MAPPING THE HUMAN GENOME

Lawrence Berkeley Laboratory
University of California

June 1989

Charter members of the LBL Human Genome Advisory Committee

Members of the Advisory Committee meet at least once each year to provide advice on long- and short-term scientific questions and on organizational and staffing issues.

BRUCE M. ALBERTS, Ph.D. (Chair)
Chairman, Department of Biochemistry and Biophysics
University of California at San Francisco

SIR WALTER BODMER, Ph.D., F.R.S.
Director of Research
Imperial Cancer Research Fund, England

DAVID BOTSTEIN, Ph.D.
Vice President–Science
Genentech, Inc.

CHARLES DELISI, Ph.D.
Chairman, Department of Biomathematical Statistics
Mt. Sinai School of Medicine

RENATO DULBECCO, M.D.
Nobel laureate
The Salk Institute

WALTER GILBERT, Ph.D.
Nobel laureate
The Biological Laboratories, Harvard University

MAXINE F. SINGER, Ph.D.
President
Carnegie Institute of Washington

NANCY WEXLER, Ph.D.
President
Hereditary Disease Foundation

FOREWORD: THE CONTEXT FOR OUR ENDEAVORS

The following pages aim to lay a foundation for understanding the excitement surrounding the "human genome project," as well as to convey a flavor of the ongoing efforts and plans at the Human Genome Center at the Lawrence Berkeley Laboratory. Our own work, of course, is only part of a broad international effort that will dramatically enhance our understanding of human molecular genetics before the end of this century. In this country, the bulk of the effort will be carried out under the auspices of the Department of Energy and the National Institutes of Health, but significant contributions have already been made both by nonprofit private foundations and by private corporations. The respective roles of the DOE and the NIH are being coordinated by an interagency committee, the aims of which are to emphasize the strengths of each agency, to facilitate cooperation, and to avoid unnecessary duplication of effort. The NIH, for example, will continue its crucial work in medical genetics and in mapping the genomes of nonhuman species. The DOE, on the other hand, has unique experience in managing large projects, and its national laboratories are repositories of expertise in physics, engineering, and computer science, as well as the life sciences. The tools and techniques the project will ultimately rely on are thus likely to be developed in multidisciplinary efforts at laboratories like LBL. Accordingly, we at LBL take great pride in this enterprise—an enterprise that will eventually transform our understanding of ourselves.

Charles R. Cantor
Director, Human Genome Center

THE HUMAN GENOME: A BRIEF INTRODUCTION

For all the manifest complexity of our species, full of creativity and contradictions, the machinery of our minds and bodies is built and run with "only" about 100,000 kinds of protein molecules. In a reductionist sense, all of our art and science, as well as the variety of the world's five billion people, are ultimately accounted for by the distribution and activity of a relatively few molecules. More surprising still, the differences between two unrelated individuals, between the man next door and Mozart, may reflect functional differences in a mere handful of these 100,000 molecules. We are far more alike than we are different. At the same time, there is room for near-infinite variety.

For each of these proteins, we can imagine a single corresponding gene (though there is sometimes some redundancy) whose job it is to ensure an adequate and timely supply. Our uniqueness as individuals, then, is due to the uniqueness of our own personal set of some 100,000 genes, copies of which are preserved in each of our body's many trillion cells. It is no overstatement to say that to decode these 100,000 genes in some fundamental way would be the most dramatic step we could take toward unraveling the manifold mysteries of life.

The *human genome* is the full complement of genetic material in a human cell. It is different for each of us, but not so different that deciphering one would not tell us a great deal about every other. The genome, in turn, is distributed among 23 pairs of *chromosomes,* which, in each of us, have been replicated and re-replicated since the fusion of sperm and egg that marked our beginning. At a more basic level, the genome is deoxyribonucleic acid, or *DNA,* a natural polymer built up of repeating *nucleotides,* each consisting of a simple sugar, a phosphate group, and one of four nitrogenous bases (Figure 1). In the chromosomes, two DNA strands are twisted together into an entwined spiral—the famous double helix—held together by weak bonds between complementary bases, adenine (A) in one strand to thymine (T) in the other, and cytosine to guanine (C-G). In the language of molecular genetics, each of these linkages constitutes a *base pair.* All told, counting only one of each pair of chromosomes, the human genome comprises about three billion base pairs.

The specificity of these base pair linkages underlies all that is wonderful about DNA. First, replication becomes straightforward. Unzipping the double helix provides unambiguous templates for the synthesis of daughter molecules. Second, by a similar template-based process, a means also becomes available for producing a DNA-like messenger to the cell cytoplasm. There, this *messenger RNA,* the faithful complement of a particular DNA nucleotide sequence, directs the synthesis of a unique

The human genome at four levels of detail.

Apart from reproductive gametes, each cell of the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA. Each strand of the DNA is a huge natural polymer consisting of repeating nucleotide units, each of which comprises a phosphate group, a sugar (deoxyribose), and a base (either guanine, cytosine, thymine, or adenine). In its "normal" state, DNA takes the form of a highly regular double-stranded helix, the strands of which are linked by hydrogen bonds between guanine and cytosine and between thymine and adenine. Each such linkage is said to constitute a "base pair"; some three billion base pairs constitute the human genome. It is the specificity of these base-pair linkages that underlies the mechanism of DNA replication illustrated here. Each strand of the double helix serves as a template for the synthesis of a new strand, the nucleotide sequence of which is strictly determined. Each daughter double helix is thus not only a twin, but also an exact replica of its sole parent.

