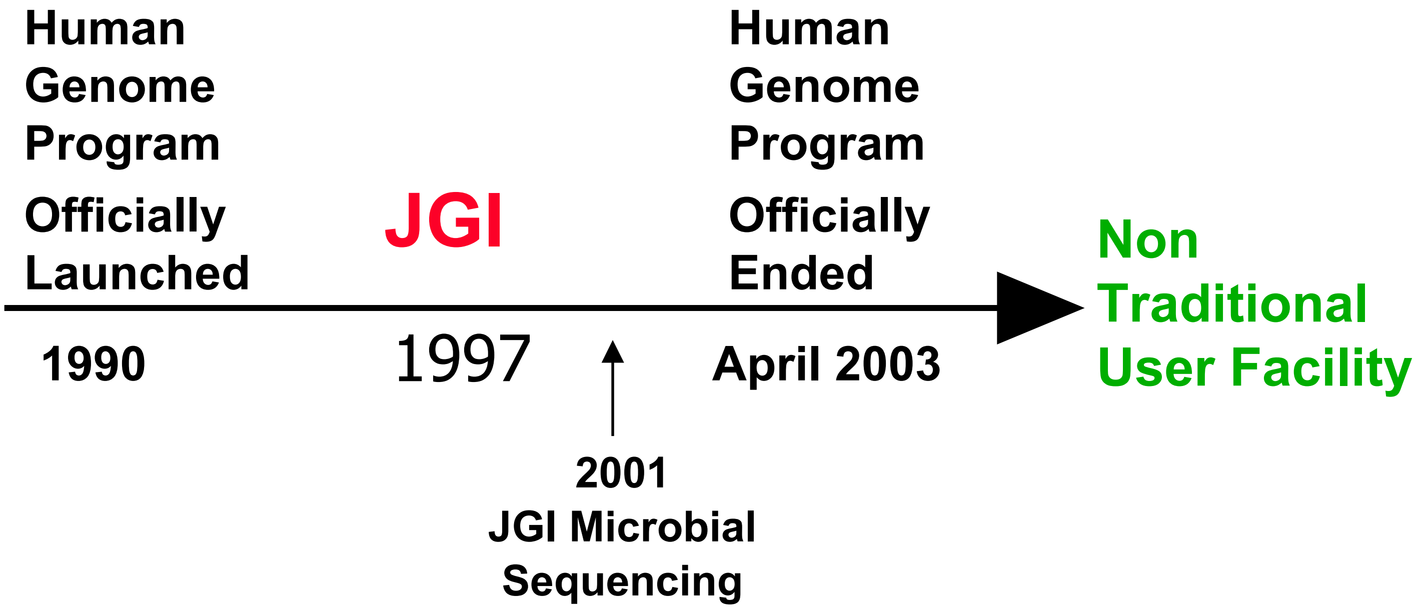
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Joint Genome Institute (JGI)

- Non-Traditional User Facility
- Microbial Genomics Program

JGI Timeline



Human Genome Program Officially Launched

1990

JGI

1997

2001
JGI Microbial
Sequencing

Human Genome Program Officially Ended

April 2003

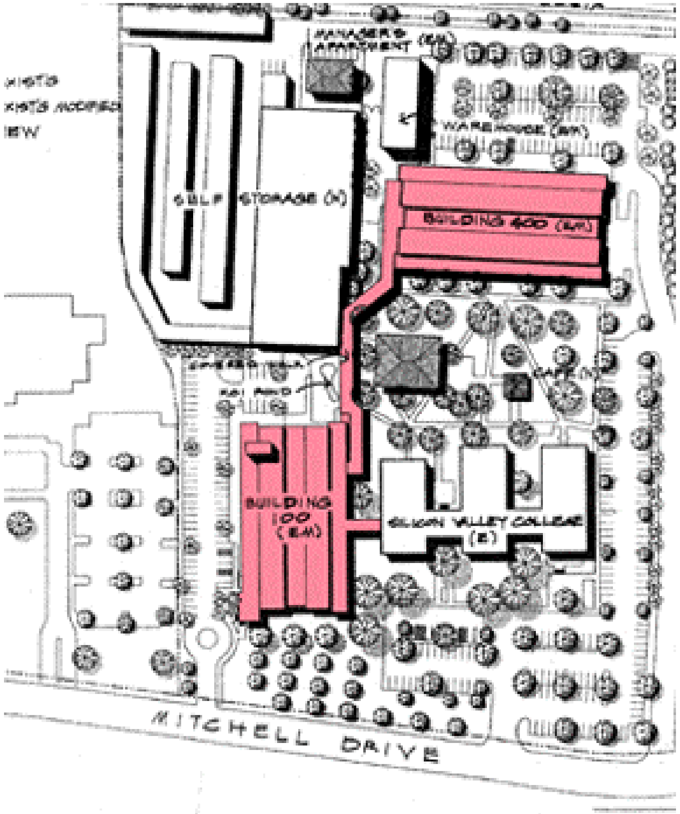
Non Traditional User Facility

US DOE Joint Genome Institute

Formed in 1997 as a MOU between DOE Labs LLNL, LBNL and LANL.

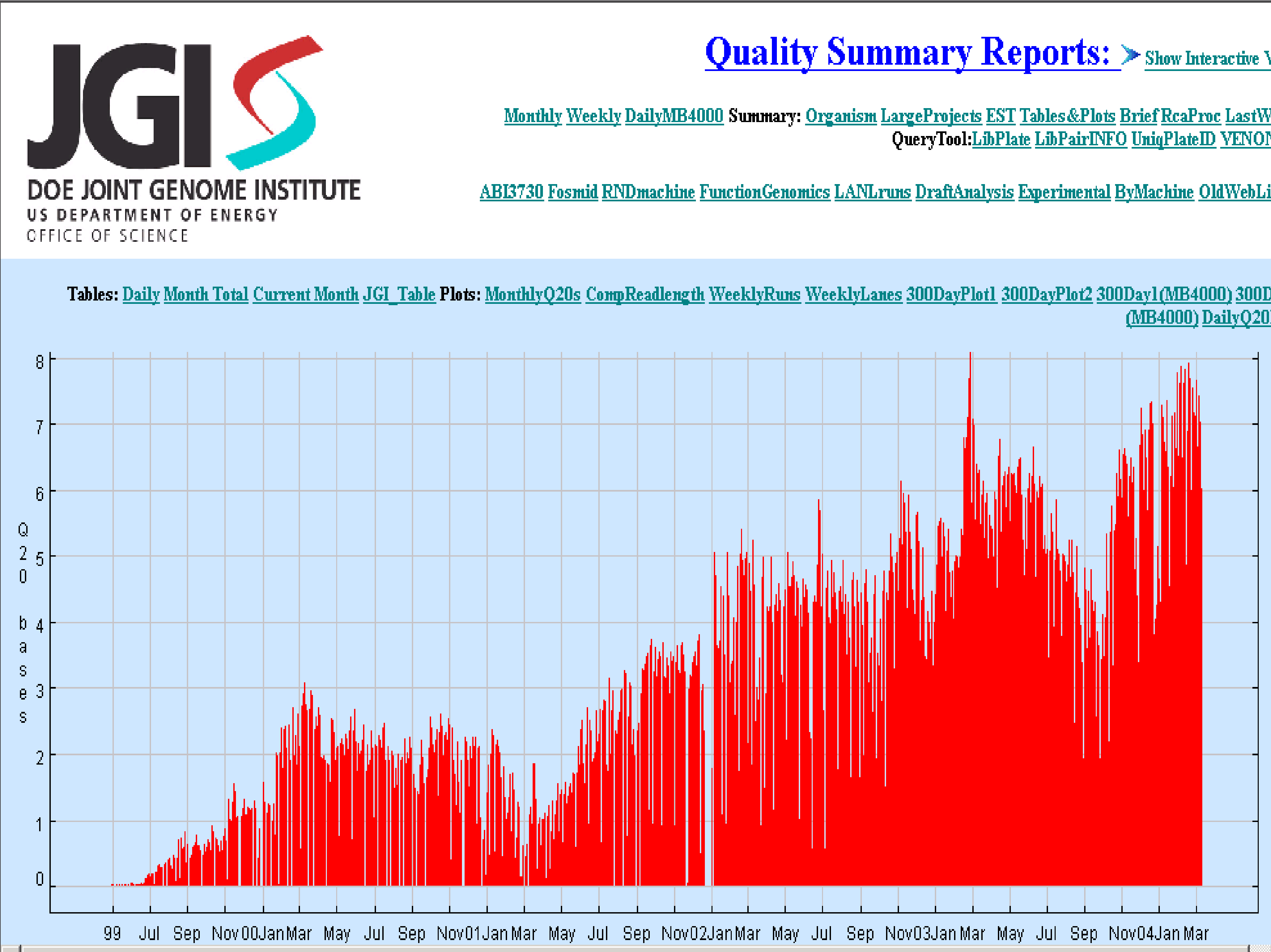
~250 FTEs

165 FTEs PGF
30 FTEs LANL
50 FTEs SHGC
5 FTEs LLNL
2-3 FTEs ORNL



PGF-Production Genomics Facility
Walnut Creek, CA
2 buildings-60,000 sq. ft.

www.jgi.doe.gov



Quality Summary Reports:

Monthly Weekly DailyMB4000 Summary: Organism LargeProjects EST Tables&Plots Brief RcaProc LastW
QueryTool:LibPlate LibPairINFO UniqPlateID YENON

ABI3730 Fosmid RNDmachine FunctionGenomics LANLruns DraftAnalysis Experimental ByMachine OldWebLi

Tables: Daily Month Total Current Month JGI_Table Plots: MonthlyQ20s CompReadlength WeeklyRuns WeeklyLanes 300DayPlot1 300DayPlot2 300Day1(MB4000) 300D
(MB4000) DailyQ20

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

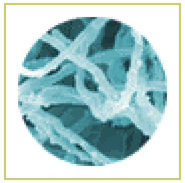
Sequencing Targets

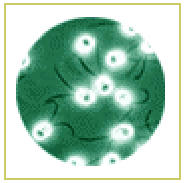
Users: DOE

Microbial Program

Other Governmental Agencies

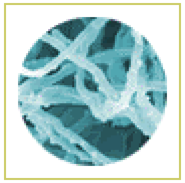
Community Sequencing Program (CSP)

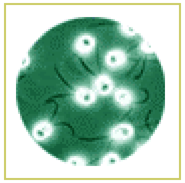




The Community Sequencing Program: (CSP)

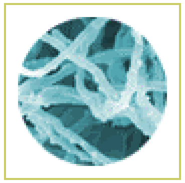
Will provide the scientific community access to high throughput sequencing at the JGI.

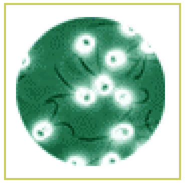




Types of projects:

A wide range of projects will be accepted. Ultimately, the most important factor in determining acceptance is a project's scientific merit.



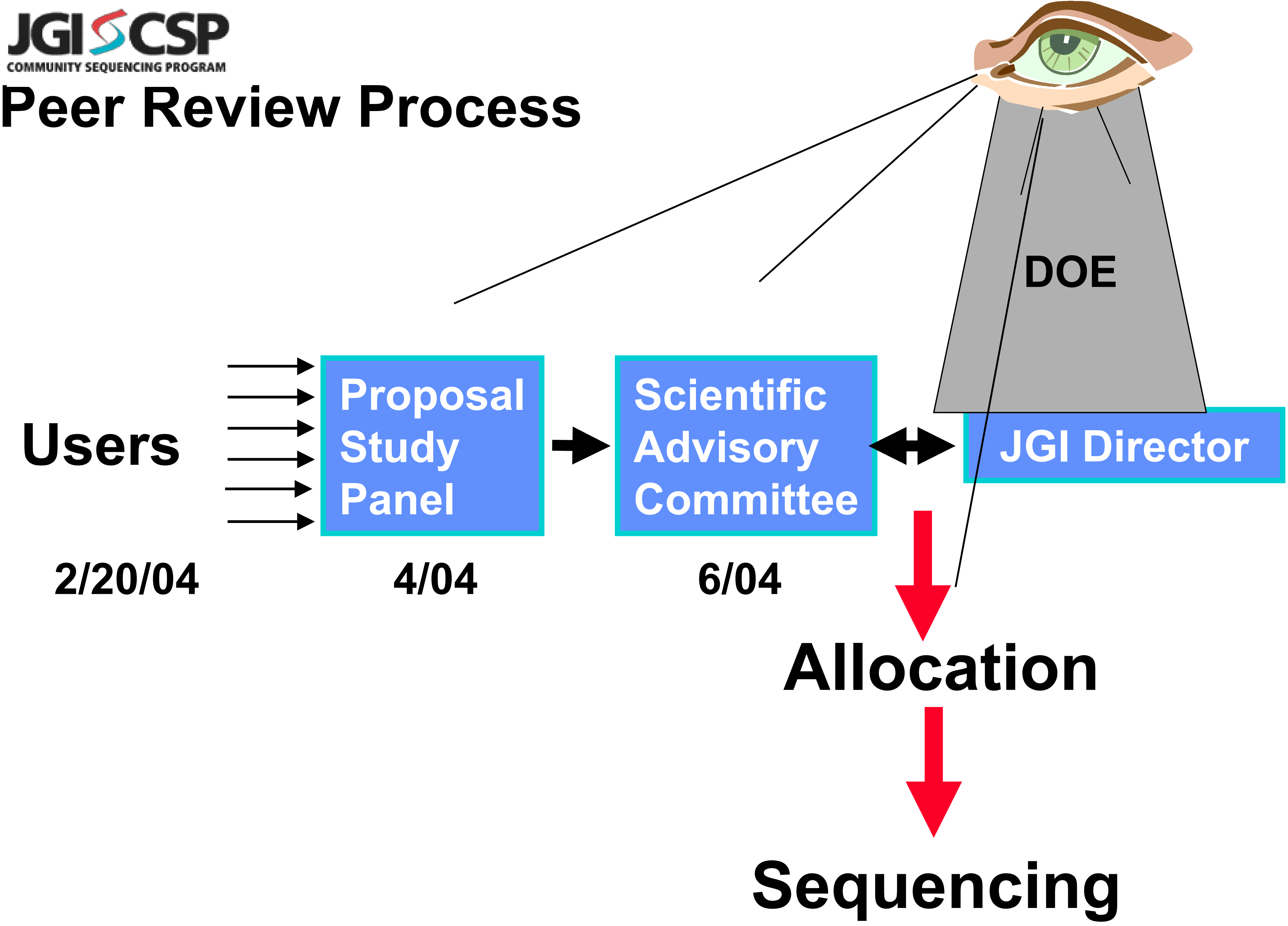


What can researchers get from the CSP program?

