

# Schatzlab Research Projects

Michael Schatz

Sept 9, 2011

Research Topics in Biology, WSBS

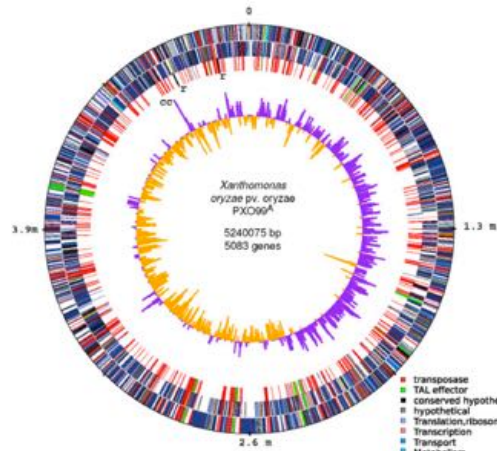


# A Little About Me

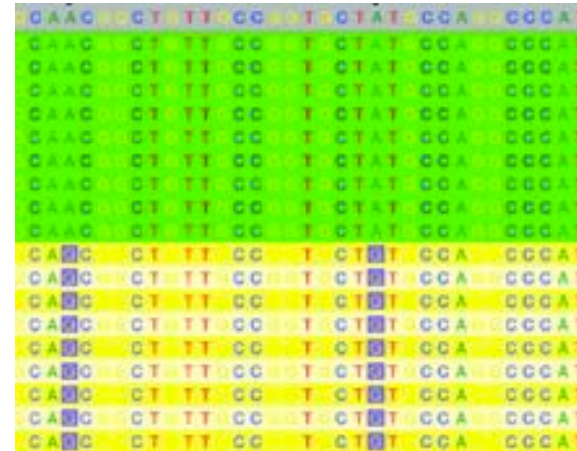


# Genomics & Quantitative Biology

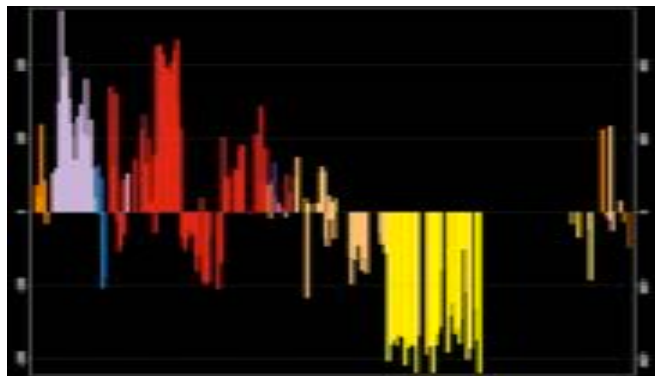
## Genome Assembly



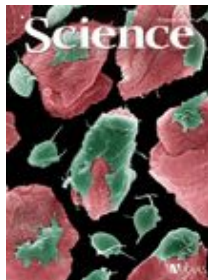
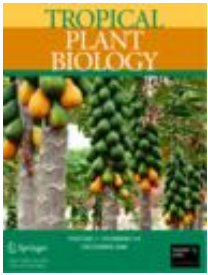
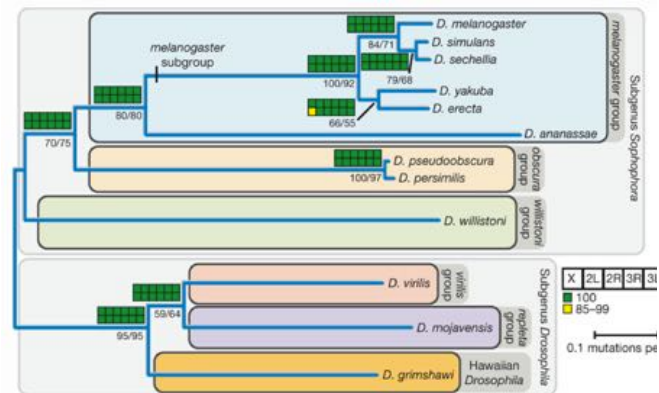
## Mutations & Disease



## 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 $\phi$ X174 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  $\phi$ 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 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  $\phi$ 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 techniques<sup>1-4</sup>, is A-B-C-D-E-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  $\phi$ X 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<sup>5</sup> (positions 2,362-2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed<sup>6,7</sup> and Schott<sup>8</sup> 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 <sup>32</sup>P-labelled triphosphates<sup>9</sup>. The ribo-substitution technique<sup>10</sup> facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method<sup>11</sup>. Suitable synthetic primers are, however, difficult to prepare and an

