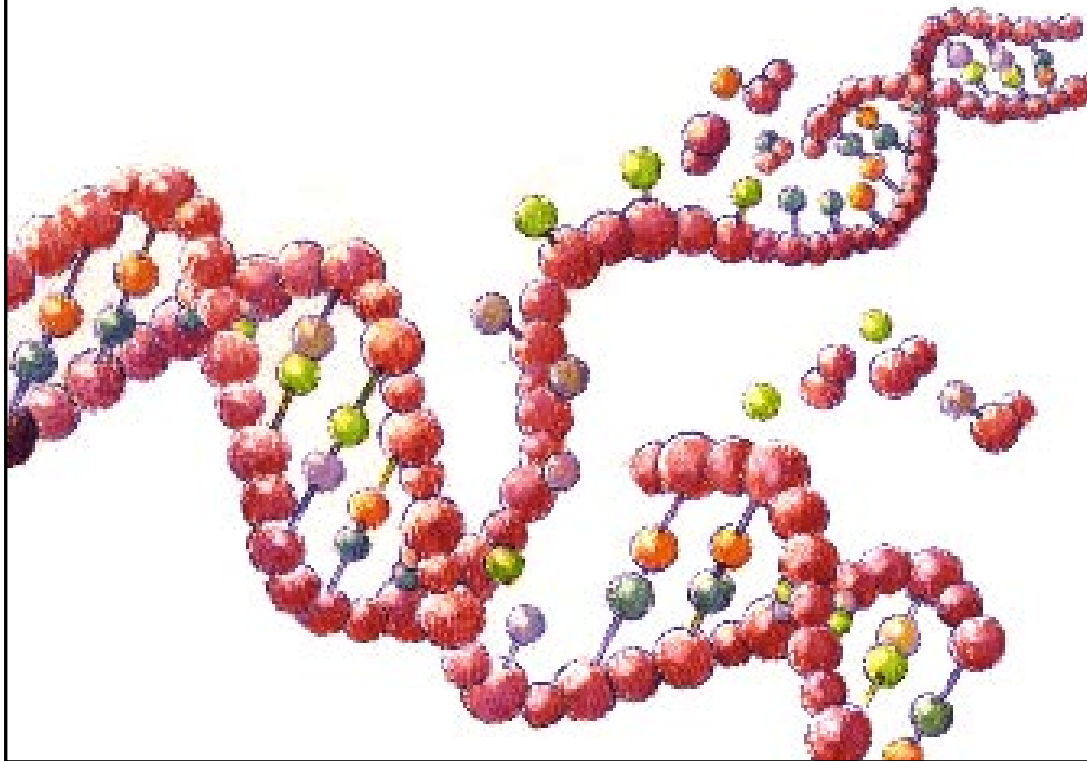


TO KNOW *OURSELVES*

—◆—
**THE U.S. DEPARTMENT OF ENERGY
AND
THE HUMAN GENOME PROJECT**

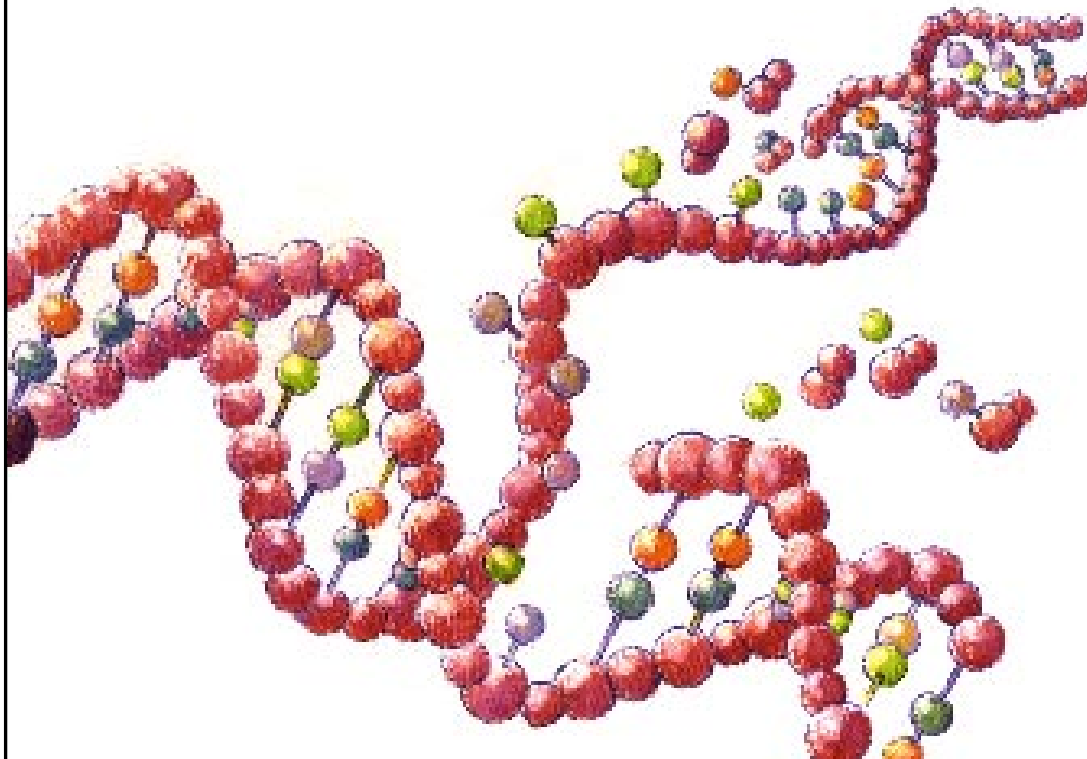


JULY 1996

TO KNOW *OURSELVES*



**THE U.S. DEPARTMENT OF ENERGY
AND
THE HUMAN GENOME PROJECT**



JULY 1996



Contents

FOREWORD	2
THE GENOME PROJECT—WHY THE DOE?	4
<i>A bold but logical step</i>	
INTRODUCING THE HUMAN GENOME	6
<i>The recipe for life</i>	
Some definitions	6
A plan of action	8
EXPLORING THE GENOMIC LANDSCAPE	10
<i>Mapping the terrain</i>	
Two giant steps: Chromosomes 16 and 19	12
Getting down to details: Sequencing the genome	16
Shotguns and transposons	20
How good is good enough?	26
Sidebar: Tools of the Trade	17
Sidebar: The Mighty Mouse	24
BEYOND BIOLOGY	27
<i>Instrumentation and informatics</i>	
Smaller is better—And other developments	27
Dealing with the data	30
ETHICAL, LEGAL, AND SOCIAL IMPLICATIONS	32
<i>An essential dimension of genome research</i>	



Foreword

AT THE END OF THE ROAD in Little Cottonwood Canyon, near Salt Lake City, Alta is a place of near-mythic renown among skiers. In time it may well assume similar status among molecular geneticists. In December 1984, a conference there, co-sponsored by the U.S. Department of Energy, pondered a single question: Does modern DNA research offer a way of detecting tiny genetic mutations—and, in particular, of observing any increase in the mutation rate among the survivors of the Hiroshima and Nagasaki bombings and their descendants? In short the answer was, Not yet. But in an atmosphere of rare intellectual fertility, the seeds were sown for a project that would make such detection possible in the future—the Human Genome Project.

In the months that followed, much deliberation and debate ensued. But in 1986, the DOE took a bold and unilateral step by announcing its Human Genome Initiative, convinced that its mission would be well served by a comprehensive picture of the human genome. The immediate response was considerable skepticism—skepticism about the scientific community's technological wherewithal for sequencing the genome at a reasonable cost and about the value of the result, even if it could be obtained economically.

Things have changed. Today, a decade later, a worldwide effort is under way to develop and apply the technologies needed to completely map and sequence the human genome, as well as the genomes of several model organisms. Technological progress

has been rapid, and it is now generally agreed that this international project will produce the complete sequence of the human genome by the year 2005.

And what is more important, the value of the project also appears beyond doubt. Genome research is revolutionizing biology and biotechnology, and providing a vital thrust to the increasingly broad scope of the biological sciences. The impact that will be felt in medicine and health care alone, once we identify all human genes, is inestimable. The project has already stimulated significant investment by large corporations and prompted the creation of new companies hoping to capitalize on its profound implications.