The deliverables can range from raw sequence traces to well-annotated assembled genomes

JGI CSP
COMMUNITY SEQUENCING PROGRAM

Peer Review Process



Users

Proposal Study Panel

Scientific Advisory Committee

DOE

JGI Director

2/20/04

4/04

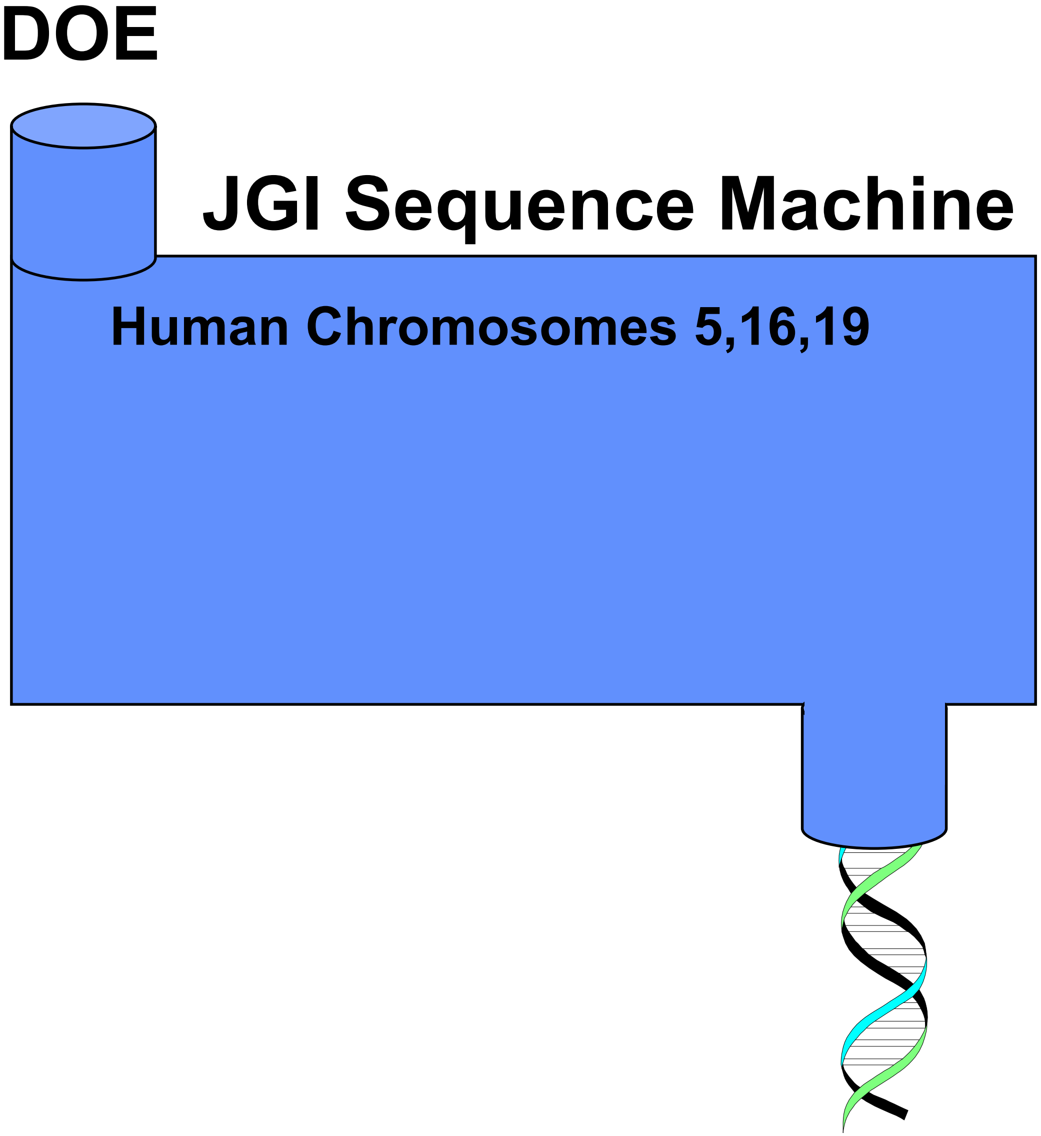
6/04

Allocation

Sequencing

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

In the Beginning….



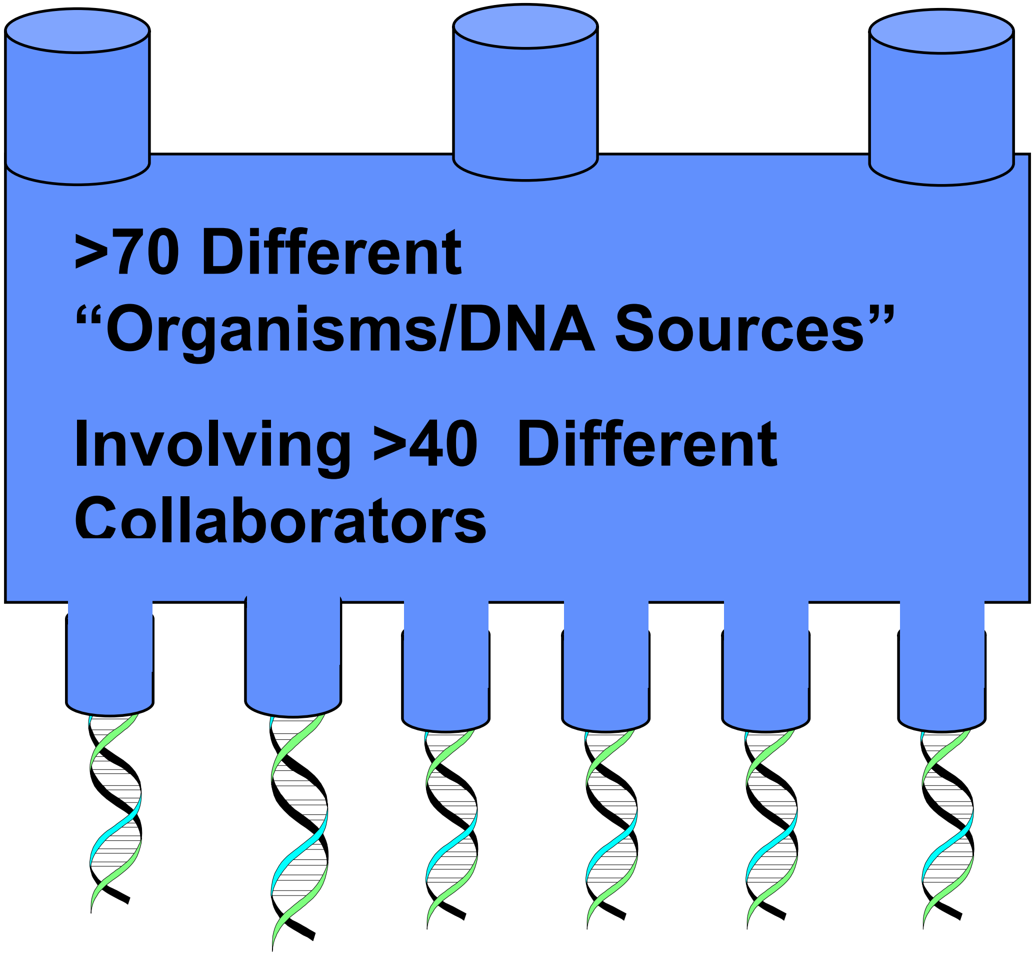
DOE

JGI Sequence Machine

Human Chromosomes 5,16,19

More Genomes

DOE CSP "Others"

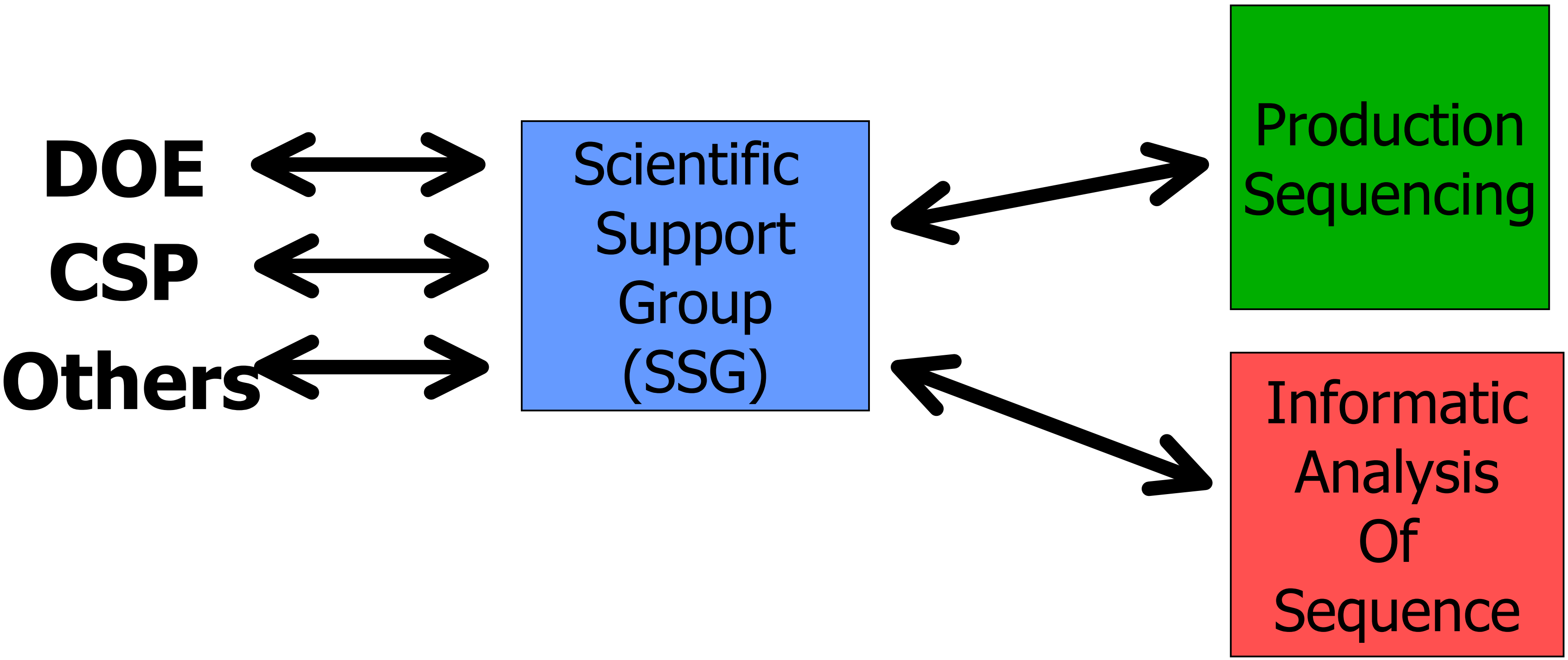


>70 Different "Organisms/DNA Sources"

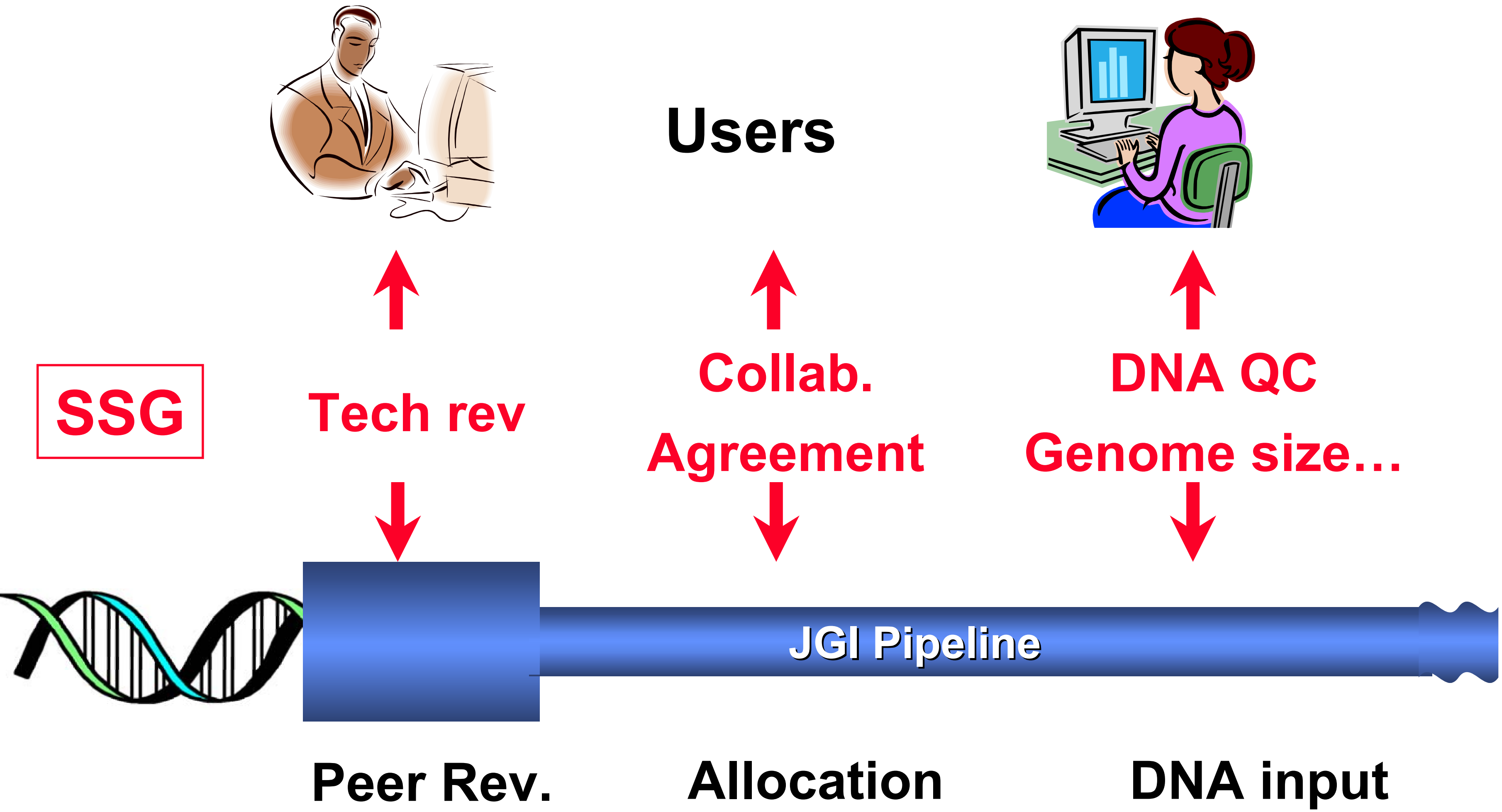
Involving >40 Different Collaborators

2004

Scientific Support Group (SSG) To provide support to "Users/Collaborators "


DOE
CSP
Others
Scientific Support Group (SSG)
Production Sequencing
Informatic Analysis Of Sequence

SSG facilitates work flow at all levels at JGI



Users

SSG

Tech rev

Collab.
Agreement

DNA QC
Genome size…

JGI Pipeline

Peer Rev.

Allocation

DNA input

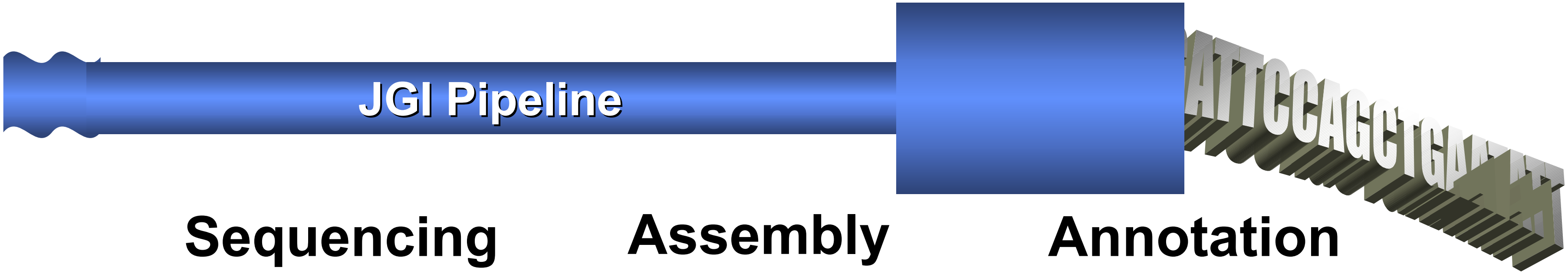
SSG the interface between "Users" and data

Users

SSG

Communication!

