DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Glomus intraradices

Status of the Genome Project

Harris Shapiro

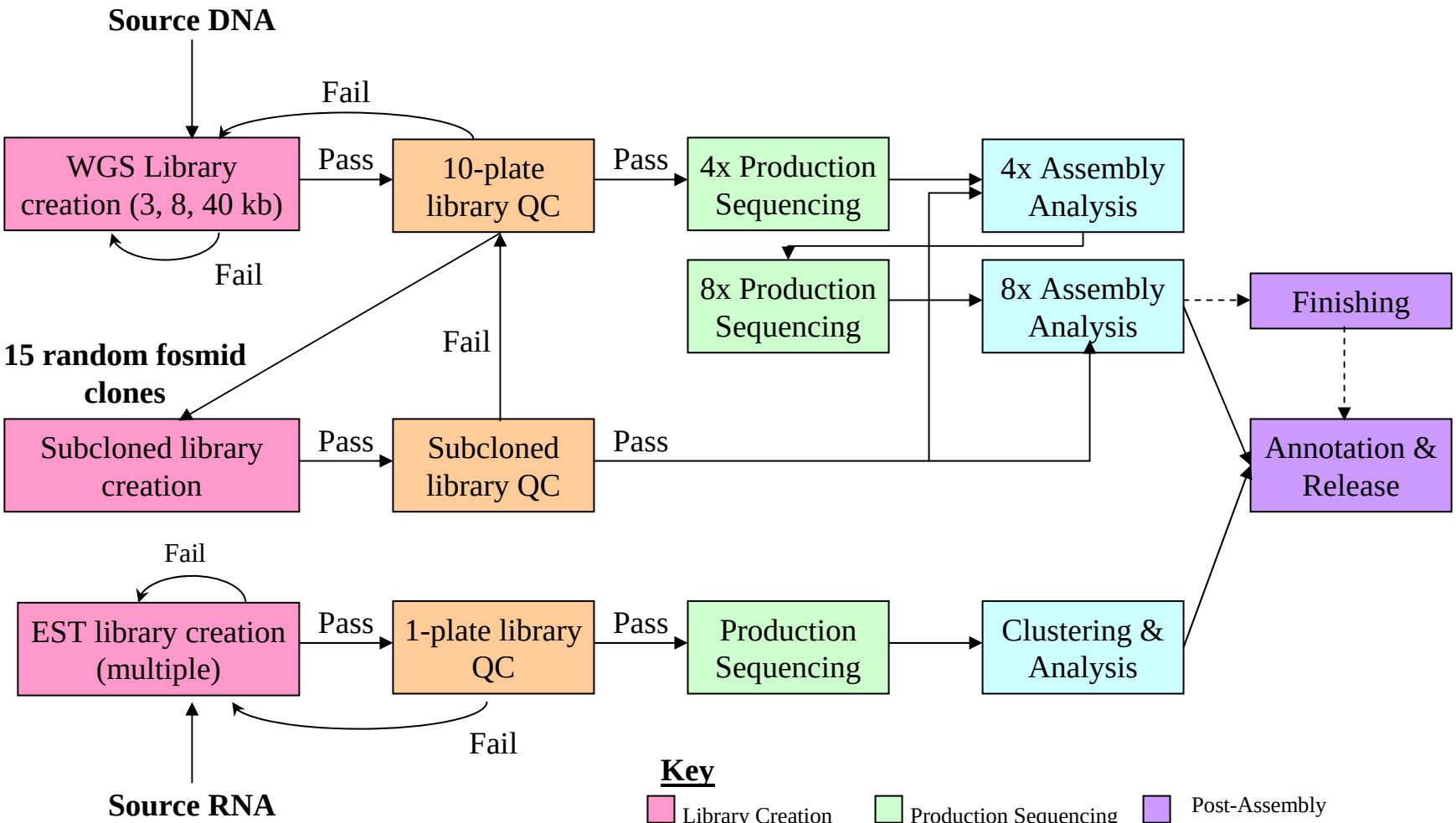
DOE Joint Genome Institute

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Overview

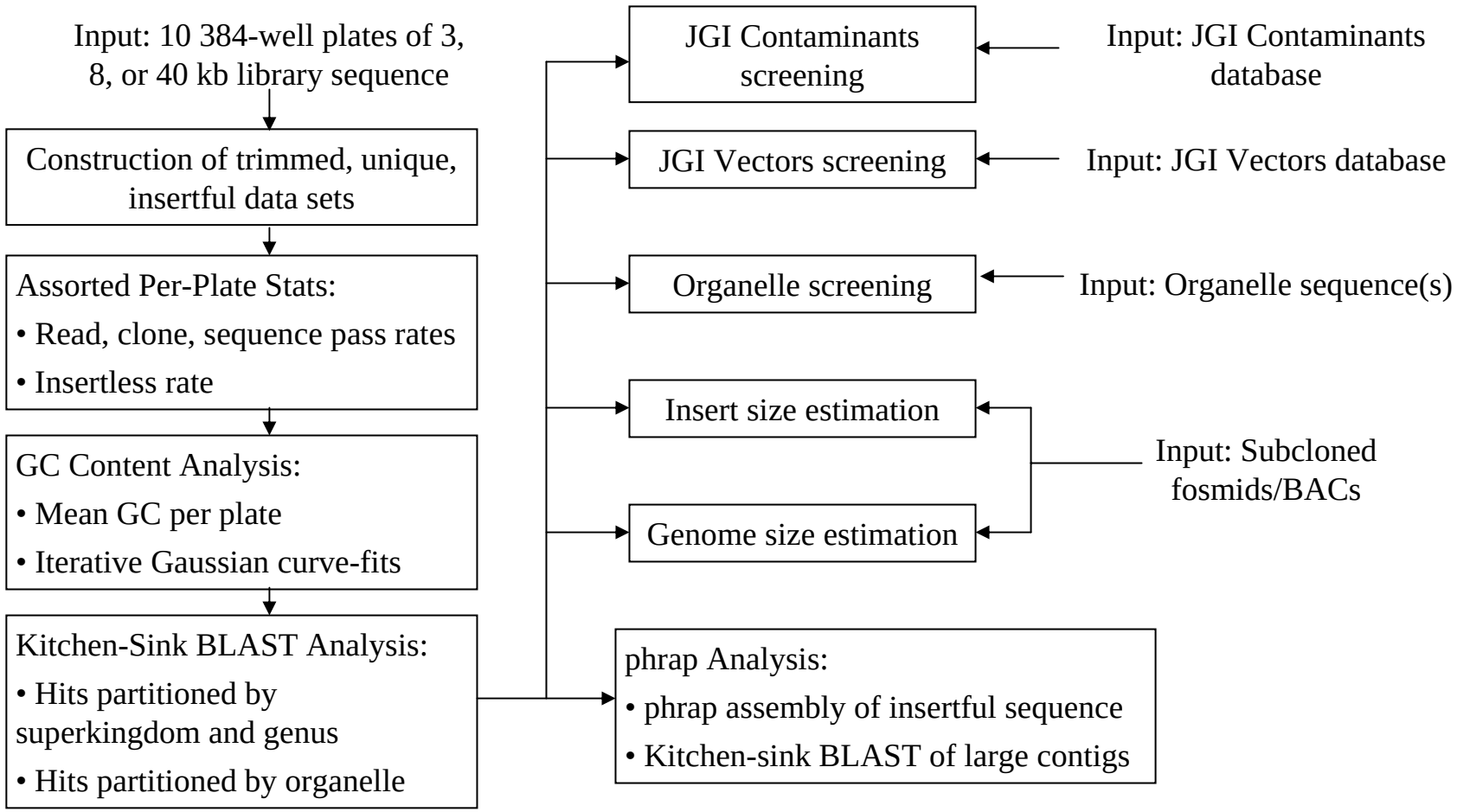
- **The standard JGI eukaryotic WGS sequencing project**
- **Data set inventory & updated QC results**
- **Depth analysis & assembly attempts**
- **What's going on?**
  - **Larger physical genome size?**
  - **Cloning bias?**
  - **Undetected contamination?**
  - **Polymorphism?**
- **Where could one go from here?**

The Standard WGS Sequencing Project



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

The Standard WGS Library QC Procedure



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

JGI WGS Data Set Inventory

Library	Insert Size	Insertful Reads	Insertful Sequence	QC Status
AHZO	3 KB	7,358	5.68 MB	Marginal Pass
BCTU	3 KB	7,012	5.28 MB	Pass
BNTH	3 KB	1,436	1.08 MB	Pass
BTPF	3 KB	4,569	2.67 MB	Marginal Pass
BWUB	4 KB	5,073	3.68 MB	Marginal Pass
AHZP	8 KB	13,519	9.82 MB	Marginal Pass
BCTW	**8 KB**	**7,467**	**5.08 MB**	**Fail**
BFTX	8 KB	7,291	5.71 MB	Marginal Pass
BTPG	8 KB	4,002	2.82 MB	Marginal Pass
AHZS	**40 KB**	**2,377**	**1.60 MB**	**Fail**
ATSX	**40 KB**	**1,395**	**0.90 MB**	**Fail**
BFTY	**40 KB**	**1,230**	**0.66 MB**	**Fail**
BTPH	40 KB	15,039	8.60 MB	Pass
FBHW	454 (GS-20)	338,216	35.98 MB	Pass

