# Microbial Metagenomics Drives a New Cyberinfrastructure

**Invited Talk**
**School of Biological Sciences**
**University of California, Irvine**
**March 3, 2006**

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

UCSD
UCIrvine

it²

# Abstract

**Calit2, in partnership with J. Craig Venter Institute in Rockville, MD, and UCSD's Center for Earth Observations and Applications at Scripps Institution of Oceanography, will build a state-of-the-art computational resource and develop software tools to decipher the genetic code of communities of microbial life in the world's oceans. The Gordon and Betty Moore Foundation has awarded $24.5 million over seven years to create the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA). Scientists will use CAMERA for metagenomics research -- analyzing microbial genomic sequence data in the context of other microbial species, as well as in comparison to a variety of other "metadata" such as the chemical and physical conditions in which microbes are sampled. The CAMERA project will contain the results of the Venter Institute's Sorcerer II Expedition, which carried out the first large-scale genomic survey of microbial life in the world's oceans to produce the largest gene catalogue ever assembled. Sorcerer II is expected to more than double the number of protein sequences currently available in the National Institutes of Health's GenBank. In addition to Sorcerer II's ecological genomic data, the CAMERA database will be augmented by the full genomes of more than 150 critical marine microbes enabling new comparative genomics studies.**

UCSD
UCIrvine

# 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 and Disease
  - Mitochondrial Evolution
  - Proteomics
  - Computational Biology
  - Information Theory and Biological Systems



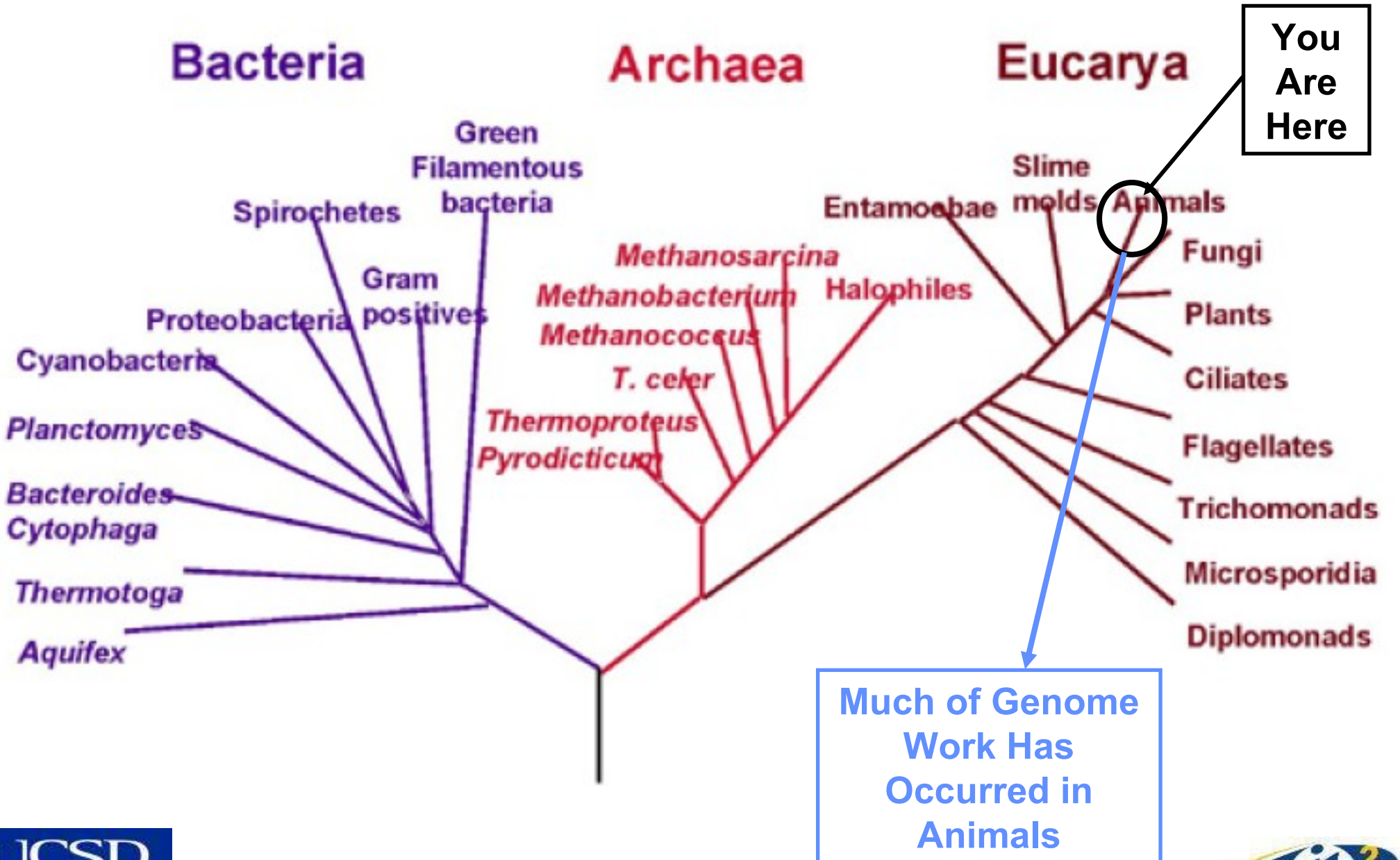




**1200 Researchers in Two Buildings**

UCSD
UCIrvine

# Evolution is the Principle of Biological Systems: Most of Evolutionary Time Was in the Microbial World




Source: Carl Woese, et al

UCSD
UCIrvine

it2

# Calit2 Researcher Eskin Collaborates with Perlegen Sciences on Map of Human Genetic Variation Across Populations




**"We have characterized whole-genome patterns of common human DNA variation by genotyping 1,586,383 single-nucleotide polymorphisms (SNPs) in 71 Americans of European, African, and Asian ancestry."**

David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, **Eleazar Eskin**, Dennis G. Ballinger, Kelly A. Frazer, David R. Cox.
"Whole-Genome Patterns of Common DNA Variation in Three Human Populations"
Science 18 February, 2005: 307(5712):1072-1079.

**"Although knowledge of a single genetic risk factor can seldom be used to predict the treatment outcome of a common disease, knowledge of a large fraction of all the major genetic risk factors contributing to a treatment response or common disease could have immediate utility, allowing existing treatment options to be matched to individual patients without requiring additional knowledge of the mechanisms by which the genetic differences lead to different outcomes ."**

**"More detailed haplotype analysis results are available at http://research.calit2.net/hap/wgha/ "**

# For Mitochondrial Diseases It Has Been More Productive
## to Classify Patients by Genetic Defect Rather than by Clinical Manifestation

**Over the past 10 years, mitochondrial defects have been implicated in a wide variety of degenerative diseases, aging, and cancer… The same mtDNA mutation can produce quite different phenotypes, and different mutations can produce similar phenotypes. …The essential role of mitochondrial oxidative phosphorylation in cellular energy production, the generation of reactive oxygen species, and the initiation of apoptosis has suggested a number of novel mechanisms for mitochondrial pathology.**

**--Douglas Wallace, Science, Vol. 283, 1482-1488, 5 March 1999**

UCSD
UCIrvine

# Comparative Genomics Can Reveal Biological Facts That Are Not Visible Within a Species

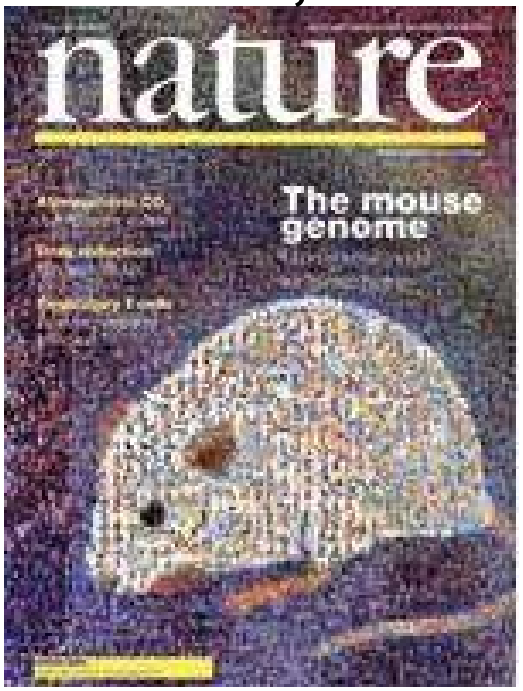
## Co-Authors Pavel Pevzner and Glenn Tesler, UCSD




December 9, 2004




April 1, 2004




December 05, 2002

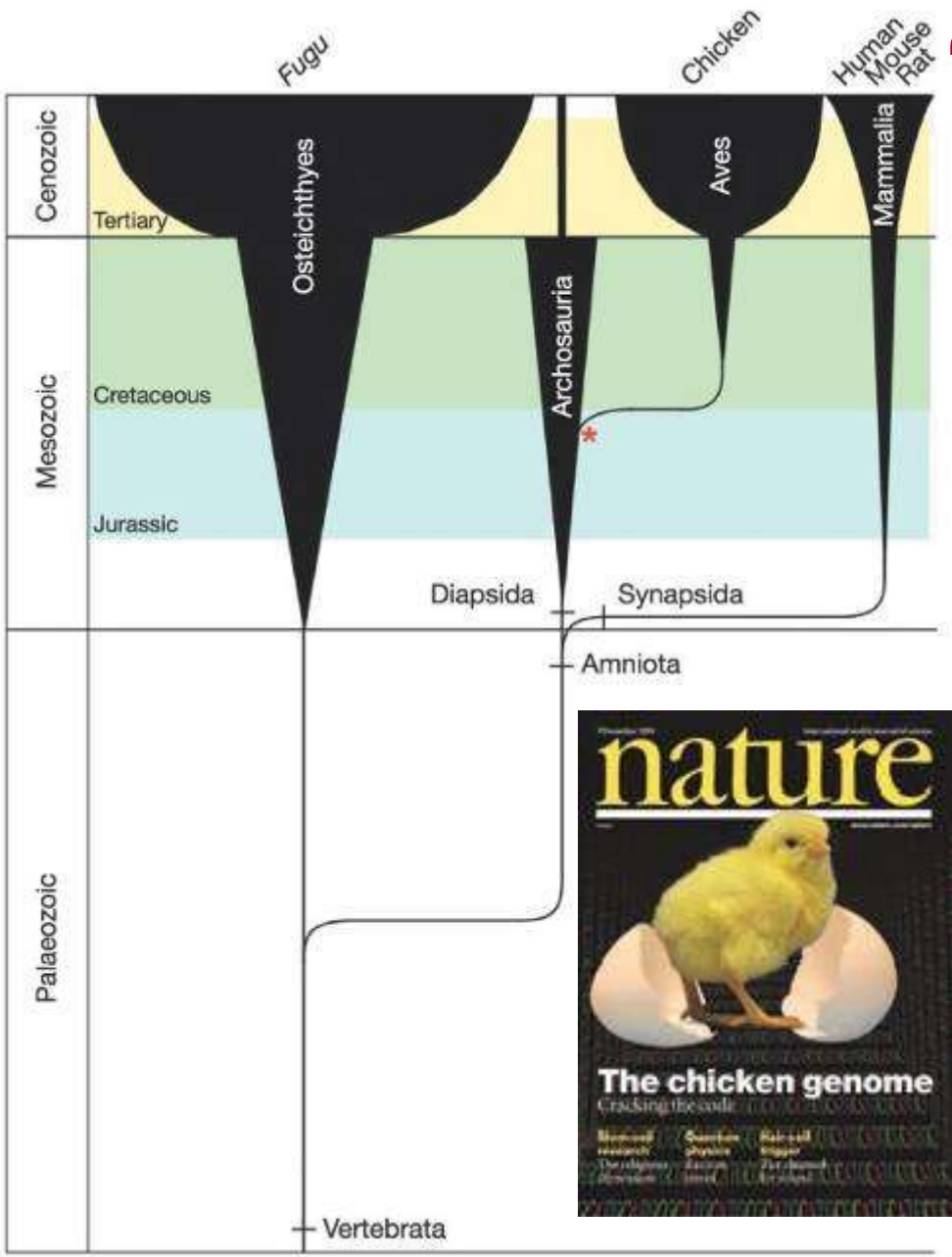
**"After sequencing these three genomes, it is clear that substantial rearrangements in the human genome happen only once in a million years, while the rate of rearrangements in the rat and mouse is much faster."**

**--Glenn Tesler, UCSD Dept. of Mathematics**

UCSD
UCIrvine

www.calit2.net/culture/features/2004/4-1_pevzner.html

# Advanced Algorithmic Techniques Reveal Unexpected Results







**"Many of the chicken–human aligned, non-coding sequences occur far from genes, frequently in clusters that seem to be under selection for functions that are not yet understood."**

**Nature 432, 695 - 716 (09 December 2004)**

UCSD
UCIrvine

# 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 ."**

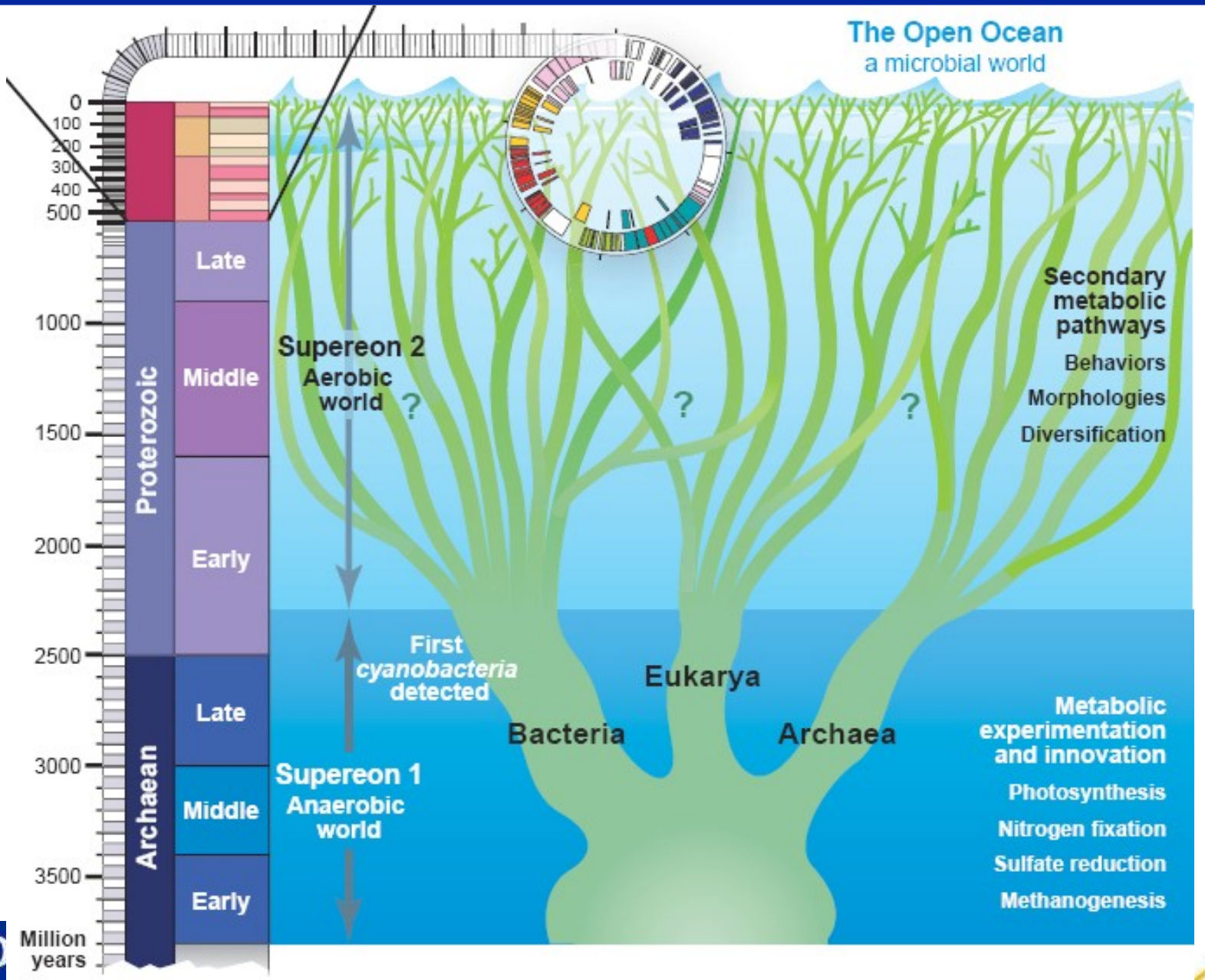
***Comparative Metagenomics of Microbial Communities***

