Commodity Computing in Genomics Research

Michael Schatz, Ben Langmead, Dan Sommer, Mihai Pop

Nov 16, 2009
Cloud Computing for Systems and Computational Biology Workshop
Supercomputing Conference '09

High Throughput Biology

1000 Genomes

Global Ocean Survey

Human Microbiome

- These studies require massive computation
  - Individual Human Genome: 3.3 Billion 35bp, 106 GB (Wang *et al.*, 2008)
  - Tens of thousands of CPU hours to analyze
- How are we going to store and analyze all that data?
  - If only there was a system for inexpensive parallel computing…

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)
  - Hadoop is the leading open source implementation

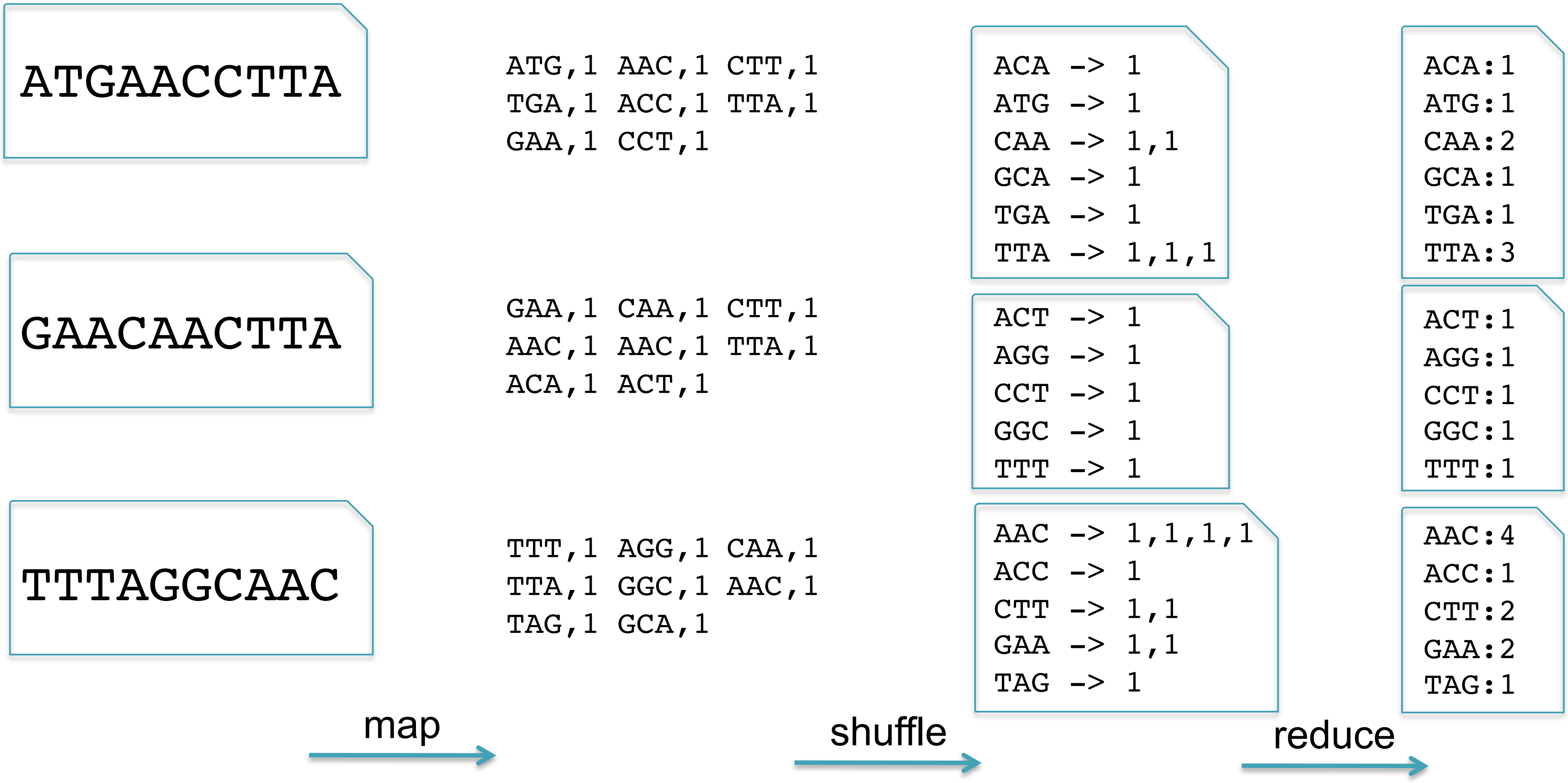
- Benefits
  - Scalable, Efficient, Reliable
  - Easy to Program
  - Runs on commodity computers

