“Building an Information Infrastructure to Support Genetic Sciences"

Invited Talk

Celebrating a Decade of Genome Sequencing

UCSD

La Jolla, CA

December 6, 2005

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology;

Harry E. Gruber Professor,

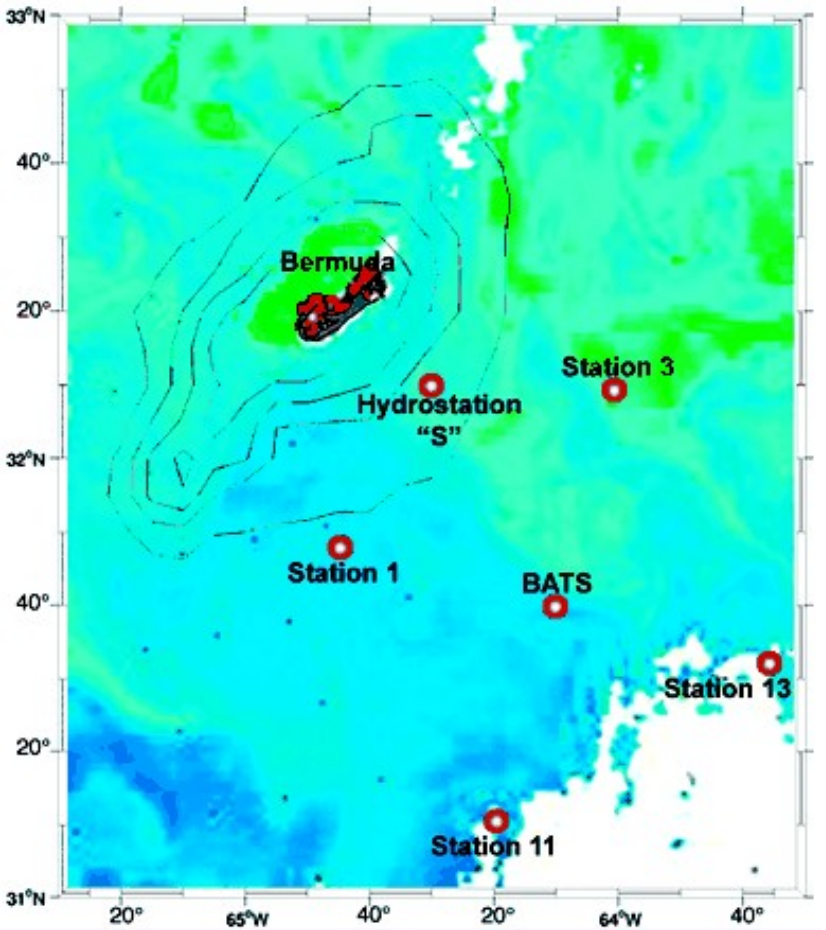
Dept. of Computer Science and Engineering

Jacobs School of Engineering, UCSD

OptIPuter

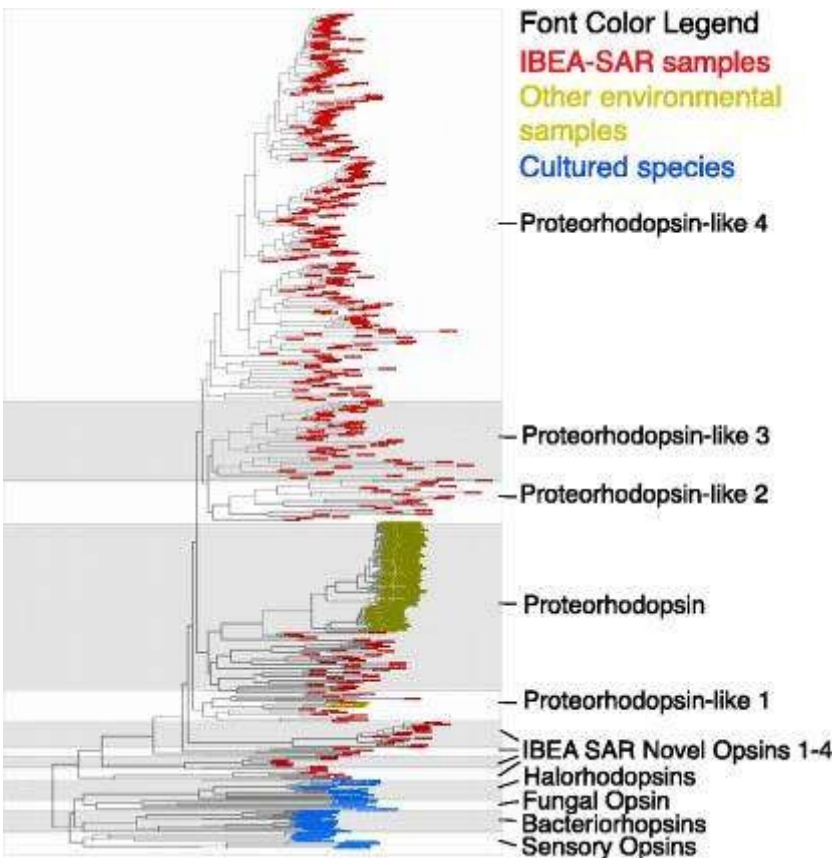
it2

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

OptIPuter

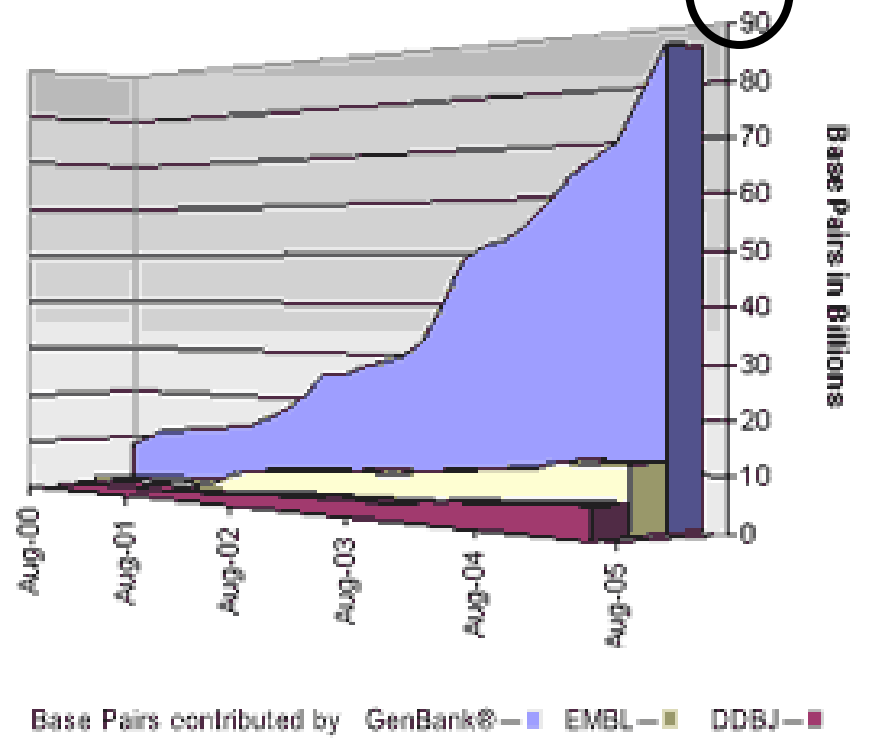
it2

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

100 Billion Bases!

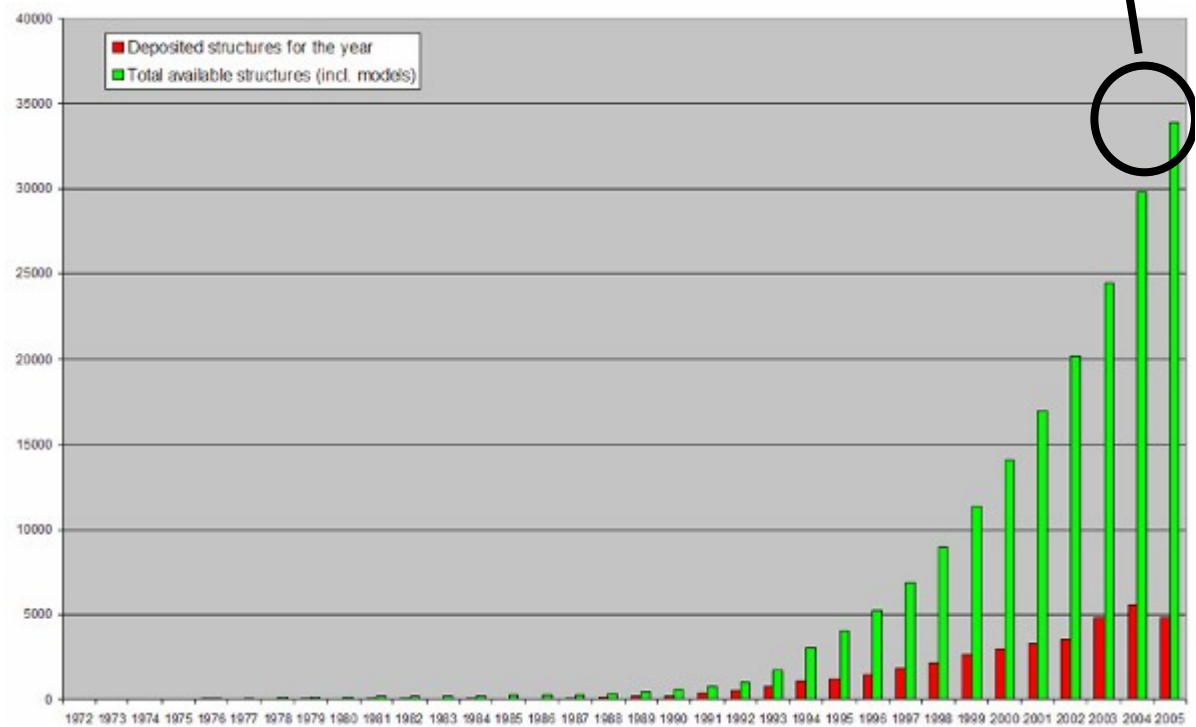
35,000 Structures

GenBank



www.ncbi.nlm.nih.gov/Genbank

Protein Data Bank



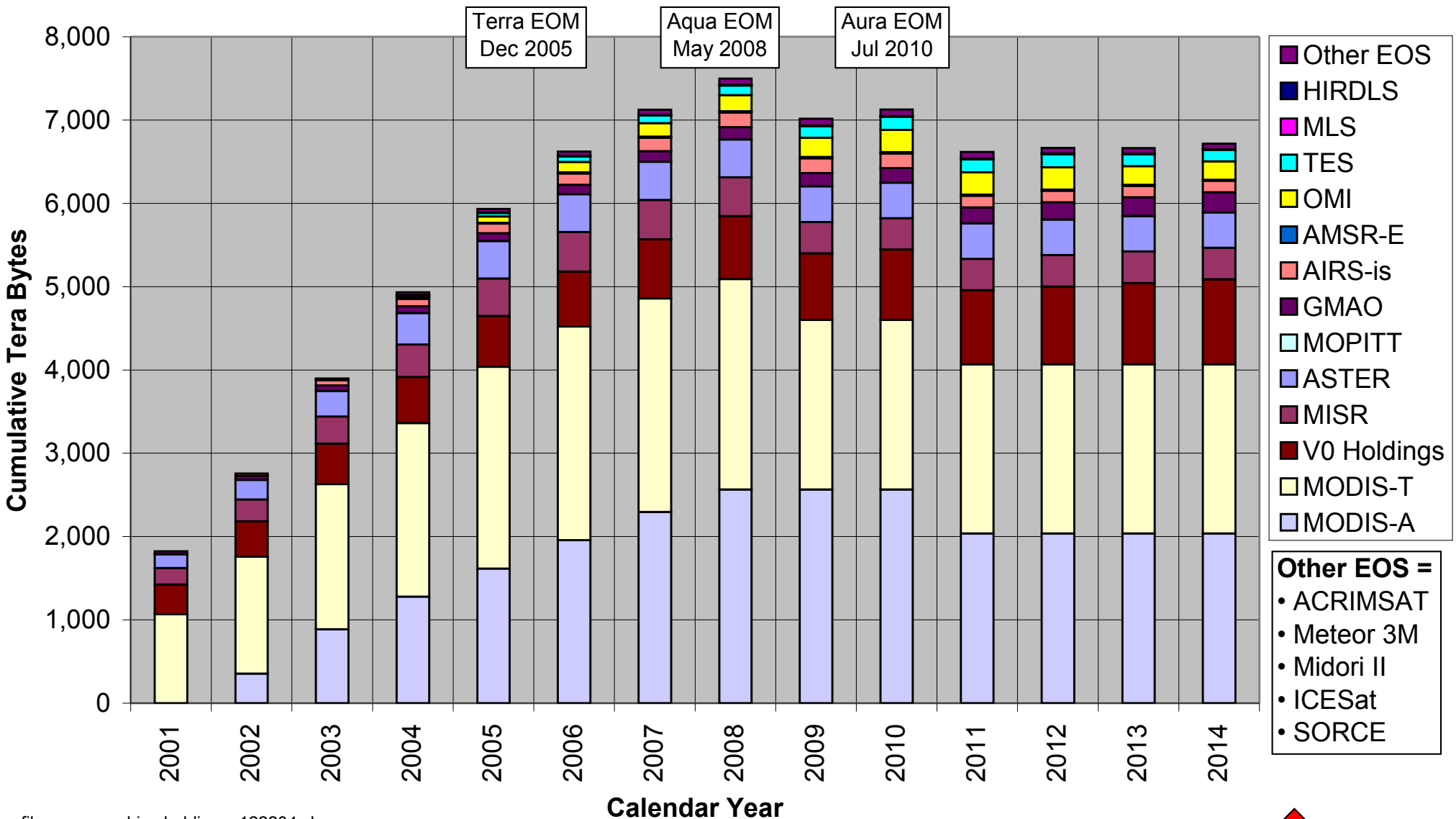
www.rcsb.org/pdb/holdings.html

OptIPuter

Total Data < 1TB

it2

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

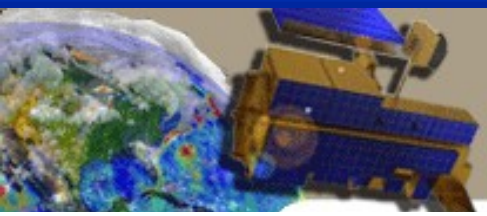


file name: archive holdings_122204.xls
tab: all instr bar

NOTE: Data remains in the archive pending transition to LTA

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

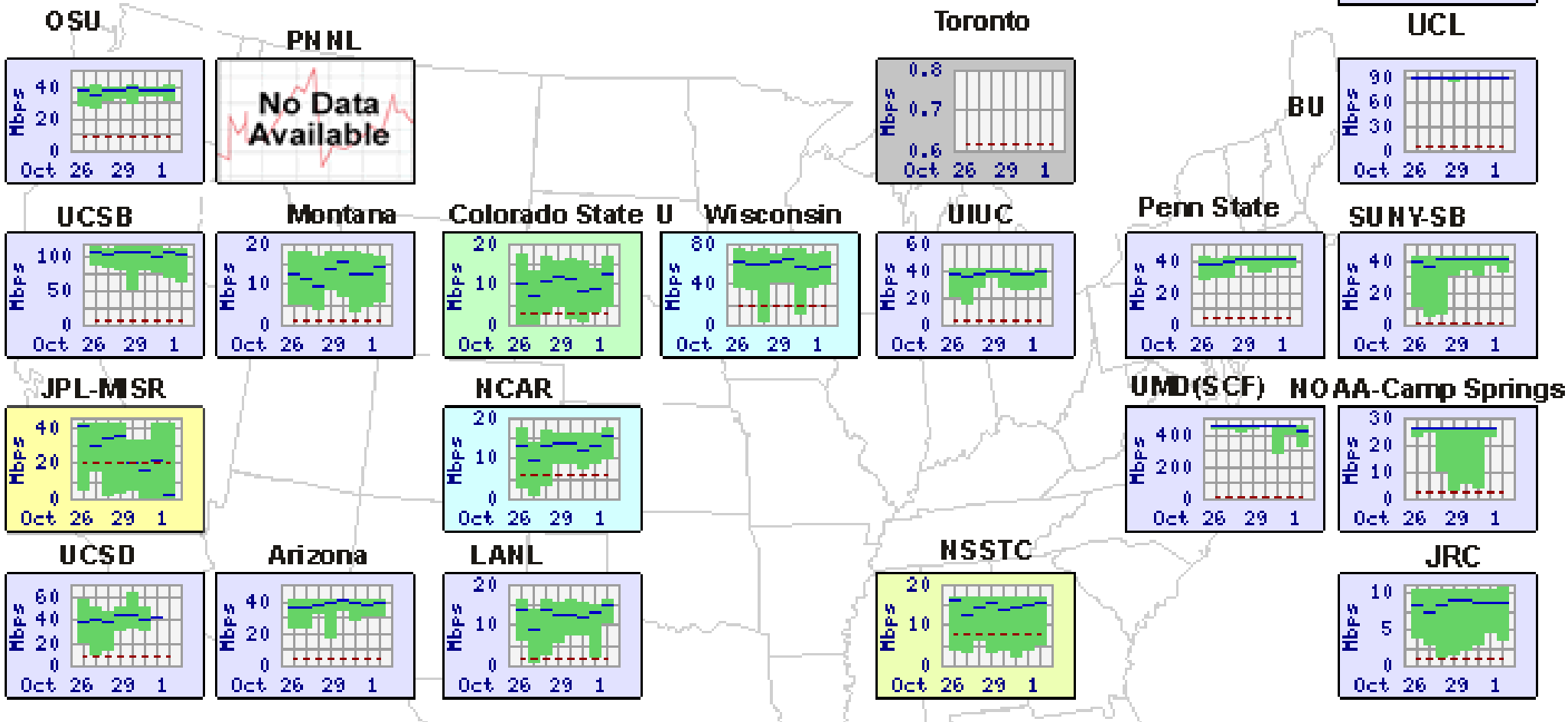
Challenge: Average Throughput of NASA Data Products to End User is < 50 Mbps



