Manual Sequence Improvement of Bacterial Genomes

Aye Wollam

Genome Sequencing Center
Washington University School of Medicine
St. Louis, MO

awollam@watson.wustl.edu

Purpose

- To produce sequence assemblies that are qualitatively better than draft assemblies.
- To do so in less cost, time and labor than those needed for finishing genomes.

awollam@watson.wustl.edu

How we are using it

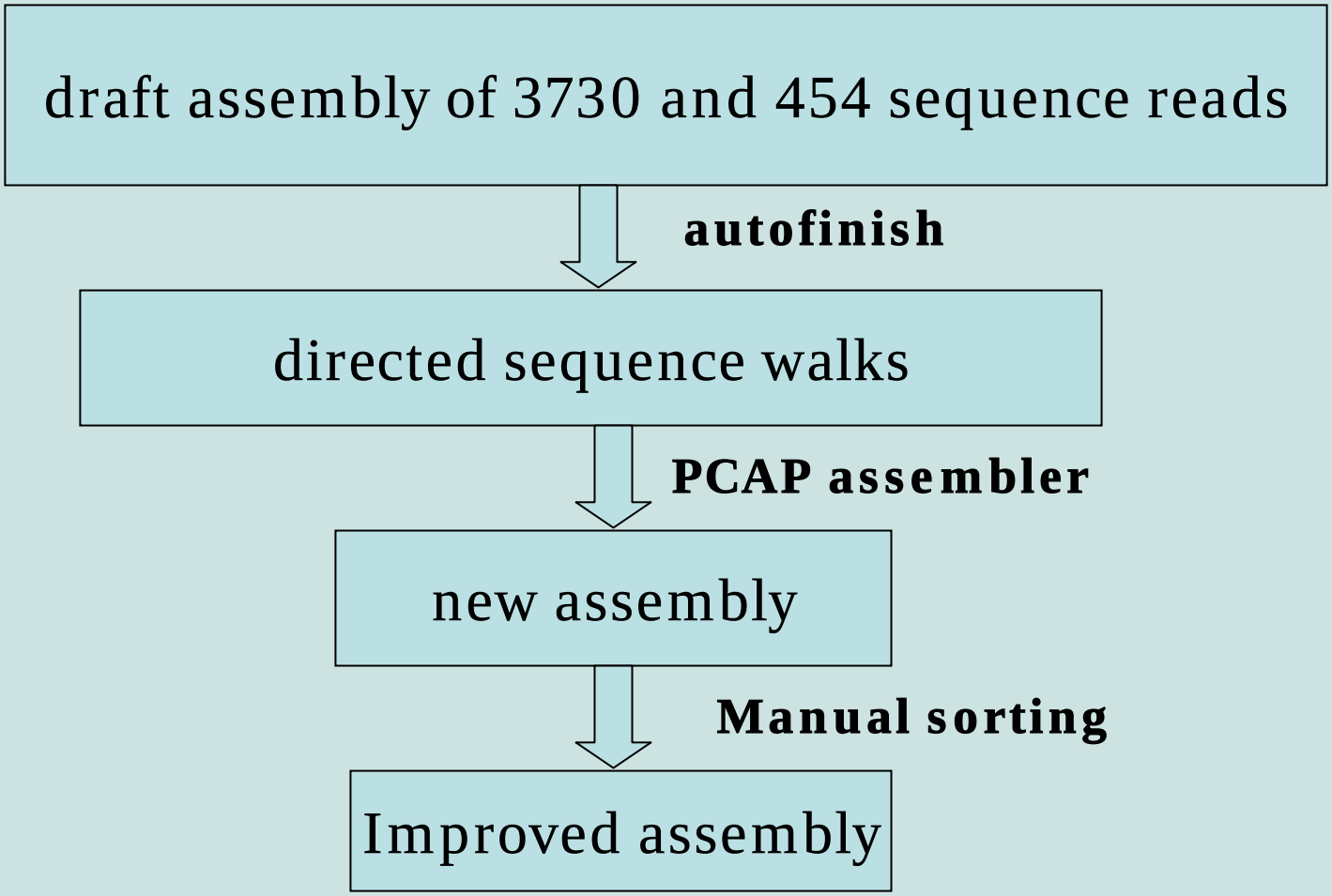
- For investigative sampling of a population, such as Human Gut Microbiome Project.
- 100 bacterial genomes of human gut are being sequenced, manually improved, and annotated.
- Selected genomes will be finished to the full.
- 1000 genomes in future pipeline

awollam@watson.wustl.edu

What are the steps in manual sequence improvement of bacterial genomes?

awollam@watson.wustl.edu

Steps involved in manual sequence improvement



Manual sorting process

- Perform joins missed by assembler
- Correct consensus errors
- Sort misplaced reads
- Identify and correct misassembled regions
- Order and orientation of contigs are provided if possible

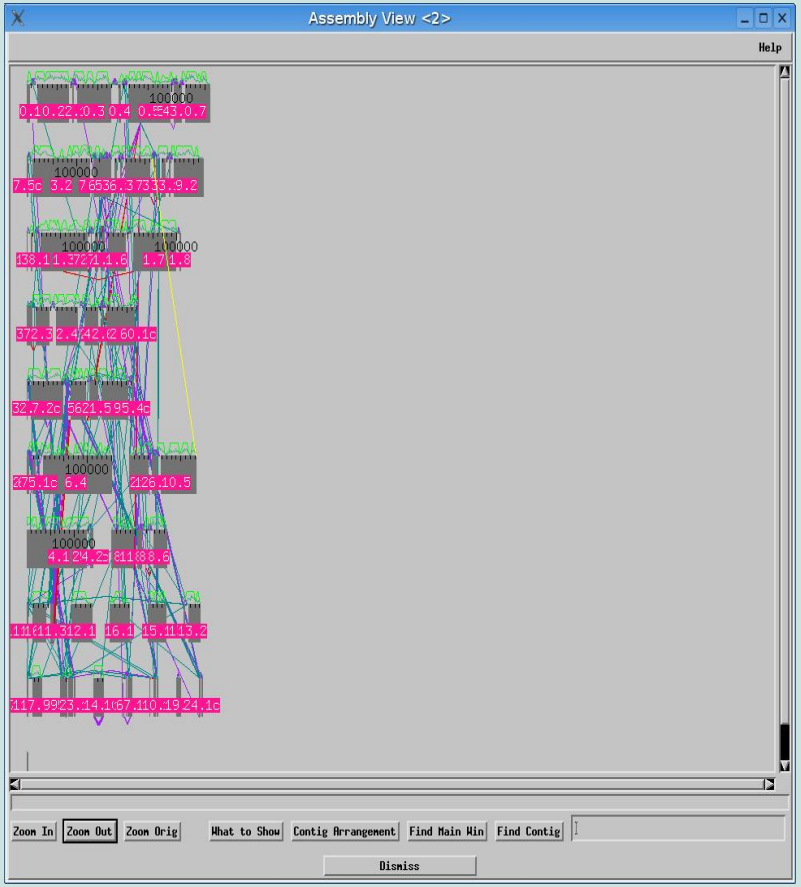
awollam@watson.wustl.edu

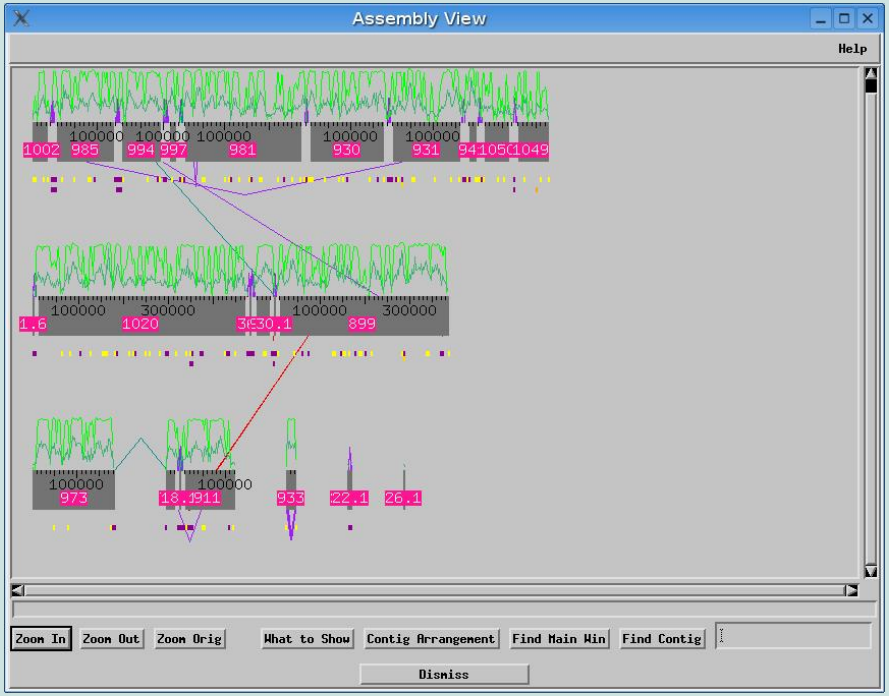
Perform joins missed by assembler

- look for missed joins
- sequence searches within a scaffold
- detect overlaps between scaffolds

