Assembly of Large Genomes using Cloud Computing

Michael Schatz

July 23, 2010

Illumina Sequencing Panel

How to compute with 1000s of cores

Michael Schatz

July 23, 2010
Illumina Sequencing Panel

Parallel Architectures

- Why Parallel?
  - CPU manufactures up against fundamental limitations
  - Need it done faster, problem is too big for a single machine

- Multi-core (2-10s of cores)
  - Familiar programming environment
  - Limited scaling

- GPU & FPGA (10s – 1000 of cores)
  - Very high performance for some applications
  - Limited/Slow memory, complicated development environment

- Cluster / Distributed Programming (10s – 1000s of machines)
  - Well suited for very large data problems
  - Scheduling, Fault tolerance & Network communication

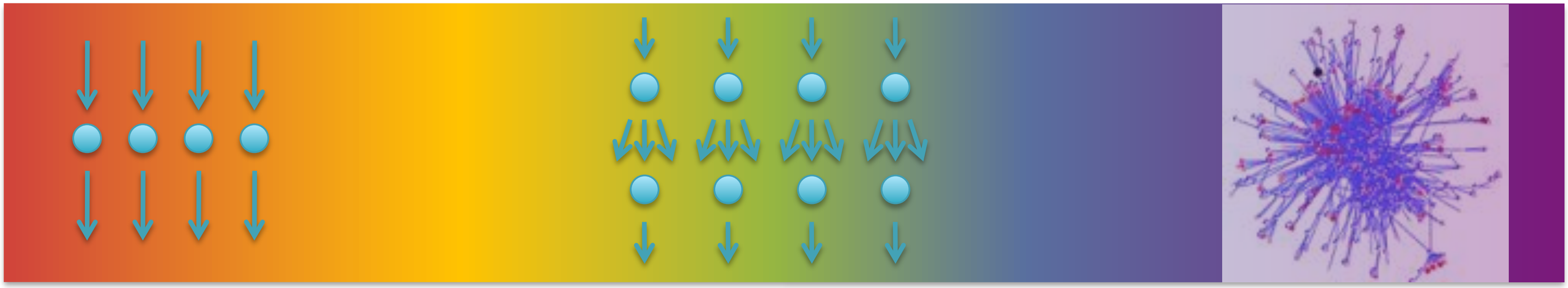
Amazon Web Services

http://aws.amazon.com

- "All you need is a credit card to use one of the largest datacenters in the world"
  - Best for large infrequent computations
- Elastic Compute Cloud (EC2)
  - On demand computing power
    - Support for Windows, Linux, & OpenSolaris
    - Starting at 8.5¢ / core / hour
- Simple Storage Service (S3)
  - Scalable data storage
    - 10¢ / GB upload fee, 15¢ / GB monthly fee

Parallel Algorithms Spectrum

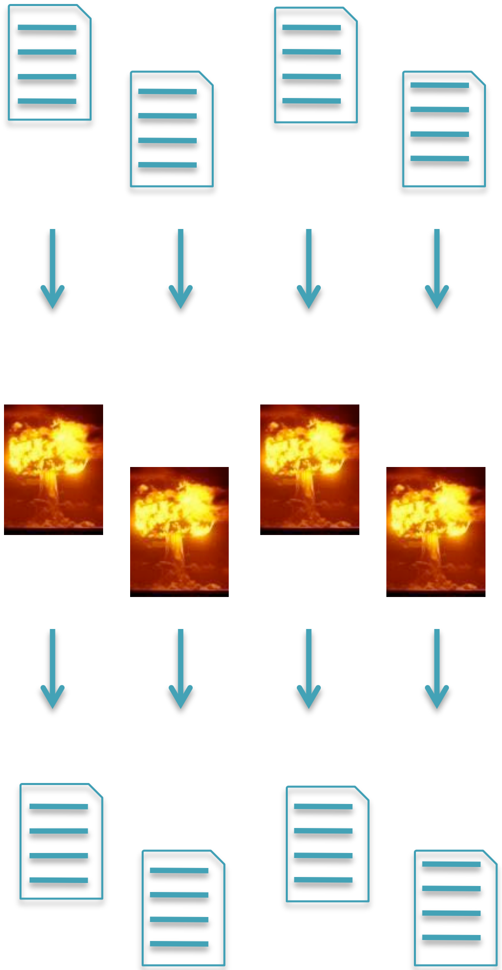
Embarrassingly Parallel	Loosely Coupled	Tightly Coupled



Batch Computing	MapReduce/DryadLINQ	MPI/MapReduce/Pregel
Alignment HMM Scoring	Genotyping K-mer Counting	Graph Analysis Genome Assembly
Scheduling + Load Balance	Embarrassingly Parallel + Parallel Communication	Loosely Coupled + Parallel Algorithm Design

Embarrassingly Parallel

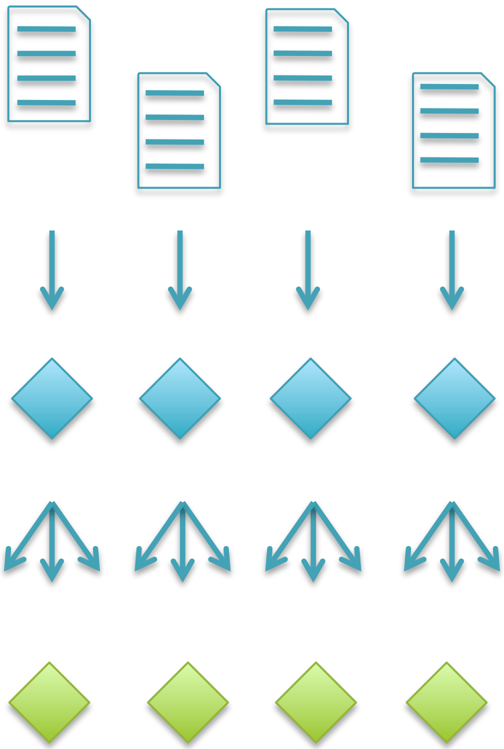
- Batch computing
  - Each item is independent
  - Split input into many chunks
  - Process each chunk separately on a different computer
- Challenges
  - Distributing work, load balancing, monitoring & restart
- Technologies
  - Condor, Sun Grid Engine
  - Amazon Simple Queue



Elementary School Dance

Loosely Coupled

- Divide and conquer
  - Independently process many items
  - Group partial results
  - Scan partial results into final answer
- Challenges
  - Batch computing challenges
  - + Shuffling of huge datasets
- Technologies
  - Hadoop, Elastic MapReduce, Dryad
  - Parallel Databases



Junior High Dance

Hadoop MapReduce

http://hadoop.apache.org

- 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
    - GATK is an alternative implementation specifically for NGS

