High-throughput SMRT Sequencing of Clinically Relevant Targets

S. Ranade[1], **L. Aro[1],** J. Harting[1],W. Rowell, I. McLaughlin[1], C. Heiner[1], P. Baybayan[1], A. Toepfer[1], B. Bowman[1], J. Ziegle, M. Seetin[5], M. Weiand[1], P.W.L. Tai[2], G. Gao[2], R. Hall[1]
[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025
[2]University of Massachusetts Medical School, 386 Plantation Street, Worcester, MA 01605

Abstract

Targeted sequencing with Sanger as well as short read based high throughput sequencing methods is standard practice in clinical genetic testing. However, many applications beyond SNP detection have remained somewhat obstructed due to technological challenges. With the advent of long reads and high consensus accuracy, SMRT Sequencing overcomes many of the technical hurdles faced by Sanger and NGS approaches, opening a broad range of untapped clinical sequencing opportunities.

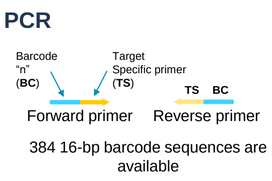
Flexible multiplexing options, highly adaptable sample preparation method and newly improved two well-developed analysis methods that generate highly-accurate sequencing results, make SMRT Sequencing an adept method for clinical grade targeted sequencing. The Circular Consensus Sequencing (CCS) analysis pipeline produces QV 30 data from each single intra-molecular multi-pass polymerase read, making it a reliable solution for detecting minor variant alleles with frequencies as low as 1 %. Long Amplicon Analysis (LAA) makes use of insert spanning full-length subreads originating from multiple individual copies of the target to generate highly accurate and phased consensus sequences (>QV50), offering a unique advantage for imputation free allele segregation and haplotype phasing.

Here we present workflows and results for a range of SMRT Sequencing clinical applications. Specifically, we illustrate how the flexible multiplexing options, simple sample preparation methods and new developments in data analysis tools offered by PacBio in support of Sequel System 5.1 can come together in a variety of experimental designs to enable applications as diverse as high throughput HLA typing, mitochondrial DNA sequencing and viral vector integrity profiling of recombinant adeno-associated viral genomes (rAAV).

Barcoding Options for Sample Multiplexing

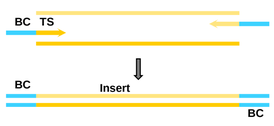
1. Barcoded Primers

PCR



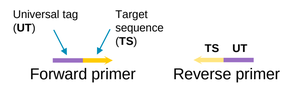
384 16-bp barcode sequences are available

Incorporate barcoded target-specific primers

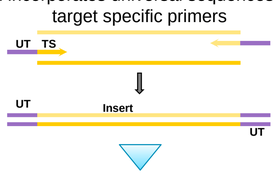


2. Barcoded Universal Primers

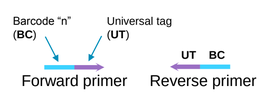
PCR 1



First PCR incorporates universal sequences tagged to target specific primers

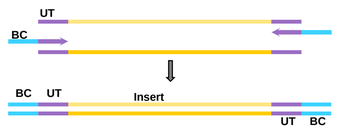


PCR 2



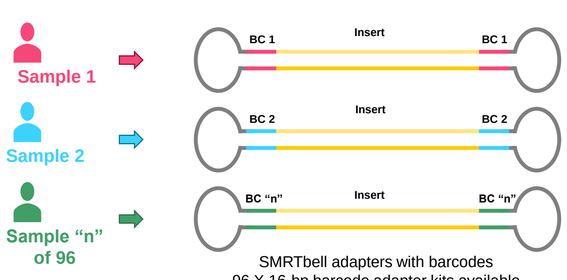
384 16-bp barcode-universal primer sequences are available