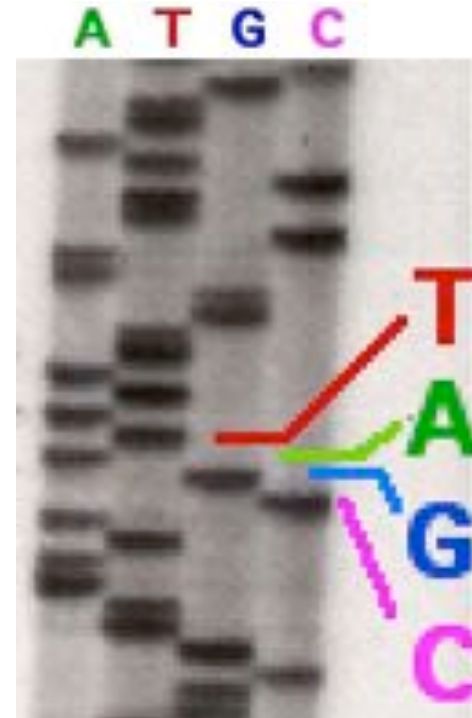
**1977**

Sanger et al.

1<sup>st</sup> Complete Organism

Bacteriophage  $\phi$  X174

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



**1995**

Fleischmann *et al.*  
1<sup>st</sup> Free Living Organism  
TIGR Assembler. 1.8Mbp



**2000**

Myers *et al.*  
1<sup>st</sup> 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

