JGI Overview Now and the Future
9/02/09

Summary of the 2008 BER S&O Review

- New JGI Management and Organizational Structure
- Changes in JGI Informatics

2009 JGI Five Year Strategic plan

JGI Science & Operations Review

Dec 3-5, 2008, Report Issued March 3, 2009

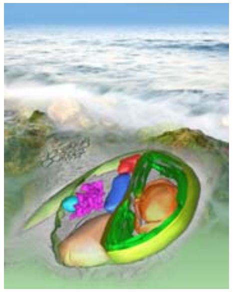
- **Science**
- Management
- Operations
- Informatics

Recent JGI Publication Metrics

	2009
Total Peer-Reviewed Publications (Science/Nature/PNAS)	81 (18)
2009 citations of JGI-Authored Papers published 2005-Present	18,919

Sorgum Genome ***Nature* 2009**

Two algal (Micromonas) Genome ***Science 2009***



Manuscripts in various stages of review

Genomic Encyclopedia of Bacteria and Archea	*(under review Nature)*
Brachypodium Genome	*(submitted to Science)*
Soybean Genome	*(soon to be submitted)*

JGI Science & Operations Review

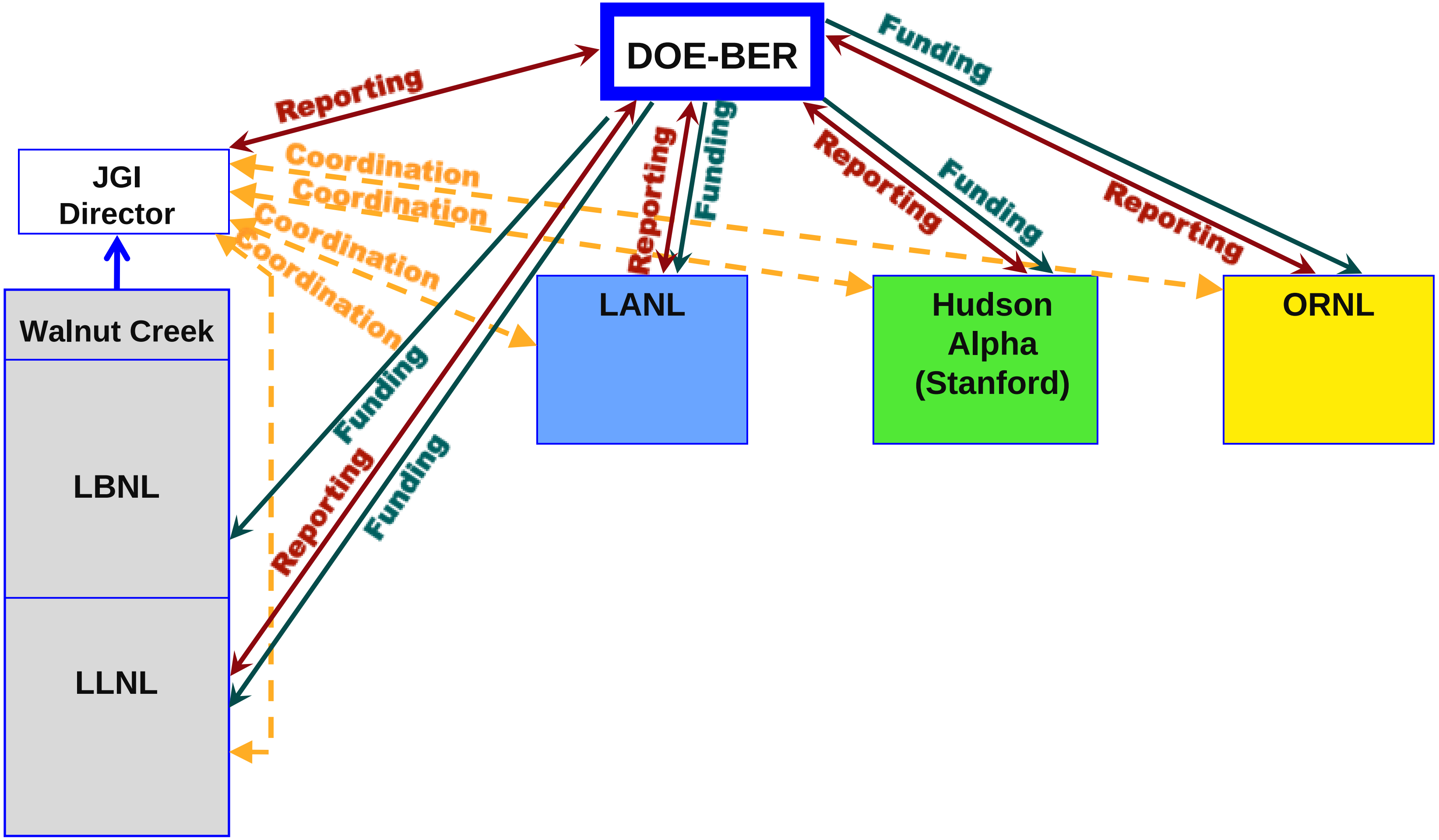
Dec 3-5, 2008, Report Issued March 3, 2009

- Science
- **Management**
- **Operations**
- Informatics

JGI Operations and Management

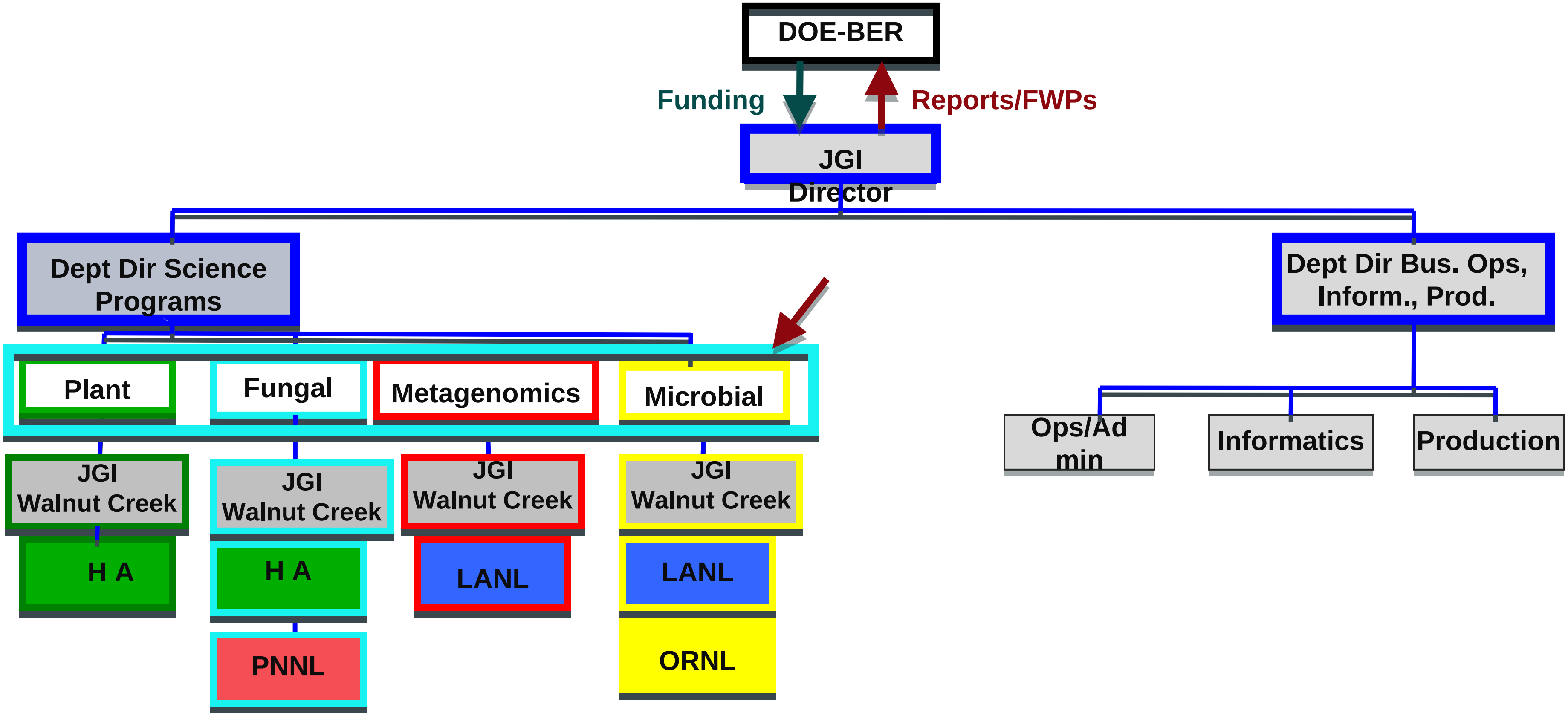
- “Establish a centralized management and organizational structure”
- “Establish funding flow from LBNL to partners to align authority and responsibility”

JGI Operations and Management



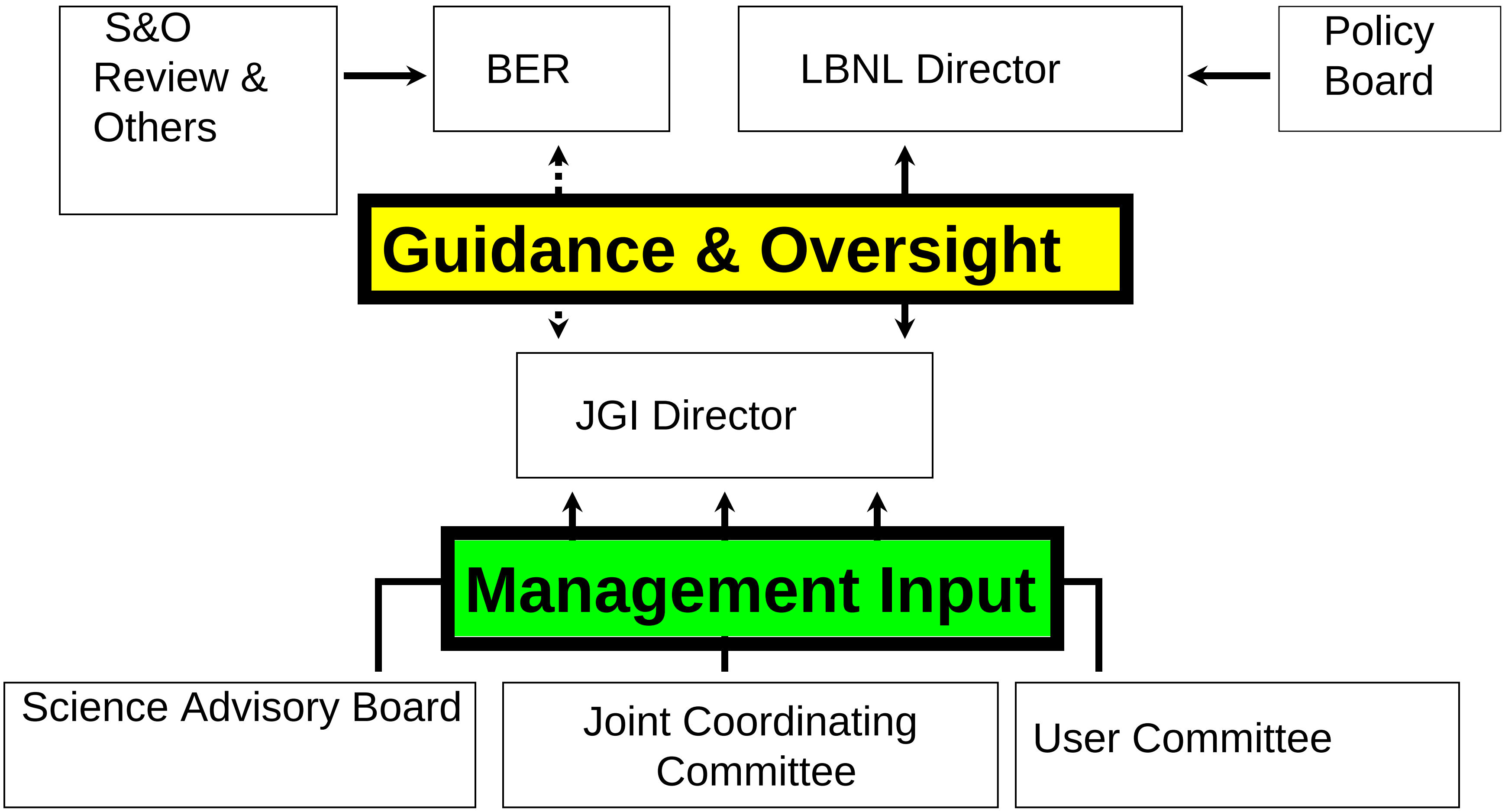
- **Coordination of JGI Scientific Goals Unlinked to Funding and Reporting**

JGI Operations and Management



- JGI Director Responsible and Accountable for JGI Productivity
- Organization Broken into Programs
- Programs Coordinate the JGI Scientific Goals Linked to Funding

Guidance/Oversight & Management Input



New Operations and Management Structure Progress

JGI Science Program Management Model Implemented

Science Program Leads Named and Functioning

Implemented Program Based Financial Planning & Reporting

(Science based programmatic funding decisions have already been made)

JGI Science & Operations Review

Dec 3-5, 2008, Report Issued March 3, 2009

- Science
- Management
- Operations
- Informatics

Informatics

“Remedy Senior Leadership Deficiencies”

Concern about the Informatic Planning and Hardware Infrastructure Faced with Dramatic Increases in Data Generation

Remedy Senior Leadership Deficiencies

JGI Informatics Department Head

Svilen Tzonev (*Illumina / Solexa , Sr. Director of Software Development and Bioinformatics)*

JGI Chief Informatic Officer/ JGI Associate Director (New position)

Jill Mesirov Part Time Consultant to the JGI Director (CIO / Associate Director Broad Harvard MIT))

JGI Informatics Project Management Program Head *(New position)*

Evi Dube (Previously Computing Division Leader (LLNL))

Remedy Senior Leadership Deficiencies

JGI Informatics Department Head
Svilen Tzonev (*Illumina / Solexa , Sr. Director of Software Development and Bioinformatics)*

JGI Chief Informatic Officer/ JGI Associate Director (New position)
Jill Mesirov Part Time Consultant to the JGI (CIO / Associate Director Broad Harvard MIT)

JGI Informatics Project Management Program Head
(New position)
Evi Dube (Previously Computing Division Leader (LLNL))

Remedy Senior Leadership Deficiencies

JGI Informatics Department Head
Svilen Tzonev (*Illumina / Solexa , Sr. Director of Software Development and Bioinformatics)*

JGI Chief Informatic Officer/ JGI Associate Director (New position)
Jill Mesirov Part Time Consultant to the JGI Director (CIO / Associate Director Broad Harvard MIT))

JGI Informatics Project Manager
(New position)
Evi Dube (Previously Computing Division Leader (LLNL))

Informatics

Informatics Advisory Committee

(New advisory group)

First meeting scheduled Dec 09

Concern about the JGI Informatic Infrastructure Faced with Dramatic Increases in Data Generation

I. ARRA Funding Targeted for IT Upgrades: $11.1M

Funding Received-August 7th

Add Infrastructure for Management, Storage and Movement of Genome-Scale Datasets

Increase Storage (100% increase)
Upgrade JGI's Central Computing Cluster (200% increase)
Upgrade ESNet Access

Concern about the JGI Management Infrastructure Faced with Dramatic Increases in Data Generation

II. Explore the DOE National Lab's high performance computing capabilities to help JGI's data intense challenges

Projects already underway or being developed

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Using ScalaBLAST for Metagenomics

- **Porting ScalaBLAST to JGI clusters; help parallelize short read assembler**

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

Working with JGI on designing computing nodes

- **Scientific engagement with JGI on computational challenges**
- **IMG update using large BLAST runs focusing on isolate genomes and potentially some of the new Titanium metagenome datasets**

Argonne NATIONAL LABORATORY

Two large-scale metagenomics problems identified

- **Goal 1: evaluate metagenome assembly by using simulated dataset**
- **Goal 2 : evaluate gene calling through a large blastx search against NR using joined Illumina pair-end reads.**

- **Argonne team identified to assist JGI team**

Lawrence Livermore National Laboratory

- **Assisted IMG in solving contentious parallel computing bug**
- **Working with JGI on designing computing nodes**
- **Possible Site for Remote Computing**

Informatics

Summary:
Recruiting leadership to strategically as well as tactically chart the JGI's informatic directions.

