DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Glomus intraradices

Initial Whole-Genome Shotgun Sequencing and Assembly Results

Harris Shapiro

DOE Joint Genome Institute

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory

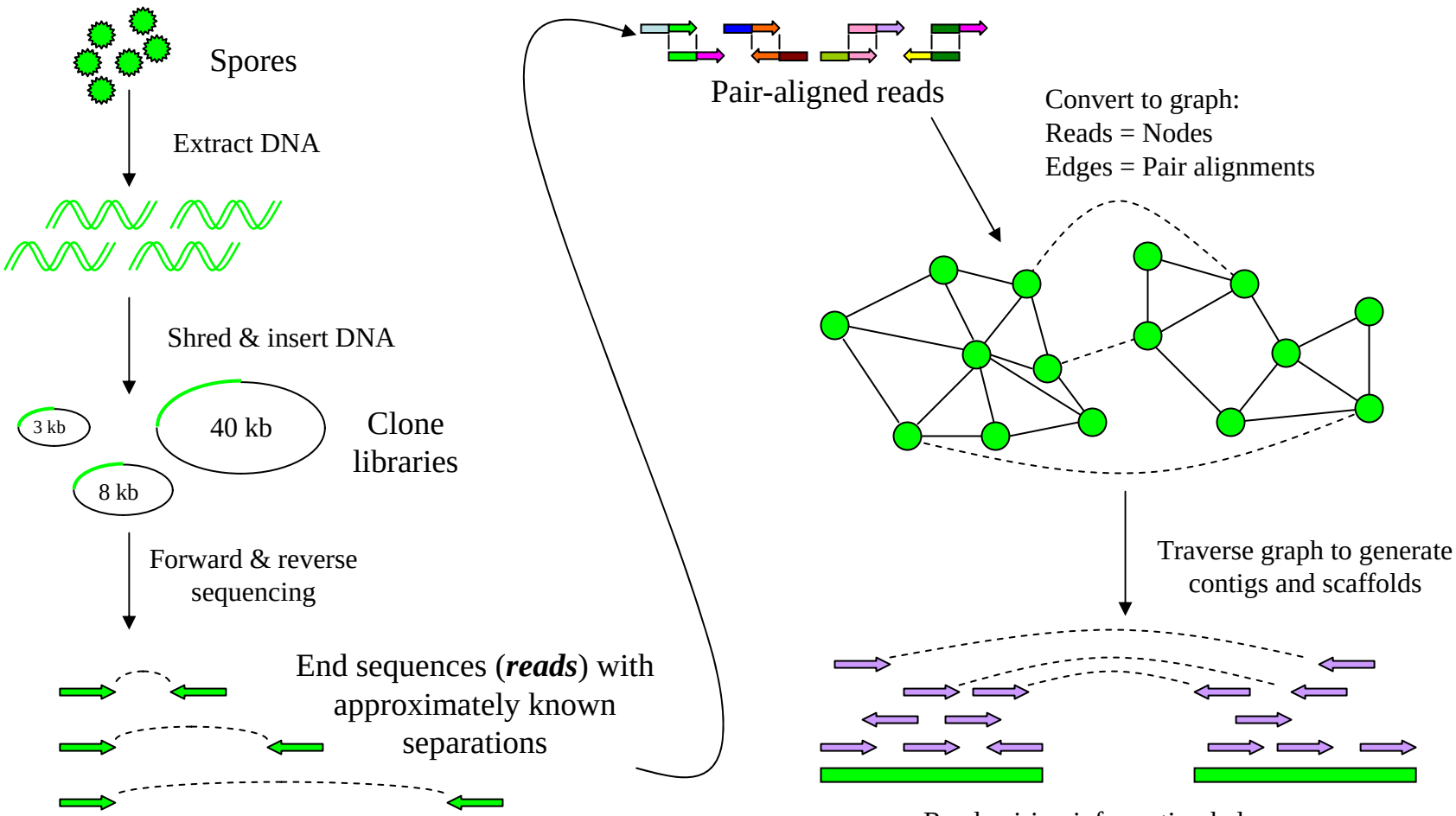
under contract No. W-7405-ENG-36.

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Overview

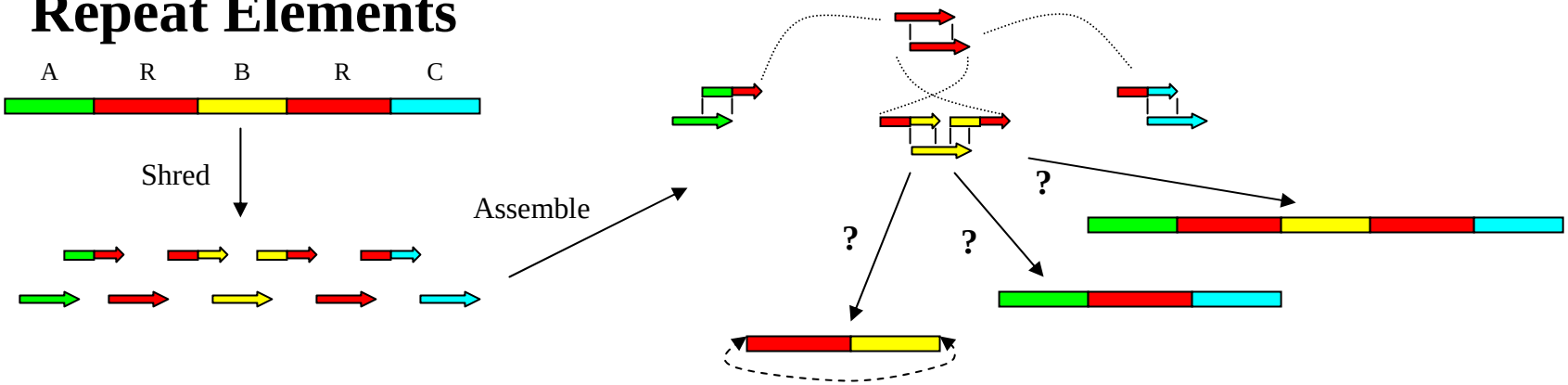
- **Review of the Whole-Genome Shotgun (WGS) assembly procedure**
- **WGS library Quality Control (QC)**
- ***G.intraradices* test assembly**
- **Polymorphism estimates**
- **Data set anomalies**
- **Where to go from here**

Outline of the Assembly Process: JAZZ, the JGI In-House Assembler

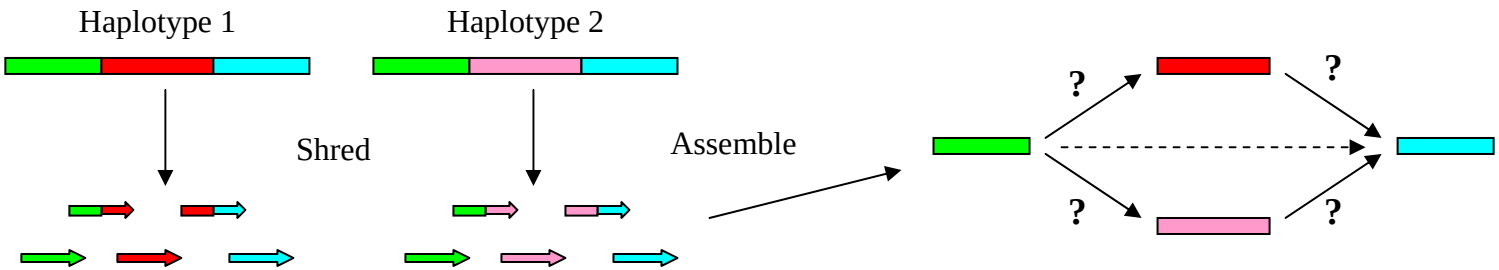


So What Could Go Wrong?

