Celera Assembler: Adapting for the Future

Granger Sutton, Brian Walenz, Art Delcher, Eli Venter, Aaron Halpern, Gennady Denisov, Jason Miller, Justin Johnson

J. Craig Venter Institute

Center for Bioinformatics and Computational Biology, University of Maryland, College Park

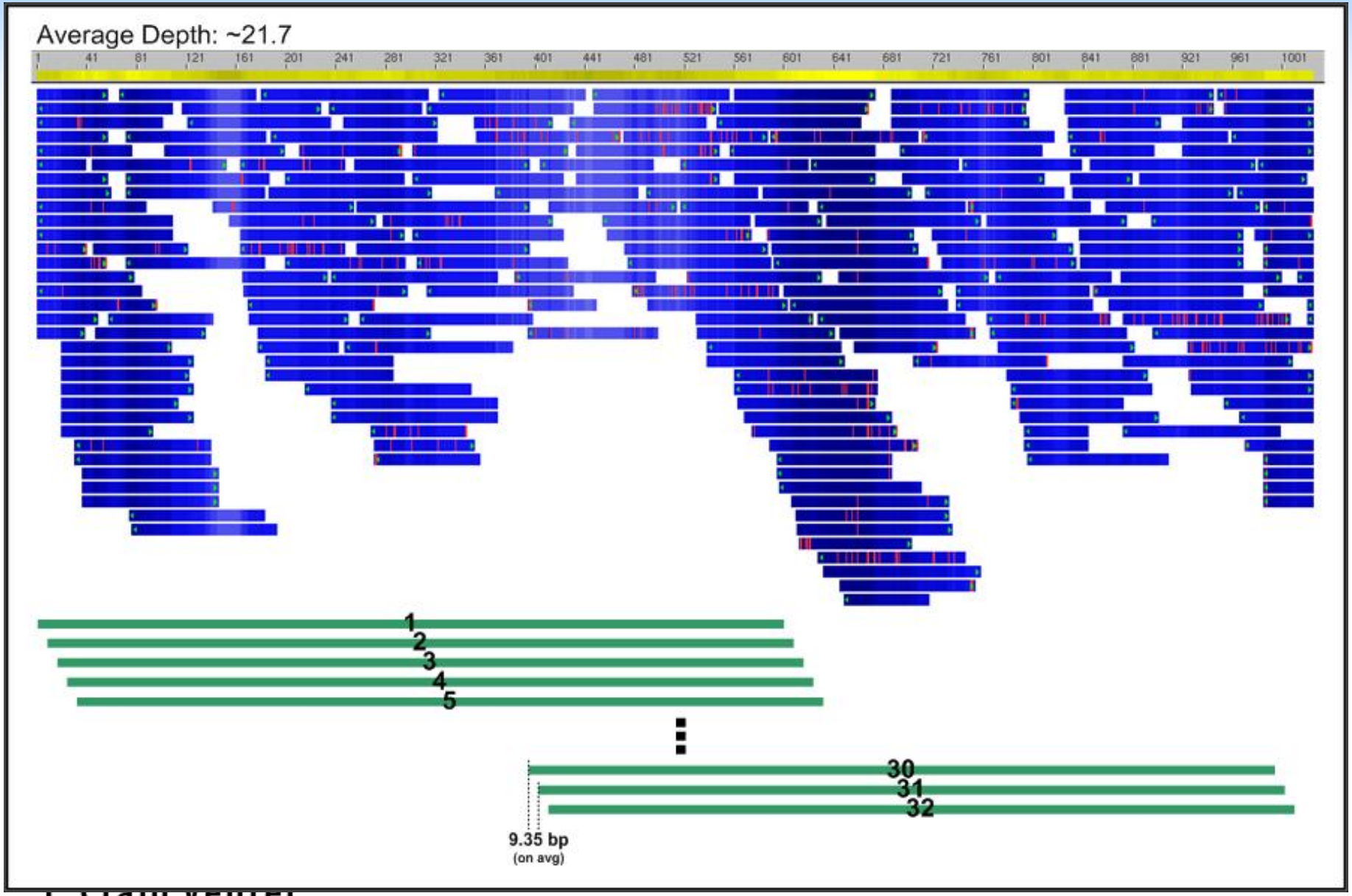
J. Craig Venter
