

Bead Capture & Washing of Probe-Bound Fragments

↓

Post-Capture Amplification

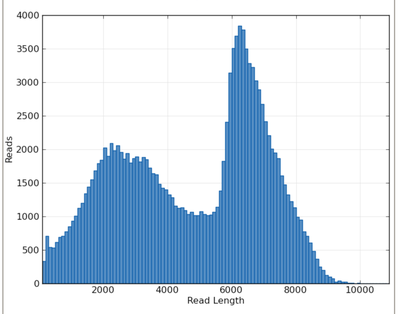


After the post-capture amplification step, Sample 1 exhibits a well-defined peak at 6,700 bp. The sample is now ready for SMRTbell library prep and sequencing.

PacBio Sequence & Analysis Workflow

SMRTbell™ Library Prep & Sequencing

↓



Mapped reads of insert plot from Sample 1 shows a primary peak around 6 kb and a secondary peak around 2 kb.

↓

Generate Reads of Insert / CCS (min 0 full pass, 75% pred. accuracy)

↓

Align to hg19

↓

Phasing with SAMtools

↓

Bin Reads by Haplotype

↓

Use Quiver to Generate Consensus per Haplotype

↓

Into Tertiary Analysis Tools

Samples Used in This Study

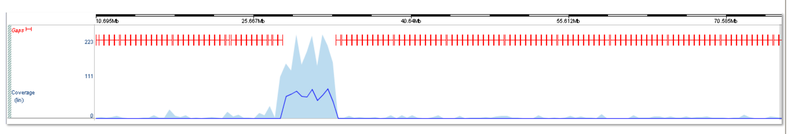
Enrichment & Sequencing Statistics from Human Samples

Sample ID	Sample	SeqCap EZ Design	Sequencing Platform	Average Fragment Size (bp)	# SMRT Cells or Runs	Mapped Reads of Insert Mean (bp)	% Reads on Target	Enrichment Factor
1	PacBio Diploid	Human MHC (5 Mb region)	PacBio® RS II	6,000	4	4,788	48.4	600x
2	PacBio Diploid	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 101 bp - paired end run	N/A	63.9	1400x

Samples were enriched with the Roche Nimblegen SeqCap EZ Human MHC or Comprehensive Cancer design and then sequenced. For the three samples sequenced by PacBio, an average of 61% of the reads were on target, representing an average enrichment factor of 1,300x.

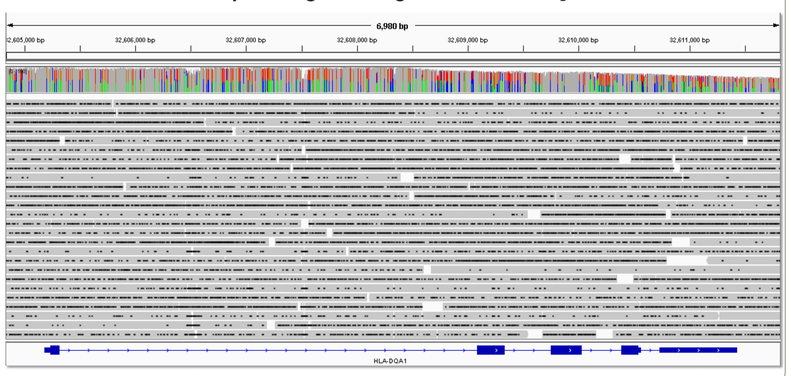
Targeting Sequencing of MHC Region

PacBio Sequencing Coverage of the MHC Region



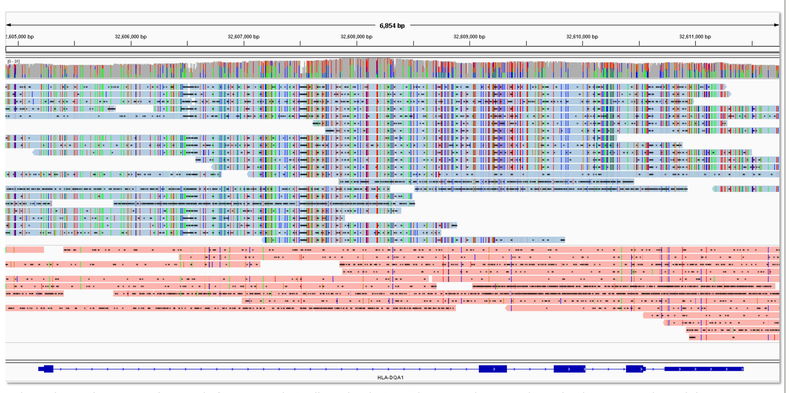
PacBio sequencing coverage plot of Chromosome 6 from Sample 1. The plot shows solid coverage of the MHC region targeted by the SeqCap EZ Human MHC design, with minimal off-target coverage in the flanking regions.