- **Repeat Elements**



- **Polymorphism**



Stricter assembly parameters can distinguish (some) repeats, but make it more likely that haplotypes will assembly separately.

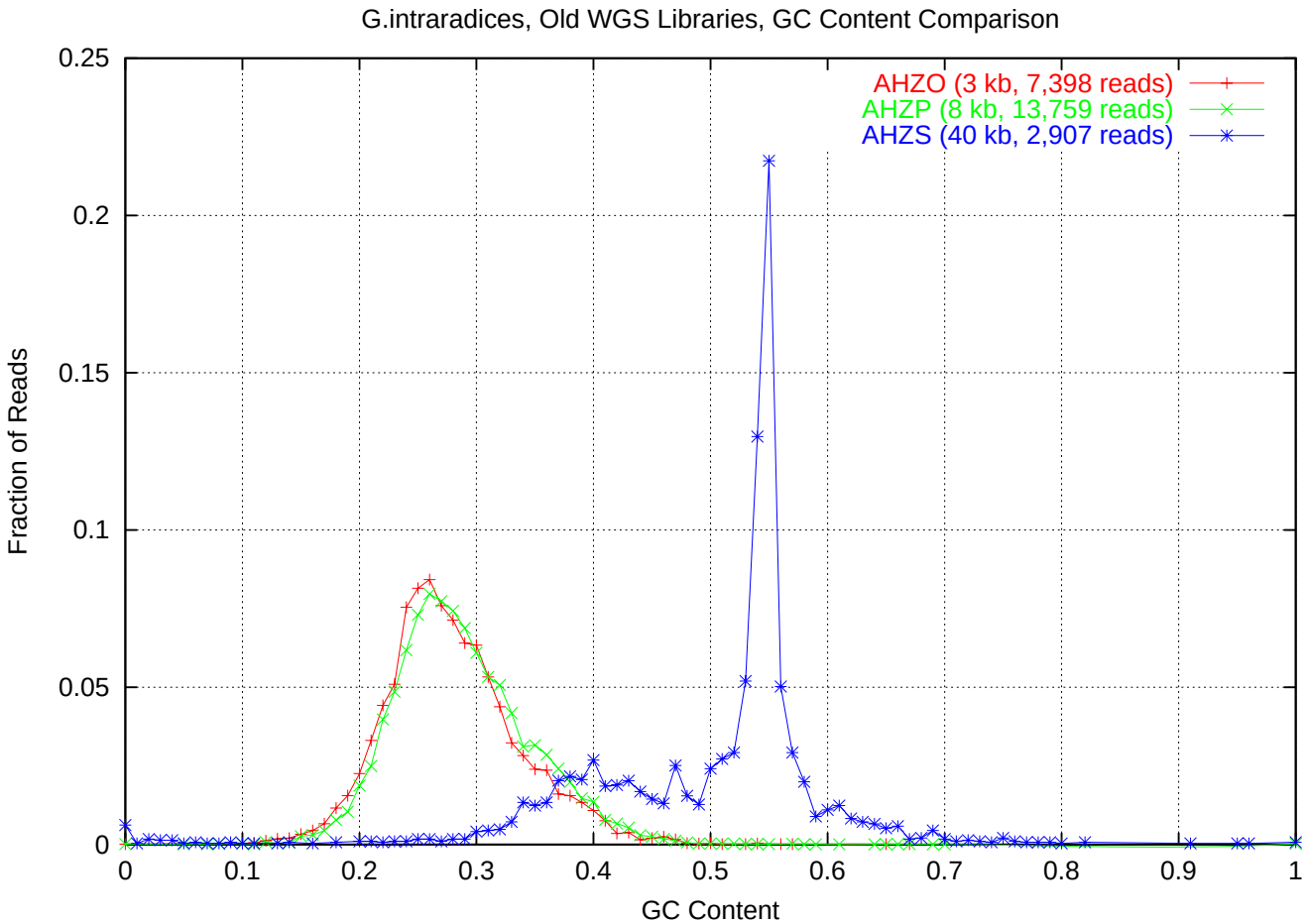
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

WGS Library Quality Control

- **Identification of insertless clones**
- **Trimming of vector and low quality sequence**
- **Examination of the GC content distribution**
- **BLAST analysis**

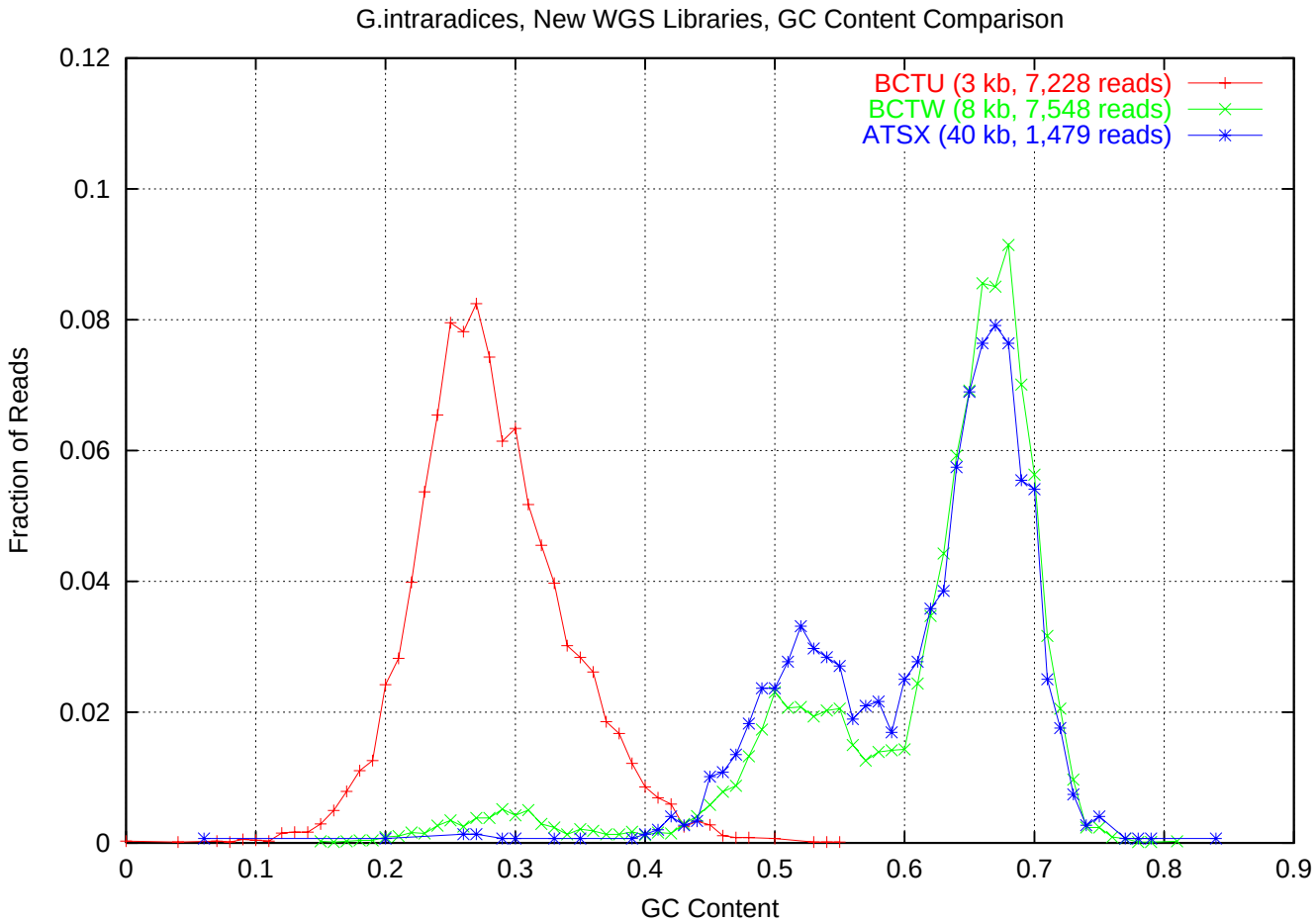
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