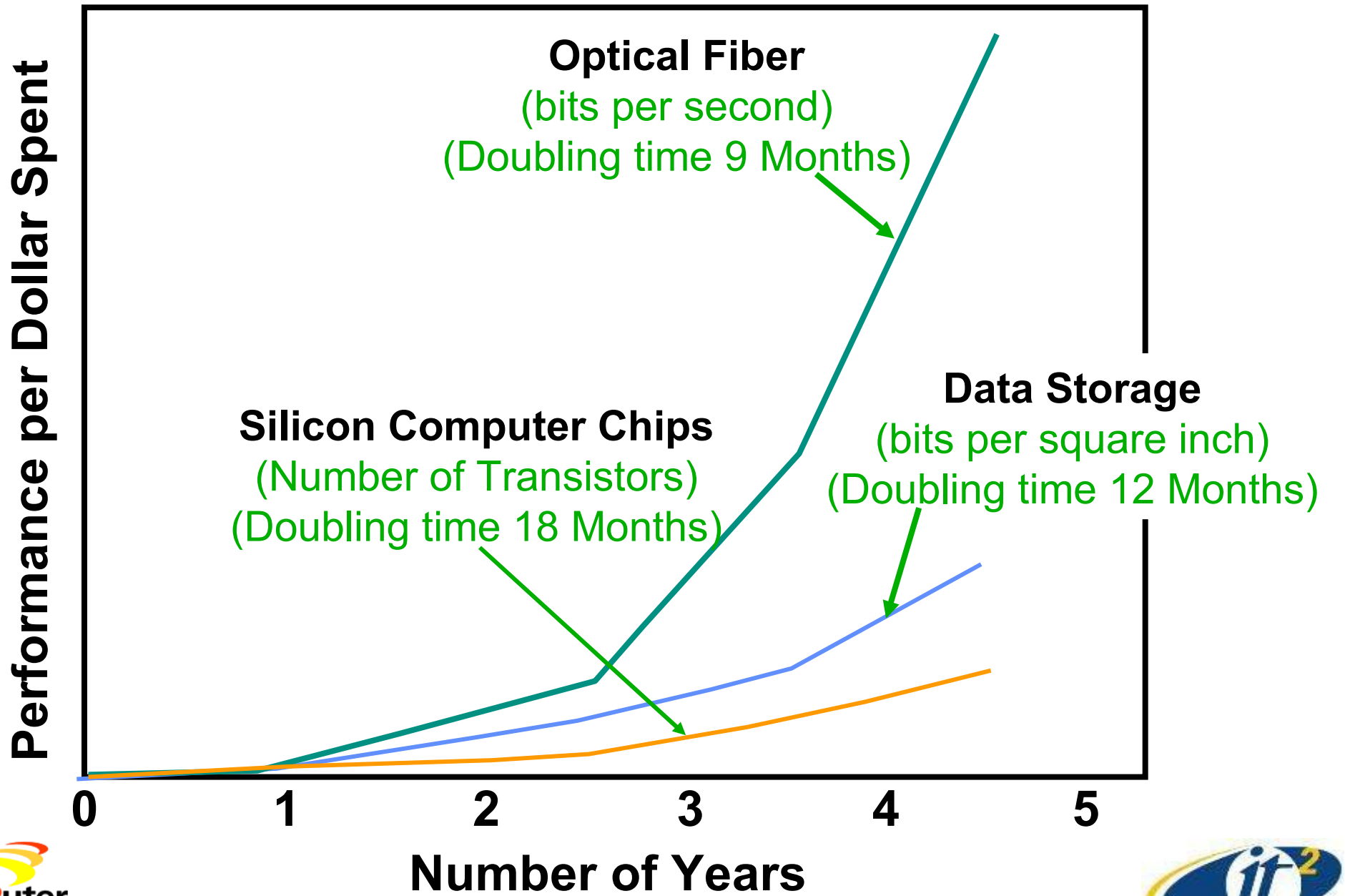
ENSIGHT Active Testing TERRA Destinations

Tested October 2005

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



Why Optical Networks Will Become the 21st Century Driver

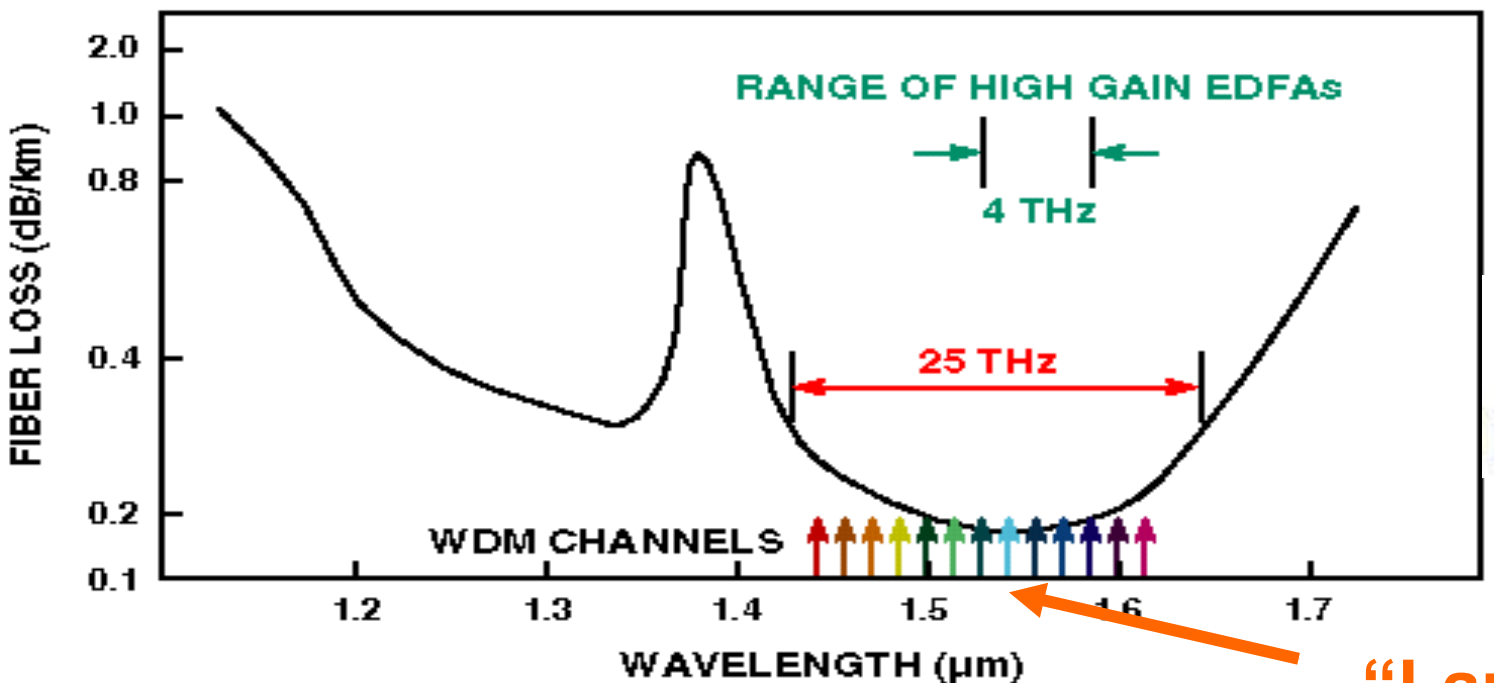


Scientific American, January 2001

Solution: Individual 1 or 10Gbps Lightpaths -- "Lambdas on Demand"

WAVELENGTH DIVISION MULTIPLEXING (WDM)

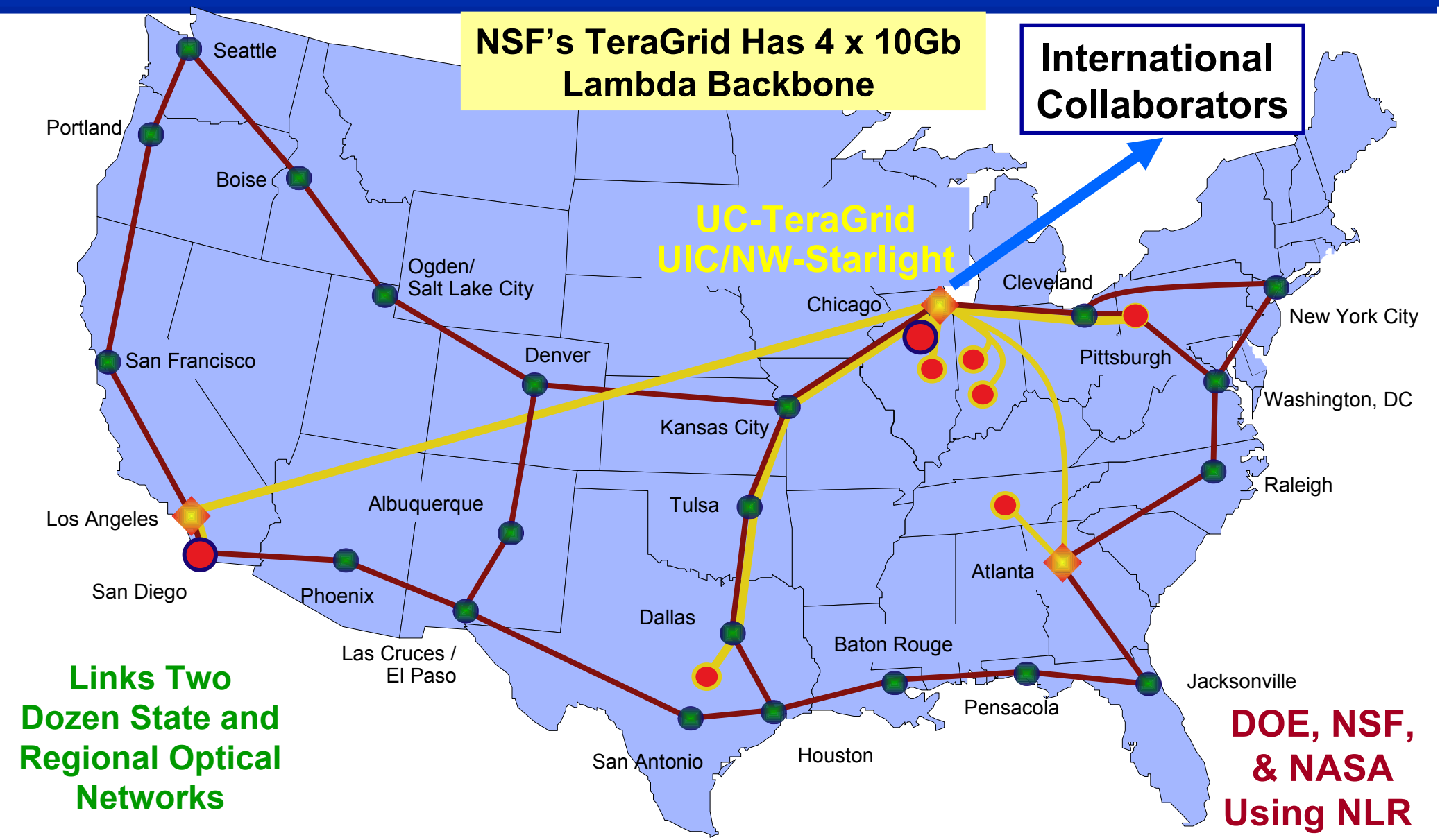
- EXPLOITS
  - ENORMOUS BANDWIDTH OF SILICA FIBER
  - HIGH-GAIN WIDEBAND OPTICAL AMPLIFIERS



