The Joint Genome Institute
Decoding the Human Genome

Oh, for a decoder ring! Divining the sequence of chromosomes 5, 16, and 19 is a 24-hour-a-day job at the Joint Genome Institute.

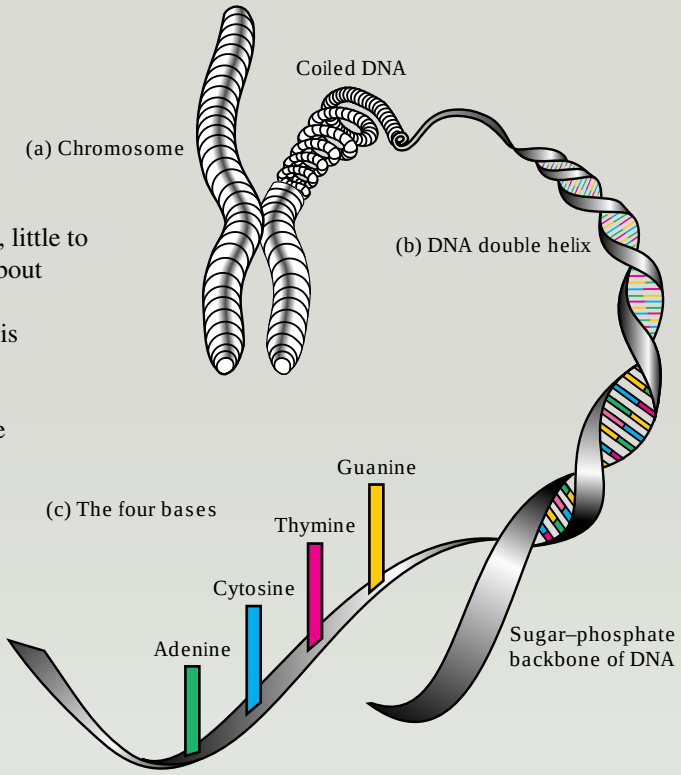
Tucked away in a light industrial park in Walnut Creek, California, about 35 miles north of Livermore, is the Joint Genome Institute (JGI), a collaboration of Lawrence Livermore, Lawrence Berkeley, and Los Alamos national laboratories funded by the Department of Energy's Office of Biological and Environmental Research. There, employees of the three institutions are working together to sequence human chromosomes 5, 16, and 19 for the worldwide Human Genome Project. This spring, the institute will announce the completion of the draft sequence of these three chromosomes, a year ahead of the schedule that was set just 18 months ago.

Sequencing is the process of decoding the 3 billion parts of our DNA, of determining the precise order of the four "bases"—adenine (A), thymine (T), guanine (G), and cytosine (C)—that comprise the entire human genetic code. By the end of 2001, the draft sequence for all 23 chromosomes will be finished, with some gaps and ambiguities still unresolved. The goal is a high-quality human DNA reference sequence by 2003. But even before the final sequence is complete, the draft sequence will be a valuable tool for researchers.

Sequencing our DNA is all about hunting genes. Each of our 100,000 genes is composed of a unique sequence of pairs of the four bases, called base pairs. Earlier research has shown that chromosome 19, for example, is home to the genes that govern lymphoid leukemia, myotonic dystrophy, diabetes mellitus, and a susceptibility to polio, along with about 2,000 others. More than 99 percent of the human DNA sequence is the same for everyone, but the variations in the remaining sequence can have huge implications. And at many places in the sequence, getting the sequence exactly right matters. A single misplaced base among the 3 billion base pairs may have lethal consequences.