- 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



ATGAACCTTA

GAACAACTTA

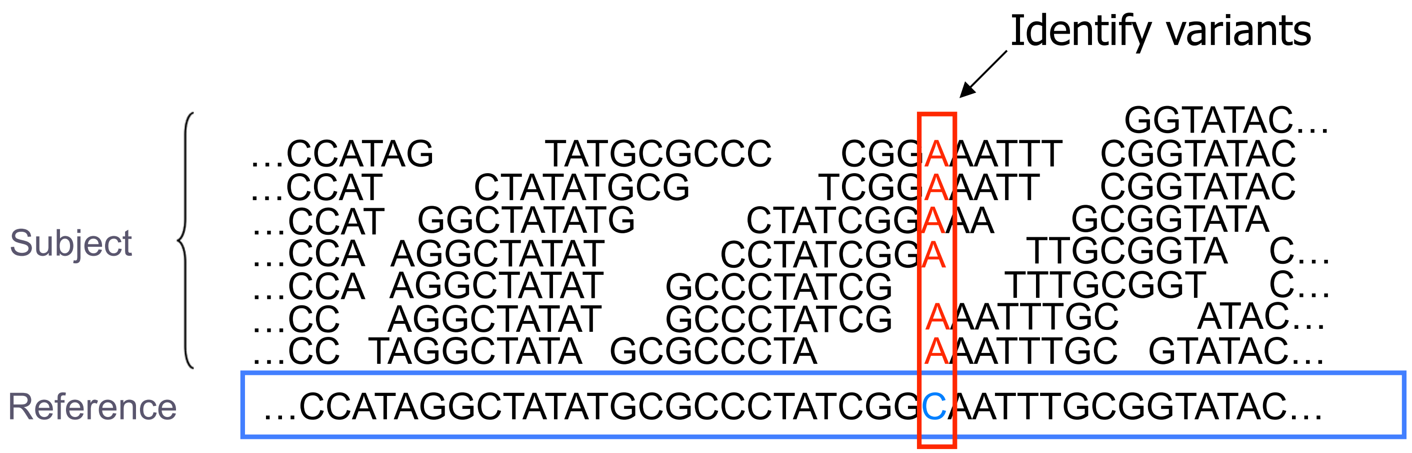
TTTAGGCAAC

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

GAA,1 CAA,1 CTT,1
AAC,1 AAC,1 TTA,1
ACA,1 ACT,1

TTT,1 AGG,1 CAA,1
TTA,1 GGC,1 AAC,1
TAG,1 GCA,1
```

```
ACA -> 1
ATG -> 1
CAA -> 1,1
GCA -> 1
TGA -> 1
TTA -> 1,1,1

ACT -> 1
AGG -> 1
CCT -> 1
GGC -> 1
TTT -> 1

AAC -> 1,1,1,1
ACC -> 1
CTT -> 1,1
GAA -> 1,1
TAG -> 1
```

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

ACT:1
AGG:1
CCT:1
GGC:1
TTT:1

AAC:4
ACC:1
CTT:2
GAA:2
TAG:1
```

Short Read Mapping with MapReduce



Identify variants

Subject

```
                                    GGTATAC...
...CCATAG    TATGCGCCC   CGGAATTT  CGGTATAC
...CCAT   CTATATGCG     TCGGAATT   CGGTATAC
...CCAT GGCTATATG    CTATCGGAAA   GCGGTATA
...CCA AGGCTATAT    CCTATCGGA   TTGCGGTA  C...
...CCA AGGCTATAT  GCCCTATCG    TTTGCGGT   C...
...CC  AGGCTATAT  GCCCTATCG AAATTTGC   ATAC...
...CC TAGGCTATA GCGCCCTA    AAATTTGC GTATAC...
```

