CIF21 DIBBs: PD: Accelerating Comparative Metagenomics through an Ocean Cloud Commons
PI: Bonnie Hurwitz, Co-PI: John Hartman, University of Arizona

Introduction: Hundreds of researchers worldwide have joined forces in the Tara Oceans Expedition to create an unprecedented planetary-scale dataset comprised of state-of-the-art next generation sequencing, microscopy, and physical/chemical metadata to explore ocean biodiversity. This summer the complete collection of data from the 2009-2013 Tara voyage was released. Yet, despite herculean efforts by the Tara Oceans Consortium to make raw data and computationally derived assemblies and gene catalogs available, most researchers are stymied by the sheer volume of the data. Specifically, the most tantalizing research questions lie in understanding the unifying principles that guide the distribution of organisms across the sea and affect climate and ecosystem function. To use the data in this capacity researchers must download, integrate, and analyze more than 7.2 trillion bases of metagenomic data and associated metadata from viruses, bacteria, archaea and small eukaryotes at their own data centers (~10 TB of raw data). Accessing large-scale data sets in this way impedes scientists from replicating and building on prior work.

The Ocean Cloud Commons: We are developing a cloud-based service called the Ocean Cloud Commons (OCC) for performing big-data analyses on the the recently released Tara Oceans Expedition data (2009-2013). The OCC service enables users to analyze the Tara Oceans data and make the results of the analysis available to subsequent OCC computations. The OCC is deployed as a large-scale prototype in OpenCloud [1], a cloud platform being deployed on Internet2 that provides an open service framework for deploying value-added services such as the OCC. As a proof-of-concept, the OCC provides Hadoop MapReduce [3,4] as a service that can be used by analyses, well as the Libra algorithm that we developed that computes all-vs-all sequence analyses. The OCC allows access remote data sets and makes its own data sets available to remote computations via the Syndicate[7] distributed storage platform.

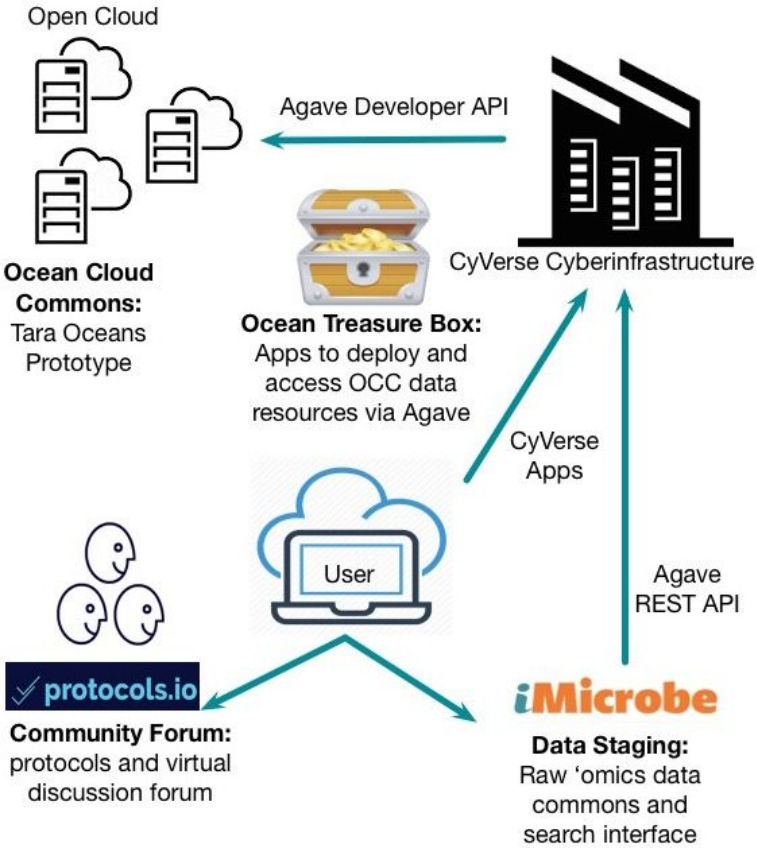

Figure 1. A cyberinfrastructure to deploy and analyze 'omics data in the Ocean Cloud Commons through: 1. Staging raw data in iMicrobe and transferring to CyVerse via the Agave REST API, 2. Developing Apps that convert raw data into persistent comparative metagenomic data clouds via the Agave Developer API (Ocean Cloud Commons and Ocean Treasure Box), 3. Developing Apps that analyze data in the Ocean Cloud Commons via the Agave Developer API (Ocean Treasure Box), 4. Develop a "cookbook" of protocols and virtual community via Protocols.io for using these resources and cyberinfrastructure.