# *Community Discussion of Community Omics Needs*

Today's discussion was motivated by OCB concerns regarding the sunset of the CAMERA database

What CAMERA aimed for: database and associated computational infrastructure providing a single system for depositing, locating, analyzing, visualizing and sharing microbial ecology data (**including geospatial information**) through an advanced web-based analysis portal (Sun et al 2011 Nucleic Acids Research).

**Data management pipeline needs for community: What are the short and long term alternatives? New opportunities for management of ocean 'omics data?**

camera PORTAL

Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis

# New opportunity to develop a plan in the ocean sciences community for the long term role of ‘omic data.

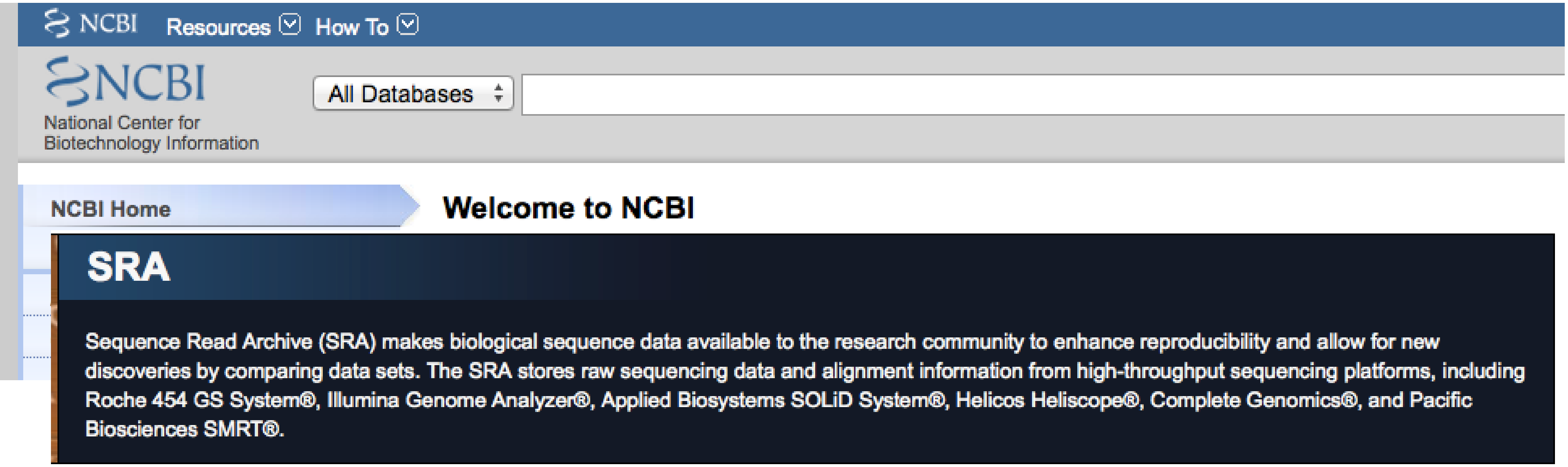
- By ‘omics we are using the broad definition of genes, RNA, proteins, metabolites, lipids, etc.

- ‘omic data record of ocean climate and biogeochemistry over space and time. Management important for understanding a changing ocean

- Accessible ‘omic data enables testable biogeochemical hypothesis:

Example application: identifying enrichment of of an enzyme for a specific process at a given location to motivate rate studies of that process at that location

