GPGPU and Cloud Computing for DNA Sequence Analysis

Michael C. Schatz

Nov. 19, 2009
Doctoral Showcase, SC09

The Evolution of DNA Sequencing

Year	Genome	Technology	Cost
2001	Venter *et al.*	Sanger (ABI)	$300,000,000
2007	Levy *et al.*	Sanger (ABI)	$10,000,000
2008	Wheeler *et al.*	Roche (454)	$2,000,000
2008	Ley *et al.*	Illumina	$1,000,000
2008	Bentley *et al.*	Illumina	$250,000
2009	Pushkarev *et al.*	Helicos	$48,000
2009	Drmanac *et al.*	Complete Genomics	$4,400

(Pushkarev *et al.*, 2009)

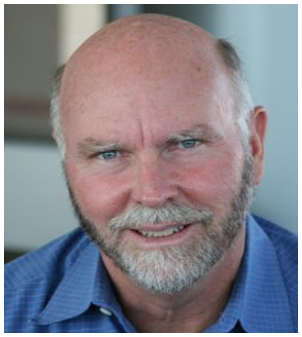
1000 Genomes

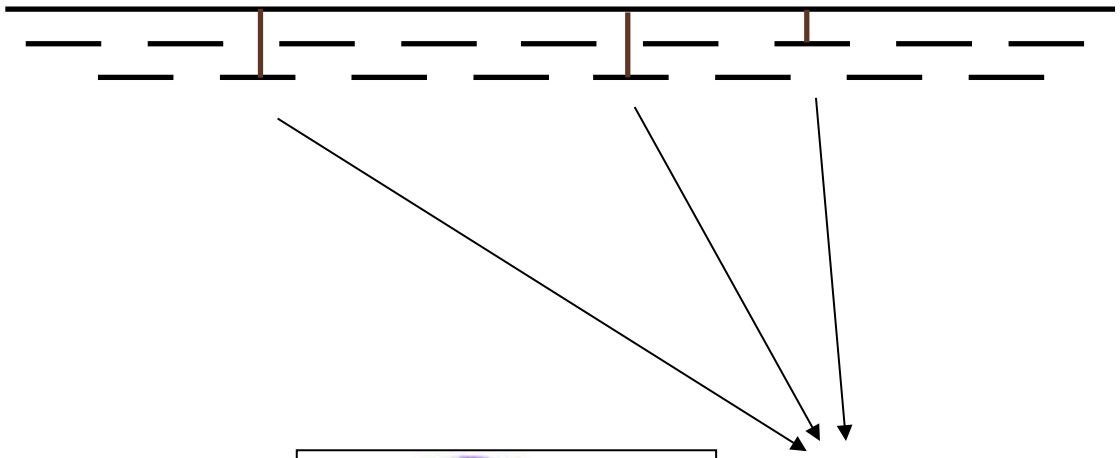
Global Ocean Survey

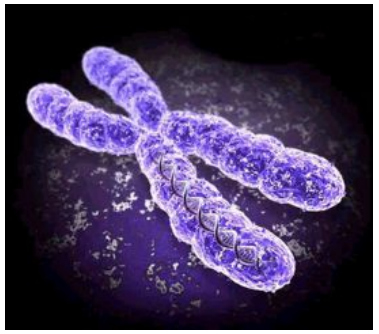
Human Microbiome

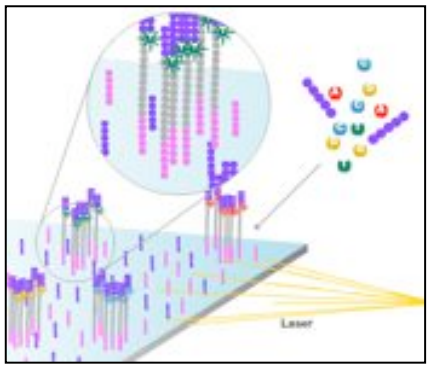
Personal Genomics

What's in your genome?









