Cloud Computing and the DNA Data Race

Michael Schatz

February 15, 2011

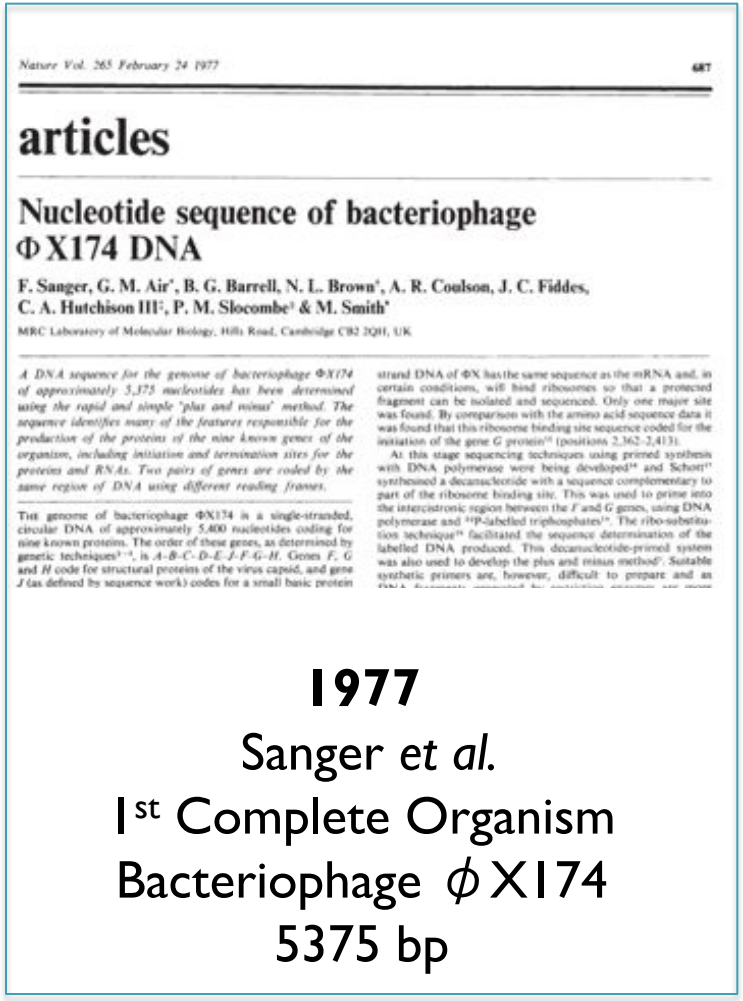
Laufer Center for Physical and Quantitative Biology

Outline

1. Milestones in DNA Sequencing
2. Hadoop & Cloud Computing
3. Sequence Analysis in the Clouds
    1. Sequence Alignment
    2. Mapping & Genotyping
    3. Genome Assembly

Milestones in DNA Sequencing

1970 | 1980 | 1990 | 2000 | 2010



Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown*, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III*, P. M. Slocombe* & M. Smith*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

THE genome of bacteriophage ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[1-4], is *A–B–C–D–E–J–F–G–H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, and gene *J* (as defined by sequence work) codes for a small basic protein

strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[10] (positions 2,362–2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[11] and Schott[12] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[13]. The ribo-substitution technique[14] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[5]. Suitable synthetic primers are, however, difficult to prepare and as

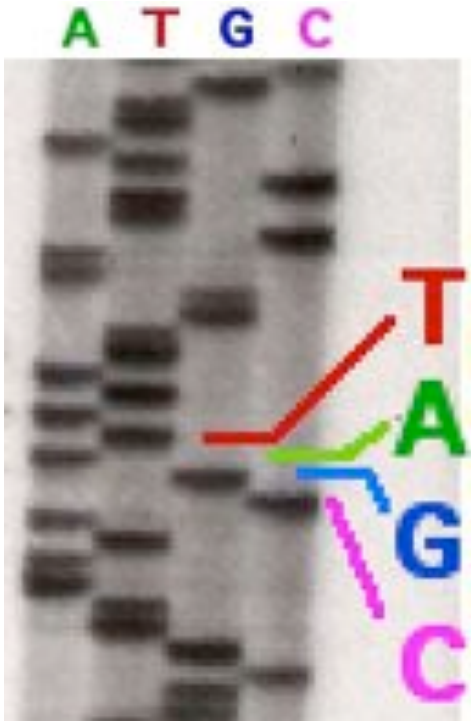
1977

Sanger *et al.*

1st Complete Organism

Bacteriophage ϕX174

5375 bp



Radioactive Chain Termination

5000bp / week / person

http://en.wikipedia.org/wiki/File:Sequencing.jpg

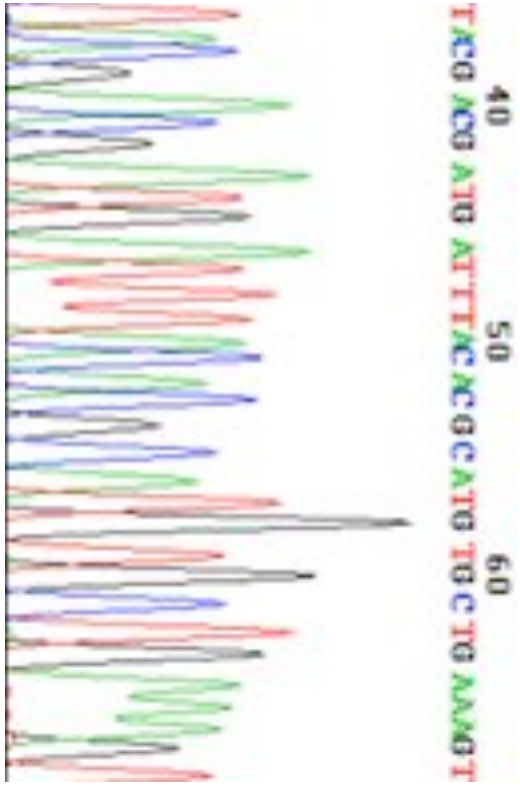
http://www.answers.com/topic/automated-sequencer

Milestones in DNA Sequencing

1970 | 1980 | 1990 | 2000 | 2010

1987
Applied Biosystems markets the ABI 370 as the first automated sequencing machine

http://commons.wikimedia.org/wiki/File:370A_automated_DNA_sequencer.jpg



Fluorescent Dye Termination
350bp / lane x 16 lanes =
5600bp / day / machine

http://www.answers.com/topic/automated-sequencer

Milestones in DNA Sequencing

1970 | 1980 | 1990 | 2000 | 2010

1995
Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp

2000
Myers *et al.*
1st Large WGS Assembly.
Celera Assembler. 116 Mbp

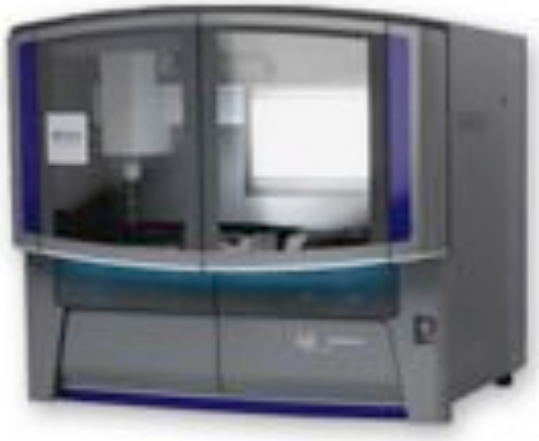
2001
Venter *et al.*,
Human Genome
Celera Assembler. 2.9 Gbp

ABI 3700: 500 bp reads x 768 samples / day = 384,000 bp / day.
"The machine was so revolutionary that it could decode in a single day the same amount of genetic material that most DNA labs could produce in a year." J. Craig Venter

Milestones in DNA Sequencing

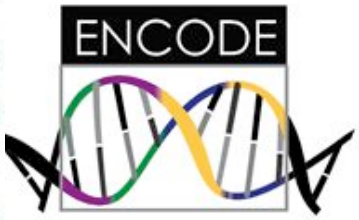
2004
454/Roche
Pyrosequencing
Current Specs (Titanium):
1M 400bp reads / run =
1Gbp / day