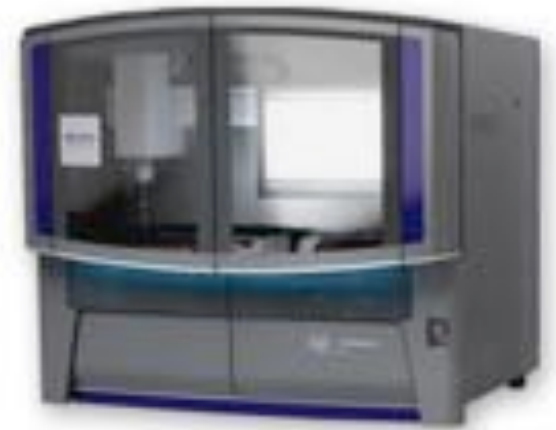
# Milestones in DNA Sequencing



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

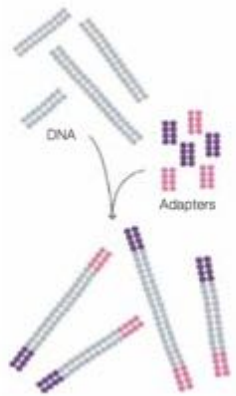


**2007**  
Illumina  
*Sequencing by Synthesis*  
Current Specs (HiSeq 2000):  
2.5B 100bp reads / run =  
60Gbp / day

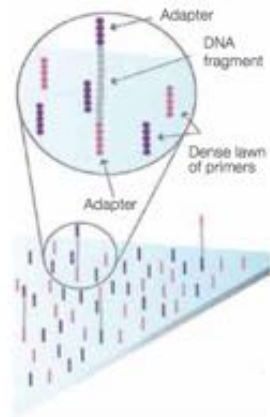


**2008**  
ABI / Life Technologies  
*SOLiD Sequencing*  
Current Specs (5500xl):  
5B 75bp reads / run =  
30Gbp / day

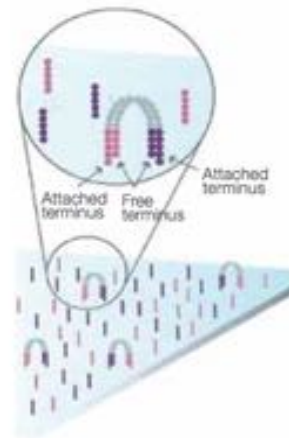
# Illumina Sequencing by Synthesis



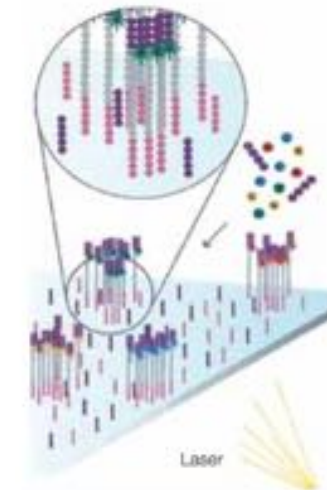
1. Prepare



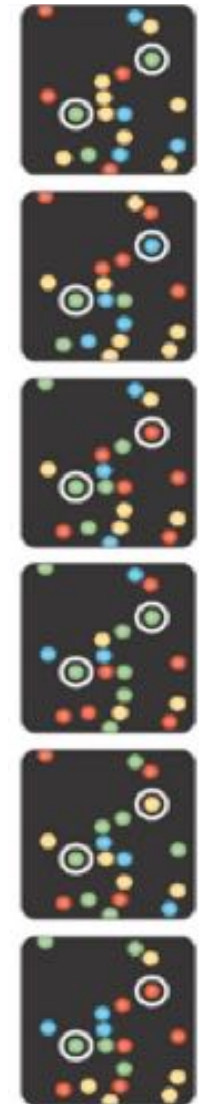
2. Attach



3. Amplify



4. Image



5. Basecall

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

[http://www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf)

# The DNA Data Race

Year	Genome	Technology	Cost
2001	Venter <i>et al.</i>	Sanger (ABI)	\$300,000,000
2007	Levy <i>et al.</i>	Sanger (ABI)	\$10,000,000
2008	Wheeler <i>et al.</i>	Roche (454)	\$2,000,000
2008	Ley <i>et al.</i>	Illumina	\$1,000,000
2008	Bentley <i>et al.</i>	Illumina	\$250,000
2009	Pushkarev <i>et al.</i>	Helicos	\$48,000
2009	Drmanac <i>et al.</i>	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



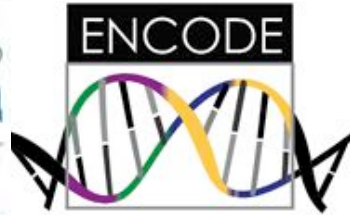


# Sequencing Centers



**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 I: 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	the worst of times, it was the	the age of wisdom, it was the	the age of foolishness, ...	
It was	the best of times, it was	the worst of times, it	was the age of wisdom, i	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 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