"Lambdas"

Source: Steve Wallach, Chiaro Networks

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



NLR 4 x 10Gb Lambdas Initially Capable of 40 x 10Gb wavelengths at Buildout

OptIPuter

it2

Calit2@UCSD Is Connected to the World at 10,000 Mbps

Maxine Brown, Tom DeFanti, Co-Chairs

iGrid 2005

THE GLOBAL LAMBDA INTEGRATED FACILITY

www.igrid2005.org



the UC Discovery Grant

September 26-30, 2005

Calit2 @ University of California, San Diego

California Institute for Telecommunications and Information Technology



50 Demonstrations, 20 Counties, 10 Gbps/Demo

Prototyping Cabled Ocean Observatories Enabling High Definition Video Exploration of Deep Sea Vents



iGrid 2005

Source John Delaney & Deborah Kelley, UWash

it2

A Near Future Metagenomics Fiber Optic Cable Observatory



Source John Delaney, UWash

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

OptIPuter

it2

Driving Cyberinfrastructure with Environmental Metagenomics

Approved Yesterday!



Proposal to the
Gordon and Betty Moore Foundation

From Organismal Genomics to Environmental Metagenomics: Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA)

Larry Smarr, UCSD/Calit2, PI
Marv Frazier, JCVI, co-investigator

Marine Microbial Metagenomics
From Species Genomes to Ecological Genomes

- **Each Sequence is a Part of an Entire Biological Community**
- **Complex Data Set Including Sequences, Genes and Gene Families, Coupled With Environmental Metadata**
  - **Tremendous Potential to Better Understand the Functioning of Natural Ecosystems**
- **Challenge**
  - **Powerful Information Infrastructure Required to Support Metagenomics and to Create Co-laboratories**

Birch Aquarium

SEA OF GENES Now Open

J. Craig Venter INSTITUTE

Scripps Institution of Oceanography
Launches Scripps Genome Center
Terry Gaasterland leads new program

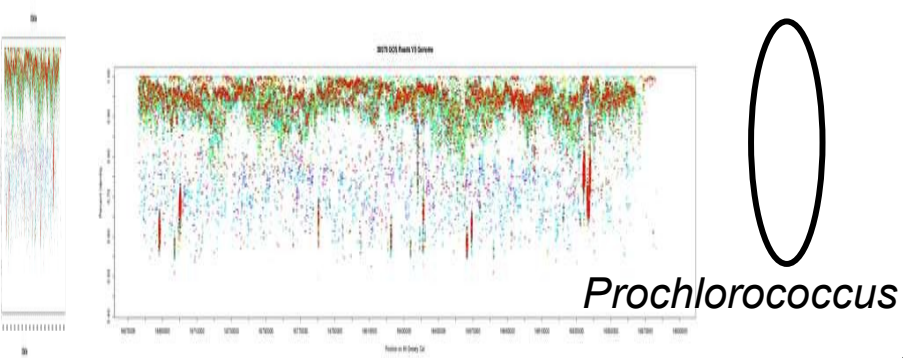
CENTER FOR EARTH OBSERVATIONS & APPLICATIONS

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

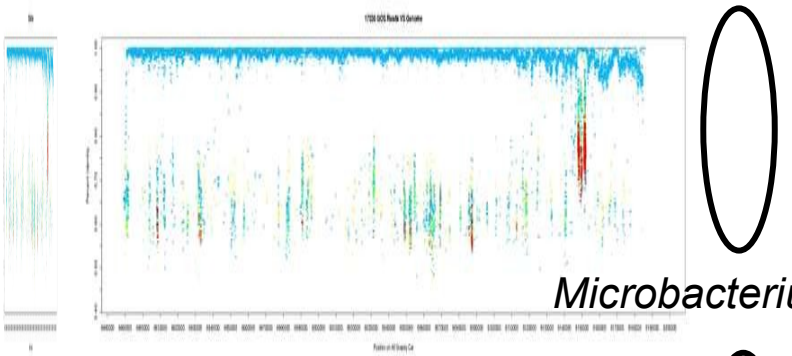
OptIPuter

it2

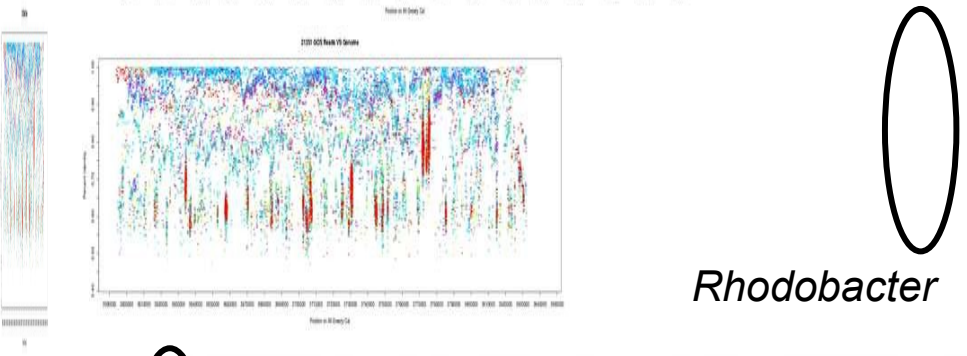
Metagenomics "Extreme Assembly" Requires Large Amount of Pixel Real Estate