Researchers know the approximate locations of genes that govern many medically important traits, but they don't know the exact location or the gene's sequence. For many other genes, their sequence (and often their location) are known, but nothing is known about what they do. And for the other tens of thousands of genes, nothing is known

Lawrence Livermore National Laboratory

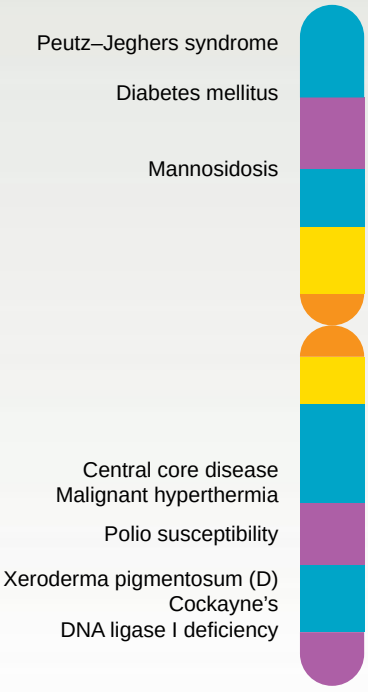


The basics of genetics. Each cell in the human body (except red blood cells) contains 23 pairs of chromosomes. Chromosomes are inherited: each parent contributes one chromosome per pair to their children. (a) Each chromosome is made up of a tightly coiled strand of DNA. The current research lies in the details of the DNA structure, which, in its uncoiled state, reveals (b) the familiar double helix shape. If we picture DNA as a twisted ladder, the sides, made of sugar and phosphate molecules, are connected by (c) rungs made of chemicals called bases. DNA has four, and only four, bases—adenine (A), thymine (T), guanine (G) and cytosine (C)—that form interlocking pairs. The order of the bases along the length of the ladder is called the DNA sequence. The hunt for genes is focused on reading the order of the bases for each DNA strand and determining which parts of the sequence constitute genes.

about their location, little to nothing is known about their sequence, and generally even less is known about their function. A goal of the Human Genome Project, the largest biological undertaking in history, is to locate all the genes on human DNA, determine their precise A, T, C, and G sequence, and then learn their function.

One of the problems in locating and sequencing genes is that among those 3 billion base pairs, only about 4 percent constitute DNA that matters to gene function. "And that 4 percent," says physicist Elbert Branscomb, director of the Joint Genome Institute, "is buried in a sea of junk—DNA whose function, if any, we do not yet understand. Furthermore, a gene does not come in a tidy package, all in one place. Pieces of a gene may be strung out along the DNA strand with lots of irrelevant DNA in between."

Even after a gene is precisely located on a chromosome and sequenced, researchers generally do not know what it does. They will usually need to determine what protein the gene produces and what those proteins do in the cell. Not only do genes and their proteins cause inheritable diseases, but they also determine how we look, how well our body metabolizes food or fights infection, and sometimes even



Research by organizations around the world had determined the location of numerous genes on chromosome 19 before Livermore began mapping it in 1987.

how we behave. In addition, some genes may become altered because of environmental factors, resulting in such maladies as heart disease, many cancers, and possibly some psychiatric disorders. For all of these genes, researchers want to determine how and why they change and what their altered proteins do to the body. For most genes, there are as yet no clear answers.

Completing the draft sequence of the human genome is just the beginning of a long, complicated chain of research events. The gene for cystic fibrosis, for example, was discovered more than four years ago. Although we are still a long way from "fixing" the cellular defect that causes this disease, unraveling the gene's secrets has allowed private industry to deal with a major symptom of the disease.

The Department of Energy is supporting the sequencing of other organisms as well. At other institutions, several viruses and bacteria have already been completely sequenced, as have such larger organisms as baker's yeast (*Saccharomyces cerevisiae*) and the roundworm (*Caenorhabditis elegans*). As part of DOE's Microbial Genome Program, numerous microbes are being sequenced that may be useful for remediation of toxic waste sites or understanding how microbes contribute to carbon sequestration and global warming. Other programs are responsible for sequencing such pathogens as anthrax and smallpox for a better understanding of ways to counter a biological terrorist attack. In addition, considerable work is under way on the mouse, about 85 percent of whose genes are identical to our own (see *S&TR*, December 1999, pp. 14–17). "The similarities indicate which parts of the genome must be really important," notes biochemist Trevor Hawkins, deputy director of the JGI. Comparative genomics—analyzing and comparing the genetic material of different species—is an important tool for studying evolution, the functions of genes, and inherited genes.

Branscomb says, "The goal of our large-scale sequencing work is to help lay down the infrastructure that allows biological scientists to answer questions as efficiently as possible. Genomic studies should soon reveal why some people are able to defend against the AIDS virus and others are not, for example.

