Sequencing and assembly of the 22-gigabase genome of loblolly pine
Daniela Puiu1, Steven L. Salzberg1, Aleksey Zimin2, James Yorke2
1 Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland.
2 Institute for Physical Science and Technology, University of Maryland College Park.

Abstract

The loblolly pine is the most widespread and commercially important pine of the Southeast US. The Loblolly Pine Genome Project (LPGP) is part of the USDA-funded PineRefSeq project whose aim is the sequencing the three largest genomes ever: loblolly pine (22Gb), sugar pine (33Gb) and Douglas-fir (18Gb). The large genome size and high repeat content makes conifer sequencing very challenging. The PineRefSeq goal is the development of high quality reference genome sequences and model approaches for sequencing other large complex genomes.

Introduction

The loblolly pine 22Gbp diploid genome (n=12) has been sequenced using the illumina technology from a combination of whole genome shotgun and pooled fosmid libraries. 13 billion reads and 1.7 trillion bases have been generated. Six universities and research institutes have been involved in this project. The assembly has been done in collaboration by the University of Maryland and Johns Hopkins University. Our main goal were the library evaluation and selection as well as the development of a high quality assembly. In this poster we are going to present the loblolly pine whole genome assembly results.

Sequencing

<table>
<thead>
<tr>
<th>Lanes</th>
<th>Mate pairs 300K</th>
<th>InsertLen bp</th>
<th>Bases 10^9</th>
<th>Cvg</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiSeq*</td>
<td>33</td>
<td>4845</td>
<td>273-565</td>
<td>1.177</td>
</tr>
<tr>
<td>GAIIx-frag*</td>
<td>35</td>
<td>975</td>
<td>209-637</td>
<td>303</td>
</tr>
<tr>
<td>GAIIx-jump</td>
<td>50</td>
<td>853</td>
<td>1304-5465</td>
<td>272</td>
</tr>
<tr>
<td>DITag</td>
<td>45</td>
<td>52</td>
<td>38000</td>
<td>16</td>
</tr>
<tr>
<td>MiSeq*</td>
<td>10</td>
<td>6,749</td>
<td>209-3000</td>
<td>1,177</td>
</tr>
</tbody>
</table>

Table 1: Dataset composition: 83% of reads come from short insert haploid libraries (pine nut); 17% of the reads come from long insert diploid libraries (needle); <1% MiSeq data were used for assembly evaluation.

Quality & base composition: fastx toolkit
- Kmer counting: jellyfish, kmerfreq => genome size estimation
- Adapter & low quality trimming: ea-utils
- Contamination removal: bwa alignment to contaminant db.
- Error correction: QuORUM (part of MaSuRCA)
- Library insert & complexity estimation

Figure 2: Quality & base composition of 100K sampled mate-pairs

Figure 3: Kmer count analysis; peak values => coverage; long tail => repeats

Data analysis

Assembly results

When the PineRefSeq project has started in 2011, no genome assembler was able to handle 3 billion reads. The MaSuRCA assembler has been developed by the PineRefSeq team at UMD specifically for assembling such large genomes. MaSuRCA is based on the Celera assembler and uses an overlap-layout-consensus approach with K-units and superReads which allows for a 100-fold data reduction.

Input: QuORUM error-corrected reads
Parameters: K=79
Runtime & resources: 3 months on a 64-core 1 TB machine

Table 2: MaSuRCA v1.0 assembly statistics; N50 was computed based on a 22Gbp genome size; The scaffolds were assembled into super-scaffolds using uncorrected long insert mates

<table>
<thead>
<tr>
<th>Contigs</th>
<th>N50bp</th>
<th>10^6</th>
<th>10^9</th>
<th>Sum 10^9</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.6</td>
<td>82</td>
<td>0.2</td>
<td>20.1</td>
<td></td>
</tr>
<tr>
<td>Scaffolds</td>
<td>16.4</td>
<td>30.7</td>
<td>5.0</td>
<td>22.5</td>
</tr>
<tr>
<td>Super-scaffolds</td>
<td>8.3</td>
<td>96</td>
<td>5.1</td>
<td>22.3</td>
</tr>
</tbody>
</table>

Table 3a: SOAPdenovo2 assembly statistics; before & after gap closing

Optimization

Using the MaSuRCA assembler we have been able to generate a 22.5Gbp loblolly pine assembly with an N50 contig size of 8.2Kbp and N50 scaffold size of 35.7Kbp.

This assembly is by far superior to the 19.5Gbp one generated by SOAPdenovo2 whose original N50 contig size was only 0.6Kbp. Some of faster SOAPdenovo2 steps were used in parameter optimization and assembly improvement.

Conclusions

3. The MaSuRCA Genome Assembler, Aleksey Zimin et al, under review

References

Acknowledgements