Entering the era of mega-genomics

Michael Schatz

March 20, 2012
JGI Users Meeting

@mike_schatz / #JGIUM7
Outline

1. Milestones in genomics
   1. Sanger to nanopore
   2. 21st Century Mega-Genomics

2. Applications of mega-genomics
   1. Cloud-scale genomics for bioenergy
   2. Single molecule sequencing & assembly
   3. De novo mutations in autism
Advances in Sequencing: Zeroth, First, Second Generation

1970s: 0th Gen
Radioactive Chain Termination
5000bp / week

1980s-1990s: 1st Gen
Automated Capillary Sequencing
384kbp / day

2000s: 2nd Gen
Pyrosequencing, SOLiD Sequencing-by-Synthesis
1Gbp+ / day
Advances in Sequencing: Now Generation Sequencing

- **Illumina HiSeq 2000**
  - Sequencing by Synthesis
  - >60 Gbp / day

- **Ion Proton**
  - Postlight Sequencing
  - >100 Gbp / day

- **Oxford Nanopore**
  - Nanopore sensing
  - Many GB / day
Collective capacity exceed 15 Pbp / year
The rise of mega-genomics

De novo Assembly

Alignment & Variations

Differential Analysis

Phylogeny, Evolution, and Modeling
The rise of mega-genomics

De novo Assembly

Alignment & Variations

Differential Analysis

Phylogeny, Evolution, and Modeling
Ingredients for a good assembly

**Coverage**

- High coverage is required
  - Oversample the genome to ensure every base is sequenced with long overlaps between reads
  - Biased coverage will also fragment assembly

**Read Length**

- Reads & mates must be longer than the repeats
  - Short reads will have false overlaps forming hairball assembly graphs
  - With long enough reads, assemble entire chromosomes into contigs

**Quality**

- Errors obscure overlaps
  - Reads are assembled by finding kmers shared in pair of reads
  - High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in de novo plant genome sequencing and assembly
Hybrid Sequencing

**Illumina**
*Sequencing by Synthesis*
- High throughput (60Gbp/day)
- High accuracy (~99%)
- Short reads (~100bp)

**Pacific Biosciences**
*SMRT Sequencing*
- Lower throughput (600Mbp/day)
- Lower accuracy (~85%)
- Long reads (10kbp+)
SMRT Sequencing Data

Yeast
(12 Mbp genome)

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

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. Error corrected reads can be easily assembled, aligned

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
Celera Assembler
http://wgs-assembler.sf.net

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
### SMRT-Assembly Results

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

***Also useful for transcriptome, repeat, and other analysis***

<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>
</tr>
</thead>
<tbody>
<tr>
<td>Lambda NEB3011</td>
<td>Illumina 100X 200bp</td>
<td>48 502</td>
<td>48 492</td>
<td>1</td>
<td>48 492 / 48 492</td>
<td>48 492 / 48 492 (100%) *</td>
</tr>
<tr>
<td></td>
<td>PacBio PBCr 25X</td>
<td></td>
<td>48 440</td>
<td>1</td>
<td>48 444 / 48 444</td>
<td>48 444 / 48 444 (100%) *</td>
</tr>
<tr>
<td>E. coli K12</td>
<td>Illumina 100X 500bp</td>
<td>4,639,675</td>
<td>4,462,836</td>
<td>61</td>
<td>221,615 / 221,553</td>
<td>100,338 / 83,037 (82.76%) *</td>
</tr>
<tr>
<td></td>
<td>PacBio PBCr 18X</td>
<td></td>
<td>4,465,533</td>
<td>77</td>
<td>239,058 / 238,224</td>
<td>71,479 / 68,309 (95.57%) *</td>
</tr>
<tr>
<td></td>
<td>Both 1X PacBio PBCr + Illumina 50X 500bp</td>
<td>4,576,046</td>
<td>65</td>
<td>238,272 / 238,224</td>
<td>93,048 / 89,431 (96.11%) *</td>
<td></td>
</tr>
<tr>
<td>E. coli C227-11</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>
</tr>
<tr>
<td></td>
<td>PacBio 25X PBCr (corrected by 25X CCS)</td>
<td></td>
<td>5,207,946</td>
<td>80</td>
<td>357,234</td>
<td>98,774</td>
</tr>
<tr>
<td></td>
<td>Both PacBio PBCr 25X + CCS 25X</td>
<td></td>
<td>5,269,158</td>
<td>39</td>
<td>647,362</td>
<td>227,302</td>
</tr>
<tr>
<td></td>
<td>PacBio 50X PBCr (corrected by 50X CCS)</td>
<td></td>
<td>5,445,466</td>
<td>35</td>
<td>1,076,027</td>
<td>376,443</td>
</tr>
<tr>
<td></td>
<td>Both PacBio PBCr 50X + CCS 25X</td>
<td></td>
<td>5,453,458</td>
<td>33</td>
<td>1,167,060</td>
<td>527,198</td>
</tr>
<tr>
<td></td>
<td>Manually Corrected ALLORA Assembly³</td>
<td></td>
<td>5,452,251</td>
<td>23</td>
<td>653,382</td>
<td>402,041</td>
</tr>
<tr>
<td>S. cerevisiae 5228c</td>
<td>Illumina 100X 300bp</td>
<td>12,157,105</td>
<td>11,034,156</td>
<td>192</td>
<td>266,528 / 227,714</td>
<td>73,871 / 49,254 (66.68%) *</td>
</tr>
<tr>
<td></td>
<td>PacBio PBCr 13X</td>
<td></td>
<td>11,110,420</td>
<td>224</td>
<td>224,478 / 217,704</td>
<td>62,898 / 54,633 (86.86%) *</td>
</tr>
<tr>
<td></td>
<td>Both PacBio PBCr 13X + Illumina 50X 300bp</td>
<td></td>
<td>11,286,932</td>
<td>177</td>
<td>262,846 / 260,794</td>
<td>82,545 / 59,792 (72.44%) *</td>
</tr>
<tr>
<td>Melopsittacus undulatus</td>
<td>Illumina 194X (220/500/800 paired-end 2/5/10Kb mate-pairs)</td>
<td>1.23 Gbp</td>
<td>1,023,532,850</td>
<td>24</td>
<td>1,090,202</td>
<td>47,383</td>
</tr>
<tr>
<td></td>
<td>454 15.4X (FLX + FLX Plus + 3/8/20Kbp paired-ends)</td>
<td></td>
<td>999,168,029</td>
<td>16</td>
<td>751,729</td>
<td>75,178</td>
</tr>
<tr>
<td></td>
<td>(median 997, max. 13 070)</td>
<td></td>
<td>1,071,356,415</td>
<td>15</td>
<td>1,238,843</td>
<td>99,573</td>
</tr>
</tbody>
</table>
The rise of mega-genomics

