SALK INSTITUTE
FOR BIOLOGICAL STUDIES

March 13, 2007

NOT FOR PUBLIC RELEASE BEFORE: MARCH 13, 2007, 12 AM PST

Darwin's famous finches and Venter's marine microbes

LA JOLLA, CA – Although the Galápagos finches were to play a pivotal role in the inception of Darwin's theory of evolution through natural selection, he had no inkling of their significance when he collected them during his voyage on the *HMS Beagle*.

Similarly, it is hard to predict the impact the vast amount of marine microbial DNA – collected during the *Sorcerer II* Global Ocean Sampling Expedition by J. Craig Venter, Ph.D., and his team – will have on our understanding of the natural world.

"If anything, this is just the beginning," says Gerard Manning, Ph.D., director of the Razavi Newman Center for Bioinformatics at the Salk Institute for Biological Studies. "We're starting to explore this trove of sequences now, but it may be decades before we fully understand it all."

Just like the famous ornithologist John Gould who had to classify the Galápagos finches before they led Darwin on the right track, Manning and many others have been busy during the last couple of months wading through roughly 7.7 million sequenced snippets of sea-borne genomic DNA to impose order on the flood of data and to classify the identified proteins.

Their findings are detailed in series of papers, published in this week's online edition of the journal *Public Library of Science Biology (www.plos.org)*.

The authors are plying the rapidly emerging trade of metagenomics (also known as environmental genomics) that seeks to examine genomic snapshots taken directly from the environment.

"Metagenomics allows us to sample the 99 percent of all bacteria that won't grow in the lab," explains Manning. "GOS opens a huge window into biological and genomic diversity and, within this diversity, to better understand many of the fundamentals of biology." he adds.

Expanding the universe of protein families

But instead of whole genomes, metagenomics produces a whole grab bag of bits and pieces for which scientists have to develop new methods to extract meaning. In one of the papers, an array of scientists, spearheaded by first author Shibu Yooseph, Ph.D., and his colleagues at the Craig

Venter Institute, compared every DNA fragment with every other available DNA fragment to produce clusters of related sequences. This exhaustive analysis predicted more than 6 million proteins in the GOS data – nearly twice the number of all proteins ever described before – and laid the groundwork for further studies.

Manning, a co-author on Yooseph's paper, looked at the other side of the coin. He ran all the public sequences and GOS data against Pfam, a collection of signature profiles for all known protein families. Each of these profiles is an average of all known members of a certain protein family.

"Instead of starting with a human kinase to find a bacterial kinase, for example, you start with all of them together, which makes the search much more sensitive, but also very computationally expensive," Manning says. "We did almost 350 million comparisons, which is probably an order of magnitude or two more than anybody has ever done before."

Manning and co-author Yufeng Zhai, Ph.D., a bioinformatics programmer in the Razavi Newman Center for Bioinformatics at the Salk, could only accomplish this rather gargantuan task with the help of Time Logic, a company in Carlsbad, California. The company specializes in hardware that accelerates genomic searches. "We only have one of their accelerators, but Time Logic stepped up and lent us eight more," says Manning. The final computation took two weeks, but would have taken well over a century on a traditional computer.

The Salk scientists could assign over half of all GOS sequences to known protein families, and discovered that certain protein profiles are more popular in the ocean or on land. For example, gram-positive bacteria are best known for their hardy spores, but this ability has been entirely lost in their marine relatives.. Flagella, whip-like extensions propelling bacteria forward and pili, short extensions used to exchange genetic material between bacteria (also known as microbial sex), are also less frequent in marine environments.

"By comparing our findings with the Yooseph clusters, we also discovered hundreds of new gene families that hadn't even been seen before," says Zhai and adds that by adding the diverse GOS data to known profiles, "we were able to make them more sensitive and diverse, and so increase their power to categorize novel sequences."