GC Content Comparison: Initial WGS Libraries



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

GC Content Comparison: Second WGS Libraries



JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Test Assembly Specification

- **All 6 WGS libraries were included: ~28.7 MB of trimmed sequence, of which perhaps 4 – 7 MB consisted of various contaminants**
- **Genome size estimate: ~15 MB (Hijri & Sanders, 2004)**
- **Estimated depth: 1 – 1.5x; set to 1.0 for the assembly**
- **Repeats: Maximum copy number of 5 times the estimated depth (5) before they can't be used to seed an alignment**
- **Polymorphism/repeat divergence: used the default value of ~3% different for a "neutral" alignment**

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Summary of Test Assembly Results

- **615 scaffolds, containing 704 KB of scaffold sequence (582 KB contig sequence = 17.3% gap).**
- **After removing short and redundant scaffolds, 165 were left, with 318 KB of scaffold sequence (197 KB contig sequence = 38.1% gap)**
- **Filtered scaffold set was easily partitioned:**
  - **Prokaryotic contaminant: GC content > 0.60**
  - **Cloning artifacts: GC content between 0.50 and 0.60 (with one exception)**
  - **Potential *G.intraradices* scaffolds: GC content < 0.50 (with one exception)**
- **EST alignments supported the identification of the low-GC scaffolds**

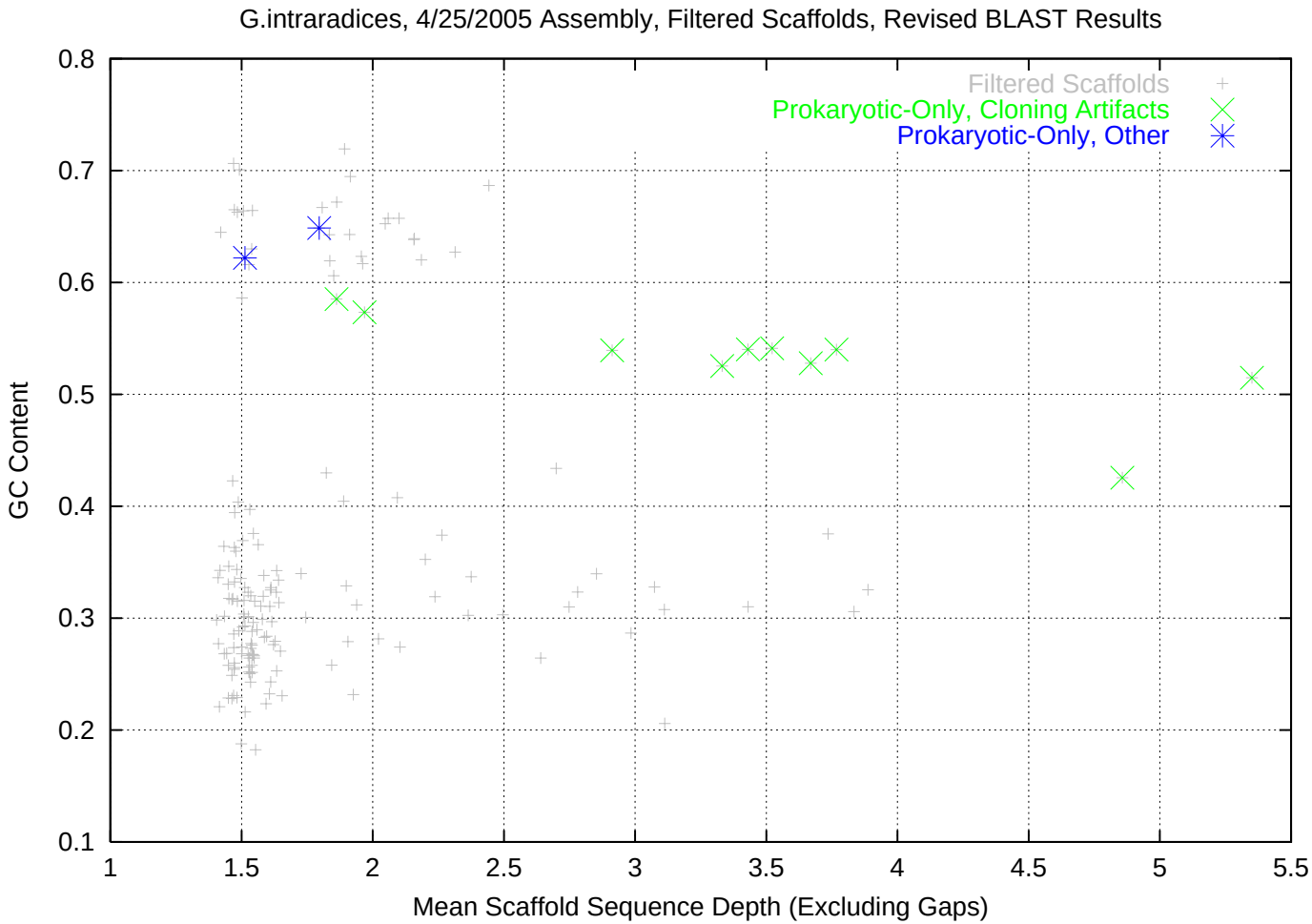
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Test Assembly Statistics

Scaffold Set	Number of Scaffolds	Scaffold Sequence Total (KB)	Contig Sequence Total (KB)
Potential *G.intraradices*	126	222.5	147.1
Cloning Artifact	10	14.3	14.3
Prokaryotic Contamination	29	81.6	35.2

