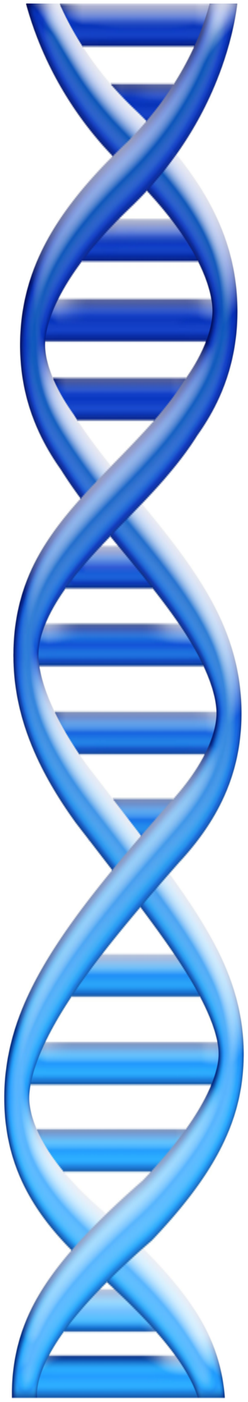
Answering the demands of digital genomics

Michael Schatz

Oct 4, 2011

Frontiers in Genomics



Outline

1. Milestones in genomics
2. The demands of genomics
3. 21st Century Genomics
   1. Parallel & Cloud Computing
   2. Hadoop and MapReduce
   3. Hadoop Applications for Genomics

Milestones in Genomics

Versuche

über

Pflanzen-Hybriden,

von

Gregor Mendel.

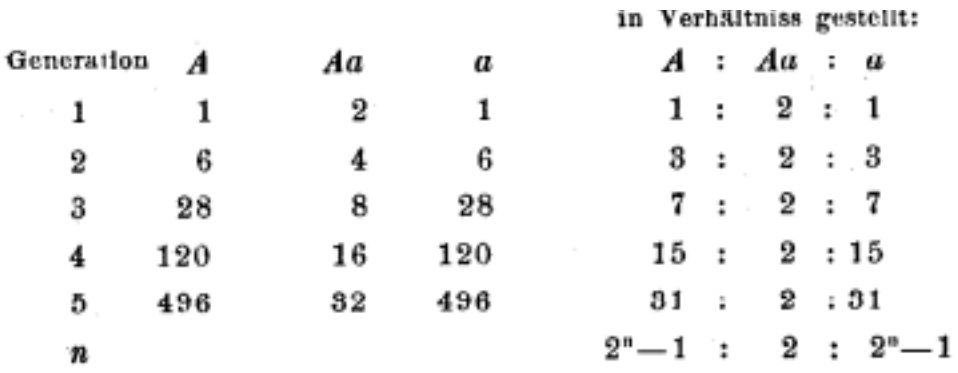
(Separatabdruck aus dem IV. Bande der Verhandlungen des naturforschenden Vereines.)

Im Verlage des Vereines.

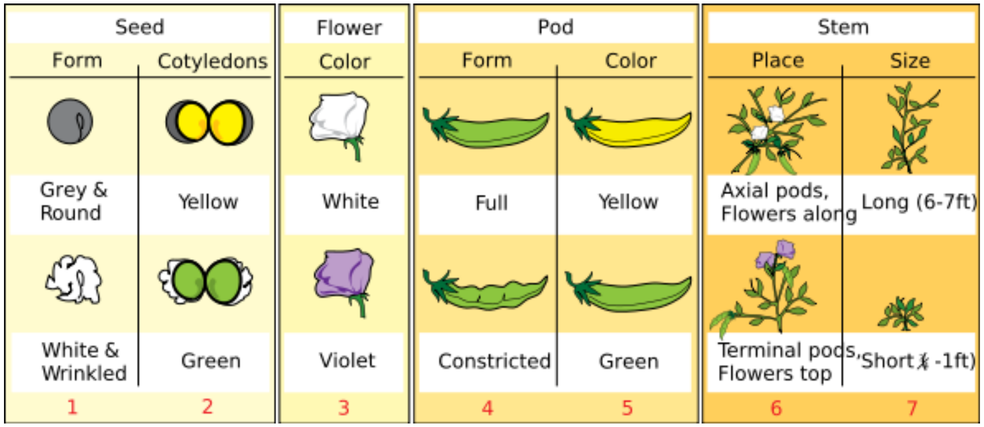
Brünn, 1866.

Aus Georg Gastl's Buchdruckerei, Postgasse Nr. 446.

Observations of 29,000 pea plants and 7 traits



Generation	A	Aa	a	in Verhältniss gestellt: $A : Aa : a$
1	1	2	1	1 : 2 : 1
2	6	4	6	3 : 2 : 3
3	28	8	28	7 : 2 : 7
4	120	16	120	15 : 2 : 15
5	496	32	496	31 : 2 : 31
n				2^n-1 : 2 : 2^n-1

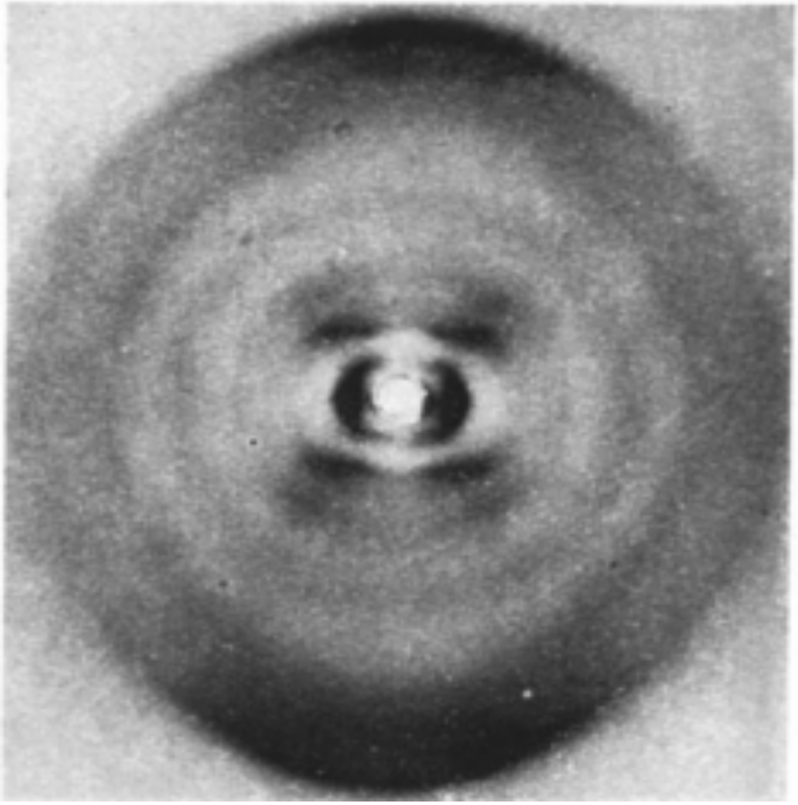


http://en.wikipedia.org/wiki/Experiments_on_Plant_Hybridization

Versuche über Pflanzen-Hybriden. Verh. Naturforsch (Experiments in Plant Hybridization)
Mendel, G. (1866). Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32).

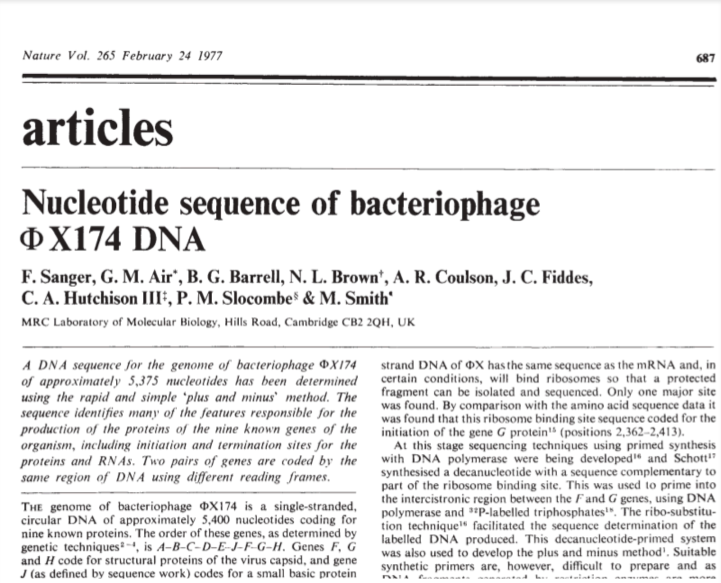
Milestones in Genomics

The origin and behavior of mutable loci in maize
McClintock, B (1950) *Proceedings of the National Academy of Sciences.* 36:344–55.



Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid
Watson JD, Crick FH (1953). *Nature* 171: 737–738.

Milestones in Genomics



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[2-4], 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[15] (positions 2,362–2,413).

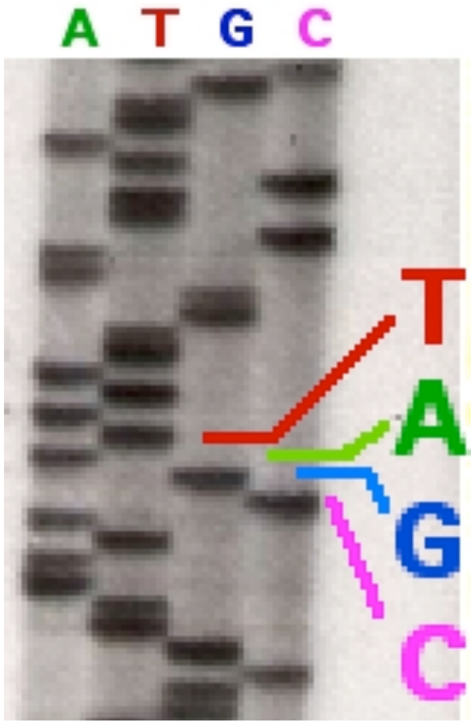
At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[16] and Schott[17] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[18]. The ribo-substitution technique[16] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[1]. Suitable synthetic primers are, however, difficult to prepare and as

1977

1st Complete Organism

Bacteriophage ϕX174

5375 bp



Radioactive Chain Termination

5000bp / week / person

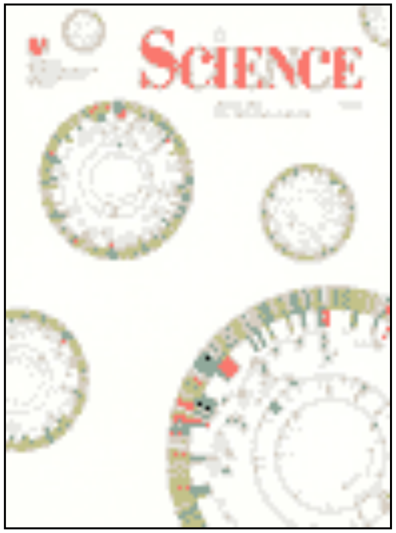
http://en.wikipedia.org/wiki/File:Sequencing.jpg

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

Nucleotide sequence of bacteriophage ϕX174 DNA

Sanger, F. et al. (1977) *Nature. 265: 687 - 695*

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

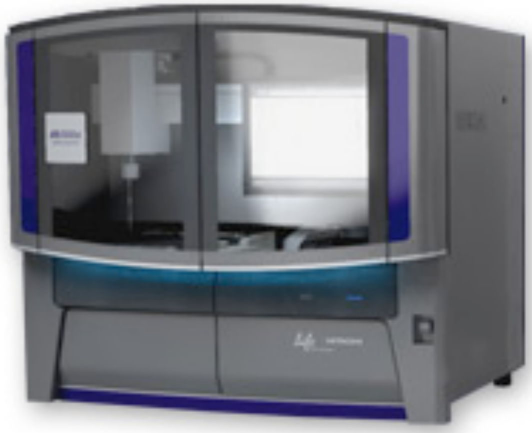
2001
Venter *et al.* / IHGSC
Human Genome
Celera Assembler. 2.9 Gbp

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

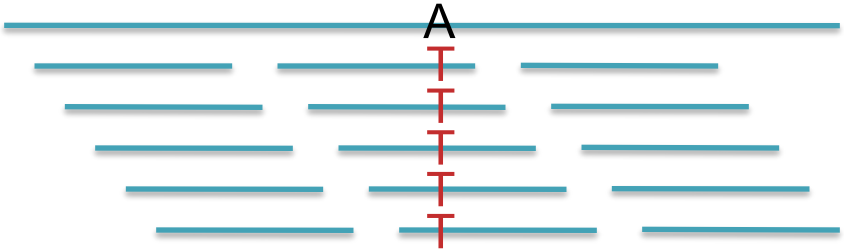
Milestones in Genomics: Third Generation Sequencing