Specific User Needs



Sequencing

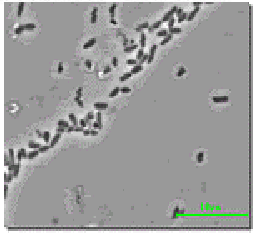
Assembly

Annotation

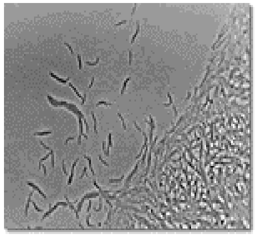
Goal for the JGI

To Operate as a Source of Genomic Infrastructure for American Science

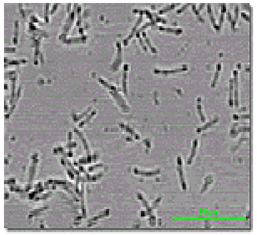
JGI 2001 Microbes



Burkholderia cepacia



Cytophaga hutchinsonii

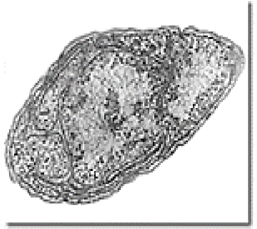


Desulfitobacterium halfniense

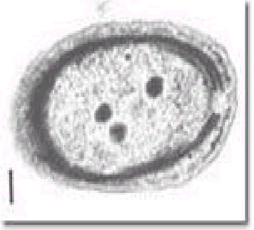
Enterococcus faecium

Ferroplasma acidarmanus

Magnetospirillum magnetotacticum



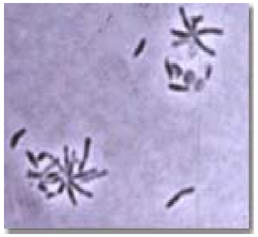
Nitrosomonas europaea



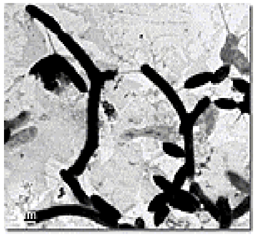
Prochlorococcus marinus

Pseudomonas fluorescens

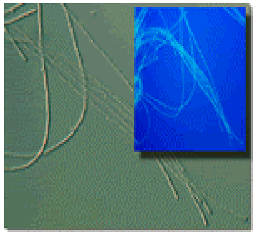
Rhodobacter sphaeroides



Rhodopseudomonas palustris



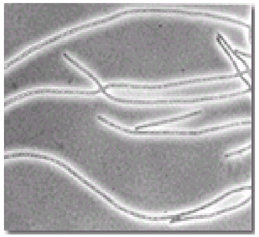
Sphingomonas aromaticivorans



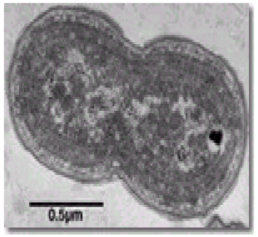
Thermomonospora fusca

Trichodesmium erythraeum

Xylella fastidiosa



Nostoc punctiforme



Marine synechococcus

Magnetococcus MC-1

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

More Microbes…

FY 2002

Lactic acid bacteria

Lactobacillus gasseri (Klaenhammer)
Lactobacillus casei (Broadbent/Steele)
Lactobacillus delbrueckii (Steele)
Lactococcus cremoris (Weimer)
Brevibacterium linens (Weimer)
Pediococcus pentosaceus (Broadbent)
Oenoccoccus oeni (Mills)
Leuconostoc mesneteroides (Breidt)
Streptococcus thermophilus (Hutkins)

Bifidobacterium longum (O'Sullivan)

Complex polysaccharide degradation

Clostridium thermocellum (Wu)
Microbulbifer degradans (Weiner)
(complements white rot fungus sequence)

Phototrophic bacteria

Rhodospirillium rubrum (Roberts)
(complements *Rhodopseudomonas palustris* and *Rhodobacter spheroides)*

Toxic waste degradation and microbial ecology

Desulfuromonas acetoxidans (Lovely)
Desulfovibrio desulfuricans
Geobacter metallicreducens (Loveley, Ciufo)
Dechloromonas aromatica
Ralstonia eutropha (Valenzuela)
Azotobacter vinelandi
Trichodesmium erythraeum

Microbes in extreme environments

Psychrobacter (Thomashow)
Exiguobacterium (Thomashow)
Methanococcoides burtonii (Sowers, Cavicchioli)

Infectious diseases of plants and animals

Erlichia chaffeensis (Yu)
Erlichia canis (Yu)
Streptococcus suis (Gottschalk)
Haemophilus somnus (Inzana)
Pseudomonas syringae (Lindow)
Agrobacterium tumefaciens

Anaerobic methane oxidizing consortium "ball of bugs" (DeLong, Monterey Bay)
one (or two?!) reverse methanogenic archaea in core plus sulfur reducing bacterium on surface

And More Microbes…

Single Microbes

Rubrobacter xylanophilus

Prochlorococcus isolate NATL2A

Kineococcus radiotolerans sp nov

Methylobacillus flagellatus, strain KT

Synechococcus elongates PCC7 942

Moorella thermoacetica ATCC39073

Anabaena variabilis ATCC 29413

Burkholderia complex (genomovar V)

Crocosphaera watsonii WH8501

Stramenopiles

Phytophthera ramorum UCD Pr4 – 2.46Mb sequence

Phytophthora sojae P6497 – 319.72Mb sequence

Microbial Consortia

Acid mine drainage from site in Iron Mountain

Chlorochromatium aggregatum

Fungus

Trichoderma reesei - 87.55Mb of Sequence Present
(Strain RUT-C30, ATCC56765)

Marine Algae

Emiliania huxleyi strain 1516

2004 DOE Microbe Projects

8 species of Chlorobia

Chlorobium limicola, DSMZ 245(T)

Chlorobium phaeobacteroides, MN1

Prosthecochloris spp.

Prosthecochloris aestuarii, SK413/DSMZ 271(t)

Chlorobium vibrioforme f. thiosulfatophilum, DSMZ 265(T)

Chlorobium phaeobacteroides, DSMZ 266(T)

Pelodictyon phaeoclathratiforme, BU-1 (DSMZ 5477(T))

Pelodictyon luteolum, DSMZ 273(T)

Model Syntrophic Consortium:

Syntrophobacter fumaroxidans, MPOB

Syntrophomonas wolfei, Göttingen (DSM 2245B)

Methanospirillum hungateii, JF1

Facultative Metal-reducing Gamma proteobacteria

Shewanella putrifaciens, CN-32

Shewanella sp., PV-4

Shewanella amazonesis

Shewanella baltica, OS1155

Shewanella frigidimarina, NCMB400

Shewanella denitrificans,OS ***217***

Shewanella putrifaciens, 200

five bacteria involved in nitrification

Nirosomonas eutropha C71

Nitrosospira multiformis Surinam

Nitrosomonas oceani

*Nitrobacter winogradskyi, Nb-*255

Nitrobacter hamburgensis

Single microbes

Synthophobacter fumaroxidans

Synthophus acidotrophicus

Arthrobacter aurescens, TC1

Thermoanaerobacter ethanolicus, X514

Frankia sp., EAN1pec

Frankia sp., CcI3

Anaeromyxobacter dehalogenans, 2CP-C

Nocardioides sp., JS614

Deinococcus geothermalis, DSM11300

Chromohalobacter salexigens, DSM3043

Clostridium beijerincki, NCIMB 8052

Acidobacterium sp., Ellin6076

Clostridium phytofermentans

Arthrobacter sp., FB24

Thiomicrospira crunogena

Thiomicrospira denitrificans

Sphingopyxis alaskensis, RB2256

Alkaliphillus metalliredigenes

Jannaschina sp.CCS1

Roseobacter sp., TM1040

Paracoccus denitrificans, 1222

Thiobacillus denitrificans, ATCC 23644

b-proteobacterium sp., JS666

Eukaryotes

Glomus intraradices

Laccaria bicolor

Pichia stipitis, CBS 6054

Pichia mRNA for cDNA libraries

Communities:

200 BACs from anaerobic bioreactor granules

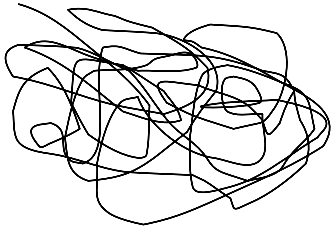
acid mine drainage community

Picoplankton BACS from HOTS site

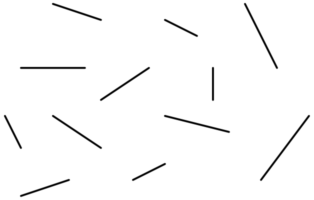
Boiling thermal pool

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

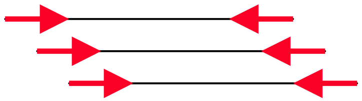
Genome Sequencing



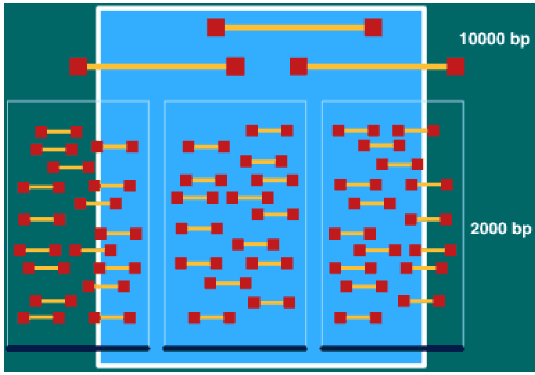
Start with genomic DNA



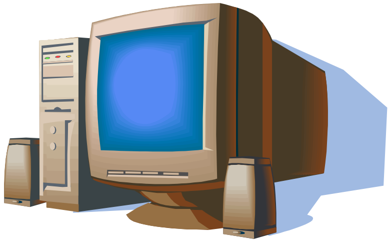
Make sheared fragments



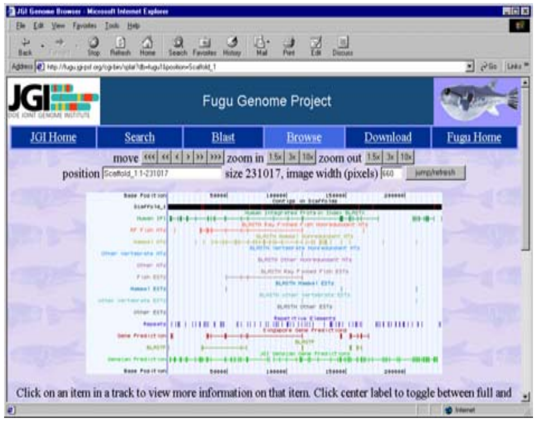
Sequence both ends of fragments



Reconstruct genome computationally



High-throughput computational analysis


JGI Genome Browser - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://fugu.jgi-psf.org/cgi-bin/splat?db=fugu1&position=Scaffold_1
Go Links
JGI DOE JOINT GENOME INSTITUTE
JGI Home Search Blast Browse Download Fugu Home
move zoom in 1.5x 3x 10x zoom out 1.5x 3x 10x
jump/refresh
Internet

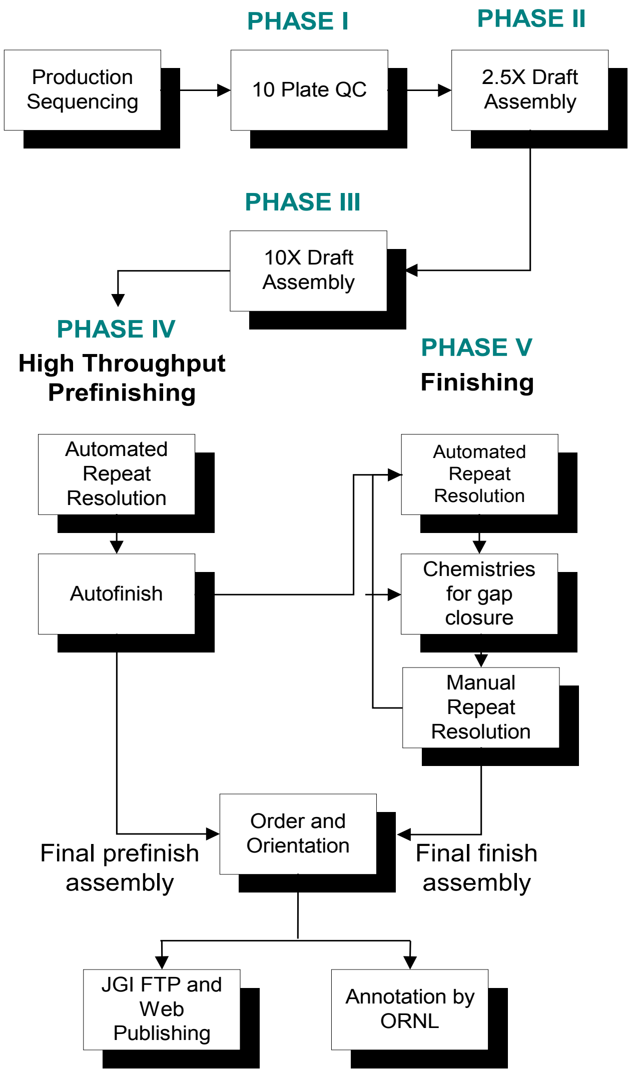
Fugu Genome Project

position Scaffold_1:1-231017 size 231017, image width (pixels) 660

Click on an item in a track to view more information on that item. Click center label to toggle between full and

Provide genome and tools to community

Life Cycle of a Microbe



PHASE I - *10 Plate QC*
10 plates are sequenced and QC performed to look for contamination.

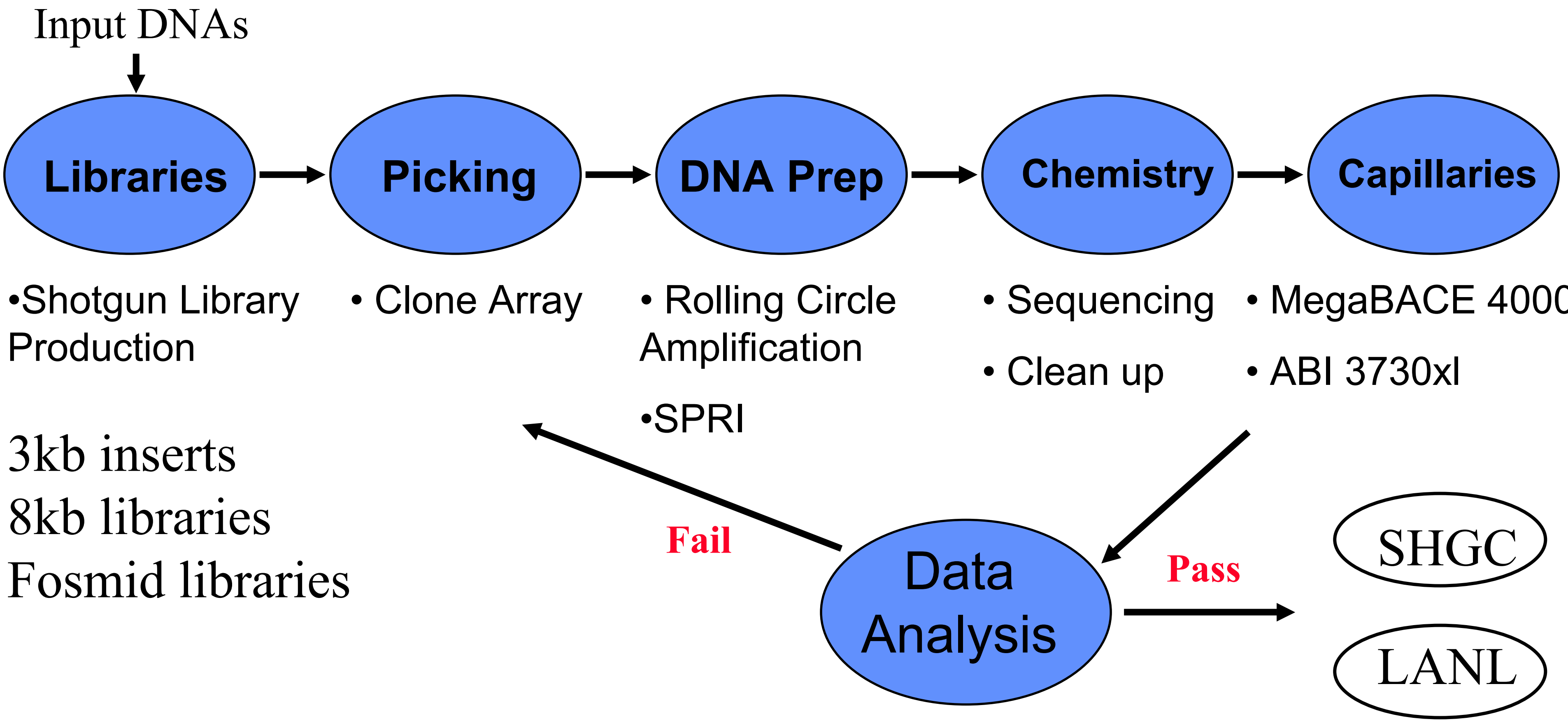
PHASE II - *2.5 Draft Assembly*
Draft sequence is peformed to 2.5X coverage. QC is performed to look for contamination.

PHASE III - *10X Draft Assembly*
Draft sequence is performed to 10X coverage. Final draft assembly is done and flagged for Finishing.

PHASE IV - *High Througput Prefinishing*
Semi-automated Prefinishing is accomplished by resolving misassemblies and closing gaps <3kb through Autofinish. Once done, the assembly is order and oriented and the results are sent to ORNL for annotation and posted on the JGI FTP site for public access.

PHASE V - *Finishing*
Assembled contigs from Phase IV are analyzed for gaps and misassemblies. Automated repeat resolution, manual repeat resolution and primer walking are performed in an iterative process to resolve misassembled regions and close remaining gaps. The final assembly is order and oriented and the results are sent to ORNL for annotation and posted on the JGI FTP site for public access.