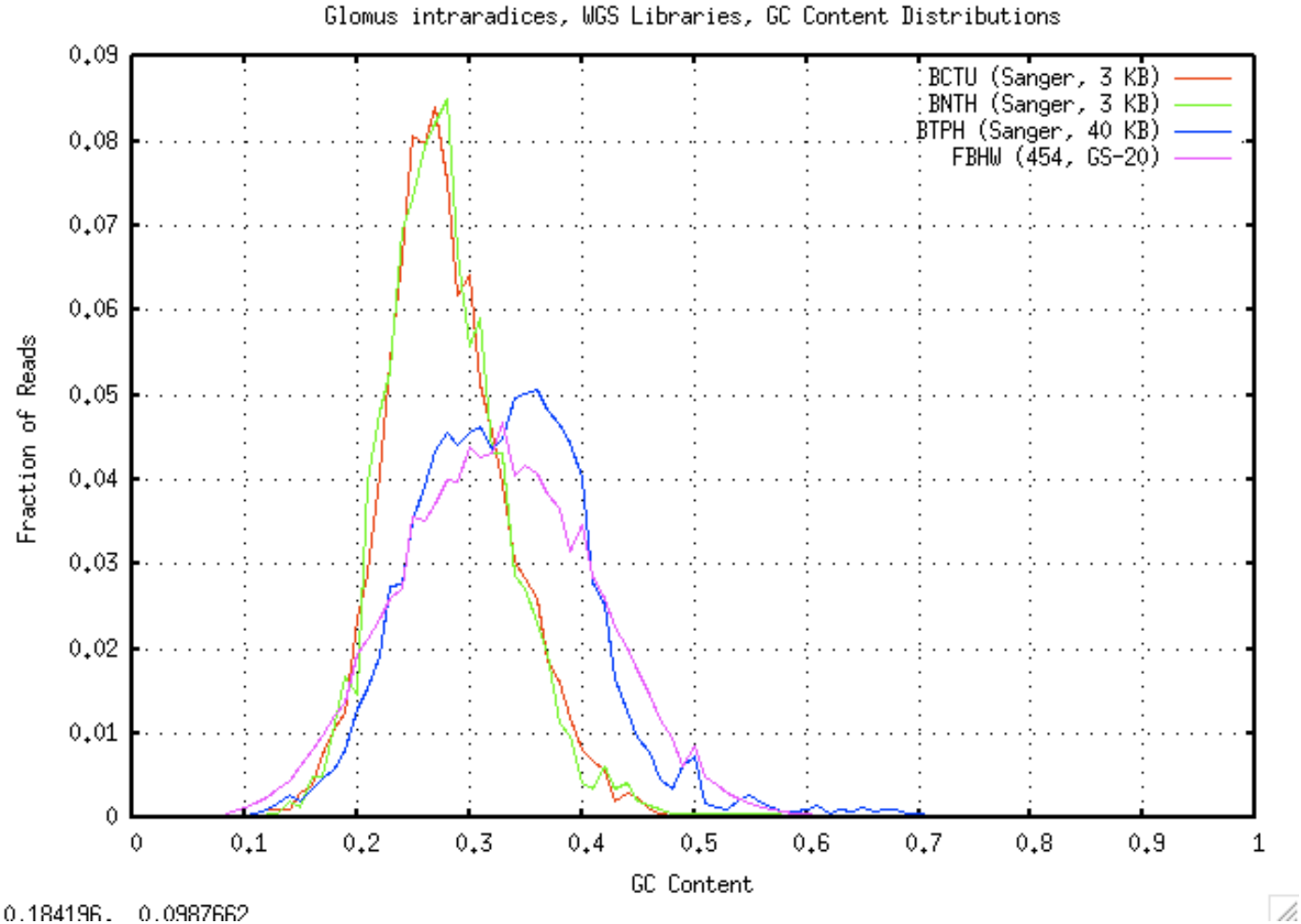
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

JGI EST Data Set Inventory

Library	Good Reads	Good Sequence	QC Status
CACE	8,741	4.64 MB	Pass
CCHU	34,602	21.93 MB	Pass

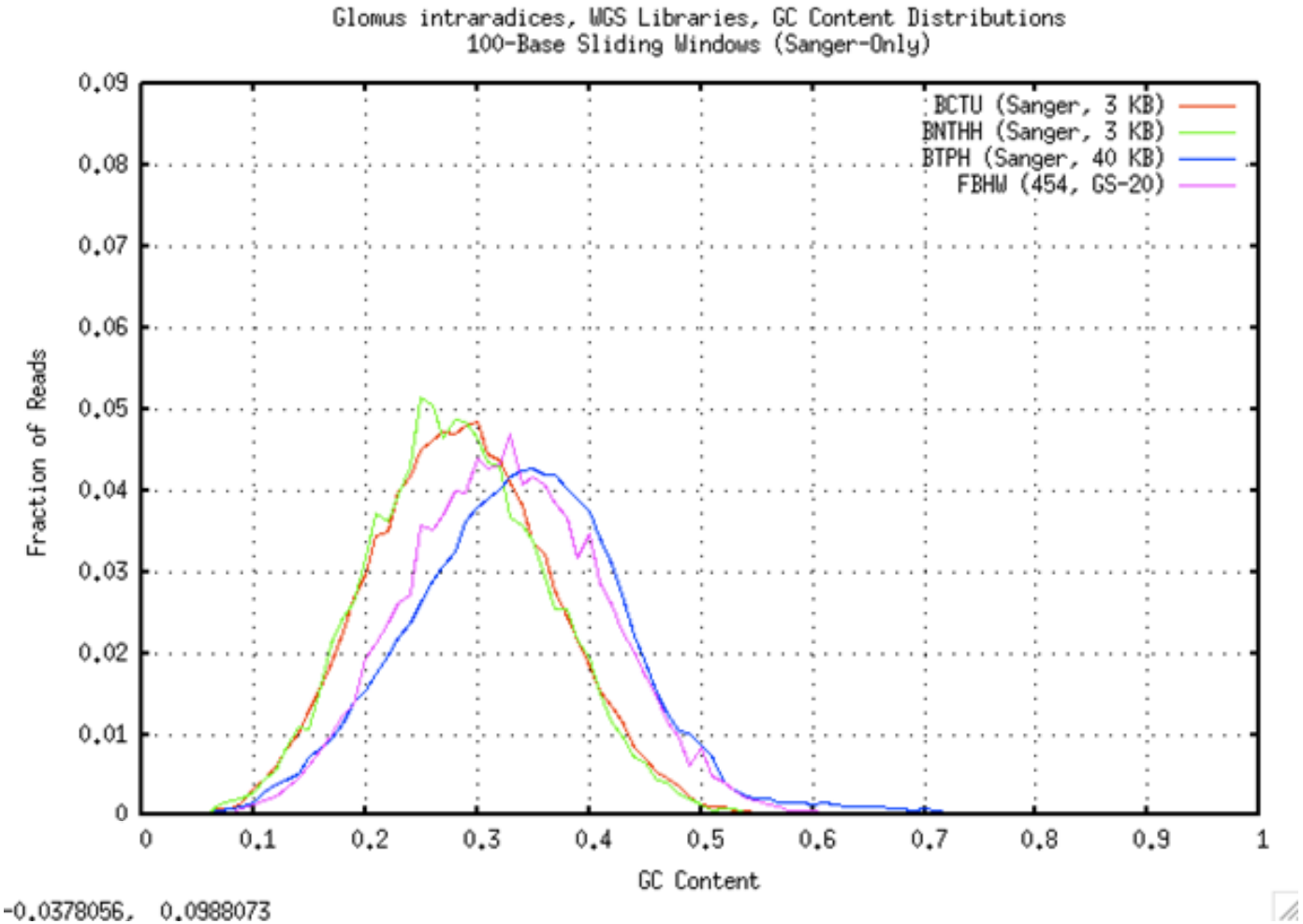
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

