# Building a Community Cyberinfrastructure to Support Marine Microbial Ecology Metagenomics

**Invited Talk**
**2006 Synthetic Biology Symposium**
**Aliso Creek Inn**
**Laguna Beach, CA**
**September 15, 2006**

**Dr. Larry Smarr**
**Director, California Institute for Telecommunications and Information Technology**
**Harry E. Gruber Professor,**
**Dept. of Computer Science and Engineering**
**Jacobs School of Engineering, UCSD**

camera
Marine Microbial Ecology

it²

# Calit2 Brings Computer Scientists and Engineers Together with Biomedical Researchers

- **Some Areas of Concentration:**
  - **Metagenomics**
  - **Genomic Analysis of Organisms**
  - **Evolution of Genomes**
  - **Cancer Genomics**
  - **Human Genomic Variation & Disease**
  - **Proteomics**
  - **Mitochondrial Evolution**
  - **Computational Biology & Bioinformatics**
  - **Information Theory & Biological Systems**





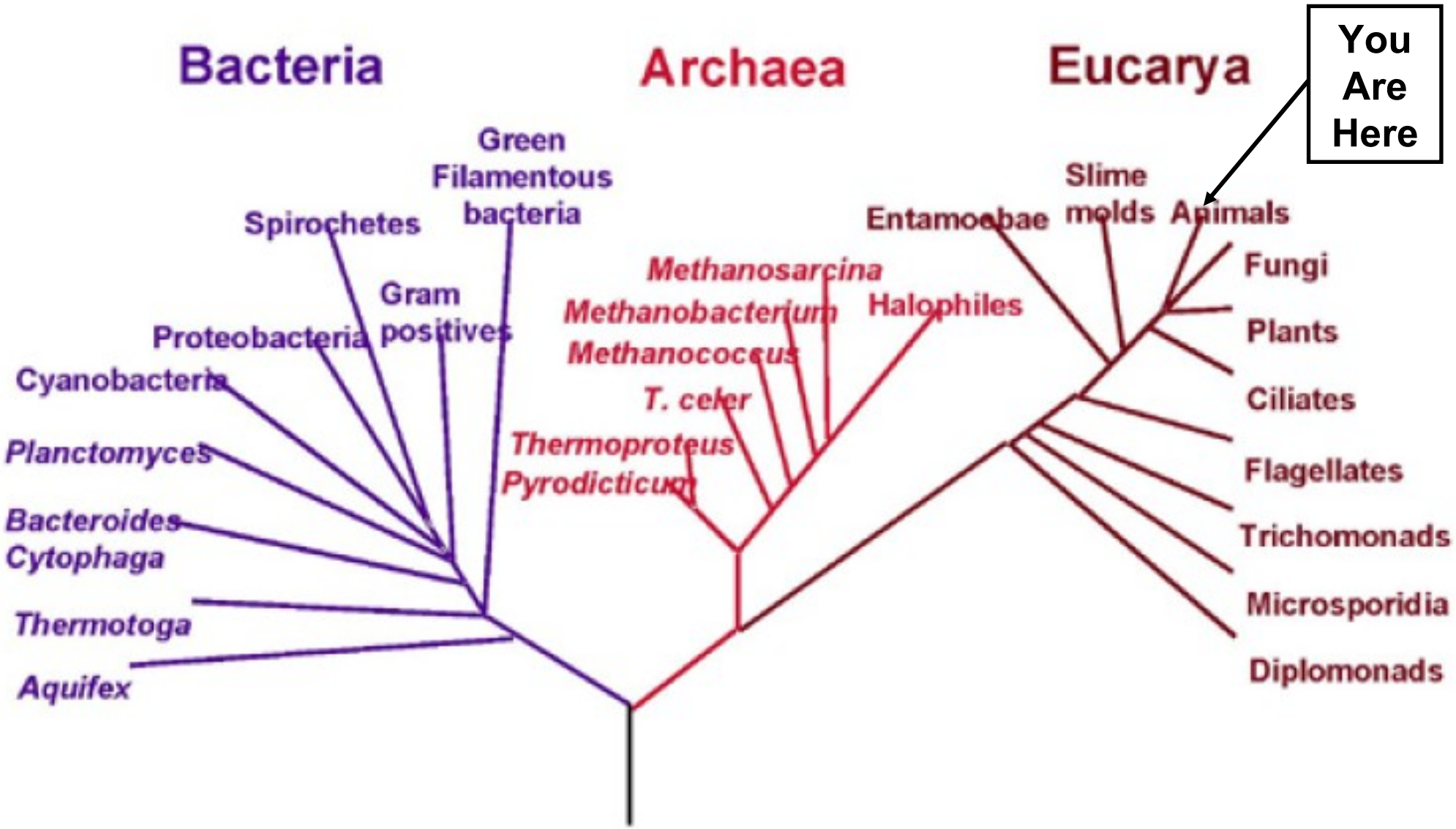
**1200 Researchers in Two Buildings**

**www.calit2.net**

camera
Marine Microbial Ecology

it²

# Most of Evolutionary Time Was in the Microbial World

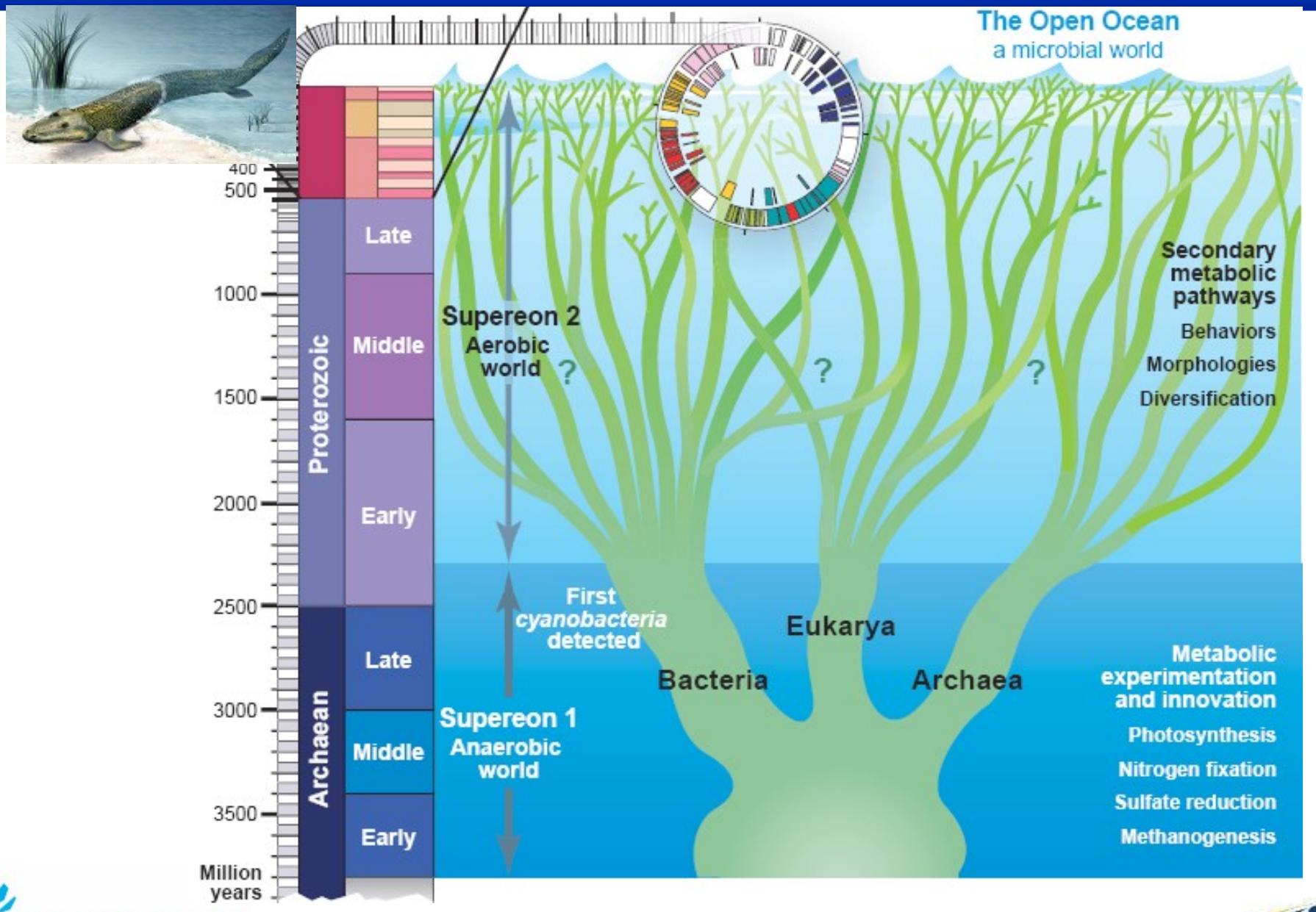



Tree of Life Derived from 16S rRNA Sequences

Source: Carl Woese, et al

camera Marine Microbial Ecology

# Microbial Genomics Let's Us Look Back Nearly 4 Billion Years In the Evolution of Life

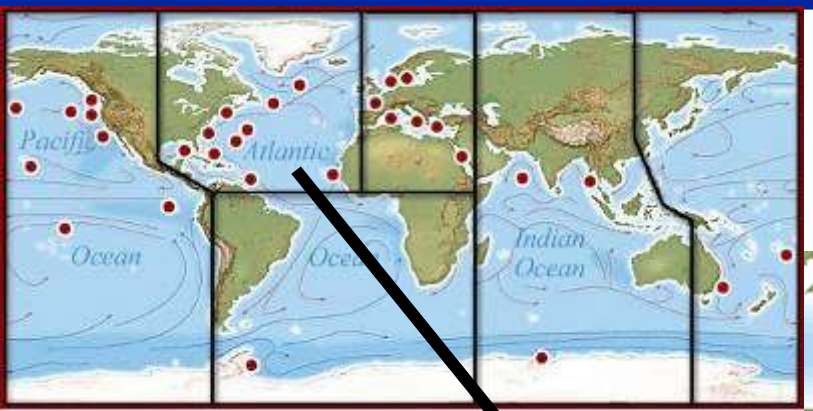



camera
Marine Microbial Ecology

Falkowski and Vargas *Science* 304 (5667) 2004

it2

# Moore Microbial Genome Sequencing Project
# Selected Microbes Throughout the World's Oceans




## Microbes Nominated by Leading Ocean Microbial Biologists

**North Atlantic**
Proteobacteria: Oligobacter sp.
Proteobacteria: Sphingomonas sp.
Proteobacteria: Loktanella vestfoldensis
Proteobacteria: Photobacterium sp.

**NE U.S. Coastal Waters**
Cyanobacteria: Microcoleus chthonoplastes
Cyanobacteria: Synechococcus sp.
Nitrospira: Nitrospira marina
Proteobacteria: Arcobacter sp.
Proteobacteria: Vibrio alginolyticus
Proteobacteria: Vibrio splendidus

**NW Atlantic**
Proteobacteria: Erythrobacter sp.

**SE U.S. Coastal Waters**
Chloroflexi: Dehalobium sp.
Proteobacteria: Roseobacter sp.
Proteobacteria: Roseovarius sp.
Proteobacteria: Sagittula stellata
Proteobacteria: Sulfitobacter sp.

**Mid-Atlantic**
Archea: Pyrolobus sp.
Archea: Thermococcus sp.
Cyanobacteria: Prochlorococcus sp.
Cyanobacteria: Synechococcus sp.
Proteobacteria: Caminibacter sp.
Proteobacteria: Sulfitobacter sp.

**Bermuda BATS Station**
Proteobacteria: Parvularcula bermudensis
Proteobacteria: Fulvimarina pelagi
Proteobacteria: Oceanicola granulosus
Proteobacteria: Oceanicola batsensis
Proteobacteria: not yet validly described
Proteobacteria: Roseovarius sp.
Proteobacteria: Oceanicaulis aliexandrii
Proteobacteria: Erythrobacter literalis
Bacteroidetes: Croceibacter atlanticus
Bacteroidetes: Robiginitalea biformata
Actinobacteria: Janibacter sp.

**Caribbean Sea**
Proteobacteria: Roseovarius nubinhibens
Cyanobacteria: Gloeothece sp.

**Gulf of Mexico**
Cyanobacteria: Synechococcus sp.
Cyanobacteria: Synechococcus sp.
Firmicutes: Bacillus sp.
Firmicutes: Bacillus sp.
Proteobacteria: Vibrionaceae sp.

**Western Tropical Atlantic**
Proteobacteria: Nitrosococcus oceani
Nitrospira: Nitrospina gracilis
Cyanobacteria: Trichodesmium thiebautii

