Floating laboratory. *Sorcerer II*, a private yacht outfitted to collect and freeze microbial samples, netted a huge bounty of DNA sequence.

METAGENOMICS

Ocean Study Yields a Tidal Wave of Microbial DNA

Data glut or unprecedented science? A global hunt for marine microbial diversity turns up a vast, underexplored world of genes, proteins, and "species"

After relishing the role of David to the Human Genome Project's Goliath, J. Craig Venter is now positioning himself as a Charles Darwin of the 21st century. Darwin's voyage aboard the H.M.S. *Beagle* 170 years ago to the Galápagos Islands netted a plethora of observations—the bedrock for his theory of evolution. Four years ago, Venter set sail for the same islands and returned 9 months later with his own cache of data—billions of bases of DNA sequence from the ocean's microbial communities. But whether that trip will prove anything more than a fishing expedition remains to be seen.

On 13 March, Venter, head of the J. Craig Venter Institute in Rockville, Maryland, and a bevy of co-authors rolled out 7.7 million snippets of sequence, dubbed the Global Ocean Sampling, in a trio of online papers in *PLoS Biology*. As a first stab at mining these data, which have just become publicly available to other scientists, Venter's team has found evidence of so many new microbial species that the researchers want to redraw the tree of microbial life. They have also translated the sequences into hypothetical proteins and made some educated guesses about their possible functions.

Some scientists are wowed by the effort. Others worry that researchers will not be able to make sense of all this information. The diversity of microbes uncovered is "overwhelming, … tantamount to trying to understand the plot of a full-length motion picture after looking at a single frame of the movie," says Mitch Sogin, a molecular evolutionary biologist at the Marine Biological Laboratory in Woods Hole, Massachusetts. And Venter doesn't necessarily disagree. In 2004, as the data were first rolling in, Venter confidently predicted that his salty DNA survey would "provide a different view of evolution." To make that happen, however, he now says, "we need even more data."

Microbial explorers. J. Craig Venter (*left*) and Anthony Knap of the Bermuda Biological Station for Research aboard Venter's yacht.

CREDITS (TOP TO BOTTOM): J. CRAIG VENTER INSTITUTE

The big trawl

This is the second time that the American millionaire genome sequencer has returned to port laden with DNA. Venter's 2004 study of microbes living in the Sargasso Sea was easily the largest DNA sequencing of environmental samples ever accomplished (*Science*, 2 April 2004, p. 66). This time around, he sailed from Halifax, Canada, through the Panama Canal and finished up 6500 kilometers southwest of the Galápagos. The funding for the $10 million project came from the Gordon and Betty Moore Foundation, the U.S. Department of Energy, and Venter's nonprofit foundation. The research vessel, the *Sorcerer II*, is Venter's private yacht tricked out as a floating laboratory.

The researchers sampled at 41 locations, isolating and subsequently freezing bacterium-sized cells. They also recorded the temperature, salinity, pH, oxygen concentration, and depth.

Back at Venter's institute, technicians extracted and sequenced the DNA. Using a whole-genome shotgun approach, they shattered all the DNA in a sample into fragments of specific sizes, sequenced each one, and then assembled these sequences together by matching the ends of the DNA with a powerful overlap-hunting computer program. In principle, this approach allows the reconstruction of entire genomes of the different organisms in a sample.

Three years and 6.3 billion bases of DNA sequence later, at least one thing is clear: The DNA in a typical community of marine microbes is so diverse that nothing close to a whole genome can be assembled, even with all the sequencing that Venter has mustered. Half of his 7.7 million DNA sequence fragments are so different that they could not be linked at all.

Nonetheless, the researchers could estimate the number of species in the samples based on slowly evolving marker genes. Judging by these glimpses of genomes, Venter's team identified more than 400 microbial species new to science, and more than 100 of those are sufficiently different to define new taxonomic families, they report. "This is a great milestone event" for environmental microbiology, says Dawn Field, a molecular evolutionary biologist at the Centre for Ecology and Hydrology in Oxford, U.K., who predicts that "these papers will become among the most highly cited of all time in biology."

Diversity deep end

The fact that Venter's brute-force sequencing approach fell short of capturing whole genomes shows that scientists are far from a full accounting of the species packed in a drop of seawater, says David Scanlan, a

marine microbiologist at the University of Warwick, U.K. And this "astounding" genetic diversity points to what Scanlan and others call the "paradox of the plankton."

Traditional ecological theory predicts that when multiple species compete for the same resources—in the case of ocean microbes, light and dissolved nutrients—then one, or a few, species should eventually outcompete the rest. If that were the case, then many of the sequences plucked from the waters by Venter's crew should map down onto a few dominant genomes.

