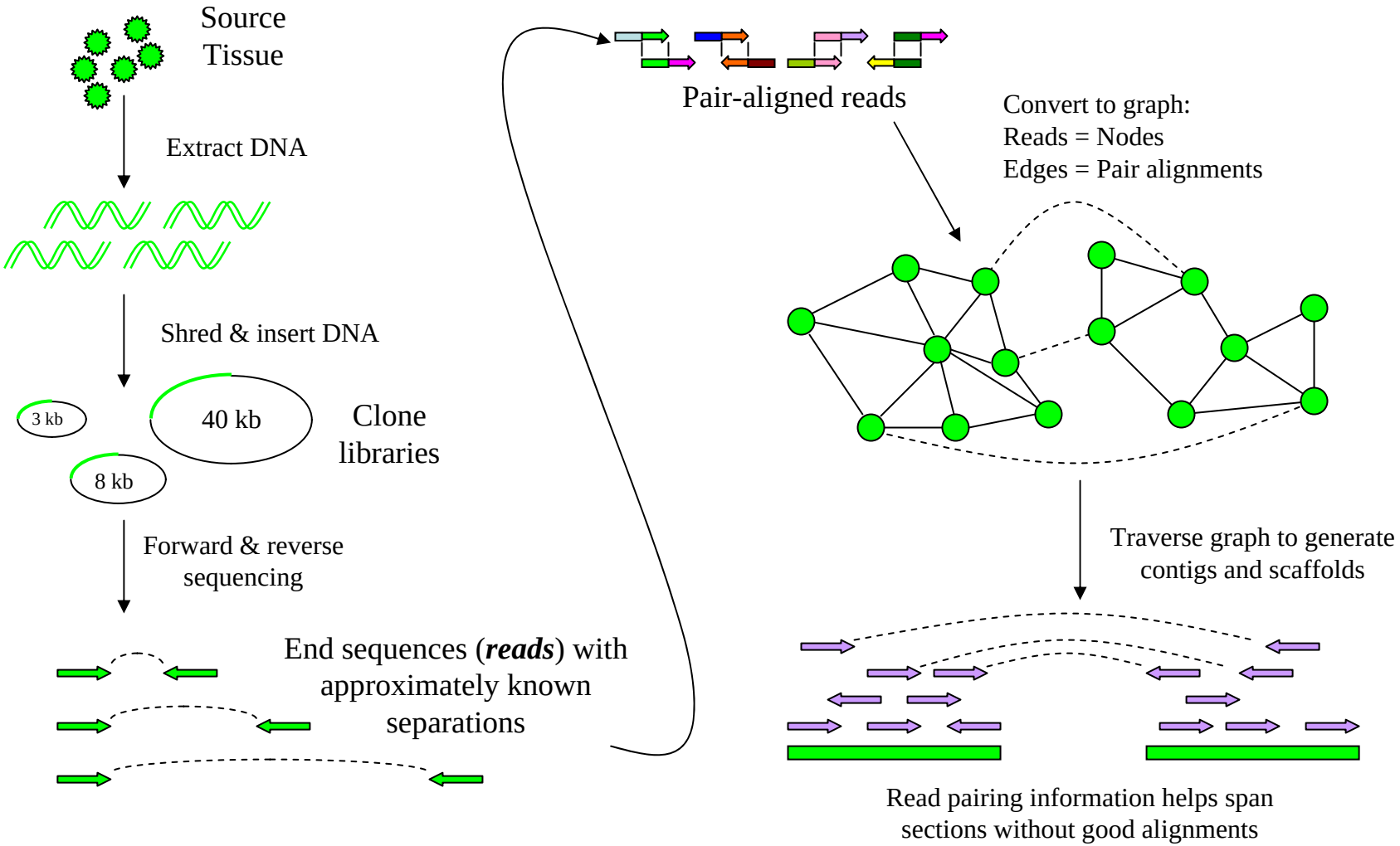
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