Arthritis [C at rs1980422]

Cancer [A at rs620861]

Bad Driver [T at rs6265]

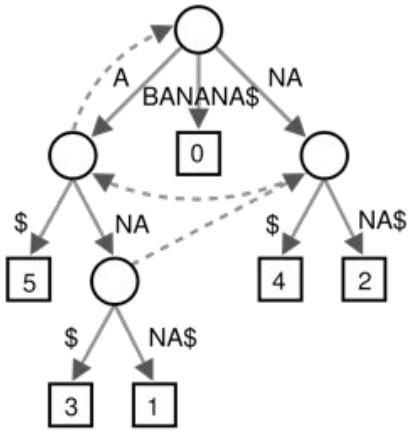
Indexing & Throughput

- Desperate need for scalable solutions
  - Individual Genome: 3.3 Billion 35bp, 106.5 GB (Wang *et al.*, 2008)
  - Read Mapping required >1,000 CPU hours / genome

MUMmerGPU

GPGPU

Suffix Tree

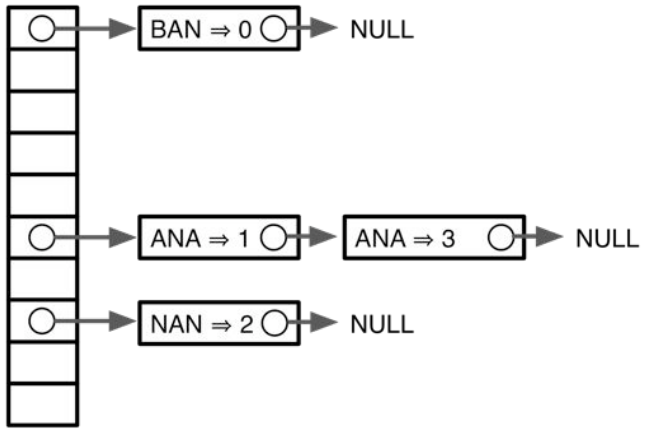


(Schatz *et al.*, 2008)

CloudBurst

Hadoop / MapReduce

Distributed Inverted Index



(Schatz, 2009)

Crossbow

Hadoop / MapReduce

Burrows-Wheeler

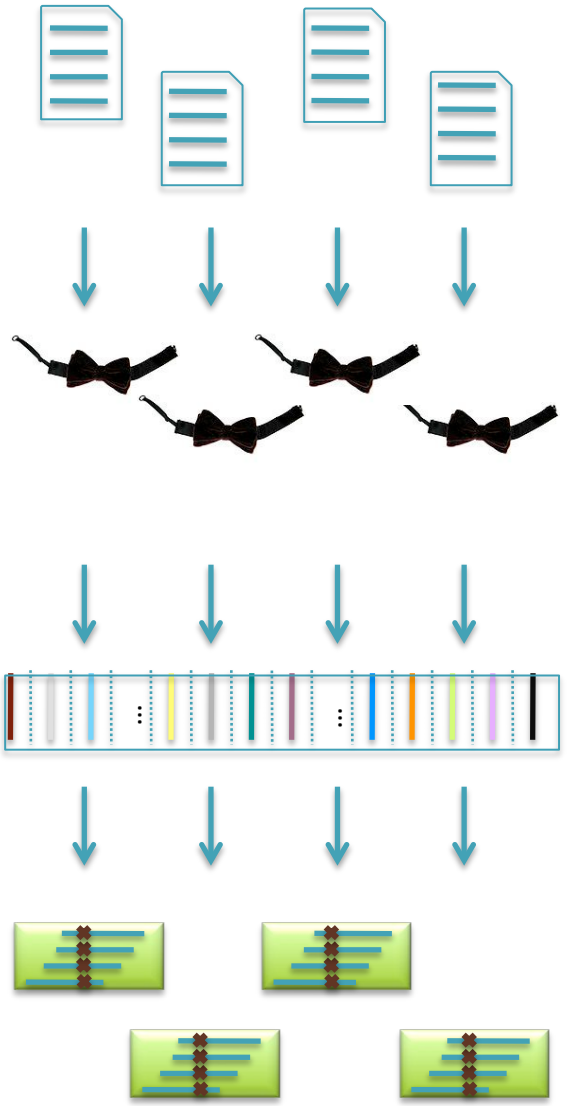
$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

(In Press)

