Metassembler: Improving de novo genome assembly

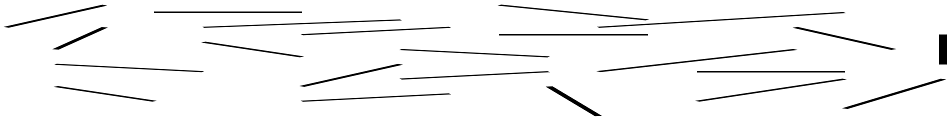
Paul Baranay, Scott Emrich, Michael Schatz

Feb 17, 2012
AGBT

@mike_schatz / #AGBT

Assembling a Genome

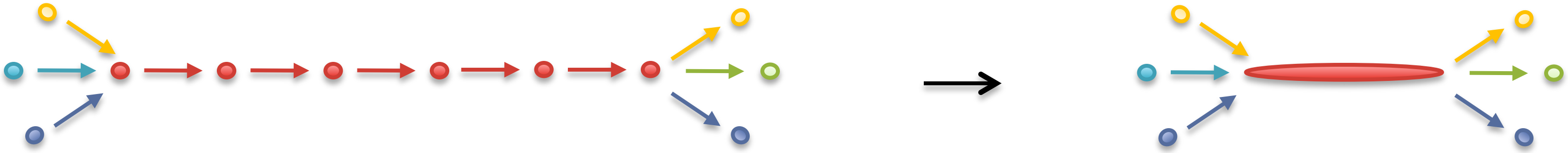
1. Shear & Sequence DNA



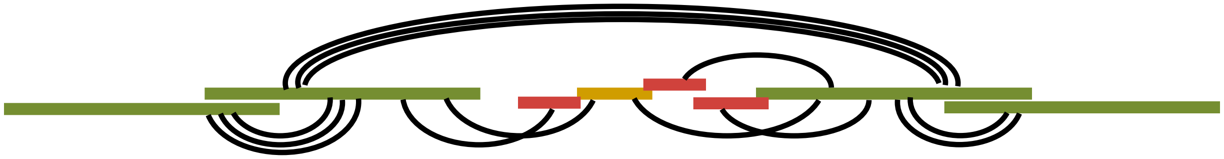
2. Construct assembly graph from overlapping reads

…AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT…

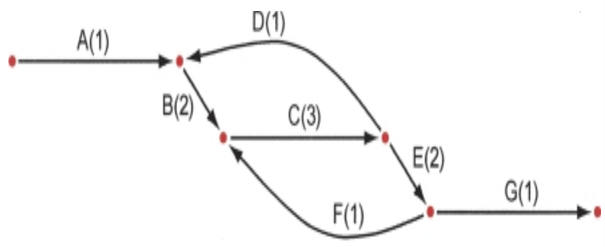
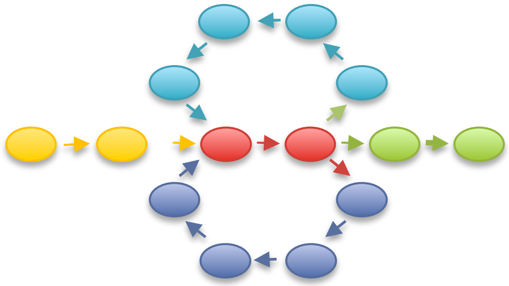
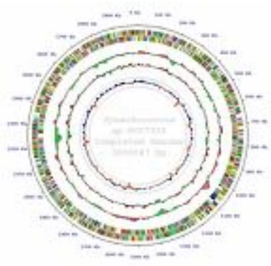
3. Simplify assembly graph



4. Detangle graph with long reads, mates, and other links



Genome Assemblers

ALLPATHS-LG	SOAPdenovo	Celera Assembler
		
Broad's assembler (Gnerre et al. 2011)	BGI's assembler (Li et al. 2010)	JCVI's assembler (Miller et al. 2008)
Unipath graph Short + PacBio (patching)	De bruijn graph Short reads	Overlap graph Medium + Long reads
Easy to run if you have compatible libraries	Most flexible, but requires a lot of tuning	Supports Illumina/454/PacBio Hybrid assemblies

Plus several dozens more

Each balancing the tension between connectivity and accuracy in a different way

2011: Year of the Assembly Bakeoff

THE ASSEMBLATHON

- Simulated genome distantly related to human chr13
- 17 labs, 50+ assemblies

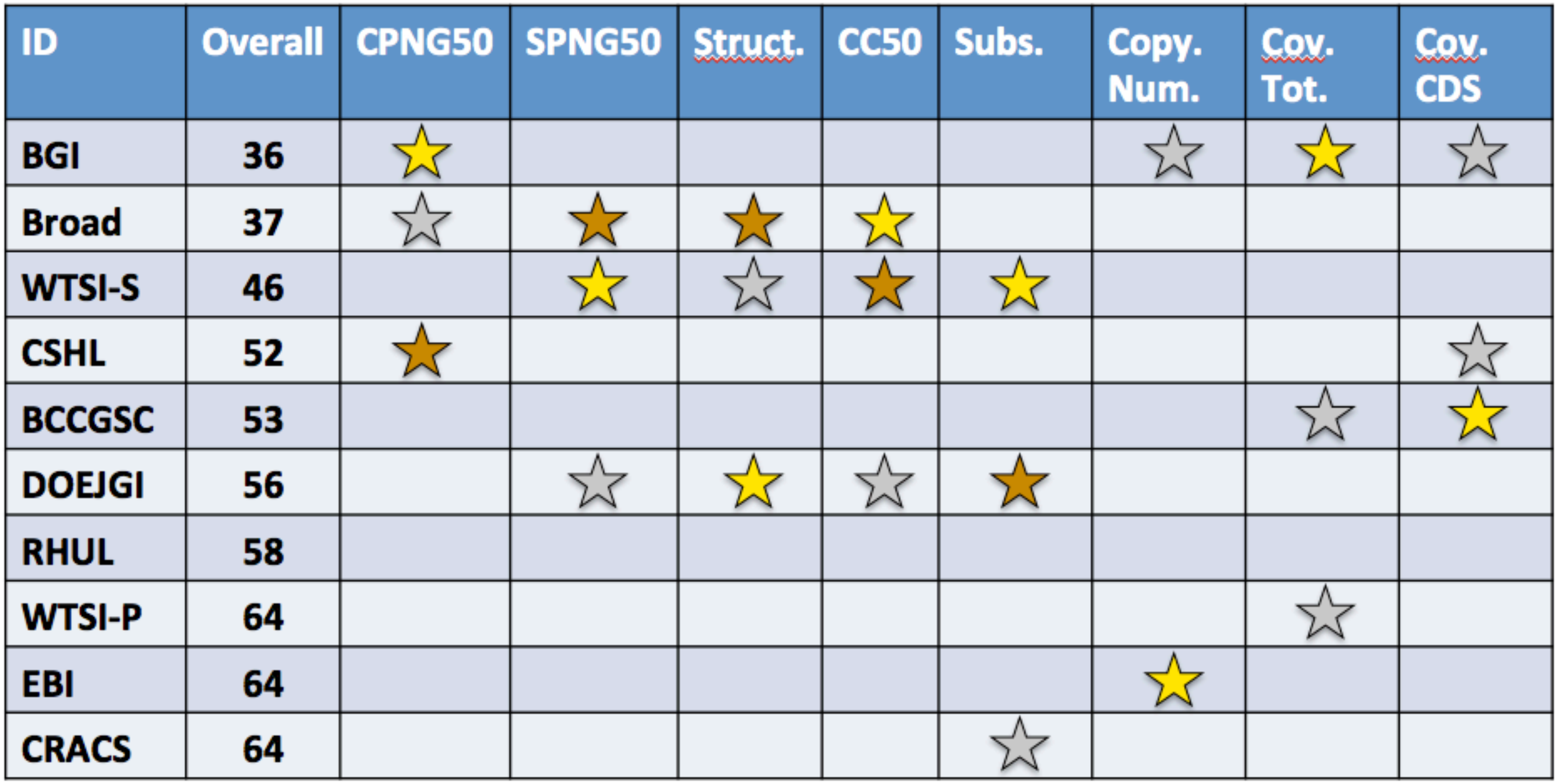
Genome Assembly Gold-Standard Evaluations