Additional informatic personnel and hardware.

Engaging the DOE National Labs to adapt their high performance computing capabilities to address JGI needs

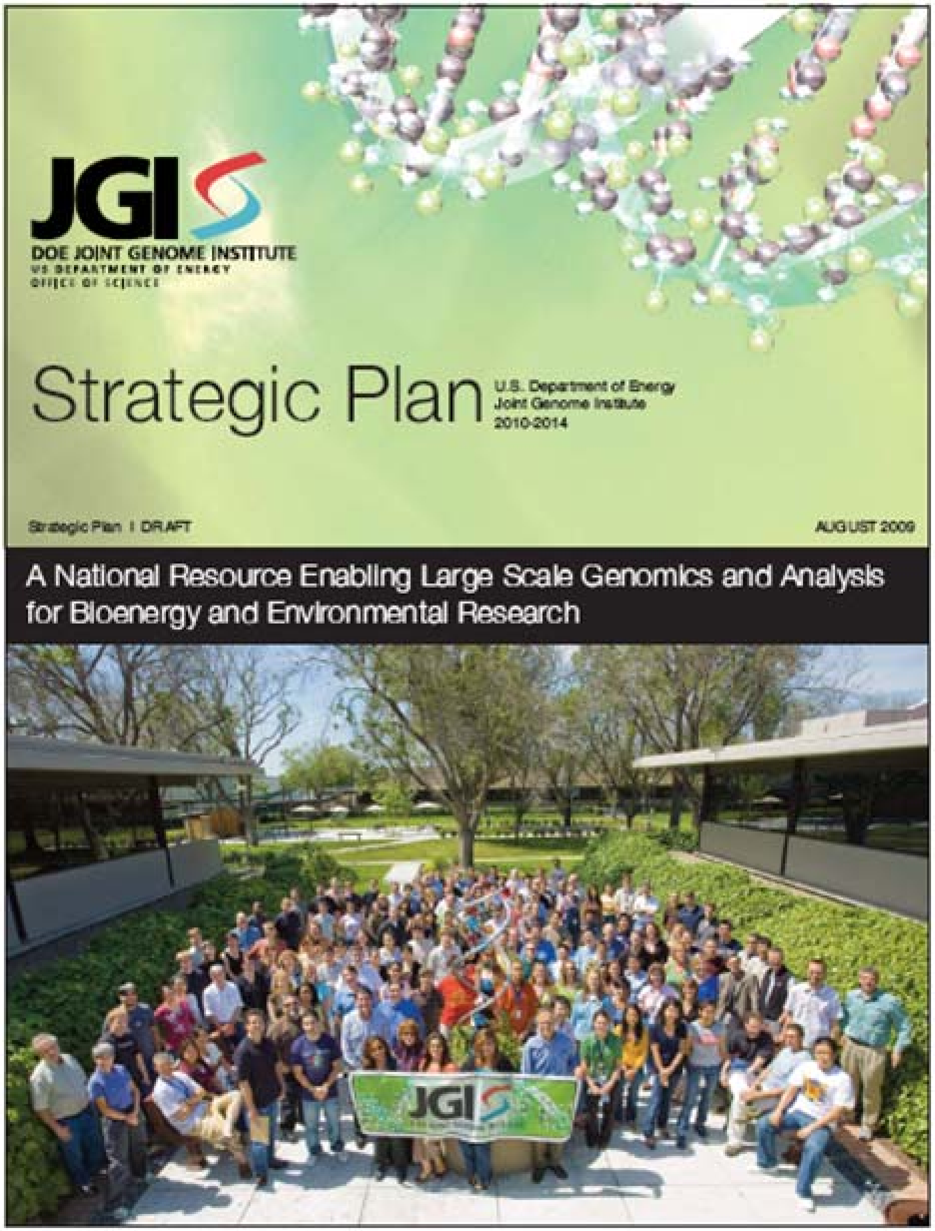
Summary of S & O Review Response

- **23 Committee Recommendations**
  - Informatics (10 Recommendations)
  - Operations & Safety (7 Recommendations)
  - Science (4 Recommendations)
  - Management (2 Management)

- **JGI Corrective Action Status**
  - All Recommendations are either fully implemented or are in the process of be implemented

- **Corrective Action Milestones-Reviewed by JGI Senior Managers and Communicated with BER**

DOE JGI 5-Year Strategic Plan



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

Strategic Plan

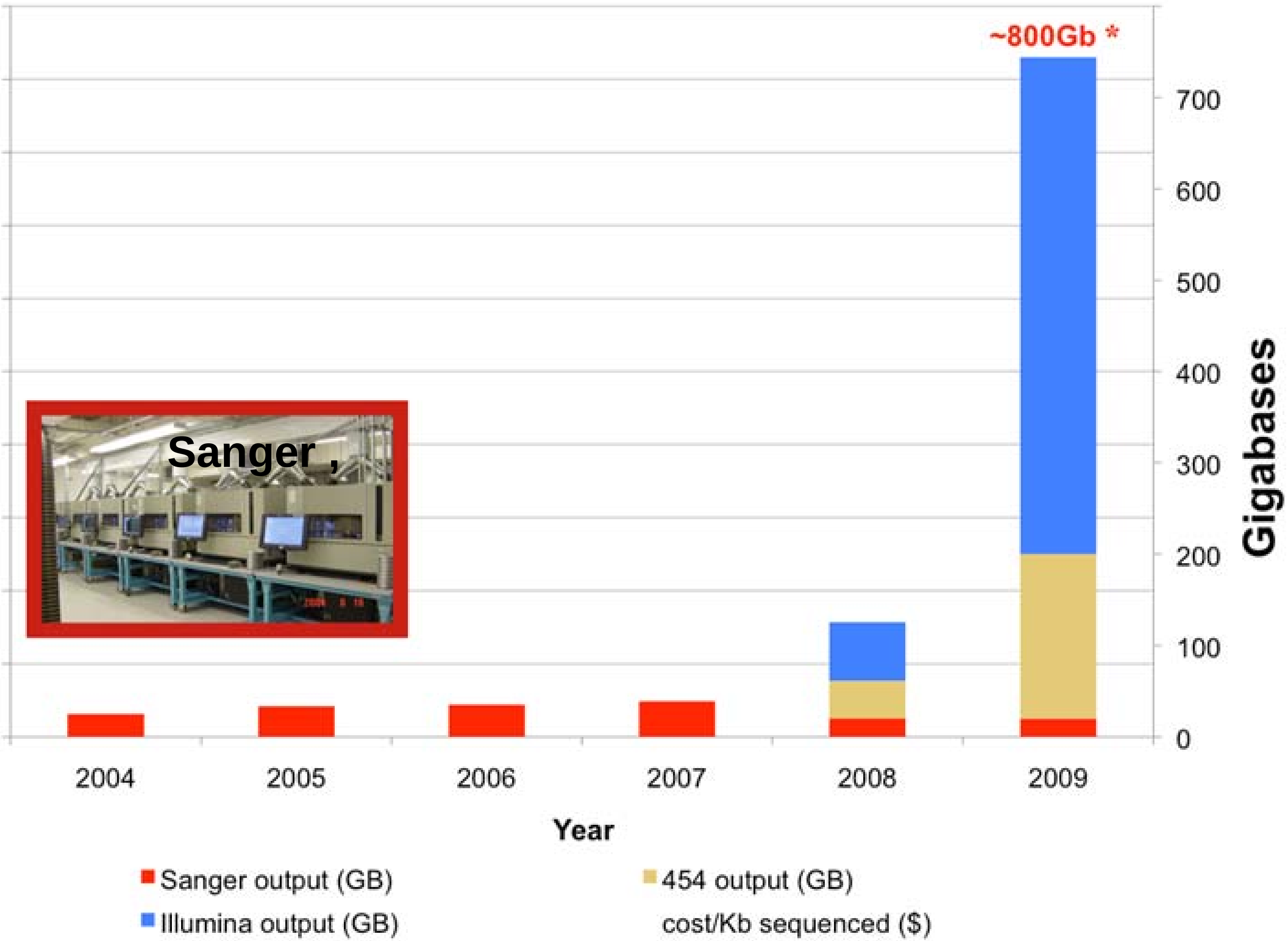
U.S. Department of Energy
Joint Genome Institute
2010-2014

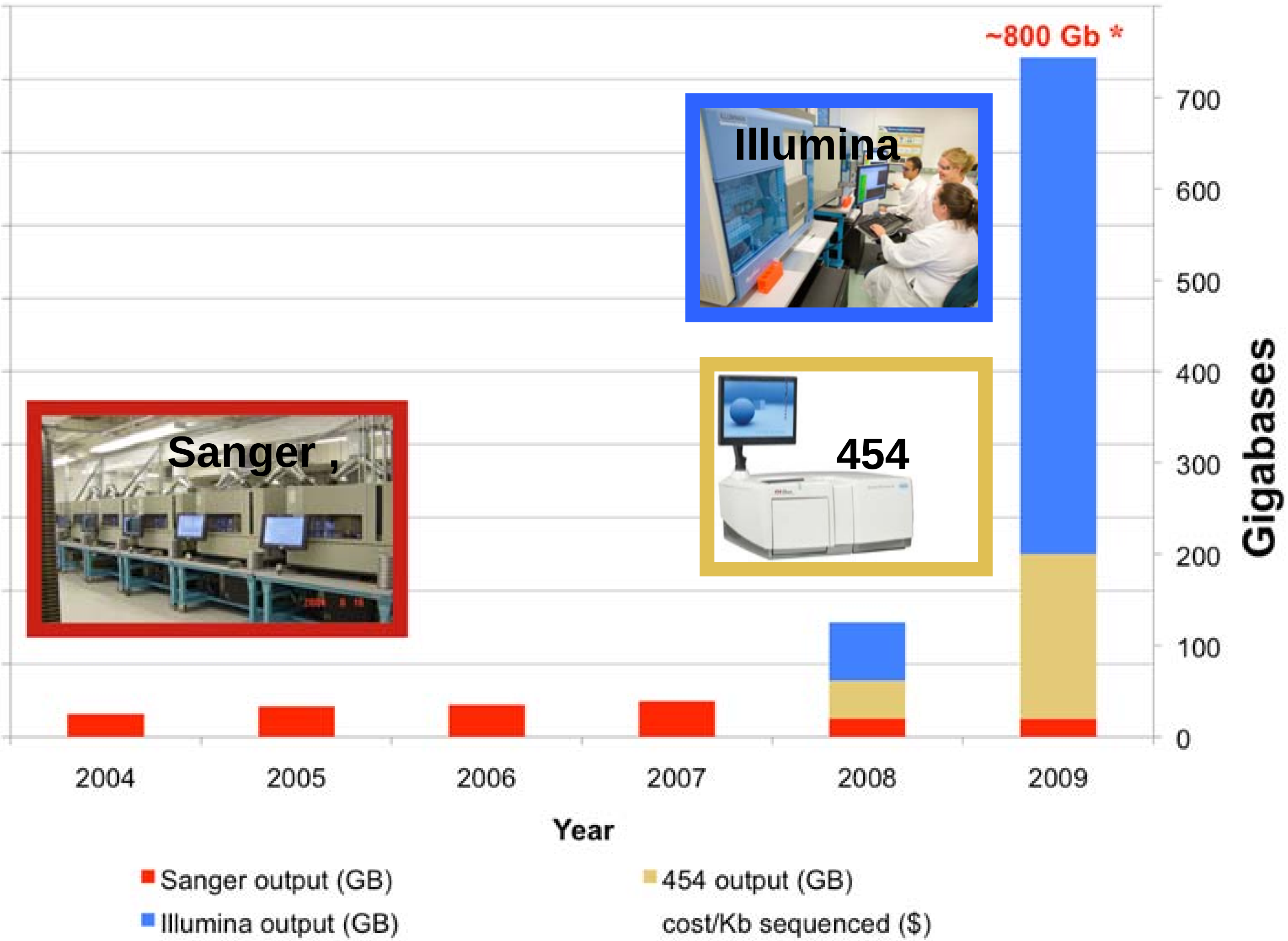
Strategic Plan | DRAFT

AUGUST 2009

A National Resource Enabling Large Scale Genomics and Analysis for Bioenergy and Environmental Research

Revolutionary Technological Changes




~800 Gb *
Illumina
454
Sanger ,
700
600
500
400
300
200
100
0
Gigabases
2004
2005
2006
2007
2008
2009
Year
🟥 Sanger output (GB)
🟨 454 output (GB)
🟦 Illumina output (GB)
cost/Kb sequenced ($)

