JGI Progress Report 2002–2005

U.S. DEPARTMENT OF ENERGY JOINT GENOME INSTITUTE

Office of Science
U.S. DEPARTMENT OF ENERGY

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

JGI's Mission

To develop and exploit new sequencing and other high-throughput, genome-scale, and computational technologies as a means for discovering and characterizing the basic principles and relationships underlying the organization, function, and evolution of living systems.

What is Sequencing?

Just as computer software is rendered in long strings of 0s and 1s, the "software" of life is represented by a string of four chemicals, abbreviated as A, T, C, and G. To understand the software of either a computer or a living organism, we must know the order, or sequence, of these informative bits.

Berkeley Lab

Los Alamos National Laboratory

Lawrence Livermore National Laboratory

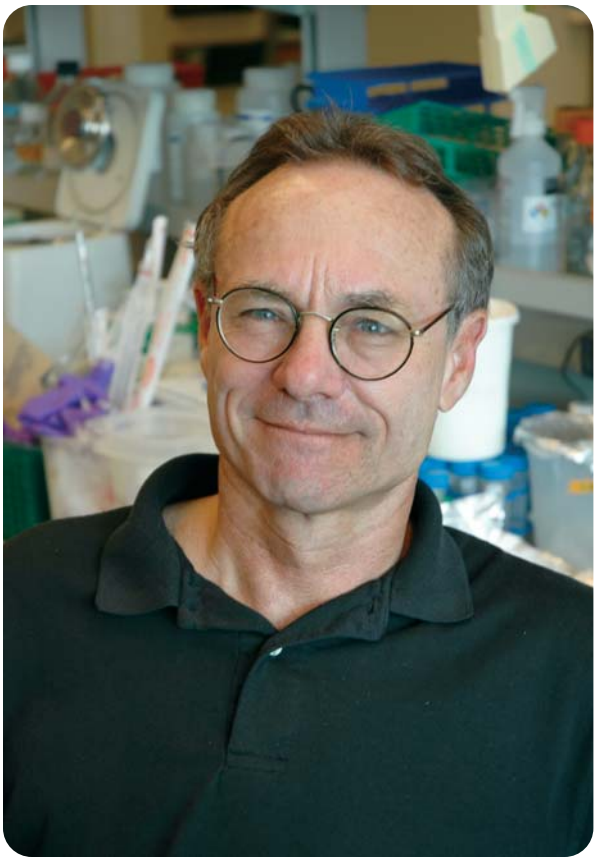
shgc

The University of California 1868

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Oak Ridge National Laboratory

table of contents

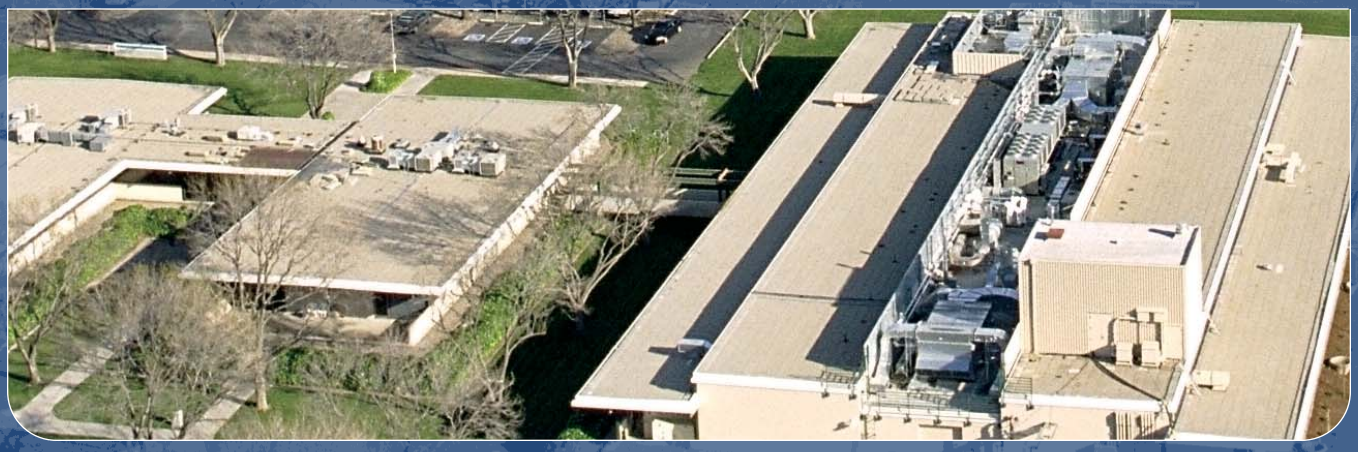


Director's Perspective

The JGI was formed virtually in 1997, and a facility established in Walnut Creek in the fall of 1999. The driving force behind the Institute's creation was to contribute to the Department of Energy's commitment to the sequencing of the human genome. With the publication of three manuscripts in the journal *Nature* during 2004, each describing a completed DOE chromosome (chromosomes 5, 16, and 19), the institution has successfully completed its human genome mission. A reasonable question accompanying the completion of the human genome is, "Should the accomplishment of the mission for which it was created remove the need for the JGI?" An answer to this question emerges from the fact that in the time between the creation of the JGI and the present, sequencing and its role in biology has extended to fields far beyond what could be imagined when the sequencing of the human genome was first begun. DNA sequence has become vital infrastructure for scientific disciplines ranging from climatology to geochemistry. One of the DOE's initial motivations for its participation in the Human Genome Project was based on its capabilities in multidisciplinary, large-scale science. These DOE National Laboratory capabilities that contributed to the success of the JGI's first phase are now joined by various drivers of the DOE's fundamental mission—carbon sequestration, energy production, and bioremediation—all of which are enabled and dependent upon the availability of sequence data. Not only has the target of the JGI's sequencing activities changed, moving from a single human genome to the genomes of countless microbes, plants, and other organisms, but the participants have also changed. Now serving as a non-traditional user facility, the JGI is empowering the science of numerous scientists, reaching across the US and beyond, carrying out sequence-based investigations to better understand the natural sciences.

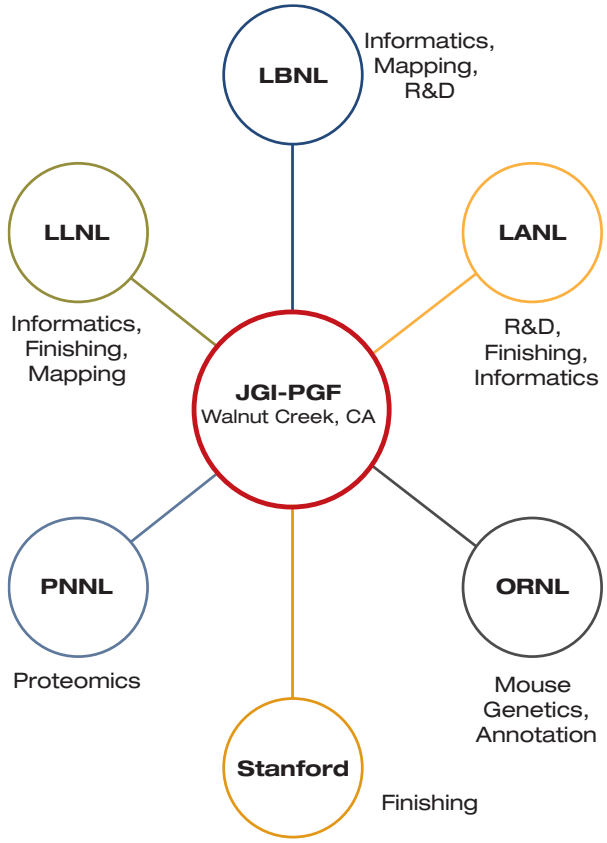
Edward M. Rubin, MD, PhD
Director, U.S. Department of Energy
Joint Genome Institute

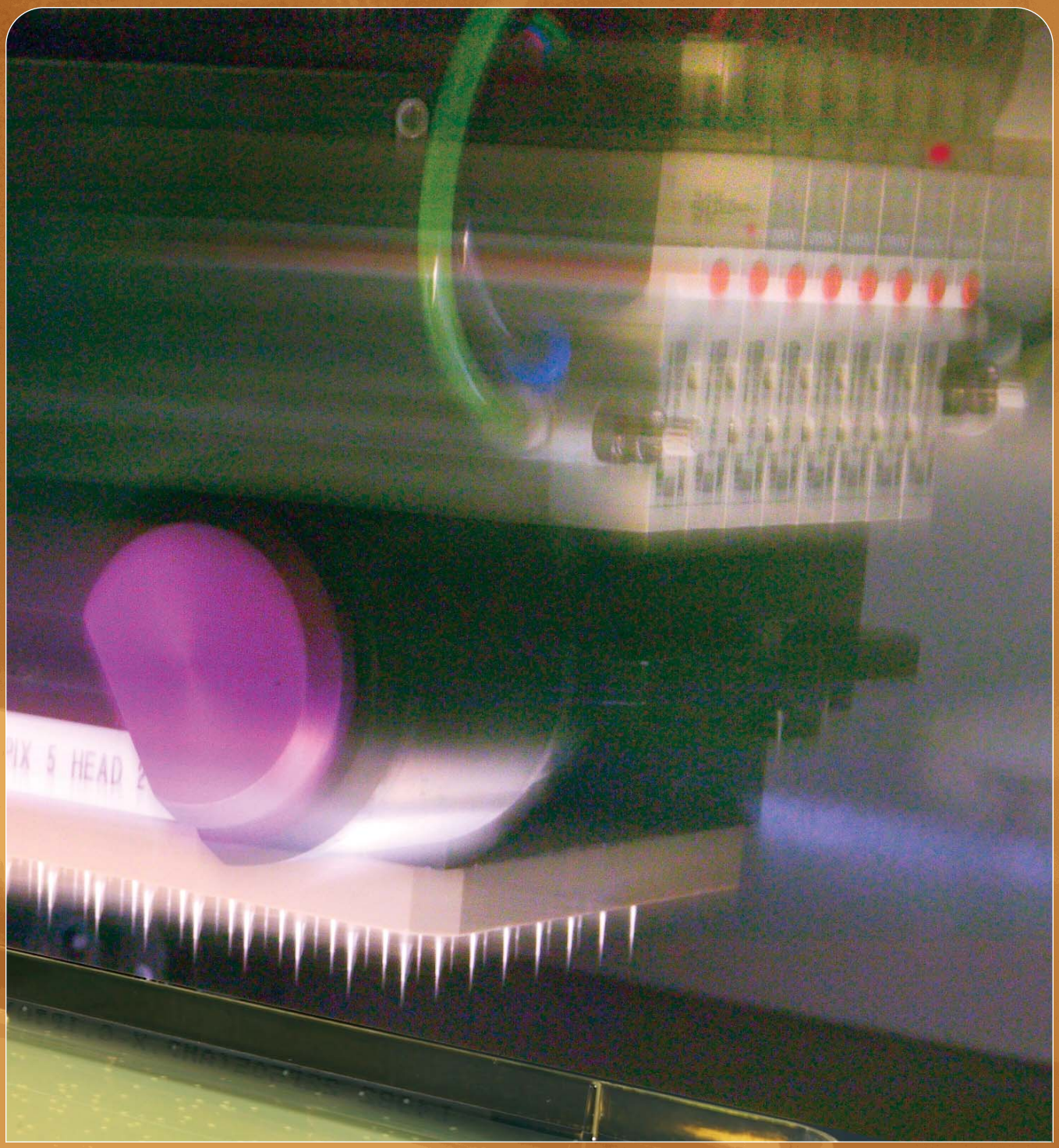
JGI History



The U.S. Department of Energy Joint Genome Institute (JGI) was created in 1997 to unite the expertise and resources in genome mapping, DNA sequencing, technology development, and information sciences pioneered at the DOE genome centers at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL). In 1999, the University of California, which manages the three national labs for the DOE, leased 60,000 square feet of laboratory and office space in a light industrial park in Walnut Creek, California, to consolidate activities and accommodate JGI's 240 employees in what is known now as the Production Genomics Facility (PGF). A partnership with the Stanford University Human Genome Center serves JGI's goal of providing high quality finished sequence to the greater scientific community. In 2005, with the successful completion of the human genome, the JGI has reoriented its mission to now serve as a traditional user facility driven by DOE mission needs, including energy production, carbon management, and bioremediation. This transition has featured broader involvement of the DOE national laboratory system, including the formal participation of Oak Ridge National Laboratory and Pacific Northwest National Laboratory in the activities of JGI.

For more information about JGI's history, facility, partners, and budget, visit our Web site at http://www.jgi.doe.gov/whoweare/index.html.

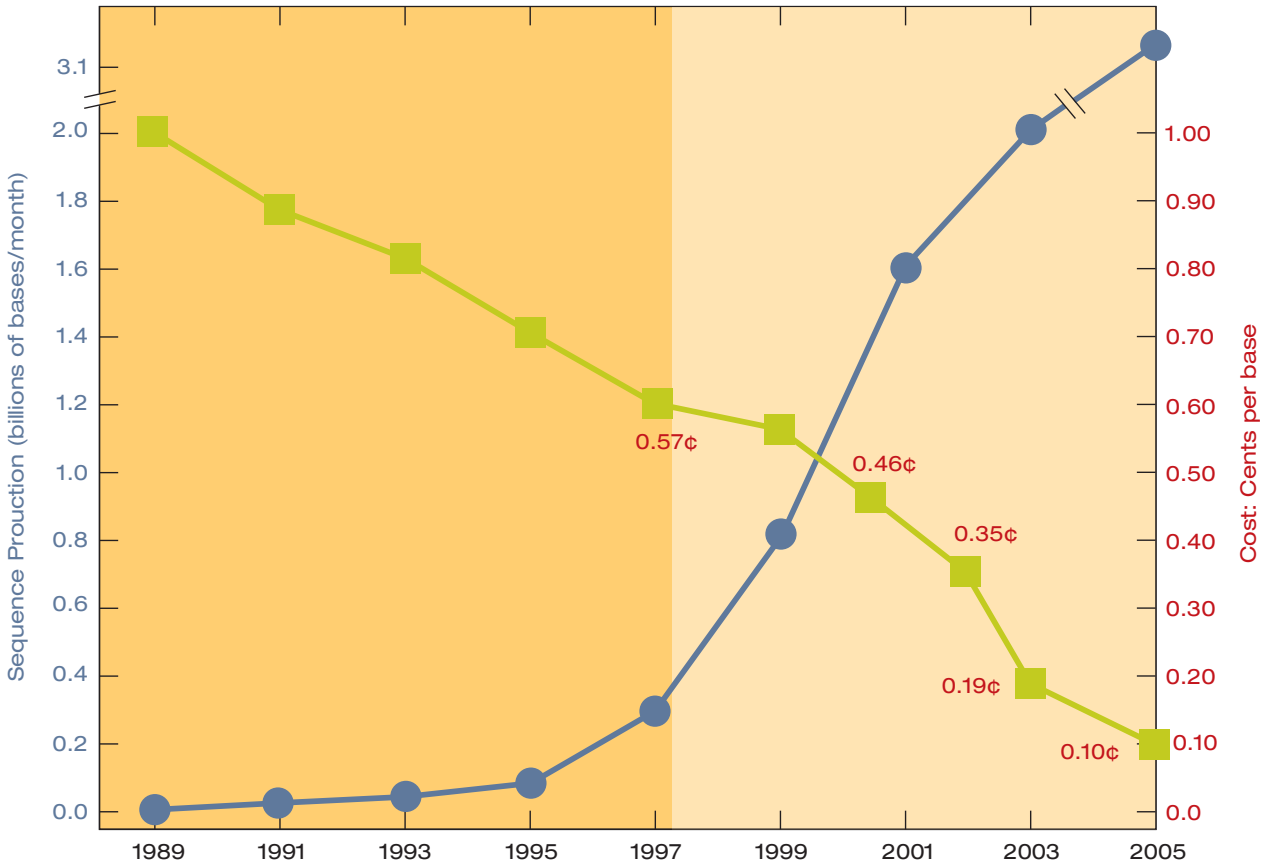




JGI Productivity—Science and Technology Advancements

Twenty years ago, when the Department of Energy first convened a meeting to explore the feasibility of sequencing the human genome and to fund a pilot program to generate that sequence, the technology was slow and laborious. Sequencing entailed mixing up a cocktail of radioisotopes to highlight the individual DNA units of information, then manually loading them on a long, thin vertical gel through which an electronic current was applied. The gel was transferred to filter paper and exposed overnight to x-ray film. The next morning, after developing the film, "ladders" of sequence would be revealed, if all went right, as little black slots corresponding to the four letters or nucleotide bases—the As, Ts, Cs, and Gs that make up an organism's genome or genetic code. These "autoradiograms" were then "read" with a ruler and pencil—a tedious proposition at best. This was a time when PhDs were awarded for characterizing stretches of just a few hundred bases. At the time, with existing technology, it would have taken 1,000 years to completely sequence the human genome.

Today, with the advent of robotics and the capillary DNA analyzer (the days of slab gels and radioactivity are long gone) sequencing has ramped up toward industrial strength at the JGI Production Genomics Facility. This elegantly simple process takes advantage of the negative chemical charge of DNA. In response to the application of an electric current, DNA migrates through the tiny glass tubes, smaller fragments before the larger ones, and as each molecule passes in front of a window at the end of the capillary, a laser excites the first fluorescently-tagged base, and its color and corresponding identity is captured by a computer. As of May 2005, the JGI completed the upgrade of its sequencing machines with the installation of 106 of the most advanced sequencers available, producing DNA sequence 24 hours per day, seven days per week.



Sequence Productivity—Economies of Scale

Dramatic improvements in sequencing efficiencies have emanated from the JGI Production Genomics Facility (PGF) largely as a result of a culture shift from one of a basic research environment to that of a production line. This shift, inspired by cultivating a better understanding among staff of the biological principles underlying the various process tasks, coupled with the advent of standard protocols for the core production groups, has rallied enthusiasm for production goals—in what was once a more relaxed academic setting.

The growth in sequence productivity over the last three years has been impressive—approaching "Moore's Law" proportions—a doubling every 18 months. In March 2005, JGI's monthly sequencing output exceeded 3.1 billion letters (bases) of genetic code generated at the PGF. That represents the equivalent of the entire human genome—or about 1,000 bases of sequence every second. From March 1999 to May 2005, JGI has contributed some 83 billion bases to the worldwide genomics community.

Quality—Prime Watchword for Production

JGI has instituted a quality control program to ensure timely review of protocols and process changes, as well as for evaluating the integrity of the reagents, and proactively troubleshooting problems in the production line.

In the enduring interdisciplinary team approach native to the national laboratory environment, the JGI Production Instrumentation Group has been established to monitor process performance and conduct preventative maintenance. In the same way that industrial-scale manufacturing or retail operations track their units, JGI has implemented a barcode tracking system for individual samples and for the 384-well microtiter plate, the common currency of the sequencing process, which has completely replaced the 96-well format. This strategy dovetails well into the real-time reporting system recently initiated at the PGF.

