

Scalable Solutions for DNA Sequence Analysis

Michael Schatz

Jan 29, 2010

NHGRI/UMD Joint Sequencing Meeting



The Evolution of DNA Sequencing

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

(Pushkarev *et al.*, 2009)

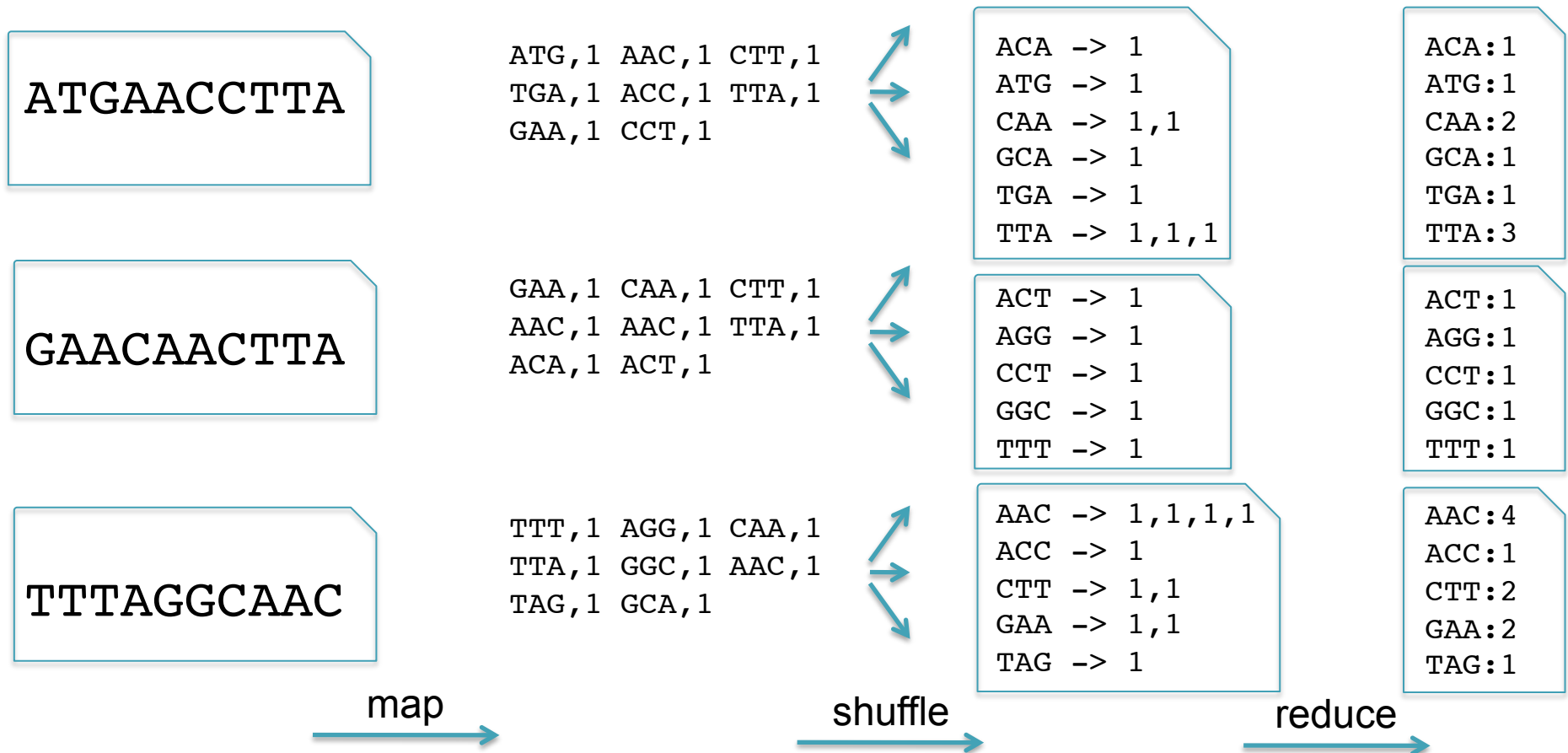


Critical Computational Challenges: Alignment and Assembly of Huge Datasets

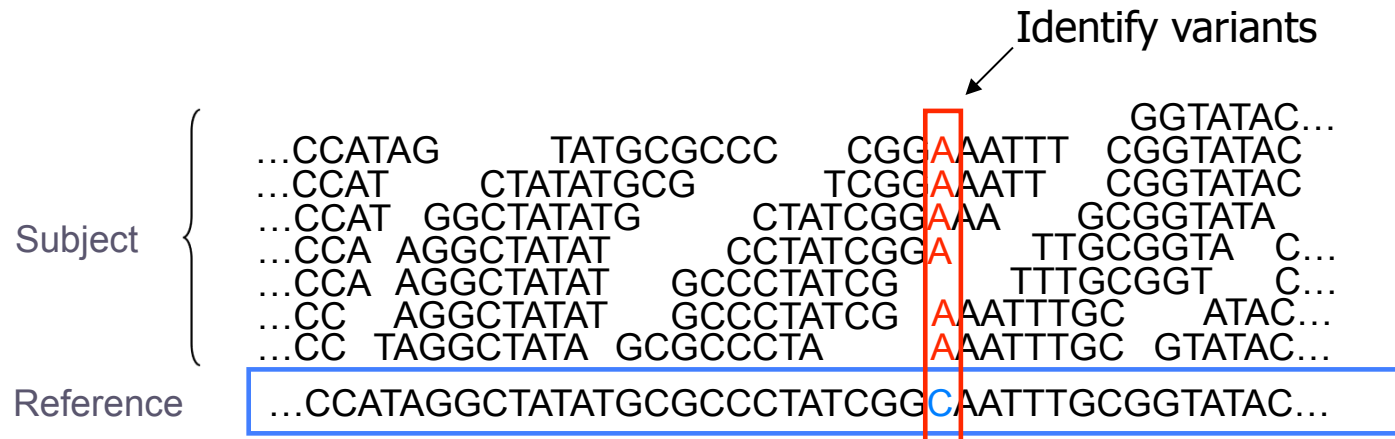
Hadoop MapReduce

- 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

Map, Shuffle & Reduce
All Run in Parallel



Short Read Mapping with MapReduce



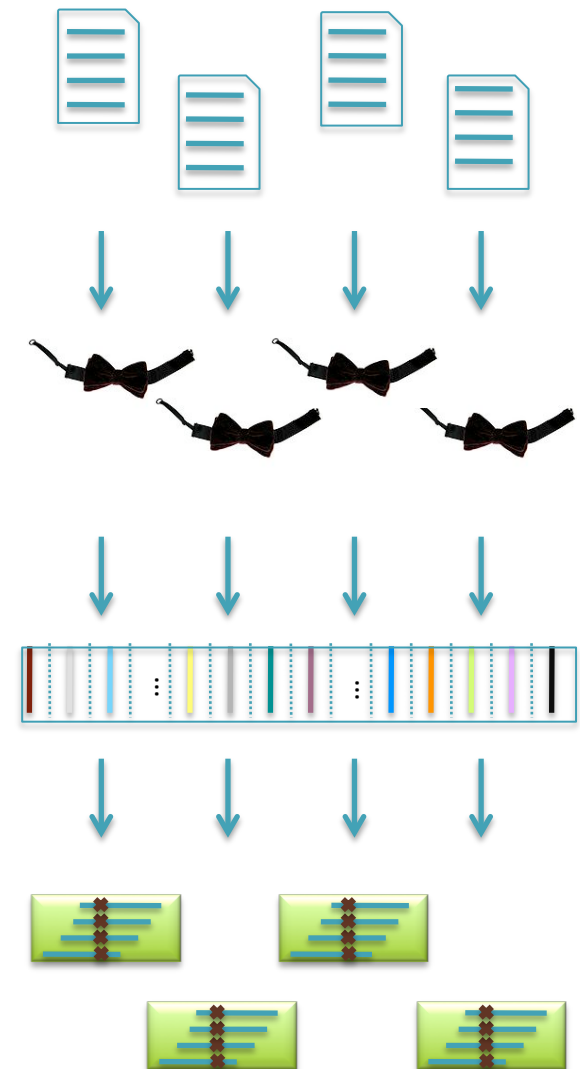
- Given a reference and many subject reads, report one or more “good” end-to-end alignments per alignable read
 - Maps the read to where it originated
- Mapping of a whole human requires ~1,000 CPU hours
 - Alignments are “embarrassingly parallel” by read
 - Variant detection is parallel by chromosome 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>

	Asian Individual Genome		
Data Loading	3.3 B reads	106.5 GB	\$10.65
Data Transfer	1h :15m	40 CPUs	\$3.40
Setup	0h : 15m	320 CPUs	\$13.94
Alignment	1h : 30m	320 CPUs	\$41.82
Variant Calling	1h : 00m	320 CPUs	\$27.88
End-to-end	4h : 00m		\$97.69

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

Searching for SNPs with Cloud Computing.

Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology*.

