Celera Assembler and Automated Finishing

Sergey Koren
Software Engineer
J. Craig Venter Institute
SFAiF 2009
Celera Assembler Object Hierarchy

Consensus

Scaffold = Contigs + Mates

Contigs = Unitigs + Mates

Unitigs = Reads + Overlaps

Reads + Paired-ends

AGGCATGACGGCTAGCGCGTA

NNNNNNNNNNCCGCAGATACGAG
Method 1: Unitig Toggling

Celera Assembler pipeline: default

Celera Assembler pipeline: toggling

Find large “repeat” unitigs placed once in scaffolds. Mark them “unique.”
Tested on Large Genomes

- Experimental run: Oil palm, genome size ≈1.93Gbp
  - Desired outcome: more placement, larger scaffolds
  - Many “repeat” unitigs are placed exactly once

- Negative control: Cucumber, genome size ≈367MBp
  - Expected outcome: little effect
  - Of “repeat” unitigs, few are placed exactly once

<table>
<thead>
<tr>
<th>Parameters</th>
<th># Toggled</th>
<th>Contig N50</th>
<th>Scaffold N50</th>
<th>Paired-End Happiness</th>
<th>Paired-End Unhappiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>N/A</td>
<td>N/A</td>
<td>46Kbp</td>
<td>1.6Mbp</td>
<td>62.64%</td>
</tr>
<tr>
<td>Toggling</td>
<td>&gt; 2,00 bp</td>
<td>19,046</td>
<td>54Kbp</td>
<td>2.1Mbp</td>
<td>70.75%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th># Toggled</th>
<th>Contig N50</th>
<th>Scaffold N50</th>
<th>Paired-End Happiness</th>
<th>Paired-End Unhappiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>N/A</td>
<td>N/A</td>
<td>87Kbp</td>
<td>832Kbp</td>
<td>12.20%</td>
</tr>
<tr>
<td>Toggling</td>
<td>&gt; 2,00 bp</td>
<td>345</td>
<td>86Kbp</td>
<td>867Kbp</td>
<td>12.57%</td>
</tr>
</tbody>
</table>

Conclusion: unitig toggling is effective and safe.
Oil Palm - Improvement in Assembly

Toggling (flagging repeats as trusted)

Toggling increases mate satisfaction with no adverse effect

N50

Cumulative bp %

Toggling increases mate satisfaction with no adverse effect

Biggest Contigs

Biggest Scaffolds

#Scaffolds
Toggled unitig placed into assembly. Inspection shows the unitig has high identity with the reference in this region, supporting the assembly.¹

Reference: Fosmid OPBP_1_311

Toggled unitig not placed into assembly. Inspection shows the unitig has low identity with the reference in this region, supporting the assembly.¹

Reference: Fosmid OPCD_1_377

Method 2: Finishing Reads

- Goals of finishing
  - Close sequencing gaps by generating new reads
  - Get higher coverage in low-coverage sections
  - Generate reads to add coverage to repeat instances
- Reads can be constrained by
  - Overlaps
  - Paired-end distance/orientation
  - Finishing anchoring sequence
  - Default assembly only aware of first two constraints
- Default assembly does not place a read into a repeat instance unless the corresponding pair agrees
  - Reads with no paired-ends are not placed into repeat instances
- Manual finishing is expensive
  - Closing only a few gaps automatically saves manual effort and expense
Finishing Read Assembly

- Test finishing constraints on:
  - Two small finished genomes
    - *Salmonella enterica* subsp. *c* Schwarzengrund str. CVM19633
    - *Escherichia coli* O157:H7 str. EC4115

<table>
<thead>
<tr>
<th>Genome Name</th>
<th>Size</th>
<th>% Paired-End Reads</th>
<th>% Masked by RepSeek(^2)</th>
<th># Closure Features</th>
<th># Closure Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. enterica</em></td>
<td>4.71MBp</td>
<td>95%</td>
<td>5.68%</td>
<td>608</td>
<td>891</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>5.57MBp</td>
<td>95%</td>
<td>34.88%</td>
<td>516</td>
<td>1,041</td>
</tr>
</tbody>
</table>

Finishing Constraints Improve Assembly

- Close more gaps in repetitive genome
- Place more reads into assembly and repetitive regions
Future Work

• Toggling
  o Automate decision to run pipeline
  o Improve set of toggled unitigs
    - Detect two-copy repeat at the ends of two scaffolds for toggling

• Finishing Constraints
  o Evaluate and improve for larger genomes
  o Improve read placement
Acknowledgements

- Jason Miller
- Brian Walenz
- Eli Venter
- Jessica Hostetler
- Justin Johnson
- Diane Radune
- JCVI IT and Grid support
- Granger Sutton