"The genome is the basis of all life," he continues. "When we get sick with an infectious disease, what's going on is a war between two sets of genes—ours and those of the virus or bacteria. Someday the medical profession will have better ways to handle these diseases, thanks to work on the genome."

Energy Secretary Bill Richardson dedicated the Production Sequencing Facility of the Joint Genome Institute on April 19, 1999.

In a Production Mode

Livermore, Berkeley, and Los Alamos had been working on the Human Genome Project for several years when the three joined forces to form the JGI in 1996. The offices in Walnut Creek house the JGI's Production Sequencing Facility (PSF). Approximately 120 people work there, half of them in sequencing and the rest in research and development, organization of the vast amount of genetic information being amassed, and other tasks.

One usually thinks of DOE's national laboratories as research and development institutions and not as industrial-scale production facilities. But the JGI has changed that, at least for DNA sequencing. The PSF was formed because the R&D facilities at Livermore, Berkeley, and Los Alamos were not expandable or adequate for the large-scale production required to meet the demanding deadlines of the Human Genome Project. In contrast, the PSF offers large, open laboratory and robotics areas that allow for high efficiency and maximum production. Power and data connections are located overhead to minimize downtime during equipment changes or production line reconfigurations. Even large pieces of equipment are on wheels to facilitate quick changes.

Employees of the Livermore and Berkeley laboratories began moving into the Walnut Creek facility in December 1998, and Energy Secretary Bill Richardson dedicated it on April 19, 1999. Los Alamos employees live a

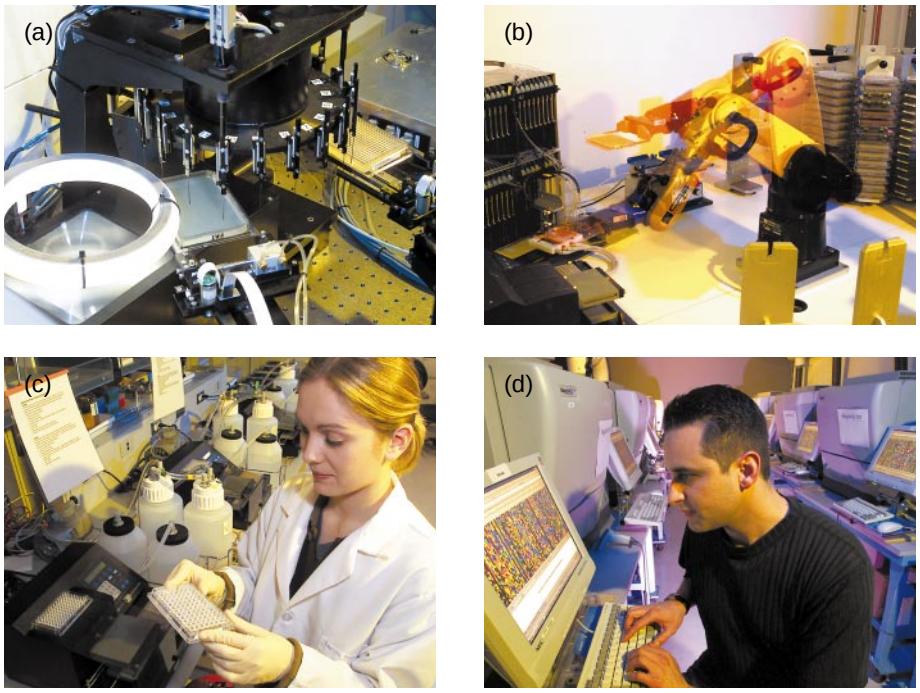
thousand miles from Walnut Creek, so some of their sequencing work continues at Los Alamos. Currently, the only Los Alamos employee at the PSF is Deputy Director Trevor Hawkins, who also serves as director of sequencing.

Some tasks of the Joint Genome Institute are still under way at Berkeley and Livermore until additional space is available at Walnut Creek. The microbial sequencing operation at Livermore led by Jane Lamerdin is, for example, scheduled to move to Walnut Creek in June when a new building opens.

