Genome Assembly: Final Results

February 24, 2014

Devin Cline
Krutika Gaonkar
Smitha Janardan
Karthikeyan Murugesan

Emily Norris
Ying Sha
Eshaw Vidyaprakash
Xingyu Yang
Topics

1. Data samples
2. Pre-processing
3. Reference Assemblers
4. De novo Assemblers
5. Merging of Reference & De novo
6. Visualization
7. PacBio
8. Pipeline
## Input Data

<table>
<thead>
<tr>
<th>No. of Samples</th>
<th>Serogroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>W</td>
</tr>
<tr>
<td>1 (M22813)</td>
<td>A</td>
</tr>
<tr>
<td>1 (M24695)</td>
<td>X</td>
</tr>
</tbody>
</table>

- *Neisseria meningitidis*
- Illumina single end and paired end data for each sample
- PacBio quiver results for 3 samples (M09261, M22189, M09293)
Preprocessing

To solve:
1. Low quality tail
2. Biased head
3. Duplicated Reads
Preprocessing

Trim biased head

awk 'NR%4==2||NR%4==0{$0=substr($0,5,)}{print}' input_file.fastq > output_file.fastq
Preprocessing

Remove duplicated Reads

run_assembly_removeDuplication.pl input_file.fastq > outputfile.fastq

Exactly same reads are considered duplicated.
For pair-end reads, both end are exactly same are considered duplicated.
Preprocessing

Remove low quality tail and
Filter reads with overall low quality

run_assembly_trimClean.pl -i <in fastq> -o <out fastq>

NNNs in original data
Reference Assembly

bowtie2

bowtie2 -k 10 -p 4 -x ../FAM18 -1 File.1.fq -2 File.2.fq -U File.fq -S File.sam

-Combine single-end and paired-end reads
-Allow multi-mapping reads
Reference Assembly
De novo Assembly

<table>
<thead>
<tr>
<th></th>
<th>M22189</th>
<th>M22748</th>
<th>M25543</th>
<th>M24695</th>
<th>M22813</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABBySS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaSuRCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAPdenovo2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAdes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velvet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Red = high assembly score
Blue = low assembly score
MaSuRCA

- Consistently MaSuRCA had higher assembly score
- Has own error correction algorithm as part of assembly software - QuorUM
  - Quality Optimized Reads from the University of Maryland
  - Based on Illumina reads
  - Suffix tree algorithm
  - Walk through tree correcting nodes with lower frequency than expected

MaSuRCA vs MaSuRCA with RACER

Red = high assembly score
Blue = low assembly score
Merging of Reference and De novo Assemblies

MAIA: Uses the overlap-layout-consensus paradigm.
- The algorithm takes as input sets of contigs, each set originating from either a de novo or a comparative assembly, i.e. from mapping against a related genome.

GAA: Uses a accordance graph data structure.
- The algorithm constructs an accordance graph to capture the mapping information between the target and query assemblies.

MINIMUS2: Uses the overlap-layout-consensus paradigm.
- It uses a nucmer based overlap detector, a tigger- a tool that identifies a cluster of reads that can be uniquely assembled, make consensus- provides a refined and precise multiple alignment of reads.

CISA: Uses step- wise splitting and merging
- Identify the representative contigs and possible extensions, remove and split contigs that may be misassembled, iteratively merge the contigs with a minimum 30% overlap and estimate the maximal size of repetitive regions, and merge of the contigs based on the size of repetitive regions.
Minimus2 vs CISA

**Minimus 2:**
It is designed to align small data-sets such as aligning reads to a specific gene.

Due to its stringency, the resulting assembly will be highly fragmented therefore overestimating the intact CDS

Reads completely contained within other reads in the input are removed from the graph.

**CISA:**
Set of contigs from at least three assemblers can be used.

Less frame-shift errors (i.e., assembly errors in indels) are generated while using CISA.