- 4 real genomes ranging from bacteria to individual human chromosome
- Internal evaluation of 8 assemblers

Assemblathon 1: A competitive assessment of de novo short read assembly methods.
Earl, DA *et al.* (2011) *Genome Research*. In press.

GAGE: A critical evaluation of genome assemblies and assembly algorithms.
Salzberg, SL *et al.* (2011) *Genome Research*. In press.

Final Rankings



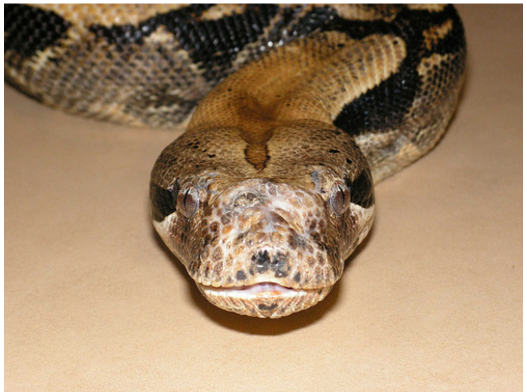
ID	Overall	CPNG50	SPNG50	Struct.	CC50	Subs.	Copy. Num.	Cov. Tot.	Cov. CDS
BGI	36	🥇					🥈	🥇	🥈
Broad	37	🥈	🥉	🥉	🥇				
WTSI-S	46		🥇	🥈	🥉	🥇			
CSHL	52	🥉							🥈
BCCGSC	53							🥈	🥇
DOEJGI	56		🥈	🥇	🥈	🥉			
RHUL	58								
WTSI-P	64							🥈	
EBI	64						🥇		
CRACS	64					🥈			

- SOAPdenovo and ALLPATHS came out neck-and-neck followed closely behind by SGA, Celera Assembler, and ABySS

- My recommendation for "typical" short read assembly is to use ALLPATHS

Assemblathon 2

- Real sequence data, *de novo* assembly



- Step 1: Apply best practices from Assemblathon 1
- Step 2: Add secret weapon for winning…

Images from Assemblathon

Forensics

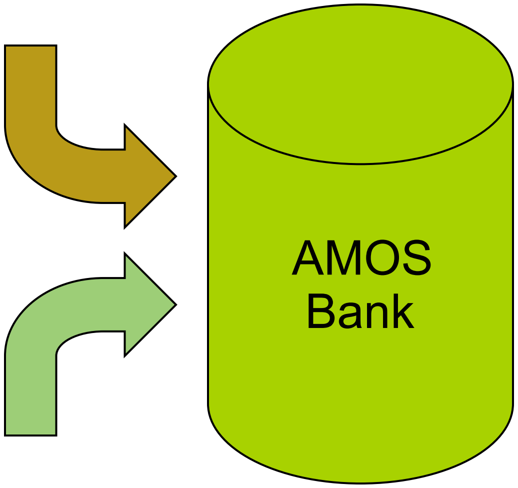
Assembly Forensics

Computationally scan an assembly for mis-assemblies.

- Data inconsistencies are indicators for mis-assembly
- Some inconsistencies are merely statistical variations

AMOSvalidate

1. Analyze Mate Pairs & Libraries
2. Analyze Depth of Coverage
3. Analyze Read Alignments
4. Analyze Read Breakpoints
5. Load Mis-assembly Signatures into Bank



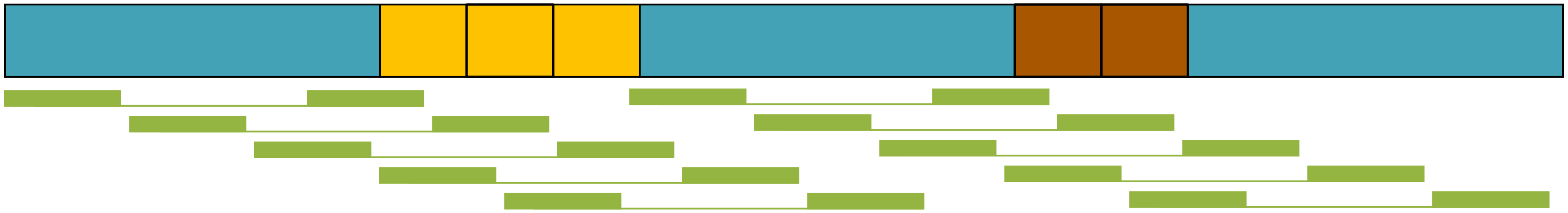
Genome Assembly forensics: finding the elusive mis-assembly.
Phillippy, AM, Schatz, MC, Pop, M. (2008) Genome Biology 9:R55.

Hawkeye & AMOS: Visualizing and assessing the quality of genome assemblies
Schatz, MC *et al.* (2012) *Briefings in Bioinformatics*. In Press.