2010
Ion Torrent
Postlight Sequencing
Current Specs (Ion 318):
11M 300bp reads / run =
>1Gbp / day

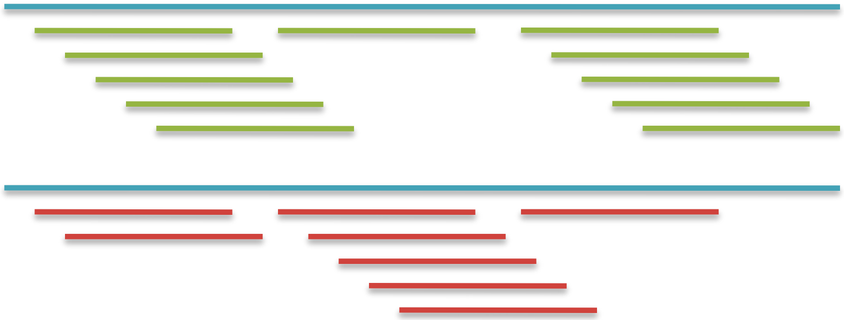
2011
Pacific Biosciences
SMRT Sequencing
Current Specs (RS):
50k 2kbp reads / run =
>200Mbp / day

Milestones in Genomics

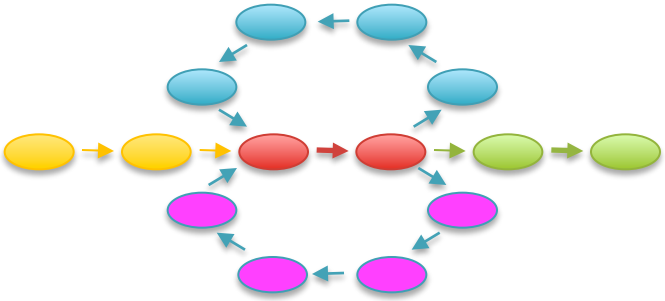
Alignment & Variations



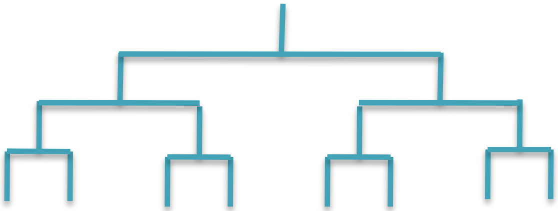
Differential Analysis



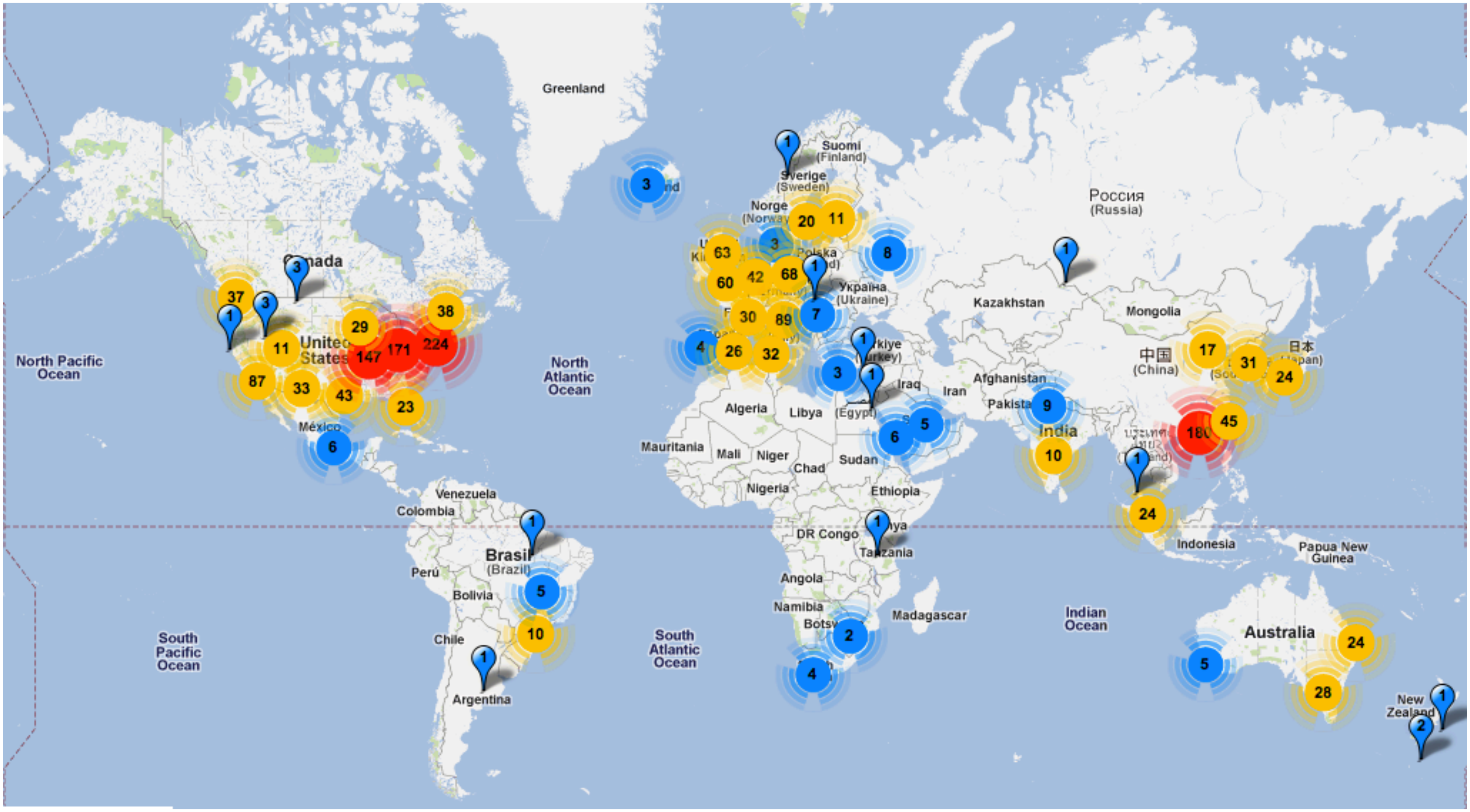
De novo Assembly



Phylogeny & Evolution



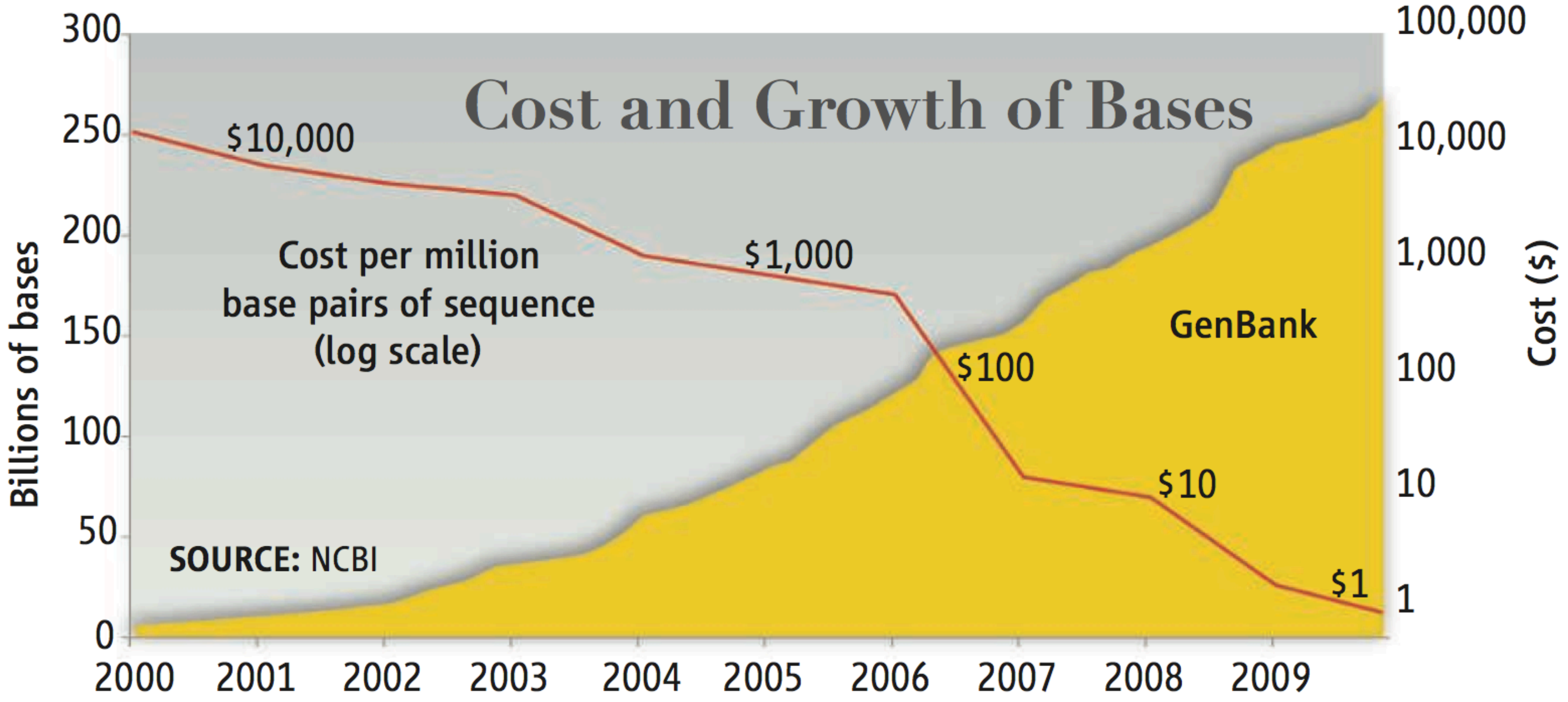
Sequencing Centers



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

DNA Data Tsunami

Current world-wide sequencing capacity exceeds 13Pbp/year and is growing at 5x per year!



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

21st Century Genomics

- The cornerstones of genomics continue to be *observation*, *experimentation*, and *interpretation* of the living world
  - Technology has and will continue to push the frontiers of genomics
  - Measurements will be made *digitally* in great quantities, at extremely high resolution, and for diverse applications

- Demands of digital genomics
  1. *Experimental design*: selection, collection, tracking & metadata
     - Ontologies, LIMS, sample databases
  2. *Observation*: measurement, storage, transfer, computation
     - Algorithms to overcome sensor errors & limitations, computing at scale
  3. *Integration*: multiple samples, multiple assays, multiple analyses
     - Reproducible workflows, common formats, resource federation
  4. *Discovery*: visualizing, interpreting, modeling
     - Clustering, data reduction, trend analysis

Genomics and Parallel Computing

Current world-wide sequencing capacity exceeds 13Pbp/year and is growing at 5x per year!

