

JAZZ: A Whole Genome Shotgun Assembler

Jarrold Chapman

Nik Putnam

Isaac Ho

Dan Rokhsar

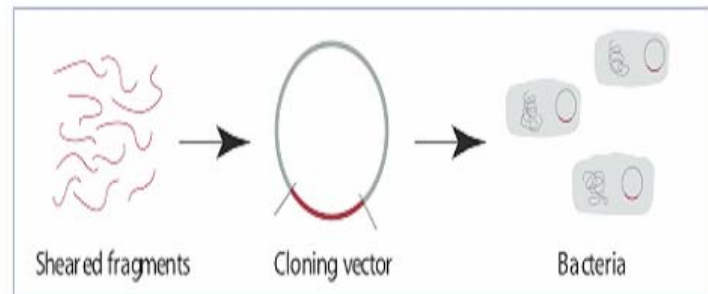


1 Genomic DNA



DNA is extracted from the cells of the organism of interest.

2 Library Creation



3 Assembly

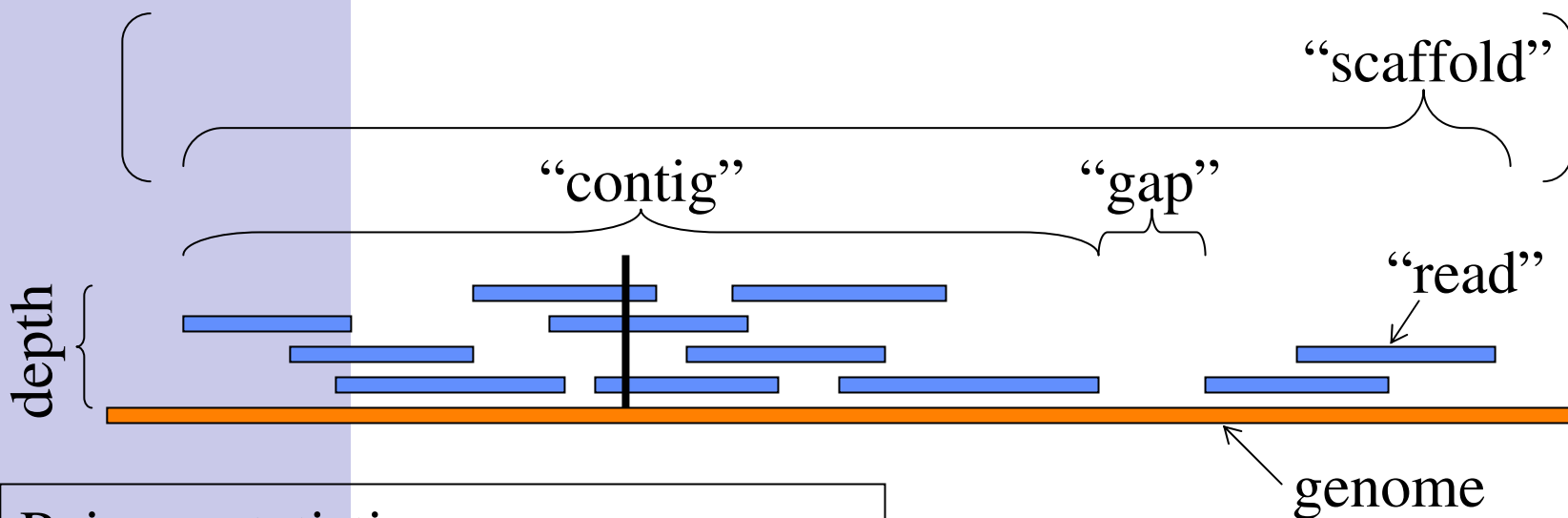
```

...ACCGTAAATGGGCTGATCATGCTTAA
                TGATCATGCTTATACCCGTGCATCCTACTG...
...ACCGTAAATGGGCTGATCATGCTTAAACCCGTGCATCCTACTG...
...ACCACCGTAAATGGGC...                               ...GCATCCTACTGTACGTAA...
    
```

4 Annotation and Analysis

A: Solving a one-dimensional jigsaw puzzle with millions of pieces (without the box)

Q: What is Whole Genome Shotgun Assembly?



Poisson statistics:

$$d = N_r L_r / G$$

$$\langle \text{reads/contig} \rangle = e^d$$

$$\langle \text{unsequenced bases} \rangle = G e^{-d}$$

[Lander and Waterman 1988
Genomics 2(3):231-9]

Idealizations:

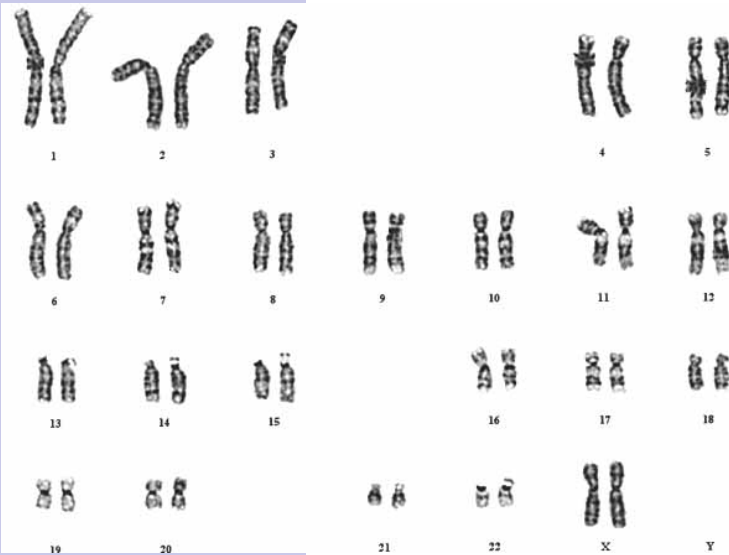
Random sampling
Random sequence
(without repeats)

- **Inherent**

- Repeats
- Paired-end reads
- Cloning bias
- Polymorphism

- **Experimental**

- Sequencing errors
- Contamination
- Tracking errors
- ???