INSTITUTE

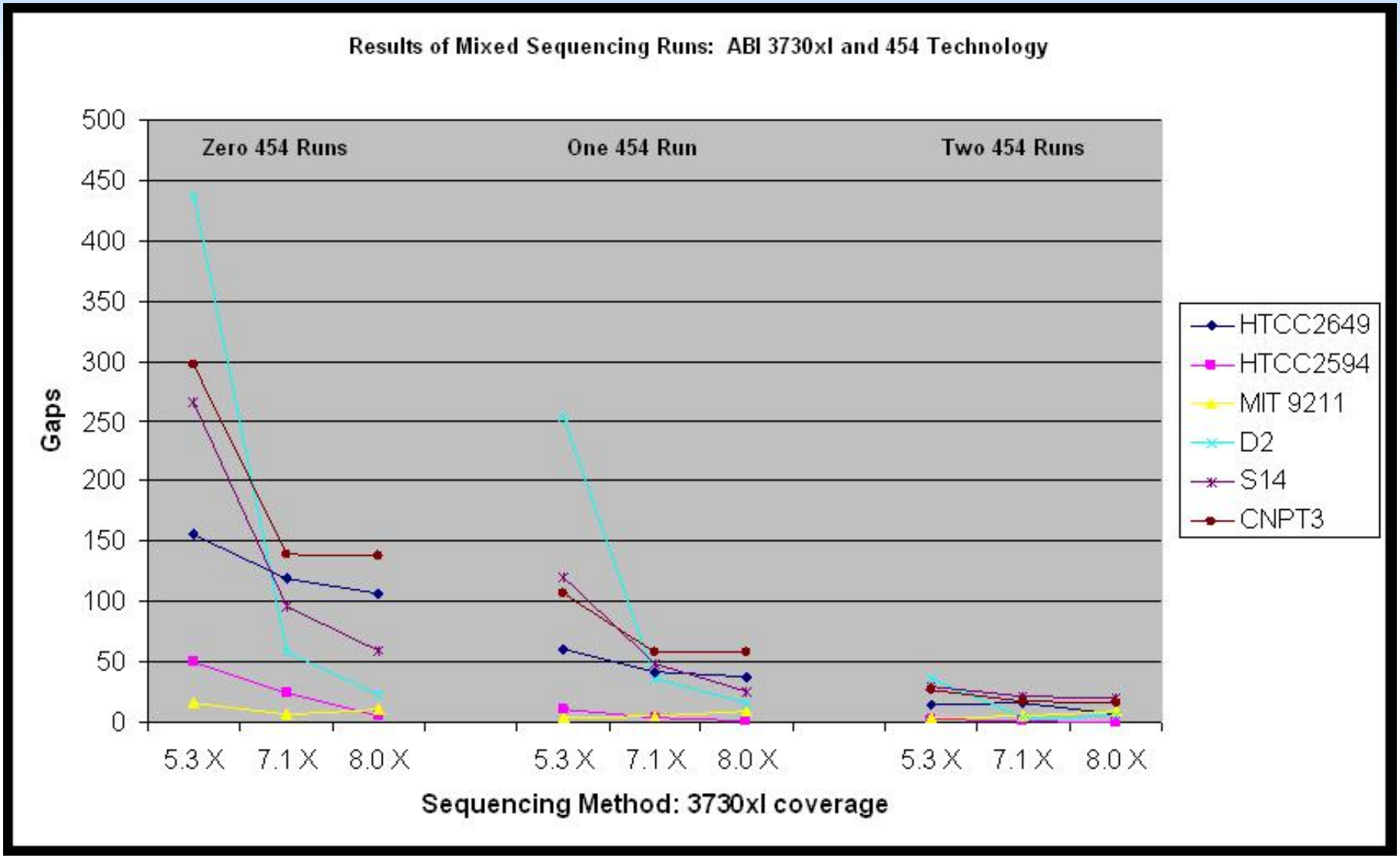
Celera Assembler

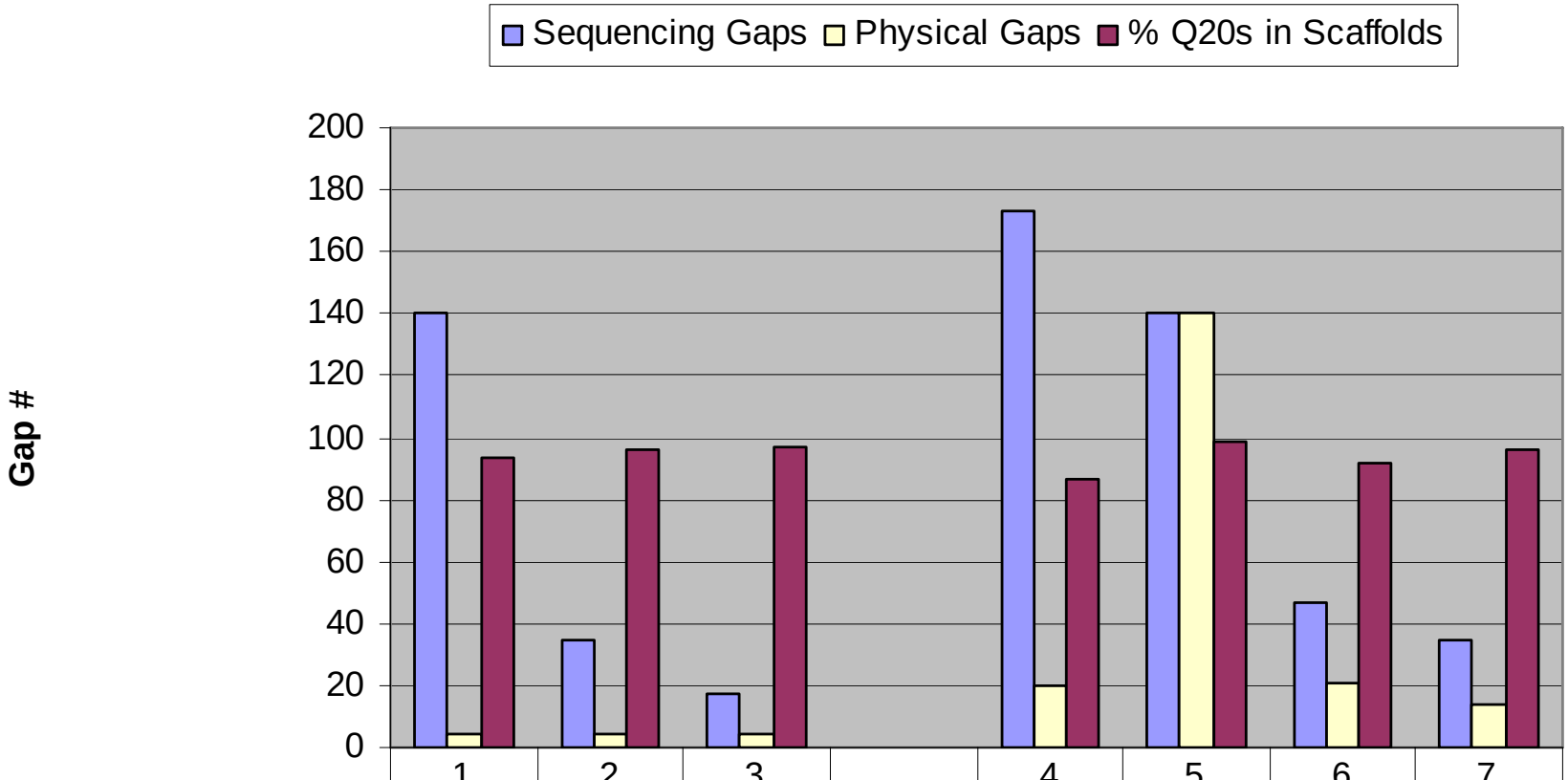
- Open source
- https://sourceforge.net/projects/wgs-assembler/
- Contact: Granger Sutton
- gsutton@jcvi.org

Current Hybrid Assembler

- Run Newbler
- Shred Newbler contigs in order to get longer fragments with fewer errors
- Input Sanger reads and shredded reads to Celera Assembler
- Maintain fragment depth of coverage
- Use low QVs for shreds so that Sanger reads dominate consensus


Average Depth: ~21.7
1
41
81
121
161
201
241
281
321
361
401
441
481
521
561
601
641
681
721
761
801
841
881
921
961
1001
1
2
3
4
5
30
31
32
9.35 bp
(on avg)