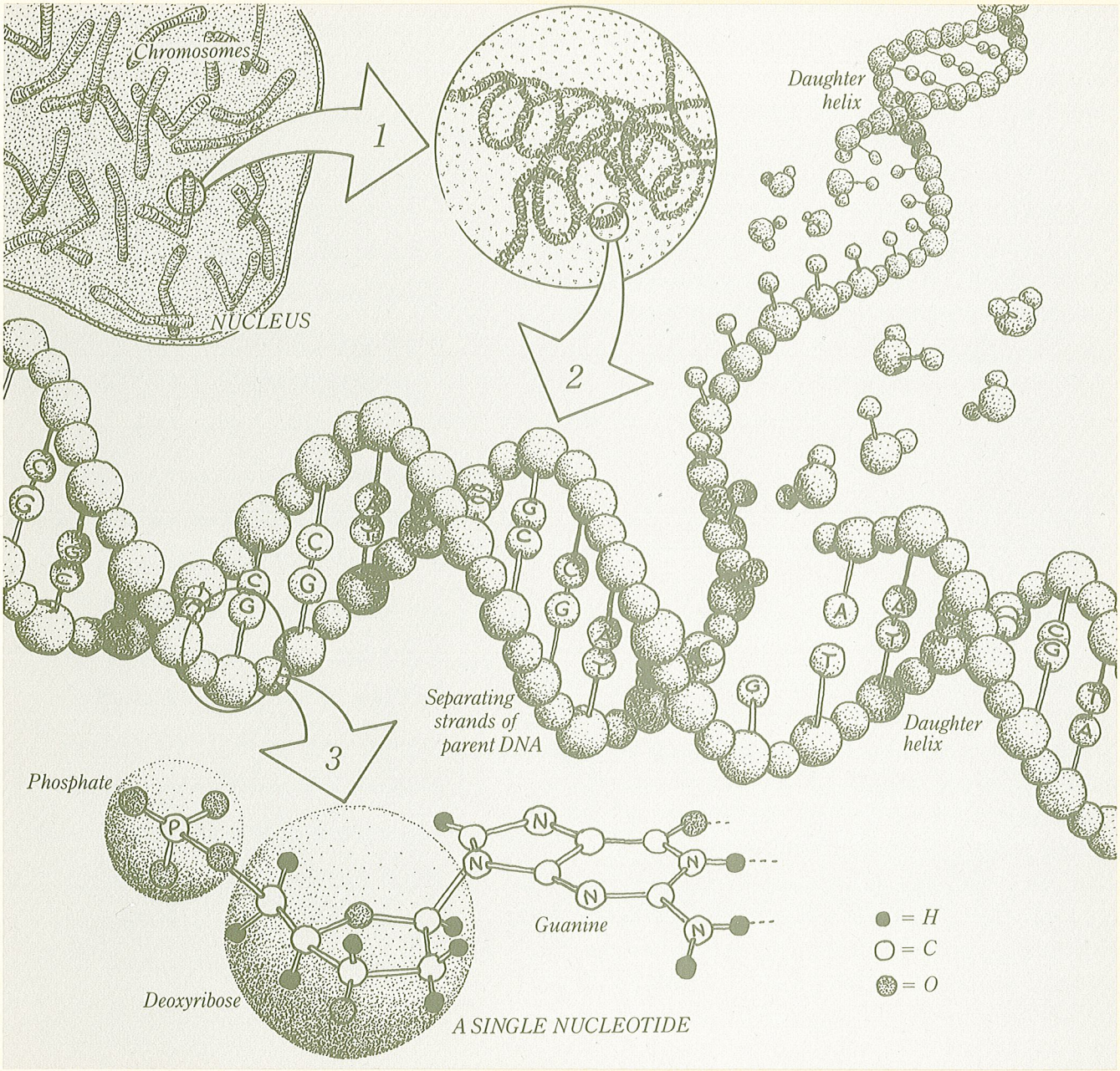

Chromosomes
NUCLEUS
1
2
3
Daughter helix
Separating strands of parent DNA
Daughter helix
Phosphate
Deoxyribose
Guanine
A SINGLE NUCLEOTIDE
= H
= C
= O

FIGURE 1

protein. Many subtleties are entailed in the synthesis of proteins, but in a schematic sense, the process is elegantly simple.

Every protein is made up of one or more *polypeptide* chains, each a series of (typically) several hundred molecules known as *amino acids,* linked by so-called peptide bonds. Remarkably, only 20 different kinds of amino acids suffice as the building blocks for all human proteins. The synthesis of a protein chain, then, is simply a matter of specifying a sequence of these different amino acids. This is the role of the messenger RNA, serving as the transcript of a nucleotide sequence within the genomic DNA. (The same nitrogenous bases are at work in RNA as in DNA, except that uracil takes the place of the DNA base thymine.) Each linear sequence of three bases (both in RNA and in DNA) corresponds uniquely to a single amino acid. The RNA sequence AAUGCA thus dictates that the amino acids asparagine and alanine should be added to a polypeptide chain (Figure 2). A segment of the chromosomal DNA that specifies the synthesis of a single type of protein constitutes a single *gene.*

Gene-directed synthesis of a protein chain.

In the cell nucleus, RNA is produced by transcription, in much the same way that DNA replicates itself. Single-stranded RNA, however, substitutes the sugar ribose for deoxyribose and the base uracil for thymine. One form of RNA, messenger RNA or mRNA, conveys the DNA recipe for protein synthesis to the cell cytoplasm. There, bound temporarily to a cytoplasmic particle known as a ribosome, each three-base "codon" of the mRNA links to a specific form of transfer RNA (tRNA) having the complementary three-base sequence. This tRNA, in turn, transfers a single amino acid to the growing protein chain. Each codon thus unambiguously directs the addition of a particular amino acid to the protein; on the other hand, the same amino acid can be added by more than one codon. In this illustration, the mRNA sequences GCA and GCC are both specifying the addition of the amino acid alanine.

MAPPING, SEQUENCING, AND UNDERSTANDING

One can imagine "understanding" the human genome at several levels of detail (Figure 3). At the coarsest level, we might seek to determine the whereabouts of the most important human genes—at least to the point that we could assign each to a specific chromosome. This kind of low-resolution *mapping* has, in fact, now been done for more than a thousand genes. At the next level, the goal might reasonably be to map the genome at some higher resolution—perhaps to establish for every human gene a chromosomal location that is accurate to within a million base pairs. In this area, we have made the barest start. At a still deeper level lies the "Holy Grail" of biology, the dream of *sequencing* the human genome. The result would be a string of three billion characters, representing the sequence of base pairs that defines our species. Included, of course, would be the sequence for every gene, as well as the sequences for the much greater lengths of DNA that have no known function. Should anyone undertake to print it all out, the result would fill 200 volumes the size of the Manhattan phone book.

In a sense, though, even a complete genome sequence is only the beginning of "understanding." The deepest mystery is how the potential of 100,000 genes is regulated and controlled, how blood cells and brain cells are able to perform their separate functions with the same appar-

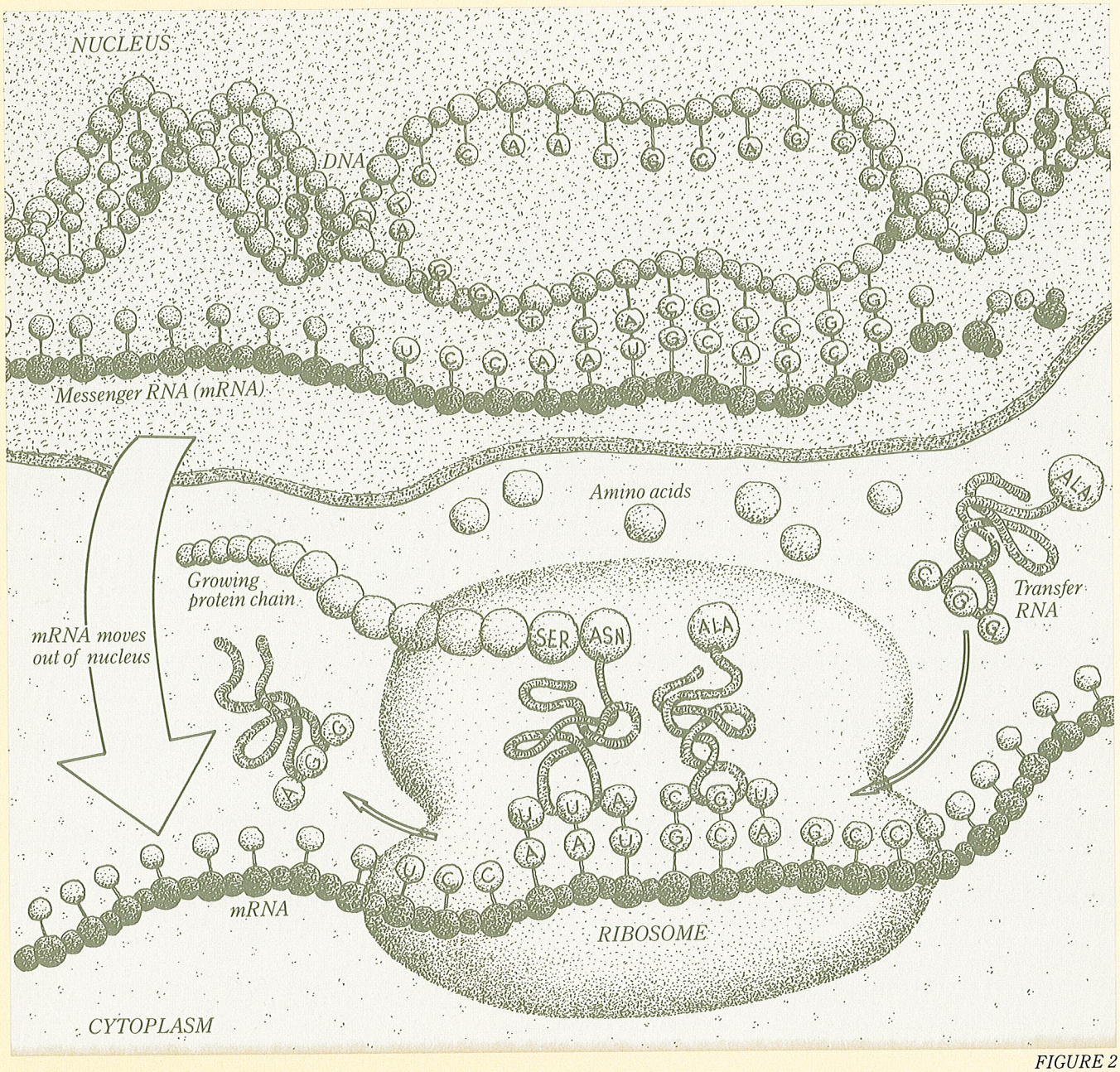

NUCLEUS
DNA
Messenger RNA (mRNA)
Amino acids
Growing protein chain
mRNA moves out of nucleus
SER
ASN
ALA
Transfer RNA
mRNA
RIBOSOME
CYTOPLASM

FIGURE 2

ent genetic program, and how these and countless other cell types arise in the first place from an undifferentiated human embryo. A first step toward solving these subtle mysteries, though, is a more complete physical picture of the master molecules that lie at the heart of it all.