**Eastern Tropical Atlantic**
Proteobacteria: Nitrobacter sp.
Cyanobacteria: Synechococcus sp.

MARINE MICROBIOLOGY INITIATIVE

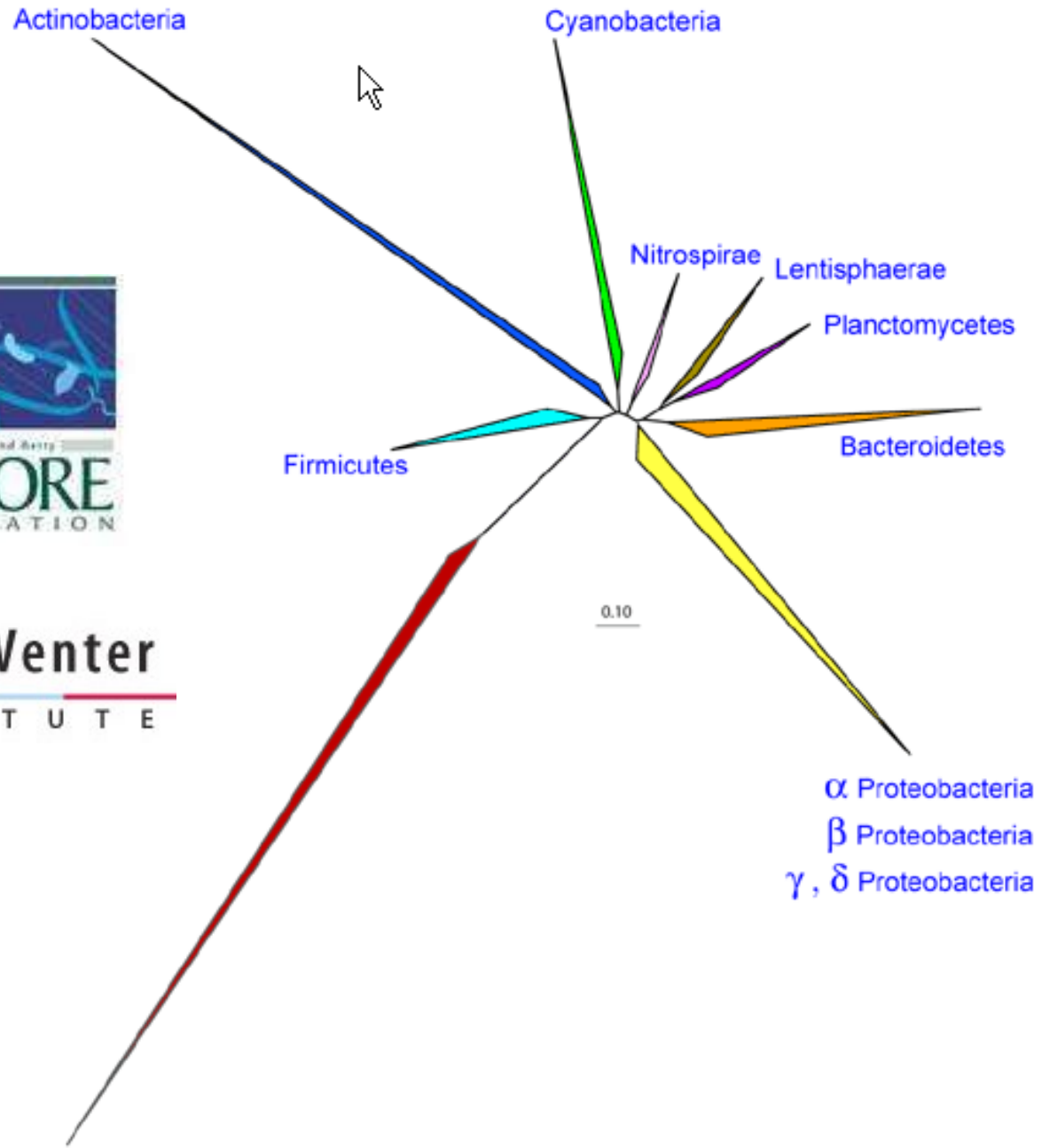
Gordon and Betty MOORE FOUNDATION

camera Marine Microbial Ecology

www.moore.org/microgenome/worldmap.asp

it2

# Moore Foundation Funded the Venter Institute to Provide the Full Genome Sequence of 150 Marine Microbes




MARINE MICROBIOLOGY INITIATIVE

Gordon and Betty MOORE FOUNDATION

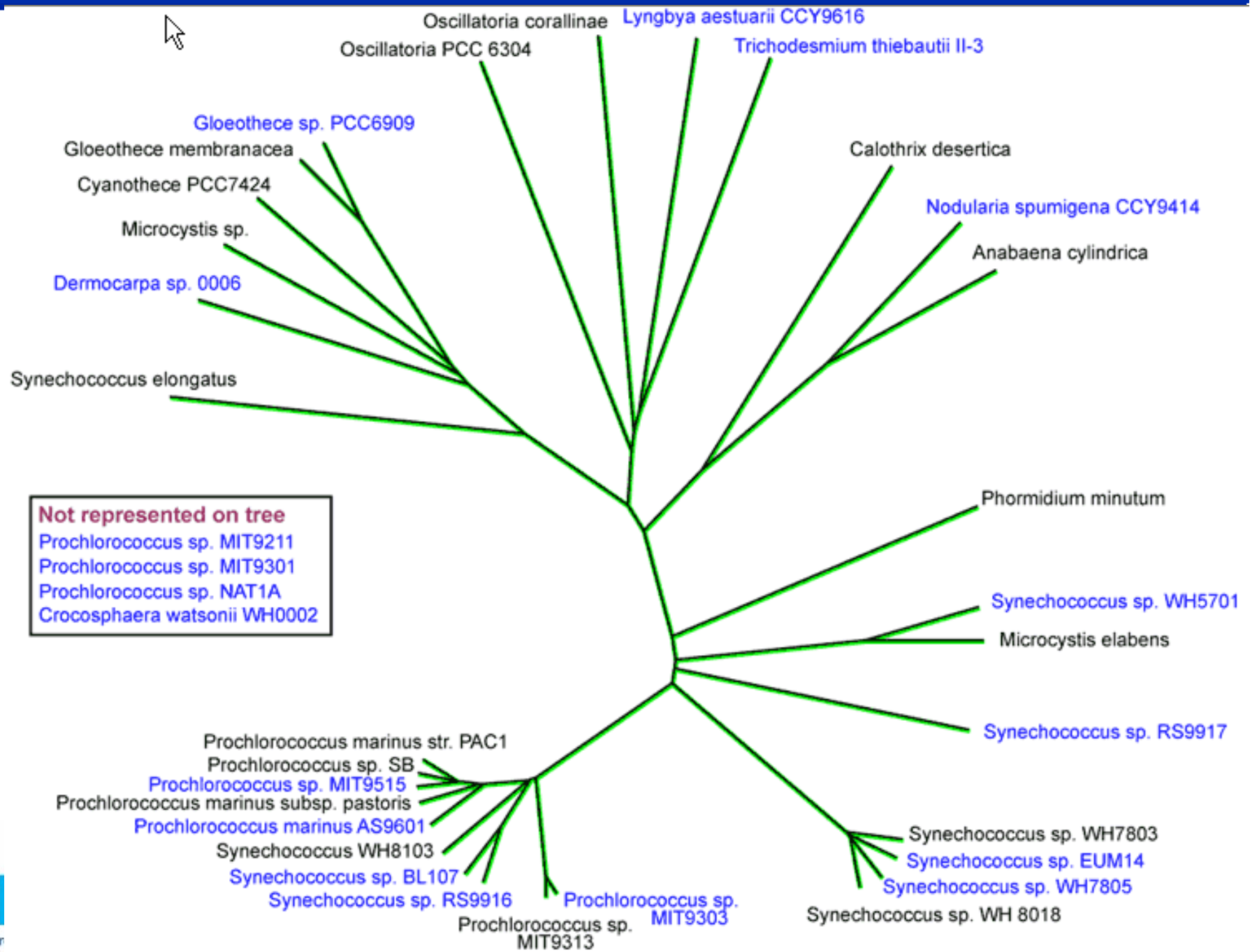
J. Craig Venter INSTITUTE

cam Marine Microbial Ec

www.moore.org/microgenome/trees_main.asp

it2

# Moore Microbial Genome Sequencing Project: Cyanobacteria Being Sequenced by Venter Institute




Marine Micr

# Full Genome Sequencing is Exploding: Most Sequenced Genomes are Bacterial

First Genome 1995
6 Genomes/ Year 2000

Completed Genon

Ongoing Genomes



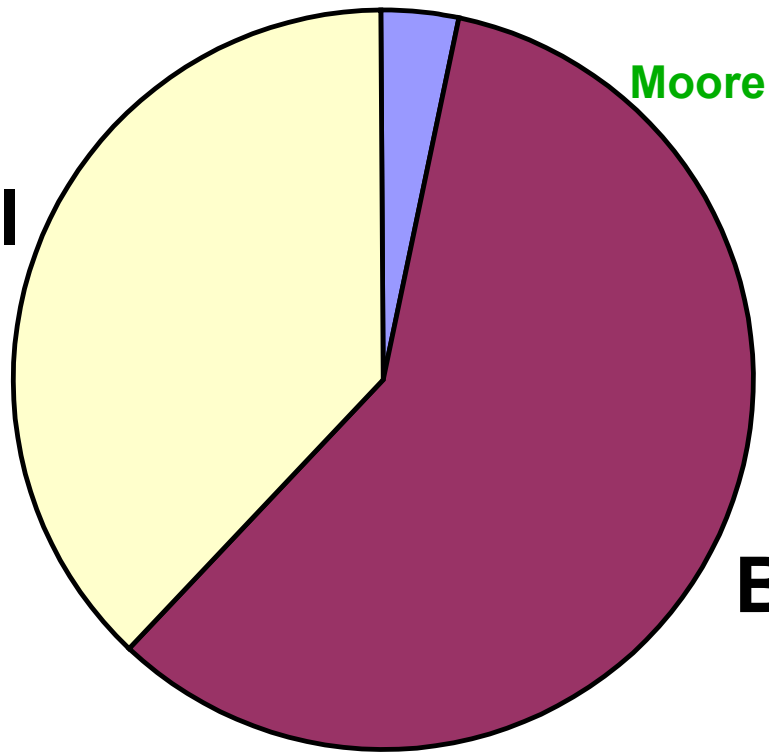