JGI 2004 Onward

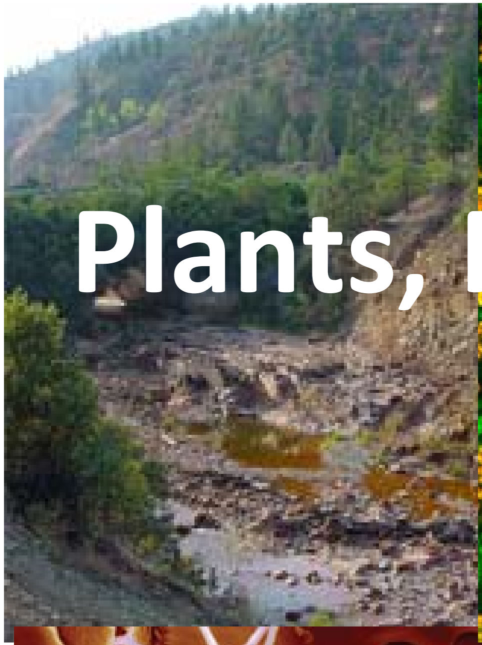
Mission:

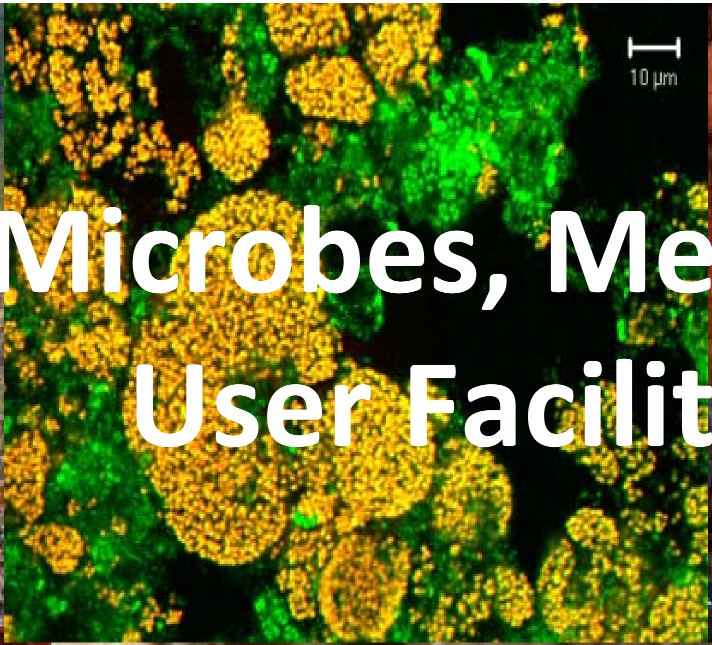
User Facility for Large Scale Genomics and Analysis
to Enable Bioenergy and Environmental Research

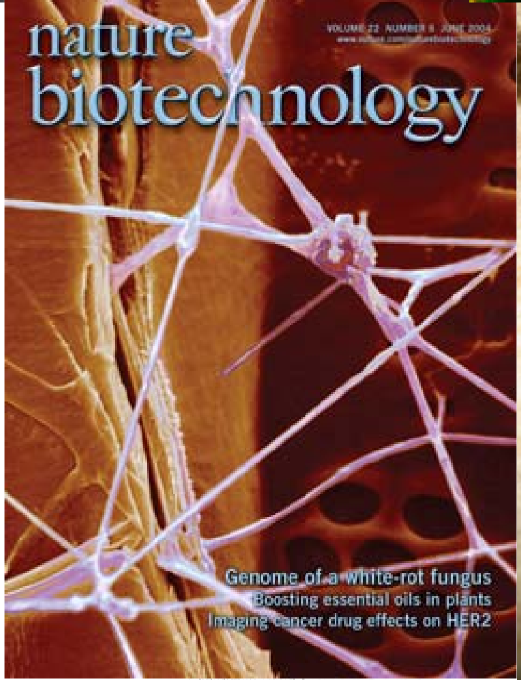
Unique Capabilities

In carrying out genomics of focus, scale, and complexity
to help users solve important but hard DOE relevant problems

Plants, Microbes, Metagenomes User Facility





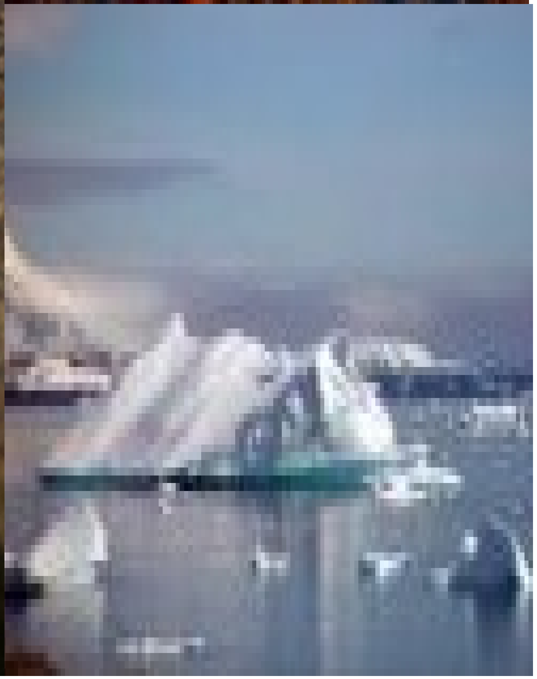


nature biotechnology

VOLUME 22 NUMBER 6 JUNE 2004
www.nature.com/naturebiotechnology

Genome of a white-rot fungus
Boosting essential oils in plants
Imaging cancer drug effects on HER2

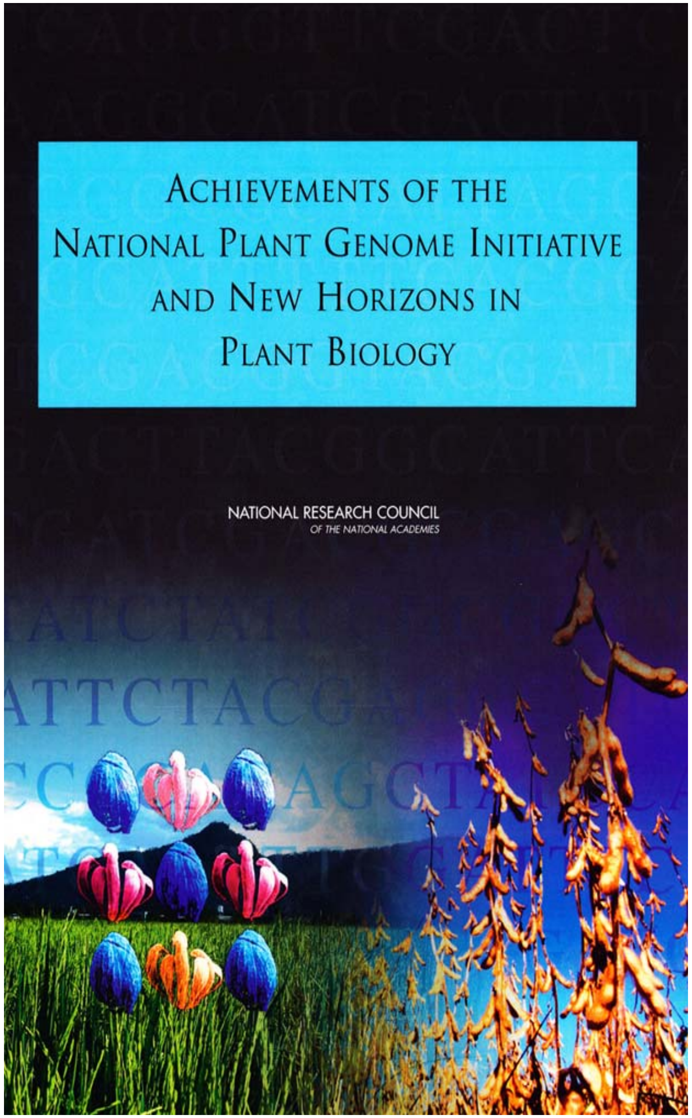
technologies were presented



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

National Research Council
Plant Genome Report 2008

U.S. DEPARTMENT OF ENERGY
ENERGY
Office of Science



Achievements of the National Plant Genome Initiative and New Horizons in Plant Biology

National Research Council
of the National Academies

“It is critical that JGI continue to serve a broad remit for sequencing and resequencing of plant genomes…”

Plant Genomes

Populus trichocarpa	***black cottonwood***
Chlamydomonas reinhardtii	***chlorophyte alga***
Physcomitrella patens	***bryophyte moss***
Sorghum bicolor	***sorghum***
Glycine max	***soybean***
Panicum virgatum	*switchgrass*
Miscanthus x giganteus	*miscanthus*
Volvox carteri	*multicellular green alga*
Arabidopsis lyrata	*Lyrate rockcress*
Selaginella moellendorffii	*lycophyte fern*
Brachypodium distachyon	*false purple brome*
Mimulus guttatus	*monkeyflower*
Manihot esculenta	*cassava*
Prunus persica	*peach*
Setaria italica	*foxtail millet*
Capsella rubella	*Pink Shepherd’s purse*
Aquilegia coerulea	*Goldsmith columbine*
Gossypium raimondii	*diploid cotton*
Eucalyptus grandis	*rose gum*
Phaseolus vulgaris	*Common bean*
Zea mays (Mo17 inbred)	*maize*

Flagship Plant Genomes
(High Priority Organisms)

- Produce higher quality accurate and well annotated genomes
- Add additional information

Sorghum Bicolor

Chlamydomonas

Brachypodium

Poplar

Switchgrass

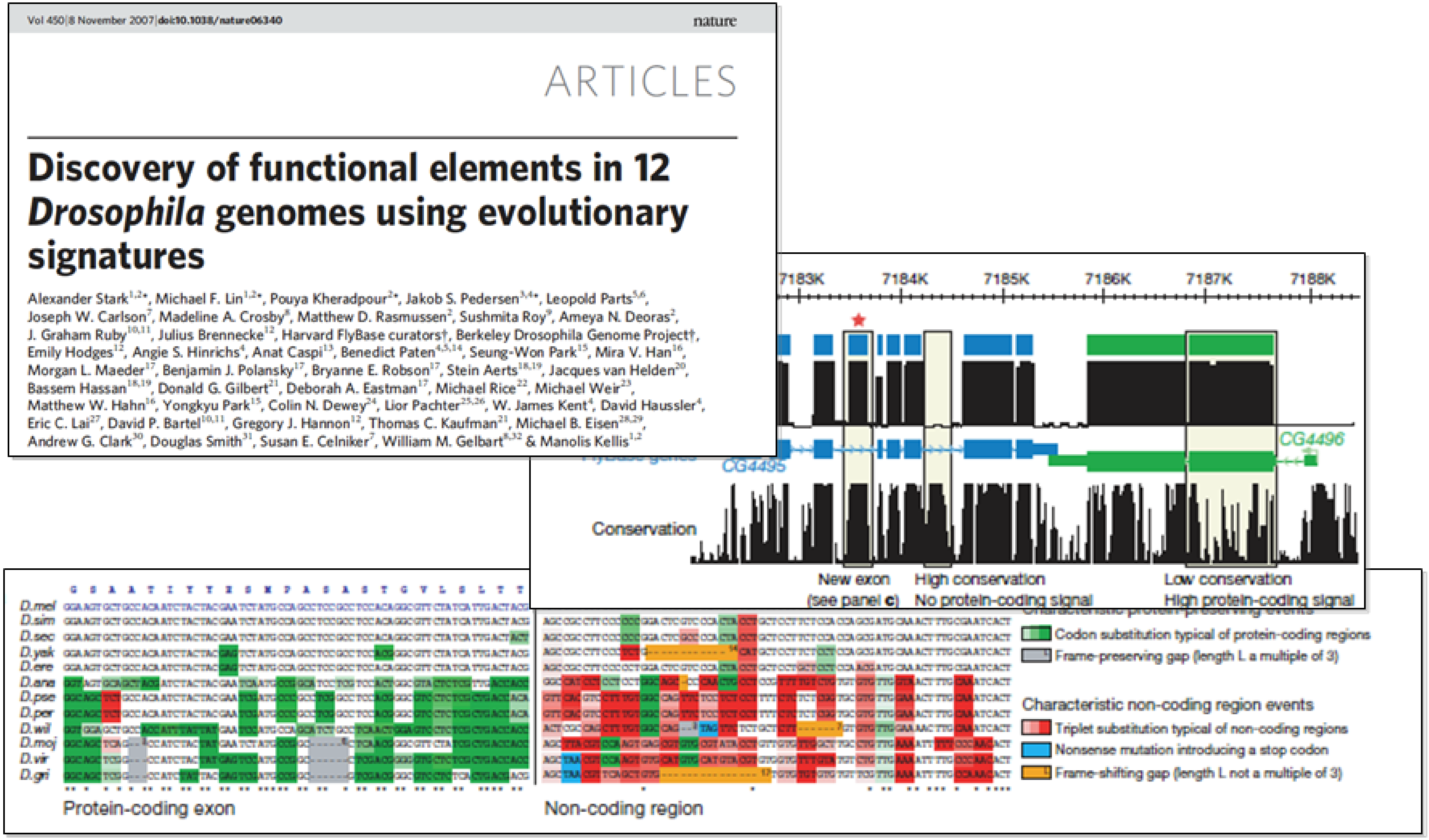
Foxtail millet

Soybean

Miscanthus

Problem with plant genomes: Though we have the sequence we cannot assign function to most of it

