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 *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

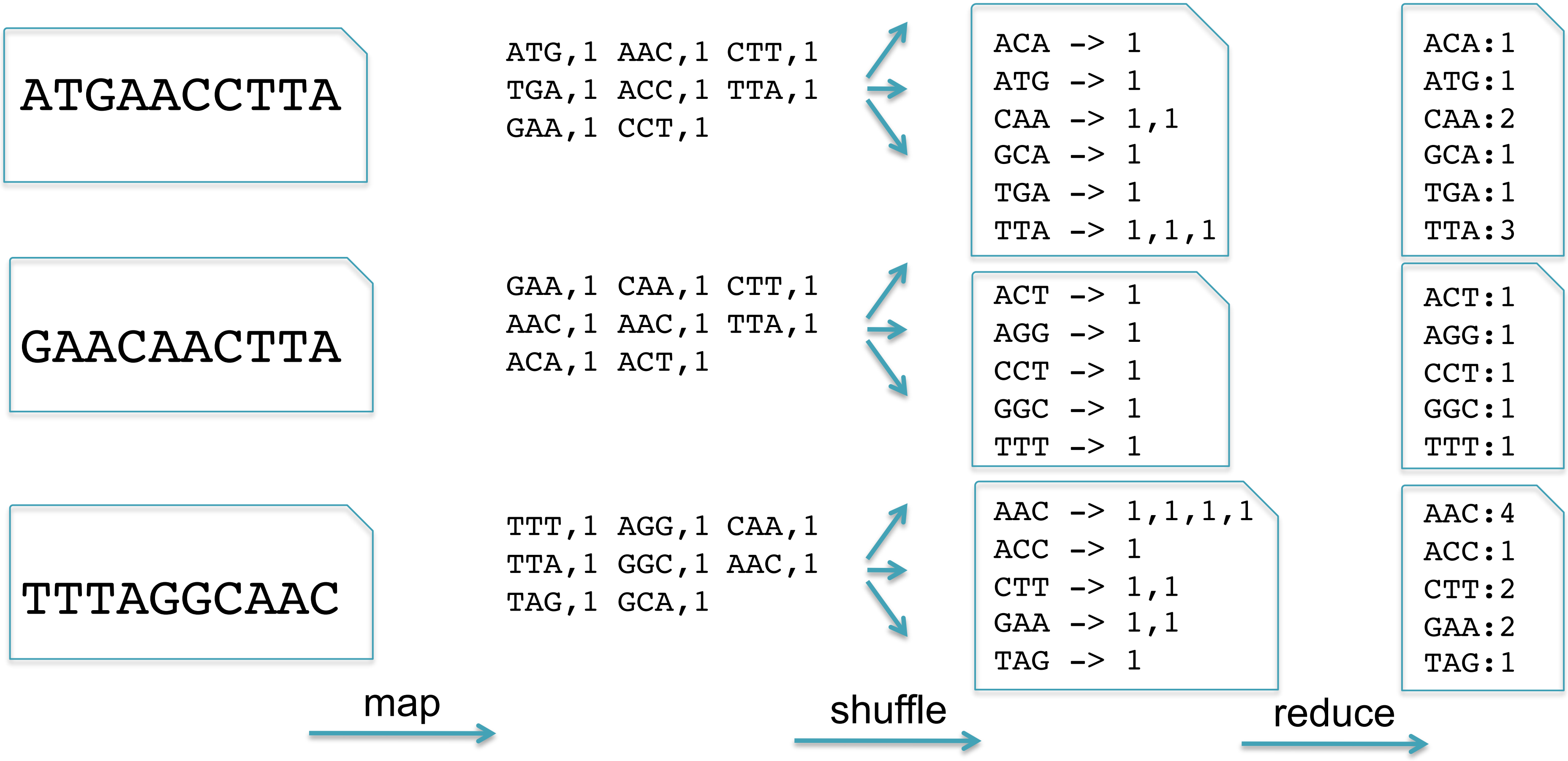
(Pushkarev *et al.*, 2009)