There are more-concrete reasons for studying the human genome as well. One influential statement was that of Nobel laureate Renato Dulbecco; in 1986 he wrote in *Science,* "We are at a turning point in the study of...cancer....If we wish to learn more..., we must now concentrate on the cellular genome.... [T]he sequence of the human DNA is the reality of the species, and everything that happens in the world depends on those sequences." As these statements appeared in print, experts from around the world were gathering in Santa Fe, New Mexico, to ponder the feasibility of sequencing the human genome by the year 2000. The overwhelming, and somewhat surprising, consensus was that such a goal was not only meritorious, but also within practical reach.

The U.S. Department of Energy, which had sponsored the Santa Fe meeting, responded to its outcome by announcing in April 1986 preliminary plans for a Human Genome Initiative. In the DOE's view, mapping the human genome would mark a major milestone in its mission to understand the health effects of energy production and use. In the fall of 1987, the availability of the first funds for this Initiative signaled a national commitment to the human genome project, and the Department designated the Lawrence Berkeley Laboratory and the Los Alamos National Laboratory as Human Genome Initiative research centers. By September 1988, the National Research Council and the Congressional Office of Technology Assessment had released reports endorsing the effort to map and sequence the human genome, and the international Human Genome Organization had been established to coordinate worldwide efforts. Within three years, a vision had become an acknowledged national, even international, priority. In the United States, the Department of Energy and the National Institutes of Health share the role of leading the effort.

Current DOE efforts aim primarily at "resource and technology development." This recognizes the current state of technology and the realistic needs of a full-scale mapping effort. With the tools now available, mapping the entire human genome at high resolution would be a mammoth, labor-intensive, and inordinately expensive undertaking. As a consequence, the emphasis today is twofold. First is the development of analytical and computational resources: better ways to physically manipulate large fragments of DNA, fast and economical (read "automated") methods for sequencing such fragments, and a greater facility for storing and handling the resulting huge quantities of data. Second is an ordered library of cloned DNA fragments. The sum of these fragments,

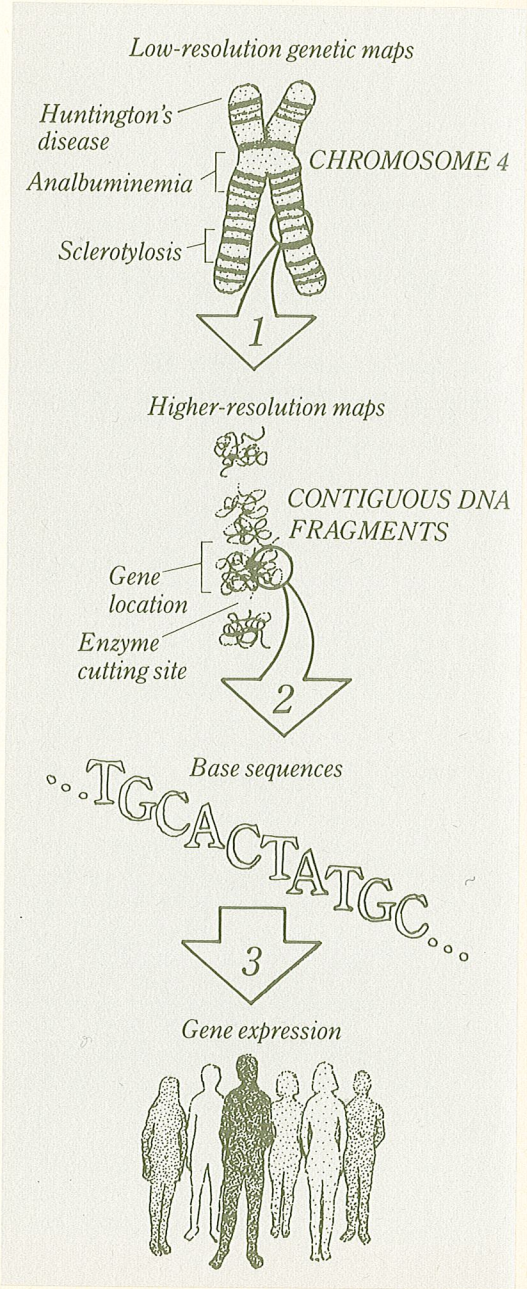


FIGURE 3

each originating in a human cell but propagated as a *clone* by genetic engineering techniques, would constitute a *reference human genome.* With the technological tools in hand and with the subject of study readily available, a full-scale mapping effort could then begin—even an effort to read off the sequence of the three billion base pairs that constitute "the reality of our species."

Levels of understanding the human genome.

The human genome may be "understood" at several levels. Geneticists have already charted the rough positions of over 1000 genes, and a patchy start has been made at establishing higher-resolution maps of the genome. Such maps can be either genetic maps, which define, at least roughly, the chromosomal locations of genes, or physical maps, which might describe, say, the cutting sites for certain DNA-cleaving enzymes. The ultimate map is the base pair sequence for the human genome, but beyond even that is an understanding of how genes regulate and express themselves.

The DOE's Mission and the Human Genome Project

Much has been written about the reasons for the effort to map and sequence the human genome. The arguments range from the esoteric to the firmly practical: As the fundamental description of the human species, the genome is the foundation of human biology; it may also be the key to conquering hundreds or even thousands of genetically linked diseases. In particular, the Department of Energy's pioneering role in exploring the human genome arises directly from its legislated mandate. Through its predecessor, the Energy Research and Development Administration (ERDA), the DOE is broadly charged with "supporting environmental, biomedical, physical, and safety research related to the development of energy resources and utilization technologies." Basic genomic data are fundamental to these responsibilities. Indeed, the old Atomic Energy Commission, with similar though narrower responsibilities, was the first federal agency to provide significant support for genetics research. Today, higher-resolution genome maps are essential to our understanding of stresses imposed by energy-related technologies and environmental factors.

To carry out its mandate, the DOE relies in large part on the resources of its national laboratories. Indeed, it was in recognition of the special capabilities of these laboratories that the strategy of the DOE Human Genome Initiative took shape: to take the lead in a multidisciplinary program of "resource and technology development, with objectives of speeding and bringing economies to the national human genome effort."

TOOLS OF THE MOLECULAR GENETICIST

Though molecular biologists are not ready today to begin the task of sequencing the entire human genome, they have a good idea of how to go about it. Indeed, several million base pairs of human DNA have already been sequenced, and it is likely that the 99.9% that remains unexplored will come to light by familiar means—much improved, but not fundamentally different.

Three tools of molecular genetics are of central importance, both for mapping and for current efforts to sequence relatively short stretches of DNA. The first is the class of DNA-breaking proteins known as *restriction enzymes.* These enzymes, the first of which were discovered in the late 1960s, cleave double-stranded DNA molecules at specific recognition sites, usually four or six nucleotides long (Figure 4). For example, a restriction enzyme called *Hae*III recognizes the single-strand sequence GGCC and invariably cuts the double helix at that site.

When digested with a particular restriction enzyme, then, identical segments of human DNA yield identical sets of *restriction fragments.* In turn, a given set of fragments yields a reproducible pattern when sorted according to size by means of a separation technique known as *gel electrophoresis.* Even a very slight change in the sequence of the original segment, however, can produce an entirely different set of restriction fragments and a different electrophoresis pattern.