**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**

UCSD
UCIrvine

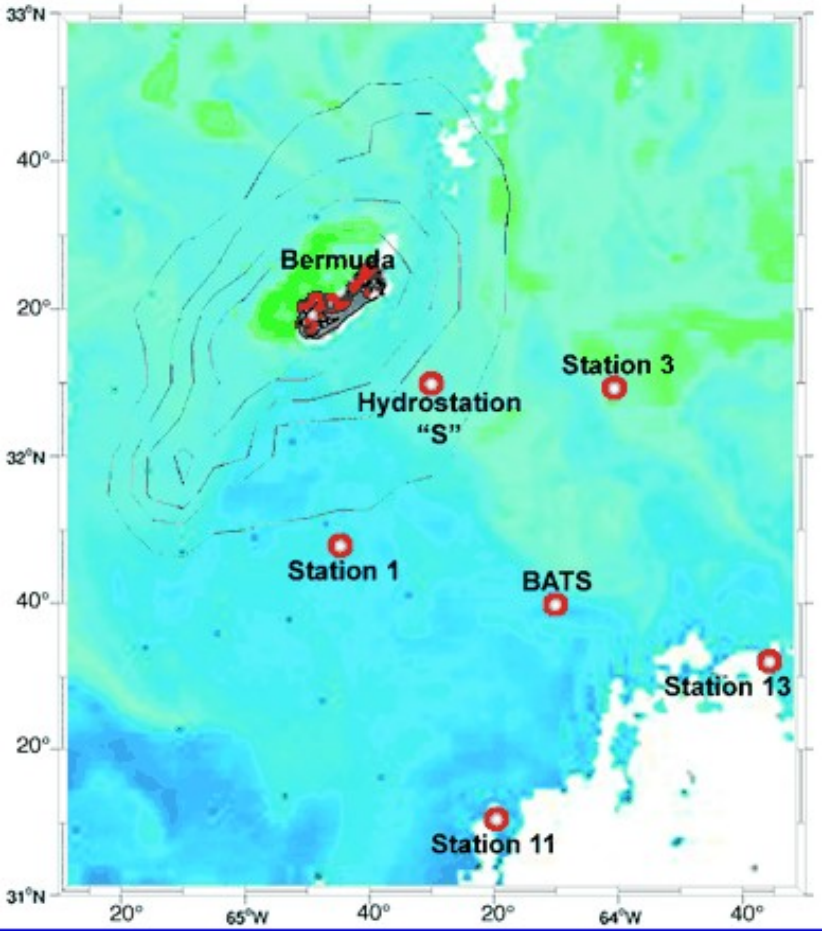
# Looking Back Nearly 4 Billion Years In the Evolution of Microbe Genomics




UCSD
UCIrvine

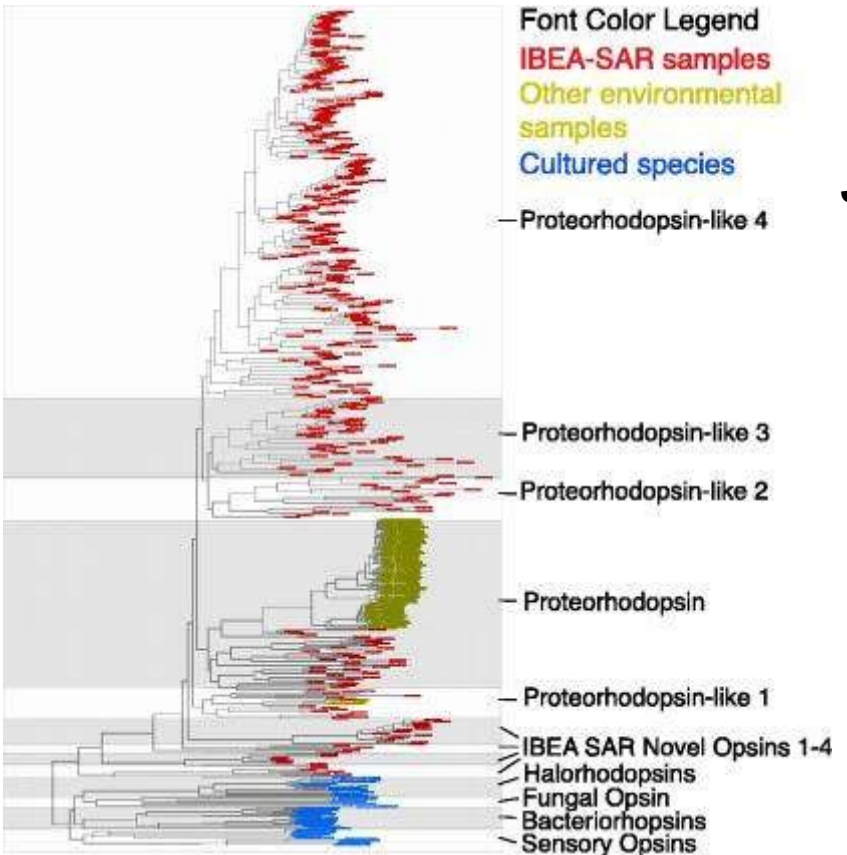
Science Falkowski and Vargas 304 (5667): 58

# 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

UCSD
UCIrvine

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

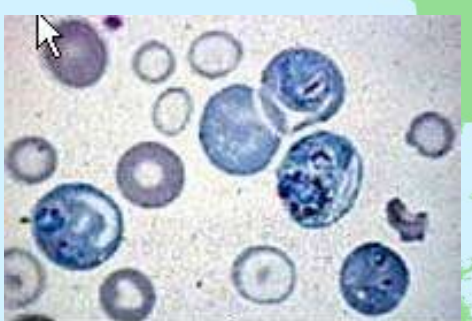
Joint Partnership of:

UCSD
UCIrvine

PI Larry Smarr

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





Canada
United States
Mexico
Puerto Rico
Honduras
Costa Rica
Panama
Galapagos Islands
x 10 samples
French Polynesia
Vanuatu
Fiji
Tonga
Cook Islands
New Caledonia
x 17 samples
Australia
Cocos-Keeling Is.
Tanzania/ Zanzibar
x 7 samples
Seychelles
Madagascar
South Africa
x 8 samples
Analyzed
Sampled, but not Analyzed

**CAMERA will include**
**All Sorcerer II Metagenomic Data**

MARINE MICROBIOLOGY INITIATIVE

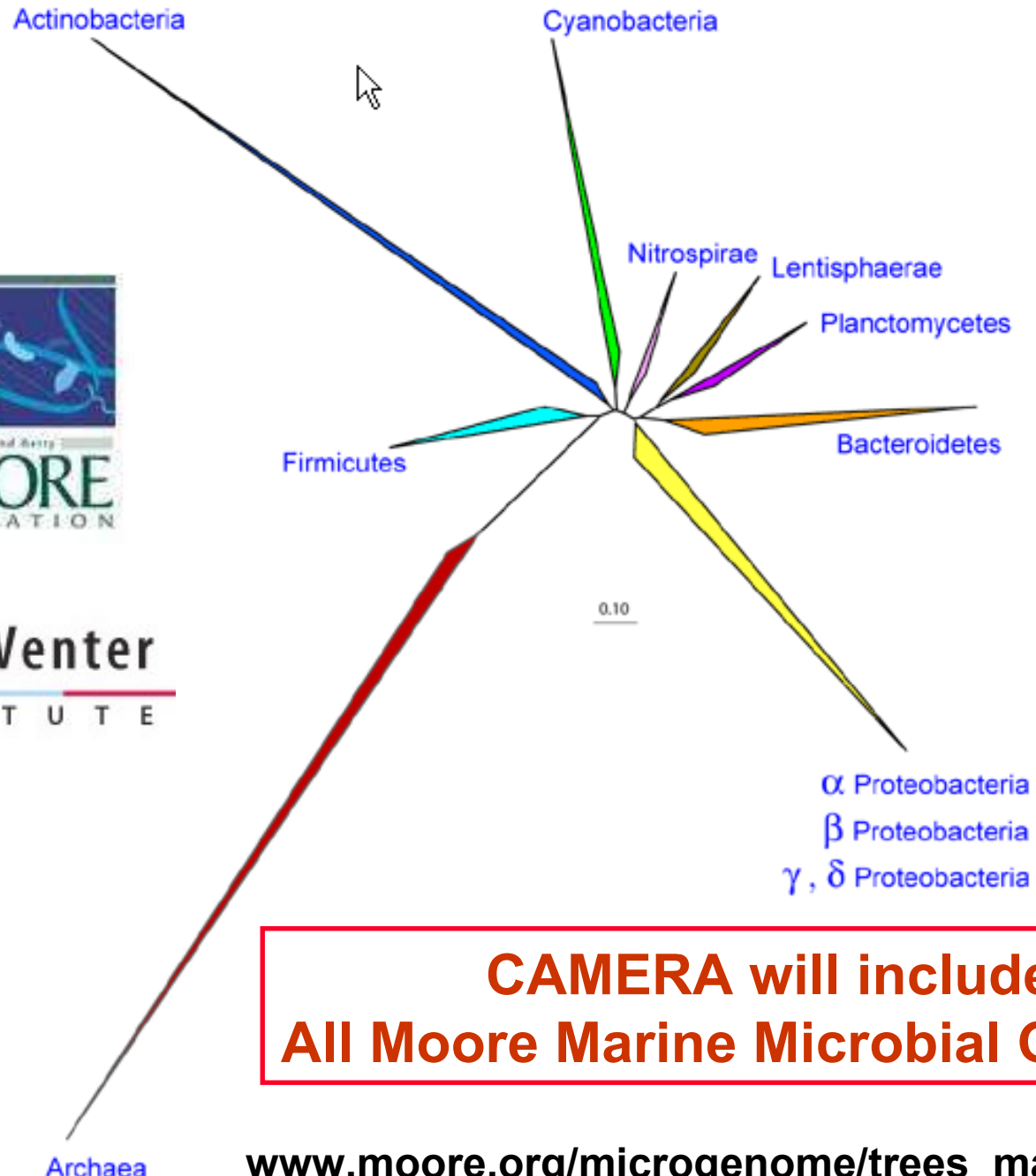
Gordon and Betty MOORE FOUNDATION

UCSD
UCIrvine

J. Craig Venter INSTITUTE

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

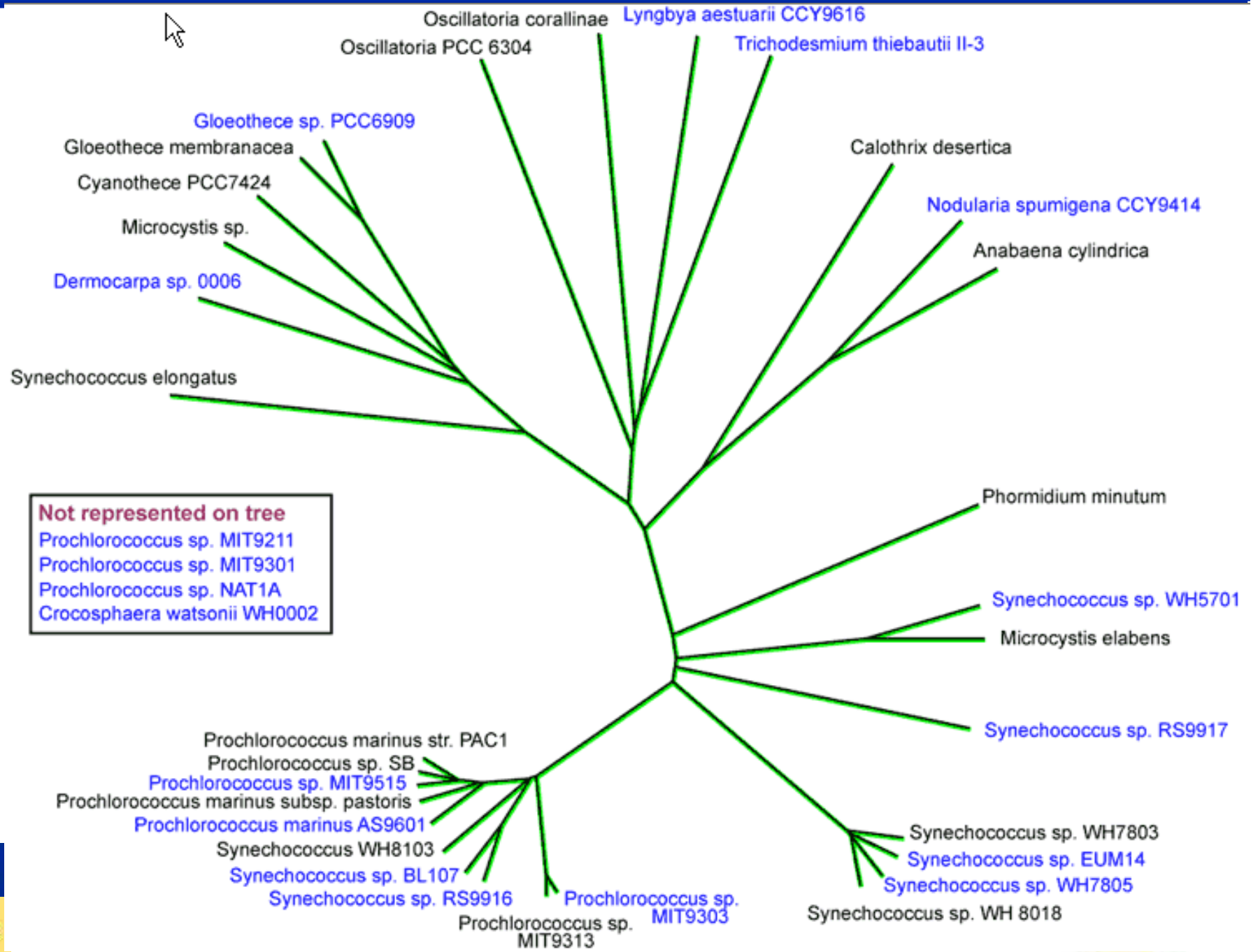
**CAMERA will include**
**All Moore Marine Microbial Genomes**