De novo Assembly

Alignment & Variations

Differential Analysis

Phylogeny, Evolution, and Modeling
A unified genetic theory for sporadic and inherited autism
Zhao et al. (2007) PNAS. 104(31)12831-12836.
Autism and de novo CNVs

Analysis of Simons Simplex Collection
- CGH arrays of 510 family quads
- 94 total de novo CNVs discovered

De novo CNVs are more common in autistic children
- 4:1 ratio in autistic kids relative to their non-autistic siblings
- Some recurrence at genes related to other psychiatric conditions

Rare de novo and transmitted copy-number variation in autism spectrum disorders. Levy et al. (2011) Neuron. 70:886-897.
Exome Sequencing Pipeline

Data (lane) FASTQ → Filtering → Family Demultiplexing → Individual Aggregation

Alignment (BWA) to reference genome → ‘Orphans’ (2nd pass homology search) → SNP (GATK) → Indel (GATK) → CNV (HMM) → Microsatellite Analysis

Completed examination of ~350 families (~1/10th total study) using published and custom code

Families prepared and captured together to minimize batch effects as much as possible
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:  ...TCAGAACAGCTGGATGAGATCTTAGCCCAACTACCAGGAGATTGTCTTTGCCCCGA...
Father:  ...TCAGAACAGCTGGATGAGATCTTAGCCCAACTACCAGGAGATTGTCTTTGCCCCGA...
Mother:  ...TCAGAACAGCTGGATGAGATCTTAGCCCAACTACCAGGAGATTGTCTTTGCCCCGA...
Sib:  ...TCAGAACAGCTGGATGAGATCTTAGCCCAACTACCAGGAGATTGTCTTTGCCCCGA...
Aut(1):  ...TCAGAACAGCTGGATGAGATCTTAGCCCAACTACCAGGAGATTGTCTTTGCCCCGA...
Aut(2):  ...TCAGAACAGCTGGATGAGATCTTACC-------CCGGAGATTGTCTTTGCCCCGA...

6bp heterozygous indel at chr13:25280526 ATP12A
De novo Genetics of Autism

• In 343 family quads so far, we see significant enrichment in de novo *likely gene killers* in the autistic kids
  – Overall rate basically 1:1 (432:396)
  – 2:1 enrichment in nonsense mutations
  – 2:1 enrichment in frameshift indels
  – 4:1 enrichment in splice-site mutations

• Observe strong overlap with the 842 genes known to be associated with fragile X mental retardation.
  – These genes relate to neuron and brain development
  – Suggest these genes are under strong purifying selection and we hypothesize particularly dosage sensitive

*Exome sequence analysis of simplex families with children on the autism spectrum* Iossifov et al. (2012) Under review
Mega-Genomics Challenges

The foundations of genomics will continue to be observation, experimentation, and interpretation

– Technology will continue to push the frontier
– Measurements will be made digitally over large populations, at extremely high resolution, and for diverse applications

Rise in Quantitative and Computational Demands

1. **Experimental design**: selection, collection & metadata
2. **Observation**: measurement, storage, transfer, computation
3. **Integration**: multiple samples, assays, analyses
4. **Discovery**: visualizing, interpreting, modeling

*Ultimately limited by the human capacity to execute extremely complex experiments and interpret results*
Acknowledgements

James Gurtowski  
Matthew Titmus  
Ware Lab  
KBase Members

Paul Baranay (CSHL/ND)  
Scott Emrich (ND)  
Steven Salzberg (JHU)  
Adam Phillippy (NBACC)  
Sergey Koren (NBACC)

Giuseppe Narzisi  
Mitch Bekritsky  
Ivan Iossifov  
Wigler Lab
Thank You!

http://schatzlab.cshl.edu/apply
@mike_schatz / #JGIUM7