But the DOE's early, catalytic decision deserves further comment. The organizers of the DOE's genome initiative recognized that the information the project would generate—both technological and genetic—would contribute not only to a new understanding of human biology, but also to a host of practical applications in the biotechnology industry and in the arenas of agriculture and environmental protection. A 1987 report by a DOE advisory committee provided some examples. The committee foresaw that the project could ultimately lead to the efficient production of biomass for fuel, to improvements in the resistance of plants to environmental stress, and to the practical use of genetically engineered microbes to neutralize toxic wastes. The Department thus saw far more to the genome project than a promised tool for assessing mutation rates. For example, understanding the human genome will have an enormous impact on our ability to assess,

individual by individual, the risk posed by environmental exposures to toxic agents. We know that genetic differences make some of us more susceptible, and others more resistant, to such agents. Far more work must be done before we understand the genetic basis of such variability, but this knowledge will directly address the DOE's long-term mission to understand the effects of low-level exposures to radiation and other energy-related agents—especially the effects of such exposure on cancer risk. And the genome project is a long stride toward such knowledge.

The Human Genome Project has other implications for the DOE as well. In 1994, taking advantage of new capabilities developed by the genome project, the DOE formulated 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, and industrial processing. As a result of this initiative, we already have complete sequences for two microbes that live under extreme conditions of temperature and pressure. Structural studies are under way to learn what is unique about the proteins of these organisms—the aim being ultimately to engineer these microbes and their enzymes for such practical purposes as waste control and environmental cleanup. (DOE-funded genetic engineering of a thermostable DNA polymerase has already produced an enzyme that has captured a large share of the several-hundred-million-dollar DNA polymerase market.)

And other little-studied microbes hint at even more intriguing possibilities. For instance, *Deinococcus radiodurans* is a species that prospers even when exposed to huge doses of ionizing radiation. This microbe has an amazing ability to repair radiation-induced damage to its DNA. Its genome is currently being sequenced with DOE support, with the hope of understanding and ultimately taking practical advantage of its unusual capabilities. For example, it might be possible to insert foreign DNA into this microbe that allows it to digest toxic organic

components found in highly radioactive waste, thus simplifying the task of further cleanup. Another approach might be to introduce metal-binding proteins onto the microbe's surface that would scavenge highly radioactive isotopes out of solution.

Biotechnology, fueled in part by insights reaped from the genome project, will also play a significant role in improving the use of fossil-based resources. Increased energy demands, projected over the next 50 years, require strategies to circumvent the many problems associated with today's dominant energy systems. Biotechnology promises to help address these needs by upgrading the fuel value of our current energy resources and by providing new means for the bioconversion of raw materials to refined products—not to mention offering the possibility of entirely new biomass-based energy sources.

We have thus seen only the dawn of a biological revolution. The practical and economic applications of biology are destined for dramatic growth. Health-related biotechnology is already a multibillion-dollar success story—and is still far from reaching its potential. Other applications of biotechnology are likely to beget similar successes in the coming decades. Among these applications are several of great importance to the DOE. We can look to improvements in waste control and an exciting era of environmental bioremediation; we will see new approaches to improving energy efficiency; and we can even hope for dramatic strides toward meeting the fuel demands of the future. The insights, the technologies, and the infrastructure that are already emerging from the genome project, together with advances in fields such as computational and structural biology, are among our most important tools in addressing these national needs.



Aristides A. N. Patrinos
Director, Human Genome Project
U.S. Department of Energy



The Genome Project— Why the DOE?

A BOLD BUT LOGICAL STEP

THE BIOSCIENCES RESEARCH community is now embarked on a program whose boldness, even audacity, has prompted comparisons with such visionary efforts as the Apollo space program and the Manhattan project. That life scientists should conceive such an ambitious project is not remarkable; what is surprising—at least at first blush—is that the project should trace its roots to the Department of Energy.