# Current Database Landscape

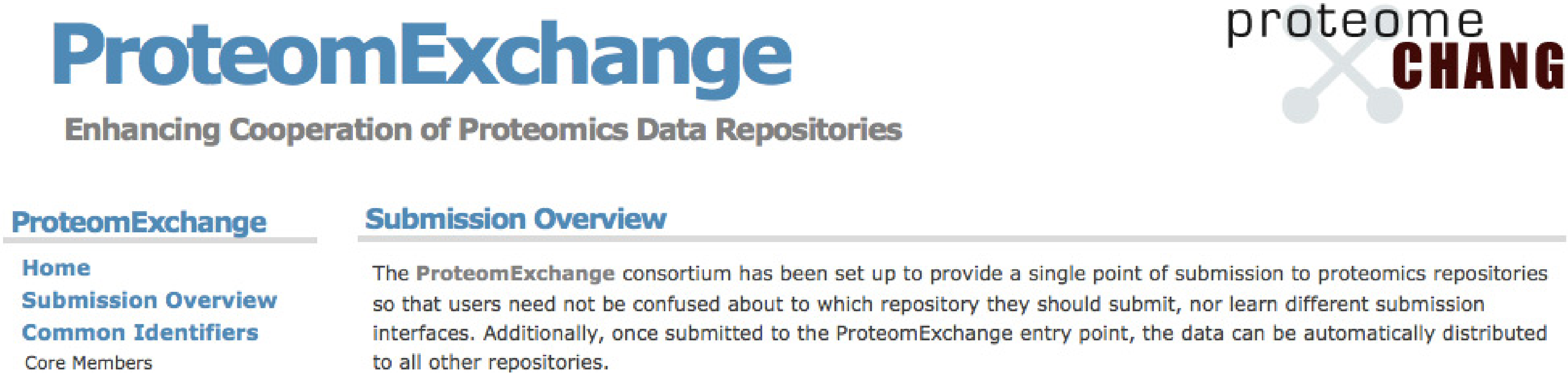


SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

The mission of the Integrated Microbial Genomes (IMG) system is to support the annotation, analysis and distribution of microbial genome and metagenome datasets sequenced at DOE's Joint Genome Institute (JGI). IMG is also open to scientists worldwide for the annotation, analysis, and distribution of their own genome and metagenome datasets, as long as they agree with the IMG data release policy and follow the metadata requirements for integrating data into IMG. Data distribution for IMG datasets is provided solely through individual genome/ metagenome data portals and is limited to **assembled** and **annotated** datasets submitted for a nnotation and integration through IMG's submission site

# Need to plan for longevity and learn from experience



# ProteomExchange

Enhancing Cooperation of Proteomics Data Repositories

## Submission Overview

The **ProteomExchange** consortium has been set up to provide a single point of submission to proteomics repositories so that users need not be confused about to which repository they should submit, nor learn different submission interfaces. Additionally, once submitted to the ProteomExchange entry point, the data can be automatically distributed to all other repositories.

proteome CHANGE

"At the time when the submitted data are declared publicly available by the submitter, all mass spectrometer output files will be deposited in the Tranche repository"

camera
PORTAL
Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis

# **OCB Community Survey Results:** Scientific interests as they pertain to generation and analysis of ‘omics data

**‘OMICS COMMUNITY**

- Metagenomics, metatranscriptomics, metaproteomics, metabolomics, phylogenetics
- Generate ‘omics data and do comparative analysis with data from others

**MODELING**

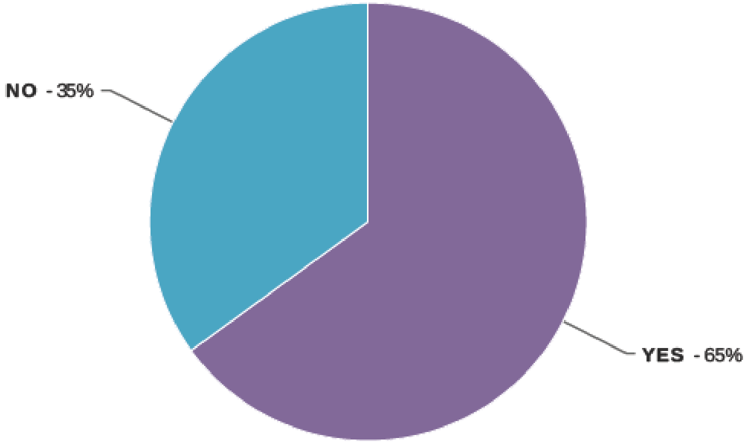
- Use in models of marine biogeochemical processes

**BIOGECHEMISTRY**

- Targeting specific organisms and biogeochemical cycles in different environments
- Generate activity data and compare rate data with CAMERA data to search for biogeochemically relevant enzymes

# How CAMERA is being used by the OCB community?

*Did you use CAMERA?*



*What did you use CAMERA for? (Listed in order of importance)*

1) Accessing unique datasets and BLAST queries
2) Access to reference datasets
3) Searching by user description of project
4) Geospatial searching
5) Data submission and annotation services

# Comments on use of CAMERA: opposite perspectives
# Key for data processing vs. used strictly as data repository

"It was/is integral to our data processing, and it made all sorts of sense to develop a community platform for such data intensive analyses, and appropriate archival. It was a great step forward, and the prospect of it shutting down seems like two steps backward. What a shame."

"I have only really used CAMERA as a data repository. I thought it was an excellent service until version 2 came out with all the bells and whistles for analysis. At this point, getting hold of other people's datasets became much more difficult.... I don't want a system that does analysis for me - I'm a bioinformatician and I hate black boxes that I can't fully describe."

# How shutdown has impacted research ?  Part 1

**Members of community that heavily rely on CAMERA resources:**

-still trying to figure out viable long term solution for the loss of the repository and querying capabilities.

-no immediate means to search environmental datasets. For now, that part of my research is on hold.

-We have spent countless hours locating resources that can be used for different steps that used to be performed on the CAMERA platform.

-no alternatives that I can use at this point: this aspect of my work has been abruptly truncated…there are no resources I know
of that enable me to easily survey a wide range of sites on a broad scale.

# How shutdown has impacted research ?  Part 2

## Members of community that use it as data repository- bioinformatic solutions elsewhere

“Well, to be honest, CAMERA has never functioned as a good place to deposit data…the group as a whole has been highly inattentive overall.”

“I used camera primarily for data archiving, the data analysis queues and wait times had slowed a while ago to the point where it was more efficient for me to use computational resources elsewhere.”

“I only used CAMERA to download data for analysis locally, so the shutdown should have minimal impact (as these data will still be available for download).”

# Immediate needs of community and minimal requirements for community ‘omics data management system

✧ 89% Search capability for a single gene or protein in environmental ‘omic data (e.g., BLAST search)

✧ 82%Text-based searching of metadata (e.g., geographical/geospatial) or annotation (e.g. gene function, taxonomic classification)

✧ 74% Storage of raw data

✧ 74% Storage of processed data

✧ 72% Geospatial searches

✧ 72% Access to datasets not available in other repositories

# Top priorities for a longer term vision for ‘omics data management

✧ 94% Ability to add user curation

✧ 94% Geospatial searches based on metadata

✧ 61% Storage of data used for biological inference (gene and protein expression)

✧ 61% comparative pathway analysis (metabolism, protein families)

✧ 50% Access to environmental reference datasets (e.g. combined from ncbi data and user supplied data

## *Challenge for the community:*
## *How to build a sustainable omics informatics capability?*

- ✧ High priority for national funding
- ✧ Collaboration with other big data efforts
- ✧ Data storage and metadata querying

“This isn't a solution, but I think a repository and an 'omics analysis pipeline should be a national funding priority given the explosion of sequence-based data over the past 5 years - we've seen this problem coming. It's my impression that there’s been a lot of wasted effort among many research groups trying to achieve similar goals that could be achieved by a national effort. The question of sustainability usually has dollars attached - I think it would be a mistake to put repositories and analytical capabilities behind additional pay-walls. If funded, sequence-based projects should have access - much like ship access is a given for oceanographic programs”

“There are a few initiatives that are currently exploring "BigData" in other fields and we should be trying to access those resources to generate tools that are beneficial to our science, but broadly applicable to other biological systems (to save resources).”