Critical Computational Challenges: Alignment and Assembly of Huge Datasets

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

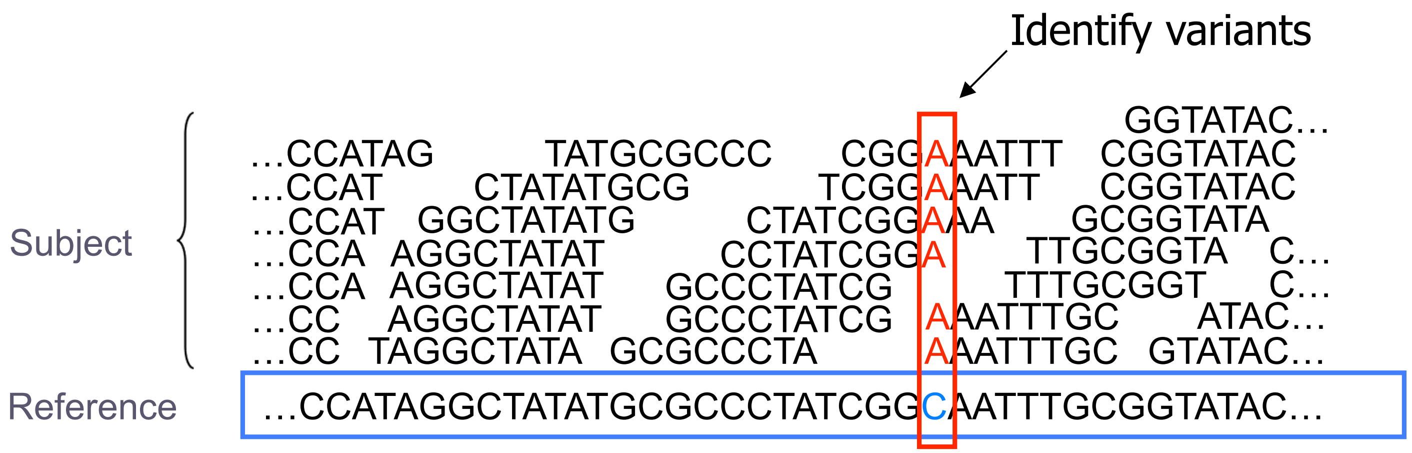


```
ATGAACCTTA      ATG,1 AAC,1 CTT,1      ACA -> 1            ACA:1
                TGA,1 ACC,1 TTA,1      ATG -> 1            ATG:1
                GAA,1 CCT,1            CAA -> 1,1          CAA:2
                                       GCA -> 1            GCA:1
                                       TGA -> 1            TGA:1
                                       TTA -> 1,1,1        TTA:3

GAACAACTTA      GAA,1 CAA,1 CTT,1      ACT -> 1            ACT:1
                AAC,1 AAC,1 TTA,1      AGG -> 1            AGG:1
                ACA,1 ACT,1            CCT -> 1            CCT:1
                                       GGC -> 1            GGC:1
                                       TTT -> 1            TTT:1

TTTAGGCAAC      TTT,1 AGG,1 CAA,1      AAC -> 1,1,1,1      AAC:4
                TTA,1 GGC,1 AAC,1      ACC -> 1            ACC:1
                TAG,1 GCA,1            CTT -> 1,1          CTT:2
                                       GAA -> 1,1          GAA:2
                                       TAG -> 1            TAG:1
```