Total 422

55 Metagenomes

Archaeal

Moore 155 In Here

Eukaryal



Bacterial

Total 1665

www.genomesonline.org

# Microbial Metagenomics is a Rapidly Emerging Field of Research

**"Despite their ubiquity, relatively little is known about the majority of environmental microorganisms, largely because of their resistance to culture under standard laboratory conditions."**

**"The application of high-throughput shotgun sequencing environmental samples has recently provided global views of those communities not obtainable from 16S rRNA or BAC clone–sequencing surveys ."**

***Comparative Metagenomics of Microbial Communities***

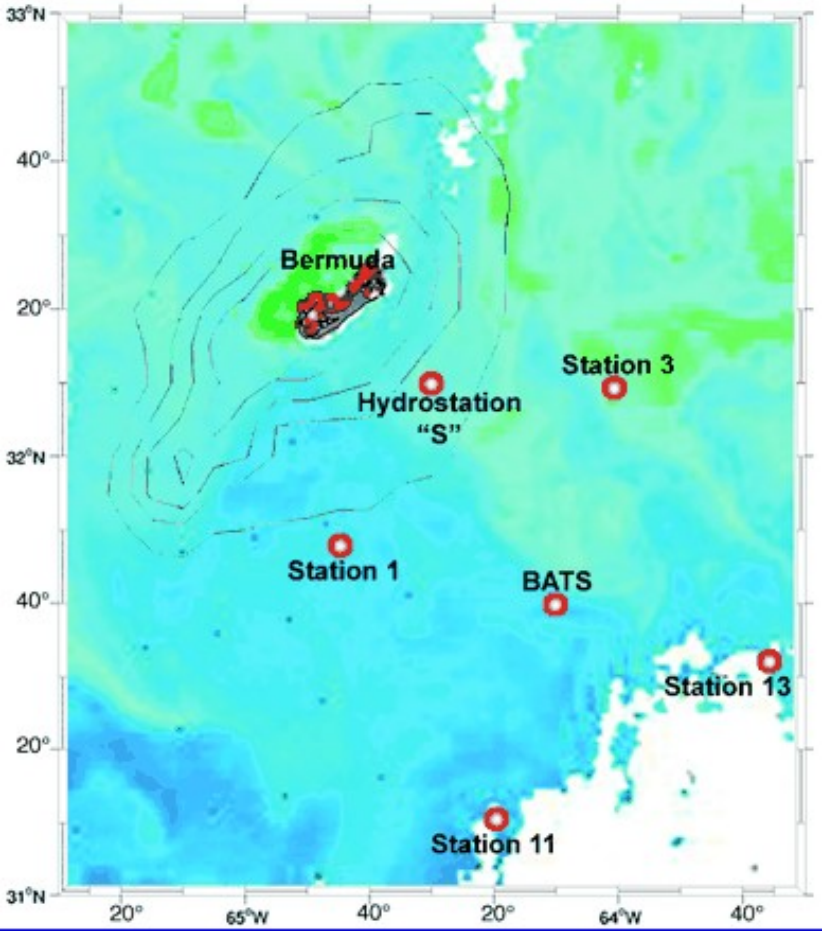
**Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A. Salamov, Kevin Chen, Hwai W. Chang, Mircea Podar, Jay M. Short, Eric J. Mathur, John C. Detter, Peer Bork, Philip Hugenholtz, Edward M. Rubin**