This leads directly to the second essential tool of modern molecular genetics, *pulsed-field gel electrophoresis,* or *PFG,* invented in 1984 by Charles Cantor. In classical gel electrophoresis, electrically charged macromolecules are caused to migrate through a polymeric gel under the influence of an imposed static electric field. In time the molecules sort themselves by size, since the smaller ones move more rapidly through the gel than do larger ones. Only with the advent of PFG, however, in which the strength and direction of the applied field is varied rapidly with time, could DNA strands of more than 50,000 base pairs be separated.

The third tool is some means of DNA "amplification." The classic example is the *cloning vector,* which may be circular DNA molecules derived from bacteria, chromosomes from bacteriophages (viruslike parasites of bacteria), or artificial chromosomes constructed from yeast genomic DNA. The characteristic all these vectors share is that fragments of "foreign" DNA can be inserted into them, whereby the inserted DNA is replicated along with the rest of the vector as the host reproduces itself. A *yeast artificial chromosome,* or *YAC,* for instance, is constructed by assembling the essential functional parts of a natural yeast chromo-

Digesting DNA with restriction enzymes.
Isolated from various bacteria, restriction enzymes serve as microscopic scalpels that cut DNA molecules at specific sites. The enzyme Hae*III, for example, cuts double-stranded DNA only where it finds the single-strand sequence GGCC. The resulting fragments can then be separated by gel electrophoresis. The electrophoresis pattern itself can be of interest, since variations in the pattern from a given chromosomal region can sometimes be associated with variations in genetic traits, including susceptibility to certain diseases. This kind of "genetic linkage" has been used to establish the locations of important genes. Knowledge of the cutting sites also yields a kind of physical map known as a restriction map.*

some — DNA sequences that initiate replication, sequences that mark the ends of the chromosomes, and sequences required for chromosome separation during cell division — then splicing in a fragment of human DNA. This engineered chromosome is then reinserted into a yeast cell, which reproduces the YAC during cell division, as if it were part of its normal complement of chromosomes. The result is a colony of yeast cells, each containing a copy, or clone, of the same fragment of human DNA.

A second, more recently developed way of amplifying DNA is the *polymerase chain reaction,* or *PCR* (Figure 5). This enzymatic replication technique requires that activators, or *PCR primers,* be attached as short complementary strands at the ends of the separated DNA strands to be replicated. An enzyme then completes the synthesis of the complementary strands. Strand separation and repetition of the polymerase reaction can amplify DNA by a hundred-thousand-fold in less than three hours. As with cloning vectors, the result is a collection of clones of the original DNA fragment. One of the important early goals of the human genome project is to establish a library of such cloned fragments that covers the entire human genome.

Indeed, if such a library could be *ordered,* that is, if the relative positions on the human chromosomes could be established for all the fragments, one would have the perfect resource for achieving the project's ultimate goal, sequencing the human genome. How this sequencing might be done can be illustrated by one among the several methods in current use, the Maxam-Gilbert procedure (Figure 6). The first step is to radioactively label one end of each identical DNA strand in a preparation of cloned fragments. The preparation is then divided into four portions, and each is subjected to a chemical agent that selectively destroys one or two of the DNA's nucleotides. Gel electrophoresis is used to sift the resulting subfragments according to size, allowing one to infer the exact nucleotide sequence for the original DNA strand.

The technical tools for sequencing the human genome are thus well in hand — in principle at least. However, if we do nothing but continue our genomic exploration at the current pace, the complete sequence will not emerge until sometime after the year 2500! The challenge is to increase the pace, at the same time reducing the cost. In 1980 a competent molecular biologist could sequence a fragment of DNA at a cost of perhaps $5–10 per base; today, under favorable conditions, the price tag for a single base has dropped to about $1. Furthermore, Japanese scientists have reported detailed plans for a sequencing "robot" that is said to cut the cost to only 17 cents. It is this kind of progress that gives practical hope to the biologists' dream.