UCSD
UCIrvine

www.moore.org/microgenome/trees_main.asp

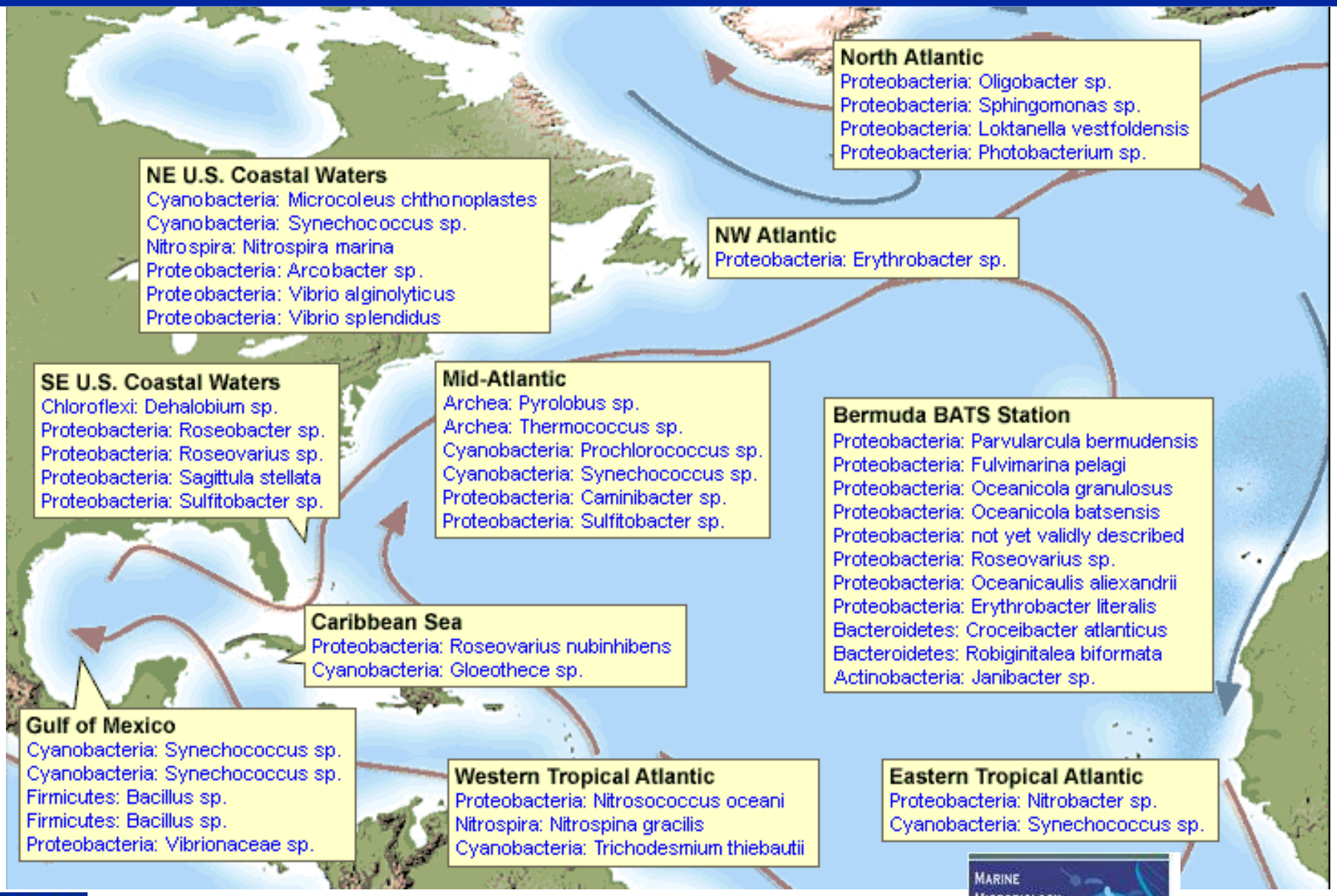
it2

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




UCSD
UCIrvine

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




UCSD
UCIrvine

www.moore.org/microgenome/worldmap.asp

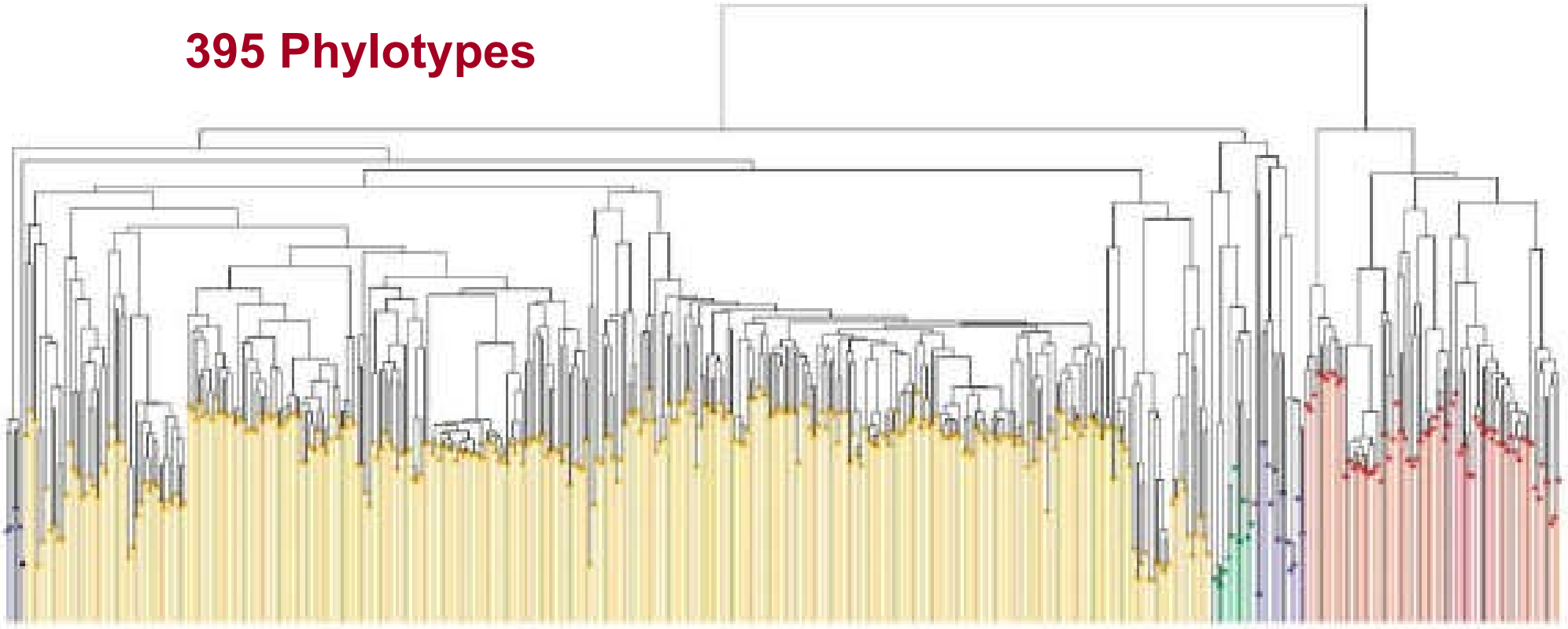
MARINE MICROBIOLOGY INITIATIVE
MOORE FOUNDATION

it2

# Calit2 is Discussing Including Other Metagenomic Data Sets

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

**395 Phylotypes**



UCSD
UCIrvine

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

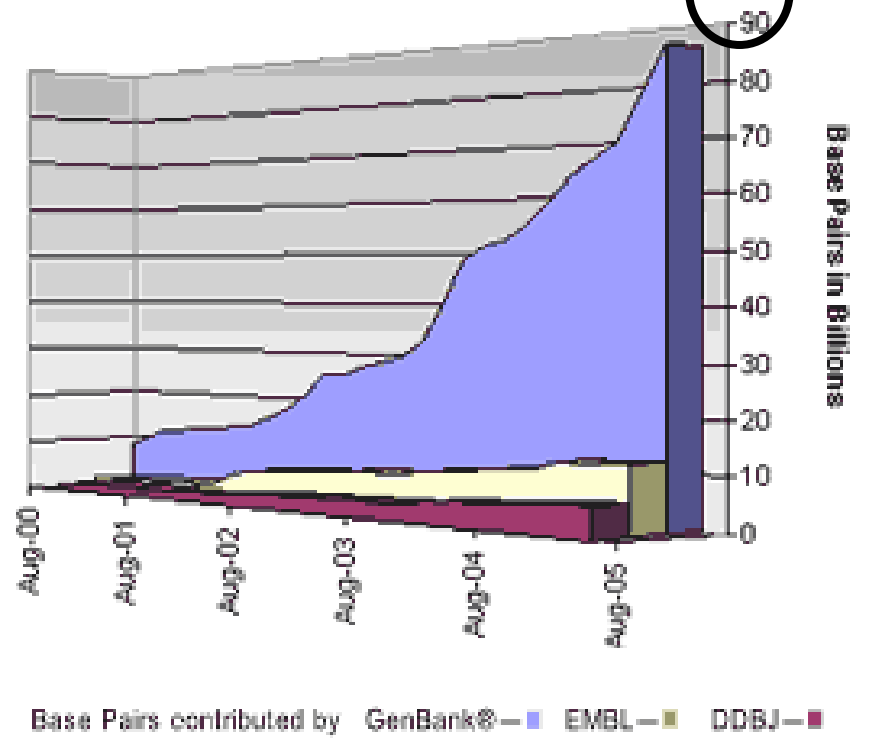
it2

# Genomic Data Is Growing Rapidly, But Metagenomics Will Vastly Increase The Scale…

100 Billion Bases!

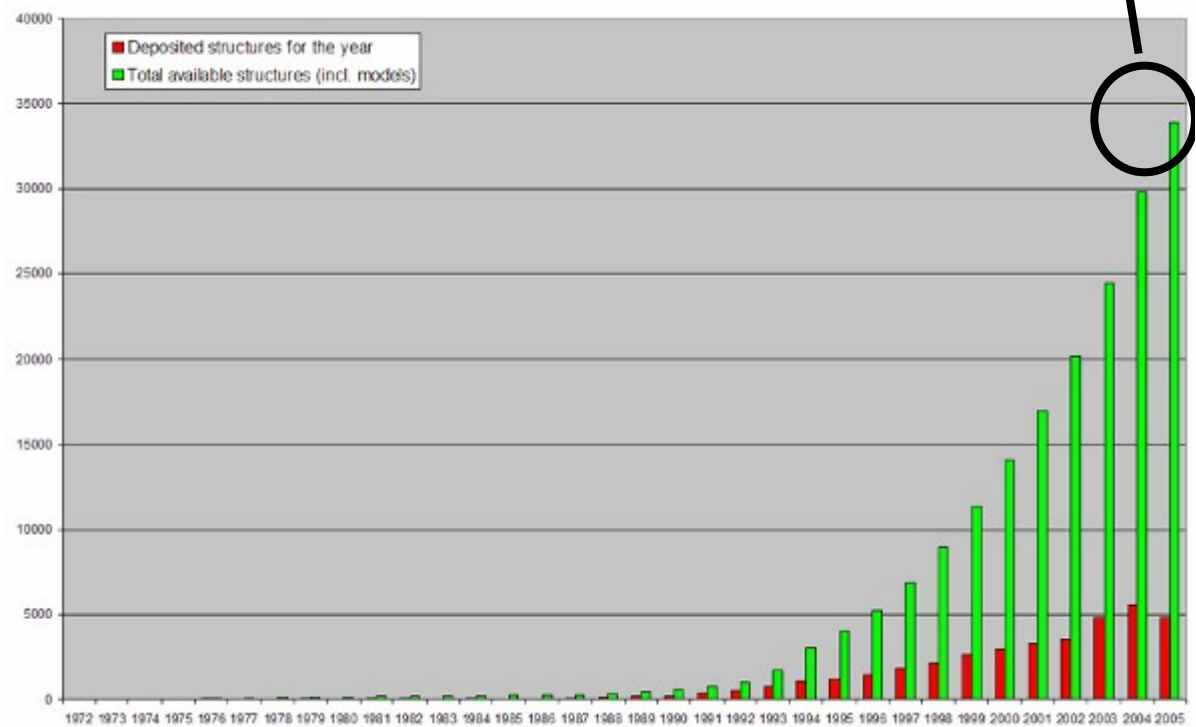
35,000 Structures

## GenBank




www.ncbi.nlm.nih.gov/Genbank

## Protein Data Bank




www.rcsb.org/pdb/holdings.html

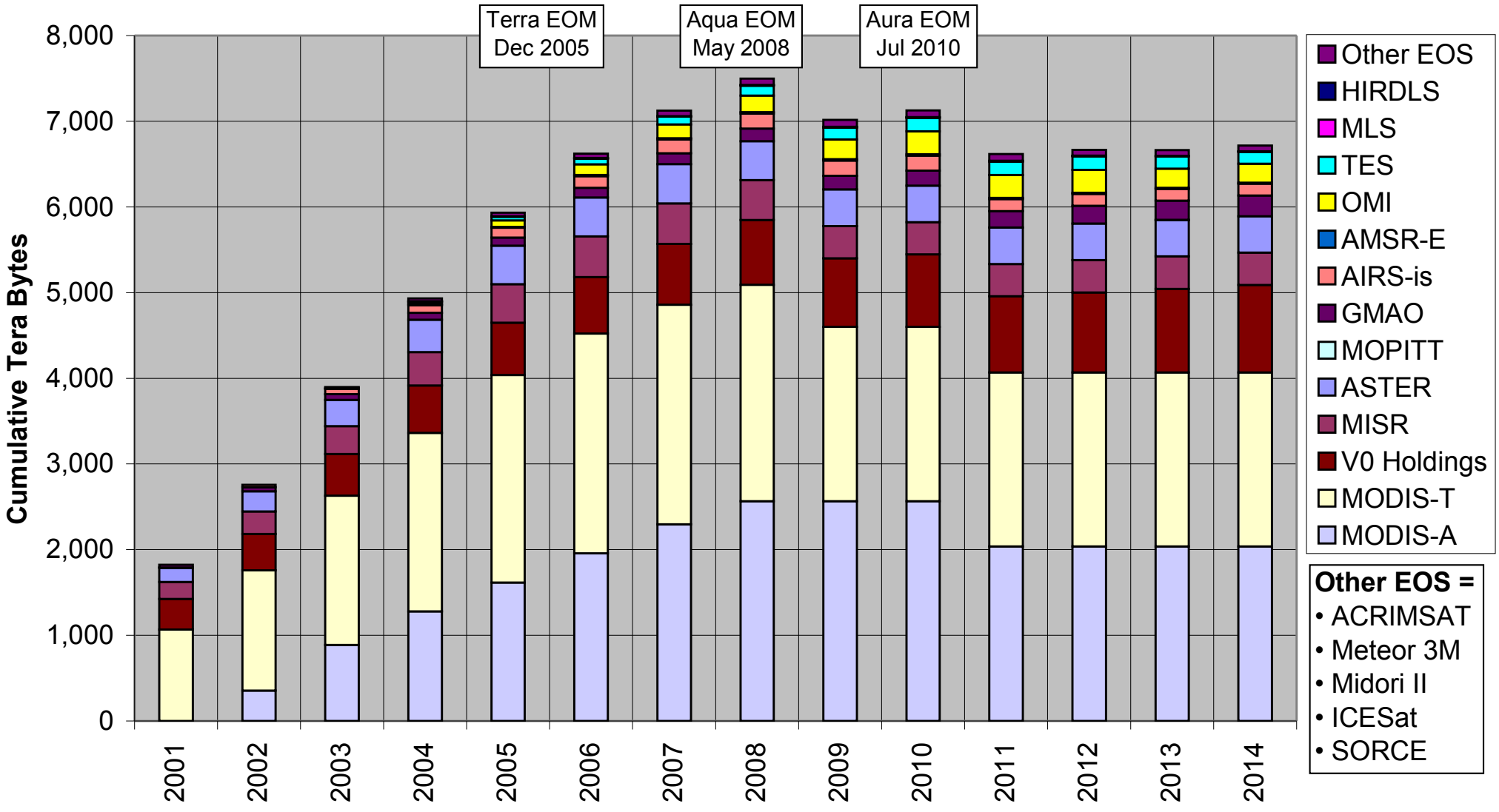
UCSD
UCIrvine

**Total Data < 1TB**

it2

# Metagenomics Will Couple to Earth Observations Which Add Several TBs/Day

## Cumulative Archive Holdings by Instruments/Missions




file name: archive holdings_122204.xls
tab: all instr bar

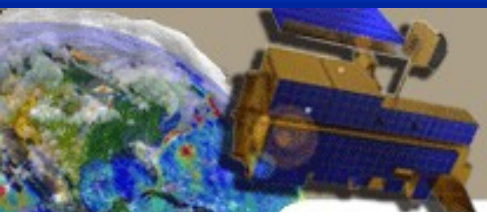
**NOTE: Data remains in the archive pending transition to LTA**