Results of Mixed Sequencing Runs: ABI 3730xl and 454 Technology
Zero 454 Runs
One 454 Run
Two 454 Runs
Gaps
500
450
400
350
300
250
200
150
100
50
0
5.3 X
7.1 X
8.0 X
5.3 X
7.1 X
8.0 X
5.3 X
7.1 X
8.0 X
Sequencing Method: 3730xl coverage
HTCC2649
HTCC2594
MIT 9211
D2
S14
CNPT3



	1	2	3		4	5	6	7
Sequencing Gaps	140	35	17		173	140	47	35
Physical Gaps	4	4	4		20	140	21	14
% Q20s in Scaffolds	93.67	95.91	97.05		86.35	98.86	92.03	96.06

Hybrid Assembly Strategy

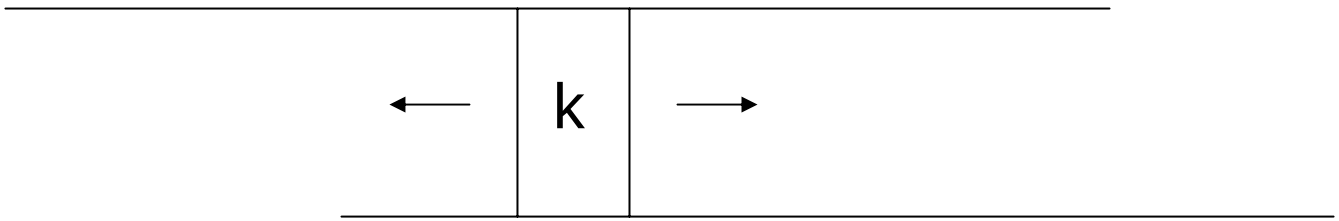
1. M063 - 0X Plasmid, .4X Fosmid, 1 FLX Run
2. M063 - 5.6X Plasmid, .4X Fosmid, 0 FLX Run
3. M063 - 5.6X Plasmid, .4X Fosmid, 1 FLX Run
4. lacto - 5X Plasmid - 0 FLX Run - 0 GS20 Run
5. lacto - 0X Plasmid - 1 FLX Run - 0GS20 Run
6. lacto - 5X Plasmid - 0 FLX Run - 1 GS20 Run
7. lacto - 5X Plasmid - 1 FLX Run - 0GS20 Run

New Hybrid Assembler

- Use 454 fragments directly
- Compress homopolymer runs for overlap detection
- Allow higher error rates in overlap extension
- Modify fragment error correction
- Modify repeat finding
- Modify consensus