Our best (only) hope is to use many computers:

- Parallel Computing aka Cloud Computing

- Now your programs will crash on 1000 computers instead of just 1

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

amazon webservices™

EC2 Architecture

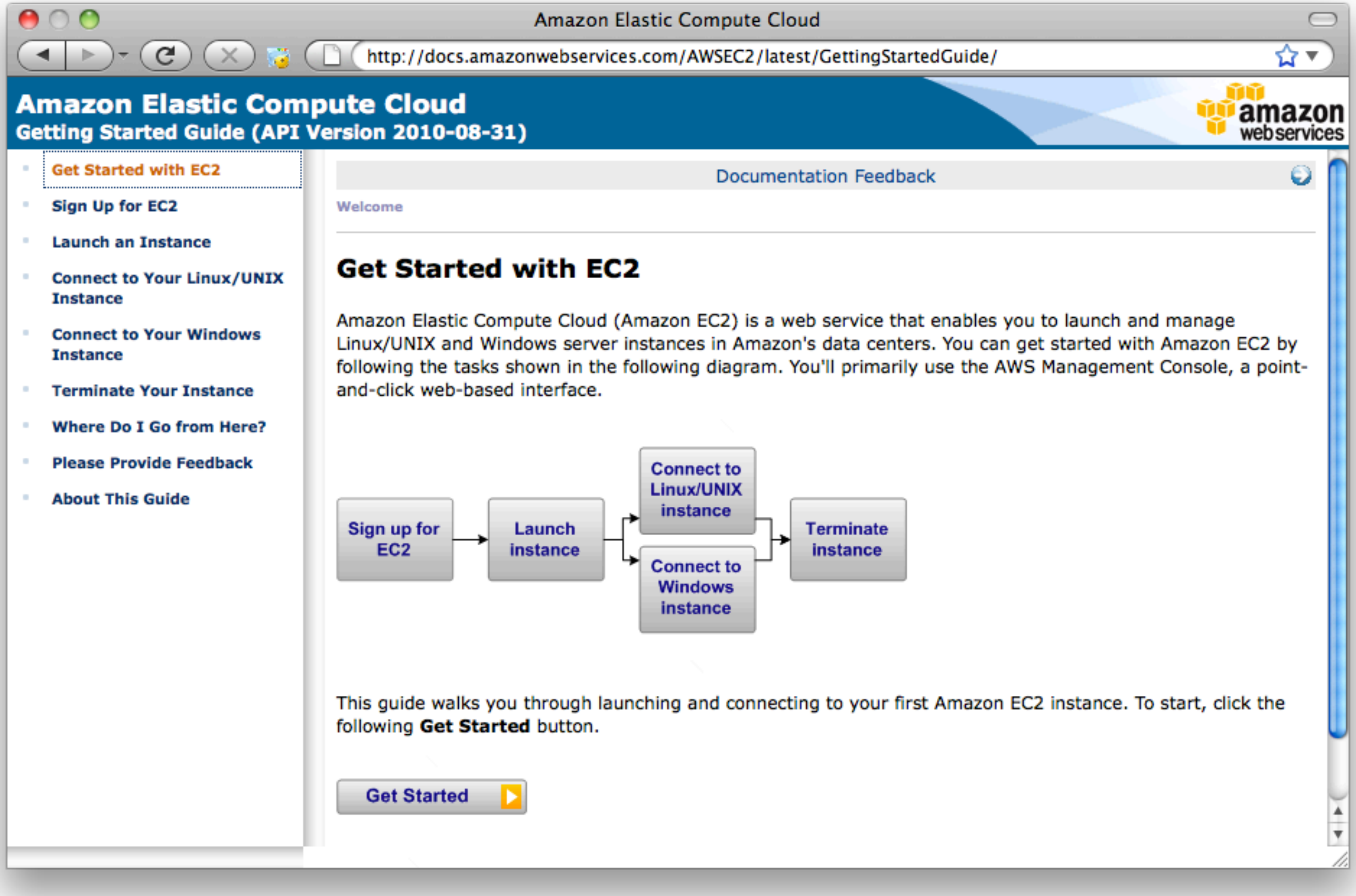
- Very large cluster of machines
  - Effectively infinite resources
  - High-end servers with many cores and many GB RAM

Windows
Linux
Solaris
Windows
Linux
Solaris

- 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

Getting Started

http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/



Amazon Elastic Compute Cloud
http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/
amazon webservices
Documentation Feedback
Welcome

Amazon Elastic Compute Cloud
Getting Started Guide (API Version 2010-08-31)

- Get Started with EC2
- Sign Up for EC2
- Launch an Instance
- Connect to Your Linux/UNIX Instance
- Connect to Your Windows Instance
- Terminate Your Instance
- Where Do I Go from Here?
- Please Provide Feedback
- About This Guide

Get Started with EC2

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that enables you to launch and manage Linux/UNIX and Windows server instances in Amazon's data centers. You can get started with Amazon EC2 by following the tasks shown in the following diagram. You'll primarily use the AWS Management Console, a point-and-click web-based interface.

This guide walks you through launching and connecting to your first Amazon EC2 instance. To start, click the following **Get Started** button.

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

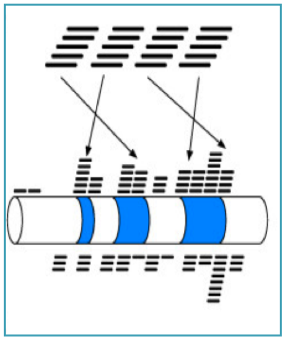
http://cloudburst-bio.sf.net

(Schatz, 2009)

Myrna

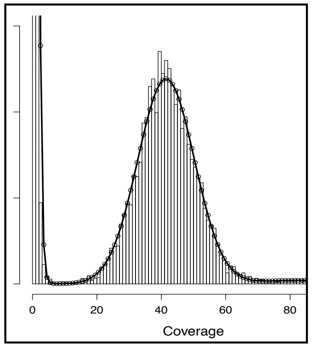
Cloud-scale differential gene expression for RNA-seq

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.cbcb.umd.edu/software/quake/

(Kelley, Schatz, Salzberg, 2010)

Genome Indexing

Rapid Parallel Construction of Genome Index

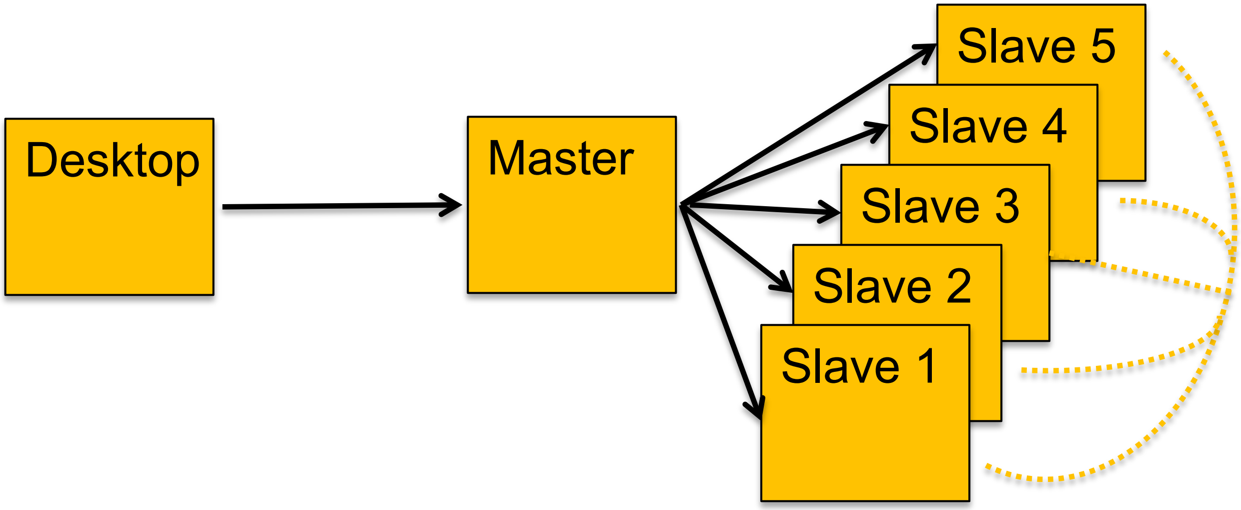
Construct the BWT of the human genome in 9 minutes

$GATTACA
A$GATTAC
ACA$GATT
ATTACA$G
CA$GATTA
GATTACA$
TACA$GAT
TTACA$GA

(Menon, Bhat, Schatz, 2011*)

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

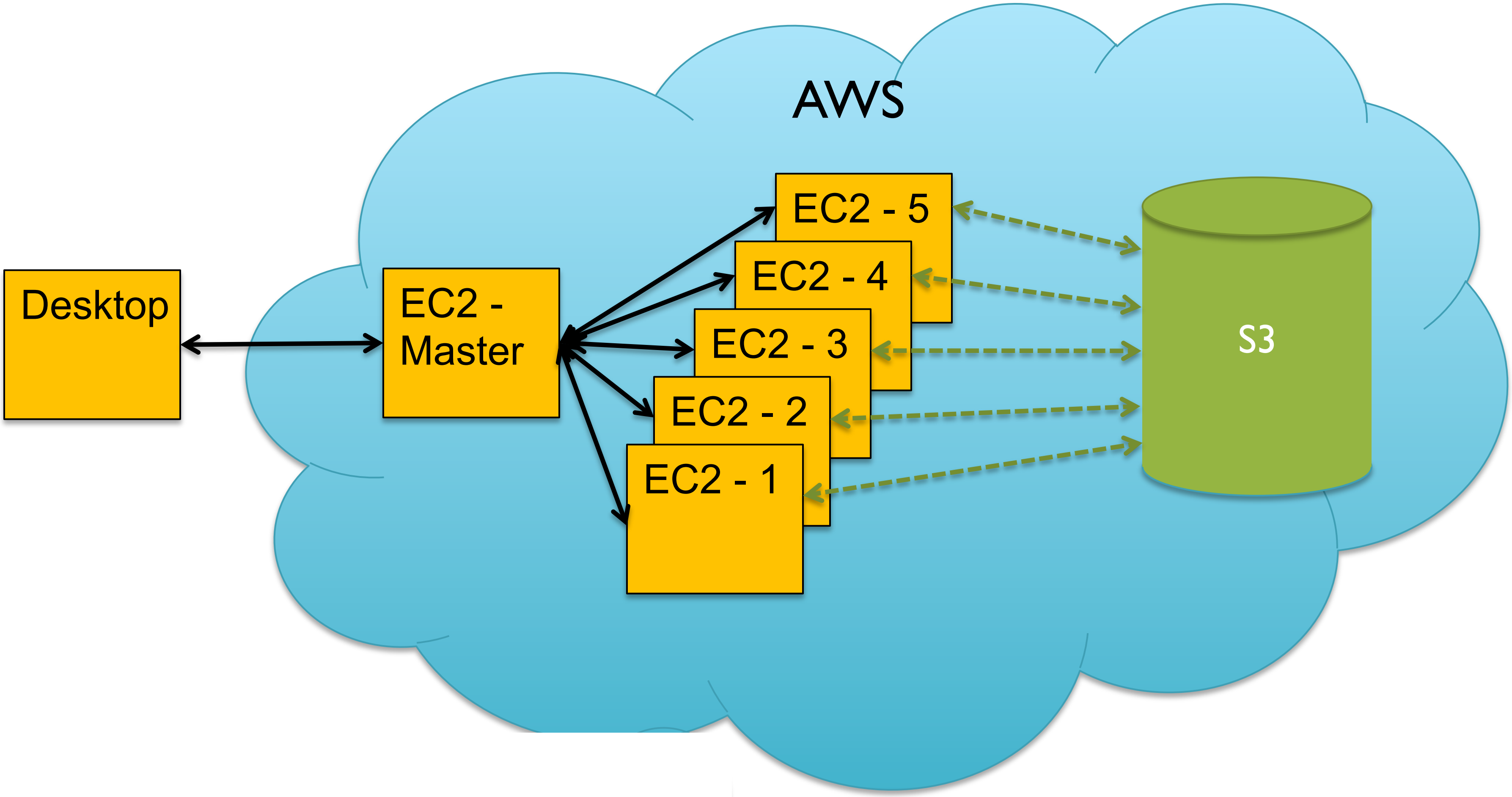
System Architecture



- 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

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.

Parallel Algorithm Spectrum

Embarrassingly Parallel

Map-only
Each item is Independent

Loosely Coupled

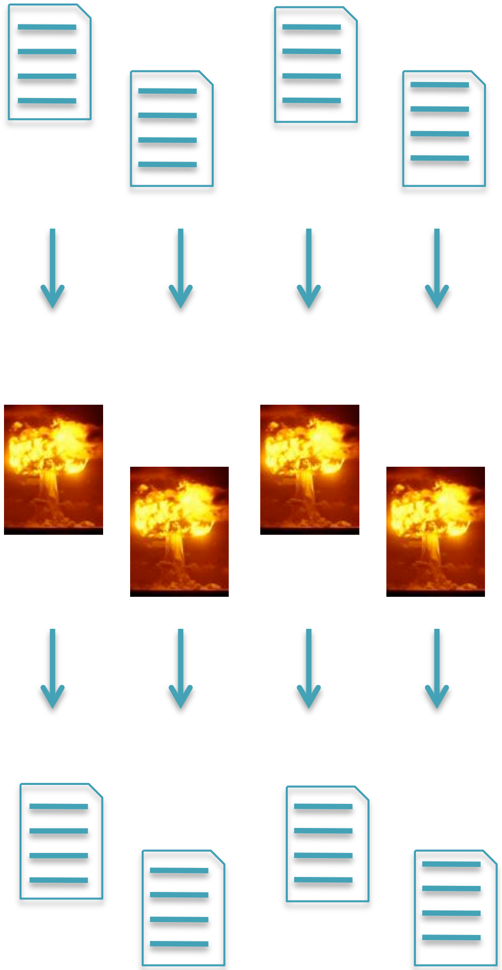
MapReduce
Independent-Sync-Independent

Tightly Coupled

Iterative MapReduce
Constant Sync

1. Embarrassingly Parallel

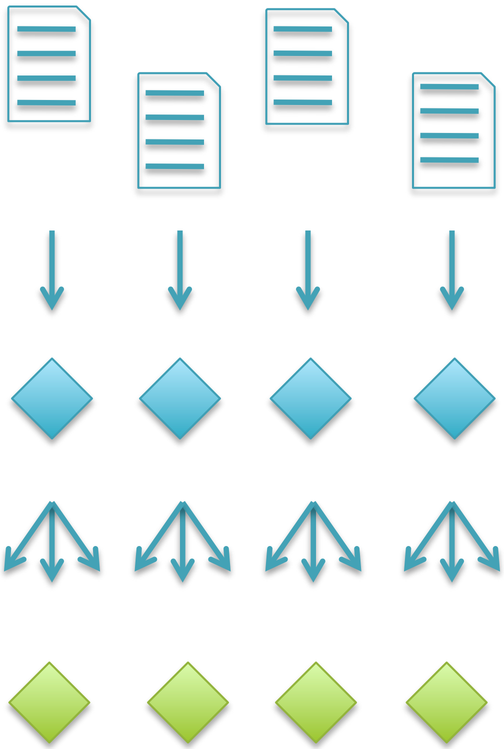
- Batch computing
  - Each item is independent
  - Split input into many chunks
  - Process each chunk separately on a different computer
- Challenges
  - Distributing work, load balancing, monitoring & restart
- Technologies
  - Condor, Sun Grid Engine
  - Amazon Simple Queue