Reference

```
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

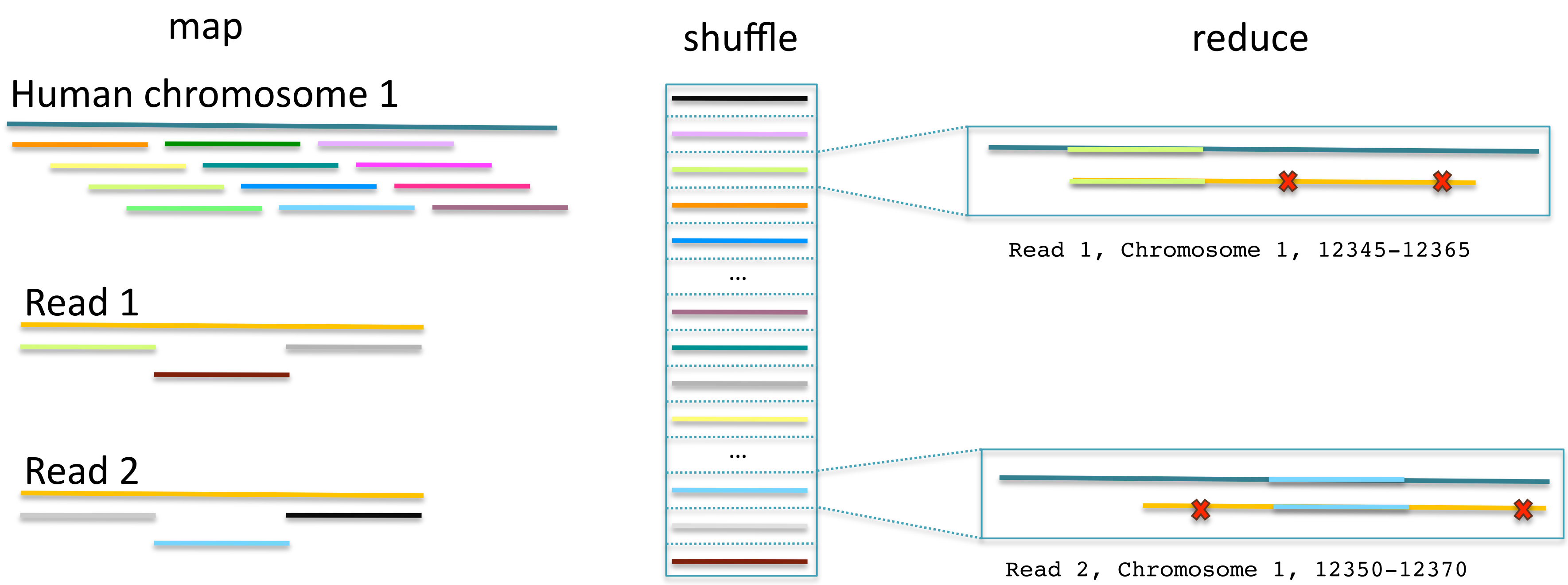
- Given a reference and many subject reads, report one or more “good” end-to-end alignments per alignable read
  - Maps the read to where it originated

- Mapping of a whole human requires ~1,000 CPU hours
  - Alignments are “embarassingly parallel” by read
  - Variant detection is parallel by chromosome region

CloudBurst

http://cloudburst-bio.sourceforge.net

- Build a distributed index of k-mers and find end-to-end alignments
- 100x speedup over RMAP (Smith *et al.*, 2008) with 96 cores in Amazon EC2


map
shuffle
reduce
Human chromosome 1
Read 1
Read 2
…
…
Read 1, Chromosome 1, 12345-12365
Read 2, Chromosome 1, 12350-12370

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

Bowtie

http://bowtie-bio.sourceforge.net

- Quality-aware search of Burrows-Wheeler Transform (BWT) to rapidly find the best alignment(s) for each read
  - 3GB BWT precomputed once, reused many times