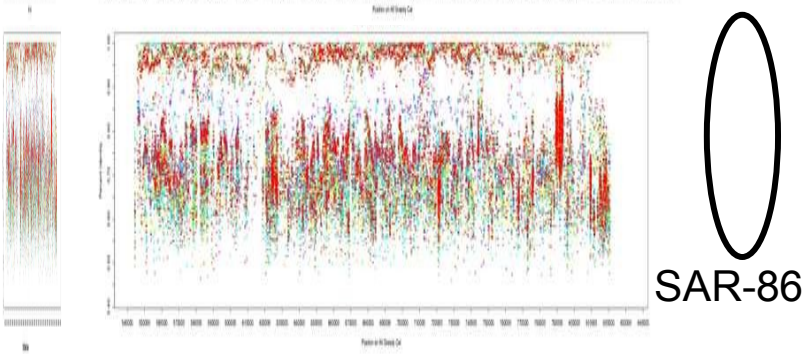
Prochlorococcus



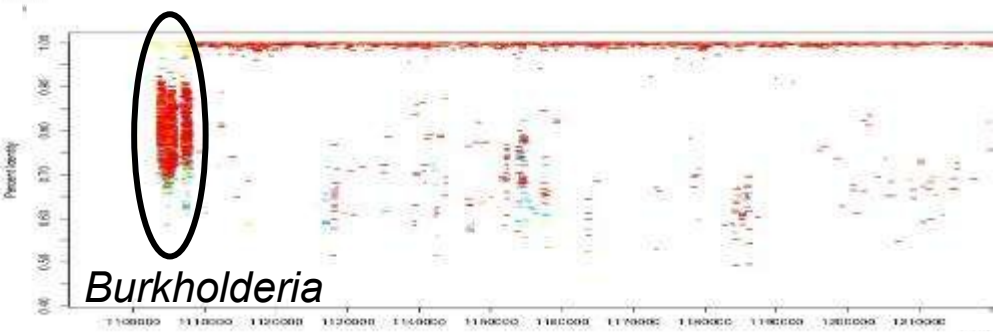
Microbacterium



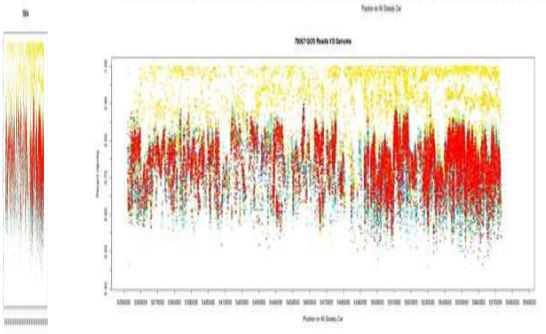
Rhodobacter



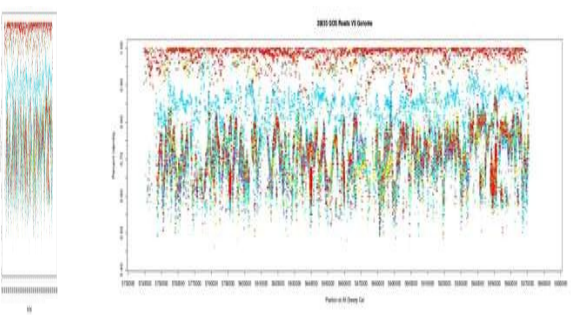
SAR-86



Burkholderia



unknown

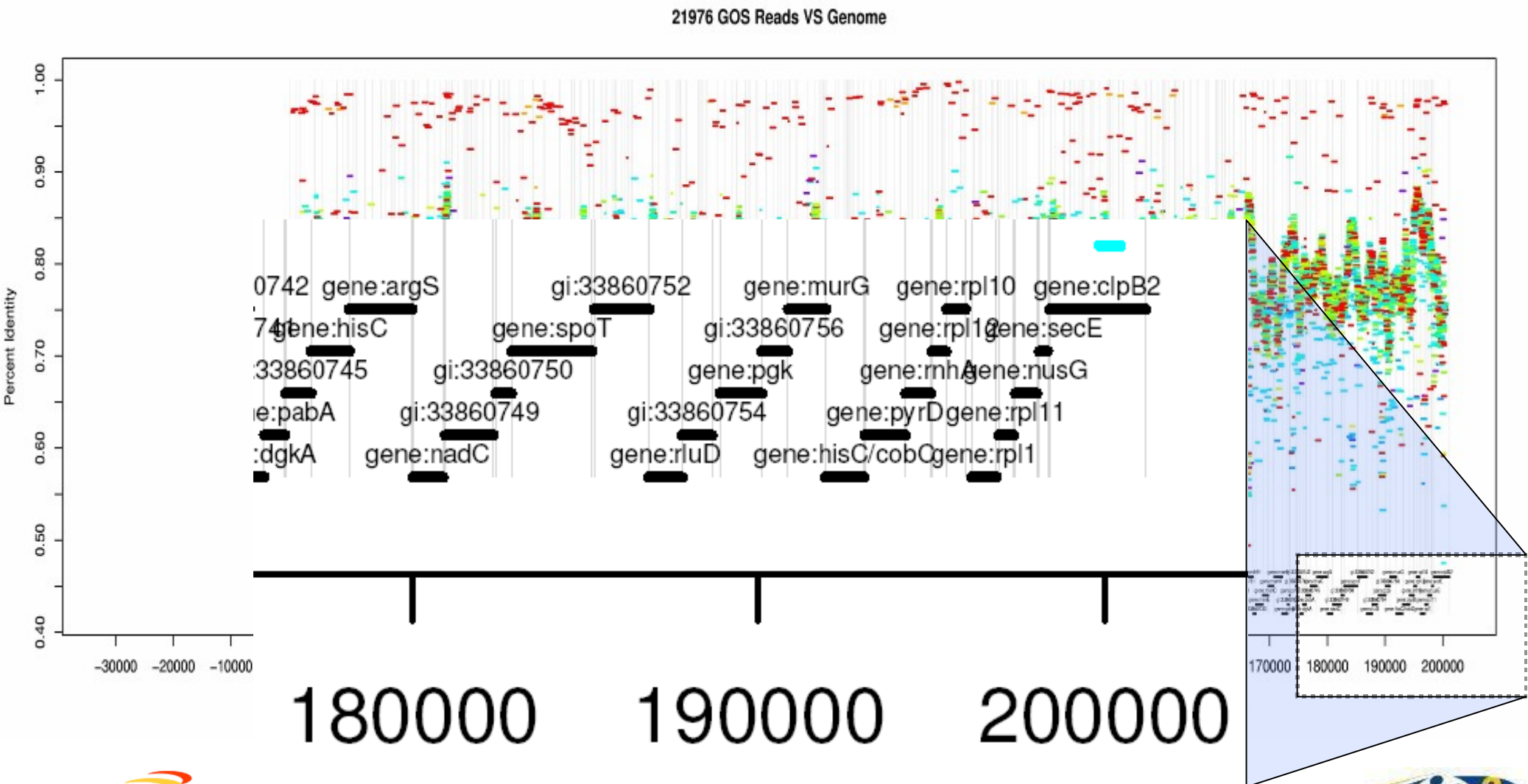


unknown

Source: Karin Remington
J. Craig Venter Institute

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

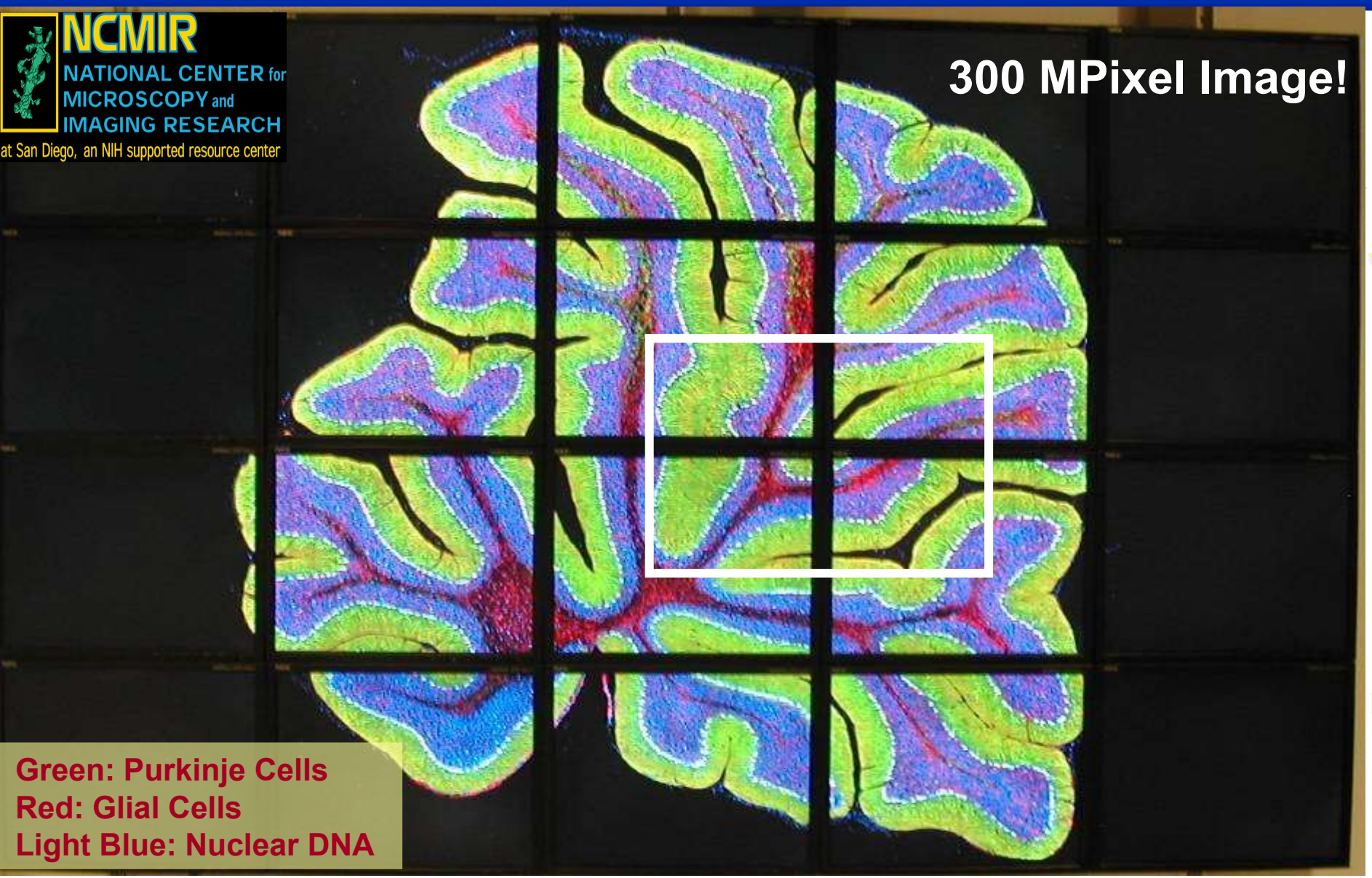
Overlay of Metagenomics Data onto Sequenced Reference Genomes (This Image: Prochloroccocus marinus MED4)



Source: Karin Remington
J. Craig Venter Institute

The OptIPuter – Creating High Resolution Portals Over Dedicated Optical Channels to Global Science Data