Assembly Goals and Motivations

paired ends

polymorphism

efficiency

scalability

visualization

- **Treat sequence overlap and paired end information on an equal footing. Build in flexibility to include BAC localization and other mapping data for mixed projects**
- **Allow for polymorphisms. Unlike microbes and flies, *fugu*, *ciona*, and other sequencing targets are neither haploid nor inbred – individuals from the wild**
- **Efficient assemblies to provide good substrates for annotation. 6X depth should statistically give good coverage (99.7%), contiguity (30 kb), and contig linking**
- **Develop parallel implementation from the start. Scale to large projects (*Fugu rubripes* @ 400 MB, *Poplar* @ 600 MB, *Xenopus tropicalis* @ 1.7 GB)**
- **Integrate assembly visualization and analysis tools for q.c. and validation – visualize multiple scales**

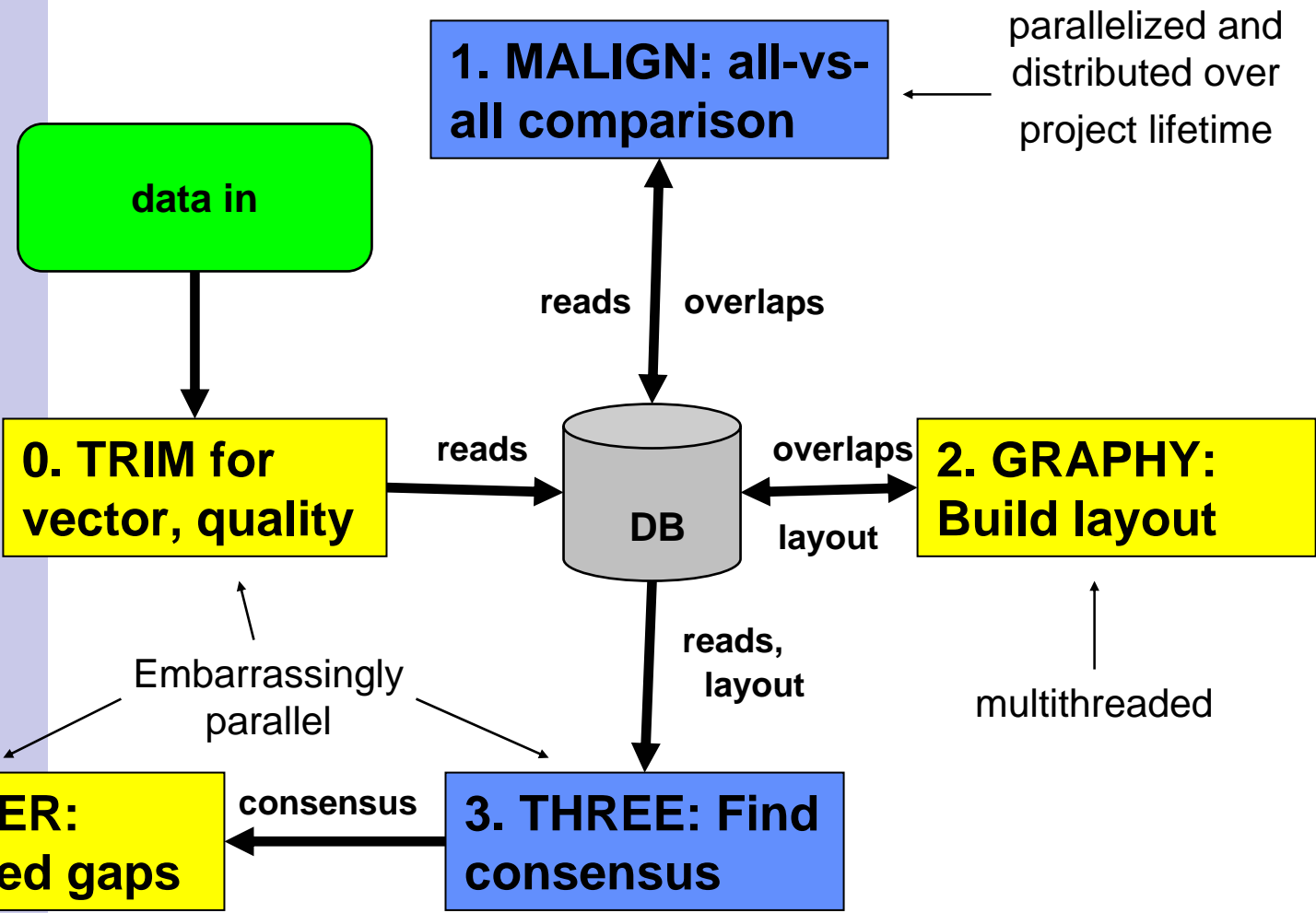
JAZZ assembly pipeline

Use

- Overlap
- Layout
- Consensus

paradigm

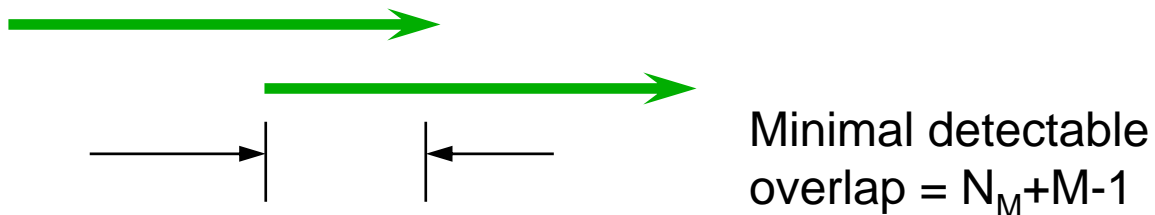
genome out



MALIGN: Rapid screening for overlaps

Screen read pairs for potential overlaps using a hashing scheme