awollam@watson.wustl.edu

Perform joins missed by assembler

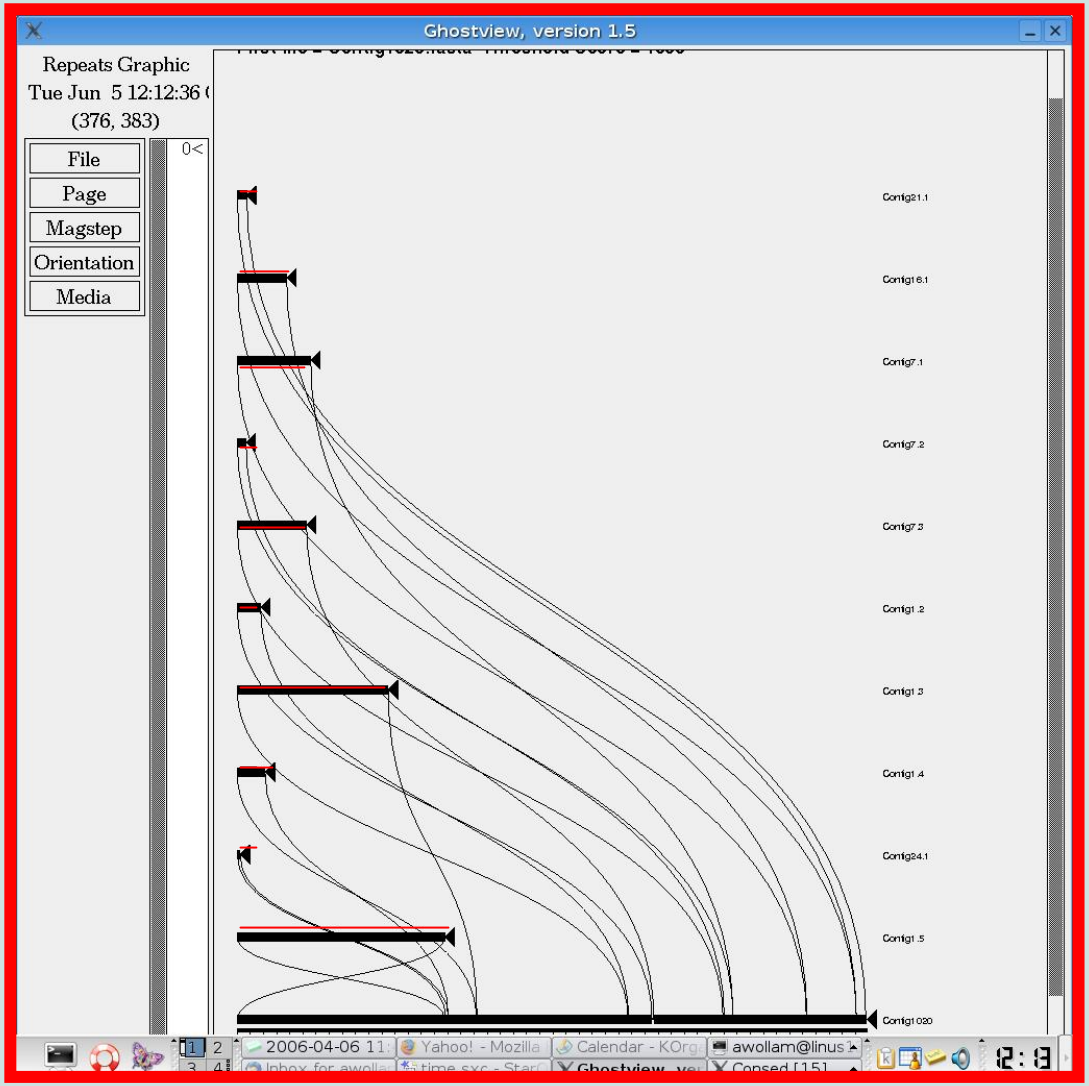




joins not made by assembler;