  - easy to distribute, fits into RAM

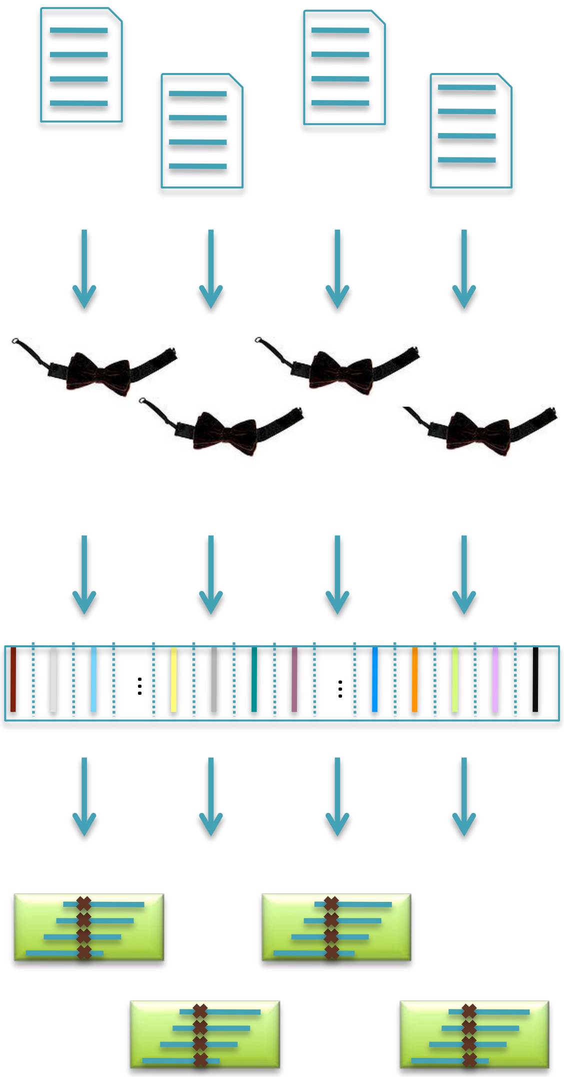
- Support for paired-end alignment, quality guarantees, uniqueness guarantees, etc…

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.
Langmead B, Trapnell C, Pop M, Salzberg SL (2009) *Genome Biology* 10:R25.

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)
  - Emit (chromome 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



Validation Results

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

	Chromosome 22			Chromosome X		
SNP Calling	**True sites**	**Sensitivity**	**Precision**	**True sites**	**Sensitivity**	**Precision**
All	46,586	99.0%	99.1%	102,219	99.0%	99.6%
only known	36,096	99.8%	99.9%	71,976	99.9%	99.9%
only novel	10,490	96.3%	96.3%	30,243	96.8%	98.8%
only homozygous	14,858	98.7%	99.9%	N/A	N/A	N/A
only heterozygous	31,728	99.2%	98.8%	N/A	N/A	N/A

- Simulate SNPs in the genome at expected rates
- Simulated 40x coverage paired-end 35bp reads with empirically derived errors, insert size distributions