**Science 22 April 2005**

camera
Marine Microbial Ecology

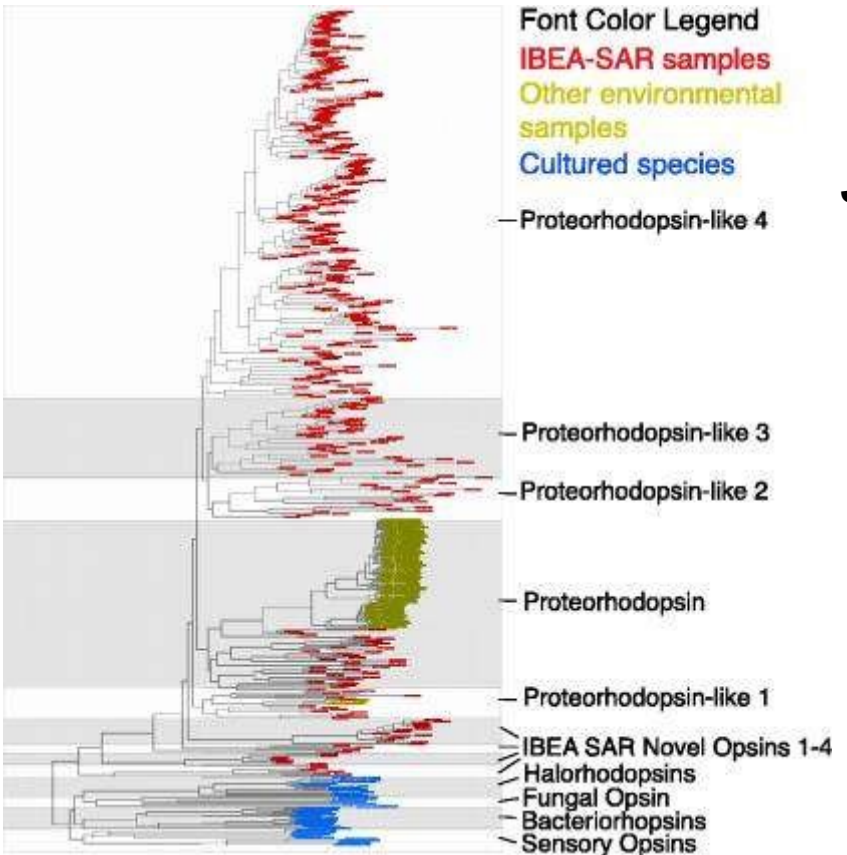
it2

# The Sargasso Sea Experiment
# The Power of Environmental Metagenomics




**MODIS-Aqua satellite image of ocean chlorophyll in the Sargasso Sea grid about the BATS site from 22 February 2003**

- **Yielded a Total of Over 1 billion Base Pairs of Non-Redundant Sequence**
- **Displayed the Gene Content, Diversity, & Relative Abundance of the Organisms**
- **Sequences from at Least 1800 Genomic Species, including 148 Previously Unknown**
- **Identified over 1.2 Million Unknown Genes**




**J. Craig Venter, et al. Science 2 April 2004: Vol. 304. pp. 66 - 74**

camera
Marine Microbial Ecology

# Marine Genome Sequencing Project – Measuring the Genetic Diversity of Ocean Microbes










MARINE MICROBIOLOGY INITIATIVE

camera
Marine Microbial Ecology

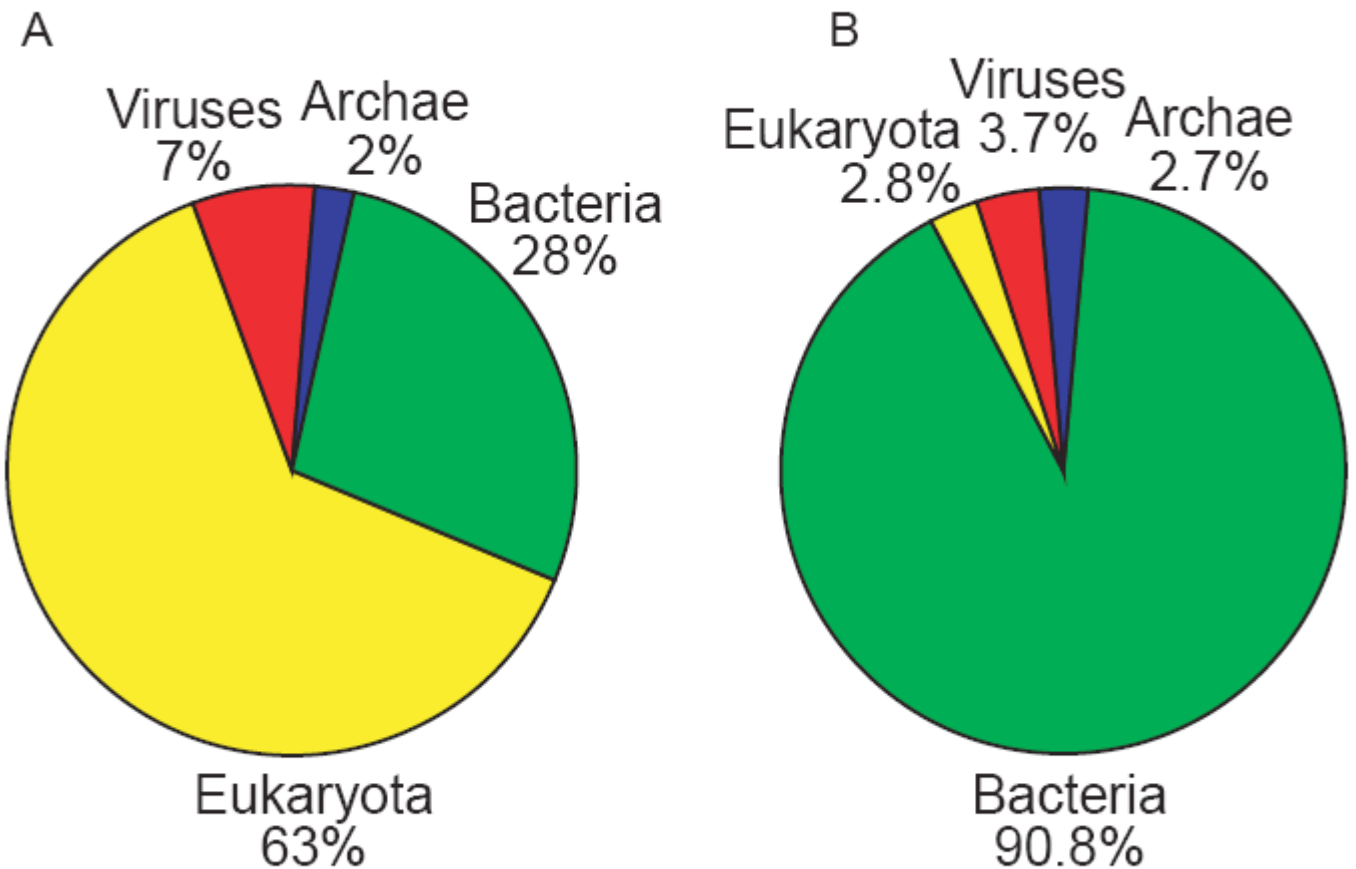
Gordon and Betty MOORE FOUNDATION

**Sorcerer II Data Will Double Number of Proteins in GenBank!**

J. Craig Venter INSTITUTE

it²

# GOS Sequences are Largely Bacterial




**~3 Million Previously Known Sequences**

**~5.6 Million GOS Sequences**

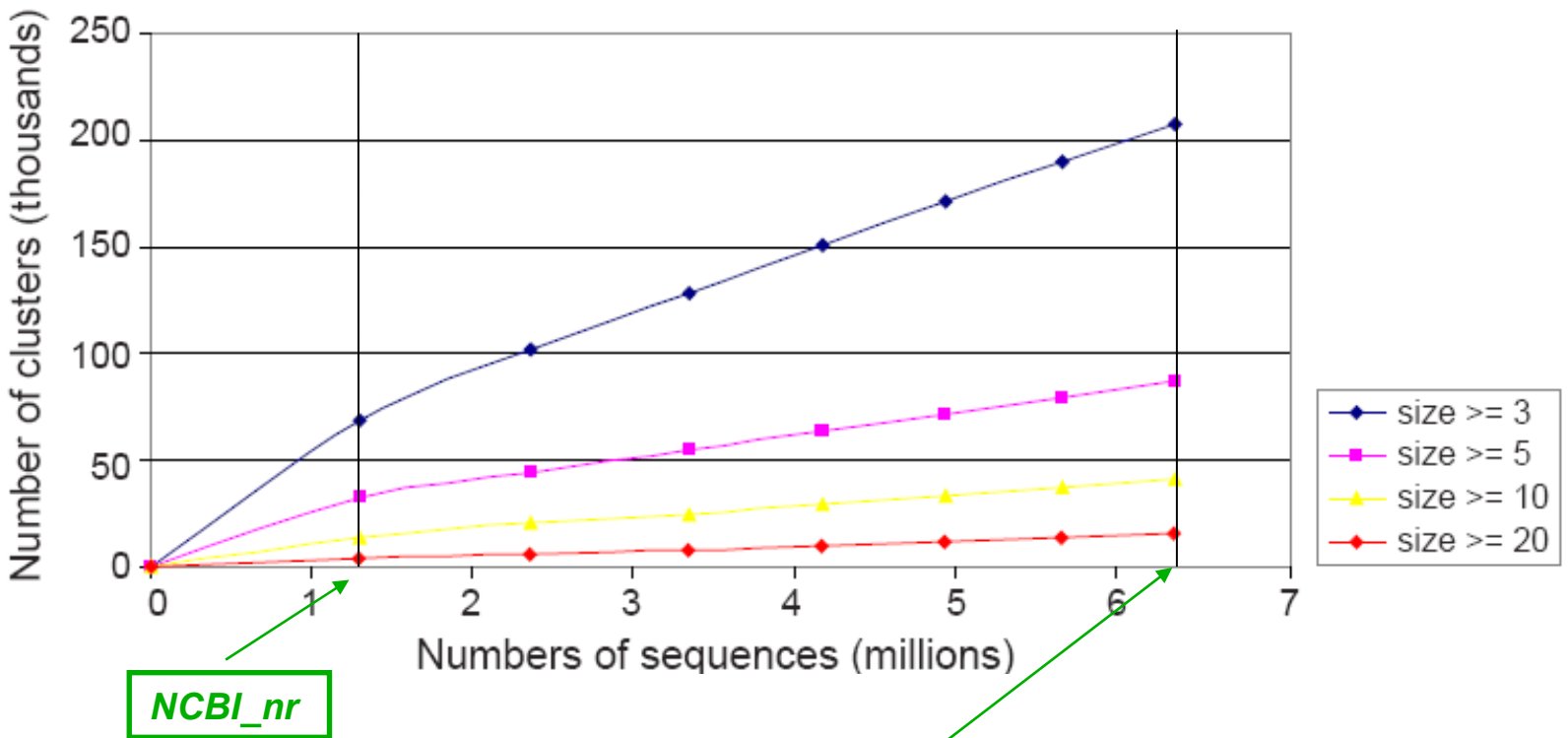
camera Marine Microbial Ecology

**Source: Shibu Yooseph, et al. (PLOS Biology in press 2006)**

it2

# GOS Analysis -- Protein Families in Nature Have Been Poorly Explored Thus Far

- Novel Sequence Similarity Clustering Process Predicts Proteins and Groups Related Sequences Into Clusters (Families)
- GOS Proteins Increase Size / Diversity of Many Protein Families
- 1,700 Novel GOS-Only Clusters Identified (>20 per Cluster)



