Sample Prep and Sequencing

- Cascade clones grown in greenhouse under controlled conditions. ~100 ug young leaves harvested and placed immediately on ice.
- DNA extracted with modified Qiagen DNAeasy mini-prep kit. Chemical precipitation and glass hooks substituted for spin columns to prevent shearing.
- 2 SMRTbell libraries constructed from unsheared DNA with 10kb size selection (BluePippin). Quality assessed with pulse field gels.

Pre-Assembly Results

- Pre-assembly (“error-correction”) generates high accuracy pre-assembled reads (“peads”) by raw read overlap and consensus (Fig 2A).
- Pre-assembled yield = 75% (total length reads / total length seed reads), used as metric for data quality and coverage.

Table 1 PacBio SMRT Sequence Data. Data were generated from Nov. 2016-May 2017 on PacBio RS II with P6-C4 chemistry, and Sequel 1.2 – 2.0 chemistry.

| PacBio SMRT Sequence Data | Sequal Cells | 83
| PacBio RS II Cells | 126
| Total Subread Bases | 342,733,627,375
| Prop. Sequel Data (% bp) | 74 %

Table 2 Summary Statistics of Reads. raw Reads. filtering: subreads >500 bp, single subreads from each ZMW. Seed Reads. raw reads longer than 10 kb (parental cut-off). Peads may be shorter than seed reads due to low coverage regions that break or truncate seed reads. Coverage based on 2.7 Gb 1-genome size.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Raw Reads</th>
<th>Seed Reads</th>
<th>Peads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Count</td>
<td>352.2 M</td>
<td>12.1 M</td>
<td>13.8 M</td>
</tr>
<tr>
<td>Total Bases</td>
<td>288 Gb</td>
<td>182 Gb</td>
<td>135 Gb</td>
</tr>
<tr>
<td>Coverage</td>
<td>19.4x</td>
<td>11.7x</td>
<td>50x</td>
</tr>
<tr>
<td>Mean Length</td>
<td>8 kb</td>
<td>15 kb</td>
<td>12 kb</td>
</tr>
<tr>
<td>N50</td>
<td>12 kb</td>
<td>15 kb</td>
<td>12 kb</td>
</tr>
</tbody>
</table>

Conclusion

- The PacBio FALCON-Unzip diploid assembly is both more complete (>2-fold increase in assembly length) and more contiguous (>600-fold increase in contig NG50) than previous short-read assemblies.
- Most assembled contigs are haplotypes due to high heterozygosity, with haplotype divergence ranging from 1-5% for SNPs and small indels.
- Haplotype deduplication removed ~450 Mb of haplotypes prior to scaffolding.

References


Acknowledgements

The authors would like to thank Emily Hatake, Greg Conopark, and Richard Hall at PacBio.