Current Production Pipeline



Input DNAs

Libraries → Picking → DNA Prep → Chemistry → Capillaries

Libraries:

- •Shotgun Library Production

Picking:

- • Clone Array

DNA Prep:

- • Rolling Circle Amplification
- •SPRI

Chemistry:

- • Sequencing
- • Clean up

Capillaries:

- • MegaBACE 4000
- • ABI 3730xl

3kb inserts
8kb libraries
Fosmid libraries

Data Analysis

Fail

Pass

SHGC

LANL

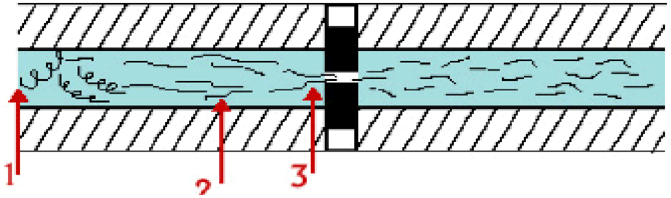
•Order and Orientation of contigs

•Assembly verification

Library Construction: Phase I

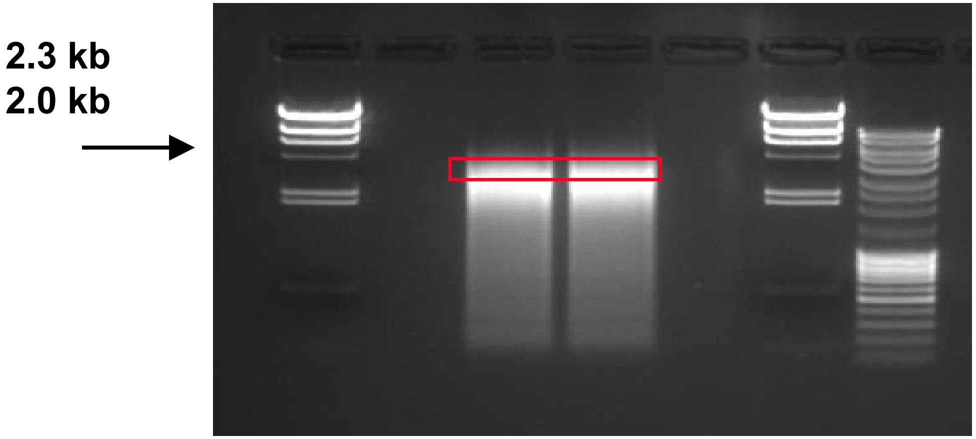
Multiple size insert libraries for each organism and sequence them to a specific depth.

4x Sequence of 2-4kbs – Small Insert
4x Sequence of 8-10kbs – Medium Insert
10x Clone coverage of Fosmid Ends



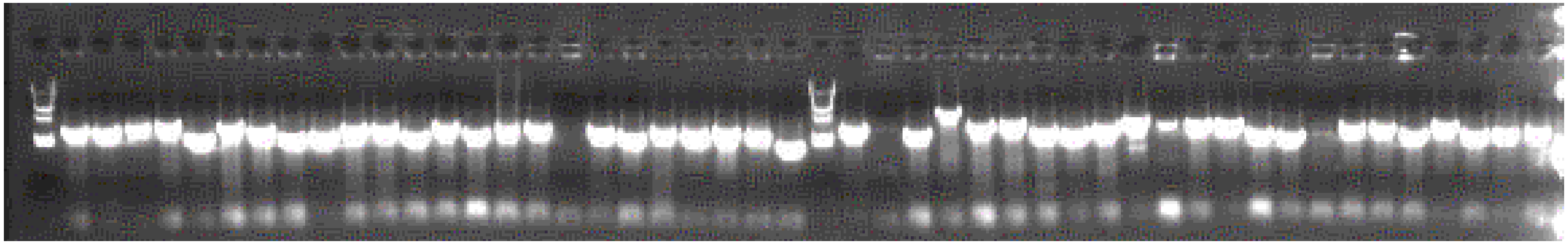
GeneMachines Hydrashear

Sheared Genomic DNA



Ensuring the Quality of Libraries

PCR QC



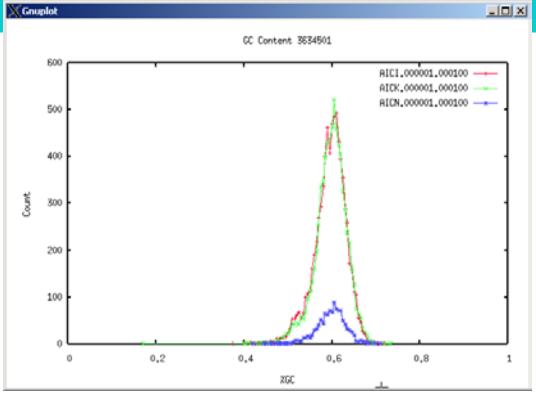
Sequence QC

First Pass Sequencing: Verify that the DNA is from the correct organism and check the following: insert size, % vector, and contamination

2x QC: Verify that there is no cloning bias within the library – both small, medium

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Library and Production QC



10 Plate QC

Project: 3634501

Library: AICI, AICK

Organism: Roseobacter sp. TM1040

Lineage: cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Roseobacter

Vector: pUC, pMCL200

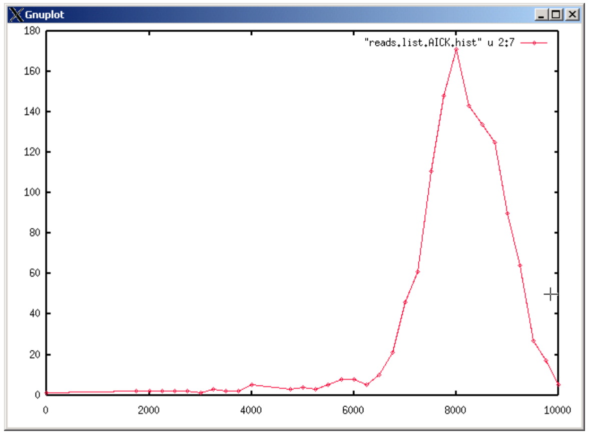
Insert Size: 3kb,8kb

Shearing Operator: CC

Date: Jan 12, 2004

QC By: HK

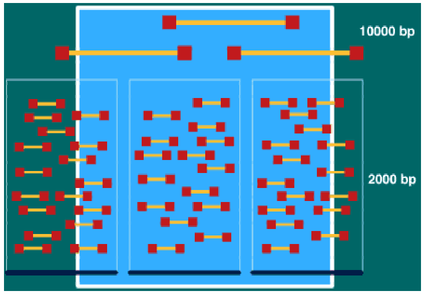
- Contamination Check
  - -Known JGI contaminants
  - -Vector
  - -GC content
  - -Correct microbe
- Library QC
  - -Read distribution
  - -Insert size distribution
  - -Compare to ideal assembly



Assembled: 3941181 (trimmed)
Phrap: 3489815
DB: 9000000
Current Depth Estimate : 4.180703 +/- 0.856594
N50 actual
About half the reads are in 105 contigs containing at least 37 reads each
N 50 analytical
N50 (analytic): About half the reads will be in 130 contigs containing at least 38 reads each (3.9 MB)
N50 (analytic): About half the reads will be in 756 contigs containing at least 7 reads each (9.0 MB)

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Assembly, Analysis and Annotation



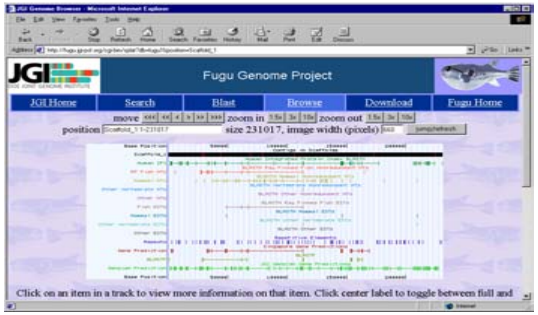
Reconstruct genome computationally

Phase III: Final Draft Assembly
All libraries sequenced to completion, data assembled and verified.

Assembly made available to the collaborator and sent to ORNL for annotation

Project transferred to JGI Microbial Group for automated Pre-finishing

Phase IV-Pre-finishing Complete
Project now “In Finishing”

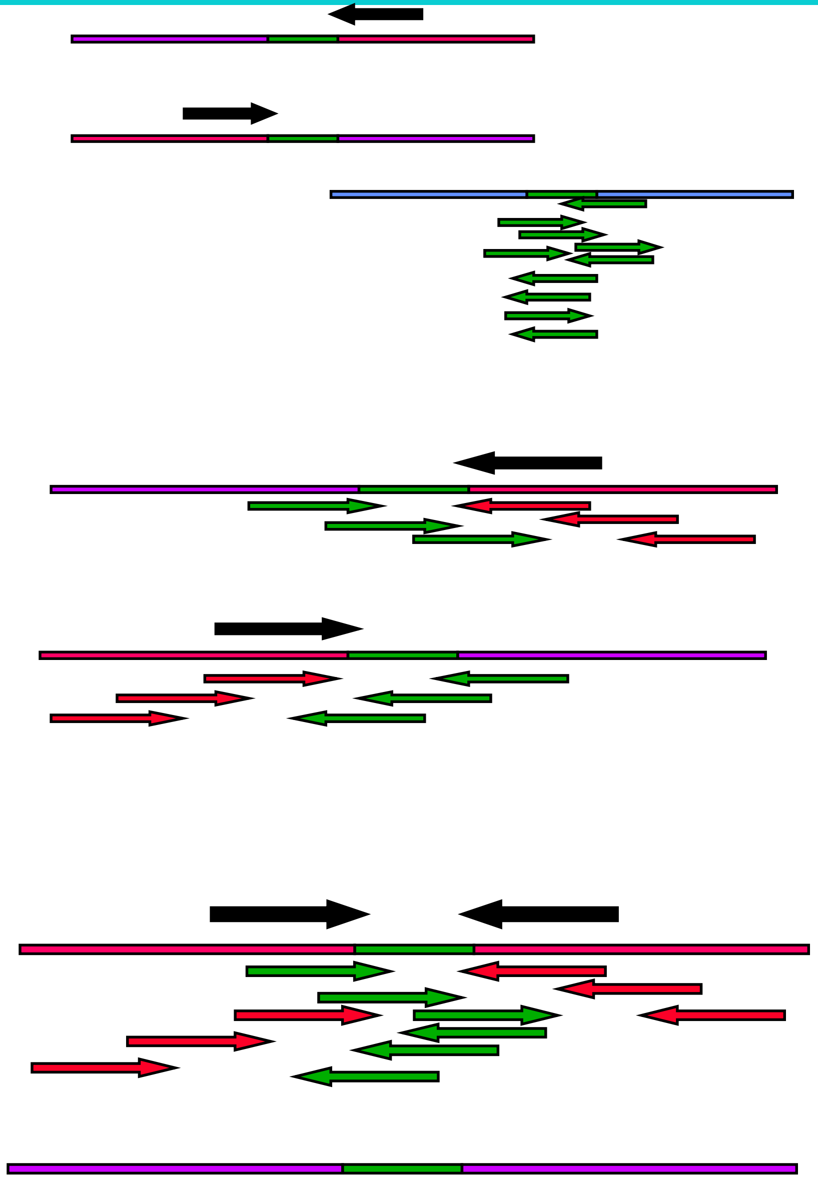


Click on an item in a track to view more information on that item. Click center label to toggle between full and

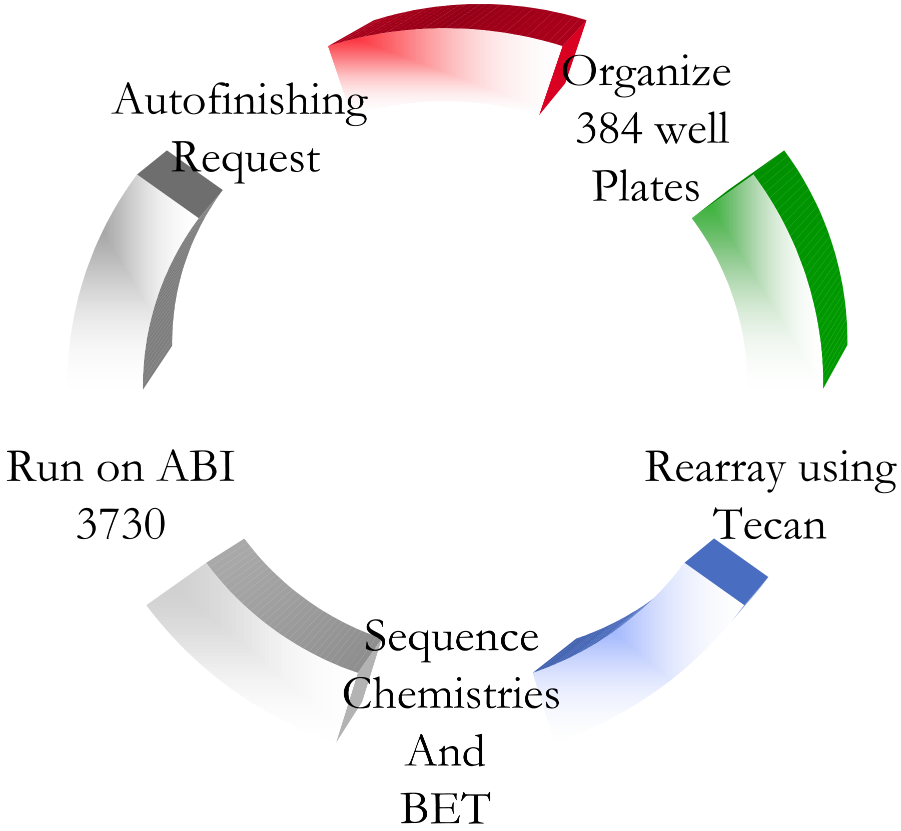
Provide genome and tools to community

Automated Repeat Resolution

- **Identify Repeats**
  - **Two types**
    - **Transposases (IS elements)**
    - **Operons (16s), phage**
- **Automated subassemblies**
  - **Group unique reads and sister pairs**
  - **Local assembly**
  - **Incorporation of new consensus seqeuence in repeat region**



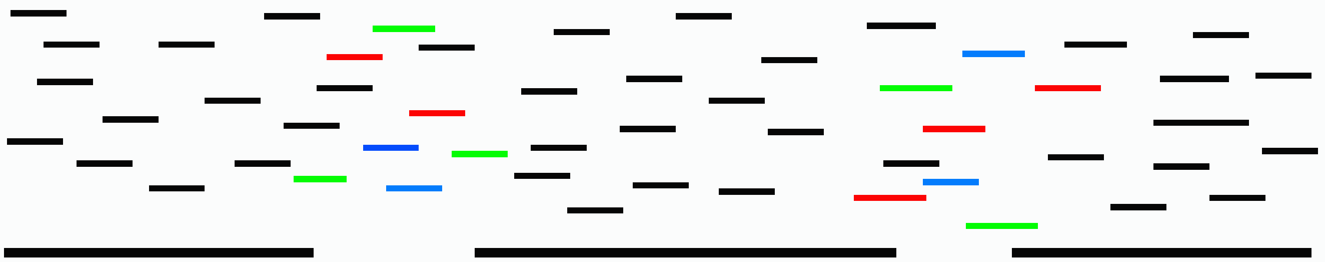
Overview of Autofinishing



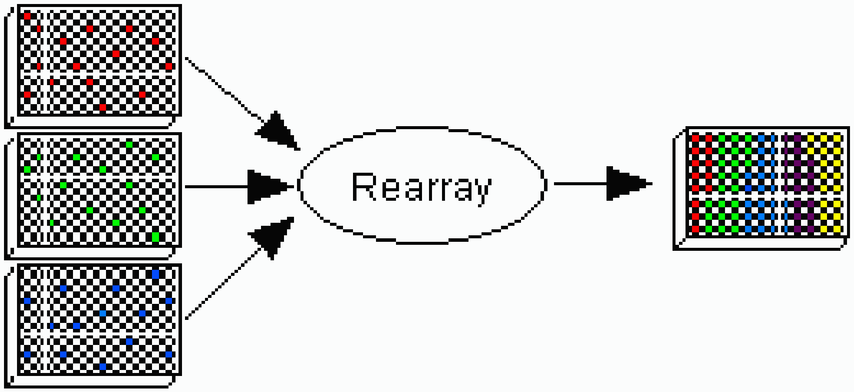
- Two or more rounds of autofinishing may be required before a genome is ready for finishing
- Every microbe is different and may require different/multiple types of chemistries

Overview of Rearray

DNA fragments are selected for re-sequence to close gaps between contigs.

