Targeted Sequencing and Chromosomal Haplotype Assembly Using Cergentis TLA Technology with SMRT® Sequencing

Introduction

Conventional, PCR-based targeted sequencing methodologies are impractical for detecting and phasing single nucleotide and structural variants over regions that are tens of kilobases in length.

The Targeted Locus Amplification (TLA) Technology[1] from Cergentis enables the targeted, hypothesis-neutral, amplification of any genomic locus of interest over 50 kb using just one primer pair complementary to a short locus-specific sequence. TLA is a strategy to selectively amplify complete loci on the basis of crosslinking physically proximal sequences. Unlike other targeted sequencing methods, TLA works without prior detailed locus information, as one primer pair is sufficient to amplify tens to hundreds of kilobases of DNA surrounding that locus. In a separate application of TLA, the unamplified template can be used for genome-wide phasing and assembly. TLA enables targeted sequencing and detection of single nucleotide and structural variants in genes or regions of interest.

Single Molecule, Real-Time (SMRT®) Sequencing provides high consensus accuracy and long read lengths. As such, it enables end-to-end sequencing of multi-kilobase TLA amplicons or unamplified TLA templates. The combination of Cergentis' TLA and SMRT Sequencing technologies allows for sequencing and haplotyping of individual genes, chromosomes and genomes.

TLA Method

In the TLA sample preparation method, the genomic DNA is first crosslinked. Because crosslinking occurs preferentially between sequences in close physical proximity, sequences predominantly from the same locus are crosslinked. The crosslinked DNA is then fragmented, religated and de-crosslinked. The resulting TLA template consists of long fragments of DNA comprising religated fragments originating from the same locus (Figure 1). At this point the procedure changes depending upon the desired application.

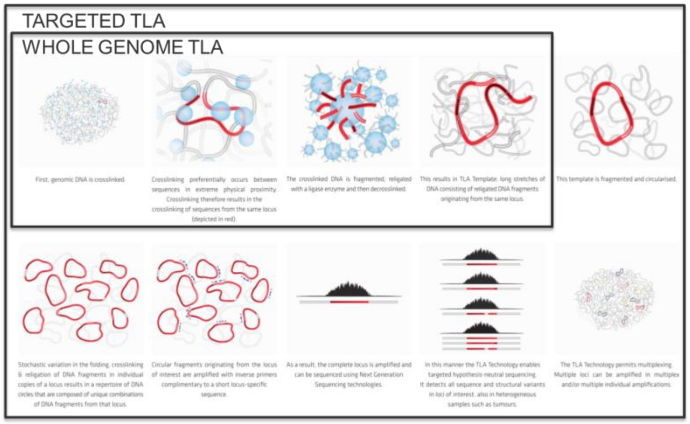
TARGETED TLA
WHOLE GENOME TLA
First, genomic DNA is crosslinked.
Crosslinking preferentially occurs between sequences in extreme physical proximity. Crosslinking therefore results in the crosslinking of sequences from the same locus (depicted in red).
The crosslinked DNA is fragmented, religated with a ligase enzyme and then decrosslinked.
This results in TLA Template: long stretches of DNA consisting of religated DNA fragments originating from the same locus.
This template is fragmented and circularised.
Stochastic variation in the folding, crosslinking & religation of DNA fragments in individual copies of a locus results in a repertoire of DNA circles that are composed of unique combinations of DNA fragments from that locus.
Circular fragments originating from the locus of interest are amplified with inverse primers complimentary to a short locus-specific sequence.
As a result, the complete locus is amplified and can be sequenced using Next Generation Sequencing technologies.
In this manner the TLA Technology enables targeted hypothesis-neutral sequencing. It detects all sequence and structural variants in loci of interest, also in heterogeneous samples such as tumours.
The TLA Technology permits multiplexing. Multiple loci can be amplified in multiplex and/or multiple individual amplifications.

Figure 1 – During the TLA process DNA is crosslinked, fragmented, religated and de-crosslinked. For Whole Genome TLA, the linear templates are sequenced. For Targeted TLA, amplicons are generated from circular templates and sequenced.