In a highly automated process, the staff at Walnut Creek is identifying human DNA sequence 24 hours a day at a rate of about 10 million Phred-20 base pairs (bp) every day. (Phred 20 is a measurement of quality, indicating a 1 in 100 chance of any base pair being incorrectly identified.) Not so long ago, sequencing 40,000 bp was considered a worthy multiyear thesis project for a Ph.D. student.

Speeding up Sequencing

The sequencing process has many steps. It begins when the DNA to be sequenced is randomly sheared into



The sequencing process at the Joint Genome Institute (JGI) has numerous steps, four of which are shown here: (a) Colonies of cells containing human DNA are selected from a cell culture plate. (b) The CRS robot system places a DNA sample plate onto a plate washer for purification of the DNA. (c) Tijana Glavina, a JGI researcher, removes a plate of purified DNA from a plate washer. (d) Aaron Avila, a JGI research technician, reviews the sequencing data produced by one of JGI's 84 DNA capillary sequencers.

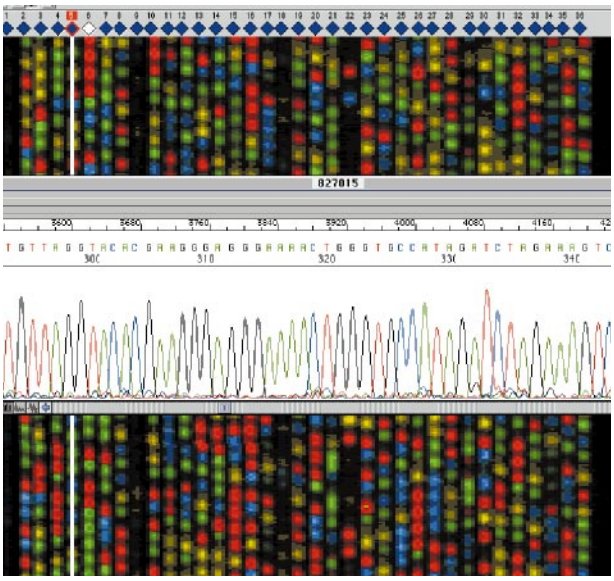
Why DOE and Genome Research?

The formation of the Joint Genome Institute was a logical outgrowth of the involvement of the national laboratories in the Human Genome Project and earlier work on human genetics. Decades ago, the U.S. Congress charged the Department of Energy's predecessor agencies (the Atomic Energy Commission and the Energy Research and Development Agency) with studying and analyzing the consequences of human genetic mutations, especially those caused by radiation and the chemical byproducts of nuclear energy production. A particular focus of research has been the attempt to detect tiny genetic mutations among the survivors of the Hiroshima and Nagasaki bombings and their descendants.

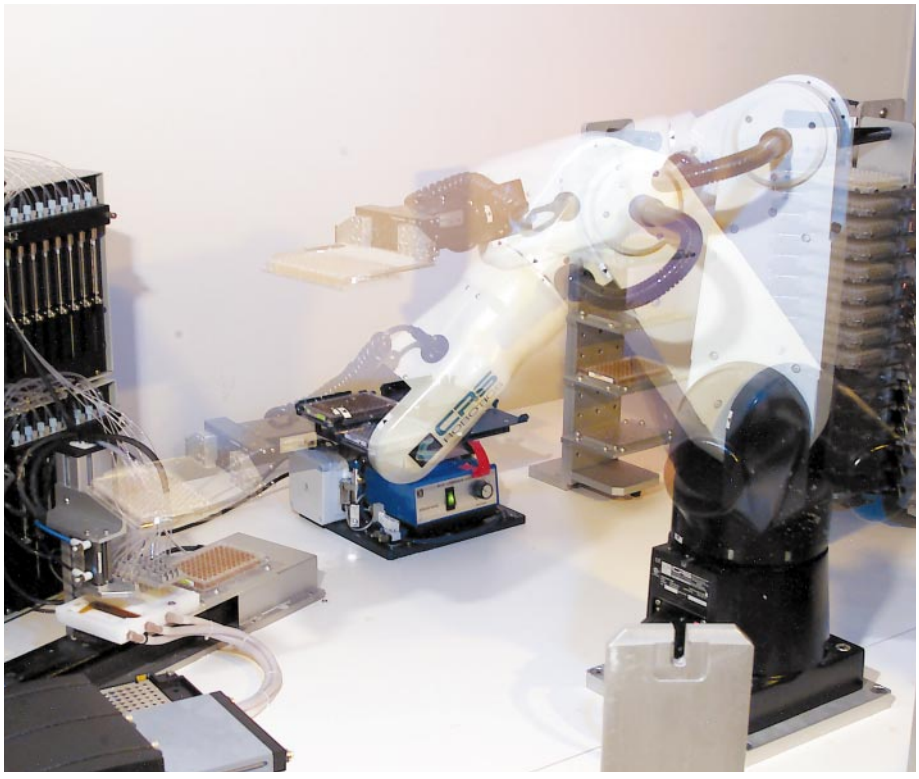
At Livermore, the first biomedical program was chartered in 1963 to study the radiation dose to humans from isotopes in the environment; a natural extension was to explore how radiation and chemicals interact with human genetic material to produce cancers, mutations, and other adverse effects. One Livermore project examined three genes on chromosome 19 that are involved in the repair of DNA damaged by radiation or chemicals.