Iteratively merges the contigs based on the size of repetitive regions (overlaps should be greater than maximum repeat size).
Evaluation for mergers

<table>
<thead>
<tr>
<th>Name</th>
<th>Num Contigs</th>
<th>Assembly Bases</th>
<th>DCJ Distance</th>
<th>%Missed</th>
<th>%Extra</th>
<th>Intact CDS</th>
<th>Max Contig</th>
<th>N50</th>
<th>Blast-based Intact CDS</th>
<th>Assembly errors (Indel&gt;=5, Inversion, Relocation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli K12 MG1655 (genome size = 4639675, number of CDS = 4320) with 36 bp paired-end reads</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAIA</td>
<td>110</td>
<td>4513348</td>
<td>96</td>
<td>2.80</td>
<td>0.02</td>
<td>4272</td>
<td>312145</td>
<td>126075</td>
<td>4212</td>
<td>5 (2, 0, 3)</td>
</tr>
<tr>
<td>minimus2#</td>
<td>155</td>
<td>4598769</td>
<td>133</td>
<td>1.94</td>
<td>0.71</td>
<td>4262</td>
<td>202745</td>
<td>86241</td>
<td>4243</td>
<td>8 (5, 1, 2)</td>
</tr>
<tr>
<td>GAA##</td>
<td>314</td>
<td>4578451</td>
<td>245</td>
<td>2.09</td>
<td>0.23</td>
<td>4248</td>
<td>157184</td>
<td>51218</td>
<td>4205</td>
<td>4 (2, 0, 2)</td>
</tr>
<tr>
<td>CISA</td>
<td>72</td>
<td>4627549</td>
<td>70</td>
<td>1.07</td>
<td>0.69</td>
<td>4276</td>
<td>312018</td>
<td>126254</td>
<td>4276</td>
<td>11 (8, 0, 3)</td>
</tr>
</tbody>
</table>
CISA Input

Good Reference Available (17 W, 1 A)
- Bowtie 2, SOAPdenovo2, MaSuRCA

Good Reference Not Available (6 W)
- SOAPdenovo2, MaSuRCA, Velvet

M24695 (X)
- SOAPdenovo2, Velvet, SPAdes
Split contigs with runs of N > 5

AACCTGTCCATGCA\text{AAAAA}TTTGTCCATGCA

AACCTGTCCATGCA

TTTGACTTACTNTCA

TTTGACTTACTNTCA
## CISA Results - Reference

<table>
<thead>
<tr>
<th>Assembly</th>
<th># Contigs</th>
<th>Total Length</th>
<th>N50</th>
<th>Ns per 100kbp</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowtie</td>
<td>68.3</td>
<td>2156444</td>
<td>94515</td>
<td>3.19</td>
<td>9.47</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>179</td>
<td>2042421</td>
<td>27236</td>
<td>159</td>
<td>8.49</td>
</tr>
<tr>
<td>MaSuRCA</td>
<td>245.3</td>
<td>2189678</td>
<td>28079</td>
<td>261</td>
<td>8.4</td>
</tr>
<tr>
<td>Merged</td>
<td>47.2</td>
<td>2229488</td>
<td>106156</td>
<td>2.66</td>
<td>9.72</td>
</tr>
</tbody>
</table>

Score = \log(N50*Total Length/#Contigs)
## CISA Results - No Reference

<table>
<thead>
<tr>
<th>Assembly</th>
<th># Contigs</th>
<th>Total Length</th>
<th>N50</th>
<th>Ns per 100kbp</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velvet</td>
<td>218.1</td>
<td>2022336</td>
<td>20674</td>
<td>278</td>
<td>8.28</td>
</tr>
<tr>
<td>SOAPdenovo</td>
<td>197.1</td>
<td>2054632</td>
<td>21567</td>
<td>192</td>
<td>8.35</td>
</tr>
<tr>
<td>MaSuRCA</td>
<td>293.9</td>
<td>2215304</td>
<td>25839</td>
<td>59.9</td>
<td>8.29</td>
</tr>
<tr>
<td>Merged</td>
<td>211.6</td>
<td>2286833</td>
<td>21471</td>
<td>8.35</td>
<td>8.35</td>
</tr>
</tbody>
</table>