Technology advancements instituted over the last three years have enabled the production line to increase throughput at significantly reduced cost, while bolstering the quality of daily sequence output.

Computational Tools of the Sequencing Trade—IMG

As the microbial world comes to light through DNA sequencing, the new Integrated Microbial Genomes (IMG) data management system has been developed to facilitate the delivery of valuable sequence information to the global research community.

A product of a collaboration between JGI and the Lawrence Berkeley National Laboratory Biological Data Management and Technology Center (BDMTC), the IMG system was launched in early 2005 as an essential means for investigators to extract information from sequence information. IMG responds to the urgent and increasing need for a means to handle the vast and growing spectrum of datasets emerging from genome projects taken on by the JGI and other public DNA sequencing centers. This important computational tool enables scientists to tap the rich diversity of microbial environments and harness the possibilities that they hold for addressing challenges in environmental cleanup, medicine, agriculture, industrial processes, and alternative energy production.

JGI is currently producing nearly one-quarter of the number of microbial genome projects worldwide, more than any other single institution. As the number of microbial genomes sequenced continues to rise, the genome analysis process becomes the rate-limiting step. By integrating publicly available microbial genome sequence with JGI sequence, the IMG system offers a powerful data management platform that supports timely analysis of genomes from a comparative functional and evolutionary perspective.

IMG's primary goal is to provide high-quality data in a comprehensible system that is diverse in terms of the number of genomes it covers. This goal follows the fundamental principle that the value of genome analysis depends on the quality of the data and increases with the number of genomes available for comparative analysis.

IMG features the ability to integrate information in an evolutionary context, critical for enabling the generation of high-quality annotations and comprehensive metabolic reconstructions—that is, determining the metabolic capabilities of an organism from its DNA sequence.

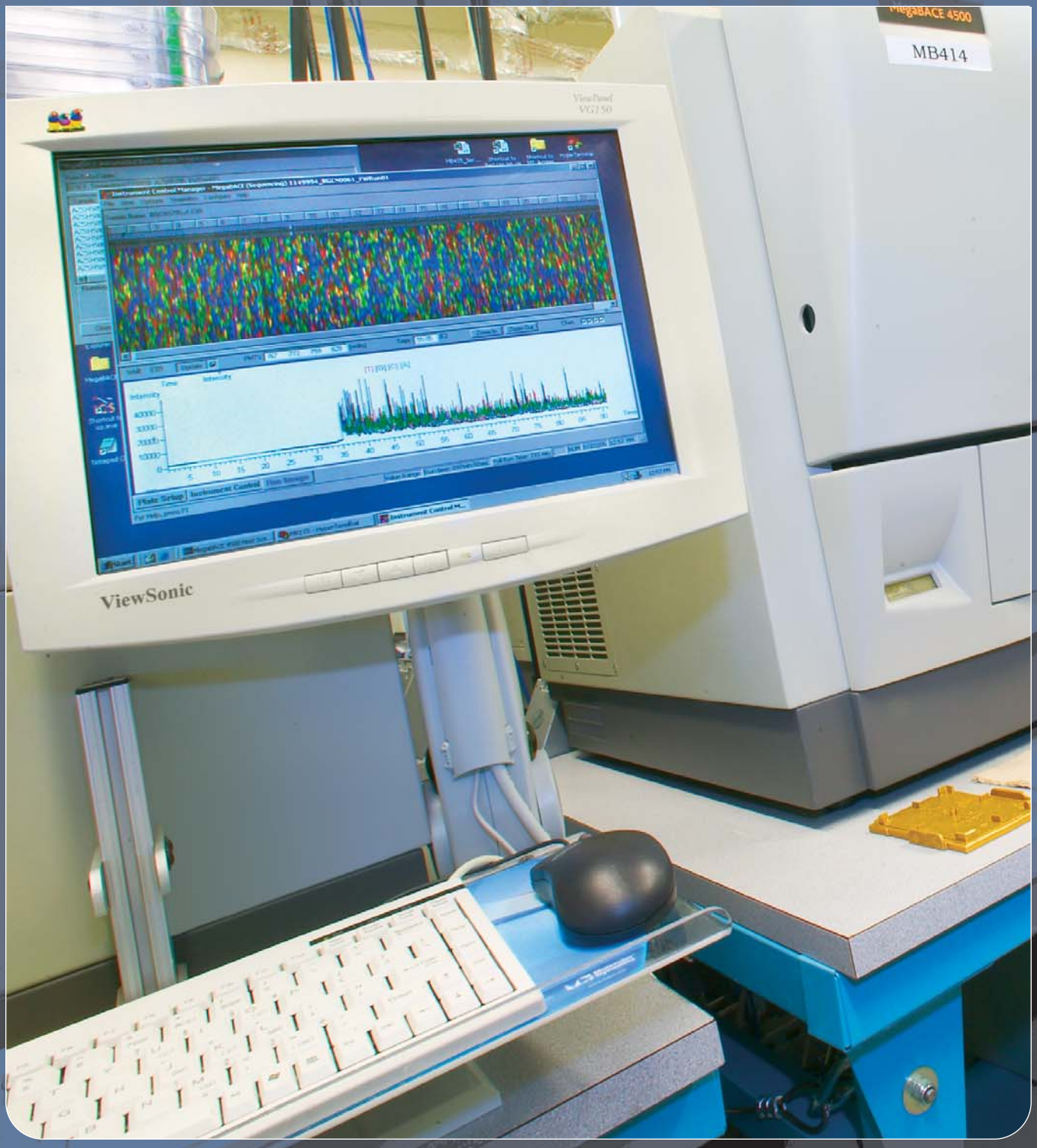
Faster and Cheaper

JGI engineers have made remarkable progress in optimizing many of the steps along the sequencing line for smaller volumes—retrofitting systems with reliable low-volume, fast dispensing of reagents. Adjusting a protocol by as little as two microliter aliquots, and the dilutions of expensive reagents, can lead to considerable savings—millions of dollars over the course of a year. These examples demonstrate ways through which the production sequencing group at the PGF is continually striving to implement improvements geared toward increasing sequencing efficiencies while lowering costs.

The PGF's current stable of capillary electrophoresis sequencing machines comprises 70 Applied Biosystems ABI 3730xls running 24 hours, seven days a week, and 36 GE Health MegaBACE 4500s running 24 hours, 5 days a week.

Safety is of paramount importance throughout every activity of the JGI Production Genomics Facility. Due to the potential for repetitive strain injuries resulting from some of the tasks associated with the sequencing line, JGI has been vigilant about conducting proactive safety and ergonomic assessments in all workstations.

User Community



Serving the DOE Microbial Genome Program

Users of JGI's considerable sequencing resources represent an increasing diversity of disciplines, dependent on sequence information to empower their science, from nearly 150 institutions across more than 40 states in the U.S., and some 90 international institutions in more than 20 countries.

Traditionally, beyond its contribution to the Human Genome Project (HGP), JGI has served as a resource to scientists funded by the DOE Microbial Genome Program. This program entirely focused on providing DNA sequence infrastructure to address issues relevant to DOE's mission, particularly in the areas of alternative energy production, carbon management, and bioremediation.

Despite the perception that the genetic diversity among animals—ranging from humans to worms—is enormous, the reality is that it pales in comparison to the diversity between the microbes that make up the bulk of the biomass on the planet.

Microbes make up around 60% of the earth's biomass. They have survived on the planet for over 3.8 billion years and have been found in every conceivable environment, surviving extremes of heat, cold, radiation, pressure, salt, and acid—often where no other forms of life can exist. This rich diversity means that microbes long ago "solved" many problems for which scientists have been actively seeking answers. A spinoff from the HGP, the goal of this program is to sequence the comparatively small genomes of microbes, which can be completely sequenced in weeks—or with recent advances in sequencing technology, even days. As of May 2005, JGI has sequenced the genomes of more than 175 microbes, more than any other sequencing center. The completed genome of one microbe, *Methanococcus jannaschii*, provided further evidence of a third major branch of life on earth, the Archaea.

Through the study and understanding of a diverse group of microbes, solutions are nearer for DOE mission challenges in environmental cleanup, medicine, agriculture, industrial processes, and energy production and use, to name a few. For example, *M. jannaschii's* ability to produce methane may have implications for new forms of fuel generation, and *Deinococcus radiodurans* has potential for the cleanup of toxic mixed-waste sites containing radionuclides, in addition to heavy metals and organic solvents, because it can survive extremely high levels of radiation and repair its own radiation-damaged DNA.

Launching the Community Sequencing Program (CSP)

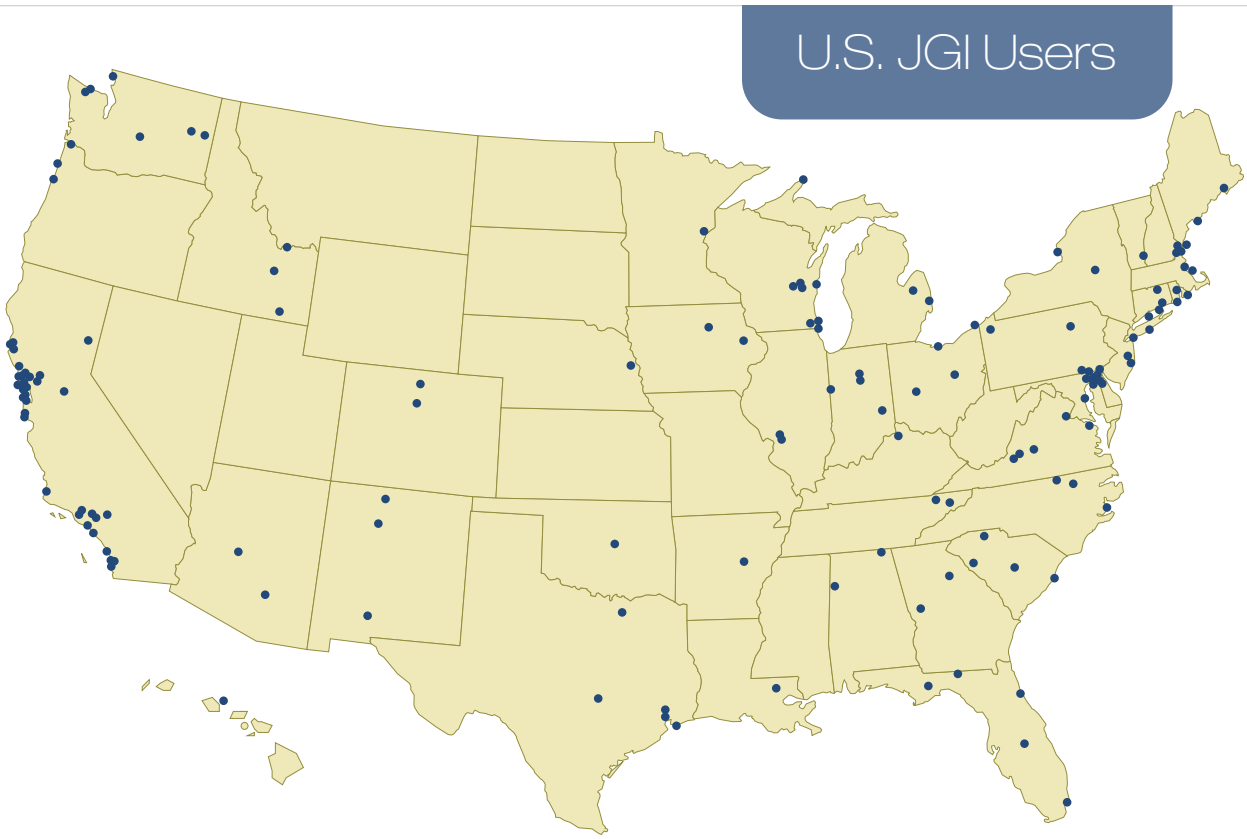
In February 2004, potential collaborators began queuing up with proposals to take advantage of the JGI's powerful DNA sequencing capacity for the debut of the Community Sequencing Program (CSP).

The primary goal of the CSP is to provide a world-class sequencing resource for expanding the diversity of disciplines—geology, oceanography, and ecology, among others—that have come to realize they can benefit from the application of genomics. The model put forward by the CSP has attracted new sequence customers not traditionally served by sequencing centers that have tried to focus on biomedical applications. Just as physicists and climatologists submit proposals to get time on accelerators and supercomputers, respectively, to address fundamental questions, investigators bring to the CSP important scientific challenges that can be informed with large amounts of sequence. The CSP, in effect, covers the biosphere of possibilities.

CSP proposals are submitted to a rigorous competitive process, and ultimately vetted by an external scientific review panel. While the major acceptance criteria for projects emphasize scientific excellence, DOE mission relevance is also stressed. As with all other sequencing projects at the JGI, data generated will be made available to the entire scientific community through the Web.

The CSP consists of two programs, one geared toward smaller projects—genomes of less than 250 million bases, or requiring less than a total commitment of one billion bases—and one dedicated to larger projects.

The small program, which includes prokaryotes and more complex multicellular organisms, also embraces metagenomic projects that seek to unravel the complex interactions of microbes inhabiting environmental communities that defy culturing in the laboratory. Large-genome program targets



must address relevance to the DOE missions of environmental remediation, carbon sequestration, and alternative energy production.

CSP PORTFOLIO 2005

Projects submitted to the CSP in 2004 for sequencing in 2005 represented a rich collection of microorganisms; higher plants and animals that inhabit both aquatic and terrestrial ecosystems.

Twenty-three proposals were deemed extremely meritorious and were therefore selected for sequencing. An example of a larger organism chosen was the moss *Physcomitrella patens*, which has a genome size of just over half a billion bases. *Physcomitrella*, submitted by collaborators at the Jepson Herbaria at University of California, Berkeley, serves as an expedient model system in that it is small, grows quickly, and is very amenable to comparative studies.

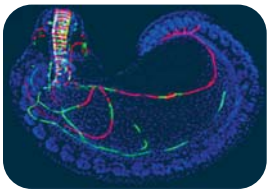
Comparing the human genome to other animals, an approach known as *comparative genomics*, has been an invaluable catalyst in deciphering features of the human genome. Human genomics has thus benefited from having a series of genome projects along the tree of life—mouse, pufferfish, fruit fly, worm—while plant genomics has suffered, since only a closely related cluster of cereals and the mustard *Arabidopsis* have been sequenced. Having the *Physcomitrella* sequence available is said by researchers to be a triumph for international plant science.

Mosses were among the first plants to colonize the land 450 million years ago. They predate the flowering plants by some 200 million years of evolutionary time. Mosses can do many of the things that the flowering plants have forgotten. For instance, some of their primitive traits—like the ability to survive extremes of dehydration—would be useful to incorporate in modern-day crops, especially in less-developed countries. By studying the genes that control these traits in the moss, researchers should be able to identify how these characteristics could be revived in flowering plants.

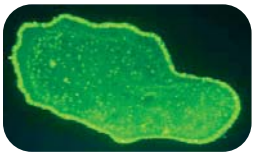
Adding to the JGI's leadership in plant genomics, an area of relevance to the DOE mission owing to the major role

plants play in carbon sequestration, was the selection of *Selaginella moellendorffii*, or the Gemmiferous Spike Moss, for sequencing. *Selaginella* and *Physcomitrella* are the first nonflowering vascular plants to be sequenced.

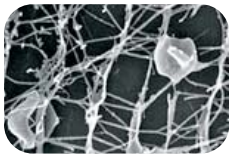
The *Selaginella* sequence leads researchers down the evolutionary path toward such traits as those that allow plants to survive and thrive on dry land. It will also enable investigators to identify proteins, metabolites, or small molecules produced by this plant that may be beneficial for human health and agriculture.



The 2005 CSP portfolio also targeted several animals selected to fill critical gaps in the tree of life. One will be the leech *Helobdella*, long used as a model system by biologists studying embryological development and functions of the nervous system. Along with the polychaete worm, *Capitella*, and the mollusk *Lottia*, also selected, these are expected to be the first representatives sequenced from the large animal group dubbed the Lophotrochozoa. This group, comprising about one third of all animal phyla, promises to reveal many of life's processes and to exemplify intermediate features underlying the patterns of genome evolution. Early branches of the tree of life will be represented by the sponge *Reniera* and an unusual organism called *Trichoplax adhaerens*. With a superficial resemblance to a giant amoeba, *Trichoplax* has the smallest animal genome, at less than 50 million base pairs.



Other organisms in the 2005 CSP queue include the cold water-dwelling microbes, *Crenarchaeota*, which offer another important foray into environmental genomics. *Crenarchaeota* have not yet been cultivated, so their biological properties are not well known. Oceanographers are particularly interested in how these microbes may be involved in the carbon and energy cycles of the deep sea. *Crenarchaeota* are members of the Archaea, a major branch of life that includes many microbial extremophiles—microbes that can live at extreme temperature, salinity, or high acidity.



Additional 2005 CSP DNA sequencing projects are summarized in Appendix B.

CSP PORTFOLIO 2006

With the 2006 CSP allocation, JGI will be making available to the greater scientific community 20 billion bases of information, roughly the equivalent of nearly seven human genomes of information. This year 135 proposals were submitted, nearly a 2.5-fold increase from the CSP's inaugural call for proposals in 2004, of which more than 40 were deemed of exceptional scientific importance to be included in the 2006 CSP allotment.

The largest single genome selected in 2005, the tropical grain *Sorghum bicolor*, was proposed by an international consortium led by researchers at the University of Georgia and Rutgers University. The *Sorghum* sequence will complement the knowledge already gleaned from rice, the only other monocot grain to have been sequenced to date. *Sorghum*, with its economic importance worldwide exceeding $69 billion per year, is expected to provide an improved blueprint for the study of other important grains such as maize, millet, and sugarcane. *Sorghum*, with a relatively compact genome of approximately 736 million bases, will also serve as a valuable reference for analyzing the four-fold larger genome of maize, the leading U.S. fuel ethanol crop. *Sorghum* is an even closer relative of sugarcane, arguably the most important biofuels crop worldwide, with annual production of about 140 million metric tons with a value approaching $30 billion.

The *Sorghum* genus also includes one of the world's most noxious weeds. The same features that make the weedy

"Johnson grass" *(S. halepense)* so tenacious are actually desirable in many forage, turf, and biomass crops. Thus, *Sorghum* offers novel learning opportunities relevant to weed biology as well as to crop improvement.

Another major CSP genome target, *Mimulus guttatus*, the common or "seep spring" monkey flower, although not a food crop, is a relative not too distant from the likes of tomato, potato, and other dicot, or broadleaf, crops. Researchers from Duke University, who proposed the project, predict that insights from deciphering the monkey flower at the genomic level will highlight its path of evolution and adaptation, a map readily transferable to crop plants.

By sequencing the monkey flower, JGI will enable genomicists to pioneer new territory, taking on one of the most difficult and fundamental questions in evolutionary biology—how new species evolve. The genus *Mimulus* is an ideal model system for this problem, because it exhibits two different types of speciation, one being the evolution of pollinator specificity and the other being the evolution of mating systems.