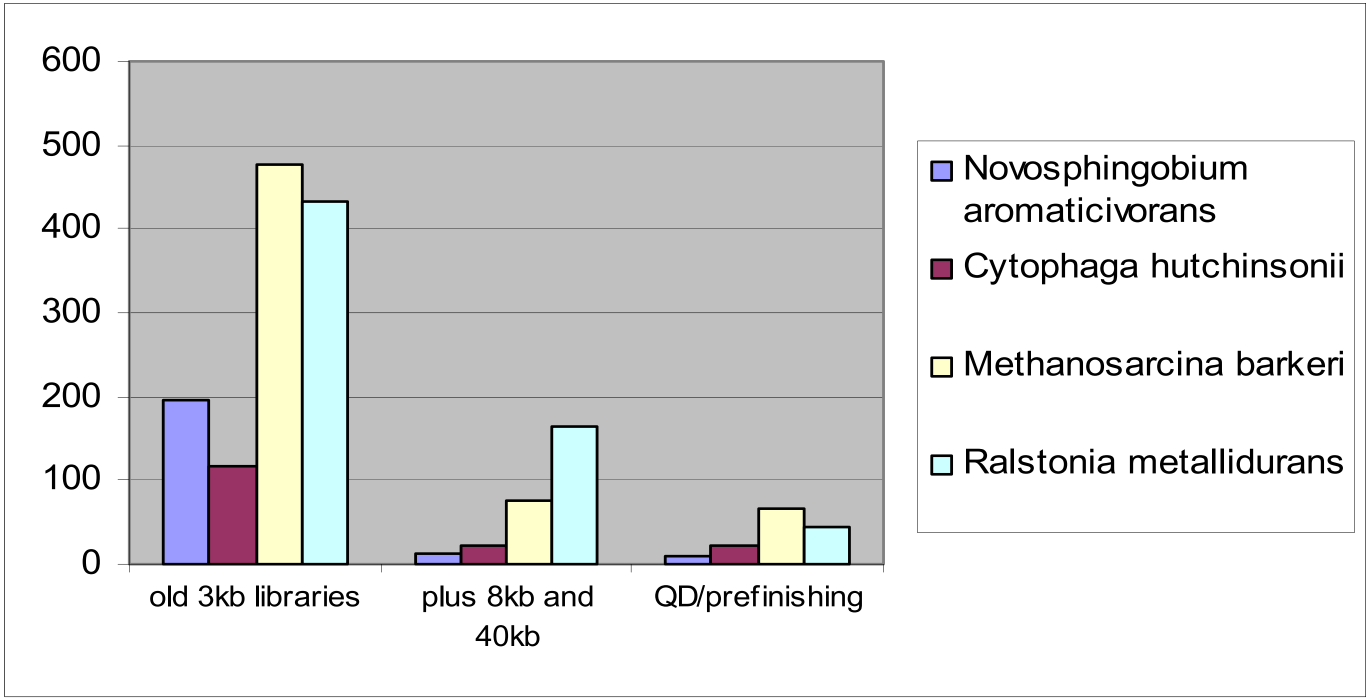


Selected samples are transferred from library plates into one plate for chemical processing.



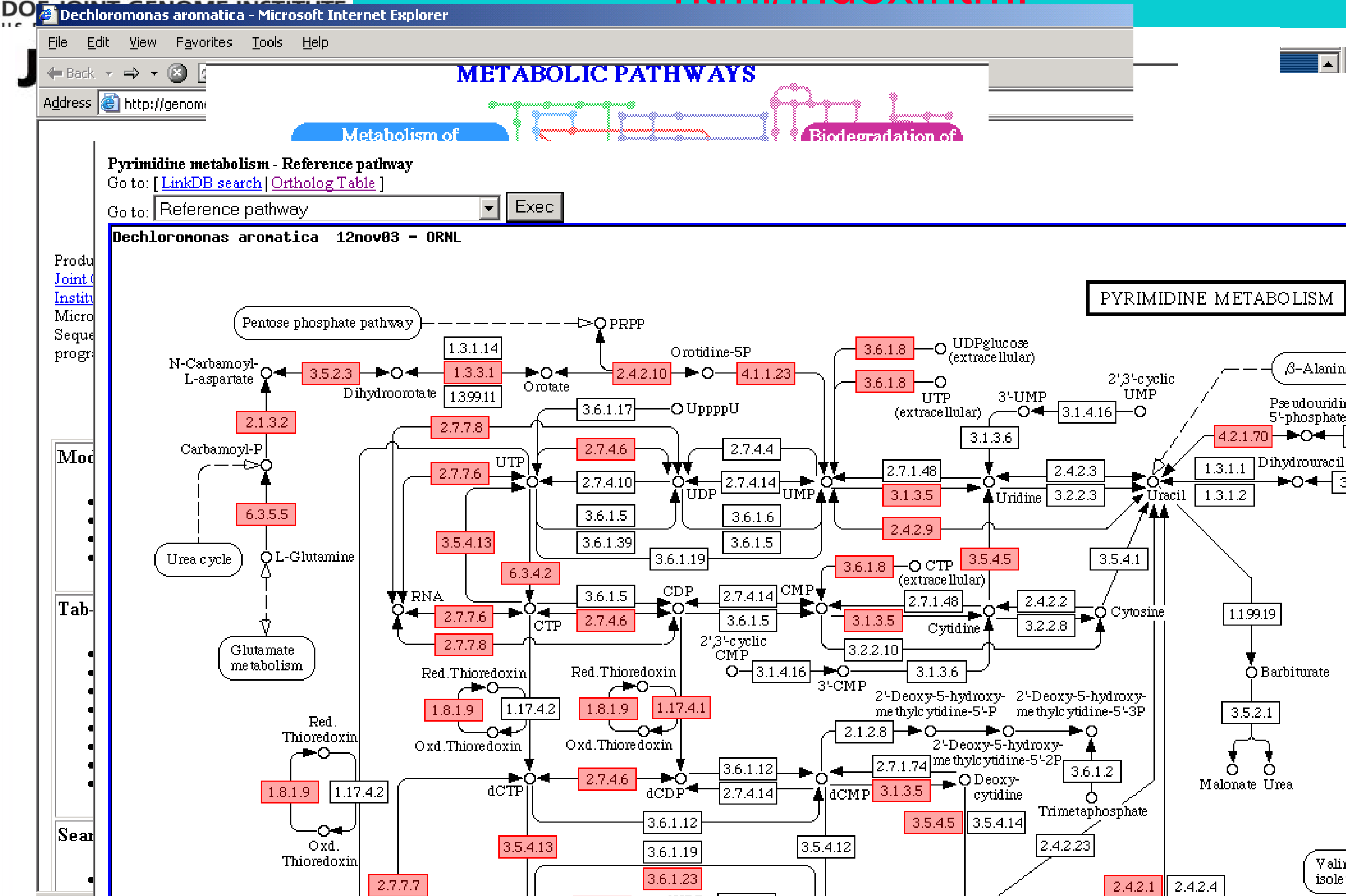
Genome Quality Improvements

	old 3kb libraries		plus 8kb and 40kb		QD/prefinishing	
	Major Contigs	Genome size (MB)	Major Contigs	Genome size (MB)	Major Contigs	Genome size (MB)
Novosphingobium aromaticivorans	197	4.17	13	4.21	9	4.215
Cytophaga hutchinsonii	118	4.36	23	4.41	22	4.41
Methanosarcina barkeri	478	3.88	77	4.83	67	4.84
Ralstonia metallidurans	432	NA	165	6.83	45	6.83



DOE JOINT GENOME INSTITUTE

http://www.jgi.doe.gov/JGI_microbial/html/index.html



Dechloromonas aromatica - Microsoft Internet Explorer

Pyrimidine metabolism - Reference pathway

Go to: [LinkDB search | Ortholog Table]

Go to: Reference pathway Exec

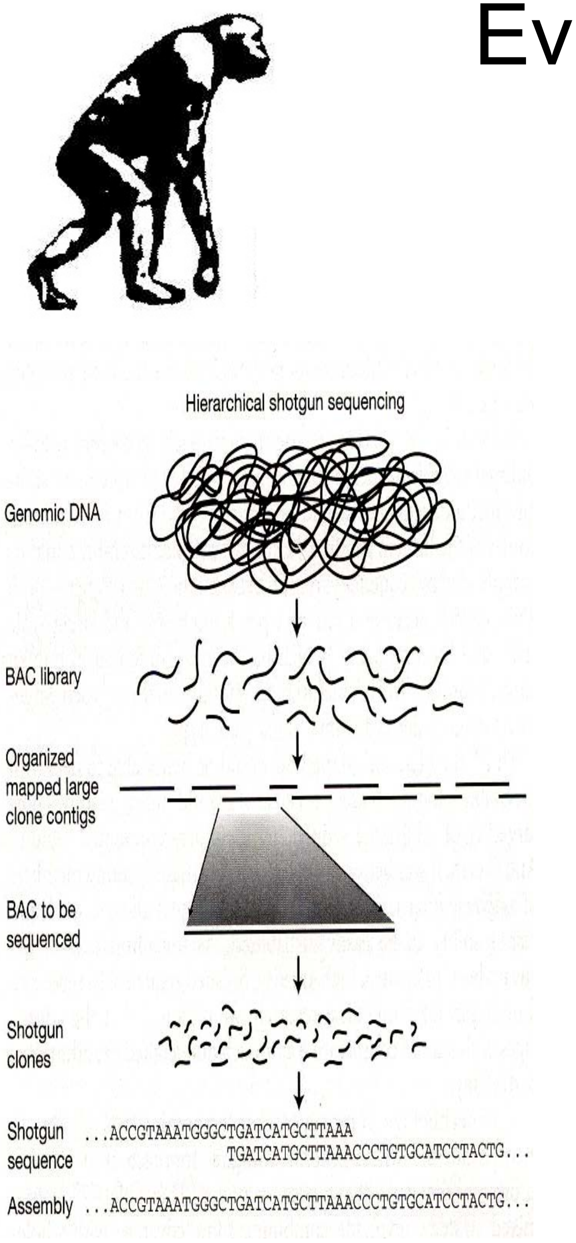
Dechloromonas aromatica 12nov03 - ORNL

JGI Capacity

- The current JGI throughput is ~2.0-2.5 billion bases per month
- In theory, JGI could sequence >400 microbes per year
- In practice, this would be very difficult to achieve
- JGI could reasonably sequence ~ 100-200 microbes per year
- This throughput depends on receiving high-quality DNA from the collaborators

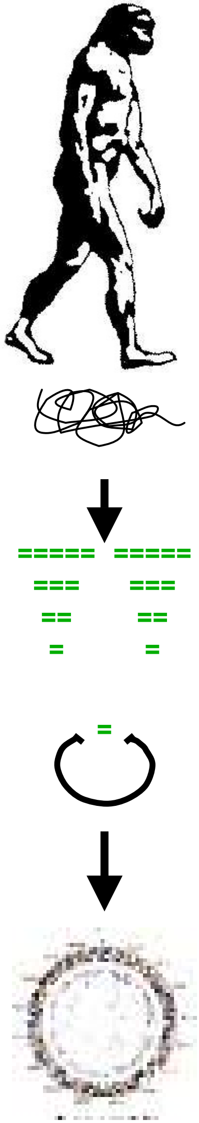
- This is the capacity for single isolates

Evolution of Sequencing

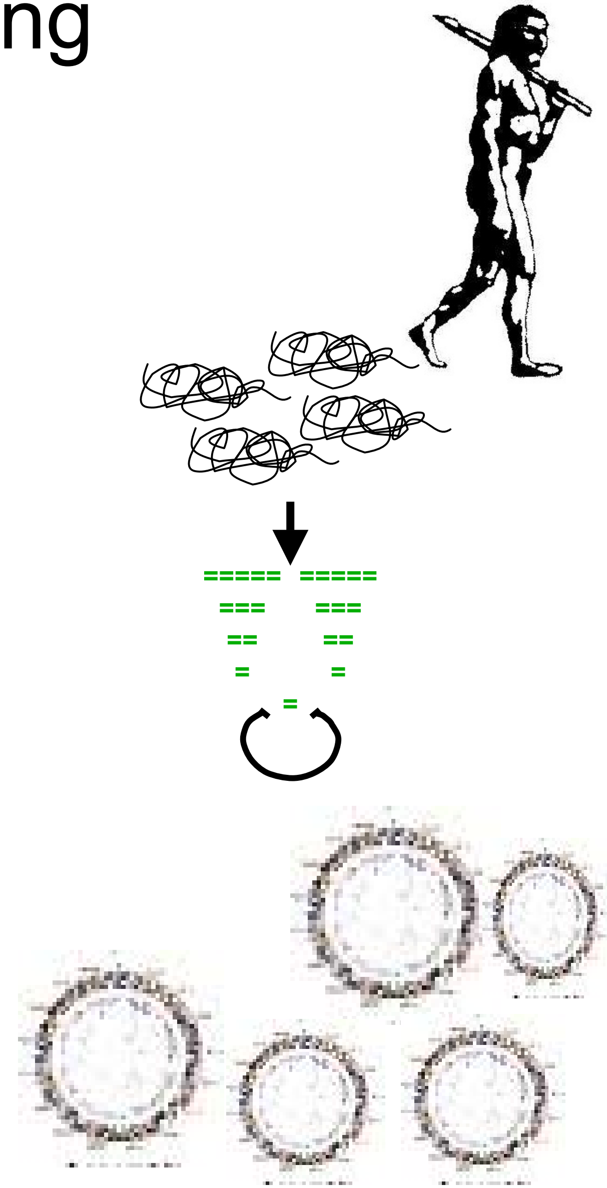

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

(Nature, 2001. 409:p863)