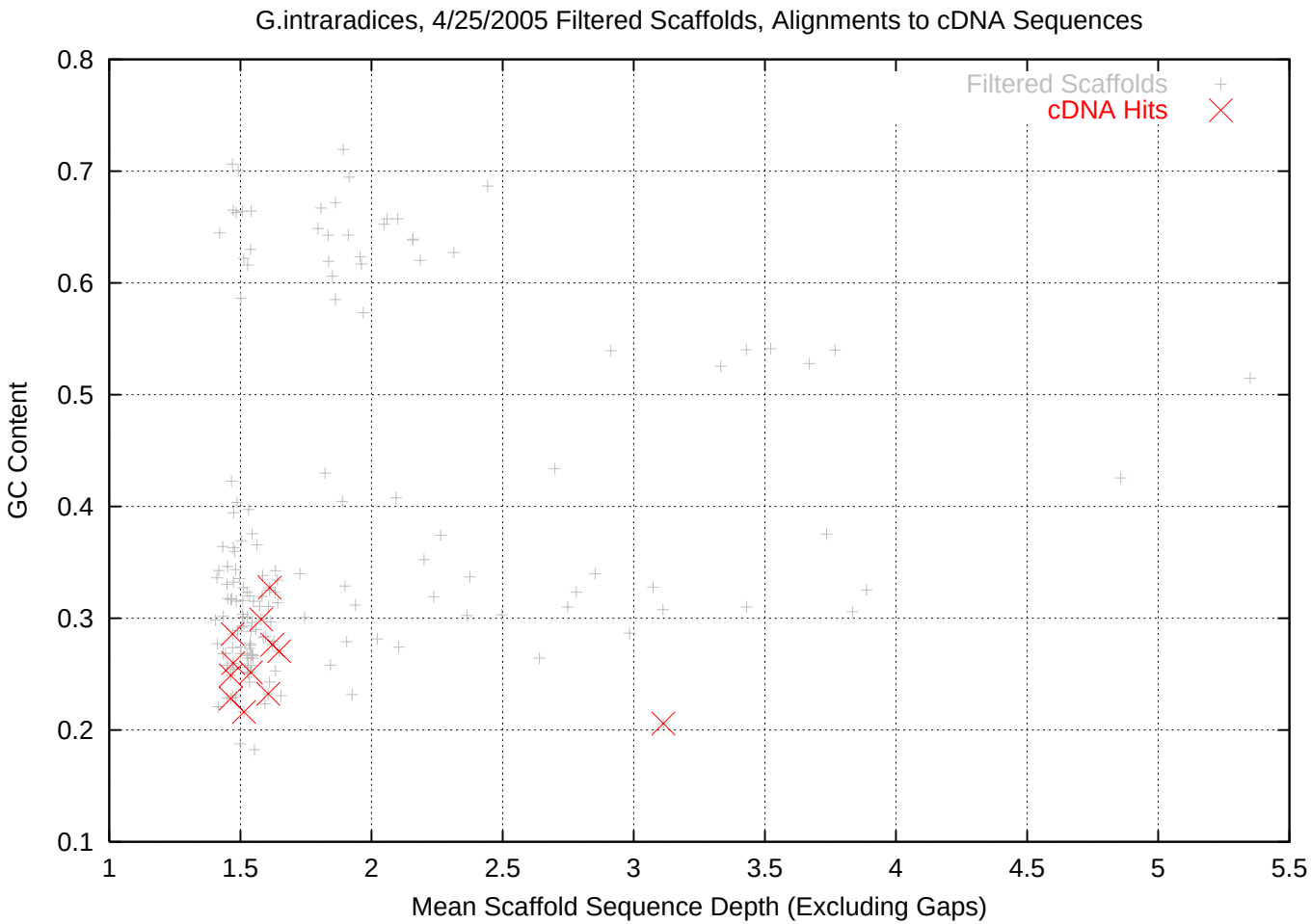
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Test Assembly: BLAST Analysis



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Test Assembly: EST Analysis



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Polymorphism Estimation Procedure

- **Generation of reference sequence: draft & finished subcloned fosmids**
- **Alignment of WGS & EST reads to the reference sequence**
- **Identification of "good" alignments**
  - **Reads with unique alignments to the reference fosmids**
  - **Greater than 97% of the read covered by the alignment**
- **Polymorphism calculations**
  - **% ID of the good alignments (Number of matches/alignment footprint)**
  - **Mismatch % of the good alignments (Number of mismatches/read length)**

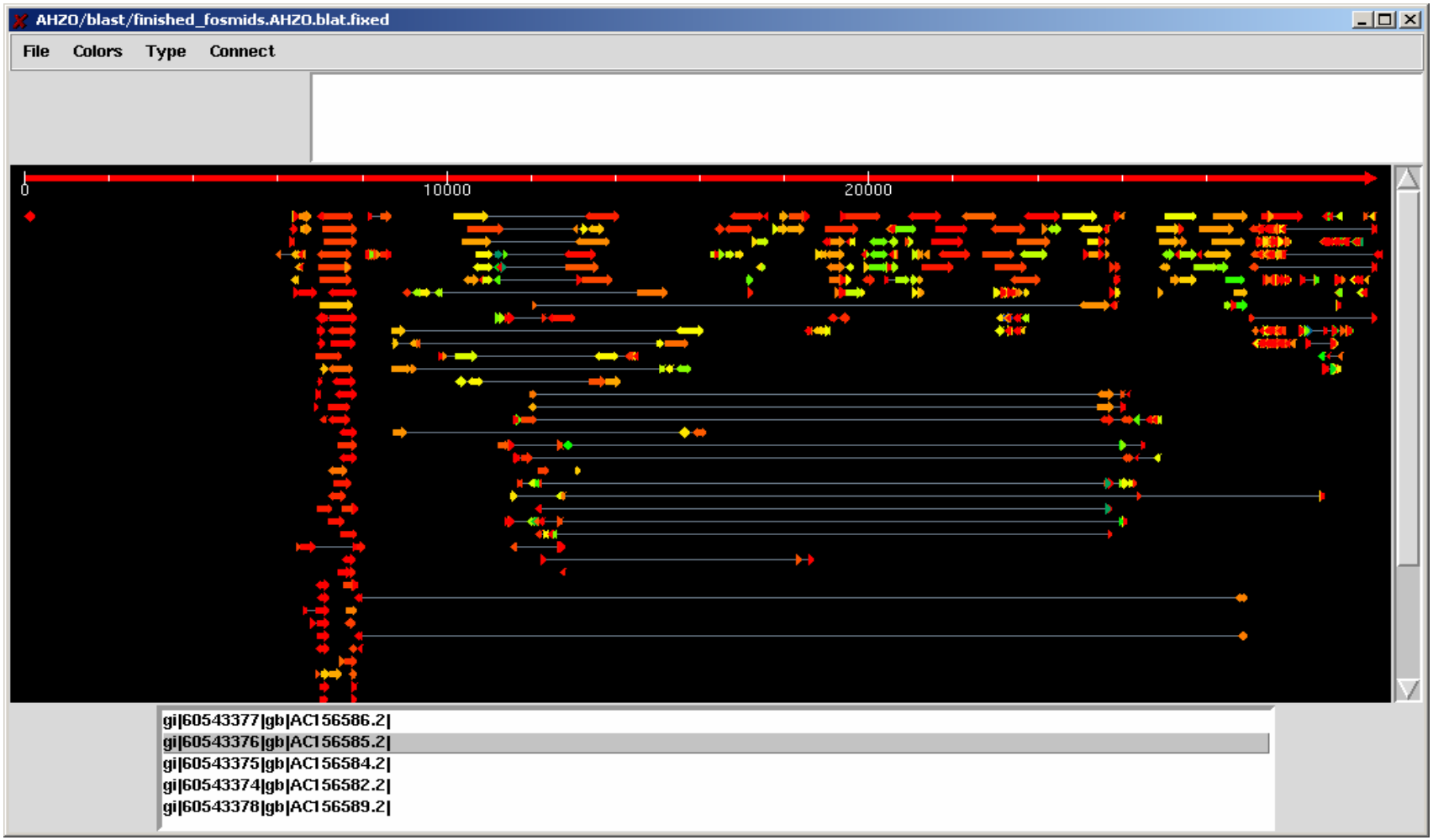
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Subcloned Fosmid Creation

- **First batch: 15 randomly selected fosmids from the AHZS library. Consisted almost entirely of cloning artifacts.**
- **Second batch: 10 targeted fosmids from the AHZS library. Due to cloning artifacts and sequencing problems, this yielded 5 finished fosmids, containing about 181 KB of reference sequence.**
- **Third batch: 10 targeted fosmids from the ATSX library. All 10 yielded draft sequences.**

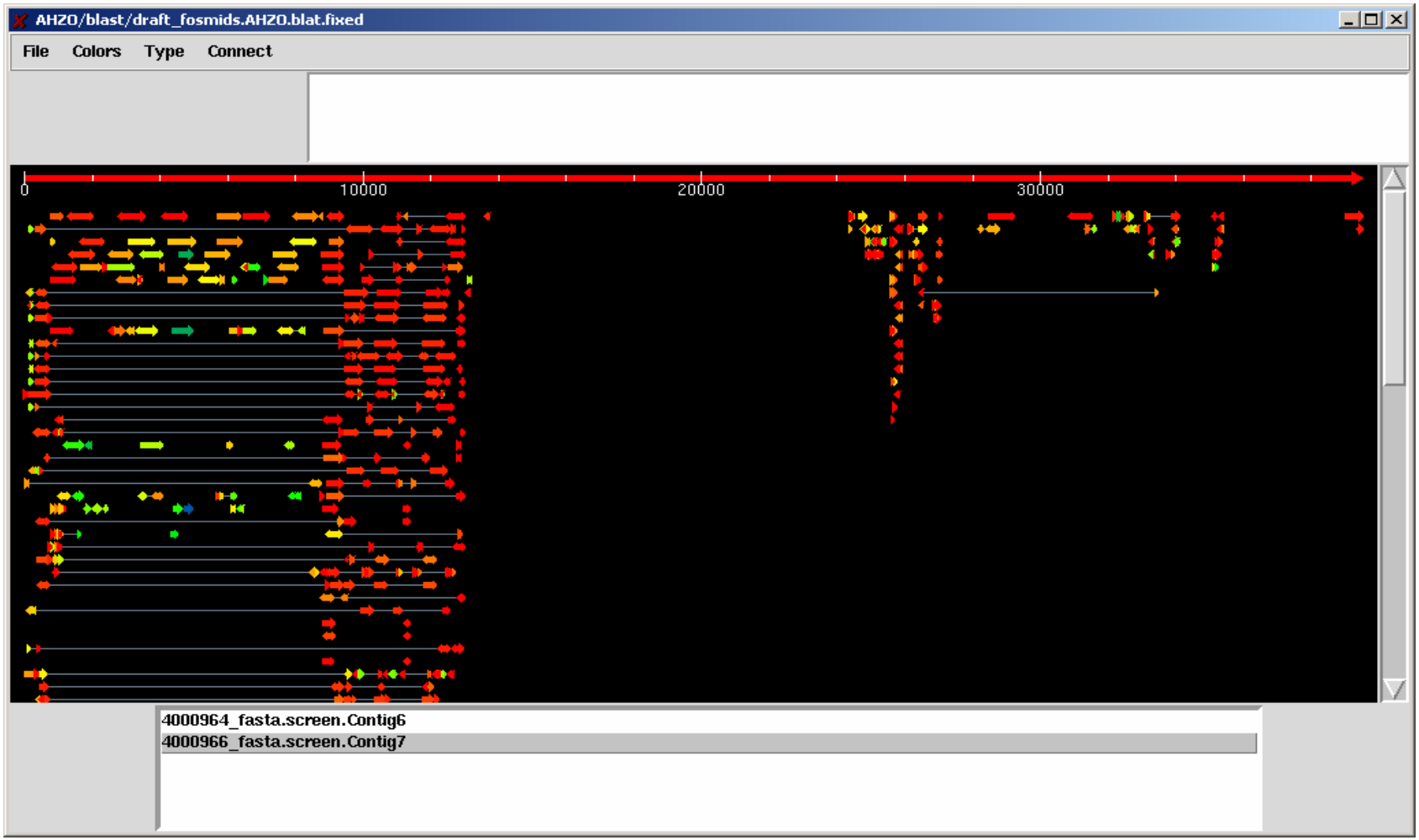
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Sample Finished Fosmid WGS Alignments



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Sample Draft Fosmid WGS Alignments



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Summary of the WGS Polymorphism Results

		Polymorphism Estimate	
Data Set	**Number of Reads**	**% ID Method**	**Mismatch Method**
AHZO	52	5.1% +/- 4.9%	3.5% +/- 2.9%
AHZP	91	3.5% +/- 4.2%	2.6% +/- 2.7%
BCTU	68	3.7% +/- 6.1%	2.3% +/- 2.5%
Overall	**211**	**4.0% +/- 5.1%**	**N/A**

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Details of the WGS Polymorphism Results