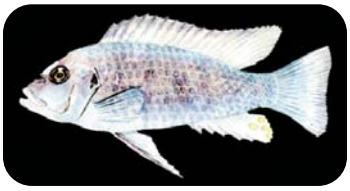
M. guttatus is also quite tolerant of soil conditions that would be toxic to other plants. For instance, the species thrives on soils composed of California's state rock—serpentine—which contains high levels of magnesium, nickel, and manganese. Sequencing the monkey flower promises a better understanding of how plants can help remediate soil contaminated with toxic metals.

Fueling the Future—Learning from Termite Hindgut Microbiota

One of DOE's most enduring goals is to replace fossil fuels with renewable sources of cleaner energy, such as hydrogen produced from plant biomass fermentation. The lowly termite is actually one of the planet's most efficient bioreactors, capable of cranking out two liters of hydrogen from fermenting just one sheet of paper. Termites accomplish this Herculean task by exploiting the metabolic capabilities of microorganisms inhabiting their hindguts. JGI will be sequencing this community of microbes to provide a better understanding of the biochemical pathways used in the termite hindgut, which may lead to more efficient strategies for converting biomass to fuels and chemicals. Similarly, an ability to harness the pathways directly involved in hydrogen production in the termite gut may one day make biological production of this alternative energy source a viable option.

Fish for Food and Links to the Past

With the 2006 allocation, JGI cast deep into the aquatic gene pool—to sequence genes from two species of catfish, the channel catfish (Ictalurus punctatus) and the blue catfish (I. furcatus). Catfish farming is a two billion dollar industry annually in the United States alone, representing 68% of all aquaculture production.



In addition, the CSP will facilitate the sequencing of five species of fish of the family Cichlidae from Lake Malawi in east Africa. Popular food fish and aquarium specimens, Cichlid fish have undergone an astonishingly rapid proliferation of species from this evolutionarily fertile source. Over the last two million years, some 700 species have emerged from the depths of Lake Malawi.

See Appendix B for additional CSP 2006 project summaries.

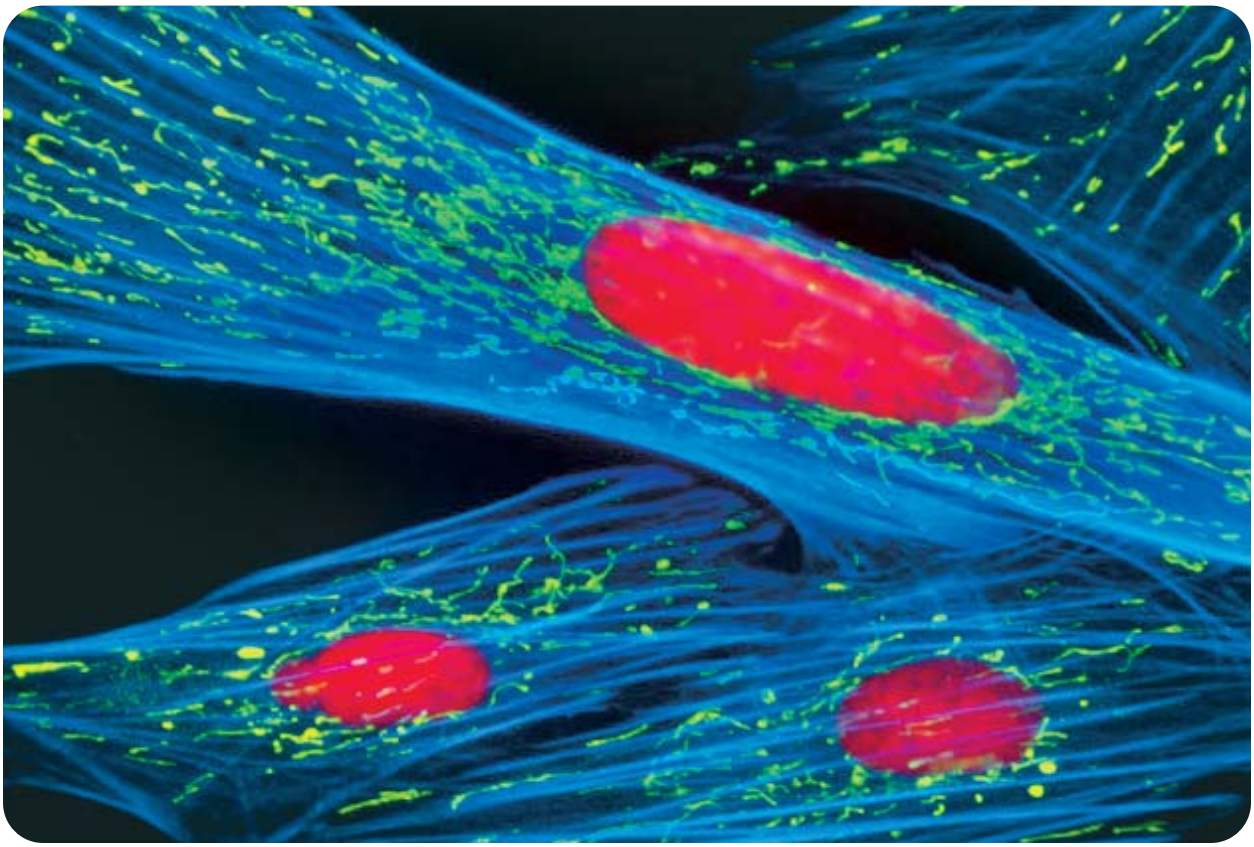
For more information about the JGI's Community Sequencing Program (CSP), visit our Web site at http://www.jgi.doe.gov/CSP/index.html.

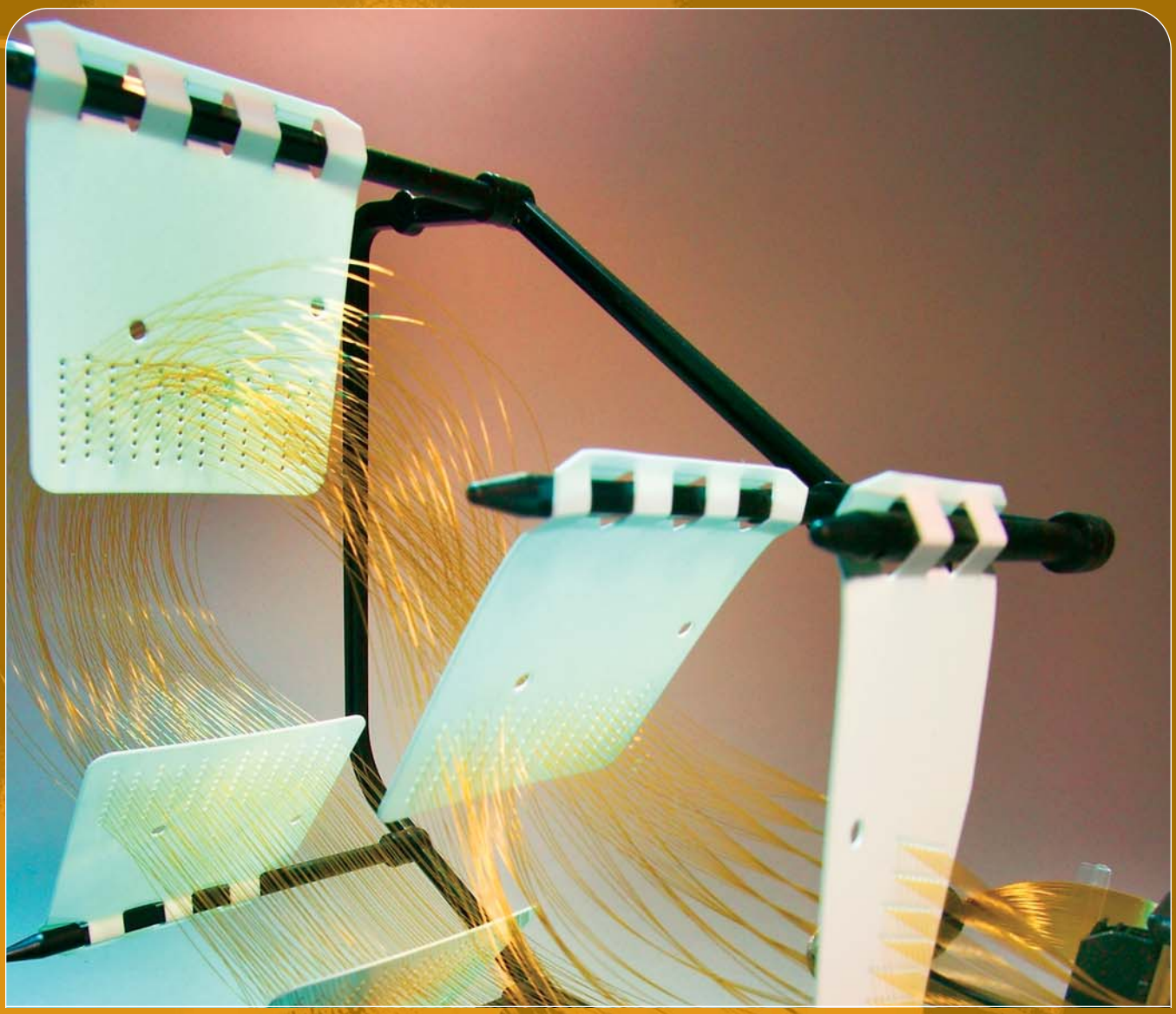
The lowly termite is actually one of the planet's most efficient bioreactors, capable of cranking out two liters of hydrogen from fermenting just one sheet of paper. JGI is seeking to harness this capability.

JGI 2002–2005—In Summary

Integral to the mission of the Department of Energy is the goal of the Joint Genome Institute to employ DNA sequencing and computational technologies as a means for discovering and characterizing the basic principles and relationships underlying the organization, function, and evolution of living systems. The JGI, having completed its significant role in the international Human Genome Project—generating the complete sequences of Chromosomes 5, 16, and 19—has moved on to contributing in other critical areas of genomics research. While NIH-funded genome sequencing activities largely continue to emphasize human-related biomedical targets and applications, the JGI has since shifted its focus to non-biomedical targets and environments that play a vital role in preserving the health of the planet. With efficiencies of scale established at the Production Genomics Facility (PGF), and progressively increasing sequencing capacity, the JGI has now tackled scores of additional genomes. These include more than 60 microbial genomes and communities of microbes that occupy important environmental niches, in addition to such important eukaryotic model systems as the pufferfish *(Fugu rubripes)* and a sea squirt *(Ciona intestinalis)*—a primitive chordate, or organism with a spinal cord. In partnership with other federal institutions and universities, the JGI played a leadership role in the sequencing of a frog *(Xenopus tropicalis)*, a green alga *(Chlamydomonas reinhardtii)*, a diatom *(Thalassiosira pseudonana)*, a white rot fungus *(Phanerochaete chrysosporium)*, the first tree ever sequenced, the black cottonwood or poplar *(Populus trichocarpa)*, and a host of agriculturally important plant pathogens.

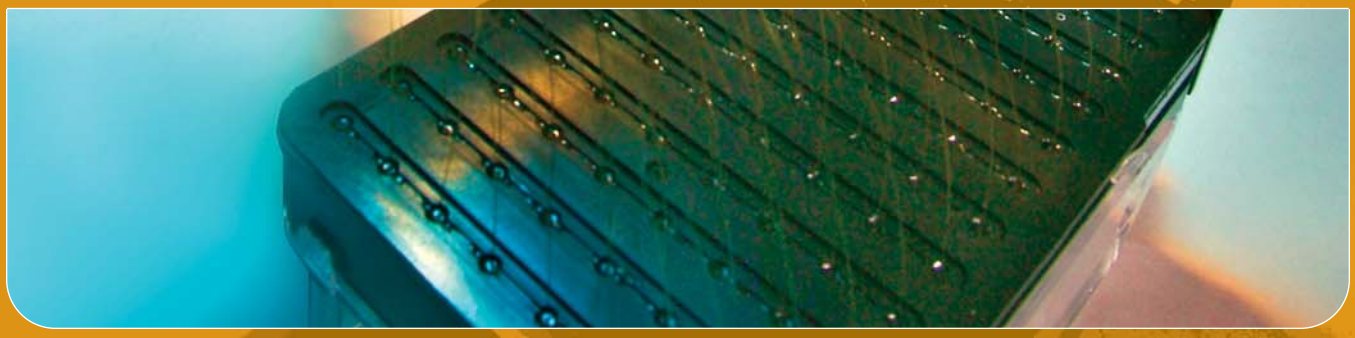
Microorganisms—for example, those that thrive under extreme conditions such as high acidity, radiation, and metal contamination—are of particular interest to the DOE and JGI, as are those that have relevance for carbon sequestration, bioremediation, and exploration of alternative energy sources. Investigations by JGI and its partners are shedding light on the cellular machinery of microbes and how they can be harnessed to clean up contaminated soil or water, capture carbon from the atmosphere, and produce potentially important sources of energy such as hydrogen and methane.

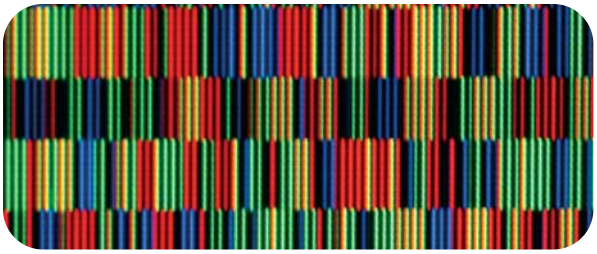




Departments & Programs

The JGI Production Genomics Facility (PGF) is comprised of teams of researchers and support staff aligned under three departments representing the infrastructure of the PGF—Operations, Sequencing, and Informatics—and six research programs: Genomic Technologies, Vertebrate, Evolutionary Genomics, Computational Genomics, Microbial Ecology, and Microbial Genome Analysis.





Sequencing Department

The JGI Sequencing Department resides at the heart of the JGI Production Genomics Facility. The department generates high-quality sequence in a cost-efficient manner, expediting DNA through the process from library creation and sequencing preparation, to capillary sequencing and analysis. As genomics is a rapidly changing field, the department constantly adapts to take advantage of new developments in technology. The JGI Sequencing Department is comprised of several subgroups, including several subgroups, including Mapping, Library Support, Sequencing Preparation, Quality Control, Sequence Assessment and Analysis, and Instrumentation.

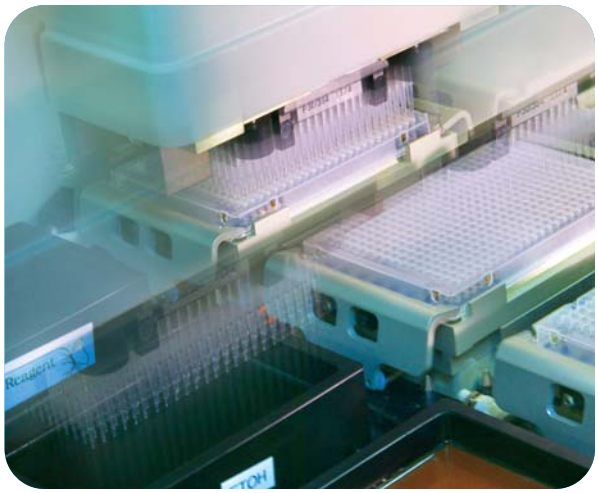
For more information about the sequencing process, visit our Web site at http://www.jgi.doe.gov/sequencing/index.html.

Informatics Department

With the wholesale improvements in the amount of DNA sequence generated monthly, managing those data and extracting the informative kernels has become increasingly challenging. The Informatics Department provides computational infrastructure and support to facilitate JGI's production and research and development efforts. Informatics is comprised of several subgroups, including Production Informatics, Assembly, Comparative Genomics, Software Support, Desktop Support and Servers, Genome Data Systems, and Genome Annotation.

Genomic Technologies Program

The Genomic Technologies Program works to make the sequencing and assembly process at JGI more efficient and to improve the quality of genomes produced. It accomplishes this by developing new protocols and evaluating new methods and instruments for use in production. Its efforts improve the overall quality of genomes produced at JGI and continually increase the Institute's capabilities. Notable initiatives include developing methods to improve read lengths, decrease reagent costs, and quickly close gaps in the sequence data.



Computational Genomics Program

The Computational Genomics Program develops new analytical tools and data management systems that transform the raw data produced by the JGI into biologically valuable information and insights. These tools are designed to facili-

tate the use of JGI data by the biological community. This work is essential for managing and visualizing the expanding body of genome-scale data, and linking it to functional and phenotypic information generated at the JGI and elsewhere.

Vertebrate Program

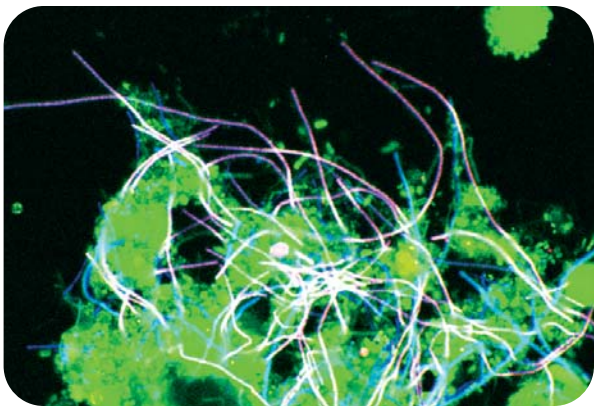
The Vertebrate Program is involved in generating and exploiting sequence data generated by JGI. These activities include studies focused on scanning large vertebrate genome assemblies, such as those produced for human *(Homo sapiens)*, frog *(Xenopus tropicalis)*, and pufferfish *(Fugu rubripes)*. Comparison of human genomic sequence to that of other vertebrates has proven to be efficient in uncovering evolutionarily conserved modules with in vivo gene-enhancer activity. In addition, the program has established a high-throughput resequencing pipeline in conjunction with the Sequencing Department to assess human genetic changes for their contributions to a variety of diseases.

Evolutionary Genomics Program

A relatively new field of study, evolutionary genomics represents the interface between evolutionary biology and genome science. Through analysis of genomic data and laboratory experimentation, this program addresses the big questions of how life evolves and genomes change over time.

Microbial Ecology Program

To date, molecular microbial ecology has relied heavily on small subunit ribosomal RNA (rRNA) sequence for culture-independent characterization of microbial communities. The Microbial Ecology Program (MEP) uses sequencing-based technologies to understand microbial communities through a combination of computational and experimental methods.

