Introduction

Insertions, deletions, duplications, translocations, inversions, and tandem repeat expansions in the structural variant (SV) size range (≥50 bp) contribute to the evolution of traits and often have significant associations with agronomically important phenotypes. However, most SVs are too small to detect with array comparative genomic hybridization and too large to reliably discover with short-read DNA sequencing. While de novo assembly is the most comprehensive way to identify variants in a genome, recent studies in human genomes show that PacBio SMRT Sequencing sensitively detects structural variants at low coverage.

Human SV Benchmark

The Genome in a Bottle Consortium has developed a benchmark set of insertion and deletion SVs in a human male, HG002/NA24385. Comparing technologies against this benchmark, PacBio has the highest precision and recall across the structural variant size range, and particularly for insertions.

Methods for SV Detection

<table>
<thead>
<tr>
<th>sequence</th>
<th>map reads</th>
<th>call variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find SV signatures</td>
<td>Cluster SV signatures</td>
<td>Filter by support</td>
</tr>
<tr>
<td>Summarize consensus</td>
<td>Genotype</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Workflow to detect structural variants from PacBio long reads. To call structural variants, pbsv identifies signatures of structural variation in alignments, clusters nearby signatures with similar length and sequence, summarizes into a consensus call, and assigns a genotype based on read support.

SV Calling in MH63 with Assembly

MH63 was re-sequenced to 30-fold PacBio read coverage and downsampled to generate lower coverage. Structural variants were called with pbsv and compared to the calls generated from de novo assembly.

Figure 5. Recall of pbsv at various coverage levels. Recall remains high for coverage ≥10-fold, with the primary limit of sensitivity being large insertions.

SV Calling in MH63 with pbsv

MH63 was re-sequenced to 30-fold PacBio read coverage and downsampled to generate lower coverage. Structural variants were called with pbsv and compared to the calls generated from de novo assembly.

Figure 5. Recall of pbsv at various coverage levels. Recall remains high for coverage ≥10-fold, with the primary limit of sensitivity being large insertions.

Conclusions

- PacBio sequencing has high precision and recall for SVs in plant genomes.
- SV calling is effective at lower coverage than is de novo assembly.
- The workflow to detect SVs is simple and efficient.

References