Discovery of functional elements through comparative genomics



Vol 450 | 8 November 2007 | doi:10.1038/nature06340

nature

ARTICLES

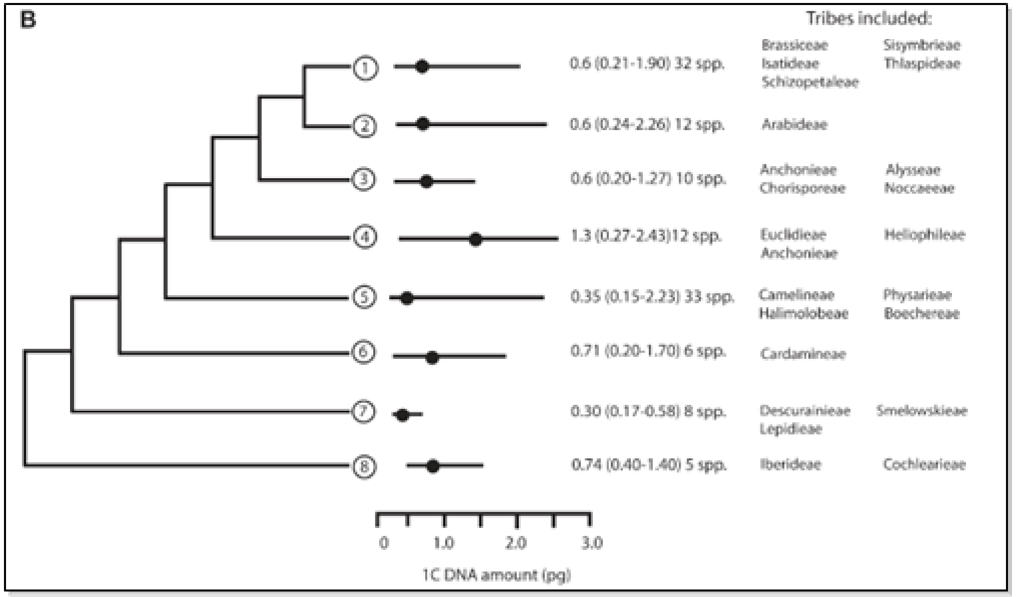
Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Alexander Stark[1,2*], Michael F. Lin[1,2*], Pouya Kheradpour[2*], Jakob S. Pedersen[3,4*], Leopold Parts[5,6], Joseph W. Carlson[7], Madeline A. Crosby[8], Matthew D. Rasmussen[2], Sushmita Roy[9], Ameya N. Deoras[2], J. Graham Ruby[10,11], Julius Brennecke[12], Harvard FlyBase curators†, Berkeley Drosophila Genome Project†, Emily Hodges[12], Angie S. Hinrichs[4], Anat Caspi[13], Benedict Paten[4,5,14], Seung-Won Park[15], Mira V. Han[16], Morgan L. Maeder[17], Benjamin J. Polansky[17], Bryanne E. Robson[17], Stein Aerts[18,19], Jacques van Helden[20], Bassem Hassan[18,19], Donald G. Gilbert[21], Deborah A. Eastman[17], Michael Rice[22], Michael Weir[23], Matthew W. Hahn[16], Yongkyu Park[15], Colin N. Dewey[24], Lior Pachter[25,26], W. James Kent[4], David Haussler[4], Eric C. Lai[27], David P. Bartel[10,11], Gregory J. Hannon[12], Thomas C. Kaufman[21], Michael B. Eisen[28,29], Andrew G. Clark[30], Douglas Smith[31], Susan E. Celniker[7], William M. Gelbart[8,32] & Manolis Kellis[1,2]

7183K
7184K
7185K
7186K
7187K
7188K
FlyBase genes
CG4495
CG4496
Conservation
New exon
(see panel c)
High conservation
No protein-coding signal
Low conservation
High protein-coding signal
D.mel
D.sim
D.sec
D.yak
D.ere
D.ana
D.pse
D.per
D.wil
D.moj
D.vir
D.gri
Protein-coding exon
Non-coding region
Characteristic protein-preserving events
🟩 Codon substitution typical of protein-coding regions
▨ Frame-preserving gap (length L a multiple of 3)
Characteristic non-coding region events
🟥 Triplet substitution typical of non-coding regions
🟦 Nonsense mutation introducing a stop codon
🟧 Frame-shifting gap (length L not a multiple of 3)

JGI Pilot Project to Sequence and Analyze 12 Arabidopsis Relatives

(Project Leaders: Joe Ecker, Detlef Weigel, Dan Rokhsar)



The 12 Arabidopsis Relatives Project will:

•Identify functional elements in these genomes based on evolutionary constraint

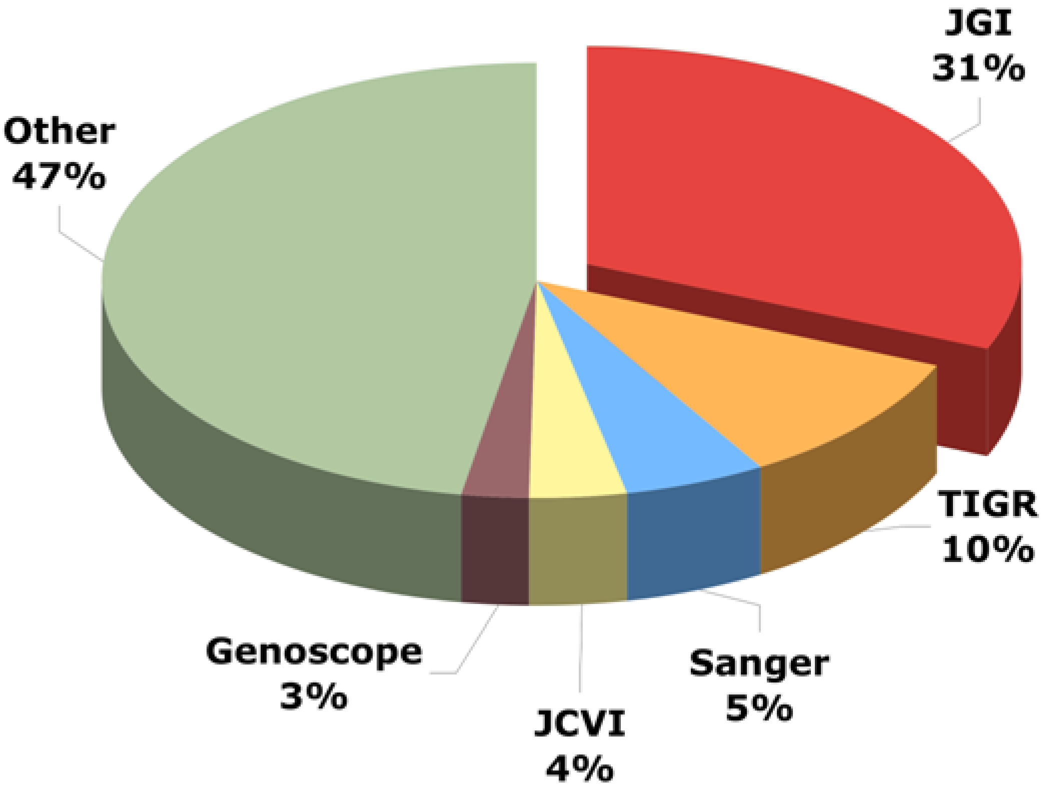
•Provide a road map for similar projects in plant groups with much larger genomes
e.g., grasses

Goal of JGI Plant Program

Produce the genomic tools necessary to understand cell wall construction, feedstock production and carbon sequestration and sunlight energy harvesting pathways

JGI Microbial Program:

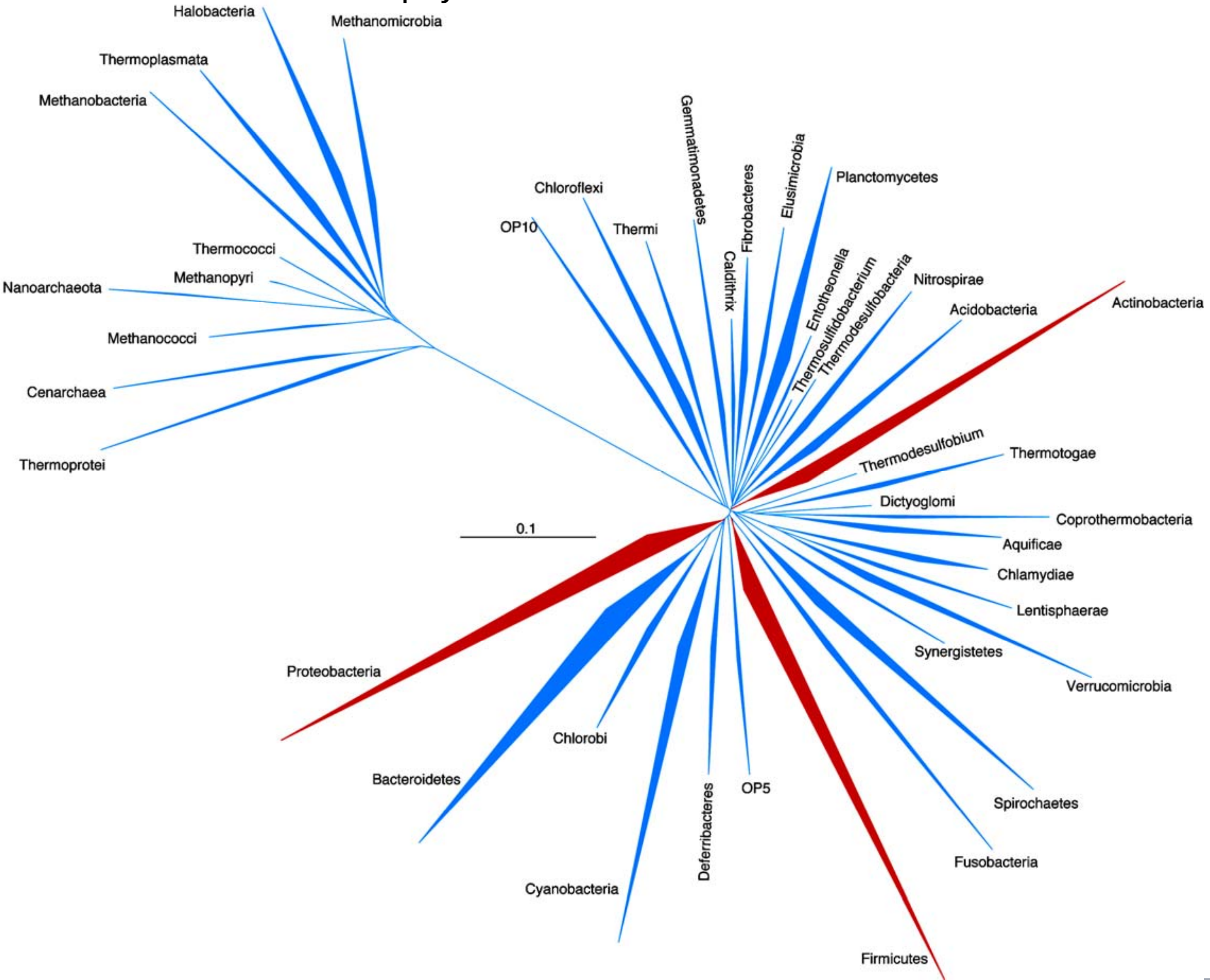
Bacteria and Archaea 939 Complete Genomes (NCBI, August 2009)



JGI has produced ~30% of the world’s complete archaeal and bacterial genomes

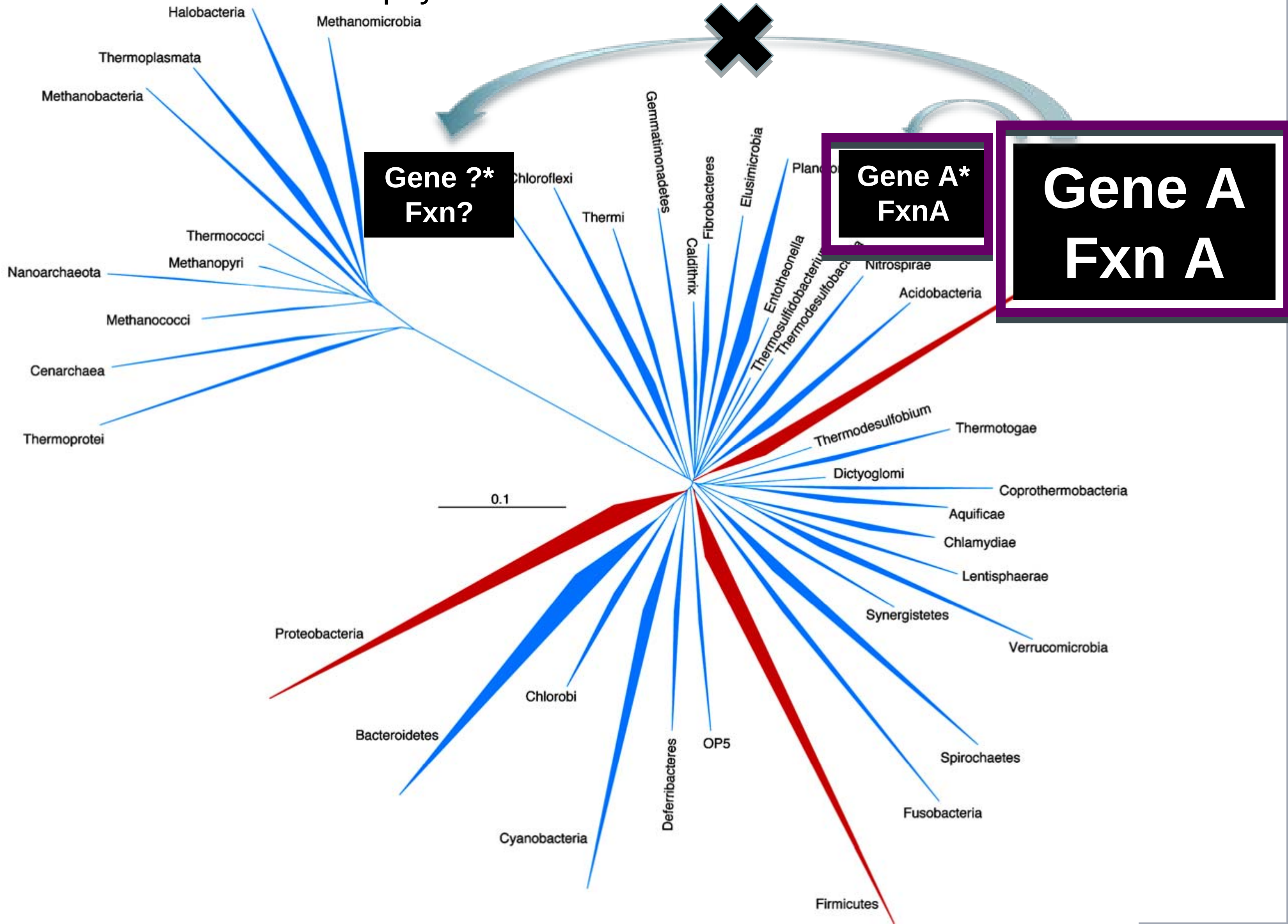
3/4 of sequenced genomes belong to just three bacterial phyla

