TLA and SMRT Sequencing

TLA is a strategy to selectively amplify and sequence complete loci on the basis of the crosslinking of physically proximal sequences. Unlike other targeted sequencing methods, TLA works without detailed prior locus information, as one primer pair is sufficient to amplify and sequence tens to hundreds of kilobases of surrounding DNA. TLA enables targeted complete sequencing and the detection of single nucleotide and structural variants in genes of interest. In addition, TLA enables the haplotyping of sequenced regions.

Unamplified TLA Template can be used for genome-wide phasing and assembly.

SMRT Sequencing enables the complete sequencing of TLA products and therefore empowers phasing and assembly.

Experiment

Here, we applied TLA on the BRCA1 gene on NA12878 with a primer pair at (hg19) Chr17:41237179-41238511 (located ~ 40 kb from the start of the 81 kb BRCA1 gene) and then sequenced the resulting 2 kb circles on the PacBio RS II instrument.

We then explored chromosomal-scale haplotype assembly by combining these data with whole-genome shotgun PacBio long reads.

Finally, by size-selecting TLA Templates >5 kb to maximize the number of segments per read and then sequencing, we targeted whole-genome haplotype assembly across all chromosomes.

PacBio SMRTbell libraries were created from the Cergentis samples following published PacBio sample prep procedures (with 6 kb BluePippin size selection and additional damage repair for the whole-genome TLA Template) and sequenced on the PacBio RS II.

TLA yields 2 kb CCS reads with ~4 segments/read, and TLA Template yields >10 kb reads with >20 segments/read. For targeted BRCA1 phasing, SNPs were de novo called using SAMtools and BCFTools. For whole-chromosome analysis, BAM (PacBio shotgun) and VCF files were obtained from GIAB.

HAPCUT was then used to phase selected regions, incorporating whole-genome PacBio shotgun data for whole-chromosome phasing.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Prep</th>
<th>Library size</th>
<th>Sequencing Chemistry</th>
<th>Fold Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA12878</td>
<td>TLA targeting BRCA1</td>
<td>2 kb</td>
<td>P6-C4</td>
<td>Variable with peak at BRCA1</td>
</tr>
<tr>
<td>NA12878</td>
<td>Whole-genome shotgun</td>
<td>~7 kb</td>
<td>PS-C3 and older</td>
<td>~40X</td>
</tr>
<tr>
<td>GM24385</td>
<td>TLA Template</td>
<td>10 kb</td>
<td>P6-C4</td>
<td>0.8X</td>
</tr>
<tr>
<td>GM24385</td>
<td>Whole-genome shotgun</td>
<td>&gt;10 kb</td>
<td>P6-C4</td>
<td>~50X</td>
</tr>
</tbody>
</table>

Whole-Chromosome Phasing

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

Because the targeted TLA data has segments aligning far outside of the BRCA1 gene region (plot on right), longer range phasing by combining those data with whole-genome shotgun PacBio data was performed. HAPCUT was able to construct a phasing block that spanned all of chromosome 17 and had low switch rates demonstrating feasibility of the approach.

Whole Genome Phasing

In a whole-genome TLA Template dataset, segments from the same read have significant distances (plot A), and many reads had >10 segments (plot B), 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 dramatically increases (table C, validation in progress).

Additional Information

References: