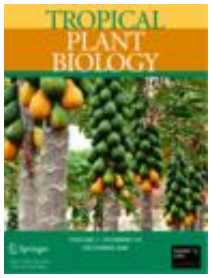
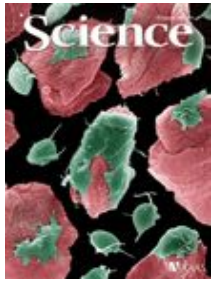
Schatzlab Research Projects

Michael Schatz

Sept 9, 2011
Research Topics in Biology, WSBS

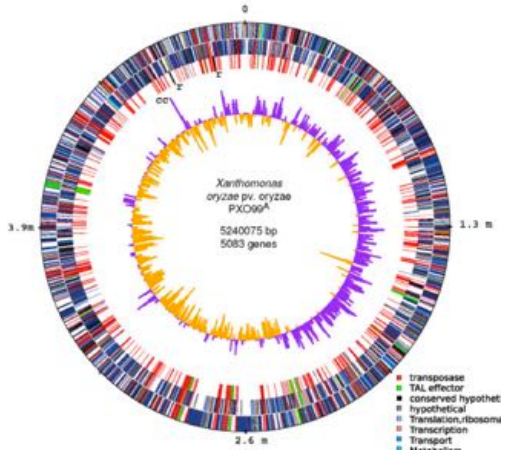
A Little About Me



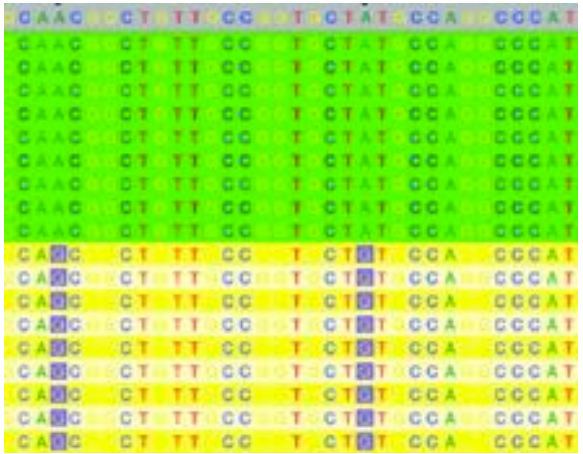


Genomics & Quantitative Biology

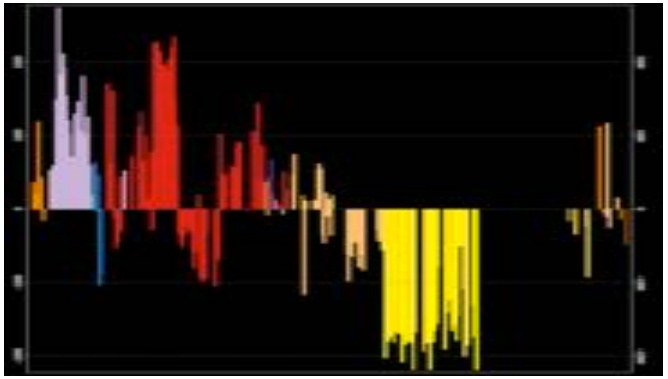
Genome Assembly



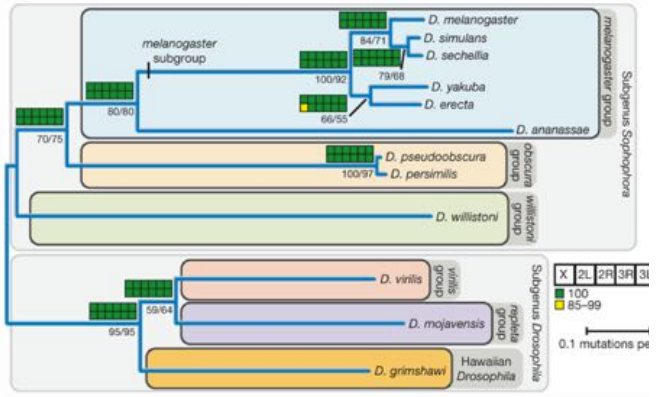
Mutations & Disease



Differential Analysis

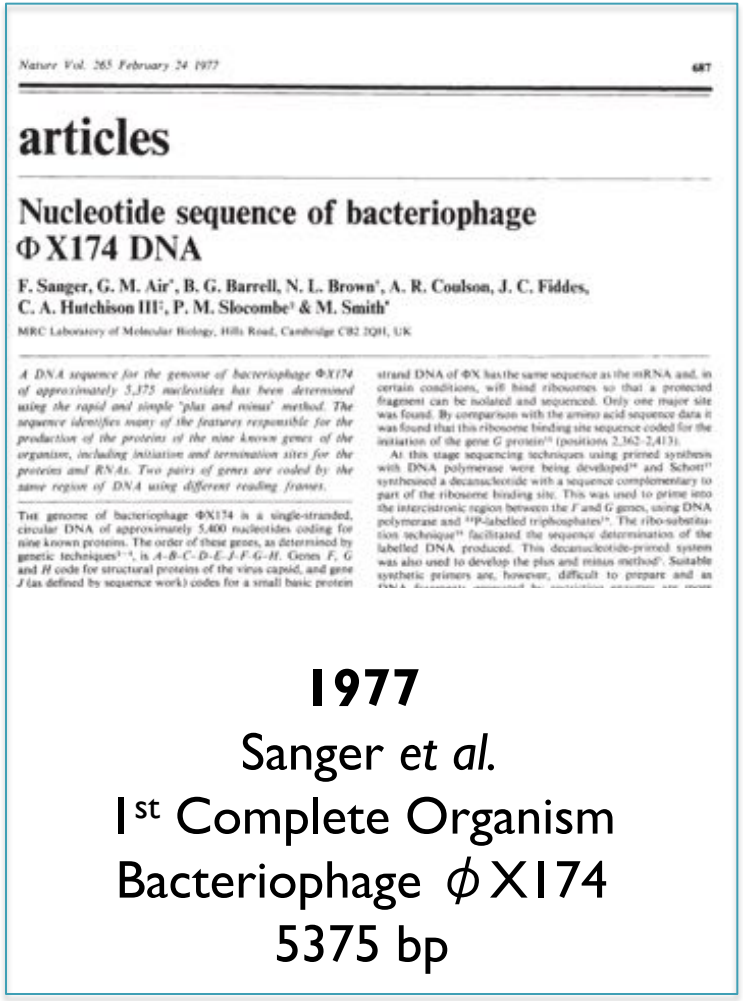


Phylogeny & Evolution



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[1–5], 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

