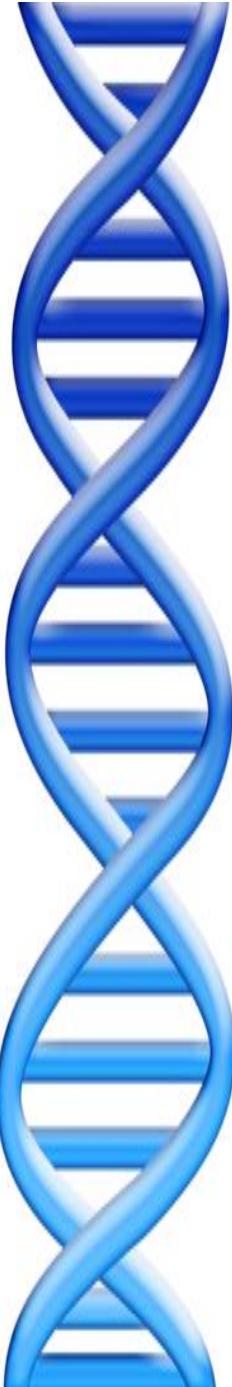


Biology in the Clouds

Michael Schatz

July 30, 2012
Science Cloud Summer School





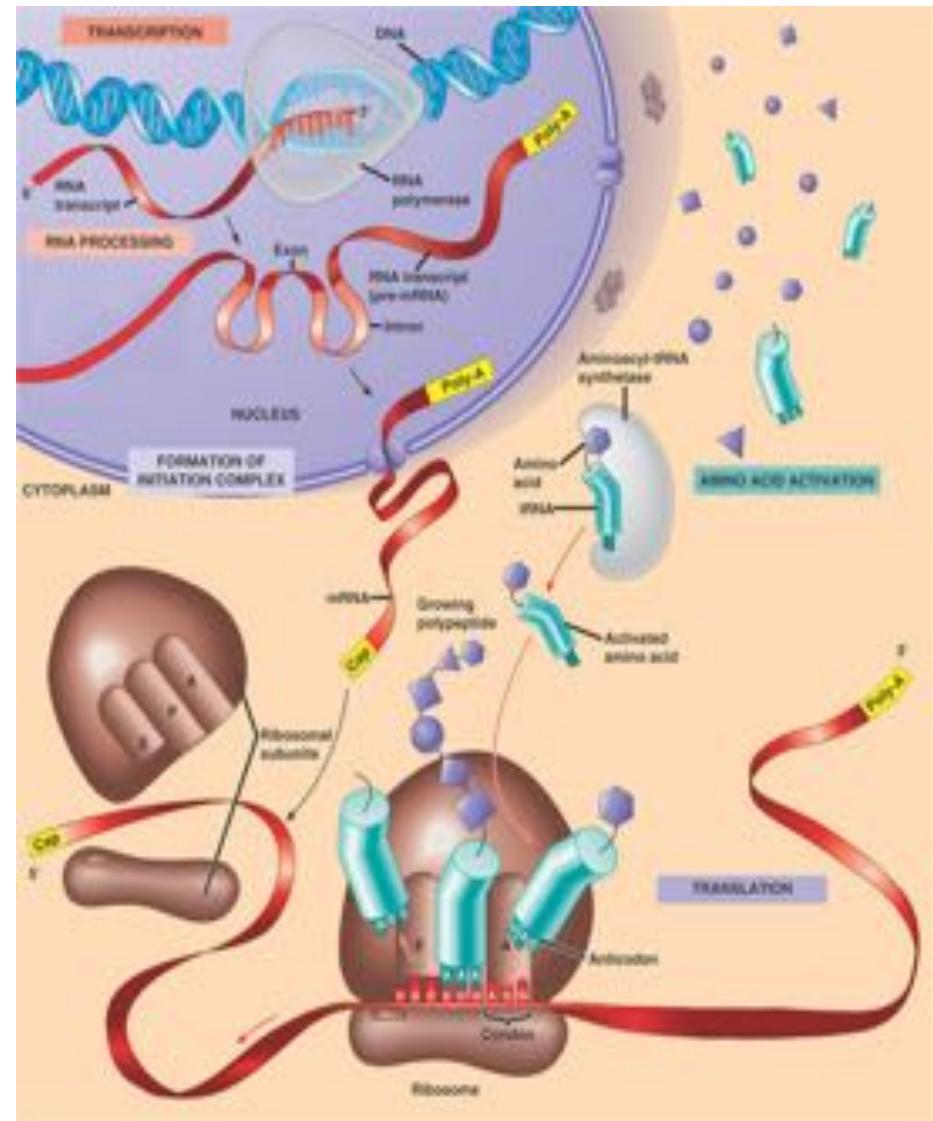
Outline

- I. Milestones in genomics
 - I. Quick primer on molecular biology
 2. The evolution of sequencing
2. Hadoop Applications for Genomics
 - I. Mapping & Jnomics
 2. Assembly & Contrail

Milestones in Genomics

In 1953 James Watson and Francis Crick determined the double helix structure of DNA as a long series of 4 different nucleotides. In 1958 Francis Crick established the **Central Dogma of Biology**:

1. Genetic information is transmitted from generation to generation by the sequence of nucleotides in your **DNA**.
2. Active regions called genes, are transcribed into **messenger RNAs** that are sent to cellular machines called ribosomes for processing
3. RNA messages are translated by the ribosomes into **proteins** that do work in the cell



Milestones in Genomics

Your genome and environment define who you are:

- Human with 5 fingers & 5 toes
- Hair, eye & skin color
- Susceptibility to diseases, responses to drugs
- Personality and social disorders
- ...

There is tremendous interest to sequence genomes:

- What is your genome sequence?
- How does your genome compare to my genome?
- Where are the genes and how active are they?
- How does gene activity change under development?
- Where do proteins bind and regulate genes?
- How has the disease mutated your genome?
- What virus and microbes are living inside you?
- What drugs should we give you?
- ...



Milestones in Genomics: Zeroth Generation Sequencing

Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ϕ X174 DNA

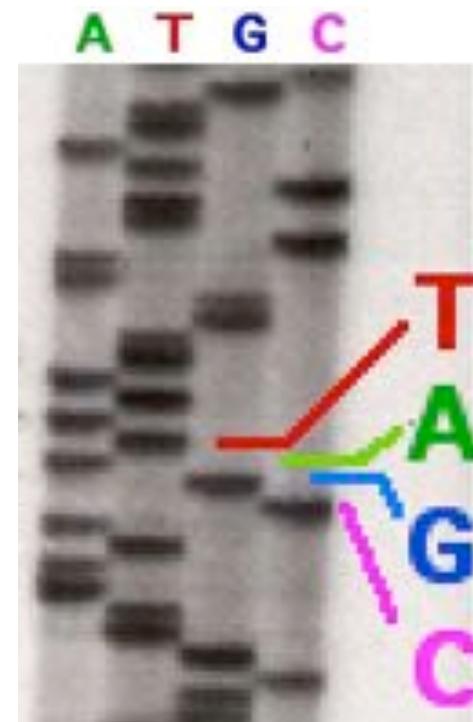
F. Sanger, G. M. Air¹, B. G. Barrell, N. L. Brown¹, A. R. Coulson, J. C. Fiddes,
C. A. Hutchison III¹, P. M. Slocombe² & M. Smith¹

¹MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage ϕ X174 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 in 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 ϕ X174 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 techniques³⁻⁵, 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.

1977
1st Complete Organism
Bacteriophage ϕ X174
5375 bp

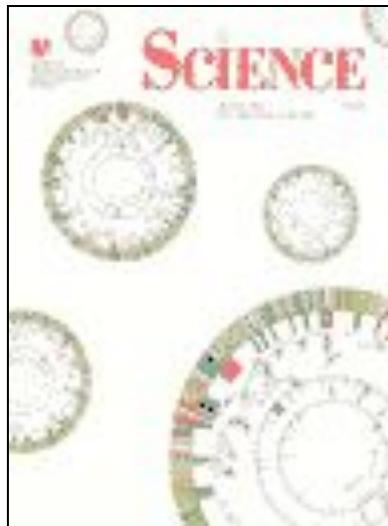


Radioactive Chain Termination
5000bp / week / person

Nucleotide sequence of bacteriophage ϕ X174 DNA
Sanger, F. et al. (1977) Nature. 265: 687 - 695

<http://en.wikipedia.org/wiki/File:Sequencing.jpg>
<http://www.answers.com/topic/automated-sequencer>

Milestones in Genomics: First Generation Sequencing



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. / IHGSC
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 Genomics: Second Generation 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 =
60Gbp / day



2008
ABI / Life Technologies
SOLiD Sequencing

Current Specs (5500xl):
5B 75bp reads / run =
30Gbp / day

Milestones in Genomics: Third Generation Sequencing



2010
Ion Torrent
Postlight Sequencing

Current Specs (Ion 318):
11M 300bp reads / run =
>1Gbp / day

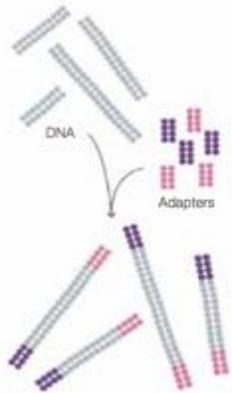
2011
Pacific Biosciences
SMRT Sequencing

Current Specs (RS):
50k 10kbp reads / run =
>500Mbp / day

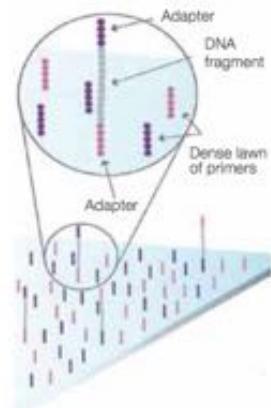
2012
Oxford Nanopore
Nanopore sensing

Many GB / day?
Very Long Reads?