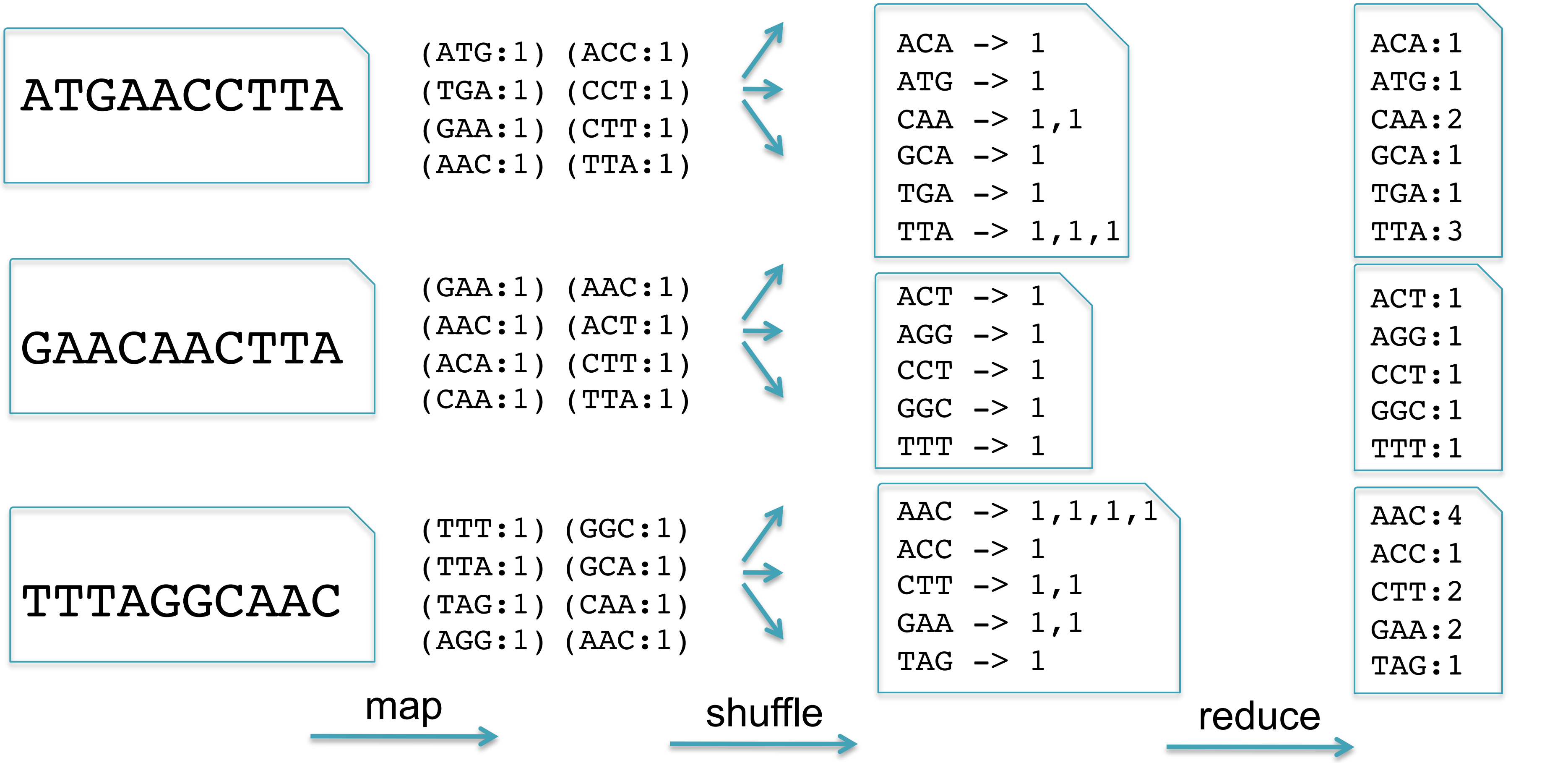
- Benefits
  - Scalable, Efficient, Reliable
  - Easy to Program
  - Runs on commodity computers

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

K-mer Counting

- 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



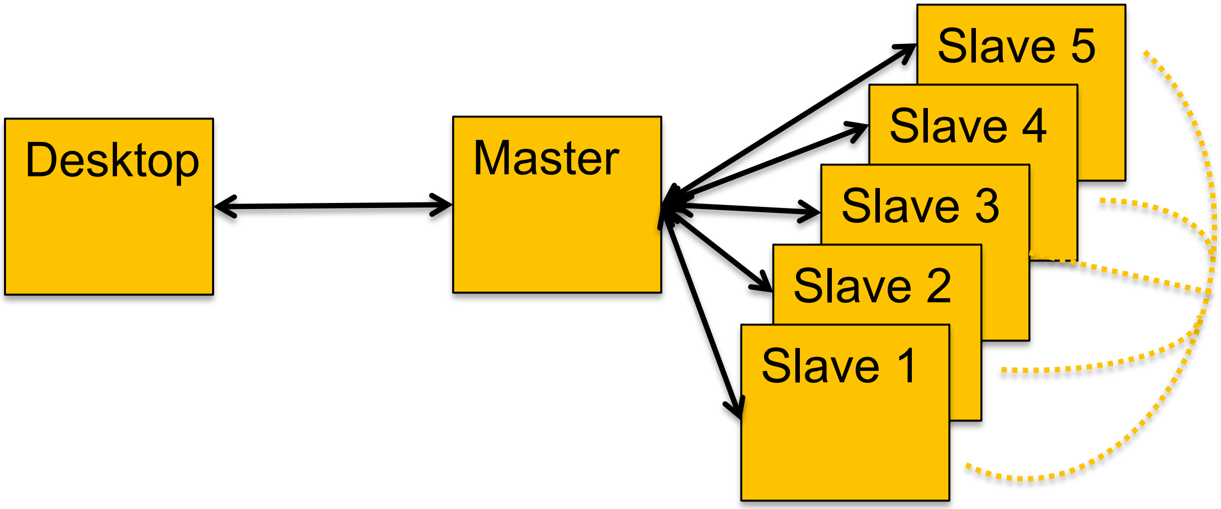
```
ATGAACCTTA

GAACAACTTA

TTTAGGCAAC
```

```
(ATG:1) (ACC:1)
(TGA:1) (CCT:1)
(GAA:1) (CTT:1)
(AAC:1) (TTA:1)

(GAA:1) (AAC:1)
(AAC:1) (ACT:1)
(ACA:1) (CTT:1)
(CAA:1) (TTA:1)

(TTT:1) (GGC:1)
(TTA:1) (GCA:1)
(TAG:1) (CAA:1)
(AGG:1) (AAC: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
```

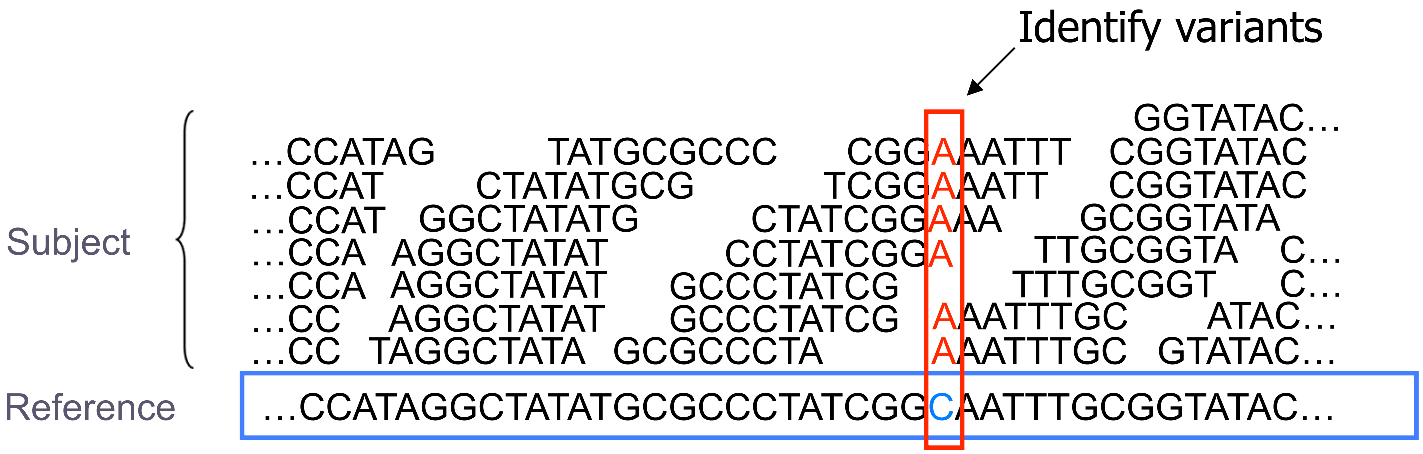
```
ACA:1
ATG:1
CAA:2
GCA:1
TGA:1
TTA:3

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

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

Hadoop Architecture



- Hadoop Distributed File System (HDFS)
  - Data files partitioned into large chunks (64MB), replicated on multiple nodes
  - Computation moves to the data, rack-aware scheduling
- Hadoop MapReduce system won the 2009 GreySort Challenge
  - Sorted 100 TB in 173 min (578 GB/min) using 3452 nodes and 4x3452 disks

Short Read Mapping

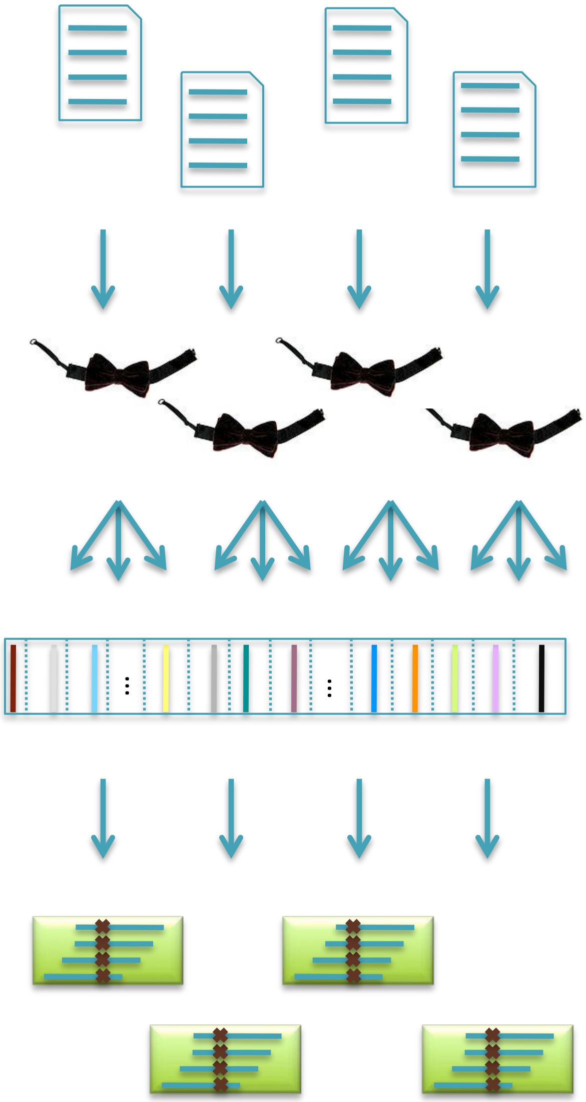