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	20+1 Medium	$3.40
Setup	0h : 15m	40+1 X-Large	$13.94
Alignment	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

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) *In Press.*

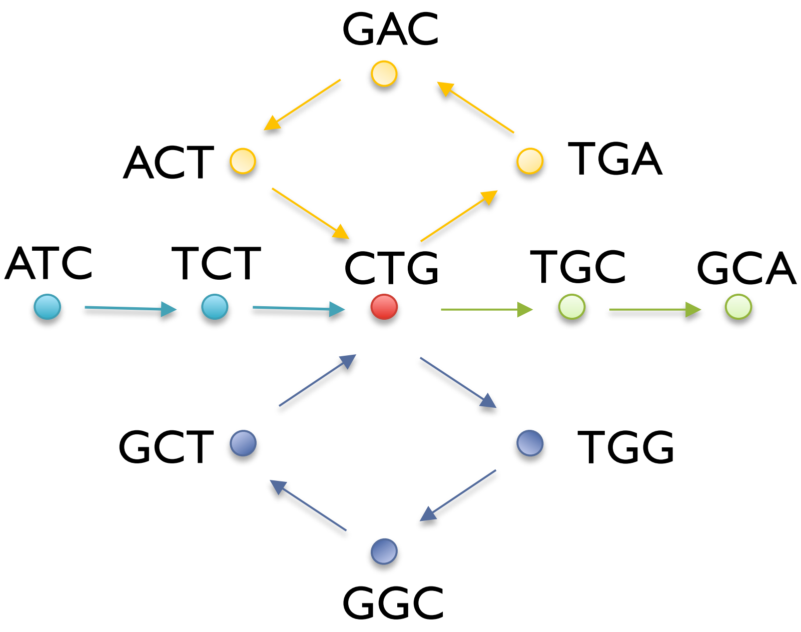
Genomics without a reference

- The new short read assemblers require tremendous computation
  - Velvet (Zerbino & Birney, 2008) on 2 Mbp *S. suis* requires > 2GB of RAM
  - ABySS (Simpson *et al.*, 2009) on human requires ~4 days on 168 cores

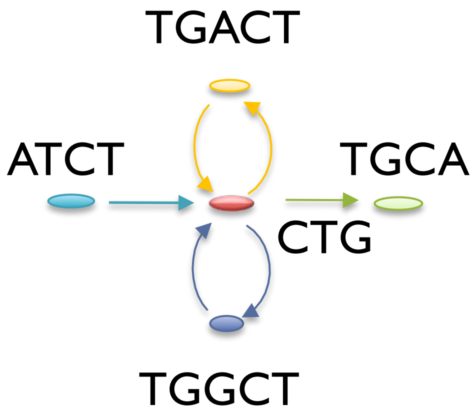
Reads

ACTG
ATCT
CTGA
CTGG
CTGC
GACT
GCTG
GGCT
TCTG
TGAC
TGCA
TGGC

de Bruijn Graph



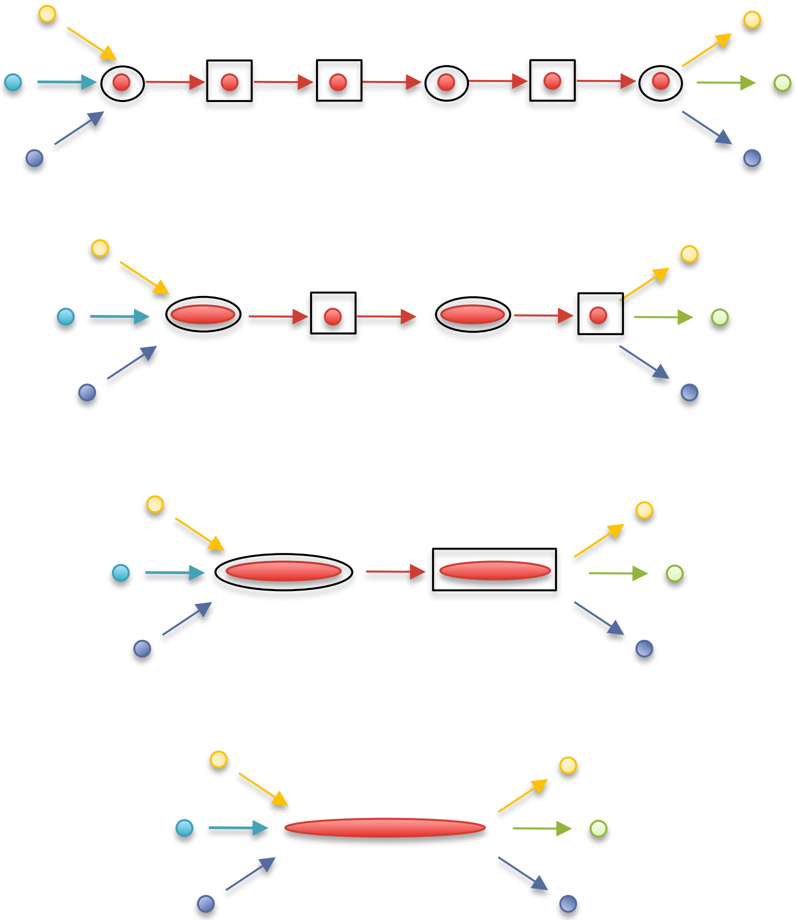
Compressed Graph



Genome Assembly with MapReduce

- Challenges
  - Nodes stored on different computers
  - Node only knows immediate neighbors

- Randomized List Ranking
  - Randomly assign Ⓗ/ $\boxed{T}$ to each compressible node
  - Compress Ⓗ-> $\boxed{T}$ links
  - E=O(log S) MapReduce cycles