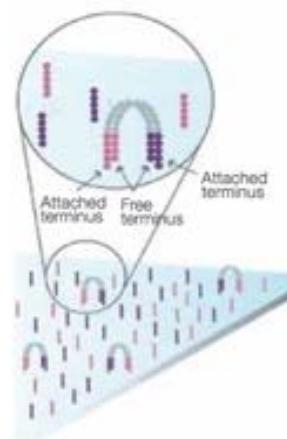
Illumina Sequencing by Synthesis



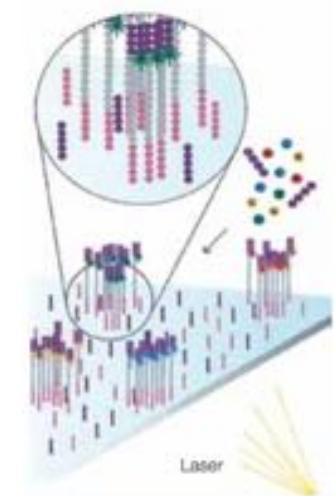
1. Prepare



2. Attach



3. Amplify



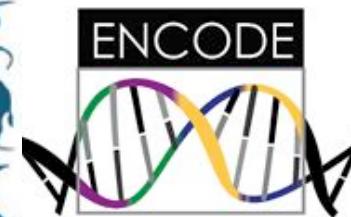
4. Image



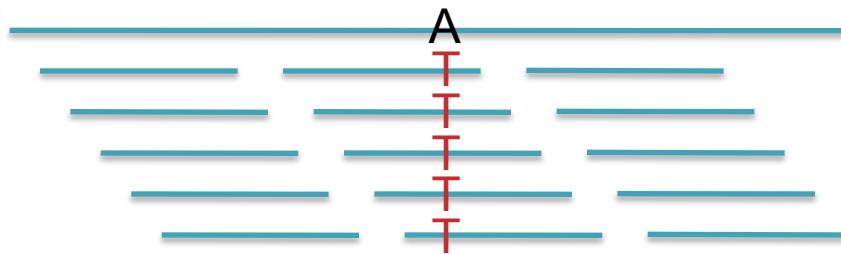
5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46
<http://www.youtube.com/watch?v=l99aKKHcxC4>

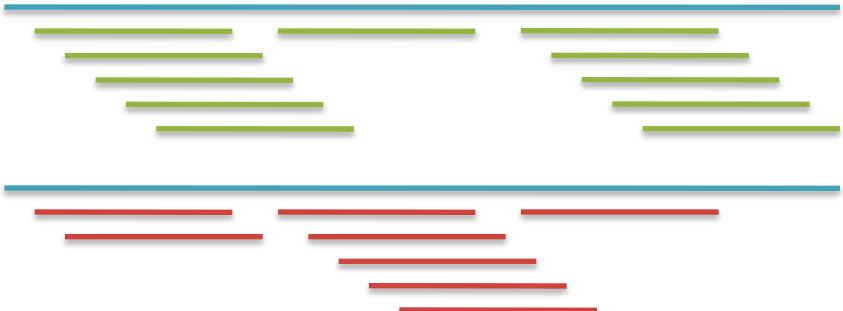
Milestones in Genomics



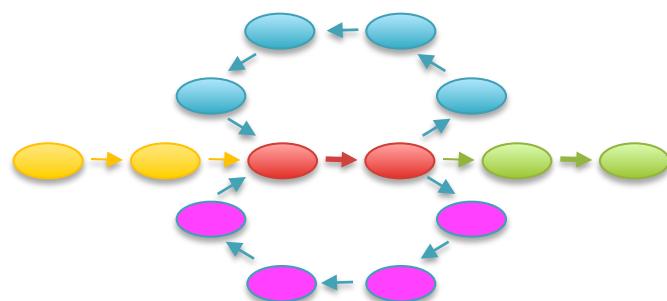
Alignment & Variations



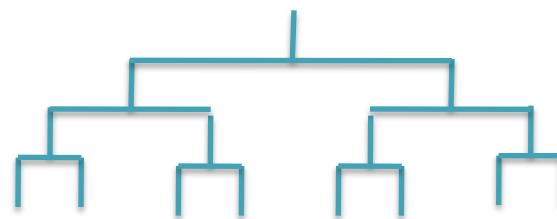
Differential Analysis



De novo Assembly

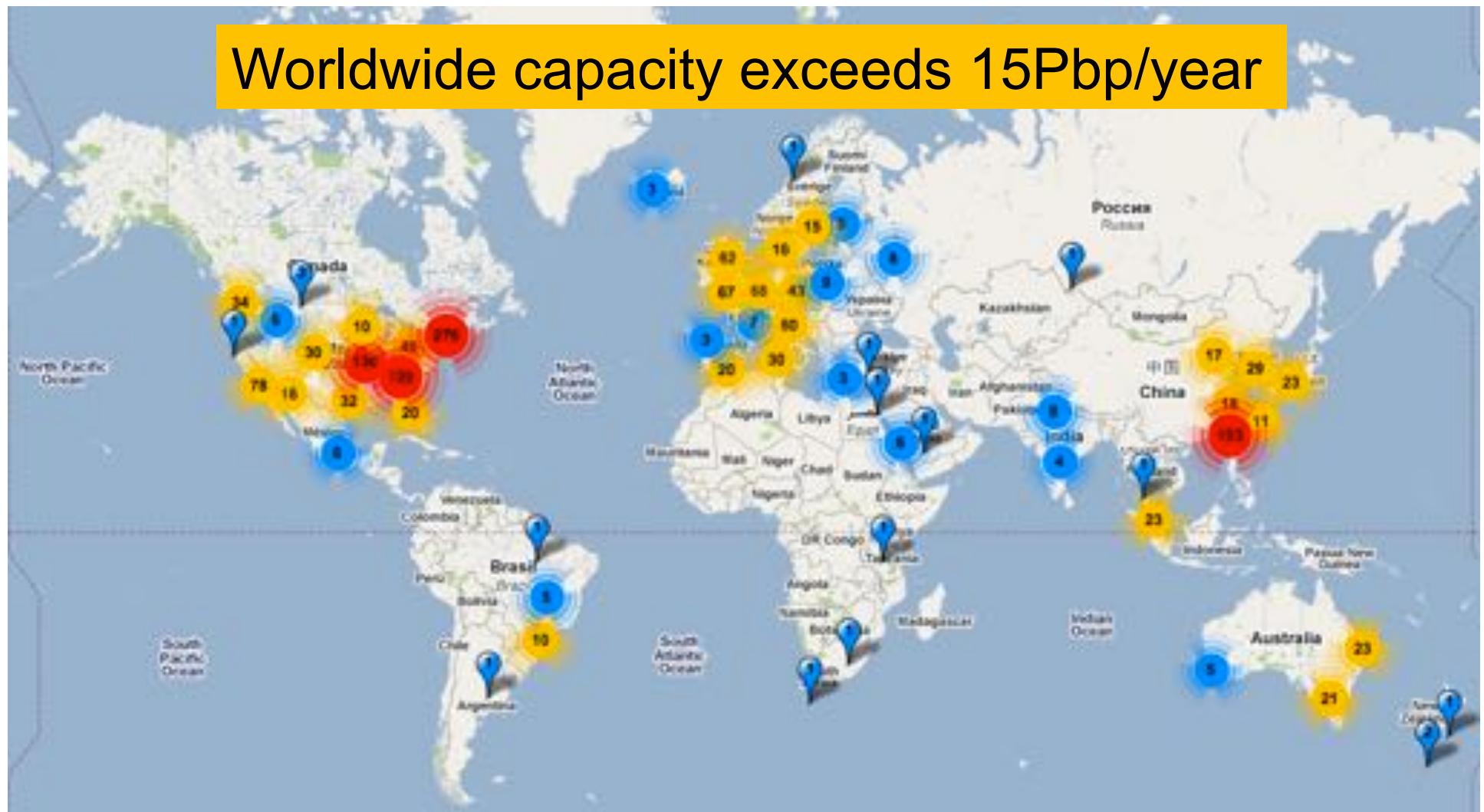


Phylogeny & Modeling



Sequencing Centers

Worldwide capacity exceeds 15Pbp/year



Next Generation Genomics: World Map of High-throughput Sequencers
<http://pathogenomics.bham.ac.uk/hts/>

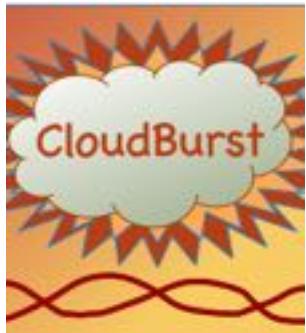
Hadoop MapReduce

<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)
 - 946PB 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



Hadoop for NGS Analysis



CloudBurst

Highly Sensitive Short Read Mapping with MapReduce

*100x speedup mapping
on 96 cores @ Amazon*

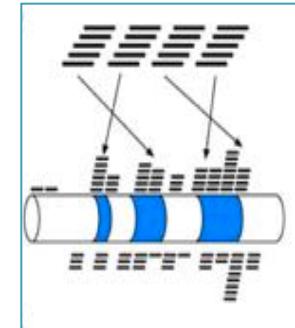
<http://cloudburst-bio.sf.net>

(Schatz, 2009)