```
                                  GGTATAC…
…CCATAG    TATGCGCCC   CGGAAATTT CGGTATAC
…CCAT   CTATATGCG     TCGGAAATT  CGGTATAC
…CCAT GGCTATATG    CTATCGGAAA   GCGGTATA
…CCA AGGCTATAT    CCTATCGGA   TTGCGGTA  C…
…CCA AGGCTATAT  GCCCTATCG    TTTGCGGT   C…
…CC  AGGCTATAT  GCCCTATCG AAATTTGC   ATAC…
…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
  - Find where the read most likely originated
  - Fundamental computation for many assays
    - Genotyping RNA-Seq Methyl-Seq
    - Structural Variations Chip-Seq Hi-C-Seq
- Desperate need for scalable solutions
  - Single human requires ~1,000 CPU hours / genome

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

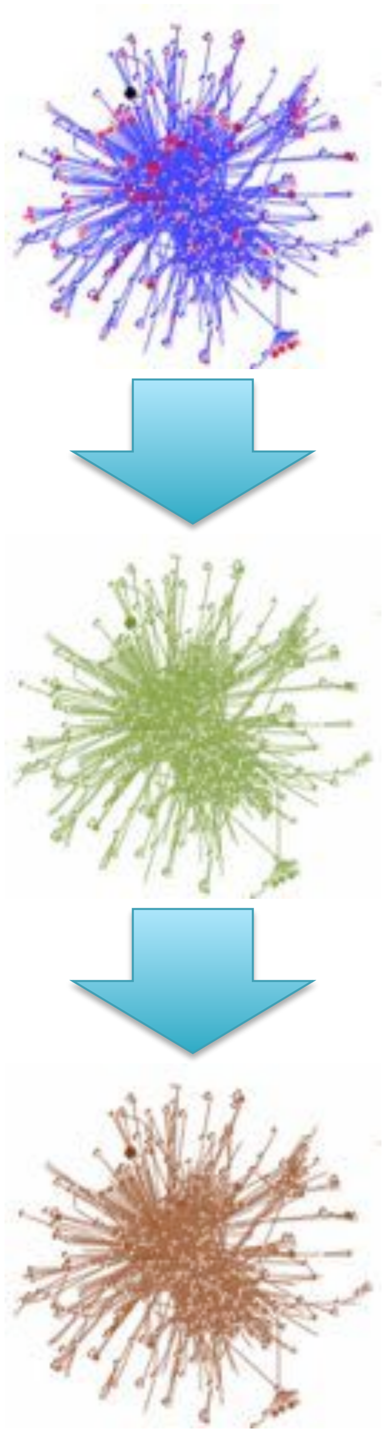
	Asian Individual Genome		
Data Loading	3.3 B reads	106.5 GB	$10.65
Data Transfer	1h :15m	40 cores	$3.40
Setup	0h : 15m	320 cores	$13.94
Alignment	1h : 30m	320 cores	$41.82
Variant Calling	1h : 00m	320 cores	$27.88
End-to-end	4h : 00m		$97.69

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.* **10**:R134

Tightly Coupled

- Computation that cannot be partitioned
  - Graph Analysis
  - Molecular Dynamics
  - Population simulations
- Challenges
  - Loosely coupled challenges
  - + Parallel algorithms design
- Technologies
  - MPI
  - MapReduce, Dryad, Pregel