“What is needed more than anything is raw data with rich metadata
for decent environmental comparisons to be made”

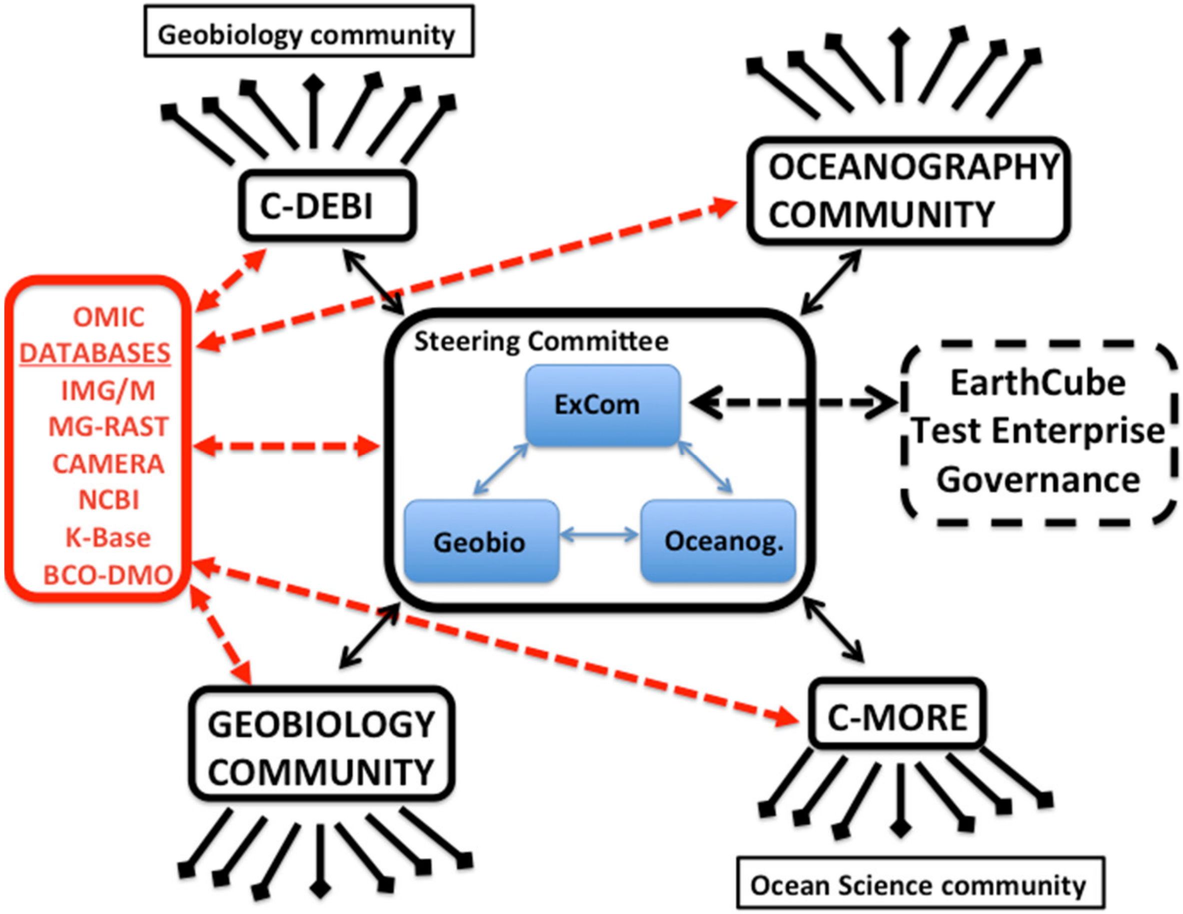
## Goals for good management of environmental sequence data

-maintain accessibility of ‘omics data to widest scientific community possible including our diverse oceanography community

e.g. Good data management will allow for better integration into models, new hypotheses for biogeochemical sampling, platform for new informatic tools to be developed by computational colleagues.

# ECOGEO RCN (Ed DeLong, PI)

# EarthCube Oceanography & Geobiology Environmental Omics Research Coordination Network



Geobiology community

C-DEBI

OCEANOGRAPHY COMMUNITY

OMIC DATABASES
IMG/M
MG-RAST
CAMERA
NCBI
K-Base
BCO-DMO

Steering Committee

ExCom

Geobio

Oceanog.