Elementary School Dance

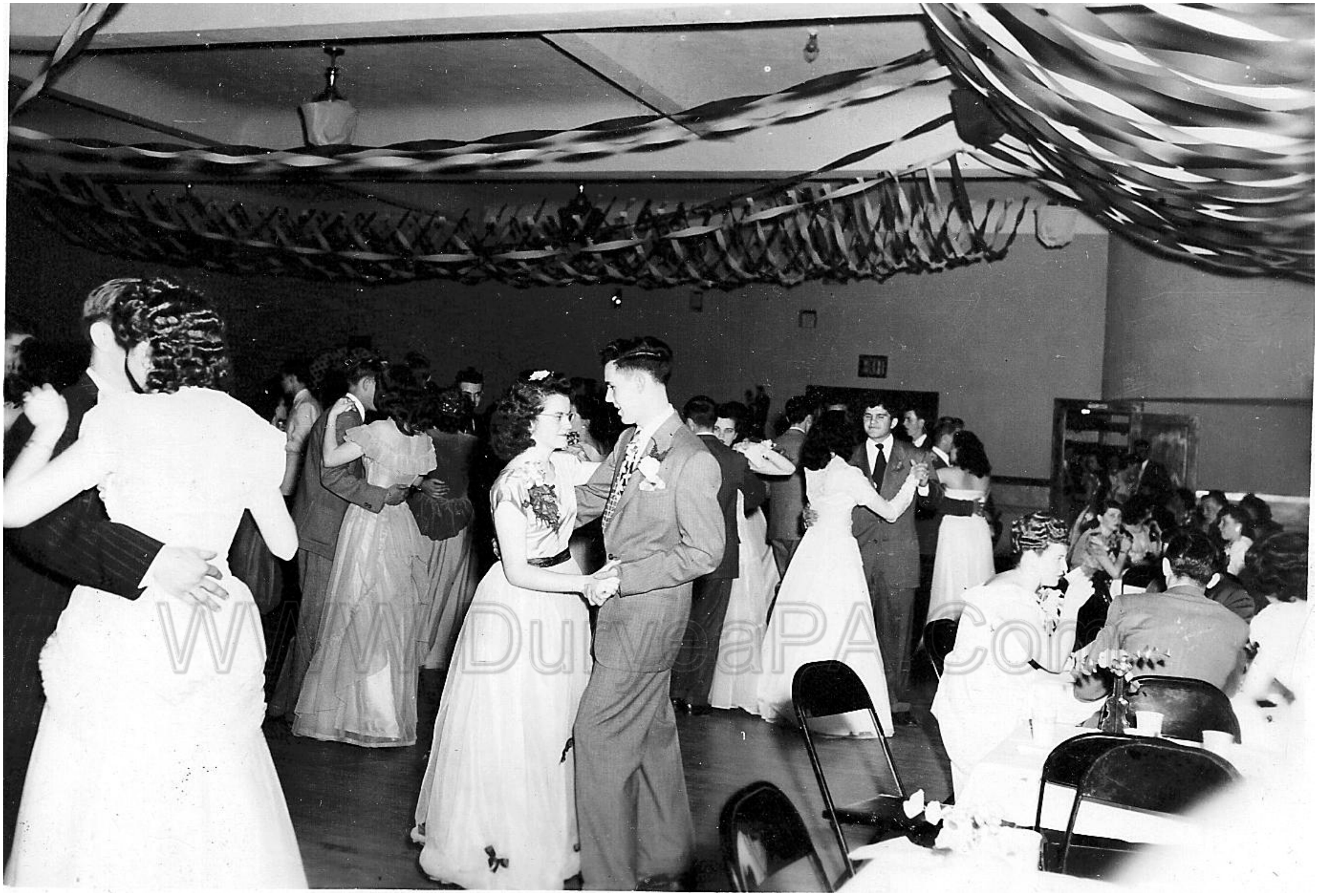
2. Loosely Coupled

- Divide and conquer
  - Independently process many items
  - Group partial results
  - Scan partial results into final answer

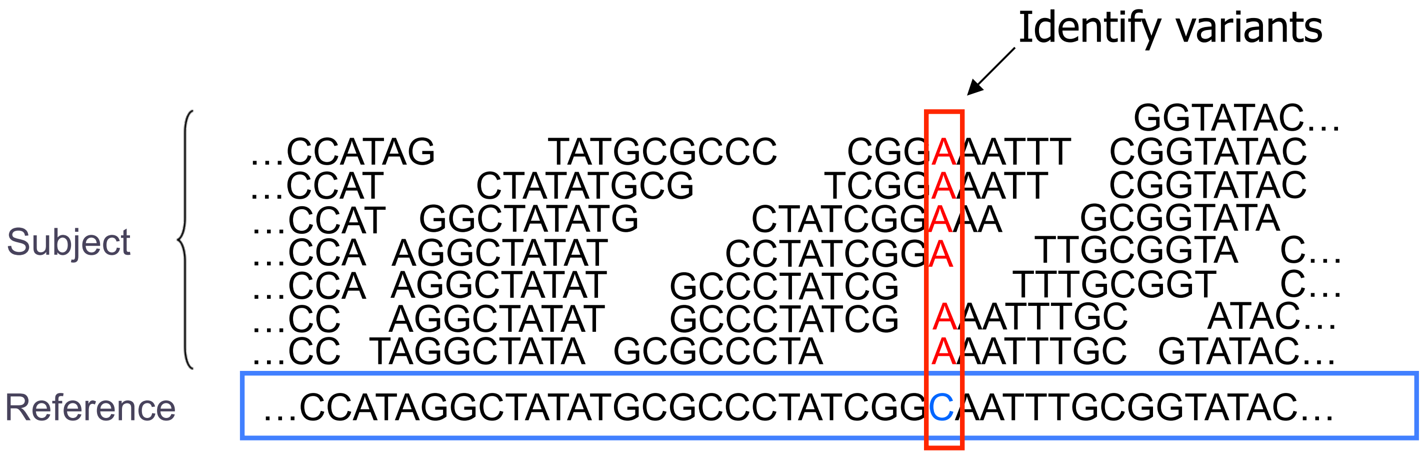


- Challenges
  - Batch computing challenges
  - + Shuffling of huge datasets
- Technologies
  - Hadoop, Elastic MapReduce, Dryad
  - Parallel Databases

Junior High Dance



Short Read Mapping



Identify variants

Subject

```
                                  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…
```

Reference

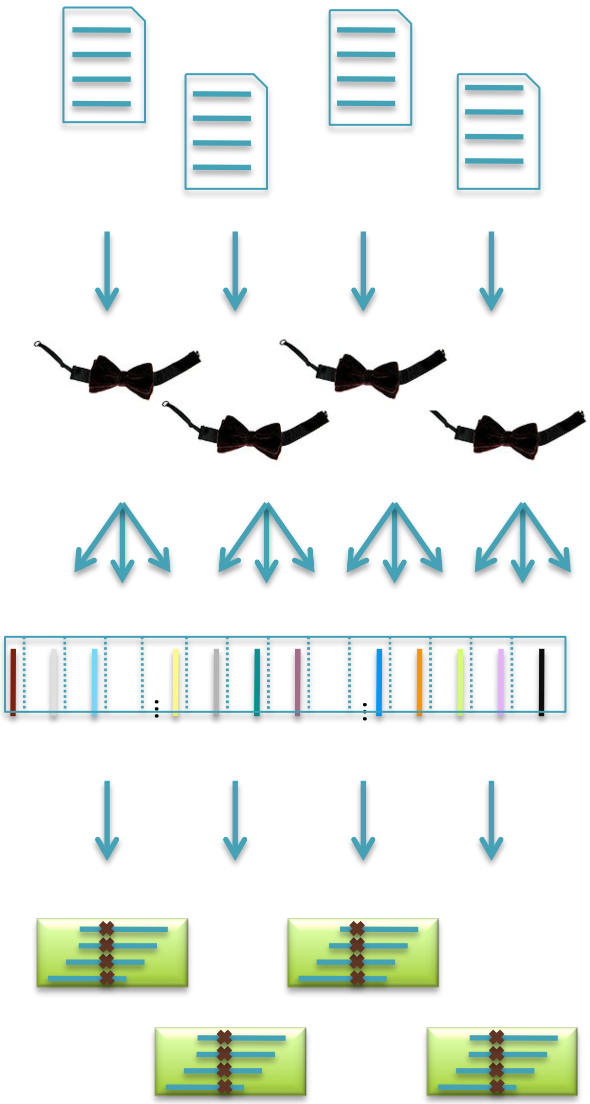
```
…CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC…
```

- 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

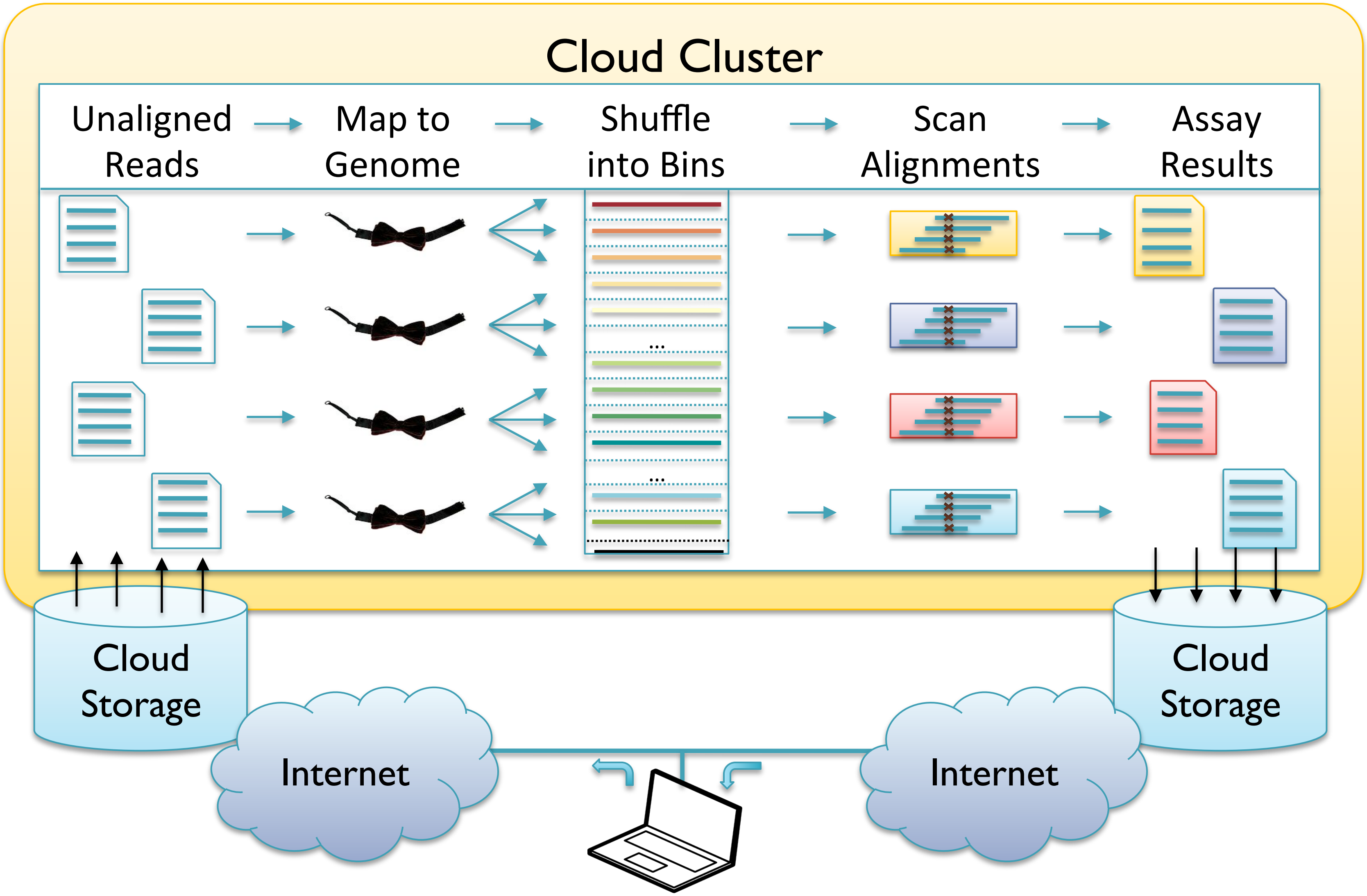
http://bowtie-bio.sourceforge.net/crossbow