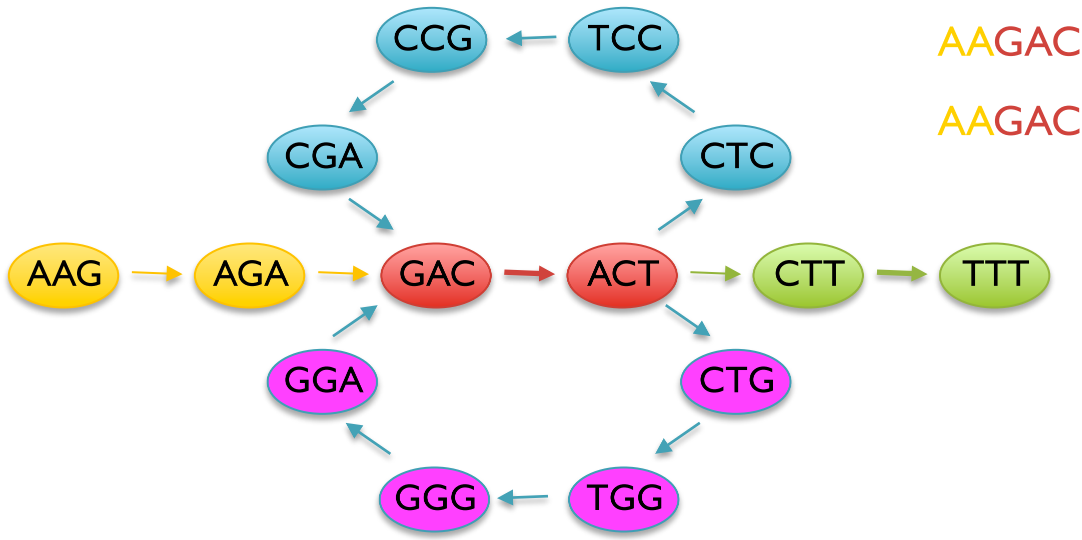
High School Dance

Short Read Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…

de Bruijn Graph



Potential Genomes

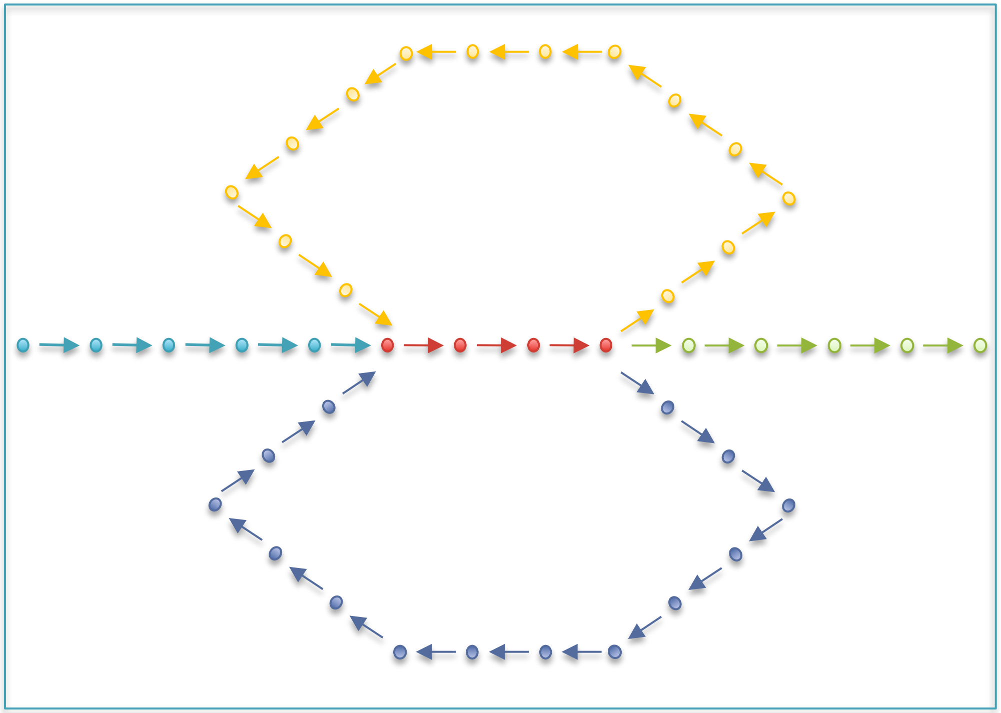
AAGACTCCGACTGGGACTTT

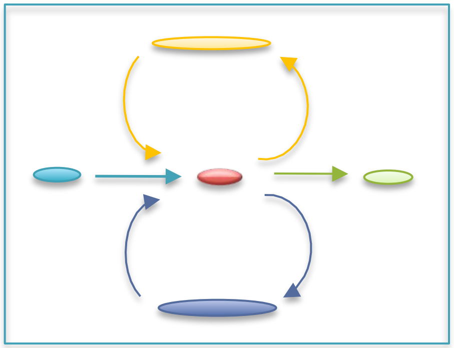
AAGACTGGGACTCCGACTTT

- 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) serial: > 2TB of RAM x weeks
  - ABySS (Simpson *et al.*, 2009) MPI: 168 cores x ~96 hours
  - SOAPdenovo (Li *et al.*, 2010) pthreads: 40 cores x 40 hours, >140 GB RAM

Graph Compression

- Graph construction straightforward in MapReduce
  - Straightforward extension to k-mer counting
- After construction, many edges are unambiguous
  - Merge together compressible nodes
  - Graph physically distributed over hundreds of computers





Warmup Exercise

- Who here was born closest to July 23?
  - You can only compare to one person at a time

2010 World Cup Bracket
Uruguay
South Korea
United States
Ghana
Netherlands
Slovakia
Brazil
Chile
Germany
England
Argentina
Mexico
Paraguay
Japan
Spain
Portugal
World Cup Champion

Find winner among 16 teams in just 4 rounds

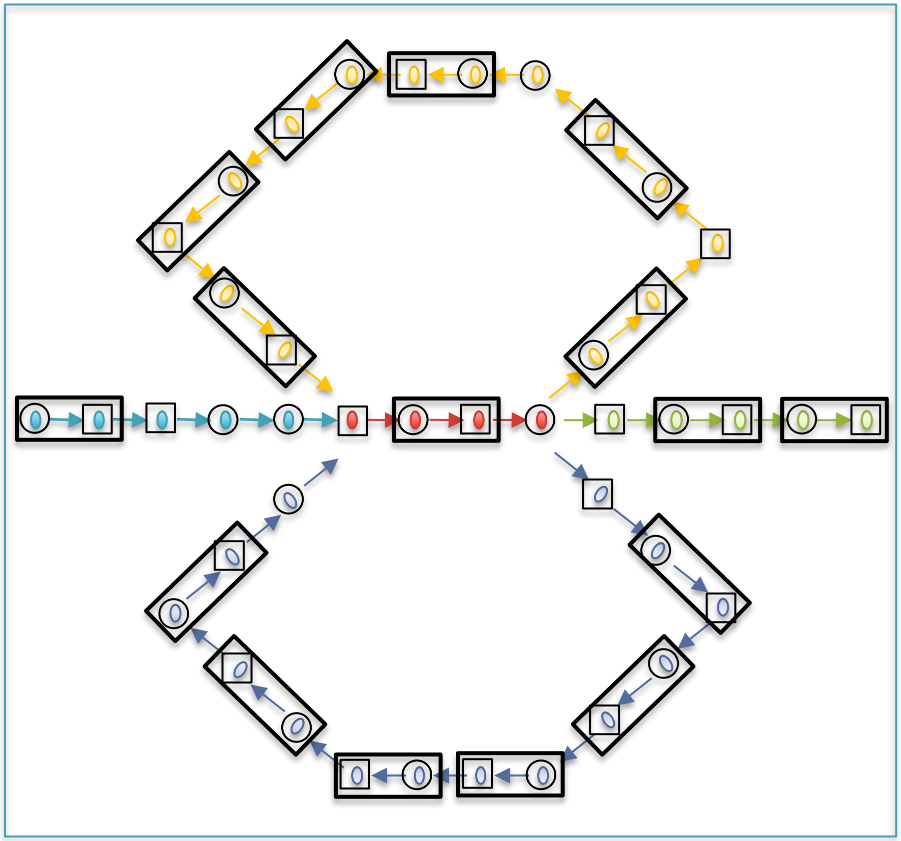
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ/$\boxed{\text{T}}$ to each compressible node
- Compress Ⓗ→$\boxed{\text{T}}$ links



Initial Graph: 42 nodes

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

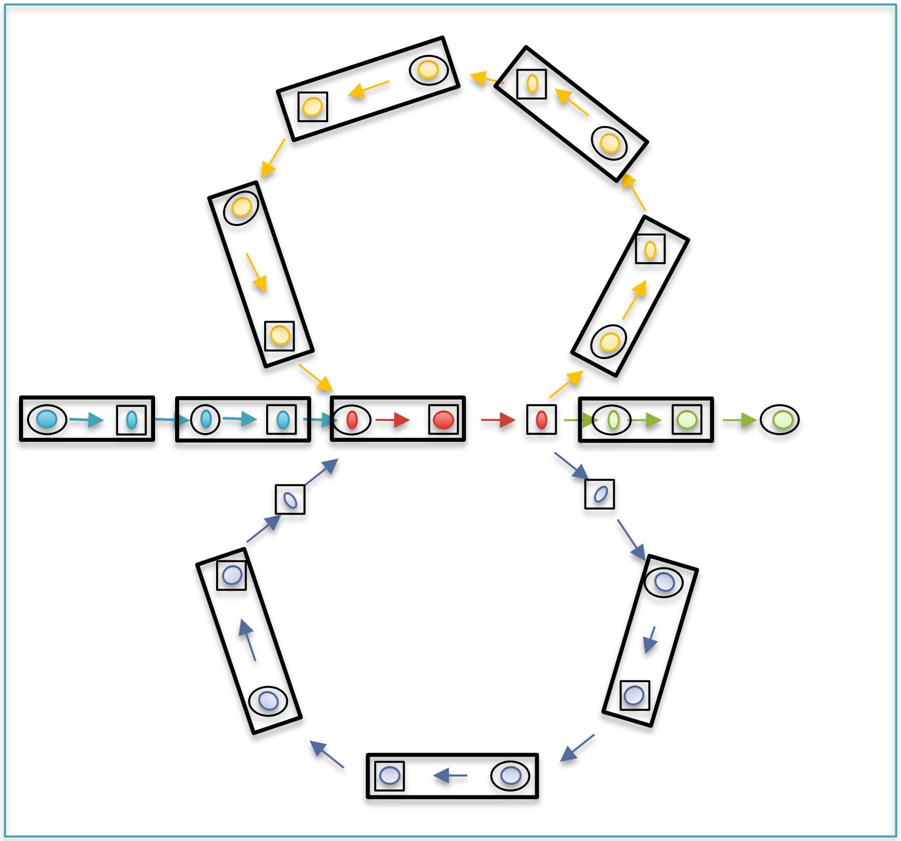
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ/ $\boxed{\text{T}}$ to each compressible node
- Compress Ⓗ→$\boxed{\text{T}}$ links



Round 1: 26 nodes (38% savings)

