Outline

1. Genome Assembly by Analogy
2. DNA Sequencing and Genomics
3. Sequence Analysis Projects
   1. Mapping & Genotyping
   2. Microsatellite Profiling
   3. De novo assembly
Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of *A Tale of Two Cities*
  - Text printed on 5 long spools

  It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

  It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

  It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

  It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

  It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

- How can he reconstruct the text?
  - 5 copies × 138,656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical
Greedy Reconstruction

The repeated sequence makes the correct reconstruction ambiguous:

- It was the best of times, it was the [worst/age]

Model sequence reconstruction as a graph problem.
de Bruijn Graph Construction

- $D_k = (V,E)$
  - $V =$ All length-$k$ subfragments ($k < l$)
  - $E =$ Directed edges between consecutive subfragments
    - Nodes overlap by $k-1$ words

Original Fragment | Directed Edge
--- | ---
It was the best of | It was the best | was the best of

- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001
de Bruijn Graph Assembly

It was the best
was the best of
the best of times,
best of times, it
of times, it was
times, it was the

it was the worst
was the worst of
the worst of times,
worst of times, it

it was the age
was the age of
the age of foolishness
the age of wisdom,
age of wisdom, it
of wisdom, it was
wisdom, it was the

After graph construction, try to simplify the graph as much as possible
It was the best of times, it was the age of wisdom, it was the age of foolishness. After graph construction, try to simplify the graph as much as possible.
Genomics & DNA Sequencing

Genome of an organism encodes the genetic information in long sequence of 4 DNA nucleotides: ACGT

- Bacteria: ~3 million bp
- Humans: ~3 billion bp

Current DNA sequencing machines can sequence millions of short (25-500bp) reads from random positions of the genome

- Per-base error rate estimated at 1-2% (Simpson et al, 2009)

Like Dickens, we can only sequence small fragments of the genome at once.

- A single human genome requires ~100 GB of raw data
- We need extremely scalable systems and algorithms
The DNA Deluge

Exponential Growth of GenBank
Dec 1982 - Oct 2010


Too much information
MapReduce is Google's framework for large data computations
- Data and computations are spread over thousands of computers
  - Indexing the Internet, PageRank, Machine Learning, etc… (Dean and Ghemawat, 2004)
  - 946,460 TB processed in May 2010 (Jeff Dean at Stanford, 11.10.2010)

- Hadoop is the leading open source implementation
  - Developed and used by Yahoo, Facebook, Twitter, Amazon, etc
  - GATK is an alternative implementation specifically for NGS

**Benefits**
- Scalable, Efficient, Reliable
- Easy to Program
- Runs on commodity computers

**Challenges**
- Redesigning / Retooling applications
  - Not Condor, Not MPI
  - Everything in MapReduce
System Architecture

- Hadoop Distributed File System (HDFS)
  - Data files partitioned into large chunks (64MB), replicated on multiple nodes
  - Computation moves to the data, rack-aware scheduling

- Hadoop MapReduce system won the 2009 GreySort Challenge
  - Sorted 100 TB in 173 min (578 GB/min) using 3452 nodes and 4x3452 disks
Programming Models

Embarrassingly Parallel
- Map-only
  - Each item is Independent
  - Traditional Batch Computing

Loosely Coupled
- MapReduce
  - Independent-Shuffle-Independent
  - Batch Computing + Data Exchange

Tightly Coupled
- Iterative MapReduce
  - Nodes interact with other nodes
  - Big Data MPI
Short Read Mapping

- Given a reference and many subject reads, report one or more “good” end-to-end alignments per alignable read
  - Find where the read most likely originated
  - Fundamental computation for many assays
    - Genotyping
    - RNA-Seq
    - Methyl-Seq
    - Structural Variations
    - Chip-Seq
    - Hi-C-Seq