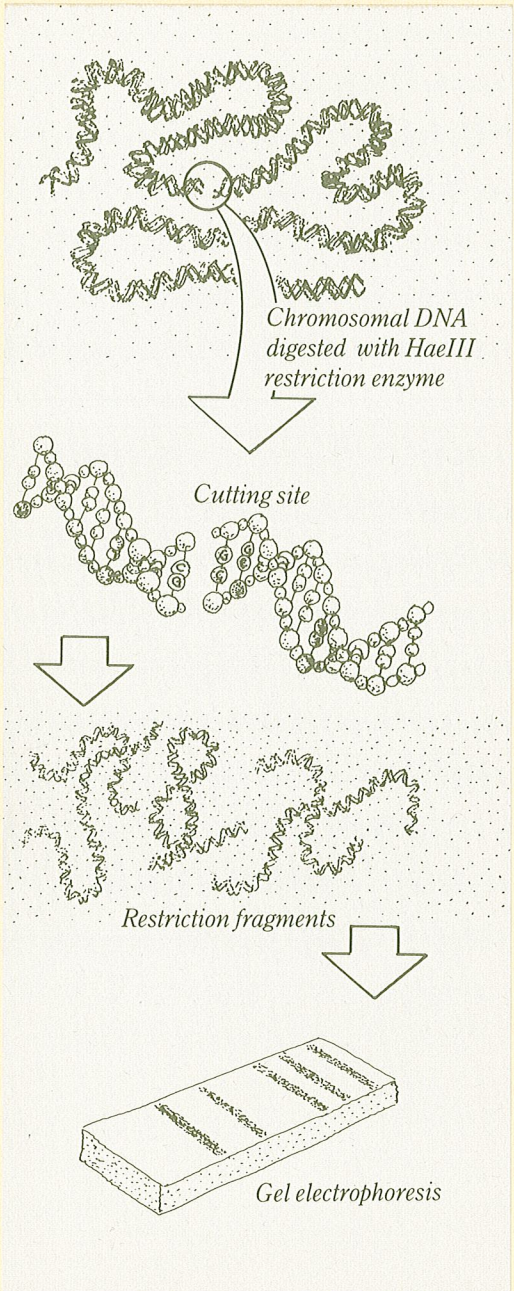


FIGURE 4

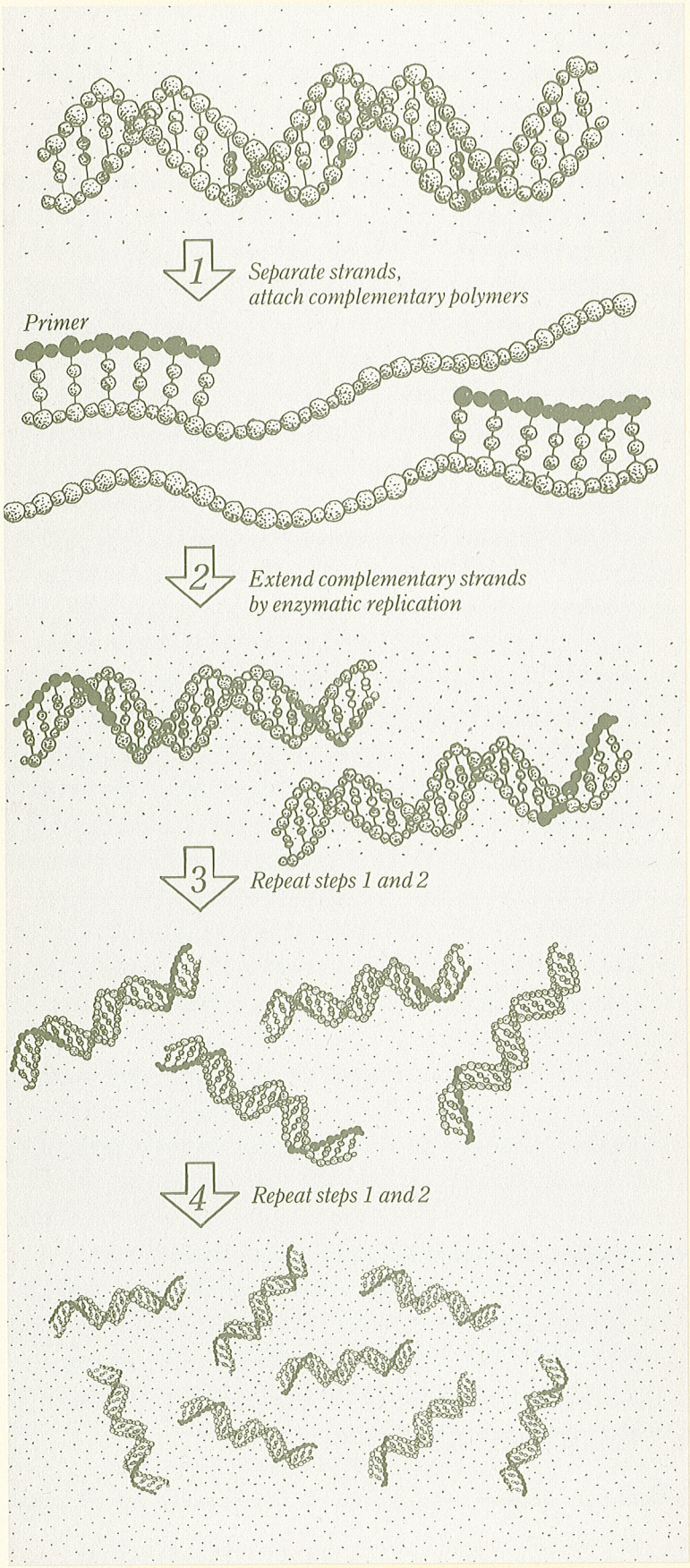


FIGURE 5

DNA amplification by enzymatic replication.

The polymerase chain reaction is a recently developed method for cloning human DNA—a necessary prelude to any intensive mapping or sequencing efforts. PCR begins with the annealing of primers to the single-stranded fragments to be replicated. With the primers attached, a DNA replication enzyme can complete the complementary strands thus started. The succeeding PCR cycle then operates on the newly produced duplicates, as well as the original fragments. Each cycle thus doubles the amount of the selected DNA fragments in the reaction mixture. Twenty-five cycles, which can be completed in less than three hours, can theoretically produce amplification by 30-million-fold; in practice, amplification by a factor of over one hundred thousand can be readily achieved.

Sequencing by the Maxam-Gilbert procedure.

In this sequencing method, single-stranded DNA is prepared with a radioactive label at one end. This preparation is then divided into four, and each is treated with a different nucleotide-destroying enzyme. Each reaction is carefully controlled so that, on average, each DNA strand is broken at only one site. Each mixture thus contains radioactively labeled fragments whose lengths reflect the positions of the nucleotide (or nucleotides) destroyed in the reaction. Gel electrophoresis—one lane per reaction mixture—then yields the sequence of the original single strand. In this illustration, labeled fragments that appear in both the "C" and the "T + C" lanes indicate the positions of cytosine; fragments in the "T + C" lane alone indicate thymine.

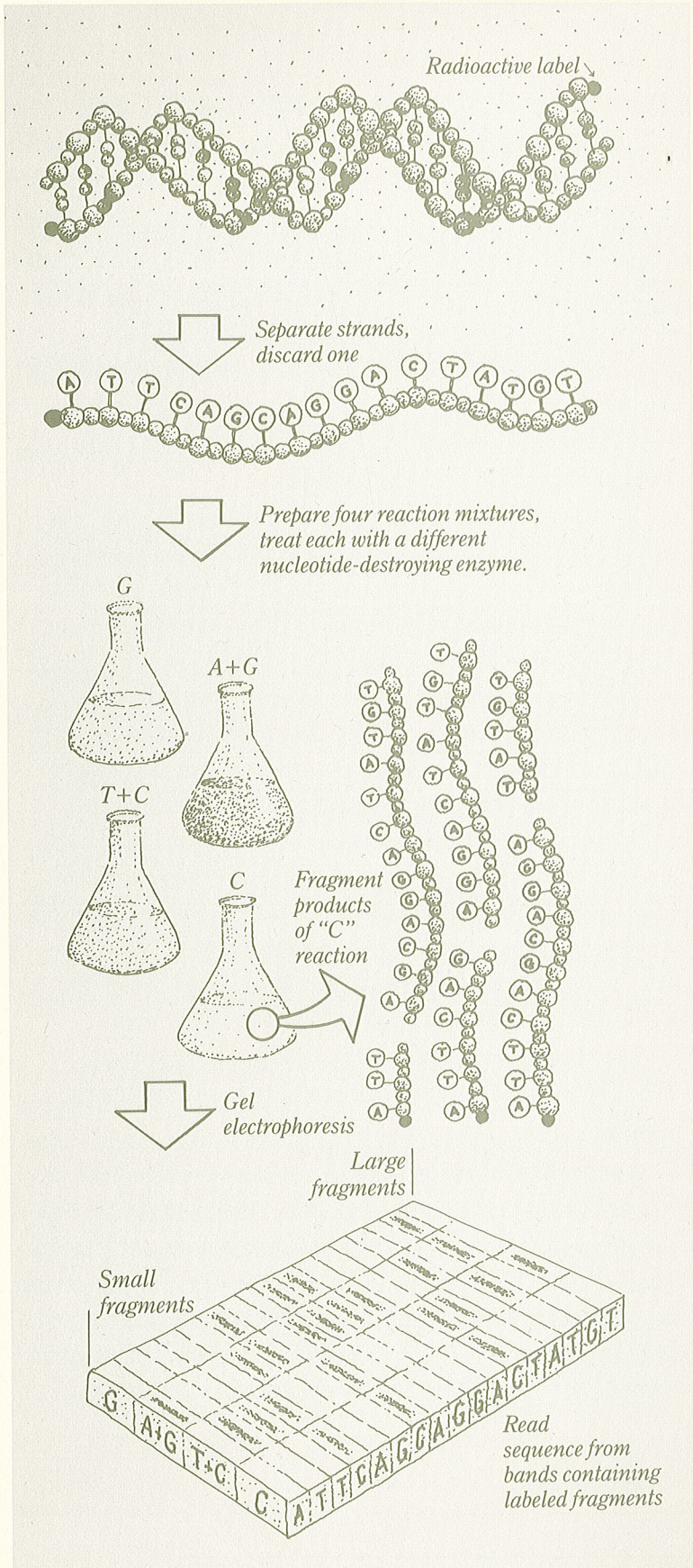


FIGURE 6

LBL'S HUMAN GENOME CENTER: ACHIEVEMENTS AND PLANS

The Lawrence Berkeley Laboratory was the birthplace not only of modern high-energy physics, but also of nuclear medicine, one of the first major cooperative enterprises between the physical and life sciences. In the past half century, research patterns have shifted and expanded, but a cooperative bond among disciplines remains an important strength of the Laboratory—and of the other national laboratories as well. In this overview of the research activities at LBL's Human Genome Center, this multidisciplinary spirit is especially evident.

In broad terms, the Center's activities are concentrated in three areas: automation of existing physical mapping methods and development of new ones, evaluation and enhancement of existing sequencing technologies, and improvement of methods for interpreting and analyzing maps and sequence data. Current efforts focus on two of the human chromosomes, numbers 21 and 22. Two principles guided the development of this research agenda. First was the strong conviction that methodological development must be accompanied by actual mapping and sequencing efforts. Hence, our program will continue to be characterized by "data-producing" projects, tightly coupled with efforts to refine laboratory methods and to develop new techniques for data handling and data analysis. The second guiding principle was the expectation that powerful automated methods will be absolutely necessary if we are to merge and reconcile research results from the many laboratories that will ultimately be involved in genome mapping and sequencing.

The following pages present a few highlights of research at the Human Genome Center, both accomplishments and continuing endeavors.

Constructing an ordered library for chromosomes 21 and 22. A major goal at the Center is to construct an ordered library of DNA fragments from human chromosomes 21 and 22. In a sense, this effort is the central project at the Center, serving as a focus for efforts to improve the tools of the trade and to develop new ones. Therefore, many of the projects described below aim for methodological improvements and technological developments that will first be tested as part of this effort.

Linking a library of cloned DNA fragments.