2007
Illumina
Sequencing by Synthesis
Current Specs (HiSeq 2000):
2.5B 100bp reads / run =
25Gbp / day

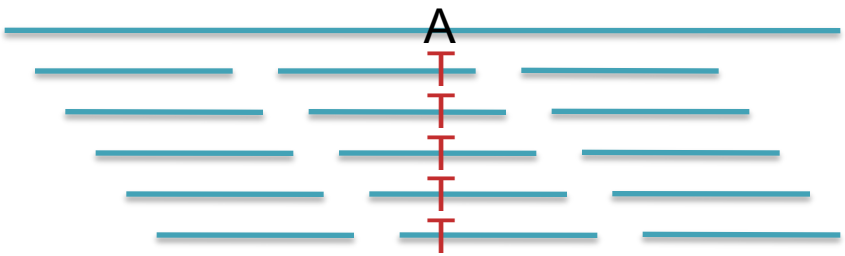


2008
ABI / Life Technologies
SOLiD Sequencing
Current Specs (5500xl):
5B 75bp reads / run =
30Gbp / day

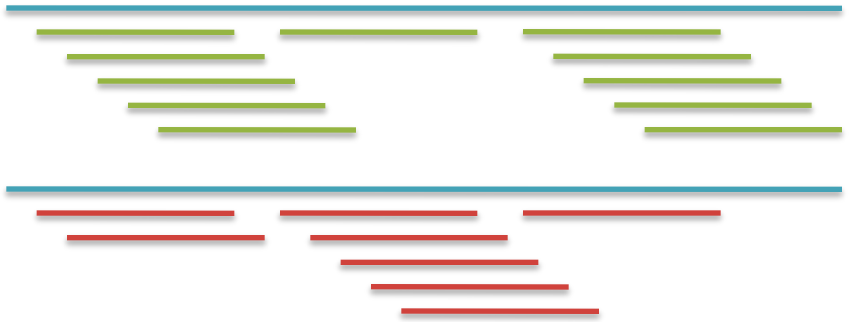
Second Generation Sequencing Applications



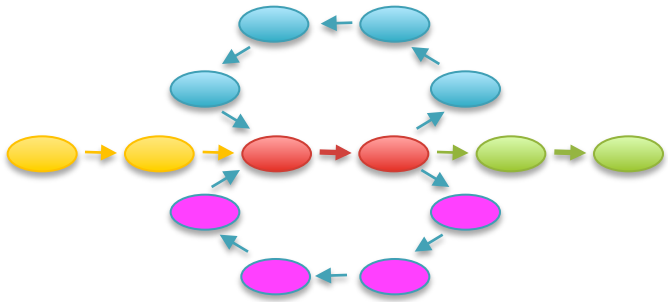
Alignment & Variations



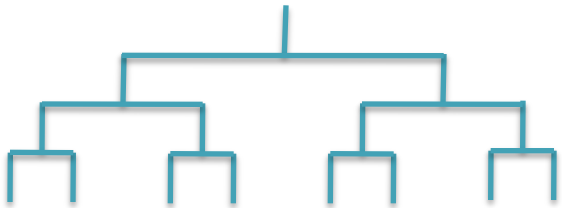
Differential Analysis



De novo Assembly

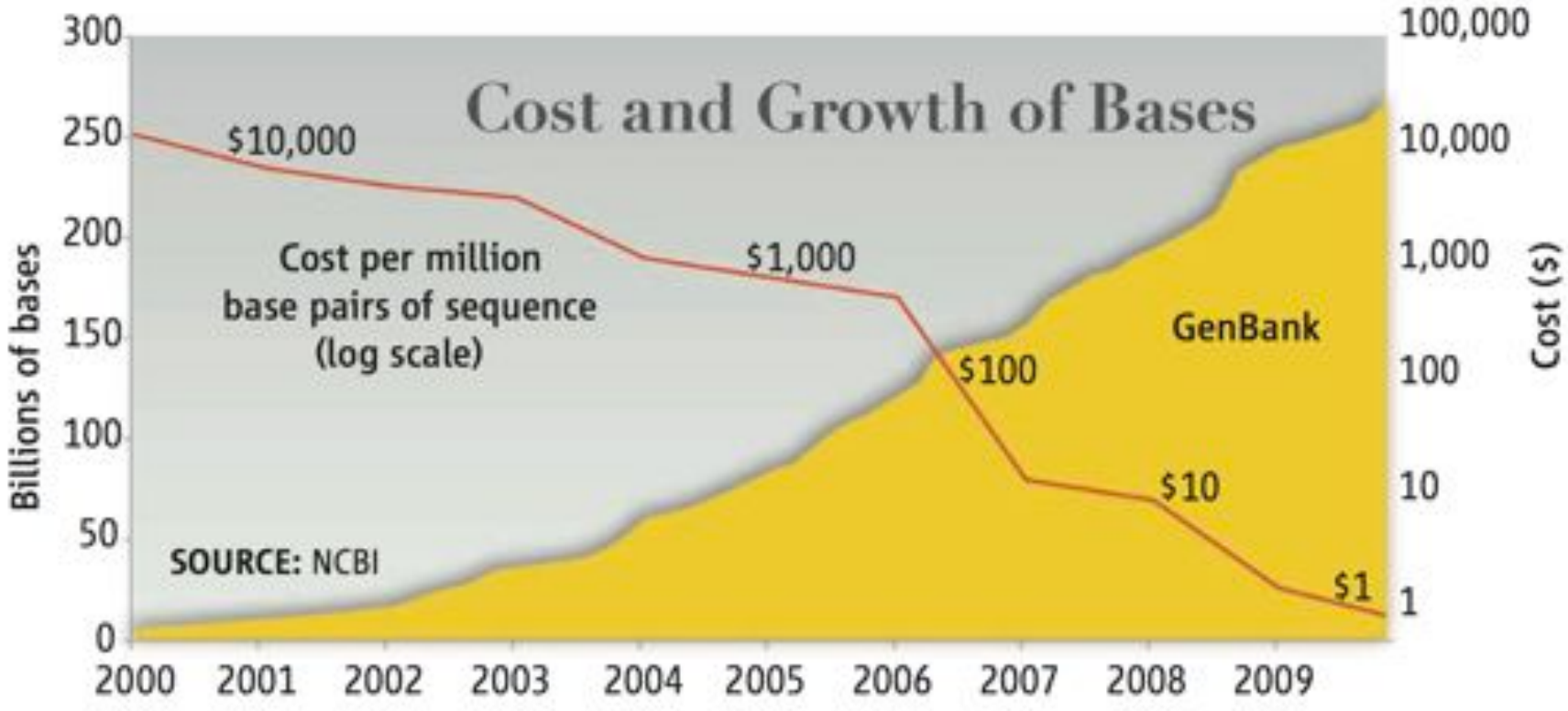


Phylogeny & Evolution



The DNA Data Tsunami

Current world-wide sequencing capacity exceeds 10Tbp/day (3.6Pbp/year) and is growing at 5x per year!



"*Will Computers Crash Genomics?*"
Elizabeth Pennisi (2011) *Science*. 331(6018): 666-668.

Hadoop MapReduce

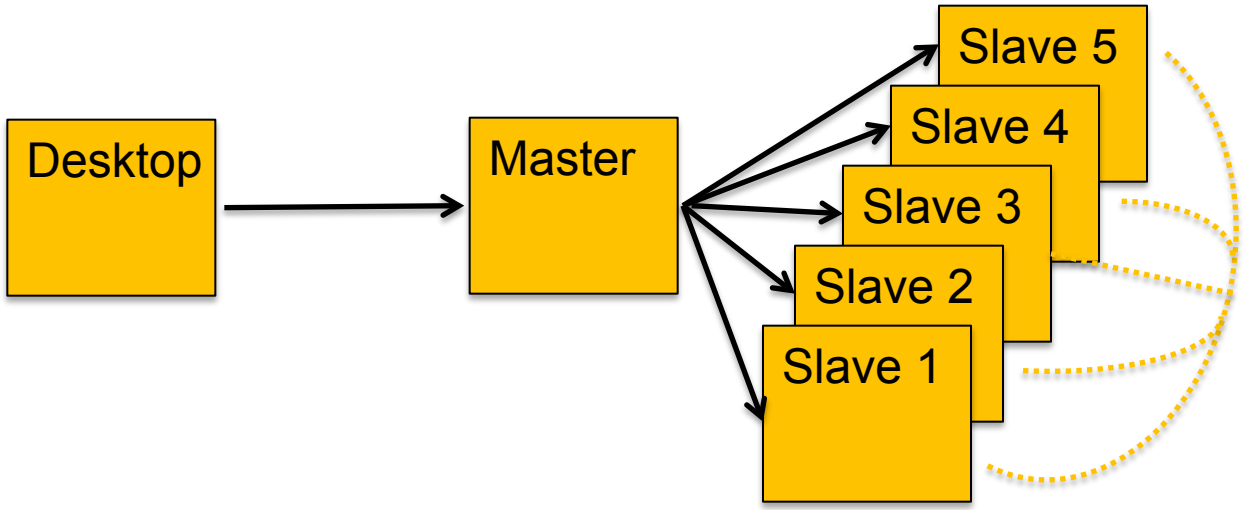
http://hadoop.apache.org

- MapReduce is Google's framework for large data computations
  - Data and computations are spread over thousands of computers
    - Indexing the Internet, PageRank, Machine Learning, etc… (Dean and Ghemawat, 2004)
    - 946,460 TB processed in May 2010 (Jeff Dean at Stanford, 11.10.2010)
  - Hadoop is the leading open source implementation
    - Developed and used by Yahoo, Facebook, Twitter, Amazon, etc
    - 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

System 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

Amazon Web Services

http://aws.amazon.com

- All you need is a credit card, and you can immediately start using one of the largest datacenters in the world

- Elastic Compute Cloud (EC2)
  - On demand computing power
- Simple Storage Service (S3)
  - Scalable data storage
- Plus many, many more

EC2 Architecture

- Very large cluster of machines
  - Effectively infinite resources
  - High-end servers with many cores and many GB RAM