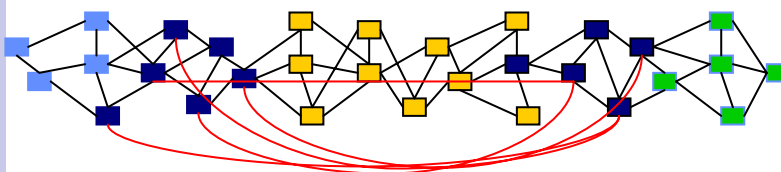
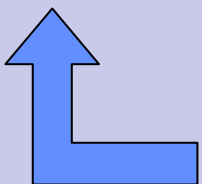
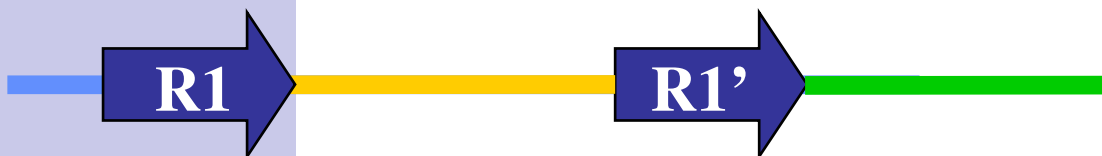
- **Identify all pairs of reads that share ten or more informative words** (reverse complement, too)
 - Use a parallel hash scheme for speed/memory.



- **Designate over-represented words in quality-trimmed data set as “unhashable.” (AAAAAAA)**
 - Their shared occurrence in two reads is not a reliable indicator of a true overlap.
- **Align candidate overlapping reads using banded Smith-Waterman algorithm. Reject low %id.**

GRAPHY: graphical layout algorithm

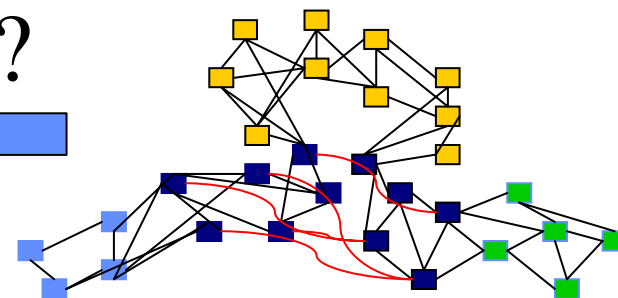
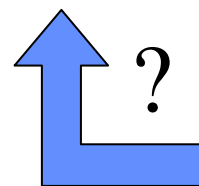
- Given a set of all-vs-all alignments (from MALIGN)
- Estimate the likelihood that each edge is true.
- Construct an initial solution from the highest confidence, unique sequence.
- Improve the solution iteratively with self-consistency requirement.



Layout Problem \equiv
Finding the sub-
graph of
'true' edges

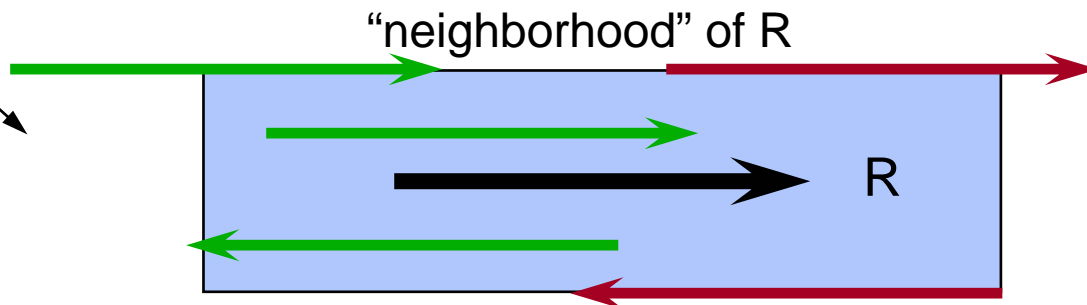
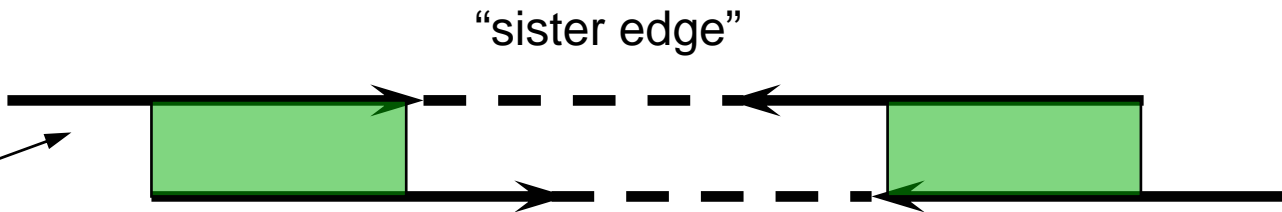
True edges join reads actually derived from overlapping portions of the genome.