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

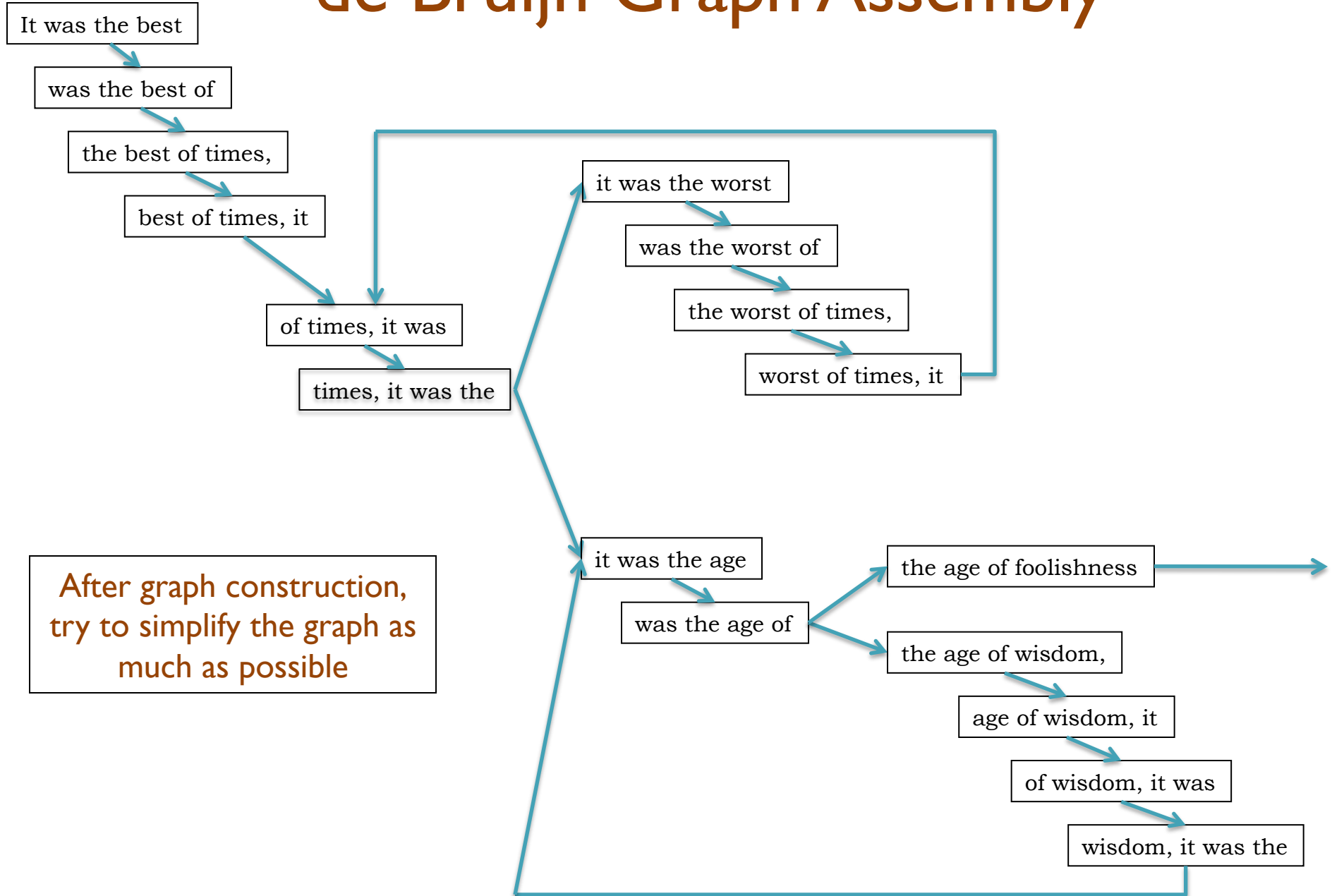
de Bruijn, 1946

Idury and Waterman, 1995

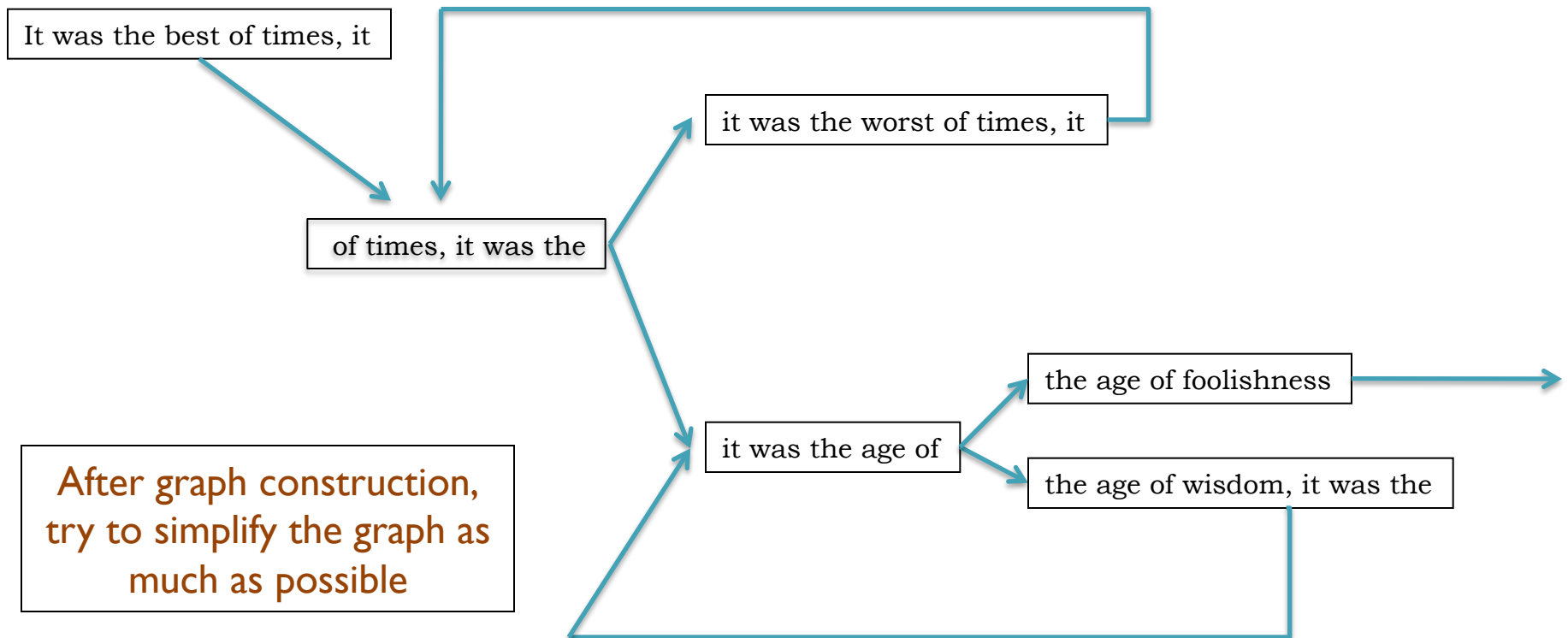
Pevzner, Tang, Waterman, 2001



# de Bruijn Graph Assembly



# de Bruijn Graph Assembly

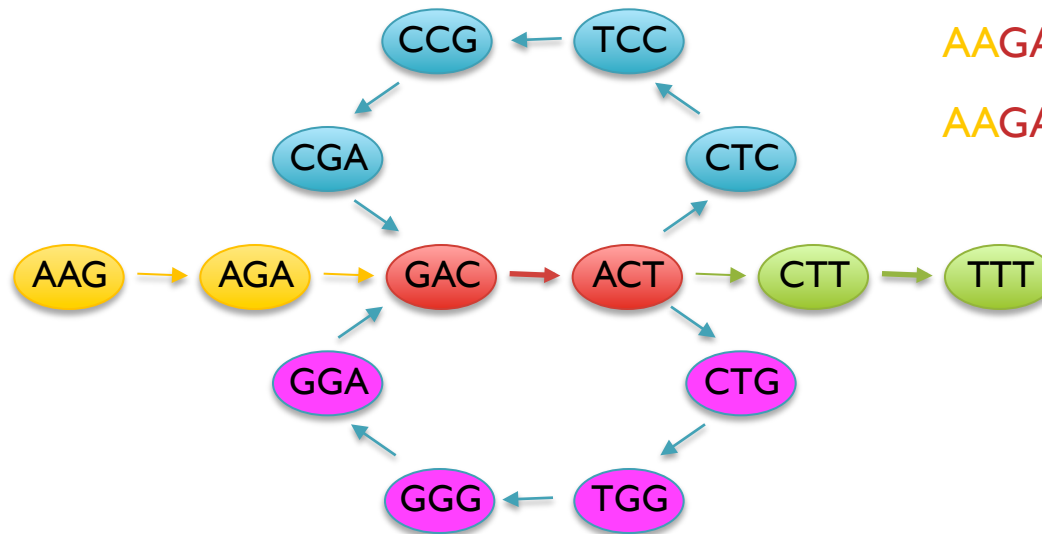