- Machines run in a virtualized environment
  - Amazon can subdivide large nodes into smaller instances
  - You are 100% protected from other users on the machine
  - You get to pick the operating system, all installed software

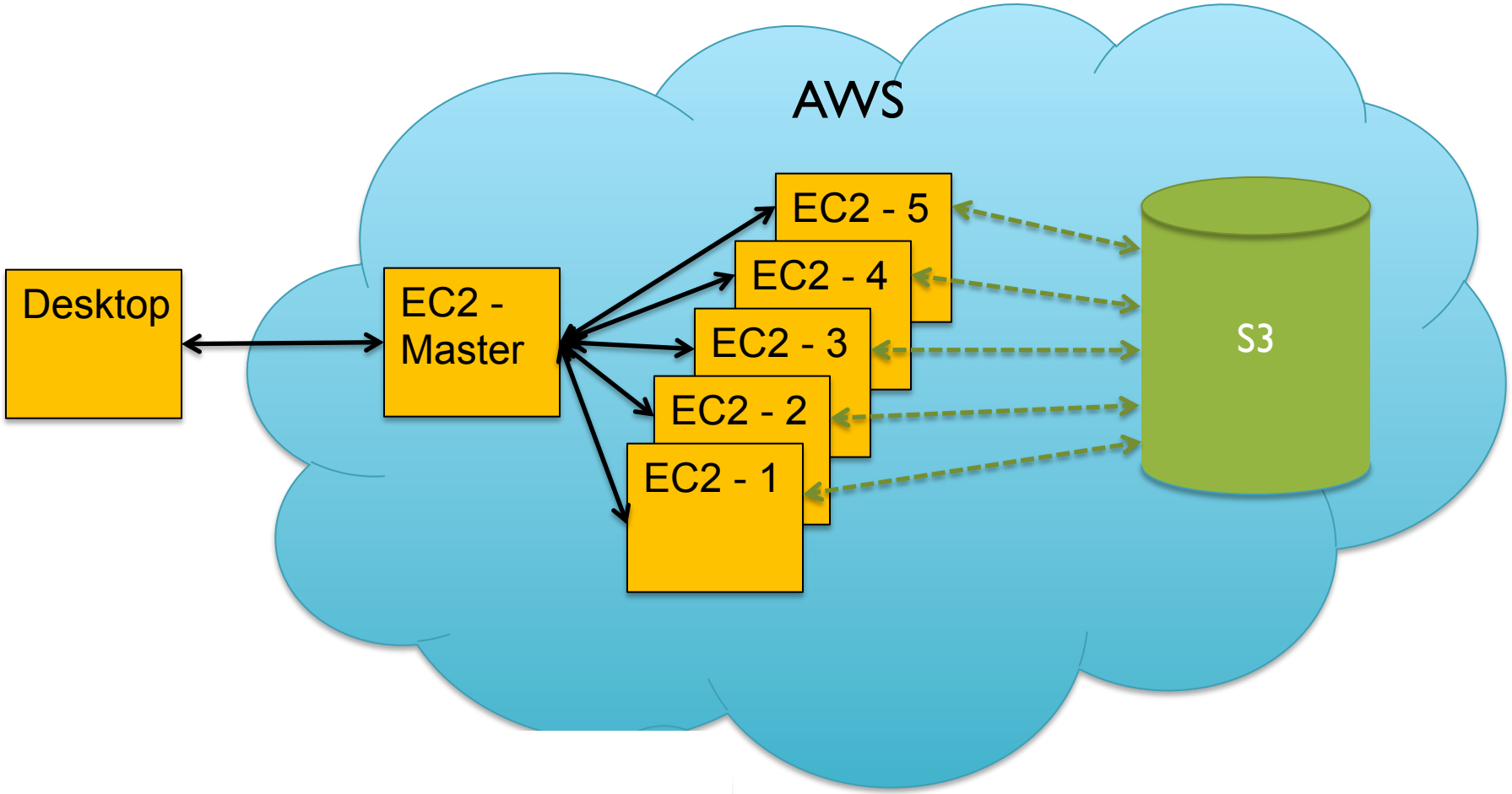
Amazon S3

- S3 provides persistent storage for large volumes of data
  - Very high speed connection from S3 to EC2 compute nodes
  - Public data sets include s3://1000genomes

– Tiered pricing by volume
  - Pricing starts at 15¢ / GB / month
  - 5.5¢ / GB / month for over 5 PB
  - Pay for transfer in and out of Amazon

– Import/Export service for large volumes
  - FedEx your drives to Amazon

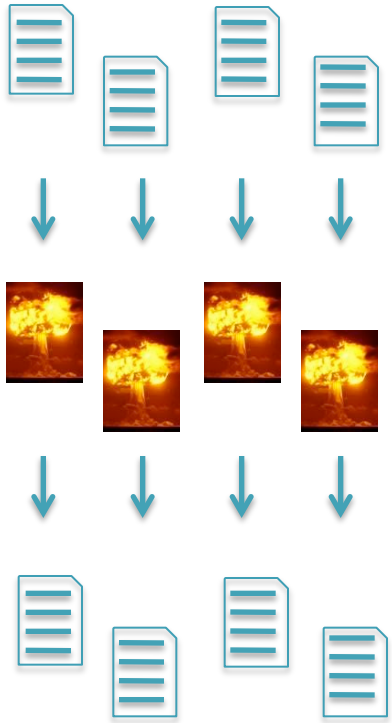
Hadoop on AWS



- If you don't have 1000s of machines, rent them from Amazon
  - After machines spool up, ssh to master as if it was a local machine.
  - Use S3 for persistent data storage, with very fast interconnect to EC2.