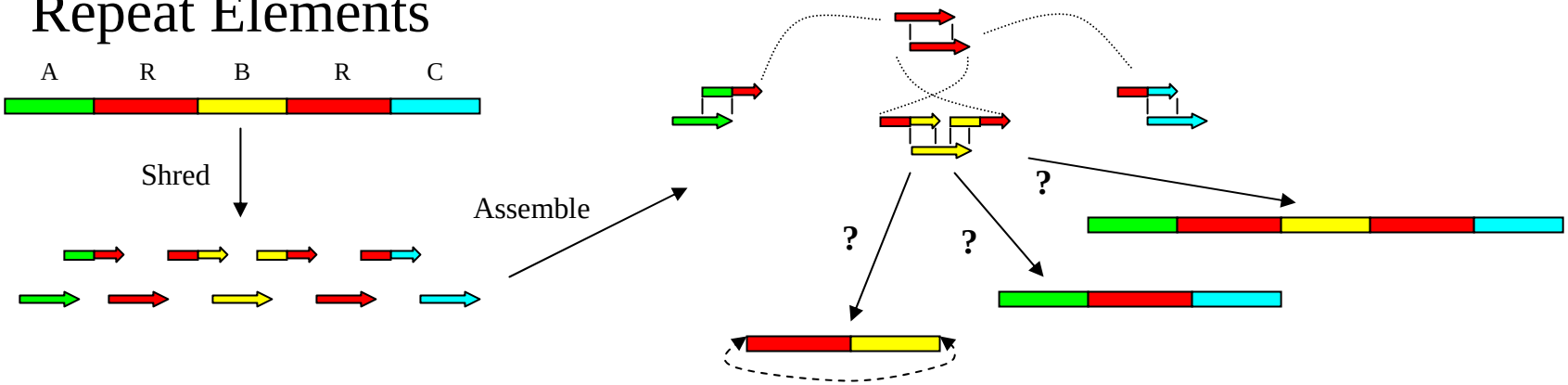
Outline of the Assembly Process: JAZZ, the JGI In-House Assembler



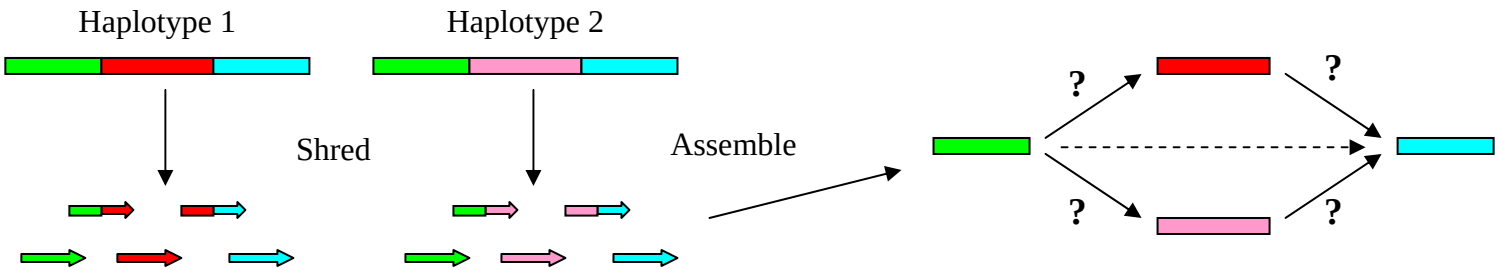
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

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.