For close to a half-century, the DOE and its governmental predecessors have been charged with pursuing a deeper understanding of the potential health risks posed by energy use and by energy-production technologies—with special interest focused on the effects of radiation on humans. Indeed, it is fair to say that most of what we know today about radiological health hazards stems from studies supported by these government agencies. Among these investigations are long-standing studies of the survivors of the atomic bombings of Hiroshima and Nagasaki, as well as any number of experimental studies using animals, cells

in culture, and nonliving systems. Much has been learned, especially about the consequences of exposure to high doses of radiation. On the other hand, many questions remain unanswered; in particular, we have

much to learn about how low doses produce their insidious effects. When present merely in low but significant amounts, toxic agents such as radiation or mutagenic chemicals work their mischief in the most subtle ways, altering only slightly the genetic instructions in our cells. The consequences can be heritable mutations too slight to produce discernible effects in a generation or two but, in their persistence and irreversibility, deeply troublesome nonetheless.

Until recently, science offered little hope for detecting at first hand these tiny changes to the DNA that encodes our genetic program. Needed was a tool that could detect a change in one “word” of the program, among perhaps a hundred million. Then, in 1984, at a meeting convened jointly by the DOE and the International Commission for Protection Against Environmental Mutagens and Carcinogens, the question was first seriously asked: Can we, should we, sequence the human genome? That is, can we develop the technology to obtain a word-by-word copy of the entire genetic script for an “average” human being, and thus to establish a benchmark for detecting the elusive mutagenic effects of radiation and cancer-causing toxins? Answering such a question was not simple. Workshops were convened in 1985 and 1986; the issue was studied by a DOE advisory group, by the Congressional Office of Technology Assessment, and by the National Academy of Sciences; and the matter was debated publicly and privately among biologists themselves. In the end, however, a consensus emerged that we should make a start.

*In 1986
the DOE
was the first
federal agency
to announce
an initiative
to pursue a
detailed under-
standing of the
human genome.*

Adding impetus to the DOE's earliest interest in the human genome was the Department's stewardship of the national laboratories, with their demonstrated ability to conduct large multidisciplinary projects—just the sort of effort that would be needed to develop and implement the technological know-how needed for the Human Genome Project. Biological research programs already in place at the national labs benefited from the contributions of engineers, physicists, chemists, computer scientists, and mathematicians, working together in teams. Thus, with the infrastructure in place and with a particular interest in the ultimate results, the Department of Energy, in 1986, was the first federal agency to announce and to fund an initiative to pursue a detailed understanding of the human genome.

Of course, interest was not restricted to the DOE. Workshops had also been sponsored by the National Institutes of Health, the Cold Spring Harbor Laboratory, and the Howard Hughes Medical Institute. In 1988 the NIH joined in the pursuit, and in the fall of that year, the DOE and the NIH signed a memorandum of understanding that laid the foundation for a concerted interagency effort. The basis for this community-wide excitement is not hard to comprehend. The first impulse behind the DOE's commitment was only one of many reasons for coveting a deeper insight into the human genetic script. Defective genes directly account for an estimated 4000 hereditary human diseases—maladies such as Huntington disease and cystic fibrosis. In some such cases, a single misplaced letter among three billion can have lethal consequences. For most of us, though, even greater interest focuses on the far more common ailments in which altered genes influence but do not prescribe. Heart disease, many cancers, and some psychiatric disorders, for example, can emerge from complicated interplays of environmental factors and genetic misinformation.

The first steps in the Human Genome Project are to develop the needed technologies, then to “map” and “sequence” the

genome. But in a sense, these well-publicized efforts aim only to provide the raw material for the next, longer strides. The ultimate goal is to exploit those resources for a truly profound molecular-level understanding of how we develop from embryo to adult, what makes us work, and what causes things to go wrong. The benefits to be reaped stretch the imagination. In the offing is a new era of molecular medicine characterized not by treating symptoms, but rather by looking to the deepest causes of disease. Rapid and more accurate diagnostic tests will make possible earlier treatment for countless maladies. Even more promising, insights into genetic susceptibilities to disease and to environmental insults, coupled with preventive therapies, will thwart some diseases altogether. New, highly targeted pharmaceuticals, not just for heritable diseases, but for communicable ailments as well, will attack diseases at their molecular foundations. And even gene therapy will become possible, in some cases actually “fixing” genetic errors. All of this in addition to a new intellectual perspective on who we are and where we came from.