Several strategies are available for ordering an unordered library of cloned restriction fragments. The one schematically illustrated here involves sequencing the ends of the fragments, together with a corresponding family of much shorter linking clones. The linking clones, themselves generated by restriction enzymes, are fragments that contain the cutting site (in this simplified example, GGCC) for the enzyme used to produce the clone library. In practice, linking clones are typically 50 to 100 base pairs long, and the fragments being linked are likely to be many thousands of base pairs long. If the library contains contiguous fragments and if the family of linkers is complete, assigning an order to the fragments is a simple matter of sequence matching—a puzzle-solving task best done by computer.

A map of restriction enzyme–cutting sites on chromosome 21 is already well-advanced, and our future efforts will complement and build on the research already complete. The result, for this chromosome, will be complete maps of increasing resolution, together with ordered libraries, in the early 1990s. Previous research has also established the approximate loci for a number of genes on chromosomes 21 and 22, so ordered libraries will point the way naturally to mapping efforts of even higher resolution that focus on chromosome regions of special interest.

Improving YAC cloning methods. One approach to an ordered library is the use of YACs. Thus, as an explicit part of our efforts on chromosomes 21 and 22, new YAC vectors will be sought and YAC cloning procedures will be improved. In particular, our aims include increasing the size of the human DNA fragments that can be cloned, the efficiency with which they can be incorporated into yeast, and the ease with which they can later be separated from the yeast's own DNA. Cross-breeding different strains of yeast—a classic "genetic engineering" technique itself—is one method that will be used to produce yeast strains receptive to large DNA fragments. To streamline the isolation of cloned fragments from a colony of yeast, we will explore methods of modifying the "uninteresting" yeast DNA—for example, by introducing additional enzyme-cutting sites, so that upon digestion the yeast genomic DNA is cut into pieces much smaller than the human DNA fragments we seek. Another approach will be to develop a different species of yeast as a cloning vector. The species being looked at has natural chromosomes much larger than, thus easily separable from, any YAC currently available.

Linking large DNA fragments and YACs by sequencing. Producing an ordered library of clones requires "linking" the many cloned fragments—that is, establishing which fragments are connected to which in the intact human chromosome. One way to establish linkage is to sequence several hundred base pairs at the ends of cloned restriction fragments, as well as the entire lengths of overlapping *linking clones,* then to look for matches (Figure 7). We plan to carry out this kind of DNA linking by means of automated sequencing procedures and gel readers, thus testing some of the tools that need the most improvement if a full-scale sequencing effort is ever to be undertaken.

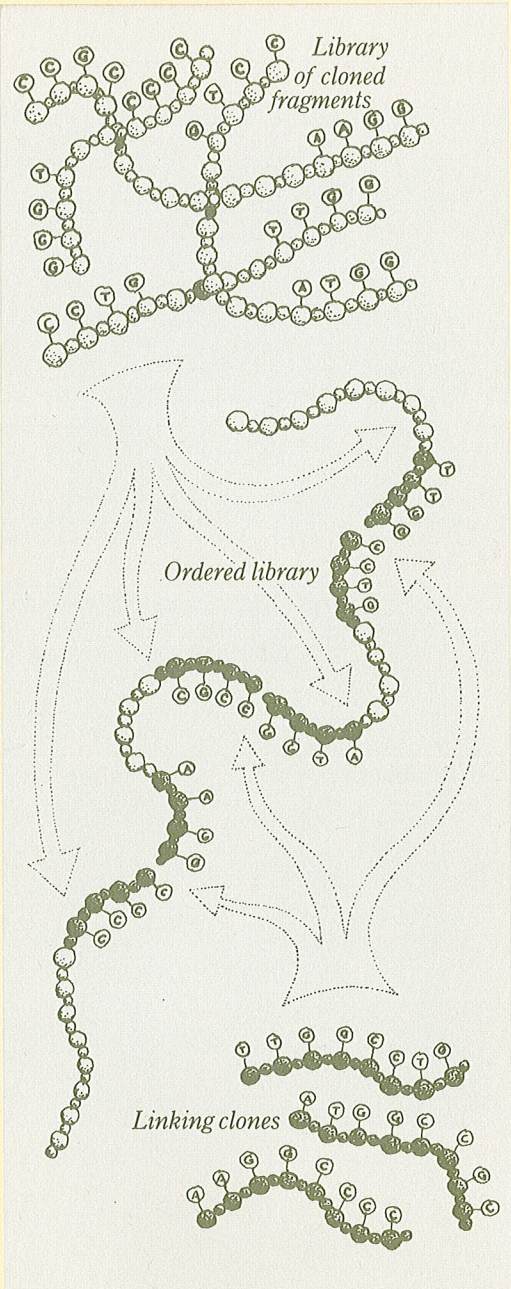
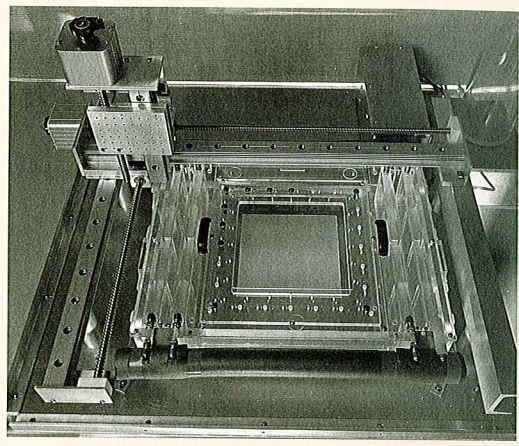


FIGURE 7

Improving pulsed-field gel electrophoresis. Pulsed-field gel electrophoresis is one of the central laboratory tools of the human genome project. In recent years, it has become the universal means for separating DNA fragments of up to 10 million base pairs. To reduce separation times and enhance resolution, LBL engineers and biologists have constructed a PFG test bed that allows conditions (such as electric field strength and direction, acidity, and temperature) to be monitored and recorded at different places in the gel (Figure 8). Analysis of these quantitative data, never before available, will provide an understanding of the physical basis of the separation and offer a route to optimizing the technique. In addition, by use of a computer controller, the PFG test bed allows active control of numerous electrode potentials, together with the capability of programming complex pulse cycles.

Developing automated methods for image enhancement, image analysis, and map construction. One of the labor-intensive elements of current mapping efforts is the analysis of electrophoresis images and the cross-correlation of results from different experiments. Furthermore, given suitable electrophoresis data, often from different sources, map construction is often no more than a tedious puzzle. Accordingly, we are working toward greater automation in several areas. First, we are applying advanced image-analysis and image-enhancement techniques



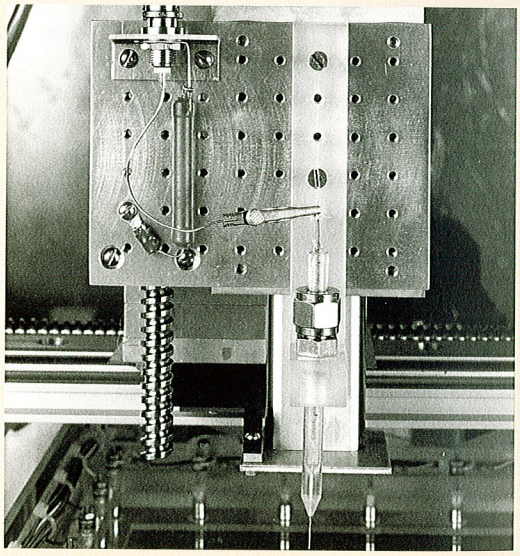


FIGURE 8

A pulsed-field gel electrophoresis test bed. *The basic DNA separation technique of modern molecular biology is gel electrophoresis, in particular, pulsed-field gel electrophoresis. In the PFG test bed shown in the upper photograph here, conditions at any point in the gel can be monitored by a probe mounted on a computer-controlled platform (lower photo). The result has been data never before available on variations in the electric field, gel temperature, and acidity, among other parameters. The test bed also allows active computer control of electrode potentials, as well as the capability for programming complex pulse cycles. The results of continuing studies with this and other test beds will be increased resolution and shorter separation times, especially for DNA fragments of more than five million base pairs.*

to the problem of reading electrophoresis gels (Figure 9). Second, we are developing both an electronic "laboratory notebook" and an "image data base" to serve as resources for eventual automated map construction. And third, we have begun work on the software for this final step of automation. We expect the elements of this automated gel analysis system to become operational in 1990 and 1991.

