Applications of micro-, mega-, and meta- assembly
Michael Schatz

Nov. 3, 2011
Genome Informatics
micro-
MicroSeq: high-throughput microsatellite genotyping
Mitch Bekritsky, Jennifer Troge, Dan Levy, Michael Wigler, Michael Schatz

• Highly variable simple sequence repeats
  – \ldots\text{GCACACACACAT}\ldots = \ldots\text{G(CA)}_5\text{C}T\ldots
  – Created and mutate primarily through slippage during replication

• Genotyping with MicroSeq:
  1. Rapidly detect MS sequences
  2. Map reads using a new MS-mapper
  3. Analyze profiles in across cells & populations
    • Loss of heterozygosity, de novo mutations
    • Development of somatic & cancer cells
    • Relations across strains, across species
    • etc…

• Currently being applied to look for de novo mutations associated with autism
  (Salipante et al. 2006)
mega- (x2)
Rapid parallel execution of NGS analysis pipelines
- FASTX, BWA, Novoalign, SAMTools, Hydra

Seamless read/write of common formats
- BAM, SAM, BED, fastq, fasta
- Sorting, merging, filtering, selection, etc

Jnomics: Cloud-scale genomics
Matt Titmus, James Gurtowski, Michael Schatz

Poster 173
Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Error Correction Results

Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina
SMRT-hybrid assembly results of 50x PacBio corrected coverage of E. coli K12
Long reads lead to **contigs** over 1Mbp
meta-
• Assembly competition with a known reference genome enables base-by-base comparison to the truth
  – But evaluating an assembly in absence of a reference is difficult
  – Once we identify differences, what can we do about them?
Forensics Pipeline

Computationally scan an assembly for mis-assemblies.
- Data inconsistencies are indicators for mis-assembly
- Some inconsistencies are merely statistical variations

AMOSvalidate

1. Load Assembly Data into Bank
2. Analyze Mate Pairs & Libraries
3. Analyze Depth of Coverage
4. Analyze Read Alignments
5. Analyze Read Breakpoints
6. Load Mis-assembly Signatures into Bank

Genome Assembly forensics: finding the elusive mis-assembly.
Mate Evaluation

- Correct: mates have expected orientation and separation

- Mis-assembled: mates have incorrect orientation and separation

- Slightly compressed/expanded mates are expected because mates are sampled from a distribution of fragments
Hidden Compression

Library size distribution
Mean: 4000, SD: 400

8 inserts: 3.2 kb-4.8kb
Local Mean: 3488
C/E Stat: \[
\frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62
\]
C/E Stat ≤ -3.0 indicates Compression
Assemblathon 2: Metassembly
Paul Baranay, Scott Emrich, Michael Schatz

Scaffold N50: 3,710,017  Contig N50: 20,183
#>1000: 2,791  #>1000: 68,591

CE Threshold: 3  Gaps closed: 595
Mis-assemblies fixed: 28  Extra bases: 529kbp

Scaffold N50: 285,413  Contig N50: 1,607
#>1000: 29,119  #>1000: 218,643

Inspired by Zimin et al. (2007) Assembly Reconciliation. Bioinformatics. 42(1) 42-45
Summary

- Assembly is moving to increasingly more complex and more diverse data types and organisms
  - PacBio error correction is my 3rd iteration of this problem
  - Assembly is useful in many different contexts, requires specialization and tuning

- There is a fundamental tension between connectivity and correctness
  - N50 is useful for evaluating connectivity but says nothing about correctness
  - CE can measure correctness at “gene-length” scale

- Metassembly is very promising for advancing assembly
  - Allows one to construct a consensus superior to the individual submissions
  - Enables one to select a locally optimal threshold
## Acknowledgements

<table>
<thead>
<tr>
<th>Schatzlab</th>
<th>CSHL</th>
<th>JHU</th>
<th>Univ. of Maryland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitch Bekritsky</td>
<td>Dick McCombie</td>
<td>Steven Salzberg</td>
<td>Mihai Pop</td>
</tr>
<tr>
<td>Matt Titmus</td>
<td>Melissa Kramer</td>
<td>Ben Langmead</td>
<td>Art Delcher</td>
</tr>
<tr>
<td>Hayan Lee</td>
<td>Eric Antonio</td>
<td>Daniela Puiu</td>
<td>David Kelley</td>
</tr>
<tr>
<td>James Gurtowski</td>
<td>Mike Wigler</td>
<td></td>
<td>Aleksey Zimin</td>
</tr>
<tr>
<td>Giuseppe Narzisi</td>
<td>Zach Lippman</td>
<td>NBACC</td>
<td></td>
</tr>
<tr>
<td>Rohith Menon</td>
<td>Doreen Ware</td>
<td>Adam Phillipy</td>
<td></td>
</tr>
<tr>
<td>Goutham Bhat</td>
<td>Ivan Iossifov</td>
<td>Sergey Koren</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ALLPATHS team</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SOAPdenovo team</td>
</tr>
</tbody>
</table>
Thank You!

