Outline

1. Assembly preliminaries
   1. De Bruijn and Overlap graph
   2. Coverage, read length, repeats, and errors

2. SMRT-assembly approaches
   1. SMRT-de novo: SMRT-only assembly
   2. SMRT-scaffolding: Long reads as links
   3. SMRT-hybrid: Short and long together

3. Review and best practices
Assembly Applications

• Novel genomes

• Metagenomes

• Sequencing assays
  – Structural variations
  – Transcript assembly
  – …
Assembling a Genome

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads

   ...AGCCTAGACCTACAGGATGCGCGACACCGT
   GGATGCGCGACACCGTCGCATATCCCGGT...

3. Simplify assembly graph

4. Detangle graph with long reads, mates, and other links
Two Paradigms for Assembly

Overlap Graph

- Long read assemblers
  - Repeats depends on read length
  - Read coherency, placements kept
  - Tangled by high coverage

de Bruijn Graph

- Short read assemblers
  - Repeats depends on word length
  - Read coherency, placements lost
  - Robust to high coverage

Assembly of Large Genomes using Second Generation Sequencing
Scaffolding

• Initial contigs (aka unipaths, unitigs) terminate at
  – Coverage gaps: especially extreme GC regions
  – Conflicts: sequencing errors, repeat boundaries

• Iteratively resolve longest, ‘most unique’ contigs
  – Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  – Uniqueness measured by a statistical test on coverage
Coverage and Read Length

Idealized Lander-Waterman model

- Reads start at perfectly random positions

- Poisson distribution in coverage
  - Contigs end when there are no overlapping reads

- Contig length is a function of coverage and read length
  - Effective coverage reduced by $o/l$
  - Short reads require much higher coverage to reach same expected contig length

Assembly of Large Genomes using Second Generation Sequencing
Repeats and Read Length

• Explore the relationship between read length and contig N50 size
  – Idealized assembly of read lengths: 25, 35, 50, 100, 250, 500, 1000
  – Contig/Read length relationship depends on specific repeat composition

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Sequencing Errors and Assembly

Sequencing errors add complexity to assembly

• Obscure overlaps, require shorter words
• Complicates graph by introducing spurs, bubbles, tips, etc
• Increases the effective repeat rate
• Potentially high error rate in consensus
Consensus Accuracy and Coverage

Coverage can overcome most random errors

- Simulate layout of 1000bp reads with random errors
- Compute accuracy of consensus call
Overlap Seeds and Error Rate

Random errors obscure overlap seeds

- Simulate layout of 20x coverage of 1000bp reads
- What fraction of overlapping reads match for at least S bp?
Approach 1: SMRT-de novo

- De novo assembly of SMRT-reads
  - Rapid sequencing and assembly
  - Long reads to span repeats

- Challenges
  - 15% error rate per read equates to 30% error rate per overlap
  - CCS reads as shorter, but higher quality reads
1. **Pre-overlap**
   - Consistency checks
2. **Trimming**
   - Quality trimming & partial overlaps
3. **Compute Overlaps**
   - Find high quality overlaps
4. **Error Correction**
   - Evaluate difference in context of overlapping reads
5. **Unitigging**
   - Merge consistent reads
6. **Scaffolding**
   - Bundle mates, Order & Orient
7. **Finalize Data**
   - Build final consensus sequences

http://wgs-assembler.sf.net
Yeast – Long reads

969,445 reads after filtering
Mean: 710 +/- 663
Median: 558 Max: 8,495

Yeast – CCS reads

731,638 reads after filtering
Mean: 306 +/- 115
Median: 279 Max: 1,425
Read Accuracy

Yeast – Long reads

94% aligned reads
48% reads aligned >100bp
7% reads aligned >1kbp

Yeast – CCS reads

99.93% aligned reads
98.2% reads aligned >100bp
38.8% reads aligned >300bp
Alignment Quality

<table>
<thead>
<tr>
<th>Match</th>
<th>83.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mismatch</td>
<td>1.4%</td>
</tr>
<tr>
<td>Insertions</td>
<td>11.5%</td>
</tr>
<tr>
<td>Deletions</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

Sample of 1M reads aligned with BLASR requiring >100bp alignment
SMRT-de novo Results

• De novo assembly of long reads
  – Experiments in progress
  – Very challenging to find good overlaps with very high error rate

• De novo assembly of CCS reads
  – Contig N50: 24,582bp

• De novo assembly of ref-corrected CCS
  – Contig N50: 65,119bp
Approach 2: SMRT-scaffolding

• Use long reads (or strobe reads) to link high quality contigs from short reads
  – Long reads (orange) span repetitive short-read contig (red)

• Challenges
  – Creating good draft assembly
  – Properly aligning reads to contigs
  – Untangling complex repeats
Error Correction with Quake

1. Count all “Q-mers” in reads
   - Fit coverage distribution to mixture model of errors and regular coverage
   - Automatically determines threshold for trusted k-mers