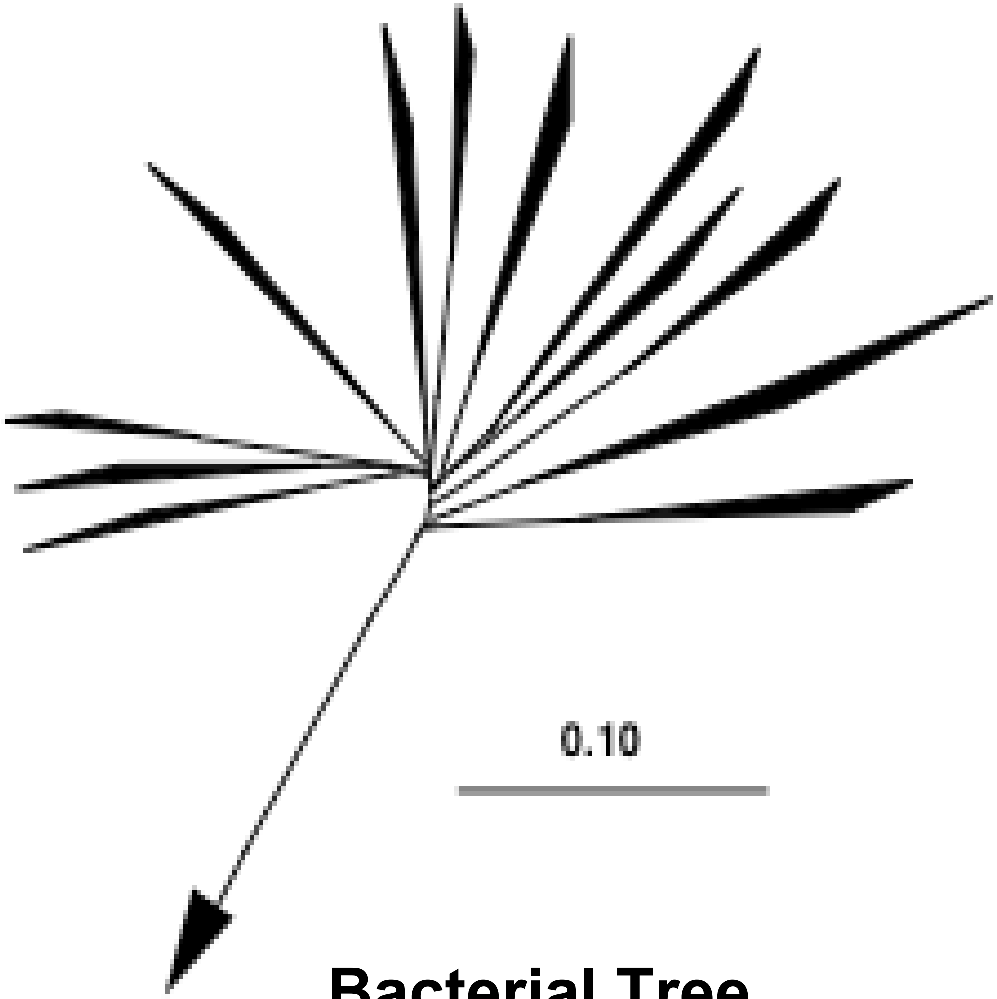
BAC Mapping



Whole Genome Shotgun



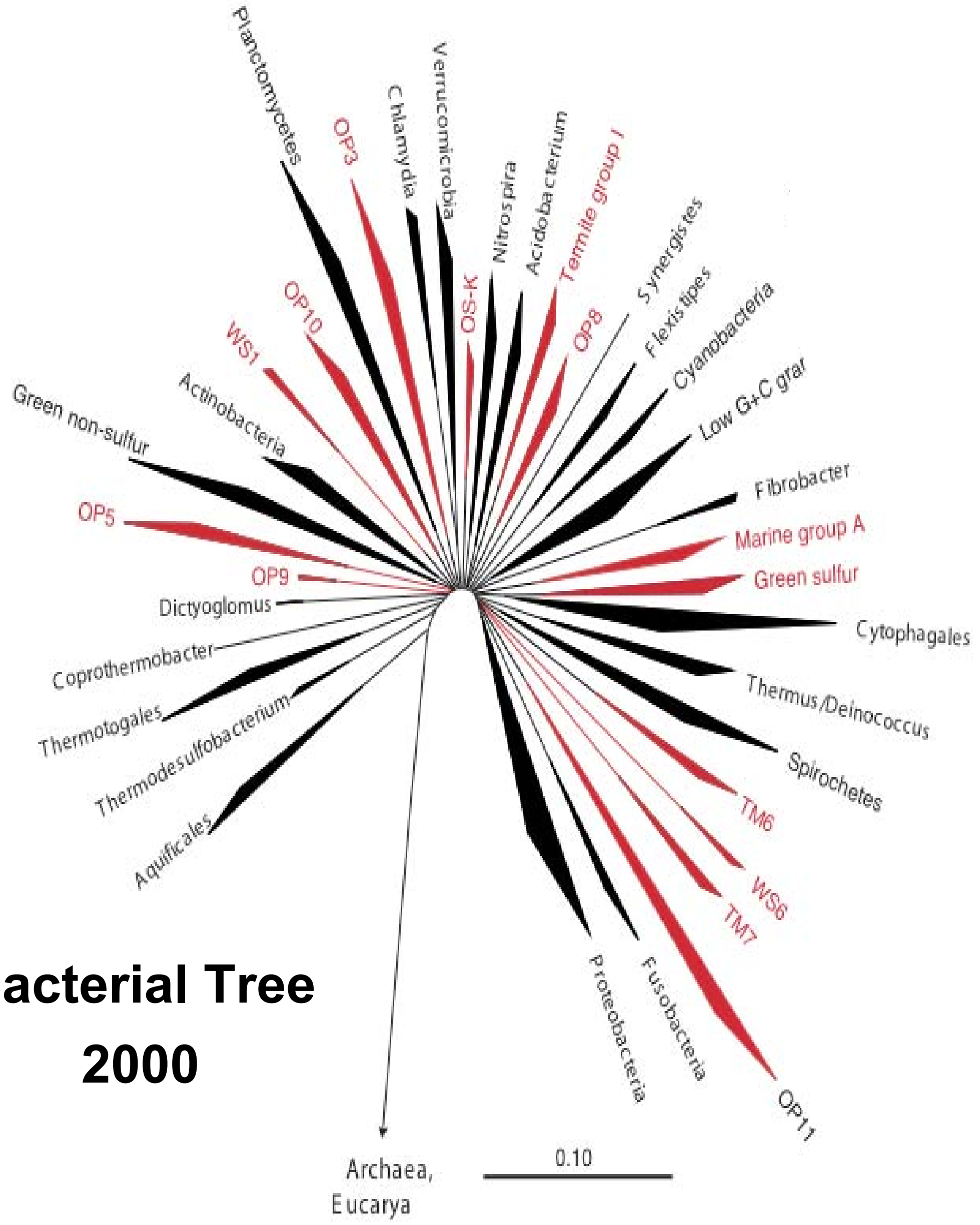
Environmental Metagenomics



Bacterial Tree

1987

Red = known only from Sequence



.987

DeLong and Pace, 2000

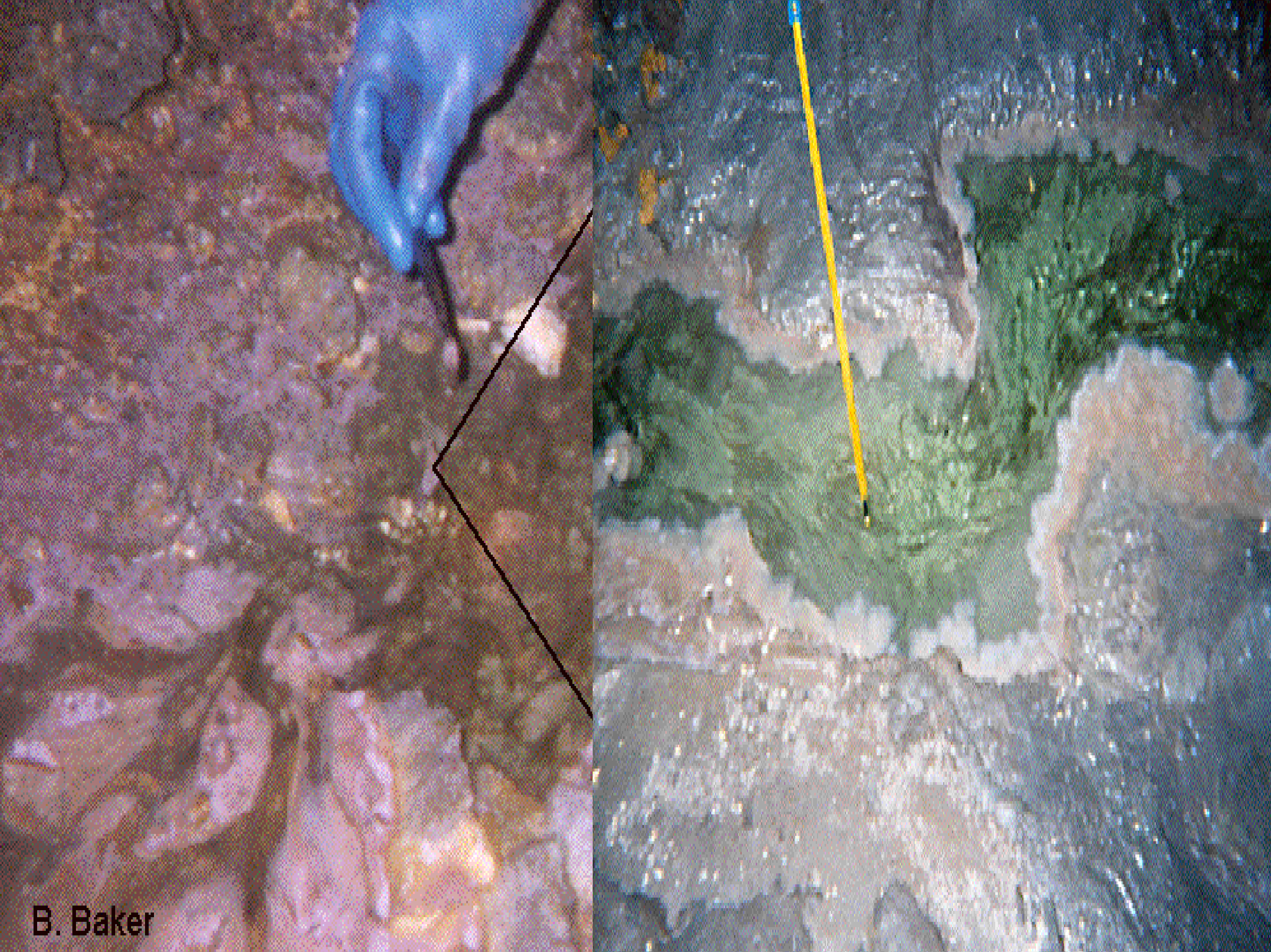
Bacterial Tree 2000

What Environments to Study?

Acid Mine Drainage Site
Iron Mountain, CA

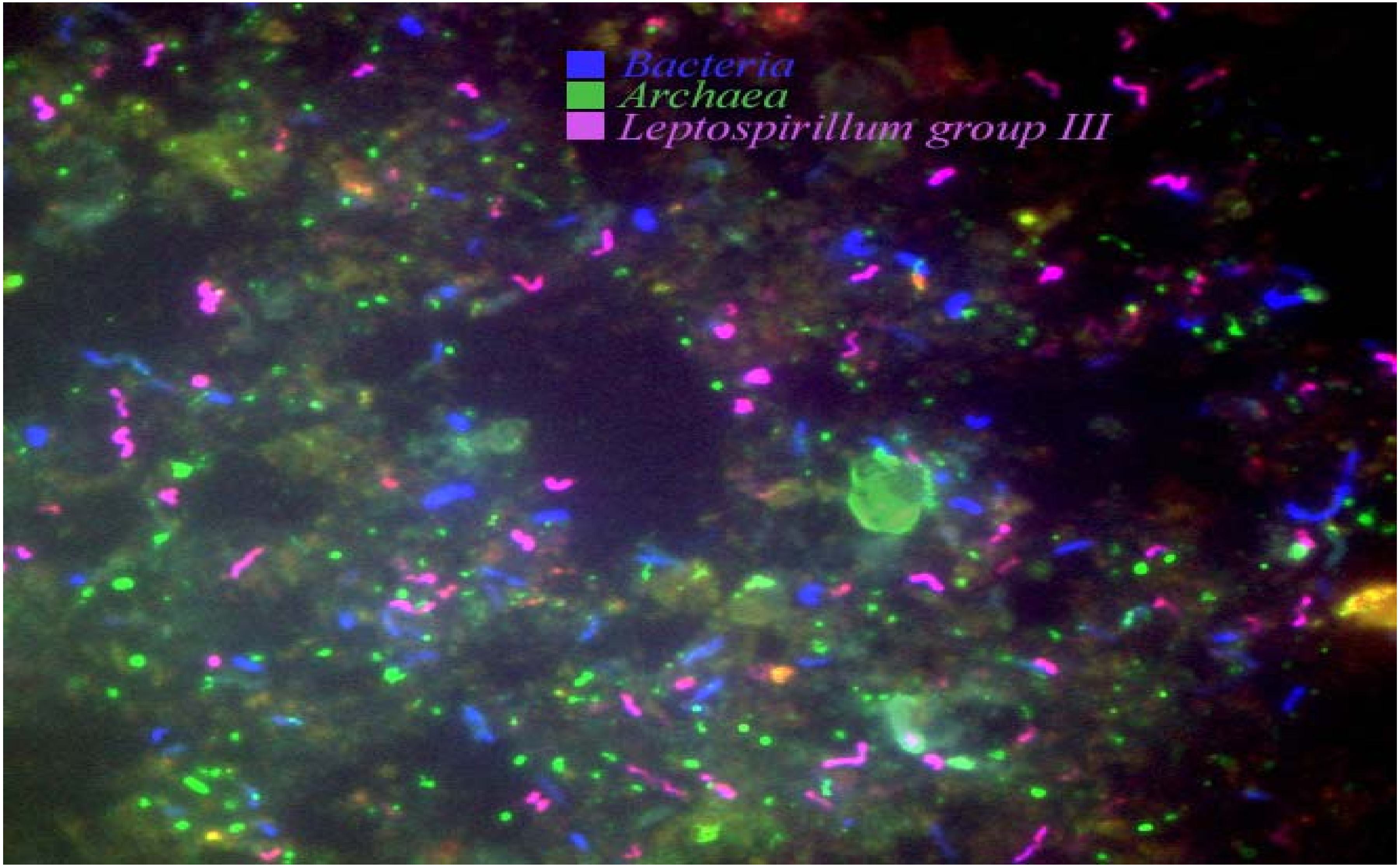
Jill Banfield et al.
UC Berkeley

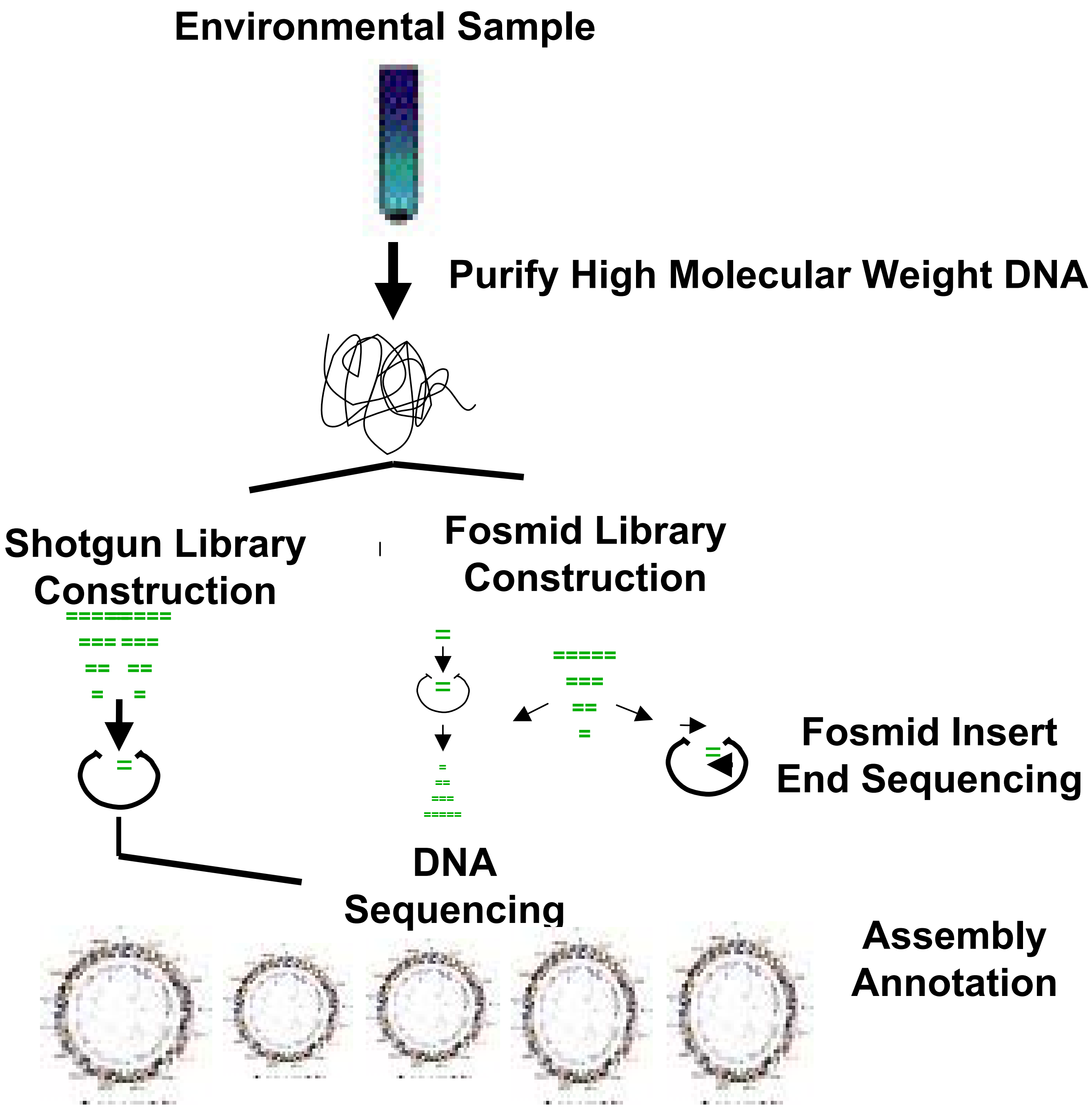
Superfund site Discharging >1 ton of toxic metals/day (pH <1) FeS_2