**Source: Glenn Iona, EOSDIS Element Evolution Technical Working Group January 6-7, 2005**

UCSD
UCIrvine

it²

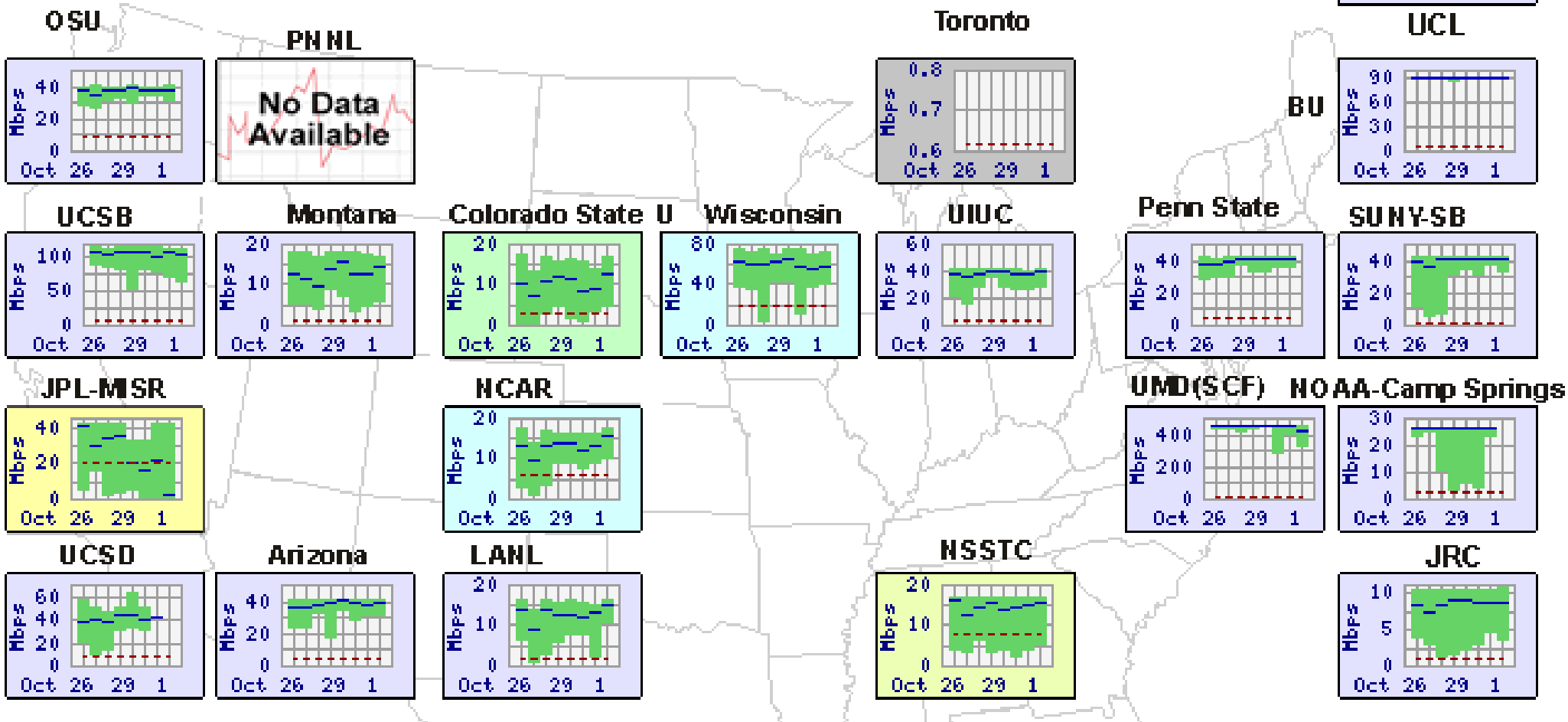
# Challenge: Average Throughput of NASA Data Products to End User is < 50 Mbps



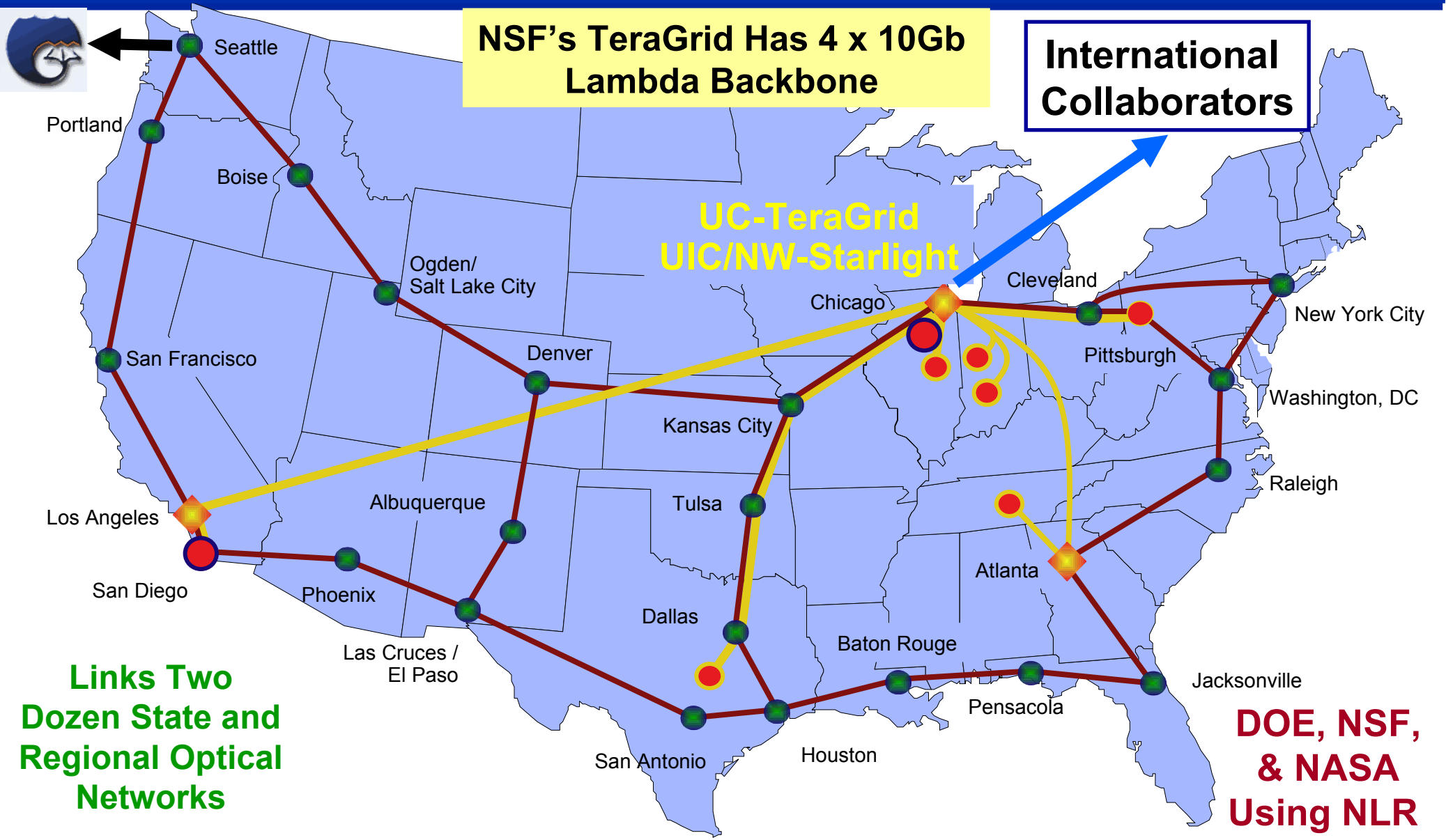
## ENSIGHT Active Testing TERRA Destinations

Tested October 2005

**Internet2 Backbone is 10,000 Mbps!**
**Throughput is < 0.5% to End User**



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



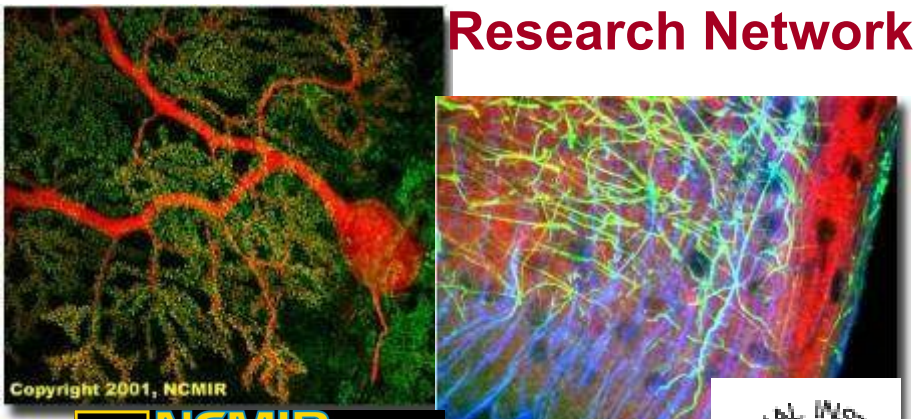

UCSD
UCIrvine

# The OptIPuter Project –

## Creating a LambdaGrid "Web" for Gigabyte Data Objects

- **NSF Large Information Technology Research Proposal**
  - Calit2 (UCSD, UCI) and UIC Lead Campuses—Larry Smarr PI
  - Partnering Campuses: USC, SDSU, NW, TA&M, UvA, SARA, NASA
- **Industrial Partners**
  - IBM, Sun, Telcordia, Chiaro, Calient, Glimmerglass, Lucent
- **$13.5 Million Over Five Years**
- **Linking Global Scale Science Projects to User's Linux Clusters**

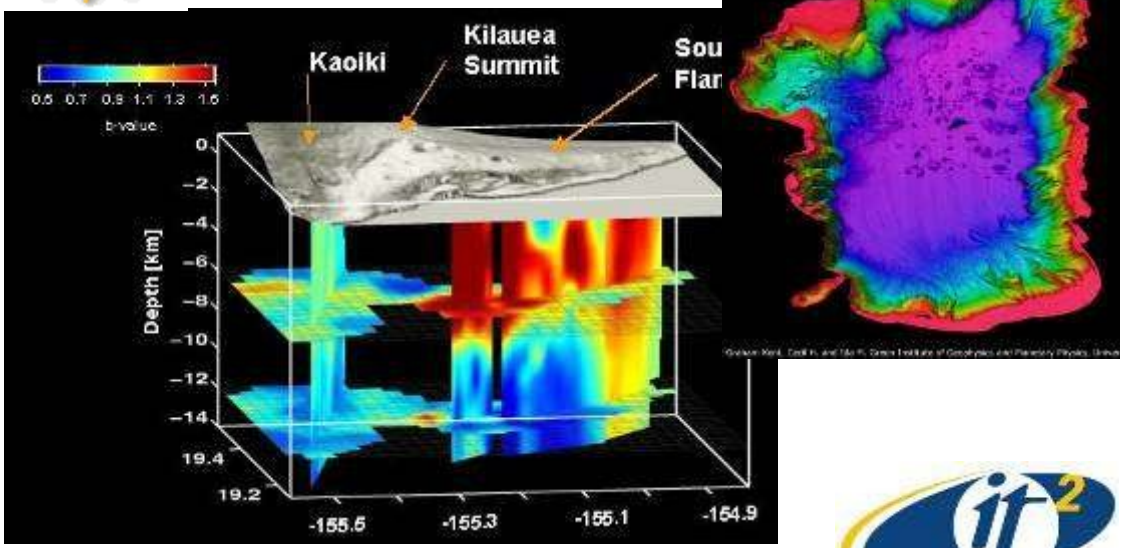
**NIH Biomedical Informatics Research Network**