	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

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

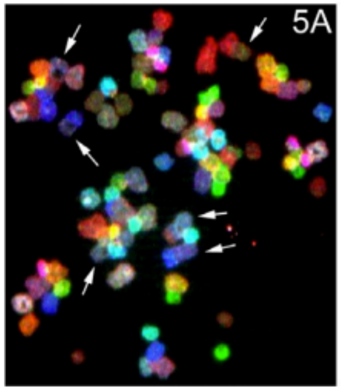
Jnomics case study:

Structural variations in esophageal cancer

- Structural variations are common to many forms of cancer
  - Indels, Inversions, CNVs, Translocations of more than a single basepair
  - "An analysis of available data shows that gene fusions occur in all malignancies, and that they account for 20% of human cancer morbidity."
    - Mitelman *et al.* (2007) The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer.* 7:223-245

- Traditionally identified through cytogenetic imaging & microarrays
  - FISH, CGH, SOMA, etc

- Recent trend is to use sequencing to identify SVs
  - Decreased cost, improved resolution
  - Potential exists for basepair resolution of events

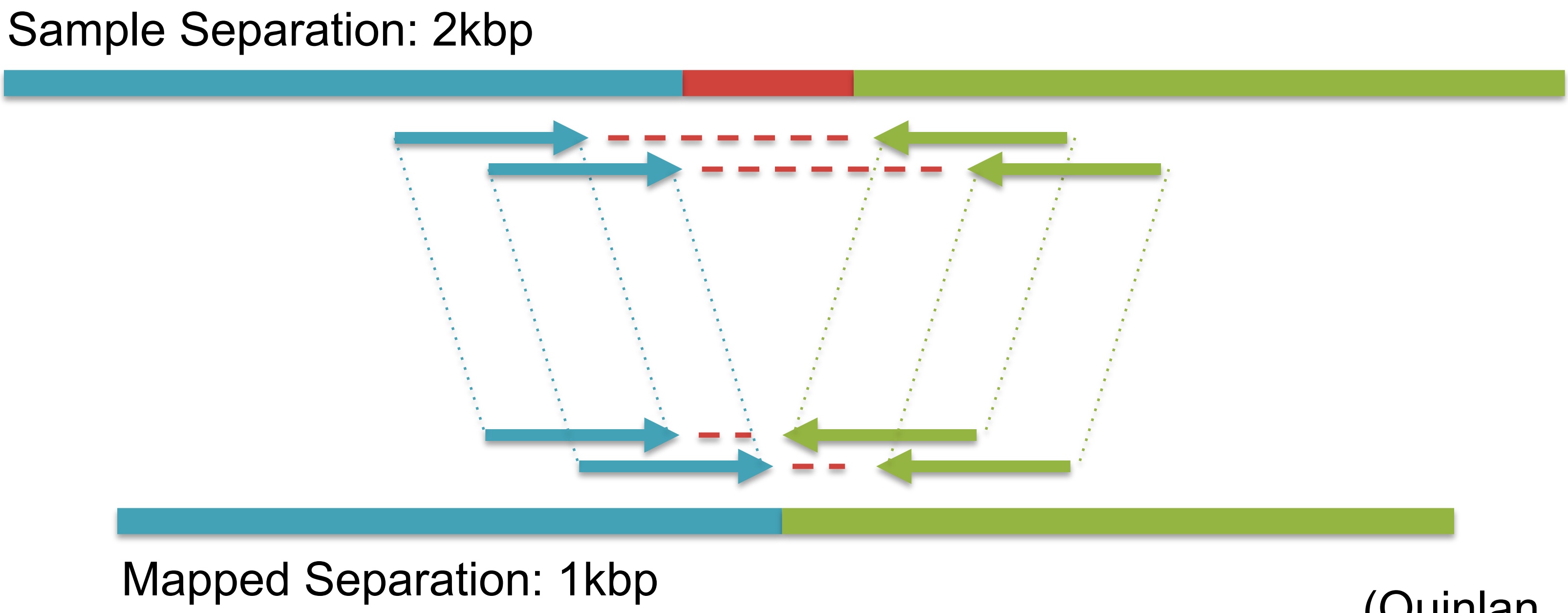


Applications of SKY in cancer cytogenetics
Bayani, JM, Squire, JA (2002) Cancer Invest. 20(3):373-86.

Hydra Discordant Pair Analysis

Illumina sequencing generates reads in pairs from both ends of a fragment with a known separation

1. Sequence diseased sample using paired-end/mate-pair protocol
2. Map reads from sample to reference genome
3. If a pair maps unexpectedly far away or with unexpected orientation, there is a SV between the reads
4. Cluster pairs to pinpoint breakpoints



(Quinlan, 2010)

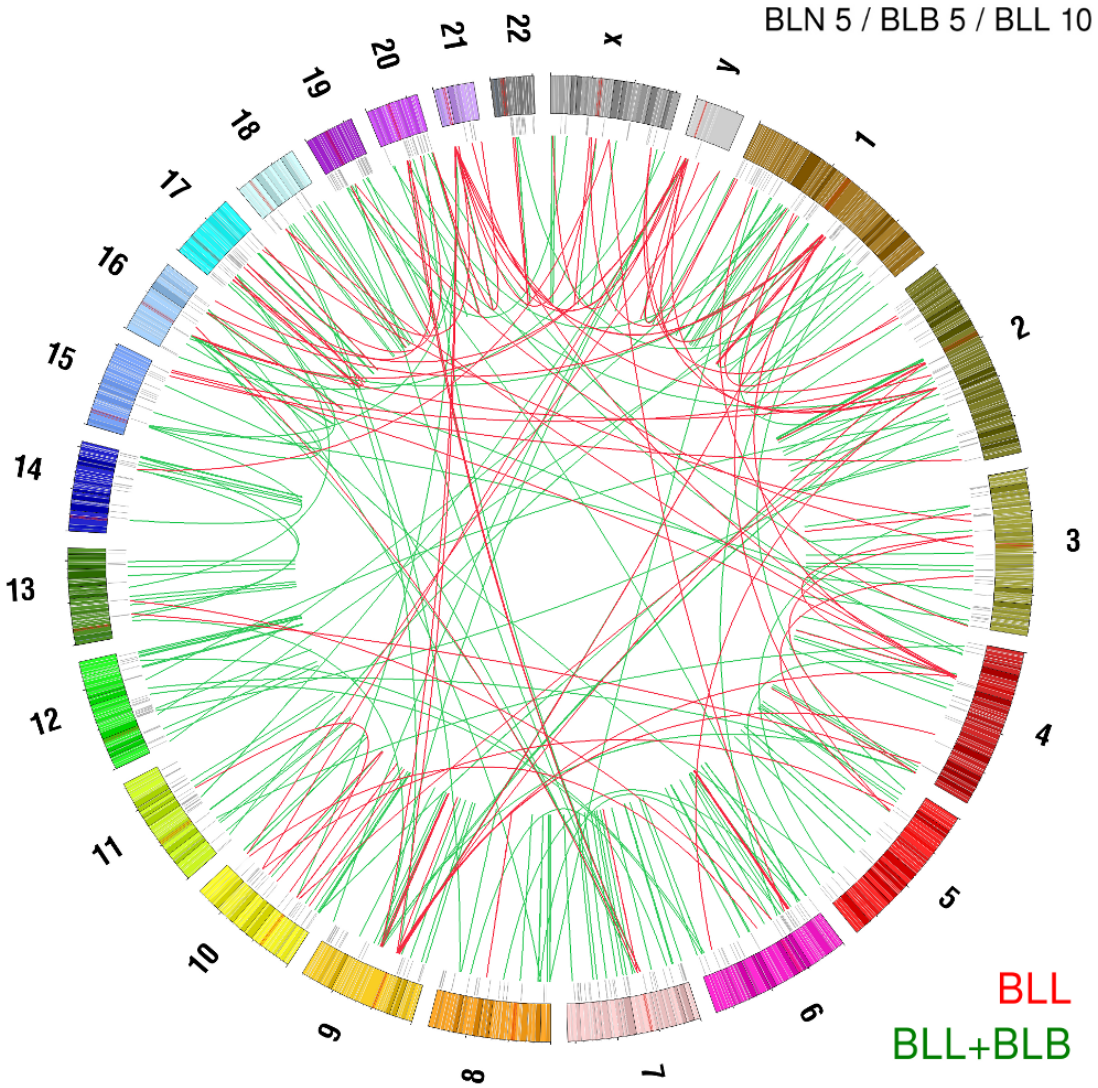
Jnomics Structural Variations

Circos plot of high confidence SVs specific to esophageal cancer sample

- Red: SVs specific to tumor
- Green: SVs in both diseased and tumor samples

Detailed analysis of disrupted genes and fusion genes in progress

- Preliminary analysis shows many promising hits to known cancer genes

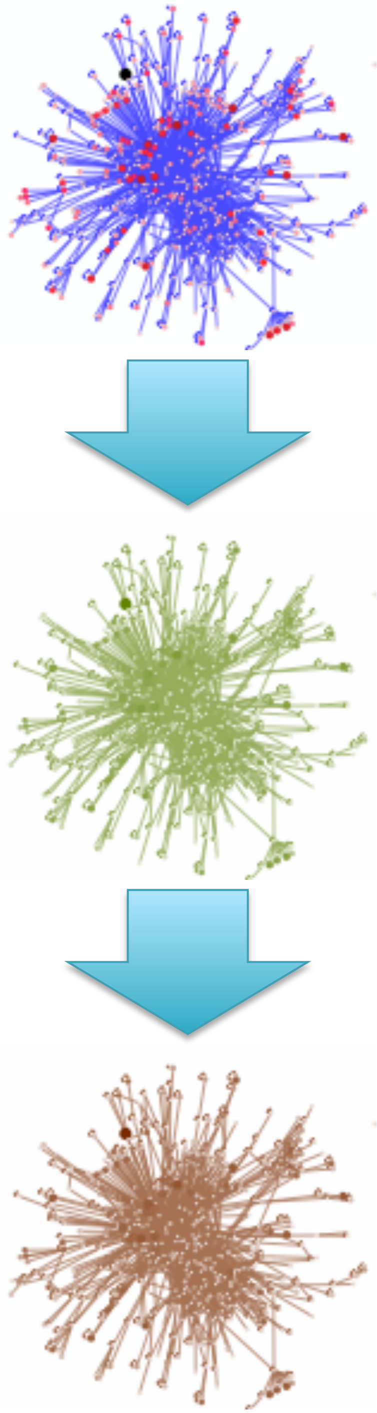


3. Tightly Coupled

- Computation that cannot be partitioned
  - Graph Analysis
  - Molecular Dynamics
  - Population simulations