The Department of Energy is proud to be playing a central role in propelling us toward these noble goals. 🌞



Introducing the Human Genome

THE RECIPE FOR LIFE

FOR ALL THE DIVERSITY of the world's five and a half billion people, full of creativity and contradictions, the machinery of every human mind and body is built and run with fewer than 100,000 kinds of protein molecules. And 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. In a material sense, then, all of the subtlety of our species, all of our art and science, is ultimately accounted for by a surprisingly small set of discrete genetic instructions. More surprising still, the differences between two unrelated individuals, between the man next door and Mozart, may reflect a mere handful of differences in their genomic recipes—perhaps one altered word in five hundred. We are far more alike than we are different. At the same time, there is room for near-infinite variety.

It is no overstatement to say that to decode our 100,000 genes in some fundamental way would be an epochal step toward unraveling the manifold mysteries of life.

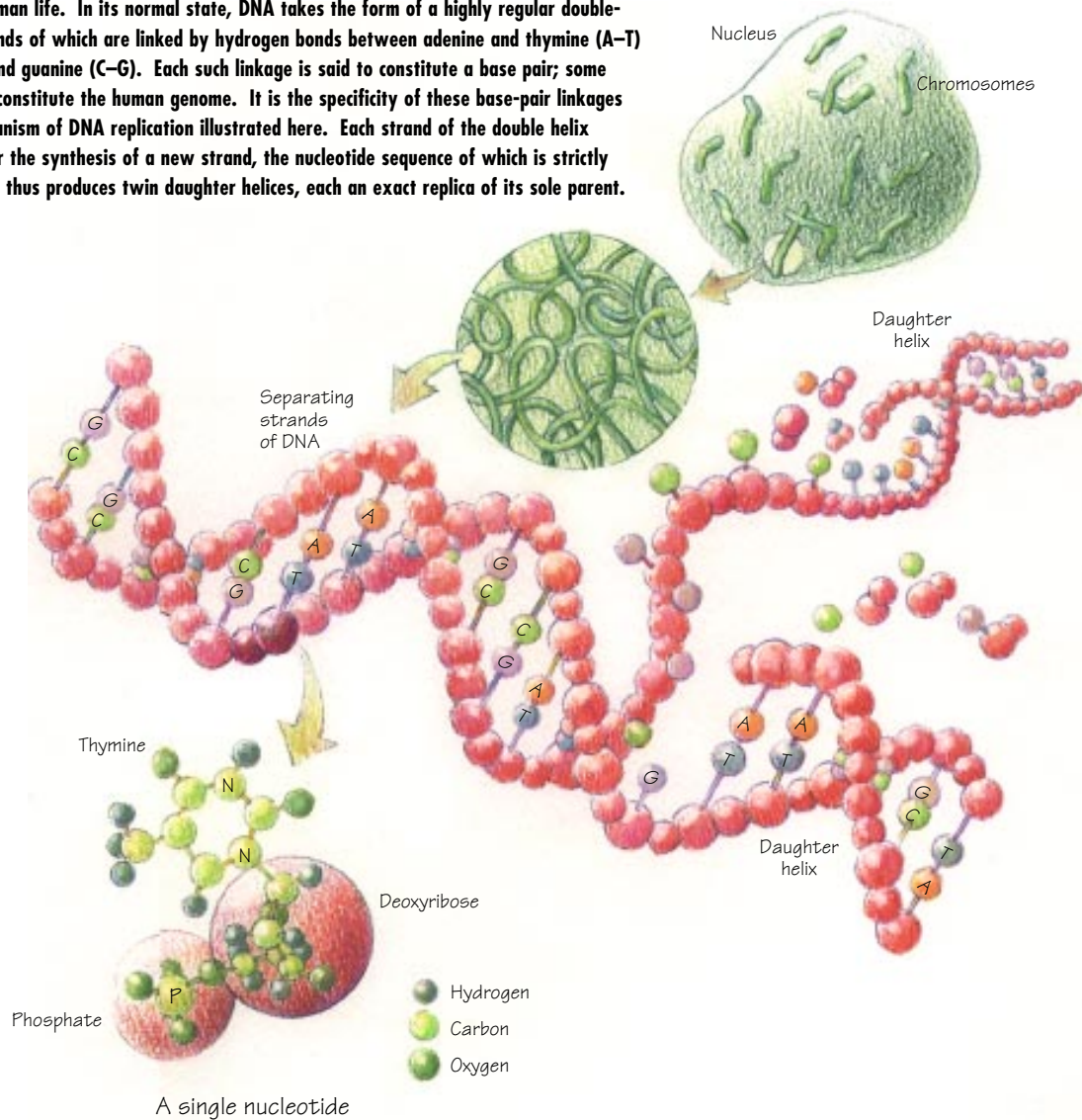
SOME DEFINITIONS

The *human genome* is the full complement of genetic material in a human cell. (Despite five and a half billion variations on a theme, the differences from one genome to the next are minute; hence, we hear about *the* human genome—as if there were only one.) The genome, in turn, is distributed among 23 sets of *chromosomes*, which, in each of us, have been replicated and re-replicated since the

fusion of sperm and egg that marked our conception. The source of our personal uniqueness, our full genome, is therefore preserved in each of our body's several trillion cells. At a more basic level, the genome is DNA, deoxyribonucleic acid, a natural polymer built up of repeating *nucleotides*, each consisting of a simple sugar, a phosphate group, and one of four nitrogenous bases. The hierarchy of structure from chromosome to nucleotide is shown in 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, if we count 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: One helix begets two with near-perfect fidelity. Second, by a similar template-based process, depicted in Figure 2, a means is also available for producing a DNA-like messenger to the cell cytoplasm. There, this *messenger RNA*, the faithful complement of a particular DNA segment, directs the synthesis of a particular protein. Many subtleties are entailed in the synthesis of proteins, but in a schematic sense, the process is elegantly simple.