Second PCR incorporates barcoded universal primers



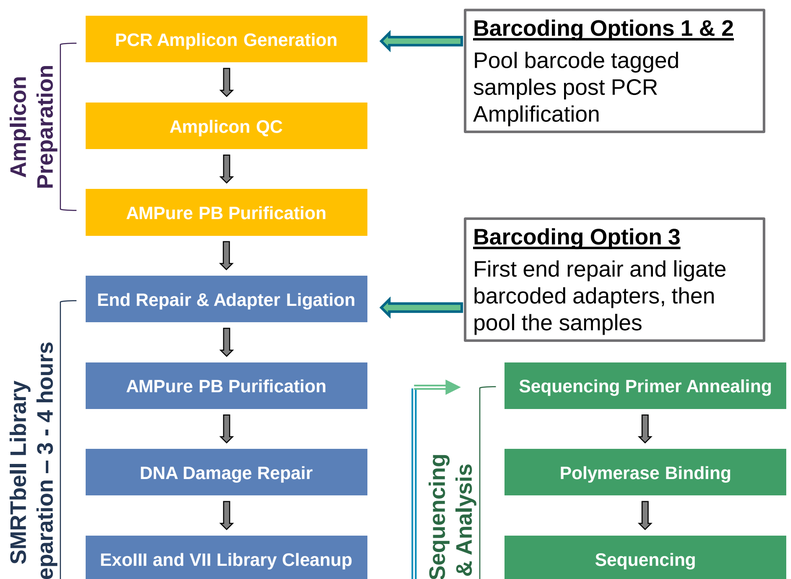
3. Barcoded Adapters

Adapter Ligation (SMRTbell Library Preparation)



SMRTbell adapters with barcodes
96 X 16-bp barcode adapter kits available

Sample Preparation for Multiplex Targeted SMRT Sequencing

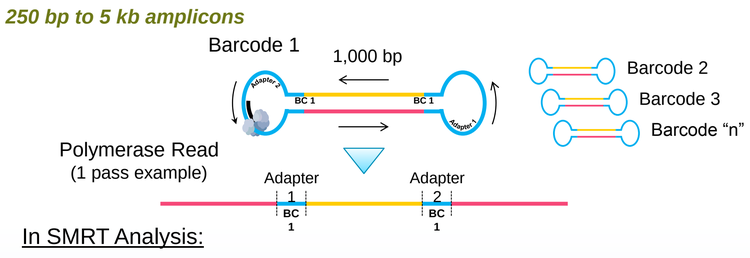


Non-specific amplicons are removed post PCR using PB AMPure bead purification or gel purification. Alternately, BluePippin or SageELF size selection may be used for SMRTbell library purification.

Analysis Workflows for Targeted SMRT Sequencing

Circular Consensus Sequencing (CCS)

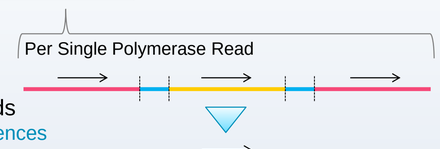
250 bp to 5 kb amplicons



In SMRT Analysis:

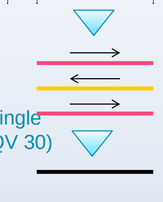
1. Pre-Process Filtering (Analysis Parameters)

2. Demultiplex



3. Generate Subreads
   Remove adapter sequences

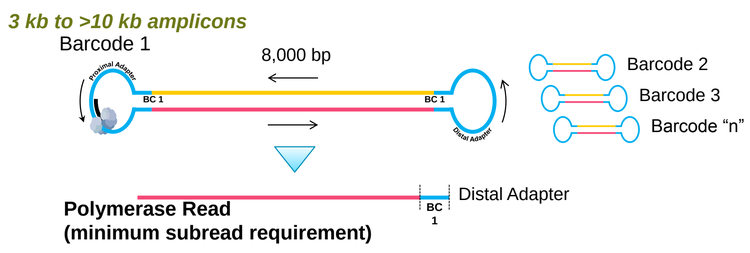
4. Generate Circular Consensus
   Combining multiple passes from a single molecule results in high accuracy (QV 30)