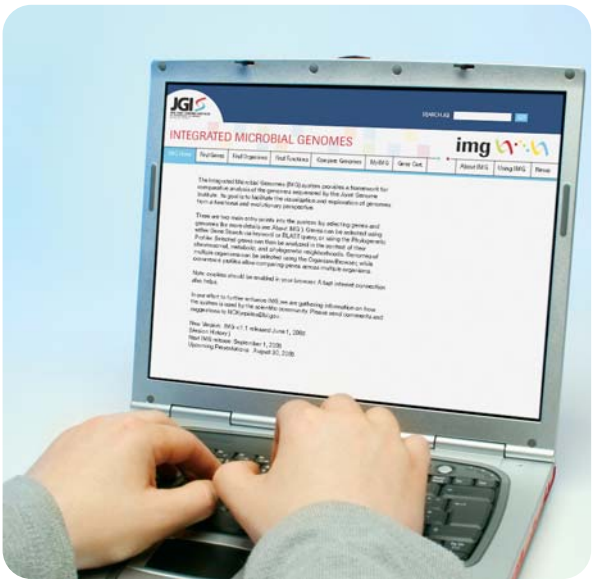


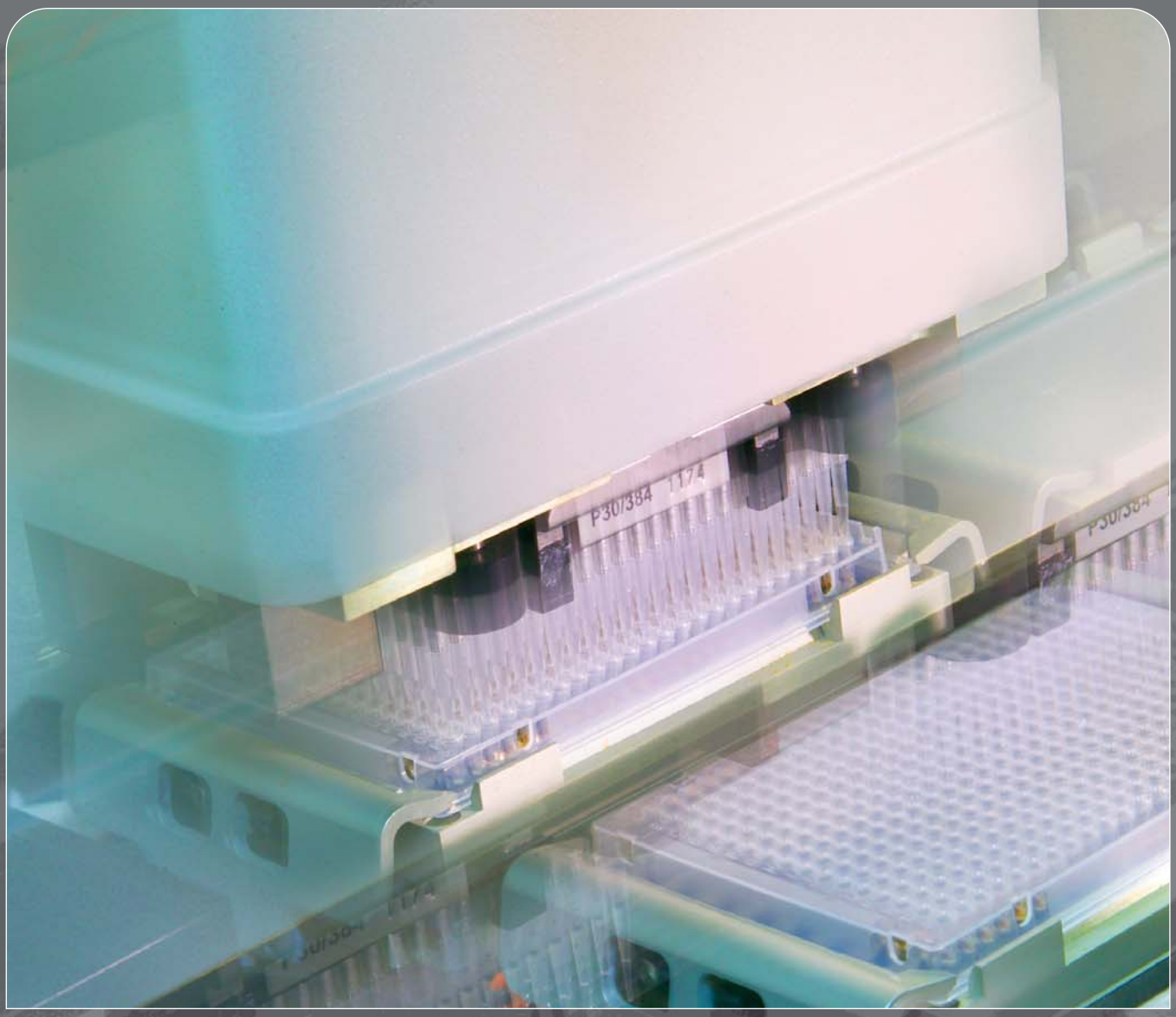
MEP has established a high-throughput pipeline for analysis of rRNA signature sequence and fluorescence in-situ hybridization (FISH) to visualize phylogenetic groups under the microscope. Building on this work, MEP is pioneering the emerging field of metagenomics—cloning, sequencing, and characterizing DNA extracted directly from environmental samples—to obtain an overview of community function and population dynamics. Since environmental shotgun sequencing is in its infancy, MEP is exploring ways to analyze and visualize metagenomic data together with the JGI's Microbial Genomics Analysis Program.

Microbial Genome Analysis Program

The identification of the complete set of functions of any organism provides the foundation upon which our understanding of the biology of that organism rests. In essence, it forms the basic framework that any genome project targets, and from which any biological interpretation originates. However, while the quality and quantity of sequencing data has dramatically increased during the last few years, their interpretation remains a major bottleneck.

In fact, as more and more microbes are sequenced, the scientific community's efforts to assign functions to genes are lagging. In addition, the importance of comparative analysis and extensive sequence integration for a comprehensive genome analysis and reconstruction of the functional cellular subsystems (e.g. metabolic pathways, information processes), has been largely overlooked by most contemporary genome databases. Among the many objectives of the Microbial Genome Analysis Program is to speed up annotation through the development of software tools for determining microbial gene function.





Science Highlights

The last three years have been particularly productive for the JGI. In parallel with completing its human genome contribution, the complete sequence of several important model organisms were published.

Refining the Three DOE Chromosomes—5, 16, & 19

The Human Genome Project (HGP) stands among the crowning achievements of 20th century biology. The Department of Energy's role as the project's initial prime mover, and its continued leadership throughout the course of the project, reminds us of the enormous wealth of talent in the DOE national laboratory system—talent that, in the case of the HGP, has served not just the nation, but all the peoples of the world.

The Department of Energy, and its predecessor agencies, have sponsored genomics research for decades, including basic studies of DNA replication, damage, and repair, and the consequences of radiation-induced heritable mutations. In 1987, recognizing its pioneering contributions to discovery in large-scale science, the "Report on the Human Genome Initiative," recommended DOE assume a major role in an expansive multidisciplinary undertaking to map and sequence the human genome. Thus, between 1988 and 1989, three genome research centers were established at Lawrence Berkeley, Lawrence Livermore, and Los Alamos National Laboratories. These were combined in 1997 into the DOE Joint Genome Institute (JGI). DOE was joined in 1990 by the National Institutes of Health, which eventually assumed a leadership role in the multi-agency, international project.

In 2004, JGI become the first of the five primary Human Genome Project sequencing sites, known as the "G5," to complete their publication of scientific articles describing each of the human chromosomes that they originally committed to sequence. DOE's commitment entailed chromosomes 5, 16, and 19, all sequenced by JGI, representing 11% of the human genome.

The effort to complete the sequence of these chromosomes included over 100 researchers from the partnership of Los Alamos, Lawrence Berkeley, and Lawrence Livermore National Laboratories, as well as the Stanford Human Genome Center, and Children's Hospital in Oakland, California.

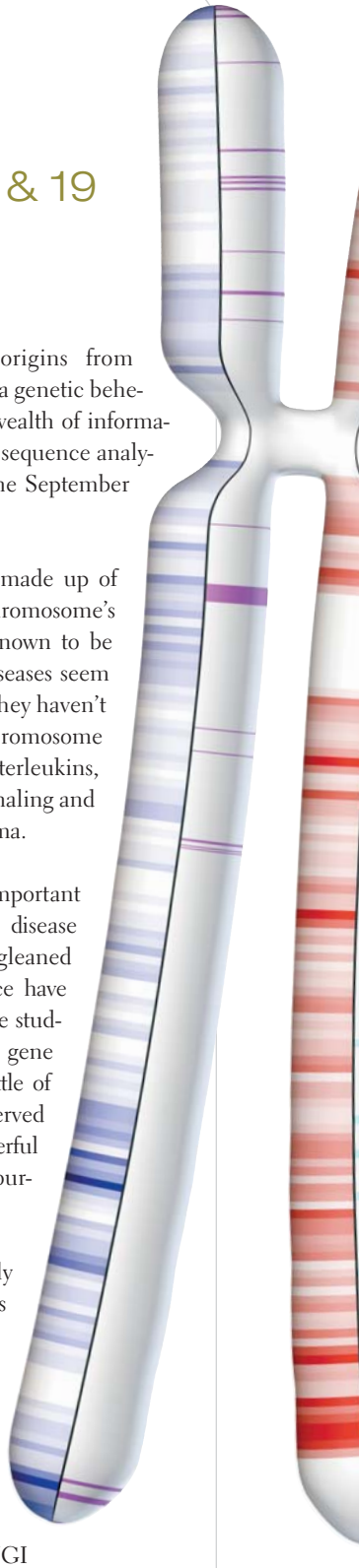
CHROMOSOME 5—VAST GENE DESERT

Chromosome 5, with its historic origins from Lawrence Berkeley National Laboratory, is a genetic behemoth containing key disease genes and a wealth of information about how humans evolved. The final sequence analysis was published by JGI investigators in the September 16, 2004 issue of *Nature*.

Chromosome 5, one of the largest, is made up of 180.9 million bases which spell out the chromosome's 923 genes, including 66 genes that are known to be involved in human disease. Another 14 diseases seem to be caused by chromosome 5 genes, but they haven't yet been linked to a specific gene. Other chromosome 5 genes include a cluster that codes for interleukins, molecules that are involved in immune signaling and maturation and are also implicated in asthma.

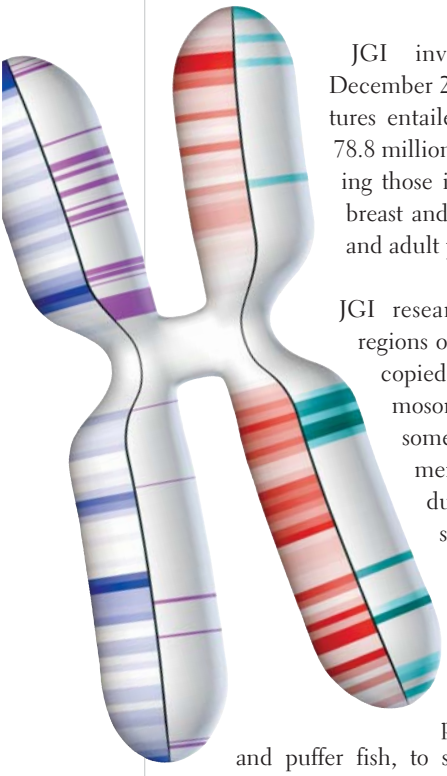
The spaces between the genes are as important as the genes themselves. In addition to disease genes, other important genetic motifs gleaned from vast stretches of noncoding sequence have been found on Chromosome 5. Comparative studies conducted by JGI scientists of the vast gene deserts, where it was thought there was little of value, have shown that these regions, conserved across many mammals, actually have powerful regulatory influence, as described in the journal *Science* from October 17, 2003.

These gene-free stretches were previously considered "junk DNA," but in recent years those seemingly barren regions have taken on greater prominence as researchers have learned that they can control the activity of distant genes. Some of the noncoding regions have also stayed remarkably consistent compared with those in mice or fish, rather than accumulating mutations over the course of evolution. This work by JGI was highlighted in a publication in the October 21, 2004 edition of *Nature*.



CHROMOSOME 16—
HIGHLY REPETITIVE TERRAIN

Chromosome 16 was the original focus of DNA repair gene studies initiated at DOE's Los Alamos National Laboratory in 1988. Additional interest stemmed from the discovery of genes on chromosome 16 implicated in the detoxification and transport of heavy metals.



JGI investigators published, in the December 23, 2004 edition of *Nature*, the features entailed in Chromosome 16, citing its 78.8 million bases, home to 880 genes including those implicated in the development of breast and prostate cancer, Crohn's disease, and adult polycystic kidney disease.

JGI researchers characterized the many regions on chromosome 16 that have been copied to other places within the chromosome, and even to the other chromosomes—a phenomenon known as segmental duplication. While segmental duplication is present on all chromosomes, it is particularly prevalent on chromosome 16. Researchers compared these human sequences to regions conserved over time in other vertebrate genomes, including chimpanzee, dog, mouse, rat, chicken, and puffer fish, to shed light on changes that have occurred since the last common ancestor, ranging from five million to 400 million years ago.

CHROMOSOME 19—
THE GENE MOTHER LODE

The complete sequence of Chromosome 19, the most gene-rich of all the human chromosomes, was described by JGI investigators in the April 1, 2004 edition of the journal *Nature*.

Embedded in this sequence information are critical regulatory networks of genes tasked with controlling such functions as repairing DNA damage caused by exposure to radiation and to other environmental pollutants. Studies of DNA-repair genes, initiated at the DOE national laboratories, are yielding insights into the development of certain cancers, many of which appear to be caused by defects in DNA-repair pathways. Also, new insights are being gleaned about other gene families implicated in detoxifying and excreting chemicals foreign to the body.

Chromosome 19, at 55.8 million bases, although representing only about 2% of the human genome, features nearly 1,500 genes. They include genes that code for such diseases as insulin-dependent diabetes, myotonic dystrophy, migraines, and familial hypercholesterolemia (an inherited form of elevated blood cholesterol), which increases the risk of cardiovascular disease.

Beyond the significant revelation that chromosome 19 has more than twice the gene density of the genome-wide average, it also offers a fertile landscape for exploring evolutionary motifs. An intriguing picture has emerged regarding conservation and divergence, revealing large blocks of gene conservation with rodents as well as segments of coding and noncoding conservation with more distant species such as the pufferfish, *Fugu rubripes*, which was also sequenced at the JGI. While these noncoding regions were considered nonsense until recently, now they are actually proving to have powerful regulatory influence over the genes that they bracket.

The DOE originally selected chromosome 19 as a sequencing target because of the agency's abiding mission of investigating the link between DNA damage from radiation exposure and human cancer. Initial work conducted by Lawrence Livermore National Laboratory in the mid-1990s led to the mapping of multiple DNA-repair genes on chromosome 19. Chromosome 19 topography is filled with such biologically interesting features as transcription factors, olfactory-receptor genes, and zinc-finger genes. Transcription factors are proteins that need to be recognized by RNA polymerase in order to initiate the elaboration of nucleotides along the DNA molecule. Zinc-finger proteins are chains of amino acids that capture a zinc ion, binding to RNA or DNA, and play a critical role in a cell's life cycle. Drug development efforts seek to disrupt these zinc-finger structures to prevent viruses from functioning.

The Pufferfish

THE SUCCINCT HUMAN GENOME

Following JGI's completion of its commitment to the Human Genome Project, it tackled the genome of the Japanese pufferfish, *Fugu rubripes*, which has the smallest known genome of any vertebrate species. As vertebrates, fish and humans share not only the defining characteristic of a backbone, but also many basic anatomical and physiological similarities. The compact *Fugu* genome contains the same basic vertebrate blueprint as the human genome in a sequence seven times shorter. This difference is primarily due to the scarcity in *Fugu* of the large repeat-filled tracts that litter the human genome. The relative compactness of the *Fugu* genome simplifies the detection and analysis of both gene sequence and gene-regulatory elements.

Beyond being a sushi delicacy in Japan—providing the tasty flesh is not tainted with a potent neurotoxin liberated by improper preparation—the analysis of the *Fugu* sequence helps illuminate the human genome. By comparing the human and *Fugu* sequences, common functional elements such as genes and regulatory sequences can be recognized as having been preserved in the two genomes during the 450 million years since the species diverged from their common ancestor. In contrast, nonfunctional sequences are randomized over this long time period. More than 30,000 *Fugu* genes have been identified in our analysis. The great majority of human genes have counterparts in *Fugu*, and vice versa, with notable exceptions including genes of the immune system, metabolic regulation, and other physiological systems that differ in fish and mammals. Nearly 1,000 previously unrecognized human genes are predicted by comparing the two genomes. Remarkable stretches of sequence were found, containing dozens of genes whose linkage is conserved between humans and *Fugu*, shedding light on the large-scale evolutionary processes that shape genomes.

The *Fugu* genome sequence analysis, presented by JGI investigators and colleagues led by Nobelist Sydney Brenner, was highlighted in the August 23, 2002 issue of *Science*.

The Sea Squirt

A MODEL FOR DEVELOPMENTAL BIOLOGY

Another important model organism sequenced by JGI was *Ciona intestinalis*, the smallest of any experimentally manipulable chordate organism with a spinal cord. This organism provides a good system for exploring the evolutionary origins of the chordate lineage, from which all vertebrates sprouted. The animal's genome shows many similarities with the human and mouse genomes. Though containing only 15,000–16,000 protein-coding genes—half the size of the human genome—the *Ciona* genome contains genes similar to human genes that code for hormones and for components of the human immune system and nervous system. About 80% of *Ciona's* genes are also found in humans and other vertebrates. Comparison of the *Ciona* genome with the genomes of other animals also provides clues to the evolutionary origins of the human brain, spine, heart, eye, and thyroid gland.

Ciona's 15,000-plus genes may be controlled by some 10,000 regulatory DNAs, including enhancers and silencers. The complete *C. intestinalis* genome sequence provides a foundation for genome-scale analysis of regulatory networks during development.

C. intestinalis has easily visualized cells and morphogenetic processes, existing methods for transient transgene expression, and is available throughout the world all year long. In addition, there is a deep classical literature on sea squirt development and an active community of researchers worldwide.

JGI researchers published the sequence of *Ciona* in the December 13, 2002 issue of the journal *Science*.

White Rot

NATURE'S OWN WOOD PULP PROCESSOR

JGI's contribution to solving the sequence of *Phanerochaete chrysosporium*, a white rot fungus, was published in the June 2004 edition of the journal *Nature Biotechnology*. What makes white rot fungi compelling for industrial application is that they are the only known microbes capable of efficiently degrading the recalcitrant aromatic plant polymer lignin, one of the most abundant natural materials on earth. The unique extracellular oxidative enzymes they produce can also be harnessed to deactivate compounds found in explosive-contaminated materials, pesticides, and toxic wastes.

Lignin plays a key role in the carbon cycle as the most abundant aromatic compound in nature, providing the protective matrix surrounding the cellulose microfibrils of plant-cell walls. Although lignin is a formidable substrate, its degradation by certain fungi was recognized and described nearly 125 years ago. The organisms responsible, collectively referred to as white rot fungi (since they degrade brown lignin and leave behind white cellulose), are basidiomycetes, a fungal group that includes edible mushrooms as well as plant pathogens such as smuts and rust.