False edges arise from repeats.

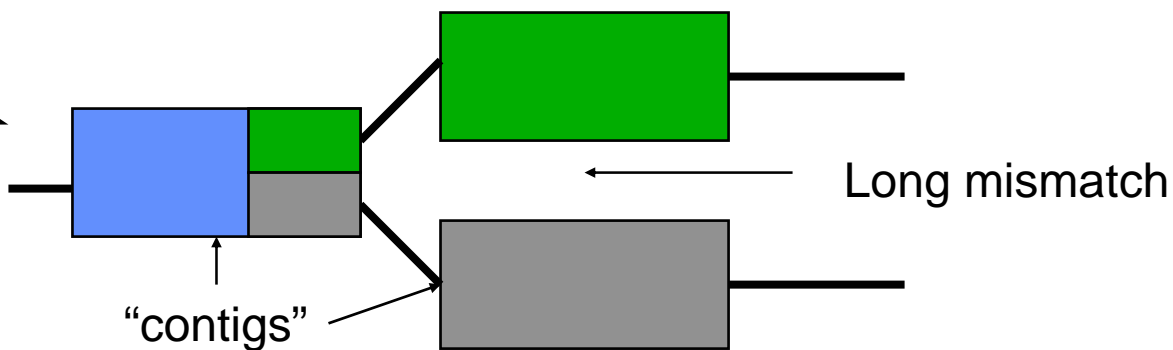


Key ingredients

- Use of “rectangles” and other local structures in read graph to corroborate overlaps and reads



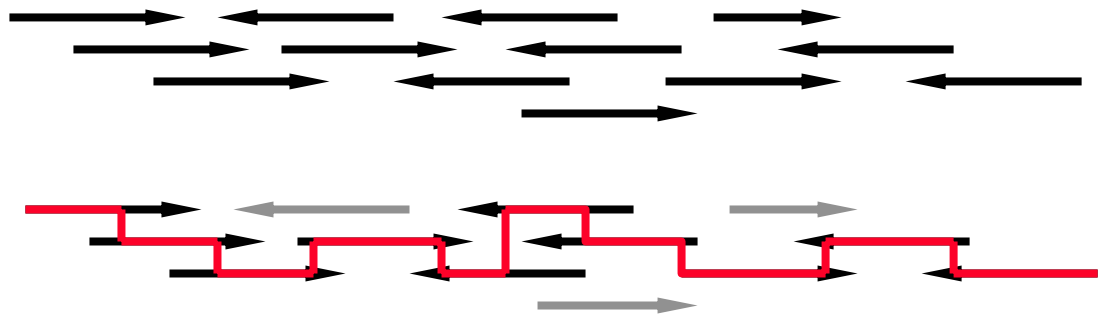
- Iterative, self-consistent formation of contigs and scaffolds



THREE: Reaching consensus

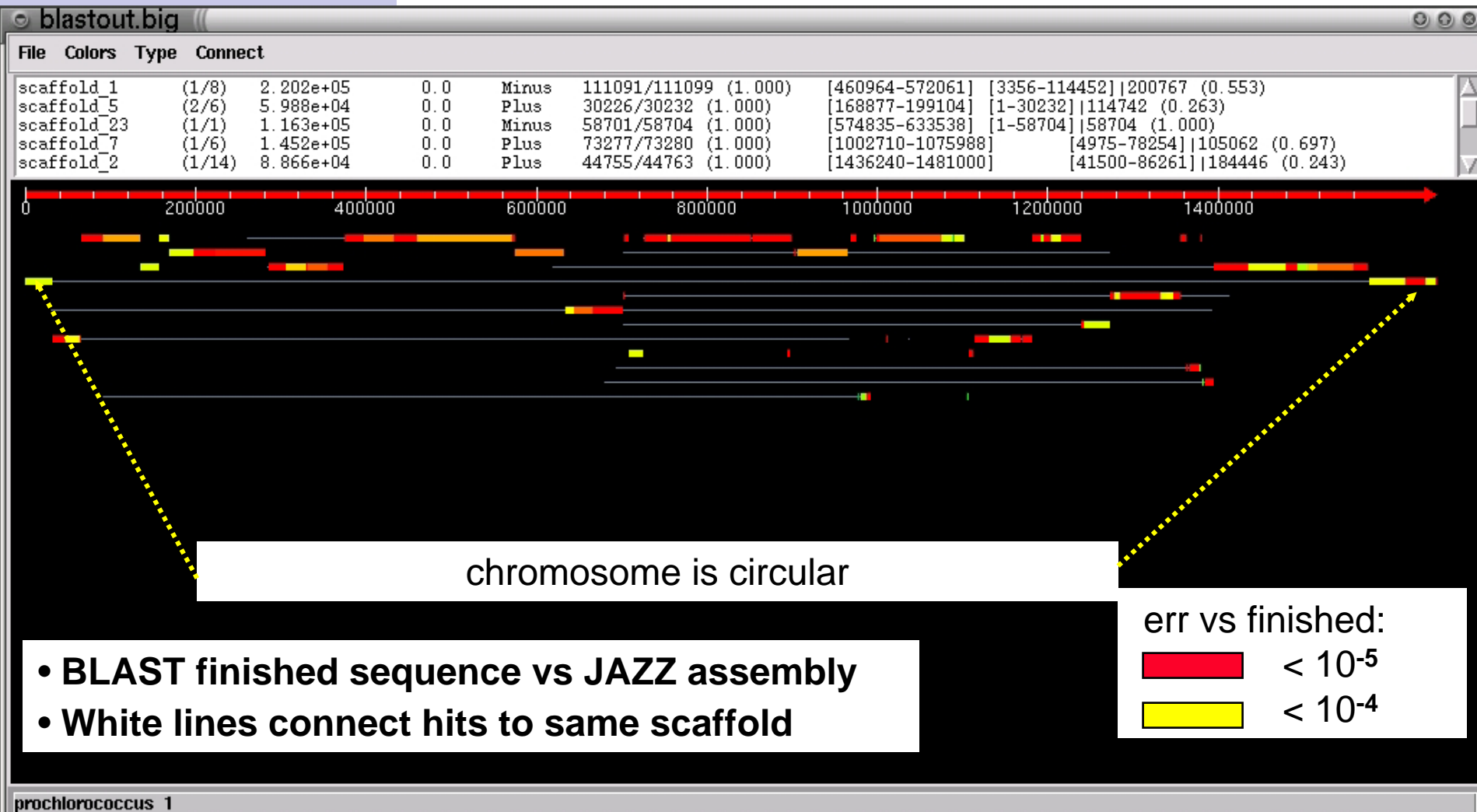
Use central high quality segments of reads as a proto-consensus