GC Content Distributions



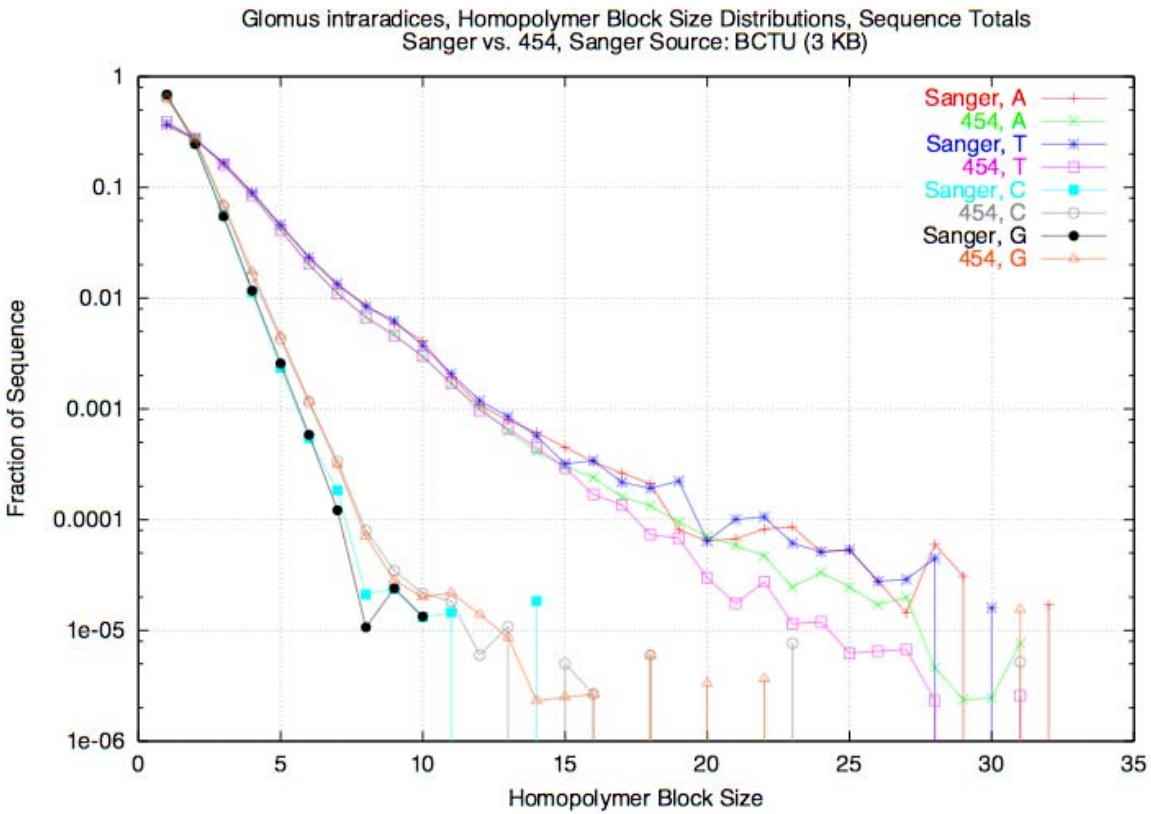
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Windowed GC Content Distributions



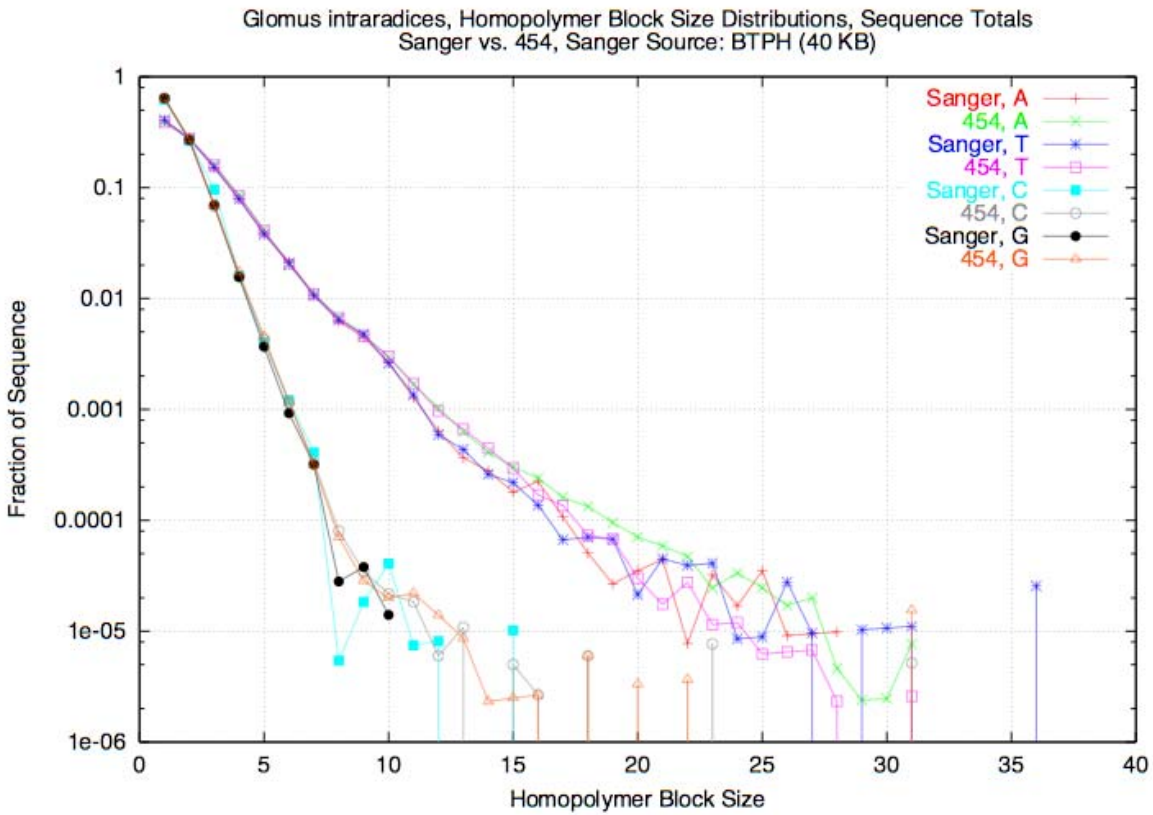
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Homopolymer Block Analysis (1)



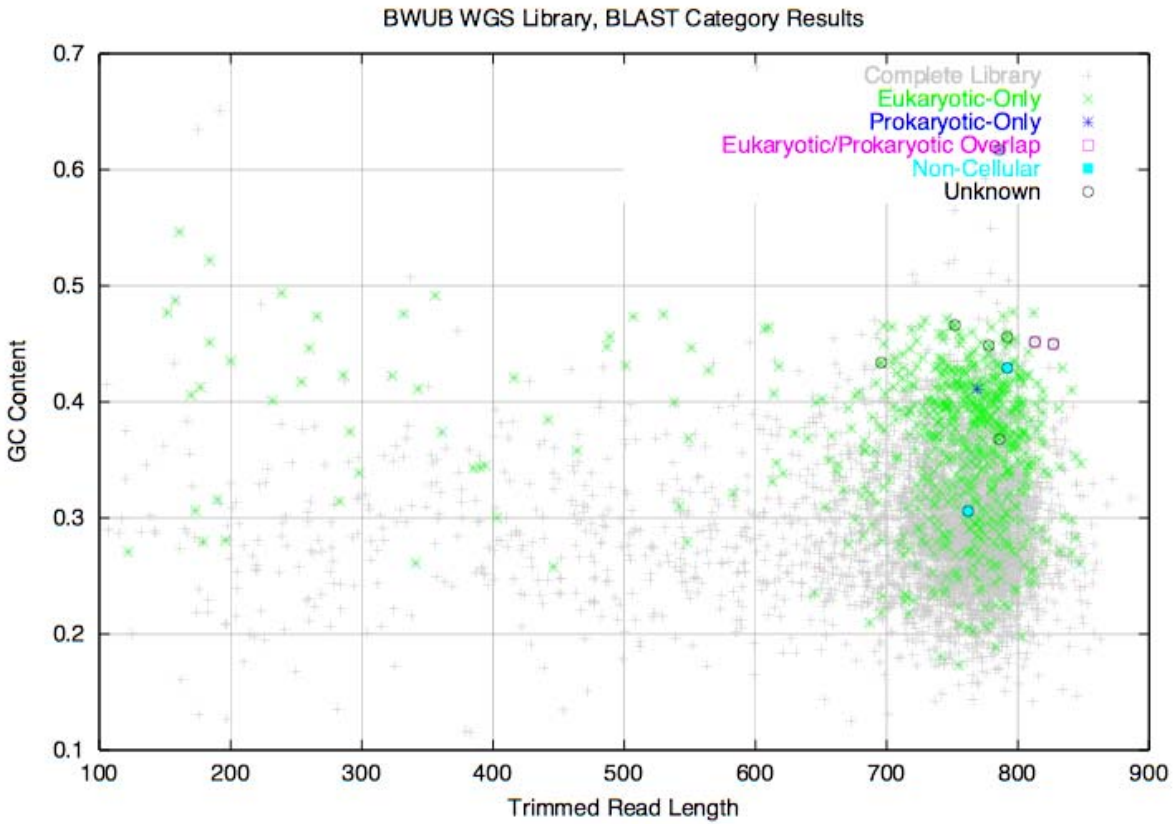
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Homopolymer Block Analysis (2)



