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

Dec. 9, 2011
CSHL In house
Schatz Lab Overview

Computation

Human Genetics

Modeling

Plant Genomics
Assembling a Genome

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads

   ...AGCCTAGACCTACA  GGATGCGCGACACGT
   GGATGCGCGACACGT   CGCATATCCGGT...

3. Simplify assembly graph

4. Detangle graph with long reads, mates, and other links
Assembly Applications

Novel genomes

Metagenomes

Sequencing assays
  • Transcript assembly
  • Structural variations
  • Haplotype analysis
  • …
Algorithms Overview

1. micro- 
   – Microsatellite mutations
   – Haplotype Microassembly

2. mega- 
   – Genome Dark Matter
   – Cloud-scale Genome analysis
   – Single Molecule Sequencing & Assembly

3. meta- 
   – Assembly Forensics & Metassembly
micro-
**MicroSeq: Microsatellite Analysis**

M. Bekritsky, J. Troge, D. Levy, M. Wigler, M. Schatz

- Highly variable simple sequence repeats
  - …GCACACACACAT… = …G(CA)₅T…
  - Mutate by slippage during replication, creating indels
  - High mutation rate makes it a useful marker for inferring phylogeny, associated with many diseases

- Genotyping with MicroSeq:
  1. Rapidly detect MS in short reads
  2. Map reads using a new MS-mapper
  3. Analyze profiles across populations

- Currently looking at de novo mutations associated with autism
**Scalpel: Haplotype Microassembly**


- Use assembly techniques to identify complex variations from short reads
  - Improved power to find indels
  - Trace candidate haplotypes sequences as paths through assembly graphs

Ref:  
Father:  
Mother_1:  
Mother_2:  
Aut_2:  
Aut_1:  
Sib_1:  
Sib_2:  

24 bp heterozygous indel at chr5:176026122 GPRIN1
mega-
Genomic Dark Matter

Hayan Lee, Michael Schatz

• Short read mapping is a essential for identifying mutations in the genome
  – Not every base of the genome can mapped equally well, especially because of repeats

• Introduced a new probabilistic metric - the Genome Mappability Score - that quantifies how reliably reads can be mapped to every position in the genome
  – We have little power to measure 11-13% of the human genome, including of known clinically relevant variations
  – Errors in variation discovery are dominated by false positives, especially in low GMS regions

Genomic Dark Matter: The reliability of short read mapping illustrated by the GMS.
**Jnomics: Cloud-scale genomics**

Matt Titmus, James Gurtowski, Michael Schatz

- Rapid parallel execution of NGS analysis pipelines
  - FASTX, BWA, Novoalign, SAMTools, Hydra
  - Sorting, merging, filtering, selection, of BAM, SAM, BED, fastq
- Case study: Structural variations in esophageal cancer

**Answering the demands of digital genomics**

Pacific Biosciences RS
Single Molecule Real Time (SMRT) Sequencing

Imaging of fluorescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).
SMRT Read Types

- **Standard sequencing**
  - Long inserts so that the polymerase can synthesize along a single strand

- **Circular consensus sequencing**
  - Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule.

- **Strobe sequencing**
  - Very long inserts, alternate the lasers in the instrument between on and off. On periods generate strobe sub-reads and the off periods determine the length of the spacing between, known as strobe advance.
Read Quality

**Yeast**  
(12 Mbp)

65 SMRT cells  
734,151 reads after filtering  
Mean: 642.3 +/- 587.3  
Median: 553 Max: 8,495

Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch
PacBio Error Correction & Assembly

http://wgs-assembler.sf.net

1. Correction Pipeline
   1. Map short reads (SR) to long reads (LR)
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR

2. Co-assemble corrected LRs and SRs
   – Celera Assembler enhanced to support 32 Kbp reads

3. Error corrected reads can be easily assembled, aligned

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina
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
Assembly Results