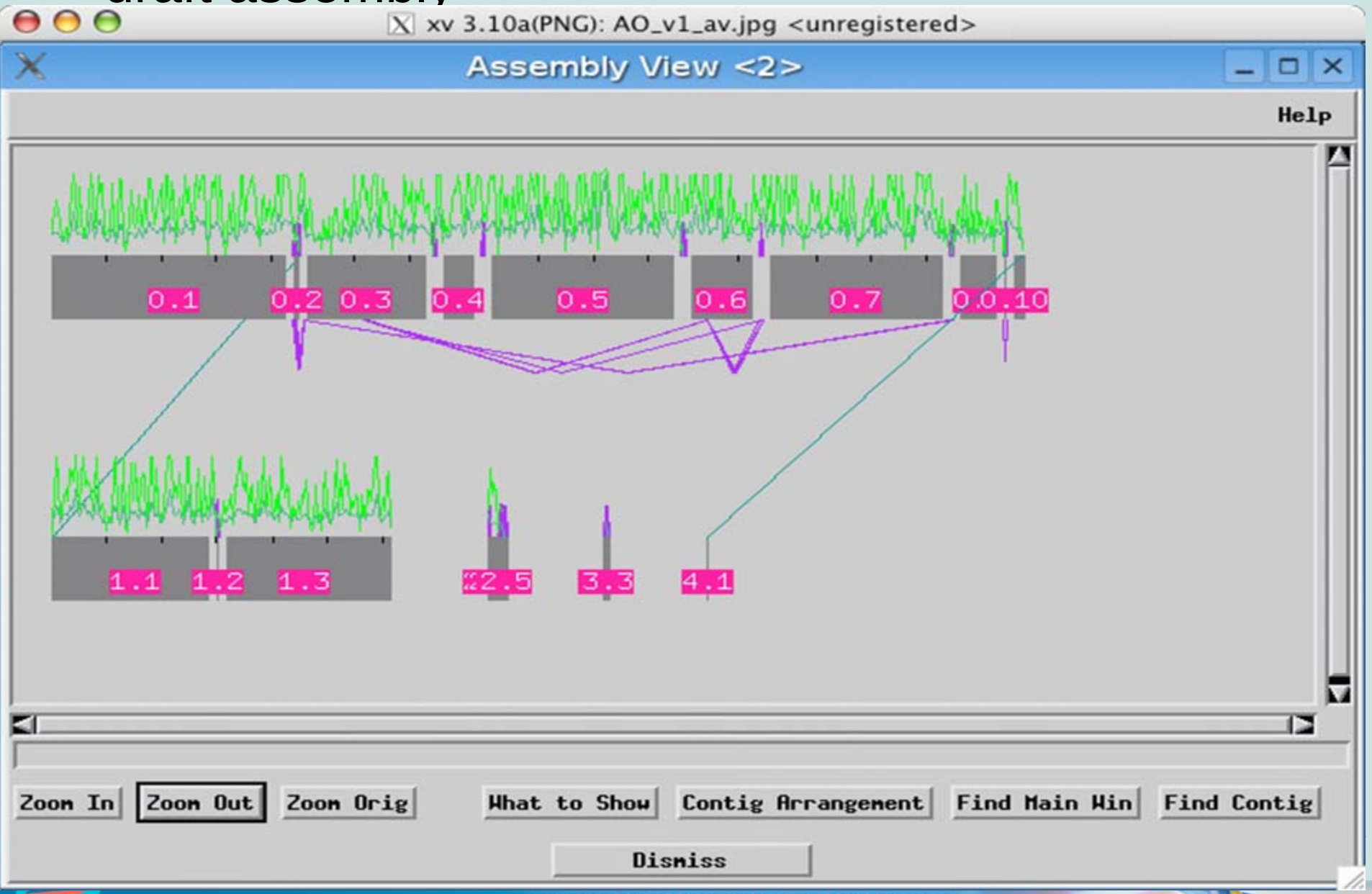
contigs sorted and joins made manually

Perform joins missed by assembler

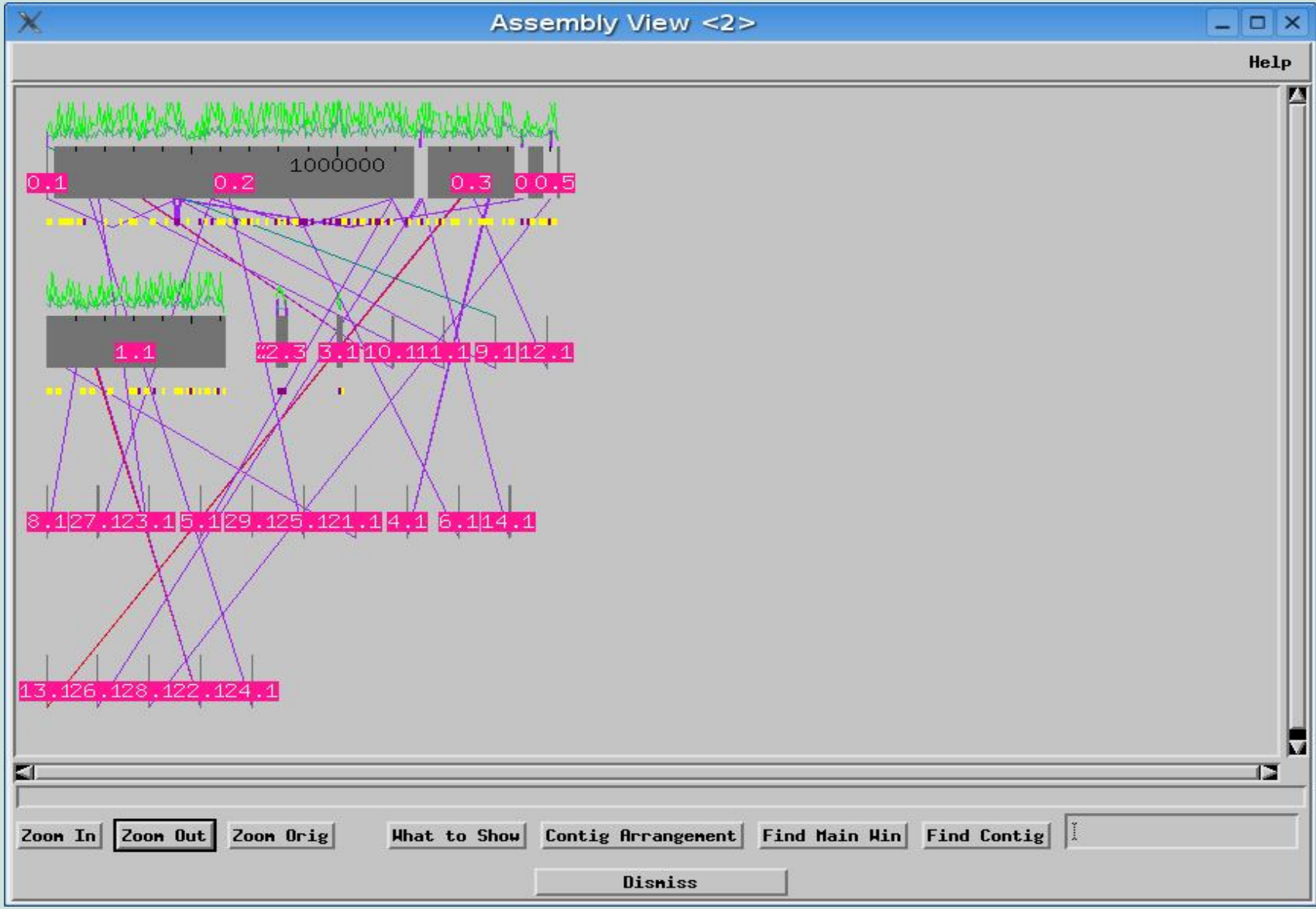


joins made manually in msi process;
red lined contigs(draft contigs) compared with MSI contig