- Identify “backbone” of forward-moving, minimally overlapping reads spanning each contig



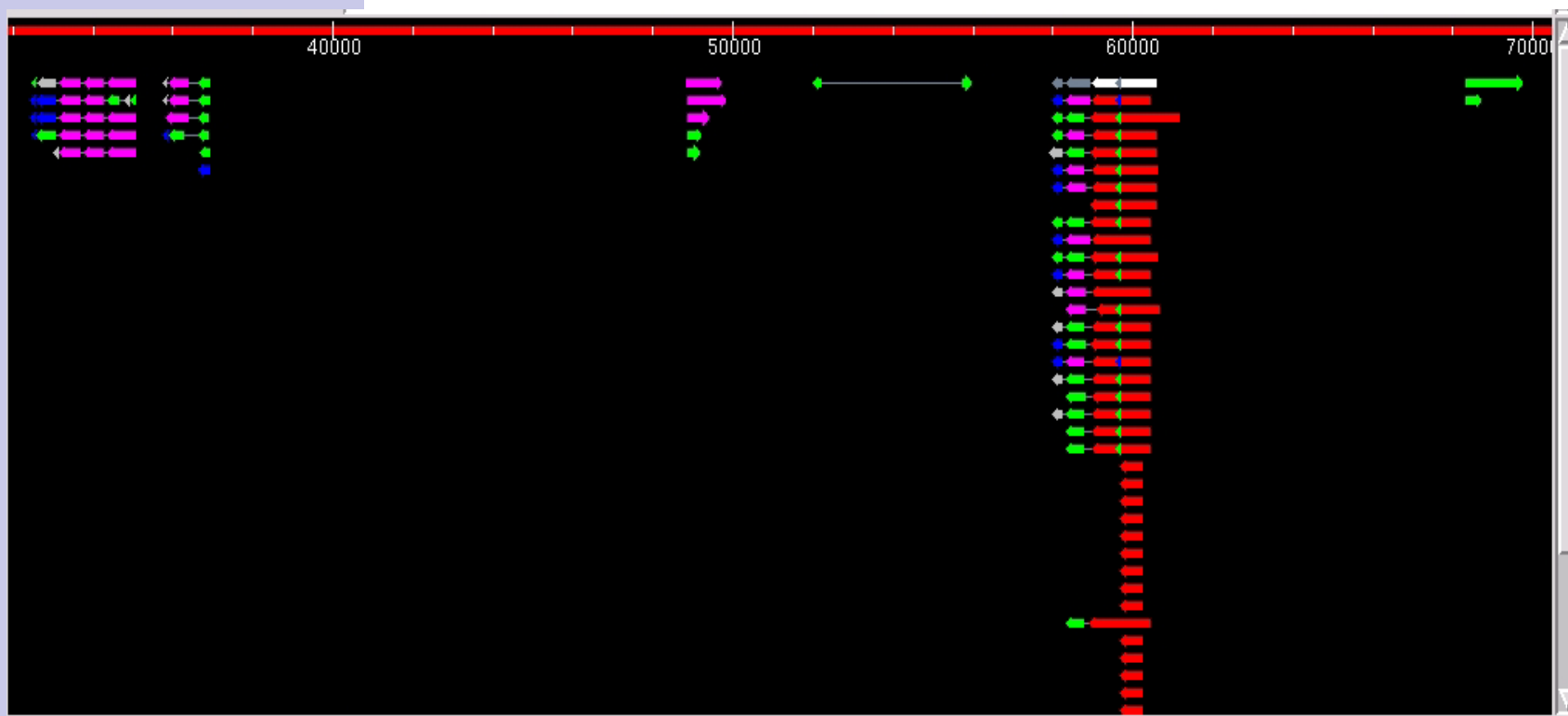
- Form “reference” made from concatenating segments closest to center of each backbone read
- Make master-slave alignment of reference segment to its overlapping reads (with quality-weighted voting)
- Mark potential polymorphisms/misassemblies in alignment