Programming Models

Embarrassingly Parallel

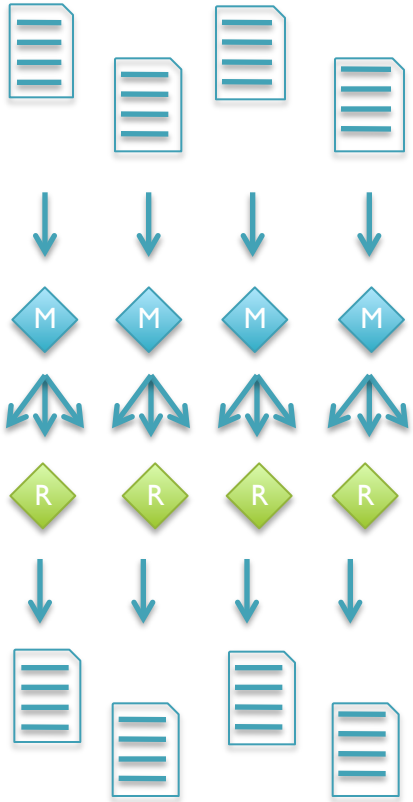


Map-only

Each item is Independent

Traditional Batch Computing

Loosely Coupled

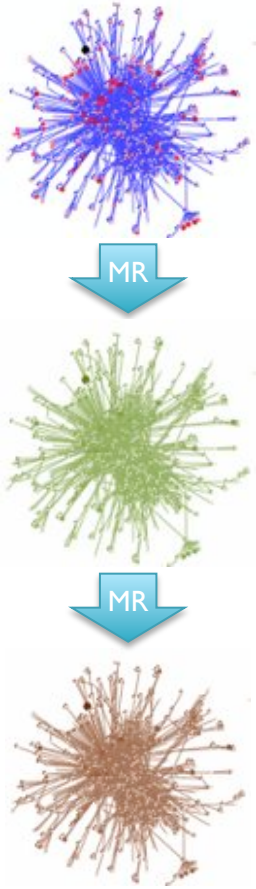


MapReduce

Independent-Shuffle-Independent

Batch Computing + Data Exchange

Tightly Coupled



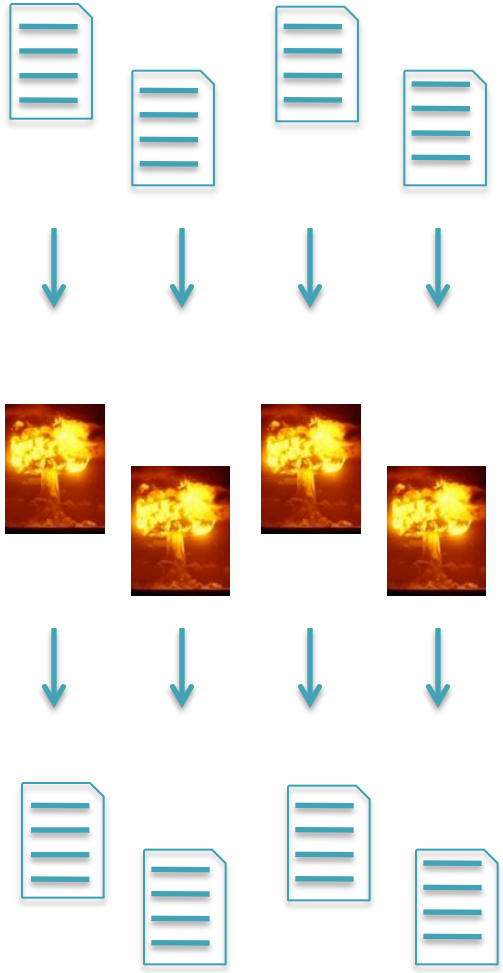
Iterative MapReduce

Nodes interact with other nodes

Big Data MPI

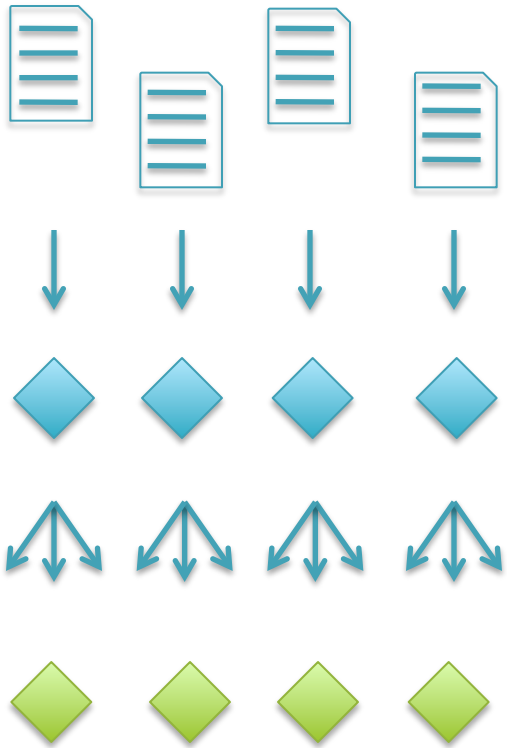
1. Embarrassingly Parallel

- 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



2. 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



Short Read Mapping

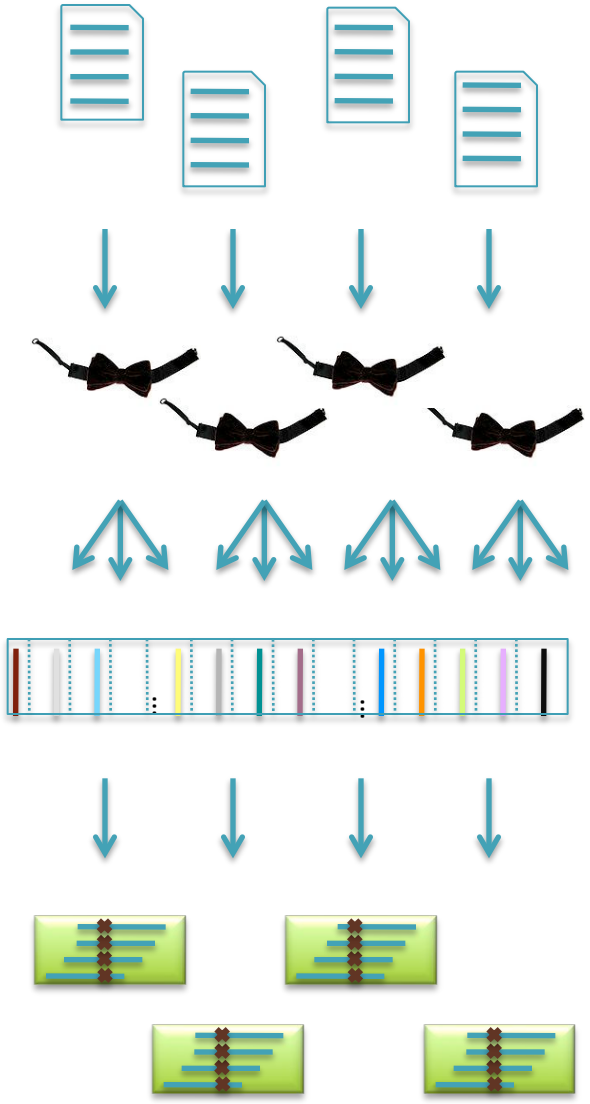
- Given a reference and many subject reads, report one or more "good" end-to-end alignments per alignable read
  - Find where the read most likely originated
  - Fundamental computation for many assays
    - Genotyping
    - Structural Variations
    - RNA-Seq
    - Chip-Seq
    - Methyl-Seq
    - Hi-C-Seq

- 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

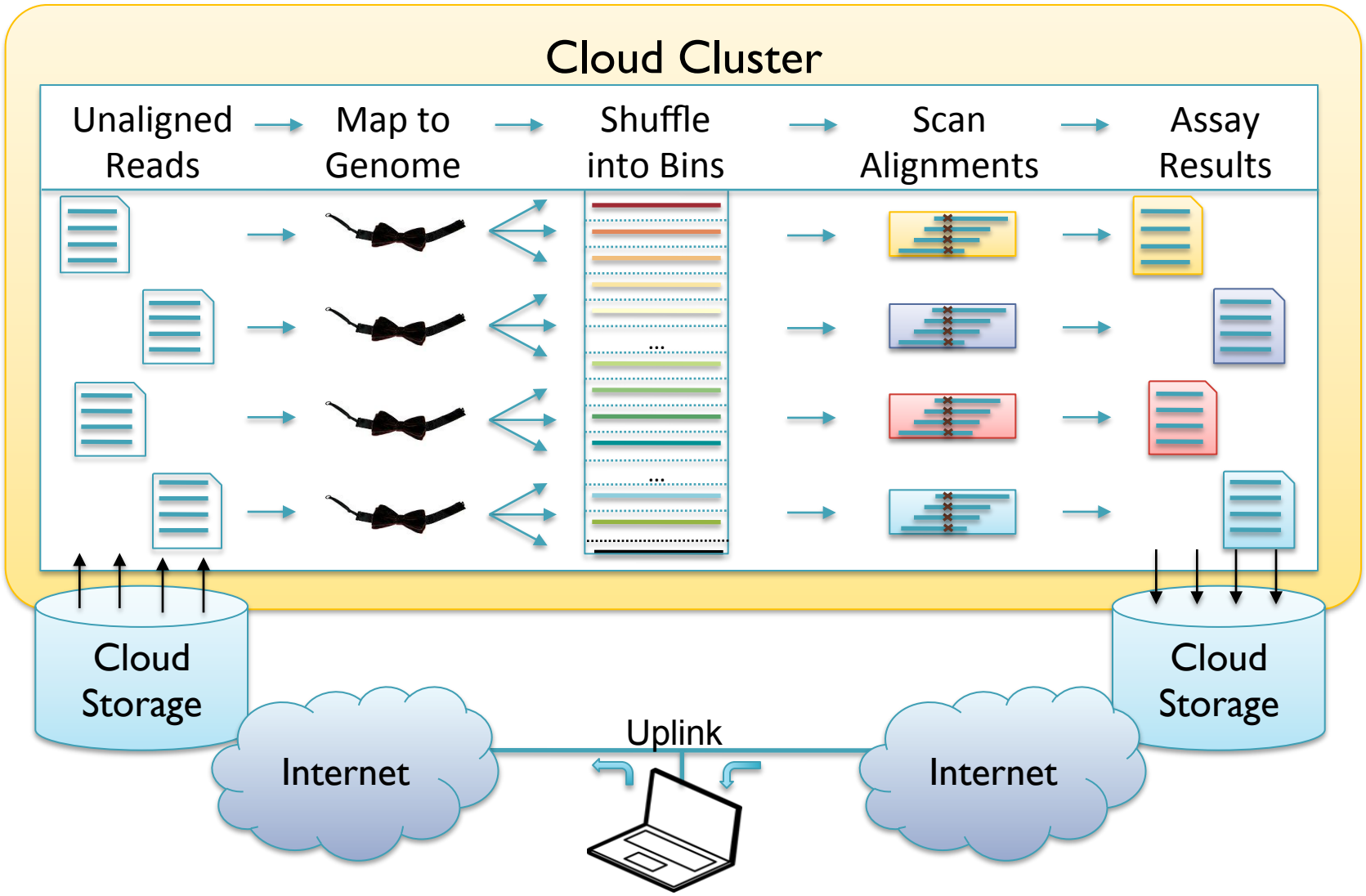
http://bowtie-bio.sourceforge.net/crossbow

	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

Discovered 3.7M SNPs in one 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