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[14] and Schott[15] 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[16]. The ribo-substitution technique[17] 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 an

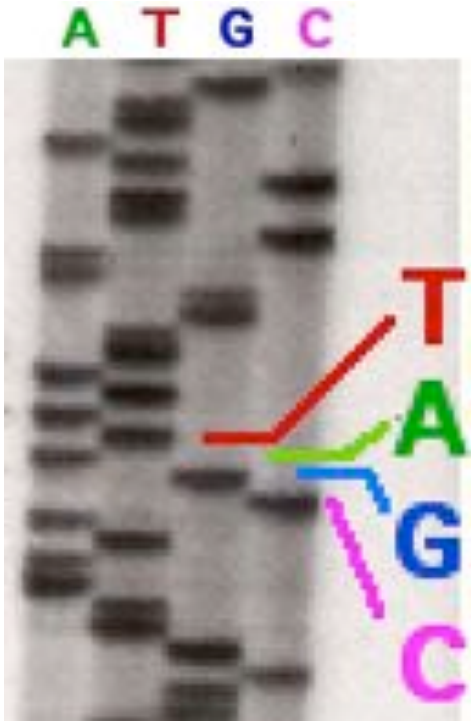
1977

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

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

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

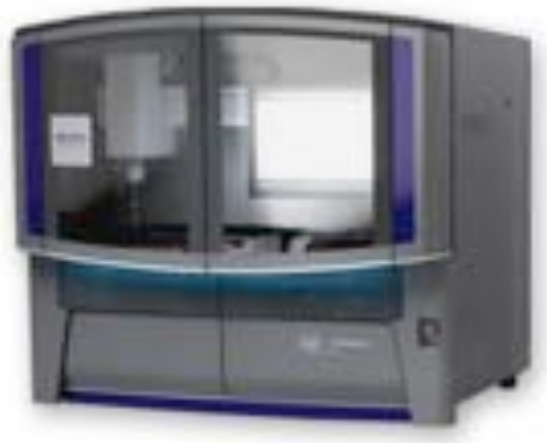
2001
Venter *et al.*,
Human Genome
Celera Assembler. 2.9 Gbp

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

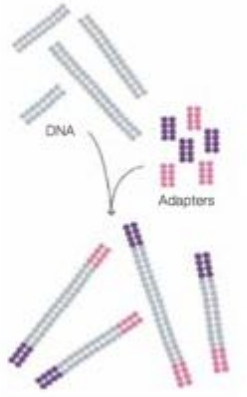
2004
454/Roche
Pyrosequencing
Current Specs (Titanium):
1M 400bp reads / run =
1Gbp / day