Diversity of microbial kinases

In a separate study, Manning, Zhai, and first author Natarajan Kannan, Ph.D., a postdoctoral researcher in the lab of HHMI investigator and UCSD professor Susan S. Taylor, Ph.D., traded the breadth of the ocean survey for the depth of a single protein domain. They zoomed in on kinases, extremely well studied enzymes, which control every aspect of eukaryotic cell biology and are important cancer drug targets. They control the activity of proteins and small molecules by attaching tiny phosphate groups to them. By contrast, much less has been known about their bacterial counterparts.

Again and again, the researchers combed the GOS data for bacterial kinases, each time rebuilding their domain profiles by including the new members found in the previous round. All

in all, they dug up 45,000 protein kinase sequences that fell into 20 distinct families, of which the eukaryotic protein kinases are just one. The additional 19 families spanned a huge range and included several that had never been described before.

“Prokaryotic protein-like kinases were considered to be some sort of niche players, but actually they are more prevalent and widespread than histidine kinases,” explains Manning. Bacteria were thought to rely mostly on histidine kinases, which are structurally different from protein kinases, for all their signaling needs.

Even though the different kinase families had very little similarity in their sequence, it emerged that 10 key residues were conserved in almost all kinase families, fingering them as being at the core of what it means to be a kinase. Seven of those had been previously known to be important in human kinases, but the other three were unexpected finds.

The other surprising finding was just how innovative and plastic the different families were, even with these core residues, as one or another family had found ways to eliminate any but one of the 10 key residues. Using structural modeling, and patterns of sequence conservation, Kannan was able to show that loss of one key residue could be compensated by changes around other conserved regions of the protein, and that some of these changes in bacterial kinases are also seen in specific human kinases.

Says Manning, “By looking at all these very distant microbial relatives we can understand more even about human kinases and their relationship to cancer and other diseases. We go out into the ocean, we find all this diversity and analyzing what’s new and what’s not new reflects back on the things we thought we knew well.”

Research done at the Salk Institute was supported by the Razavi-Newman Foundation.

Sorcerer II Global Ocean Sampling Expedition

The circumnavigating *Sorcerer II* Expedition, named after the sailboat J. Craig Venter transformed into a marine research vessel, was inspired in part by the journeys of the HMS Beagle and the HMS Challenger in the nineteenth century. But unlike those pioneering expeditions, the *Sorcerer II* team led by J. Craig Venter and a globe-spanning network of collaborators are after tiny microbes, classifying the species they encounter not by their appearance but by their unique genetic code.

The current studies analyze samples collected from surface waters during the first phase (or first third) of the voyage, which led the *Sorcerer II* from Newfoundland through the Panama Canal and Galapagos Island on to French Polynesia. Venter’s crew siphoned seawater through a series of increasingly fine filters to collect the microbes, which they sent back to the J. Craig Venter Institute in Maryland.

In the lab, the scientists shredded the collected genetic material in millions of random snippets and then determined their sequence. Based on overlapping sequences, computer programs can then assemble longer stretches and merge them into longer pieces of a genome. These so-called

"scaffolds" are a treasure trove for a diverse group of scientists, who try to squeeze as much information as possible from the largest metagenomic dataset ever collected. For more information about the expedition, please go to: the J. Craig Venter Institute (URL: www.sorcerer2expedition.org and www.venterinstitute.org)

About the Salk Institute:

The Salk Institute for Biological Studies in La Jolla, California, is an independent nonprofit organization dedicated to fundamental discoveries in the life sciences, the improvement of human health, and the training of future generations of researchers. Jonas Salk, M.D., whose polio vaccine all but eradicated the crippling disease poliomyelitis in 1955, opened the Institute in 1965 with a gift of land from the City of San Diego and the financial support of the March of Dimes.

###

Media contacts:

Salk Institute:
Gina Kirchweger
kirchweger@salk.edu
858.453.4100 x1340

Gerard Manning
manning@salk.edu
858-453-4100 x 1757

UCSD
Sherry Seethaler
sseethaler@ucsd.edu
858-534-4656

10010 North Torrey Pines Road, La Jolla, CA 92037-1099