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)
  - Align reads to BWT index of reference
  - Emit (chromosome region, alignment)
- Shuffle: Hadoop
  - Group and sort alignments by region
- Reduce: SOAPsnp (Li *et al.*, 2009)
  - Scan alignments for divergent columns
  - Output all SNPs



Crossbow at Amazon EC2

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

	Asian Individual Genome		
Data Loading	3.3 B reads	106.5 GB	$10.65
Data Transfer	1h :15m	20+1 Medium	$3.40
Setup	0h : 15m	40+1 X-Large	$13.94
Mapping	1h : 30m	40+1 X-Large	$41.82
Variant Calling	1h : 00m	40+1 X-Large	$27.88
End-to-end	4h : 00m		$97.69

Raw sequences to SNPs for ~$100 in an afternoon.
Accuracy validated at better than 99%

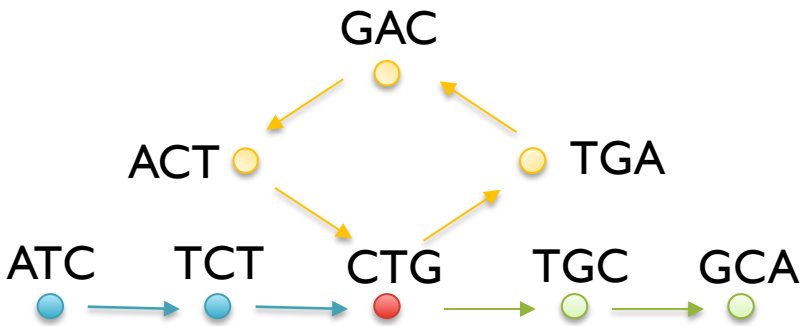
Searching for SNPs with Cloud Computing.
Langmead, B, Schatz, MC, Lin, J, Pop, M, Salzberg, SL (2009) *In Press*.

Genomics without a reference

Reads

ACTG
ATCT
CTGA
CTGC
GACT
TCTG
TGAC
TGCA

de Bruijn Graph



Genome Sequence

ATCTGACTGCA

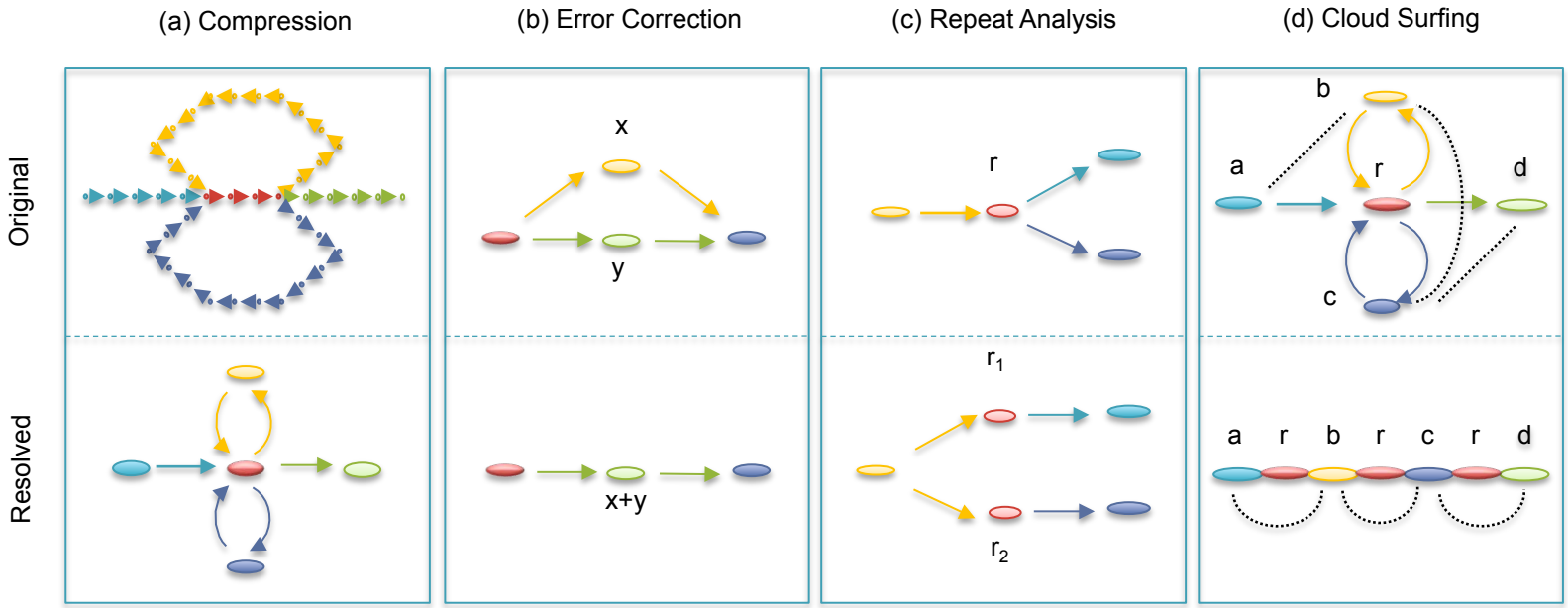
- Graph assembly modeled as finding an Eulerian tour through the de Bruijn graph
  - Human genome: ~3B nodes, ~10B edges
- The new short read assemblers require tremendous computation
  - Velvet (Zerbino & Birney, 2008) on human > 2 TB of RAM
  - ABySS (Simpson *et al.*, 2009) on human ~4 days on 168 cores