bacterial and archaeal phyla with cultured isolate



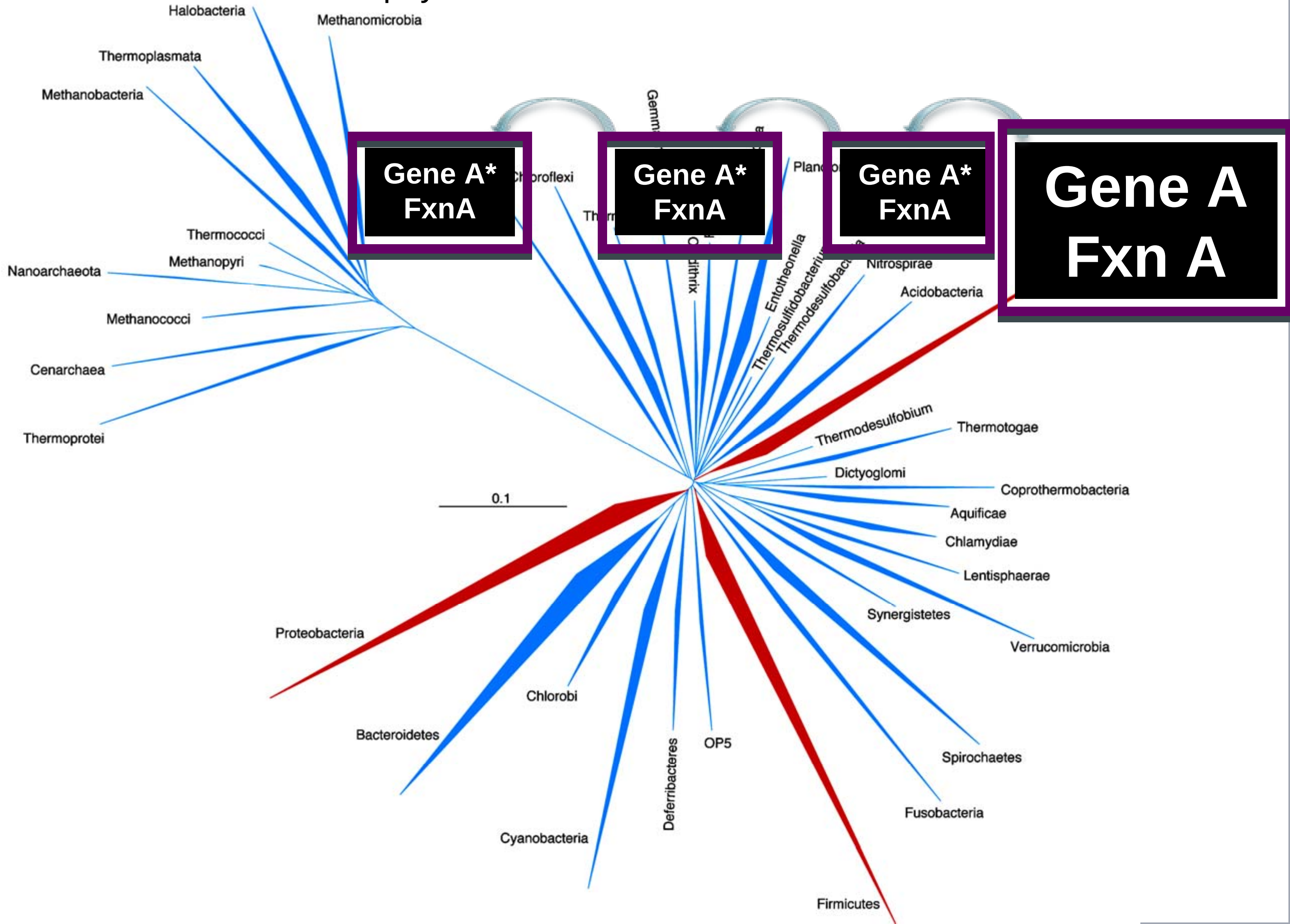
3/4 of sequenced genomes belong to just three bacterial phyla

bacterial and archaeal phyla with cultured isolate



3/4 of sequenced genomes belong to just three bacterial phyla

bacterial and archaeal phyla with cultured isolate



Gene A*
FxnA

Gene A*
FxnA

Gene A*
FxnA

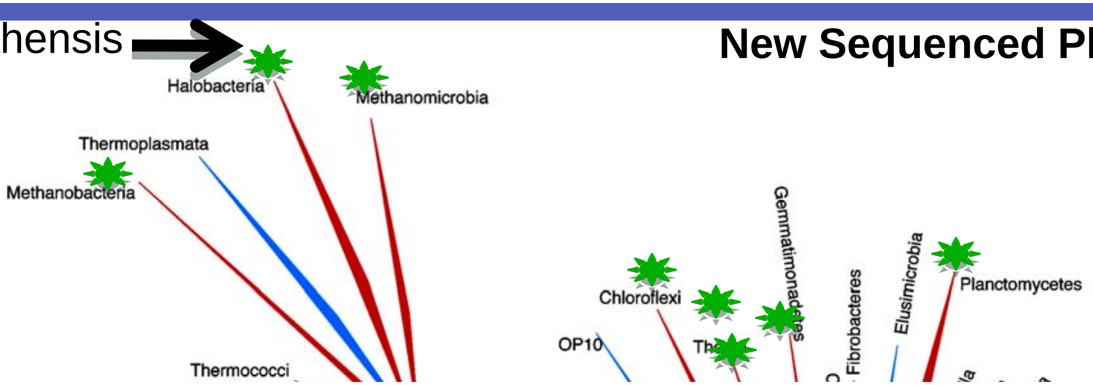
Gene A
Fxn A

‘

Pilot Genomic Encyclopedia of Bacteria and Archaea
GEBA (Sequencing and analysis of 64 genomes)

H. Utahensis

New Sequenced Phyla GEBA 🟩

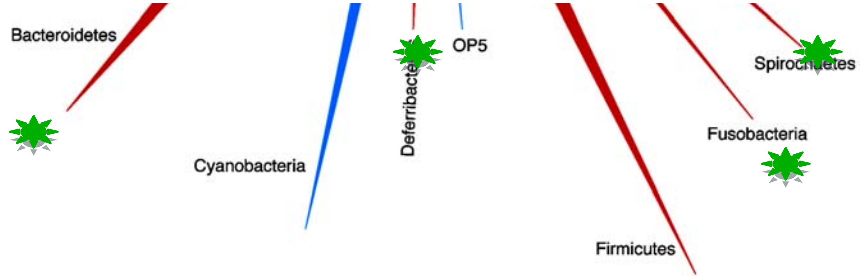


(Final stages of Nature review)

- **“Significantly improves analysis of metagenomic data”**

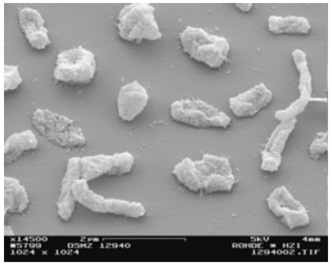
- **“Markedly improves the identification of new gene families and members of gene families”**

including those of DOE relevance



H. Utahensis

Great Salt Lake

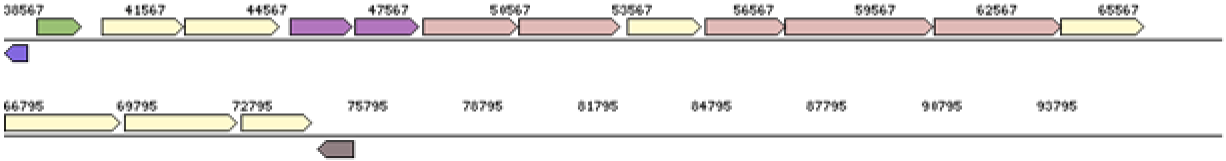


Halorhabdus utahensis

27% NaCl

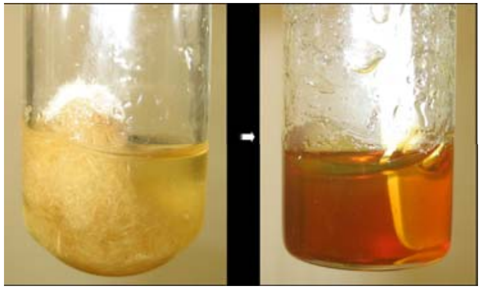
6x saltier than ocean

A cellulase gene cluster



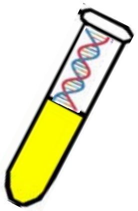
Optimized Cellulases

Ionic Liquid Detergents for Cellulose Dissolution

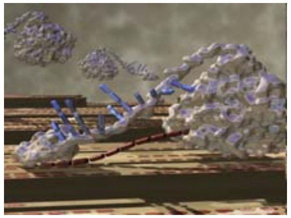


before after

DOE Bioenegy Center (JBEI) of Synthesized H Utahensis Cellulases

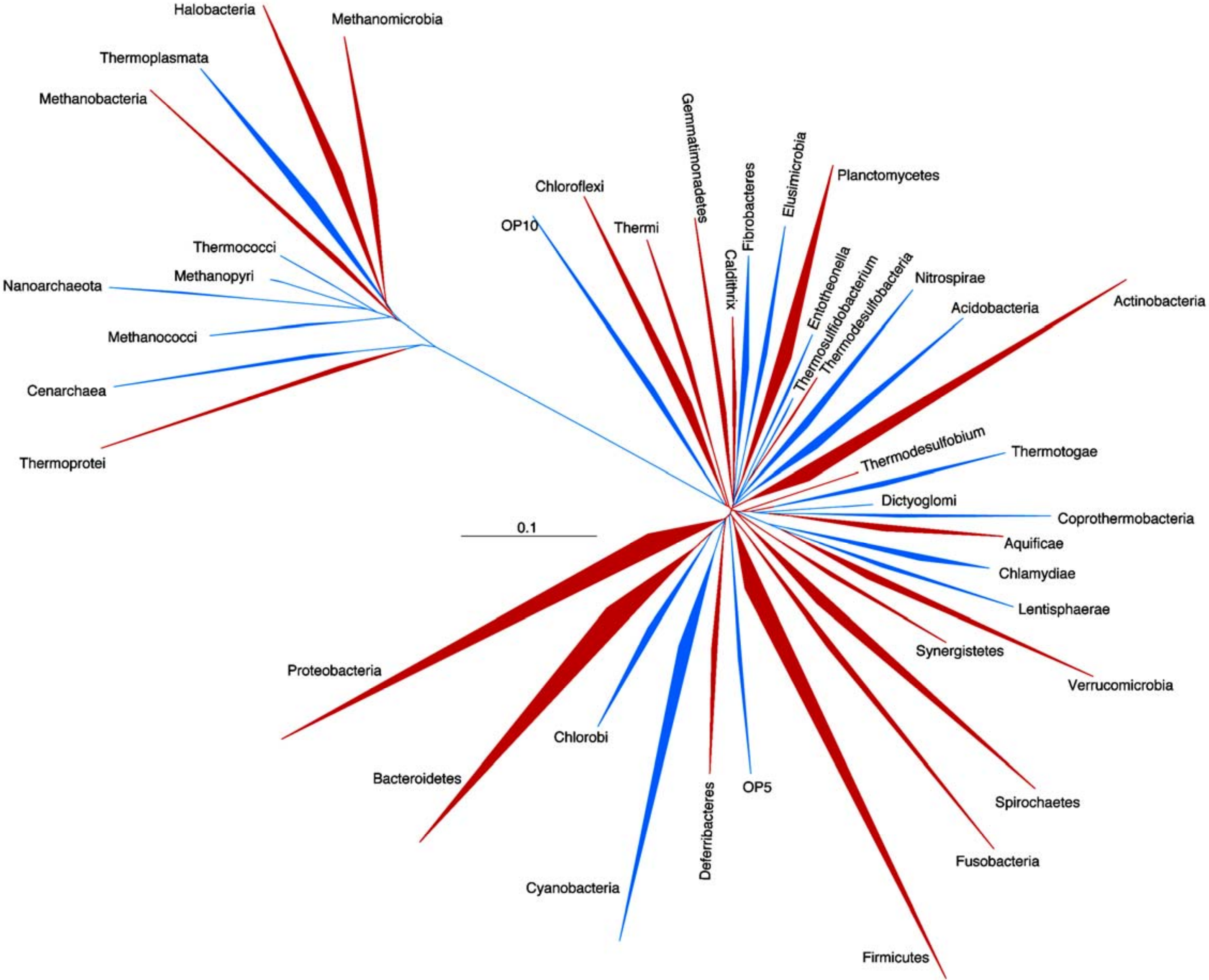


High Salt Cellulase Activity



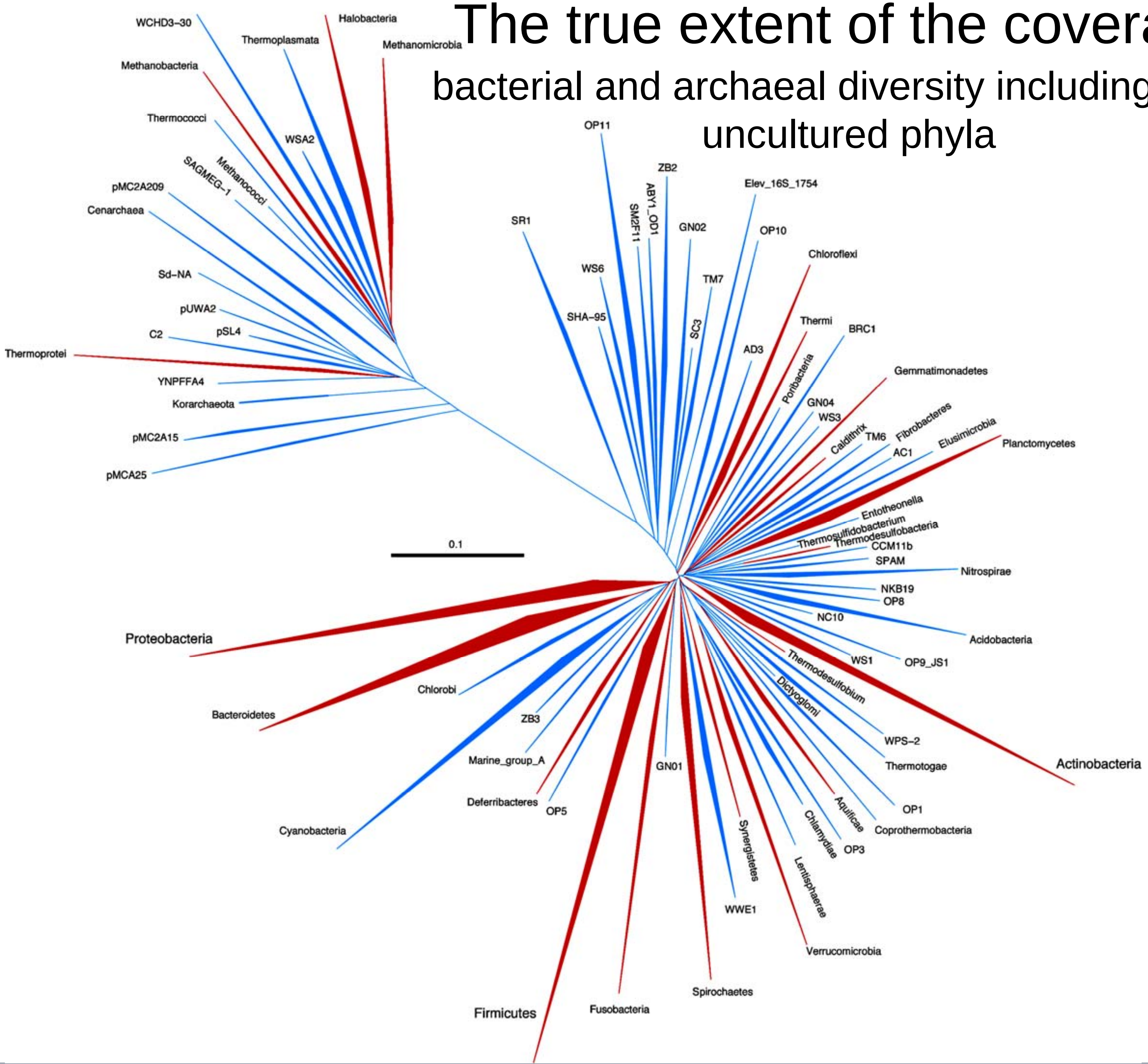
Present State of Bacteria and Archaeal Genome Coverage *(post GEBA)*

