Scalable Solutions for DNA Sequence Analysis

Michael Schatz

Jan 29, 2010
NHGRI/UMD Joint Sequencing Meeting
The Evolution of DNA Sequencing

<table>
<thead>
<tr>
<th>Year</th>
<th>Genome</th>
<th>Technology</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Venter et al.</td>
<td>Sanger (ABI)</td>
<td>$300,000,000</td>
</tr>
<tr>
<td>2007</td>
<td>Levy et al.</td>
<td>Sanger (ABI)</td>
<td>$10,000,000</td>
</tr>
<tr>
<td>2008</td>
<td>Wheeler et al.</td>
<td>Roche (454)</td>
<td>$2,000,000</td>
</tr>
<tr>
<td>2008</td>
<td>Ley et al.</td>
<td>Illumina</td>
<td>$1,000,000</td>
</tr>
<tr>
<td>2008</td>
<td>Bentley et al.</td>
<td>Illumina</td>
<td>$250,000</td>
</tr>
<tr>
<td>2009</td>
<td>Pushkarev et al.</td>
<td>Helicos</td>
<td>$48,000</td>
</tr>
<tr>
<td>2009</td>
<td>Drmanac et al.</td>
<td>Complete Genomics</td>
<td>$4,400</td>
</tr>
</tbody>
</table>

(Pushkarev et al., 2009)

Critical Computational Challenges: Alignment and Assembly of Huge Datasets
Hadoop MapReduce

- Application developers focus on 2 (+1 internal) functions
  - Map: input ➔ key, value pairs
  - Shuffle: Group together pairs with same key
  - Reduce: key, value-lists ➔ output

ATG,1 AAC,1 CTT,1
TGA,1 ACC,1 TTA,1
GAA,1 CCT,1

ACA,1
ATG,1
CAA,1
GCA,1
TGA,1
TTA,1

ACT,1
AGG,1
CCT,1
GGC,1
TTT,1

AAC,1
ACC,1
CTT,1
GAA,1
TAG,1

ACA:1
ATG:1
CAA:2
GCA:1
TGA:1
TTA:3

Map, Shuffle & Reduce
All Run in Parallel
Short Read Mapping with MapReduce

- Given a reference and many subject reads, report one or more “good” end-to-end alignments per alignable read
  - Maps the read to where it originated

- Mapping of a whole human requires ~1,000 CPU hours
  - Alignments are “embarrassingly parallel” by read
  - Variant detection is parallel by chromosome region
Crossbow
http://bowtie-bio.sourceforge.net/crossbow

• Align billions of reads and find SNPs
  – Reuse software components: Hadoop Streaming

• Map: Bowtie (Langmead et al., 2009)
  – Find best alignment for each read
  – Emit (chromosome region, alignment)

• Shuffle: Hadoop
  – Group and sort alignments by region

• Reduce: SOAPsnp (Li et al., 2009)
  – Scan alignments for divergent columns
  – Accounts for sequencing error, known SNPs
Performance in Amazon EC2

http://bowtie-bio.sourceforge.net/crossbow

<table>
<thead>
<tr>
<th></th>
<th>Asian Individual Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Loading</td>
<td>3.3 B reads</td>
</tr>
<tr>
<td></td>
<td>106.5 GB</td>
</tr>
<tr>
<td></td>
<td>$10.65</td>
</tr>
<tr>
<td>Data Transfer</td>
<td>1h :15m</td>
</tr>
<tr>
<td></td>
<td>40 CPUs</td>
</tr>
<tr>
<td></td>
<td>$3.40</td>
</tr>
<tr>
<td>Setup</td>
<td>0h :15m</td>
</tr>
<tr>
<td></td>
<td>320 CPUs</td>
</tr>
<tr>
<td></td>
<td>$13.94</td>
</tr>
<tr>
<td>Alignment</td>
<td>1h :30m</td>
</tr>
<tr>
<td></td>
<td>320 CPUs</td>
</tr>
<tr>
<td></td>
<td>$41.82</td>
</tr>
<tr>
<td>Variant Calling</td>
<td>1h :00m</td>
</tr>
<tr>
<td></td>
<td>320 CPUs</td>
</tr>
<tr>
<td></td>
<td>$27.88</td>
</tr>
<tr>
<td>End-to-end</td>
<td>4h :00m</td>
</tr>
<tr>
<td></td>
<td>$97.69</td>
</tr>
</tbody>
</table>

Analyze an entire human genome for ~$100 in an afternoon.
Accuracy validated at 99%

Searching for SNPs with Cloud Computing.
Short Read Assembly

Genome assembly as finding an Eulerian tour of the de Bruijn graph
  - Human genome: ~3B nodes, ~10B edges

The new short read assemblers require tremendous computation
  - Velvet (Zerbino & Birney, 2008) on human > 2 TB of RAM
  - ABYSS (Simpson et al., 2009) on human ~4 days on 168 cores
Genome Assembly with MapReduce

1. Build Compressed de Bruijn Graph
2. Correct Errors & Resolve Short Repeats
3. Cloud Surfing: Mate directed repeat resolution & scaffolding

Assembly of Large Genomes with Cloud Computing.
Acknowledgements

Advisor

Steven Salzberg

UMD Faculty

Mihai Pop, Art Delcher, Amitabh Varshney, Carl Kingsford, Ben Shneiderman, James Yorke, Jimmy Lin, Dan Sommer

CBCB Students

Adam Phillippy, Cole Trapnell, Saket Navlakha, Ben Langmead, James White, David Kelley
Thank You!

http://www.cbcb.umd.edu/~mschatz