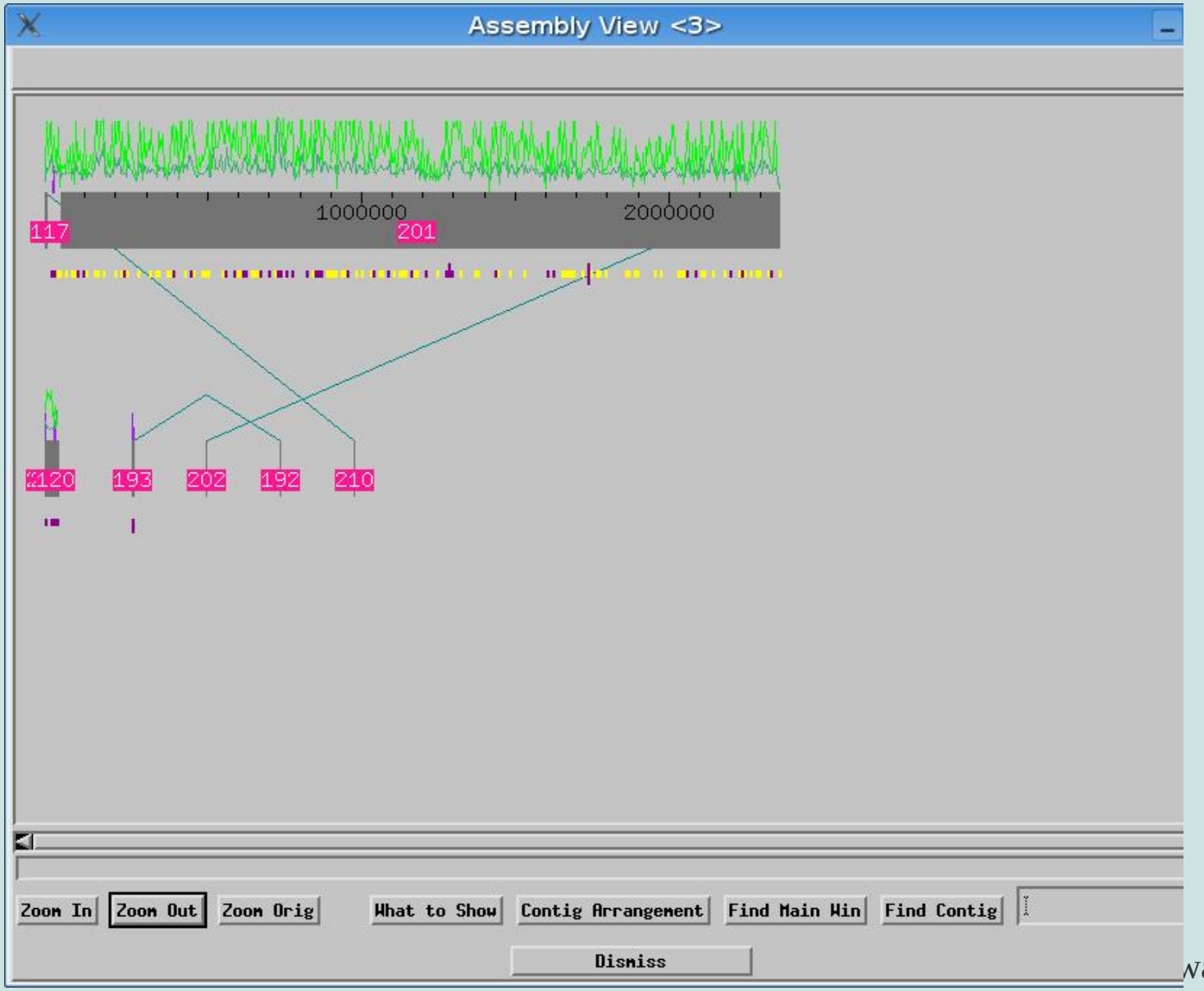
draft assembly



2nd assembly with autofinish data



After manual improvement



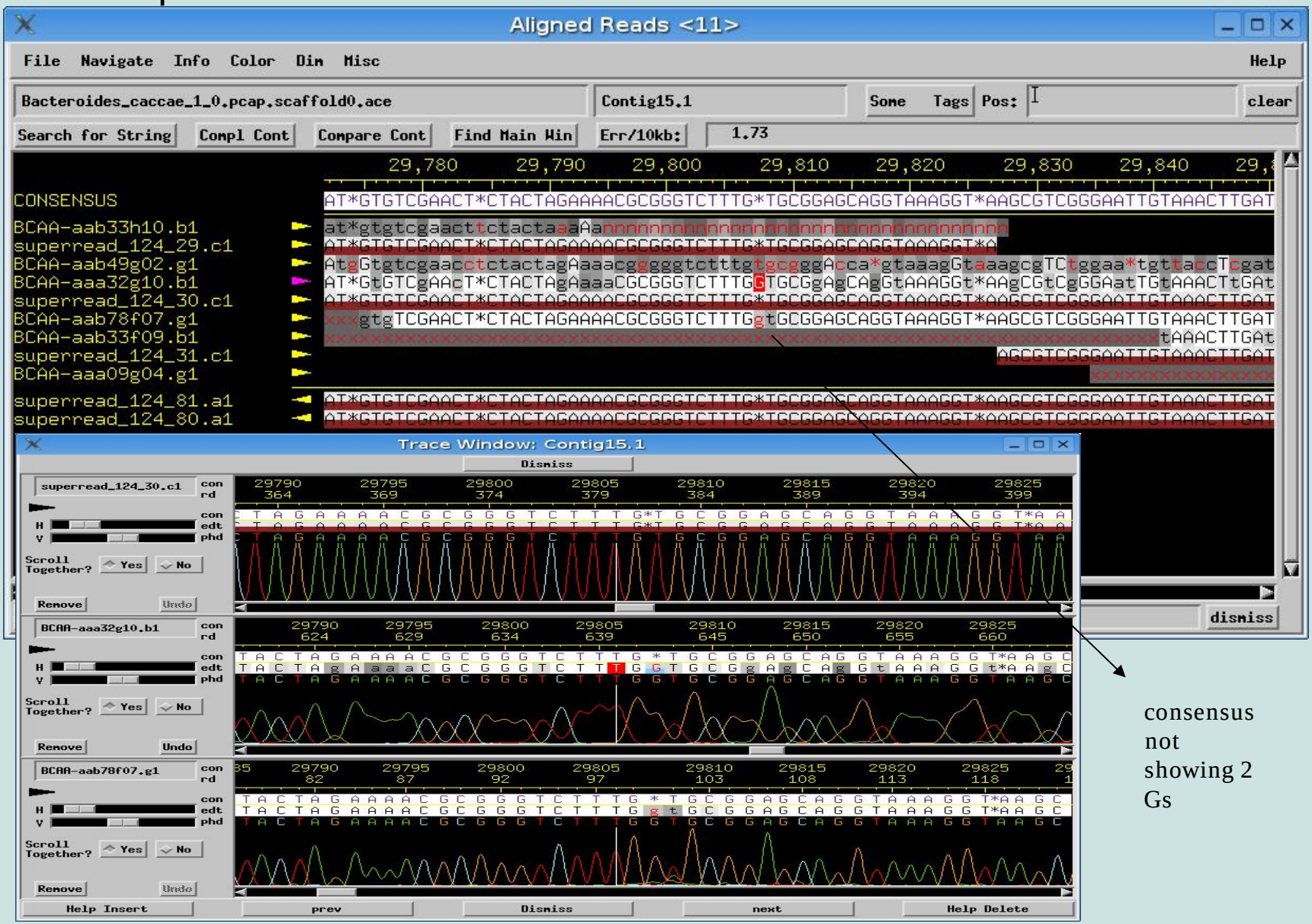
watson.wustl.edu

Correct sequence errors

- Navigate by low quality regions and high quality discrepant bases
- correct consensus errors
- trim off irrelevant sequences at contig ends