Short Read Mapping with MapReduce



```
                                                          GGTATAC...
...CCATAG       TATGCGCCC      CGGAAATTT   CGGTATAC
...CCAT      CTATATGCG       TCGGAAATT    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...
 ...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

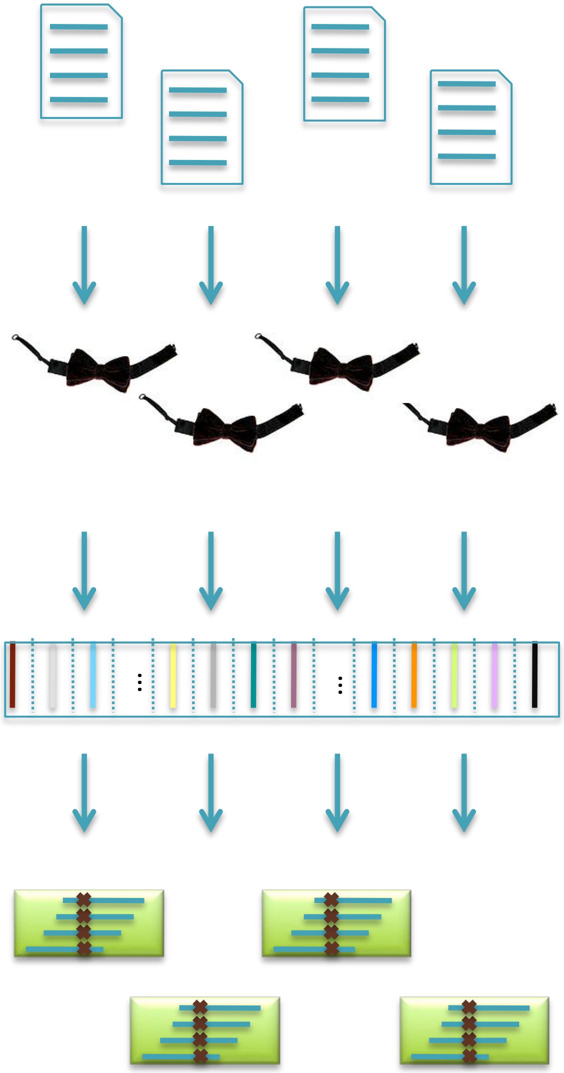
- Given a reference and many subject reads, report one or more “good” end-to-end alignments per alignable read
  - Maps the read to where it originated

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

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%

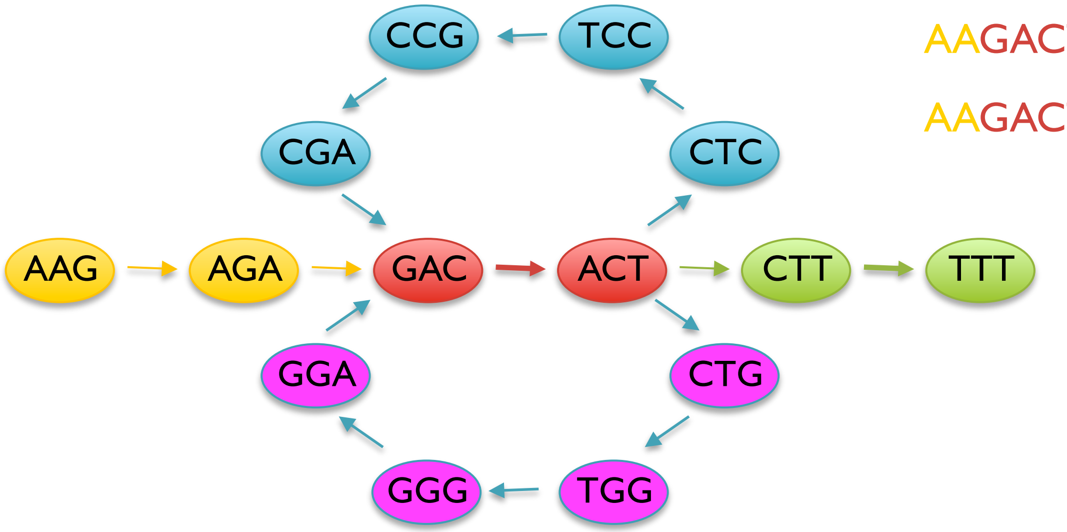
Searching for SNPs with Cloud Computing.
Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology.*

Short Read Assembly

Reads

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

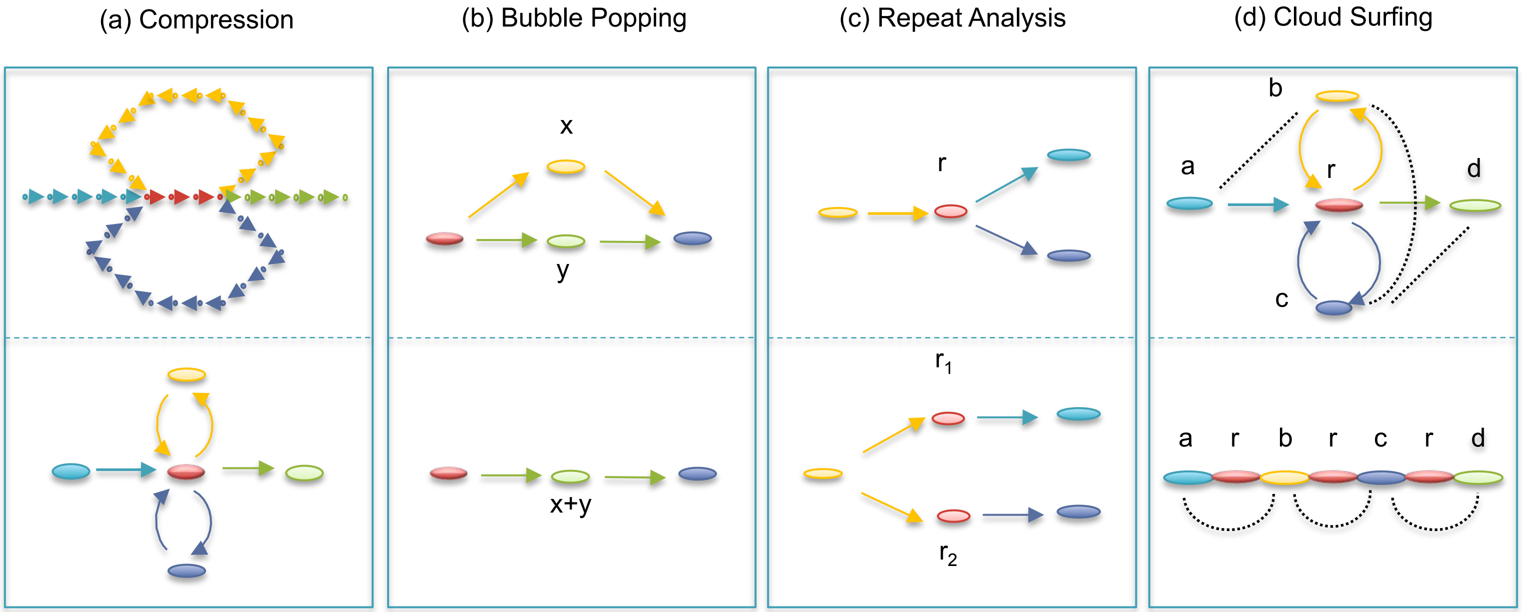
(Chaisson, 2009)

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



(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.*

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