Accuracy (*Prochlorococcus* @8X, 3kb only)

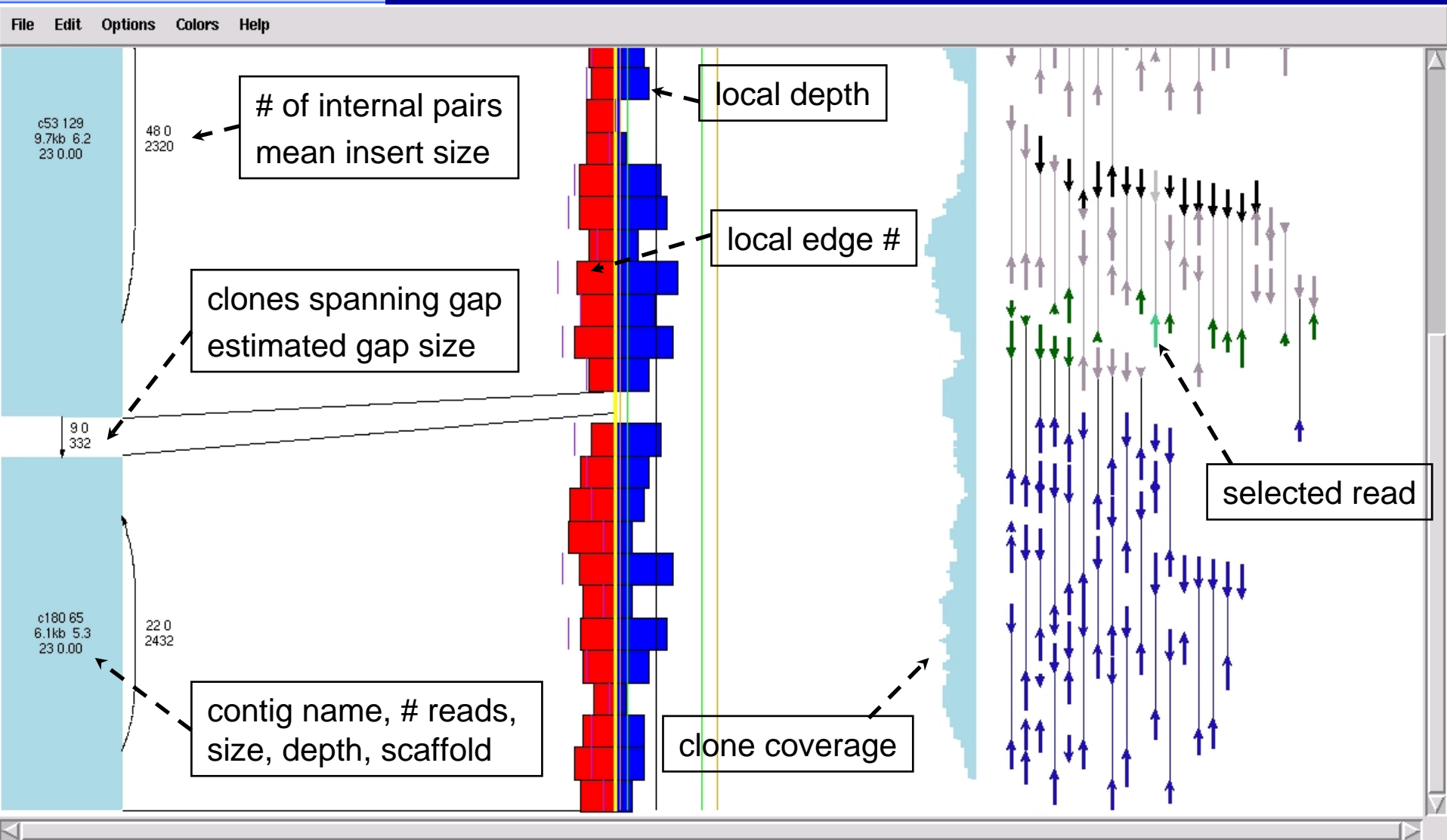


JAZZ scaffolds are a good substrate for annotation

- GeneWise models introduce few or no indels
- Approximate error rate: < 1 indels/10 kb as expected at 6X depth