awollam@watson.wustl.edu

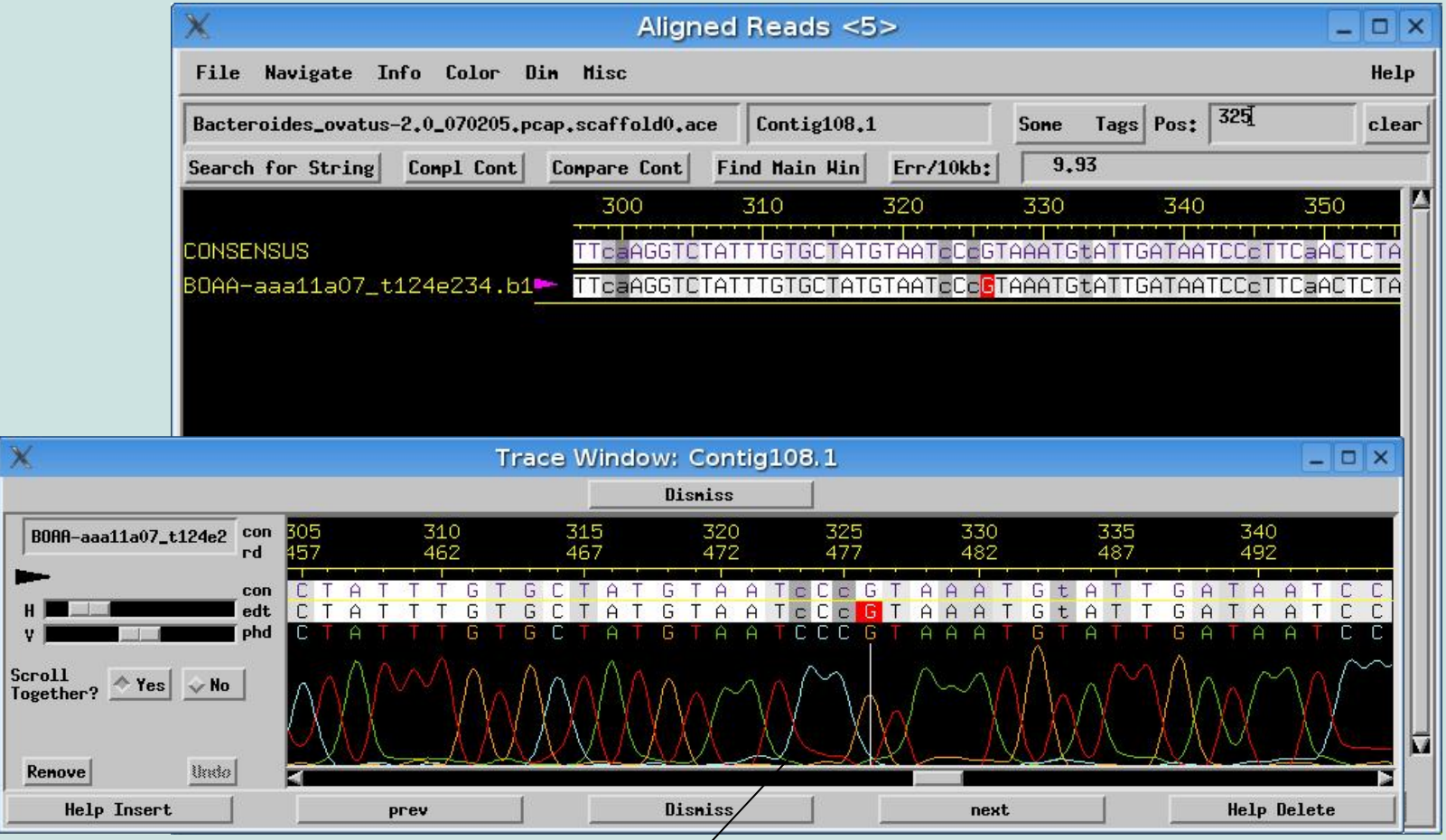
Correct sequence errors



consensus not showing 2 Gs

awollam@watson.wustl.edu

Correct sequence errors

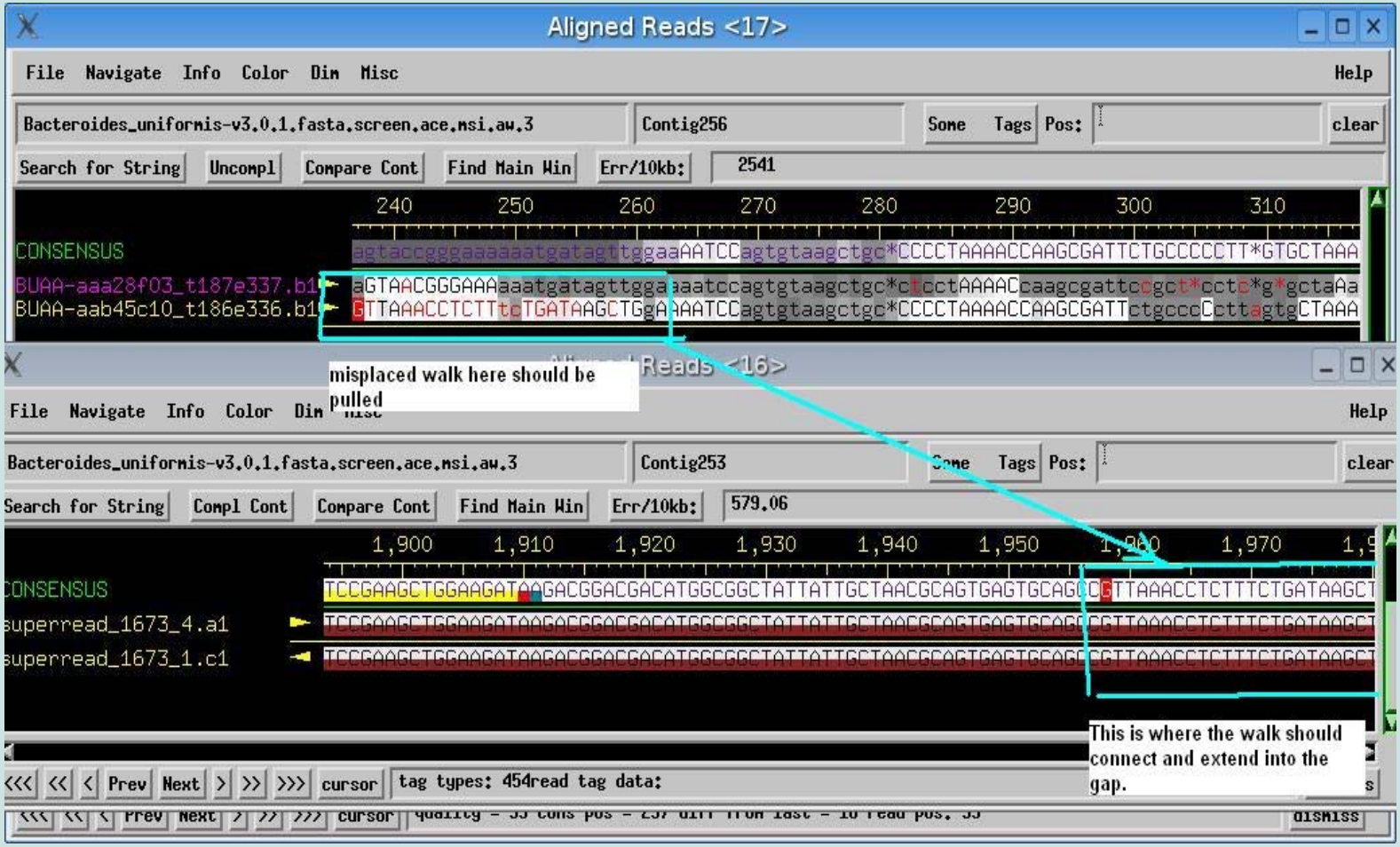


wrong base calls (3 Cs) instead of 2 Cs here

awollam@watson.wustl.edu

Sort misplaced reads

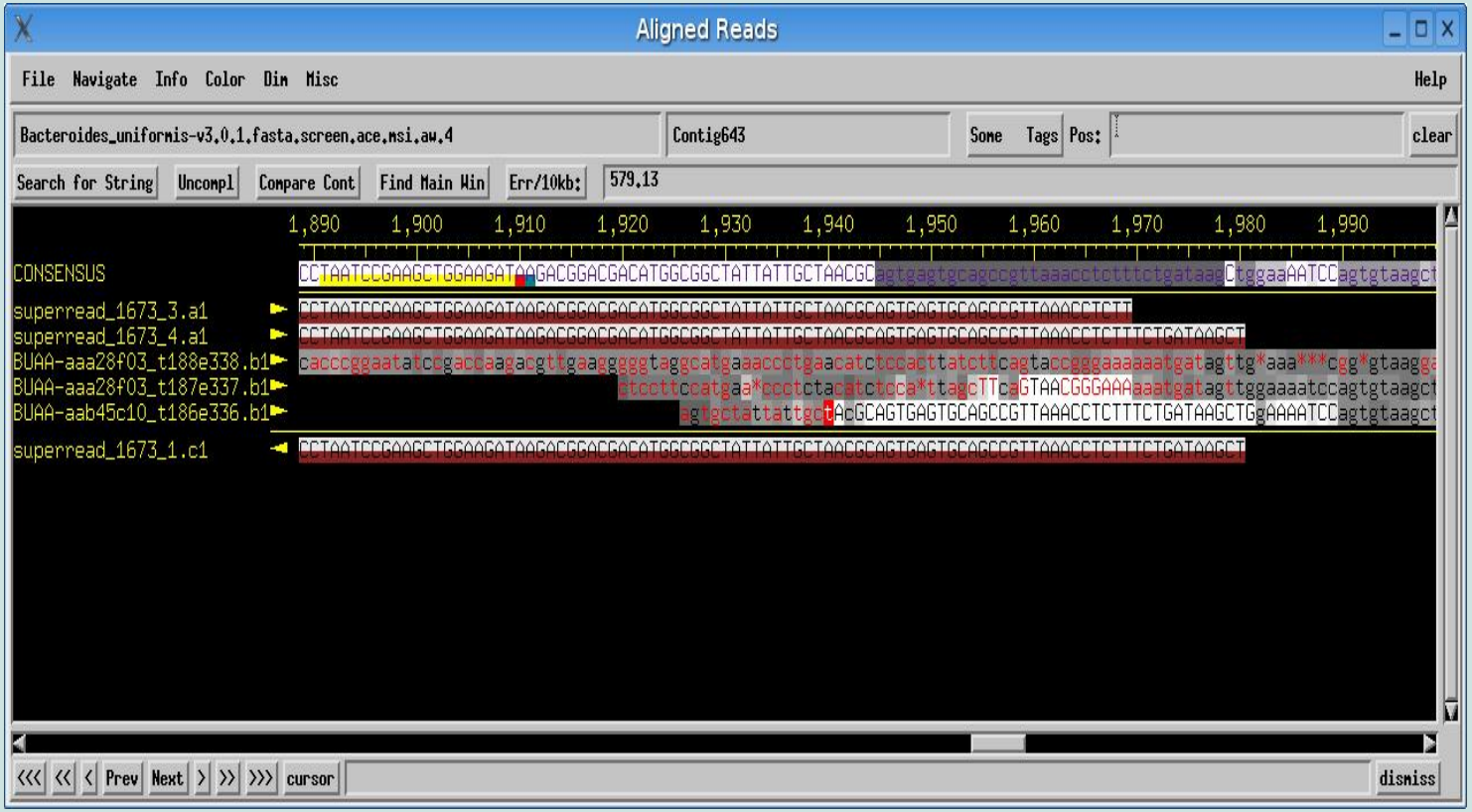
example 1a: sequence walks disconnected to the priming site



awollam@watson.wustl.edu

Sort misplaced reads

example 1b: sequence walks placed to the priming site. Extended 850 base pair into the gap



awollam@watson.wustl.edu

Identify and correct misassembled regions

- look for evidence of misassemblies based on read pair inconsistencies and high quality discrepant reads
- rearrange read placement, tear and join contigs as needed.

awollam@watson.wustl.edu

high quality discrepant bases indicating collapsed repeat

Aligned Reads <2>

File Navigate Info Color Dim Misc Help

Eubacterium_ventriosum-2.0_061027.pcap.scaffold0. | Contig2.1 | Some Tags | Pos: | clear

Search for String | Compl Cont | Compare Cont | Find Main Win | Err/10kb: 0.34

377,810 377,820 377,830 377,840 377,850 377,860

CONSENSUS CTTTTTTCTTGGAGTACCAGCCATCAAAAGTATATCCTTCACGTCTGATGGCAGGAGCGT

EVAA-aaf21f08.g1 cTTTTttctggaagtaacaatccatcaagt

EVAA-aah04h08.g1 CTTTTTTCTTGgAGTACCAGCCATCAAAAGTATATCCTTCaCGTCTGAtGGCAgGAgCGt

EVAA-aaf44f07.g1 CttTtttcttGgaga*ccagccatcatatgtatgtcctTCacgtCtGAtGgtaggtgcgt

EVAA-aag44c01.g1 CTTTTTTCTTGGAGTACCAGCCATCAAAAGTATATCCTTCACGTCTGATGGCAGGAGCGT

superead_1193_1.c1 CTTTTTTCTTGGAGTACCAGCCATCAAAAATATATCCTTCACGTCTGATGGCAGGAGCGT

EVAA-aah29a12.g1

EVAA-aaf10h03.b1 CTTTTTTCTTGGAGTaCCAgccatcaaaaatataTCCTTCACGTCTGATgGCAGgagctt

EVAA-aaf82e05.b1 CTTTTTTCTTGGAGTACCAGCCATCAAAAGTATATCCTTCACGTCTGATGGCAGGAGCGT

EVAA-aaf56d03.b1 CTTTTTTCTTGGAGTACCAGCCATCAAAAATATATCCTTCACGTCTGATGGCAGGAGCGT

EVAA-aah24b02.b1 *tTTTTTcTtGGAgTaCCAgCCATCAAAaGTATatctTtCAcGTctGatggCAGGAGCGT

superead_1193_1.a1 CTTTTTTCTTGGAGTACCAGCCATCAAAAATATATCCTTCACGTCTGATGGCAGGAGCGT

Trace Window: Contig2.1

Dismiss

EVAA-aaf56d03.b1 con rd 377815 238 377820 233 377825 228 377830 223 377835 218 377840 213 377845 208 377850 203

con GGAGTACCAGCCATCAAAAgTATATCCTTCACGTCTGATG

edt GGAGTACCAGCCATCAAAAATATATCCTTCACGTCTGATG

phd GGAGTACCAGCCATCAAAAATATATCCTTCACGTCTGATG

H V

Scroll Together? Yes No

Remove Undo

EVAA-aaf82e05.b1 con rd 377815 242 377820 237 377825 232 377830 227 377835 222 377840 217 377845 212 377850 207