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

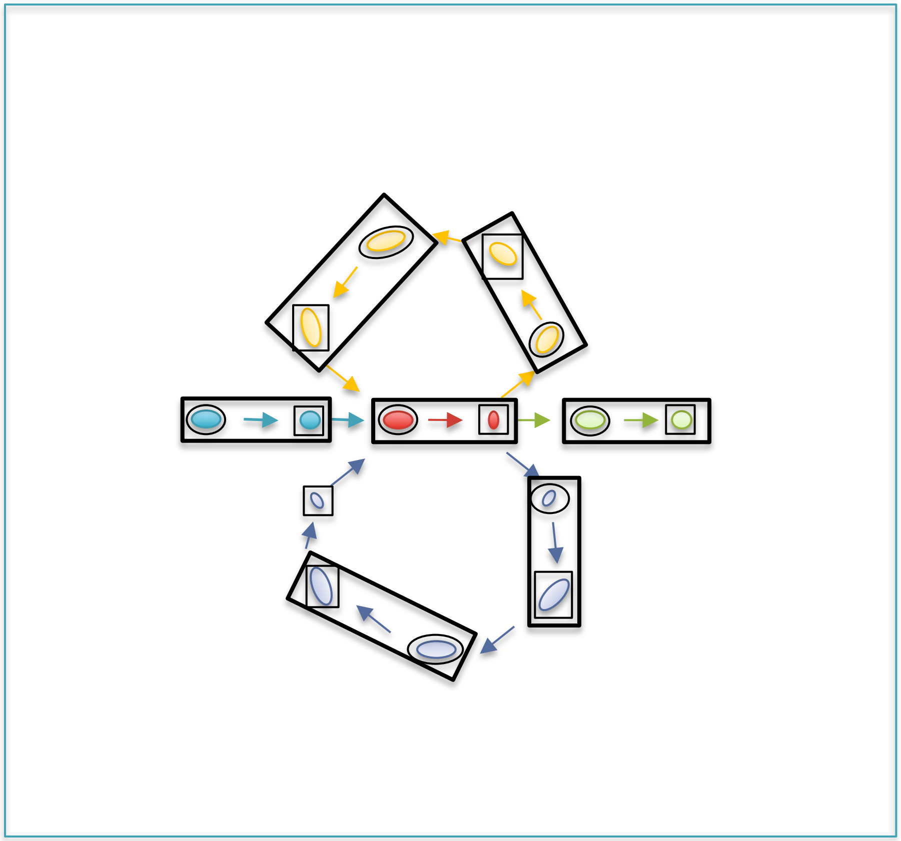
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ/ $\boxed{\text{T}}$ to each compressible node
- Compress Ⓗ➔$\boxed{\text{T}}$ links



Round 2: 15 nodes (64% savings)

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

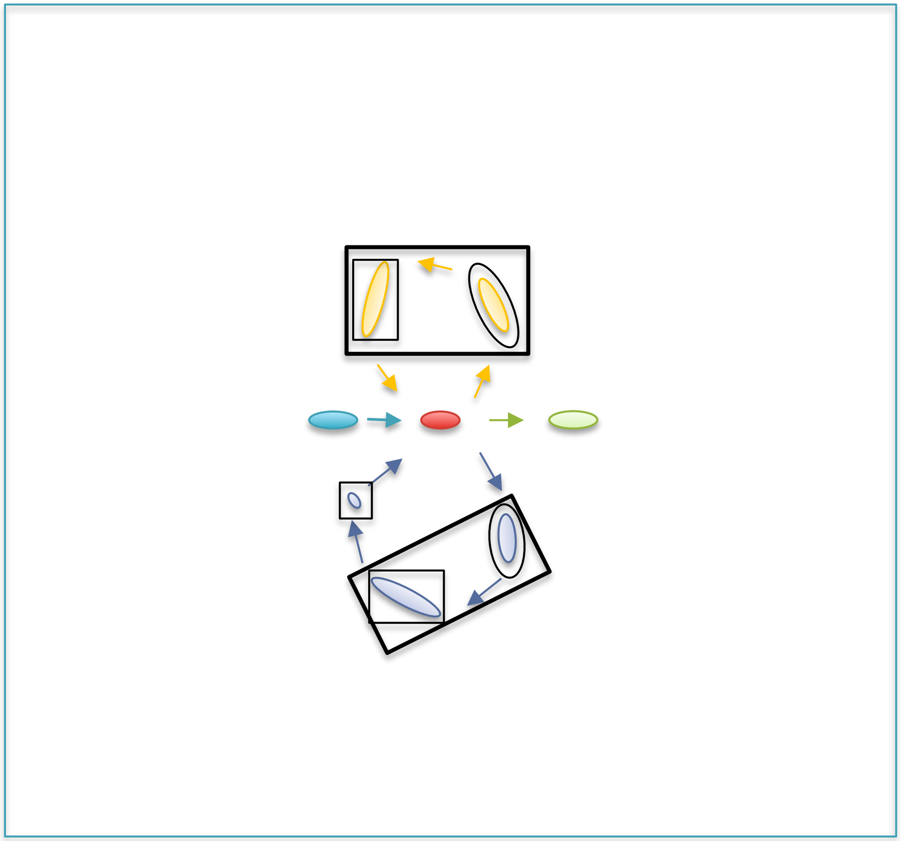
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ/ 🅃 to each compressible node
- Compress Ⓗ→🅃 links



Round 2: 8 nodes (81% savings)

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

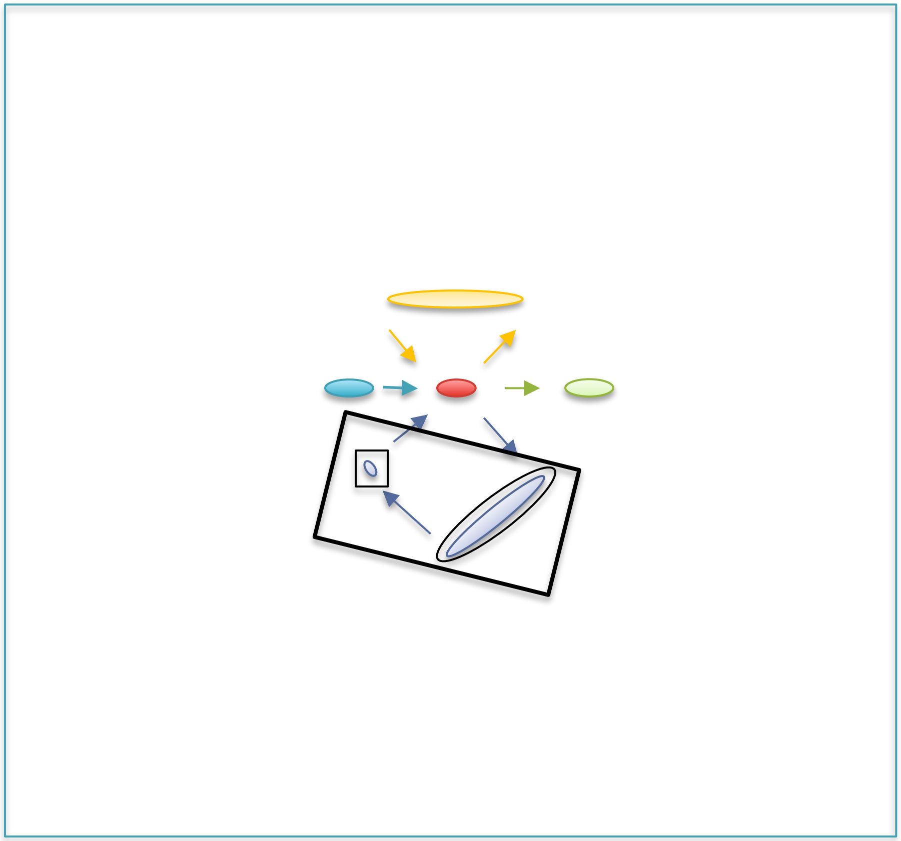
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ/ $\boxed{\text{T}}$ to each compressible node
- Compress Ⓗ ➔ $\boxed{\text{T}}$ links



Round 3: 6 nodes (86% savings)