Phanerochaete chrysosporium, the first white rot fungus genome to be sequenced, secretes an array of peroxidases and oxidases that act nonspecifically via the generation of lignin-free radicals, which then undergo spontaneous cleavage reactions. The nonspecific nature and exceptional oxidation potential of the enzymes has attracted considerable interest for application in bioprocesses such as organopollutant degradation and fiber bleaching.

Phanerochaete chrysosporium has several useful features. First, unlike some white rot fungi, it leaves the cellulose of the wood virtually untouched. Second, it has a very high optimum temperature (about 40° C), which means it can grow on wood chips in compost piles that attain a very high temperature. These characteristics point to some possible roles in biotechnology.

The Diatom

KEY PLAYER IN GLOBAL CARBON MANAGEMENT

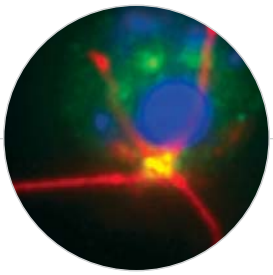
In 2004, JGI produced the first genetic instruction manual of a diatom from a family of microscopic ocean algae that are among the Earth's most prolific carbon dioxide assimilators. This sequence information yielded important insights into how the creature uses nitrogen, fats, and silica. The diatom DNA sequencing project provided insight into how the diatom species *Thalassiosira pseudonana* prospers in the marine environment while it contributes to absorbing the major greenhouse gas CO_2, in amounts comparable to all the world's tropical rain forests combined.

This critical information enables a better understanding of the vital role that diatoms and other phytoplankton play in mediating global warming. The research was summarized in the October 1, 2004 issue of *Science*.

All together, these single-celled organisms, or diatoms, generate as much as 40% of the 50 billion to 55 billion tons of organic carbon produced each year in the sea, while also using carbon dioxide and producing oxygen. And they are an important food source for many other marine organisms.

Scientists would like to better understand how these organisms react to changes in sea temperatures, the amount of light penetrating the oceans, and nutrients. Oceanographers thought they understood how diatoms use nitrogen, but illustrating what can be gained with the organism's genome in hand, the project revealed something unexpected—that diatoms have a urea cycle. A urea cycle is a nitrogen waste pathway found in animals and has never before been seen in a photosynthetic eukaryote like a diatom. Nitrogen is crucial for diatom growth and is often in short supply in seawater, depending on ocean conditions. The genome work revealed that the diatom *Thalassiosira pseudonana* has the genes to produce urea-cycle enzymes that may help to reduce its dependence on nitrogen from the surrounding waters.

The genome work also shed additional light on how this diatom species uses fats, or lipids, which it is known to store in huge amounts. Learning the actual pathways diatoms use to metabolize their fats helps explain the ability of diatoms to withstand long periods with little sunlight—even to overwinter—and then start growing rapidly once they return to sunlight.



Green Algae

THE OCEAN'S PHOTOSYNTHETIC POWERHOUSE

Chlamydomonas reinhardtii is a single-celled chlorophyte, or green alga. Highly adaptable, it lives in many different environments throughout the world. Normally deriving energy from photosynthesis, *C. reinhardtii* can also thrive in total darkness if acetate is available as a carbon source.

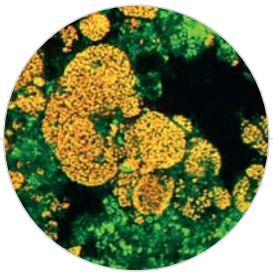
The relative adaptability and quick generation time has made *Chlamydomonas* an important model for biological research. Over the years, studies of *Chlamydomonas* have provided major research contributions in the areas of photosynthesis and molecular biology.

This green alga serves as the principal photosynthetic unicellular eukaryote model organism for genetic studies of photosynthesis and carbon assimilation.

By analyzing the genomes of several other microscopic ocean-dwelling organisms, JGI has provided new information into how the oceans affect its climate.

Comparative studies of four types of cyanobacteria—"photosynthetic" microbes that derive energy from sunlight, just like plants—were published in 2002 by JGI researchers in the journals *Nature* and *Proceedings of the National Academy of Sciences* (PNAS). Three of the microbes—two strains of *Prochlorococcus* and one of *Synechococcus*—were among the first organisms to have their DNA sequenced at JGI in the late 1990s, and are the first ocean bacteria to be sequenced.

Cyanobacteria are important, in part, because of their ability to turn sunlight and carbon into organic material. As the smallest yet most abundant photosynthetic organisms in the oceans, cyanobacteria play a critical role in regulating atmospheric carbon dioxide, a chief contributor to global climate change. Scientists estimate that *Prochlorococcus* and *Synechococcus* remove about 10 billion tons of carbon from the air each year—as much as two-thirds of the total carbon fixation that occurs in the oceans.



Metagenomics

DNA SEQUENCE ADDRESSING ENVIRONMENTAL CHALLENGES

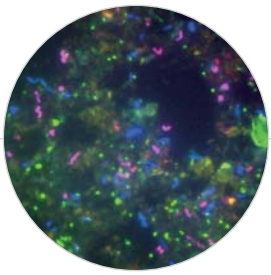
Microbes, the oldest and most abundant form of life on Earth, inhabit nearly every environment and can thrive under extreme conditions of heat, cold, pressure, and radiation. Although microbes represent the vast majority of life on the planet, more than 99% have not been cultured, and consequently their genomic diversity has been largely unrecognized and unutilized. While traditionally, pure cultures of microbes served as the substrate for genome sequencing projects, most microbes in the environment are not amenable to this approach. Accordingly, a new method that exploited the decreasing cost of sequencing, and increasing computational capabilities, was pioneered at the JGI. In this strategy, DNA isolated from environmental samples containing numerous organisms was sequenced as a mixture. Since all the genomes present in an environment are analyzed simultaneously, the approached has been dubbed, "metagenomics."

By studying the genomes of microbial populations of various environmental niches, scientists hope to find ways to use microbes to develop new pharmaceutical and agricultural products, energy sources, industrial processes, and solutions to a variety of environmental problems.

Genomics has much to offer the world of ecology. To expedite the application of genomics to ecology, in 2004 JGI established programs in microbial ecology and microbial genome analysis.

The goal is to enable scientists to better understand how organisms interact with their environment, with an initial emphasis on microbial activity in extreme and degraded environments—from hot springs to acid mine drainage and sewage.

Metagenomics frequently represents a synergistic relationship between the disciplines of microbial ecology, hydrology, geology, geochemistry, process engineering, and bioinformatics. Besides gaining insight into the workings of cellular machinery, through metagenomics researchers hope to discover how microbial processes can be used to clean up contaminated soil or water, and be harnessed for industrial applications.



Acid Mine Drainage

RECONSTRUCTING A TOXIC SITE

An Environmental Protection Agency superfund site, which for a century has been churning out toxic levels of sulfuric acid, may some day gain relief from the new strategy of metagenomics. JGI scientists teamed with collaborators from the University of California, Berkeley, to characterize the genomes of organisms from the site's acid mine drainage (AMD) microbial community. The work, published in the February 1, 2004 issue of the journal *Nature*, represents the first sequencing of a microbial community directly from the environment.

Decades of earthmoving activity at the Richmond Mine, located outside the northern California town of Iron Mountain, created opportunities for microbes to colonize the site. The complex interaction of microbes, water, and exposed iron ore has generated dangerously high levels of sulfuric acid and toxic heavy metals that ultimately find their way into the upper Sacramento River ecosystem. From this work, it became clear that the microbial community thriving in the water of the mine was responsible for the production of sulfuric acid.

The microbial community in the mine plays an integral role in the ecosystem. Examination of individual members of this community has been difficult because the vast majority of microorganisms resist cultivation in the laboratory, so their actual function within the environment may remain cryptic. By applying the shotgun sequencing strategy to the problem, information about these communities, and their interactions with their environment, is surfacing.

This work has provided a fascinating window into not only the diversity of the species involved, but their interdependency. This technique of ecological community genome reconstruction holds great promise for the study of the vast majority of microbes that are not culturable. Besides offering important clues to metabolic networks and community structure, such reconstructions may reveal the conditions that would make culturing possible.

Sudden Oak Death

GENOMICS TAKES TO THE HILLS

Researchers are closer now to thwarting two related plant pathogens, one causing "Sudden Oak Death" (SOD) and another responsible for a devastating soybean disease, thanks to the DNA sequence. JGI, in collaboration with the Virginia Bioinformatics Institute (VBI), with support from the U.S. Department of Agriculture (USDA), the National Science Foundation (NSF), and the DOE, conducted this multi-agency effort.

The genome sequences for these two *Phytophthora* [pronounced Fy-TOFF-thor-uh] species provide a framework for understanding how these plant pathogens cause disease and what can be done to control them. The aptly named genus *Phytophthora* derives its moniker from the Greek words for "plant destroyer." Part of a fungus-like group of organisms known as oomycetes, or water molds, they are relatives of such aquatic algae as diatoms and kelp. The pathogens survive as thick-walled spores that can persist in soil for years. Of the 59 recognized *Phytophthora* species, it was *P. infestans* that was responsible for the mid-19th century Irish potato famine.

Phytophthora species attack a wide variety of plants, including agricultural crops as well as trees and shrubs of native ecosystems. *Phytophthora sojae* attacks primarily soybeans, and *Phytophthora ramorum* is the pathogen causing Sudden Oak Death. This sequencing project, completed in 2004, will advance the battle against these devastating diseases and enable the identification of cellular processes that can be targeted for novel detection systems and for safe and effective means of chemical or biological control.

Sudden Oak Death was first reported in 1995, but the agent responsible for the disease was discovered by University of California scientists in 2000. The pathogen is known to be present in more than a dozen California counties and also in Southern Oregon. It has also been detected at scores of nurseries across the nation, elevating concerns about the pathogen to an all-time high.

The economic impact of *Phytophthora sojae* has been far-reaching. The U.S. produces almost half the world's soybeans. Losses attributed to *P. sojae* infestation, known as *Phytophthora* root rot of soybean, a post-emergence disease of the field, exceeded $1 billion in 2003.

Poplar: The First Tree Sequenced

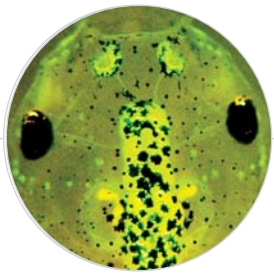
ADVANCING ALTERNATIVE ENERGY SOURCES

Forest trees are the dominant life form in many ecosystems and contain more than 90% of the Earth's terrestrial biomass. Forests throughout the world provide such environmental benefits as carbon sequestration, renewable energy supplies, watershed protection, improved air quality, biodiversity, habitat for endangered species, and access to recreational opportunities.

Despite the importance of forest trees for natural ecosystems and the world economy, little is known about their biology in comparison with the detailed information available for crop plants and model organisms such as *Arabidopsis.* As a result, the forest science community has engaged the resources of the JGI to help sequence *Populus trichocarpa*, the black cottonwood.

Traditional genetic breeding approaches in forestry are limited by the large size, long generation interval, and outcrossing mating system of most trees. Sequence information will enable forest tree biologists to begin large-scale, thorough analyses of genes and other genetic motifs. This will not only shed light on basic science questions, but will also lead to improved plant materials for the forest products industry and ultimately allow selection of novel traits that could be used to address questions related to the energy-related mission of the Department of Energy. *Populus* (poplar) species are used in activities ranging from carbon sequestration research, to free-air CO_2 enrichment studies, and to the development of fast-growing trees as a renewable bioenergy resource. The sequencing effort will also inform applications of phytoremediation, where trees can be used to remediate hazardous waste sites.

Efforts to sequence and study the poplar are being led by JGI scientists and involve a large team of scientists from Oak Ridge National Laboratory, Genome Canada, and the Umeå Plant Science Centre in Sweden.



Xenopus tropicalis

FROG AS ENVIRONMENTAL SENTINEL

The largest whole genome project tackled by the JGI to date is the western clawed frog *Xenopus tropicalis* (below, right). This sentinel species, a preferred system for the study of environmental toxicology by the Environment Protection Agency, is slated to be completed in 2005.

Xenopus tropicalis is a unique resource for two critical areas in vertebrate biology: early embryonic development and cell biology. In the former, a close relative, *Xenopus laevis* has led the way as a major model organism for developmental biology, helping to identify the mechanisms of early fate decisions, patterning of the basic vertebrate body plan, and early organogenesis.

Contributions in cell biology and biochemistry include seminal work on chromosome replication, chromatin and nuclear assembly, control of the cell-cycle components, in-vitro reconstruction of cytoskeletal element dynamics, and signaling pathways. In fact, *Xenopus* has become a major vertebrate model for the cellular and developmental biology research that is supported by most of the institutes of the National Institutes of Health.

Sequence-Based Science at JGI

From DOE's extensive experience running user facilities, it has become clear that the presence of an in-house science program provides an essential impetus for ensuring the highest quality of services in support of the greater user community. Accordingly, there are several sequence-based science programs operating at the PGF. The following pages provide a sampling of some of the higher profile projects.

The Noncoding DNA Conundrum

Can you lose scores of pages from a novel and still follow the story line? In the case of the mouse genome—and perhaps even our own—the answer appears to be "yes." JGI researchers demonstrated that, after deleting large swaths of DNA sequence shared by mice and humans, they were still able to generate mice that suffered no apparent ills from their genomes being millions of letters lighter. These research findings were published in the October 21, 2004 edition of the journal *Nature*.

After completing the sequencing of the human genome, a question still lingers: is all the noncoding DNA (sometimes called "junk DNA")—which makes up nearly 98% of the genome—required, or is some of it potentially disposable?

In these studies, scientists were looking for sequences that might not be essential, but were surprised, given the magnitude of the information being deleted from the genome, by the complete lack of impact noted. From these results, researchers concluded that some noncoding sequences have minimal function. To see what these noncoding sequences were doing, the investigators took a brute-force approach. To use an engineering analogy, they asked which walls in the room actually support the ceiling above. By removing the walls, the truth was soon revealed.

Through molecular techniques, a total of 2.3 million letters of DNA code from the 2.7-billion-base-pair mouse genome were deleted. To do this, embryonic cells were genetically engineered to contain the newly compact mouse genome. Mice were subsequently generated from these stem cells. The research team then compared the resulting mice with the abridged genome to mice with the full-length version. A variety of features were analyzed, ranging from viability, growth, and longevity to numerous other biochemical and molecular features. Despite the researchers' efforts to detect differences in the mice with the abridged genome, none were found. However, the ability to test for a particular characteristic in mice is currently limited.

The negligible impact of removing these sequences suggests that the mammalian genome may not be densely encoded. Similar-sized regions have previously been removed from the mouse genome, invariably resulting in mice that did not survive, because the missing sequences contained important genes and their deletion had severe consequences for the animal.

Fingerprinting Environmental Communities

In a study published in the April 22, 2004 edition of the journal *Science*, JGI scientists demonstrated for the first time that the signatures of the genes alone in terrestrial and aquatic samples can accurately diagnose the health of the sampled environments. This work positions large-scale genome sequencing to accelerate advances in environmental sciences akin to the contributions DNA sequencing has made to biomedical sciences.

These DNA sequence fingerprints, or Environmental Genomic Tags (EGTs), can be used to provide highly accurate assessments of the vitality of extremely diverse environments and can reveal environments under stress as well as signal progress in remediating contaminated environments. EGTs capture the DNA profile of a particular niche and reflect the presence and levels of nutrients, pollutants, and other environmental features. The EGT approach employed in the study shares similarities with aspects of the Human Genome Project research. In the early 1990s, incomplete fragments of human genes called Expressed Sequence Tags (ESTs) were used as diagnostic fingerprints for human tissues to determine their unique features and disease status. These information-rich data allowed researchers to forge ahead with studying genes important in disease processes, long before the completion of the entire human genome.

With EGTs, researchers do not need a complete genome's worth of data to understand the functions required of the organisms living in a particular setting. Rather, the quantity of genes present in the EGT data reflect the demands of the setting to can inform what is happening in an environment without knowing the actual identities of the microbes living there.

By using EGTs, researchers illustrated how genes involved in breaking down plant material are overrepresented in soil and absent in sea water, while in sea water, genes involved in the passage of sodium, a major chemical component of salt water, were particularly abundant. As light is a major energy source for microbes living in surface water, there was an abundance of genes involved in photosynthesis in samples collected from shallow water. These differences in the abundances of genes involved in particular functions provide DNA clues to features of the environments from where the samples were taken. Importantly the DNA clues were easy to find despite the vast numbers of different individual microbial species within the samples.

Recovering Ancient DNA Samples

THE CAVE BEAR AS PROOF-OF-PRINCIPLE FOR CHARACTERIZING HUMAN PREDECESSORS

Genomic DNA of an extinct Pleistocene cave bear species—the kind of stuff once reserved for science fiction—has been captured, sequenced, and logged into scientific literature by JGI investigators. The study, published in the June 2, 2005 online edition of the journal *Science*, has set the research community's sights on traveling back in time through the vehicle of DNA sequencing to reveal the story of other extinct species, including our nearest relatives, the Neandertals.

Until now, researchers have been stymied in attempts to sequence genomes of extinct species. The JGI scientists overcame many of the difficulties normally associated with recovery of DNA from ancient samples. DNA starts degrading at death, while microbes attack the decaying carcass to utilize the nutrients present in the dead organism as an energy source. What remains and confounds the efforts to sequence and characterize these artifacts is an overabundance of microbial contaminants, along with the occasional DNA fingerprints contributed unwittingly by the modern fossil hunters or lab workers.

JGI applied the standard techniques normally used for modern genome projects to ancient DNA specimens, a brute force high-throughput sequencing approach. With this strategy, researchers weren't terribly concerned that much of what they were sequencing were microbial contaminants, but hoped that among the large amounts of generated sequence would be some of the ancient DNA they were really interested in. They were looking for the proverbial needle in the haystack and it worked. Among the expected lion's share of contaminants, they recovered reasonable amounts of 40,000-year-old cave bear DNA, and applied the powerful magnet of industrial-strength computing to tease out the interesting data from a hodgepodge of different DNAs.

It turned out that about 6% of the sequence from the sample yielded cave bear sequence—the rest represented a mosaic of microbial contaminants. Nevertheless, within that fraction, there was a range of genomic sequence types, including fragments of 21 genes, identified by comparing the cave bear sample to the complete dog genome sequence that exists in the public databases. Dogs and bears, which diverged some 50 million years ago, are 92% similar on the sequence level.

The samples of cave bear bones and teeth from the study were collected from two cave locations in Austria. Extinct for more than 10,000 years, these particular cave bears, *Ursus spelaeus*, whose remains are found in abundance, were related to the ancestors of modern brown bears and polar bears. Fossil and cave-painting evidence supports that ancient humans interacted with the cave bears.

Currently, the theoretical age limit is about 100,000 years for sequencing viable DNA from samples preserved in the same conditions in which the cave bear specimens were found. Barring some fundamental misunderstanding about the nature of DNA decay, it is highly unlikely that viable DNA will be recovered from 150- to 200-million-year-old Jurassic age samples.