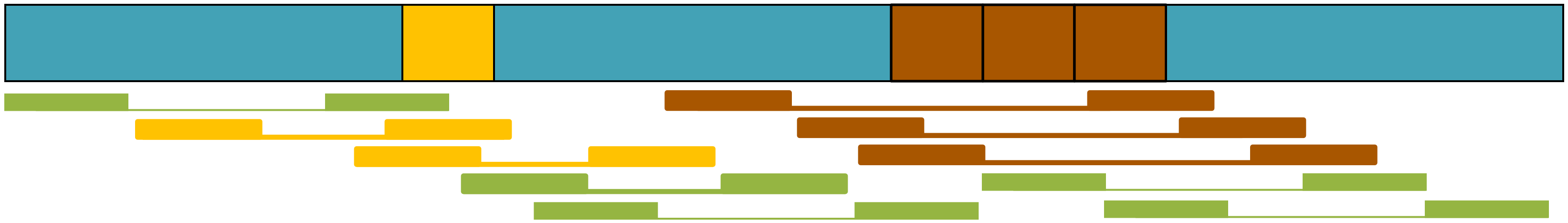
Forensics

Mate Evaluation

- Correct: mates have expected orientation and separation



- Mis-assembled: mates have incorrect orientation and separation



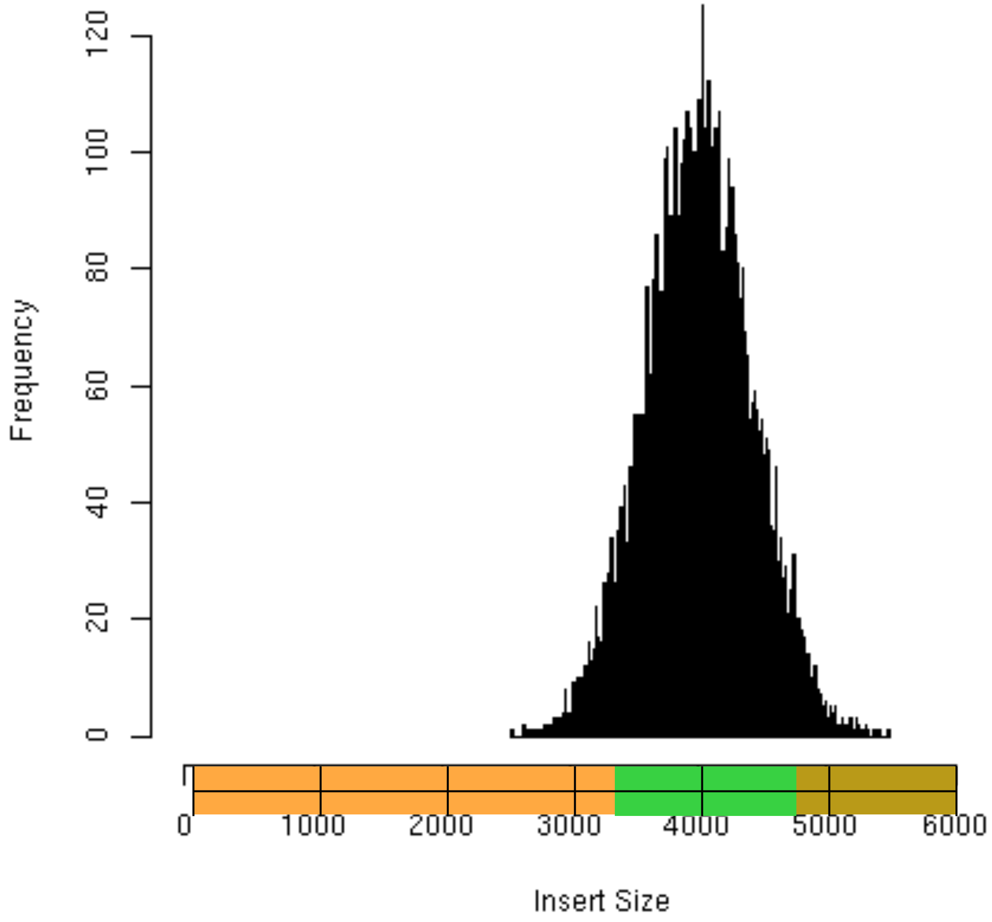
- Slightly compressed/expanded mates are expected because mates are sampled from a distribution of fragments