con GGAGTACCAGCCATCAAAAgTATATCCTTCACGTCTGATG

edt GGAGTACCAGCCATCAAAAGTATATCCTTCACGTCTGATG

phd GGAGTACCAGCCATCAAAAGTATATCCTTCACGTCTGATG

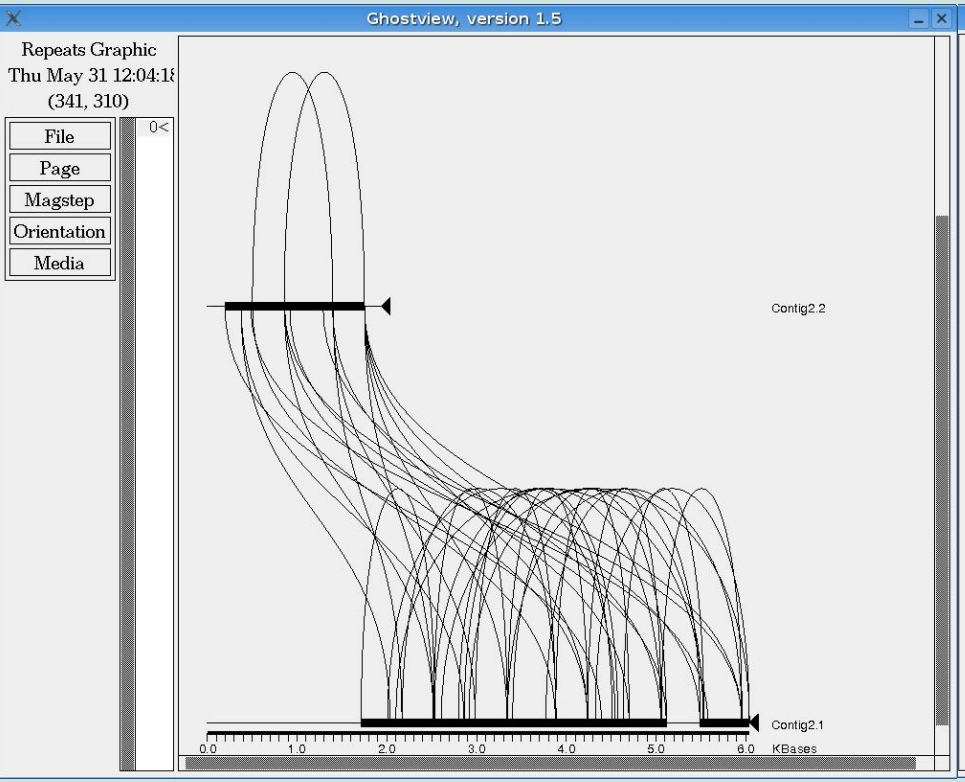
H V

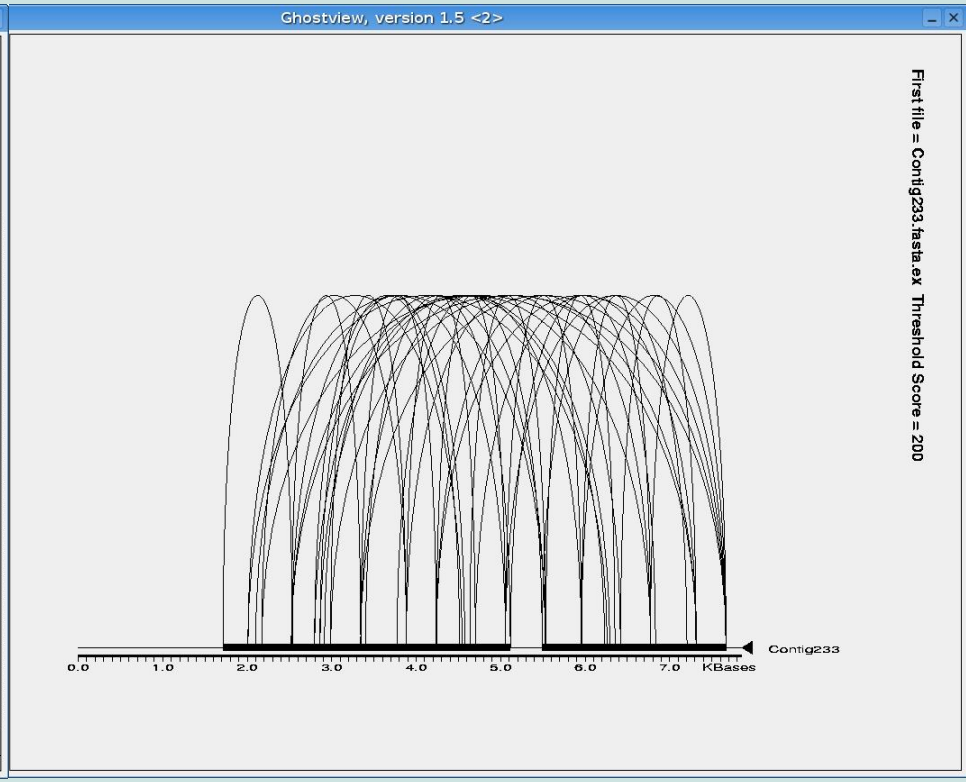
Scroll Together? Yes No

Remove Undo

Help Insert | prev | Dismiss | next | Help Delete

Sort and correct misassembled regions

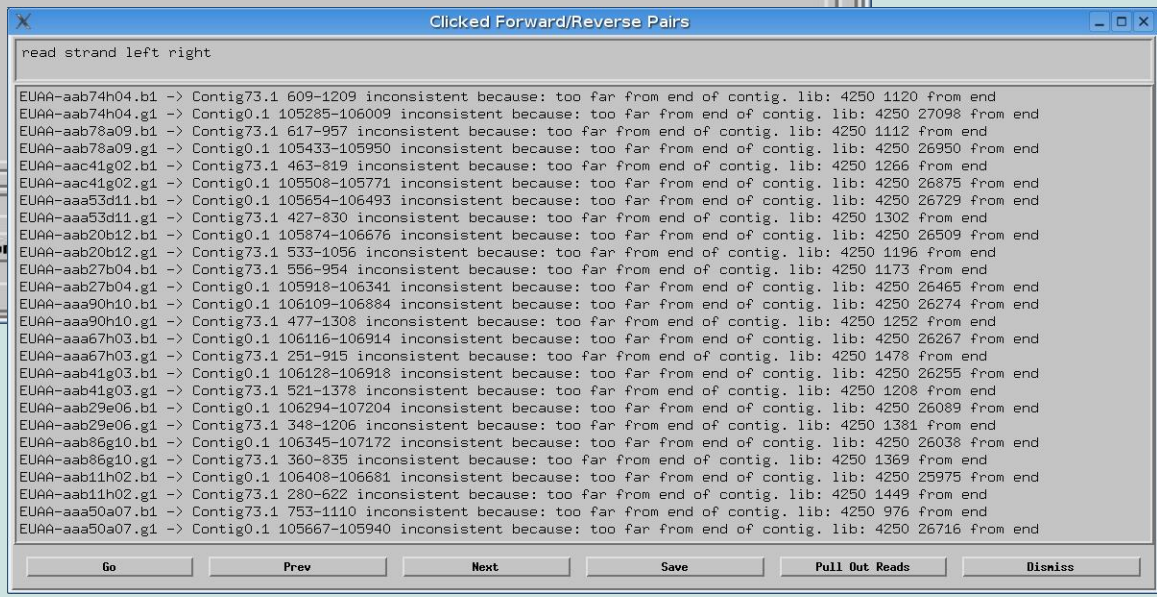




- two contigs with collapsed repeats are sorted and the result is one contig with repeats separated after manual improvement. No high quality mismatched bases in this contig.

awollam@watson.wustl.edu

Sort and correct misassembled regions



inconsistent read pairs

Sort and correct misassembled regions

- read pairs are now consistent
- repeat copies separated

awollam@watson.wustl.edu

What we don’t do in MSI

- additional directed walks $$
- PCR for unoriented contigs $$
- mini library (short insert libraries) for hard to close gaps $$$