The cave bear was used as an ancient DNA test case because the samples used in the study were roughly the same age as the extinct Neandertal—which will soon be pursued for comparative studies with the data gleaned from the Human Genome Project.

For more information about current scientific research at JGI, visit our Web site at http://www.jgi.doe.gov/science/index.html.

Dogs and bears, which diverged some 50 million years ago, are 92% similar on the sequence level. Such comparative analyses help shed light on previously uncharacterized genomes.

Jamborees—Bringing the Scientific Community Together

JGI has advanced the concept of the genome annotation jamboree. These jamborees are working meetings at which members of a scientific community gather to discuss and annotate the genome of an organism of common interest. The focus can be a single organism or a family of organisms. Similar creatures are also typically studied in order to draw comparisons and contrasts. The goal is to identify genes and generate high-quality annotations.

Since 2002, JGI has hosted seven such jamborees, including those to annotate *Ciona*, lactic acid bacteria; the diatom, *Chlamydomonas*; *Phytophthora sojae* and *P. ramorum* (sources of soybean "damping off" disease and sudden oak death); and one for the poplar tree.

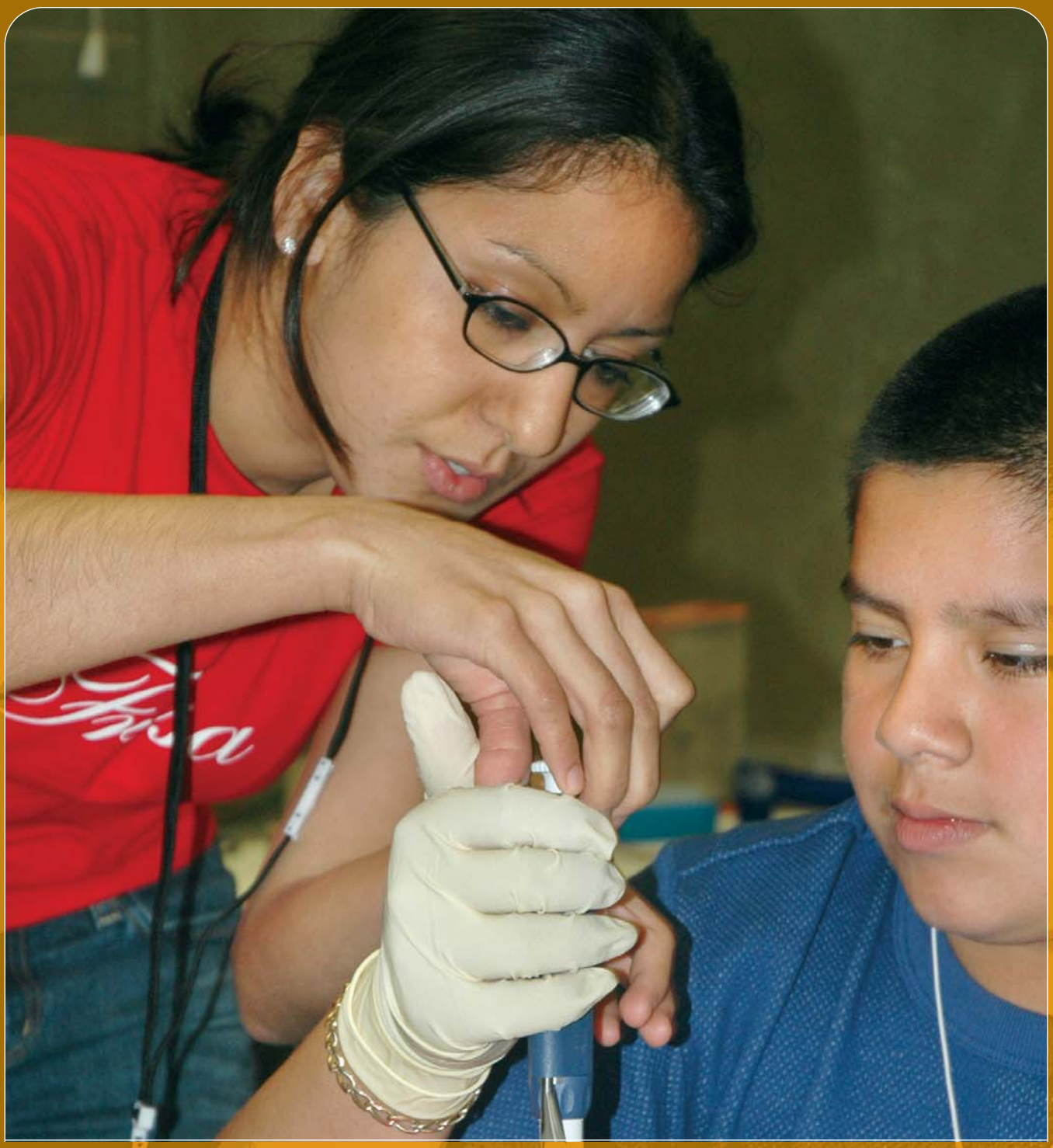
As an example of how JGI jamborees contribute to advancing important science, in April 2004, on behalf of the DOE GTL: Genomics Program, JGI hosted a sulfate-reducing bacteria (SRB) jamboree. Some 40 microbiologists, biochemists, and computational scientists representing a dozen different labs, charted the unseen metabolic processes of the stinky "bug" that goes by the name of *Desulfovibrio desulfuricans* G20. Not a bug per se, G20 is a microbe with a robust appetite for such toxic metals as uranium and chromium, from a family known as sulfate-reducing bacteria, or SRB. Isolated from a corroded oil well in the late 1980s, the G20 strain was sequenced by JGI.

Sulfate-reducing bacteria are conspicuous because of the product of their respiration, hydrogen sulfide (H2S), smells of rotten eggs. Producing this extremely reactive gas—toxic to plants and animals—these bacteria thrive in anaerobic conditions of deep marine sediments, where oil can be found. Another idiosyncrasy—one that researchers are seeking to tap—is their generosity in donating electrons, a process known as reduction. In association with organics in the soil, SRB convert sulfate to sulfide—hence the big stink.

DOE's interest in these microbes hinges on their ability to reduce uranium from its highly oxidized state. This reaction detoxifies the metal so that it is not as harmful to humans while making it insoluble in water, effectively immobilizing it in the soil. There are significant economic and environmental pressures compelling the study of SRB. The oil industry wants to contain these bacteria because they cause souring, or production of H2S. The problem is exacerbated when drillers inject seawater, rich in sulfur, into the wells to force out the oil. Useful for liberating reserves, this practice degrades oil quality and stimulates the bugs to corrode pipes, all revealed in the microbial physiology of down-hole conditions.

As oil and gas become more depleted, there is increasing pressure to consider alternatives, such as harnessing photosynthesis to convert light into hydrogen. As a biological pathway, *Desulfovibrio* can serve as a workhorse organism. The more is known about how it generates energy, transports things, repairs DNA, and responds to environmental assaults, the more researchers can make reasonable predictions about such mechanisms as horizontal gene transfer—believed to be essential for genetic plasticity, the ability of bacteria to adapt to new environments.

For more information about the JGI Jamborees, visit our Web site at http://www.jgi.doe.gov/jamborees/index.html.



Education, Outreach, & Diversity Efforts

Education, Outreach, & Diversity Efforts

The JGI Production Genomics Facility is a magnet for educational and community partnerships. There is a strong interest on behalf of the PGF's neighbors, from high school students to retired citizens, in learning more about the power of genomics and the relevance it has for their health and that of the environment. Given the opportunity, visitors delight at witnessing industrial-scale DNA sequencing in action. In addition, JGI researchers and administrators actively reach out to the community in such opportunities as judging local science fairs and addressing Rotary, Lions, and other community service organizations. For those unable to visit in person, the JGI Web site (http://www.jgi.doe.gov/index.html) offers detailed descriptions of the sequencing process and its applications. Highlights of JGI's education, outreach, and diversity efforts include:

- During the 2004-2005 school year alone, the JGI Production Genomics Facility hosted over 40 tours, totaling some 800 visitors, mostly high school and community college students. Of those, approximately 25% of the visitors were underrepresented minorities (that is, African American and Hispanic).
- In 2004 and 2005, JGI participated in the Contra Costa Summer Biotech Summer Camp, a weeklong set of activities for 60 students, 30% of whom were underrepresented minorities.
- JGI participated in the October 2, 2004 Tri-Valley Expanding Your Horizons in Science and Mathematics Conference (a conference for 6th–12th grade youth) at the University of the Pacific in Stockton, CA. Over 200 students experienced a three-part education module including gel electrophoresis, DNA structure model assembly, and DNA sequence simulation.
- JGI participated in the October 21, 2004 Department of Energy "What's Next," which included a hands-on DNA extraction activity serving over 200 Chicagoland youth, of which over 50% were underrepresented minorities.
- JGI organized an "All About DNA Day," where eighty seventh-grade students from Concord, California's Glenbrook Middle School received an emergence course in genome science.

The event was organized into concurrent sessions offering hands-on exercises in:

› Extracting DNA From Fruit

› Building a Giant DNA Model

› DNA Gel Electrophoresis

› Identifying Mystery "Critters" by DNA Sequence

› Predicting Susceptibility to Sickle Cell Anemia

› Making a Transgenic Mouse

The event was covered by the Contra Costa Times, and the article is on their Web site at http://www.contracostatimes.com/mld/cctimes/email/news/11199375.htm

- JGI developed a diversity Web site (http://www.lbl.gov/Workplace/GD-LS-Diversity/) with the dual purpose of attracting under-represented minorities into all levels of job classifications and enabling researchers to more effectively tap into the existing pool of candidates through various mechanisms.

Appendices

APPENDIX A: Genomics Glossary

ANNOTATION: The process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do.

ARCHAEA: One of the three domains of life (Eukaryotes and Bacteria being the others) that subsume primitive microorganisms that can tolerate extreme (temperature, acid, etc.) environmental conditions.

ASSEMBLY: Compilation of overlapping sequence from one or more related genes that have been clustered together based on their degree of sequence identity or similarity.

BAC: (Bacterial Artificial Chromosome) An artificially-created chromosome in which medium-sized segments of foreign DNA are cloned into bacteria. Once the foreign DNA has been cloned into the bacteria's chromosome, many copies of it can be made and sequenced.

BASE: A unit of the DNA. There are 4 bases: adenine (A), guanine (G), thymine (T), and cytosine (C). The sequence of bases is the genetic code.

BASE PAIR: Two DNA bases complementary to one another (A and T or G and C) that join the complementary strands of DNA to form the double helix characteristic of DNA.

CLONING: Using specialized DNA technology to produce multiple, exact copies of a single gene or other segment of DNA to obtain enough material for further study.

CONTIG: Group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome.

COVERAGE: The amount of times a fragment or the entire genome has been sequenced.

COSMID: A cosmid is an artificially constructed structure made from bacterial DNA into which is spliced a small fragment of a genome to be amplified and sequenced and used in cloning (copying) pieces of target DNA.

DRAFT GENOME SEQUENCE: An automated assembly resulting from alignment of redundant sequencing reads produced from whole-genome shotgun libraries. Aligned reads are used to generate a consensus sequence and form contigs (contiguous sequence) that do not contain gaps. Scaffolds are groups of contigs that are ordered and oriented relative to one another based on subclone paired-end linking information. As this sequence is generated automatically, it may contain sequence errors and misassemblies.

ELECTROPHORESIS: A process by which molecules (such as proteins, DNA, or RNA fragments) can be separated according to size and electrical charge by applying an electric current to them. Each kind of molecule travels through the medium at a different rate, depending on its electrical charge and molecular size.

EUKARYOTES: The domain of life of organisms that consist of one or more cells, each with a nucleus and other well-developed intracellular compartments. Eukaryotes include all animals, plants, and fungi.

FINISHED GENOME SEQUENCE: Draft sequence is finished by closing or sequencing all gaps in a draft sequence, while also improving low-quality regions and misassemblies. The final sequence should contain no more than one base error in 10,000.

FOSMID: A bacterial cloning vector suitable for cloning genomic inserts approximately 40 kilobases in size.

LIBRARY: An unordered collection of clones (i.e., cloned DNA from a particular organism), whose relationship to each other can be established by physical mapping.

MAPPING: Charting the location of genes on chromosomes.

METAGENOMICS: (also Environmental Genomics or Community Genomics) is the study of genomes recovered from environmental samples. This relatively new field of genetic research allows the genomic study of organisms that are not easily cultured in a laboratory.

PASS RATE: The efficiency at which viable information is derived from each of the 384 sample wells in a microtiter plate.

PCR: Acronym for Polymerase Chain Reaction, a method of DNA amplification.

PHYLOGENY: The evolutionary history of a molecule such as gene or protein, or a species.

PLASMID: Autonomously replicating, extrachromosomal, circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as cloning vectors.

POLYMERASE: Enzyme that copies RNA or DNA. RNA polymerase uses preexisting nucleic acid templates and assembles the RNA from ribonucleotides. DNA polymerase uses preexisting nucleic acid templates and assembles the DNA from deoxyribonucleotides.

PROKARYOTES: Unlike Eukaryotes, these organisms, (e.g., bacteria) are characterized by the absence of a nuclear membrane and by DNA that is not organized into chromosomes.

RCA: Acronym for Rolling Circle Amplification, a randomly primed method of making multiple copies of DNA fragments, which employs a proprietary polymerase enzyme, where the amplified DNA does not need to be purified before being added to the sequencing reaction.

READ LENGTH: The amount of nucleotides tabulated by the DNA analyzer per DNA reaction well.

SEQUENCE: Order of nucleotides (base sequence) in a DNA molecule. In the case of DNA sequence, it is the precise ordering of the bases (A, T, G, C) from which the DNA is composed.

SUBCLONING: The process of transferring a cloned DNA fragment from one vector to another.

TRANSFORMATION: A process by which the genetic material carried by an individual cell is altered by the incorporation of foreign DNA into its genome.

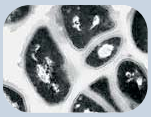
VECTOR: DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where it can be reproduced in large quantities. Examples are plasmids, cosmids, Bacterial Artificial Chromosomes (BACs), or Yeast Artificial Chromosomes (YACs).

WHOLE GENOME SHOTGUN: Semi-automated technique for sequencing long DNA strands in which DNA is randomly fragmented and sequenced in pieces that are later reconstructed by a computer.

APPENDIX B: 2005 & 2006 CSP DNA Sequencing Projects

CSP 2005

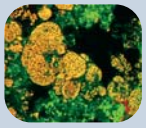
The complex bacterial community that lives under the skin of the gutless marine worm, *Olavius algarvensis*, and provide the energy source for the host.



Staphylococcus aureus, a food-borne pathogen that is implicated in thousands of infections in the U.S. alone. Sequencing of an antibiotic resistant form of this organism will inform how antibiotic resistance occurs.

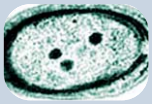
Sequencing of groundwater samples from contaminated sites within the Oak Ridge National Laboratory Y12 Security Complex. The site contains one of the highest-concentration plumes of mobile uranium along with volatile organics, technetium, nitrate, aluminum, thorium, zinc, and nickel. Sequence generated will complement biogeochemistry, hydrology, microbiology, and engineering studies to help evaluate the impacts of contaminants and remediation treatments on microbial community dynamics.

Karenia brevis, a single-celled alga responsible for the natural saltwater phenomenon known as "red tide," which can pose a human health risk and detrimentally affect regional marine economies.



Accumulibacter phosphatis stores huge amounts of phosphorus inside its cell. Engineers have exploited this capability to remove excess phosphorus—initially from fertilizers and detergents—from wastewater. Too much phosphorous can over-stimulate microbial growth, resulting in eutrophication, where oxygen in the water is depleted and fish and other organisms residing in the habitat die.

Alvinella pompejana, the marine Pompeii worm, which has adapted to thrive at super-hot hydrothermal vents.



Prochlorococcus, a marine phytoplankton that plays a critical role in regulating the dynamics of the global carbon cycle, responsible for a significant fraction of photosynthesis in the world's oceans.

Lactobacillus rhamnosus is considered a potential probiotic, which can protect its host and prevent disease.

CSP 2006

A metagenomic community of waste-degrading bacteria capable of treating industrial streams contaminated with terephthalate, a major byproduct of plastics manufacturing.

Arabidopsis lyrata and *Capsella rubella*, two mustard relatives whose sequence will shed light on the genetics, physiology, development, and structure of plants in general and how they respond to disease and environmental stresses.

A community of *Korarchaeota*, a group of Archaea, the least well characterized of the three domains of life, obtained from Obsidian Pool hot spring in Yellowstone National Park.

Six members of the Crenarchaeota group of Archaea, including *Methanocorpusculum labreanum*, isolated from surface sediments of La Brea Tar Pits in Los Angeles, which present features allowing proteins to function at extremes of temperature, acid, and salinity.

A powerful fungal pathogen—*Mycosphae-rella fijiensis*—cause of black Sigatoka—currently regarded as one of the most serious threats to world banana production.

Mytilus californianus, the edible pacific mussel that is a sentinel species for environmental pollution.

Triphysaria versicolor, a parasitic plant that releases chemicals into the soil that affect the growth and development of nearby plants, a phenomenon known as allelopathy, which could be used to control unwanted vegetation.

The soil-dwelling fungal microorganism *Trichoderma virens* that also has promise for biological weed control.

Petrolisthes cinctipes, the porcelain crab, whose heat and cold tolerance will help inform climate change research.

Bicyclus anynana, a butterfly whose sequence encodes wing patterns that should reveal key issues in evolutionary-developmental biology, and provide information that will bolster efforts to understand biological diversity.