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

NATIONAL INSTITUTES OF HEALTH

**NSF EarthScope and ORION**



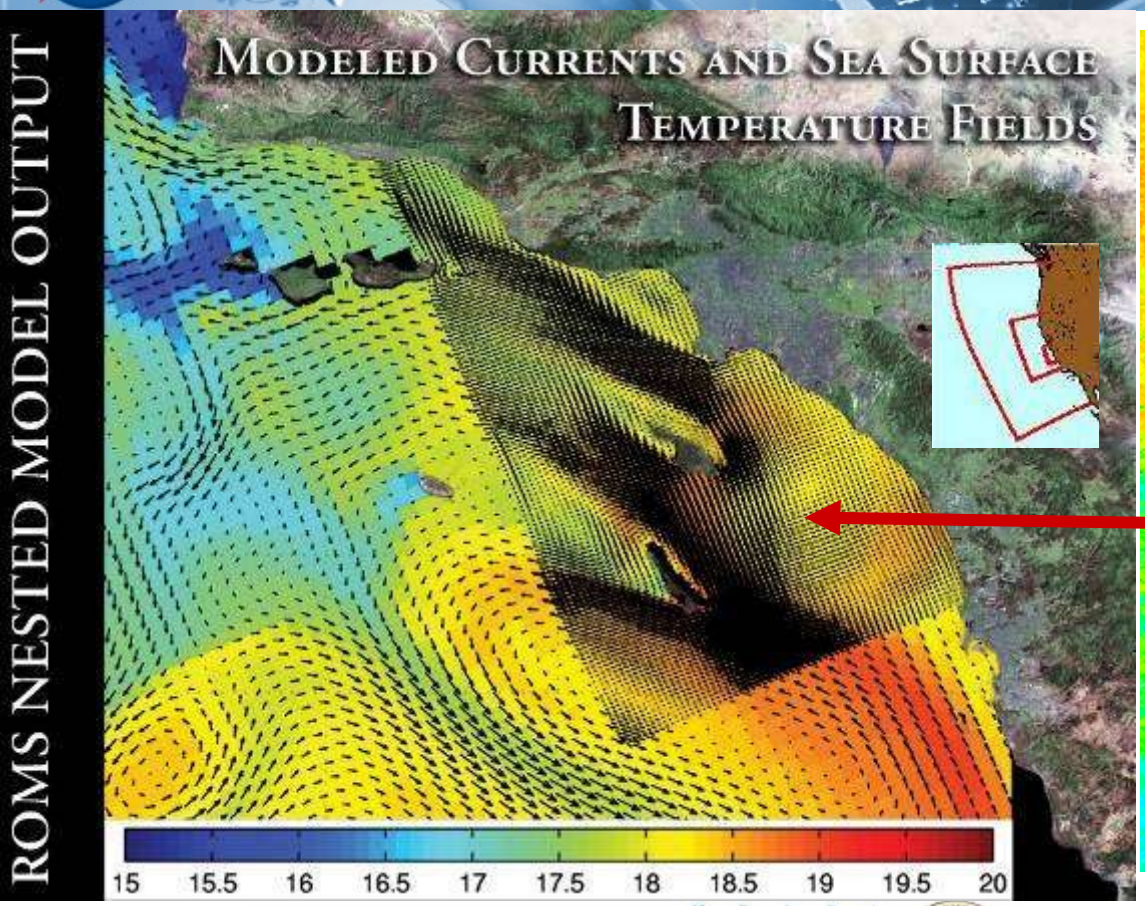

UCSD
UCIrvine

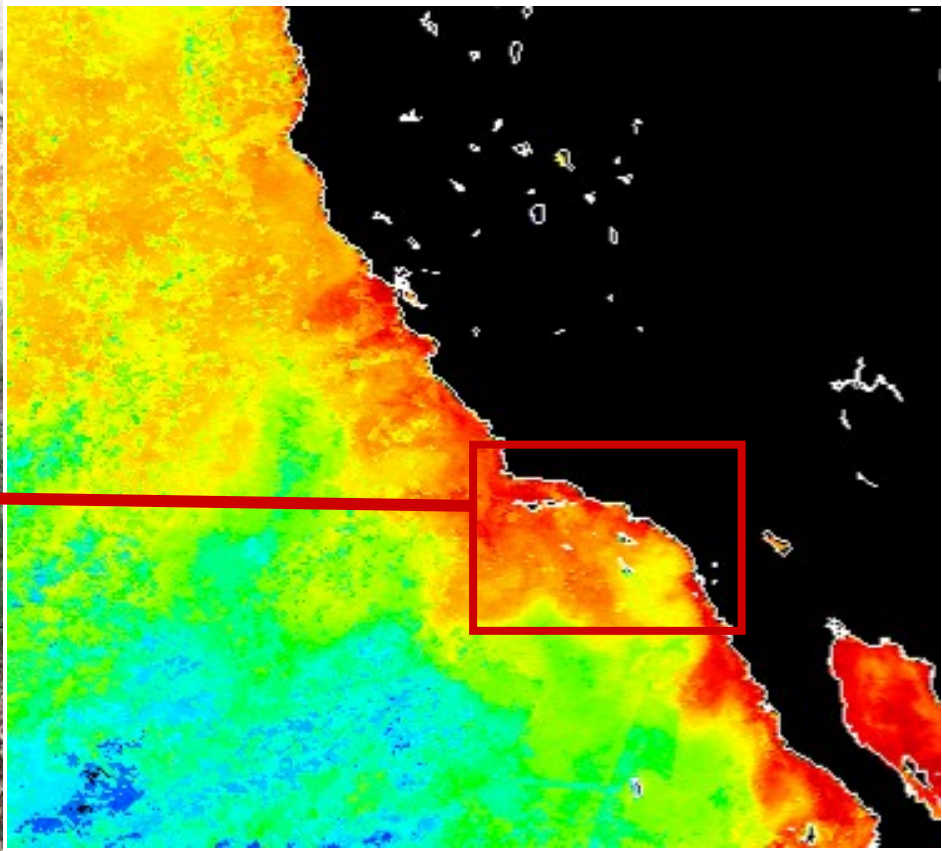
SDSC
SAN DIEGO SUPERCOMPUTER CENTER

it2

# Using the OptIPuter to Couple Data Assimilation Models to Remote Data Sources Including Biology



**NASA MODIS Mean Primary Productivity for April 2001 in California Current System**

JPL ONR MBARI

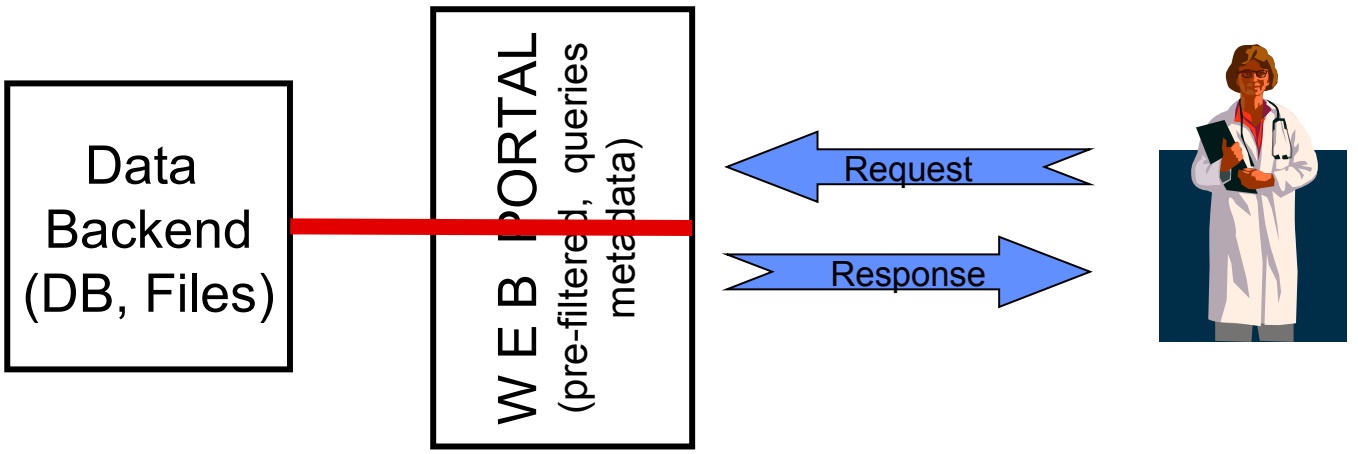
UCSD
UCIrvine

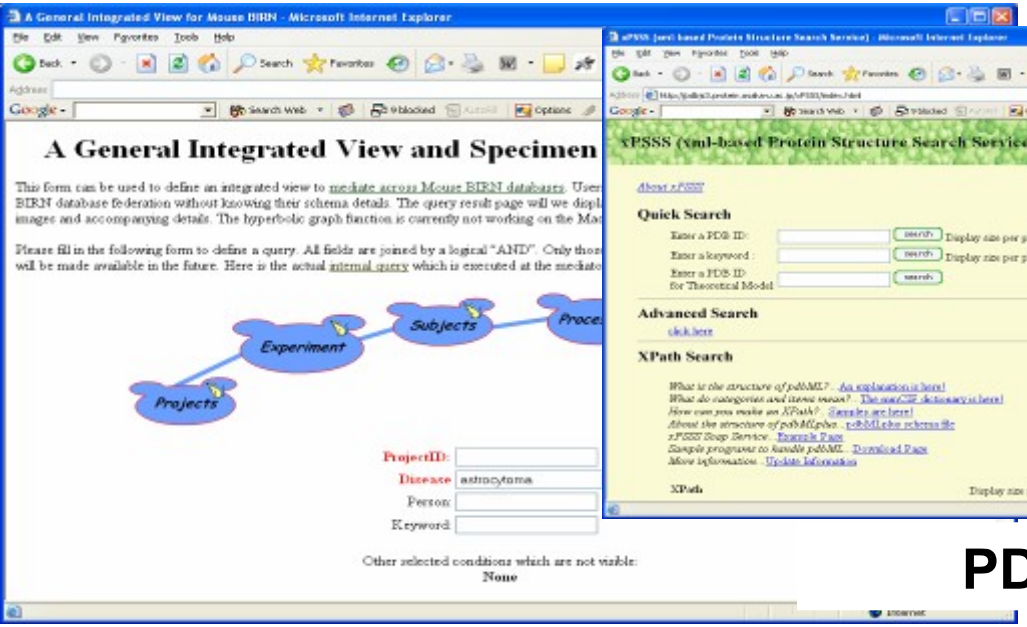
**Regional Ocean Modeling System (ROMS)**

http://ourocean.jpl.nasa.gov/

it2

# Calit2 Intends to Jump Beyond Traditional Web-Accessible Databases

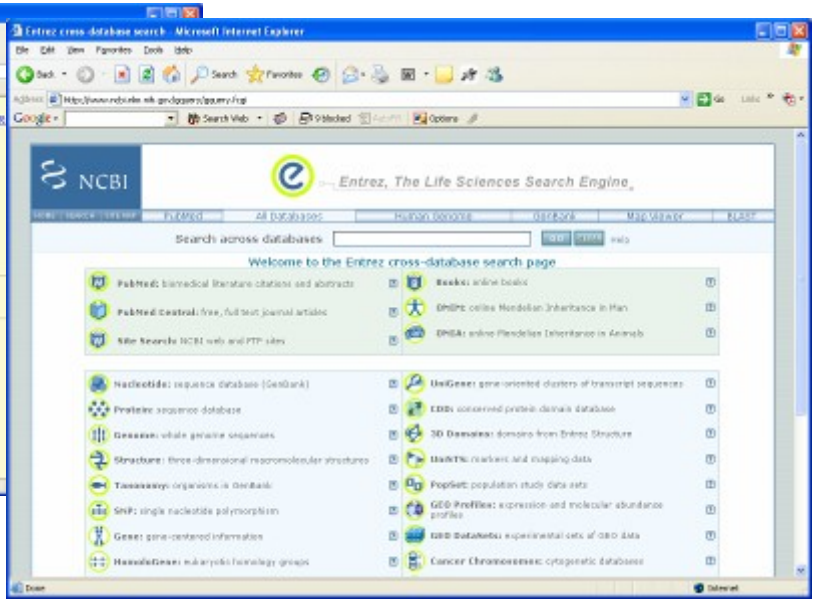



BIRN




PDB




NCBI Genbank

**+ many others**

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

UCSD UCIrvine

it2

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

**Sargasso Sea Data**

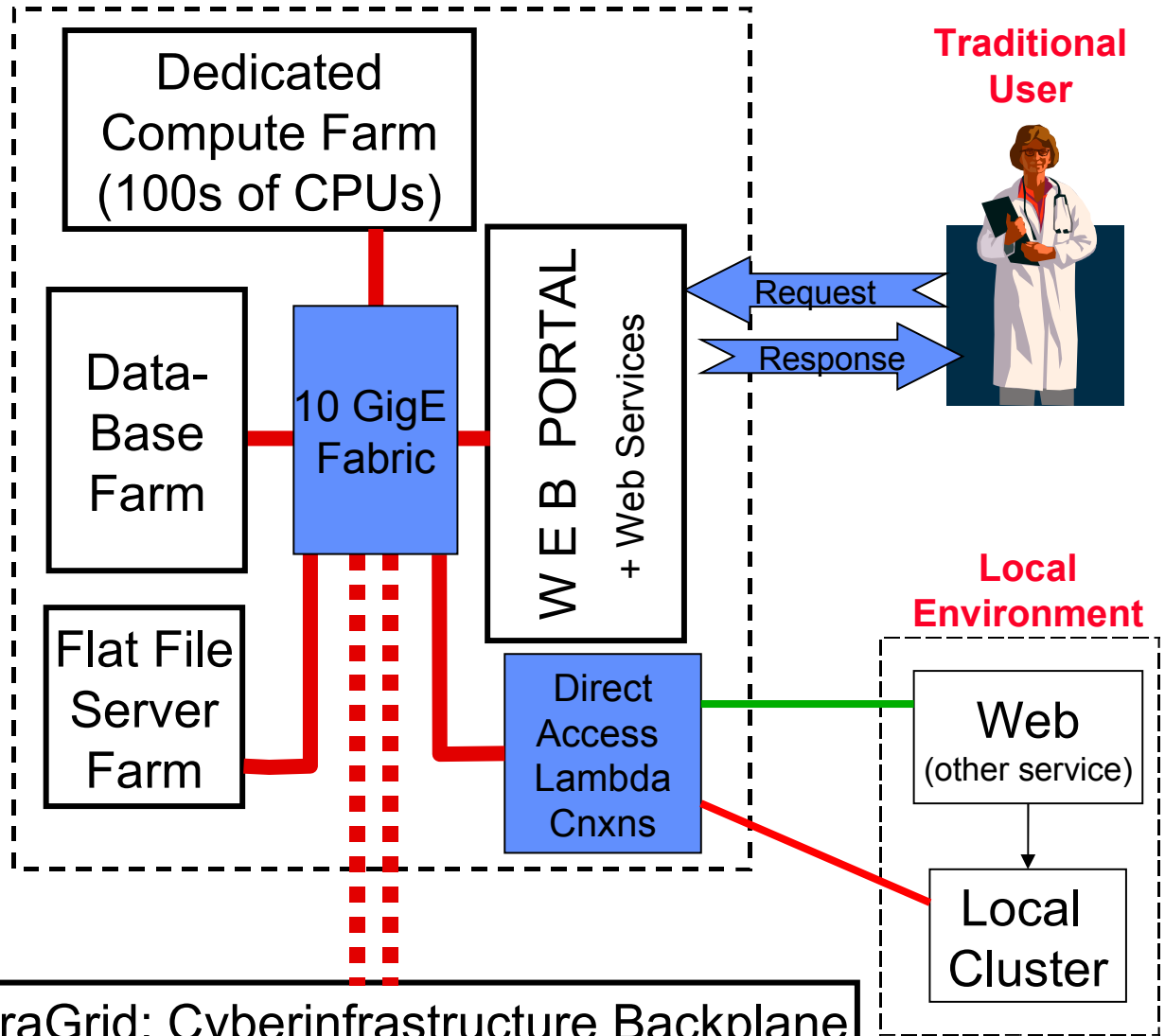
**Sorcerer II Expedition (GOS)**

**JGI Community Sequencing Project**

**Moore Marine Microbial Project**

**NASA Goddard Satellite Data**

**Community Microbial Metagenomics Data**




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

CENTER FOR EARTH OBSERVATIONS & APPLICATIONS

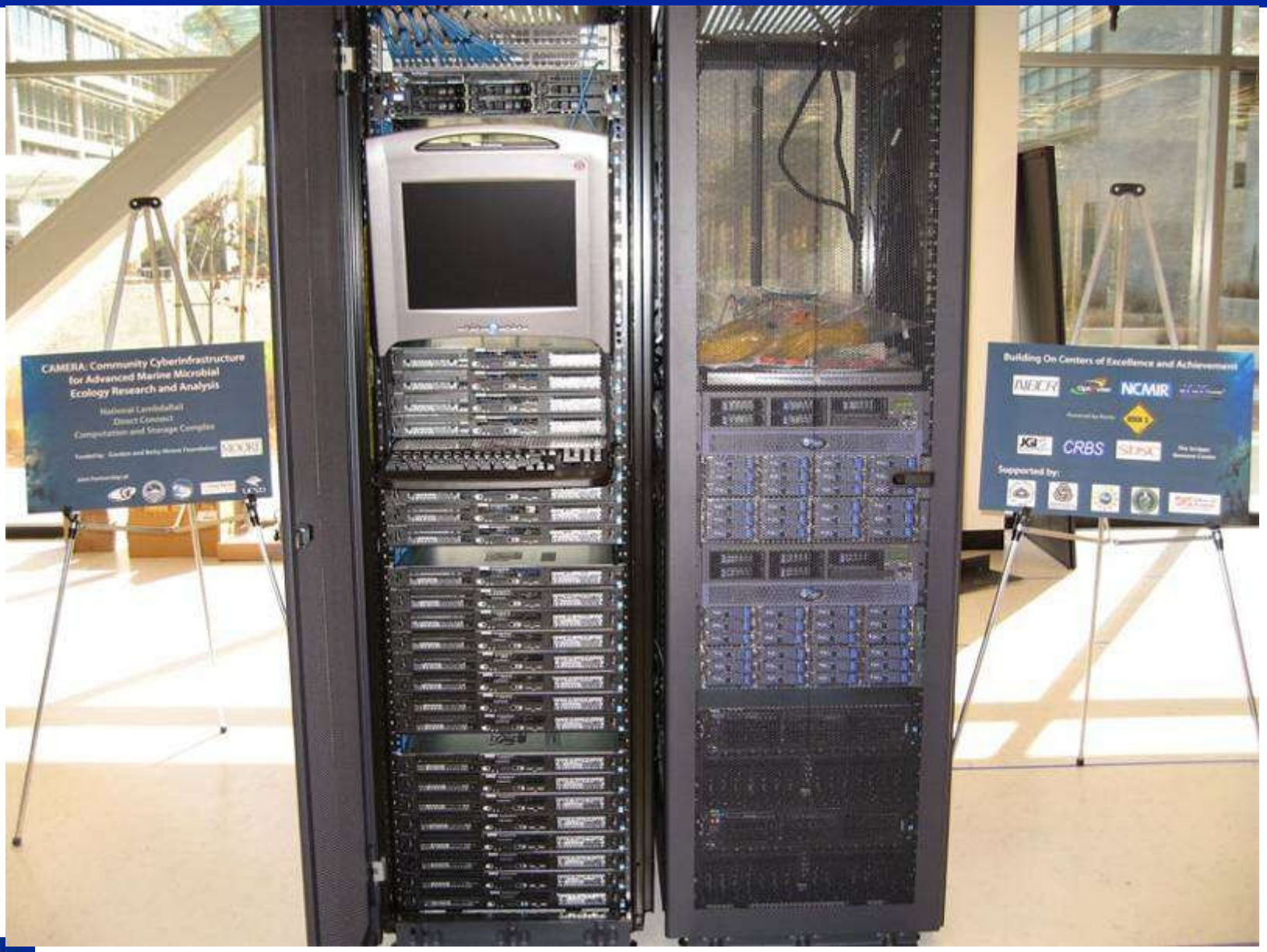
J. Craig Venter INSTITUTE

UCSD UCIrvine

SDSC SAN DIEGO SUPERCOMPUTER CENTER

it2

# First Implementation of the CAMERA Complex




Compute

Database & Storage

UCSD
UCIrvine

# 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

UCSD
UCIrvine

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

it2

# CAMERA Timeline

- **Release 1: Mid-2006**
  - Majority of GOS + Moore Microbe Genome Data
    - 6 Gbp Has Been Assembled
  - Initial Versions of Core Tools
    - BLAST, Reference Alignment Viewer
- **Release 2: Early-2007**
  - Additional Data
  - Additional/Improved Tools
  - Improved Usability
- **Subsequent**
  - Move Towards Semantic DB, Direct Access
  - Additional Tools & Data Based on Community Feedback