bacterial and archaeal phyla with **cultured isolate**



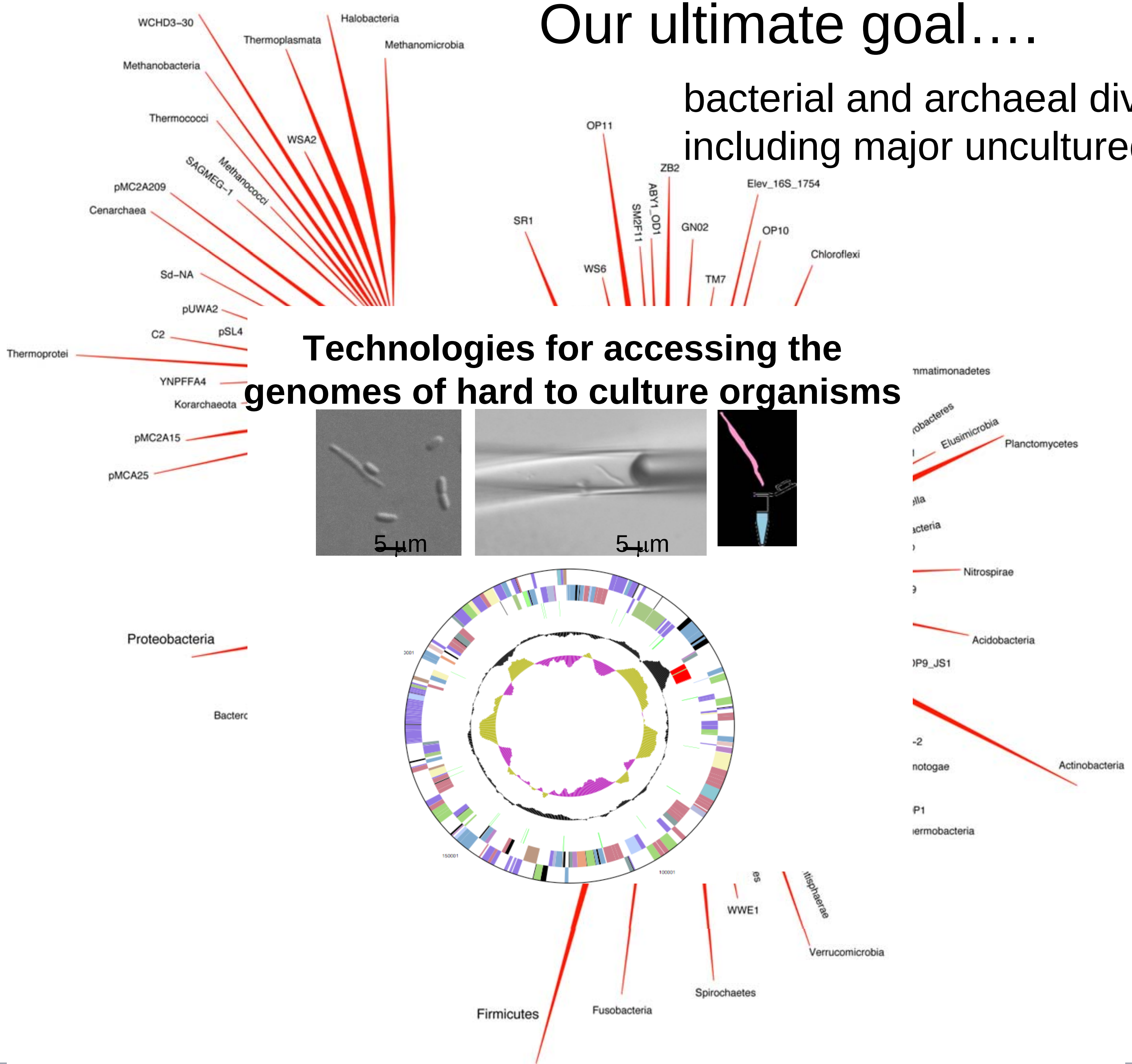
The true extent of the coverage….

bacterial and archaeal diversity including major uncultured phyla



Our ultimate goal….

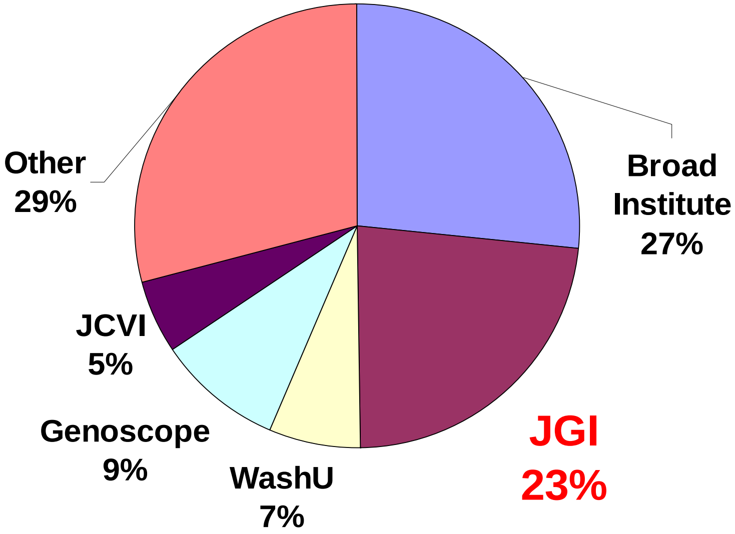
bacterial and archaeal diversity
including major uncultured phyla



Technologies for accessing the genomes of hard to culture organisms

JGI Fungal Program

Fungal genome projects worldwide



Large Scale Genomic Analysis of Fungi

Vol.1. **Plant feedstock health**

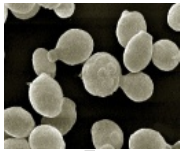
Part 1. Mycorrhizal Symbionts (Basidiomycota)

Part 2. Plant Pathogens (Dothideomycota)

Vol.2. **Lignocellulose degradation**

White and brown rot (Basidiomycota)

Vol.3. **Fermentation**

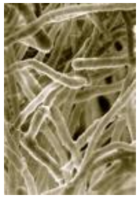


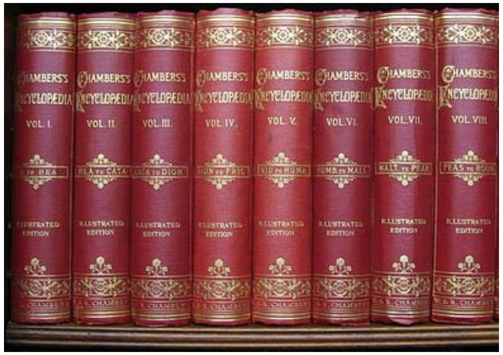
Genomic Encyclopedia of Fungi (GEF)

Vol.4. **Industrial organisms**

Part 1. Trichoderma spp.

Part 2. Aspegillus nigri





JGI Metagenomics

2004

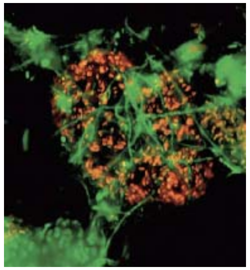
2006

2008

2009

Acid Mine Drainage
Banfield/Chapman
Nature

Marine Archaeal
Methane Oxidation
DeLong/Putnam
Science



Fossil
Noonan
Science

Soil
Tringe
Science

Olavius algarvensis
Dubilier/Woyke
Nature

Marine Planktonic
DeLong
Science

Wastewater EBR
Hugenholtz
Nature Biotech

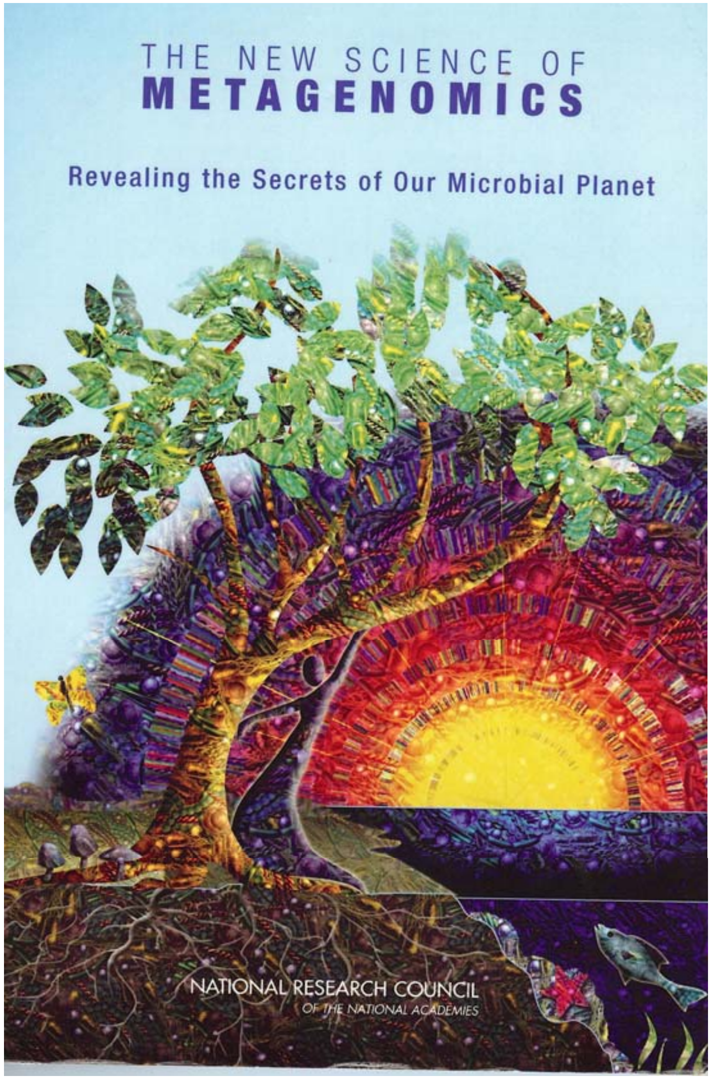
Deep Mine
Chivian
Science

Termite Gut
Leadbetter/
Hugenholtz
Nature

Oral TM7
Quake
PNAS

Lake Washington
Methylotrophic
Chistoserdova
Nature Biotech

Background for Metagenomics Plans


THE NEW SCIENCE OF METAGENOMICS
Revealing the Secrets of Our Microbial Planet
NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

recommended the establishment of a small number of large-scale projects to study particular habitats

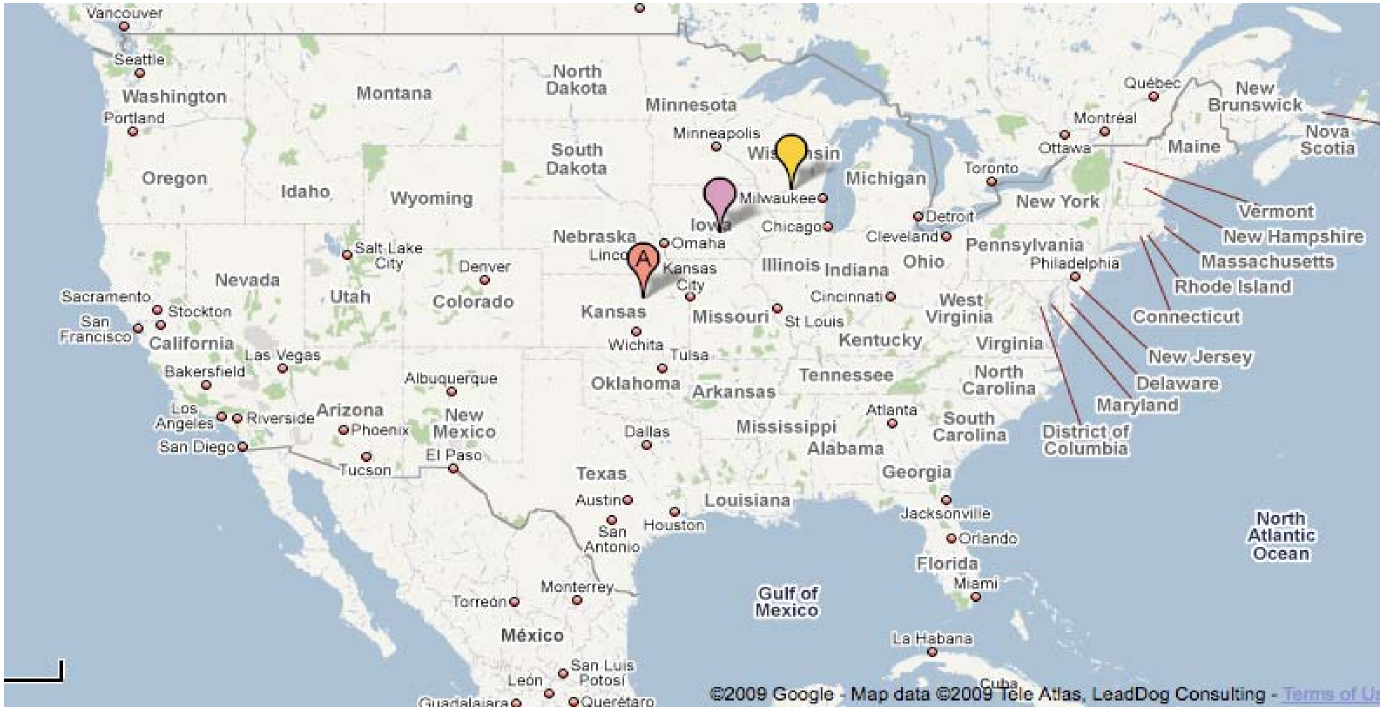
Soil

Represents a habitat that is important due to its role in carbon and nitrogen cycling and biomass production. It is a particularly “genomicaly” challenging environment due to its complexity.