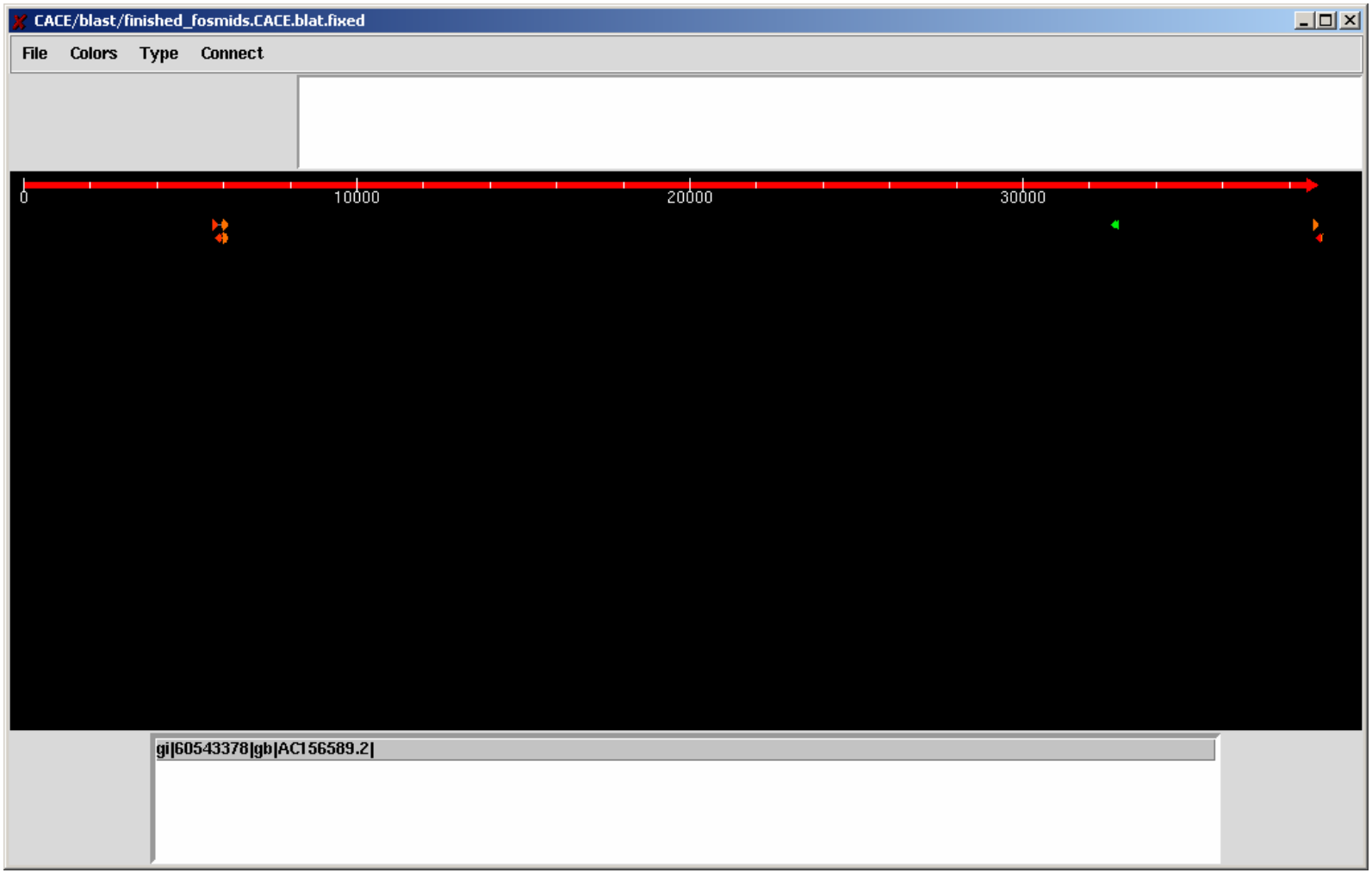
- **With the % ID Method:**
  - **All three WGS libraries had their main polymorphism peak at 0%-1%.**
  - **All three WGS libraries had a secondary peak at 7%-8%.**
  - **AHZO may have had an additional secondary peak at 5%.**
- **With the Mismatch Method:**
  - **All three WGS libraries had their main polymorphism peak at 0%-1%.**
  - **AHZO and BCTU still had secondary peaks at 7% - 8%.**
  - **AHZO still had a secondary peak at 5%.**
  - **AHZP no longer had any secondary peaks, instead having its mismatch rate decline smoothly to a background value.**

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

WGS Data Set Issues

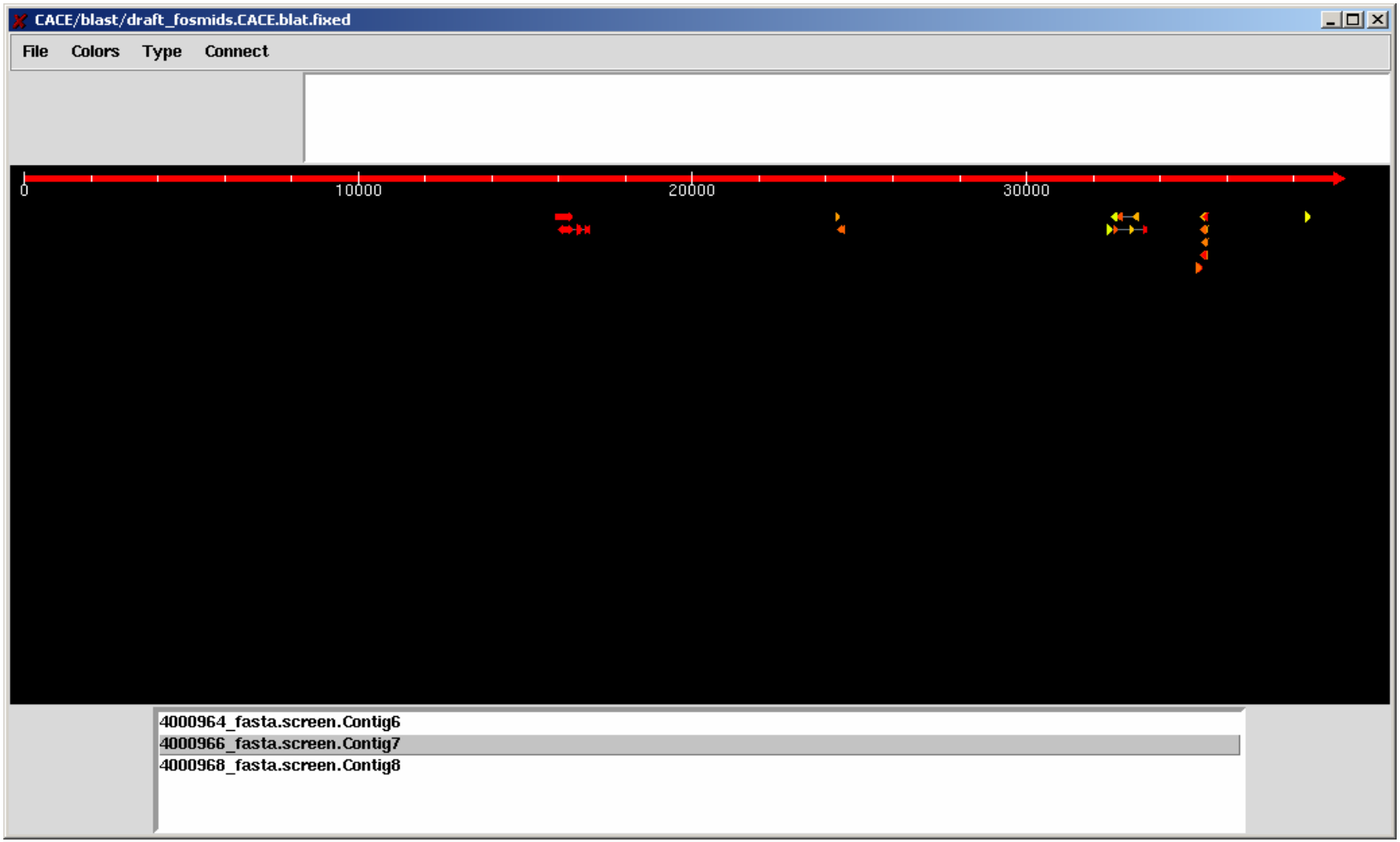
- **Only produced plausible alignments to 2 of the 10 subcloned fosmids in the latest batch**
- **Very non-uniform coverage of some of the draft and finished fosmids**
- **Potential causes:**
  - **Mis-assembled or –finished fosmids**
  - **Chimeric fosmid clones**
  - **Undetected non-Glomus contamination**
  - **Library bias**
  - **Statistical artifact**