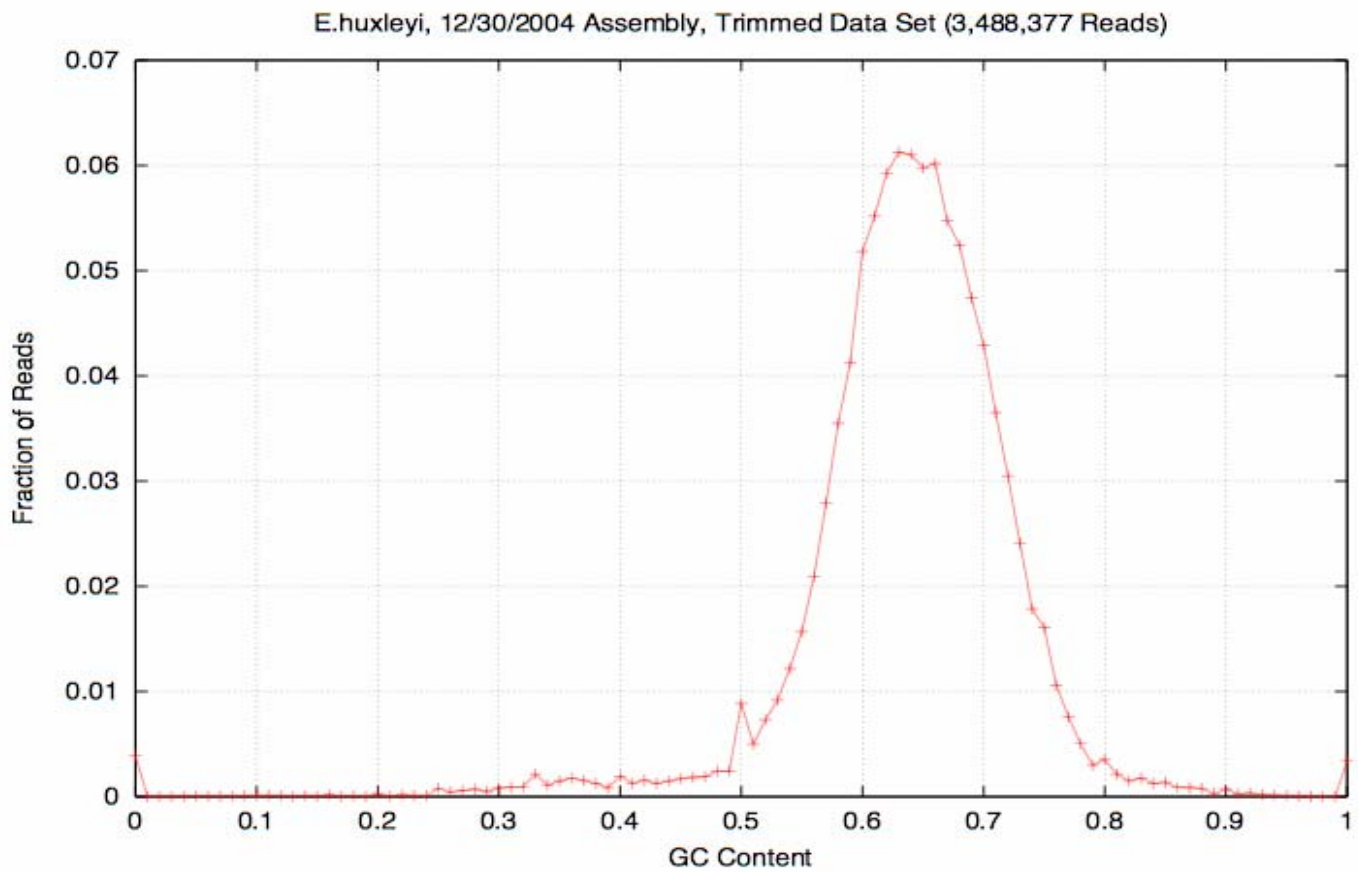
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Summary of Assembly Data Sets

Quantity	9/30/2004 Assembly	12/30/2004 Assembly
Number of Untrimmed Reads	2,291,871	3,894,600
Amount of Untrimmed Sequence	2,284 MB	3,910 MB
Number of Used Reads	1,689,410	3,087,548
Amount of Used Sequence	966 MB	1,917 MB

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

WGS Library GC Content Distribution



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Assembly Results 1: Scaffold & Contig Statistics

Quantity	9/30/2004 Assembly	12/30/2004 Assembly
Scaffold Total	10,849	10,514
Scaffold Sequence Total	213 MB	290 MB
Scaffold N50	277	408
Scaffold L50	173 KB	173 KB
Contig Total	48,572	54,652
Contig Sequence Total	139 MB (35.1% gap)	196 MB (32.2% gap)
Contig N50	6,753	5,680
Contig L50	4,700	7,786