Targeted TLA

When using the targeted TLA approach, the linear TLA templates are formed into DNA circles that are composed of unique combinations of DNA fragments from that locus. Inverse primers are designed for the locus of interest and only circles containing the complementary region are amplified. As a result, the complete locus is amplified and the resulting amplicons can be sequenced.

Whole Genome TLA

When using the Whole Genome TLA approach, the linear fragments are not circularized. Instead, the linear TLA templates (Panel D in Figure 1), are directly converted into SMRTbell™ templates, size-selected and sequenced. Each individual TLA template contains fragments that may be ten to hundreds of kilobases apart in base space, but are all close enough to be cross-linked. Because it is not a targeted approach, this strategy results in long range information that covers the entire genome.

Materials and Methods

SMRTbell libraries were created from the TLA templates following PacBio's published sample preparation procedures (with BluePippin™ system size-selection and additional damage repair for the Whole Genome TLA Template) and sequenced on the PacBio® RS II system. For targeted TLA phasing, Reads of Insert were generated from SMRT Analysis v2.3 using default parameters and iteratively mapped using BLASR; SNPs were *de novo* called using SAMtools and BCFtools. For Whole Genome TLA analysis, BAM (PacBio shotgun) and VCF files were obtained from GIAB, and CCS parameters were adjusted to obtain one read per molecule (0 full pass, 75% minimum quality). HAPCUT was then used to phase selected regions, incorporating whole genome PacBio shotgun data for whole-chromosome phasing.

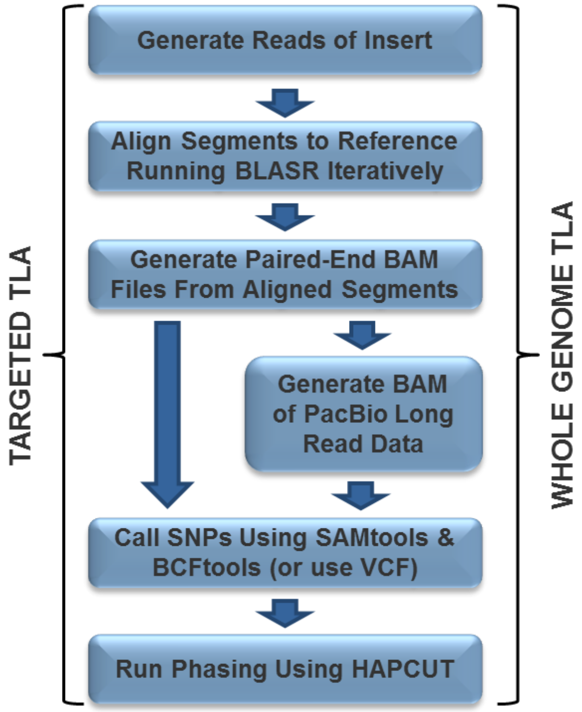
TARGETED TLA
Generate Reads of Insert
Align Segments to Reference Running BLASR Iteratively
Generate Paired-End BAM Files From Aligned Segments
Generate BAM of PacBio Long Read Data
Call SNPs Using SAMtools & BCFtools (or use VCF)
Run Phasing Using HAPCUT
WHOLE GENOME TLA

Figure 2 – PacBio data analysis workflow for Targeted and Whole Genome TLA templates.

Here, Targeted TLA was used to enrich for the BRCA1 gene in NA12878. After generating TLA template circles, a single primer pair at (hg19) Chr17:41237179-41236511 (located ~ 40 kb from the start of the ~ 81 kb long BRCA1 gene) was used to amplify this locus. The resulting amplicons were then sequenced on the PacBio RS II system.

Next, the data was combined with whole genome shotgun PacBio long reads from NA12878 to explore chromosomal-scale haplotype assembly.

In the final experiment, Whole Genome TLA was used on GM24385 and the data was combined with whole genome shotgun PacBio long reads (from the same sample) to explore whole genome haplotype assembly across all chromosomes.