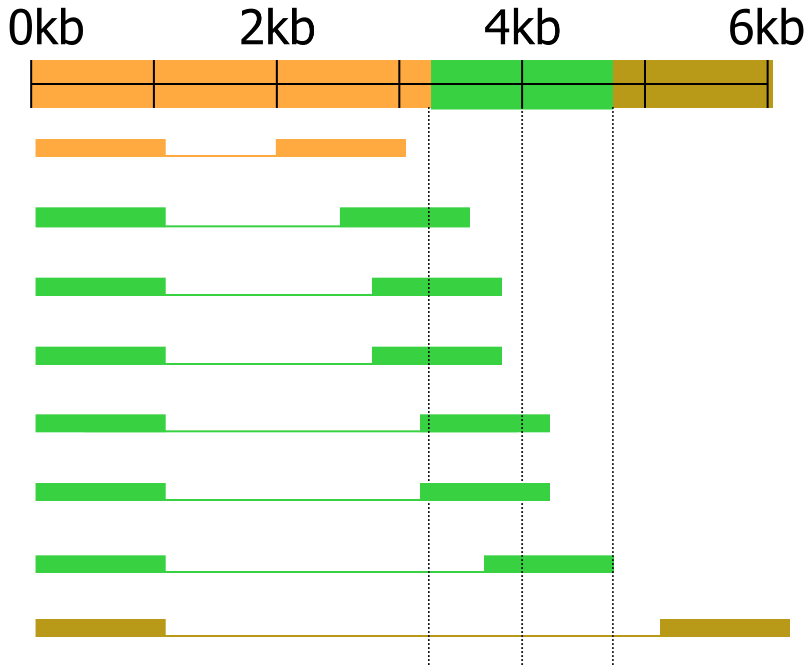
Forensics

Compression/Expansion Statistic

Library size distribution

Mean: 4000, SD: 400





8 inserts: 3kb-6kb

Local Mean: 4048

C/E Stat: $\frac{(4048-4000)}{(400 / \sqrt{8})} = +0.33$

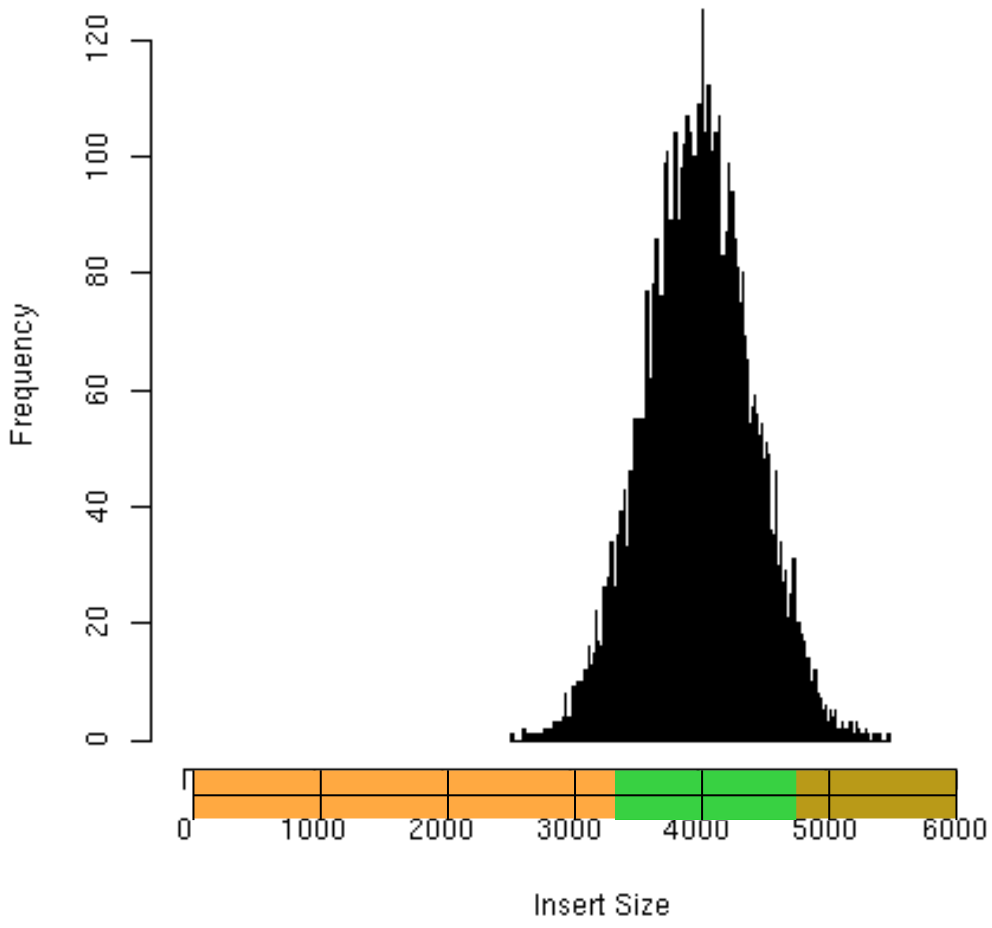
Near 0 indicates overall happiness

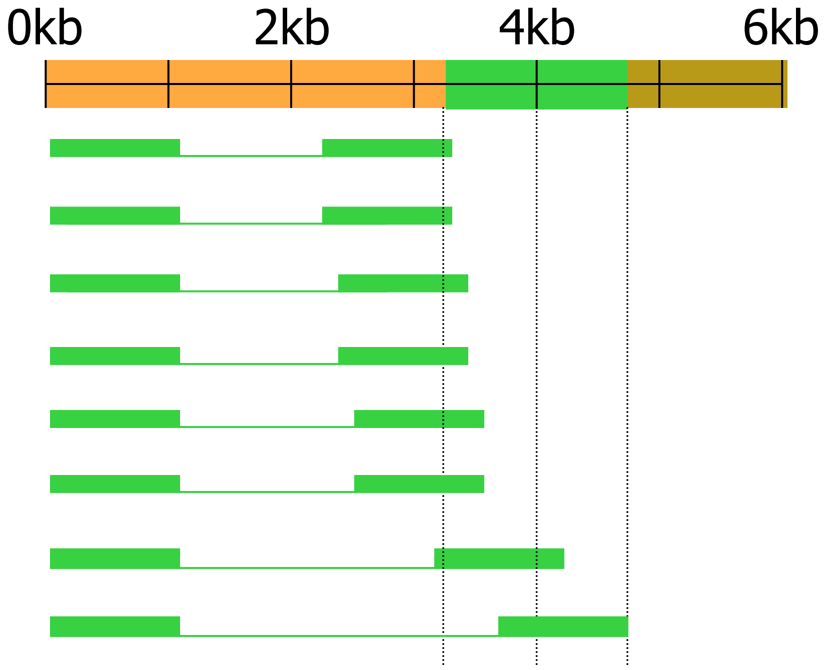
Forensics

Hidden Compression

Library size distribution

Mean: 4000, SD: 400





8 inserts: 3.2 kb-4.8kb

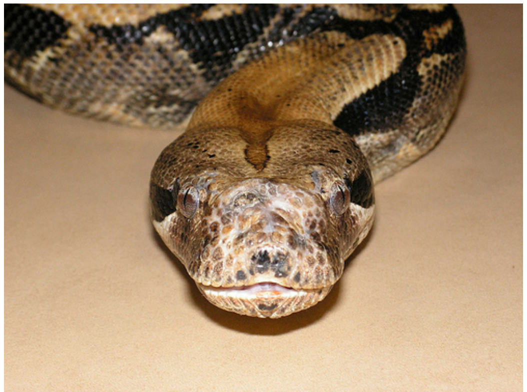
Local Mean: 3488

C/E Stat: $\frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62$

C/E Stat ≤ -3.0 indicates Compression

Assemblathon 2

- Real sequence data, *de novo* assembly



- Step 1: Apply best practices from Assemblathon 1
- Step 2: Add secret weapon for winning…