2. Correction Algorithm
   - Considers editing erroneous kmers into trusted kmers in decreasing likelihood
   - Includes quality values, nucleotide/nucleotide substitution rate

**Quake:** quality-aware detection and correction of sequencing reads.
**Illumina Sequencing & Assembly**

**Quake Results**
2x76bp @ 275bp
2x36bp @ 3400bp

**SOAPdenovo Results**

<table>
<thead>
<tr>
<th></th>
<th># ≥ 100bp</th>
<th>N50 (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffolds</td>
<td>2,340</td>
<td>253,186</td>
</tr>
<tr>
<td>Contigs</td>
<td>2,782</td>
<td>56,374</td>
</tr>
<tr>
<td>Unitigs</td>
<td>4,151</td>
<td>20,772</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Validated</th>
<th>51,243,281</th>
<th>88.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Corrected</td>
<td>2,763,380</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>Trim Only</td>
<td>3,273,428</td>
<td>5.6%</td>
</tr>
<tr>
<td></td>
<td>Removed</td>
<td>606,251</td>
<td>1.0%</td>
</tr>
</tbody>
</table>
SMRT-scaffolding results

SMRTpipe hybrid scaffold of SOAPdenovo assembly + >2kbp long reads
- Scaffold N50: 310,246bp (+22% improvement)
- Scaffold cnt: 2246 (4% reduction)

SMRTpipe hybrid scaffold of ref-CCS assembly + >2kbp long reads
- Scaffold N50: 97,414bp (+50% improvement)
- Scaffold cnt: 6,610 (3% reduction)
Approach 3: SMRT-hybrid

• Co-assemble long reads and short reads
  – Long reads natively span repeats (red)
  – Guards against mis-assemblies in draft assembly
  – Use all available data at once

• Challenges
  – Long reads have too high of an error rate to assemble directly
  – Assembler must support a wide mix of read lengths
SMRT-hybrid Error Correction & Assembly

1. Trim/correct SR sequence

2. Compute an SR layout for each LR
   1. Map SRs to LRs
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR

3. Co-assemble corrected LRs and SRs
   - Celera Assembler enhanced to support 16 Kbp reads

Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina
### 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>39 197</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>454 50X</td>
<td>4 714 344</td>
<td>66</td>
<td>308 060</td>
<td>161 109</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td>5 077 294</td>
<td>23</td>
<td>1 412 332</td>
<td>356 148</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both PacBio 25X + 454 50X</td>
<td>5 049 276</td>
<td>21</td>
<td>1 207 754</td>
<td>551 820</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>1.23Gbp</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
Conclusions

• SMRT-sequencing extremely promising for de novo assembly
  – Compute high quality consensus sequence from error prone reads
  – Long reads are challenging to use alone, but very effective when combined with high quality CCS or short reads

• Error correction is key to unlocking potential of SMRT-sequencing
  – The leading second generation sequence assemblers aggressively compensate for the platform specific error model
  – Easy prediction: same will be true for 3\textsuperscript{rd} generation assemblers

• Significant challenges ahead
  – Technology: Throughput, accuracy, read length, cost
  – Protocols: Sample preparation, library construction, read types
  – Informatics: Robust computational analysis methods
Acknowledgements

CSHL
Dick McCombie
Melissa Kramer
Eric Antoniou
Laura Gelley
Stephanie Muller

NBACC
Adam Phillippy
Sergey Koren

UMD
Steven Salzberg
Mihai Pop
David Kelley

JCVI
Brian Walenz

JGI
Jeffrey Martin

Duke
Erich Jarvis

UMD SOM
David Rasko
Thank You

http://schatzlab.cshl.edu
@mike_schatz
Repeats and Coverage Statistics

If $n$ reads are a uniform random sample of the genome of length $G$, we expect $k = n \Delta / G$ reads to start in a region of length $\Delta$.

- If we see many more reads than $k$ (if the arrival rate is $> A$), it is likely to be a collapsed repeat.
- Requires an accurate genome size estimate.

\[
\Pr(X - \text{copy}) = \binom{n}{k} \left( \frac{X \Delta}{G} \right)^k \left( \frac{G - X \Delta}{G} \right)^{n-k}
\]

\[
A(\Delta, k) = \ln \left( \frac{\Pr(1 - \text{copy})}{\Pr(2 - \text{copy})} \right) = \ln \left( \frac{(\Delta n / G)^k e^{-\Delta n / G}}{k!} \right) = \frac{n \Delta}{G} - k \ln 2
\]
Lander-Waterman statistics

L = read length
T = minimum overlap
G = genome size
N = number of reads
c = coverage \(\frac{NL}{G}\)
\(\sigma = 1 - \frac{T}{L}\)

\[E(\#\text{islands}) = Ne^{-c \sigma}\]
\[E(\text{island size}) = \frac{L(e^{c \sigma} - 1)}{c + 1 - \sigma}\]
contig = island with 2 or more reads