From studies such as these grew the recognition that the best way to study genetic changes was to analyze the entire human genome to obtain a reference sequence. In 1986, DOE was the first federal agency to launch a major initiative to completely decipher the entire human genetic code. A year later, Livermore researchers began studying all of chromosome 19. In 1990, DOE joined forces with the National Institutes of Health, which had its own research under way, to kick off the Human Genome Project.

In 1994, DOE expanded its genomic research with the Microbial Genome Initiative to sequence the genomes of bacteria of likely interest in the areas of energy production and use, environmental remediation, and waste reduction. Such microbes live under extreme conditions of temperature and pressure and could be engineered for such practical purposes as waste control and environmental cleanup.



The four colors in this chromatogram represent the four bases that make up our DNA: green is adenine (A), blue is cytosine (C), yellow is guanine (G), and red is thymine (T). Each fragment of DNA differs from the next fragment by one base, and the dye indicates the terminal base of each fragment. The order of the colors indicates the order of the bases and hence the sequence of the DNA.



An automated platform centered around an articulated robot made by the robotics firm CRS is used to isolate DNA samples from cellular debris and purify them for sequencing. The Joint Genome Institute uses two of these robots, which, together, can process nearly 31,000 samples every day. Here, one of the CRS robots is placing one plate of 96 DNA samples onto a plate washer to remove cellular debris.

overlapping chunks of about 150,000 to 200,000 bp. Many clones or identical copies are made, which are then cut into smaller overlapping fragments of 2,000 to 4,000 bp and cloned again in a culture of *E. coli* bacteria. The DNA is then purified to remove all cellular debris.

Next, four nearly identical reactions are run in a plate with 96 tiny wells. During the reaction, an enzyme copies the input DNA to produce fragments that differ from one another by one base. Four fluorescent dyes are added to the mix. Each dye is attracted to one and only one base.

Next, in a process known as electrophoresis, an electric current is applied to the labeled fragments. Because the DNA itself has a negative charge, the many fragments migrate at different rates depending on their size: smaller ones move faster than larger ones. As the fragments migrate past a certain point, a scanning laser beam excites the dyes on the DNA fragments to indicate which base is on the end of each fragment. Then another group of labeled DNA fragments, one base longer than the previous group, passes the laser scanner, indicating the identity of that terminal base. The process continues with each set of fragments one base longer than the previous set. The fluorescent signals are captured and digitized, resulting in a four-color chromatogram showing peaks that represent each of the four DNA bases. At this point, the order of the colors is the order of the bases and therefore the sequence of the DNA.

About a year ago, the PSF replaced a labor-intensive and space-consumming electrophoresis process with a commercially developed, automated one. It employs machines with tiny capillary tubes that fit into the 96 wells and through which the samples pass as the laser scans them. Hawkins says, “If the time and labor savings weren’t

enough, the capillary machines also produce higher quality data and can read longer fragment lengths."