Data Sets Used

Sample	Prep Method	Library Size	Sequencing Chemistry	Fold Coverage
GM12878	Targeted TLA	2 kb	P6-C4	Variable with peak at BRCA1
NA12878	Whole Genome Shotgun	~ 7 kb	P5-C3 and older	~ 40X
GM24385	Whole Genome TLA	10 kb	P6-C4	0.8X
NA24385	Whole Genome Shotgun	>10 kb	P6-C4	~50X

Table 1 – Targeted TLA of GM12878 was used for BRCA1 targeting and Chromosome 17 haplotyping. Whole Genome TLA of GM24385 was used for whole genome haplotyping.

Results

Targeted TLA of BRCA1

Targeted TLA templates generated ~ 2 kb reads of inserts with ~4 segments per read. The schematic in Figure 3 illustrates the general structure of the reads and how they mapped to the reference. Mapped reads generated with the Targeted TLA to BRCA1 fully cover the BRCA1 region (Figure 4), with heterozygous SNPs clearly visible from the TLA data (Figure 5), allowing excellent phasing performance (Table 2). As can be seen, the 81 kb length of BRCA1 is represented by a single haplotype block (haplotyping was validated against a reference dataset).

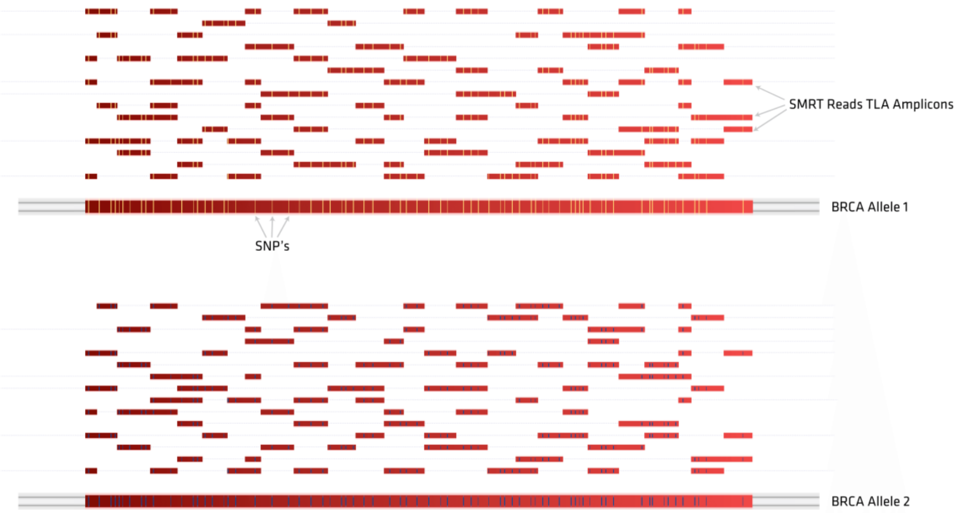

Figure 3 – Schematic representation of the TLA reads and how they map to the reference. Heterozygous SNPs can be used to assign each read to the correct allele.

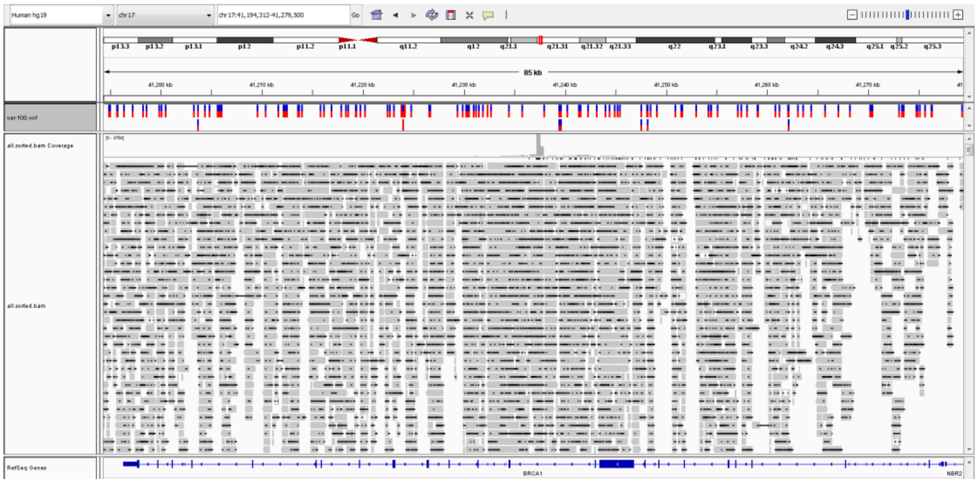

Figure 4 – Targeted TLA reads provide coverage of the entire 81 kb BRCA1 gene.