Sample Finished Fosmid EST Alignments



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Sample Draft Fosmid EST Alignments



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

EST Polymorphism Results

- **Very few EST sequences aligned to the subcloned fosmid sequences.**
  - **Only 1 of the 5 finished fosmids had any alignments**
  - **Only 2 of the 10 draft fosmids had plausible alignments**
- **The longest alignments all seemed to be missing sections of the ESTs, on the order of hundreds of bases.**
- **Due to the small number of alignments and the anomalies associated with them, it was not possible to estimate the polymorphism rate using this data set.**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

EST Data Set Issues

- **Only produced plausible alignments to 2 of the 10 subcloned fosmids in the latest batch**
- **Very small number of alignments overall**
- **Alignments that were present did not cover large portions of the ESTs**
- **Potential causes:**
  - **Mis-assembled or –finished fosmids**
  - **Chimeric fosmid clones**
  - **Undetected non-Glomus contamination**
  - **Library bias**
  - **Spurious alignments**
  - **The haplotypes actually differ greatly from each other**
  - **Statistical artifact**

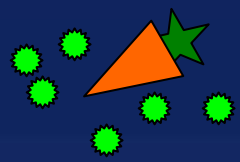
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Where Do We Go From Here?

- **Investigation of the WGS and EST subcloned fosmid anomalies**
- **Creation of a third round of WGS libraries**
- **Further attempts at WGS assembly**
  - **Requires the production of a good set of WGS libraries**
  - **Analysis of the possible repeats, which is complicated by high polymorphism**
  - **Initially strict parameters to force the haplotypes to assemble separately**
  - **Distribution of polymorphic elements might require a relaxed-parameter assembly, to try to force the haplotypes to assemble together**
  - **Pre-screening of repeat elements may allow some sidestepping of the repeat/polymorphism trade-off**

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Acknowledgements

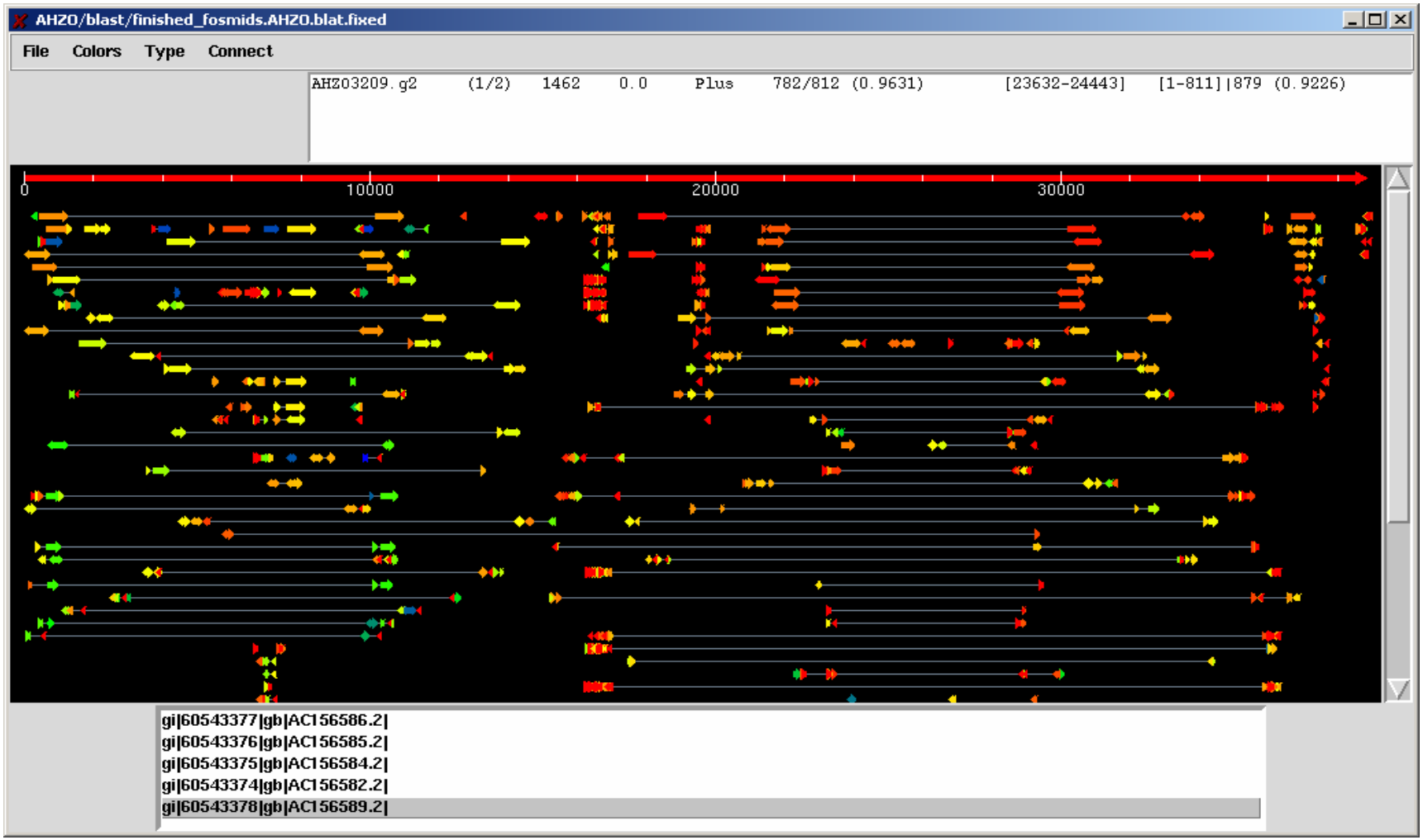


- **Joint Genome Institute**
  - **Library Creation: Eileen Dalin, Chris Detter, Paul Richardson**
  - **Production Sequencing: Tijana Glavina, Miranda Harmon-Smith, Susan Lucas**
  - **Genome Assembly: Nik Putnam, Jarrod Chapman, Isaac Ho, Dan Rokhsar**
  - **ESTs: Erika Lindquist, Peter Brokstein**
  - **Subcloned Fosmids: Jane Grimwood**
- **Peter Lammers & his group**
- **Ian Sanders & his group**