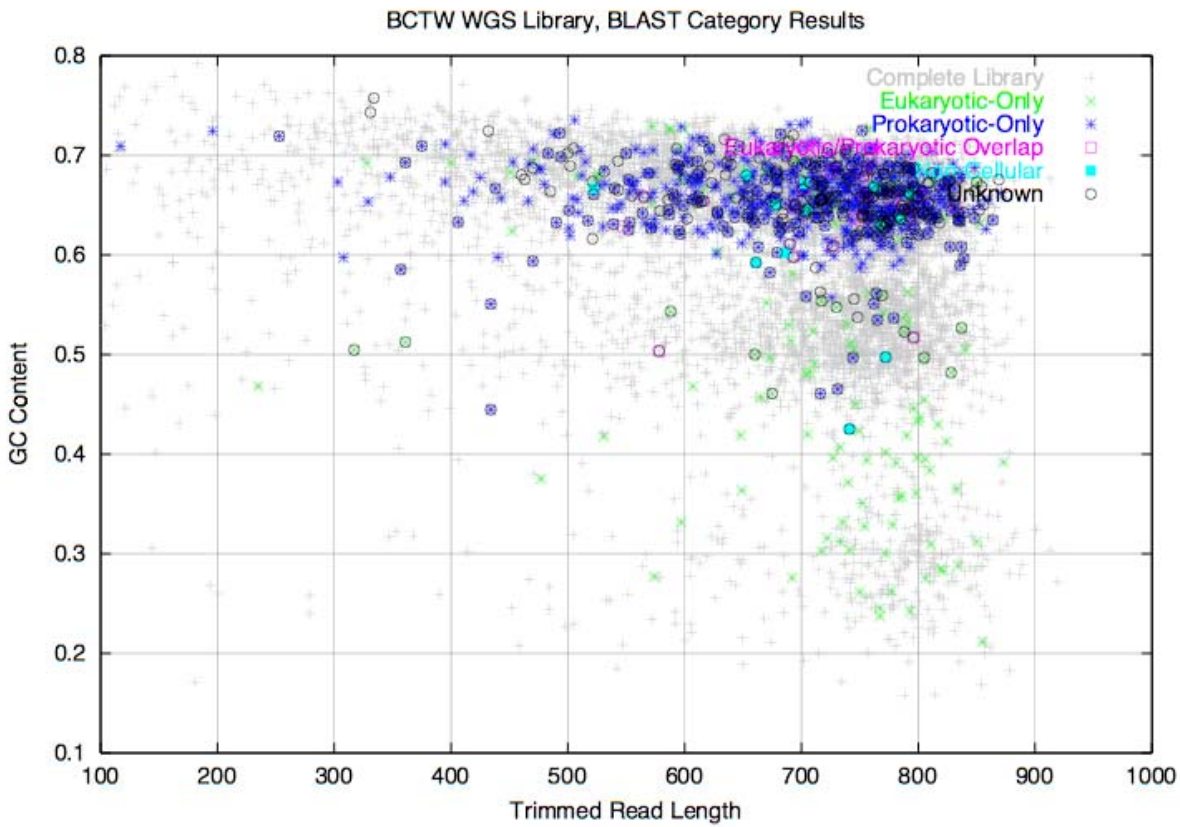
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Kitchen-Sink BLAST Results: A Good Library



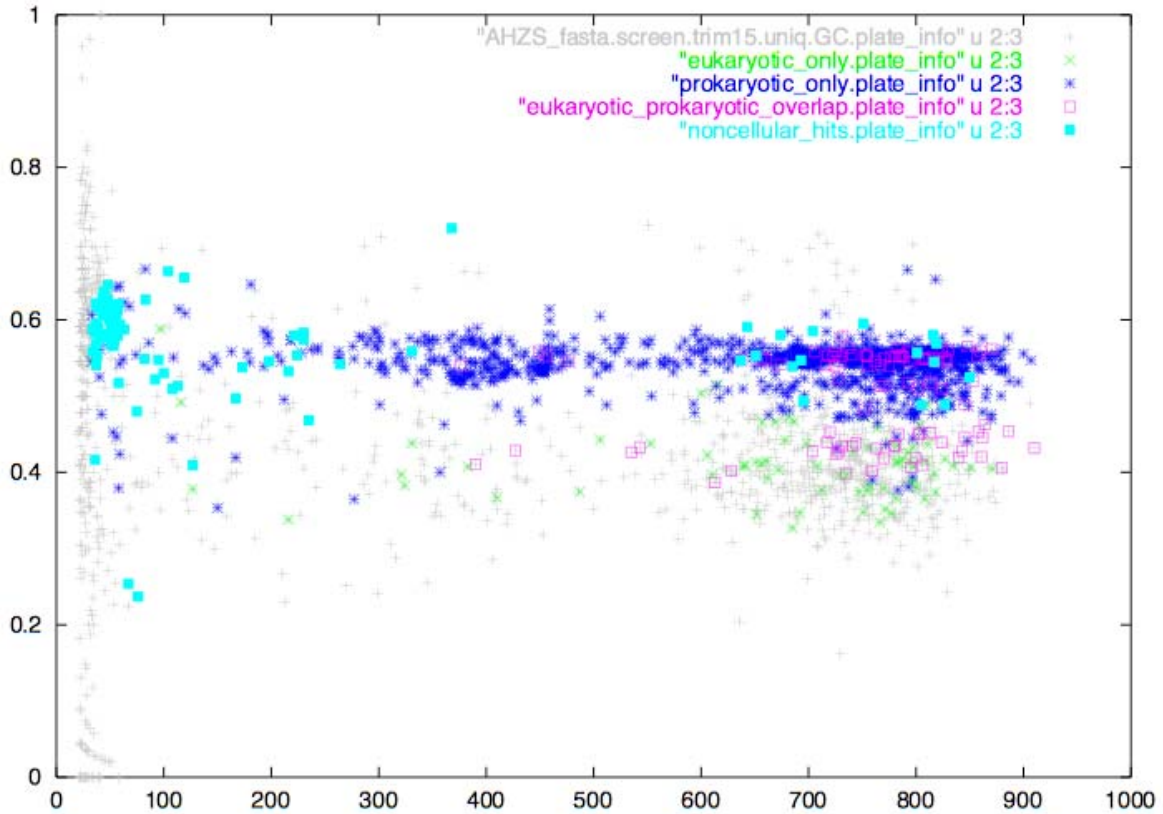
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Kitchen-Sink BLAST Results: A Bad Library



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Kitchen-Sink BLAST Results: A Really Bad Library



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

Subcloned Fosmid QC Procedure

- **Standard QC procedure**
  - **Random selection based on fosmid end sequences**
  - **Assembly with phrap**
  - **Kitchen-sink BLAST of phrap contigs to screen out contaminants**
- **Extended QC procedure (August 2008)**
  - **Kitchen-sink BLAST of phrap contigs against the current NCBI nt database**
  - **Screening with good WGS libraries (Sanger & 454)**
  - **Screening with five EST data sets**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Subcloned Fosmid Inventory

Batch (Date)	Total Subclones	Initial QC (Pass/Provisional/Fail)	Revised QC (Pass//Fail)
1 (8/2004)	15	0/1/14	0/15
2 (2/2005)	10	0/7/3	7/3
3 (5/2005)	10	0/10/0	2/8
4 (11/ 2006)	25	20/4/1	22/3
5 (7/2008)	12	11/1/0	11/1

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

How Did Bad Fosmids Slip Through?

