

January 17, 2006 UC San Diego Partners with Venter Institute to Build Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

Gordon and Betty Moore Foundation Awards \$24.5 Million Grant to Advance Metagenomics Research

January 17, 2006

Doug Ramsey

Researchers at UCSD 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 new resource will help scientists understand how microbes function in their natural ecosystems, enable studies on the effect humans are having on the environment, as well as permit insight into the evolution of life on Earth. The UCSD Division of the California Institute for Telecommunications and Information Technology (Calit2) will lead the project in partnership with J. Craig Venter Institute (Venter Institute) in Rockville, MD, and UCSD's Center for Earth Observations and Applications (CEOA) at Scripps Institution of Oceanography.

"This prototype cyberinfrastructure will be used by scientists studying marine life and ecosystems to examine in an unprecedented manner the genomic complexities of natural communities of micro-organisms as they have evolved in their local environments," said UCSD Chancellor Marye Anne Fox. "This project will change the way large-scale science can be conducted and we are proud to develop this world-class and pioneering facility on our UCSD campus."

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 explosion of data from the collection and sequencing of marine microbes requires a completely novel approach to storing, accessing, mining, analyzing, and drawing conclusions from this rich new wealth of information," said co-investigator J. Craig Venter, Ph.D., president of Venter Institute. "The goal is to create a community resource to house all metagenomic data that will facilitate and advance knowledge of marine microbial ecology, other natural environments, and evolutionary biology."

The CAMERA project builds on pioneering efforts in metagenomics to sequence the genomes of entire microbial communities, often comprising thousands of species. The largest such effort is Venter Institute's Sorcerer II Expedition, for which sequencing is funded by the Moore Foundation. The Expedition is developing 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. The metagenomics database will include new sequences, genes and gene families, together with their annotations and associated environmental metadata.

The move from traditional organism genome databases to the CAMERA-based environmental metagenomics data storage and computational complex requires development of a more complex cyber-architecture. Using

dedicated optical circuits, CAMERA will permit scientists to connect their local laboratory PC clusters directly to the CAMERA database and tools using the National LambdaRail or international optical circuits, resulting in up to a hundred-fold increase in bandwidth over current standards.

The enhanced connectivity is based on a model pioneered by the OptIPuter project and funded by the National Science Foundation. "Linking Venter Institute to Calit2 will be the first persistent application of the OptIPuter high-performance 'collaboratory'," said Calit2 director Larry Smarr, Ph.D., principal investigator on both the OptIPuter and CAMERA projects. "The architecture is quite general and will be quickly adaptable to other areas of data-intensive science." Collaboratories are virtual laboratories where scientists can collaborate on research from dispersed locations-interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information from remote digital libraries.

Calit2 will also partner with UCSD's San Diego Supercomputer Center (SDSC) to create a next-generation science data server complex, which couples the Calit2 and SDSC middleware, compute, and storage capabilities with the NSF's TeraGrid distributed, high-performance computing facility in a unified Service Oriented Architecture. SDSC's Philip Papadopoulos noted that, "the CAMERA complex will have a thousand processors of dedicated local cluster computing and several hundred terabytes of replicated data storage, backed up by the SDSC and TeraGrid high performance compute and storage complexes." This will enable "scalable computing" resources to be applied to a wide range of computational tools to tackle the computationally intense questions derived from the larger metagenomic data collection.

Calit2 and Venter Institute will also support a series of training sessions and specialized seminars on this emerging discipline, as well as provide space for environmental metagenomics visitors to collaborate with CAMERA specialists. Over the next few years, CAMERA is expected to include other environmental or medical metagenomic datasets, as the novel cyberinfrastructure enables research in other disciplines.

The Moore Foundation grant, in part, contributes to the \$1 billion fundraising goal of The Campaign for UCSD: *Imagine What's Next*.

Media Contacts:

Doug Ramsey, Calit2, (858) 822-5825 Melanie Wranaker, Venter Institute, (301) 943-8879 Mario Aguilera or Cindy Clark, Scripps/CEOA, (858) 534-3624 Ashley Wood, SDSC, (858) 534-8363 Alex Barnum, Moore Foundation, (415) 561-7414 Debra Kain, UCSD Health Sciences, (619) 543-6163

Background

The goal of the CAMERA project is to create important advances in the knowledge of evolutionary biology and microbial ecology in marine and other natural environments.

"Metagenomics has the potential to revolutionize our understanding of microbial ecology in a large number of environments," said David Kingsbury, Ph.D., chief program officer for science at the Moore Foundation. "The major factor limiting its further progress has been the management of the very large quantities of data. We are delighted to be able to support the development of this community resource at one of the world's premier sites for high-speed networking and high-performance scientific computing."

In addition to Sorcerer II's ecological genomic data, the CAMERA database will be augmented by the soon-to-be-completed genomes of more than 150 critical marine microbes, also funded by the Moore Foundation, for comparative genomics studies. Venter Institute's Marv Frazier, Ph.D., co-principal investigator with Larry Smarr, said "We are looking forward to providing a metagenomics server complex for the data produced by our colleagues at the Department of Energy's Joint Genome Institute."

Scripps researchers will contribute expertise in modeling, analysis and information management across Earth-science observing systems. "We also have a set of world-class researchers in microbial ecology and annotating marine genomic data," said Scripps Deputy Director John Orcutt, who directs the CEOA and is a co-investigator on the new project along with Terry Gaasterland, director of the Scripps Genome Center launched last October. Experts at Venter Institute and the Scripps Genome Center will create annotations for much of the CAMERA genomic data.

The project brings together new technologies of high-throughput DNA sequencing and metagenomic analysis tools on the one hand, and cyberinfrastructure innovations on the other. Together, they will provide new tools to help marine microbial ecologists access and derive inferences from the massive data sets. The tools will allow ecologists, for example, to analyze families of proteins and conduct comparative analyses across multiple genomes.

"Each individual sequence is no longer just a piece of a genome. It is part of an entire biological community," said Peter Arzberger, Ph.D., director of the NIH-funded National Biomedical Computation Resources (NBCR), and lead author of the CAMERA grant proposal. "CAMERA will build on the NBCR software tools and user portal to explore the metagenomics data."

NBCR has links to the UCSD School of Medicine, and co-investigator John Wooley is affiliated with the Skaggs School of Pharmacy and Pharmaceutical Sciences. "Along with providing a novel approach to advancing fundamental biological knowledge, analysis of the marine genome data will allow us insight into natural marine products and how they can be applied for pharmacy and medicine," said Wooley, the university's Associate Vice Chancellor for Research. "The new resource will greatly enhance our health science researchers' ability to advance the development of new drugs and therapies from the ocean's resources to combat cancer and neurodegenerative and other diseases."

Venter Institute will make available a large collection of community-developed genome analysis software tools. The CAMERA tools will address the needs of two groups of users. The first group comprises potential users with little programming ability using web-based tools to explore data and visualization tools to interpret the results. The second group comprises bioinformatics experts with their own tools and programming. CAMERA will encourage this later group to contribute their software analysis tools, thus engaging the broader community in strengthening this international resource.

Other co-investigators on the CAMERA project include Venter Institute's Saul Kravitz, Aaron Halpern and Jonathan Eisen, as well as UCSD-based scientists Tom DeFanti and Ingolf Krueger.