Images from Assemblathon

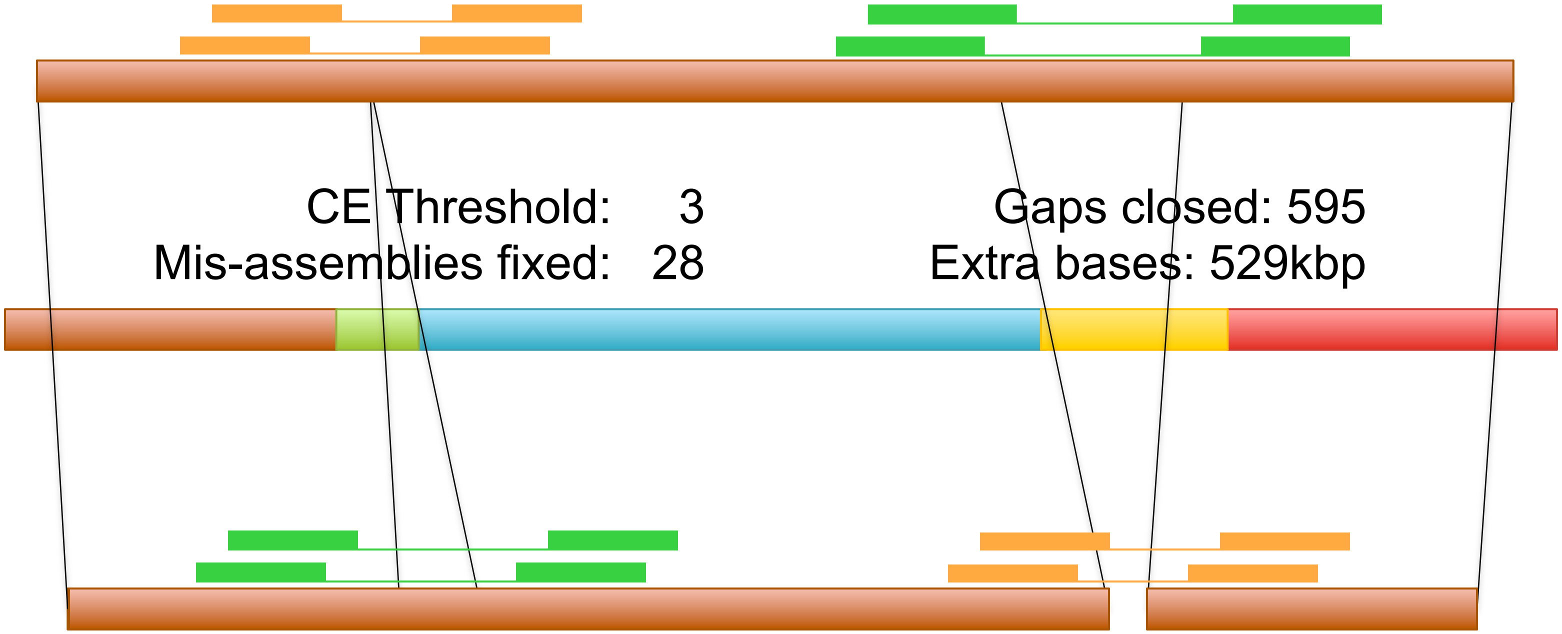
Fish Metassembly

http://metassembler.sf.net

ALLPATHS-LG

Scaffold N50: 3,710,017
#>1000: 2,791

Contig N50: 20,183
#>1000: 68,591



CE Threshold: 3
Mis-assemblies fixed: 28

Gaps closed: 595
Extra bases: 529kbp

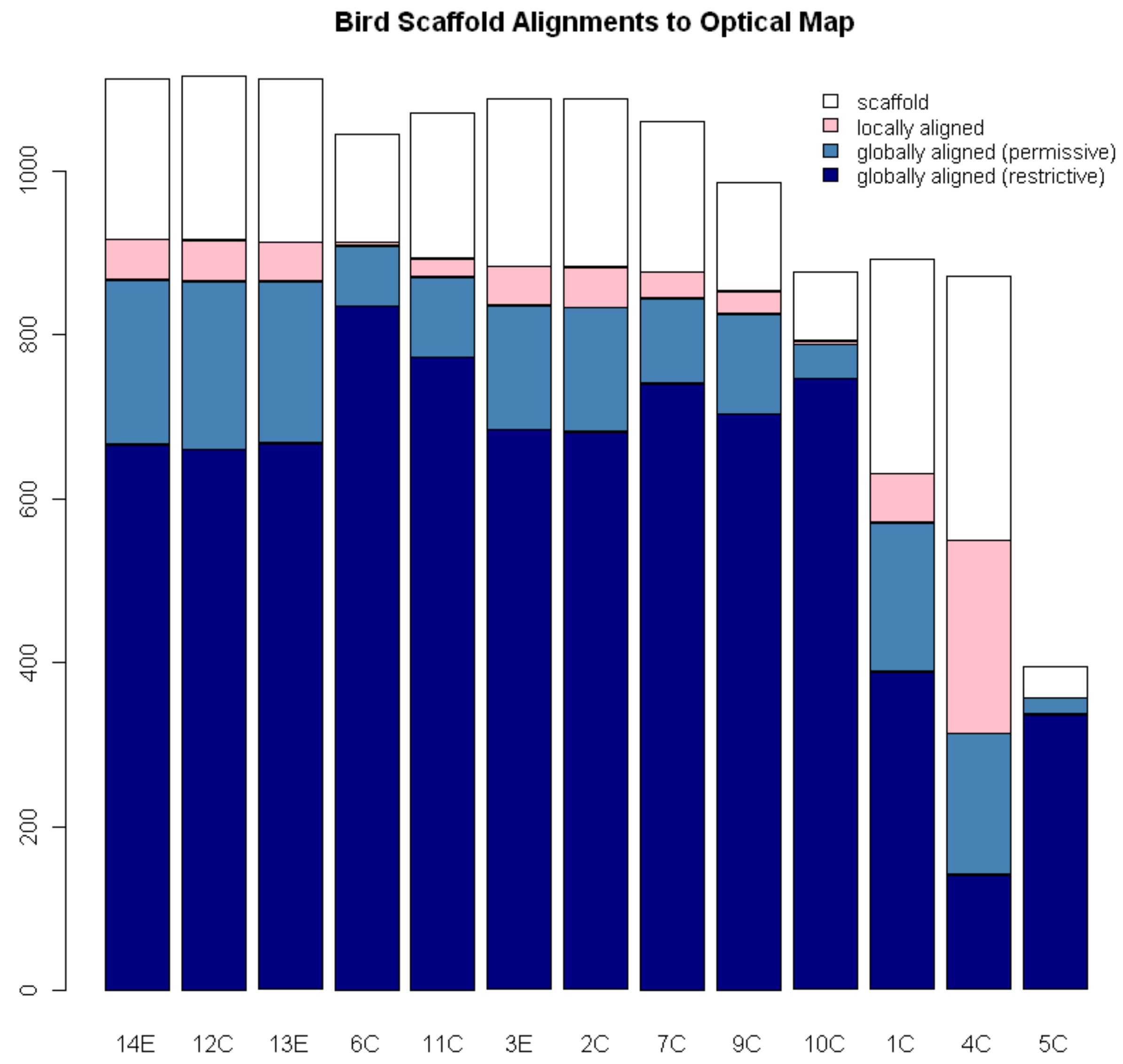
SOAPdenovo
+ FLASH
+ Quake
+ AMOS

Scaffold N50: 285,413
#>1000: 29,119

Contig N50: 1,607
#>1000: 218,643

Parrot Metassembly

http://metassembler.sf.net


Bird Scaffold Alignments to Optical Map
□ scaffold
▨ locally aligned
▨ globally aligned (permissive)
■ globally aligned (restrictive)
0
200
400
600
800
1000
14E
12C
13E
6C
11C
3E
2C
7C
9C
10C
1C
4C
5C

- Crowd-source individual assemblies
  - 13 submissions (including variants of same basic assembly)

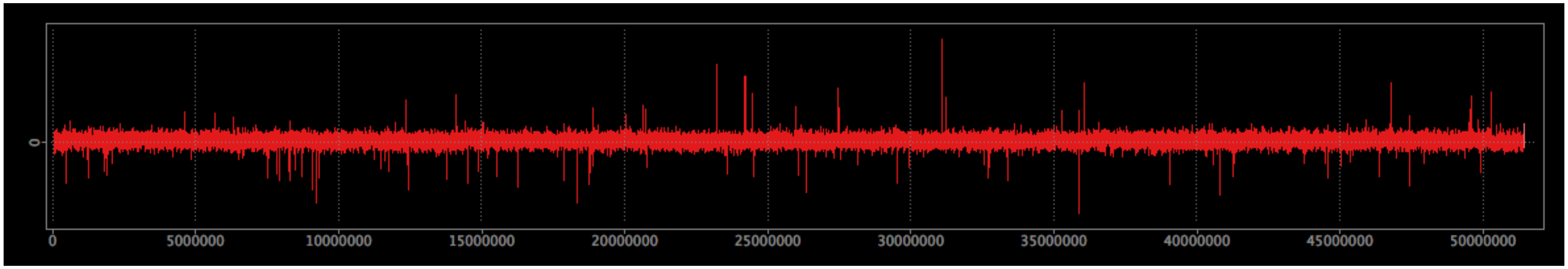
- Use optical maps to evaluate long range consistency as the gold standard