Contrail

http://contrail-bio.sourceforge.net

Scalable Genome Assembly with MapReduce

- *Parallel Randomized List Ranking*: merge non-branching nodes
- *Parallel Network Motif Finding*: recognize graph topology
- *Parallel Frontier Search*: breadth-first-search of neighborhood



Assembly of Large Genomes with Cloud Computing.
Schatz, MC, Sommer, D, Pop, M, *et al. In Preparation.*

Genomics across the Tree of Life

Genomes

- *N. ceranae* (Cornman *et al.*, 2009)
- *B. taurus* (Zimin *et al.*, 2009)
- *G. indiensis* (Desjardins *et al.*, 2009)
- *C. papaya* (Ming *et al.*, 2008)
- *C. papaya* (Suzuki *et al.*, 2008)
- *X. oryzae* (Salzberg *et al.*, 2008)
- *T. vaginalis* (Carlton *et al.*, 2007)
- Drosophila (Drosophila 12 genomes consortium, 2007)
- *A. aegypti* (Nene *et al.*, 2007)
- *B. malayi* (Ghedin *et al.*, 2007)
- *G. indiensis* (Desjardins *et al.*, 2007)
- Campylobacter (Fouts *et al.*, 2005)

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

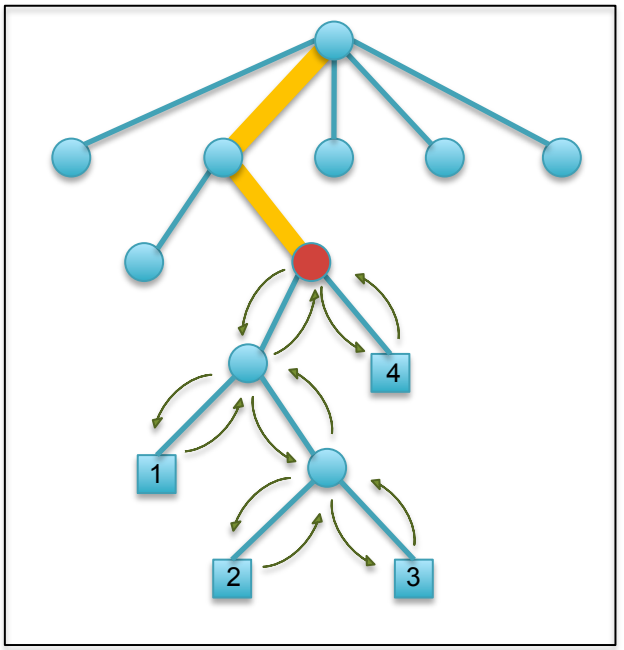
Thank You!