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

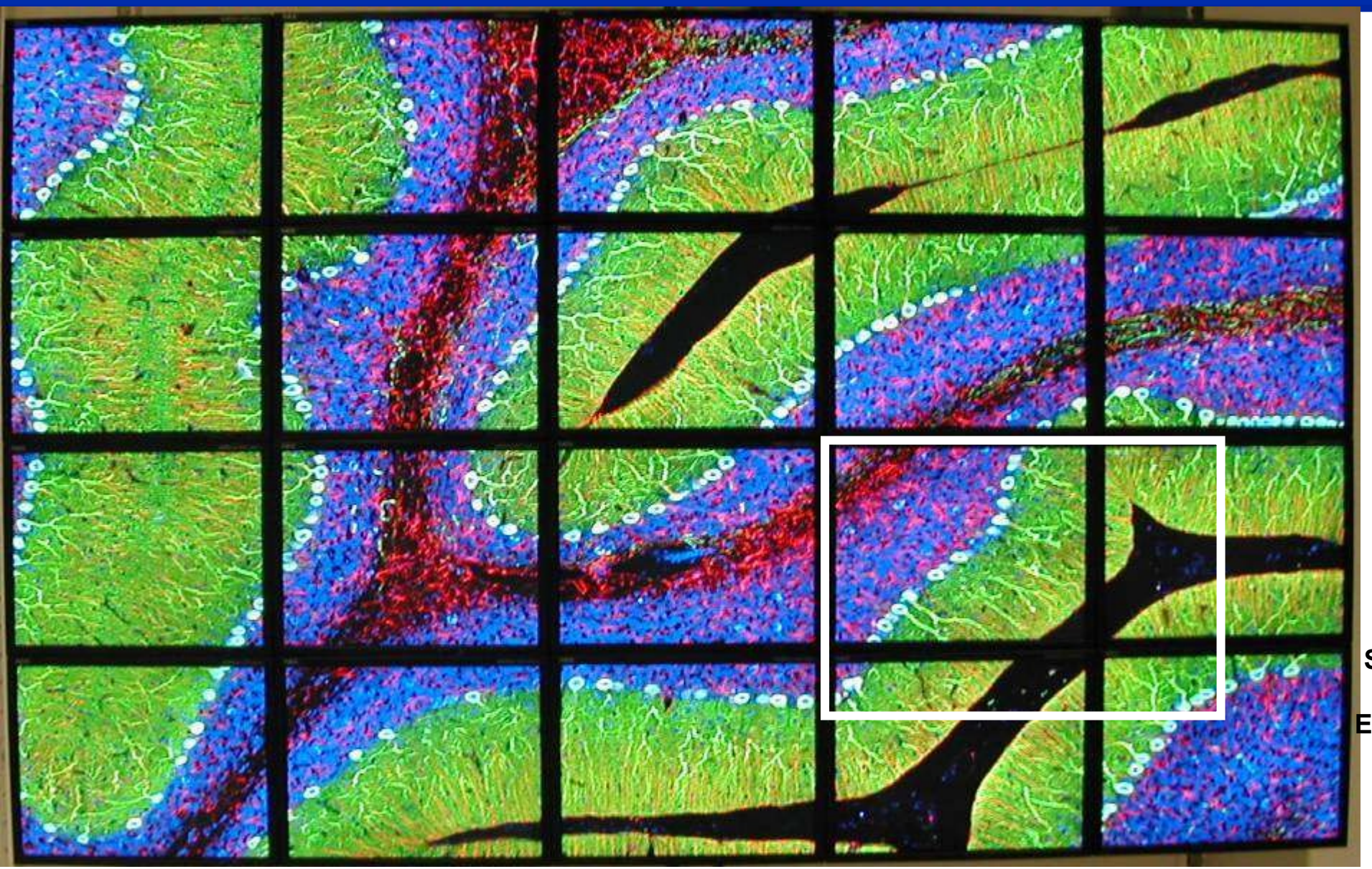


Source: Mark Ellisman, David Lee, Jason Leigh

OptIPuter

Calit2 (UCSD, UCI) and UIC Lead Campuses—Larry Smarr PI
Partners: SDSC, USC, SDSU, NW, TA&M, UvA, SARA, KISTI, AIST

Scalable Displays Allow Both Global Content and Fine Detail



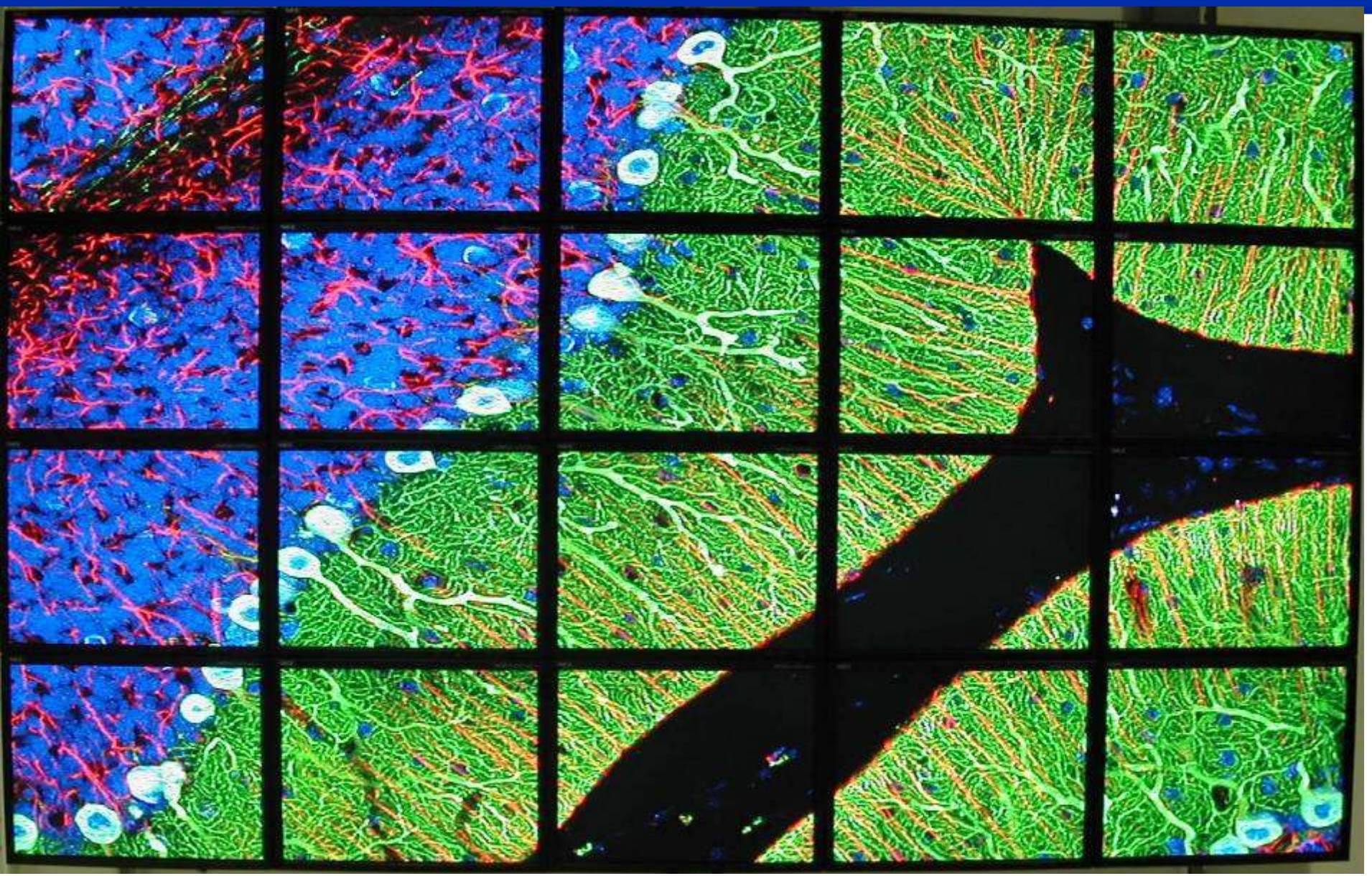
Source: Mark Ellisman, David Lee, Jason Leigh