Myrna

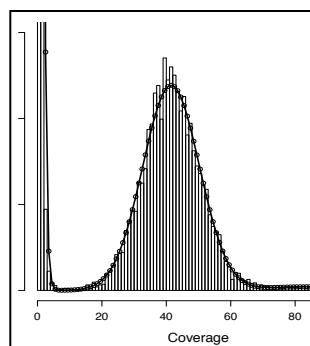
Cloud-scale differential gene expression for RNA-seq

*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.cbcu.umd.edu/software/quake/>

(Kelley, Schatz,
Salzberg, 2010)

Genome Indexing

Rapid Parallel Construction
of Genome Index

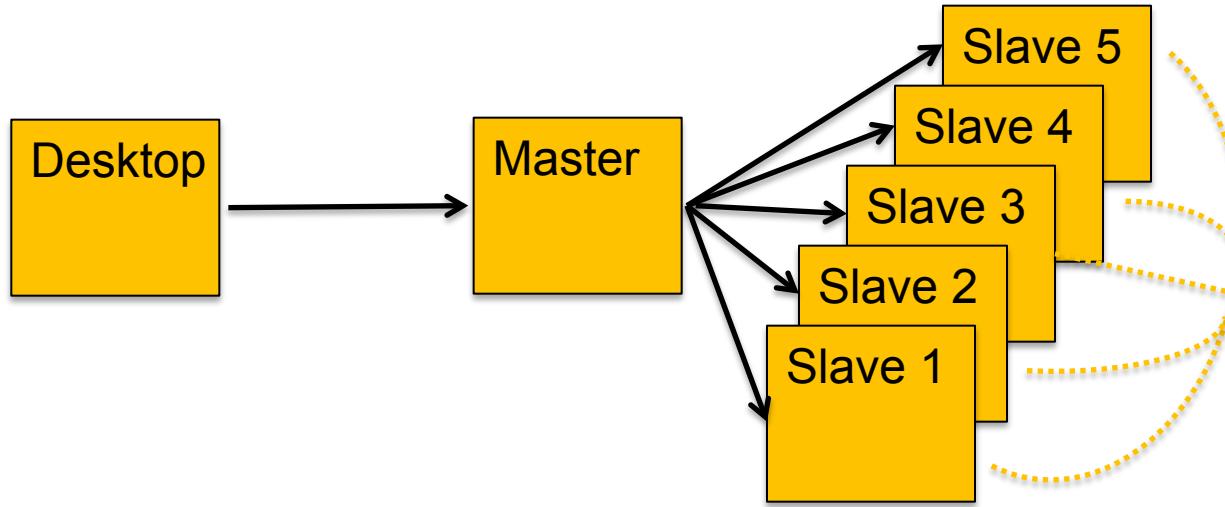
*Construct the BWT of
the human genome in 9 minutes*

\$GATTACA
A\$GATTAC
ACA\$GATT
ATTACA\$G
CA\$GATTA
GATTACA£
TACA\$GA T
TTACA\$GA

(Menon,
Bhat, Schatz, 2011*)

[http://code.google.com/p/
genome-indexing/](http://code.google.com/p/genome-indexing/)

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

Parallel Algorithm Spectrum

Embarrassingly Parallel



Map-only
Each item is Independent

Loosely Coupled



MapReduce
Independent-Sync-Independent

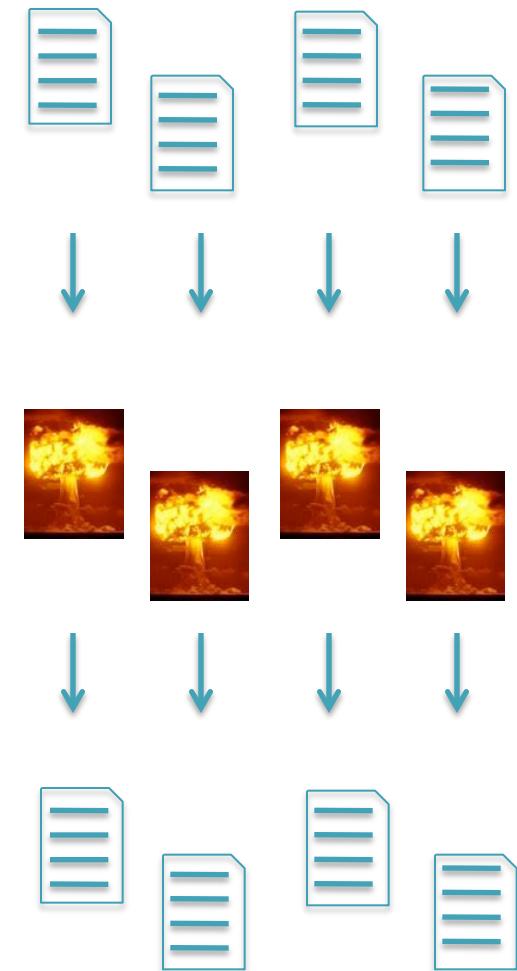
Tightly Coupled



Iterative MapReduce
Constant Sync

I. 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

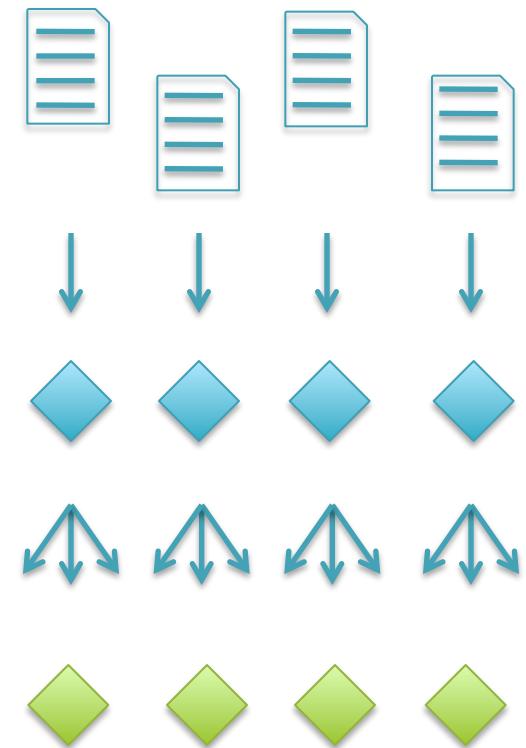


Elementary School Dance



2. Loosely Coupled

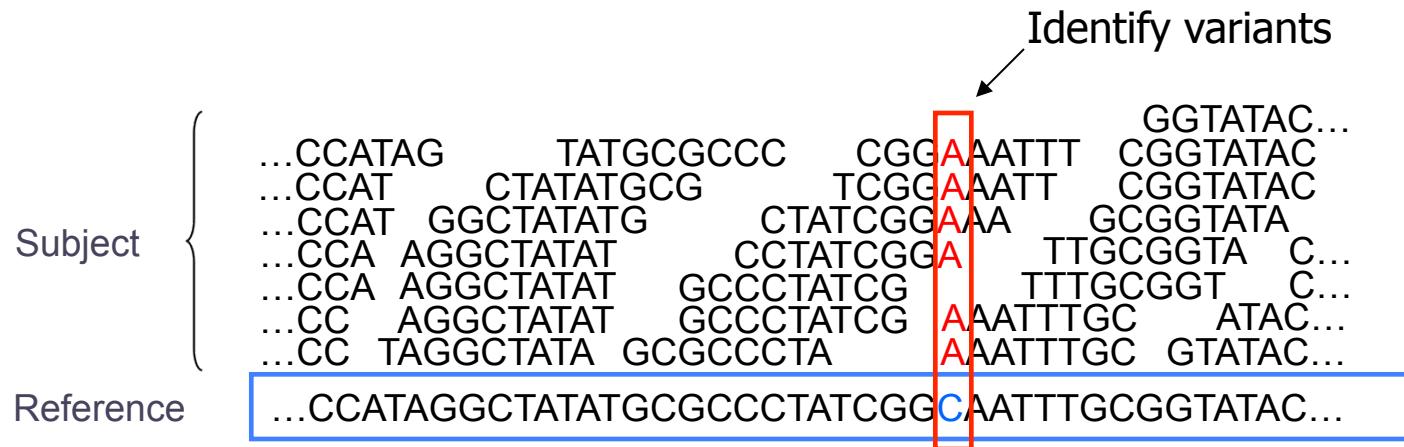
- Divide and conquer
 - Independently process many items
 - Group partial results
 - Scan partial results into final answer
- Challenges
 - Batch computing challenges
 - + Shuffling of huge datasets
- Technologies
 - Hadoop, Elastic MapReduce, Dryad
 - Parallel Databases