5. Accuracy: **>99.99% accuracy at 40-fold coverage**.
   CCS reads alignment

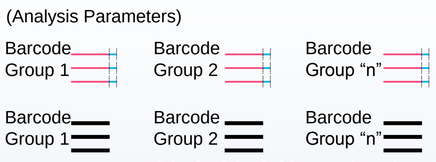
Long Amplicon Analysis (LAA)

3 kb to >10 kb amplicons



In SMRT Analysis: *De novo* LAA
In Command Line Analysis: LAA with Guided Clustering (LAAgc)

1. Pre-Process Filtering (Analysis Parameters)



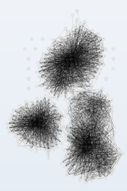
2. Demultiplex

3. Generate Subreads

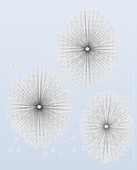
De novo LAA
(Unguided Waterfall Clustering)

LAA with Guided Clustering (LAA-gc)*NEW**
*Dynamic alignment and phasing of locus specific subreads (Step 3 to 6)

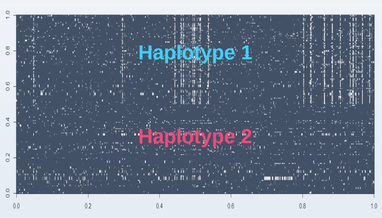
4. Overlap
   Sequencing homology



5. Clustering
   Large-scale differences



6. Phasing
   Small-scale differences



7. Phased Haplotype Consensus
   Combining subreads from multiple ZMWs results in high accuracy **>99.999% accuracy at 40-fold coverage**.

8. Post-Process Filtering (noise and chimeras)
   Consensus reads ready for alignment

HLA Sequencing

High throughput imputation free HLA Typing

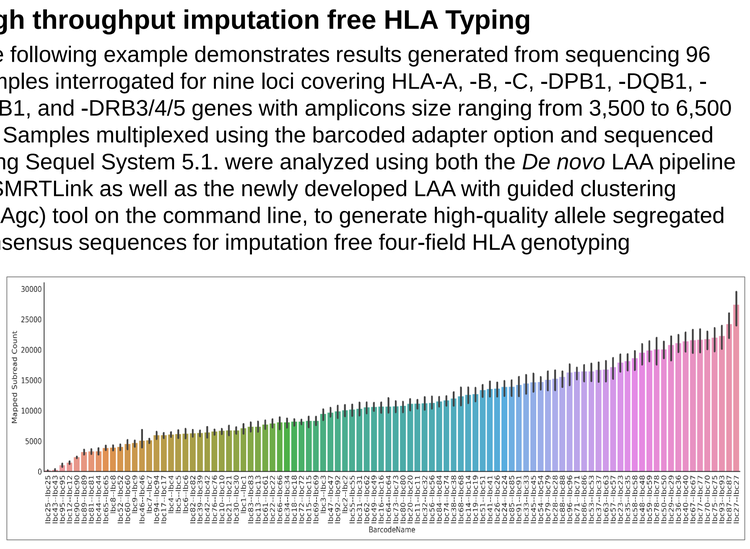
The following example demonstrates results generated from sequencing 96 samples interrogated for nine loci covering HLA-A, -B, -C, -DPB1, -DQB1, -DRB1, and -DRB3/4/5 genes with amplicons size ranging from 3,500 to 6,500 bp. Samples multiplexed using the barcoded adapter option and sequenced using Sequel System 5.1. were analyzed using both the *De novo* LAA pipeline in SMRTLink as well as the newly developed LAA with guided clustering (LAAgc) tool on the command line, to generate high-quality allele segregated consensus sequences for imputation free four-field HLA genotyping

A

B

A. Imbalances in mapped subreads for all loci for each sample compared across 96 samples
B. Imbalances in mapped subreads across the nine loci within a sample due to PCR or amplicon pooling related biases