OptIPuter

30 MPixel SunScreen Display Driven by a 20-node Sun Opteron Visualization Cluster

it2

Allows for Interactive Zooming from Cerebellum to Individual Neurons

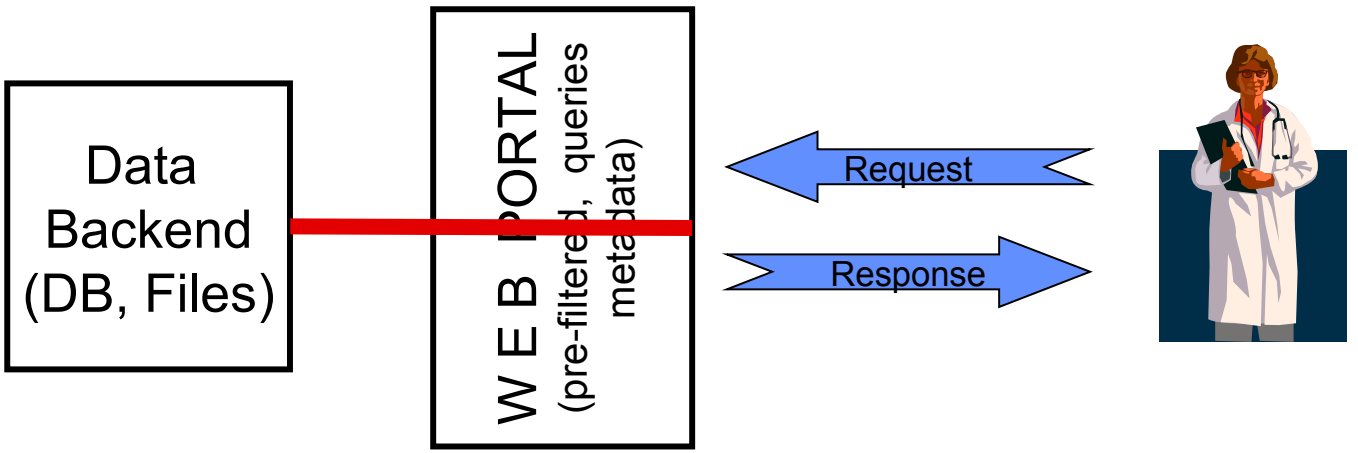


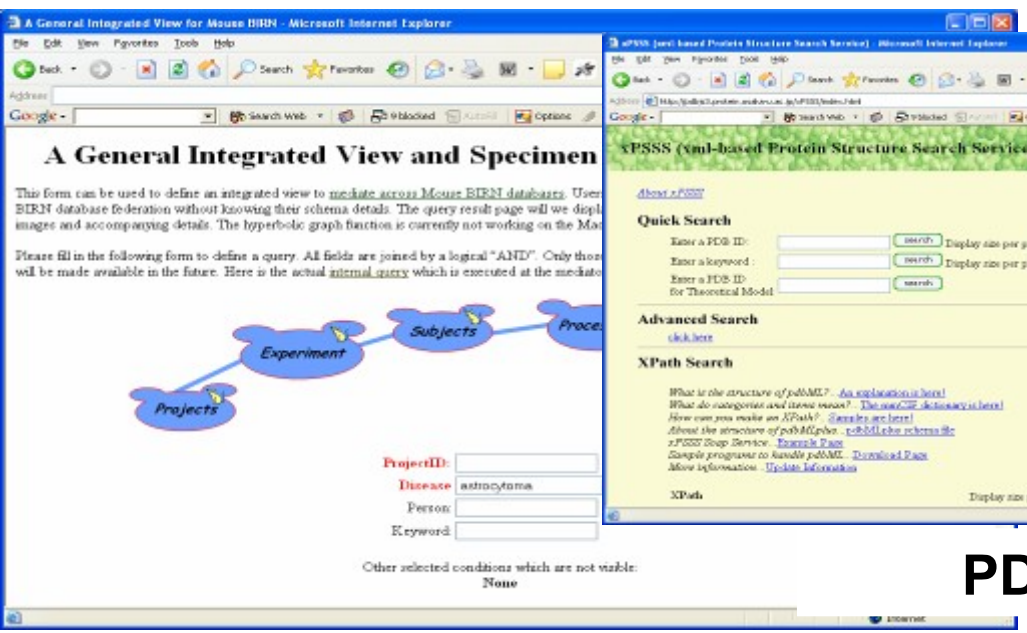
OptIPuter

Source: Mark Ellisman, David Lee, Jason Leigh

it2

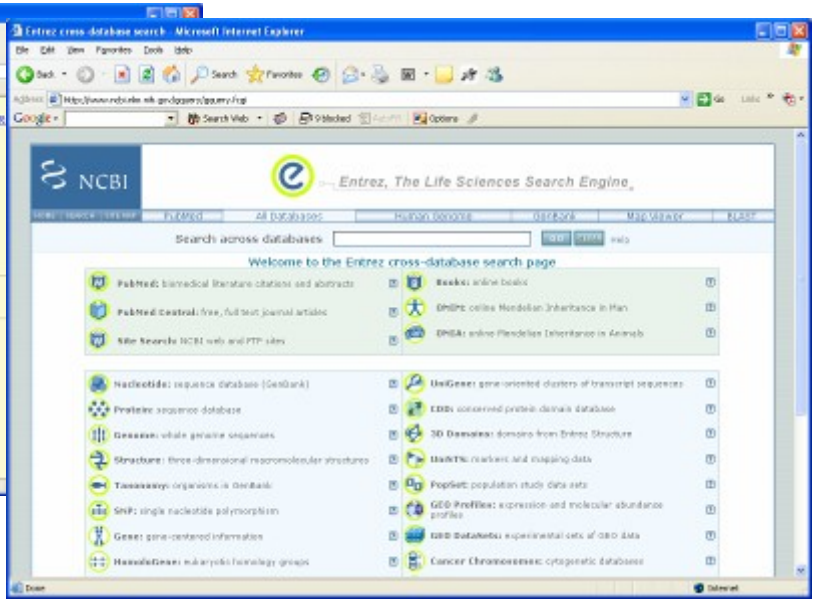
Calit2 Intends to Jump Beyond Traditional Web-Accessible Databases





BIRN

PDB



NCBI Genbank

+ many others

Source: Phil Papadopoulos, SDSC, Calit2

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

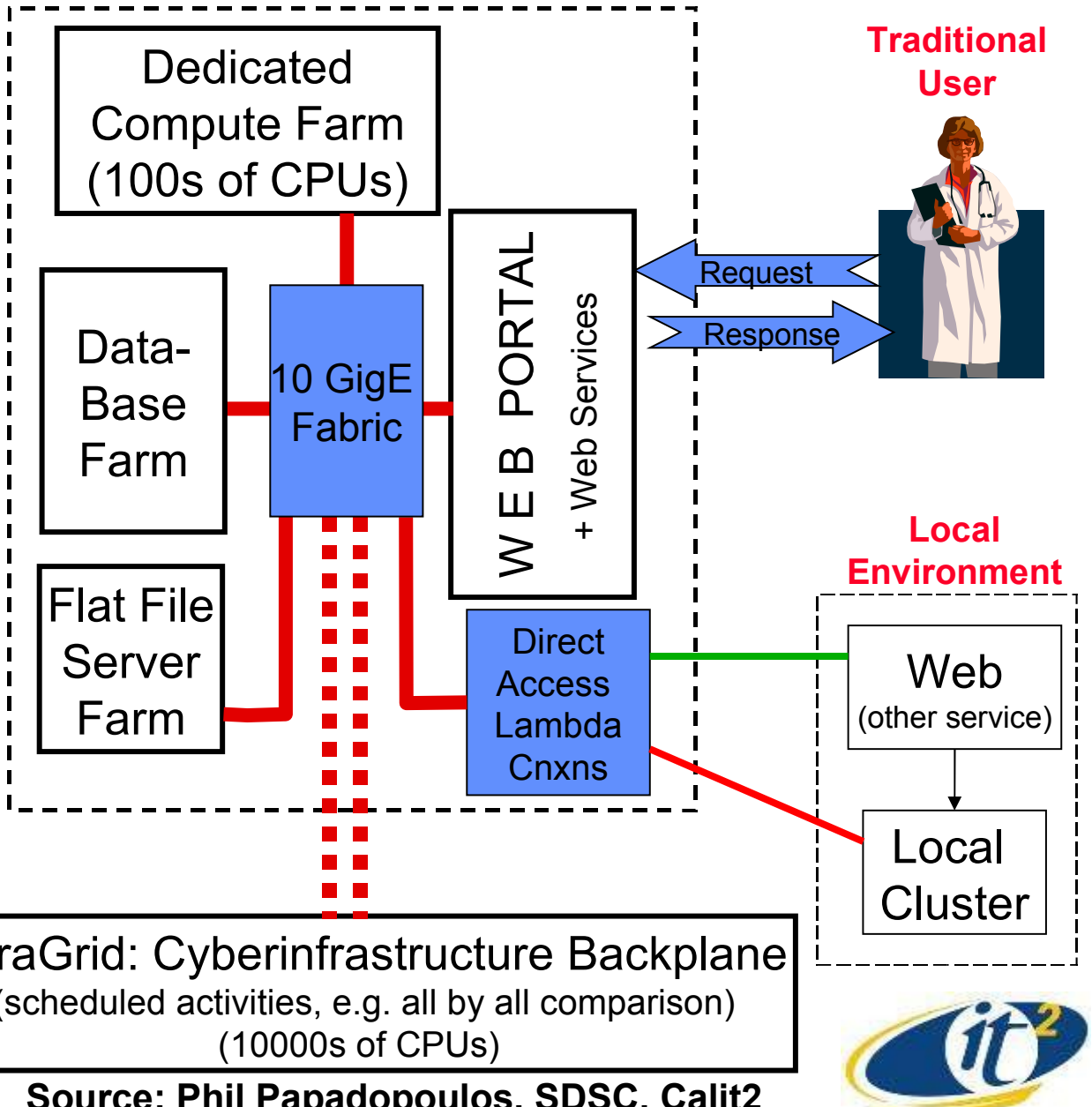
Sargasso Sea Data

Sorcerer II Expedition (GOS)

JGI Community Sequencing Project

Moore Marine Microbial Project

NASA Goddard Satellite Data



Source: Phil Papadopoulos, SDSC, Calit2

CENTER FOR EARTH OBSERVATIONS & APPLICATIONS

J. Craig Venter INSTITUTE

OptIPuter

SDSC SAN DIEGO SUPERCOMPUTER CENTER

it2

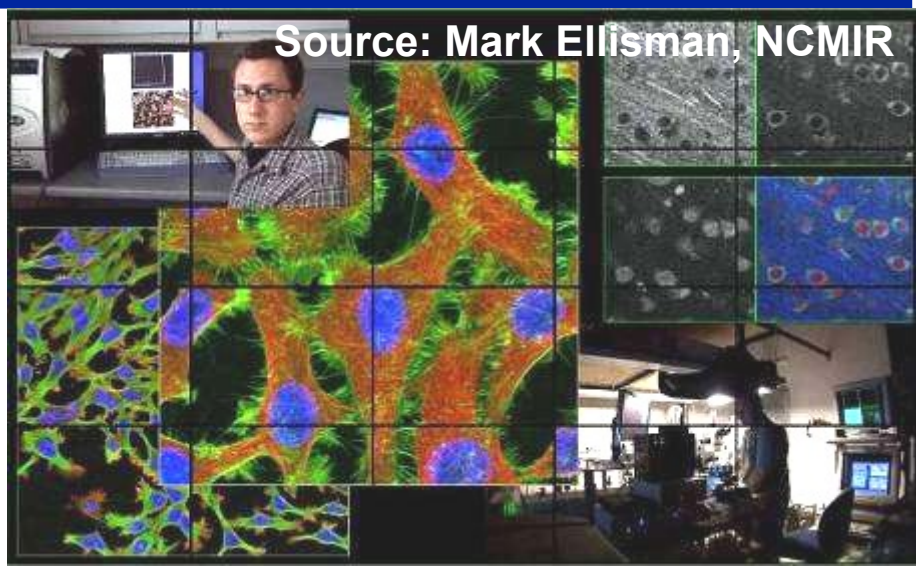
Analysis Data Sets, Data Services, Tools, and Workflows

- 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

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

The OptIPuter Enabled Collaboratory: Remote Researchers Jointly Exploring Complex Data

