Structural Variant Detection in Crops Using PacBio SMRT Sequencing

Greg Concepcion1, Shreyasee Chakraborty1, Michelle Vierra1, Emily Hatas1, Aaron Wenger1
1. PacBio, 1305 O’Brien Drive, Menlo Park, CA 94025

Introduction

Structural variants (genomic differences ≥50 base pairs) contribute to the evolution of traits and disease. Most structural variants (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 coverage1. Here we present SV characterization in two major crop species grown worldwide, *Zea mays* (Maize) and *Glycine max* (Soy).

Datasets

Figure 2. Illustrations; A) *Zea mays* B) *Glycine max*

<table>
<thead>
<tr>
<th>Species</th>
<th>Cultivar</th>
<th>SRA Accession</th>
<th>Seq Platform</th>
<th>Gbp</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Zea mays</em></td>
<td>Mo17</td>
<td>SRR58643</td>
<td>PacBio RSII</td>
<td>48.0</td>
<td>22.7</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>Mo17</td>
<td>SRR5826129</td>
<td>Illumina</td>
<td>89.9</td>
<td>42.7</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>Williams</td>
<td>unreleased</td>
<td>PacBio Sequel</td>
<td>24.3</td>
<td>24.9</td>
</tr>
<tr>
<td><em>Glycine max</em></td>
<td>Wm82</td>
<td>SRR425302</td>
<td>Illumina</td>
<td>34.8</td>
<td>35.6</td>
</tr>
</tbody>
</table>

Table 1. For both species, *Glycine max* and *Zea mays*, two parallel datasets, consisting of both long and short reads, were acquired for comparison.

Methods

Structural variation detection was performed using two parallel pipelines appropriate for the respective technologies. For PacBio long reads, NGMLR2 was used for mapping and SV detection performed with pbsv, while the standard BWA3 mapping tool was used followed by SV detection using manta4 for short reads. Additionally, a subset of the long read data for each dataset was used to investigate the sensitivity of SV detection with low-fold coverage.

Results

Figure 5. *Zea mays* data mapped to *AGPv4* reference DNA from the same strain of *Zea mays* (Mo17) was sequenced in parallel with both PacBio and Illumina and subsequently analyzed for structural variant detection. Despite being at a coverage disadvantage, more than 5 times the number of structural variants were detected with PacBio long read technology.

Aligned reads in IGV

Figure 7. Venn diagrams showing overlap between A) *Zea mays* B) *Glycine max* structural variation call sets

Figure 8. Structural variation in *Zea mays* visualized with IGV IGV 2.4 makes it easy to visualize structural variants in haplotypes. On *Zea Mays* *AGPv4* Chromosome 1, SNP locations between Illumina and PacBio alignments are in agreement. In addition, at low-fold coverage PacBio alignments also highlight one large insertion and one large deletion in that are not detected in Illumina alignments.

Conclusions

- Structural variant annotation performed with PacBio long reads detects many more variants than short reads in both maize and soy.
- In soy, SVs account for ~6.4 Mb of sequence while for maize the number is much higher at ~492 Mb. Part of this large number is likely due to strain differences from the reference.
- SV detection with low-fold coverage PacBio data is a viable approach for genomic characterization of crops.

References