Assembly of “large” metagenomes using a hybrid core computer

“How I turned 600 billion base pairs of reads into 9Gb of contigs”

Alex Copeland
DOE Joint Genome Institute,
Walnut Creek, CA

SFAF
June 7, 2012
• Motivation
• System
• Data
• Results
• Conclusions
Current Production Pipeline

Trim_illumina_reads.pl → quality_report.pl (TBD)

postTrimIllumina.pl

SOAPdenovo 105 hash → SOAPdenovo 101 hash → SOAPdenovo 97 hash → SOAPdenovo 93 hash → SOAPdenovo 89 hash → SOAPdenovo 85 hash

dereplicate_contigs.pl

SequentialRunCombineAssemblies.pl

dereplicate_contigs.pl → quality_report.pl (TBD)
• Cluster: 3 1Tb nodes; 4 0.5Tb nodes
• 0.3-1Tb RAM / assembly
  — ~40% need > 0.5Tb RAM
• RAM prediction
  — avoids core dumps and swapping
  — more efficient use of resources
Using 50 assemblies (6 hash lengths; 9 libraries) all models predict >500G assemblies with less than 5% error.
Main genome contig total: 954375
Main genome contig sequence total: 648.6 MB (→ 0.0% gap)
Main genome contig N/L50: 128384/953
Number of contigs > 50 KB: 110
% main genome in contigs > 50 KB: 1.2%

<table>
<thead>
<tr>
<th>Minimum Scaffold Length</th>
<th>Number of Scaffolds</th>
<th>Number of Contigs</th>
<th>Total Scaffold Length</th>
<th>Total Contig Length</th>
<th>Scaffold Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>954,375</td>
<td>954,375</td>
<td>648,619,883</td>
<td>648,619,883</td>
<td>100.00%</td>
</tr>
<tr>
<td>1 kb</td>
<td>120,873</td>
<td>120,873</td>
<td>316,979,541</td>
<td>316,979,541</td>
<td>100.00%</td>
</tr>
<tr>
<td>2.5 kb</td>
<td>33,102</td>
<td>33,102</td>
<td>184,133,672</td>
<td>184,133,672</td>
<td>100.00%</td>
</tr>
<tr>
<td>5 kb</td>
<td>10,410</td>
<td>10,410</td>
<td>107,138,152</td>
<td>107,138,152</td>
<td>100.00%</td>
</tr>
<tr>
<td>10 kb</td>
<td>2,984</td>
<td>2,984</td>
<td>56,991,775</td>
<td>56,991,775</td>
<td>100.00%</td>
</tr>
<tr>
<td>25 kb</td>
<td>506</td>
<td>506</td>
<td>21,379,196</td>
<td>21,379,196</td>
<td>100.00%</td>
</tr>
<tr>
<td>50 kb</td>
<td>110</td>
<td>110</td>
<td>8,036,371</td>
<td>8,036,371</td>
<td>100.00%</td>
</tr>
<tr>
<td>100 kb</td>
<td>14</td>
<td>14</td>
<td>1,714,605</td>
<td>1,714,605</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
• >100 completed metagenomes
• Typical assembly 2%-30% reads map
• Can handle 1-2 HiSeq lanes (<120Gbp)
No problem
Convey Hybrid-core Architecture

“Commodity” Intel Server

- Intel® Xeon® processor
- Intel® Memory Controller Hub (MCH)
- Intel® I/O Subsystem
- Standard Intel® x86-64 Server
  - x86-64 Linux

Intel Xeon 2.13 GHz quad-core processor; 128GB RAM

Convey FPGA-based coprocessor

- Application Engine Hub (AEH)
- Application Engines (AEs)
- Xilinx Virtex 5 LX330 FPGA; 64 GB (Convey SG DIMMs); 16 custom memory controllers

Convey coprocessor
- FPGA-based
- Highly parallel memory: 8192 simultaneous requests
- 76.8 GB/sec bandwidth

Convey Computer
Hardware Acceleration

- **RAM reduced 2/3**
  - 38GB to 13GB
  - Data types / structures
  - Automated roadmap partitioning
- **Speed up depends on**
  - Data set size
  - Kmer space complexity

X86/host: 2 Intel X5670 2.93GHz processors (12 cores total), stripe 4 @ 600GB SATA disks
HC-1: 128GB (host), 64GB (coproc), stripe 2 @ 1TB SATA disks
HC-2: 96GB DDR3 (host), 16GB SG (coprocessor)
Velvet on the HC1

velvetg
- Graph
- Scaffolding
- Cleanup Reports

Cnygc
- Unify
- Roadmap
- Graph
- Cleanup Reports

Input Sequences
- CnyUnifiedSeq
- CnyRoadMap
- Graph or Graph2

LastGraph
- Contig Stats
- AMOS

Convey Computer
• Roadmap - list of kmer matches between sequences
  • Construct in parallel on hardware
    – transparent to user
    – hash used to assign kmers to one of 32 match engines
• Reduces memory
CnyRoadMap - Partitioning