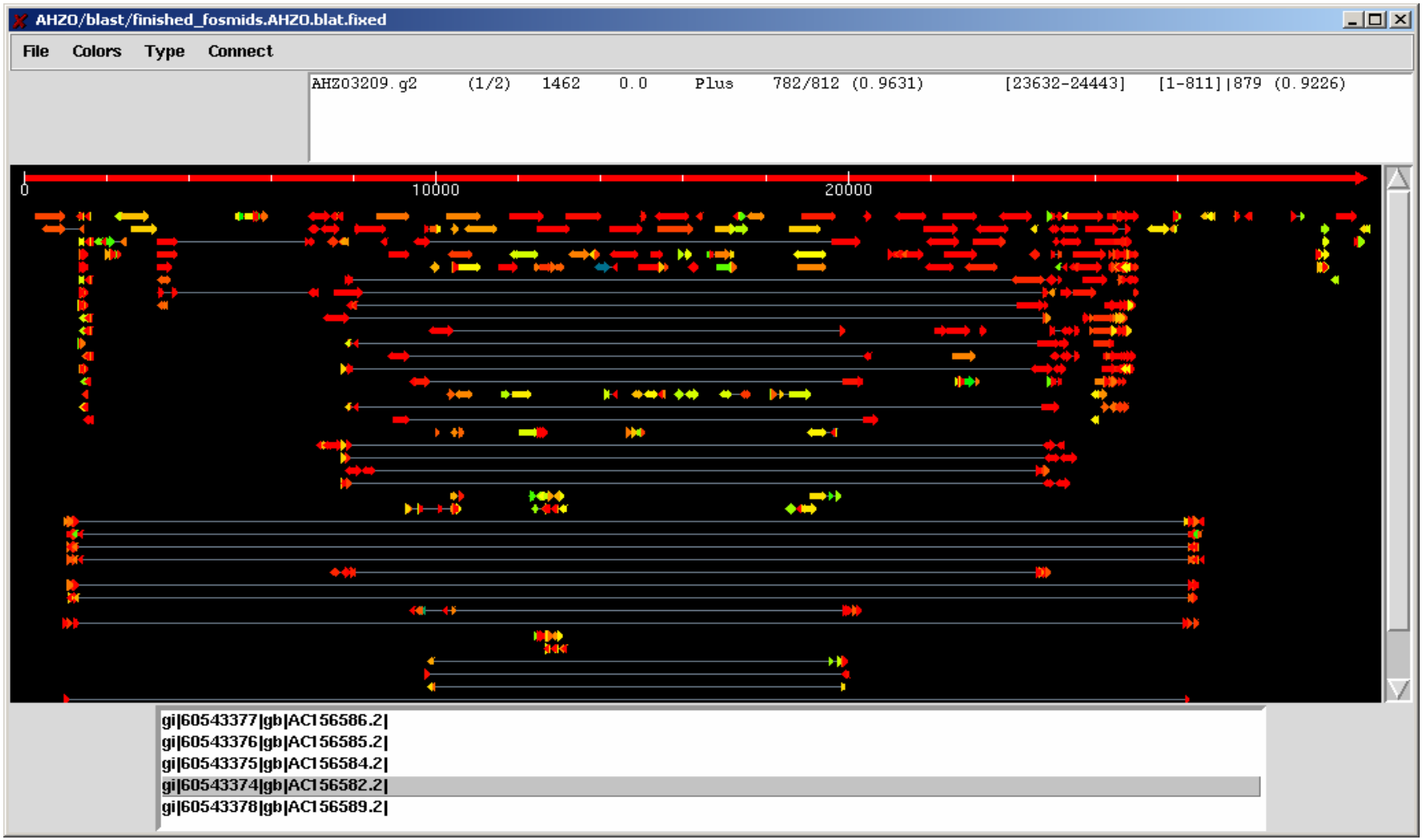
This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

AHZO: ATYW (Finished Fosmid) Alignments

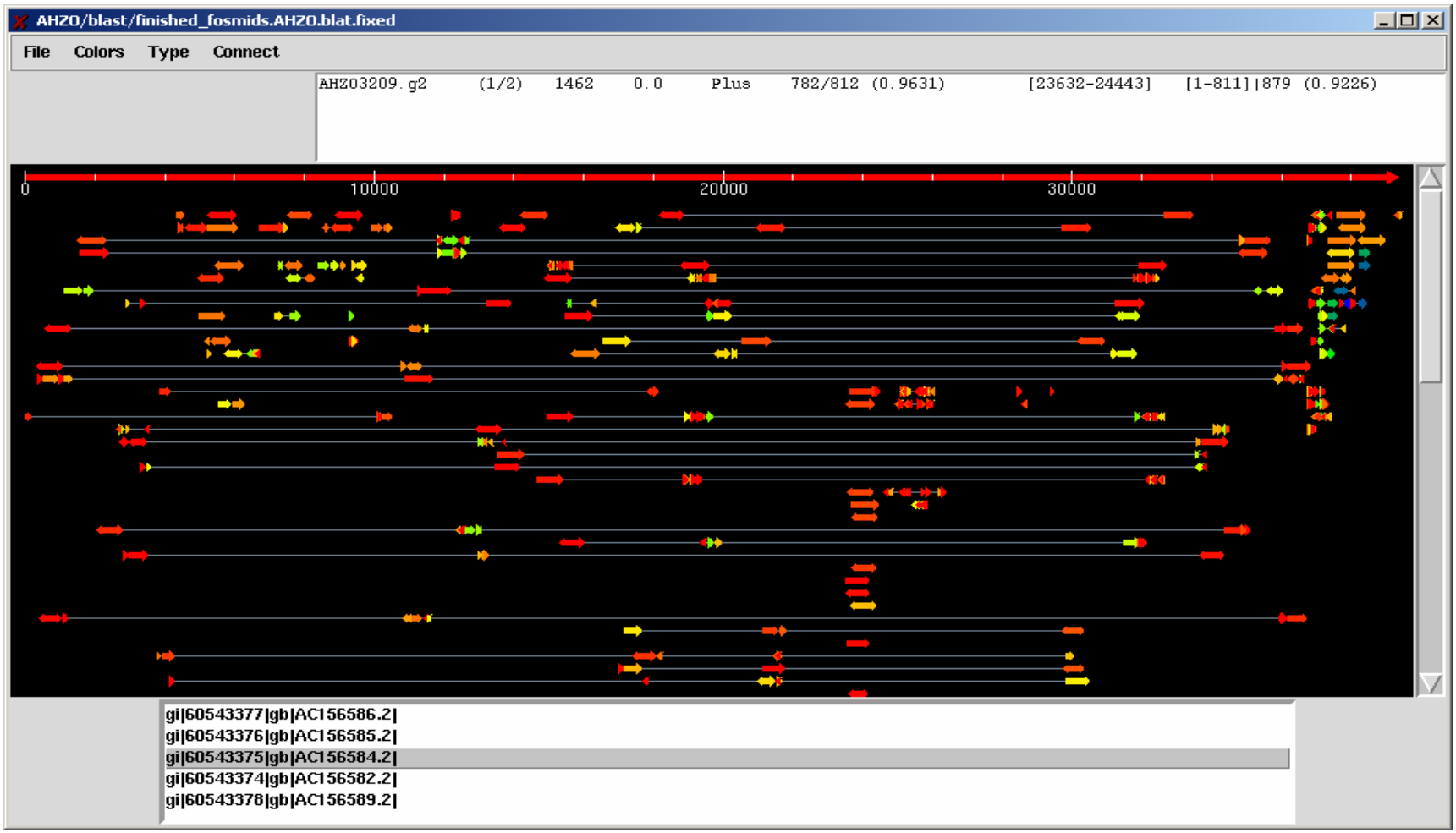


AHZO: ATYY (Finished Fosmid) Alignments



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

AHZO: ATYZ (Finished Fosmid) Alignments



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

AHZO: ATZC (Finished Fosmid) Alignments