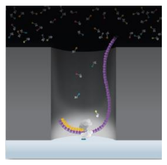
Multiplex Target Enrichment Using Barcoded Multi-Kilobase Fragments and Probe-Based Capture Technologies

Jenny Gu, Kevin Eng, Anand Sethuraman, and Steve Kujawa
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

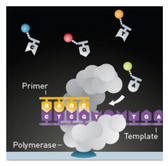
ABSTRACT

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 haplotyes. In addition, SMRT® Sequencing offers the lowest GC-bias and can sequence through repetitive regions. We demonstrate the additional insights possible by using in-depth long read capture sequencing for key immunology, drug metabolizing, and disease causing genes such as HLA, filaggrin, and cancer associated genes.

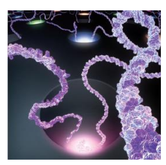
SMRT® Technology



Zero-Mode Waveguides



Phosopholinked Nucleotides



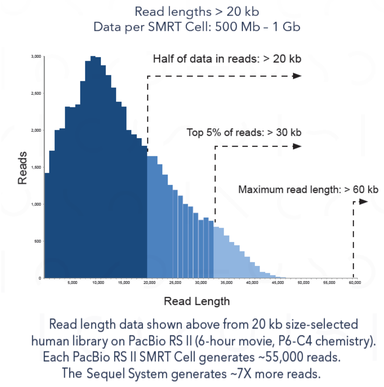
Up to a million ZMWs / SMRT Cell

SMRT Cells containing up to a million ZMWs are processed on PacBio® Systems which simultaneously monitor each of the waveguides in real time.

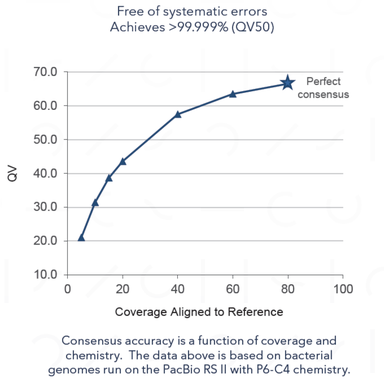
SMRT Sequencing Advantage

- Longest Read Lengths
- Highest Consensus Accuracy
- Uniform Coverage
- Simultaneous Epigenetic Characterization

Typical Results



Read length data shown above from 20 kb size-selected human library on PacBio RS II (6-hour movie, P6-C4 chemistry). Each PacBio RS II SMRT Cell generates ~55,000 reads. The Sequel System generates ~7X more reads.

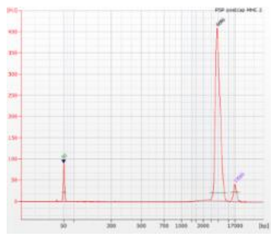


Consensus accuracy is a function of coverage and chemistry. The data above is based on bacterial genomes run on the PacBio RS II with P6-C4 chemistry.

Multiplex Target Capture Workflow

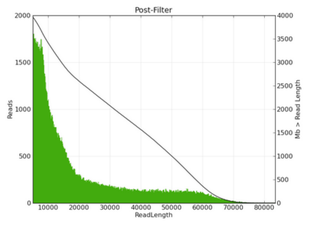
Multiplex & Capture

- Shear gDNA to 7 kb
- Linear Barcoded Adapter Ligation
- Pre-Capture Amplification
- Size-Selection
- Hybridization w/ Probes
- Bead Capture & Washing
- Post-Capture Amplification



SMRT Sequencing

- SMRTbell™ Library Prep
- Size-Selection



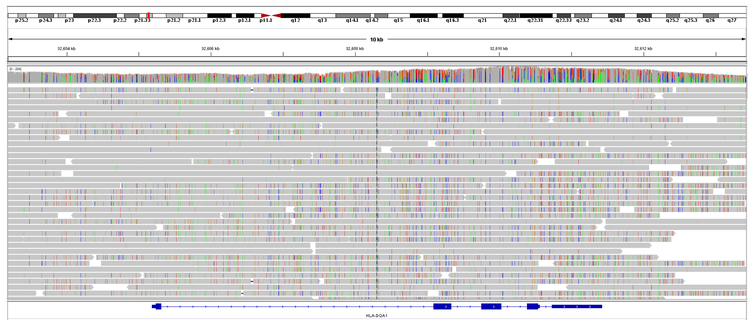
(Sequencing results on RSII)

Data Analysis

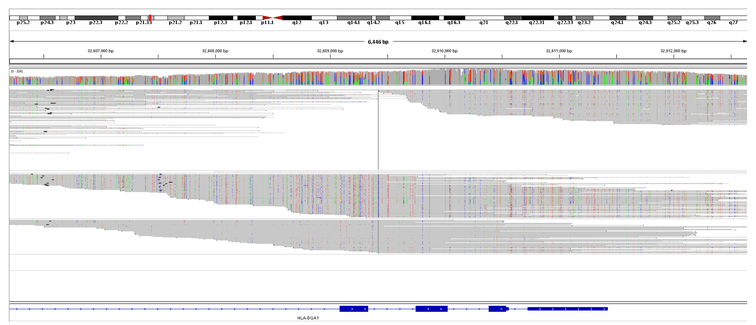
- Generate Reads of Insert
- Demultiplex Samples
- Align to Genome Reference
- Tertiary Analysis Tools

SNP Phasing with Direct Data Evidence

12-Plex MHC capture results show even coverage and phased SNPs to distinguish haplotypes.