Map-Shuffle-Scan for Genomics



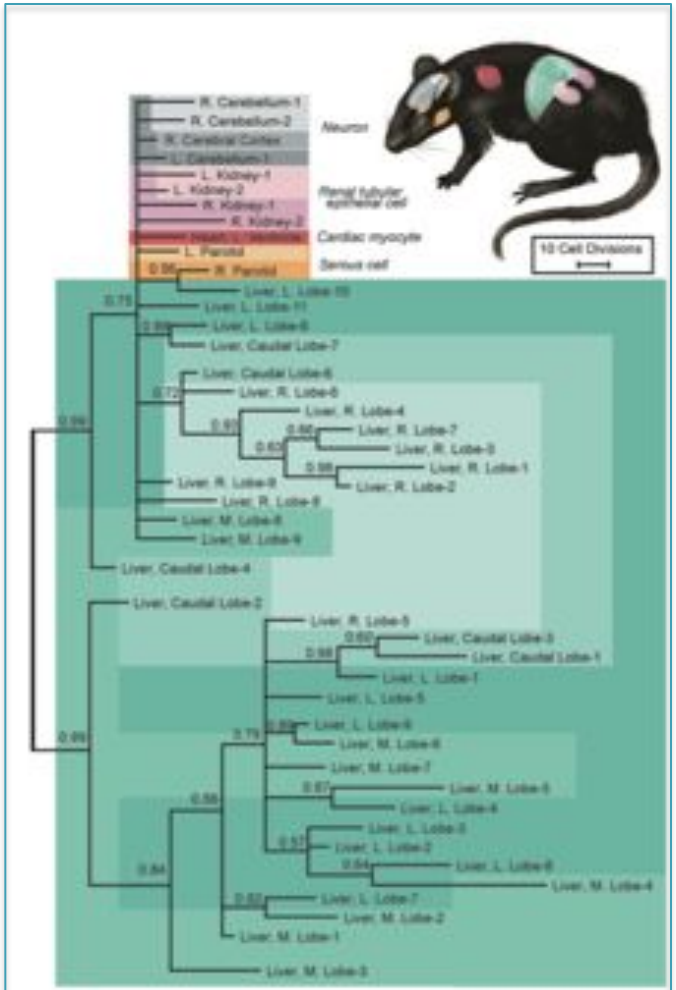
Cloud Computing and the DNA Data Race.
Schatz, MC, Langmead B, Salzberg SL (2010) *Nature Biotechnology.* **28**:691-693

MicroSeq: NextGen Microsatellite Profiling

Mitchell Bekritsky, WSBS

- Class of simple sequence repeats
  - …GCACACACACAT… = …G(CA)$_5$T…
  - Created and mutate primarily through slippage during replication
  - Highly variable & ubiquitous

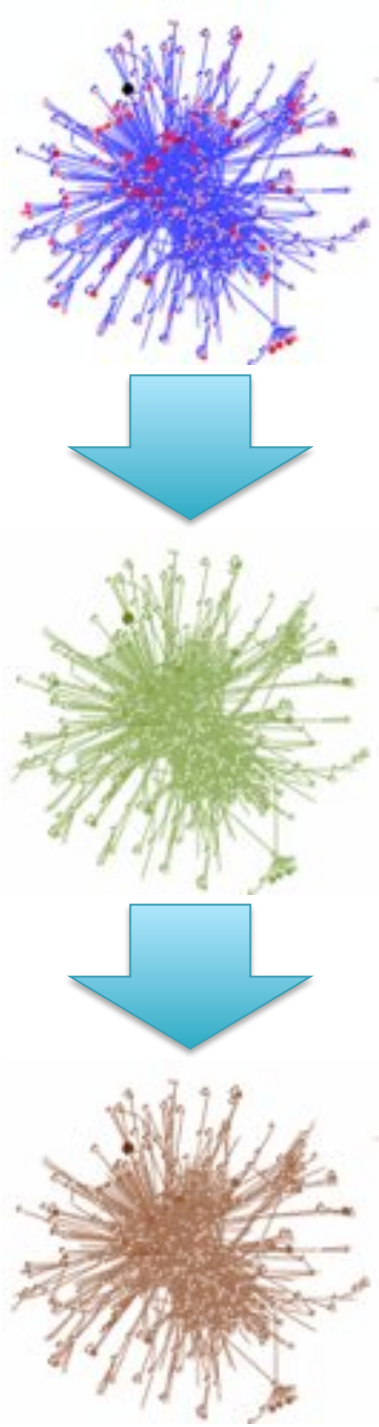
- Genotyping with MicroSeq
  - Map reads using a new MS-mapper
  - Collect MS-reads into MS-genotypes
  - Analyze profiles in cells, across cells, & across populations
    - Loss of heterozygosity
    - Development of somatic & cancer cells
    - Relations across strains, across species
    - etc…



(Salipante *et al.* 2006)

3. 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

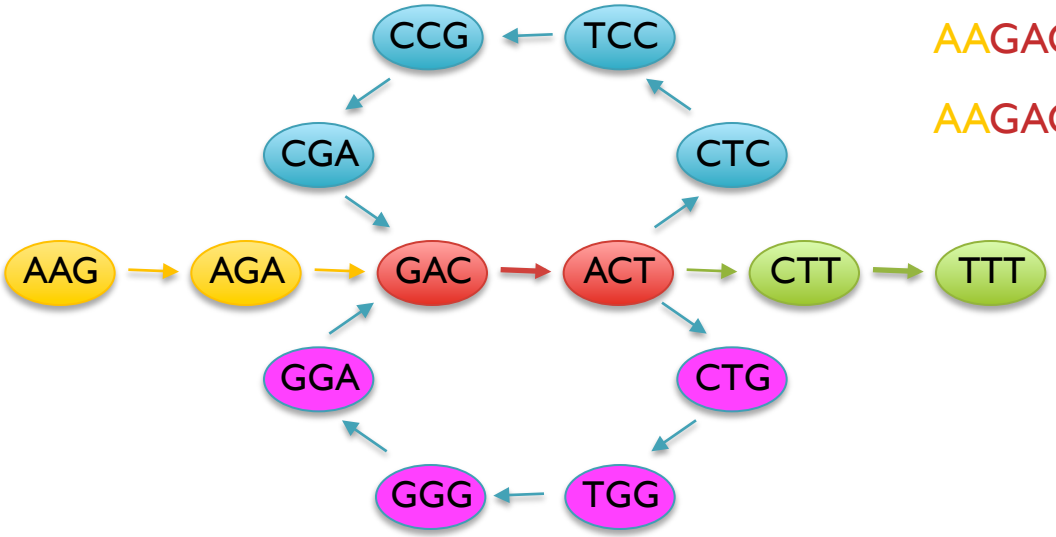


Short Read Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…

de Bruijn Graph



Potential Genomes

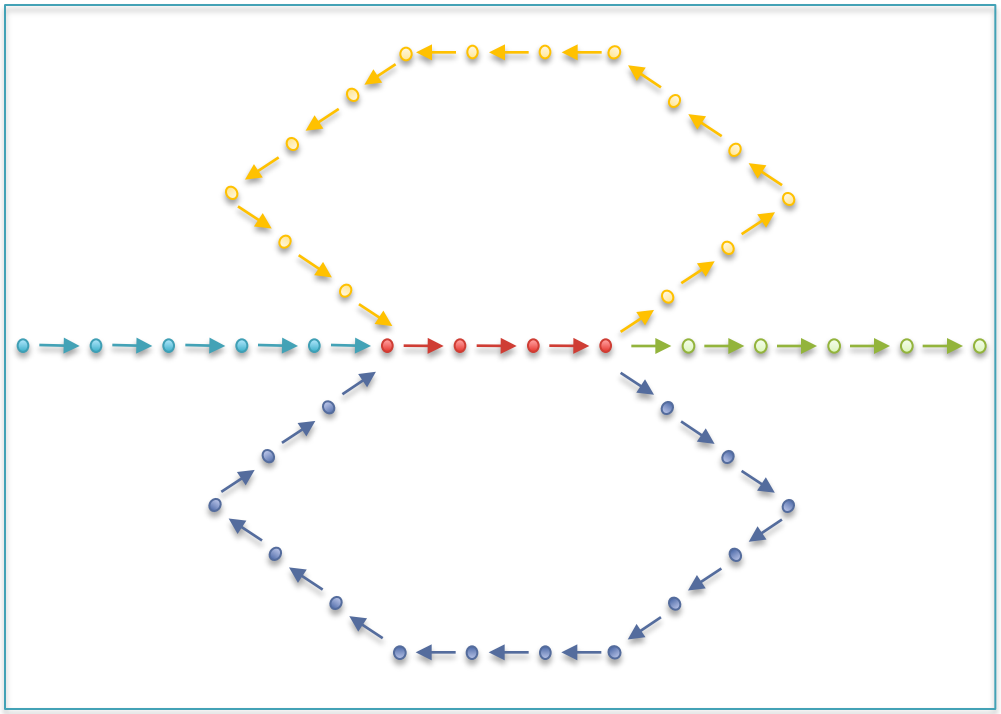
AAGACTCCGACTGGGACTTT

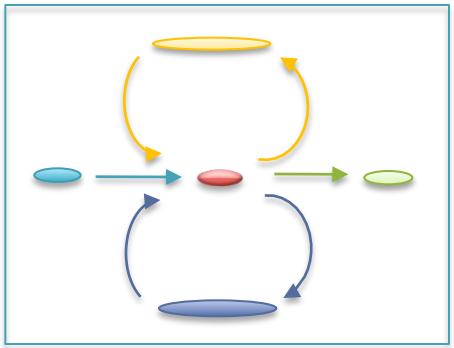
AAGACTGGGACTCCGACTTT

- 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
  - 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

- After construction, many edges are unambiguous
  - Merge together compressible nodes
  - Graph physically distributed over hundreds of computers