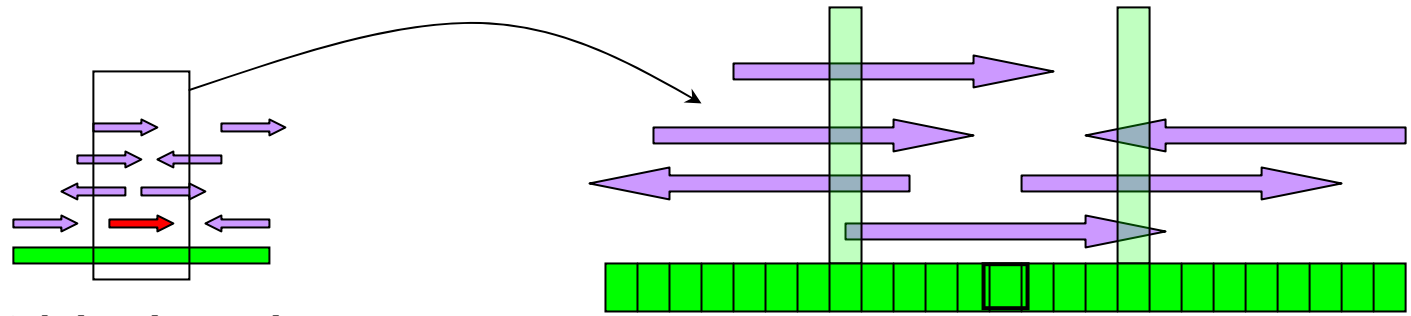
- **Batch 1: Legacy fosmid selection technique; all but one caught during initial QC**
- **Batch 2: Attempt to generate useful data from a failed fosmid library; all failed subclones caught during initial QC/finishing**
- **Batch 3: Lack of confirmed Glomus data for comparison**
- **Batches 4, 5: Difficulty of ruling out plant sequence based on fosmid ends alone**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Sequence Depth Estimation

- **Assembly-based methods:**
  - **Contig base coverage**
- **Non-assembly methods:**
  - **Subcloned fosmid coverage**
  - **EST coverage**
  - **k-mer frequency distribution (not done)**

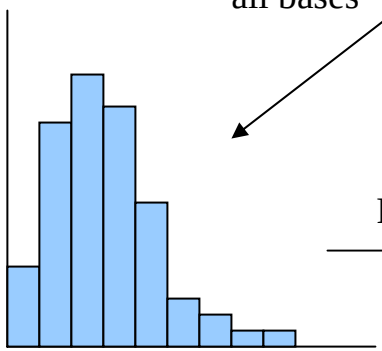
Subcloned Fosmid Coverage Method



Subcloned Fosmids

Count the alignments covering each subcloned fosmid base

Histogram of all bases



Number of Bases

Number of Alignments

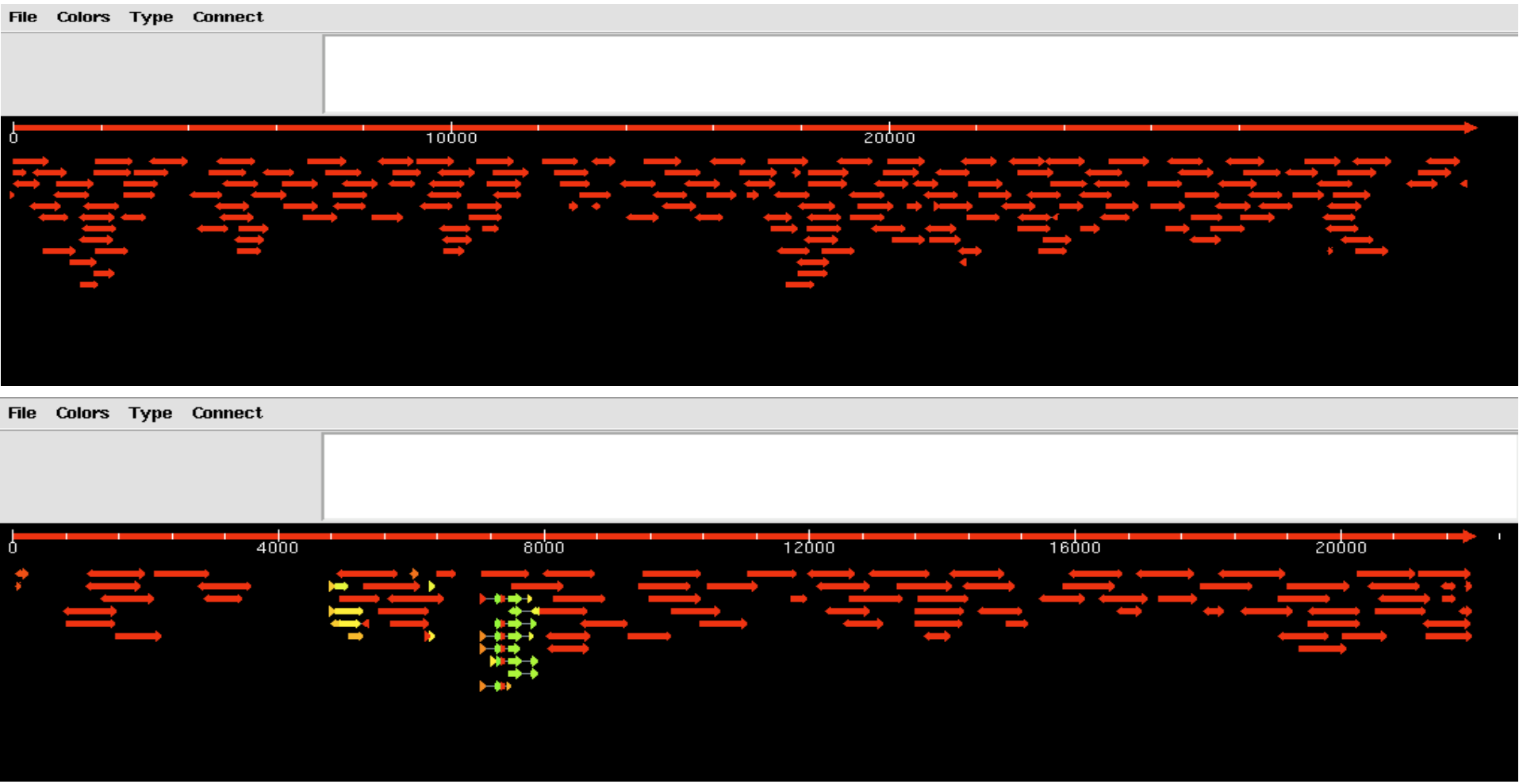
Poisson curve-fit

$$f(x) = N \cdot \exp[-d] \cdot \frac{d^x}{x!}$$

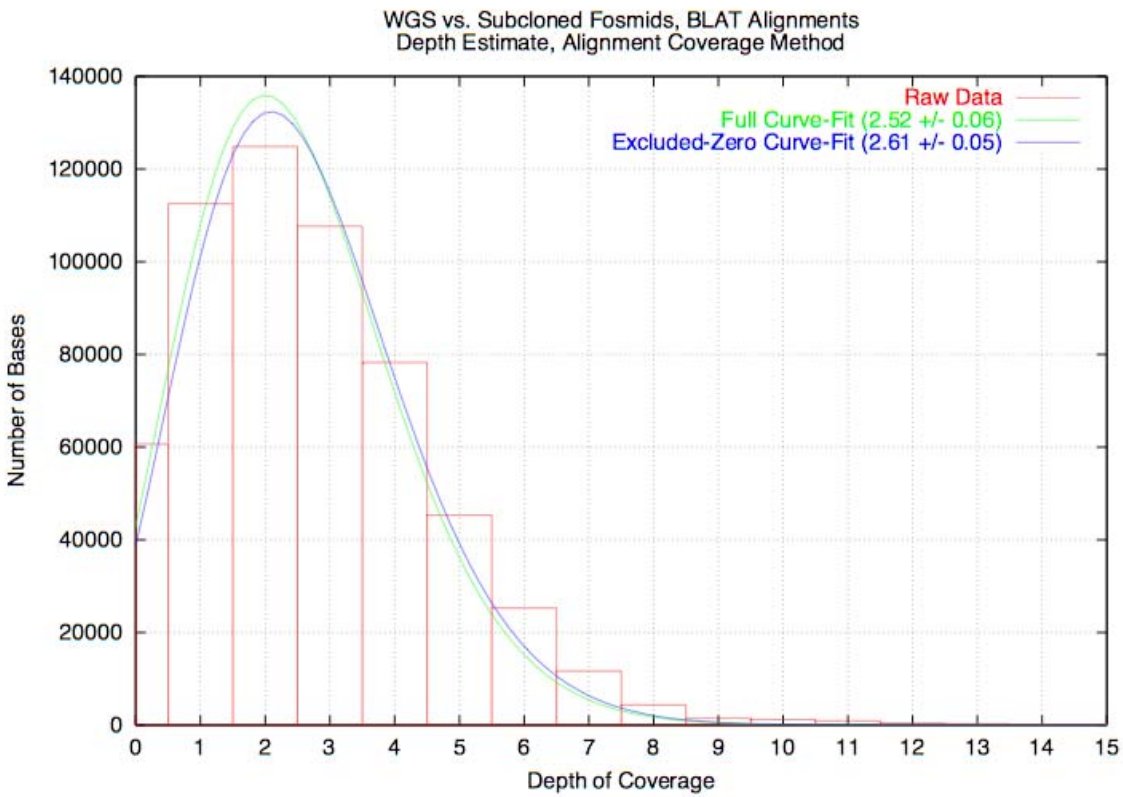
d = Estimated sequence depth

N = Number of read ends

Subcloned Fosmid Method: Good Genome (1)



Subcloned Fosmid Method: Good Genome (2)