After the bases are read, computers reassemble the overlapping fragments into long continuous stretches of sequenced DNA that are analyzed for errors, gene-coding regions, and other characteristics. This process is repeated over and over for all of the chunks of DNA that make up a chromosome.

The front end of the operation, where the chunks of DNA are cut, involves the most skilled handwork. Virtually all other facets of the process have been automated, using robots such as the one shown on p. 8.

Molecular biologist Paul Predki manages the draft sequencing operation at the PSF. Paul Richardson, manager of research and development and also a molecular biologist, works with Predki's staff to find ways to improve the production process. Richardson and his team are on the lookout for ways to cut out steps, develop new materials, and increase automation. They are also working to reduce the volume of reagents used. A major development has been new plates with 384 wells, four times the capacity of the 96-well plates.

These changes in instrumentation and methodology, combined with a physical reorganization of laboratory operations more in line with an industrial production setting, have resulted in remarkable increases in the amount of DNA that can be sequenced. Production is measured in the number of lanes run through the capillary machines (or earlier through the gel plates). Production has increased spectacularly during the months that the JGI has been in operation. In January 1999, 113,000 lanes were run; in December 1999, 823,000 lanes were run.

Because of a continuing stream of improvements, the schedule for the Human Genome Project has been revamped several times. The goal now is for a finished, fully sequenced genome by 2003, two years earlier than originally planned.

Genomes of Microbes, Too

For the overall genomic effort, sequencing the genomes of microbes has both short-term and long-term benefits. In the short term, researchers such as Jane Lamerdin and her team are studying specific microbes that may be helpful in environmental remediation. Looking to the longer term, we will learn more about the microbial role in the overall "metabolism" of Earth. For an increasing number of microorganisms, microbiologists can proceed from a complete knowledge of an organism's genomic blueprint to its consequent activities and behaviors. "With our information, biologists will have a better understanding of how organisms interact and work together in a given environmental niche," says Lamerdin.

Microbes are thought to make up more than 60 percent of the Earth's biomass. They are found in every environment, thriving in extremes of heat, cold, radiation, pressure, salinity, acidity, and darkness, often where no other forms of life are found and where nutrients come only from inorganic matter.

It is precisely this ability to thrive in apparently bizarre environments that makes microbes potentially so

Ethical Concerns about Genomic Discoveries

Even before the formal beginning of the Human Genome Project in 1990, project managers, researchers, and lawmakers recognized that increasing knowledge about human biology and personal genetic information would raise a number of complex issues for individuals and society. In response to Congressional mandates for identifying and defining such issues and developing effective policies to address them, the Department of Energy and the National Institutes of Health have devoted 3 to 5 percent of their annual Human Genome Project budgets to studies of the project's ethical, legal, and social implications (ELSI).

Such implications include the ability to predict future illnesses well before any symptoms or medical therapies exist; the privacy and fair use of genetic information with respect to employers, insurers, direct marketers, banks, credit raters, law enforcement agencies, and others; the availability of genetic information in largely unprotected data banks; and the possible discriminatory misuse of genetic information. One possible misuse of the Human Genome Project is that genome research and the wide use of genetic screening could foster a new genetic underclass, leading to a host of new societal conflicts.