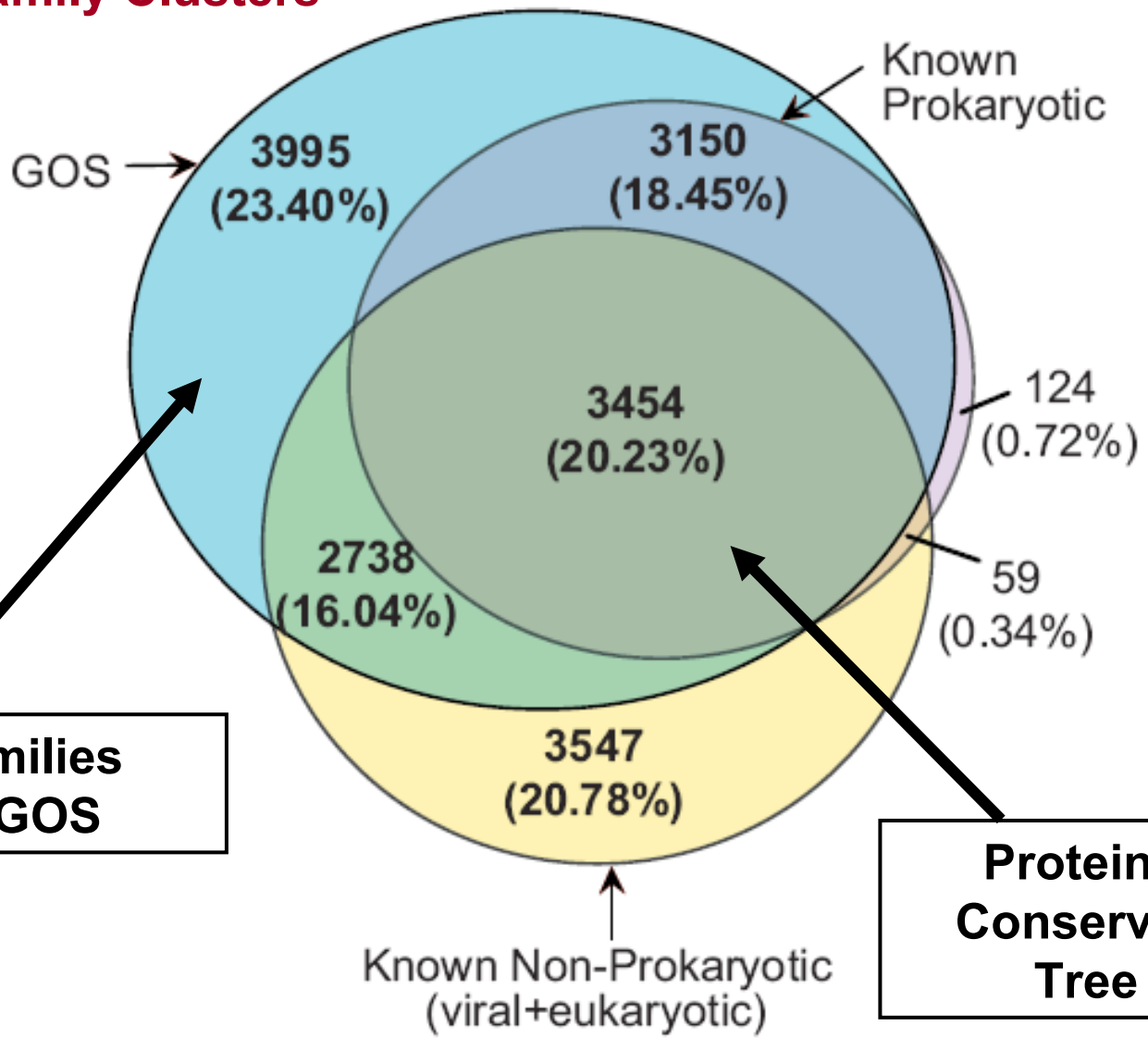

camera Marine Microbial Ecology

Source: Shibu Yooseph, Granger Sutton, --JCVI

it2

# Current Universe of Medium/ Large Protein Families

**17,067 Protein Family Clusters**




**Protein Families Unique to GOS**

**Protein Families Conserved Across Tree of Life**

camera Marine Microbial Ecology

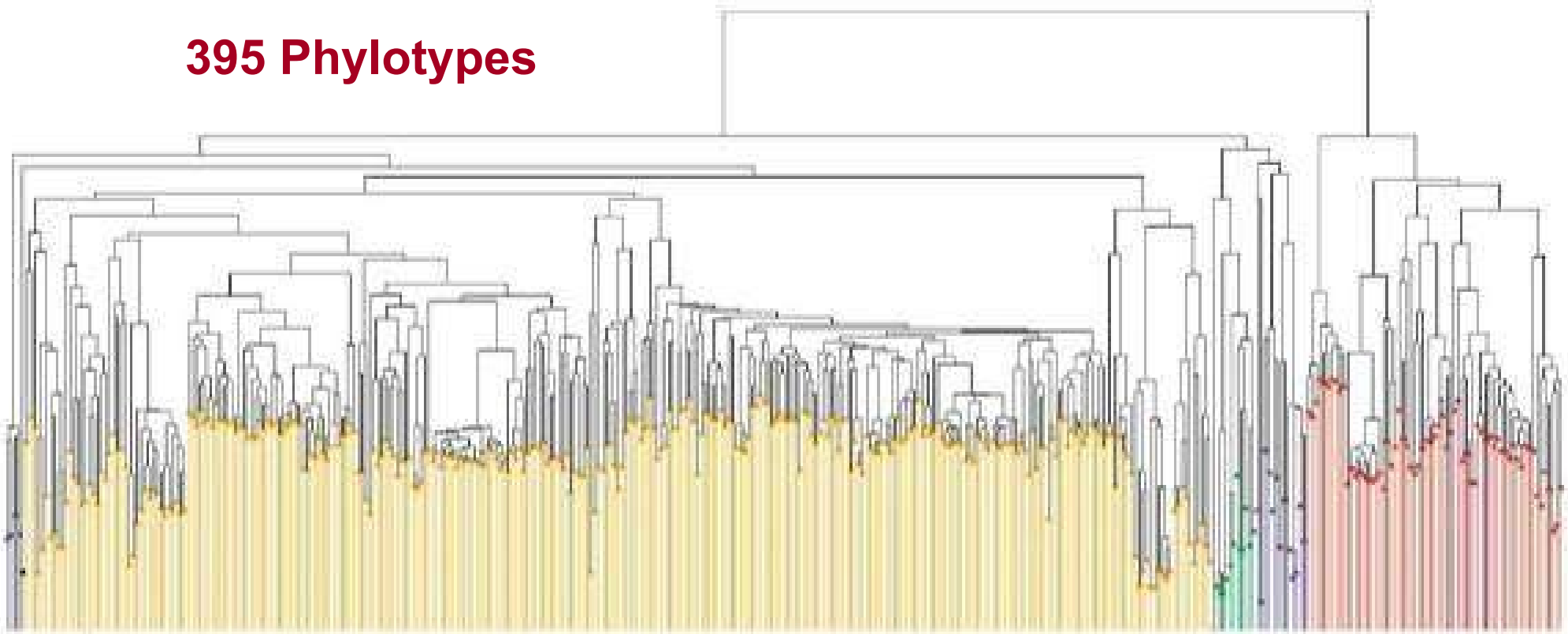
**Source: Shibu Yooseph, et al. (PLOS Biology in press 2006)**

it2

# Metagenomic Data Sets Are Rapidly Being Accumulated

- "A majority of the bacterial sequences corresponded to uncultivated species and novel microorganisms."
- "We discovered significant inter-subject variability."
- "Characterization of this immensely diverse ecosystem is the first step in elucidating its role in health and disease."