Score = \( \log(\text{N50} \times \text{Total Length} / \#\text{Contigs}) \)
<table>
<thead>
<tr>
<th>Assembly</th>
<th>Serogroup</th>
<th># contigs</th>
<th>Largest contig</th>
<th>Total length</th>
<th>N50</th>
<th># N's per 100 kbp</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>7149</td>
<td>W</td>
<td>62</td>
<td>296622</td>
<td>2368518</td>
<td>85380</td>
<td>3.72</td>
<td>9.51</td>
</tr>
<tr>
<td>7165</td>
<td>W</td>
<td>53</td>
<td>204959</td>
<td>2213595</td>
<td>90033</td>
<td>4.88</td>
<td>9.58</td>
</tr>
<tr>
<td>9261</td>
<td>W</td>
<td>40</td>
<td>329153</td>
<td>2326427</td>
<td>179980</td>
<td>2.97</td>
<td>10.02</td>
</tr>
<tr>
<td>9293</td>
<td>W</td>
<td>35</td>
<td>325056</td>
<td>2295972</td>
<td>118300</td>
<td>1.87</td>
<td>9.89</td>
</tr>
<tr>
<td>22189</td>
<td>W</td>
<td>20</td>
<td>325069</td>
<td>2048957</td>
<td>176935</td>
<td>2.15</td>
<td>10.26</td>
</tr>
<tr>
<td>22718</td>
<td>W</td>
<td>27</td>
<td>329249</td>
<td>2231911</td>
<td>132154</td>
<td>2.37</td>
<td>10.04</td>
</tr>
<tr>
<td>22740</td>
<td>W</td>
<td>25</td>
<td>325095</td>
<td>2229592</td>
<td>118977</td>
<td>0.9</td>
<td>10.03</td>
</tr>
<tr>
<td>22748</td>
<td>W</td>
<td>31</td>
<td>262407</td>
<td>2141647</td>
<td>105719</td>
<td>1.63</td>
<td>9.86</td>
</tr>
<tr>
<td>22751</td>
<td>W</td>
<td>48</td>
<td>231389</td>
<td>2260604</td>
<td>82788</td>
<td>3.89</td>
<td>9.59</td>
</tr>
<tr>
<td>22772</td>
<td>W</td>
<td>31</td>
<td>325067</td>
<td>2058872</td>
<td>123630</td>
<td>2.77</td>
<td>9.91</td>
</tr>
<tr>
<td>22801</td>
<td>W</td>
<td>53</td>
<td>325131</td>
<td>2241643</td>
<td>143112</td>
<td>5.44</td>
<td>9.78</td>
</tr>
<tr>
<td>22803</td>
<td>W</td>
<td>61</td>
<td>325109</td>
<td>2252787</td>
<td>85380</td>
<td>4.31</td>
<td>9.5</td>
</tr>
<tr>
<td>22813</td>
<td>A</td>
<td>91</td>
<td>150196</td>
<td>2202940</td>
<td>37750</td>
<td>0.14</td>
<td>8.96</td>
</tr>
<tr>
<td>23347</td>
<td>W</td>
<td>62</td>
<td>231424</td>
<td>2245906</td>
<td>65748</td>
<td>2.67</td>
<td>9.38</td>
</tr>
<tr>
<td>23413</td>
<td>W</td>
<td>47</td>
<td>325056</td>
<td>2298167</td>
<td>80047</td>
<td>2.74</td>
<td>9.59</td>
</tr>
<tr>
<td>25419</td>
<td>W</td>
<td>114</td>
<td>109235</td>
<td>2280894</td>
<td>49311</td>
<td>2.98</td>
<td>8.99</td>
</tr>
<tr>
<td>25543</td>
<td>W</td>
<td>23</td>
<td>257168</td>
<td>2295565</td>
<td>119358</td>
<td>1.22</td>
<td>10.08</td>
</tr>
<tr>
<td>25807</td>
<td>W</td>
<td>26</td>
<td>294413</td>
<td>2136794</td>
<td>116201</td>
<td>1.31</td>
<td>9.98</td>
</tr>
</tbody>
</table>