UCSD
UCIrvine

# Building On Centers of Excellence and Achievement












Powered by Rocks













The Scripps Genome Center

## Supported by:















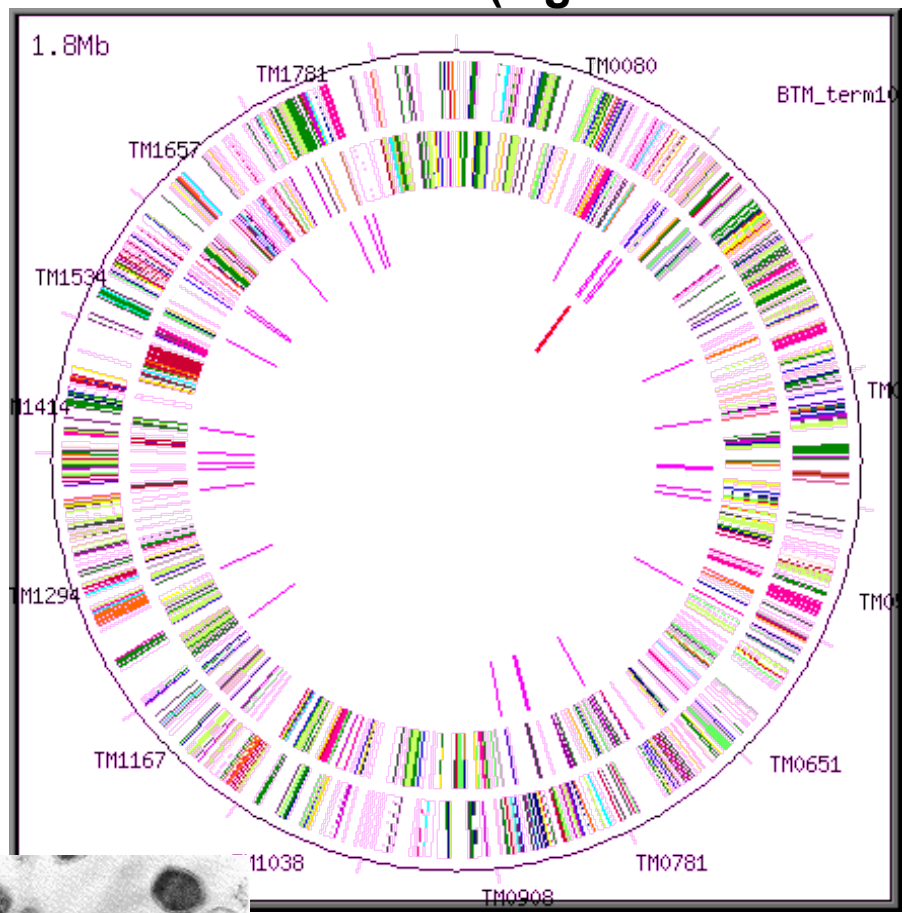
UCSD
UCIrvine

**Announcing Tuesday January 17, 2006**

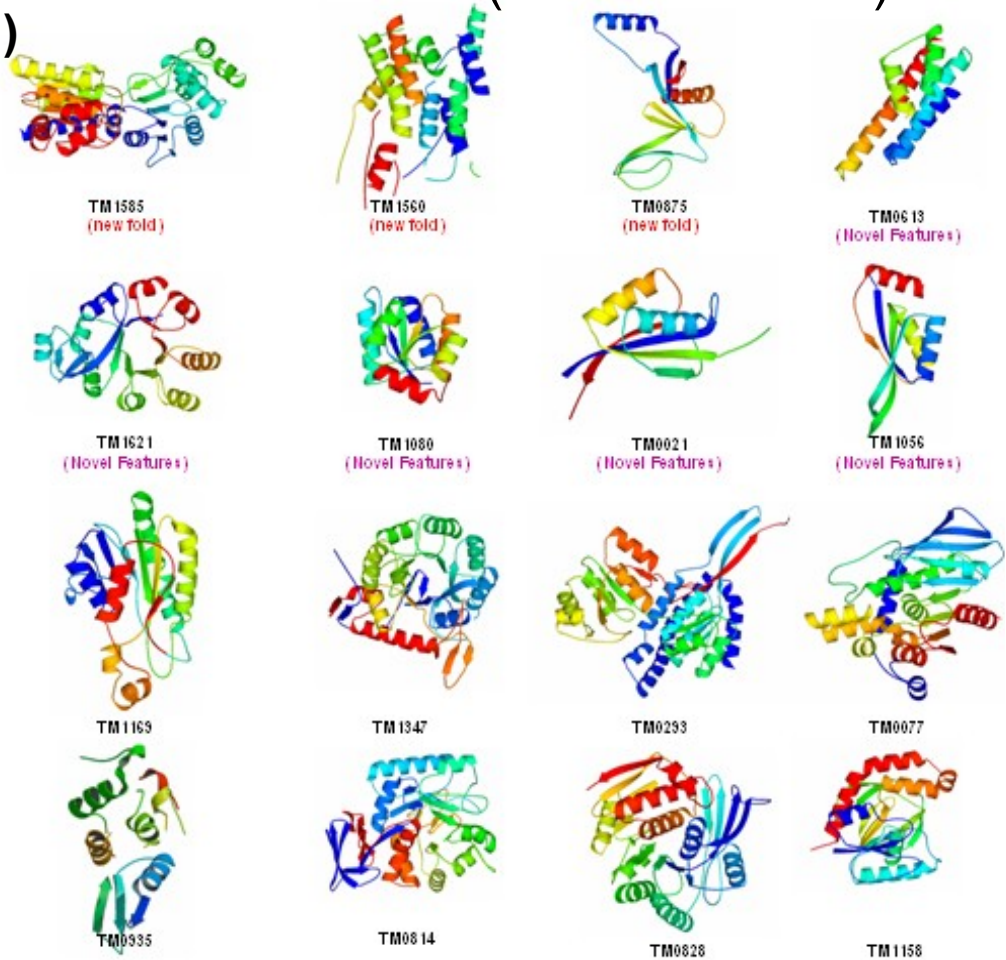
it²

# The Bioinformatics Core of the Joint Center for Structural Genomics will be Housed in the Calit2@UCSD Building

**Extremely Thermostable -- Useful for Many Industrial Processes (e.g. Chemical and Food)**




173 Structures (122 from JCSG)





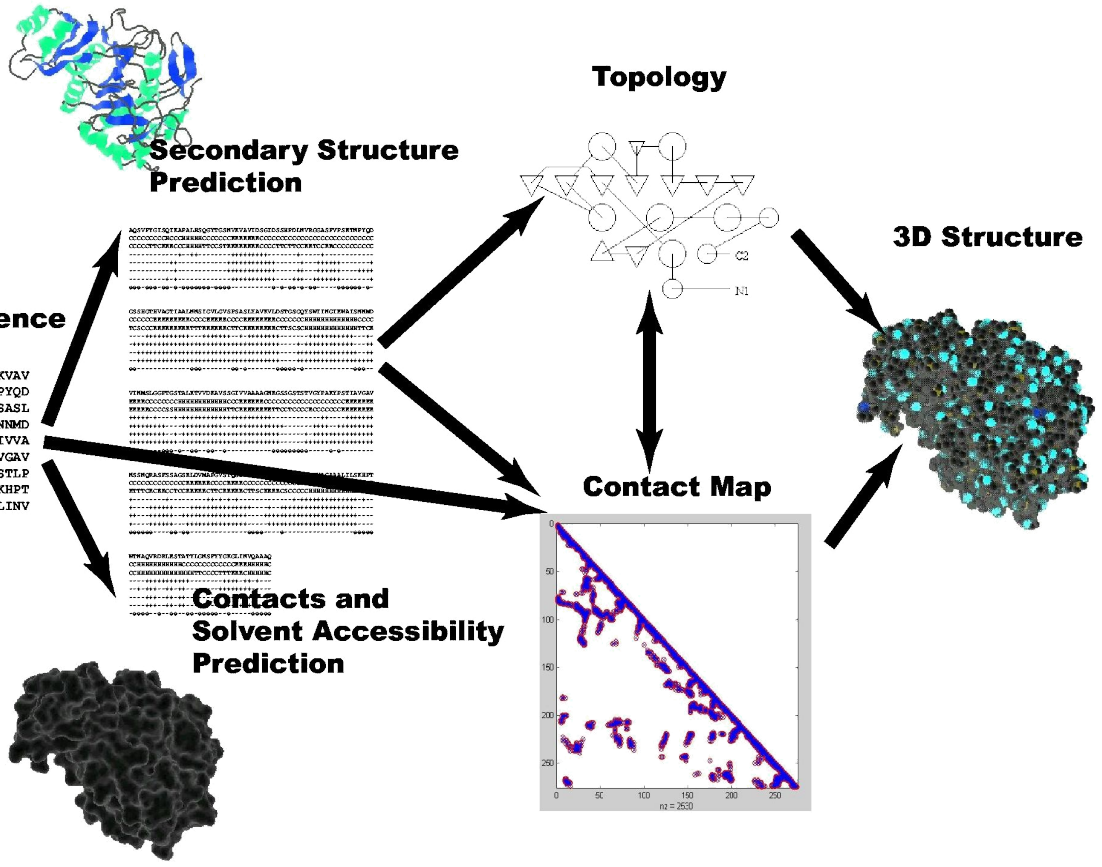

- Determining the Protein Structures of the Thermotoga Maritima Genome
- 122 T.M. Structures Solved by JCSG (75 Unique In The PDB)
- Direct Structural Coverage of 25% of the Expressed Soluble Proteins
- Probably Represents the Highest Structural Coverage of Any Organism

**Source: John Wooley, UCSD**

UCSD
UCIrvine

# UCI's IGB Develops a Suite of Programs and Servers for Protein Structure and Structural Feature Prediction

INSTITUTE FOR GENOMICS AND BIOINFORMATICS
University of California, Irvine






CODA Genomics Inc. Obtains License on Patents for Technology Developed at UCI

4.28.2005 - Orange County's CODA Genomics Inc. and UC Irvine's Donald Bren School of Information and Computer Sciences announced today that CODA was granted an exclusive license to certain issued patents (US 5,082,767) and patent applications of the university related to Computer Optimized DNA Assembly (CODA) and protein expression optimization.
[More]

**Sixty Affiliated IGB Labs at UCI e.g.:**

The Center for Virus Research

**www.igb.uci.edu/tools.htm**

UCSD
UCIrvine

**Source: Pierre Baldi, UCI**

# CAMERA Builds on Cyberinfrastructure Grid, Workflow, and Portal Projects in a Service Oriented Architecture

NBCR

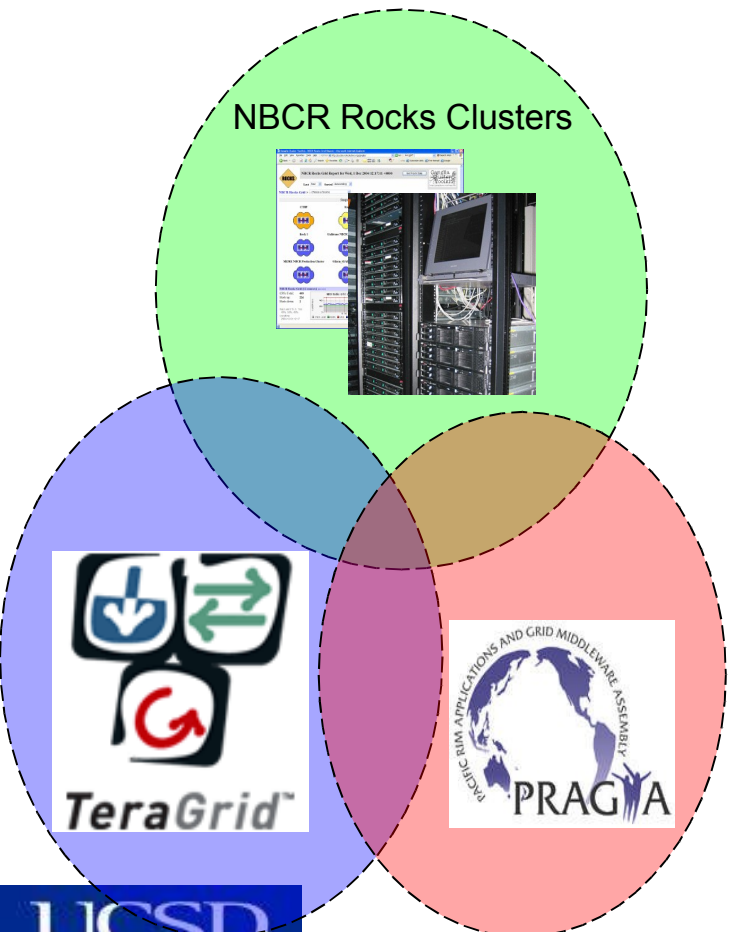
National Biomedical Computation Resource

*an NIH supported resource center*

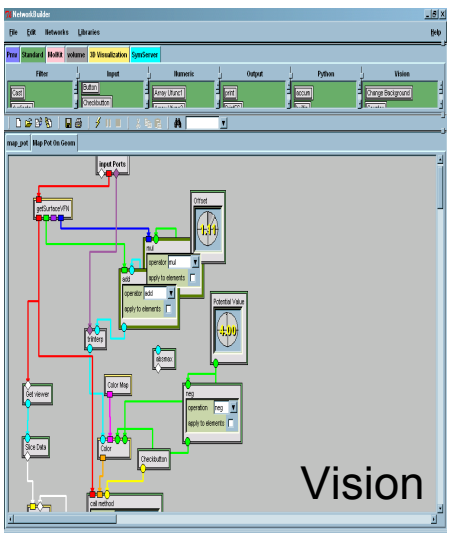
**Located in Calit2@UCSD Building**

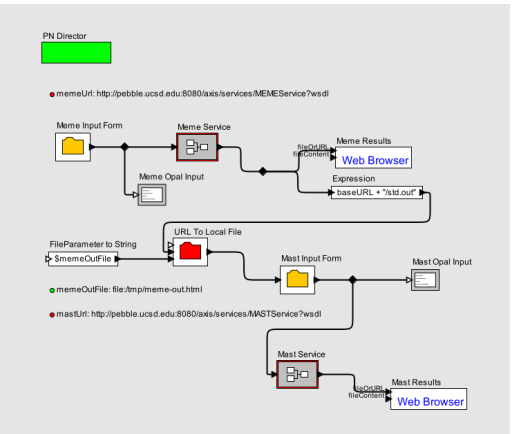
## yberinfrastructure: Raw Resources, Middleware & Execution Environment