JAZZ view of assembly

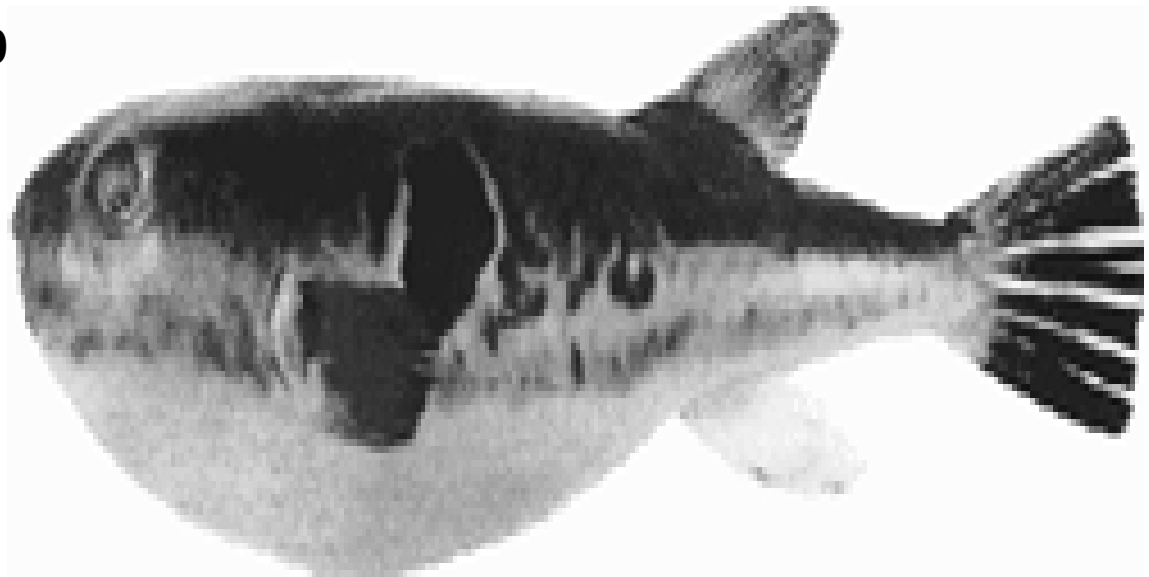


Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu Rubripes*

Aparicio, Chapman, ..., Putnam, ..., Rokhsar, Brenner

Science 23 August 2002

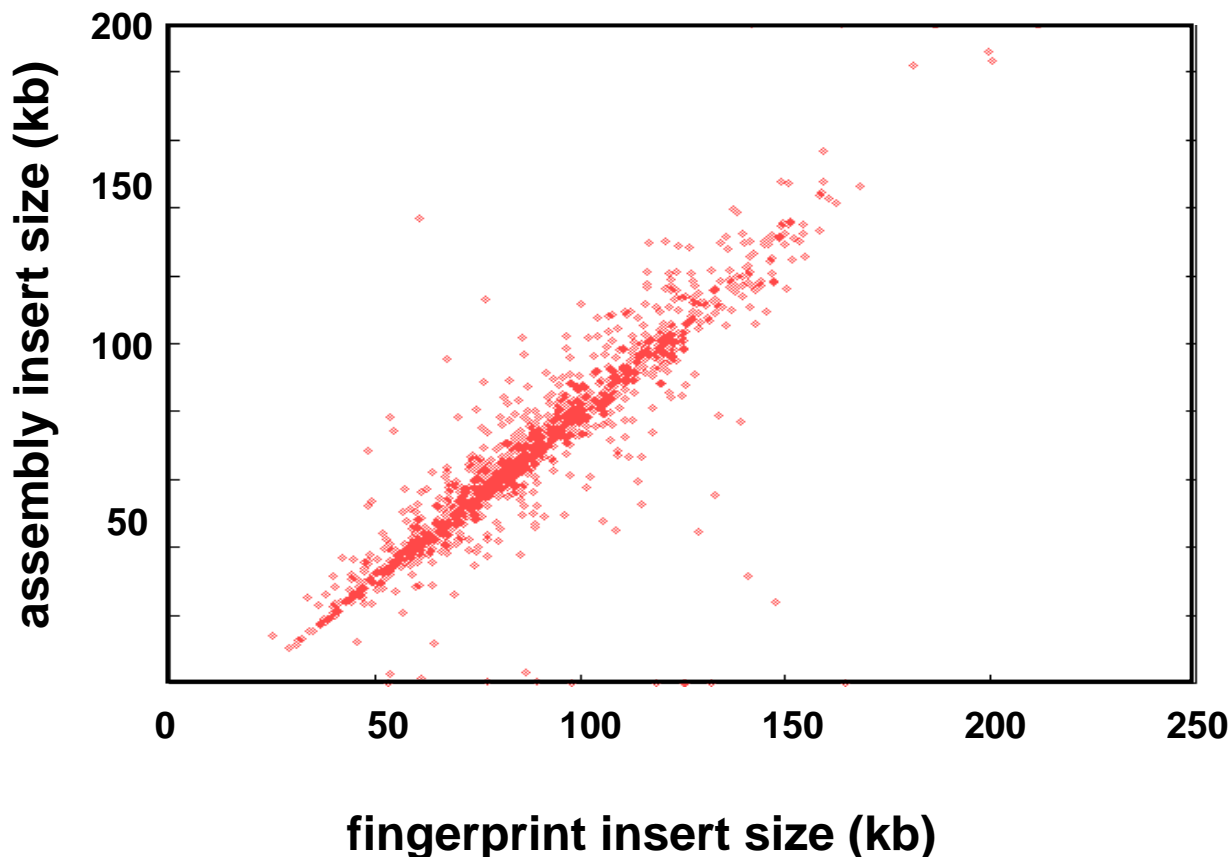
Vol.297 No. 5585 1301-1310



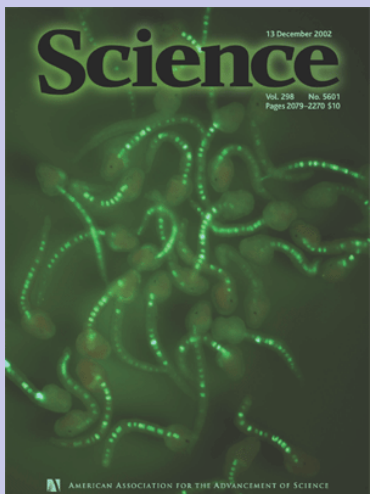
<http://genome.jgi-psf.org/fugu6/fugu6.home.html>