B. Baker

Iron Mtn biofilm





Environmental Sample

Purify High Molecular Weight DNA

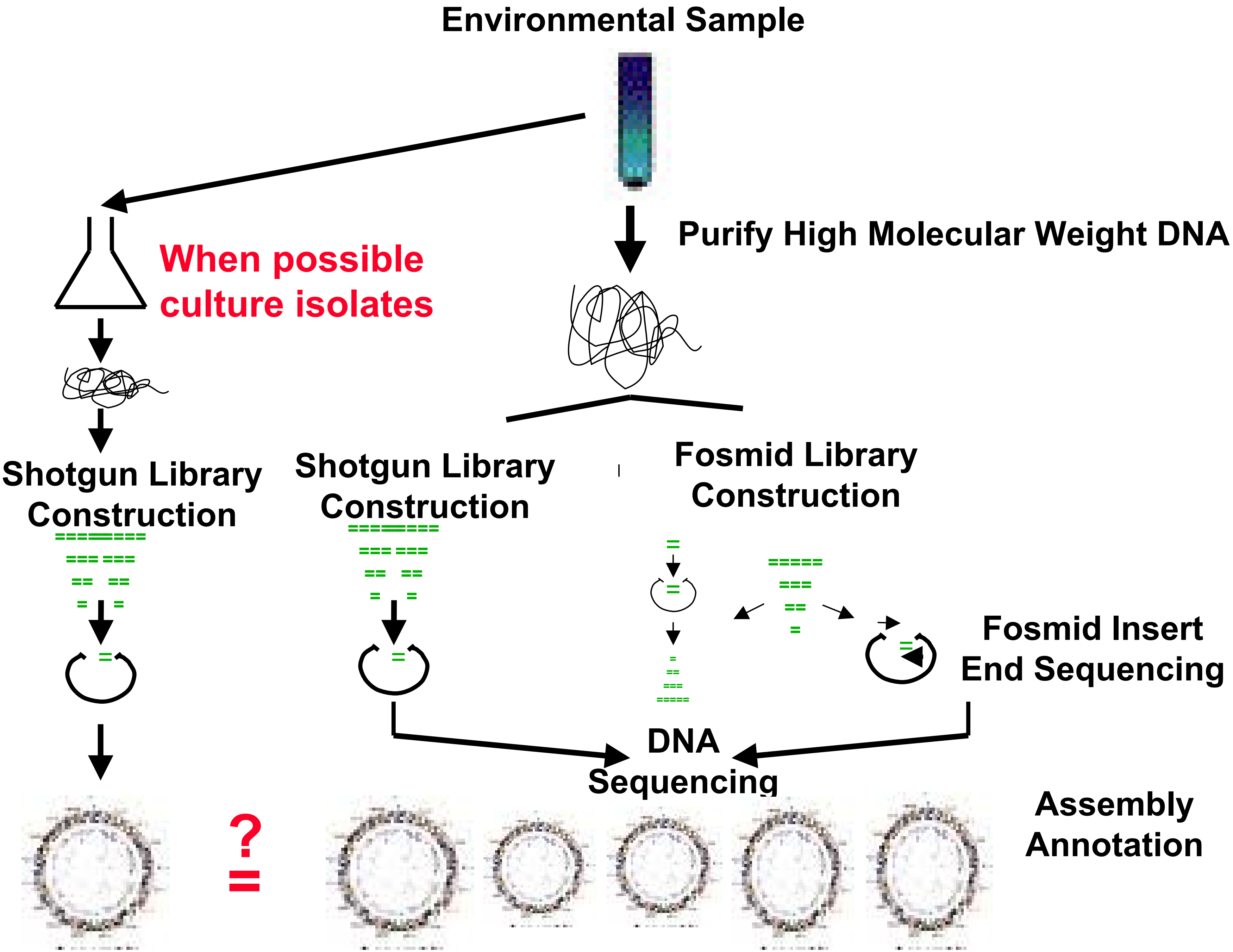
Shotgun Library Construction

Fosmid Library Construction

Fosmid Insert End Sequencing

DNA Sequencing

Assembly Annotation



Environmental Sample

Purify High Molecular Weight DNA

When possible
culture isolates

Shotgun Library
Construction

Shotgun Library
Construction

Fosmid Library
Construction

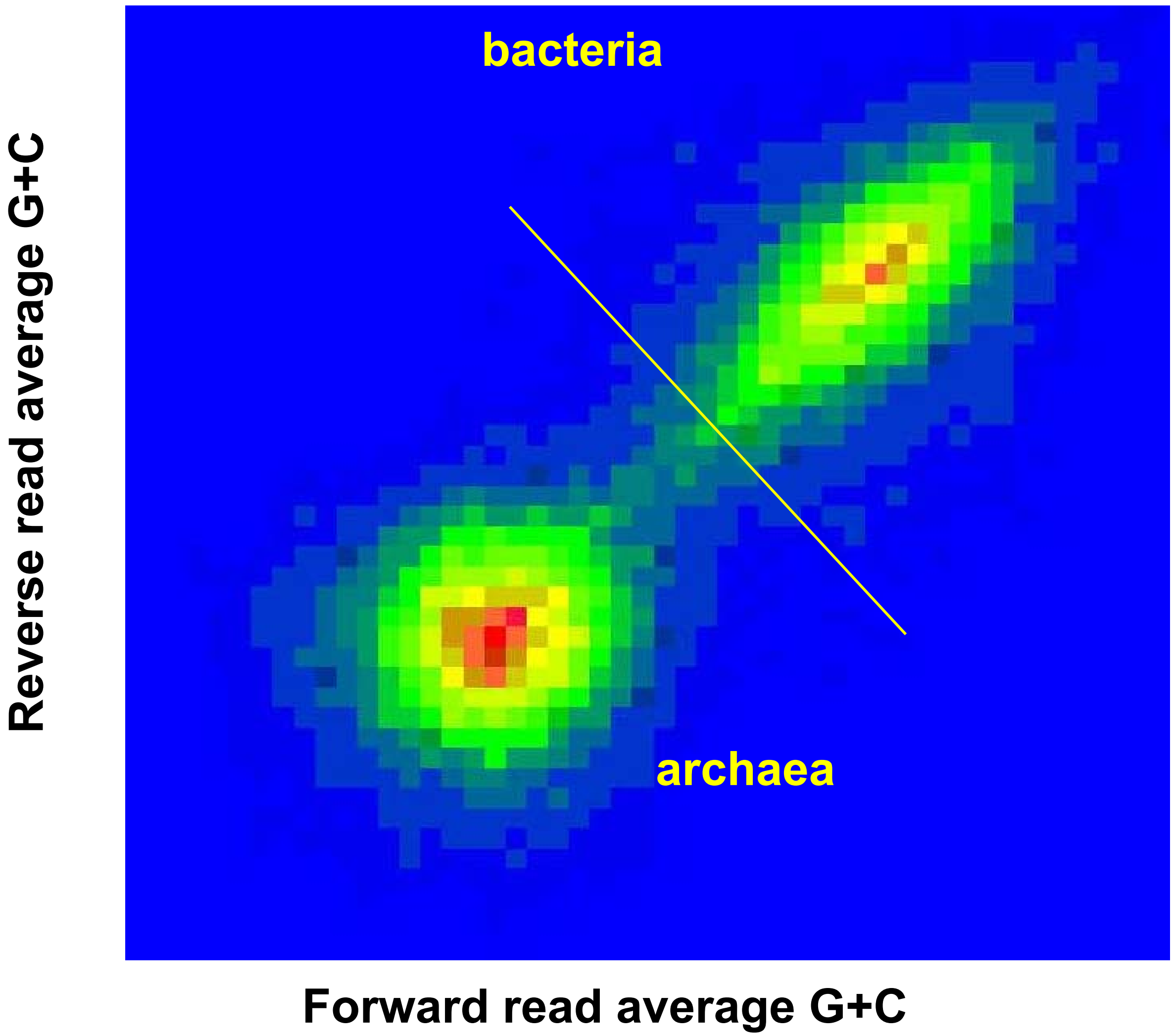
Fosmid Insert
End Sequencing

DNA
Sequencing

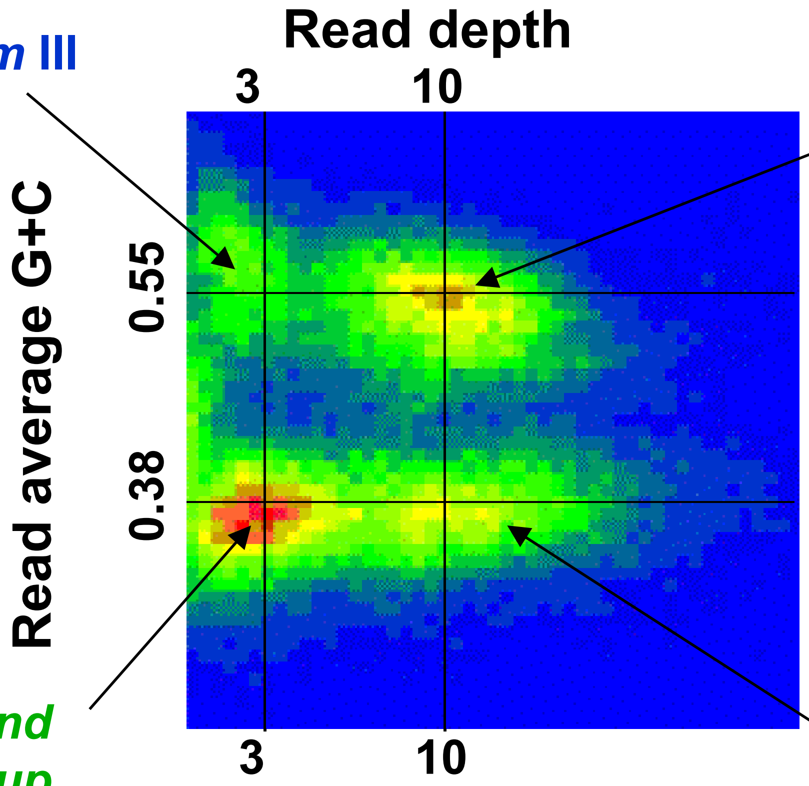
Assembly
Annotation

?
=

GC content separates into two components



Iron Mountain Microbial Community Includes Both Bacterial and Archael Groups



***Leptospirillum* III**

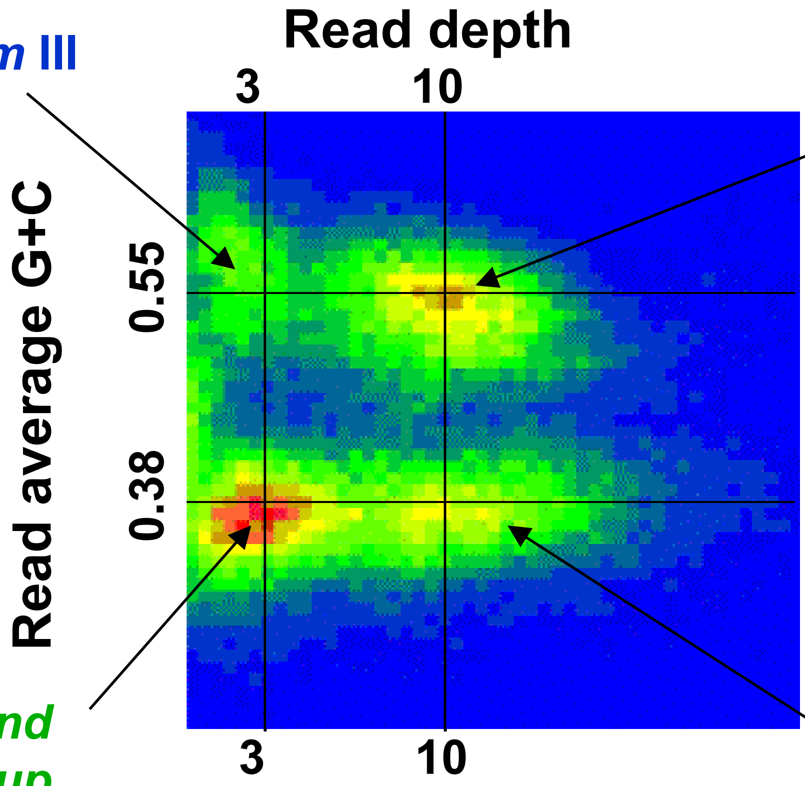
***Leptospirillum* II**

Bacteria

Archaea

Ferroplasma I, and
"G-Plasma" group

Ferroplasma II



Leptospirillum III

Leptospirillum II

Bacteria

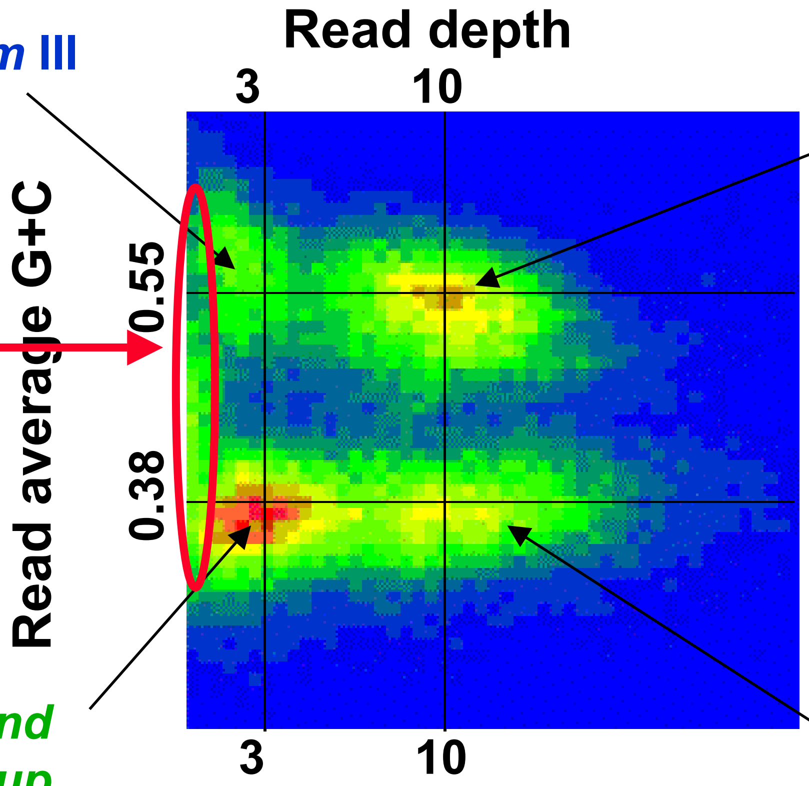
Archaea

Cultured and Sequenced

Ferroplasma I, and
“G-Plasma” group

Ferroplasma II

Leptospirillum III



Leptospirillum II

Bacteria

Archaea

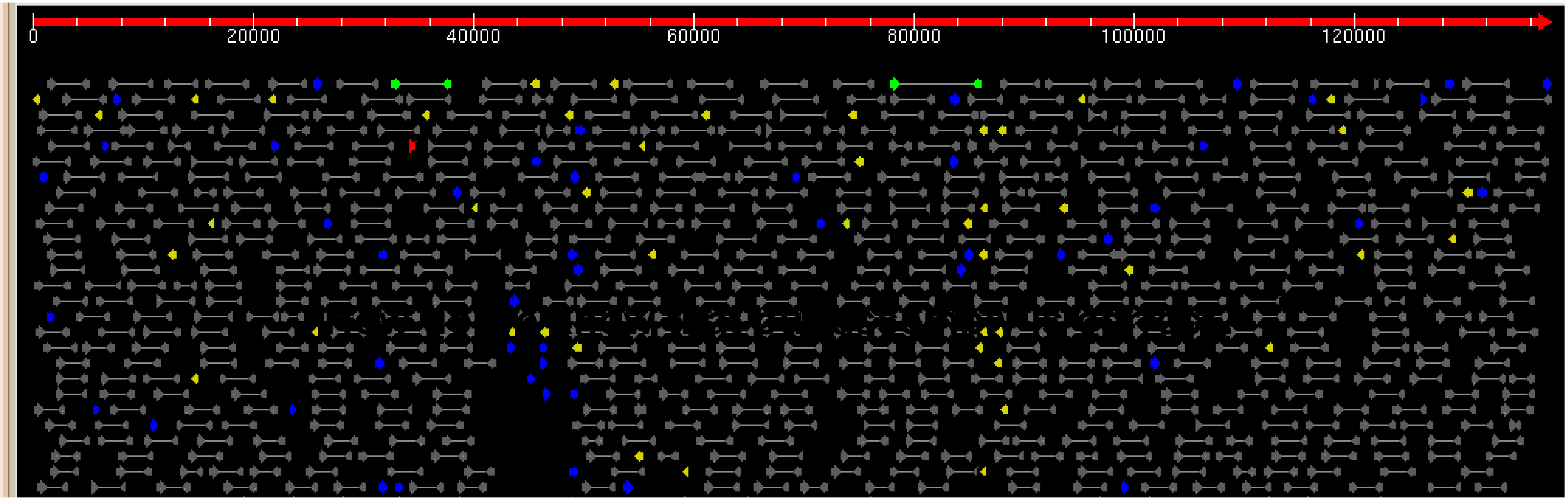
Other sampled genomes at low depth (including eukaryotes) 15% of reads

Ferroplasma I, and
"G-Plasma" group

Ferroplasma II

How do we know that our assembly is correct?