    - *B. anthracis* 268,925 -> 19 cycles
    - Human: 37,172 ->16 cycles



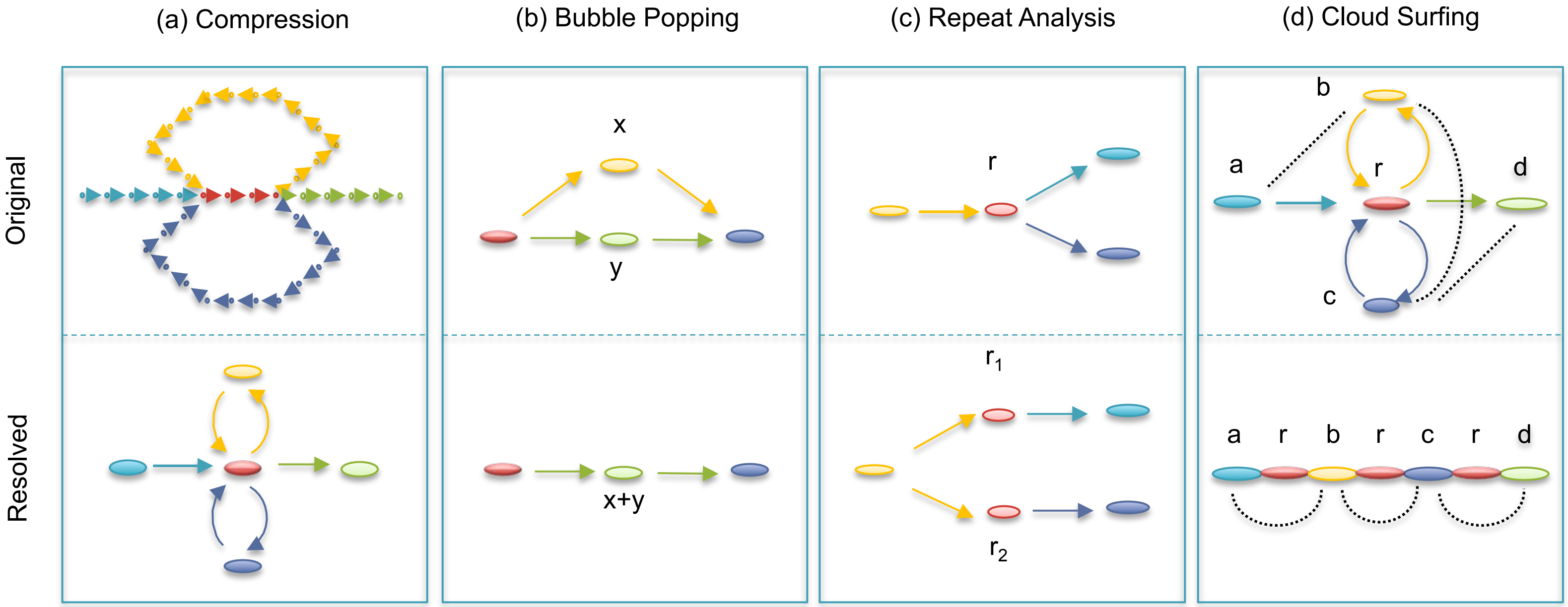
Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

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. Hadoop is well suited to big data biological computation
2. Hadoop Streaming for easy scaling of existing software
3. Cloud computing is an attractive platform to augment resources
4. Look for many cloud computing & MapReduce solutions this year

Acknowledgements

Ben Langmead

Mihai Pop

Dan Sommer

Jimmy Lin

Steven Salzberg

hadoop

CBCB
Center for Bioinformatics & Computational Biology

amazon web services™

Thank You!

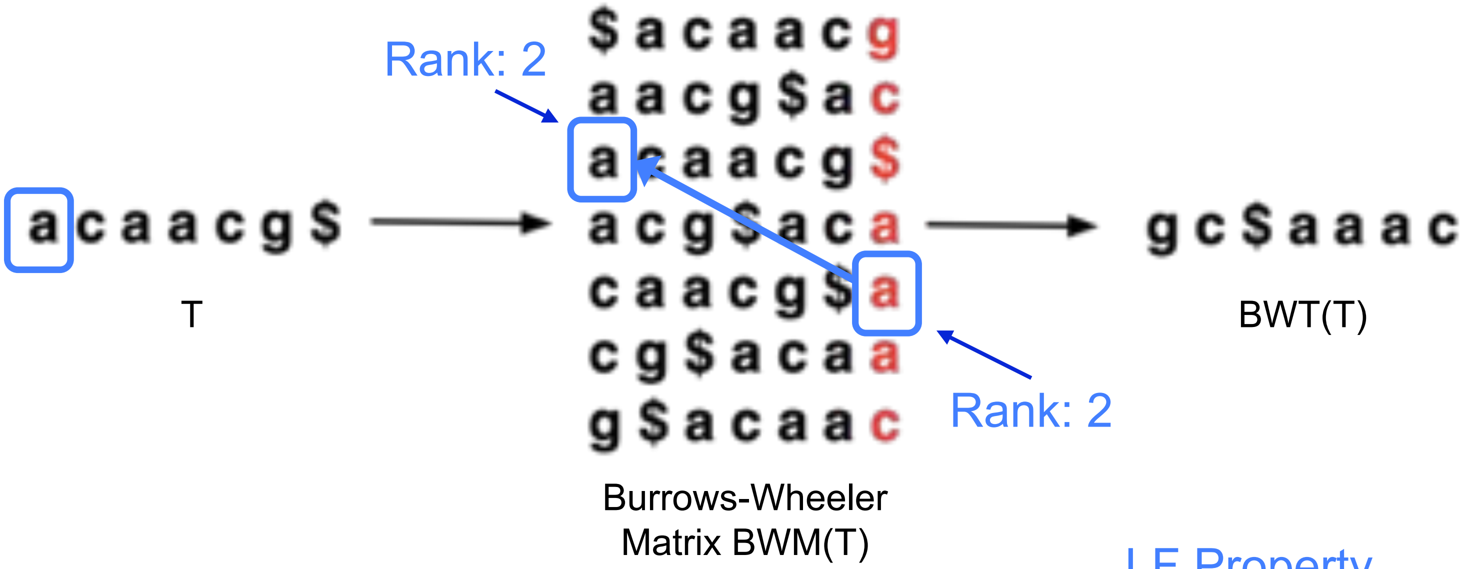
Crossbow Poster:
Tuesday, 5:15PM - 7:00PM
Oregon Ballroom Lobby

Doctoral Showcase:
Thursday, 3:45PM - 4:00PM
Room PB251

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

Burrows-Wheeler Transform

- Reversible permutation of the characters in a text



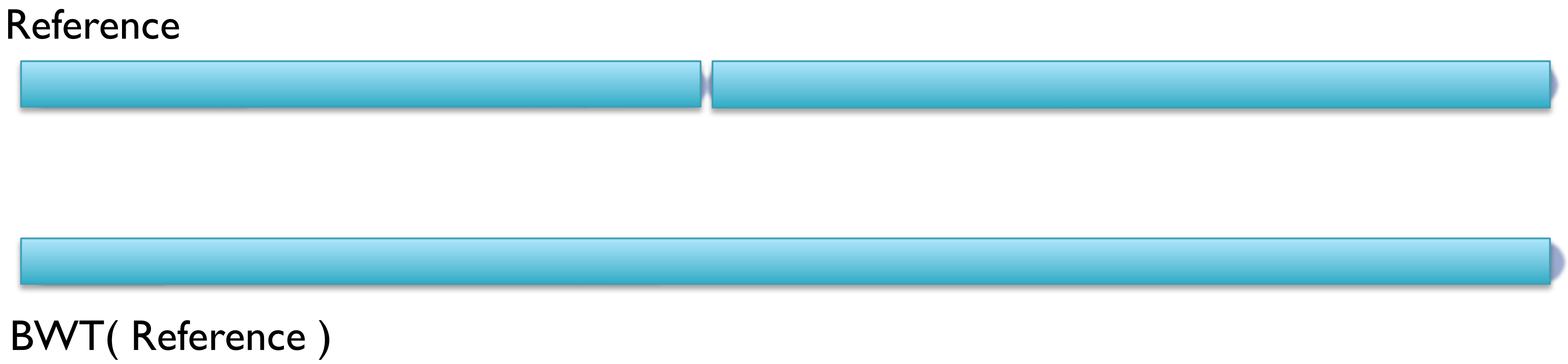
```
acaacg$