Junior High Dance



Short Read Mapping



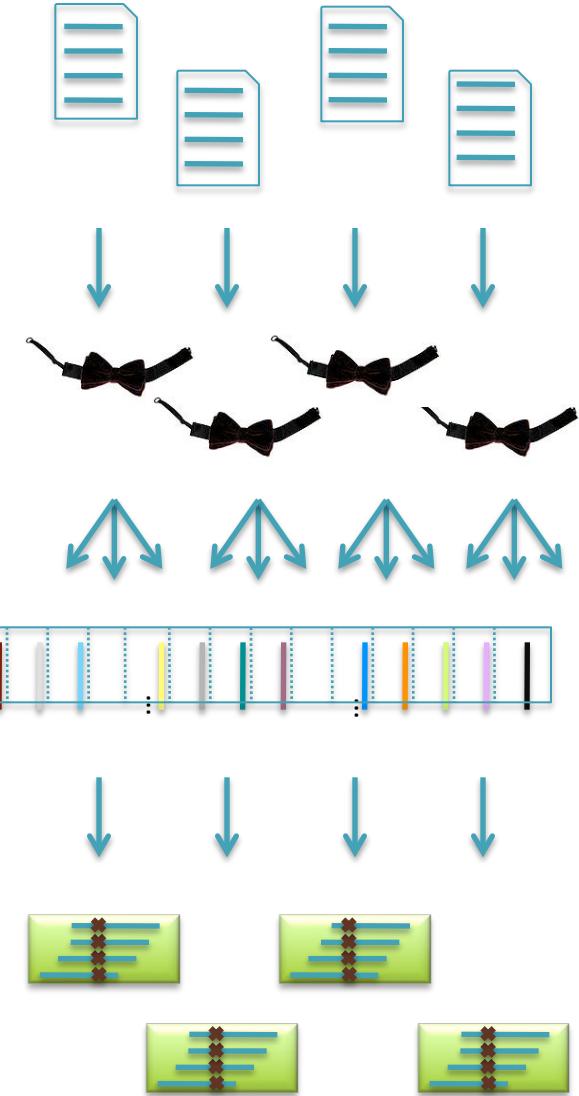
- Given a reference and many subject reads, report one or more “good” end-to-end alignments per alignable read
 - Find where the read most likely originated
 - Fundamental computation for many assays
 - Genotyping RNA-Seq Methyl-Seq
 - Structural Variations Chip-Seq Hi-C-Seq
 - Desperate need for scalable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

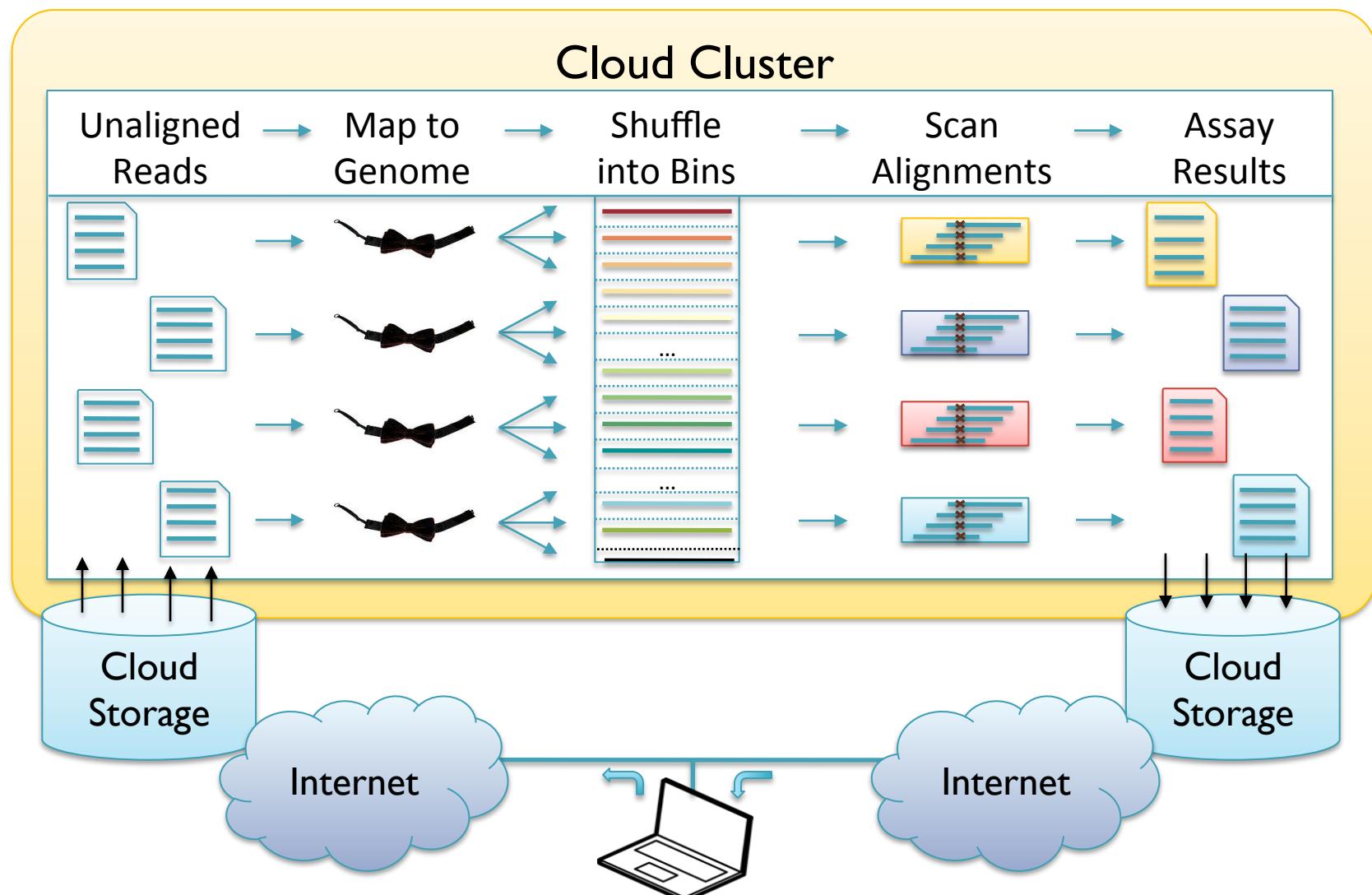
Asian Individual Genome			
Data Loading	3.3 B reads	106.5 GB	\$10.65
Data Transfer	1h :15m	40 cores	\$3.40
Setup	0h :15m	320 cores	\$13.94
Alignment	1h :30m	320 cores	\$41.82
Variant Calling	1h :00m	320 cores	\$27.88
End-to-end	4h :00m		\$97.69

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

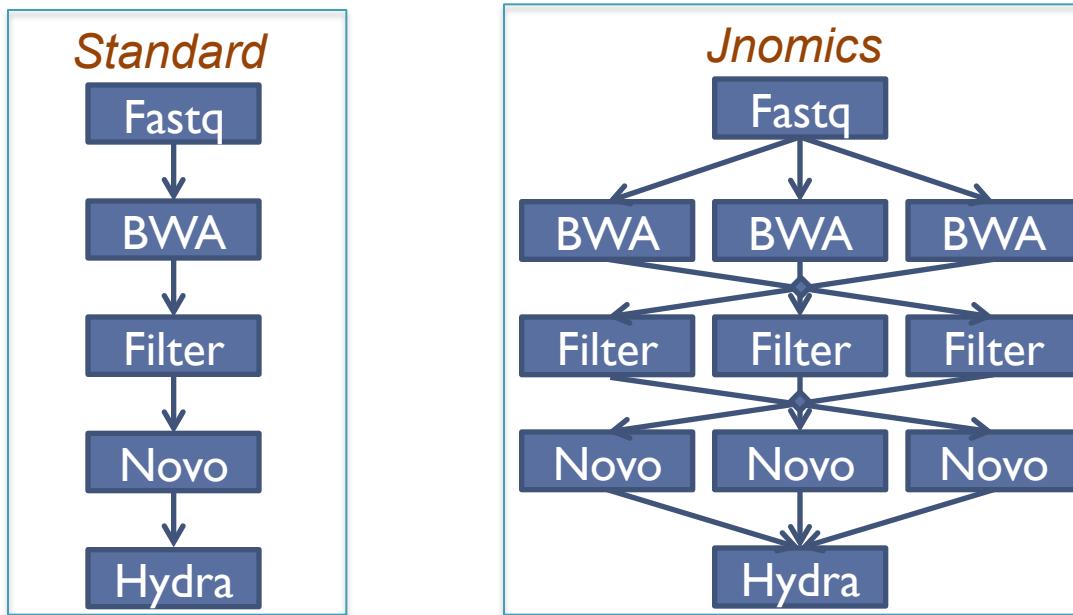


Cloud Computing and the DNA Data Race.

Schatz, MC, Langmead B, Salzberg SL (2010) *Nature Biotechnology*. **28**:691-693

Jnomics: Cloud-scale genomics

James Gurtowski, Matt Titmus, Michael Schatz



- Rapid parallel execution of NGS analysis pipelines
 - FASTX, BWA, Bowtie, Novoalign, SAMTools, Hydra
 - Sorting, merging, filtering, selection, of BAM, SAM, BED, fastq
 - Population analysis: Clustering, GWAS, Trait Inference
- Used for rapidly analyzing human diseases and plants

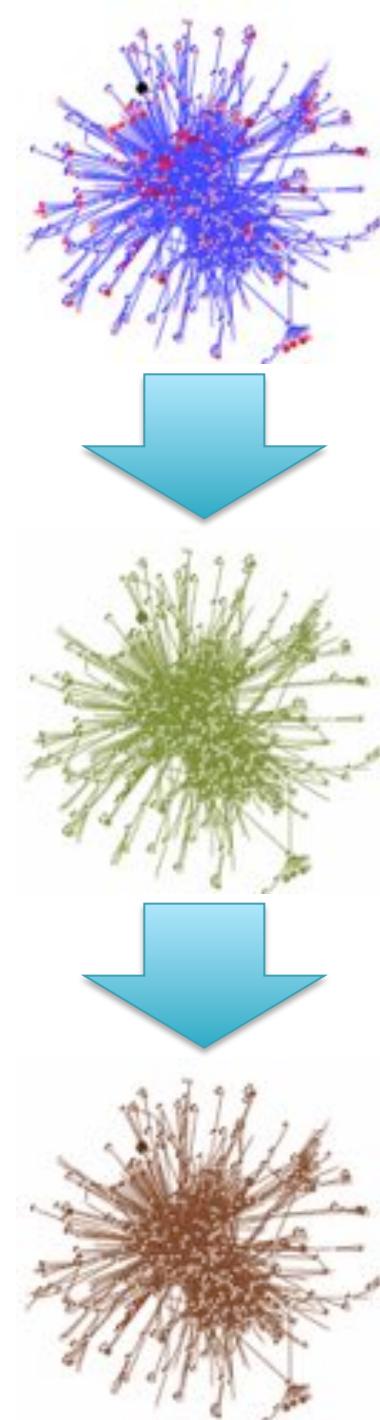


Answering the demands of digital genomics

Titmus, M.A., Gurtowski, J, Schatz, M.C.. (2012) Under Review

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



High School Dance



Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of A Tale of Two Cities
 - Text printed on 5 long spools

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...