- Challenges
  - Loosely coupled challenges
  - + Parallel algorithms design

- Technologies
  - MPI
  - MapReduce, Dryad, Pregel



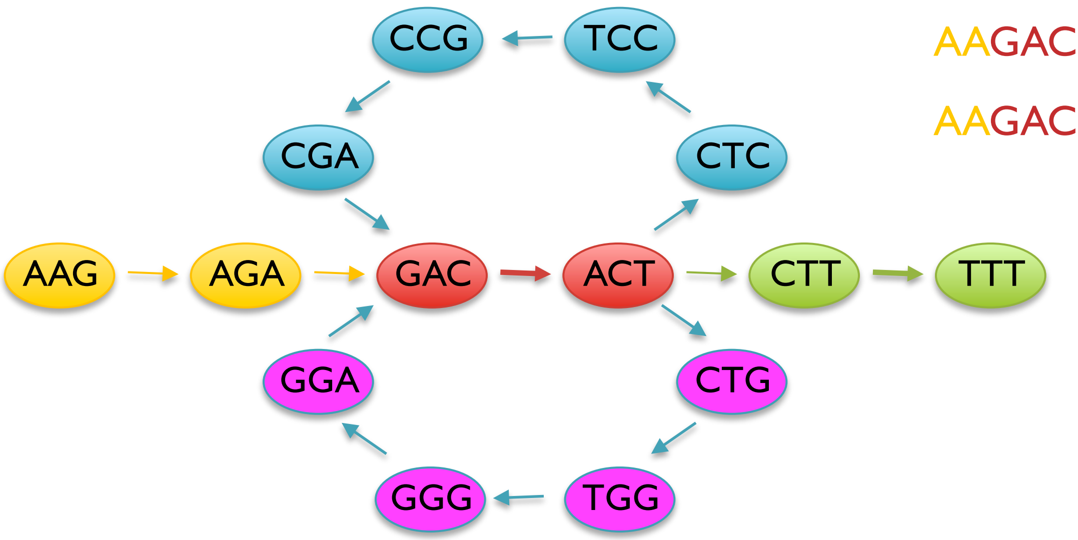
High School Dance

Short Read Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…

de Bruijn Graph



Potential Genomes

AAGACTCCGACTGGGACTTT

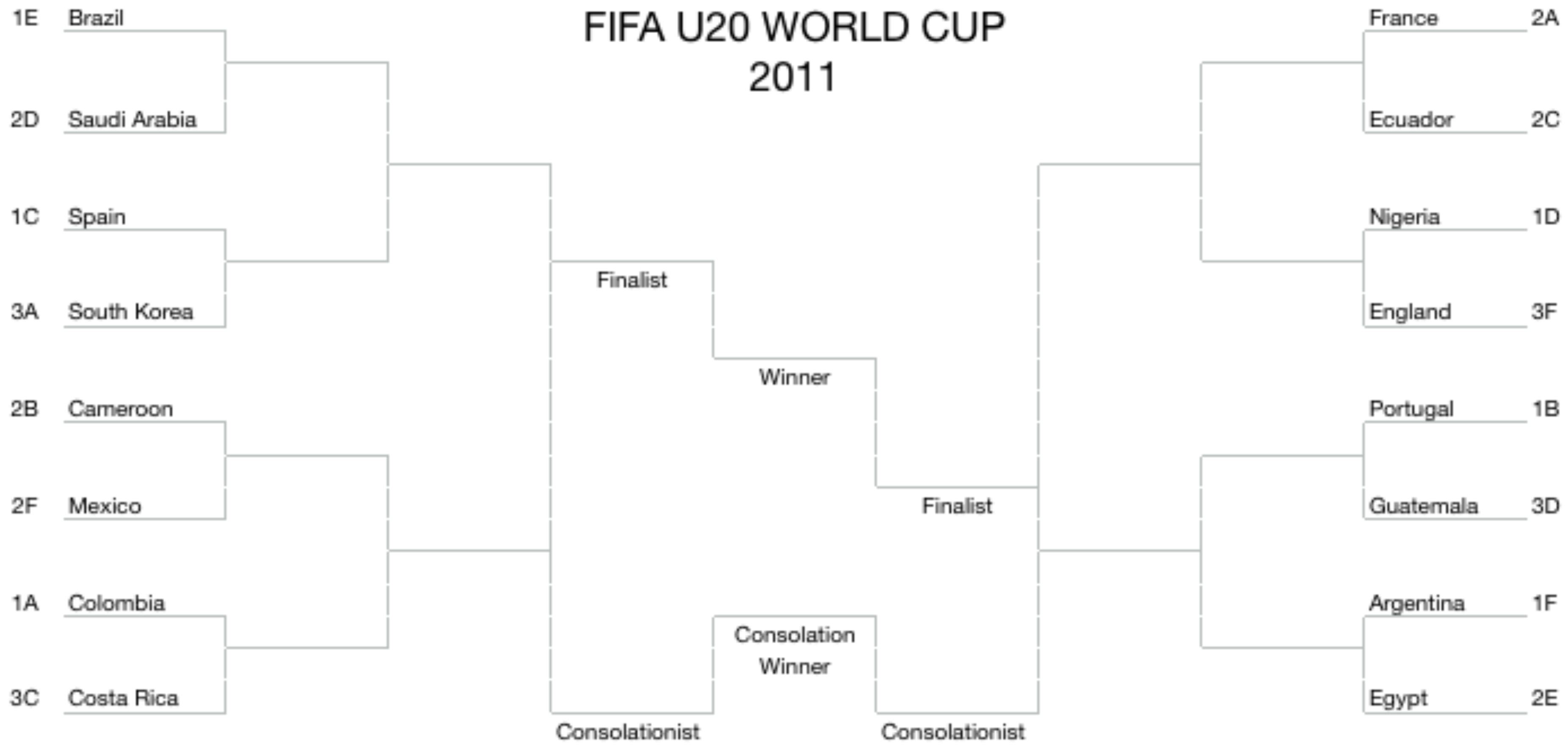
AAGACTGGGACTCCGACTTT

- 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

Warmup Exercise

Who here was born closest to Oct 4?

– You can only compare to 1 other person at a time



FIFA U20 WORLD CUP
2011

1E Brazil

2D Saudi Arabia

1C Spain

3A South Korea

2B Cameroon

2F Mexico

1A Colombia

3C Costa Rica

France 2A

Ecuador 2C

Nigeria 1D

England 3F

Portugal 1B

Guatemala 3D

Argentina 1F

Egypt 2E

Finalist

Winner

Finalist

Consolation
Winner

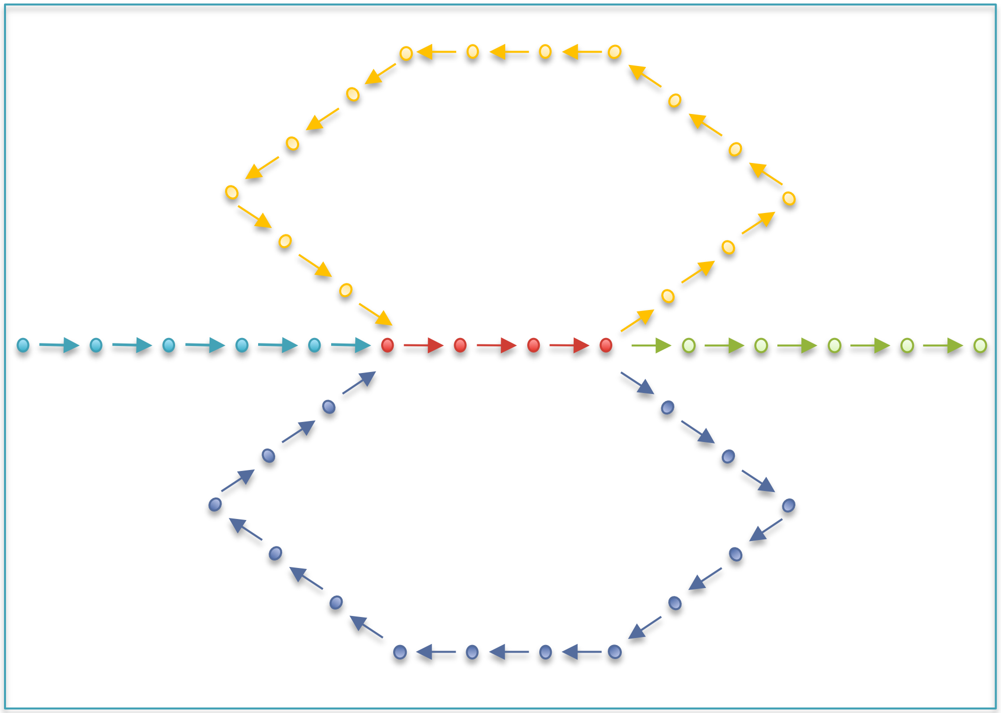
Consolationist

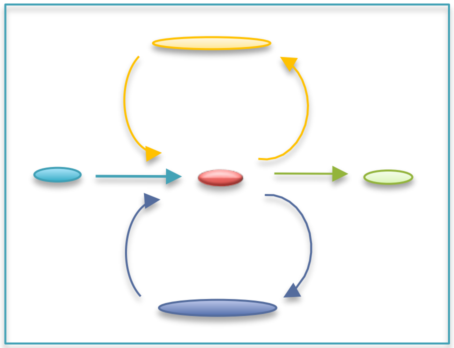
Consolationist

Find winner among 16 teams in just 4 rounds

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)