Reference Average

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Serogroup</th>
<th># contigs</th>
<th>Largest contig</th>
<th>Total length</th>
<th>N50</th>
<th># N's per 100 kbp</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>7293</td>
<td>W</td>
<td>169</td>
<td>113468</td>
<td>2385110</td>
<td>25292</td>
<td>2.52</td>
<td>8.55</td>
</tr>
<tr>
<td>22471</td>
<td>W</td>
<td>161</td>
<td>111494</td>
<td>2388666</td>
<td>33735</td>
<td>3.43</td>
<td>8.7</td>
</tr>
<tr>
<td>22702</td>
<td>W</td>
<td>320</td>
<td>44596</td>
<td>2247750</td>
<td>11221</td>
<td>4.4</td>
<td>7.9</td>
</tr>
<tr>
<td>22811</td>
<td>W</td>
<td>195</td>
<td>60212</td>
<td>2262857</td>
<td>18205</td>
<td>4.11</td>
<td>8.32</td>
</tr>
<tr>
<td>22819</td>
<td>W</td>
<td>166</td>
<td>101110</td>
<td>2317991</td>
<td>23979</td>
<td>3.88</td>
<td>8.52</td>
</tr>
<tr>
<td>22822</td>
<td>W</td>
<td>170</td>
<td>87803</td>
<td>2324236</td>
<td>27479</td>
<td>1.59</td>
<td>8.57</td>
</tr>
<tr>
<td>24695</td>
<td>X</td>
<td>300</td>
<td>54272</td>
<td>2081223</td>
<td>10385</td>
<td>11.72</td>
<td>7.86</td>
</tr>
</tbody>
</table>

No Reference Average

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Serogroup</th>
<th># contigs</th>
<th>Largest contig</th>
<th>Total length</th>
<th>N50</th>
<th># N's per 100 kbp</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>211.6</td>
<td>81851</td>
<td>2286833</td>
<td>21471</td>
<td>4.52</td>
<td>8.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.34</td>
</tr>
</tbody>
</table>
Visualization: Mauve Contig Aligner

Given a set of genomes:

Genome 1:

\[
\begin{array}{c}
A \quad R_1 \quad B \quad R_2 \quad C
\end{array}
\]

Genome 2:

\[
\begin{array}{c}
A \quad R_1 \quad B \quad D \quad R_2 \quad C
\end{array}
\]

Genome 3:

\[
\begin{array}{c}
B \quad D \quad R_2 \quad C
\end{array}
\]

Ideal *positional homology* multiple genome alignment:

Block 1:

\[
\begin{array}{c}
A \quad R_1 \\
A \quad R_1 \\
A \quad R_1
\end{array}
\]

Block 2:

\[
\begin{array}{c}
B \quad D \quad R_2 \quad C \\
B \quad D \quad R_2 \quad C
\end{array}
\]

Ideal *glocal* multiple genome alignment:

Block 1:

\[
\begin{array}{c}
A \\
A \\
A
\end{array}
\]

Block 2:

\[
\begin{array}{c}
A \\
A
\end{array}
\]

Block 3:

\[
\begin{array}{c}
B \quad D \\
B \quad D
\end{array}
\]

Block 4:

\[
\begin{array}{c}
C \\
C
\end{array}
\]
Visualization of Assemblies
Draft_Assembly_22189
Visualization of Assemblies

Draft_Assembly_22702
Visualization of Assemblies

Draft_Assembly22813(Serogroup A)
Assembly evaluation using PacBio

Aligning assembled genomes with PacBio quiver assemblies
M22189