Scalable Solutions for DNA Sequence Analysis

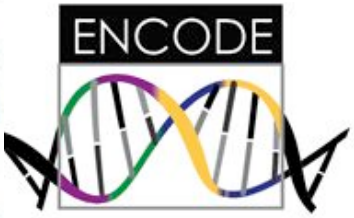
Michael Schatz

Dec 4, 2009
JHU/UMD Joint Sequencing Meeting

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)



Critical Computational Challenges: Alignment and Assembly of Huge Datasets

Research Highlights

Alignment

Crossbow

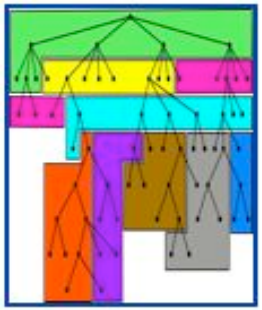
Searching for SNPs with Cloud Computing

(Langmead, Schatz, Lin, Pop, Salzberg, 2009)

CloudBurst

Highly Sensitive Read Mapping with MapReduce

(Schatz, 2009)



MUMmerGPU

High Throughput Sequence Alignment Using GPUs

(Schatz, Trapnell, Varshney, Delcher, 2007)
(Trapnell, Schatz, 2009)

Assembly

Hawkeye

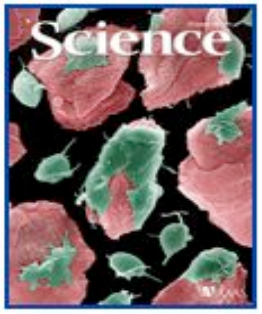
Assembly Visualization & Analytics

(Schatz, Phillippy, Shneiderman, Salzberg, 2007)

AutoEditor & AutoJoiner

Improving Genome Assemblies without Resequencing

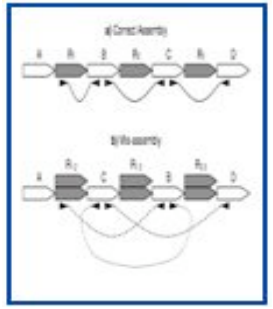
(Gajer, Schatz, Salzberg, 2004)
(Carlton *et al.*, 2007)



Assembly Forensics

Finding the Elusive Mis-assembly

(Phillippy, Schatz, Pop, 2008)

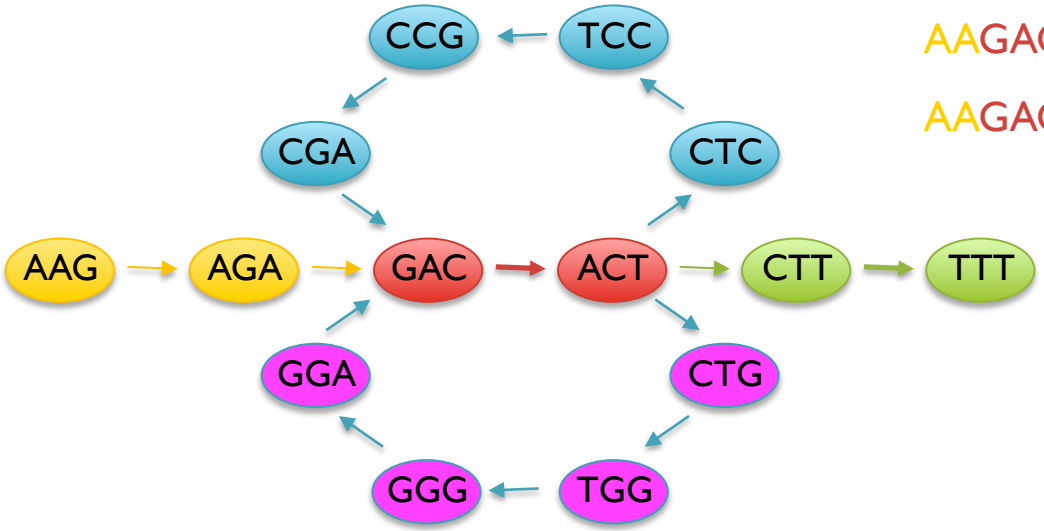


Short Read Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…

de Bruijn Graph



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) on human > 2 TB of RAM
  - ABySS (Simpson *et al.*, 2009) on human ~4 days on 168 cores

Hadoop MapReduce

- MapReduce is the parallel distributed framework invented by Google for large data computations.
  - Data and computations are spread over thousands of computers, processing petabytes of data each day (Dean and Ghemawat, 2004)
  - Indexing the Internet, PageRank, Machine Learning, etc…
  - Hadoop is the leading open source implementation