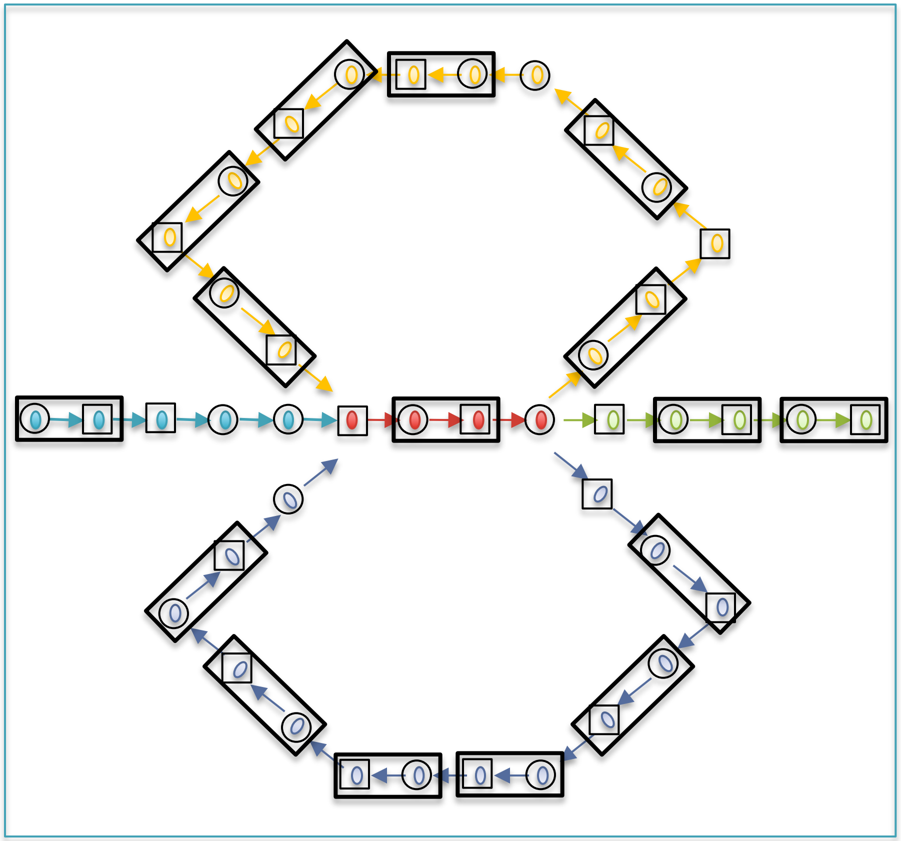
Fast Path Compression

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



Initial Graph: 42 nodes

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

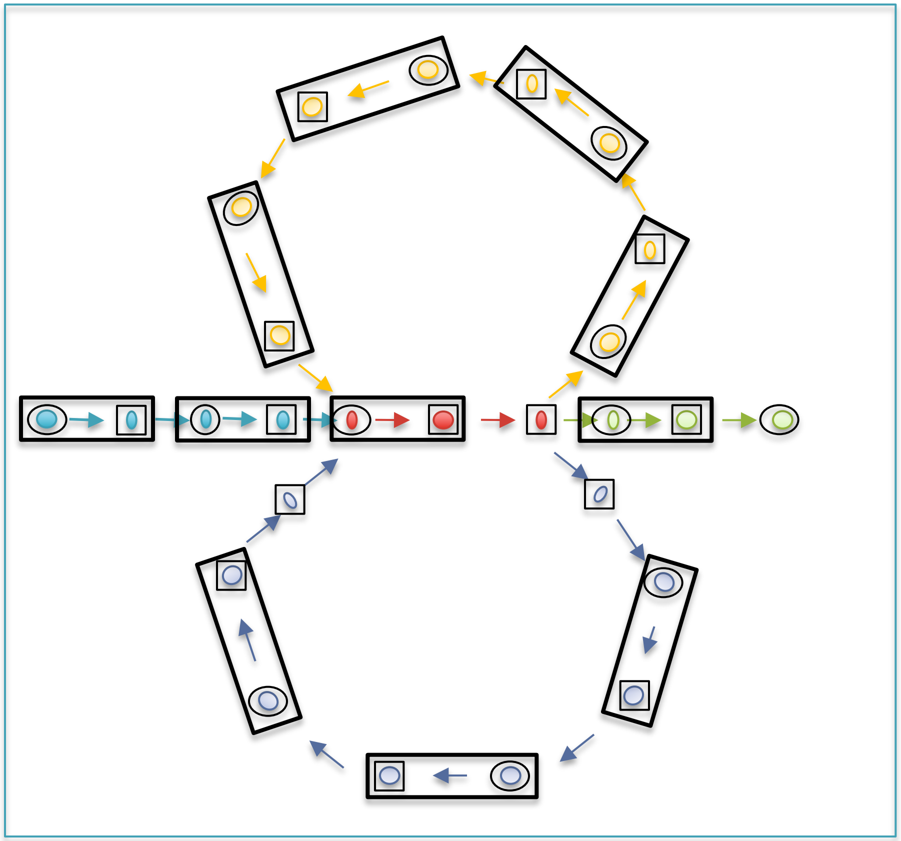
Fast Path Compression

Challenges

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

Randomized List Ranking

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

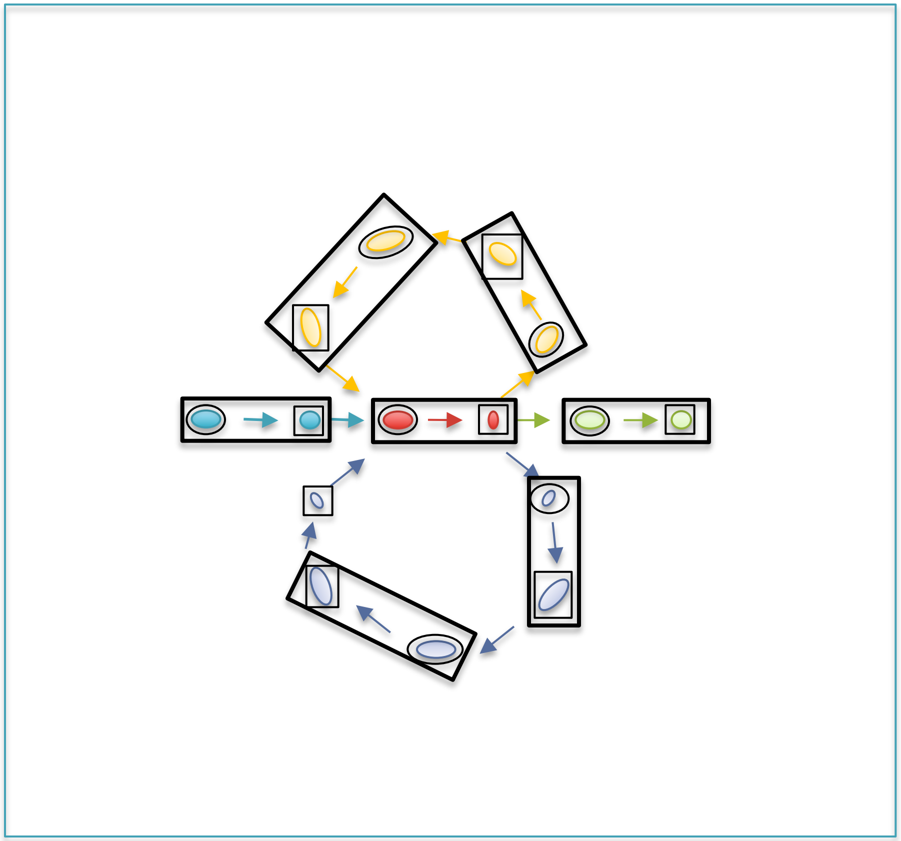
Fast Path Compression

Challenges

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

Randomized List Ranking

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



Round 2: 15 nodes (64% savings)

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

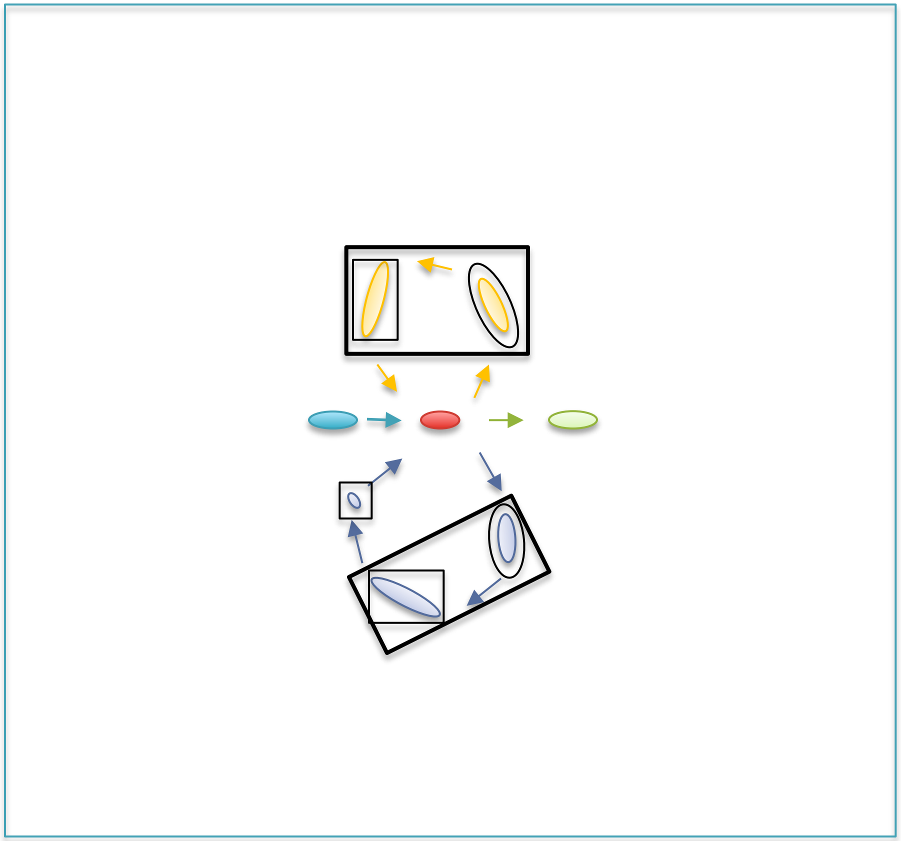
Fast Path Compression

Challenges

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

Randomized List Ranking

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



Round 2: 8 nodes (81% savings)

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

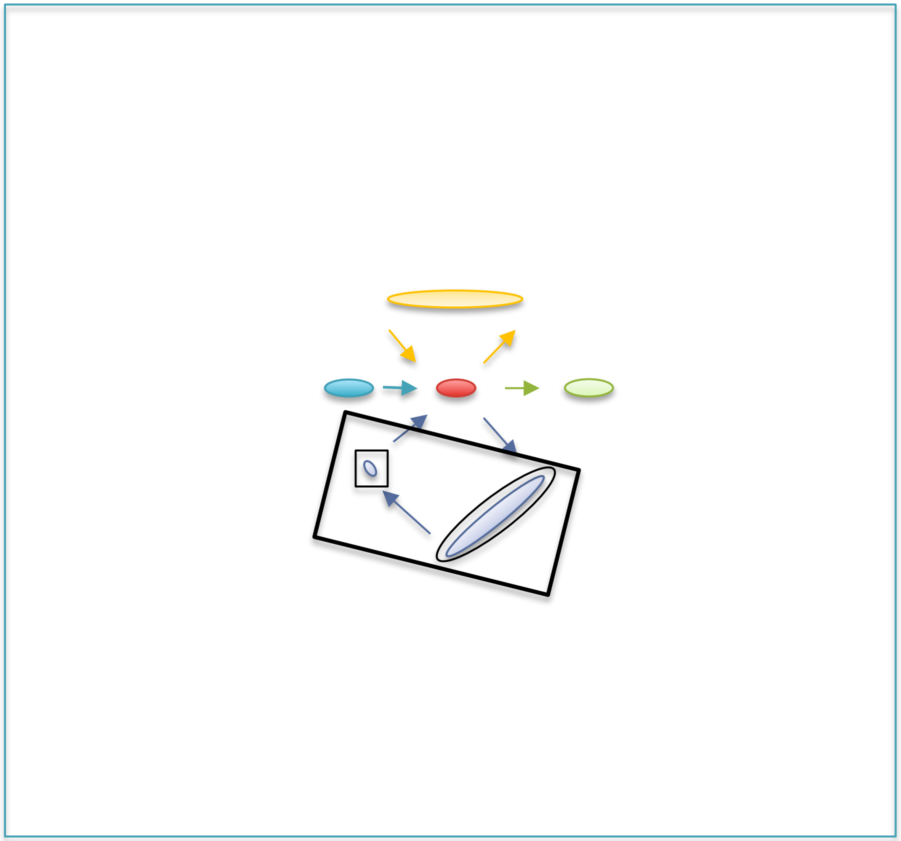
Fast Path Compression

Challenges

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

Randomized List Ranking

- Randomly assign Ⓗ/ $\boxed{\text{T}}$ 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

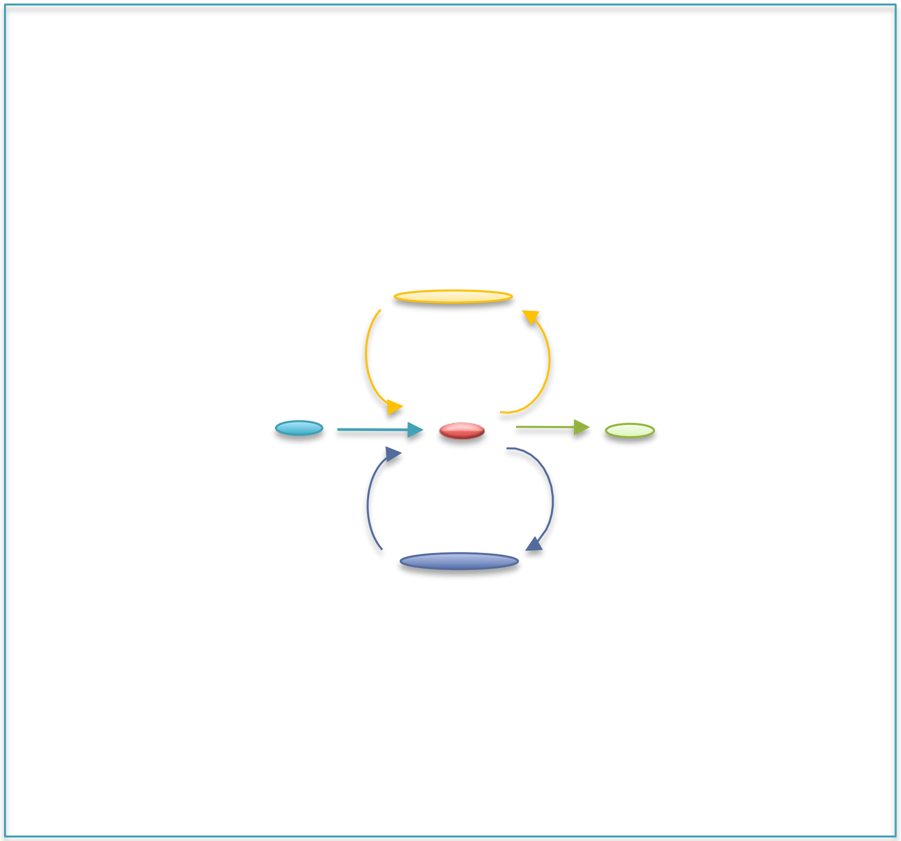
- Nodes stored on different computers
- Nodes can only access direct neighbors

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)

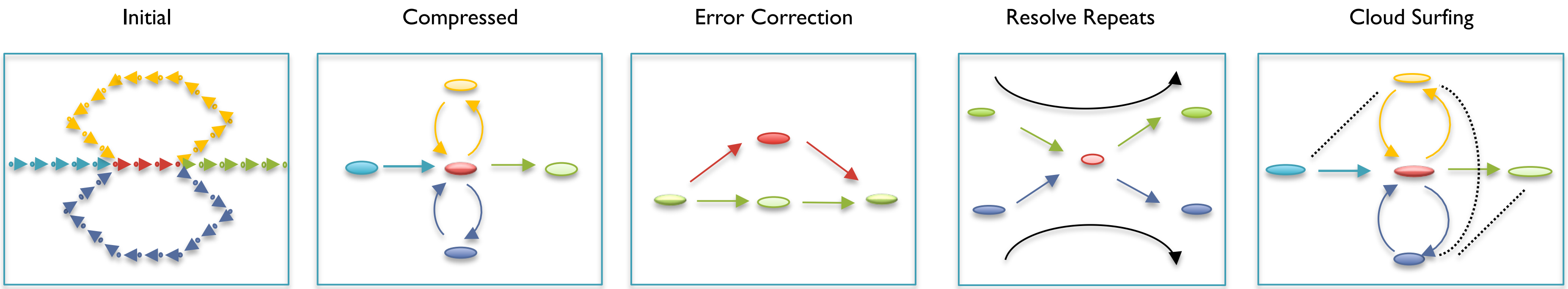
Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