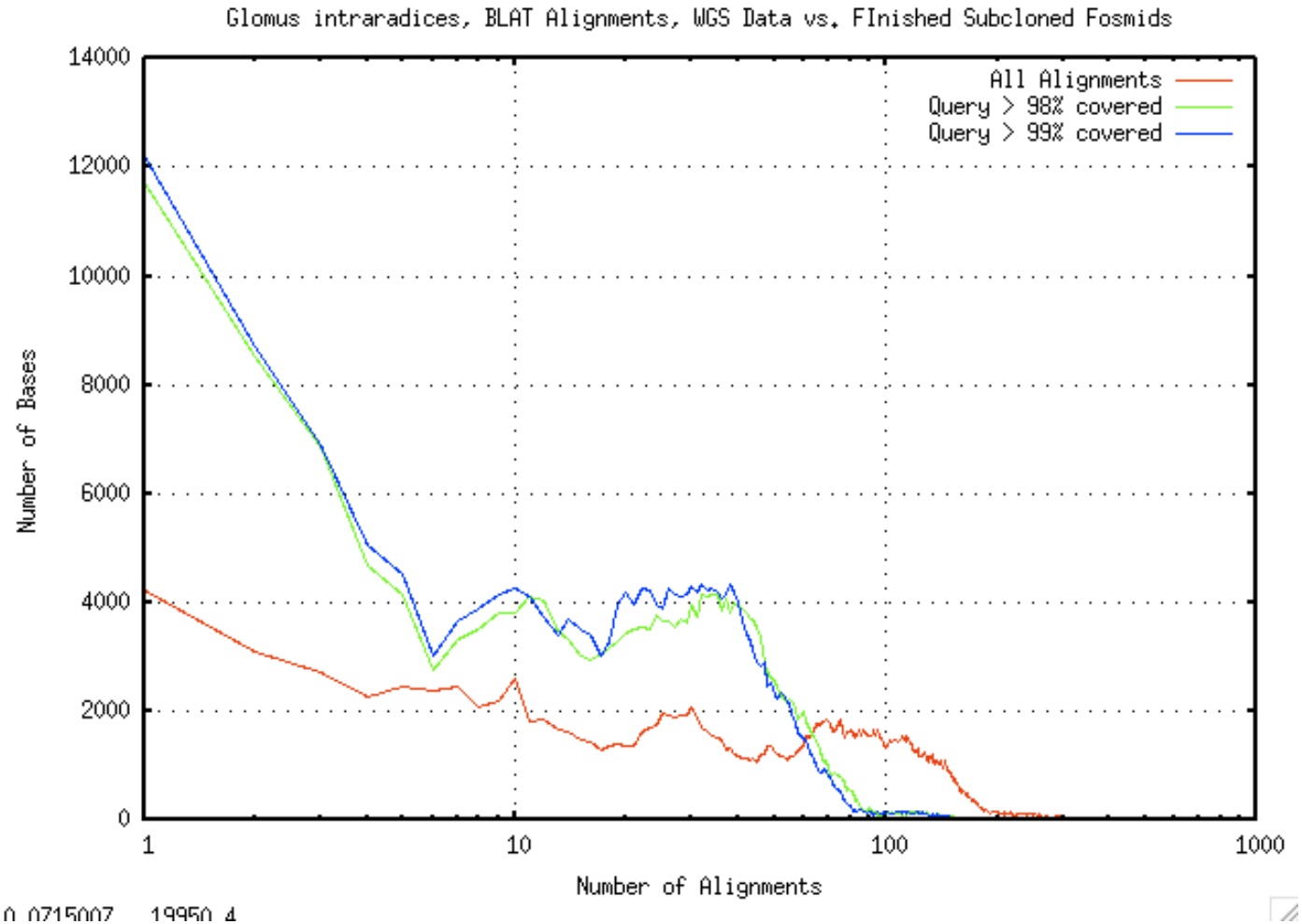


DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Subcloned Fosmid Method: Sample *Glomus* Alignment Results

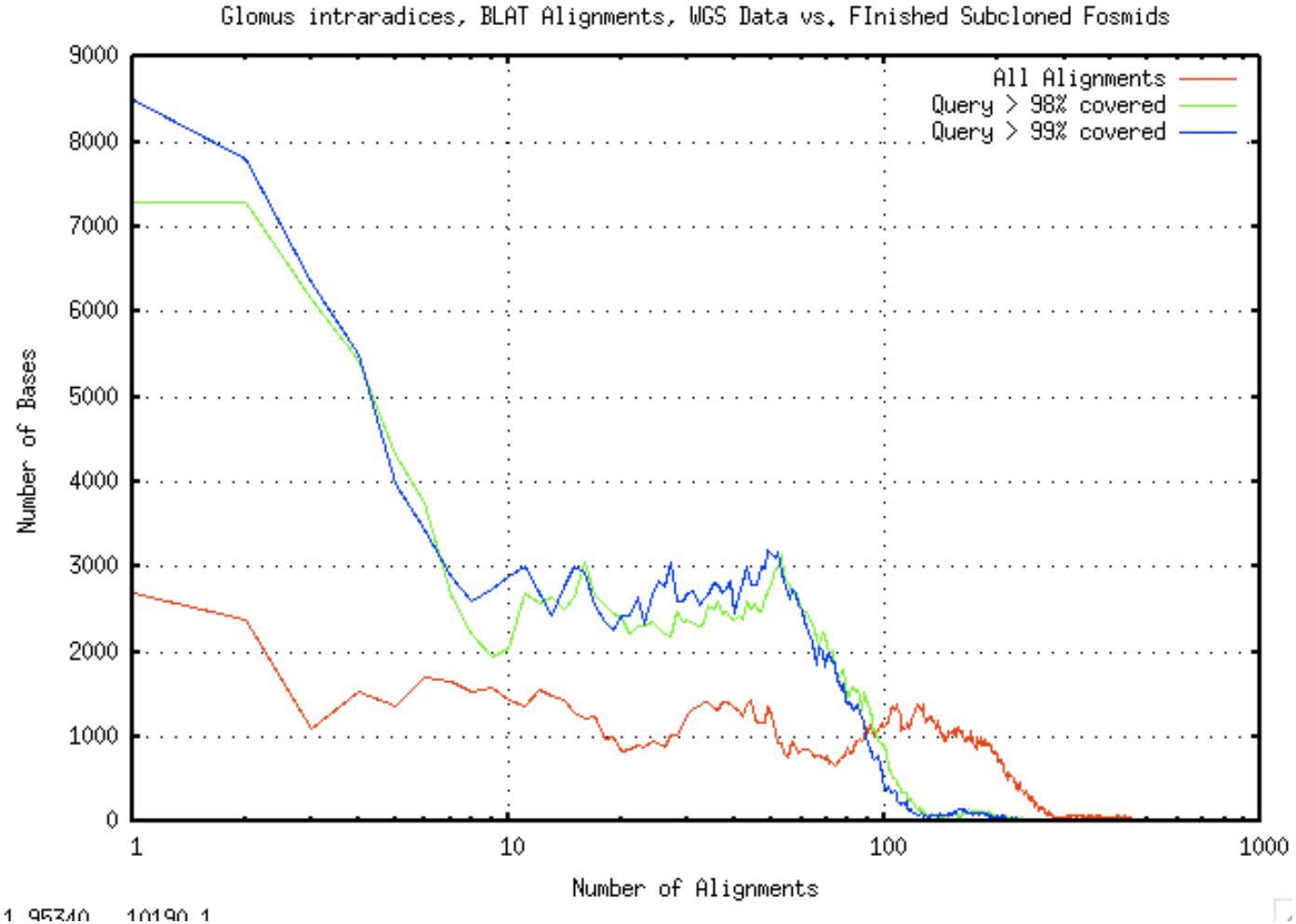
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Subcloned Fosmid Method: Finished Fosmid Results (1)