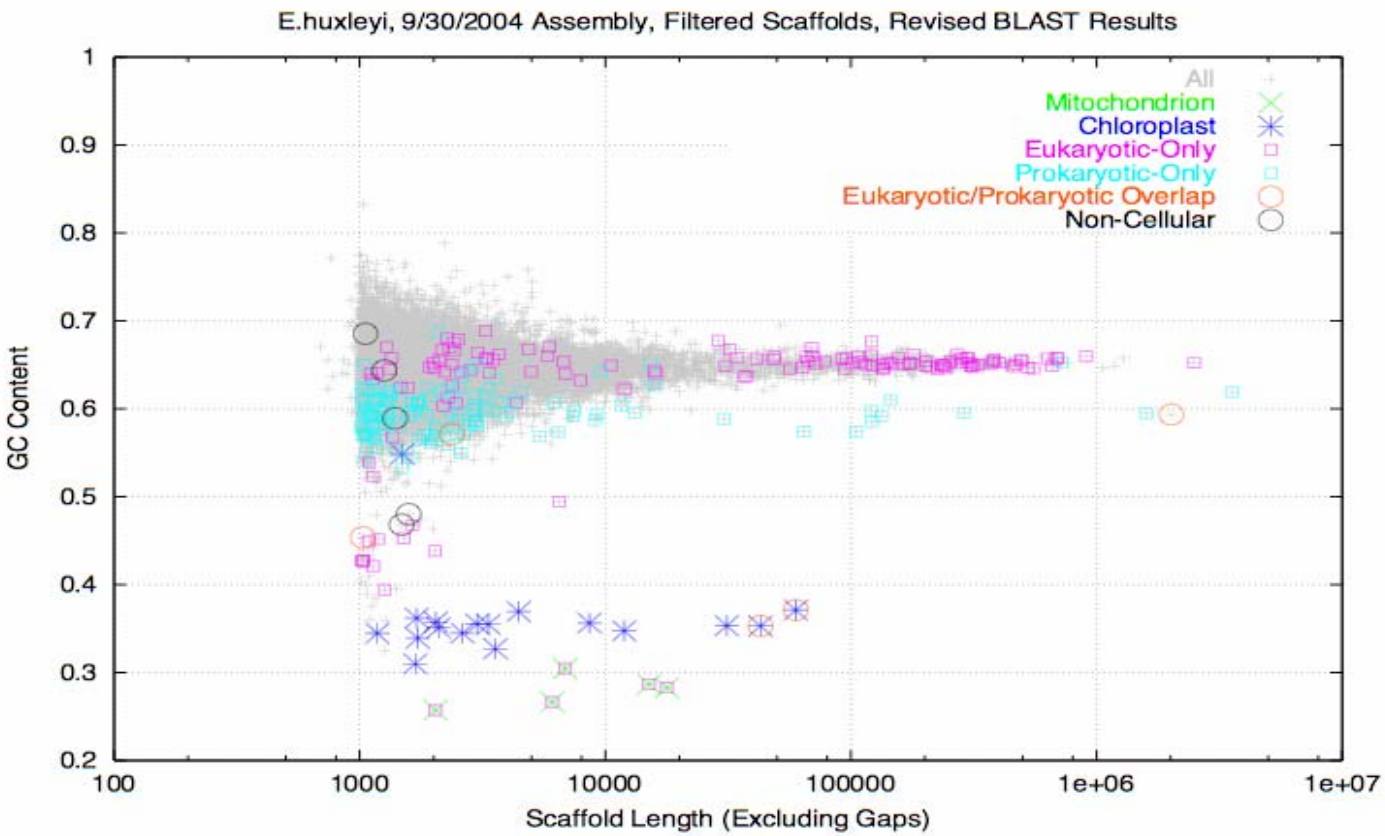
JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

Assembly Results 2: Depth & Completeness Estimates

Quantity	9/30/2004 Assembly	12/30/2004 Assembly
Estimated Depth	7.2 +/- 0.3	13.8 +/- 0.5
Data Completeness (> 20% Covered)	97.6%	98.2%
Data Completeness (> 50% Covered)	96.9%	97.7%
Data Completeness (> 80% Covered)	93.9%	96.5%
Scaffold Completeness	96.1%	97.4%