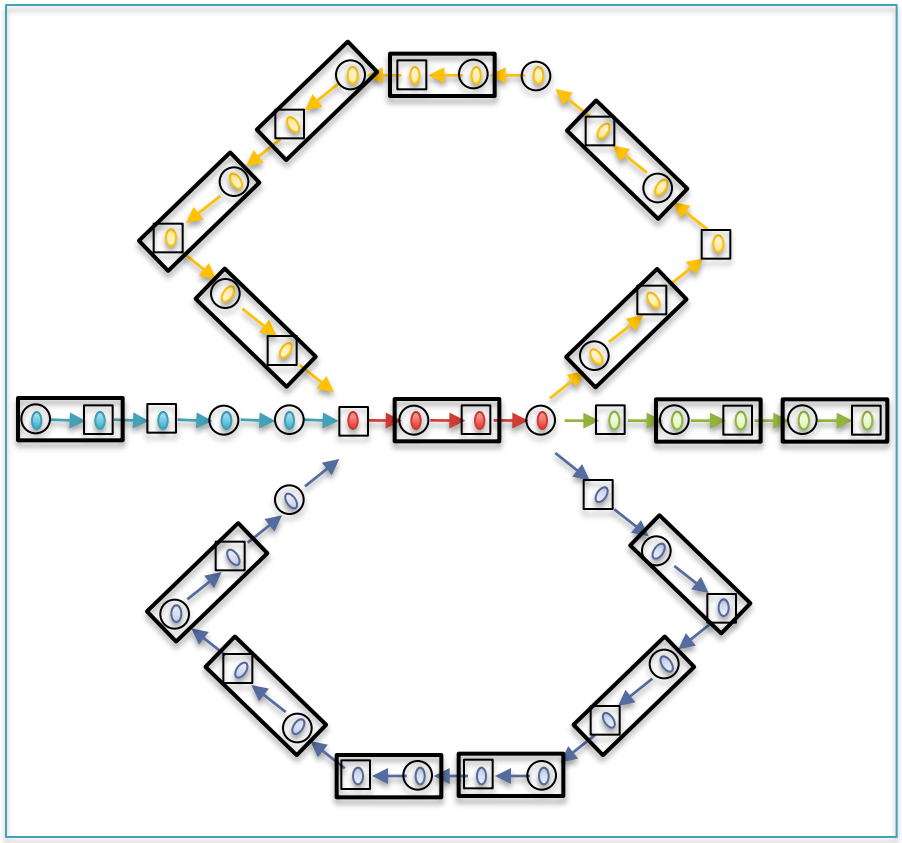
Fast Path Compression

Challenges

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

Randomized List Ranking

- Randomly assign (H) / [T] to each compressible node
- Compress (H) → [T] links



Initial Graph: 42 nodes

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

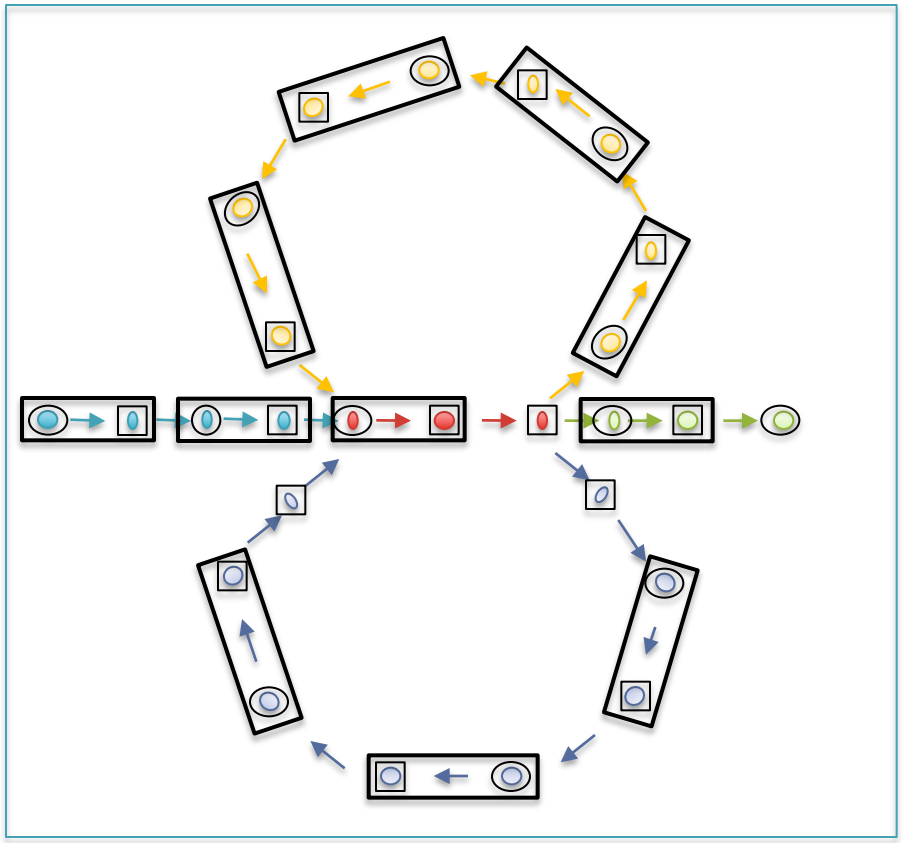
Fast Path Compression

Challenges

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

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ → [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.*

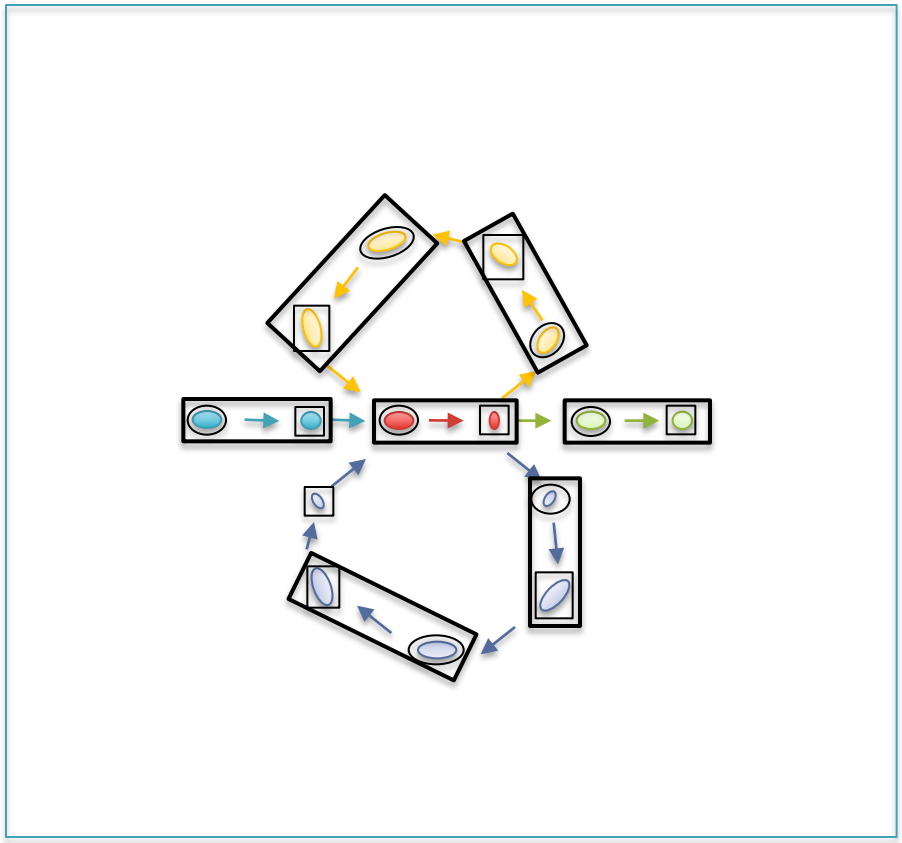
Fast Path Compression

Challenges

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

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ → [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.*

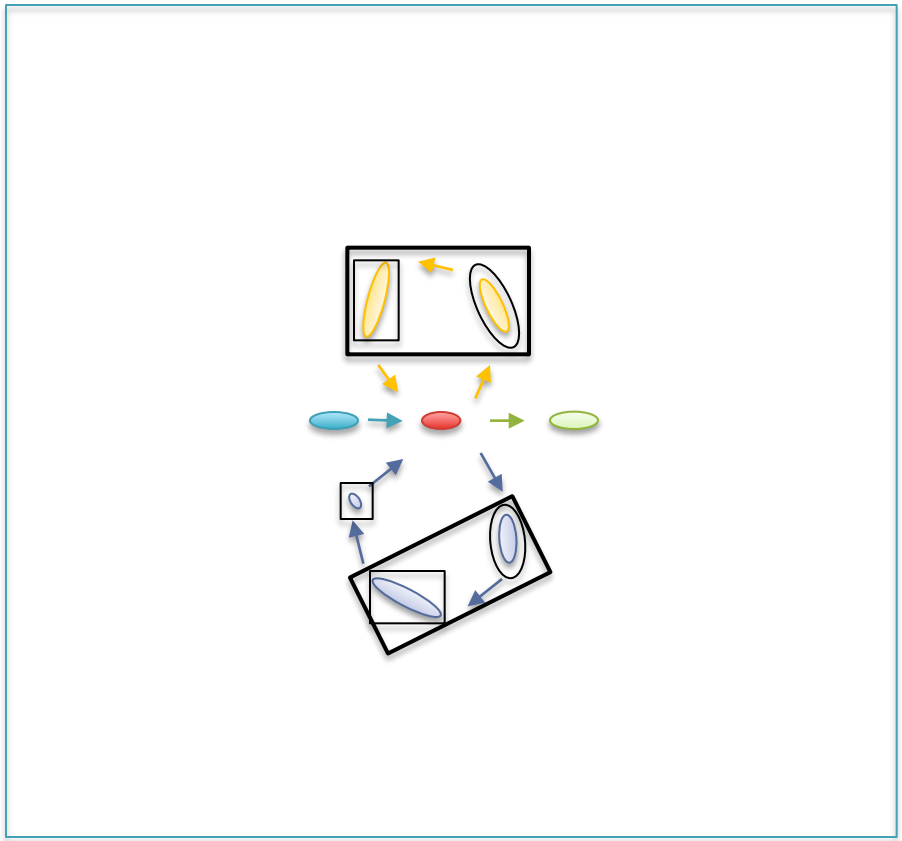
Fast Path Compression

Challenges

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

Randomized List Ranking

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



Round 2: 8 nodes (81% savings)