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

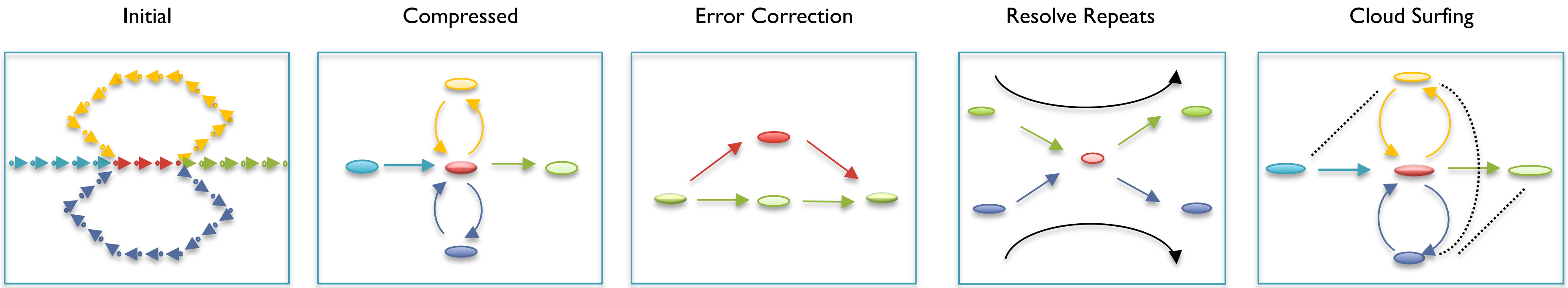
Assembly of Large Genomes with Cloud Computing.
Schatz MC, Sommer D, Kelley D, Pop M, *et al. In Preparation.*

Contrail

http://contrail-bio.sourceforge.net

De novo Assembly of the Human Genome

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



Initial

Compressed

Error Correction

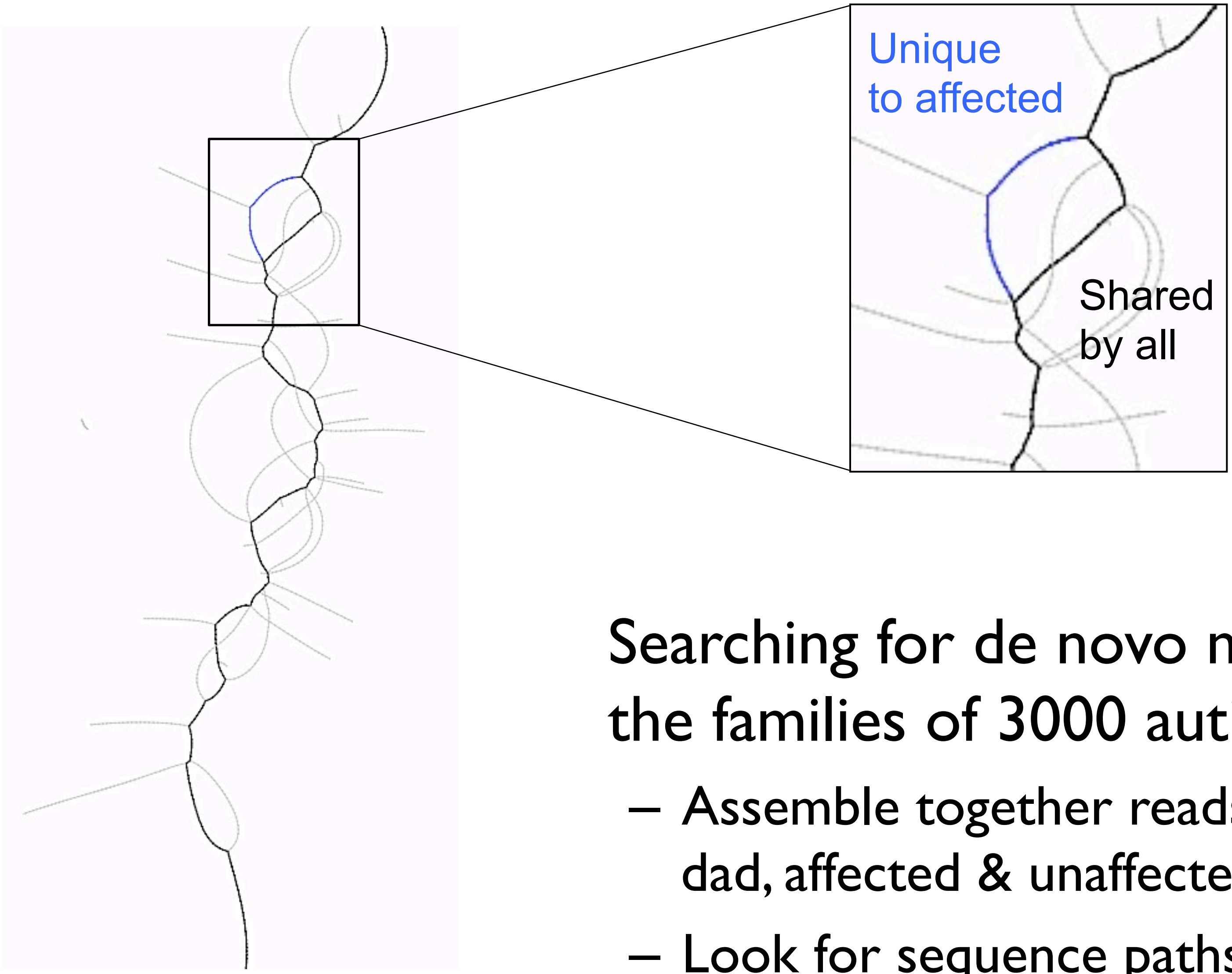
Resolve Repeats

Cloud Surfing

N	>7 B	>1 B	4.2 M	4.1 M	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

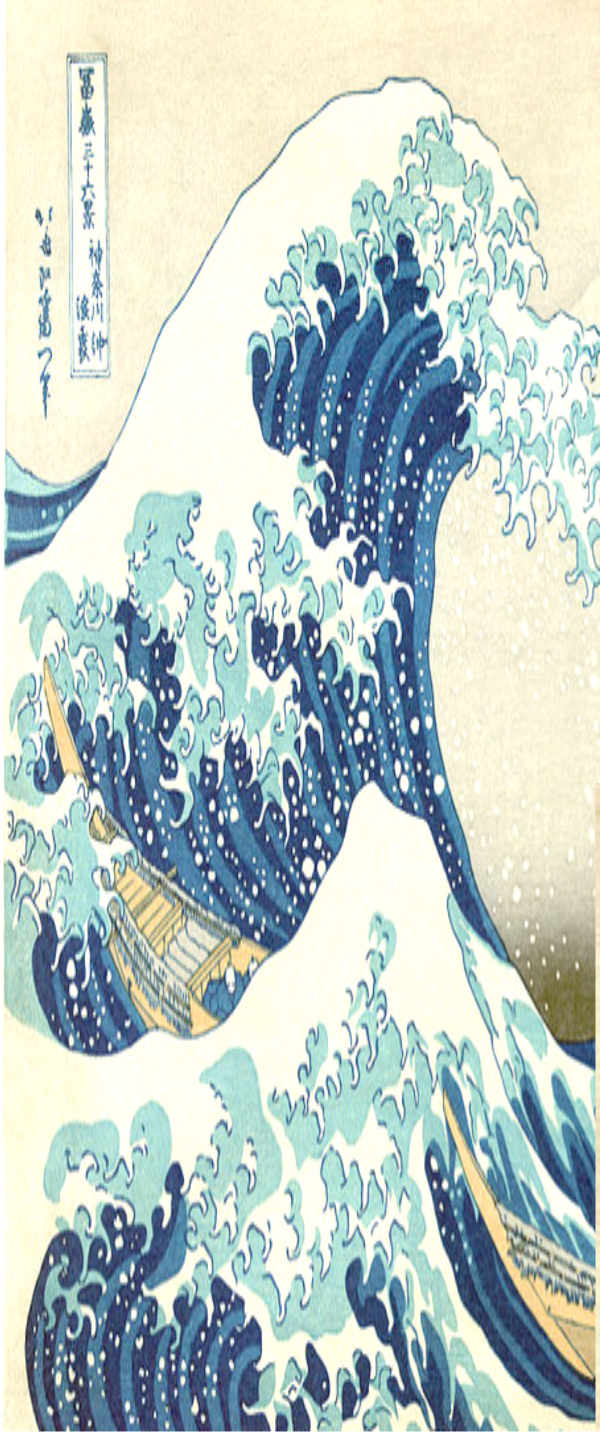


COLEC12
C->A

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



Summary

- We are entering 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

- Emerging technologies are a great start, but we need continued research
  - Need integration across disciplines

Acknowledgements

Schatzlab

Mitch Bekritsky
Matt Titmus
Hayan Lee
James Gurtowski
Anirudh Aithal
Rohith Menon
Goutham Bhat

CSHL

Dick McCombie
Melissa Kramer
Eric Antonio
Mike Wigler
Zach Lippman
Doreen Ware
Ivan Iossifov

JHU

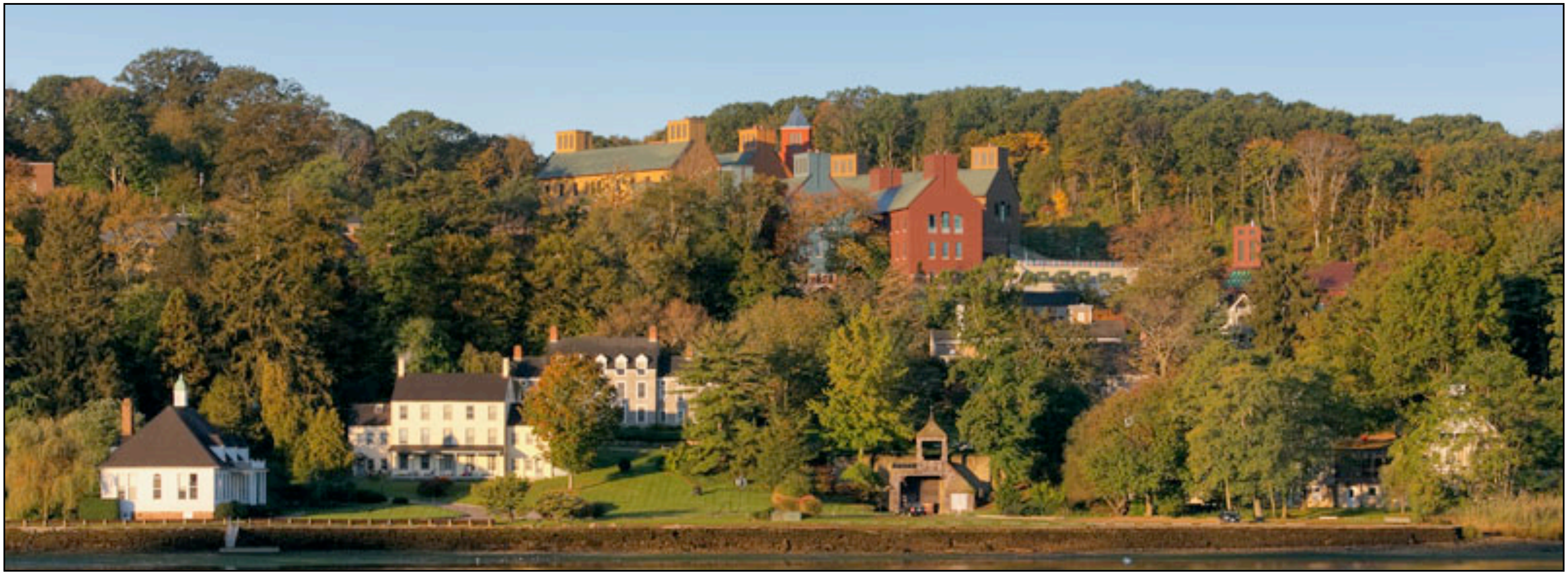
Steven Salzberg
Ben Langmead
Jeff Leek

NBACC

Adam Phillipy
Sergey Koren

Univ. of Maryland

Mihai Pop
Art Delcher
Jimmy Lin
David Kelley
Dan Sommer
Cole Trapnell



Thank You!

http://schatzlab.cshl.edu
@mike_schatz