# Genome Assembly

## Reads

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

## de Bruijn Graph



## Potential Genomes

AAGACTCCGACTGGGACTTTT

AAGACTGGGACTCCGACTTTT

- 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

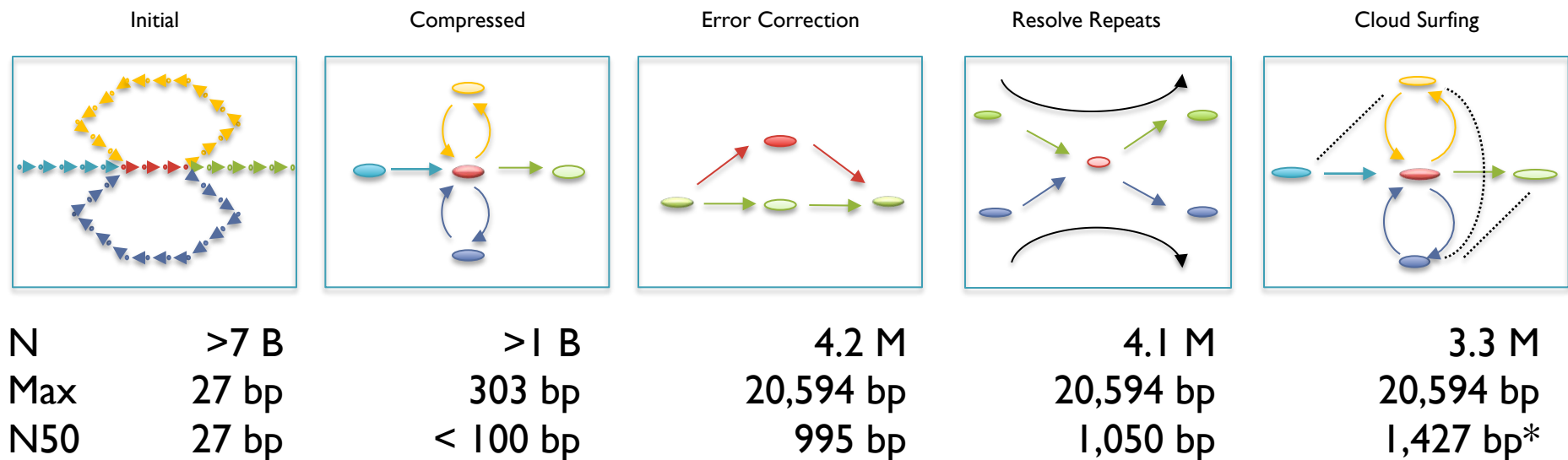
# Contrail

<http://contrail-bio.sourceforge.net>