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

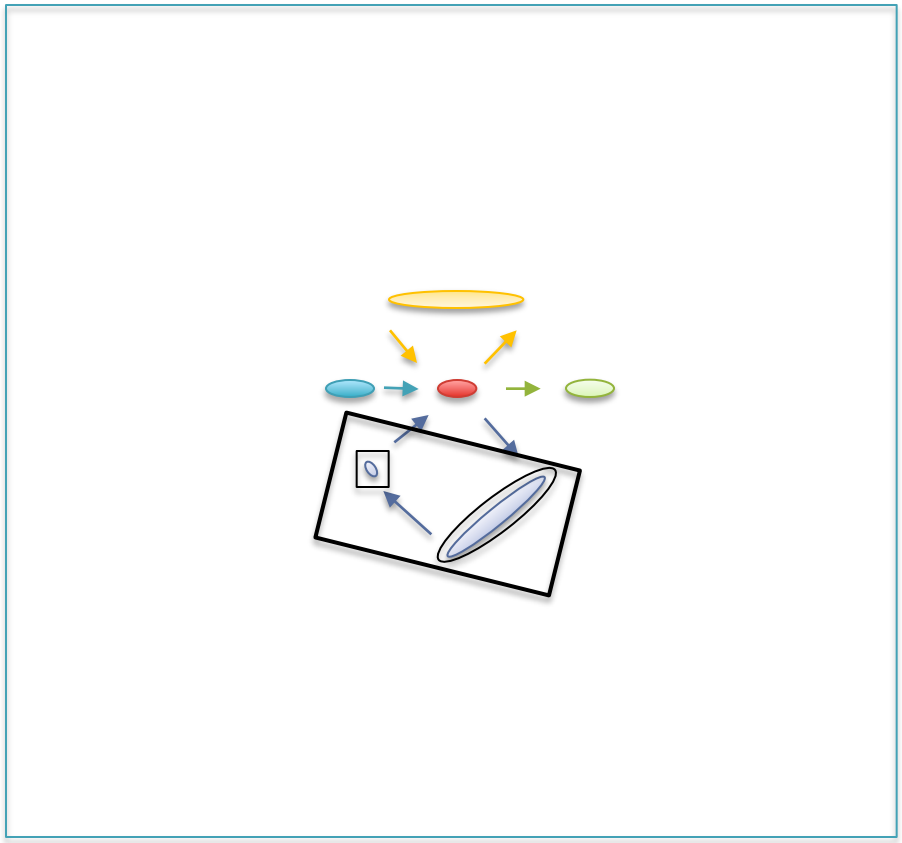
Fast Path Compression

Challenges

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

Randomized List Ranking

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

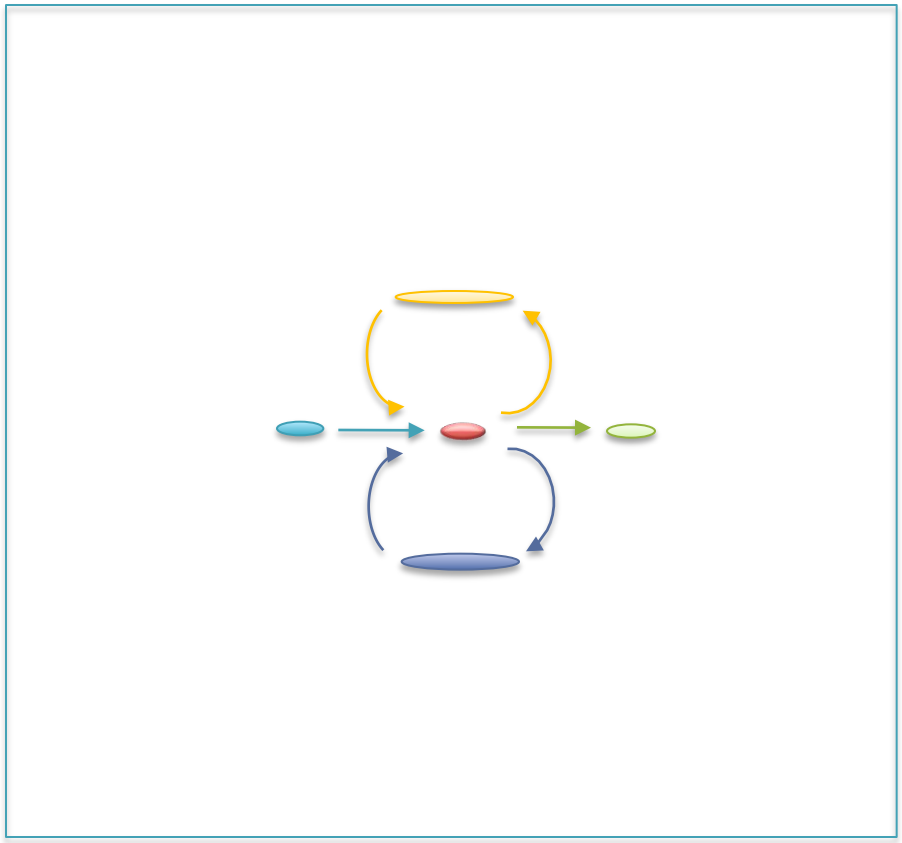
- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

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

Performance

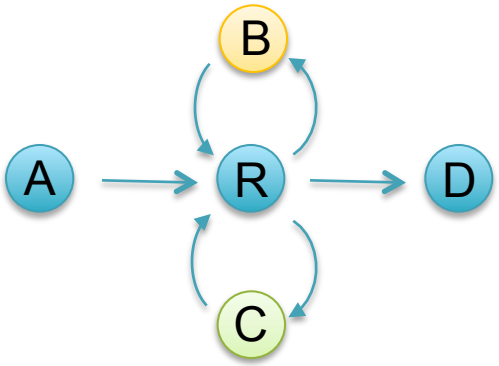
- Compress all chains in log(S) rounds



Round 4: 5 nodes (88% savings)

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

Counting Eulerian Tours



ARBRCRD
or
ARCRBRD

Generally an exponential number of compatible sequences

– Value computed by application of the BEST theorem (Hutchinson, 1975)

$$\mathcal{W}(G,t) = (\det L)\left\{\prod_{u\in V}(r_u-1)!\right\}\left\{\prod_{(u,v)\in E} a_{uv}!\right\}^{-1}$$

L = *n x n* matrix with r_u-a_{uu} along the diagonal and -a_{uv} in entry uv

r_u = $d^+(u)+1$ if $u=t$, or $d^+(u)$ otherwise

a_{uv} = multiplicity of edge from *u* to *v*

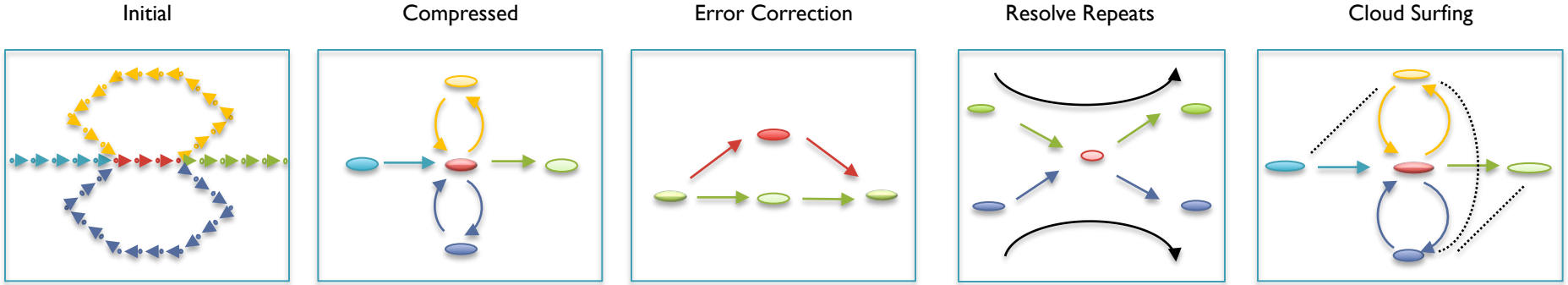
Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.