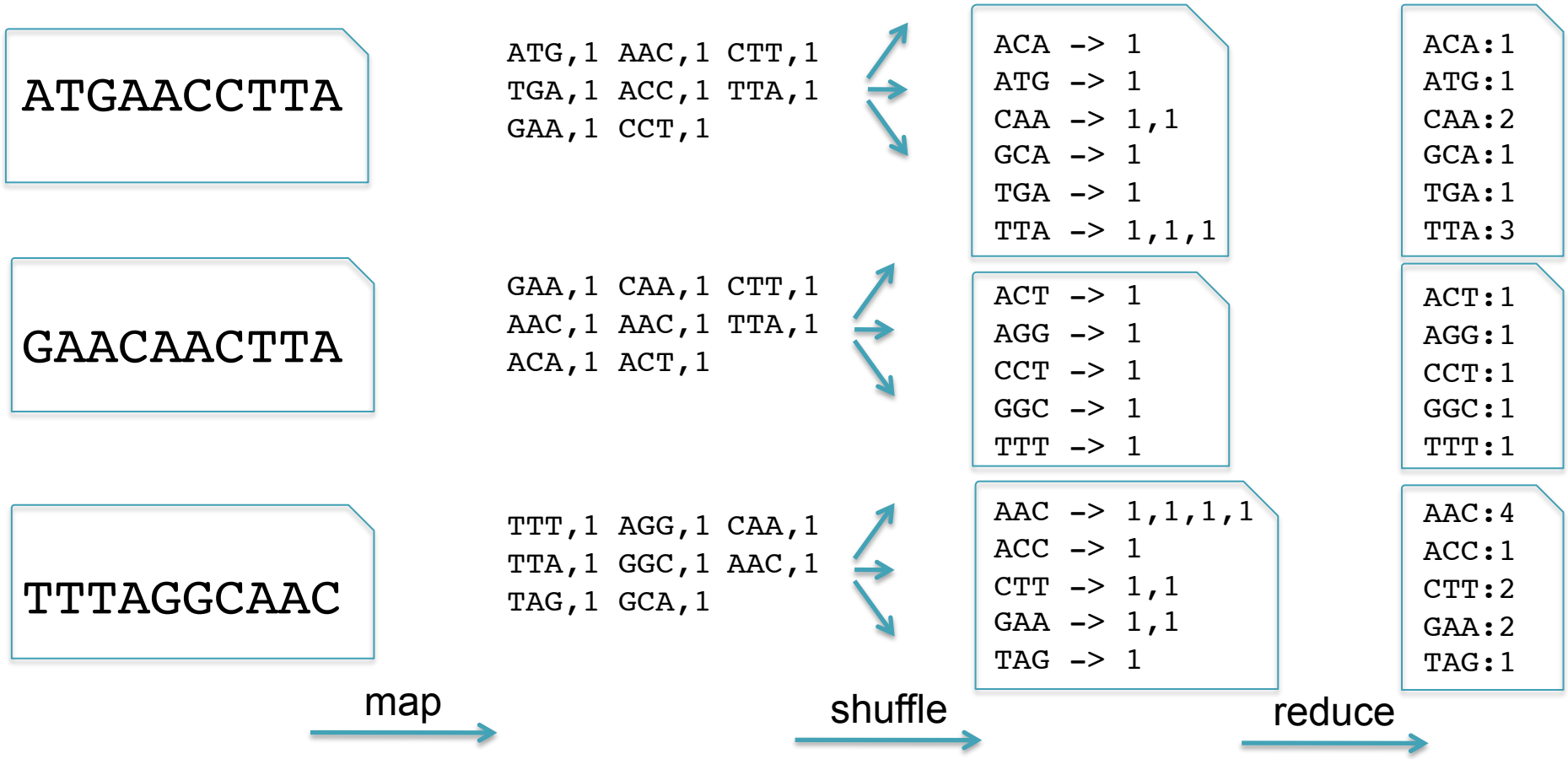
- Benefits
  - Scalable, Efficient, Reliable
  - Easy to Program
  - Runs on commodity computers

- Challenges
  - Redesigning / Retooling applications
    - Not SunGrid, Not MPI
    - Everything in MapReduce