2007
Illumina
Sequencing by Synthesis
Current Specs (HiSeq 2000):
2.5B 100bp reads / run =
60Gbp / day

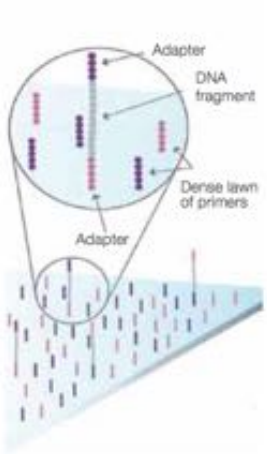


2008
ABI / Life Technologies
SOLiD Sequencing
Current Specs (5500xl):
5B 75bp reads / run =
30Gbp / day

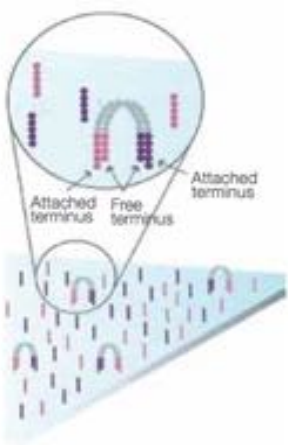
Illumina Sequencing by Synthesis



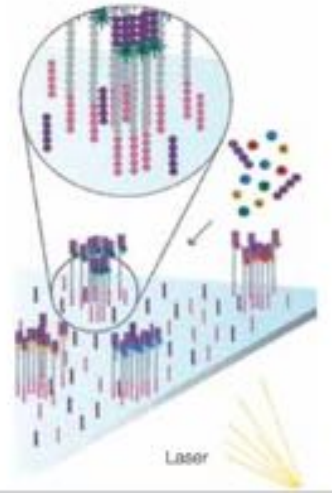
1. Prepare



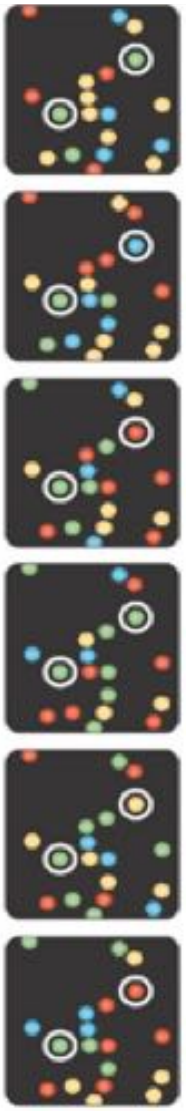
2. Attach



3. Amplify



4. Image



5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

The DNA Data Race

Year	Genome	Technology	Cost
2001	Venter *et al.*	Sanger (ABI)	$300,000,000
2007	Levy *et al.*	Sanger (ABI)	$10,000,000
2008	Wheeler *et al.*	Roche (454)	$2,000,000
2008	Ley *et al.*	Illumina	$1,000,000
2008	Bentley *et al.*	Illumina	$250,000
2009	Pushkarev *et al.*	Helicos	$48,000
2009	Drmanac *et al.*	Complete Genomics	$4,400

(Pushkarev *et al.*, 2009)

Sequencing a single human genome uses ~100 GB of compressed sequence data in billions of short reads.

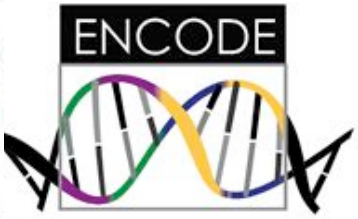
~20 DVDs / genome

Sequencing Centers

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

The DNA Data Tsunami





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

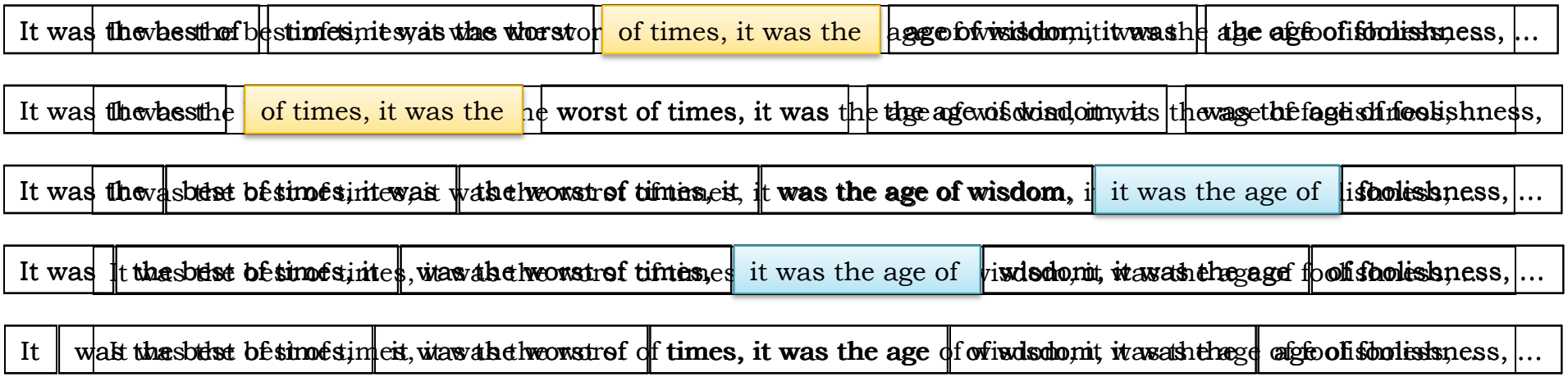
Our best (only) hope is to use many computers:

- Parallel Computing aka Cloud Computing

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

Warmup 1: Shredded Book Reconstruction

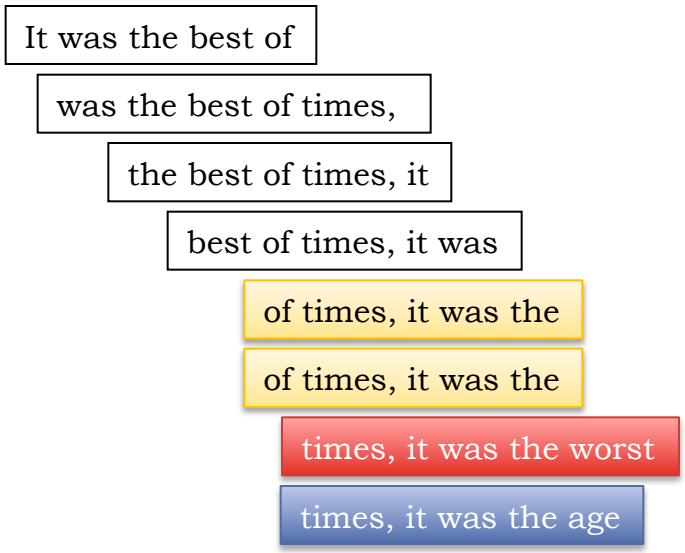
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