SMRT-hybrid assembly results of 50x PacBio corrected coverage of E. coli K12
Long reads lead to **contigs** over 1Mbp
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><em>Lambda</em> NEB3011</td>
<td>Illumina 50X 200bp</td>
<td>48 502</td>
<td>48 452</td>
<td>1</td>
<td>48 452</td>
<td>48 440</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td></td>
<td>48 440</td>
<td>1</td>
<td>48 440</td>
<td>48 440</td>
<td>0</td>
</tr>
<tr>
<td><em>E. coli</em> K12</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></td>
<td>4 473 206</td>
<td>79</td>
<td>222 024</td>
<td>66 408</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Both 20X PacBio + Illumina 50X 500bp</td>
<td>4 516 224</td>
<td>4 516 224</td>
<td>67</td>
<td>374 849</td>
<td>93 148</td>
<td>8</td>
</tr>
<tr>
<td><em>E. coli</em> 227-11</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></td>
<td>5 252 618</td>
<td>56</td>
<td>379 516</td>
<td>162 597</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td></td>
<td>5 397 525</td>
<td>41</td>
<td>596 739</td>
<td>216 129</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>PacBio 50X</td>
<td></td>
<td>5 476 824</td>
<td>39</td>
<td>1 057 326</td>
<td>365 964</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>PacBio 75X</td>
<td></td>
<td>5 601 310</td>
<td>55</td>
<td>642 068</td>
<td>308 312</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Both PacBio 50X + CSS 25X</td>
<td></td>
<td>5 453 558</td>
<td>33</td>
<td>1 167 060</td>
<td>527 198</td>
<td>8</td>
</tr>
<tr>
<td><em>E. coli</em> 17-2</td>
<td>Illumina 50X 500bp</td>
<td>4 929 374</td>
<td>3 018 23</td>
<td>71</td>
<td>301 823</td>
<td>108 581</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Illumina 50X 500bp + 50X 3Kbp</td>
<td>5 138 293</td>
<td>3 914 61</td>
<td>58</td>
<td>391 461</td>
<td>190 996</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Illumina 50X 500bp + 50X 6Kbp</td>
<td>5 157 771</td>
<td>4 391 168</td>
<td>46</td>
<td>403 168</td>
<td>186 135</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Illumina 50X 500bp + 50X 3Kbp + 50X 6Kbp</td>
<td>5 140 142</td>
<td>3 972 948</td>
<td>60</td>
<td>397 294</td>
<td>153 941</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>PacBio 25X</td>
<td></td>
<td>5 277 371</td>
<td>38</td>
<td>424 482</td>
<td>285 861</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Both PacBio 25X + Illumina 50X 500bp</td>
<td>5 410 343</td>
<td>3 912 608</td>
<td>41</td>
<td>912 608</td>
<td>286 829</td>
<td>9</td>
</tr>
<tr>
<td><em>E. coli</em> JM211</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></td>
<td>4 912 923</td>
<td>57</td>
<td>420 268</td>
<td>118 962</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Both PacBio 25X + Illumina 50X 300bp</td>
<td>4 995 486</td>
<td>4 995 486</td>
<td>54</td>
<td>423 420</td>
<td>125 900</td>
<td>-</td>
</tr>
<tr>
<td><em>S. cerevisiae</em> S228c</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></td>
<td>11 101 617</td>
<td>226</td>
<td>191 587</td>
<td>63 095</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Both PacBio 13X + Illumina 50X 300bp</td>
<td>12 157 105</td>
<td>12 157 105</td>
<td>207</td>
<td>323 716</td>
<td>67 117</td>
<td>21</td>
</tr>
<tr>
<td><em>Melopsittacus undulatus</em></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></td>
<td>882 984 450</td>
<td>237 121</td>
<td>51 333</td>
<td>3 250</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Lander Waterman 3X Prediction</td>
<td></td>
<td>1 153 148 167</td>
<td>173 565</td>
<td>69 663</td>
<td>9 026</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
Transcript Alignment

- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
  - Raw reads and raw alignments (red) have many spurious indels inducing false frameshifts and other artifacts
  - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing

- New collaboration with Gingeras Lab looking at splicing in human
meta-
2011: Year of the Assembly Bakeoff


- Simulated genome distantly related to human chr13
- 17 labs, 50+ assemblies
- 4 real genomes ranging from bacteria to individual human chromosome
- Internal evaluation of 8 assemblers
Assemblathon Results

Scaffolds
- Broad
- DOE/JGI
- CSHL

Contig Paths
- BGI
- Broad
- CSHL

Mis-assembly Markers
- BGI.1
- WTSI-P.1
- BCCGSC.5
Final Rankings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BGI</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WTSI-S</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSHL</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCCGSC</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOEJGI</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RHUL</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WTSI-P</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBI</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRACS</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- SOAPdenovo and ALLPATHS came out neck-and-neck followed closely behind by SGA, Celera Assembler, and ABySS
- My recommendation for “typical” short read assembly is to use ALLPATHS
Assemblathon 2

• Real sequence data, *de novo* assembly

• Step 1: Apply best practices from Assemblathon 1
• Step 2: Add secret weapon for winning...

Images from Assemblathon
Assembly Forensics

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.
Basic mis-assemblies can be combined into more complicated patterns: Insertions, Deletions, Giant Hairballs
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
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
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 \leq -3.0 indicates Compression
Hawkeye & AMOS: Visualizing and assessing the quality of genome assemblies
Assemblathon 2

• Real sequence data, \textit{de novo} assembly

• Step 1: Apply best practices from Assemblathon 1
• Step 2: Add secret weapon for winning...

Images from Assemblathon
Assemblathon 2: Metassembly
Paul Baranay, Scott Emrich, Michael Schatz

Scaffold N50: 3,710,017
#>1000: 2,791

Contig N50: 20,183
#>1000: 68,591

Scaffold N50: 285,413
#>1000: 29,119

Contig N50: 1,607
#>1000: 218,643

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

Assembly is a powerful tool for analyzing sequences, and is moving to increasingly more complex genomes and data types.

- Microassembly is a powerful tool needed to fully understand the genetics of autism and other diseases.

- A global analysis of the genome requires new statistics and computational methods to understand the patterns that we observe.

- Metassembly lets us maximize connectivity without sacrificing the quality of a de novo assembly.
Acknowledgements

Mitch Bekritsky
Giuseppe Narzisi

Ivan Iossifov
Wigler Lab

Hayan Lee
Matt Titmus
James Gurtowski
Ware Lab
McCombie Lab

Adam Phillippy (NBACC)
Sergey Koren (NBACC)

Paul Baranay (CSHL/ND)
Scott Emrich (ND)
Steven Salzberg (JHU)
Mihai Pop (UMD)
Thank You

http://schatzlab.cshl.edu
@mike_schatz