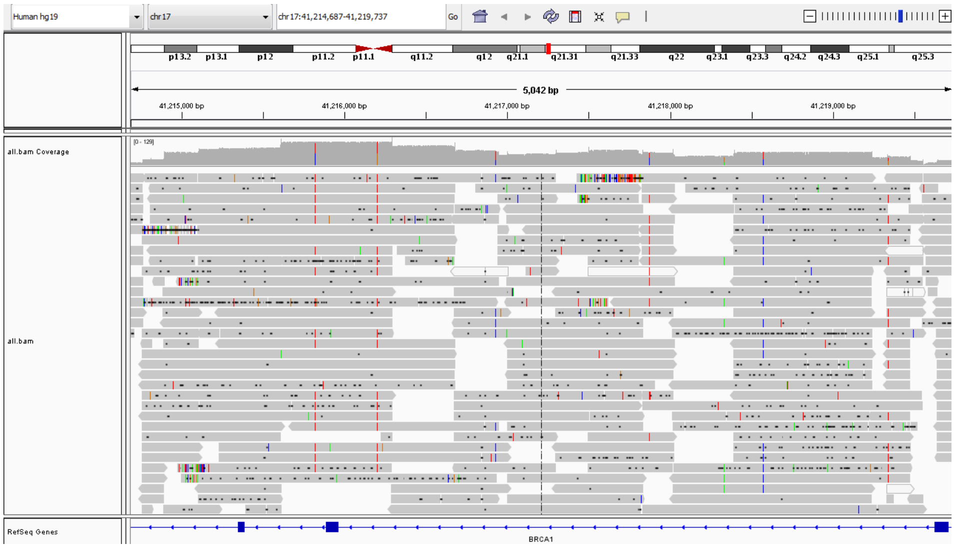

Figure 5 – Heterozygous SNPs are clearly visible in the reads and enable haplotyping of the entire gene.

Phasing Information for BRCA1	
#Haplotype Blocks	1
Block Span	81,463 bp
# hetSNPS Phased	116
# hetSNPs in Validation Set	117
Switch Errors	0

Table 2 – Targeted TLA data demonstrates high accuracy phasing information.

Targeted TLA for Chromosomal Haplotype Assembly

The targeted TLA data had segments aligning far outside of the BRCA1 gene region that was targeted (Figure 6). In addition, segments from the same TLA molecule in most cases were from the same haplotype (Figure 7).

Given the additional information contained in the data, longer range phasing was explored by combining those data with whole genome shotgun PacBio data. HAPCUT was used to construct a phasing block that spanned all of chromosome 17 and had low switch rates, demonstrating the feasibility of this approach.

Statistics of Longest Phasing Block on Chr17	
Block Span	79,628,306 bp
Chromosome 17 Size	81,628,306 bp
# Phased Bases	28,133,018 bp
# hetSNPs Phased	21,762
Long Switch Rate	0.4%
Short Switch Rate	0.8%

Table 3 – The long-range information contained in the targeted TLA data resulted in a phasing block spanning chromosome 17.