### Virtual Organizations



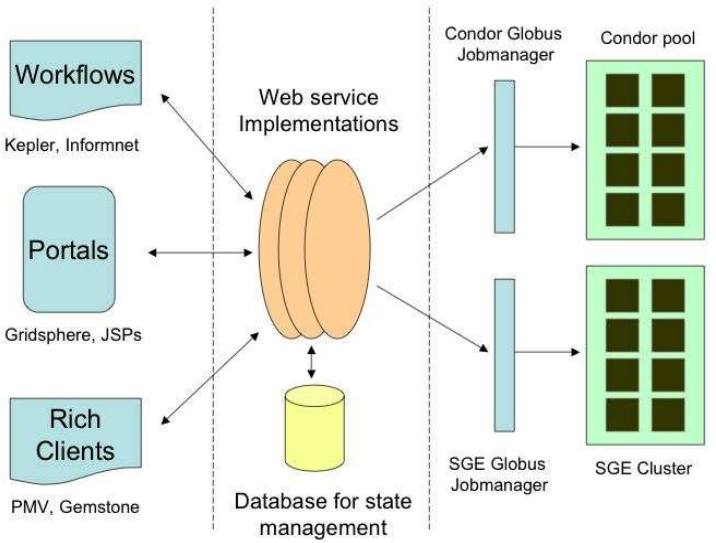

### Workflow Management




KEPLER

### Web Services




### Telescience Portal

telescience project

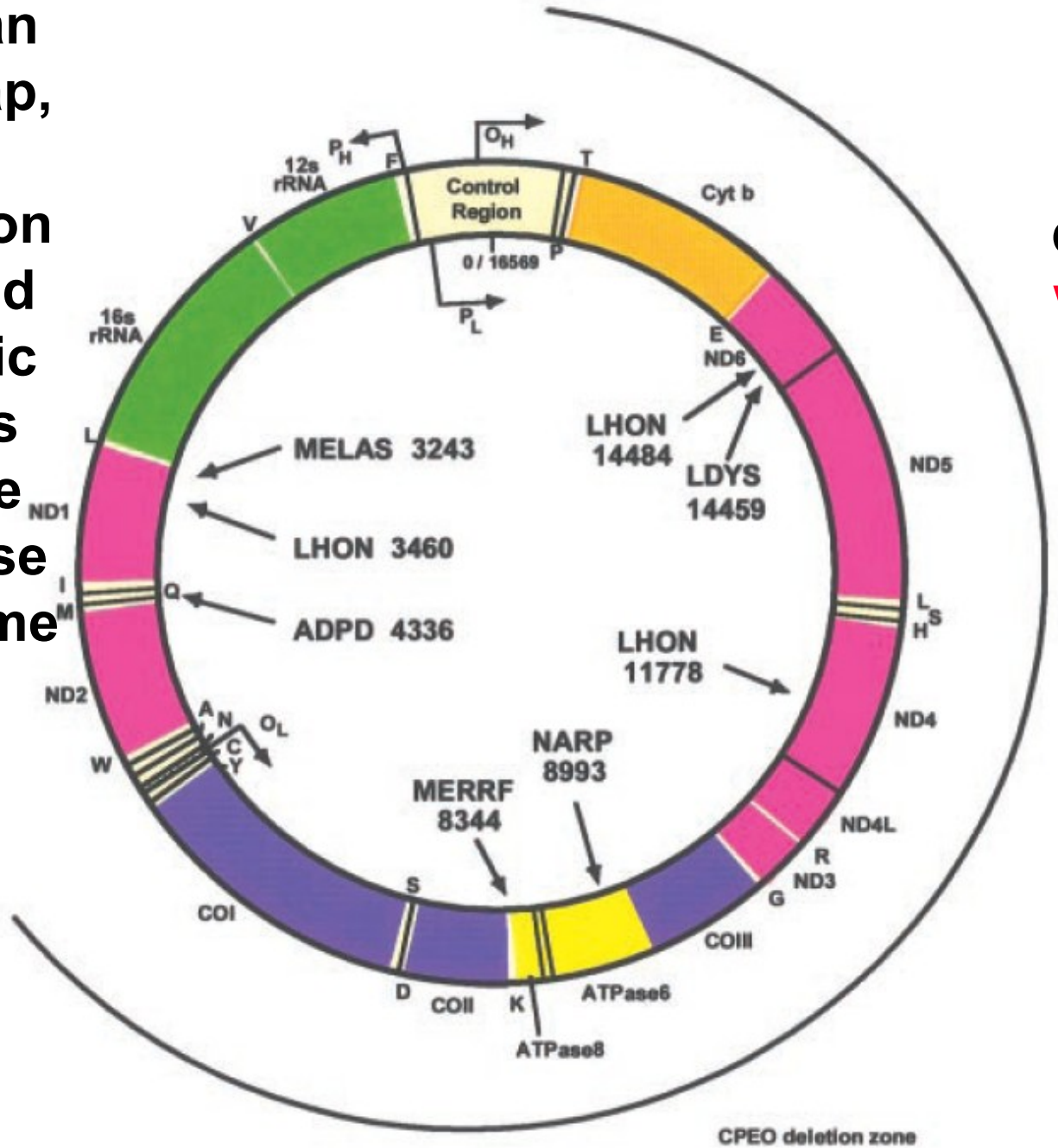
UCSD

UCIrvine

it2

# Calit2 is Collaborating with Douglas Wallace-- Planning to Bring MITOMAP into Calit2 Domain

**The Human mtDNA Map, Showing the Location of Selected Pathogenic Mutations Within the 16,569-Base Pair Genome**




**MITOMAP: A Human Mitochondrial Genome Database. www.mitomap.org, 2005**




**5 March 1999**

UCSD
UCIrvine

it2

# Displaying Images from Electron Microscope



**Zeiss Scanning Electron Microscope in Calit2@ UCI**

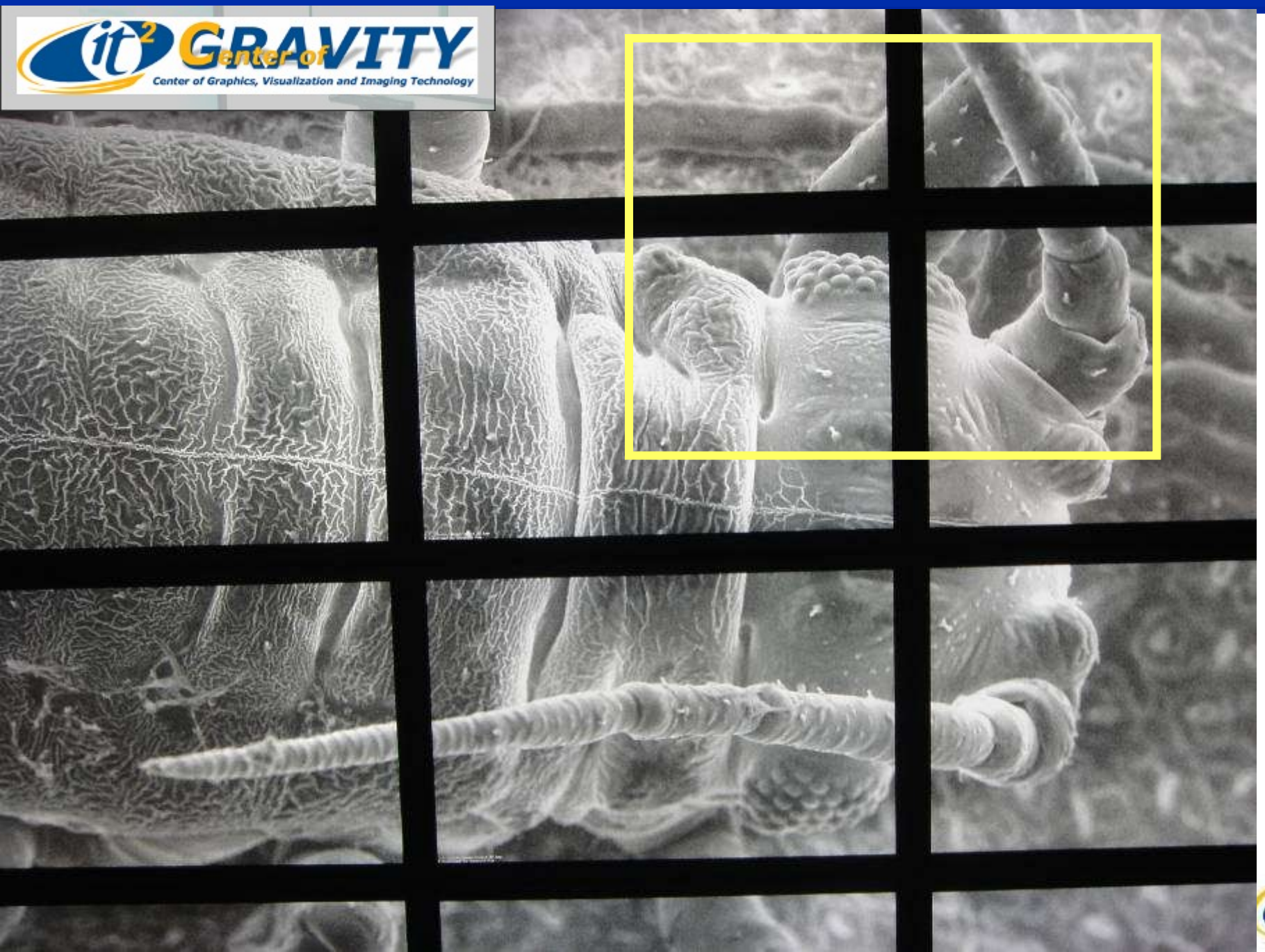



UCSD
UCIrvine

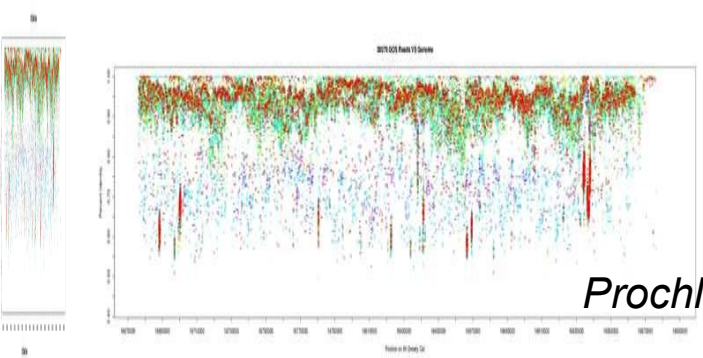
it² Center of GRAVITY
Center of Graphics, Visualization and Imaging Technology

it²

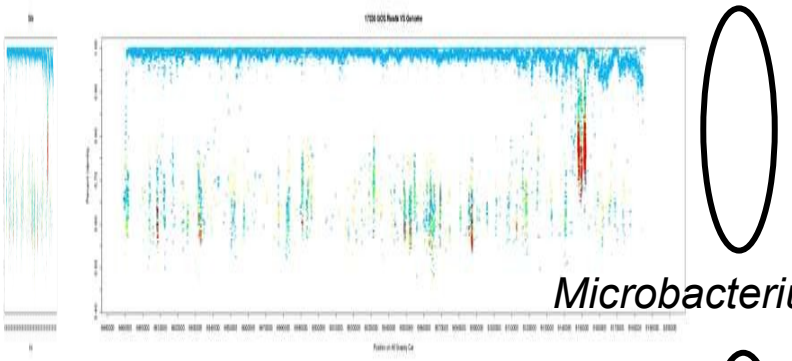
# Zooming In


GRAVITY
Center of Graphics, Visualization and Imaging Technology

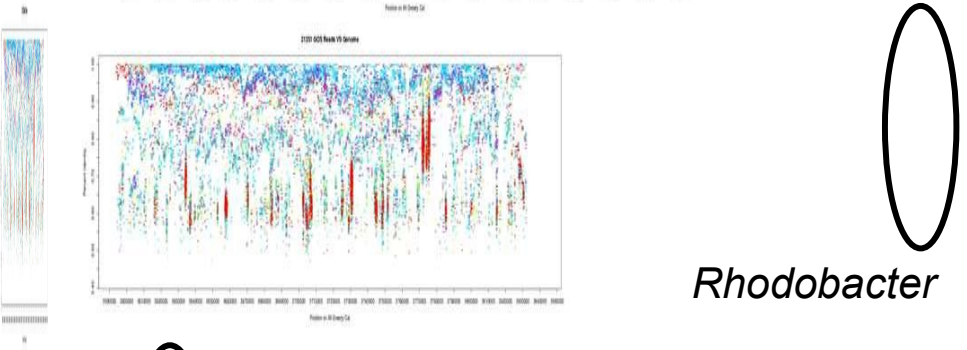
# Metagenomics "Extreme Assembly" Requires Large Amount of Pixel Real Estate