Several thousand fingerprinted BACs have both ends placed in same (cosmid-O&O'd) scaffold
With small calibration correction, distance between ends on (small-insert-only) assembly equals sum of restriction fragment lengths

BAC sizes: assembly vs fingerprint

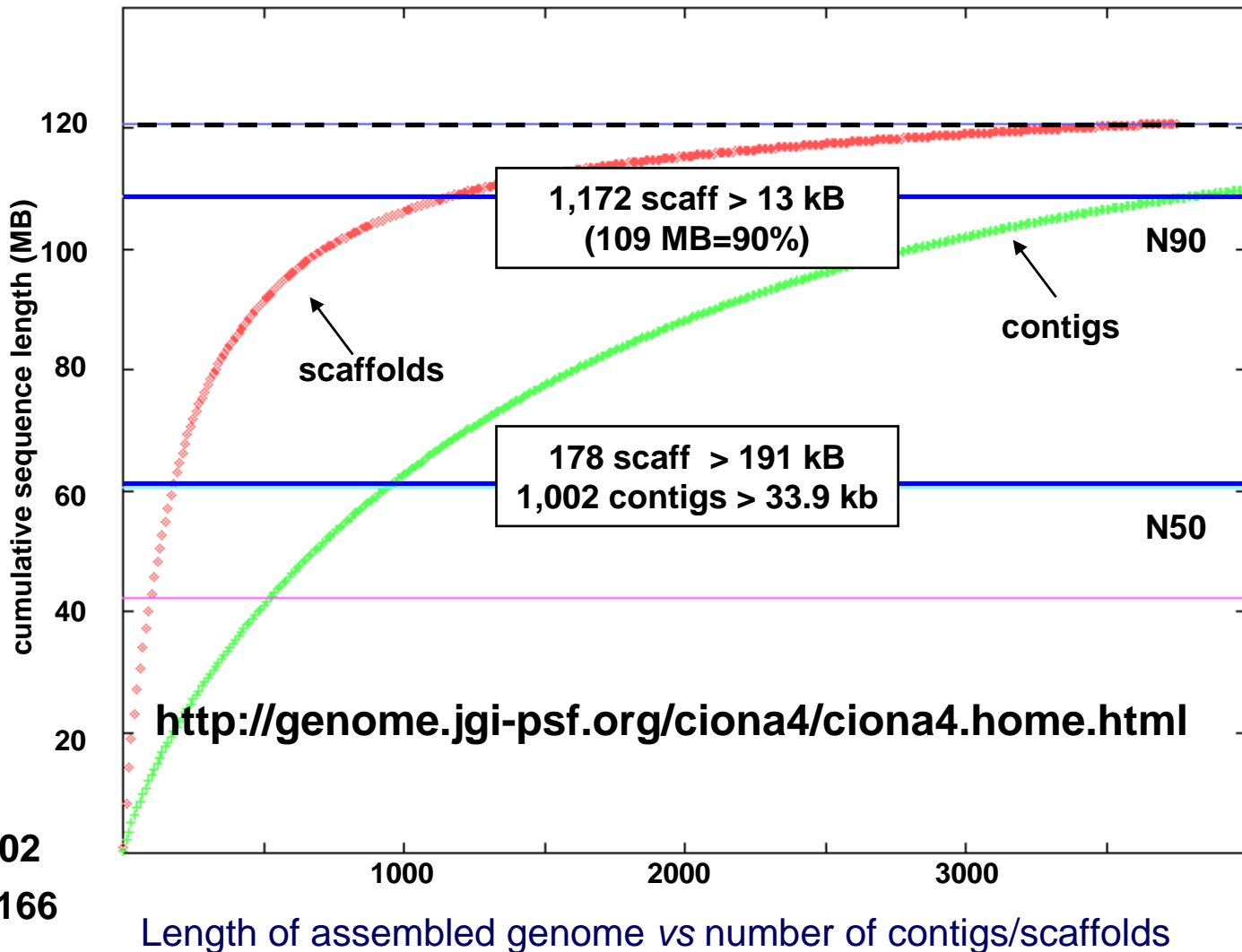


Assembly summary



The Draft Genome
of *Ciona intestinalis*:
Insights into Chordate
And Vertebrate Origins

Science 13 December 2002
Vol. 298 No. 5601 2157-2166



Assembly Goals Revisited

paired ends

✓ **Plasmid-, cosmid-, BAC-end data used to achieve good contiguity**

polymorphism

✓ **1.5% *Ciona intestinalis*, 0.4% *Fugu***

efficiency

✓ **Successful annotation of *fugu*, *ciona* – assembly provides good annotation substrate**

scalability

✓ **400MB *fugu* assembled – gearing up for poplar (600 MB) *Xenopus* (1.7GB)**

visualization

✓ **JAZZ view allows read, contig, scaffold level visualization**

To come...

microbial consortia; larger polymorphic genomes; general availability

- **DOE Joint Genome Institute (JGI)**
- **JGI Assembly Team**
 - **Nik Putnam, Isaac Ho, Dan Rokhsar**
- **Susan Lucas, Paul Richardson,
Chris Detter**
- **Fugu and Ciona Genome Consortia**
- **DOE CSGF / Krell Institute**

JGI: <http://www.jgi.doe.gov>