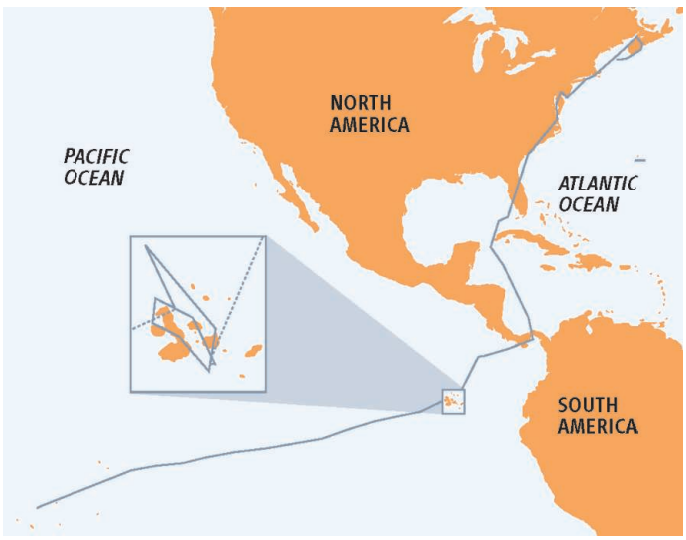
But rather than a sharp portrait of a few different microbes, the data create a pointillist painting of a countless mob. The vast majority of the microbes that found themselves snared in Venter's filters were genetically unique, says Scanlan: "It's a clear message that there's a tremendous gene pool in the ocean."

The diversity itself could be the solution to the paradox, according to Douglas Rusch, a computational biologist at the Venter Institute, and his colleagues. The staggering variety of genes may endow each species with sufficiently different metabolic tool kits to take advantage of slightly different combinations of resources, including the waste products of others, such that they can all coexist.

The newly detailed diversity also suggests that microbial taxonomy needs a major overhaul, says Ian Joint, a marine microbiologist at the Plymouth Marine Laboratory in the U.K. The current taxonomy carves up microbes into different "ribotypes" by comparing the sequence of the highly conserved genes of the protein-synthesizing ribosome. Because there is so much diversity within the DNA even after dividing them into ribotypes, Venter's team proposes to throw out ribotyping altogether. Instead, they are defining groups of microbes based on the environment in which they were collected and how well their DNA matches a reference set of fully sequenced marine microbial genomes. Doing so has allowed Venter's team to group sequence fragments into different "subtypes." Venter's team says that each of these subtypes represents a "distinct, closely related population" of microbes that fill a particular niche in their local environment. However, many more marine microbial genomes must be sequenced to make this scheme work, says Joint.

Marine data-mining

The samples brought to port by *Sorcerer II* do more than shake up microbial taxonomy. Based on their best guess as to the beginning and end of each gene teased out from the DNA sequences, Venter Institute computational biologist Shibu Yooseph and his colleagues have concluded that the DNA encodes 6.12 million hypothetical proteins. That finding almost doubles the number of known proteins in a single stroke. It also shows that the end of protein diversity is not in sight, says David O'Connor, a molecular biologist at the University of Southampton, U.K. Most of the predicted proteins are of unknown function, and a quarter of them have no similarity to any known proteins. Venter expects that some of these can be exploited to develop new synthetic materials, clean up pollution, or bioengineer fuel production.



Taking stock. *Sorcerer II* collected bacteria at dozens of sites in the Atlantic and Pacific, particularly around the Galápagos Islands (*inset*).

SOURCE: J. CRAIG VENTER INSTITUTE

But the hypothetical proteins are already offering a new view of basic microbial biology. A team led by Venter and Gerard Manning, a computational biologist at the Salk Institute for Biological Studies in San Diego, California, says that the current picture of the proteins responsible for coordinating marine microbes' gene expression and metabolism is off the mark. By comparing predicted amino acid sequences with those of known proteins, they found a surprising abundance of signaling proteins thought to be used only by multicellular organisms. Among the hypothetical proteins from their marine samples, the researchers found 28,000 of the so-called eukaryotic protein kinases, as well as another 19,000 of a group that are highly similar to these kinases—triple the number previously known.

These analyses of Venter's metagenomic data hint at the work that lies ahead for protein researchers. "Claims by some biologists that complete catalogs of the protein universe would be attainable within a decade now look naïve," O'Connor points out.

Thus to some, the data produced by Venter's voyage are an exciting starting point for protein, gene, and microbe discovery. It's something "people will be working on for quite some time," says Howard Ochman, a molecular evolutionary biologist at the University of Arizona in Tucson. But for others, the value of this tidal wave of data is uncertain. James Prosser, a molecular biologist at the University of Aberdeen, U.K., worries that adding all of this sequence to the existing gene and protein databases could "swamp" the system, cluttering the results of searches for well-characterized genes.

To help researchers deal with not just Venter's 100 gigabytes of sequence data but also other relevant information about a microbe's environment and location, Venter's team and Larry Smarr, a computer scientist at the California Institute for Telecommunications and Information Technology in San Diego, have built a metagenomics version of GenBank, the online genetic database curated by the National Center for Biotechnology Information in Bethesda, Maryland. In addition to doing the typical gene searches and genome comparisons, the new system, known as the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA), can hunt for correlations between DNA sequence and environment for clues about co-occurring microbes. So far, however, CAMERA has only a few active users.

A more serious drawback of Venter's study, says Prosser, is that the samplings do not appear to have been carried out with any specific scientific hypotheses or aims in mind. The cynical view is that these are little more than "fishing trips," he says. "There would be greater potential for scientific advances if more focused, better designed studies were carried out."

Will the voyage of the *Sorcerer II* live up to Venter's hopes? It took Darwin 25 years after returning from his expedition to publish his theory of evolution. With the three papers online this week, Venter, at least, has hopped on the fast track. But in terms of synthesizing the big picture of marine microbiology, he and his colleagues are still out to sea. **–JOHN BOHANNON**