- How can he reconstruct the text?
 - $5 \text{ copies} \times 138,656 \text{ words} / 5 \text{ words per fragment} = 138k \text{ fragments}$
 - The short fragments from every copy are mixed together
 - Some fragments are identical

It was the best of

age of wisdom, it was

best of times, it was

it was the age of

it was the age of

it was the worst of

of times, it was the

of times, it was the

of wisdom, it was the

the age of wisdom, it

the best of times, it

the worst of times, it

times, it was the age

times, it was the worst

was the age of wisdom,

was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was

Greedy Reconstruction

It was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

times, it was the worst

times, it was the age

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

- $D_k = (V, E)$
 - V = All length- k subfragments ($k < l$)
 - E = Directed edges between consecutive subfragments
 - Nodes overlap by $k-1$ words

Original Fragment

It was the best of

Directed Edge

It was the best → was the best of

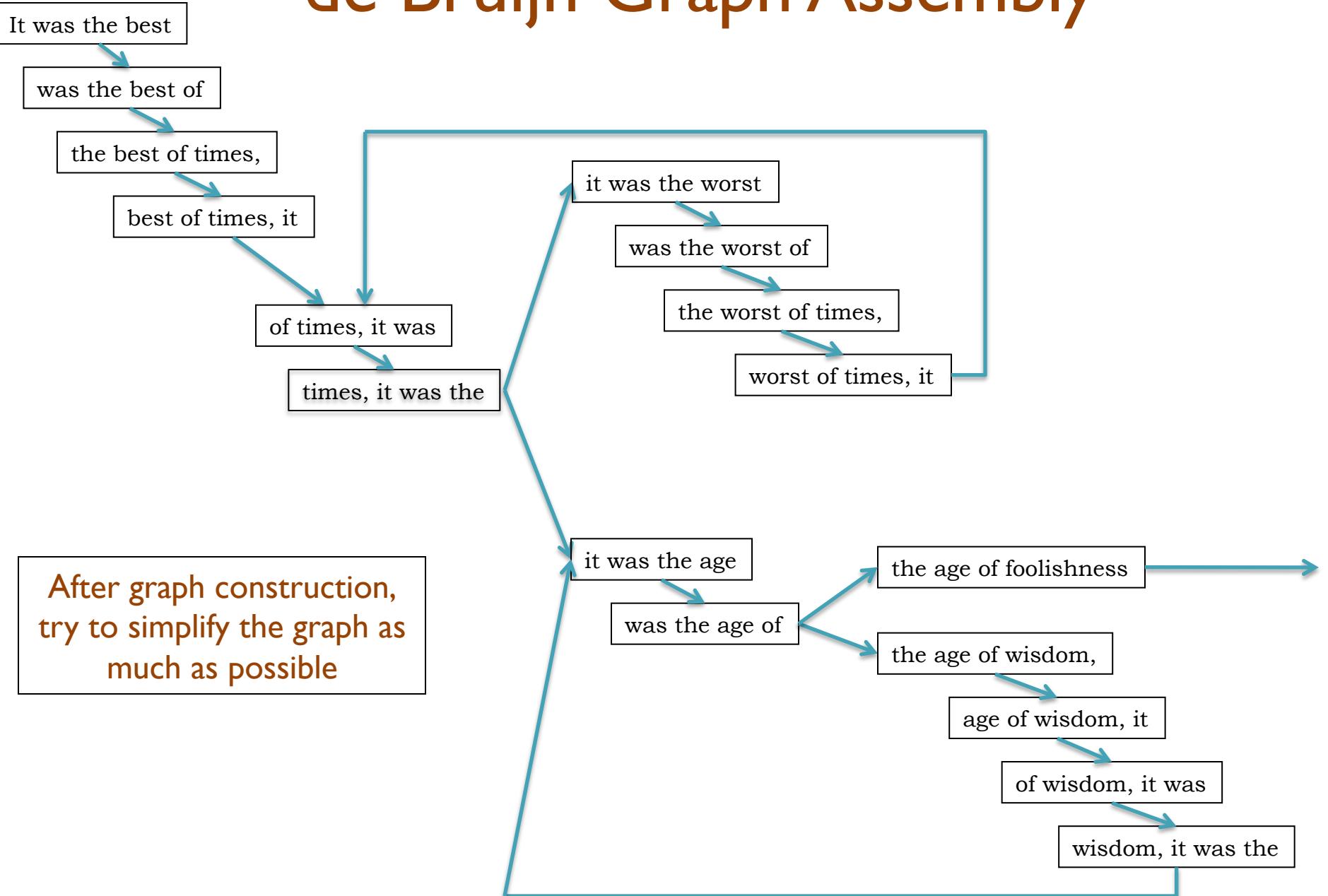
- Locally constructed graph reveals the global sequence structure
 - Overlaps between sequences implicitly computed

de Bruijn, 1946

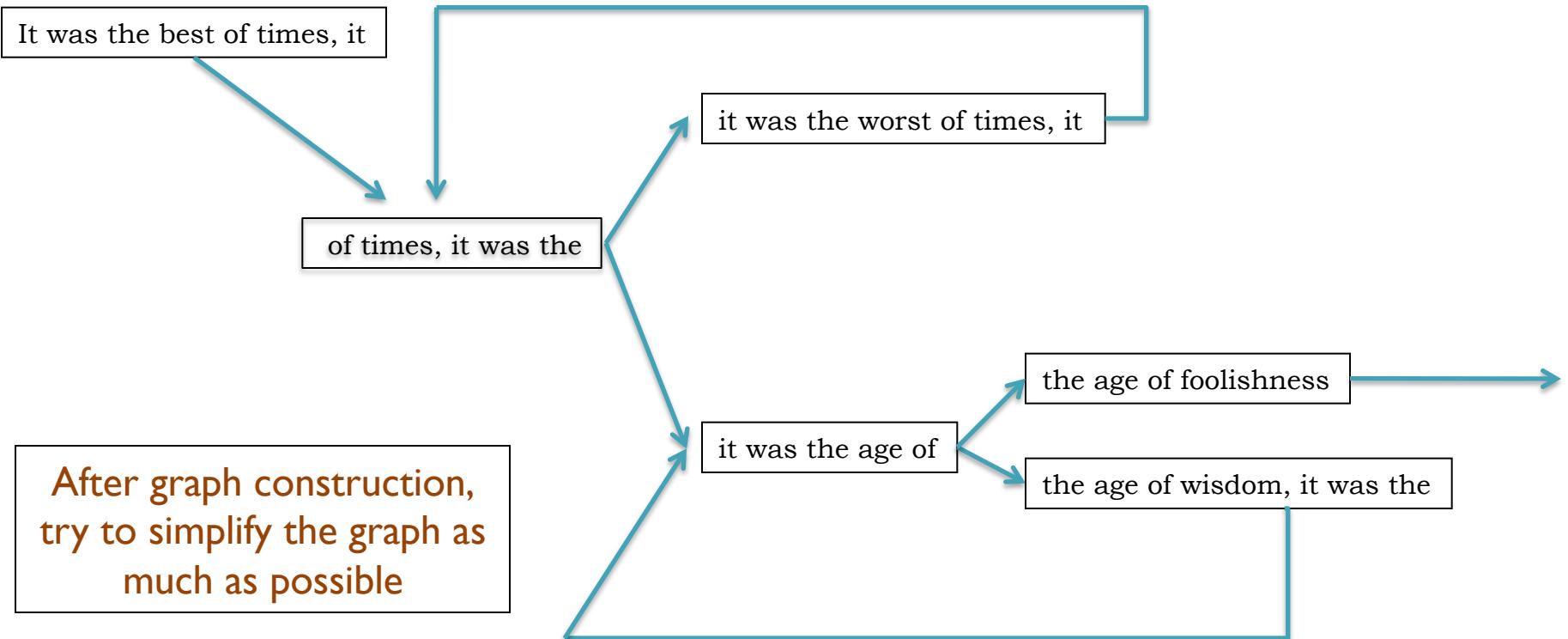
Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



de Bruijn Graph Assembly

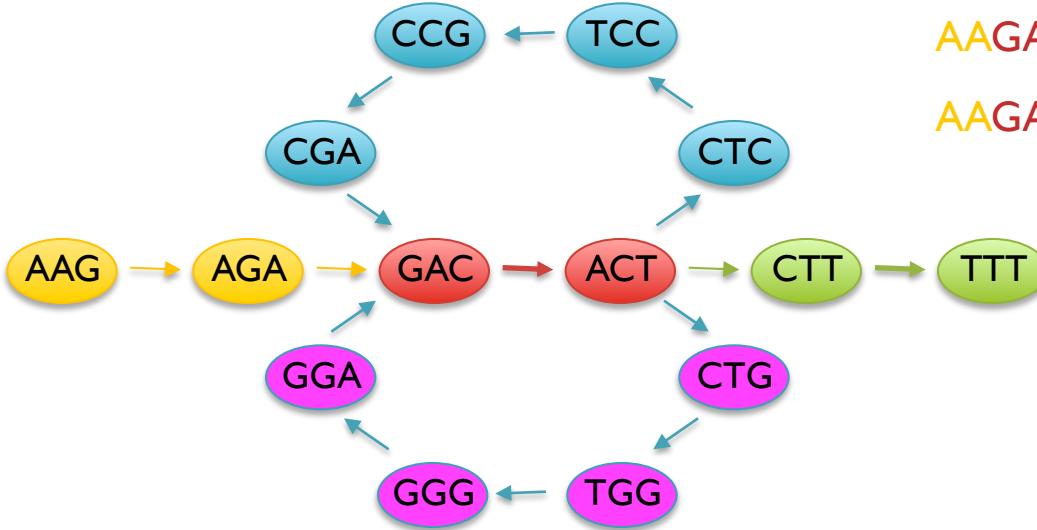


Genome Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
...