- Desperate need for scalable solutions
  - Single human requires >1,000 CPU hours / genome
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>Step</th>
<th>Time</th>
<th>Cores</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Loading</td>
<td>3.3 B reads</td>
<td>106.5 GB</td>
<td>$10.65</td>
</tr>
<tr>
<td>Data Transfer</td>
<td>1h :15m</td>
<td>40 cores</td>
<td>$3.40</td>
</tr>
<tr>
<td>Setup</td>
<td>0h : 15m</td>
<td>320 cores</td>
<td>$13.94</td>
</tr>
<tr>
<td>Alignment</td>
<td>1h : 30m</td>
<td>320 cores</td>
<td>$41.82</td>
</tr>
<tr>
<td>Variant Calling</td>
<td>1h : 00m</td>
<td>320 cores</td>
<td>$27.88</td>
</tr>
<tr>
<td>End-to-end</td>
<td>4h : 00m</td>
<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.**
Hadoop for NGS Analysis

**Quake**
Quality-aware error correction of short reads

Correct 97.9% of errors with 99.9% accuracy

(Kelly, Schatz, Salzberg, 2010*)

http://www.cbcb.umd.edu/software/quake/

**CloudBurst**
Highly Sensitive Short Read Mapping with MapReduce

100x speedup mapping on 96 cores @ Amazon

http://cloudburst-bio.sf.net

(Schatz, 2009)

**Myrna**
Cloud-scale differential gene expression for RNA-seq

Expression of 1.1 billion RNA-Seq reads in ~2 hours for ~$66

(Langmead, Hansen, Leek, 2010)

http://bowtie-bio.sf.net/myrna/

**AMOS**
Searching for SNPs in the Turkey Genome

Scan the de novo assembly to find 920k heterozygous alleles

(Dalloul et al, 2010)

http://amos.sf.net
SeqMS: NextGen Microsatellite Profiling

Mitchell Bekritsky, WSBS

• Class of simple sequence repeats
  – ...GCACACACACAT... = ...G(CA)₅T...
  – Created and mutate primarily through slippage during replication
  – Highly variable & ubiquitous

• Genotyping with SeqMS
  – Rapidly detect MS sequences
  – Map reads using a new MS-mapper
  – Analyze profiles in cells, across cells, & across populations
    • Loss of heterozygosity
    • Development of somatic & cancer cells
    • Relations across strains, across species
    • etc…

(Salipante et al. 2006)
Short Read Assembly

Reads
AAGA
ACTT
ACTC
ACTG
AGAC
CCGA
CGAC
CTCC
CTGG
CTTT
...

de Bruijn Graph

Potential Genomes
AAGACTCCGACTGGGACTTT
AAGACTGGGACTCCGACTTT

• 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) serial: > 2TB of RAM
  – ABYSS (Simpson et al., 2009) MPI: 168 cores x ~96 hours
  – SOAPdenovo (Li et al., 2010) pthreads: 40 cores x 40 hours, >140 GB RAM
Graph Compression

- After construction, many edges are unambiguous
  - Merge together compressible nodes
  - Graph randomly distributed over hundreds of computers

Design Patterns for Efficient Graph Algorithms in MapReduce.
Lin, J, Schatz, MC (2010) *Workshop on Mining and Learning with Graphs Workshop (KDD/MLG-2010)*
Fast Path Compression

Challenges
- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking
- Randomly assign \( \overline{H}/T \) to each compressible node
- Compress \( \overline{H} \rightarrow T \) links

Performance
- Compress all chains in \( \log(S) \) rounds
- <30 rounds for human genome

Randomized Speed-ups in Parallel Computation.
Node Types

Isolated nodes (10%)

Tips (46%)

Bubbles/Non-branch (9%)

Dead Ends (.2%)

Half Branch (25%)

Full Branch (10%)

(Chaisson, 2009)
De novo bacterial assembly

- **Genome:** *E. coli* K12 MG1655, 4.6Mbp
- **Input:** 20.8M 36bp reads, 200bp insert (~150x coverage)
- **Preprocessor:** Quake Error Correction