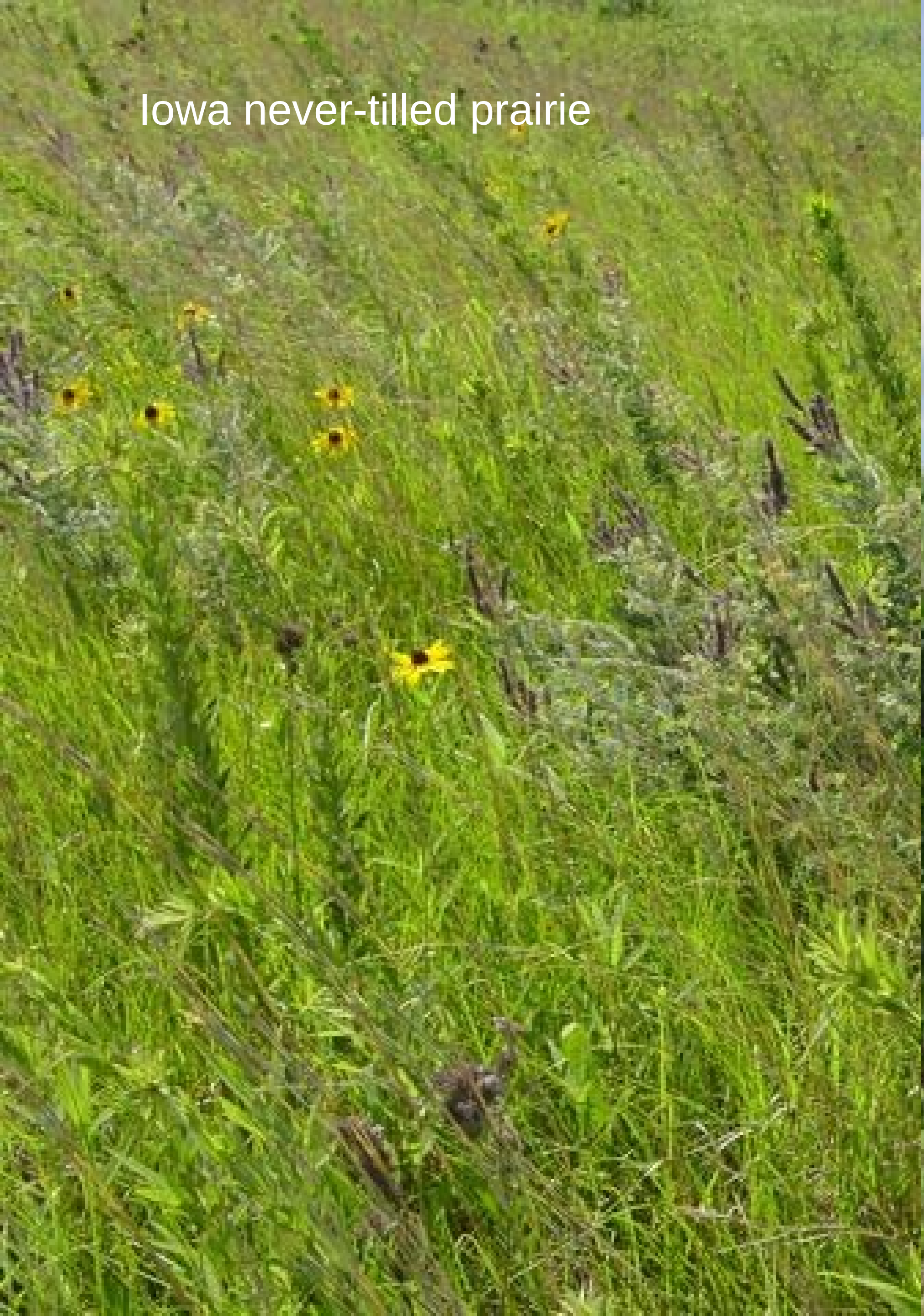
Great Prairie Soil Metagenome Project
(Jim Tiedge, Phil Hugenholtz)

Multiple samplings of 3 Different Prairie Sites

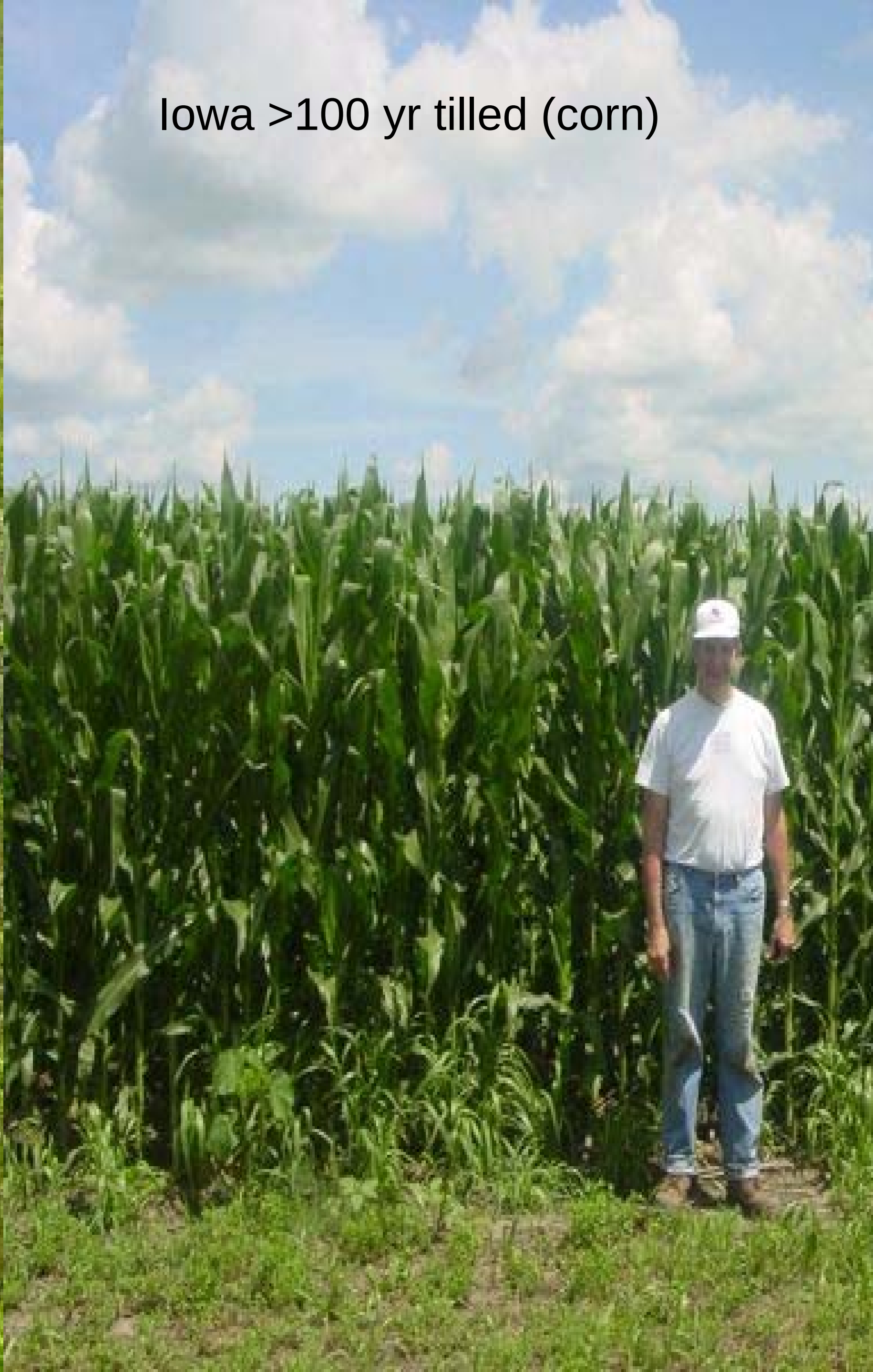
- Midwest prairie represents largest expanse of world's most fertile soils
- Sequesters the most carbon of any soil system in the U.S.
- Produces large amounts of biomass annually: Key for biofuels, carbon sequestration



©2009 Google - Map data ©2009 Tele Atlas, LeadDog Consulting - Terms of Use



Iowa never-tilled prairie



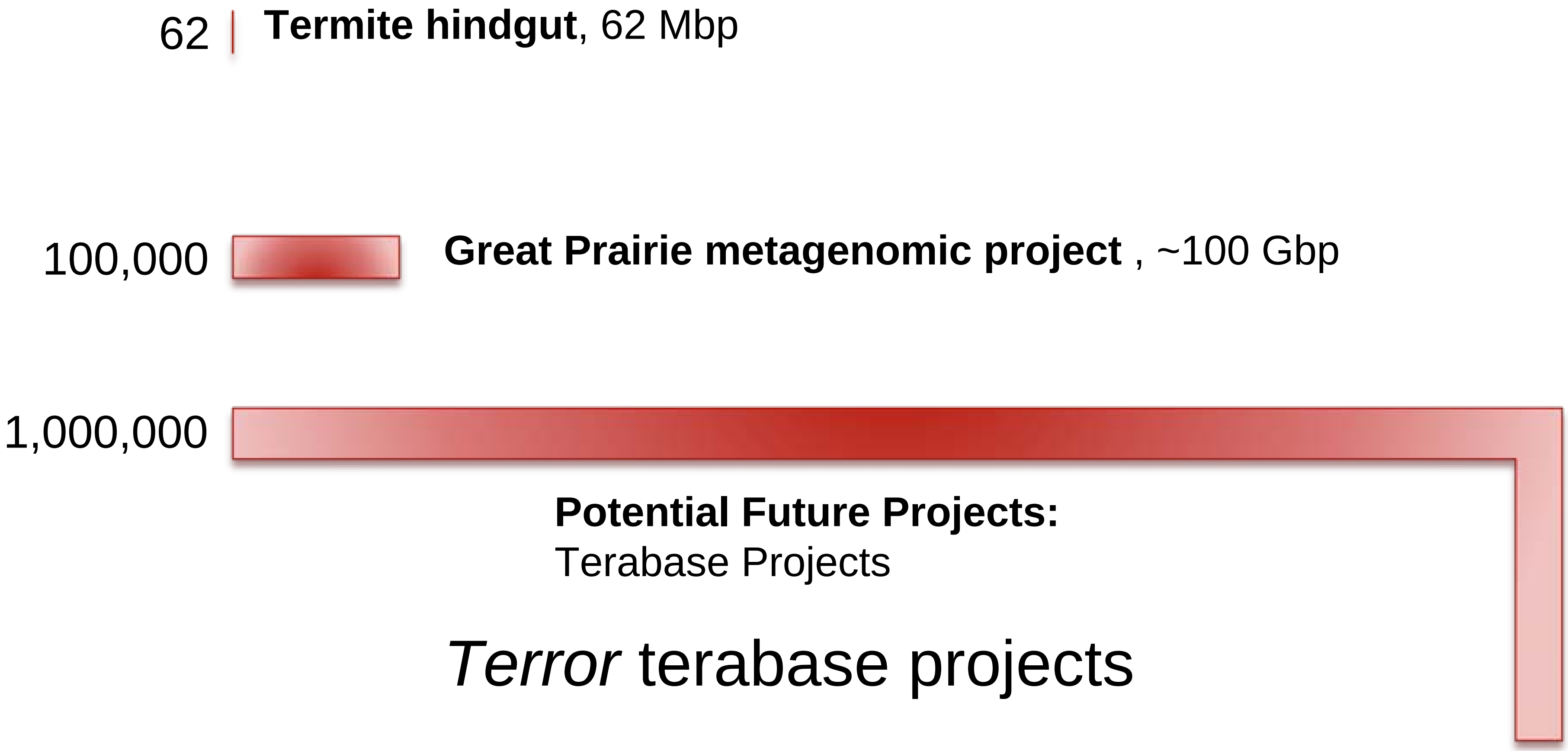
Iowa >100 yr tilled (corn)

How Do We hope to Use the Information:
Great Prairie Soil Metagenome Project

- To improve soil management, carbon sequestration,
- Through genetic understanding maybe able to potentially manage traits such as green house gas fluxes, and carbon stability

Established Switchgrass (Wisconsin)

Increasing Scale of Metagenomic Projects in the Future



Termite hindgut, 62 Mbp

Great Prairie metagenomic project , ~100 Gbp

Potential Future Projects:
Terabase Projects

Terror terabase projects

Potential foci for 1Tb Grand Challenge

- **Sequence the 12 Great Soil Orders (the major soil types)**
  - Covers the extremes in soil diversity, e.g. permafrost to tropical desert to alkali flats to taiga forest to wetlands to discover gene adaptation to environmental extremes.

Sequence the State Soil of each State

These official state soils share the same distinction as state flowers and birds.