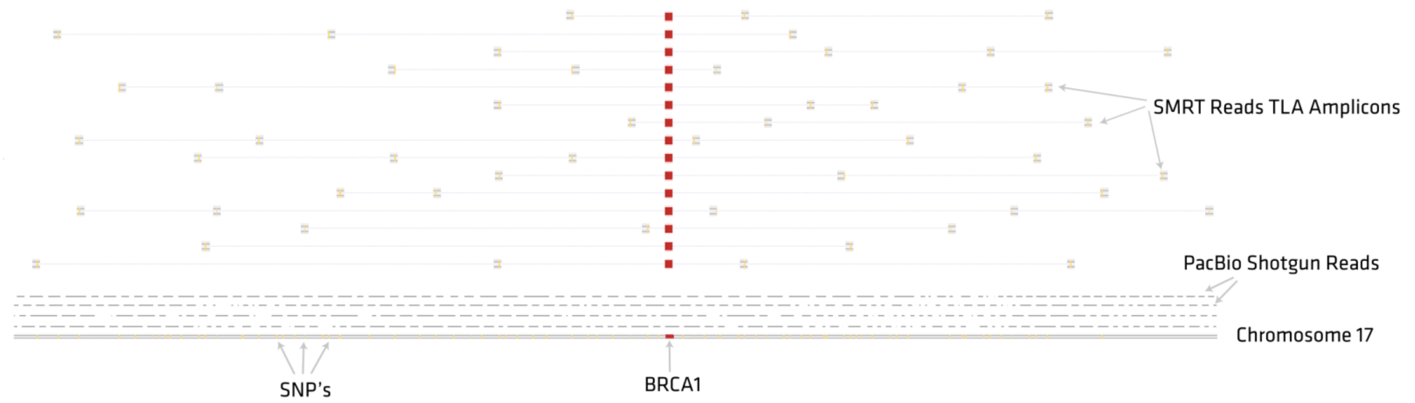

Figure 6 – Schematic depiction of TLA BRCA1 SMRT Sequencing-based phasing of chromosome 17 (one allele shown).

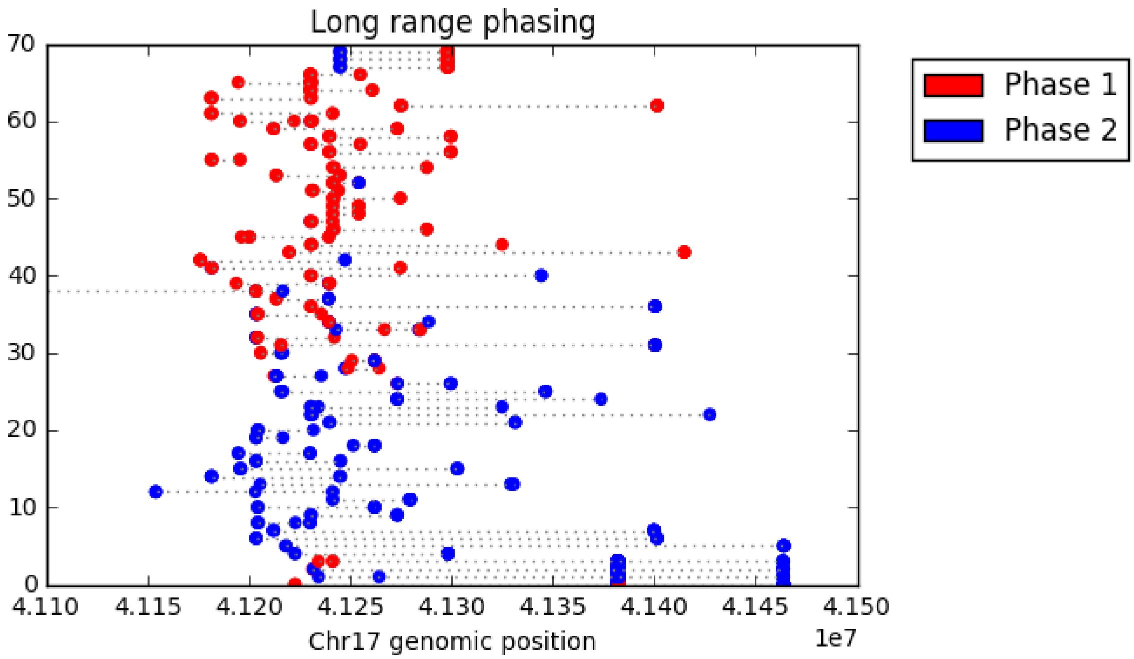

Figure 7 – Each row corresponds to a single TLA molecule that maps to chromosome 17, with each point representing a phased segment. The TLA molecules in general have a consistent phase.