Exploring new methods for sequence analysis, data handling, and sample distribution. The one certain outcome of the international human genome project is an enormous quantity of data; one of its undeniable needs is a coordinated means for handling and sharing this data as it accumulates. Automated methods will likewise be required for evaluating, maintaining, and distributing the archives of clones that will be the basic experimental resource for genomic research. We are approaching these needs on two fronts. First, we have begun development of a workstation that we hope will provide a uniform means of data access for all involved in the human genome project (Figure 10). As a logical extension of our electronic laboratory notebook, we are designing the workstation as a means for integrating all existing data bases, the result of which could then be displayed hierarchically at different levels of resolution. As part of this integrative function, we plan to develop sophisticated programs for comparing physical and genetic maps from different sources and, in the long term, for reconstructing three-dimensional genomic structures from sequence data. Significant elements of this system will be in place by 1991.

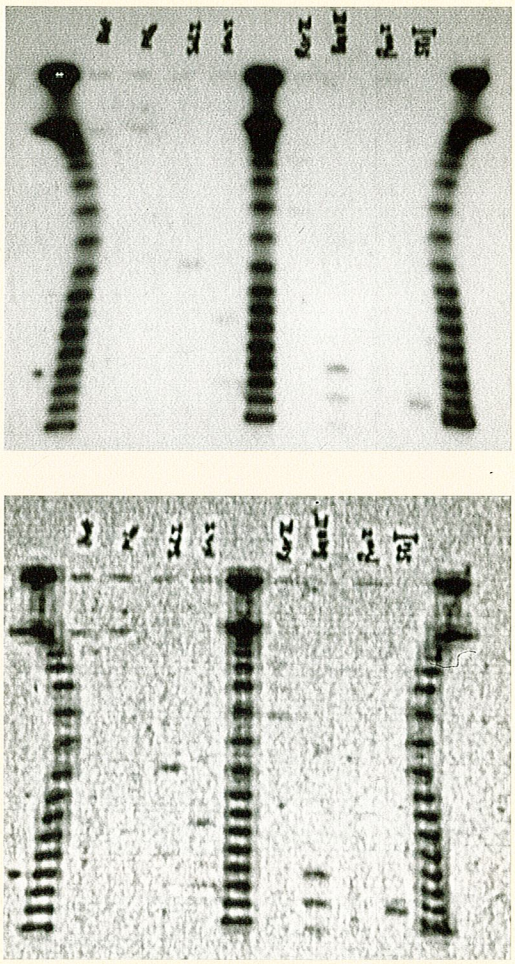


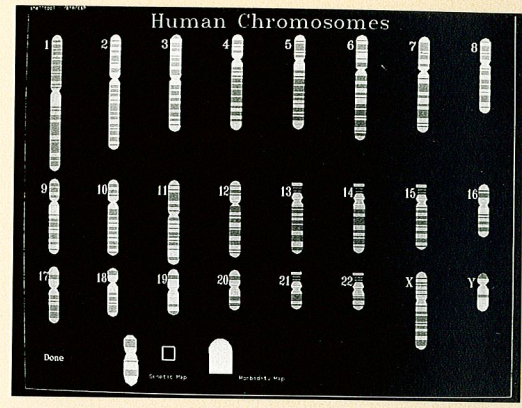
FIGURE 9

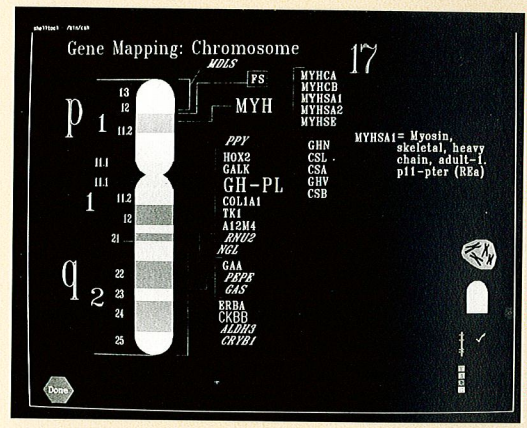
Image enhancement of electrophoresis data.
Preliminary but dramatically successful efforts have been made to apply digital enhancement techniques to gel electrophoresis images. The results have included the resolution of bands from apparently featureless dark areas and the detection of weak bands previously undetected. In the upper illustration here, the three dark lanes contain standards, which overwhelmed the response of the recording film, whereas the dilute samples produced faint or undetectable bands in the other lanes. The enhanced lower image shows resolved bands in the standards lanes, as well as quantifiable bands in the sample lanes.

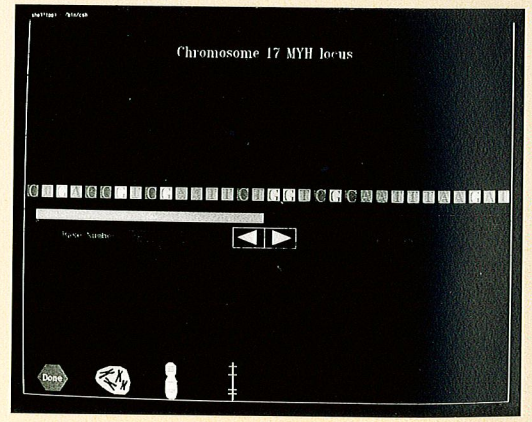
On the second front, we plan an effort to develop automated methods for acquiring, evaluating, maintaining, and distributing an archive of clones and its associated data. Coordinating access to the many thousands of DNA samples that must eventually be accumulated will give rise to problems in fields ranging from data management and inventory control to robotics.

Developing advanced techniques for DNA manipulation, amplification, and imaging. The polymerase chain reaction is most easily carried out if the desired fragment can be isolated and purified. Thus, the challenge is to obtain the pure fragment one wishes to replicate. One approach is to develop tools for orienting and slicing isolated DNA molecules. To this end, we have already succeeded in imaging single fluorescently labeled DNA molecules in the light microscope and in manipulating them in soft gels by means of electric fields. Further work will aim at cutting molecules at prescribed sites by methods that include locally delivered restriction enzymes, focused x-rays, and mechanical cleavage.

In addition, efforts will be made to broaden the applications of PCR itself. Currently, it can be used only in aqueous solution. It would be of much greater value if it could be applied, say, to DNA fragments in a gel or on a membrane surface. As a result, we plan to develop methods to make possible regionally localized PCR. One of the key steps will be to develop PCR primers that can be delivered to DNA molecules localized on a substrate rather than in solution.



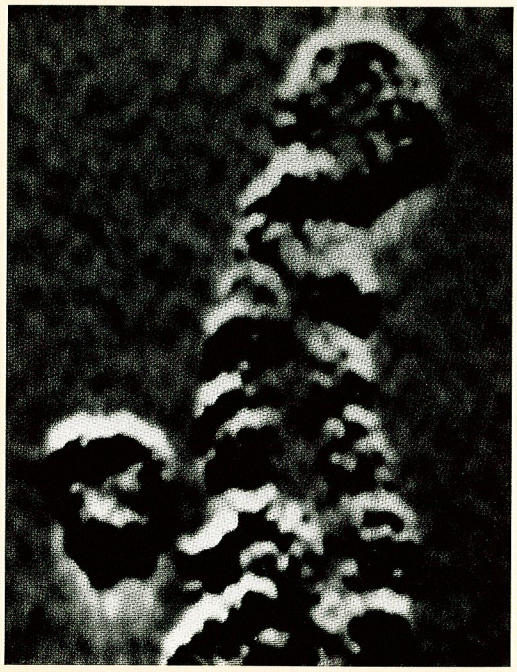




Data access through interactive workstations. *One of the great challenges of the human genome project is coordination of effort. A particularly vexing problem is how to integrate and provide access to the growing mass of genomic data. One solution is a sophisticated workstation, now under development, that would provide a uniform user interface with all mapping and sequencing data bases. With the prototype developed at LBL's Human Genome Center, the user examines data at increasing resolution by "enlarging" selected regions of successive displays. The illustrations here show displays of the full complement of chromosomes, a single chromosome with mapped gene locations, and the sequence for a selected region.*

FIGURE 10

A final aspect of LBL's Human Genome Center is its openness to revolutionary developments, which, by their very nature, cannot be anticipated. One technology that may presage such a revolution is the *scanning tunneling microscope,* or *STM.* By passing a fine stylus over a sample on a conductive substrate and monitoring the amount by which the stylus tip must be retracted to maintain a constant current through the sample, the STM has provided topographic images of material surfaces at unprecedented resolutions. Recently, in a collaboration between scientists at LBL and the Lawrence Livermore National Laboratory, this technique was used to image a single, unstained DNA molecule (Figure 11). Is it possible that with the STM we will someday be able to "visually" read DNA sequences? We don't know, but the pursuit of such promising technologies is an essential element in the quest to unlock the secrets of the human genome.