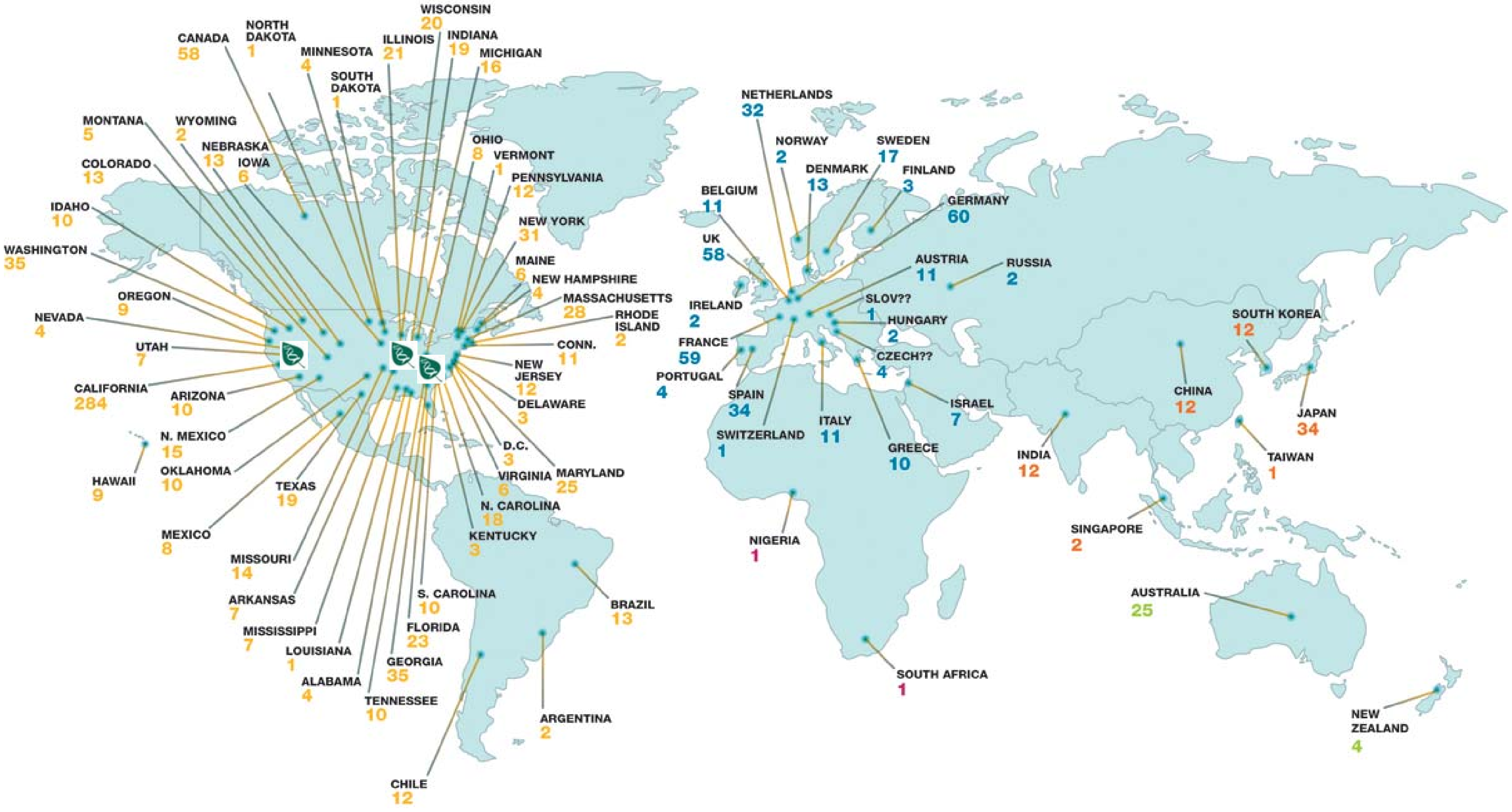
JGI User Programs

1351

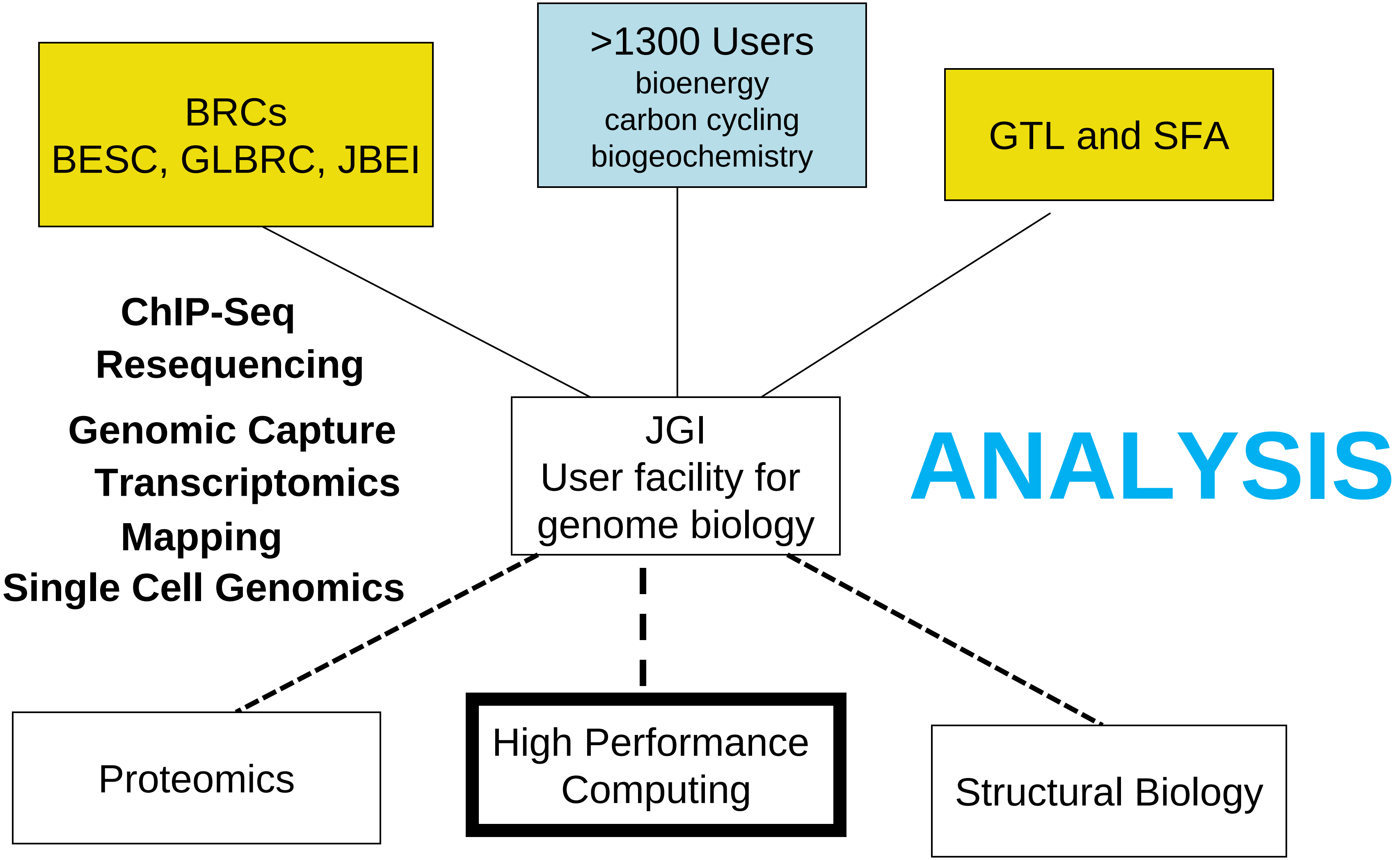
JGI Users Worldwide in 2008

AMERICAS	EUROPE	AFRICA	ASIA	AUSTRALASIA
856	336	2	64	29

DOE Bioenergy Research Centers



JGI User Program Looking Forward



>1300 Users
bioenergy
carbon cycling
biogeochemistry

BRCs
BESC, GLBRC, JBEI

GTL and SFA

ChIP-Seq

Resequencing

Genomic Capture

Transcriptomics

Mapping

Single Cell Genomics

JGI
User facility for
genome biology

ANALYSIS

Proteomics

High Performance Computing

Structural Biology

Opportunities and Challenges

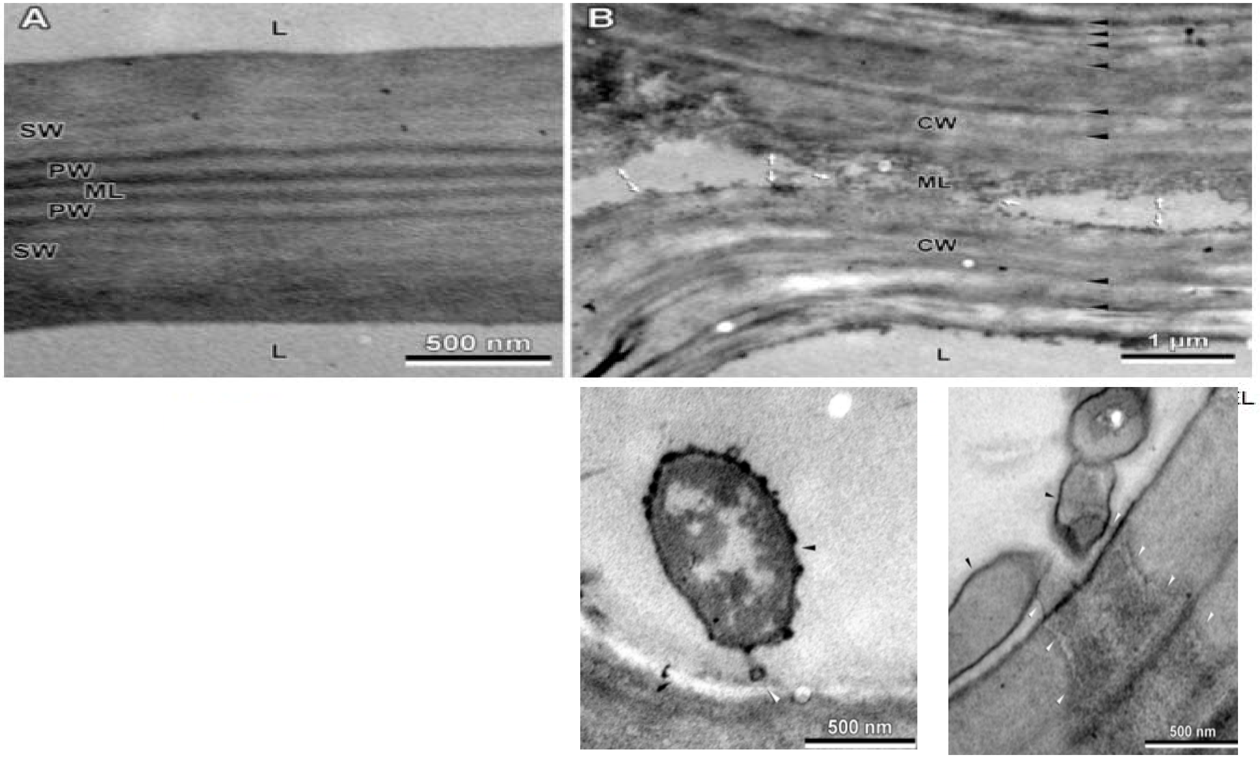
Mining for biomass deconstruction enzymes within the cow rumen

Degradation of switchgrass in the rumen (72 hr incubation)

Pre Incubation

Post Incubation

Transmission Electron Microscopy

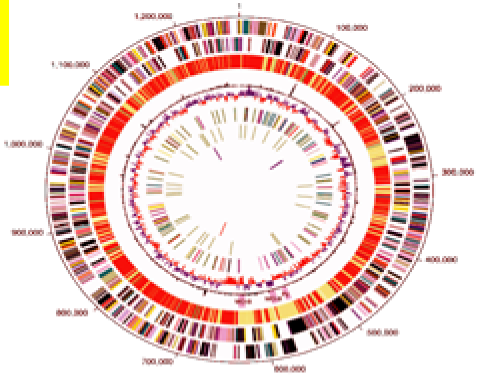


Opportunities from generating 18 GB of Data

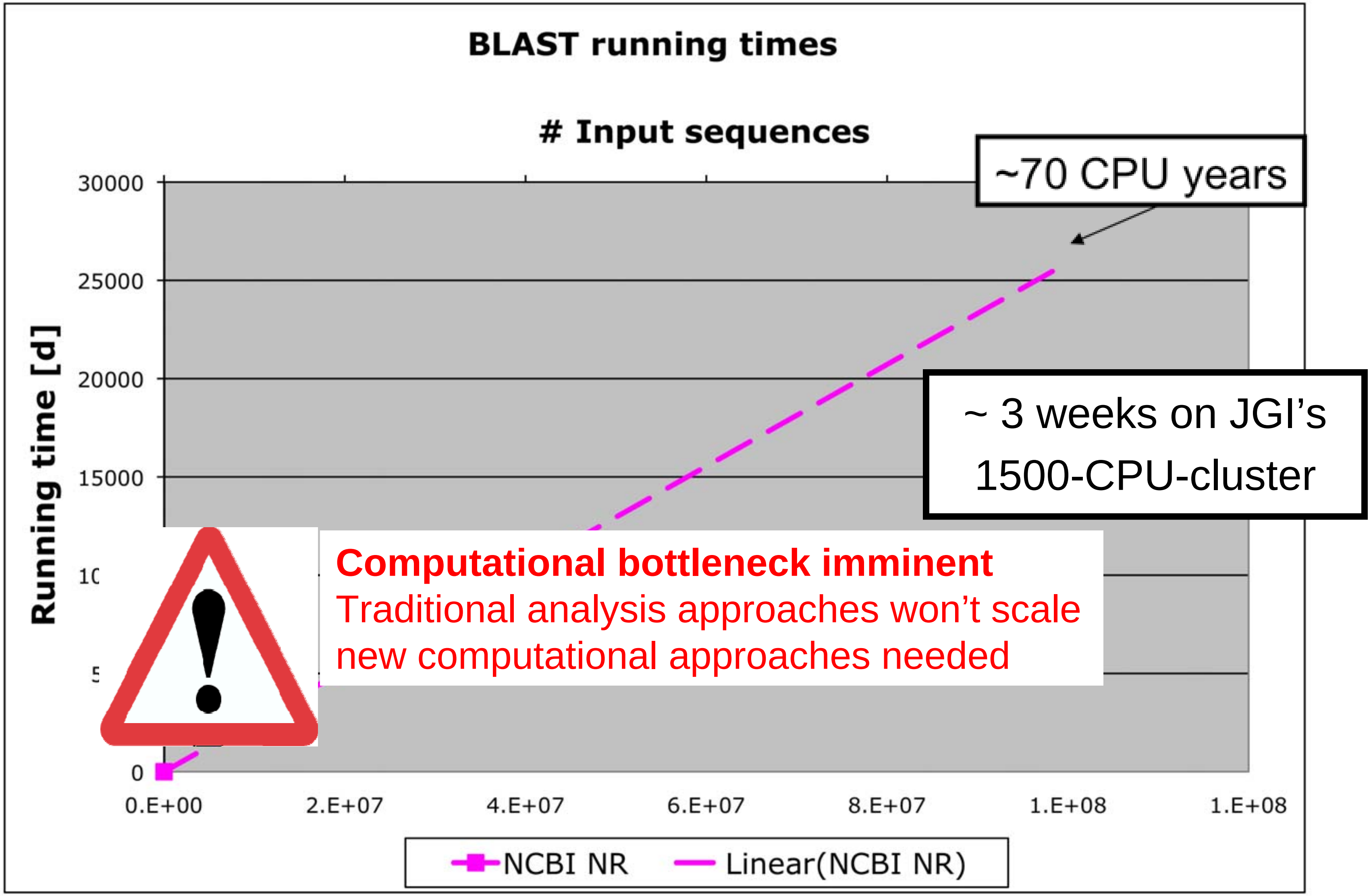
(Termite project total size: ~62 MB)

Rumen project total size: ~18 GB

Full Length Lignocellulolytic Genes



Challenge is the analysis of the 18 GB Cow Rumen Data



~70 CPU years

~ 3 weeks on JGI's 1500-CPU-cluster

Computational bottleneck imminent

Traditional analysis approaches won't scale

new computational approaches needed

Coupling the JGI’s data production with DOE Nat Lab HPC resources will advance energy and environmental science

JGI-Sponsored Workshop on HPC for Next Gen Sequencing

Focus on bringing National Lab computational infrastructure and expertise to bear on problems associated with the volume and nature of next gen sequencer out-put

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

It is not about the machines

U.S. DEPARTMENT OF ENERGY Office of Science