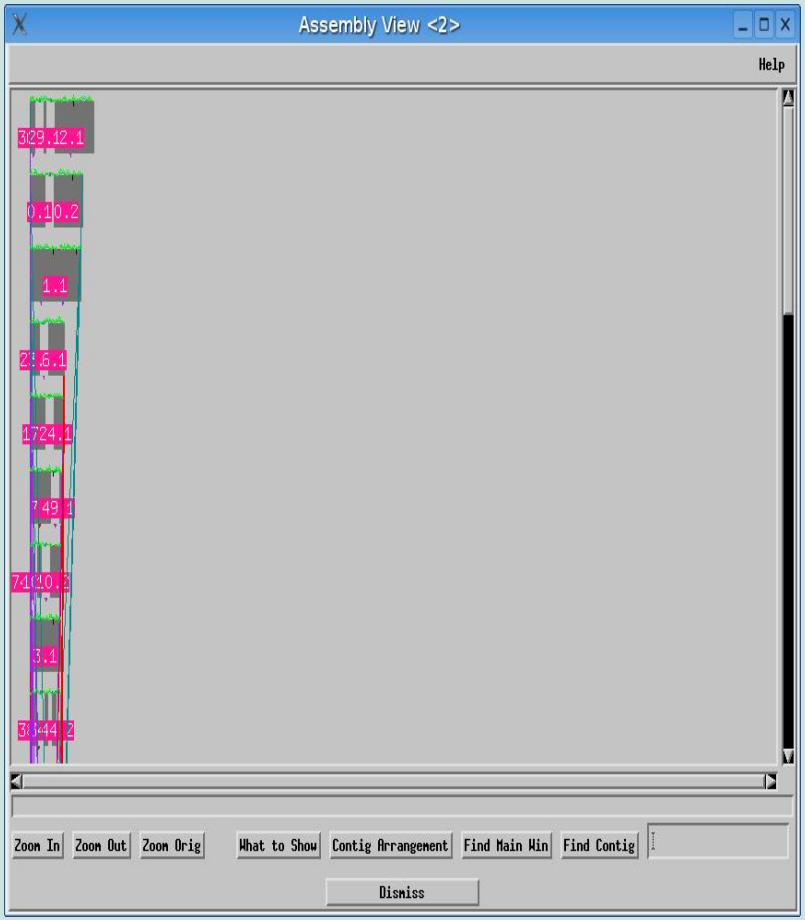
awollam@watson.wustl.edu

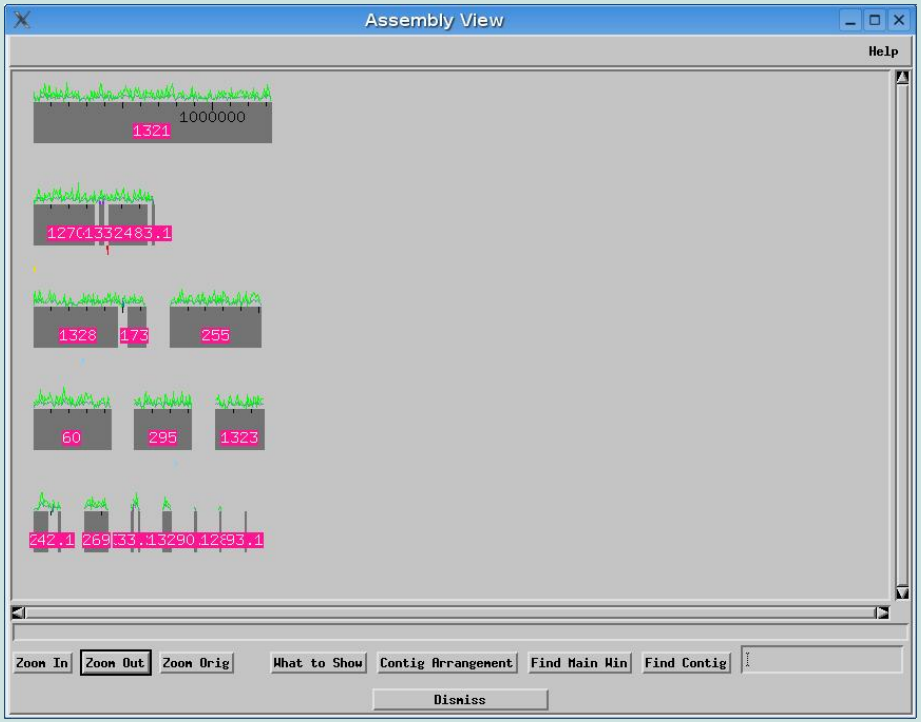
Results

- Thus far, twelve bacterial genomes have been improved
- Results show that MSI assemblies have longer N50 contig lengths, more contiguity, and more consistent read pair placement indicating better representation of the genome.

awollam@watson.wustl.edu

Increased contiguity (example1: BC genome)