**395 Phylotypes**

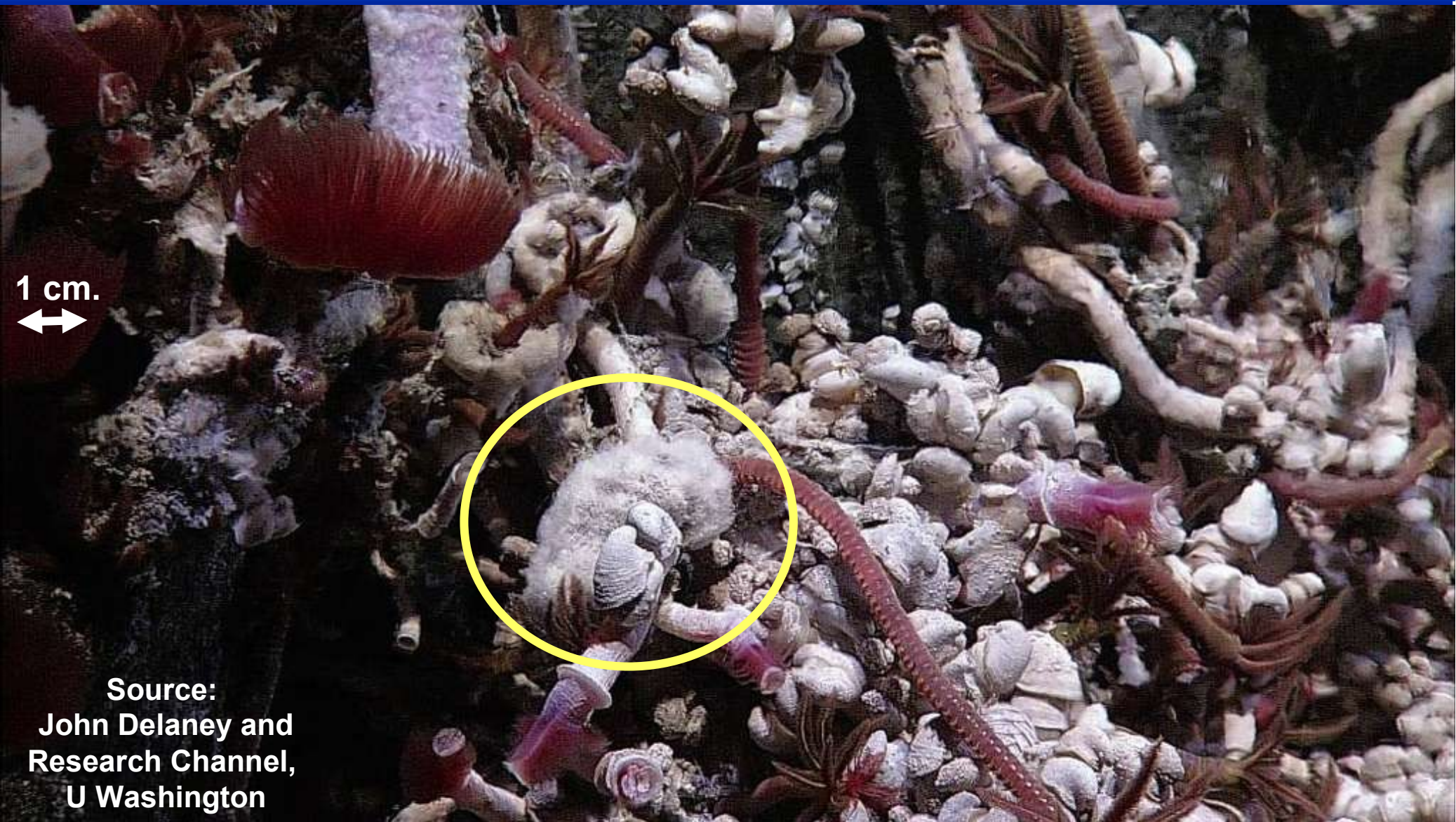


camera Marine Microbial Ecology

"Diversity of the Human Intestinal Microbial Flora"
Paul B. Eckburg, et al *Science* (10 June 2005)

it²

# Microbes Form the Base of the Living World




**White Filamentous Bacteria on 'Pill Bug' Outer Carapace**

**High Definition Still Frame**
**of Hydrothermal Vent Ecology 2.3 Km Deep**

camera
Marine Microbial Ecology

it2

# CAMERA: Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

## National LambdaRail
## Direct Connect
## Computation and Storage Complex

Funded by: Gordon and Betty Moore Foundation

**PI Larry Smarr**

Joint Partnership of:

camera
Marine Microbial Ecology

**Announced January 17, 2006**
**$24.5M Over Seven Years**

# Paul Gilna Has Been Recruited from Los Alamos to Become Calit2's Executive Director of CAMERA

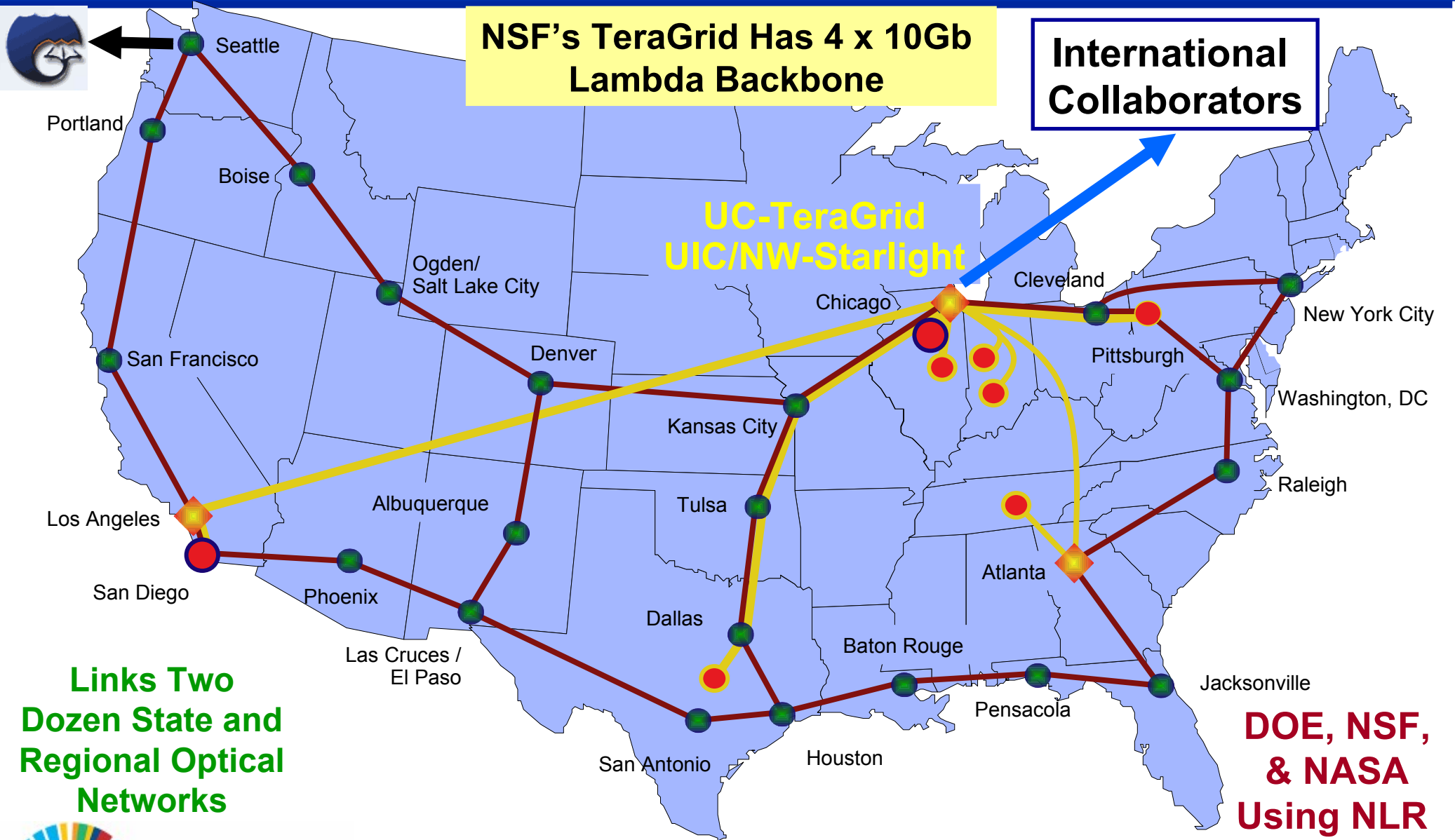
- **Formerly**
  - Former Director of the Department of Energy's Joint Genome Institute (JGI) Operations at Los Alamos National Laboratory (LANL)
  - Group Leader of Genomic Science and Computational Biology in LANL's Bioscience Division
- **JGI**
  - A $70-million-per-Year Collaboration:
    - Lawrence Berkeley,
    - Lawrence Livermore,
    - Los Alamos,
    - Oak Ridge, and
    - Pacific Northwest
    - and the Stanford Human Genome Center
  - Working at The Frontiers of Genome Sequencing and Biosciences