K-mer Counting with MapReduce

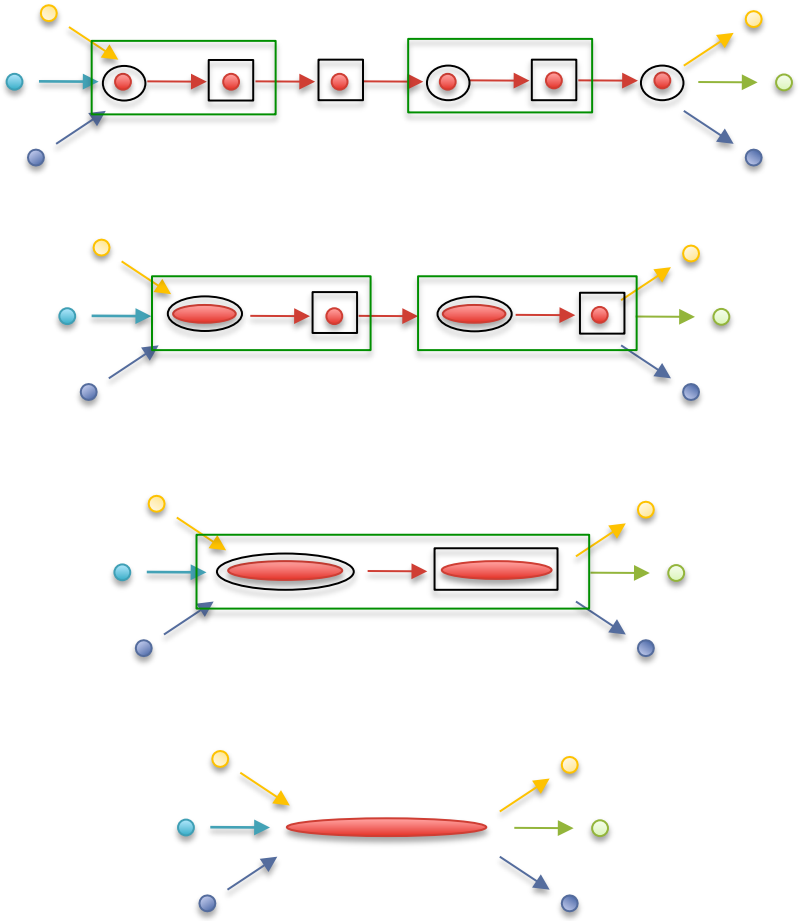
- Application developers focus on 2 (+1 internal) functions
  - Map: input ➔ key, value pairs
  - Shuffle: Group together pairs with same key
  - Reduce: key, value-lists ➔ output

Map, Shuffle & Reduce
All Run in Parallel



Genome Assembly with MapReduce

- Challenges
  - Nodes stored on different computers
  - Node only knows immediate neighbors
- Randomized List Ranking
  - Randomly assign (H) / [T] to each compressible node
  - Compress (H) -> [T] links



Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

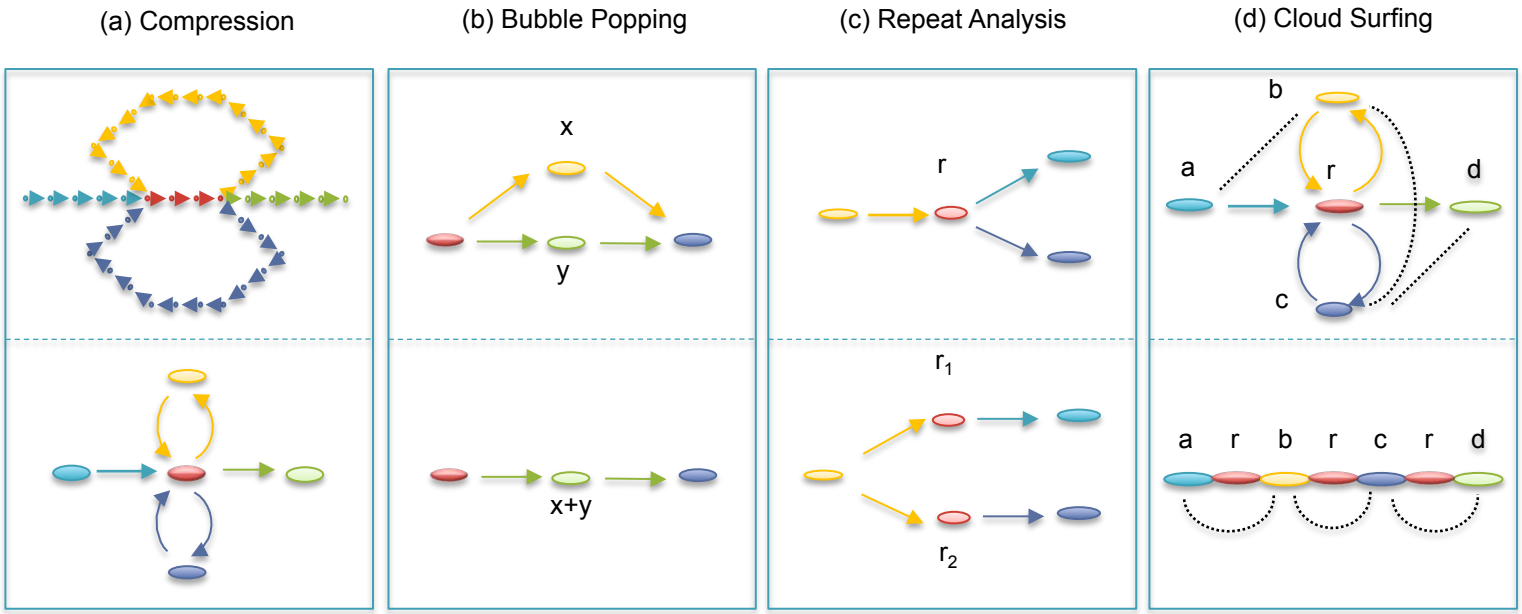
(Chaisson, 2009)

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



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

Summary

1. Managing the tidal wave of NextGen sequence data is a central challenge in biology

2. Hadoop is well suited towards scaling up biological computation

3. Cloud computing is an attractive platform to augment resources

4. Look for many cloud computing & MapReduce solutions this year

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