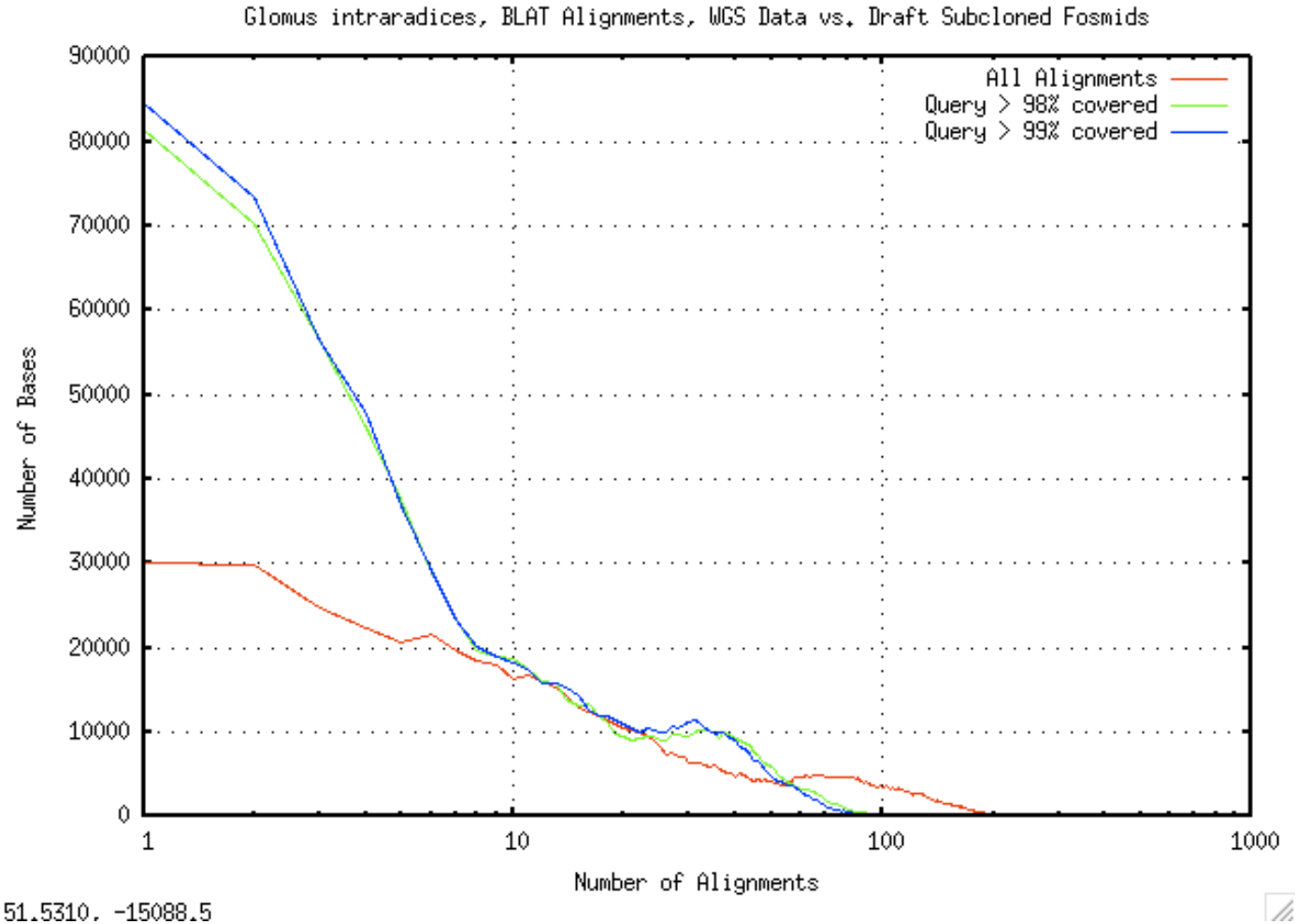
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Subcloned Fosmid Method: Finished Fosmid Results (2)



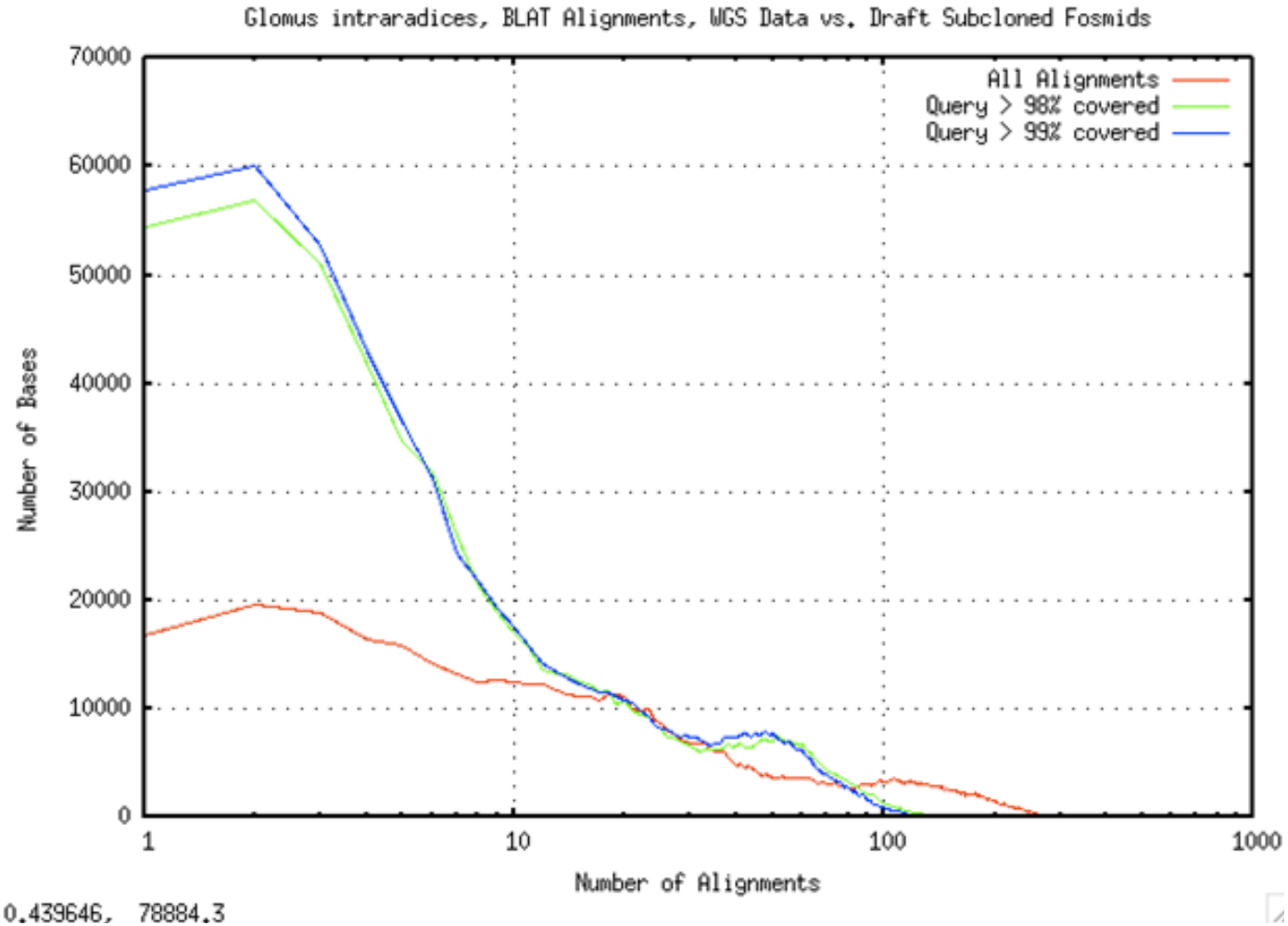
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Subcloned Fosmid Method: Draft Fosmid Results (1)



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Subcloned Fosmid Method: Draft Fosmid Results (2)



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

EST Coverage Analysis Procedure

- **BLAT-aligned the available WGS data sets against five different EST sets**
- **Calculated the fractions of each EST set not hit by WGS sequences, using three different thresholds**
- **Assuming a Poisson distribution, inferred a sequence depth from the fractions of each set without coverage**
- **Available EST data sets:**
  - **JGI Sanger libraries (CACE, CCHU)**
  - **INRA (454)**
  - **MSU (454)**
  - **SAMS (Consensus sequences)**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

EST Coverage Analysis Results

EST Data Set	< 1% Covered (Inferred Depth)	< 10% Covered (Inferred Depth)	< 20% Covered (Inferred Depth)
CACE	5.0% (3.0)	5.5% (2.9)	8.4% (2.5)
CCHU	14.4% (1.95)	14.9% (1.9)	16.7% (1.8)
INRA	55.7% (0.6)	55.7% (0.6)	57.0% (0.55)
MSU	16.4% (1.8)	16.4% (1.8)	18.6% (1.7)
SAMS	11.8% (2.15)	12.3% (2.1)	14.9% (1.9)