## De novo Assembly of the Human Genome

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

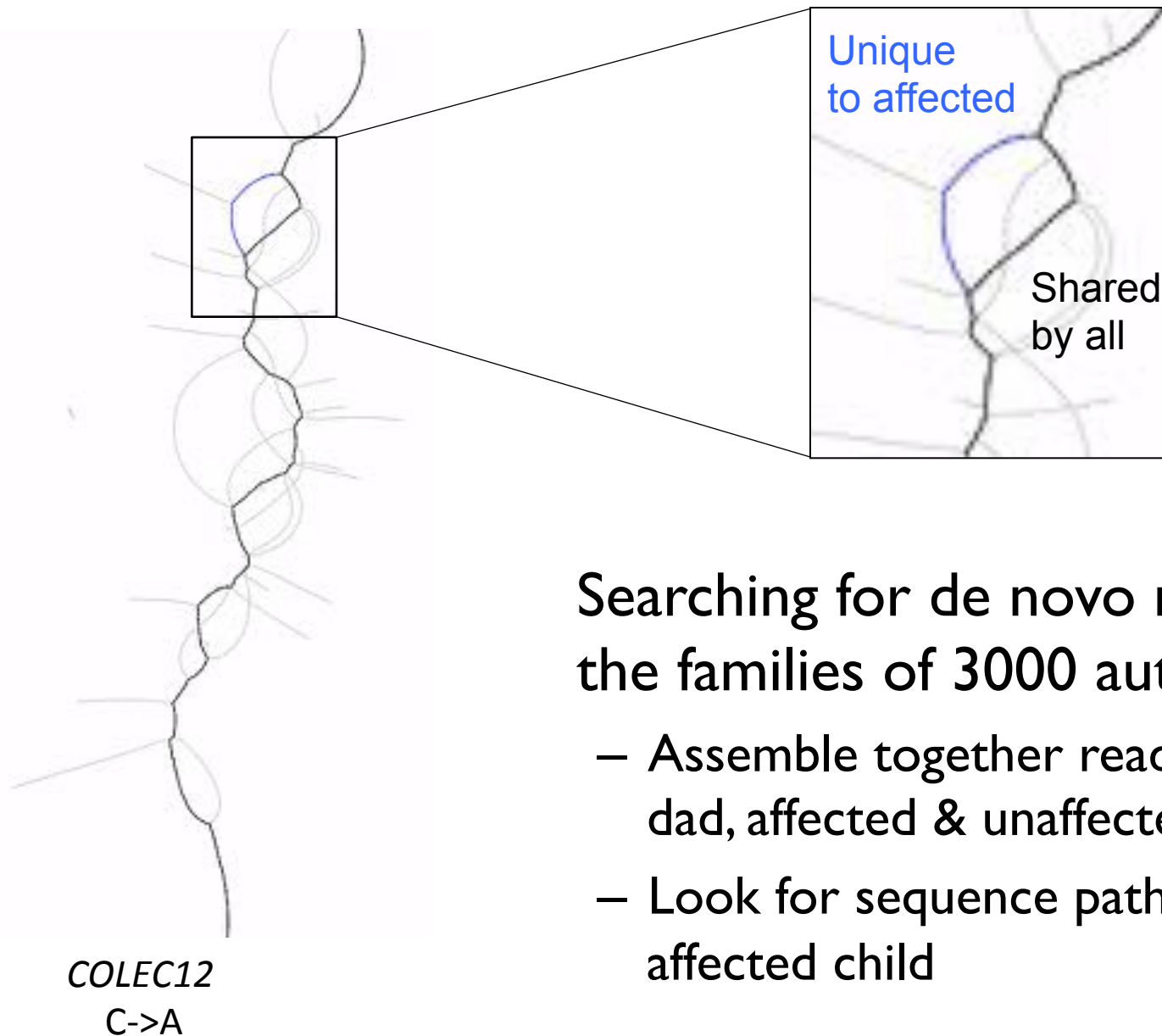


## Assembly of Large Genomes with Cloud Computing.

Schatz MC, Sommer D, Kelley D, Pop M, *et al.* *In Preparation.*



# De novo mutations and de Bruijn Graphs



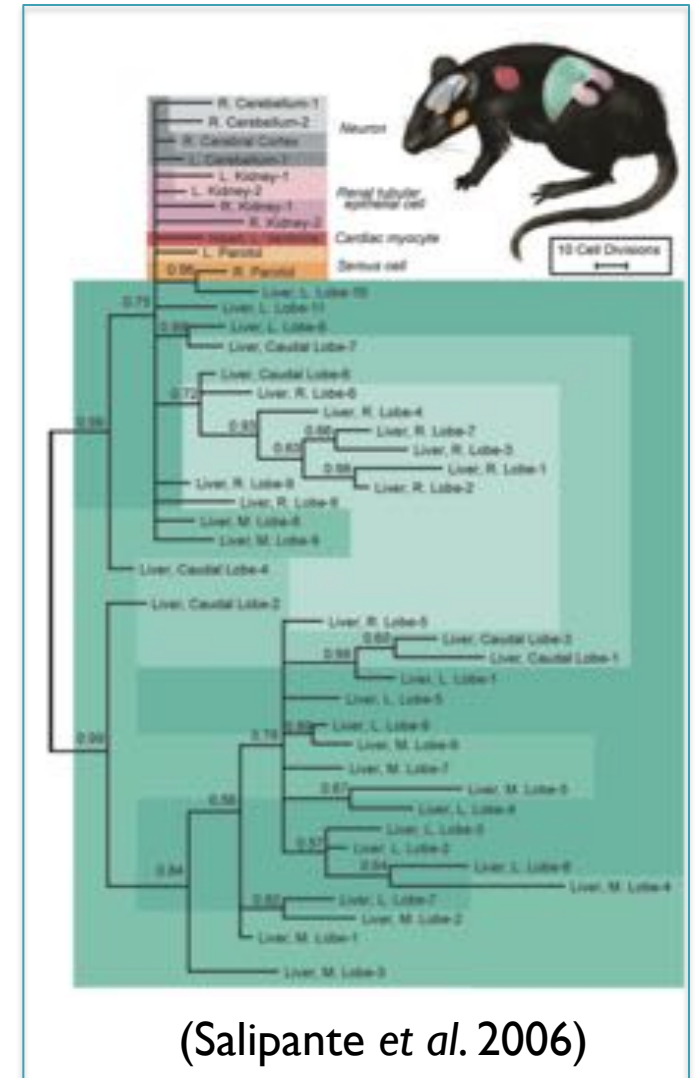
Searching for de novo mutations in the families of 3000 autistic children.

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

# MicroSeq: NextGen Microsatellite Profiling

Mitchell Bekritsky, WSBS

- Class of simple sequence repeats
  - ...GCACACACACAT... = ...G(CA)<sub>5</sub>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...