EarthCube Test Enterprise Governance

GEOBIOLOGY COMMUNITY

C-MORE

Ocean Science community

**Newly funded. Anticipated Start date: September 1, 2014**

## Slide 14

**1** edward delong, 7/22/2014

# WHAT IS NSF EARTH CUBE ?

EarthCube

## Community-Driven Data & Knowledge Environment for Geosciences

“ EarthCube is a collaboration between the U.S. National Science Foundation and Earth, atmosphere, ocean, computer, information, and social scientists, educators, data managers, and more. EarthCube aims to transform the conduct of research through the development of community-guided cyberinfrastructure to integrate information and data across the geosciences”

# ECOGEO RCN was inspired by August 2013 EarthCube End user workshop @ Catalina Is.

*“The overall goal of this EarthCube workshop was to bring together a group of leaders in ocean omic science & computer science, to help identify and prioritize a set of unifying scientific drivers and cyberinfrastructure requirements necessary to enable the storage, curation, federation, & comparative analyses of large and small scale ocean omic datasets, that are emerging from recent scientific efforts.”*

**Meeting Executive Summary:**
http://workspace.earthcube.org/sites/default/files/files/document-repository/Ocean%20'Omics%20EarthCube%20End-User%20Workshop%20-%20Executive%20Summary.pdf

**Meeting report :**
http://www.standardsingenomics.org/index.php/sigen/article/view/dsigs.5749944/1133

# ECOGEO RCN Goals

1) Build and strengthen partnerships and collaboration between geobiologists, microbial oceanographers and cyber/computer scientists interested in managing sharing and analyzing large scale omic datasets;

2) Define the requirements for necessary components for an ocean and geoscience interoperable omics cyberinfrastructure framework (e.g. dataset curation methods, search engines, high performance compute facilities, workflows, user analytical facilities, developers, etc.). (Specifically we will identify those systems that are operational and in use now, determine what is working, and what is not, and define further improvements and developments for a federated, networked next generation cyberinfrastructure to serve the ocean and geoscience omics data-user community).

3) Organize and collaborate with other organizations like the Genome Standards Consortium (http://gensc.org), and the EarthCube Test Enterprise Governance organization, to facilitate omic data and metadata exchange and integration between ocean and geo scientists and the world.

# Next steps and activities for ECOGEO RCN

- Organize and conduct annual meetings to define community needs and mechanisms for achieving them

- Hold technical and educational “virtual” discussions and workshops to further refine community needs & ways to achieve them

- Link people/community to data, analytics and EarthCube opportunities

- *Set the stage and assemble the expertise to engage EarthCube resources,  to build community cyberinfrastructures for ocean omics*