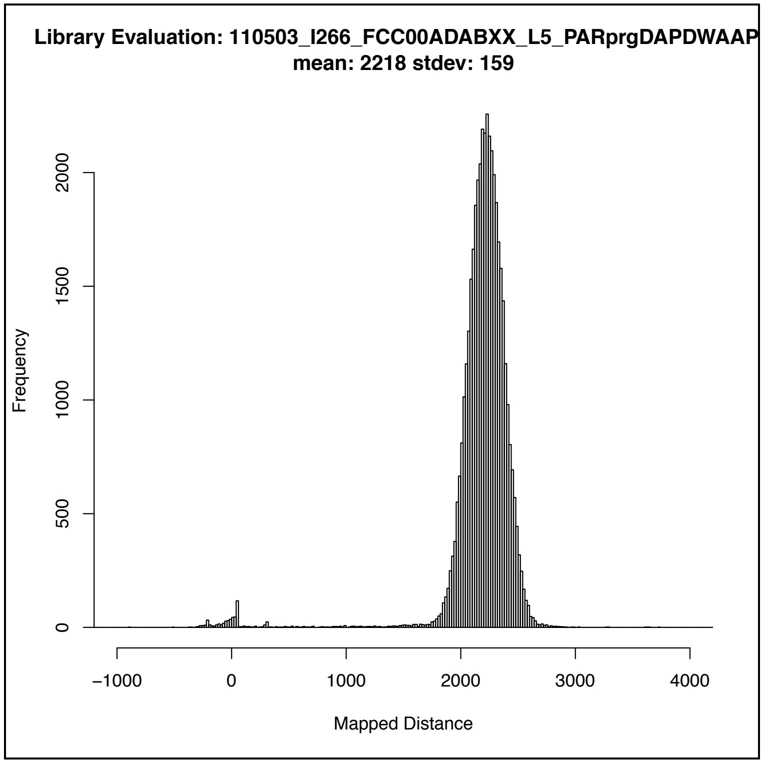
Fig. from Steve Goldstein

Parrot Metassembly

CE statistic (projected) across 51.1 Mbp scaffold

6C

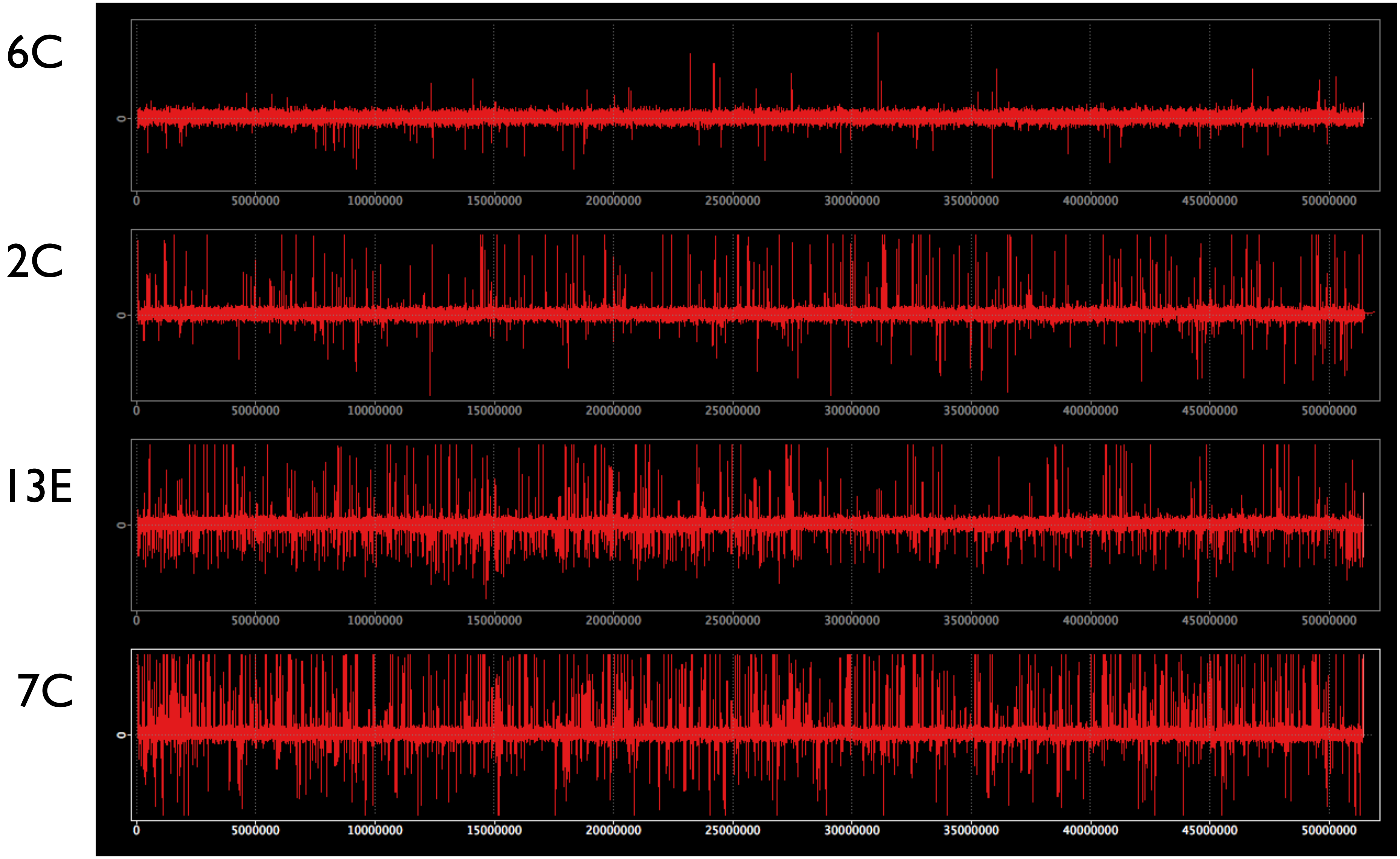




- Re-map 2kbp mates to each draft assembly, compute CE statistic at every position
- Extreme CE values are likely to be mis-assemblies
  - Can also look at coverage, mis-oriented mates, and other forensics features
  - Approximately 1.4 major events per Mbp