FIGURE 1. SOME DNA DETAILS. Apart from reproductive gametes, each cell of the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA. Every strand of the DNA is a huge natural polymer of repeating nucleotide units, each of which comprises a phosphate group, a sugar (deoxyribose), and a base (either adenine, thymine, cytosine, or guanine). Every strand thus embodies a code of four characters (A's, T's, C's, and G's), the recipe for the machinery of human life. 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 adenine and thymine (A-T) and between cytosine and guanine (C-G). 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. Replication thus produces twin daughter helices, each an exact replica of its sole parent.



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 particular sequence of amino acids. This is the role of the messenger RNA. (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 AAU thus dictates that the amino acid asparagine should be added to a polypeptide chain, GCA specifies alanine—and so on. A segment of the chromosomal DNA that directs the synthesis of a single type of protein constitutes a single *gene*.

A PLAN OF ACTION

In 1990 the Department of Energy and the National Institutes of Health developed a joint research plan for their genome programs, outlining specific goals for the ensuing five years. Three years later, emboldened by progress that was on track or even ahead of schedule, the two agencies put forth an updated five-year plan. Improvements in technology, together with the experience of three years, allowed an even more ambitious prospect.

In broad terms, the revised plan includes goals for genetic and physical mapping of the genome, DNA sequencing, identifying and locating genes, and pursuing further developments in technology and informatics. To a large extent, the following pages are devoted to a discussion of just what these goals mean, and what part the DOE is playing in pursuing them. In addition, the plan emphasizes the continuing importance of the ethical, legal, and social implications of genome research, and it underscores the critical roles of scientific training, technology transfer, and public access to research data and materials. Most of the goals focus on the human genome, but the importance of continuing research on widely studied “model organisms” is also explicitly recognized.

Among the scientific goals of human genome research, several are especially notable, as they provide clear milestones for future progress. In reciting them, however, it is important to note an underlying assumption of adequate research support. Such support is obviously crucial if the joint plan is to succeed. Some of the central goals for 1993–98 follow:

The plan includes goals for genetic and physical mapping, DNA sequencing, identifying and locating genes, and pursuing further developments in technology and informatics.

