EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS
Catalina Island, USC Marine Station August 20-23, 2013
Final Draft: Ed DeLong and Workshop Participants, August 29, 2014

Earth Cube Workshop Title: Ocean 'omics science and technology cyberinfrastructure : current challenges and future requirements.

Introduction: The overall goal of this EarthCube workshop was to bring together a group of leaders in ocean 'omic science and computer science, to help identify and prioritize a set of unifying scientific drivers and cyberinfrastructure requirements necessary to enable the storage, curation, federation, and comparative analyses of large and small scale ocean 'omic datasets, that are emerging from many recent scientific efforts. Applications of these data have great potential for improving our understanding ecosystem processes and predicting their future trajectories, but the necessary computational tools for doing so are still lacking.

A large group of ocean scientists and oceanographers are now employing 'omics approaches to characterize and quantify the nature, distribution and function of organisms in ocean ecosystems. "Omics" is defined here as the collective molecular or biochemical characterization of pools of biological molecules, such as genes and genomes, transcripts and transcriptomes, proteins and proteomes, and small molecules, metabolites and metabolomes, that together encode the structure, function, dynamics and activities of an organism or organisms. The tools and datasets that encompass 'omics science are diverse, complex, and rapidly expanding, and require the construction, curation, and query of diverse federated databases, as well as development of shared interoperable, "big-data capable" analytical tools.

To achieve the workshop goals, participants (46 in total, predominantly U.S. citizens) represented a variety of relevant disciplines including microbial oceanography and genomics (35%), phytoplankton genomics and ecology (15%), deep-sea microbiology (24 %), cyberinfrastructure and genomic scientists (15%) and Foundation representatives (11%). The condensed outcomes of Ocean Omics EarthCube workshop discussions are summarized below. A more detailed report will be provided after the vetting of comments and breakout group summaries with the workshop attendees.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.

    - How do physical and chemical oceanographic parameters and biological population structure and function co-vary within and between different oceanographic provinces? Do steep physical and chemical gradients result in

steep microbial functional gradients and drive changes in microbial biodiversity? Do feedbacks exist in both directions?

- How does 'omic and population plasticity in microbes bolster ecosystem resilience to disturbances? How does global change and environmental disturbance impact genomic repertoires, transcriptional organization, protein and metabolome content, and biogeochemical activity?
- What are the underlying molecular and biochemical mechanisms that regulate the physiological responses of microbes to environmental change, and their downstream biogeochemical consequences and feedbacks?
- How do microbial communities in the ocean fluctuate as a function of distance from land, seafloor spreading centers, gyres, and upwelling zones? How do they change as a function of geochemistry, currents, and crustal age? How does this affect the flux of matter and energy in the surface and deep sea?
- By what microbially-mediated mechanisms does rapid polar climate change affect the budget of greenhouse gases in the context of permafrost thawing and dissolved organic carbon release and transport, in time and space?
- How can 'omics data be more effectively leveraged into predictive frameworks for understanding ecosystem processes and their future trajectories ? How can 'omics data be distilled into tools useful to managers and stakeholders for efficiently monitoring ecosystem change and detecting ecosystem impairment?

2. **Current challenges to high-impact, interdisciplinary science:** Several themes emerged as consistent challenges faced within/across the involved discipline(s).

- It is still a challenge for the community to develop, validate and implement standardized and federated procedures for sample collection schemes, sample QC/QA, data formats, annotation workflows, and data analyses, and to integrate those with geochemical, biological, and physical oceanographic data over multiple nested spatiotemporal scales.
- The community currently has limited access, storage space, and transfer mechanisms for sharing and archiving of raw data, processed data, data products from workflows, and records of the provenance of data analyses.
- The community generally has limited access to large scale, high performance compute capabilities necessary for the annotation, comparison, statistical analyses and other workflows required for analyses of large scale ocean 'omic datasets.
- There are new non-sequence-based datatypes (e.g. mass spectrometry used in metabolomics) emerging that will need to be stored, accessed and analyzed and federated with other environmental and 'omic datastreams.

- The community lacks sufficient tools for simultaneous visualization and intercomparison of heterogeneous datatypes (e.g., environmental, 'omic and oceanographic datasets).
- It is currently difficult to integrate emerging 'omics datasets and analyses with existing and developing physical and biogeochemical models. This is partly an analytical problem (e.g., the mapping of genes and pathways onto their respective biogeochemical activities), and partly an integration problem, requiring the combination of quantitative 'omics-derived biogeochemical information, with quantitative geophysical and geochemical models.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

1. **Desired tools, databases, etc. needed for pursuing key science questions with brief elaboration:**

- Omic database development is required for curation, maintenance and data standardization that will allow for easy data submission, extraction and query. As well, tools for rapid and simple data query and metadata association are necessary. This includes federation with non-sequence-based datasets (e.g. metabolomics and lipidomics) into existing/emerging oceanographic 'omics database/analysis/visualization platforms. Environmental 'omic databases need to be: (1) federated (i.e., all datasets are interoperably queriable and transparently accessible), (2) curated (validated and updated, as for example NCBI nr datasets), (3) sustained (i.e. a five-year commitment of support is not sufficient), and importantly, (4) intuitively accessible to a broad range of scientists, and the public.

- The ocean 'omics community would benefit from “Google-like” or “Kayak-like” search and suggestion functions/engines, that could query across complex and heterogeneous, federated environmental, oceanographic and 'omic databases.

- Tools and mechanisms are required for access to high performance computing and statistical analyses of large scale 'omic datasets, that could accommodate both naïve users as well as experienced “power users”. One possibility is a user facility that functions similarly to UNOLS oceanographic facilities, that would provide access to software developers, bioinformaticians, and analytical tools, as well as the hardware required (storage facilities, servers, clouds, etc) required for 'omic analyses. Researchers could request access to this facility in association with successful grant applications, as with UNOLS. Extending the capabilities of BCO-DMO or similar services also seems another tractable model.
- Tools are required for more intuitive, accessible and integrated visualization of linked environmental, 'omic and oceanographic (and other interdisciplinary) data

sets. Statistical tools and techniques for dataset inter-comparison and spatiotemporal modeling also are critical and need further development.

- The community would benefit from access to a web clearing house/portal with links to standard “ocean 'omics” best practices, algorithms, software and workflows, as well as analytical and statistical methods under development, with entry points for both naïve and power users, would be a useful resource for the community.

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward and how it can use EarthCube to achieve those goals:**

- Cross train and educate computer scientists and engineers, and ocean and earth scientists to improve communication and collaboration among disciplines. This includes training and education to develop cross-disciplinary expertise within and between bioinformatics, the Earth sciences, and the Ocean sciences.
- Facilitate access, availability and utilization of NSF supercomputers for the Earth and Ocean sciences communities. (Using government supercomputers should be as technically easy, and as feasible as accessing the Amazon EC2 grid).
- Plan and initiate a community Research Coordination Network to support cyberinfrastructure technology and infrastructure development and education in ocean 'omics.
- Promote the development of an EarthCube system that would combine the facilitative role of the BCO-DMO database (or similar), with novel and flexible analyses and visualization services for analyzing and exploring ocean omics oceanographic data (e.g., Ocean Data View-like software and tools, for ocean ‘omics data).
- Further identify ocean 'omics cyberinfrastructure “parts” (e.g. dataset curators, search engines, high performance compute facilities, workflows, user analytical facilities, developers, etc.) that are operational and in use now, and determine which ones might be further improved, developed, federated, and networked into a functional EarthCube community ocean 'omics cyberinfrastructure solution.