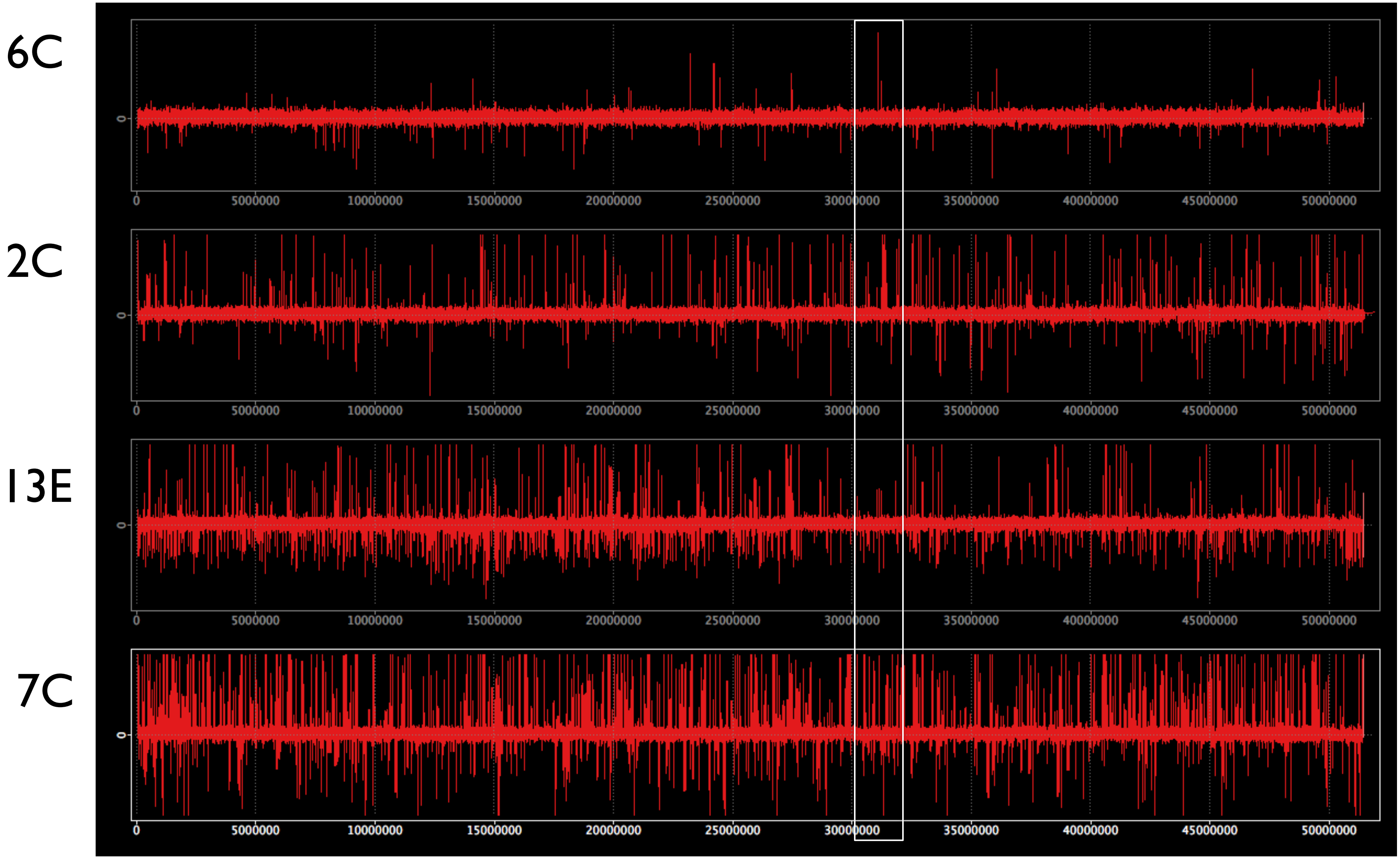
Parrot Metassembly

CE statistic (projected) across 51.1 Mbp scaffold



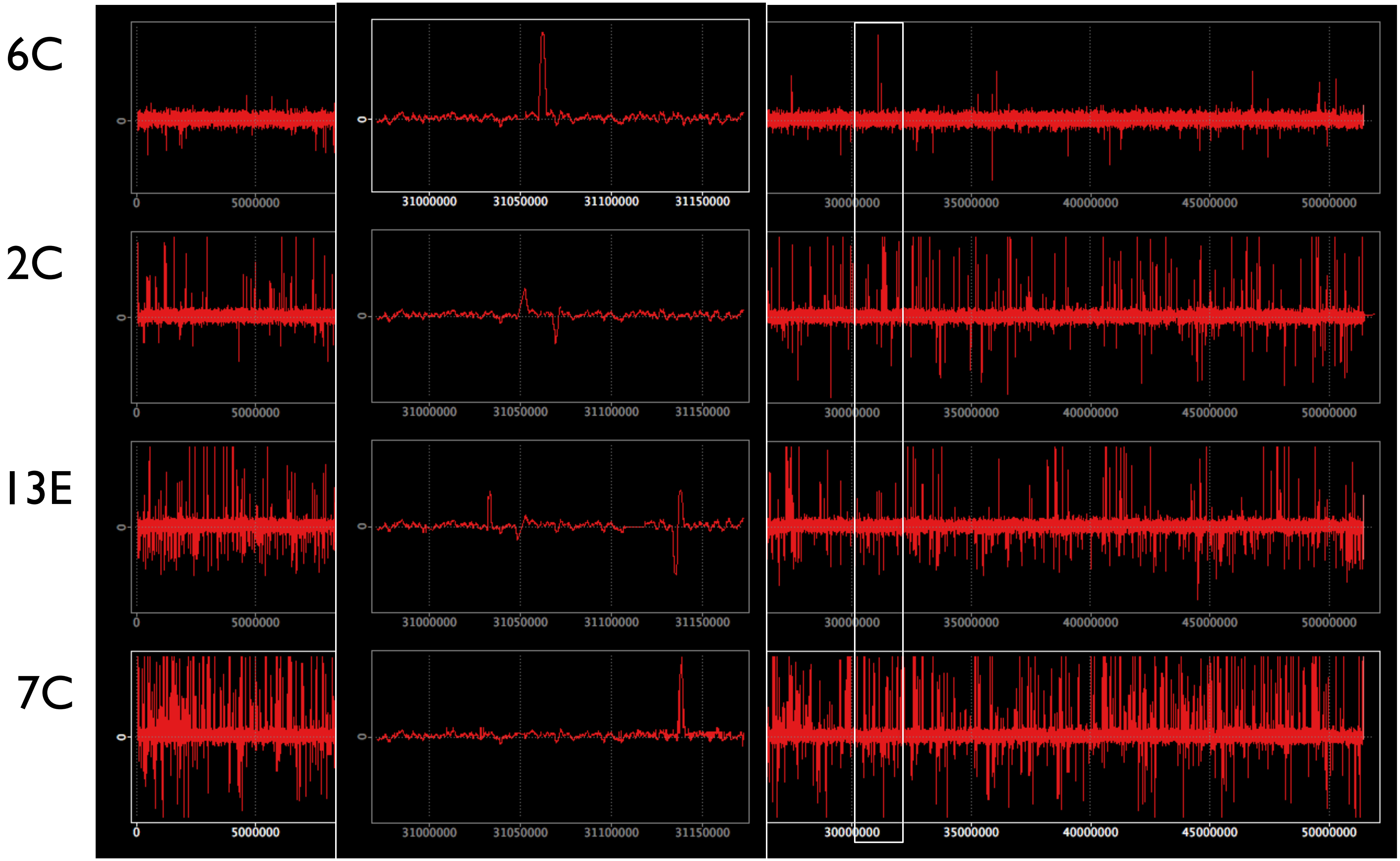
Parrot Metassembly

CE statistic (projected) across 51.1 Mbp scaffold

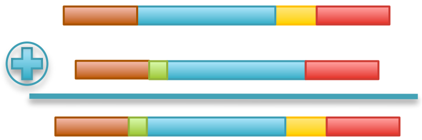


Parrot Metassembly

CE statistic (projected) across 51.1 Mbp scaffold



Summary



- Metassembly can correct nearly every mis-assembly and small gap in the parrot genome
  - Sliding window to select best representation along the 6C backbone