- User selects partition to process
  - partitions are merged later in input to graph
  - partitions are independent and can be generated in parallel on multiple systems
- Can divide data into 32 partitions x 32 subpart. = 1024 total chunks
- Can generate roadmaps for hundreds of GBp of sequence
Soil Metagenomes

Great Prairie

Pilot study for DOE Grand Challenge

-soil

-Major carbon sink and source
-Stores 3X more carbon than vegetation
-Highly complex communities

Science leads: Jim Tiedje, Janet Jansson
Prairie soils
• 27% of continental US land surface
• 31-39% of soil organic carbon stocks

J. Jansson
Sampling sites

• Wisconsin
  – Native prairie (Goose Pond, Audubon)
  – Long term cultivation (corn)
  – Switchgrass rotation (previously corn)
  – Restored prairie (from 1998)
• Iowa
  – Native prairie (Morris prairie)
  – Long term cultivation (corn)
• Kansas
  – Native prairie (Konza prairie)
  – Long term cultivation (corn)
<table>
<thead>
<tr>
<th>Location</th>
<th>Bases (B)</th>
<th>Reads (B)</th>
<th>Readlen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kansas Native Prairie</td>
<td>598</td>
<td>5.12</td>
<td>95</td>
</tr>
<tr>
<td>Kansas Continuous Corn</td>
<td>271</td>
<td>2.67</td>
<td>78</td>
</tr>
<tr>
<td>Iowa Native Prairie</td>
<td>367</td>
<td>3.75</td>
<td>69</td>
</tr>
<tr>
<td>Iowa Continuous Corn</td>
<td>197</td>
<td>2.05</td>
<td>74</td>
</tr>
<tr>
<td>Wisconsin Native Prairie</td>
<td>199</td>
<td>2.09</td>
<td>80</td>
</tr>
<tr>
<td>Wisconsin Continuous Corn</td>
<td>192</td>
<td>1.91</td>
<td>84</td>
</tr>
<tr>
<td>Wisconsin Restored Prairie</td>
<td>53</td>
<td>0.35</td>
<td>107</td>
</tr>
<tr>
<td>Wetland Surface Sediment</td>
<td>772</td>
<td>4.75</td>
<td>120</td>
</tr>
<tr>
<td>Location</td>
<td>Bases (B)</td>
<td>Reads (B)</td>
<td>Contigs (M)</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>Kansas Native Prairie</td>
<td>598</td>
<td>5.12</td>
<td>27.8</td>
</tr>
<tr>
<td>Kansas Continuous Corn</td>
<td>271</td>
<td>2.67</td>
<td>4.9</td>
</tr>
<tr>
<td>Iowa Native Prairie</td>
<td>367</td>
<td>3.75</td>
<td>10.7</td>
</tr>
<tr>
<td>Iowa Continuous Corn</td>
<td>197</td>
<td>2.05</td>
<td>2.7</td>
</tr>
<tr>
<td>Wisconsin Native Prairie</td>
<td>199</td>
<td>2.09</td>
<td>7.1</td>
</tr>
<tr>
<td>Wisconsin Continuous Corn</td>
<td>192</td>
<td>1.91</td>
<td>3.9</td>
</tr>
<tr>
<td>Wisconsin Restored Prairie</td>
<td>53</td>
<td>0.35</td>
<td>0.6</td>
</tr>
<tr>
<td>Wetland Surface Sediment</td>
<td>772</td>
<td>4.75</td>
<td>…</td>
</tr>
</tbody>
</table>
Kansas Native Prairie
Kansas Continuous Corn

Contigs average fold coverage vs. GC
Contigs average fold coverage vs. GC
Conclusions

• Using the HC1/Velvet
  – Assembly of >500Gbp possible
  – Assembly times are quite reasonable
  – No prefiltering required (probably beneficial)

• Next steps
  – Continue with larger datasets until something breaks
  – Improve assemblies
    • Prefilter data
    • Tune parameters
  – Analysis of contigs
Acknowledgements

• JGI
  – Brian Foster, Kurt Labutti
  – Susannah Tringe, Janet Jansson

• Convey Computer
  – George Vandegrift, George Vacek, Kirby Collins, Mike D’Jamoos, Wesley Hart

• U.S. Department of Energy
  – Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231