C



	Denovo	Guided
poorAlignment	25	29
duplicates	9	13
dropped	136	110

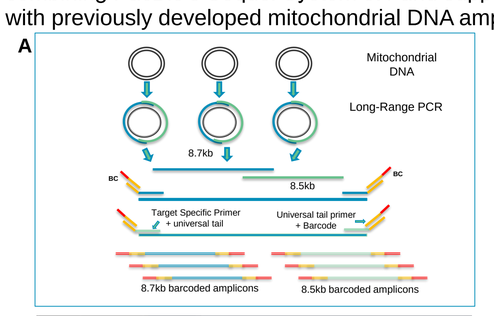
C. 96 samples X 9 loci were analyzed using both the *De novo* LAA as well as the newly developed LAA with guided clustering approach:
- *De novo* pipeline randomly seeds a sampling of subreads per barcode into the analysis pipeline and is affected by PCR imbalances as well as amplicon and or sample pooling biases
- Guided Clustering ensures maximum available coverage per sample-locus allele into the LAA pipeline, informatically compensating for sample preparation issues
- 20% decrease in dropped alleles observed (110 Vs. 136 dropped alleles in guided Vs. *De novo* analysis)
- Of the 1235 expected alleles, the guided clustering method did not miss any alleles with ≥ 50-fold mapped subreads available in the entire data
- >1M mapped subreads for the whole cell

Mitochondrial DNA Sequencing

Targeted sequencing of complete mitochondrial genomes

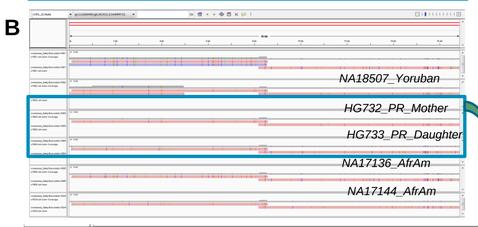
Mitochondrial DNA mutations make important contributions to an array of human diseases and get routinely tested for clinical, ancestry and forensic analysis. Several recent publications have demonstrated the advantage of SMRT Sequencing for mitochondrial DNA somatic & germline variants[1,2]. The following results demonstrate easy segue from NGS to SMRT Sequencing, by simply combining PacBio's Sequel System 5.1 and supporting multiplexing products with previously developed mitochondrial DNA amplification protocol[3]

A



A. The Segue from NGS to SMRT Sequencing:
   - Barcoded universal primer for multiplexing
   - Two overlapping amplicon targets for complete mitochondrial genome coverage[4]

B



B. 24 individuals sequenced in duplicate using two different BCs. The samples included:
   - Mother/ Daughter pairs

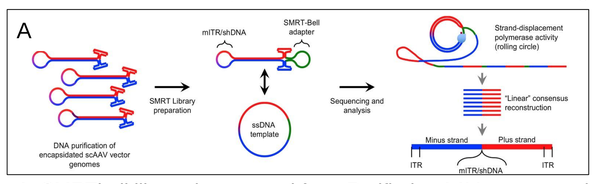
C

C. Conserved mitochondrial variants between mother and daughter in comparison to Cambridge Reference Seq ((CRS) = white male)

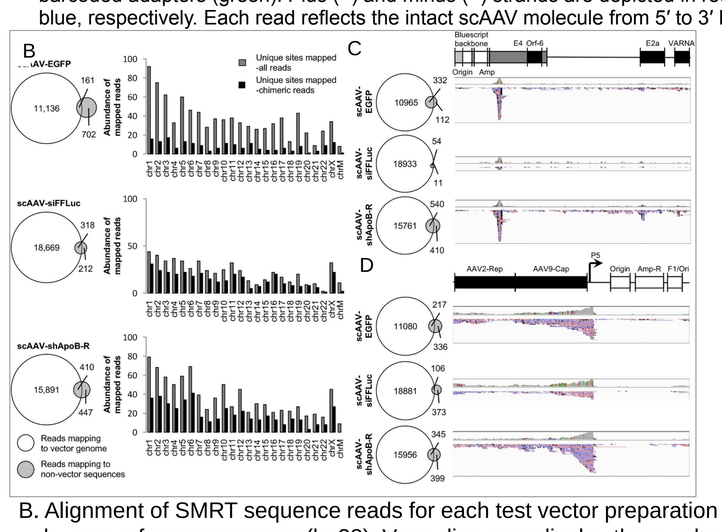
Recombinant Adeno-Associated Virus (rAAV) Vector Integrity Profiling