de Bruijn Graph



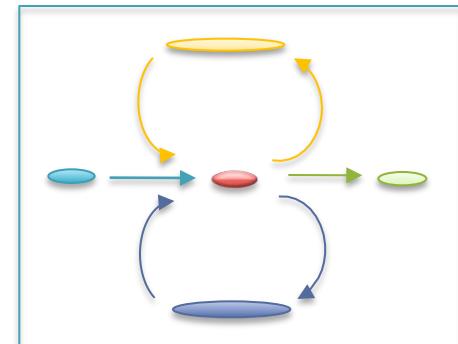
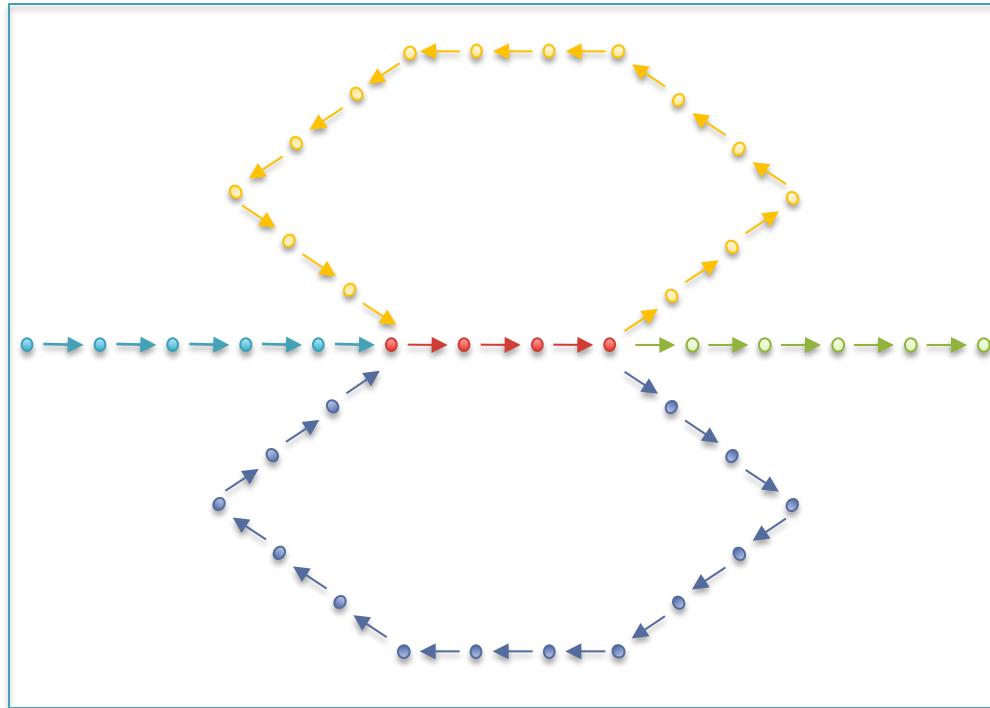
Potential Genomes

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



Design Patterns for Efficient Graph Algorithms in MapReduce.

Lin, J., Schatz, M.C. (2010) Workshop on Mining and Learning with Graphs Workshop (KDD-2010)

Warmup Exercise

- Who here was born closest to July 30?
 - You can only compare to 1 other person at a time



Find winner among 64 teams in just 6 rounds

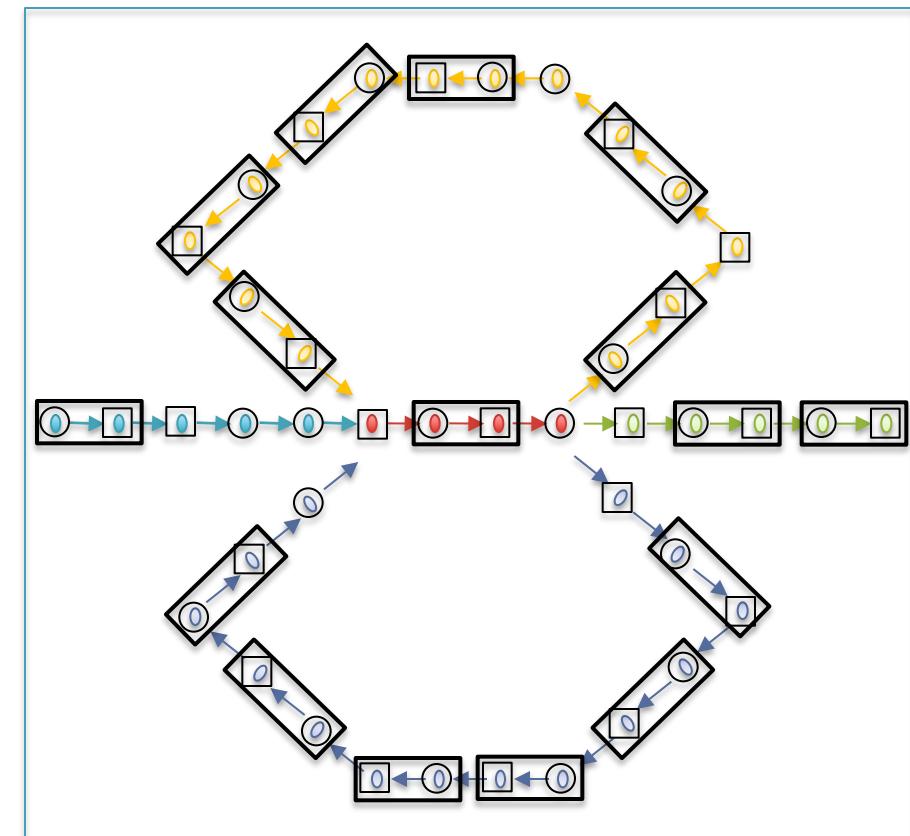
Fast Path Compression

Challenges

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

Randomized List Ranking

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



Initial Graph: 42 nodes

Randomized Speed-ups in Parallel Computation.

Vishkin U. (1984) ACM Symposium on Theory of Computation. 230-239.

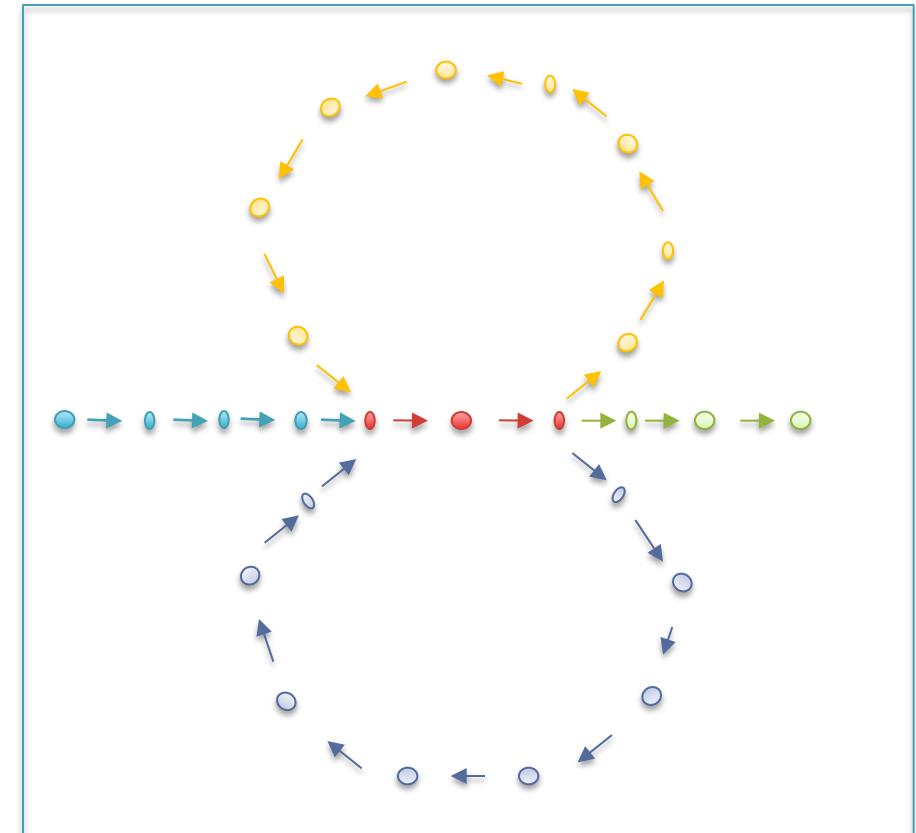
Fast Path Compression

Challenges

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

Randomized List Ranking

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



Round 1: 26 nodes (38% savings)

Randomized Speed-ups in Parallel Computation.