$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac

gc$aaac
```

LF Property implicitly encodes Suffix Array

- BWT(T) is the index for T

A block sorting lossless data compression algorithm.
Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation.* Technical Report 124

Bowtie algorithm

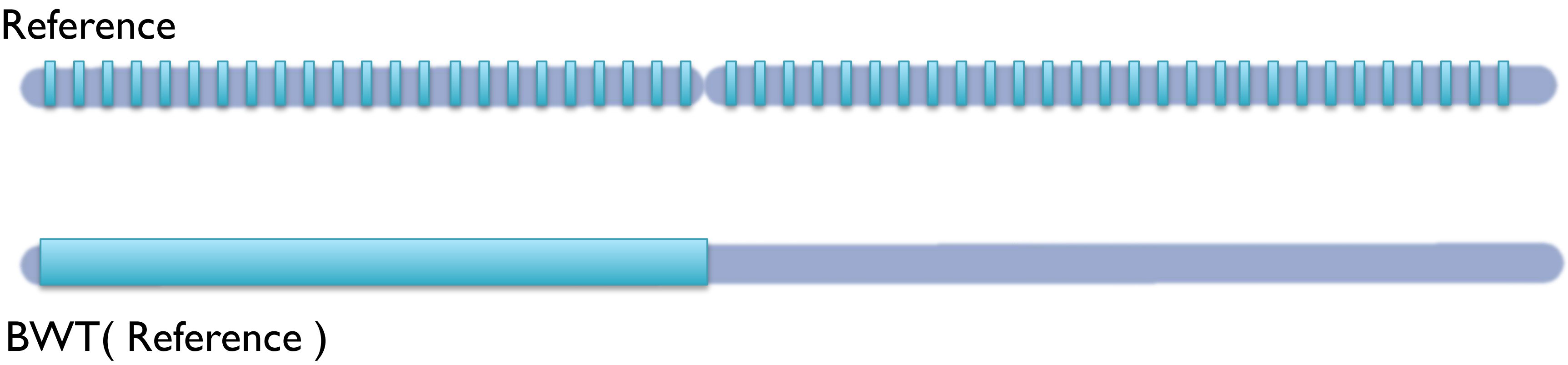

Reference

BWT(Reference)

Query:

AATGATACGGCGACCACCGAGATCTA

Bowtie algorithm



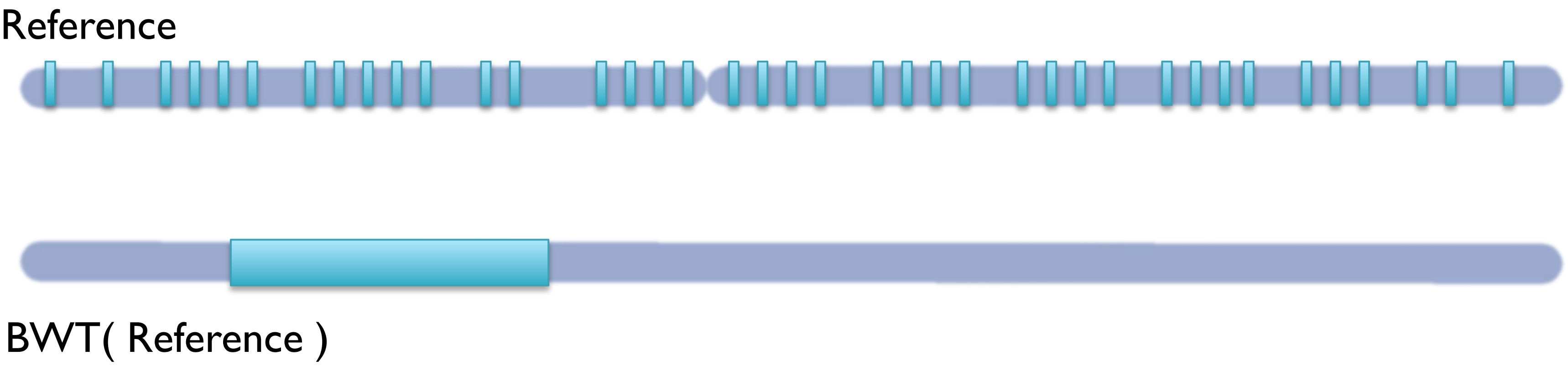
Reference

BWT(Reference)

Query:

AATGATACGGCGACCACCGAGATCTA

Bowtie algorithm