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

MUMmerGPU

http://mummergpu.sourceforge.net

- Index reference using a suffix tree
  - Each suffix represented by path from root
  - Reorder tree along space filling curve
- Map many reads simultaneously on GPU
  - Find matches by walking the tree
  - Find coordinates with depth first search
- Performance on nVidia GTX 8800
  - Match kernel was ~10x faster than CPU
  - Search kernel was ~4x faster than CPU
  - End-to-end runtime ~4x faster than CPU

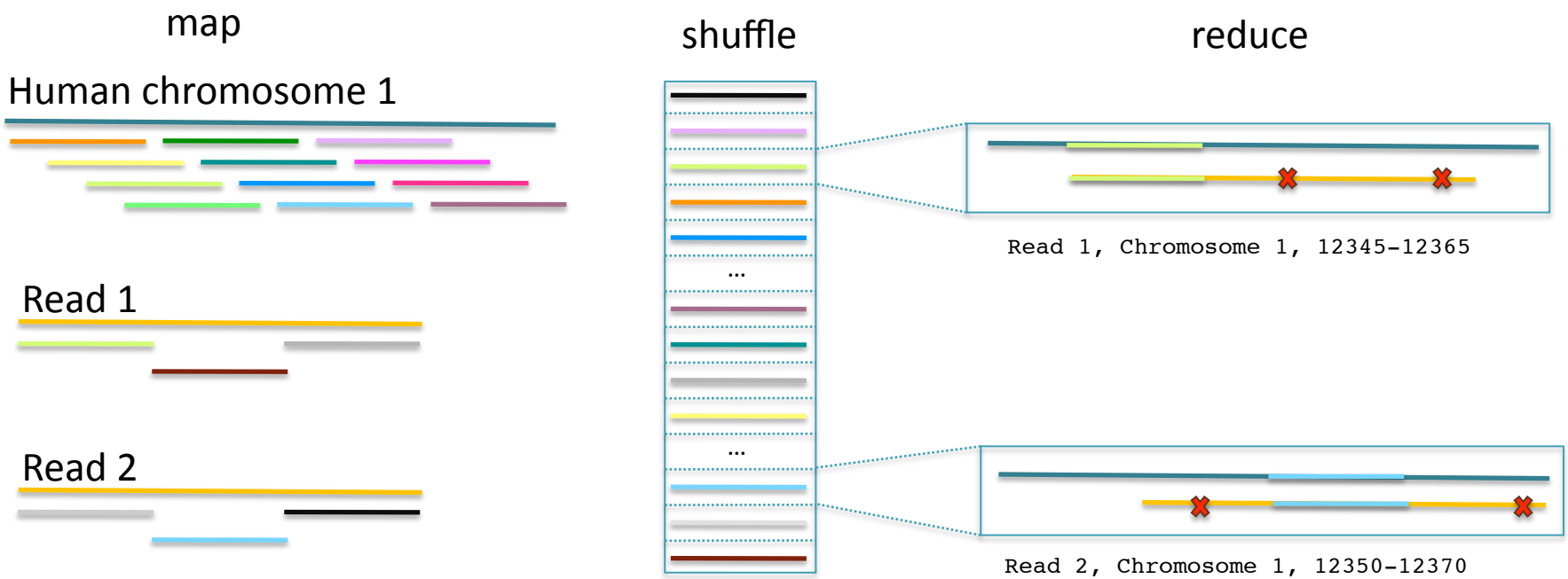


Optimizing data intensive GPGPU computations for DNA sequence alignment.
Trapnell C, Schatz MC. (2009) *Parallel Computing.* 35(8-9):429-440.

CloudBurst

http://cloudburst-bio.sourceforge.net

- Leverage Hadoop to build a distributed inverted index of k-mers and find end-to-end alignments
- 100x speedup over RMAP with 96 cores at Amazon EC2



CloudBurst: Highly Sensitive Read Mapping with MapReduce.
Schatz MC (2009) *Bioinformatics*. 25:1363-1369

Grand Challenge of Biology

"NextGen sequencing has completely outrun the ability of good bioinformatics people to keep up with the data and use it well… We need a MASSIVE effort in the development of tools for "normal" biologists to make better use of massive sequence databases."

Jonathan Eisen – JGI Users Meeting – 3/28/09

- Computational Biology
  - Make the analysis of large genomes accessible to individual researchers
- HPC
  - Research parallel algorithms for MapReduce and multicore systems