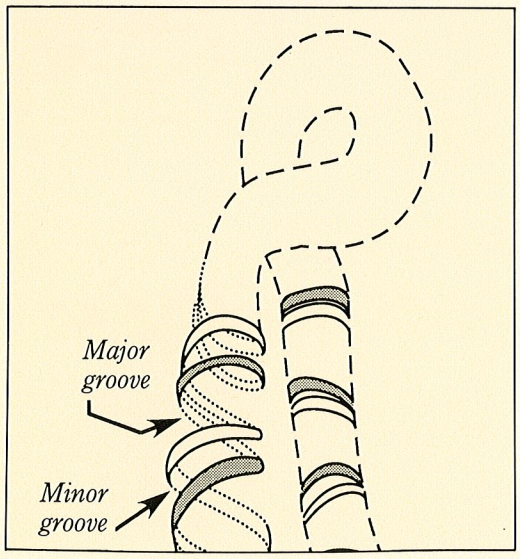


FIGURE 11

Imaging by scanning tunneling microscopy.
In a collaborative effort with the Lawrence Livermore National Laboratory, LBL scientists have used STM to image native, unstained DNA molecules under conditions of normal atmospheric pressure. The image here, as illustrated in the schematic depiction below it, is sufficiently resolved to show the major and minor grooves of the DNA double helix. Some researchers have even suggested that with further development STM imaging may itself someday be used to sequence DNA.

LOOKING TO THE FUTURE

Two characteristics especially distinguish LBL's Human Genome Center: effective leadership and closely linked efforts by a multidisciplinary team. The Center is under the direction of Charles R. Cantor, who is also Professor of Molecular Biology at the University of California at Berkeley. Dr. Cantor is one of the pioneers of modern molecular genetics, the developer of pulsed-field gel electrophoresis, an advisor to the DOE and to private industry, and a member of the executive council of the international Human Genome Organization (HUGO). In addition, the Center's research direction is broadly overseen by the LBL Human Genome Advisory Committee, comprising leading geneticists and molecular biologists from the U.S. and Europe.

The interdisciplinary team is also already in place, as evidenced by many of the early accomplishments described on the previous pages. Contributors to the effort include members of five LBL research divisions: Cell and Molecular Biology, Chemical Biodynamics, Information and Computing Sciences, Engineering, and Materials and Chemical Sciences. In addition, collaborative research will involve faculty members at several University of California campuses and other universities.

Members of the Center, however, remain dispersed, pending construction of a home for their efforts. A Human Genome Laboratory has thus been proposed to provide not only the physical space for this growing effort but also the fertile, interactive environment essential to productive multidisciplinary research. This integration of molecular genetics, engineering, and computer science into a critical research mass is a key ingredient of the Human Genome Center—and of the DOE's strategy for resource and technology development. The new building, which would provide lab and office space for about 60 professional and support personnel, would also provide a core facility for the common use of expensive instrumentation and technologies (Figure 12). It would thus serve as an important user facility for the biomedical scientific community committed to this DOE-inspired national project.

The proposed Human Genome Laboratory.

This 32,000-square-foot building at LBL would provide the lab and office space needed for the anticipated increase in effort at the Human Genome Center. It would also serve as a user facility for the biomedical scientific community and would ensure a critical interactive mass of researchers in engineering and the physical and biological sciences.

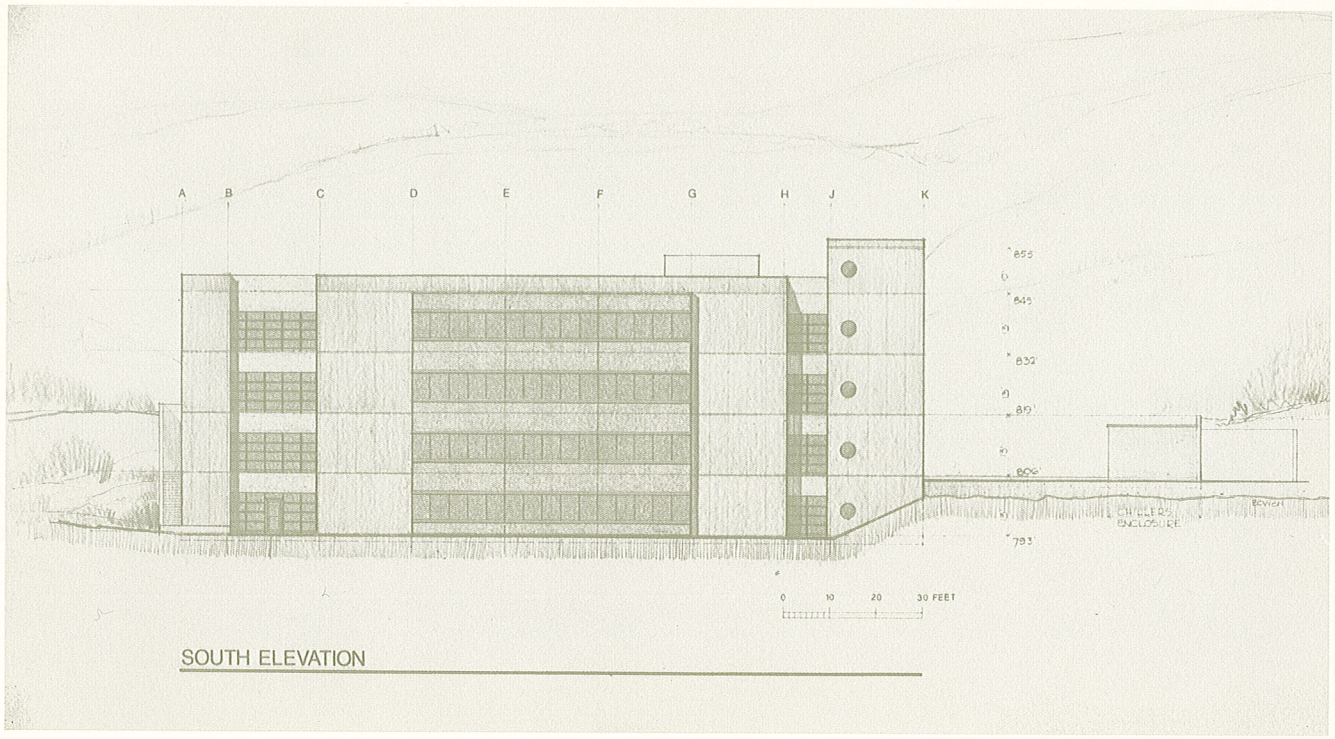

A
B
C
D
E
F
G
H
J
K
855
845
832
819
806
793
CHILLERS ENCLOSURE
0
10
20
30 FEET
SOUTH ELEVATION

FIGURE 12

This document was prepared as an account of work sponsored by the United States Government. Neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California and shall not be used for advertising or product endorsement purposes.

Lawrence Berkeley Laboratory is an equal opportunity employer. Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098. PUB-657/6-89.

HUMAN GENOME CENTER
LAWRENCE BERKELEY LABORATORY
ONE CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720