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*: Quake Error Correction



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

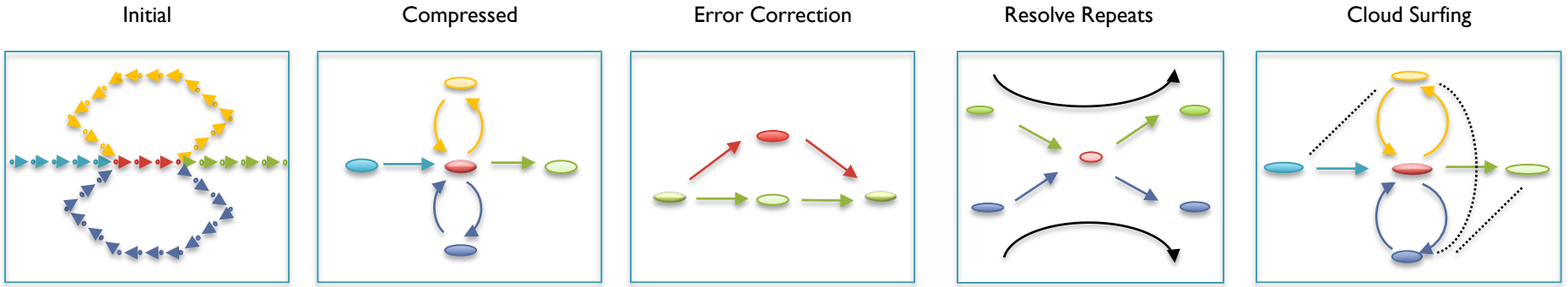
Assembly of Large Genomes with Cloud Computing.
Schatz MC, Sommer D, Kelley D, Pop M, *et al. In Preparation.*

Contrail

http://contrail-bio.sourceforge.net

De novo Assembly of the Human Genome

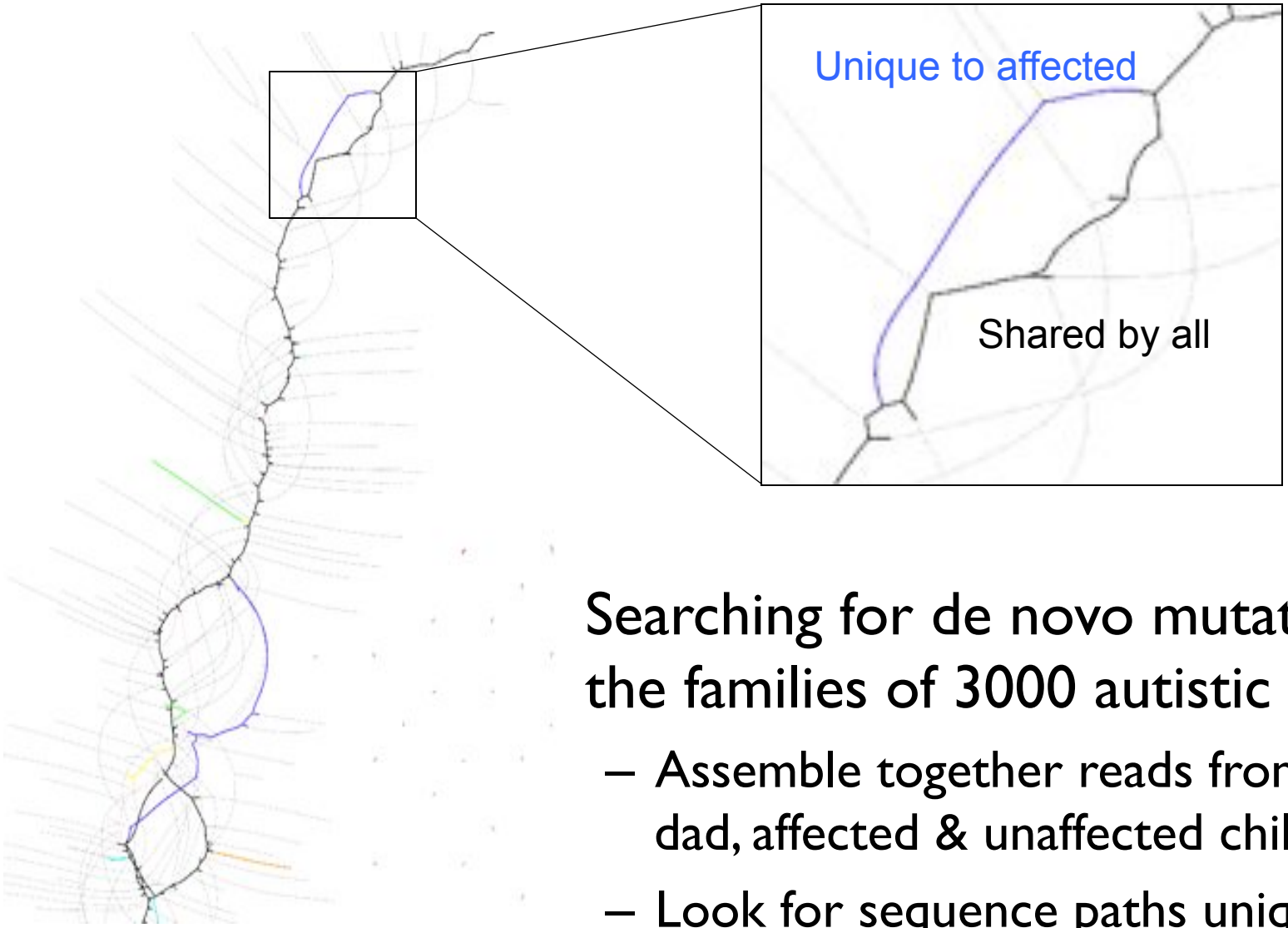
- *Genome:* African male NA18507 (SRA000271, Bentley *et al.*, 2008)
- *Input: 3.5*B 36bp reads, 210bp insert (~40x coverage)



N	>7 B	>1 B	4.2 M	4.1 M	3.3 M
Max	27 bp	303 bp	20,594 bp	20,594 bp	20,594 bp
N50	27 bp	< 100 bp	995 bp	1,050 bp	1,427 bp*

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

De novo mutations and de Bruijn Graphs



Searching for de novo mutations in the families of 3000 autistic children.

- Assemble together reads from mom, dad, affected & unaffected children
- Look for sequence paths unique to affected child

Hadoop for NGS Analysis

CloudBurst

Highly Sensitive Short Read Mapping with MapReduce

100x speedup mapping on 96 cores @ Amazon

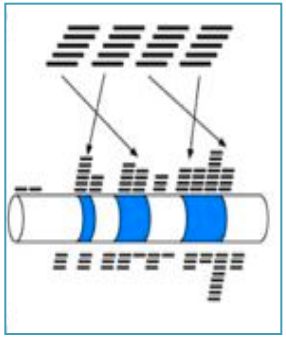
http://cloudburst-bio.sf.net

(Schatz, 2009)

Myrna

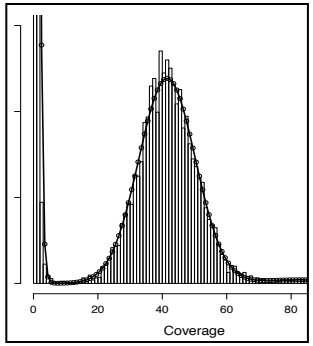
Cloud-scale differential gene expression for RNA-seq

Expression of 1.1 billion RNA-Seq reads in ~2 hours for ~$66



(Langmead, Hansen, Leek, 2010)

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



Quake

Quality-aware error correction of short reads

Correct 97.9% of errors with 99.9% accuracy

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

(Kelley, Schatz, Salzberg, 2010)

Genome Indexing

Rapid Parallel Construction of Genome Index

Construct the BWT of the human genome in 9 minutes

(Menom, Bhat, Schatz, 2011*)

http://code.google.com/p/genome-indexing/

Research Directions

- Scalable Sequencing
  - Genomes, Metagenomes, *-Seq, Personalized Medicine
  - How do we survive the tsunami of sequence data?
    - Improved indexing & algorithms, multi-core & multi-disk systems

- Practically Parallel
  - Managing n-tier memory hierarchies, crossing the PRAM chasm
  - How do we solve problems with 1000s of cores?
    - Locality, Fault Tolerance, Programming Languages & Parallel Systems

- Computational Discovery
  - Abundant data and computation are necessary, but not sufficient
  - How do we gain insight?
    - Statistics & Modeling, Machine Learning, Databases, Visualization & HCI

Summary

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

Acknowledgements

CSHL
Mike Wigler
Zach Lippman
Dick McCombie
Doreen Ware
Mitch Bekritsky

SBU
Steve Skiena
Matt Titmus
Rohith Menon
Goutham Bhat
Hayan Lee

JHU
Ben Langmead
Jeff Leek

Univ. of Maryland
Steven Salzberg
Mihai Pop
Art Delcher
Jimmy Lin
Adam Phillippy
David Kelley
Dan Sommer

Thank You!

http://schatzlab.cshl.edu
@mike_schatz