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

Greedy Reconstruction

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



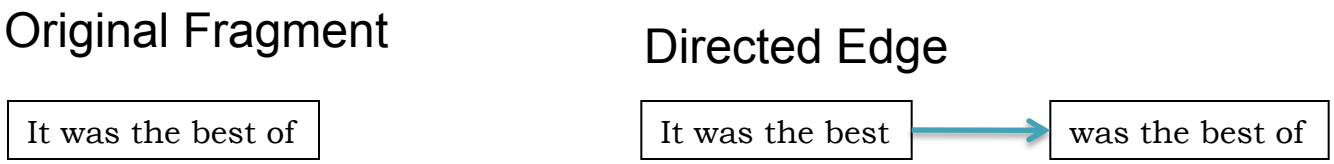
The repeated sequence make the correct reconstruction ambiguous

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

Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

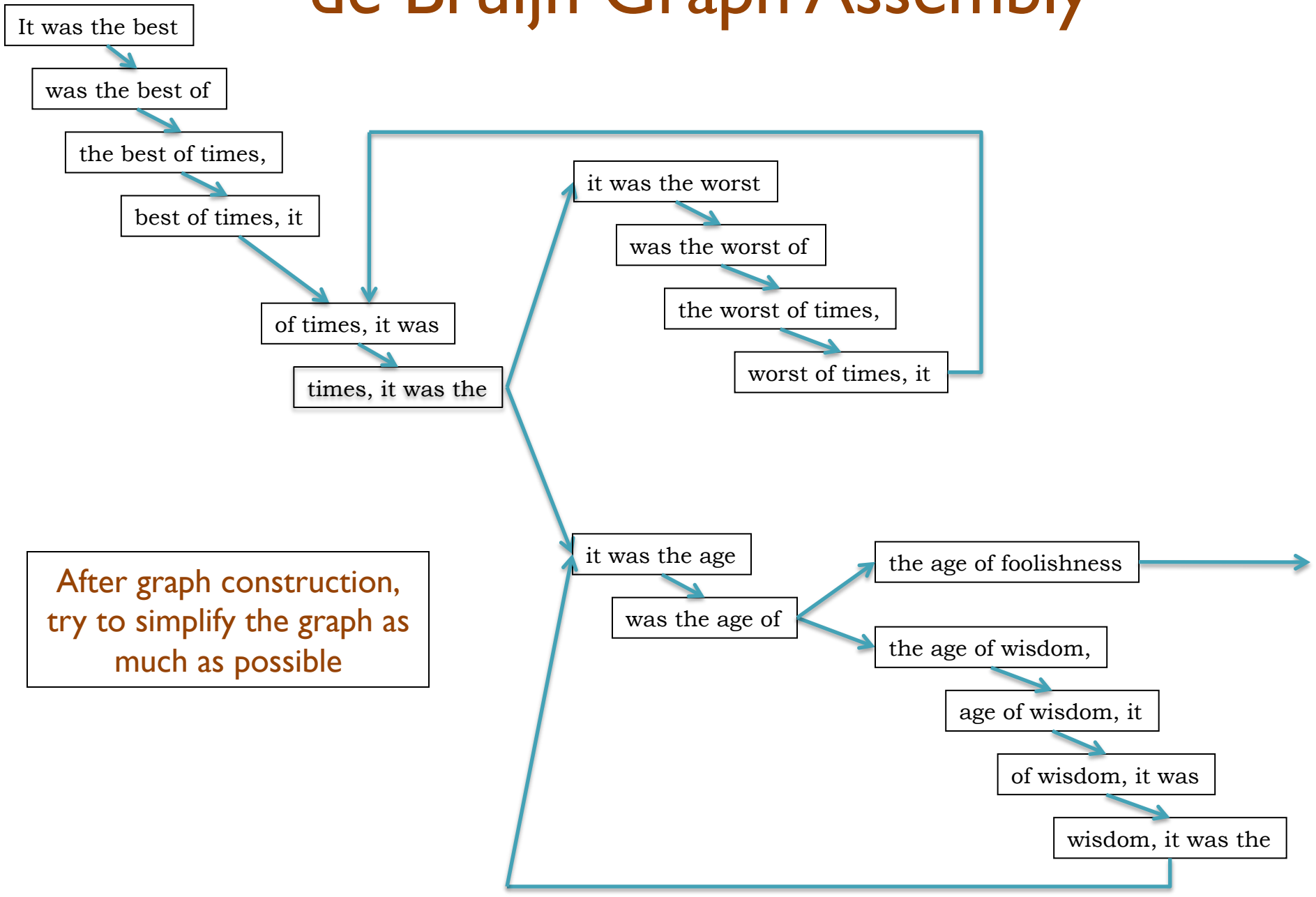
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