PacBio Sequencing Coverage of the HLA-DQA1 Gene



PacBio sequencing reads of insert from Sample 1 aligned to hg19. The plot shows even coverage across the entire gene.

PacBio Phasing of the HLA-DQA1 Gene



Phased PacBio sequencing reads from Sample 1 aligned to hg19. These reads were phased using SAMtools and then separated and grouped by haplotype (blue for one haplotype, pink for the other). 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. The haplotyped sequences from this gene differed from the haplotyped sequences of a separate PacBio de novo assembly of the same sample by only one base pair.

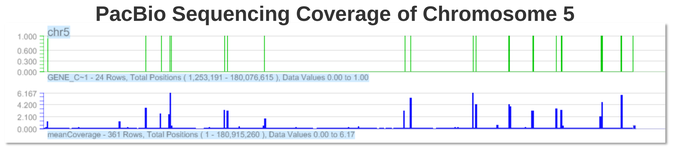
Typing of the HLA-DQA1 Gene

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

Sample 1 was enriched using the Roche NimbleGen SeqCap EZ method and subsequently phased and typed at the HLA-DQA1 locus. The same sample was independently typed using sequence data from a Sanger-based assembly. Each method produced the identical type.

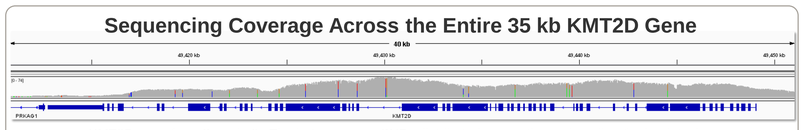
Targeted Sequencing of Cancer Genes

PacBio Sequencing Coverage of Chromosome 5



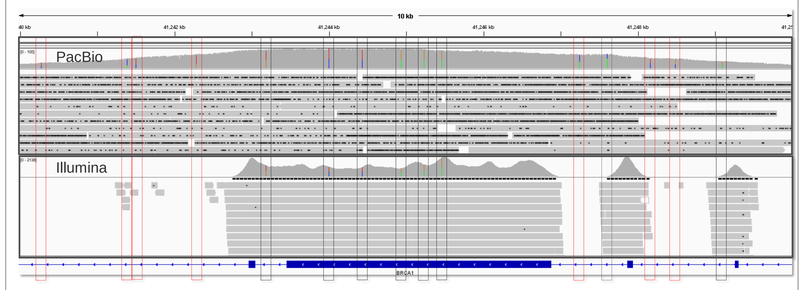
The top track (green) indicates the location of the genes targeted by the Comprehensive Cancer design. The lower track (blue) indicates the PacBio sequencing coverage. Chromosome 5 is representative of the coverage in the other chromosomes.

Sequencing Coverage Across the Entire 35 kb KMT2D Gene



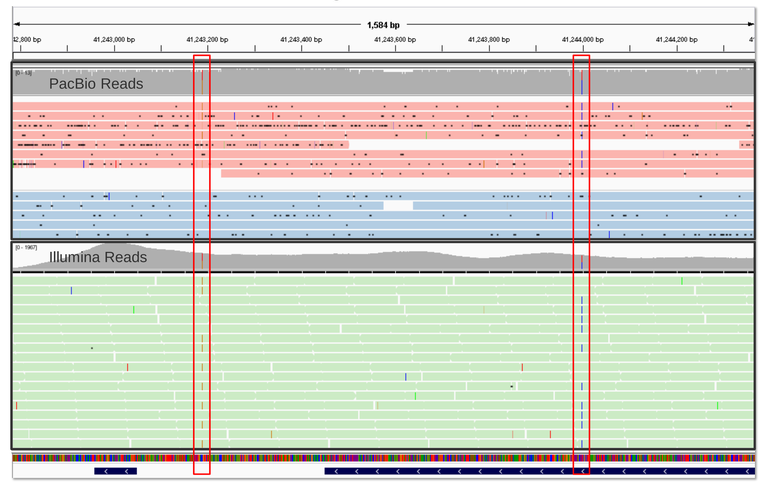
KMT2D gene from Sample 2. The heterozygous SNPs distributed across the region allow for haplotyping.

PacBio Detects Additional Variants in the BRCA1 Gene



Sequencing results from Samples 3 and 4 (NA12762 enriched with the Comprehensive Cancer design) comparing PacBio to Illumina. Across this 10 kb region of BRCA1, Illumina detected 8 SNPs. PacBio was able to detect these same SNPs (black boxes) and an additional 7 SNPs that were either not detected, or detected with insufficient coverage, by Illumina (red boxes).

Phasing SNPs in BRCA1



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 100 bp Illumina reads.

Summary and Resources

Summary:

- 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.

PacBio Targeted Sequencing Information Available Here:
http://www.pacificbiosciences.com/applications/target/

Data & Analysis Information Available Here:
https://github.com/lhon/targeted-phasing-consensus