Whole Genome TLA for Whole Genome Haplotype Assembly

A Whole Genome TLA Template dataset was expected to increase haplotype performance even more, due to several favorable properties of the data. First, segments from the same read mapped to locations on the genome with significant distances (Figure 8). In addition, many reads had >10 segments (Figure 9), which greatly increases the chance that two segments from one read will each have a heterozygous SNP. Combining these data with shotgun data from the same individual, the number of phased SNPs increased dramatically (Table 4, validation in progress).

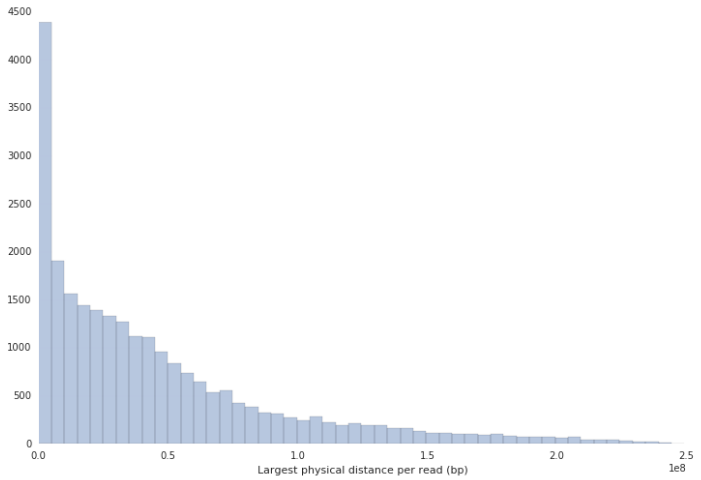

Figure 8 – The longest distance between uniquely mapped segments was calculated for each read with greater than 10 mapped segments. The distribution of distances shows a large percentage of reads having multi-megabase spanning distances.

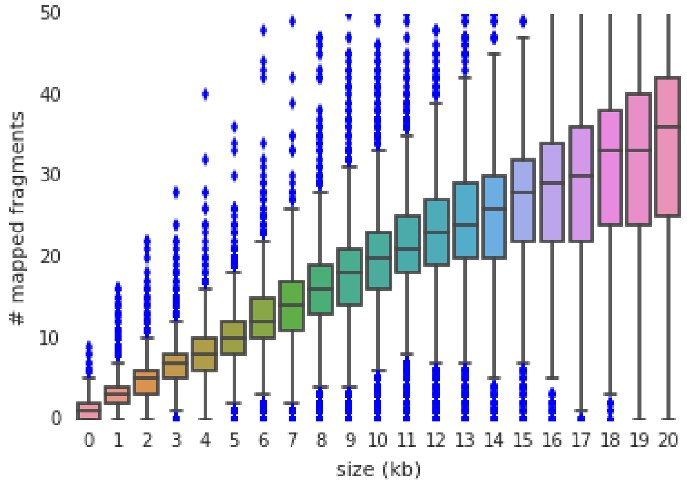

Figure 9 – Longer reads on average have more mapped segments. For example, a 10 kb read has about 20 mapped segments.

Statistics of Longest Phasing Block on Chr17	
Block Span	81,121,761 bp
Chromosome 17 Size	81,195,210 bp
# Phased Bases	70,906,325 bp
# hetSNPs Phased	48,349

Table 4 – The Whole Genome TLA data enabled a much higher number of phased SNPs across all of chromosome 17.

References

1. de Vree, JP et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nature Biotechnology, 2014.
2. Cergentis website
3. PacBio targeted sequencing webpage
4. GitHub site for iterative mapping and phasing scripts
5. Bansal, V and Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics, 2008.
6. PacBio Assembly Data for NA12878
7. PacBio Assembly Data for NA24385

For Research Use Only. Not for use in diagnostic procedures.
PN 100-880-400-02