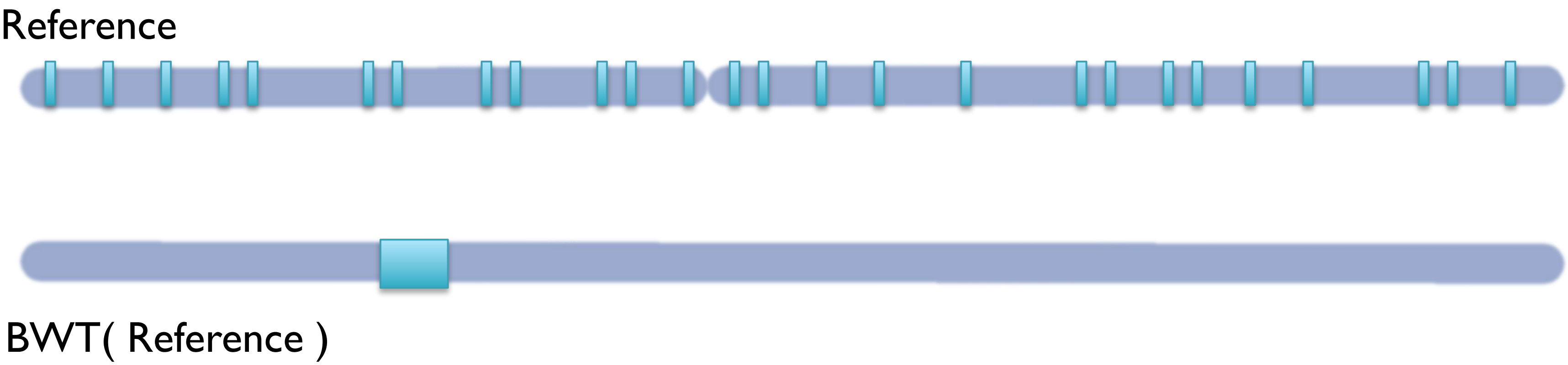
Reference

BWT(Reference)

Query:

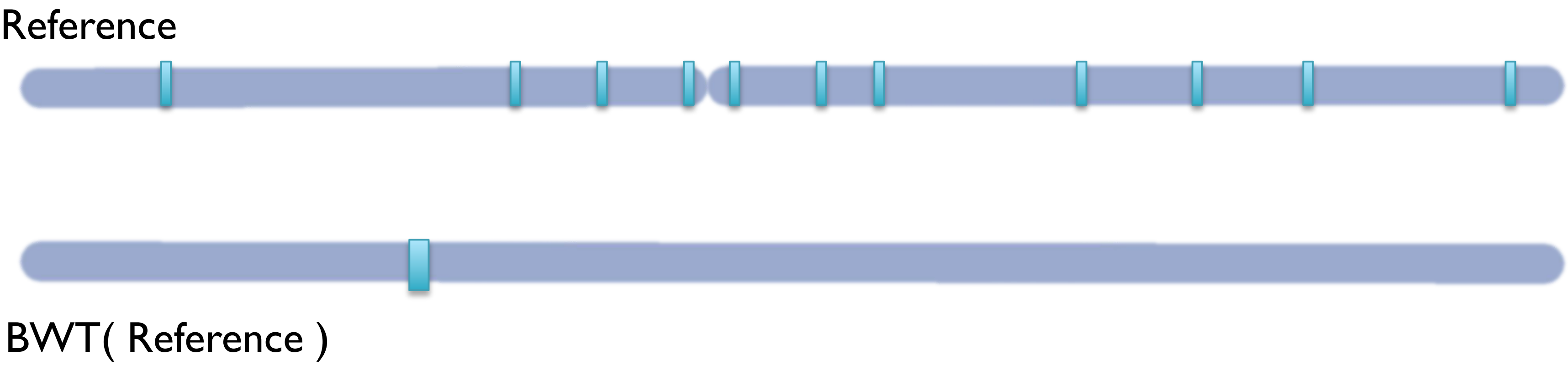
AATGATACGGCGACCACCGAGATCTA

Bowtie algorithm



Query:
AATGATACGGCGACCACCGAGATCTA

Bowtie algorithm



Query:
AATGATACGGCGACCACCGAGATCTA

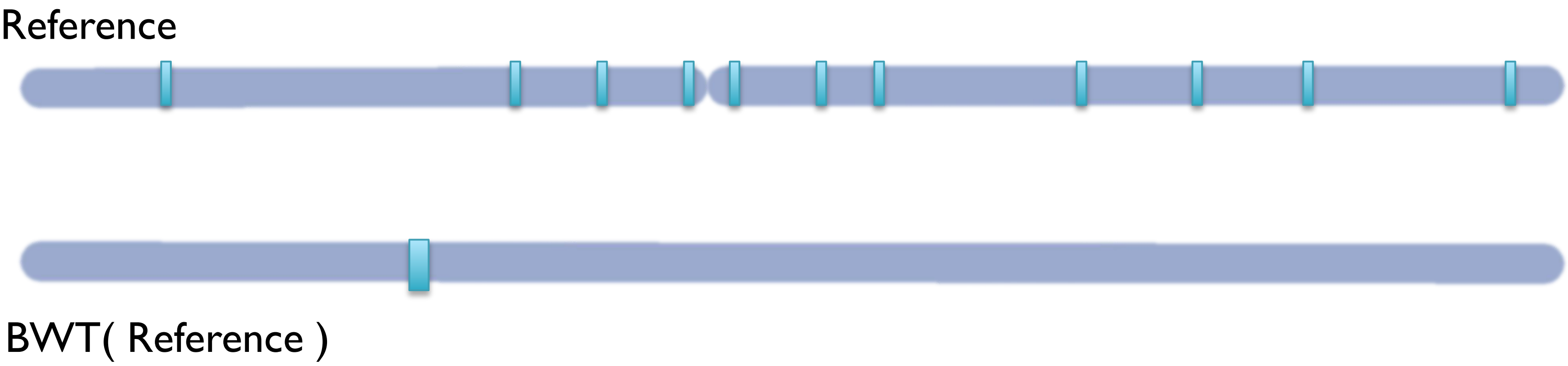
Bowtie algorithm

Reference

BWT(Reference)

Query:
AATGATACGGCGACCACCGAGATCTA

Bowtie algorithm



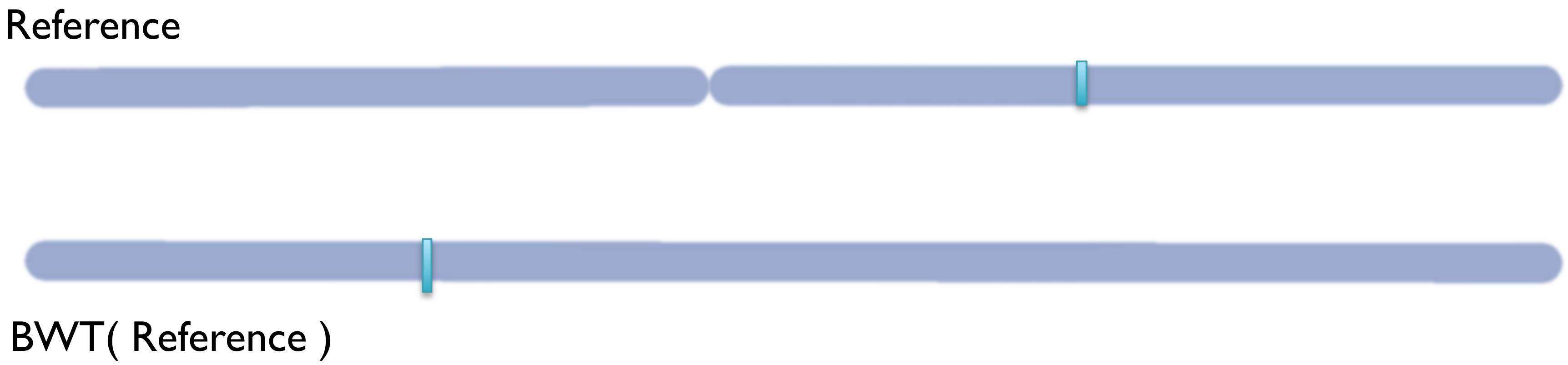
Reference

BWT(Reference)

Query:

AATGTTACGGCGACCACCGAGATCTA

Bowtie algorithm



Reference

BWT(Reference)

Query:

AATGTTACGGCGACCACCGAGATCTA