- ◆ Complete a genetic linkage map at a resolution of two to five centimorgans by 1995—As discussed on page 10, this goal was far surpassed by the fall of 1994.
- ◆ Complete a physical map at a resolution of 100 kilobases by 1998—This implies a genome map with 30,000 “signposts,” separated by an average of 100,000 base pairs. Further, each signpost will be a *sequence-tagged site*, a stretch of DNA with a unique and well-defined DNA sequence. Such a map will greatly facilitate “production sequencing” of the entire genome. By the end of 1995, molecular biologists were halfway to this goal: A physical map was announced with 15,000 sequence-tagged signposts. Physical mapping is discussed on pages 10–16.
- ◆ By 1998 develop the capacity to sequence 50 million base pairs per year in long continuous segments—Adequate fiscal investment and continuing progress beyond 1998 should then produce a fully sequenced human genome by the year 2005 or earlier. Sequencing is the subject of pages 16–26.
- ◆ Develop efficient methods for identifying and locating known genes on physical maps or sequenced DNA—The goals here are less quantifiable, but the aim is central to the Human Genome Project: to home in on and ultimately to understand the most important human genes, namely, the ones responsible for serious diseases and those crucial for healthy development and normal functions.
- ◆ Pursue technological developments in areas such as automation and robotics—A continuing emphasis on technological advance is critical. Innovative technologies, such as those described on pages 27–30, are the necessary underpinnings of future large-scale sequencing efforts.
- ◆ Continue the development of database tools and software for managing and interpreting genome data—This is the area of informatics, discussed on pages 30–31. The challenge is not so much the volume of data, but rather the need to