Coverage for HLA-DQA1. PacBio SMRT Sequencing achieves full gene coverage and spans intergenic regions

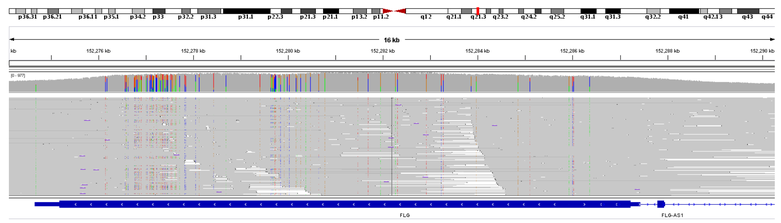


Haplotype phasing by SNP sorting in IGV Viewer. Aligned reads are grouped and sorted by SNP position : Chr 6:32,609,427 bp (C:56%, T:43%).Two distinct haplotypes with different SNP profiles can be observed.

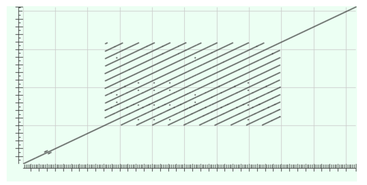
Viewed using IGV Viewer updated with features for PacBio long reads.

Resolve Complex Structural Variation

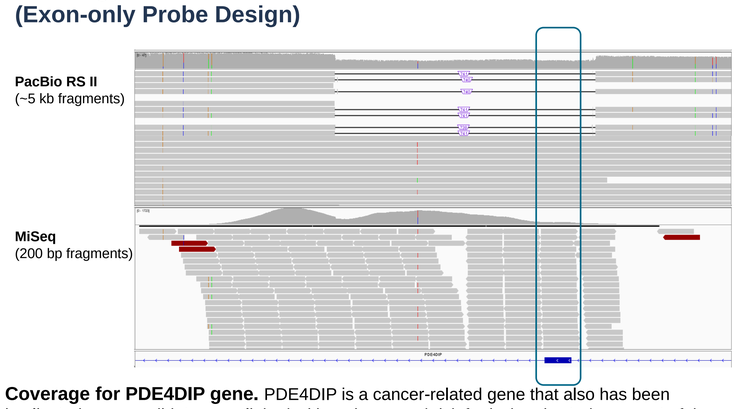
Filaggrin Gene Capture with Neurology Gene Panel (Exon-Only Probe Design)



Coverage for Filaggrin Gene. PacBio reads span 12 repeats in exon 1 of the Filaggrin gene. Dot plot (right) between the two phased alleles show complexity of this repetitive region. The number of Filaggrin repeats varies between 10, 11, and 12 copies in the human population, and this variation has been associated with dry skin phenotype (Ginger R.S., 2005), as well as atopic dermatitis (Brown S.J., 2012).



Oncology Gene Panel Capture of NA12762 (Exon-only Probe Design)



Coverage for PDE4DIP gene. PDE4DIP is a cancer-related gene that also has been implicated as a candidate gene linked with an increased risk for ischemic stroke as part of the NHLBI Exome Sequence Project (Auer PL, 2015). Illumina misses a heterozygous 740 bp deletion containing an entire exon!

Key Benefits

- Full gene capture to characterize introns, exons, and intergenic regions.
- Improve variant calling with less mapping ambiguity.
- Phase SNPs over >7kb regions to distinguish alleles and identify haplotypes.
- Detect homologs and resolve duplicated genes.
- Characterize genes with repetitive elements and complex structural variation.
- Opportunity for improved diagnostic yield for diseases difficult to identify with standalone SNVs.
- Fully reconstruct clinically relevant regions.

Reference & Resources

Ginger RS, et. al. Filaggrin repeat number polymorphism is associated with a dry skin phenotype. *Arch Dermatol Res.* (2005)
Brown SJ, et. al. Intragenic copy number variation within filaggrin contributes to the risk of atopic dermatitis with a dose-dependent effect. *J Invest Dermatology* (2012)
Auer PL, et. al. Rare and coding region genetic variants associated with risk of ischemic stroke: The NHLBI Exome Sequence Project. *JAMA Neurol* (2015)
Application Note: Multiplex target enrichment using barcoded multi-kilobase fragments and probe-based capture technologies.
Blog: Precision medicine review highlights need for accuracy and comprehensiveness in genome sequencing.
Towards precision medicine. Euan A. Ashley. *Nature Review Genetics* **17**, 507-522 (2016)
IGV Viewer with features for PacBio data.
PacBio Targeted Enrichment Solutions

Acknowledgements

The authors would like to thank our collaborators at Roche Nimblegen and IDT technologies who have helped to generate this data.