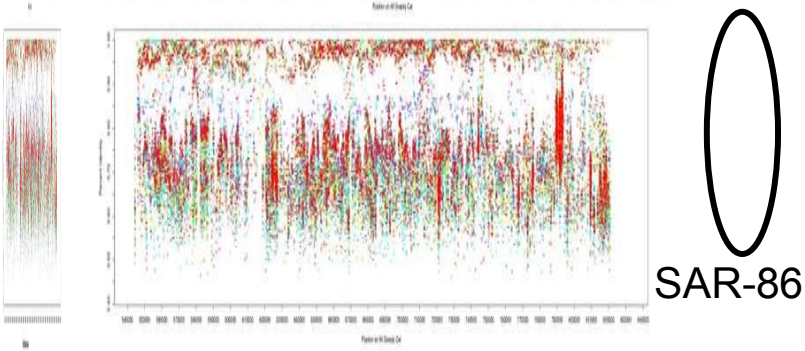
*Prochlorococcus*



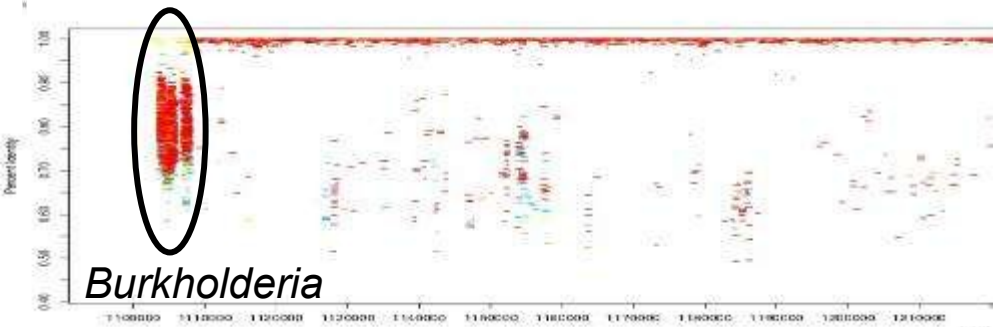
*Microbacterium*



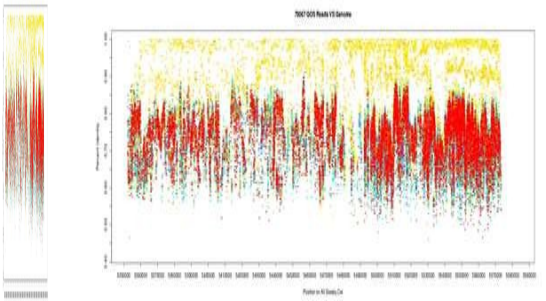
*Rhodobacter*



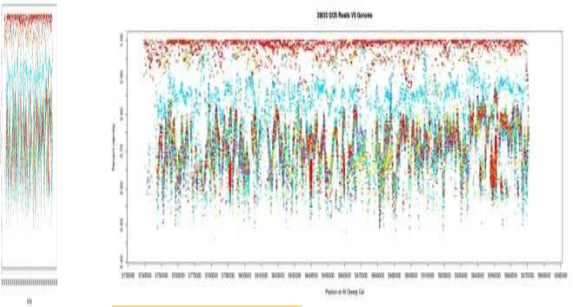
SAR-86



*Burkholderia*



unknown

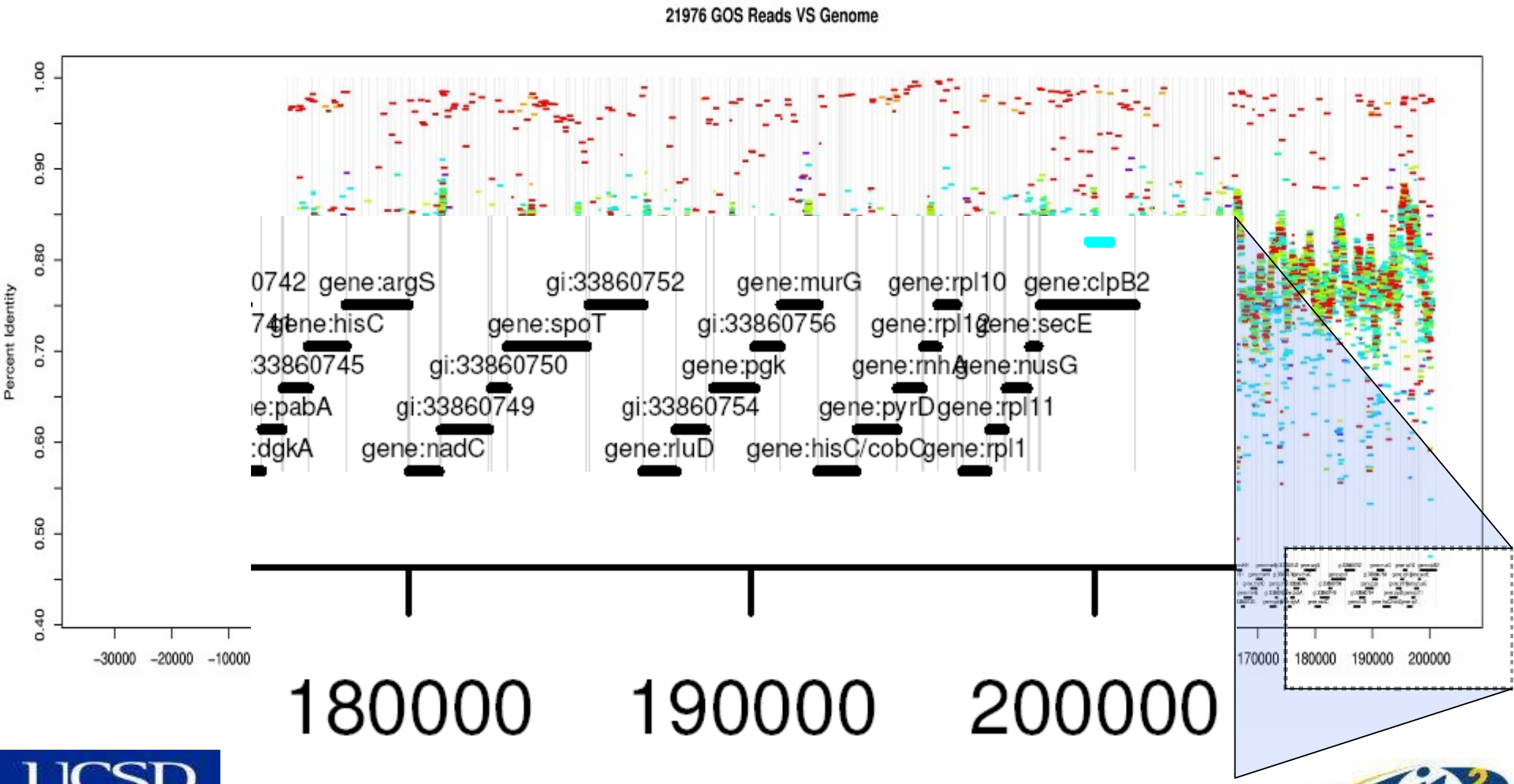


unknown

**Source: Karin Remington**
**J. Craig Venter Institute**

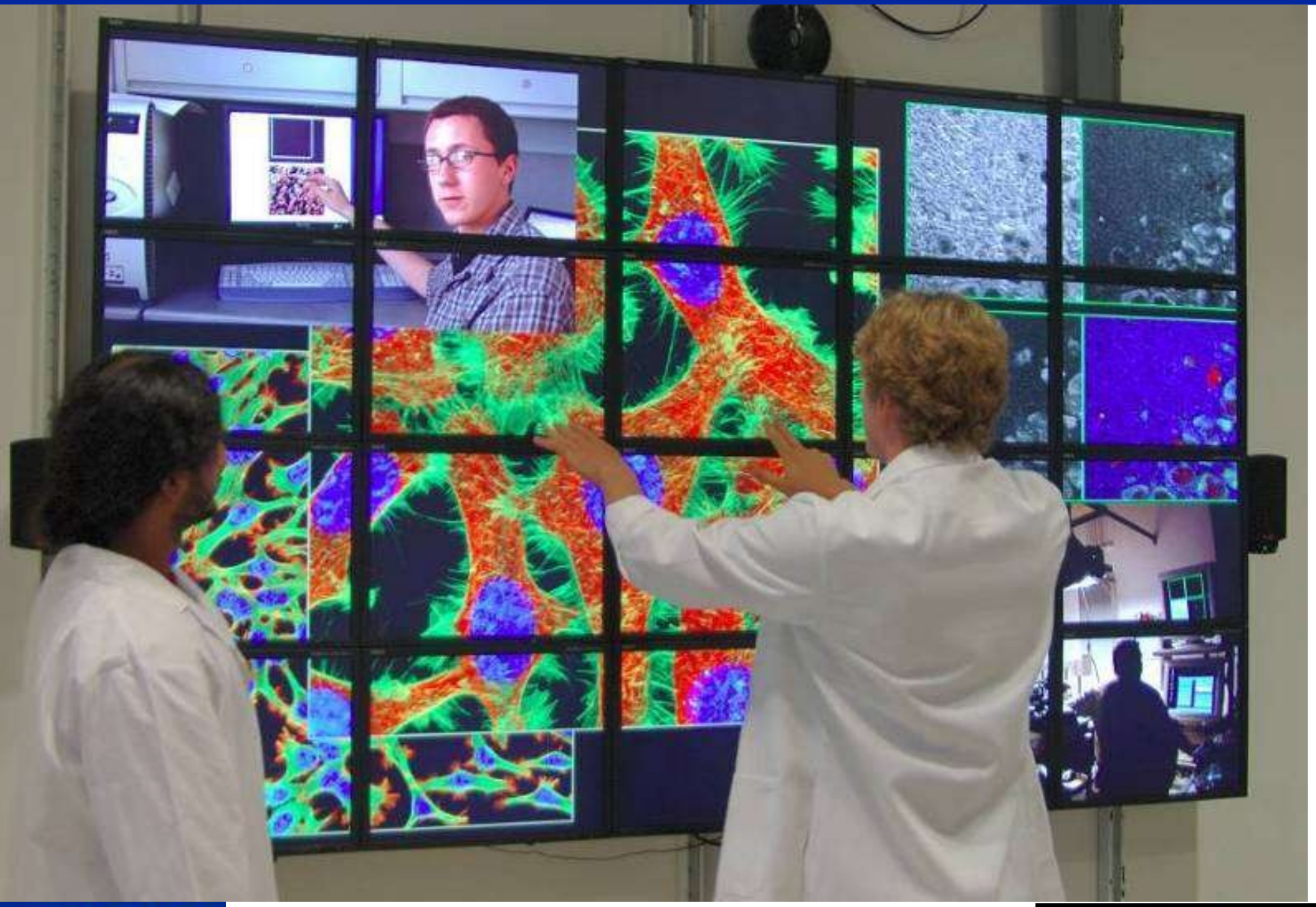
# Metagenomics Requires a Global View of Data and the Ability to Zoom Into Detail Interactively

## Overlay of Metagenomics Data onto Sequenced Reference Genomes (This Image: Prochloroccocus marinus MED4)




Source: Karin Remington
J. Craig Venter Institute

# OptIPuter Scalable Adaptive Graphics Environment (SAGE) Allows Integration of HD Streams



UCSD UCIrvine

evl electronic visualization laboratory

**Source: David Lee, NCMIR, UCSD**

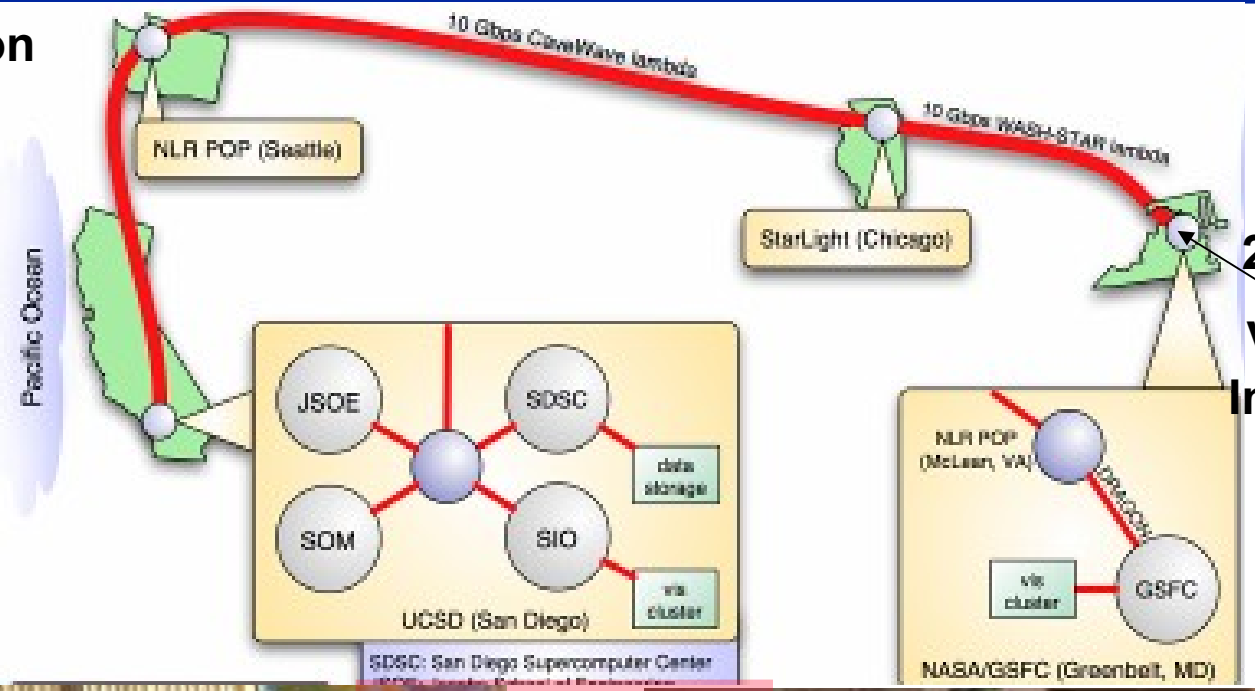
NCMIR NATIONAL CENTER for MICROSCOPY and IMAGING RESEARCH at San Diego, an NIH supported resource center

it2

# Calit2 and the Venter Institute Will Combine Telepresence with Remote Interactive Analysis

**Live Demonstration**

**of 21st Century National-Scale Team Science**




**25 Miles**

**Venter Institute**



**OptIPuter Visualized Data**

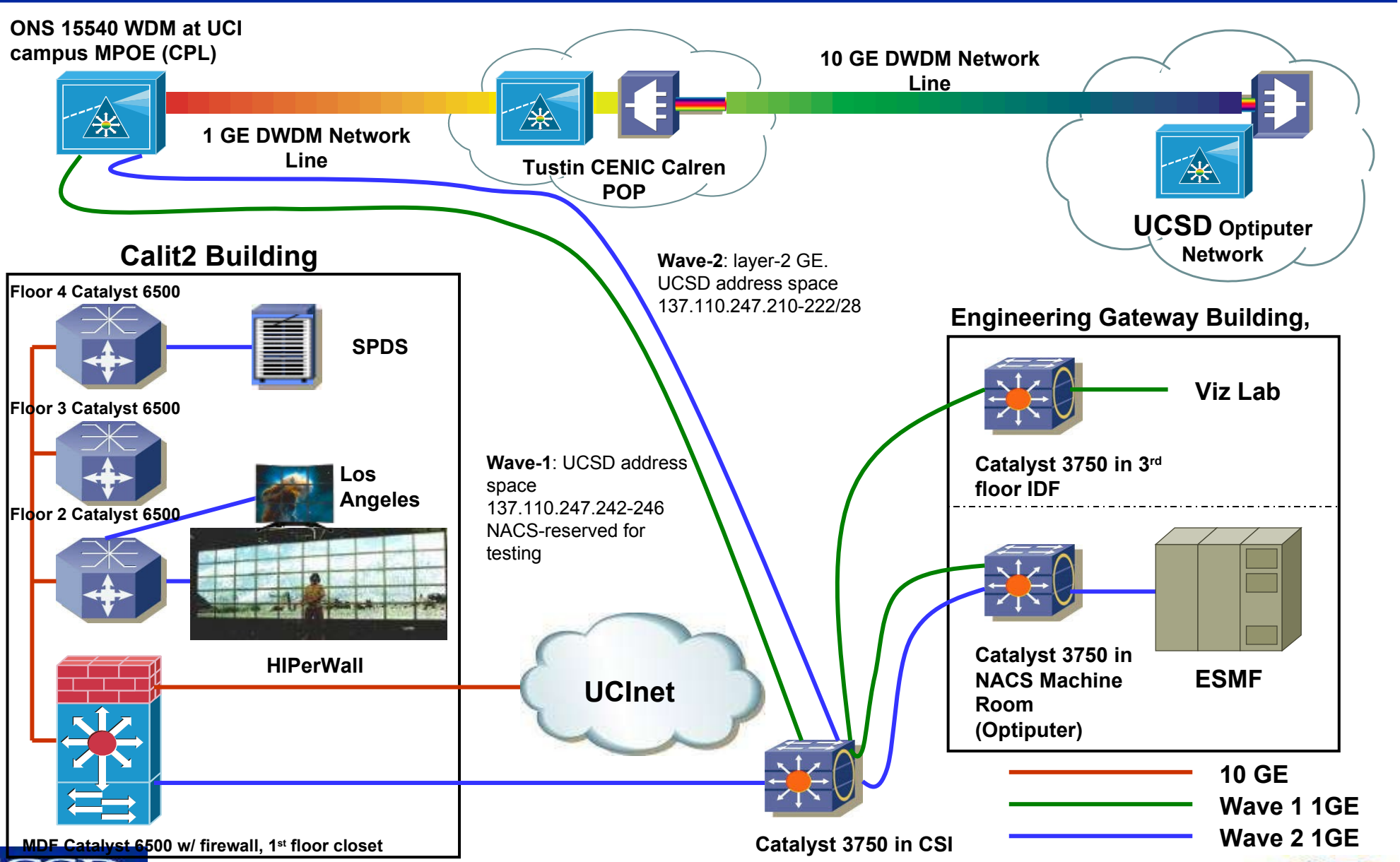
**HDTV Over Lambda**



UCSD

UCIrvine

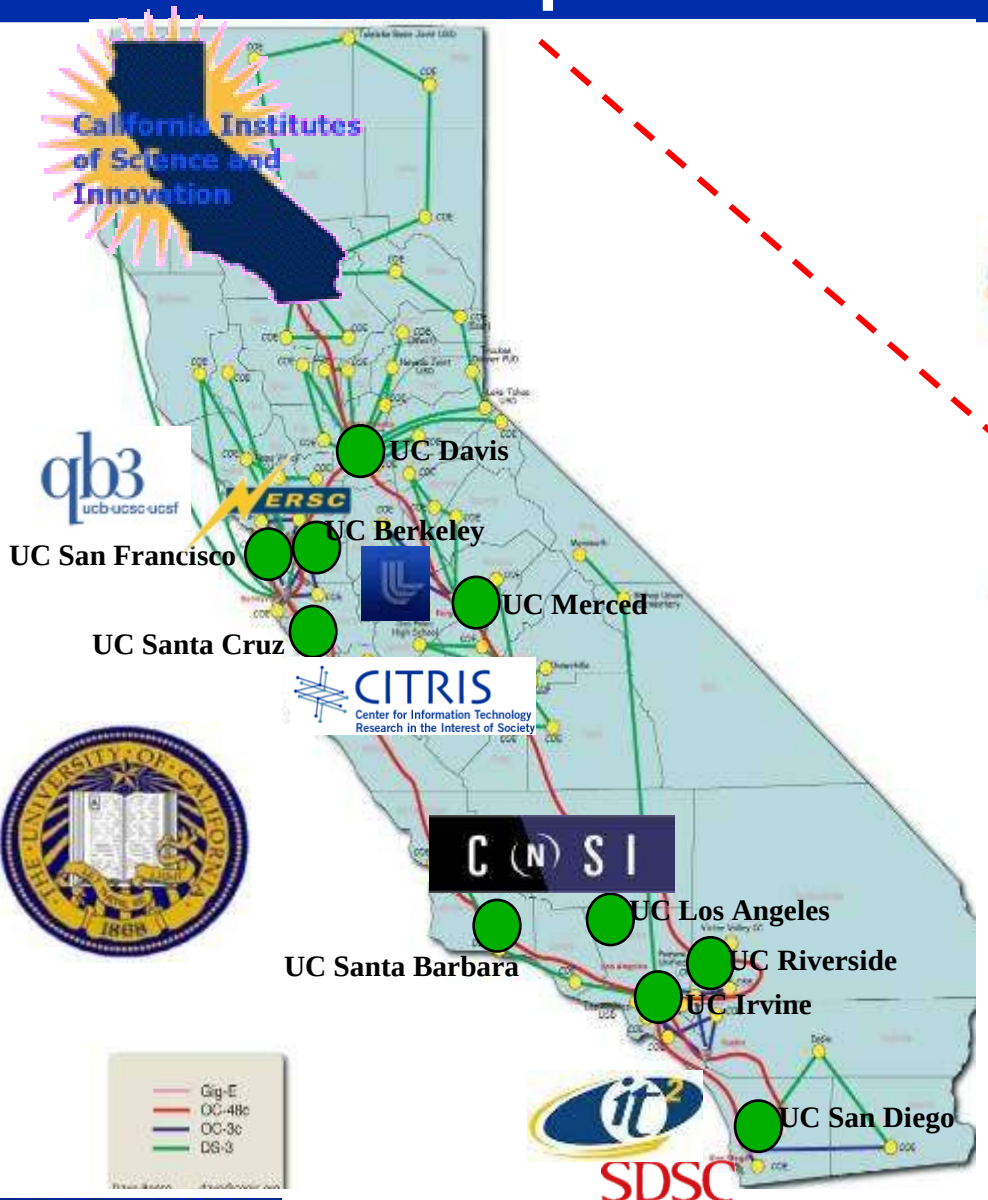
# OptIPuter@UCI is Up and Working




Created 09-27-2005 by Garrett Hildebrand

Modified 11-03-2005 by Jessica Yu

# Calit2/SDSC Proposal to Create a UC Cyberinfrastructure of "On-Ramps" to National LambdaRail Resources




## OptIPuter + CalREN-XD + TeraGrid = "OptiGrid"




**Creating a Critical Mass of End Users on a Secure LambdaGrid**

UCSD
UCIrvine

Source: Fran Berman, SDSC , Larry Smarr, Calit2