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

Fast Path Compression

Challenges

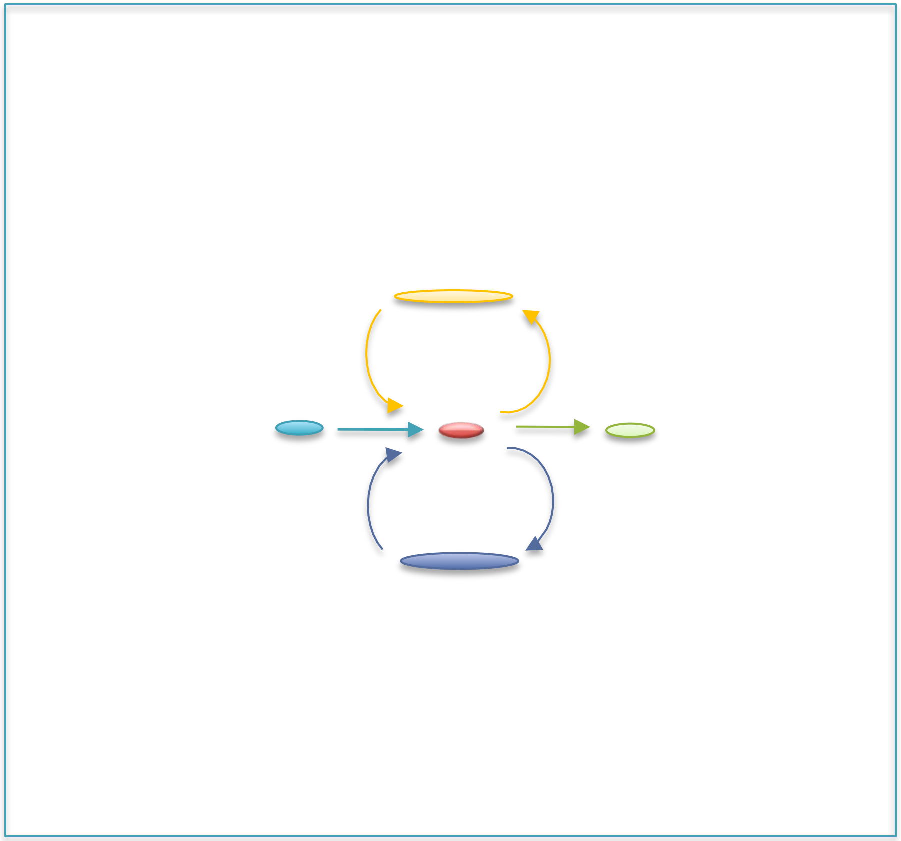
- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ/ $\boxed{\text{T}}$ to each compressible node
- Compress Ⓗ→$\boxed{\text{T}}$ links

Performance

- Compress all chains in log(S) rounds

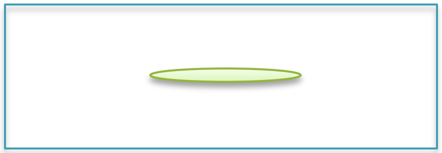


Round 4: 5 nodes (88% savings)

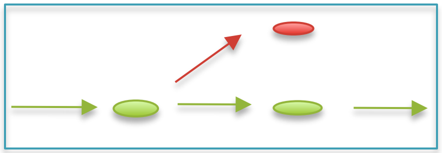
Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

Node Types

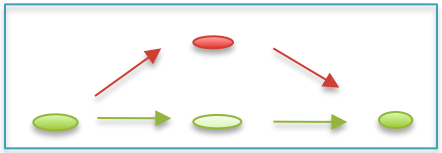
(Chaisson, 2009)



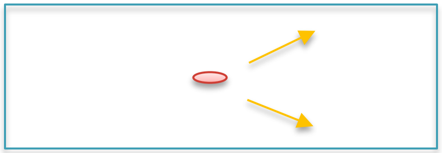
Isolated nodes (10%)



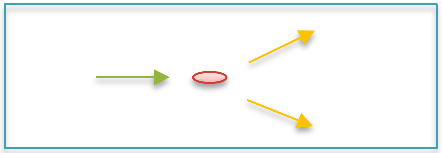
Tips (46%)



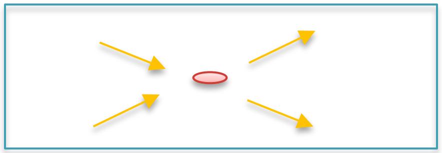
Bubbles/Non-branch (9%)



Dead Ends (.2%)



Half Branch (25%)



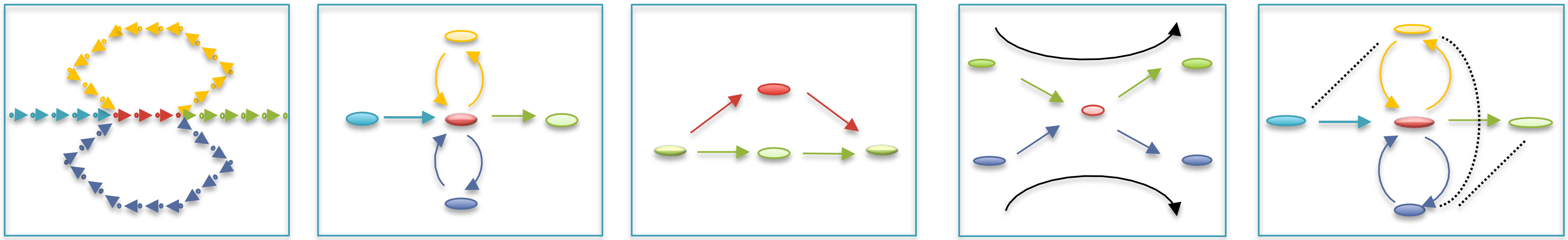
Full Branch (10%)

Contrail

http://contrail-bio.sourceforge.net

De novo bacterial assembly

- *Genome: E. coli* K12 MG1655, 4.6Mbp
- *Input:* 20.8M 36bp reads, 200bp insert (~150x coverage)
- *Preprocessor*: Quality-Aware Error Correction



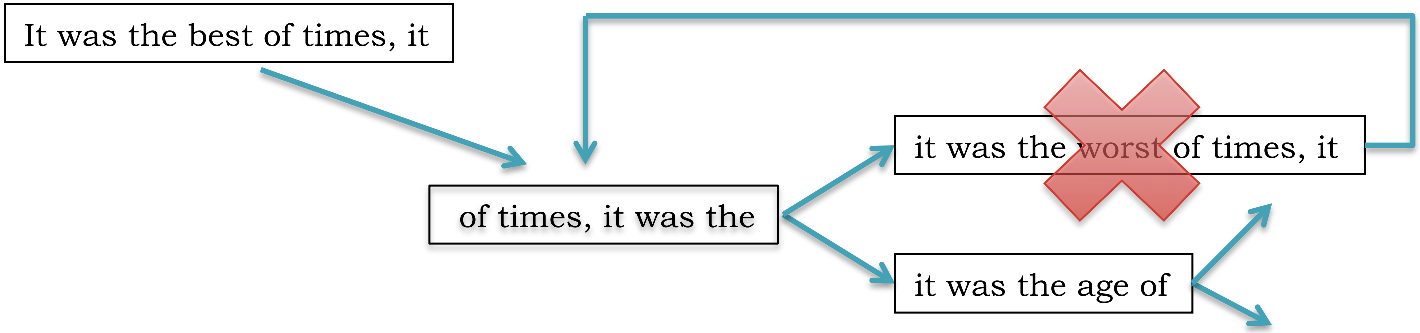
	Initial	Compressed	Error Correction	Resolve Repeats	Cloud Surfing
N	5.1 M	245,131	2,769	1,909	300
Max	27 bp	1,079 bp	70,725 bp	90,088 bp	149,006 bp
N50	27 bp	156 bp	15,023 bp	20,062 bp	54,807 bp

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

E. coli Assembly Quality

Incorrect contigs: Align at < 95% identity or < 95% of their length

Assembler	Contigs ≥ 100bp	N50 (bp)	Incorrect contigs
Contrail PE	300	54,807	4
Contrail SE	529	20,062	0
SOAPdenovo PE	182	89,000	5
ABySS PE	233	45,362	13
Velvet PE	286	54,459	9
EULER-SR PE	216	57,497	26
SSAKE SE	931	11,450	38
Edena SE	680	16,430	6