APPENDIX B: 2005 & 2006 CSP DNA Sequencing Projects *cont.*

	ORGANISM		COLLABORATOR	INSTITUTION
	2005			
MICROBES	*Olavius*	*algarvensis*	Dubilier	Max Planck Inst. of Marine Microbiology
	Crenarchaeota		Delong	Massachusetts Inst. of Technology
	Marinobacter	aquaeolei	Edwards	Woods Hole Oceanographic Inst.
	Staphylococcous	*aureus* VISA	Tomasz	Rockefeller Univ.
	Prochlorococcus		Chisholm	Massachusetts Inst. of Technology
	Rhodocyclus-like	polyphosphate	Hugenholtz	Joint Genome Inst.
	Rhodobacter		Kaplan	Univ. of Texas, Houston
	Contaminated	groundwater	Zhou	Oak Ridge National Laboratory
	Lactobacillus	*reuteri* (2 strains)	Tannock	Univ. of Otago, Dunedin, NZ
	Bacillus	*cereus*	Sorokin	INRA, France
	Synechococcus	2 strains	Palenik	Scripps Inst. of Oceanography
	Syntrophobacter	*fumaroxidans*	McInerney	Univ. of Oklahoma
	Thermoanaerobacter	*ethanolicus*	Fields	Miami Univ.
	Thiomicrospira	2 strains	Scott	Univ. of South Florida
BASAL ORGANISMS	*Selaginella*	*moellendorffii*	Banks	Purdue Univ.
	Trichoplax	*adhaerens*	DellaPorta	Yale Univ.
	Sporobolomyces	*roseus*	Wolfe	Trinity College, Dublin
	Reniera		Degnan	Univ. of Queensland, Australia
	Mycosphaerella	2 sp.	Goodwin	Purdue Univ.
	Spironucleus	*vortens*	Cande	Univ. of California, Berkeley
	Naegleria	*gruberi*	Cande	Univ. of California, Berkeley
HIGHER ANIMALS & PLANTS	*Physcomitrella*	*patens*	Mishler	Univ. of California, Berkeley
	Lottia	*gigantea* (limpet)	Edsinger-Gonzalez	Univ. of California, Berkeley
	Helobdella	*robusta* (leech)	Weisblat	Univ. of California, Berkeley
	Capitella	*capitata*	Cande	Univ. of California, Berkeley
ESTs & TARGETED SEQUENCING	*Alvinella*	*pompejana*	Tainer	Scripps Research Inst.
	Mitochondria	seed plant	Palmer	Indiana Univ., Bloomington
	Karenia	*brevis*	Bhattacharya	Univ. of Iowa
	Dipteran	fosmids	Eisen	Lawrence Berkeley National Laboratory
	2006			
EUKARYOTES	*Sorghum*	*sp.*	Paterson	Univ. of Georgia
	Arabidopsis	*lyrata*	Weigel	Max Planck Inst. of Developmental Biology
	Mimulus	*guttatus*	Willis	Duke Univ.
	Piromyces	sp. E2	Baker	Pacific Northwest National Laboratory
	Hydractinia	*symbiolongicarpus*	Buss	Yale Univ.
	Phycomyces	*blakesleeanus*	Corrochano	Univ. of Seville
	Xanthoria	*parietina*	Crittendon	Univ. of Nottingham
	Trichoderma	*virens*	Ebbole	Texas A&M Univ.
	Mycosphaerella	*fijiensis*	Goodwin	US Dept. of Agriculture—ARS, Purdue Univ.
	Mytilus	*californianus*	Gracey	Stanford Univ.
	Phytophthora	*capsici*	Kingsmore	National Center for Genome Resources
	Campanulales		Knox	Indiana Univ.
	Cichlid	Lake Malawi	Kocher	Univ. of New Hampshire
	Ciona	*intestinales*	Lemaire	CNRS, France
	Ictalurus	*punctatus*	Liu-J	Auburn Univ.
	Ictalurus	*furcatus*	Liu-J	Auburn Univ.
	Bicyclus	*anynana*	Long	Univ. of California, Irvine
	Melampsora	*larici-populina*	Martin	Inst. National de la Recherche Agronomique
	Ostreococcus	low-light strain	Palenik	Univ. of California, San Diego
	Parhyale	*hawaiensis*	Patel	Univ. of California, Berkeley
	Jassa	*slatteryi*	Patel	Univ. of California, Berkeley
	Petrolisthes	*cinctipes*	Stillman	Univ. of Hawaii
	Batrachochytrium	*dendrobatidis*	Taylor	Univ. of California, Berkeley
	Triphysaria		Yoder	Univ. of California, Davis

APPENDIX B: 2005 & 2006 CSP DNA Sequencing Projects *cont.*

	ORGANISM		COLLABORATOR	INSTITUTION
	2006 *cont.*			
BACTERIA AND ARCHAEA	*Euryarchaeota*	community	Baker	Univ. of California, Berkeley
	Polynucleobacter		Hahn	Inst. for Limnology, Austria
	Microbial soil	community Alaska	Handelsman	Univ. of Wisconsin, Madison
	Salinospora	*tropicalis*	Jensen	Scripps Inst. of Oceanography
	Salinospora	*arenicola*	Jensen	Scripps Inst. of Oceanography
	Termite gut microbial	community	Leadbetter	California Inst. of Technology
	Terephthalate (TA)	community	Liu	National Univ. of Singapore
	Archaeal	sp. hyperthermo.	Lowe	Univ. of California, Santa Cruz
	Bacterioplankton	Antarctic marine	Murray	Desert Research Inst.
	Thermotogales	7 hyperthermo.	Noll	Univ. of Connecticut
	Nitrosomonas		Norton	Utah State Univ.
	Microbial mats	hypersaline	Pace	Univ. of Colorado
	Sinorhizobium	*medicae*	Reeve	Murdoch Univ.
	Verrucomicrobium		Schmidt	Michigan State Univ.
	Bacillus	*coagulans*	Shanmugam	Univ. of Florida
	Crenarchaeote	community	Simon	Oregon Health & Science Univ.
	Verrucomicrobia	5 strains	Smidt	Wageningen Univ.
	Acidovorax	sp.	Stahl	Univ. of Washington
	Caulobacter	2 strains	Stephens	Santa Clara Univ.
	Korarchaeota	community	Stetter	Univ. Regensberg, Diversa Corp.
	Archaea	6 strains	Woese	Univ. of Illinois, Urbana-Champaign

APPENDIX C: DOE Microbial Genome Support Projects

GENUS	SPECIES	COLLABORATOR	INSTITUTION
2002			
Dechloromonas	*aromatica*	Coates	Univ. of California, Berkeley
Desulfuromonas	*acetoxidans*	Lovely	Univ. of Massachusetts
Ehrlichia	2 strains	McBride	Univ. of Texas, Medical Branch
Geobacter	*metallireducens*	Lovely	Univ. of Massachusetts
Methanococcoides	*burtonii*	Sowers	Univ. of Maryland
Pseudomonas	*syringae*	Lindow	Univ. of California, Berkeley
Psychrobacter	sp.	Tiedje	Michigan State Univ.
Ralstonia	*eutropha*	Gonzalez	Pontificia Univ. Catolica de Chile
Streptococcus	*suis*	Gottschalk	Univ. of Montreal, Canada
2003			
Burkholderia	2 strains	Tiedje	Michigan State Univ.
Methylobium	*petroleophilum*	Kane	Lawrence Livermore National Laboratory
Prochlorococcus	sp.	Chisholm	Massachusetts Institute of Technology
Synechococcus	*elongatus*	Golden	Texas A&M Univ.
2004			
Burkholderia	2 strains	Tiedje	Michigan State Univ.
Clostridium	*phytofermentans*	Leschine	Univ. of Massachusetts, Amherst
Frankia	sp.	Tisa	Univ. of New Hampshire
Nitrobacter	*hamburgensis*	Arp	Oregon State Univ.
Nitrobacter	*winogradskyi*	Arp	Oregon State Univ.
Nitrosococcus	*oceani*	Arp	Oregon State Univ.
Nitrosomonas	*eutropha*	Arp	Oregon State Univ.
Nitrosospira	*multiformis*	Arp	Oregon State Univ.
Prochlorococcus	sp.	Chisholm	Massachusetts Institute of Technology
Shewanella	2 strains	Fredrickson	Pacific Northwest National Laboratory
Synechococcus	2 strains	Palenik	Scripps Institution of Oceanography
Syntrophobacter	*fumaroxidans*	McInerney	Univ. of Oklahoma
Thermoanaerobacter	*ethanolicus*	Fields	Miami Univ.
Thiomicrospira	2 strains	Scott	Univ. of South Florida
2005			
Acidiphilium	*cryptum*	Magnuson	Idaho State Univ.
Acidobacterium	Ellin345	Kuske	Los Alamos National Laboratory
Acidothermus	*cellulolyticus*	Berry	Univ. of California, Davis
Actinobacillus	*succinogenes*	Vieille	Michigan State Univ.
Aspergillus	*niger*	Baker	Pacific Northwest National Laboratory
Aureococcus	*anophageggerens*	Gobler	Southhampton College of Long Island Univ.
Bacillus	*selenitireducens*	Stolz	Duquesne Univ.
Bradyrhizobium	sp.	Sadowsky	Univ. of Minnesota
Burkholderia	*ambifaria*	Tiedje	Michigan State Univ.
Caldicellulosiruptor	*saccharolyticus*	Kelly	North Carolina State Univ.
Calyptogena	*magnifica*	Cavanaugh	Harvard Univ.
Chloroflexus	*aggregans*	Bryant	Pennsylvania State Univ.
Chloronema	sp.	Bryant	Pennsylvania State Univ.
Chlorothrix	*halophila*	Bryant	Pennsylvania State Univ.
Clostridium	sp.	Stolz	Duquesne Univ.
Dehalococcoides	2 strains	Spormann	Stanford University
Desulfotomaculum	*reducens*	Tebo	Univ. of California, San Diego
Flavobacterium	*johnsoniae*	McBride	Univ. of Texas, Medical Branch
Geobacter	sp.	Kostka	Florida State Univ.
Halorhodospira	*halophila*	Hoff	Univ. of Chicago
Heliothrix	*oregonensis*	Bryant	Pennsylvania State Univ.
Herpetosiphon	*aurantiacus*	Bryant	Pennsylvania State Univ.

APPENDIX C: DOE Microbial Genome Support Projects *cont.*

GENUS	SPECIES	COLLABORATOR	INSTITUTION
2005 cont.			
Iron Mountain AMD	Site 1	Banfield	Univ. of California, Berkeley
Iron Mountain AMD	Site 2	Banfield	Univ. of California, Berkeley
Lake Washington	formaldehyde	Lidstrom	Univ. of Washington
Lake Washington	formate	Lidstrom	Univ. of Washington
Lake Washington	methane	Lidstrom	Univ. of Washington
Lake Washington	methanol	Lidstrom	Univ. of Washington
Lake Washington	methylamine	Lidstrom	Univ. of Washington
Methanosaeta	*thermophila*	Smith	Clemson Univ.
Micromonas	2 strains	Worden	Univ. of Miami
Mono Lake deltapro-teobacter		Stolz	Duquesne Univ.
Mono Lake gammapro-teobacter		Stolz	Duquesne Univ.
Mycobacterium	5 strains	Miller	Utah State Univ.
Nectria	*haematococca*	VanEtten	Univ. of Arizona
Obsidian Hot Spring		Mead	Lucigen
Phycovirus	11 strains	Wommack	Delaware Biotechnology Institute
Polaromonas	*napthalenivorans*	Madsen	Cornell Univ.
Postia	*placenta*	Cullen	U.S. Department of Agriculture
Pseudoalteromonas	*atlantica*	Karls	Univ. of Georgia
Pseudomonas	*putida*	Parales	Univ. of California, Davis
Psychromonas	*ingrahamii*	Staley	Univ. of Washington
Rhodopseudomonas	4 strains	Harwood	Univ. of Iowa
Roseiflexus	2 strains	Bryant	Pennsylvania State Univ.
Shewanella	7 strains	Fredrickson	Pacific Northwest National Laboratory
The Cedars Alkaline Springs		Nealson	Univ. of Southern California
Ultra back	C level 1	Banfield	Univ. of California, Berkeley
Xanthobacter	*autotrophicus*	Ensign	Los Alamos National Laboratory

JGI Publications 2002–2005

2002

Aparicio S, et al.
Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297 (5585): 1301–10 (2002).

Banerjee P, et al.
SNPs in putative regulatory regions identified by human mouse comparative sequencing and transcription factor binding site data. *Mammalian Genome* Oct 13 (10): 554–7 (2002).

Boore JL, et al.
The mitochondrial genome of the sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. *Mol. Biol. Evol.* 19 (2): 127–137 (2002).

Callow M, et al.
Expression profiling and comparative sequence derived insights into lipid metabolism. *Current Opinion in Lipidology* (13): 173–179 (2002).

Dehal P, et al.
The draft genome of *Ciona intestinalis:* insights into chordate and vertebrate origins. *Science* 298: 2157–2167 (2002).

Detter JC, et al.
Isothermal strand displacement amplification applications for high-throughput genomics. *Genomics* 80 (6): 691–698 (2002).

Doyle SA, et al.
High throughput Proteomics: A Flexible and Efficient Pipeline for Protein Production. *Journal of Proteome Research* 1 (6): 531–536 (2002).

Dubchak,I, et al.
The computational challenges of applying comparative-based computational methods to whole genomes. *Briefings in Bioinformatics* 3, 18 (2002).

Elkin C, et al.
Magnetic bead purification of labeled DNA fragments for high-throughput capillary electrophoresis sequencing. *Biotechniques* Jun 32 (6):1296, 1298–1300, 1302 (2002).

Francino MP, et al.
Phylogenetic relationships of bacteria with special reference to enteric species.
In *The Prokaryotes 3rd edition, a Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications* Dworkin M (ed.) Springer-Verlag, New York (2002).

Hawkins TL, et al.
Whole genome amplification – applications and advances. *Current Opinion in Biotechnology* 13 (1): 65–67 (2002).

Harafuji N, et al.
Genome-wide identification of tissue-specific enhancers in the *Ciona* tadpole. *Proceedings of the National Academy of Sciences* USA (99): 6802–6805 (2002).

Keys DN*, et al.
Control of intercalation is cell-autonomous in the notochord of *Ciona intestinalis. Developmental Biology* 246: 329–340 (2002).

Lavrov DV, et al.
Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: Duplication and non-random loss. *Mol. Biol. Evol.* 19 (2): 163–169 (2002).

Loots GG, et al.
rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research* 2255–0 (2002).

Lucas S, et al.
The US DOE Joint Genome Institute's High-Throughput Production Sequencing Program. DOE Genome Contractor-Grantee Workshop IX, Oakland, CA (2002).

Ochman H, et al.
In *The Encyclopedia of Molecular Medicine*. Creighton T (ed.) Academic Press (2002).

Olsen A, et al.
Assembly and Analysis of Finished Sequence for Human Chromosome 19. The US DOE Genome Contractor-Grantee Workshop IX, Oakland, CA (2002).

Pennacchio LA, et al.
Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. *Human Molecular Genetics* 11(24): 3031–38 (2002).

Schulte II JA, et al.
2002 Rostral horn evolution among agamid lizards of the genus Ceratophora endemic to Sri Lanka. *Molecular Phylogenetics and Evolution* 22: 111-117 (2002).

Talmud P J, et al.
Relative contribution of variation within the APOC3-A4-A5 gene cluster in determining plasma triglycerides. *Human Molecular Genetics* 11(24): 3039–46 (2002).

Waterston et al.
Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 (6915): 520–562 (2002).

Wei L, et al.
Comparative genomics approaches to study organism similarities and differences. *J Biomed Inform.* 35: 142–50 (2002).

2003

Azumi K, et al.
Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: "waiting for Godot". *Immunogenetics* 55(8): 570–81 (2003).

Boffelli D, et al.
Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* Feb 28 (299): 1391–1394 (2003).

Bray N, et al.
AVID: A Global Alignment Program. *Genome Res.* Jan 13 (1): 97–102 (2003).

Brudno M, et al.
Glocal alignment: finding rearrangements during alignment. *Bioinformatics* Jul 19 Suppl 1:I54–I62 (2003).

Chain P, et al.
Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *J Bacteriol.* May 185 (9): 2759–73 (2003).

Cheng J-F, et al.
Comparative and Functional Analysis of

Cardiovascular-Related Genes. *Pharmacogenomics* 4 (5): 571–82 (2003).

Couronne O, et al.
Strategies and Tools for Whole Genome Alignments. *Genome Res.* Jan 13 (1): 73–80 (2003).

Dubchak I, et al.
Multi-species sequence comparison: the next frontier in genome annotation. *Genome Biol.* 2003 4 (12): 122. Epub Nov 27 (2003).

Feil H, et al.
Site-Directed Disruption of the fimA and fimF Fimbrial Genes of *Xylella fastidiosa*. *Phytopathology* 93 (6): 675–682 (2003).

Francino MP, et al.
Phylogenetic relationships of bacteria with special reference to enteric species. In *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*, 3rd edition, release 3.13, Dworkin M (ed.) Springer-Verlag, New York, http://link.springer-ny.com/link/service/books/10125/ (2003).

Frazer KA, et al.
Cross-species Sequence Comparisons: A Review of Methods and Available Resources. *Genome Res.* 2003 Jan 13 (1): 1–12 (2003).

Gibson-Brown JJ, et al.
A proposal to sequence the amphioxus genome submitted to the Joint Genome Institute of the US Department of Energy. *J Exp Zoology Part B Mol Dev Evol.* 300 (1): 5–22 (2003).

Grimwood et al.
The DNA sequence and biology of human chromosome 19. *Nature* 428: 529–35 (2003).

Grogan JL, et al.
Basal chromatin modification at the IL-4 gene in helper T cells. *J Immunol.* Dec 15; 171(12): 6672–9, (2003).

Grossman AR, et al.
Chlamydomonas reinhardtii at the crossroads of genomics. *Eukaryotic Cell* 2(6): 1137–50 (2003).

Hallam SJ, et al.
Identification of methyl coenzyme M reductase A (mcr A) genes associated with methane-oxidizing Archaea. *Appl Env. Micro.* 69:583–91 (2003).

Huby T, et al.
Regulation of the expression of the apolipoprotein(a) gene : Evidence for a regulatory role of the 5' distal ACR enhancer in YAC transgenic mice. *Arteriosclerosis, Thrombosis and Vascular Biology* Sept 1, 23(9): 1633–9 (2003).

Ishida T, et al.
Endothelial lipase is a major determinant of HDL level. *Journal of Clinical Investigation* 111(3): 347–55 (2003).

Medina M, et al.
Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. *International Journal of Astrobiology* 2:203–211 (2003).

Medina M, et al.
The Role of Molecules in Understanding Molluscan Evolution. Pp-14–44. In *Molecular Systematics and Phylogeography of Mollusks*. Lydeard C and Lindberg D (ed.) Smithsonian Institution Press (2003).

Medina M.
Techniques in Molecular Systematics and Evolution. Methods and Tools In Biosciences and Medicine. Edited by DeSalle R, Giribet G, and Wheeler W. *Quarterly Review of Biology* 78(2): 219 (2003).

Murphy MB, et al.
An Improved Method for the in vitro Evolution of Aptamers and Applications in Protein Detection and Purification. *Nucleic Acids Research* 31 (18) e110 (2003).

Nardi F, et al.
Hexapod origins, monophyletic or paraphyletic? *Science* 299: 1887-1889 (2003).

Nardi F, et al.
Technical Comment: Response to Comment on "Hexapod origins: Monophyletic or paraphyletic?" *Science* 301: 1482e (2003).

Nobrega MA, et al.
Scanning Human Gene Deserts for Long-Range Enhancers. *Science* 302: 413 (2003).

Parolini et al.
Targeted replacement of mouse apolipoprotein A-I with the human apo-I or the mutant apoa-Imilano: Evidence of apoa-IM impaired hepatic secretion. *Journal of Biological Chemistry* 278 (7): 4740–6 (2003).

Passamonti M, et al.
Molecular evolution and recombination in gender-associated mitochondrial DNAs of the Manila clam *Tapes philippinarum*. *Genetics* 164: 603–611 (2003).

Pennacchio LA, et al.
Comparative sequence tools and databases providing insights into the human genome. *Journal of Clinical Investigation* 111 (8): 1099–1106 (2003).