camera
Marine Microbial Ecology

it2

# National Lambda Rail (NLR) and TeraGrid Provides Cyberinfrastructure Backbone for U.S. Researchers




camer
Marine Microbial Ecology

# Calit2's Direct Access Core Architecture Will Create Next Generation Metagenomics Server

**Sargasso Sea Data**

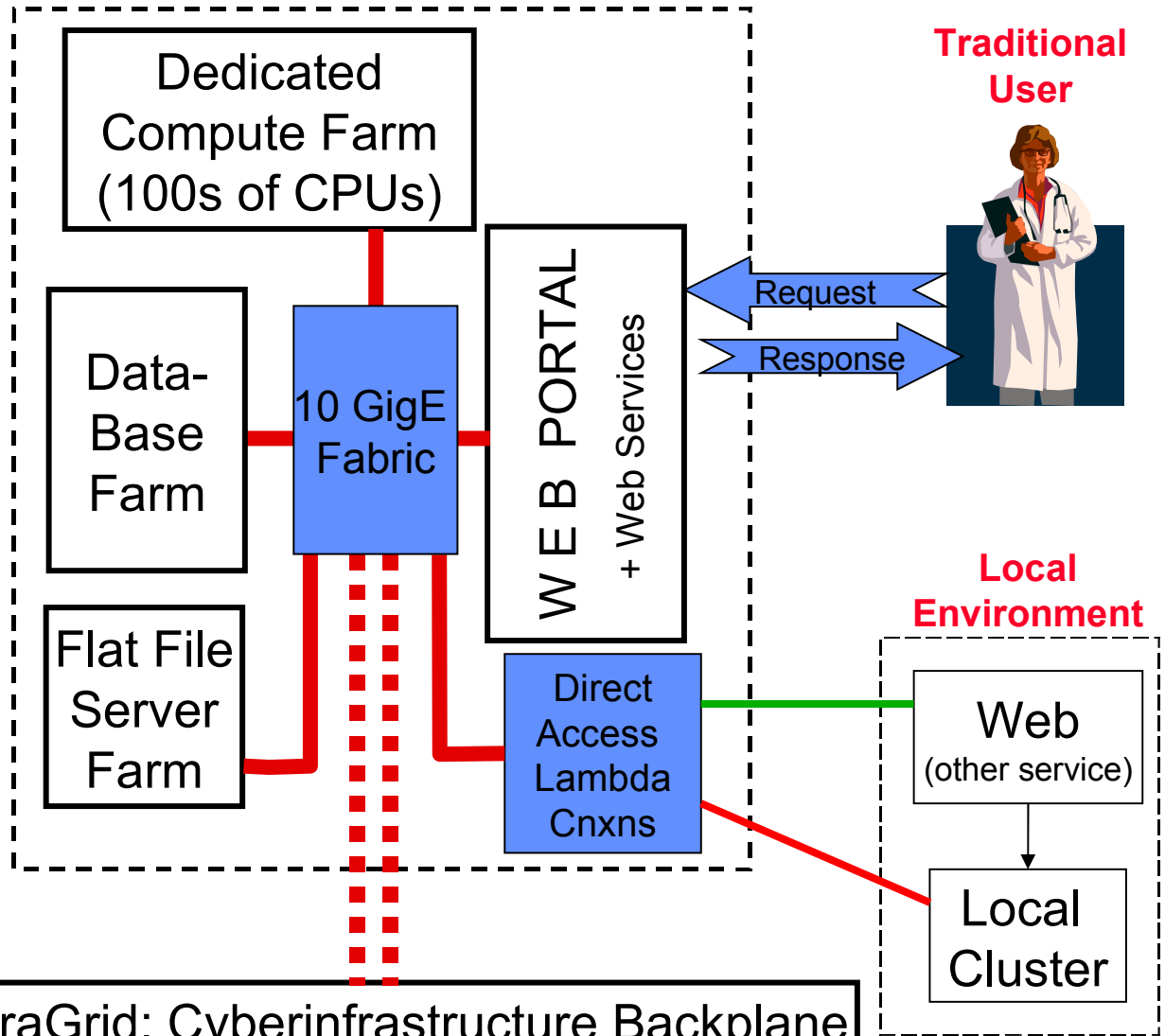
**Sorcerer II Expedition (GOS)**

**JGI Community Sequencing Project**

**Moore Marine Microbial Project**

**NASA and NOAA Satellite Data**

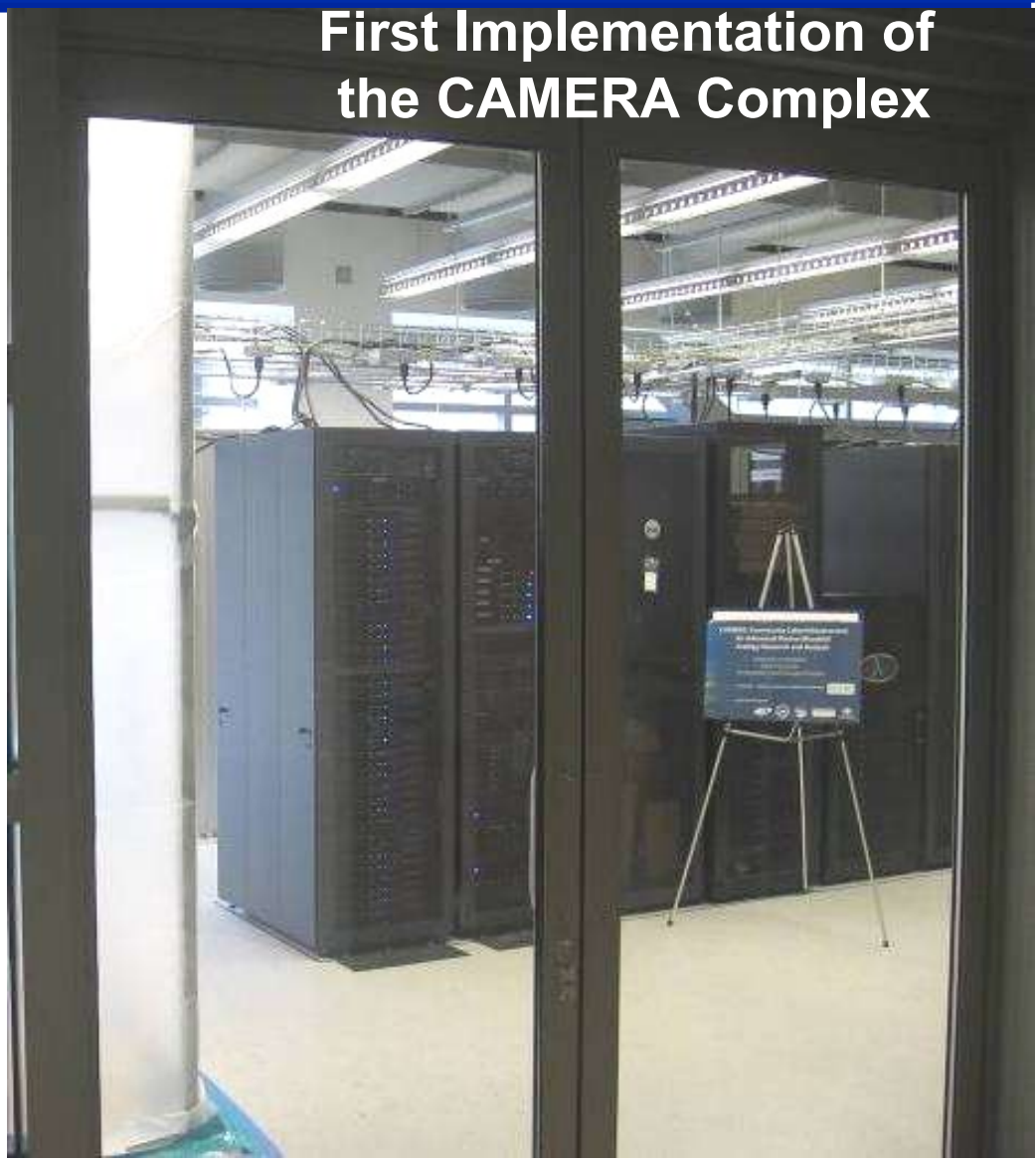
**Community Microbial Metagenomics Data**