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

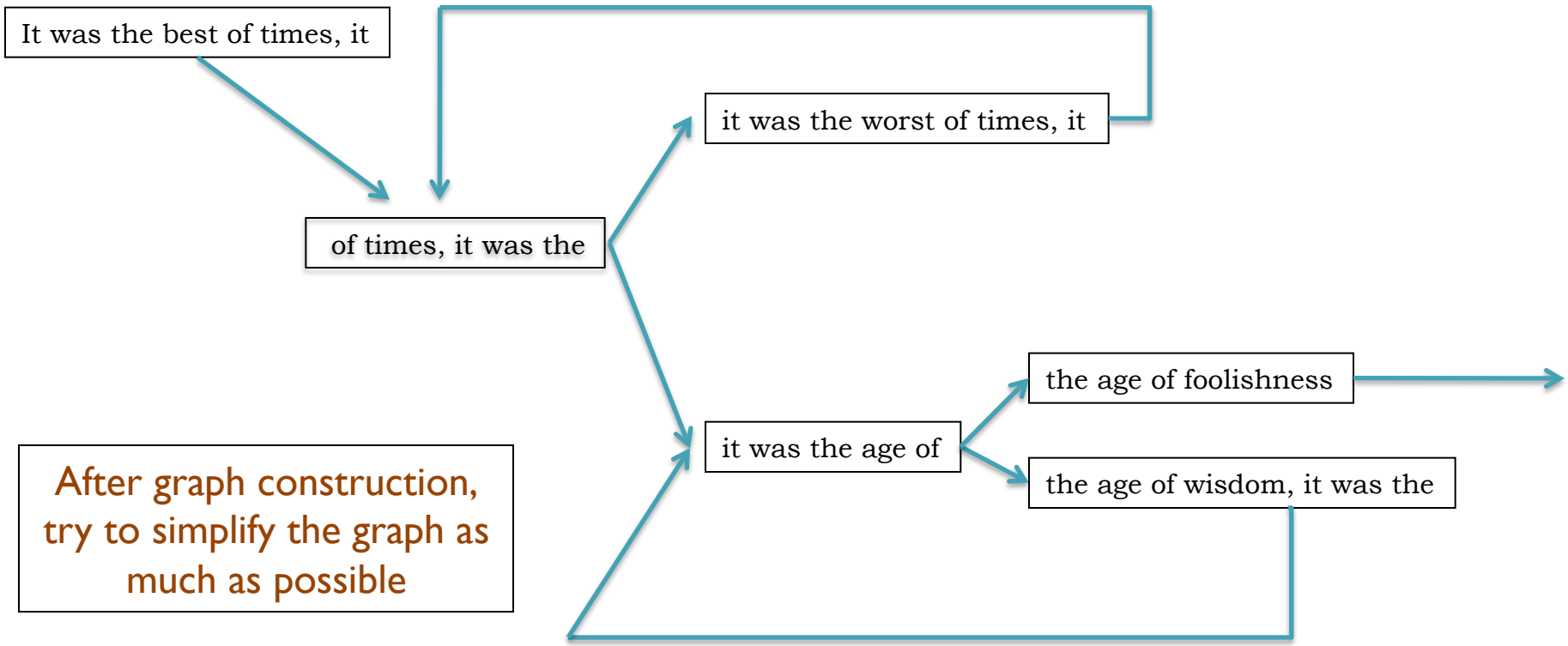
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



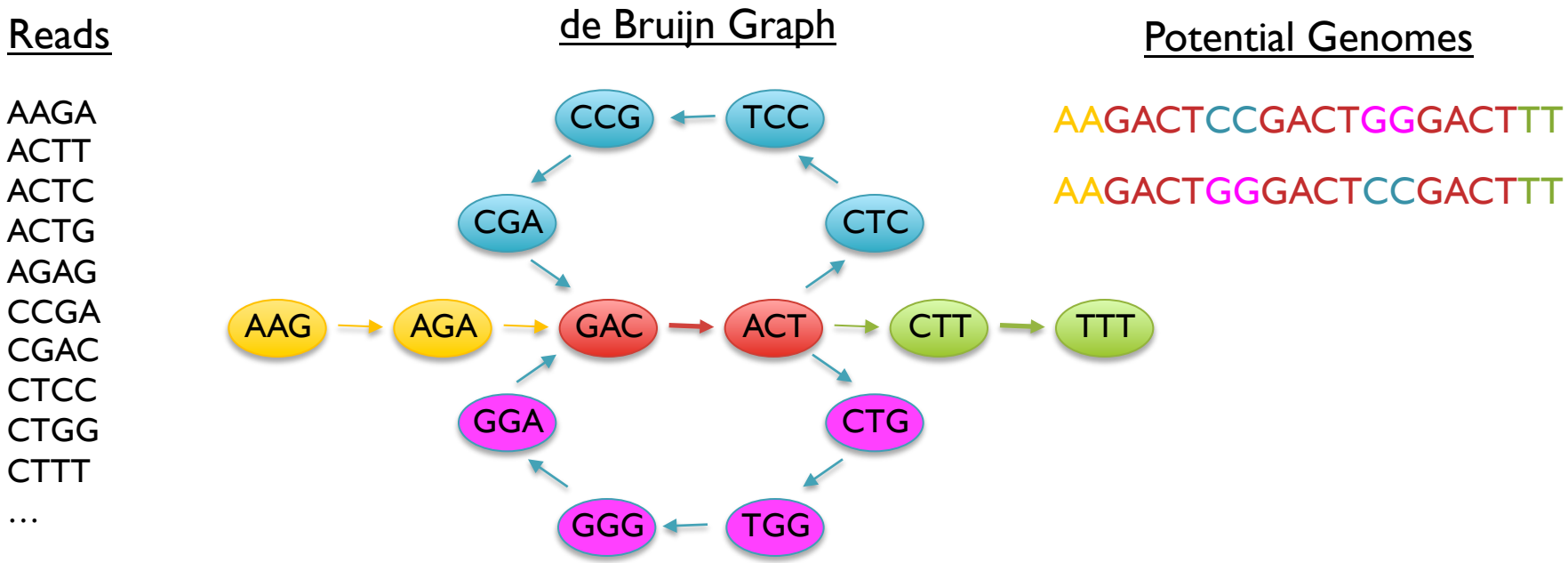
After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