Vishkin U. (1984) ACM Symposium on Theory of Computation. 230-239.

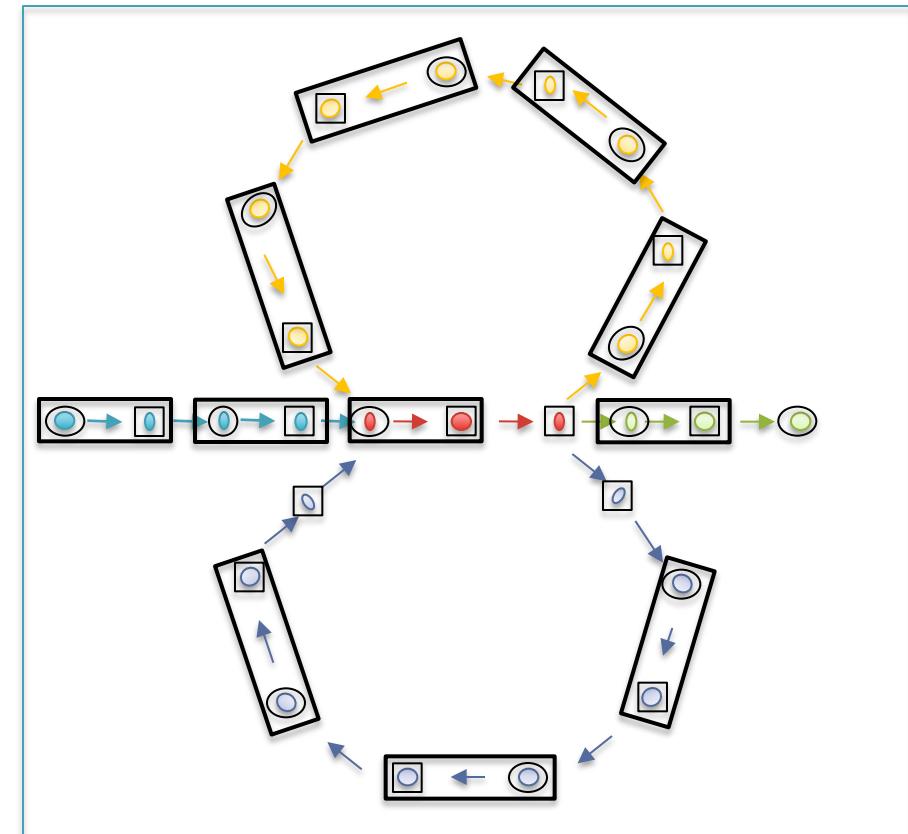
Fast Path Compression

Challenges

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

Randomized List Ranking

- Randomly assign H / T to each compressible node
- Compress $H \rightarrow T$ links



Round 1: 26 nodes (38% savings)

Randomized Speed-ups in Parallel Computation.

Vishkin U. (1984) ACM Symposium on Theory of Computation. 230-239.

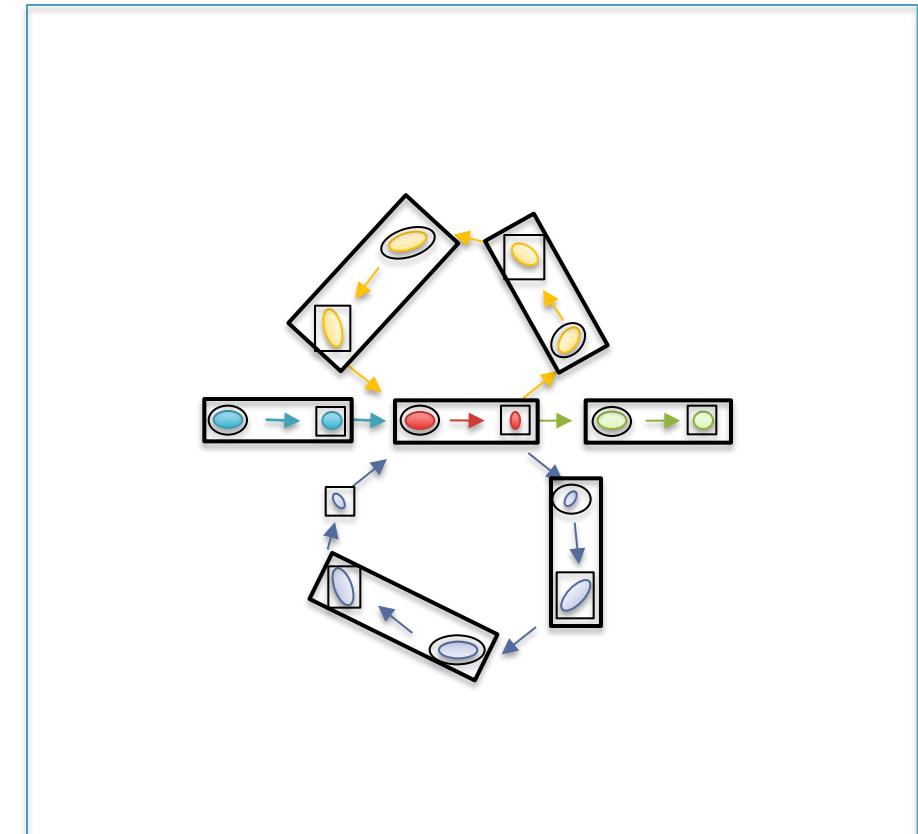
Fast Path Compression

Challenges

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

Randomized List Ranking

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



Round 2: 15 nodes (64% savings)

Randomized Speed-ups in Parallel Computation.

Vishkin U. (1984) ACM Symposium on Theory of Computation. 230-239.

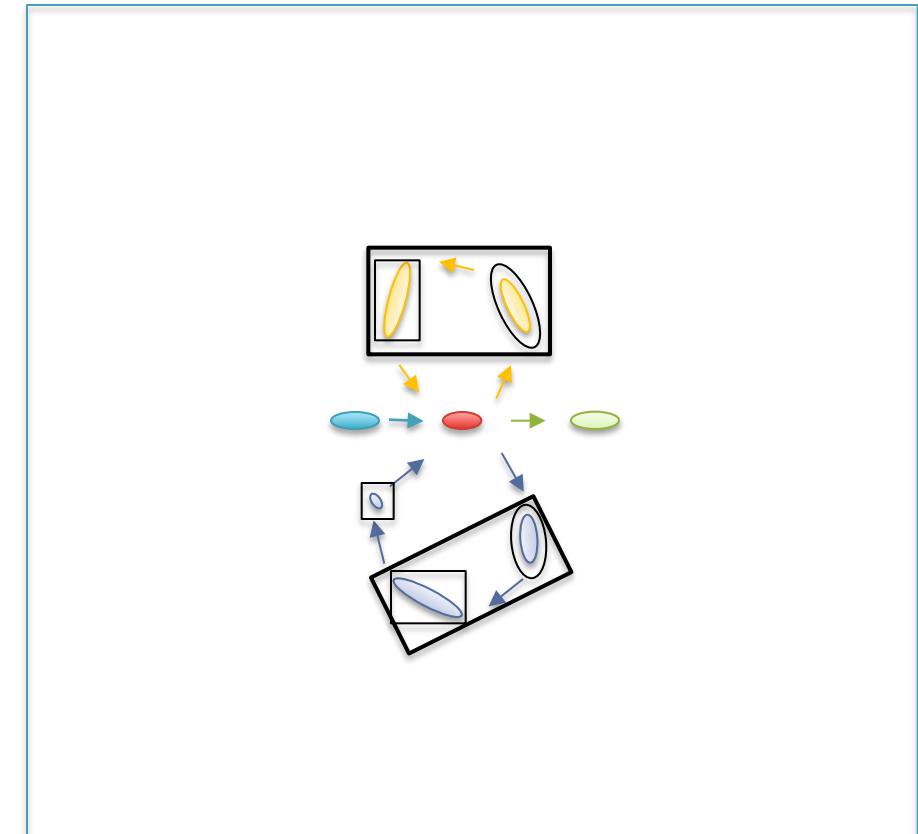
Fast Path Compression

Challenges

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

Randomized List Ranking

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



Round 2: 8 nodes (81% savings)

Randomized Speed-ups in Parallel Computation.

Vishkin U. (1984) ACM Symposium on Theory of Computation. 230-239.