draft assembly: fragmented

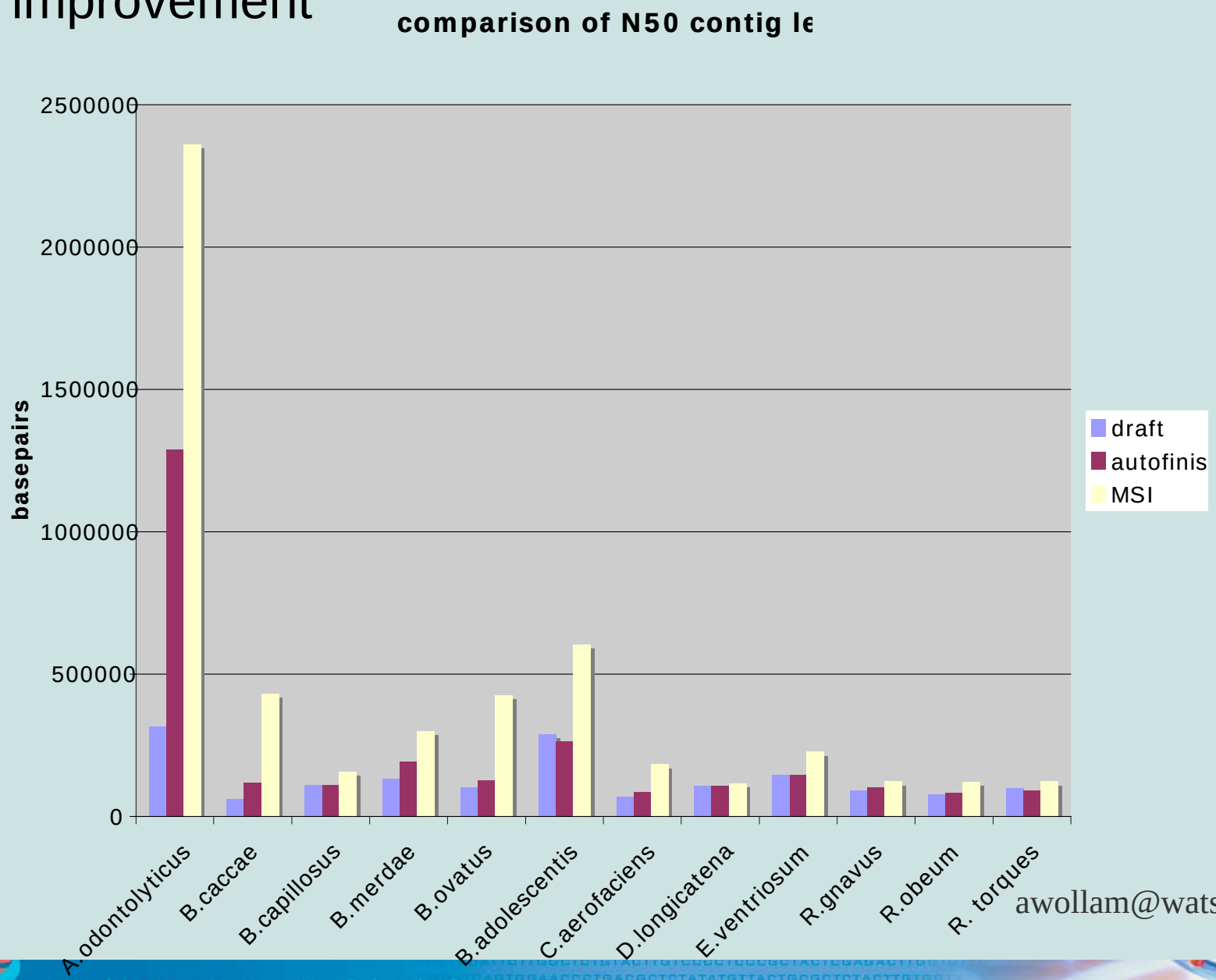
maximum contig: 231,267 bp

MSI assembly: more contiguous

max contig:1,329,168 bp

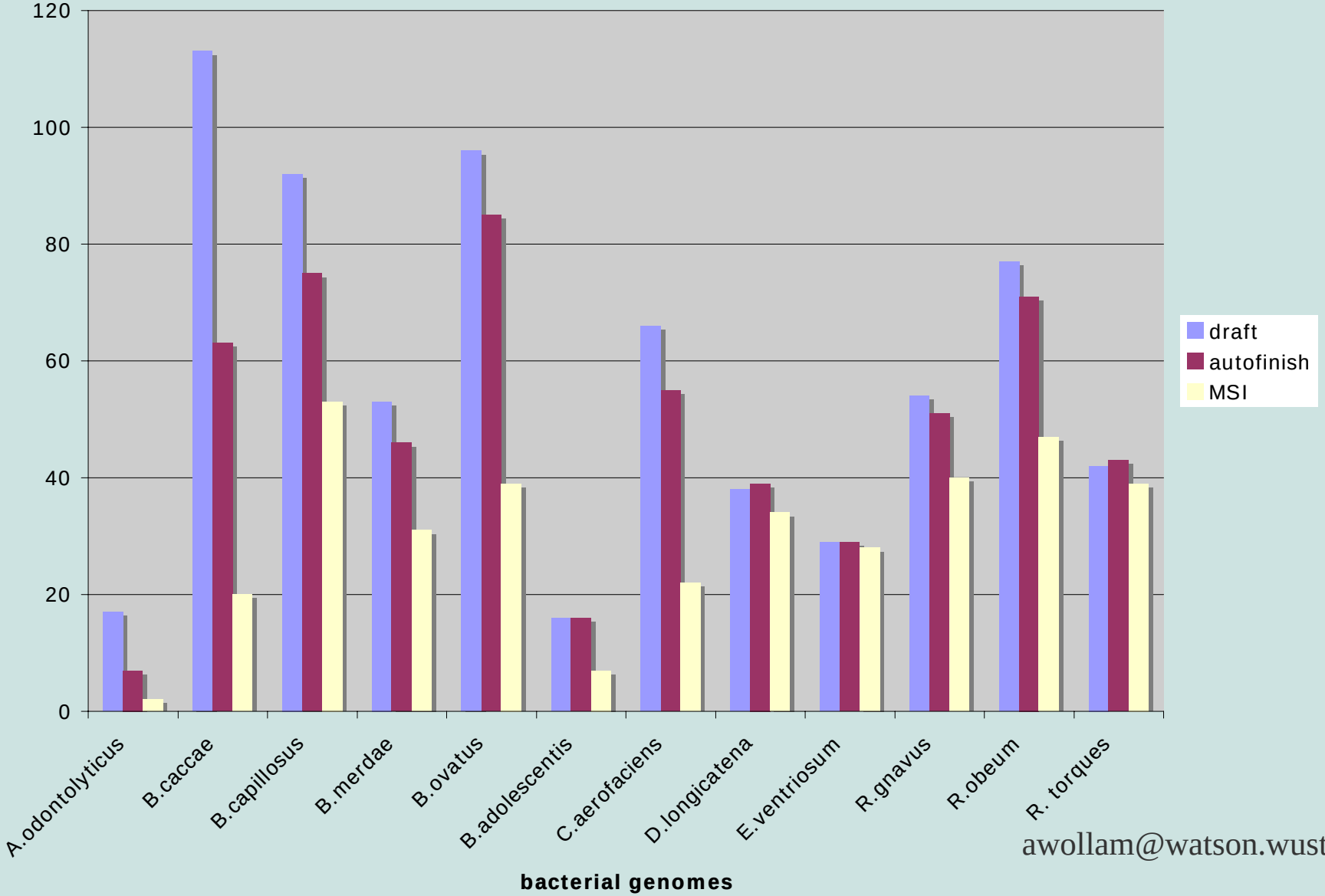
awollam@watson.wustl.edu

Comparison of assemblies at varying stages of improvement



awollam@watson.wustl.edu

Comparison of assemblies at varying stages of improvement (total major contigs)



awollam@watson.wustl.edu

Summary of comparison between draft and MSI assemblies

- N50 contig length in twelve genomes = 1.2 - 7.5 folds increase
- average N50 contig length increase for twelve genomes = 3 folds
- Percentage decrease in total major contigs (more than 2 kb in size) = 3.5 -88.2%
- average decrease in contigs for twelve genomes = 44 %

awollam@watson.wustl.edu

finishing process versus whole genome MSI process

- more time needed as hard to finish regions can linger depending on the difficulty of genomes (5 Mb bacterial genome may take ~6 plus months)
- less high throughput (capacity of 2 Mb/month per finisher in BAC finishing)
- cost includes reagents, special techniques and more human resources.

- requires less time (5 Mb bacterial genome takes ~10 working days). Difficult regions are sorted as much as possible but not completely resolved.
- more high throughput (capacity of 20-25 Mb/month per finisher)
- approximately less than 10% of finishing cost (no additional reagent cost, less personnel)

awollam@watson.wustl.edu

annotation notes

- We will be looking at number of gene discovery, and annotation yield (confidence in functional assignments; number of functionally annotated versus conserved hypothetical and hypothetical proteins) as part of the parameters to be used to judge advantages of MSI.
- preliminary data on BC genome indicates MSI process eliminates false/split gene hits compared to draft assembly by 81.5%. (i.e. split genes were replaced by whole functional genes)
- 61% of draft genes are different than MSI genes, most of which go from smaller to larger sizes.
- detailed analysis of annotated regions between draft versus MSI assemblies are underway. Phylogenomic analysis of 1000 ORFS are currently in process.

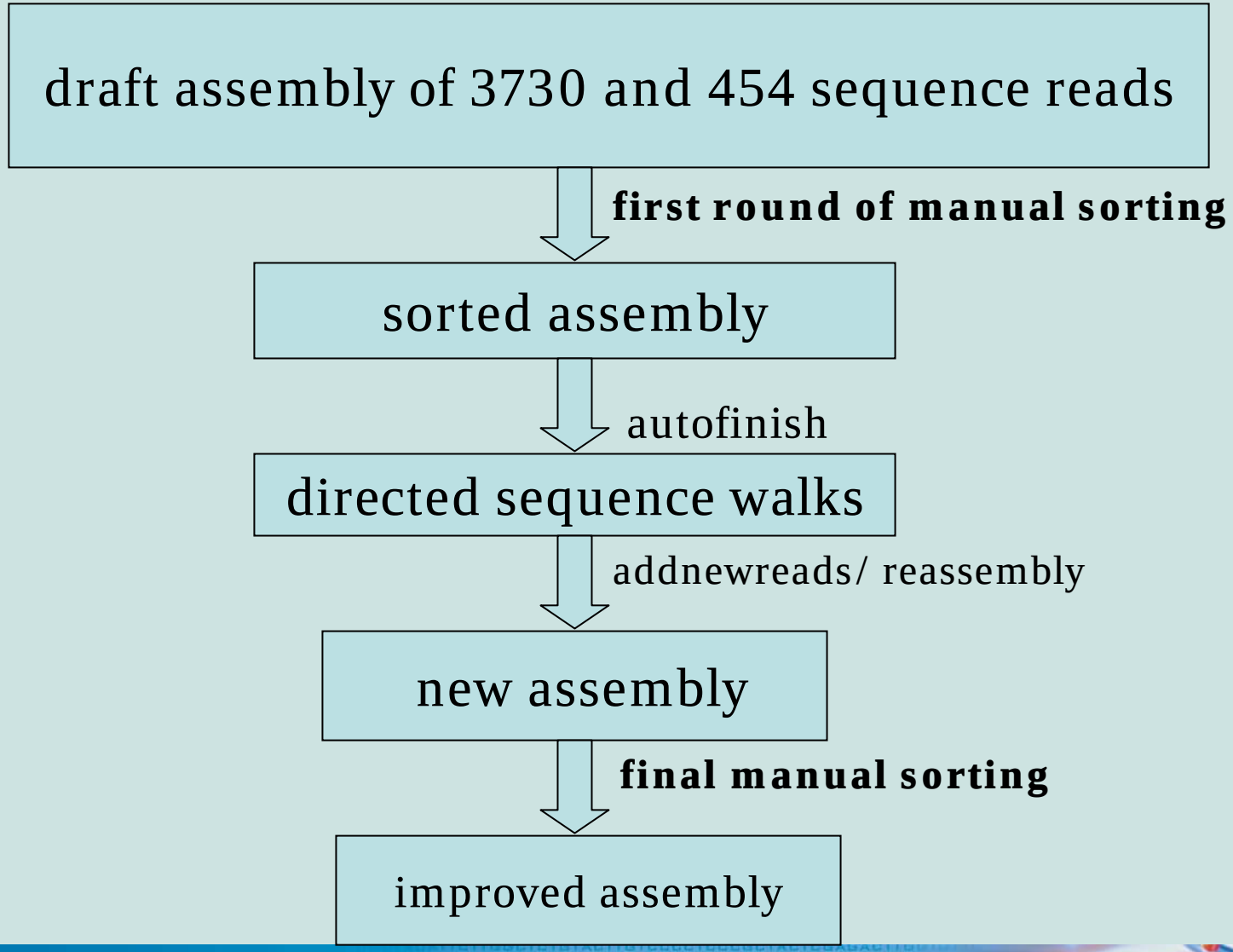
awollam@watson.wustl.edu

What's next?

- Modify MSI pipeline.
- Apply MSI process to large genomes for improving regions of interest.
- Auto-joining tools: development of software for determining possible joins between contigs, and completing joins.

awollam@watson.wustl.edu

modified MSI pipeline



summary

- Manual sequence improvement approach offers a better alternative to draft assemblies.
- To date, twelve bacterial genomes have been manually improved in significantly less time, cost and effort than those of finishing process.
- The result shows a sharp increase in N50 contig length, increased contiguity, less misassembled regions and better annotation when compared to the draft assembly.
- It is a better suited approach when dealing with large number of genomes

awollam@watson.wustl.edu

Whole genome finishing and manual improvement team

Bob Fulton

Aye Wollam

Neha Shah

Tom Wang

Matt Cordes

Kelsi Rotter (part time)

Jennifer Hodges (part time)

awollam@watson.wustl.edu