**Source: Phil Papadopoulos, SDSC, Calit2**

# The Future Home of the Moore Foundation Funded Marine Microbial Ecology Metagenomics Complex




camera
Marine Microbial Ecology

Photo Courtesy Joe Keefe, Calit2

it2

# Analysis Data Sets, Data Services, Tools, and Workflows

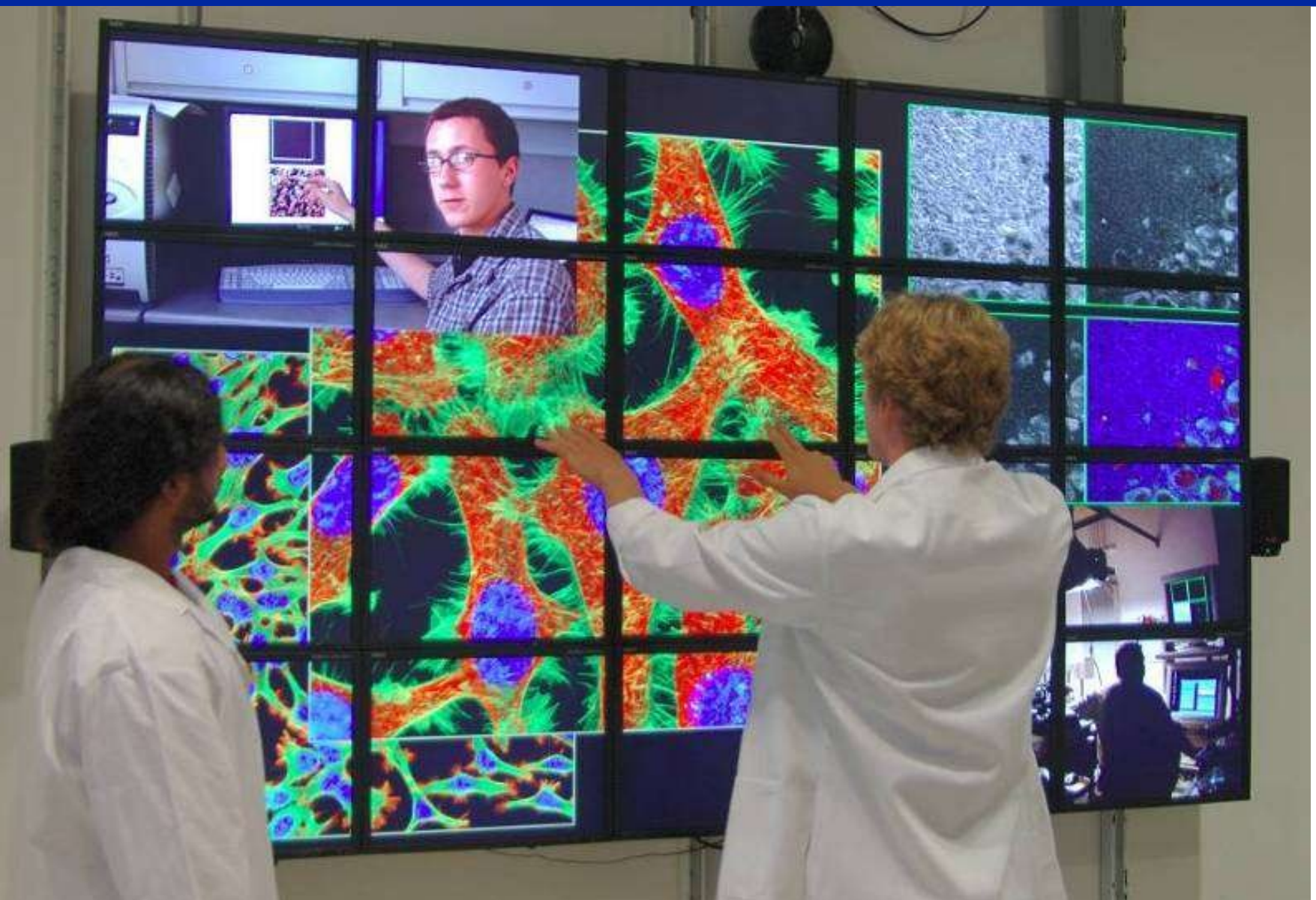
- Assemblies of Metagenomic Data
  - e.g, GOS, JGI CSP
- Annotations
  - Genomic and Metagenomic Data
- "All-against-all" Alignments of ORFs
  - Updated Periodically
- Gene Clusters and Associated Data
  - Profiles, Multiple-Sequence Alignments,
  - HMMs, Phylogenies, Peptide Sequences
- Data Services
  - 'Raw' and Specialized Analysis Data
  - Rich Query Facilities
- Tools and Workflows
  - Navigate and Sift Raw and Analysis Data
  - Publish Workflows and Develop New Ones
  - Prioritize Features via Dialogue with Community

camera
Marine Microbial Ecology

**Source: Saul Kravitz**
**Director of Software Engineering**
**J. Craig Venter Institute**

it2

# OptIPortal–Termination Device for the Dedicated Gigabit/sec Lightpaths



**Collaborative Analysis of Large Scale Images of Cancer Cells**

**Integration of High Definition Video Streams with Large Scale Image Display Walls**

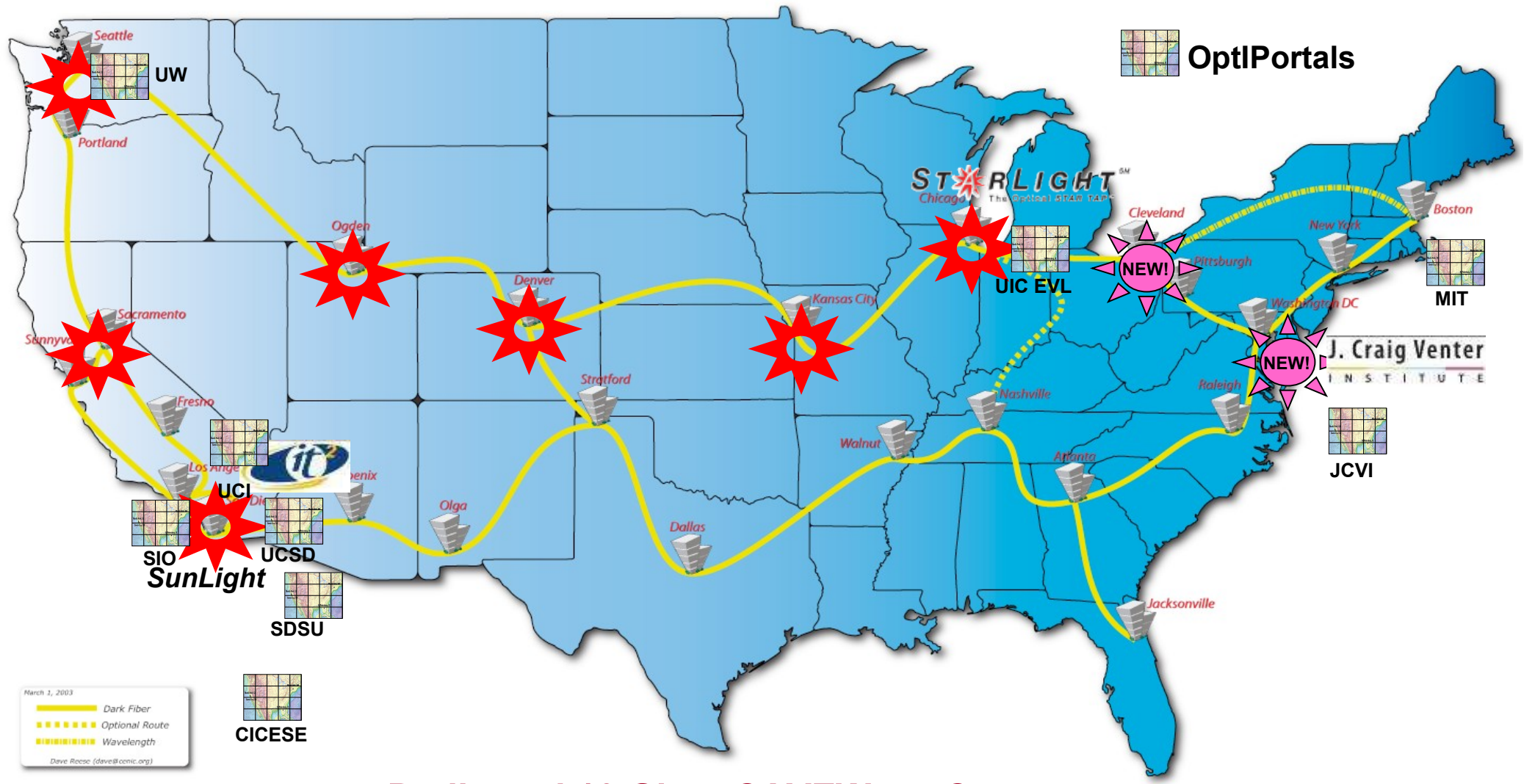
calit2 electronic visualization laboratory
Marine Microbial Ecology

**Photo Source: David Lee, Mark Ellisman NCMIR, UCSD**

NCMIR NATIONAL CENTER for MICROSCOPY and IMAGING RESEARCH
at San Diego, an NIH supported resource center

it2

# Emerging OptIPortal Sites on the National LambdaRail




**Dedicated 10 Gbps CAVEWave Connects**
**San Diego to Seattle to Chicago to Washington D.C.**

camera
Marine Microbial Ecology

# CAMERA Outreach Modes

- **Scientific Advisory Board**
  - Early Adopters – OptIPortal End Points
- **Targeted Workshops**
  - User Forums
  - User Software Testing
  - Viz Tool Brainstorming
- **Presentations at Scientific Meetings**
  - e.g. Demonstration Booth at JCVI Genomes, Medicine, and the Environment Conference October 2006
- **Partnerships With Metagenomics Projects**
  - E.g. DoE's Joint Genome Institute (JGI)
- **Training and User Services Team**

camera
Marine Microbial Ecology

it²

# Timeline: Sprint and Marathon

- **Sprint**
  - **Release 0.0: April 2006**
    - **Test Cluster for UCSD/JCVI Collaboration**
  - **Release 1.0: Late Fall 2006**
    - **Initial Data and Core Tools Release**
    - **Supports Publication of GOS Papers**
- **Marathon**
  - **Release 2.0: Fall 2007**
    - **Additional/Improved Tools & Better Usability**
  - **Beyond 2.0**
    - **Move Towards Semantic DB**
    - **Additional Tools Based on Community Feedback**

camera
Marine Microbial Ecology

it2