Check pair ends against scaffold



At the gross level: check pairs (expect few % due to failing/chimeric clones)

Align all reads back against assembled scaffolds

scaffolds end where there is no clone coverage in 3kb plasmids

Identifies potentially repetitive areas and/or rearrangements

What does it mean to assemble a community genome?

What is a “species” in the environment?

- In Metagenomics, Each Sequencing Read

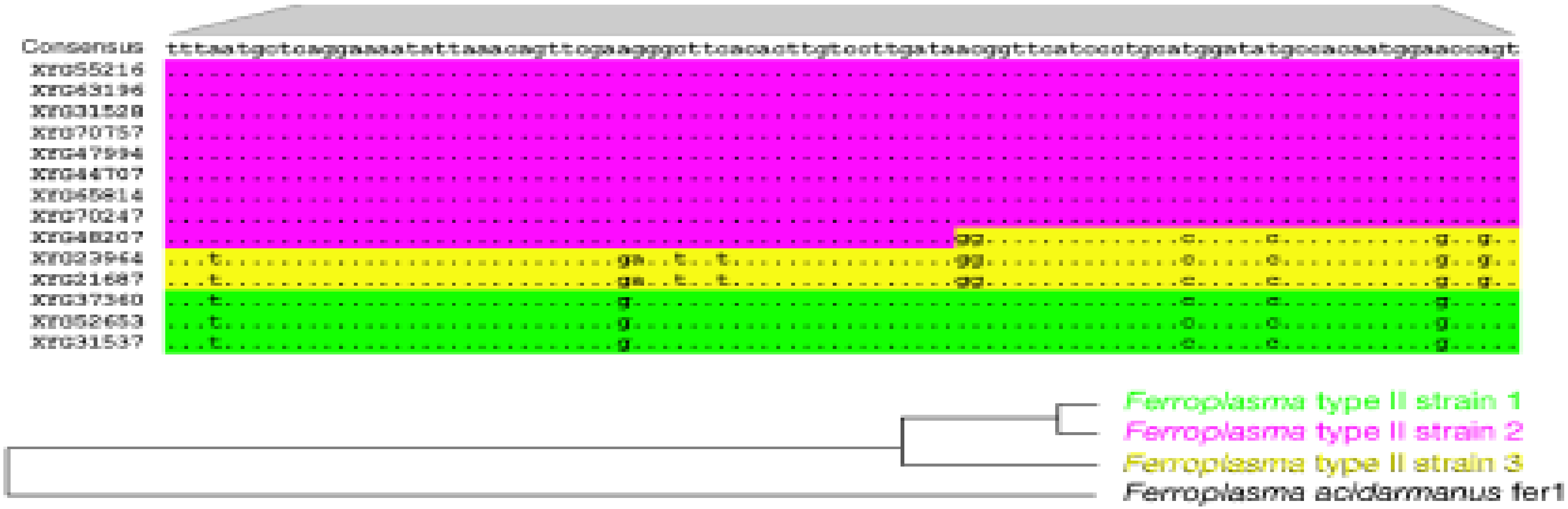
  is a Different Organism

Nucleotide Polymorphism Rates

•*Leptospirillum* gp. II 0.08%

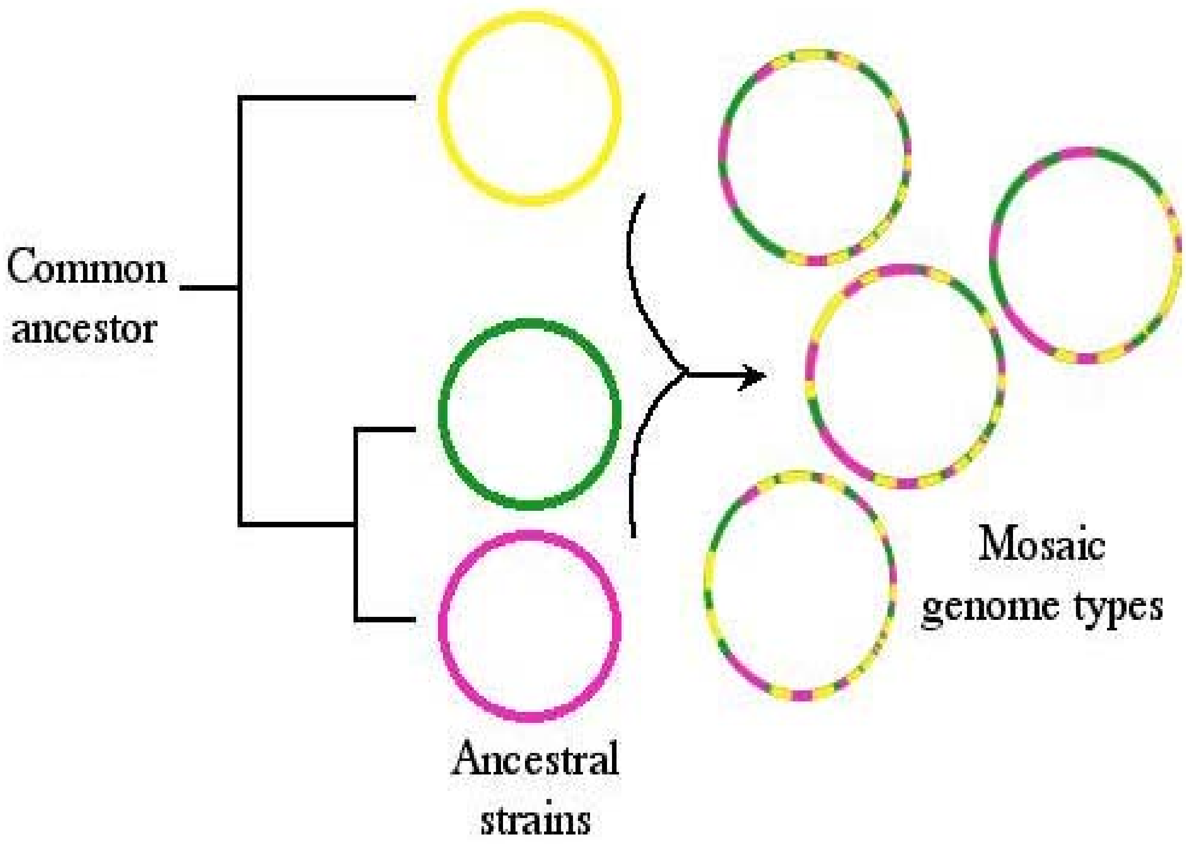
•*Ferroplasma* type II 2.2%

FerII Variation Occurs in Shared Blocks


Consensus tttaatgctcaggaaaatattaaacagttcgaagggcttcacacttgtccttgataacggttcatccctgcatggatatgccacaatggaacagt
XFG55216
XFG63196
XFG31528
XFG70757
XFG47994
XFG44707
XFG65814
XFG70247
XFG48207
XFG23964
XFG21687
XFG37360
XFG52653
XFG31537
Ferroplasma type II strain 1
Ferroplasma type II strain 2
Ferroplasma type II strain 3
Ferroplasma acidarmanus fer1

Fer II variations are non-random

3 basic haplotypes



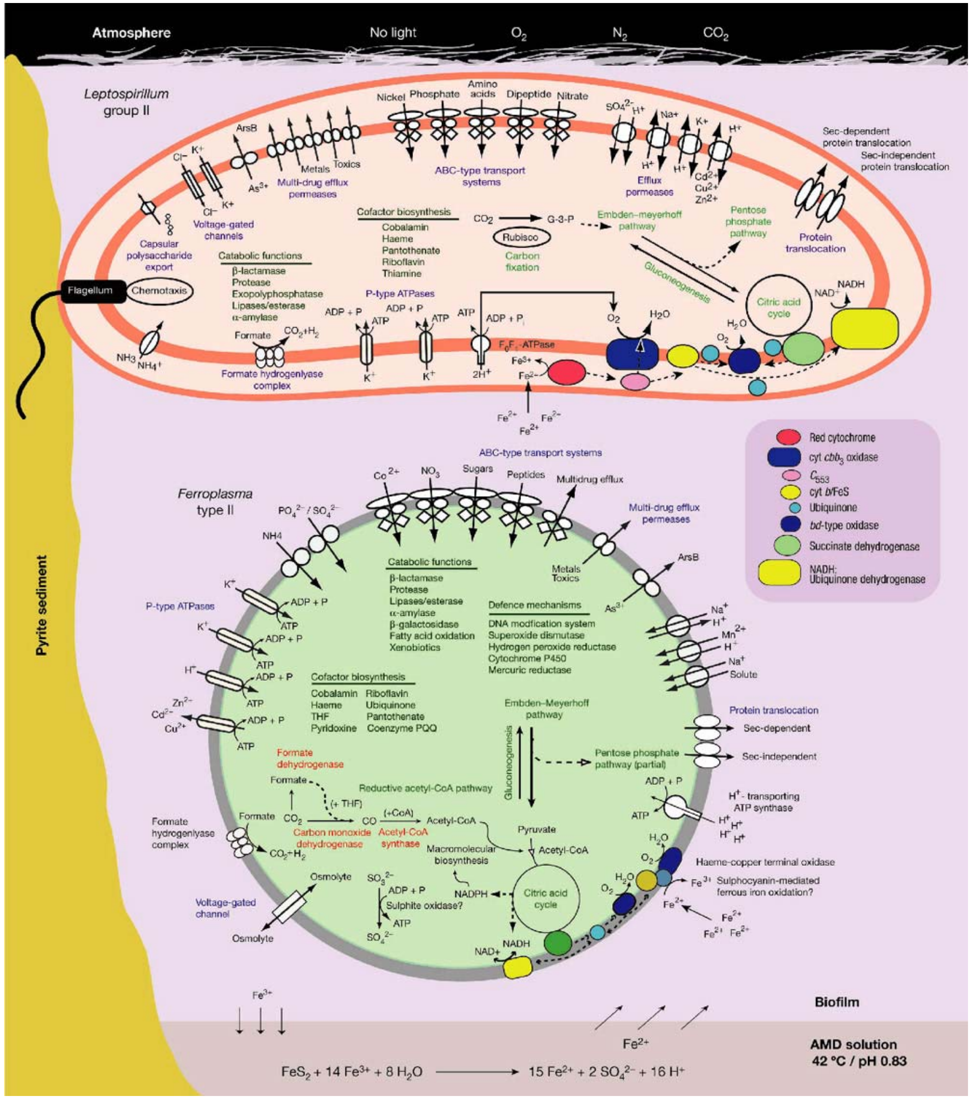
Genes of Each Species Correlated to Function: Interdependence Identified

All organisms contain pathways for carbon fixation

~~*Leptospirillum* II~~
~~Dominant bacteria~~

Nitrogen Fixation
Leptospirillum III

Ferroplasma I
Role—carbon catabolism

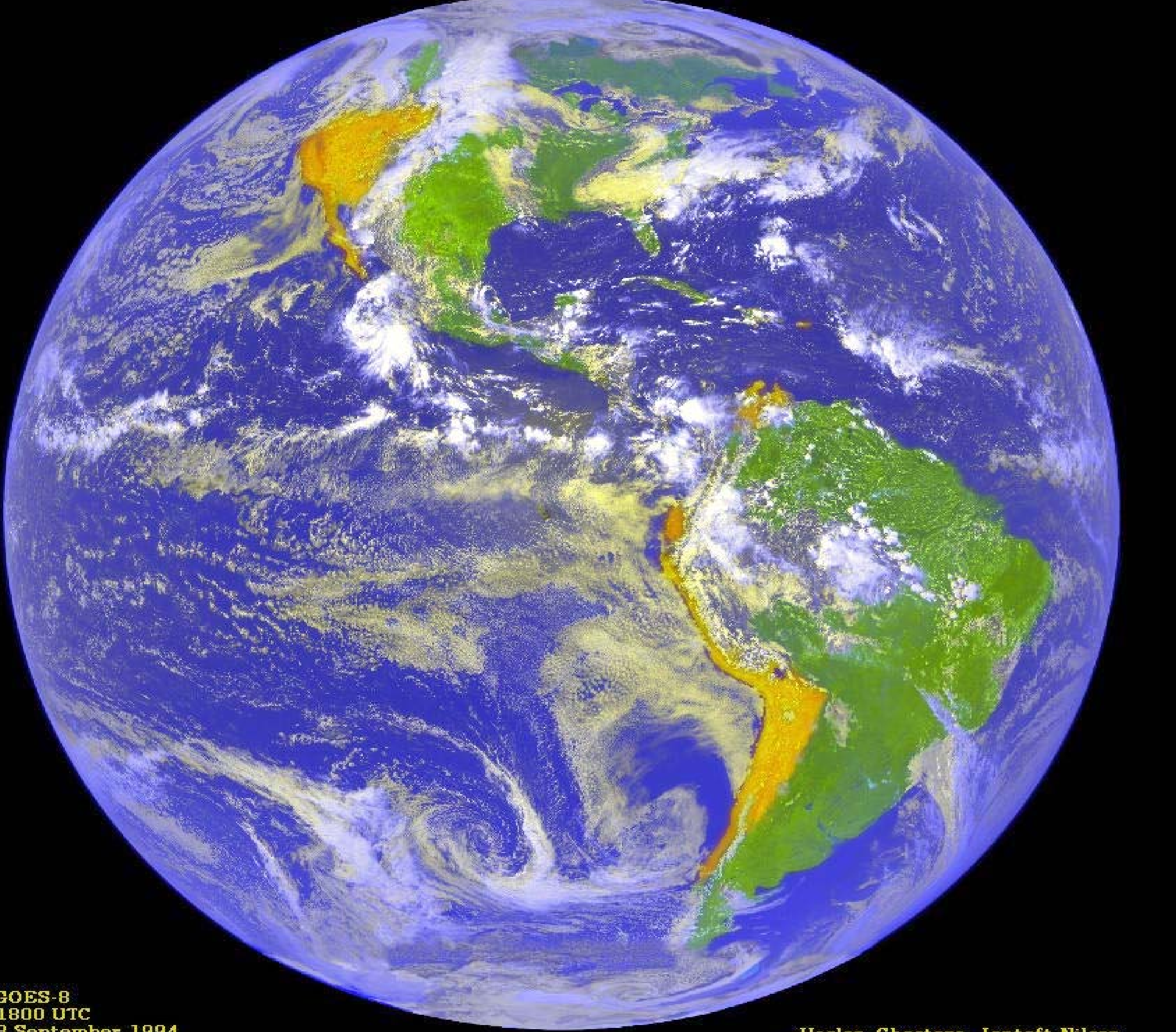


Summary of Iron Mountain Biofilm

- Limited number of predominant species present in biofilm the majority have never been cultured
- Evidence suggests correct genome assembly
- Simplicity of community suggests removal of most variants by natural selection
- Insights into metabolic capabilities of community offer "potential" approach to remediation

Next Steps

- **DNA Isolation Methods**
- **Cloning Techniques**
- **Amplification Technologies**
- **Assembly Algorithms**
- **More Complex Communities…**



GOES-8
1800 UTC
3 September 1994
Red: Visible
Green: Visible
Blue: Inverted 11 µm Infrared

Hasler, Chesters, Jentoft-Nilsen
NASA Goddard Lab. for Atmospheres
&
Nielsen
University of Hawaii

U.S.
DEPARTMENT
OF ENERGY
JGI