Pennacchio LA.
Insights from human/mouse genome comparisons. *Mammalian Genome* 14 (7): 429–36 (2003).

Pennacchio LA, et al.
Apolipoprotein A5: A newly identified gene impacting plasma triglyceride levels in humans and mice. *Arteriosclerosis, Thrombosis, and Vascular Biology.* 23 (4): 529–34 (2003).

Pennacchio LA, et al.
Comparative sequence tools and databases providing insights into the human genome. *Journal of Clinical Investigation* 111(8): (2003).

Richardson PM, et al.
The *Xenopus tropicalis* genome project. Book Chapter in *Current Genomics 4*. (2003).

Richardson PM, et al.
Practical Applications of Rolling Circle Amplification of DNA Templates in Genetic Engineering (25):51–63. Eds. Setlow JK. Kluwer/Plenum Press (2003).

Santini S, et al.
Evolutionary conservation of regulatory elements in vertebrate HOX gene clusters. *Genome Res.* 13: 1111–1122 (2003).

Satou Y, et al.
A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. I. Genes for bHLH transcription factors. *Dev Genes Evol.* 213 (5–6):213–21 (2003).

Saunders NF, et al.
Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res.* Jul 13 (7): 1580–8. Epub 2003 Jun 12 (2003).

Shen X, et al.
Genome-wide examination of myoblast cell cycle withdrawal during differentiation. *Developmental Dynamics* 226 (1): 128–138 (2003)

Terajima D, et al.
Identification of candidate genes encoding the core components of the cell death machinery in the *Ciona intestinalis* genome. *Cell Death Differentiation* 10 (6):749–53 (2003).

Vu-Dac N, et al.
Apolipoprotein A5, a Crucial Determinant of Plasma Triglyceride Levels, isHighly Responsive to PPARa Activators. *Journal of Biological Chemistry* 278 (20): 17982–85 (2003).

Wada S, et al.
A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. II. Genes for homeobox transcription factors. *Dev Genes Evol.* 213(5–6): 222–34 (2003).

Walsh PJ, et al.
A second glutamine synthetase gene with expression in the gills of the ureotelic gulf toadfish *(Opsanus beta)*. *Journal of Experimental Biology* 206: 1523–1533 (2003).

Wyman SK, et al.
Annotating animal mitochondrial tRNAs: An experimental evaluation of four methods. pp. 44–46 in *Proc. European Conf. Computational Biol.* (ECCB).

Yagi K, et al.
A genomewide survey of developmentally relevant genes in Ciona intestinalis. III. Genes for Fox, ETS, nuclear receptors and NFkappaB. *Development, Genes, and Evolution* 213 (5–6): 235–44 (2003).

2004

Alexandrino J, et al.
Strong selection against hybrids at a hybrid zone in the Ensatina ring species complex and its evolutionary implications. *Evolution* 59(6): 1334-1347 (2004).

Ahituv N, et al.
Exploiting human-fish genome comparisons for deciphering gene regulation. *Hum Mol Genet.* Oct 1 13 Spec No 2:R261–6 (2004).

Armbrust EV, et al.
The genome of the diatom *Thalassiosira pseudonana:* ecology, evolution, and metabolism. *Science* Oct 1 306 (5693): 79–86 (2004).

Baldauf SL, et al.
The Tree of Life: An Overview. In *Assembling the Tree of Life* Cracraft J, Donoghue M (ed.) 592p. (2004).

Baroukh N, et al.
Analysis of apolipoprotein A5, C3 and plasma triglyceride concentrations in genetically engineered mice. *Artheriosclerosis Thrombosis and Vascular Biology* 24 (7): 1297–1302 Jul (2004).

Boore JL.
Complete mitochondrial genome sequence of *Urechis caupo*, a representative of the phylum Echiura. *BMC Genomic* Sep 15 5 (1): 67 (2004).

Bensasson D, et al.
Genes without frontiers? *J. Heredity* 92: 483–489 (2004).

Boffelli D, et al.
Convergent evolution in primates and an insectivore. *Genomics* Jan 83 (1): 19–23 (2004).

Boffelli D, et al.
Comparative genomics at the vertebrate extremes. *Nature Review Genetics* June 5 (6):456–65 (2004).

Boffelli D, et al.
Intra-species sequence comparisons for annotating genomes. *Genome Research*, 14: 2406–2411 (2004).

Bolotin A, et al.
Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus. Nat. Biotechnol.*, Dec; 22(12):1554-8 (2004)

Boore, JL, et al.
Complete sequences of two highly rearranged molluscan mitochondrial genomes, those of the scaphopod *Graptacme eborea* and of the bivalve *Mytilus edulis. Molecular Biology and Evolution* 21(8): 1492–15003 (2004).

Brudno M, et al.
Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.* Apr 14 (4): 685–92 (2004).

Cooper GM, et al.
Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res.* Apr 14 (4): 539–48 (2004).

Detter JC, et al.
Phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. Book Chapter (p.245–260) in *DNA Amplification: Current Technologies and Applications* Horizon Scientific Press (2004).

Eichenbaum-Voline S, et al.
Linkage and Association of Haplotypes at the APOA1/C3/A4/A5 Gene Cluster to Familial Combined Hyperlipidemia. *Arteriosclerosis, Thrombosis, and Vascular Biology* 24 (1): 167–74 (2004).

Epting CL, et al.
Stem cell antigen-1 is necessary for cell-cycle withdrawal and myoblast differentiation in C2C12 cells. J. *Of Cell Science* 117 (25): 6185–6195 (2004).

Feldman CR, et al.
Molecular systematics of Old World stripe-necked turtles (Testudines: Mauremys). *Asiatic Herpetological Research* 10: 28-37 (2004).

Frazer KA, et al.
VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32: W273–W279 Suppl. 2, JUL 1 (2004).

Gatesy J, et al.
Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of Crocodylia. *Systematic Biology* 53 (2): 342-55 (2004).

Gibbs et al.
Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* Apr 1 428 (6982): 493–521 (2004).

Grimwood J, et al.
The DNA sequence and biology of human chromosome 19. *Nature* Apr 1 428 (6982): 529–35 (2004).

Ginige MP, et al.
Use of stable-isotope probing, full-cycle rRNA analysis, and fluorescence in situ hybridization-microautoradiography to study a methanol-fed denitrifying microbial community. *Applied and Environmental Microbiology* 70 (1): 588–596 Jan (2004).

Hallam SJ, et al.
Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science* Sep 3 305 (5689): 1457–62 (2004).

Helfenbein KG, et al.
The mitochondrial genome of *Phoronis architecta*—Comparisons demonstrate that phoronids are lophotrochozoan protostomes. *Mol. Biol. Evol.* 21 (1): 153–157 (2004).

Helfenbein KG, et al.
The mitochondrial genome of *Paraspadella gotoi* is highly reduced and reveals that chaetognaths are a sister-group to protostomes. *Proceedings of the National Academy of Sciences* USA 101 (29): 10639-10643 (2004).

Huber T, et al.
Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* 20(14): 2317–2319 (2004).

Hugenholtz P, et al.
Reclassification of *Sphaerobacter thermophilus* from the subclass Sphaerobacteridae in the phylum Actinobacteria to the class *Thermomicrobia* (emended description) in the phylum Chloroflexi (amended description). *International Journal of Systematic and Evolutionary Microbiology* 54: 2049–2051 (2004).

Imanishi T, et al.
Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *PLoS Biol.* 2004 Apr 20 (2004).

Joyce WG, et al.
Developing a protocol for the conversion of rank-based taxon names to phylogenetically defined clade names, as exemplified by turtles. *Journal of Paleontology* 78 (5): 989-1013 (2004).

Kim J, et al.
Lineage-specific imprinting and evolution of the zinc finger gene ZIM2. *Genomics* Ju: 84 (1): 47–58 (2004).

Lavrov D, et al.
The demise of a phylum: mtDNA analyses indicate strongly that Pentastomida is a group within the Pancrustacea. *Proc Royal Soc London B* 271(1538): 537–544 (2004).

Leem SH, et al.
Closing the gaps on human chromosome 19 revealed genes with a high density of repetitive tandemly arrayed elements. *Genome Research* 14 (2): 239–246 (2004).

Li TT, et al.
Genetic variation responsible for mouse strain differences in integrin a2 expression is associated with altered platelet responses to collagen. *Blood* May 1; 103(9):3396-402 (2004)

Macey JR, et al.
Genetic variation among agamid lizards of the Trapelus agilis complex in the Caspian-Aral Basin. *Asiatic Herpetolgical Research* 10: 208-214 (2004).

Macey JR, et al.
Phylogenetic relationships among amphisbaenian reptiles based on complete mitochondrial genomic sequences. *Mol Phylogenet Evol.* Oct 33 (1): 22–31 (2004).

Martin J, et al.
The sequence and analysis of duplication-rich human chromosome 16. *Nature* Dec 23 432 (7020): 988–94 (2004).

Martinez D, et al.
Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nature Biotechnology* 22 (6): 695–700 (2004).

Masta SE, et al.
The complete mitochondrial genome sequence of the spider *Habronattus oregonensis* reveals rearranged and extremely truncated tRNAs. *Mol. Biol. Evol.* 21: 893–902 (2004).

Mueller RL, et al.
Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders: Novel insights from complete mitochondrial genome sequences. *Proceedings of the National Academy of Sciences USA* 101 (38): 13820-13825 (2004).

Nobrega MA, et al.
Megabase deletions of gene deserts result in viable mice. *Nature* Oct 21 431(7011): 988–93 (2004).

Nobrega M, et al.
Application and utility of comparative genomic analysis for biological discovery. *Journal of Physiology* 554(1): 31–9 (2004).

Olivier M, et al.
Haplotype analysis of the apolipoprotein gene cluster on human chromosome 11. *Genomics* 83 (5): 912–923 May (2004).

Parra M, et al.
Differential domain evolution and complex RNA processing in a family of paralogous EPB41 (protein 4.1) genes facilitate expression of diverse tissue-specific isoforms. *Genomics* Oct 84 (4): 637–46 (2004).

Parham JF, et al.
The evolutionary distinctiveness of the extinct Yunnan box turtle revealed by DNA from an old museum specimen. *Proceedings of the Royal Society* 271: S391-S394 (2004).

Pennachio L, International Human Genome Sequencing Consortium.
Finishing the Euchromatic Portion of the Human Genome. *Nature* 431 (7011): 931–945 (2004).

Pennacchio LA, et al.
Human-Mouse Comparative Genomics: Successes and Failures to Reveal Functional Regions of the Human Genome. *Symposia on Quantitative Biology: The Genome of Homo sapiens* 68: 303–9, Cold Spring Harbor Press.

Predki PF, et al.
Rolling circle amplification for sequencing templates. Production Sequencing Department DOE Joint Genome Institute, Walnut Creek, CA USA *Methods Mol Biol.* 255: 189–96 Humana press (2004).

Rodionov DA, et al.
Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. *Genome Biol.* 5 (11): R90. Epub Oct 22 (2004).

Ruiz-Trillo I, et al.Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Mol. Phylogenet. Evol.* 33 (2): 321–332 Nov (2004).

Sangwan P, et al.
Chthoniobacter flavus gen. nov., sp. nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of the phylum Verrucomicrobia. *Applied and Environmental Microbiology* 70 (10): 5875–5881 (2004).

Schmutz J, et al.
The DNA sequence and comparative analysis of human chromosome 5. *Nature* Sep 16 431 (7006): 268–74 (2004).

Schoenborn L, et al.
Liquid serial dilution is inferior to solid media for isolation of cultures representative of the phylum-level diversity of soil bacteria. *Applied and Environmental Microbiology* 70 (7): 4363–4366 (2004).

Shah N, et al.
Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics* Mar 22 20 (5): 636–43. Epub Jan 22 (2004).

Stuart BL, et al.
Molecular phylogeny of the critically endangered Indochinese box turtle Cuora galbinifrons. *Molecular Phylogenetics and Evolution* 31: 164-177 (2004).

Symula DJ, et al.
Functional annotation of mouse mutations in embryonic stem cells by use of expression profiling. *Mammalian Genome* Jan 15 (1): 1–13 (2004).

Townsend T, et al.
Molecular phylogenetics of squamata: The position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. *Systematic Biology* 53 (5): 735-757 (2004).

Tyson GW, et al.
Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* Mar 4 428 (6978): 37–43 (2004).

Wang Q-F, et al.
Haplotypes in the ApoA1-C3-A4-A5 gene cluster affect plasma lipids in both humans and baboons. *Human Molecular Genetics* 13 (10): 1049–56. (2004).

Wong GK, et al.
A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* Dec 9 432 (7018): 717–22 (2004).

Wyman S, et al.
Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20 (17): 3252–3255 Nov 22 (2004).

Zeng P, et al.
Identification of a novel enhancer of brain expression near the apoE gene cluster by comparative genomics. *Biochimica et Biophysica Acta-Gene Structure and Expression* 1676 (1): 41–50 Jan 5 (2004).

2005

Boore JL, et al.
In *Molecular Evolution: Producing the Biochemical Data, Part B* Zimmer EA, Roalson E (ed.) Volume 395 of the *Methods in Enzymology* series, Elsevier, Burlington, Massachusetts, 311-348 (2005).

Carapelli A, et al.
Relationships between hexapods and crustaceans based on mitochondrial genomics. In *Crustacean and Arthropod Relationships* Koenemann S, Jenner RA (ed.) 295-306 (2005).

Collins AG, et al.
Cnidarian phylogeny and character evolution clarified by new large and small subunit rDNA data and an assessment of the utility of phylogenetic mixture models. *Systematic Biology* (in press).

Dehal P, et al.
Two rounds of genome duplication in the ancestral vertebrate genome. *PLoS Biology* (in press).

Doyle SA.
High-throughput Cloning for Protein Expression Studies. *Chemical Genomics: Reviews and Protocols* John Walker, Ed. The Humana Press Inc. (in press)

Doyle SA.
Screening for Soluble Expression of Recombinant Proteins. *Chemical Genomics: Reviews and Protocols* John Walker, Ed. The Humana Press Inc. (in press)

Foster J, et al.
Evolution of the genome of the obligate Wolbachia endosymbiont of *Brugia malayi*, a pathogenic nematode responsible for human lymphatic filariasis. *PloS Biology* (in press)

Francino MP.
An adaptive radiation model for the origin of new gene functions. *Nature Genetics* 37 (6): 573–577 Jun (2005).

Gatesy J, Baker R.
Hidden likelihood support in genomic data: Can forty-five wrongs make a right? *Systematic Biology* 54 (3): 483-492 (2005).

Jansen RK, et al.
Methods for obtaining and analyzing whole chloroplast genome sequences. In *Molecular Evolution: Producing the Biochemical Data, Part B* Zimmer EA, Roalson E (ed.) Volume 395 of the *Methods in Enzymology* series, Elsevier, Burlington, Massachusetts, 348-384 (2005).

Keys DN, et al.
A saturation screen for cis-acting regulatory DNA in the Hox genes of *Ciona intestinalis*. *Proceedings of the National Academy of Sciences, USA* 102: 679–683.

Leebens-Mack, et al.
Identifying the basal angiosperms in chloroplast genome phylogenies: Avoiding the Felsenstein zone. *Molecular Biology and Evolution* (in press).

Macey JR.
Plethodontid salamander mitochondrial genome: A parsimony evaluation of character conflict and implications for historical biogeography. *Cladistics* 21: 194-202 (2005).

Macey JR, et al.
The complete mitochondrial genome of a gecko and the phylogenetic position of the Middle Eastern *Teratoscincus keyserlingii*. *Molecular Phylogenetics and Evolution* 36: 188-193 (2005).

Medina M.
Genomes, phylogeny and evolutionary systems biology. *Proceedings of the National Academy of Sciences* 102: 6630-6635 (2005).

Medina M, et al.
Phylogeny of the sea hares in the Aplysia clade based on mitochondrial DNA sequence data. *Bulletin of Marine Science* (in press).

Mueller RL, et al.
Molecular mechanisms and evolutionary causes of extensive mitochondrial gene rearrangement in plethodontid salamanders. *Molecular Biology and Evolution* (in press).

Murphy MB, et al.
High-throughput Protein Production for Proteomic Studies. *Chemical Genomics: Reviews and Protocols* John Walker, Ed. The Humana Press Inc. (in press).

Noonan JP, et al.
Genomic Sequencing of Pleistocene Cave Bears. *Science* 1113485: Jun 2 (2005)

Parham JF.
A reassessment of the referral of sea turtle skulls to Osteopygis (Late Cretaceous, New Jersey, USA). *Journal of Vertebrate Paleontology* 25 (1): 71-77 (2005).

Place AR, et al.
Genetic markers in Blue Crabs *(Callinectes sapidus)* II: Complete mitochondrial genome sequence and characterization of genetic variation. *Journal of Experimental Marine Biology and Ecology* 319: 15–27 (2005).

Richards S, et al.
Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution. *Genome Res.* Jan 15 (1): 1–18 (2005).

Tringe S, et al.
Comparative metagenomics of microbial communities. *Science* 308 (5721): 554–557 Apr 22 (2005).

Vrdoljak G, et al.
Characterization of a Diesel Sludge Microbial Consortia for Bioremediation. *Scanning* 27 (1): 8–14 Jan–Feb (2005).

Rojas A, et al.
Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element. *Development* June 29 (2005).

Schulte JA, et al.
A genetic perspective on the geographic association of taxa among arid North American lizards of the *Sceloporus magister* complex (Squamata: Iguanidae: Phrynosomatinae). *Molecular Phylogenetics and Evolution* (in press).

Schwarz J, et al.
Coral reef genomics: Developing tools for functional genomics of coral symbiosis. *Proceedings of International Coral Reef Symposium* Okinawa, Japan (in press).

Shi H, et al.
A report on the wild hybridization between two species of threatened Asian box turtles (Testudines: Cuora) on Hainan Island, China. *Amphibia-Reptilia* (in press).

Sly LI, et al.
Blastobacter Zavarzin 1961, 962 AL emend. Sly 1985, 44. *Bergey's Manual of Systematic Bacteriology, 2nd ed, Volume 2 The Proteobacteria* Brenner DJ, Kreig NR, Staley JT, Garrity GM (ed.) Springer, New York. (in press).

Tyson GW, et al.
Environmental shotgun sequencing. *Encyclopedia of Genomics, Proteomics and Bioinformatics* John Wiley & Sons, Ltd. (2005).

Weng L, et al.
Lack of MEF2A Mutations in Coronary Artery Disease. *Journal of Clincial Investigation* 115 (4): 1016–1020 Apr (2005).

Wolf PG, et al.
The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae): Implications for land plant phylogeny. *Gene* 350 (2): 117-128 (2005).

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231, and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.

For more information about JGI, contact:
David Gilbert, Public Affairs Manager
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
email: gilbert21@llnl.gov phone: (925) 296-5643

JGI Web site: http://www.jgi.doe.gov/

Published by the Berkeley Lab Creative Services Office in collaboration with JGI researchers and staff for the U.S. Department of Energy. LBNL-58383

CSOJO10897