**WE NEED & ENCOURAGE OPEN PARTICIPATION OF BROAD OCEANOGRAPHIC AND GEOBIOLOGY COMMUNITIES !**

# QUESTIONS ?

## Please contact Ed DeLong &/or Steering Committee members

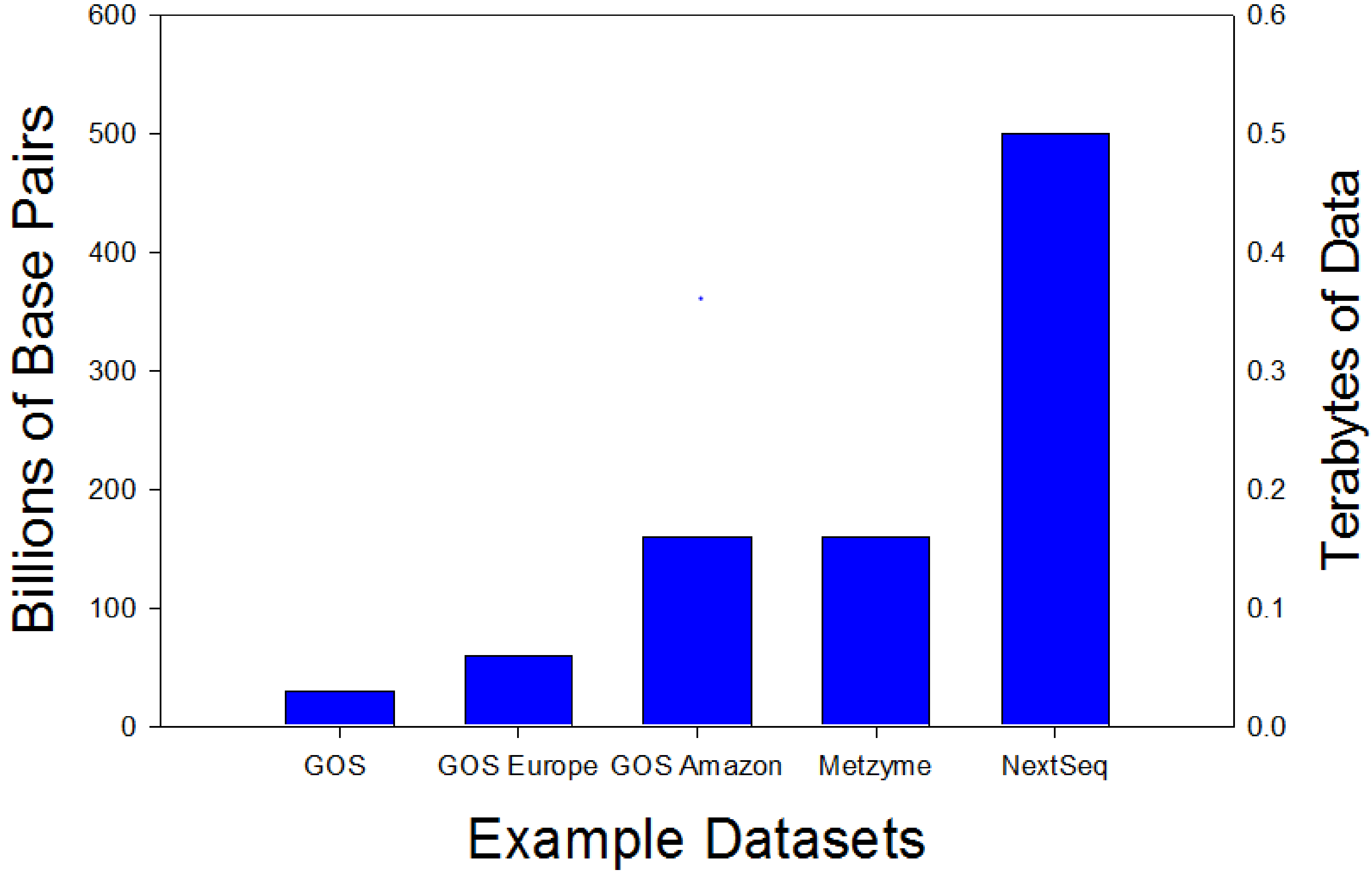
| | | |
|---|---|---|
| Ginger | Armbrust | University of Washington |
| Eric | Allen | Scripps Institution of Oceanography/CAMERA |
| Dylan | Chivian | Lawrence Berkeley National Laboratory/K-Base |
| Ed | DeLong | MIT, University of Hawaii/C-MORE |
| Greg | Dick | University of Michigan |
| Jack | Gilbert | Argonne National Laboratory/Genome Stand. Consort. |
| John | Heidelberg | University of Southern California/C-DEBI |
| Bethany | Jenkins | University of Rhode Island |
| Nikos | Kyrpides | Lawrence Berkeley National Laboratory/IMG |
| Mak | Saito | Woods Hole Oceanographic Institution |
| Erik | Zinser | University of Tennessee |
| Folker | Meyer | Argonne National Laboratory/MG-RAST |

# *Ocean Omics Discussion Topics*

1. What are the short-term and long-term ocean ‘omics needs?

2. What funding and scoping models are sustainable?
   - Data Repository
   - Data Analysis:  software/algorithms/processing
     - Open access/Open source and commercial packages
     - Computing Infrastructure
   - Possibilities and limitations of leveraging biomedical resources

3. Open Discussion


Billions of Base Pairs
600
500
400
300
200
100
0
Terabytes of Data
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GOS
GOS Europe
GOS Amazon
Metzyme
NextSeq
Example Datasets

## What about metadata requirements?

78% thought metadata standards recommended by the genome standards consortium were insufficient for minimal metadata associated with ‘omics data

Repository needs to include

- Oceanographic data (biogeochemical, physical)
- Sample processing method

## Needs for and barriers to good metadata

- Community buy-in
- Rapid response to maintain database as technology moves
- Good cross referencing of linked samples and datasets
- Funding to manage metadata not rewarded