### Assembly of Large Genomes with Cloud Computing.
### E. coli Assembly Quality

Incorrect contigs: Align at < 95% identity or < 95% of their length

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Contigs ≥ 100bp</th>
<th>N50 (bp)</th>
<th>Incorrect contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrail PE</td>
<td>300</td>
<td>54,807</td>
<td>4</td>
</tr>
<tr>
<td>Contrail SE</td>
<td>529</td>
<td>20,062</td>
<td>0</td>
</tr>
<tr>
<td>SOAPdenovo PE</td>
<td>182</td>
<td>89,000</td>
<td>5</td>
</tr>
<tr>
<td>ABySS PE</td>
<td>233</td>
<td>45,362</td>
<td>13</td>
</tr>
<tr>
<td>Velvet PE</td>
<td>286</td>
<td>54,459</td>
<td>9</td>
</tr>
<tr>
<td>EULER-SR PE</td>
<td>216</td>
<td>57,497</td>
<td>26</td>
</tr>
<tr>
<td>SSAKE SE</td>
<td>931</td>
<td>11,450</td>
<td>38</td>
</tr>
<tr>
<td>Edena SE</td>
<td>680</td>
<td>16,430</td>
<td>6</td>
</tr>
</tbody>
</table>

It was the best of times, it was the age of

...
**Contrail**
http://contrail-bio.sourceforge.net

De novo assembly of the Human Genome
- *Genome*: African male NA18507 (SRA000271, Bentley *et al.*, 2008)
- *Input*: 3.5B 36bp reads, 210bp insert (~40x coverage)

<table>
<thead>
<tr>
<th>Stage</th>
<th>N</th>
<th>Max</th>
<th>N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>~7 B</td>
<td>27 bp</td>
<td>27 bp</td>
</tr>
<tr>
<td>Compressed</td>
<td>&gt;1 B</td>
<td>303 bp</td>
<td>&lt; 100 bp</td>
</tr>
<tr>
<td>Error Correction</td>
<td>4.2 M</td>
<td>20,594 bp</td>
<td>995 bp</td>
</tr>
<tr>
<td>Resolve Repeats</td>
<td>4.1 M</td>
<td>20,594 bp</td>
<td>1,050 bp</td>
</tr>
<tr>
<td>Cloud Surfing</td>
<td>3.3 M</td>
<td>20,594 bp</td>
<td>1,427 bp*</td>
</tr>
</tbody>
</table>

**Assembly of Large Genomes with Cloud Computing.**
Variations and de Bruijn Graphs

Searching for de novo mutations in the families of 3000 autistic children.

- Assemble together reads from mom, dad, affected & unaffected children
- Look for sequence paths unique to affected child

MRC1LI
Summary

• Staying afloat in the data deluge means computing in parallel
  – Hadoop + Cloud computing is an attractive platform for large scale sequence analysis and computation

• Significant obstacles ahead
  – Price
  – Transfer time
  – Privacy / security requirements
  – Time and expertise required for development
    (Schatz et al., Nature Biotechnology, 2010)

• Emerging technologies are a great start, but we need continued research
  – A word of caution: new technologies are new
Acknowledgements

CSHL
Mike Wigler
Ivan Iossifov
Mike Ronemus
Jude Kendall
Dan Levy

Zach Lippman
Dick McCombie
Doreen Ware

Univ. of Maryland
Steven Salzberg
Mihai Pop
Carl Kingsford
Art Delcher
Jimmy Lin
Dan Sommer
David Kelley

JHU
Ben Langmead

Mitch Bekritsky
Matt Titmus
Thank You!

http://schatzlab.cshl.edu

@mike_schatz
Often an astronomical number of possible assemblies

- Value computed by application of the BEST theorem (Hutchinson, 1975)

\[
\mathcal{W}(G, t) = (\det L) \left\{ \prod_{u \in V} (r_u - 1)! \right\} \left\{ \prod_{(u,v) \in E} a_{uv}! \right\}^{-1}
\]

L = n x n matrix with \(r_u - a_{uu}\) along the diagonal and \(-a_{uv}\) in entry uv

\(r_u = d^+(u) + l\) if \(u=t\), or \(d^+(u)\) otherwise

\(a_{uv}\) = multiplicity of edge from \(u\) to \(v\)

**Assembly Complexity of Prokaryotic Genomes using Short Reads.**
Why HPC?

- Moore’s Law is valid in 2010
  - But CPU speed is flat
  - Vendors adopting parallel solutions instead

- Parallel Environments
  - Many cores, including GPUs
  - Many computers
  - Many disks

- Why parallel
  - Need results faster
  - Doesn’t fit on one machine

The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software
Cloud Computing and the DNA Data Race.