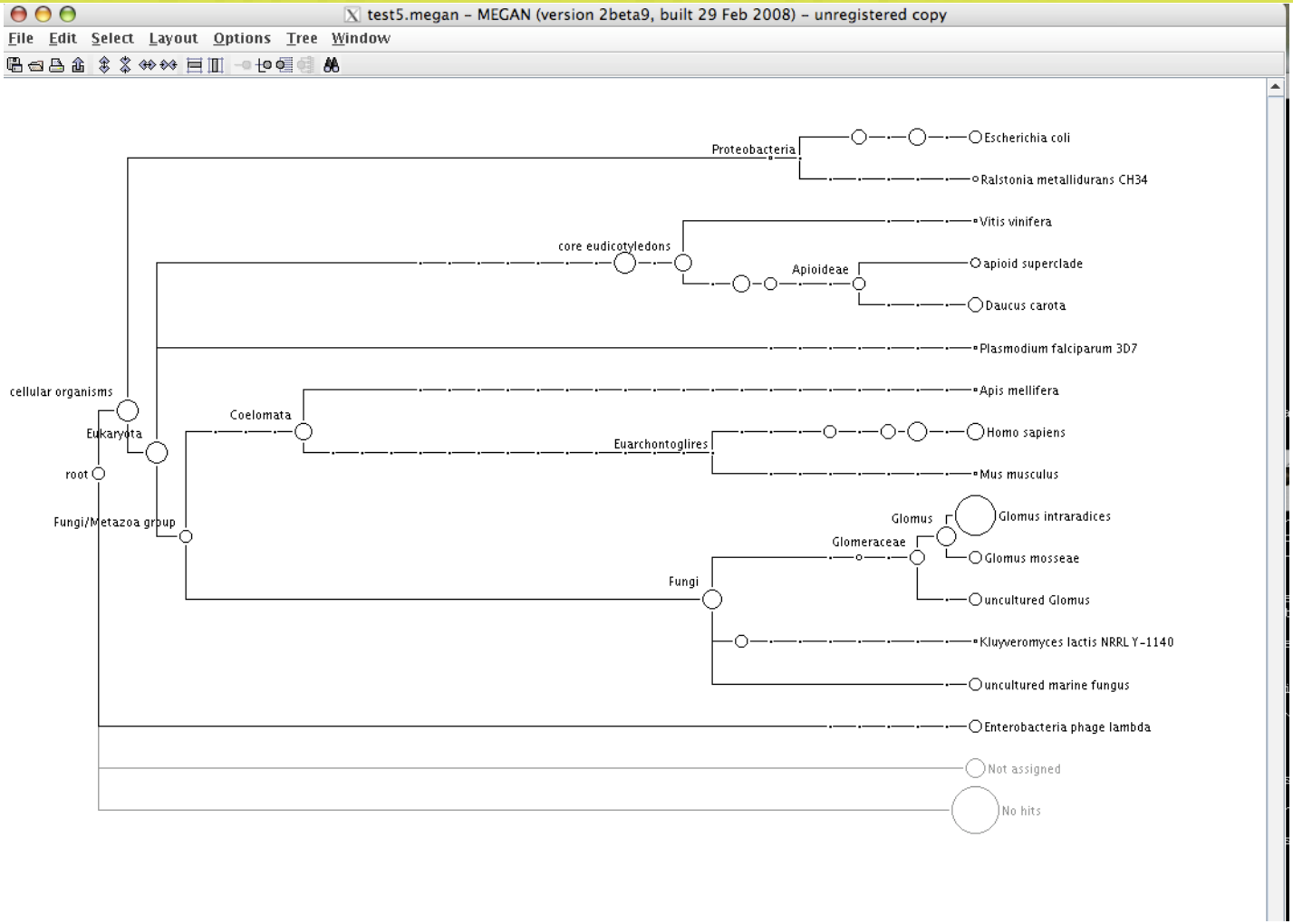
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Newbler Assembly Statistics

Assembly	Contig Total	Contig L50	Largest Contig	Large Scaffold Total	Scaffold L50	Largest Scaffold	Estimated Depth
test2	10.9 MB	813	70,783	7.6 MB	831	72,961	1.88 +/- 0.02
test3	10.7 MB	814	70,783	7.6 MB	836	102,082	1.88 +/- 0.02
test4	10.7 MB	809	55,799	7.4 MB	828	70,803	1.88 +/- 0.18
test5	12.4 MB	800	70,741	7.8 MB	814	70,741	1.73 +/- 0.12
test6	12.3 MB	795	8,834	7.6 MB	803	13,271	1.73 +/- 0.12
test7	27.0 MB	815	6,957	14.1 MB	819	17,331	2.36 +/- 0.22
test8	28.0 MB	815	6,957	14.1 MB	819	17,331	2.38 +/- 0.22

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Newbler Assembly BLAST Results



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

Why Won't *Glomus* Assemble?

- **Substantially larger physical genome size**
- **Cloning bias**
- **Substantial amounts of undetected contamination**
- **Polymorphism**

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

Substantial Undetected Contamination?

- **The WGS data sets would need to be almost entirely non-*Glomus***
- **Prokarytic contamination**
- **Eukaryotic contamination**
  - **Plant**
  - **Fungus**
  - **Other?**

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

Ribosome PCR Results

- **Three ribosome PCR libraries were created, using general 18S primers**
- **192 reads were sequenced from each, and processed using the standard library QC procedure**
- **For all three libraries, the kitchen-sink BLAST mainly yielded hits to *Glomus*.**
- **All three libraries had a few reads hit the Apiineae suborder**

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

So What Does That Leave?

- **The fundamental issue is a lack of effective sequence depth, due to a genome space that is much larger than the per-nucleus genome size**
- **Three separate estimates (one assembly-based, two others not) are consistent with an effective sequence depth of ~2**
- **Some type of polymorphism is about the only plausible explanation left**
  - **SNPs**
  - **Rearrangements**
  - **Other?**

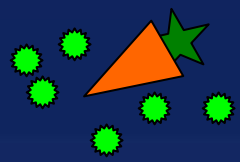
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Whither the *Glomus* Project?

- **Generation of additional sequence depth**
  - **Additional 454 FLX sequencing**
  - **454 Titanium sequencing**
- **Alternate methods**
  - **Large scale 454 sequencing of pooled, bar-coded subclones**
  - **Single-nucleus amplification**

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

Acknowledgements



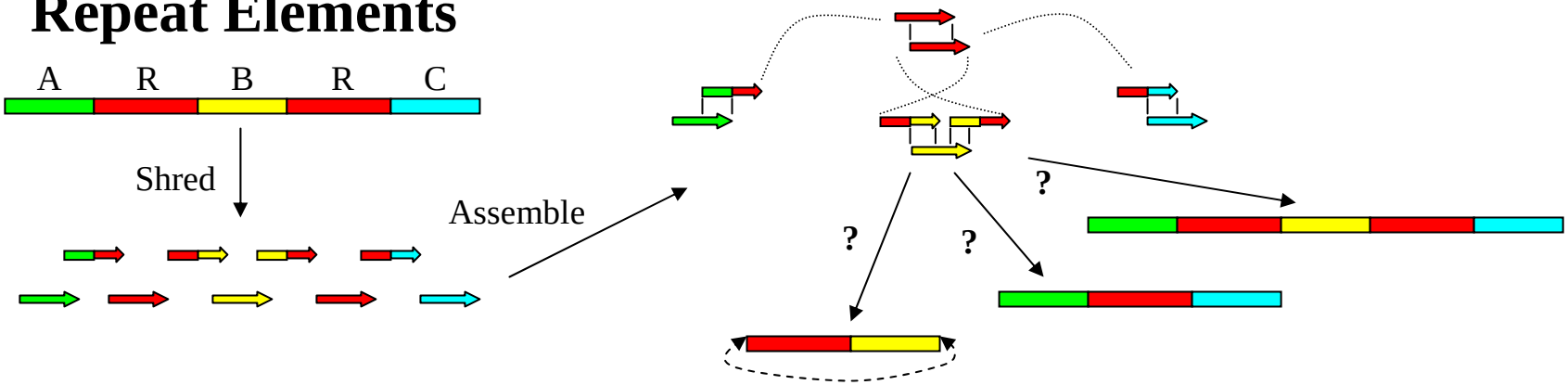
- **The *Glomus* project consortium**
- **JGI Cloning Technology group; Jane Grimwood**
- **JGI EST support: Erika Lindquist, Jasmyn Pangilinan**
- **Jan-Fang Chang, Feng Chen (Ed Kirton)**
- **Alex Copeland (Production QA/QC)**

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 Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

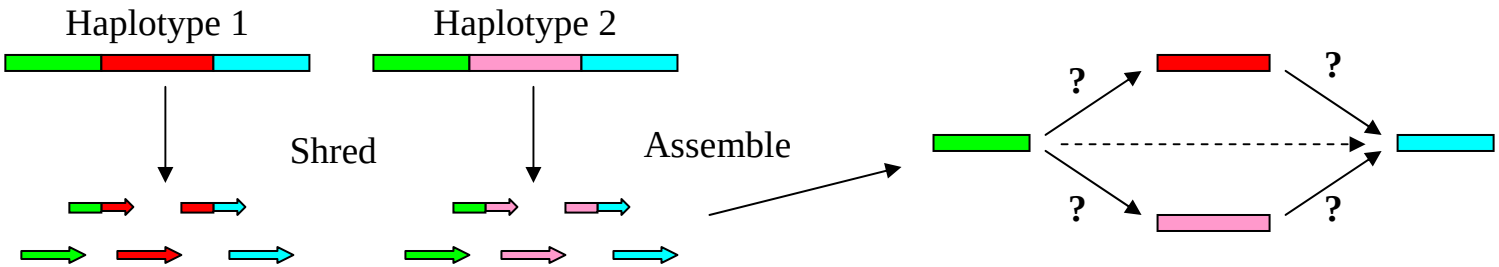
JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

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

Auspice Statement

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 Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.