Contrail

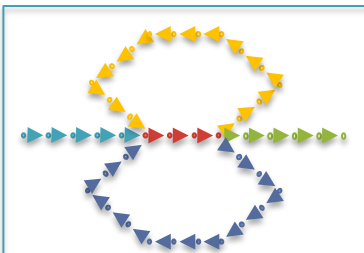
<http://contrail-bio.sourceforge.net>



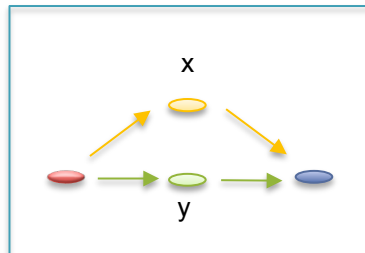
Genome Assembly with MapReduce

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

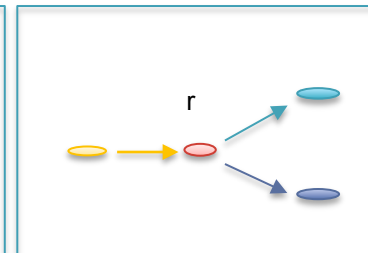
(a) Compression



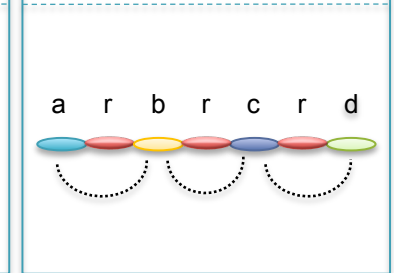
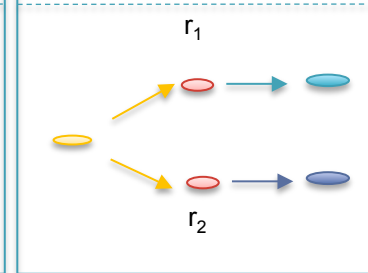
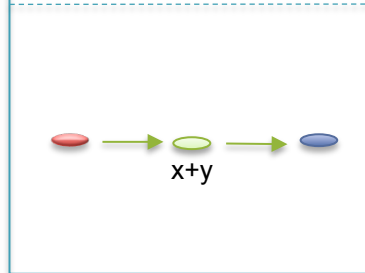
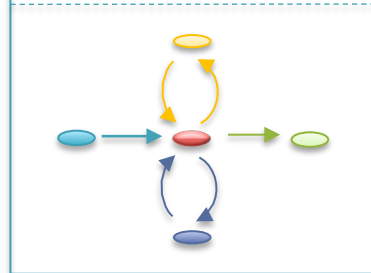
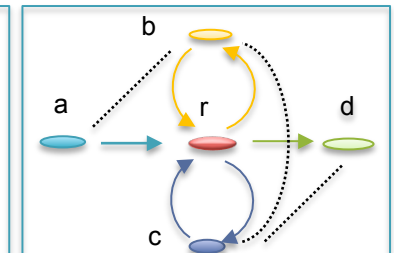
(b) Bubble Popping



(c) Repeat Analysis

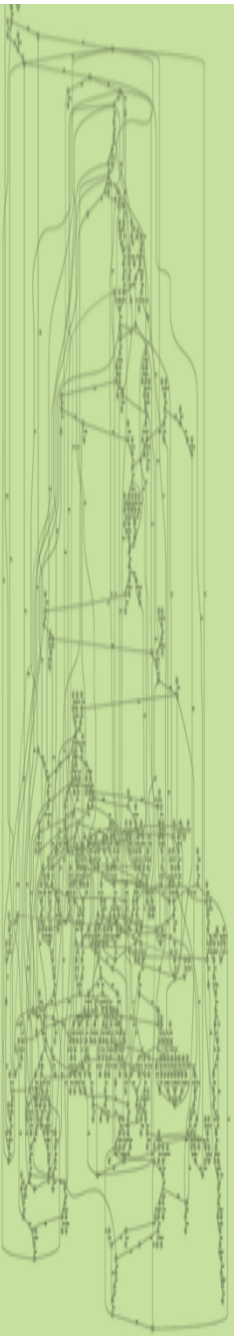


(d) Cloud Surfing



Assembly of Large Genomes with Cloud Computing.

Schatz MC, Sommer D, Pop M, *et al.* *In Preparation.*



(Chaisson, 2009)

Acknowledgements

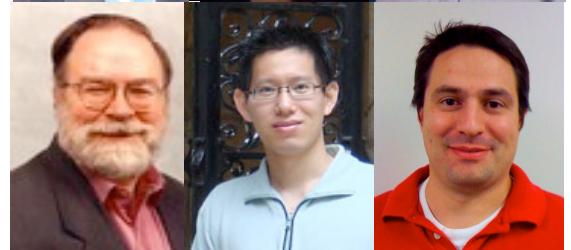
Advisor

Steven Salzberg



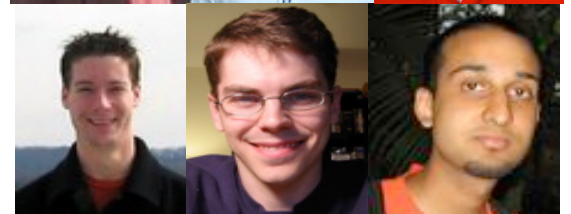
UMD Faculty

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.cbcu.umd.edu/~mschatz>