# Structural Variations in Cancer

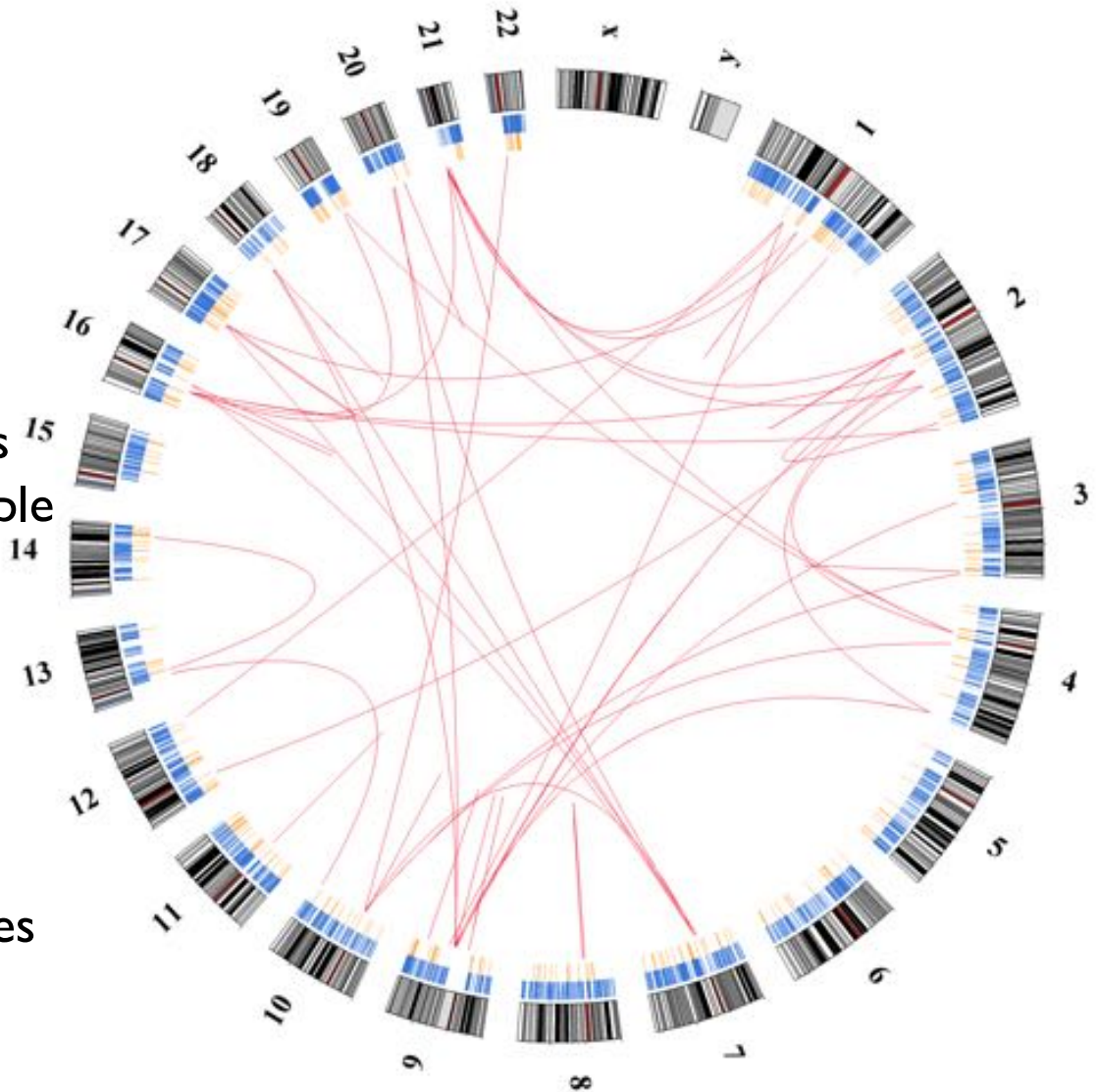
Use short reads to discover large scale variations

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

Circos plot of high confidence SVs specific to esophageal cancer sample

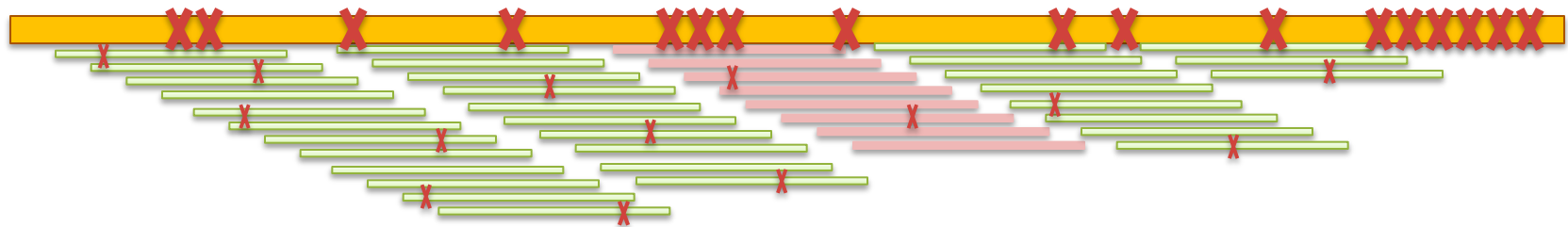
- Red: SV links
- Orange: 375 cancer genes
- Blue: 4950 disease genes

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, VWR, 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