Correcting k-mer Seeds

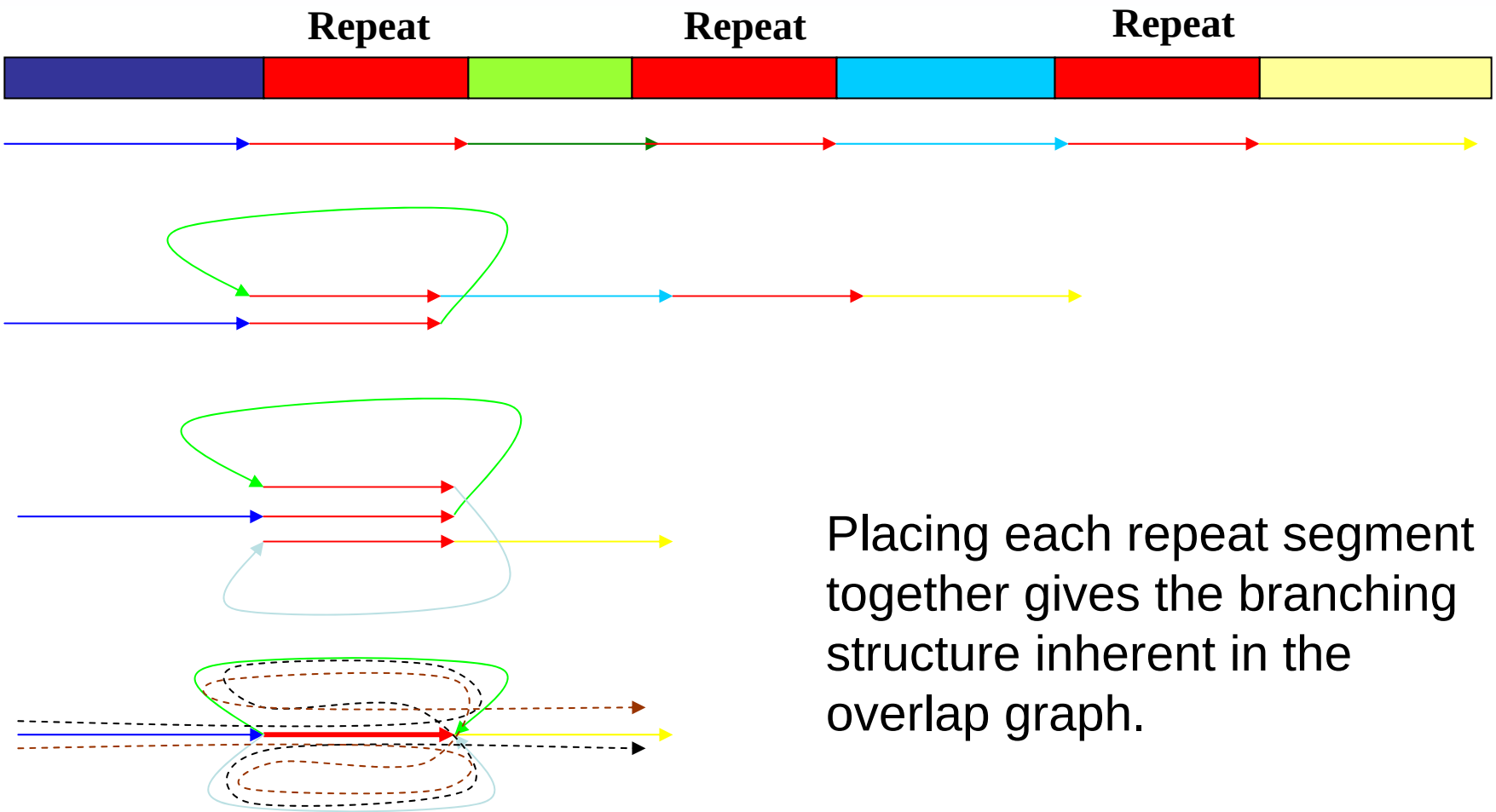
- Homopolymer run errors corrupt k-mers
- Truncate all homopolymer runs > t to length t
- Recompute k-mers to use for seeds
- Test on compressed D. willistoni fragments



Error Correction

```
1 ...AGT...AACT...TTACC...
2 ...AGT...AACT...TTACC...
3 ...AGT...AACT...TTACC...
4 ...AAT...AACT...TTACC...
5 ...AGT...AACT...TTACC...
-------------------------
6 ...AGT...AAAT...TGACC...
7 ...AGT...AAAT...TGACC...
8 ...AGT...AAAT...TGACC...
9 ...AGT...AAAT...TGACC...
```

Repeat / Chunk / Unitig / String Graph



Placing each repeat segment together gives the branching structure inherent in the overlap graph.

Consensus

A

```
...AGGCGA-TGAG...
...AGGCG-ATGAG...
...AGGCGA-TGAG...
...AGGCG-ATGAG...
...AGGCG-ATGAG...
...AGGCGA-TGAG...
...AGGCG-ATGAG...
...AGGCGAATGAG...
```

B

```
...AGGCGA-TGAG...
...AGGCGA-TGAG...
...AGGCGA-TGAG...
...AGGCGA-TGAG...
...AGGCGA-TGAG...
...AGGCGA-TGAG...
...AGGCGA-TGAG...
...AGGCGAATGAG...
```

Auto-Closure

- Will output gap or problematic region spanning clones
- Can design primers or output flanking sequence
- Can output consistently problematic mate pairs
- Established formats?

Leveraging Auto-Closure

- Assembler must respect constraints on finishing reads
- Assembler should allow and respect previously assembled contigs if desired
- Transposon mate pairs should be supported

Metagenomics

- Compilation macro and parameter settings for metagenomic assembly
- Shorter k-mers
- Higher error in overlaps allowed (10-12%)
- Significant increase in assembly for deep metagenomic data
- Significant increase in assembly for some polymorphic genomes like body louse

Celera Assembler

- https://sourceforge.net/projects/wgs-assembler/
- Granger Sutton gsutton@jcvi.org
- Generates .ace .agp files
- What other files needed?
- What other input must be supported?