Genome Assembly



- 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 2: Birthday Matching

- Who here was born closest to Sept 9?
  - You can only compare to 1 other person at a time

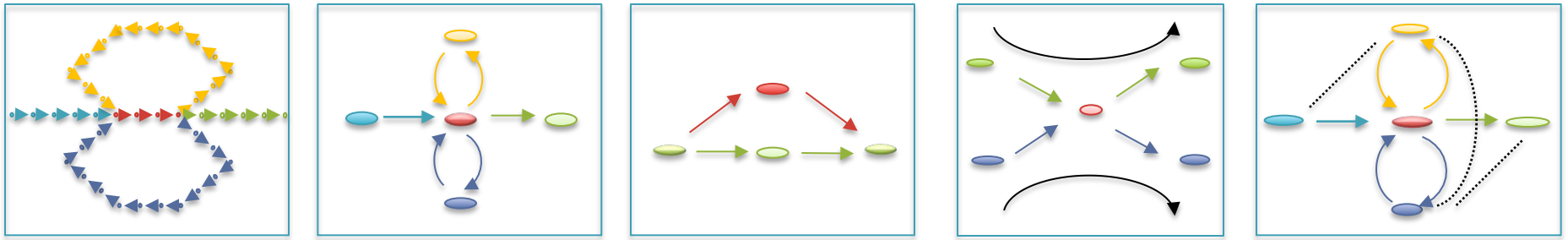
Find winner among 64 teams in just 6 rounds

Contrail

http://contrail-bio.sourceforge.net

De novo Assembly of the Human Genome

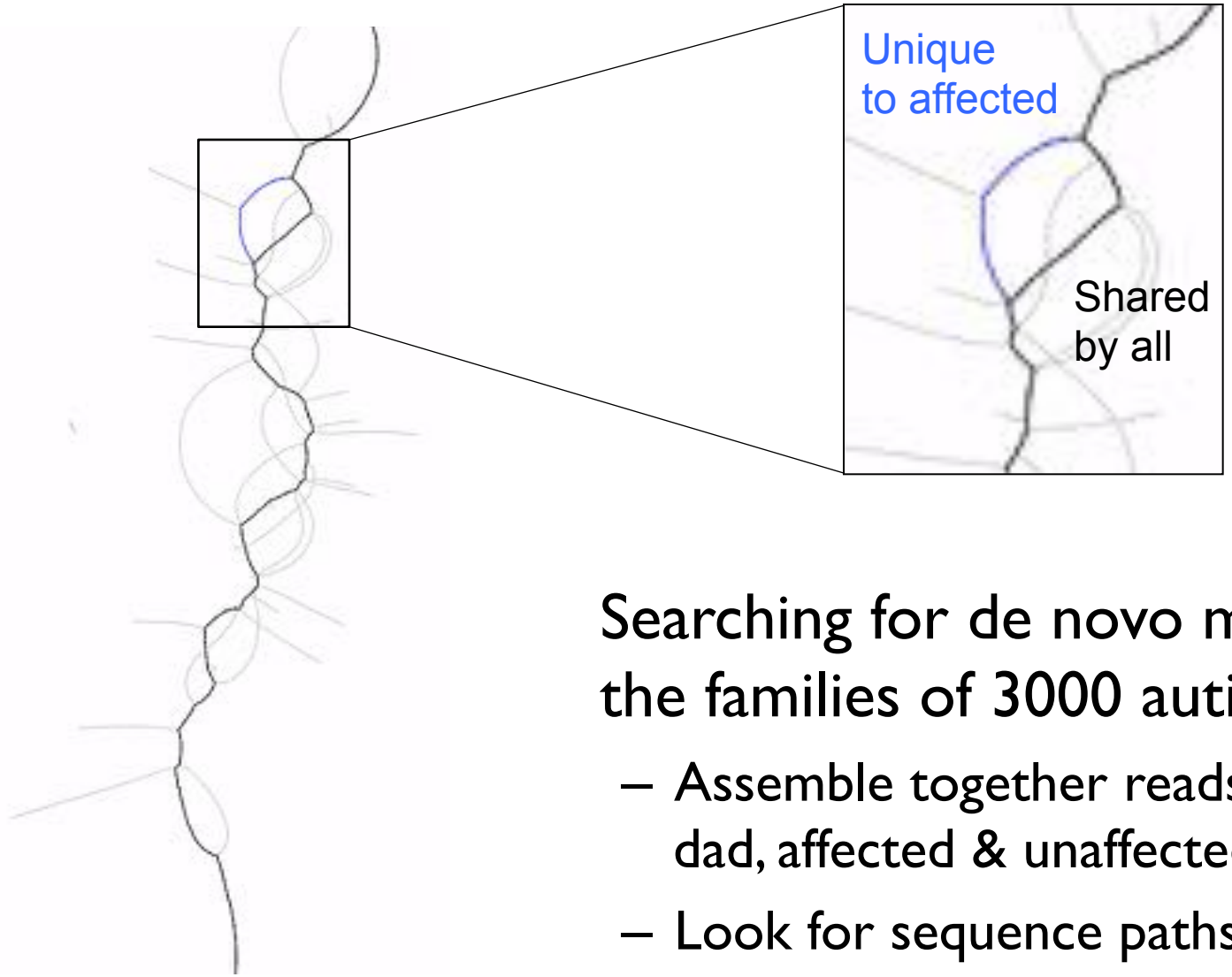
- *Genome:* African male NA18507 (SRA000271, Bentley *et al.*, 2008)
- *Input: 3.5*B 36bp reads, 210bp insert (~40x coverage)