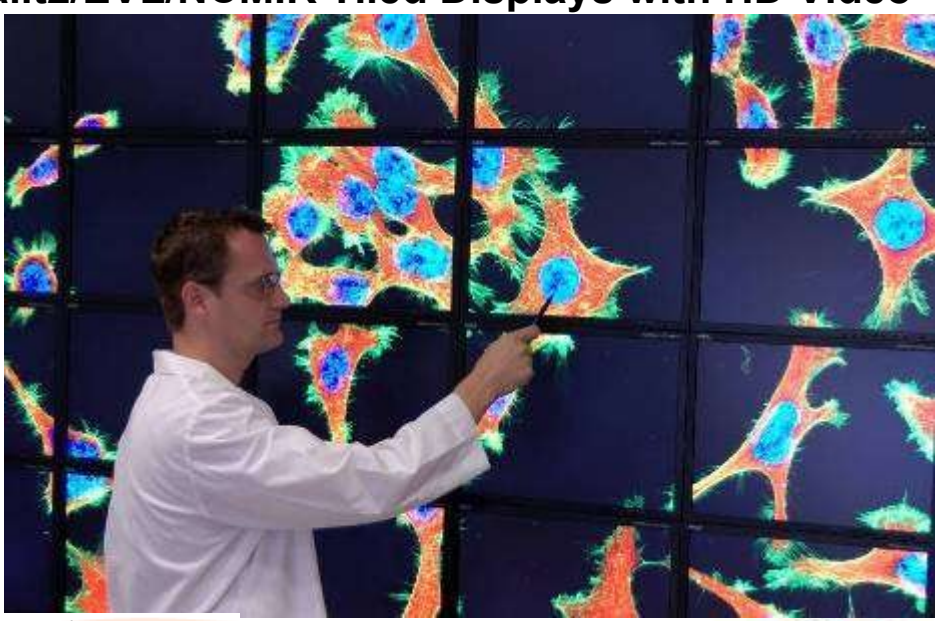




Calit2/EVL/NCMIR Tiled Displays with HD Video

New Home of SDSC/Calit2 Synthesis Center





OptIPuter

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

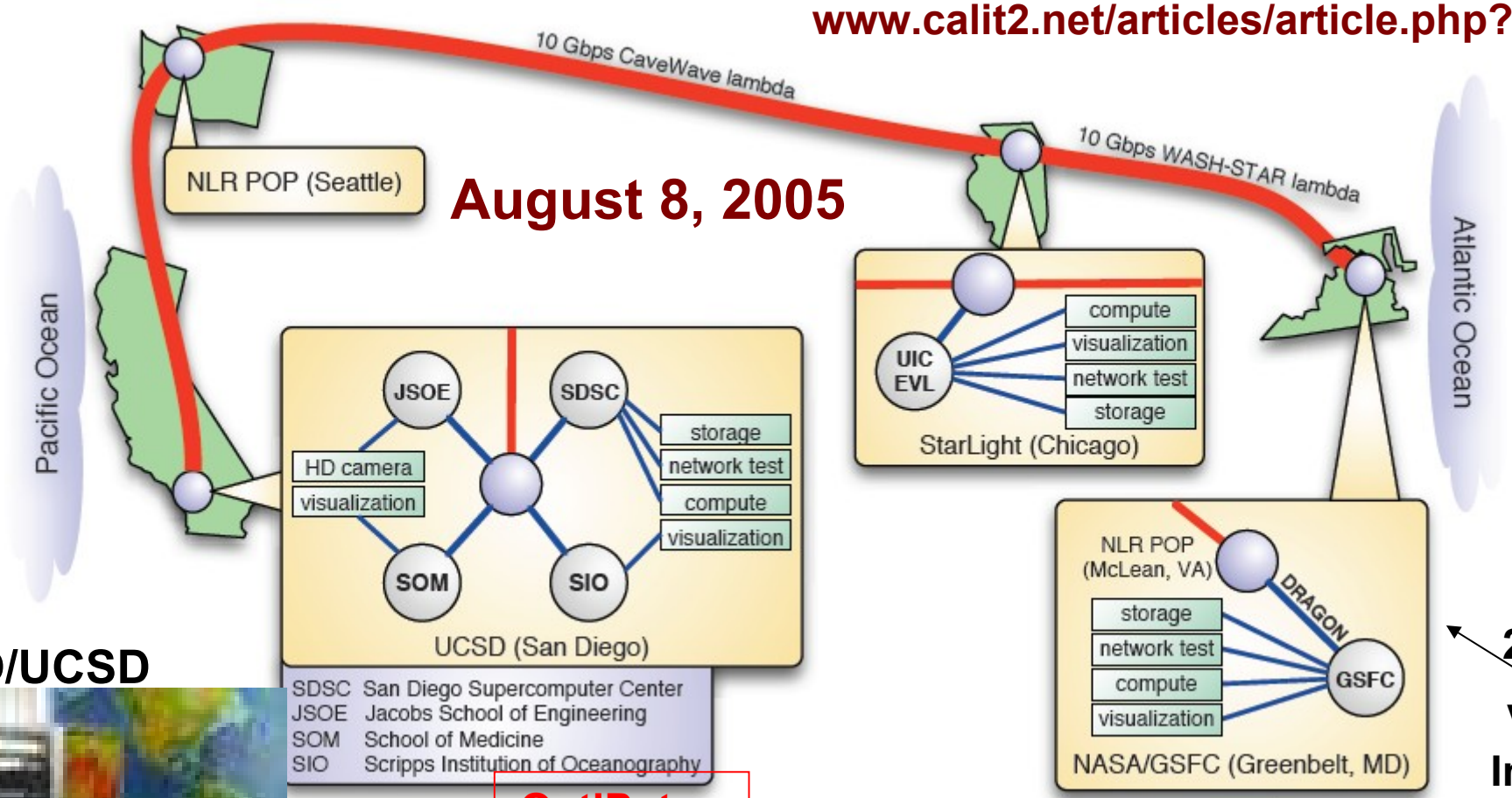
evl electronic visualization laboratory

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

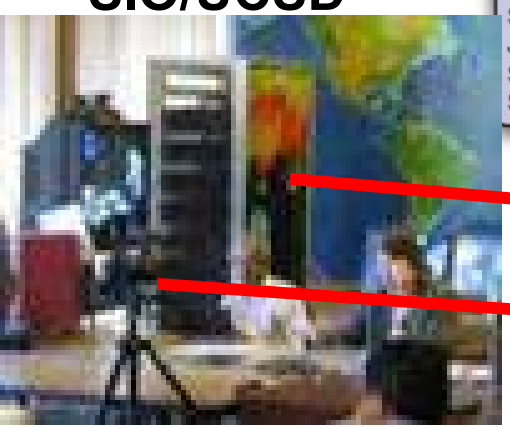
it2

Eliminating Distance to Unify Remote Laboratories

www.calit2.net/articles/article.php?id=660



SIO/UCSD



OptIPuter Visualized Data

HDTV Over Lambda

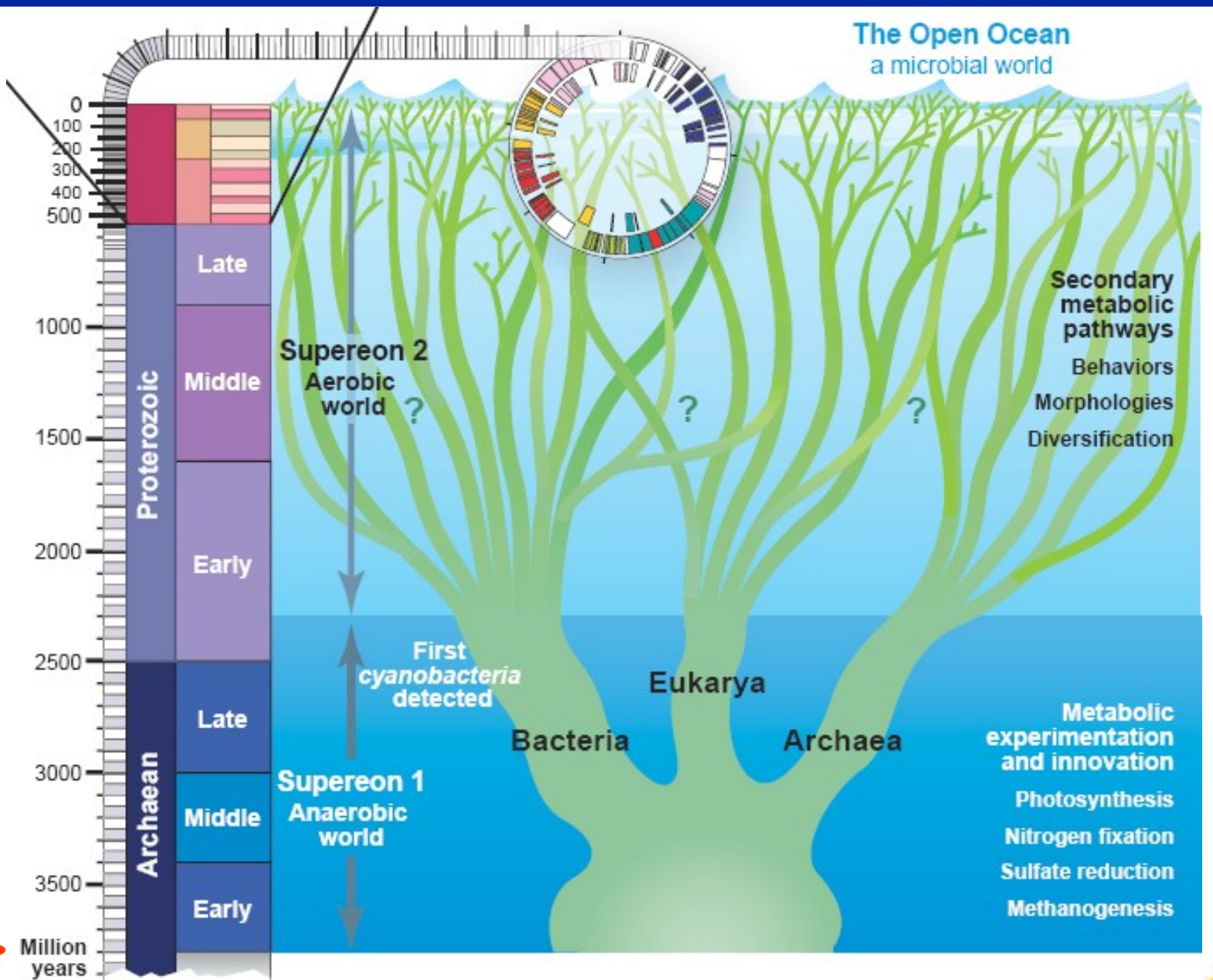


NASA Goddard

OptIPuter

it2

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



Science Falkowski and Vargas 304 (5667): 58