- Metassembly draws on individual strengths of each submission to locally optimize the problem
  - Different sequencing technologies
  - Different algorithms
  - Different parameters

- Summary/Consensus methods extremely powerful in virtually every complex optimization computation

Acknowledgements

Schatzlab
Paul Baranay
Rob Aboukhalil
Mitch Bekritsky
Hayan Lee
James Gurtowski
Giuseppe Narzisi

ND
Scott Emrich

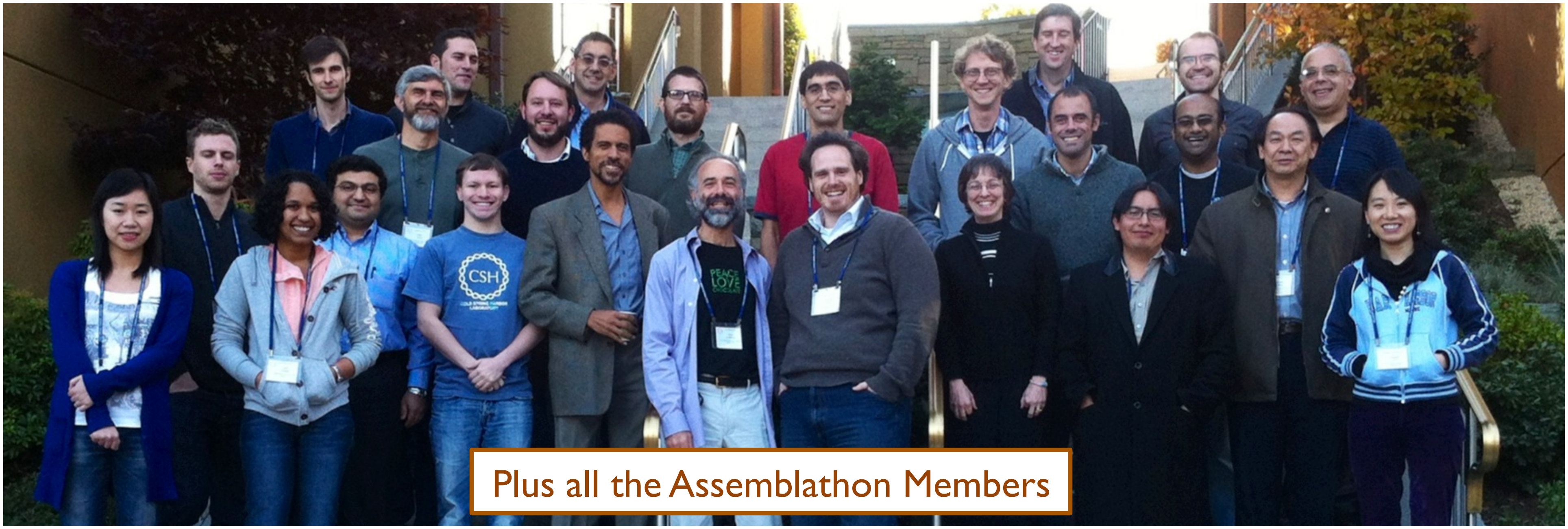
CSHL
McCombie Lab
Wigler Lab
Iossifov Lab

NBACC
Adam Phillipy
Sergey Koren

JHU
Steven Salzberg
Ben Langmead

Univ. of Maryland
Mihai Pop
Art Delcher
David Kelley
Cole Trapnell

Duke
Erich Jarvis



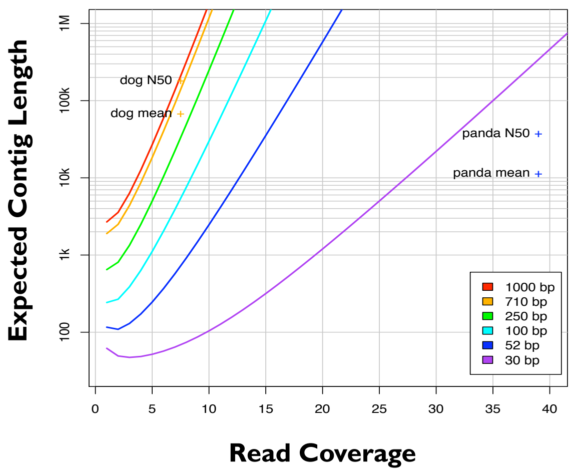
Plus all the Assemblathon Members

Thank You

http://schatzlab.cshl.edu
@mike_schatz / #AGBT

Ingredients for a good assembly

Coverage

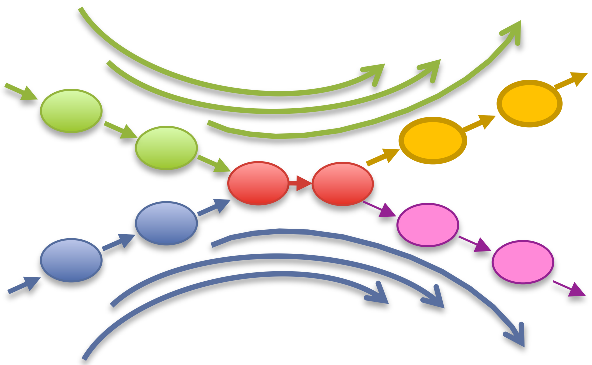


High coverage is required

– Oversample the genome to ensure every base is sequenced with long overlaps between reads

– Biased coverage will also fragment assembly

Read Length

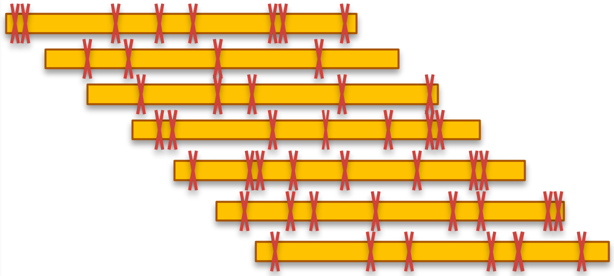


Reads & mates must be longer than the repeats

– Short reads will have ***false overlaps*** forming hairball assembly graphs

– With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

– Reads are assembled by finding kmers shared in pair of reads

– High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.