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



COLEC12
C->A

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

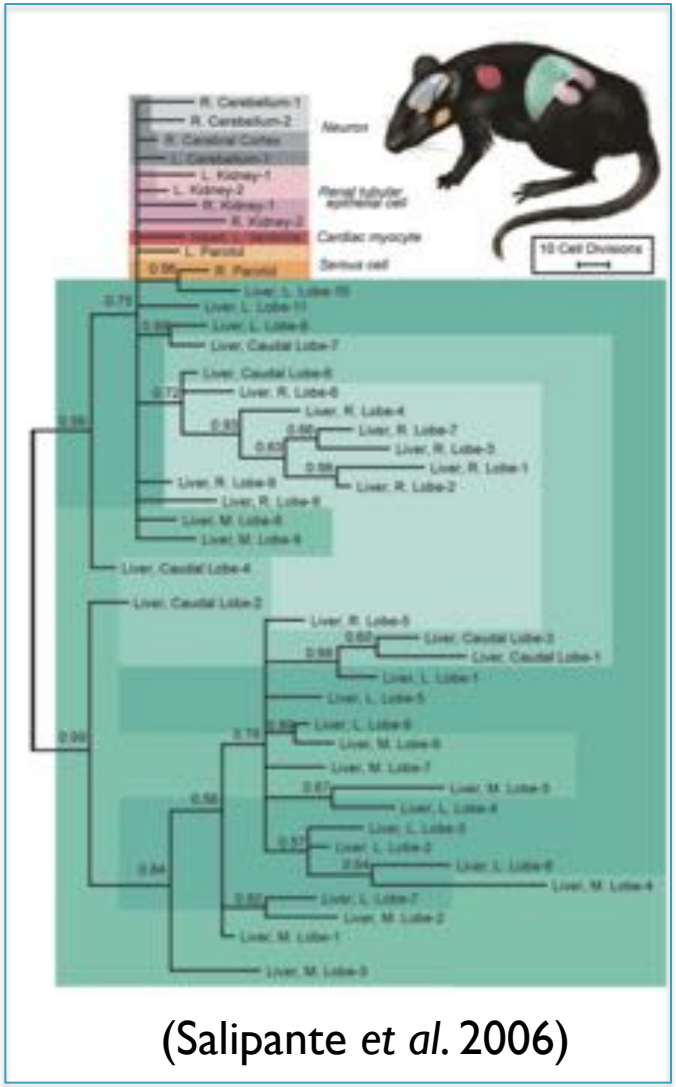
- Assemble together reads from mom, dad, affected & unaffected children
- Look for sequence paths unique to affected child

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 SeqMS
  - Rapidly detect MS sequences
  - Map reads using a new MS-mapper
  - 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)