rAAV Population Genome Sequencing

The following results demonstrate the ability of AAV-GPseq[4] a comprehensive rAAV genome profiling method for clinical grade QC of gene therapy vectors facilitated by SMRT Sequencing. Packaged genomes were comprehensively profiled as single intact molecules and directly assessed for vector integrity without extensive sample preparation to establish clinical grade vector QC.



A. SMRTbell library is prepared from Purified scAAV genome using single-adaptering to barcoded adapters (green). Plus (+) and minus (−) strands are depicted in red and blue, respectively. Each read reflects the intact scAAV molecule from 5′ to 3′ ITR.



B. Alignment of SMRT sequence reads for each test vector preparation to the human reference genome (hg38). Venn diagrams display the number of reads mapping to the vector genome (white circles) and to the human genome (gray circles). Histograms display the abundance of uniquely mapped sites on each chromosome (gray bars) and the abundance of unique sites that are mapped by reads that also contain vector genome sequences (chimeras, black bars). (C) Alignment data of reads mapping to the Ad-helper plasmid and (D) to the AAV-Rep/Cap plasmid. (Venn diagrams again display reads mapping to the vector genome (white) and to either the Ad-helper or AAV-rep/cap plasmids (gray). Right, IGV displays showing individual read alignments to their respective references diagrammed above as a linear strand. Reads mapping in the forward and reverse orientations are indicated in red and blue, respectively.

Conclusion

- Targeted amplicon sequencing is a fully supported application on the Sequel System 5.1
- Flexible multiplexing options enable cost effective solutions for a broad range of clinical applications
- Two analysis workflows, CCS and LAA, support target insert sizes from 250 bp to >10 kb
- Highly accurate results: >99.99% accuracy for CCS and >99.999% accuracy for LAA, both at 40-fold coverage allow both inter and intramolecular variation analysis

References & PacBio Resources

1. Weerts M.J.A. et al. (2018). Sensitive detection of mitochondrial DNA variants for analysis of mitochondrial DNA enriched extracts from frozen tumor tissue. SCIENTIfIC REPOrTS | *8:2261* | DOI:10.1038/s41598-018-20623-7
2. Li M. et al. (2016). High frequency of mitochondrial DNA mutations in HIV infected treatment-experienced individuals 90 HIV Medicine, DOI: 10.1111/hiv.123
3. Clarke A. et al. (2014). From cheek swabs to consensus sequences: an A to Z protocol for high-throughput DNA sequencing of complete human mitochondrial genomes. BMC Genomics,15: 68
4. Tai P.W.L. et al. (2018). Adeno-Associated Virus Genome Population Sequencing Achieves Full Vector Genome Resolution and Reveals Human-Vector Chimeras. Molecular Therapy: Methods & Clinical Development open-access article (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Targeted Sequencing
PacBio Webpage: www.pacb.com/applications/targeted-sequencing/

Barcoding
Product Note, Barcoding Solutions: Multiplexing Amplicons Up To 10 kb
Document: SMRT Analysis Barcoding Overview

Circular Consensus Sequencing
Tutorial: Circular Consensus Sequence analysis application

Long Amplicon Analysis
Tutorial: Long Amplicon Analysis application

Acknowledgements

The authors would like to thank everyone who has helped in samples and data generation for this poster.