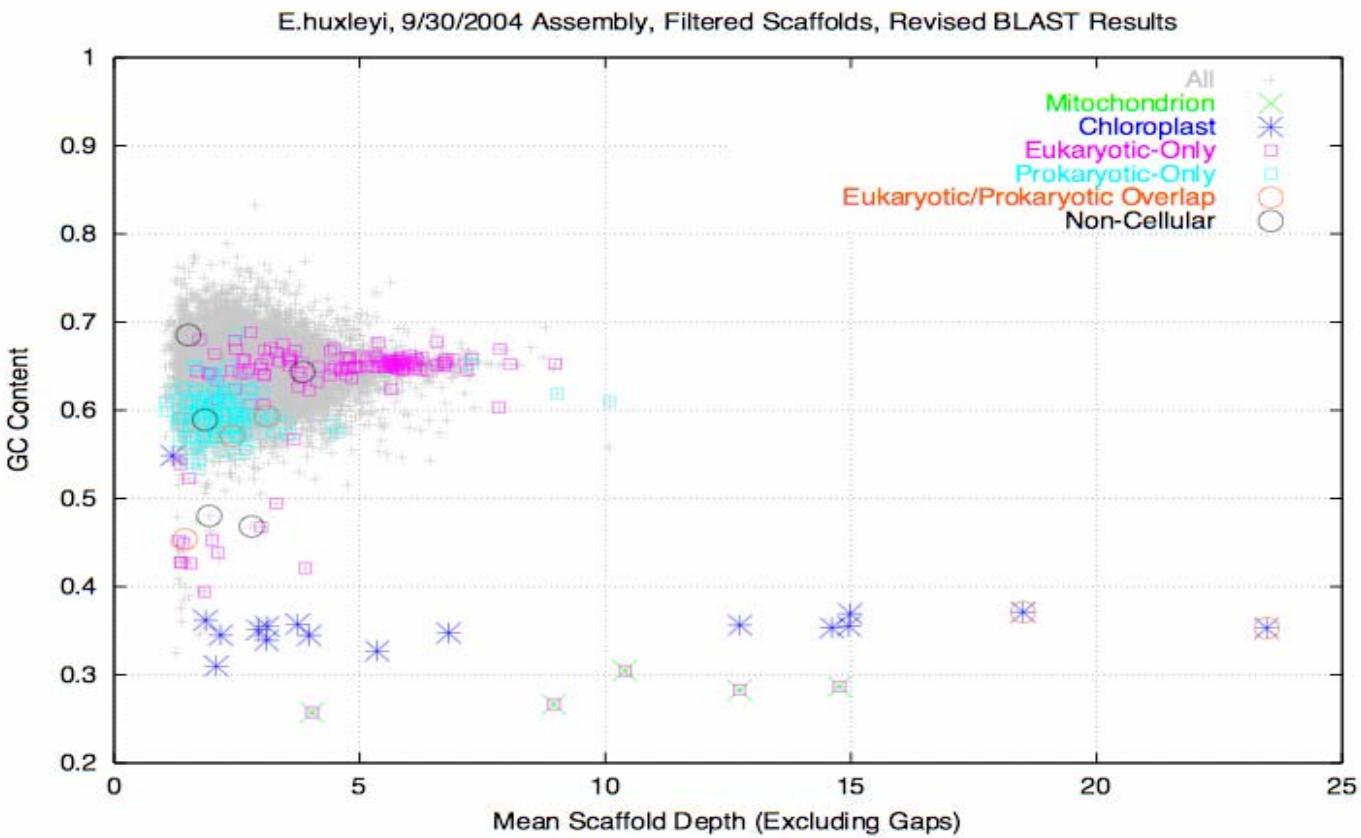
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Assembly Results 3: BLAST Analysis Results, GC vs. Length



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

Assembly Results 4: BLAST Analysis Results, GC vs. Depth



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

Summary of Prokaryotic Contaminant Results

- Analysis based on the analysis of the 9/30/20004 assembly; results from the 12/30/2004 assembly are not yet available.
- A 3.6 MB scaffold at a depth of about 9x, whose best BLAST hits were to the Erythrobacter genus.
- A set of scaffolds at a depth of 2x - 3 x, whose best hits were to the Agrobacterium genus.
- A complete 145 KB plasmid, of unknown origin.

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Summary of Assembly Results

- The E.huxleyi has proven to be very challenging to assemble.
- With the standard JGI sequencing protocol, the high-GC content of the genome could result in systematic errors at particular sequences. About half of the data set was produced with the regular protocol.
- A high-GC sequencing protocol was developed and used for the second half of the sequencing. Combining this sequence with the potential systematic biases of the “regular” set could produce an increase in the apparent polymorphism rate.
- Alignment to a set of draft subcloned fosmids suggested that the genome might be highly repetitive.
- Analysis of the consensus sequences for the 12/30/2004 suggested a possible polymorphism rate of about 1.5% However, the method used for this estimate was subject to non-trivial errors in both directions.

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

Where Do We Go From Here?

- Assemble using only the high-GC protocol sequence
- Attempt to estimate the polymorphism rate using the high-GC protocol sequence and the draft subcloned fosmids; adjust the assembly parameters accordingly
- Attempt to identify repeat elements in the genome, and exclude from the initial rounds of assembly

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Acknowledgements

- Jazz assembler: Jarrod Chapman, Nik Putnam, Isaac Ho, Dan Rokhsar
- E.huxleyi assemblies and post-assembly analysis: Harris Shapiro
- Prokaryotic contaminant analysis: Susannah Tringe

- 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.