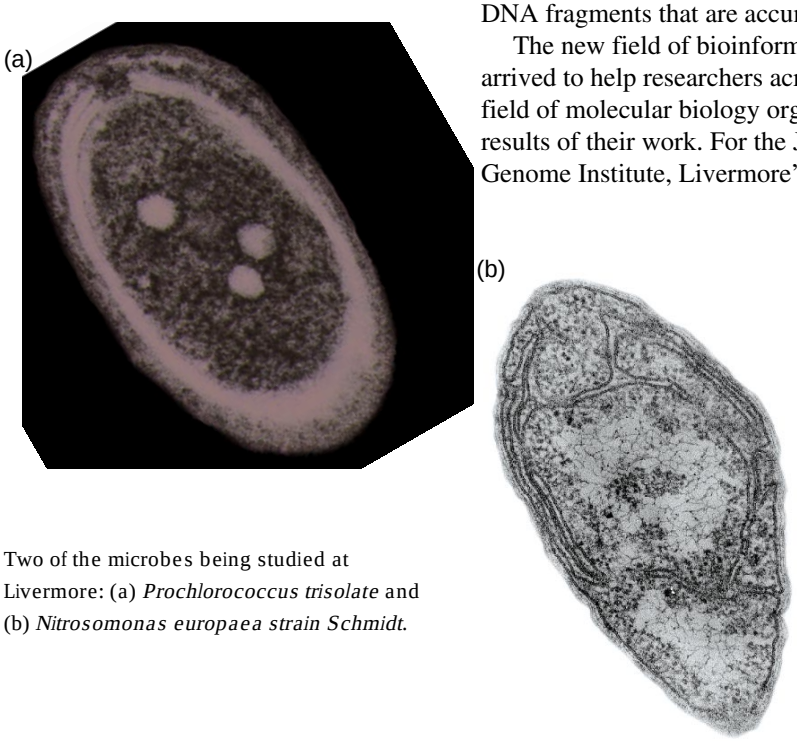
With these concerns in mind, the ELSI program emphasizes the privacy of genetic information, its safe and effective introduction into the clinical setting, fairness in its use, and professional and public education. One of DOE's major commitments is to the Human Genome Management Information System, which disseminates information on all aspects of genome research.

The ELSI program has become a model for others around the world and has led to the establishment of similar programs as part of other research activities.

The Gene Letter is a free, online quarterly newsletter on ethical, legal, and social issues in genetics for interested professionals and consumers. See www.genesage.com/geneletter/.

useful in environmental remediation. Of particular interest to DOE is the growing amount of excess carbon dioxide (CO_2) in the atmosphere and the role of microorganisms in global carbon sequestration. Lamerdin's team is sequencing five microorganisms that all use CO_2 as their sole carbon source (as opposed to an organic carbon source) and that are fairly common within their respective ecosystems. Two soil-dwelling microbes, *Nitrosomonas europaea* and *Rhodopseudomonas palustritis*, carry out chemical functions that make them possible candidates for use in the treatment of contaminated soil and water. This particular species of *Rhodopseudomonas* is also important in carbon cycling because it degrades and recycles components of wood, the most abundant polymer on Earth. A third terrestrial species of microbe being sequenced (*Nostoc punctiforme*) enters into symbiotic associations with fungi and lichens, relationships that are relevant to carbon cycling and sequestration in tundra.

The JGI is also sequencing two ubiquitous marine bacteria, *Prochlorococcus marinus* and *Synechococcus*. The former is intriguing because it has adapted to a wide range of light conditions at the various depths of its ocean habitat. It is also speculated to be the most abundant photosynthetic organism on the planet. In these microbes' genomes, researchers are looking for more information on the way they use CO_2 and nutrients in their environment to better understand their growth properties, which are often affected by global climate changes.



Two of the microbes being studied at Livermore: (a) *Prochlorococcus trisolate* and (b) *Nitrosomonas europaea strain Schmidt.*

All That Information!

What happens to the ever-expanding sequence data for all of the chromosomes in our DNA? Estimates are that if the sequence for the whole genome were printed out, it would fill 200 Manhattan-size telephone directories. And that does not count the annotation and additional data about specific genes and DNA fragments that are accumulating.

The new field of bioinformatics has arrived to help researchers across the field of molecular biology organize the results of their work. For the Joint Genome Institute, Livermore's Tom Slezak is responsible for the flood of data. Slezak, a computer scientist, has been working on the Human Genome Project for many years and notes, "The challenge is keeping up with the extraordinary rate of change and growing masses of information within the industry."

The institute participates in the most widely used database for genomic data, which is at the National Institutes of Health. The NIH's National Center for Biotechnology Information creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information. GenBank is in partnership with two other major repositories, the European Molecular Biology Laboratory and the DNA Data Bank of Japan, to exchange data. Submissions to any one of them show up at the other two within a day, with consistent accession numbers. Last December, chromosome 22 made history as the first human chromosome to be completely sequenced and deposited in GenBank.