Structural Variations in Cancer

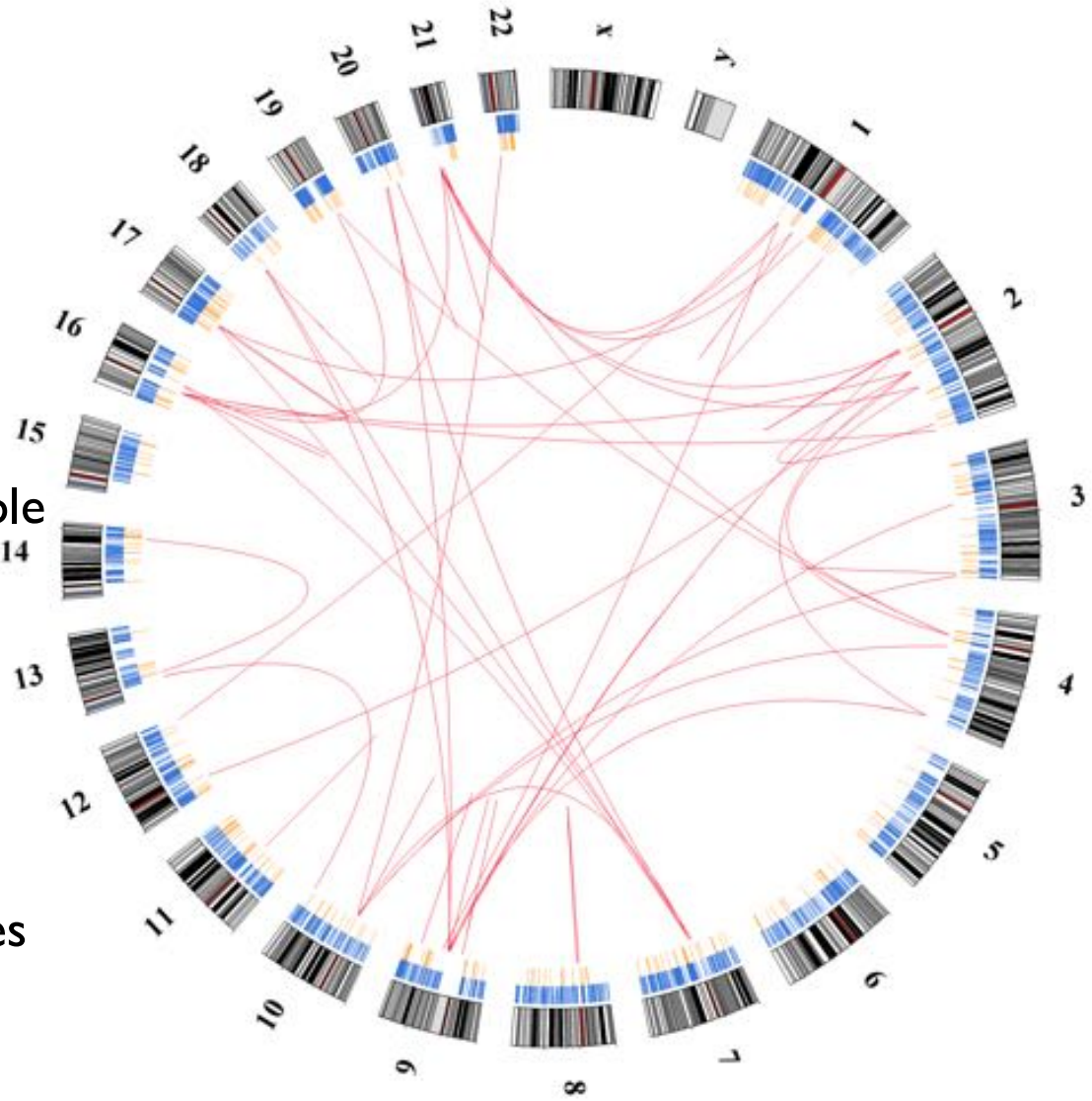
Use short reads to discover large scale variations

- Discordant Pairs Analysis with Hydra (Quinlan et al. 2010)

Circos plot of high confidence SVs specific to esophageal cancer sample

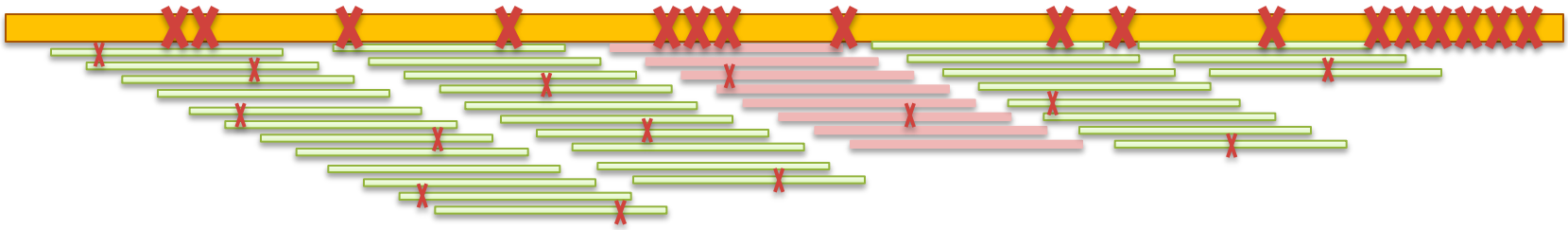
- Red: SV links
- Orange: 375 cancer genes
- Blue: 4950 disease genes

Detailed analysis of disrupted genes and fusion genes in progress



Illumina/PacBio Hybrid Assembly

1. Trim/correct SR sequence
2. Compute an SR layout for each LR
   1. Map SRs to LRs
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR
3. Co-assemble corrected LRs and SRs
   – Celera Assembler enhanced to support 16 Kbp reads



A hybrid strategy for utilizing single-molecule sequencing data for genome assembly and RNA-Seq. Koren, S, Walenz, BP, Martin, J, Jarvis, ED, Rasko, DA, Schatz, MC, McCombie, WR, Phillippy, AM. (2011) *In preparation.*

Summary

- We are entering the digital age of biology
  - Next generation sequencing, microarrays, mass spectrometry, microscopy, ecology, etc
- 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

Schatzlab
Matt Titmus
Hayan Lee
Mitch Bekritsky
Paul Baranay
Rohith Menon
Goutham Bhat
James Gurtowski

CSHL
Lippman Lab
McCombie Lab
Ware Lab
Wigler Lab

NBACC
Adam Phillippy
Sergey Koren

UMD
Steven Salzberg
Mihai Pop
Ben Langmead
Cole Trapnell

Thank You