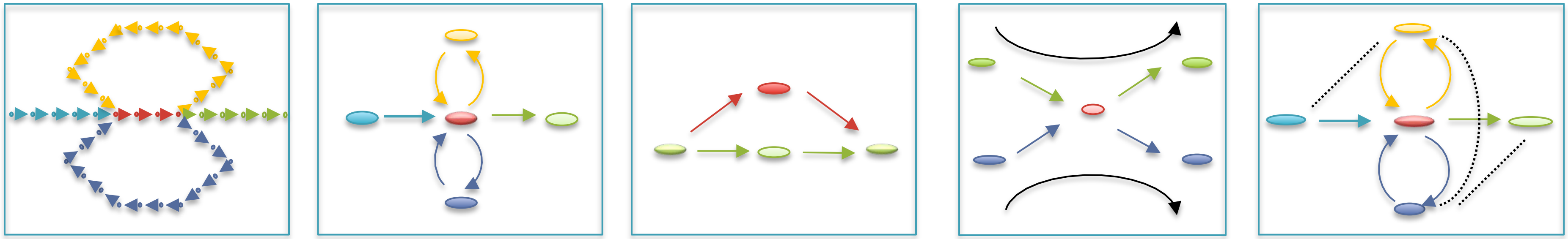
One more thing...

Contrail

http://contrail-bio.sourceforge.net

De novo Assembly of the Human Genome

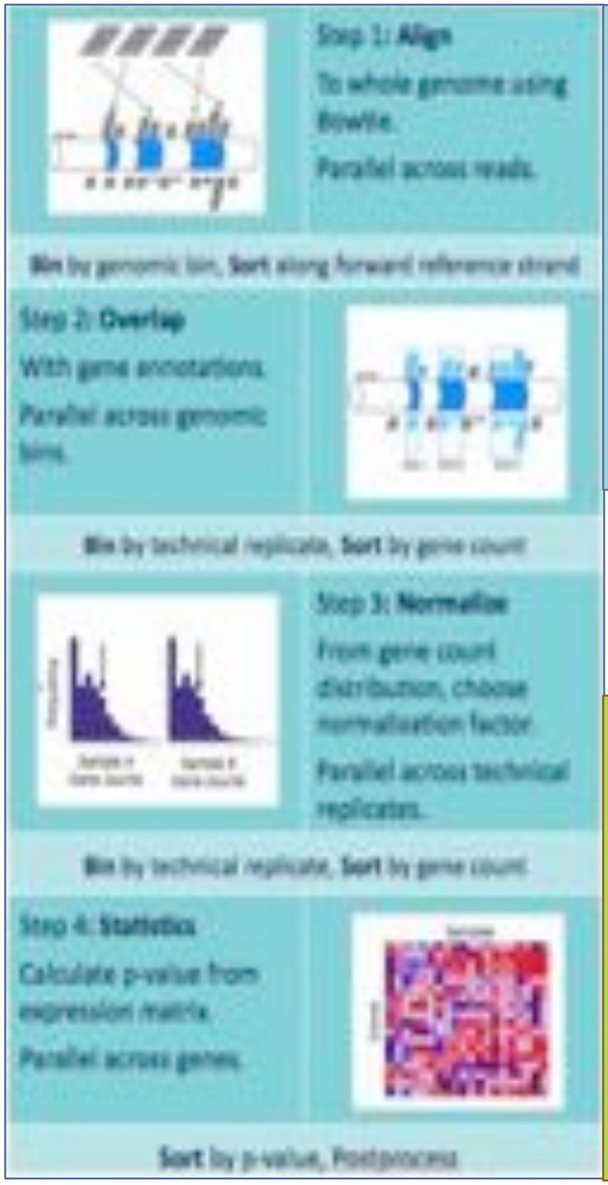
- *Genome:* African male NA18507 (SRA000271, Bentley *et al.*, 2008)
- *Input:* 3.5B 36bp reads, 210bp insert (~40x coverage)



	Initial	Compressed	Error Correction	Resolve Repeats	Cloud Surfing
N	>7 B	>1 B	4.2 M	4.1 M	
Max	27 bp	303 bp	20,594 bp	20,594 bp	In progress
N50	27 bp	< 100 bp	995 bp	1,050 bp	

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

Scalable Solutions for DNA Sequence Analysis



Step 1: **Align**

To whole genome using Bowtie.

Parallel across reads.

Bin by genomic bin, **Sort** along forward reference strand

Step 2: **Overlap**

With gene annotations.

Parallel across genomic bins.

Bin by technical replicate, **Sort** by gene count

Step 3: **Normalize**

From gene count distribution, choose normalization factor.

Parallel across technical replicates.

Bin by technical replicate, **Sort** by gene count

Step 4: **Statistics**

Calculate p-value from expression matrix.

Parallel across genes.

Sort by p-value, Postprocess

Myrna

http://bowtie-bio.sf.net/myrna

Cloud-scale differential gene expression from RNA-seq

Ben Langmead,
Kasper Hansen, Jeff Leek

Quake

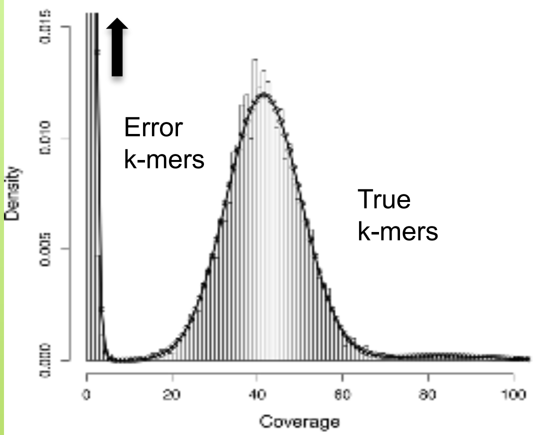
http://www.cbcb.umd.edu/software/quake

Quality-aware error correction of sequencing reads

David Kelley,
Michael Schatz, Steven Salzberg

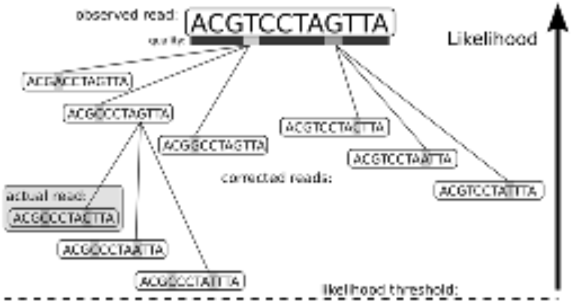
Step 1: **Compute Q-mer Distribution**

Compute in parallel across reads, merging together results across files



Step 2: **Correct reads**

Untrusted k-mers are evaluated in order of decreasing likelihood.



Summary

- Surviving the data deluge means computing in parallel
  - Cloud computing is an attractive platform for large scale sequence analysis and computation
- Significant obstacles ahead
  - Transfer time
  - Privacy / security requirements
  - Time and expertise required for development
  - Price
  - What are the alternatives?
- Emerging technologies are a great start, but we need continued research
  - A word of caution: new technologies are new

Acknowledgements

Steven Salzberg

Mihai Pop

Jimmy Lin

Ben Langmead

Dan Sommer

David Kelley

amazon webservices™

Thank You!