All parties recognize that continued investment in current and new databases and analytical tools is critical to the success of the Human Genome Project and to the future usefulness of the data it produces. These databases are already helping bioscientists in their quest to diagnose and treat disease.

After the Draft

By mid-2000 when the draft sequence is complete, work on finishing the sequence for chromosomes 5, 16, and 19 will shift to the Stanford Genome Center where the sequencing process will continue to fill in gaps. Stanford joined the JGI in October 1999.

Generating the draft sequence is requiring researchers to sequence each

piece of DNA about five times. Producing the highest quality sequence will require an additional two to five times so that each piece is sequenced seven to ten times. To ensure the highest level of confidence—and perhaps to uncover important individual differences—researchers may eventually sequence most biologically or medically important regions even more exhaustively. The error-rate goal for the finished sequence is 1 error in 10,000 bases.

Back in Walnut Creek, post-draft sequencing efforts will shift to the mouse, whose genome is remarkably similar to ours. There is an almost one-to-one correspondence between genes in the two species, although they sometimes occur in different places in the two genomes. The human genome is, of course, also nearly identical to that of chimpanzees, and it even shares many common elements with the genome of the lowly fruit fly. But the mouse, with its small size, high fertility rate, and experimental manipulability, offers great promise for studying the genetic causes and pathological progress of diseases. Studies of the mouse will also help us better understand the genetic role in disease susceptibility.

The JGI will continue collaborative sequencing work with researchers at other institutions. Various projects with scientists at Harvard, Yale, the Massachusetts Institute of Technology, Johns Hopkins, and University of California campuses at San Francisco and Davis involve in-depth sequencing of chromosomes 5, 16, and 19 as well as work on the mouse and other organisms to learn more about cancer, liver disease, autoimmune disorders, and other diseases. Tim Andriese, the institute's collaborative liaison, notes, "Producing sequence data for collaborators follows our primary goal—to provide the research community with essential sequence data."

Another focus of future work at the JGI is functional genomics, which, as its name implies, interprets the functions of human genes and other DNA sequences. This work requires that resources and strategies be developed for large-scale investigations across whole genomes. At the JGI, Edward Rubin manages this effort, which now involves several small pilot projects being carried out at many different DOE laboratories.

The ultimate goal of the work at the JGI and elsewhere is to develop a molecular-level understanding of how we develop from embryo to adult, what makes us work, and what causes things to go wrong. "Solving the genetic code of humans and other creatures is a huge, important quest," says Branscomb, "It will allow us to solve the mystery of mysteries: how does life work? Then we can really begin to address human suffering."

—Katie Walter

Key Words: bioinformatics, DNA, functional genomics, Human Genome Project, Joint Genome Institute (JGI), microbial genetics, mouse genome, sequencing.

For further information contact Elbert Branscomb (925) 296-5700 (branscomb1@llnl.gov).

Also see the following Web sites:
- *The Joint Genome Institute, jgi.doe.gov/*
- *The Human Genome Project, www.ornl.gov/hgmis/*
- *The DOE's Human Genome News, www.ornl.gov/hgmis/publicat/hgn/hgn.html*

About the Scientist

ELBERT BRANSCOMB is the director of the Department of Energy's Joint Genome Institute (JGI), a collaboration of Lawrence Livermore, Lawrence Berkeley, and Los Alamos national laboratories responsible for the sequencing of human chromosomes 5, 16, and 19. Branscomb received his B.A. in physics from Reed College in 1957 and his Ph.D. in theoretical physics from Syracuse University in 1964. He joined Lawrence Livermore in 1964 as a theoretical physicist and became a senior biomedical scientist in 1969. In 1996, he became director of JGI.

Branscomb's professional activities include being a member of the Editorial Board of the *Journal of Computational Biology* and of the National Cancer Institute's Cancer Genetics Working Group. From 1996 to 1998, he was a member of the Panel of Scientific Advisors for the National Institutes of Health–National Council of Human Genome Research's Pilot Project for Large-Scale Sequencing of the Human Genome. He is also the coauthor of numerous scholarly articles, primarily on scientific research related to the human genome.