http://schatzlab.cshl.edu
@mike_schatz / #GI2011
Compression/Expansion Statistic

Library size distribution
Mean: 4000, SD: 400

8 inserts: 3kb-6kb
Local Mean: 4048

C/E Stat: \[
\frac{(4048-4000)}{(400 / \sqrt{8})} = +0.33
\]

Near 0 indicates overall happiness
Hybrid Assembly Results

<table>
<thead>
<tr>
<th>Organism</th>
<th>Technology</th>
<th>Reference bp</th>
<th>Assembly bp</th>
<th># Contigs</th>
<th>Max Contig Length</th>
<th>N50</th>
<th>Assembly Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lambda NEB3011</strong></td>
<td>Illumina 50X 200bp</td>
<td>48 502</td>
<td>48 452</td>
<td>1</td>
<td>48 452</td>
<td>48 452</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td>48 440</td>
<td>48 440</td>
<td>1</td>
<td>48 440</td>
<td>48 440</td>
<td>0</td>
</tr>
<tr>
<td><strong>E. coli K12</strong></td>
<td>Illumina 50X 500bp</td>
<td>4 639 675</td>
<td>4 438 989</td>
<td>75</td>
<td>222 538</td>
<td>80 168</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>PacBio 20X</td>
<td>4 473 206</td>
<td>79</td>
<td>222 024</td>
<td>66 408</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both 20X PacBio + Illumina 50X 500bp</td>
<td>4 516 224</td>
<td>67</td>
<td>374 849</td>
<td>93 148</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td><strong>E. coli C227-11</strong></td>
<td>PacBio CCS 50X</td>
<td>5 504 407</td>
<td>4 917 717</td>
<td>76</td>
<td>249 515</td>
<td>100 322</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>PacBio 10X</td>
<td>5 252 618</td>
<td>56</td>
<td>379 516</td>
<td>162 597</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td>5 397 525</td>
<td>41</td>
<td>596 739</td>
<td>216 129</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PacBio 50X</td>
<td>5 476 824</td>
<td>39</td>
<td>1 057 326</td>
<td>365 964</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PacBio 75X</td>
<td>5 601 310</td>
<td>55</td>
<td>642 068</td>
<td>308 312</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both PacBio 50X + CSS 25X</td>
<td>5 453 558</td>
<td>33</td>
<td>1 167 060</td>
<td>527 198</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td><strong>E. coli 17-2</strong></td>
<td>Illumina 50X 500bp</td>
<td>4 929 374</td>
<td>71</td>
<td>301 823</td>
<td>108 581</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Illumina 50X 500bp + 50X 3Kbp</td>
<td>5 138 293</td>
<td>58</td>
<td>391 461</td>
<td>190 996</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Illumina 50X 3Kbp + 50X 6Kbp</td>
<td>5 157 771</td>
<td>46</td>
<td>403 168</td>
<td>186 135</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Illumina 50X 500bp + 50X 3Kbp + 50X 6Kbp</td>
<td>5 140 142</td>
<td>60</td>
<td>397 294</td>
<td>153 941</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td>5 277 371</td>
<td>38</td>
<td>424 482</td>
<td>285 861</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both PacBio 25X + Illumina 50X 500bp</td>
<td>5 410 343</td>
<td>41</td>
<td>912 608</td>
<td>286 829</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td><strong>E. coli JM211</strong></td>
<td>Illumina 50X 300bp</td>
<td>5 000 000</td>
<td>4 643 234</td>
<td>123</td>
<td>197 547</td>
<td>39 917</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td>4 912 923</td>
<td>57</td>
<td>420 268</td>
<td>118 962</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both PacBio 25X + Illumina 50X 300bp</td>
<td>4 995 486</td>
<td>54</td>
<td>423 420</td>
<td>125 900</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>S. cerevisiae S228c</strong></td>
<td>Illumina 50X 300bp</td>
<td>12 157 105</td>
<td>10 528 780</td>
<td>271</td>
<td>150 618</td>
<td>44 174</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>PacBio 13X</td>
<td>11 101 617</td>
<td>226</td>
<td>191 587</td>
<td>63 095</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both PacBio 13X + Illumina 50X 300bp</td>
<td>12 157 105</td>
<td>207</td>
<td>323 716</td>
<td>67 117</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td><strong>Melopsittacus undulatus</strong></td>
<td>Illumina 50X 500bp</td>
<td>1.23 Gbp</td>
<td>349 472 172</td>
<td>212 581</td>
<td>11 572</td>
<td>465</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PacBio 3X</td>
<td>882 984 450</td>
<td>237 121</td>
<td>51 333</td>
<td>3 250</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lander Waterman 3X Prediction</td>
<td>1 153 148 167</td>
<td>173 565</td>
<td>69 663</td>
<td>9 026</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case