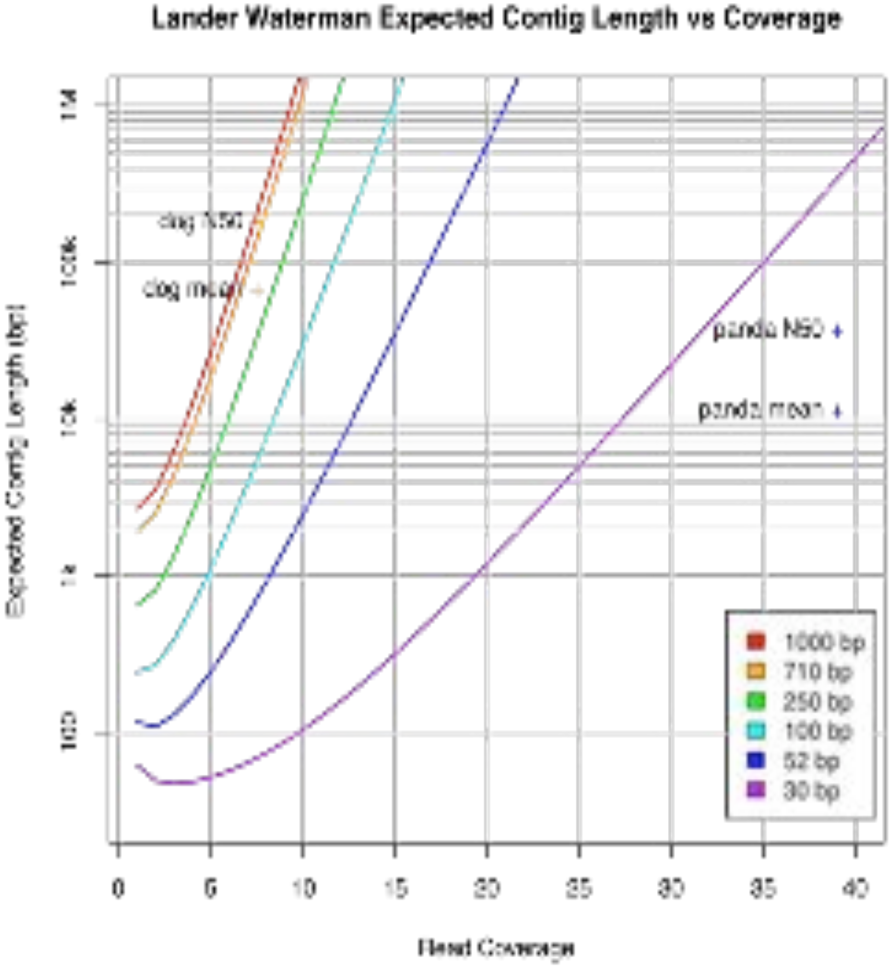
http://www.cbcb.umd.edu/~mschatz

@mike_schatz

Genome Coverage

Idealized assembly

- Uniform probability of a read starting at a given position
  - p = G/N
- Poisson distribution in coverage along genome
  - Contigs end when there is no overlapping read
- Contig length is a function of coverage and read length
  - Short reads require much higher coverage

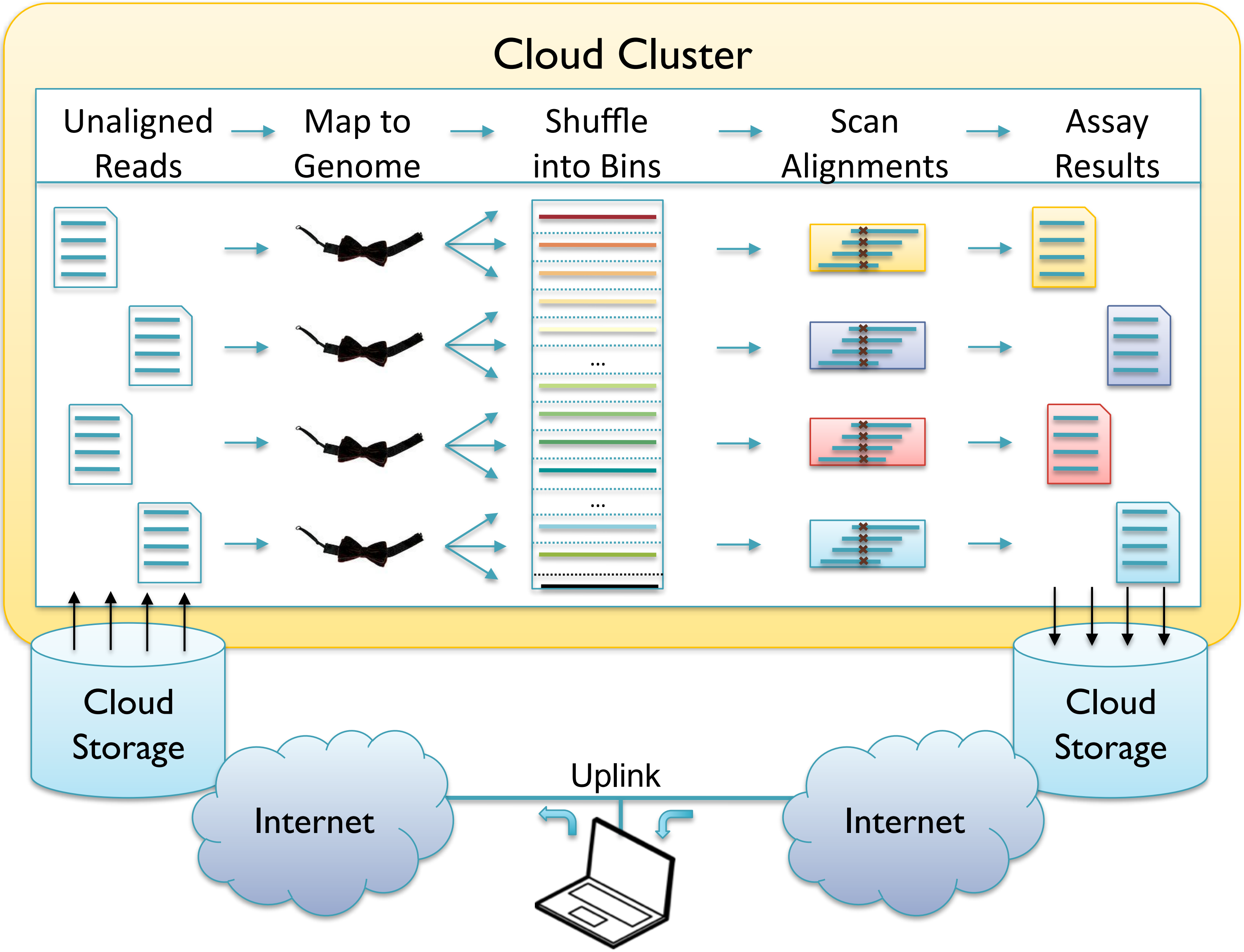


Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.*

Recent Large Assemblies

Table 1. De novo assemblies of second generation sequencing projects.

		Input Sequence						Assembly								
Organism/ Genome Size	Assembler/ Status	Type	Pair Size	Avg Read(bp)	# Reads	Read Cov	Phys Cov	Contigs				Scaffolds				[illegible]
								#	N50	Max	Total	#	N50	Max	Total	
[illegible]	[illegible]	SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	NA	NA	NA	NA	
[illegible]	[illegible]	Sanger	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		Sanger	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		Sanger	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		454	[illegible]	[illegible]	[illegible]	[illegible]	-									
[illegible]	[illegible]	Sanger	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	NA	[illegible]	[illegible]	[illegible]	NA	[illegible]	
		Sanger	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		Sanger	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	NA	[illegible]	NA	[illegible]	NA	[illegible]	NA	[illegible]	[illegible]
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	NA	[illegible]	NA	[illegible]	NA	[illegible]	NA	[illegible]	[illegible]
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
[illegible]	[illegible]	SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
[illegible]	[illegible]	454	[illegible]	[illegible]	[illegible]	[illegible]	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
		454	[illegible]	[illegible]	[illegible]	[illegible]	-									
		454	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		454	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		SA	[illegible]	[illegible]	[illegible]	[illegible]	-									
		SOLiD	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
[illegible]	[illegible]	454	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	NA	
		454	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		454	[illegible]	[illegible]	[illegible]	[illegible]	-									
		SA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]									
		SA	[illegible]	[illegible]	[illegible]	[illegible]	-									



Cloud Computing and the DNA Data Race.

Schatz, MC, Langmead, B, Salzberg SL (2010) *Nature Biotechnology.*

Human Assembly Quality

Assembler	Contigs ≥ 100bp	N50 (bp)	Total Length (Gbp)
Contrail SE	4,285,080	1,050	2.13
SOAPdenovo PE	NA	4,611	2.63
SOAPdenovo SE	NA	886	2.10
ABySS PE	2,762,173	1,499	2.18
ABySS SE	4,348,132	870	2.10