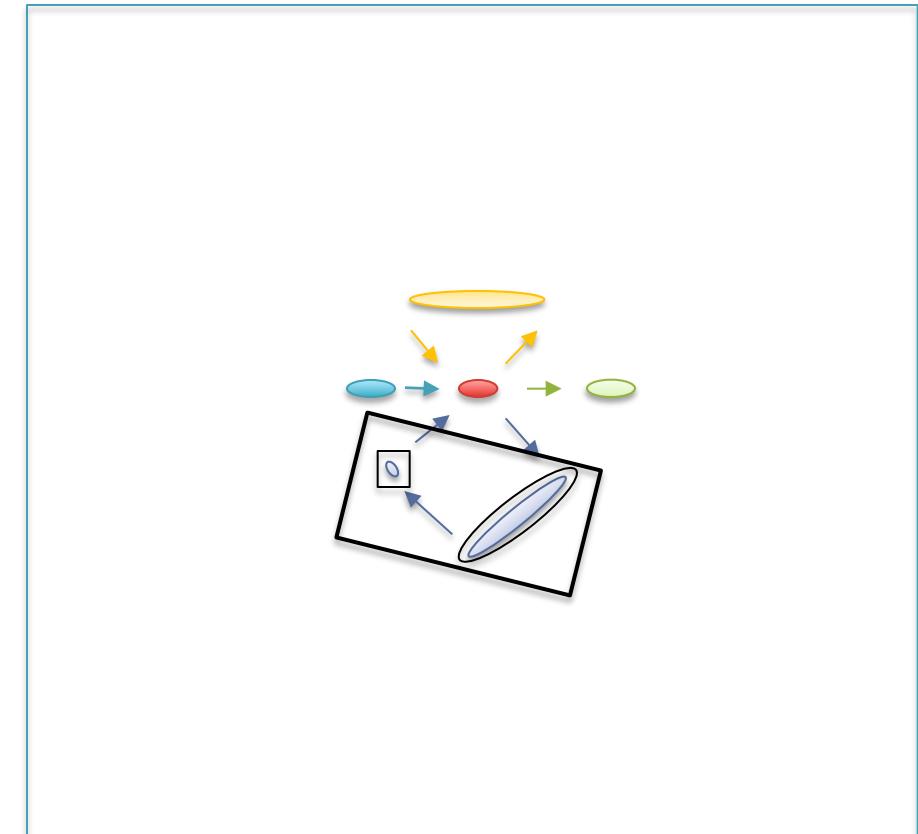
Fast Path Compression

Challenges

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

Randomized List Ranking

- Randomly assign \textcircled{H} / \boxed{T} to each compressible node
- Compress $\textcircled{H} \rightarrow \boxed{T}$ 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

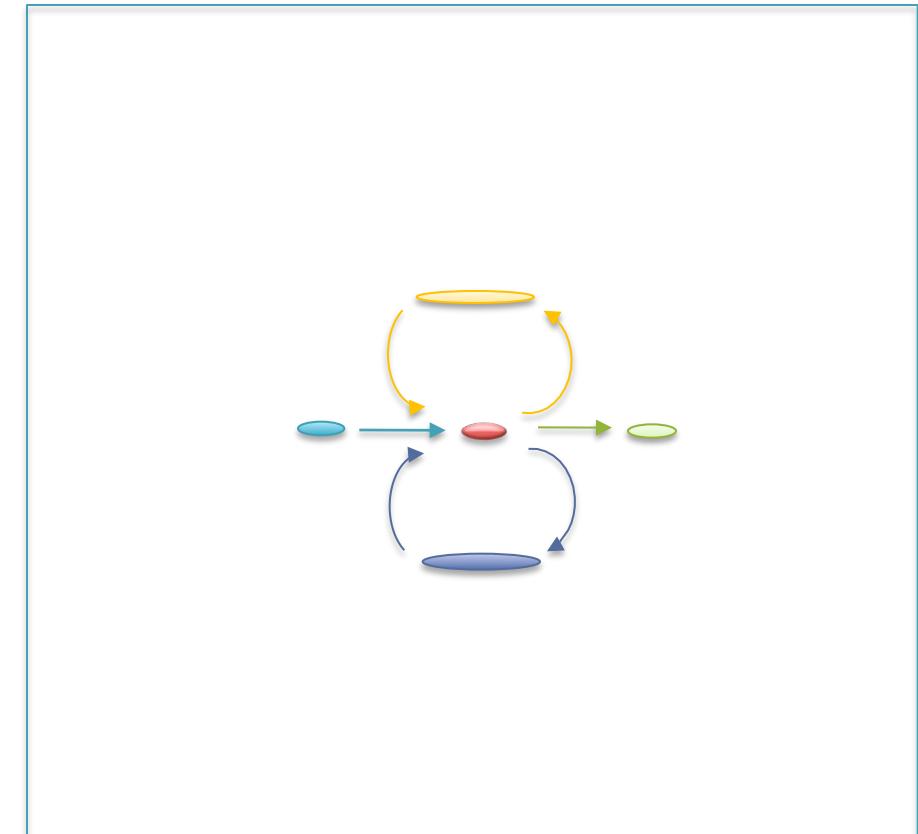
- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign \textcircled{H} / \boxed{T} to each compressible node
- Compress $\textcircled{H} \rightarrow \boxed{T}$ 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.

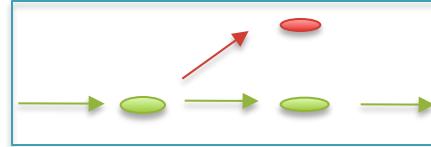
Node Types



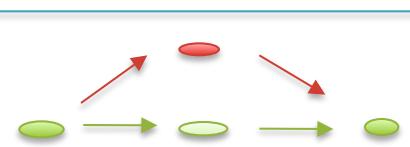
Isolated nodes (10%)



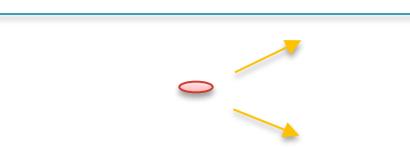
Tips (46%)



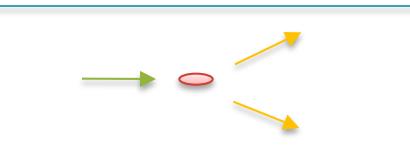
Bubbles/Non-branch (9%)



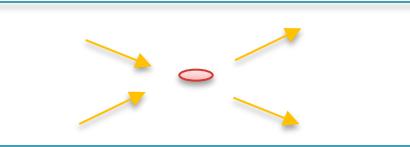
Dead Ends (.2%)



Half Branch (25%)



Full Branch (10%)



(Chaisson, 2009)

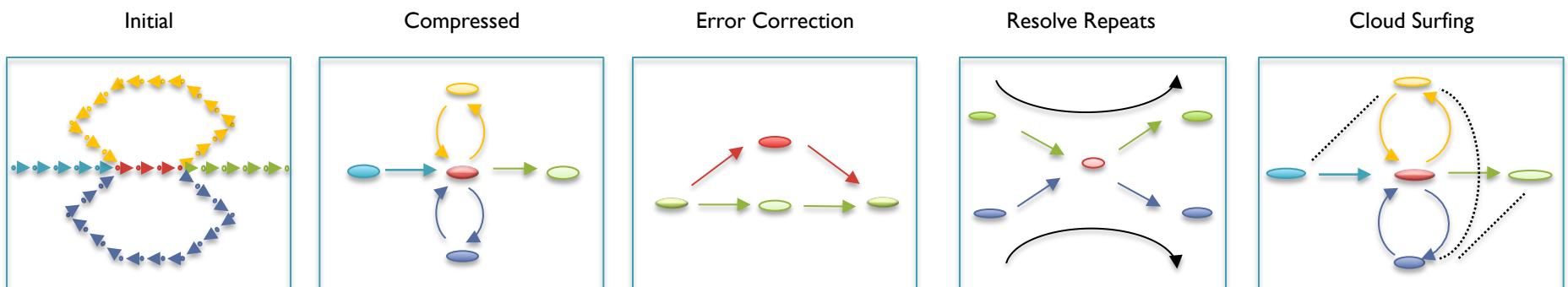
Contrail

<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)



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 *et al.* *In Preparation.*

Scalpel: Haplotype Microassembly

G. Narzisi, D. Levy, I. Iossifov, J. Kendall, M. Wigler, M. Schatz



- Use assembly techniques to identify complex variations from short reads
 - Improved power to find indels
 - Trace candidate haplotypes sequences as paths through assembly graphs



Ref: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Father: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Mother: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Sib: . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Aut(1): . . . TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTGCCCGGA . . .

Aut(2): . . . TCAGAACAGCTGGATGAGATCTTACC-----CC**G**GGAGATTGTCTTGCCCGGA . . .

6bp heterozygous indel at chr13:25280526 ATP12A



Summary

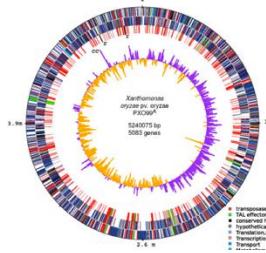
- We are in the digital age of biology
 - Next generation sequencing, microarrays, mass spectrometry, microscopy, ecology, etc
 - Parallel computing may be our only hope for keeping up with the pace of advance
- Modern biology requires (is) quantitative biology
 - Computational, mathematical, and statistical techniques applied to analyze, integrate, and interpret biological sensor data
- Don't let the data tsunami crash on you
 - Study, practice, collaborate with quantitative techniques

Acknowledgements



Mitch Bekritsky
Giuseppe Narzisi

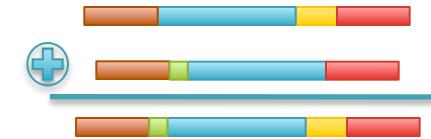
Ivan Iossifov
Wigler Lab



Hayan Lee
James Gurtowski

Ware Lab
McCombie Lab

Adam Phillippy (NBACC)
Sergey Koren (NBACC)

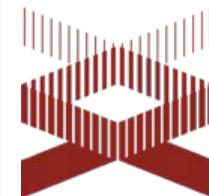


Paul Baranay
Eric Biggers
Robert Aboukhalil

Scott Emrich (ND)
Steven Salzberg (JHU)
Mihai Pop (UMD)
Ben Langmead (JHU)

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

KBASE
predictive biology
DOE Systems Biology Knowledgebase



National Human
Genome Research
Institute

Thank You!

<http://schatzlab.cshl.edu/>
@mike_schatz