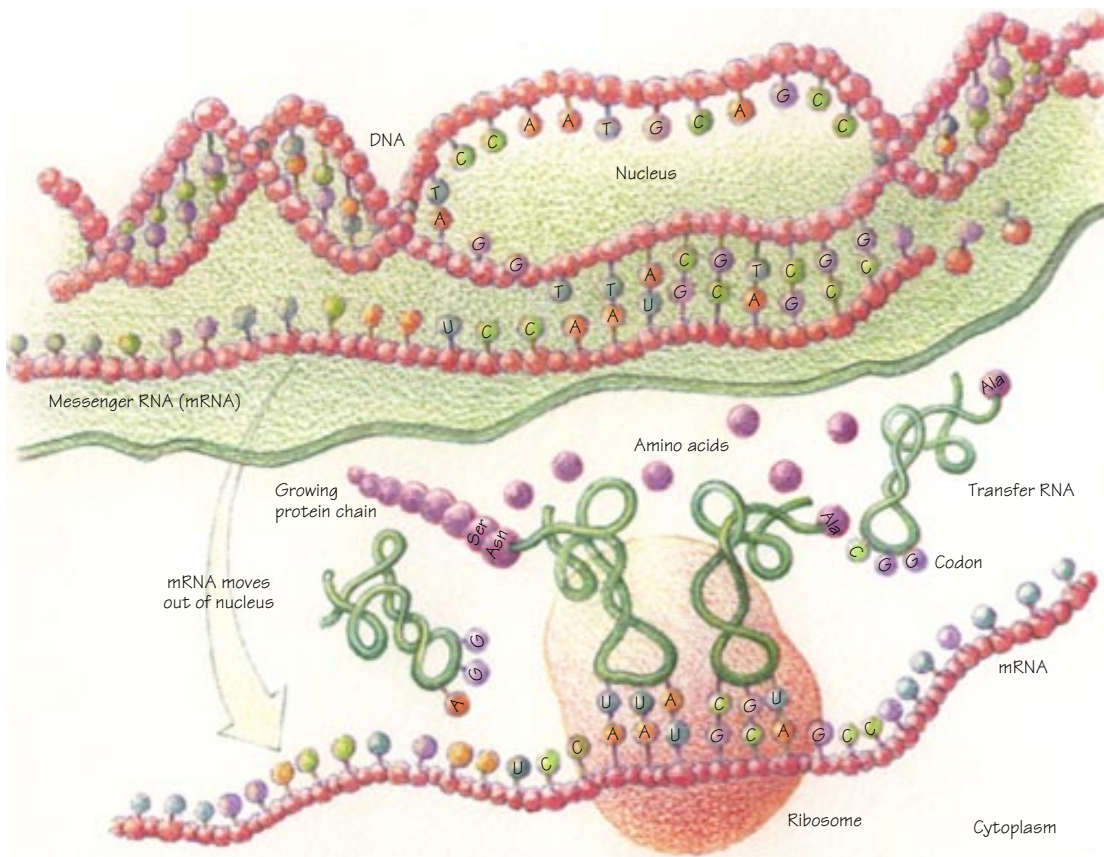


FIGURE 2. FROM GENES TO PROTEINS. In the cell nucleus, RNA is produced by transcription, in much the same way that DNA replicates itself. RNA, however, substitutes the sugar ribose for deoxyribose and the base uracil for thymine, and is usually single-stranded. One form of RNA, messenger RNA or mRNA, 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) containing the complementary three-base sequence. This tRNA, in turn, transfers a single amino acid to a growing protein chain. Each codon thus unambiguously directs the addition of one amino acid to the protein. On the other hand, the same amino acid can be added by different codons; in this illustration, the mRNA sequences GCA and GCC are both specifying the addition of the amino acid alanine (Ala).

mount a system compatible with researchers around the world, and one that will allow scientists to contribute new data and to freely interrogate the existing databases. The ultimate measure of success will be the ease with which biologists can fruitfully use the information produced by the genome project.

- ◆ Continue to explore the ethical, legal, and social implications of genome research—Much emphasis continues to be placed on issues of privacy and the fair use of genetic information. New goals focus on defining additional pertinent issues and

developing policy responses to them, disseminating policy options regarding genetic testing services, fostering greater acceptance of human genetic variation, and enhancing public and professional education that is sensitive to sociocultural and psychological issues. This side of the genome project is discussed on pages 32–33. 🌞



Exploring the Genomic Landscape

MAPPING THE TERRAIN

ONE OF THE CENTRAL GOALS of the Human Genome Project is to produce a detailed “map” of the human genome. But, just as there are topographic maps and political maps and highway maps of the United States, so there are different kinds of genome maps, the variety of which is suggested in Figure 3. One type, a *genetic linkage map*, is based on careful analyses of human inheritance patterns. It indicates

Just as there are topographic maps and political maps and highway maps, so there are different kinds of genome maps.

for each chromosome the whereabouts of genes or other “heritable markers,” with distances measured in centimorgans, a measure of recombination frequency. During the formation of sperm and egg cells, a process of genetic recombination—or “crossing over”—occurs in which pieces of genetic material are swapped between paired chromosomes. This process of chromosomal scrambling accounts for the differences invariably seen even in siblings (apart from identical twins). Logically, the closer two genes are to each other on a single chromosome, the less likely they are to get split up during genetic recombination. When they are close enough that the chances of being separated are only one in a hundred, they are said to be separated by a distance of one centimorgan.

The role of human pedigrees now becomes clear. By studying family trees and tracing the inheritance of diseases and physical traits, or even unique segments of DNA identifiable only in the laboratory, geneticists can begin to pin down the relative positions of these genetic markers. By the end of 1994, a comprehensive map was available that included more than 5800 such markers, including genes implicated in cystic fibrosis, myotonic dystrophy, Huntington disease, Tay-Sachs disease, several cancers, and many other maladies. The average gap between markers was about 0.7 centimorgan.

Other maps are known as *physical maps*, so called because the distances between features are measured not in genetic terms, but in “real” physical units, typically, numbers of base pairs. A close analogy can thus be drawn between physical maps and the road maps familiar to us all. Indeed, the analogy can be extended further. Just as small-scale road maps may show only large cities and indicate distances only between major features, so a low-resolution physical map includes only a relative sprinkling of chromosomal landmarks. A well-known low-resolution physical map, for example, is the familiar chromosomal map, showing the distinctive staining patterns that can be seen in the light microscope. Further, by a process known as *in situ hybridization*, specific segments of DNA can be targeted in intact chromosomes by using complementary strands synthesized in the laboratory. These laboratory-made “probes” carry a fluorescent or radioactive

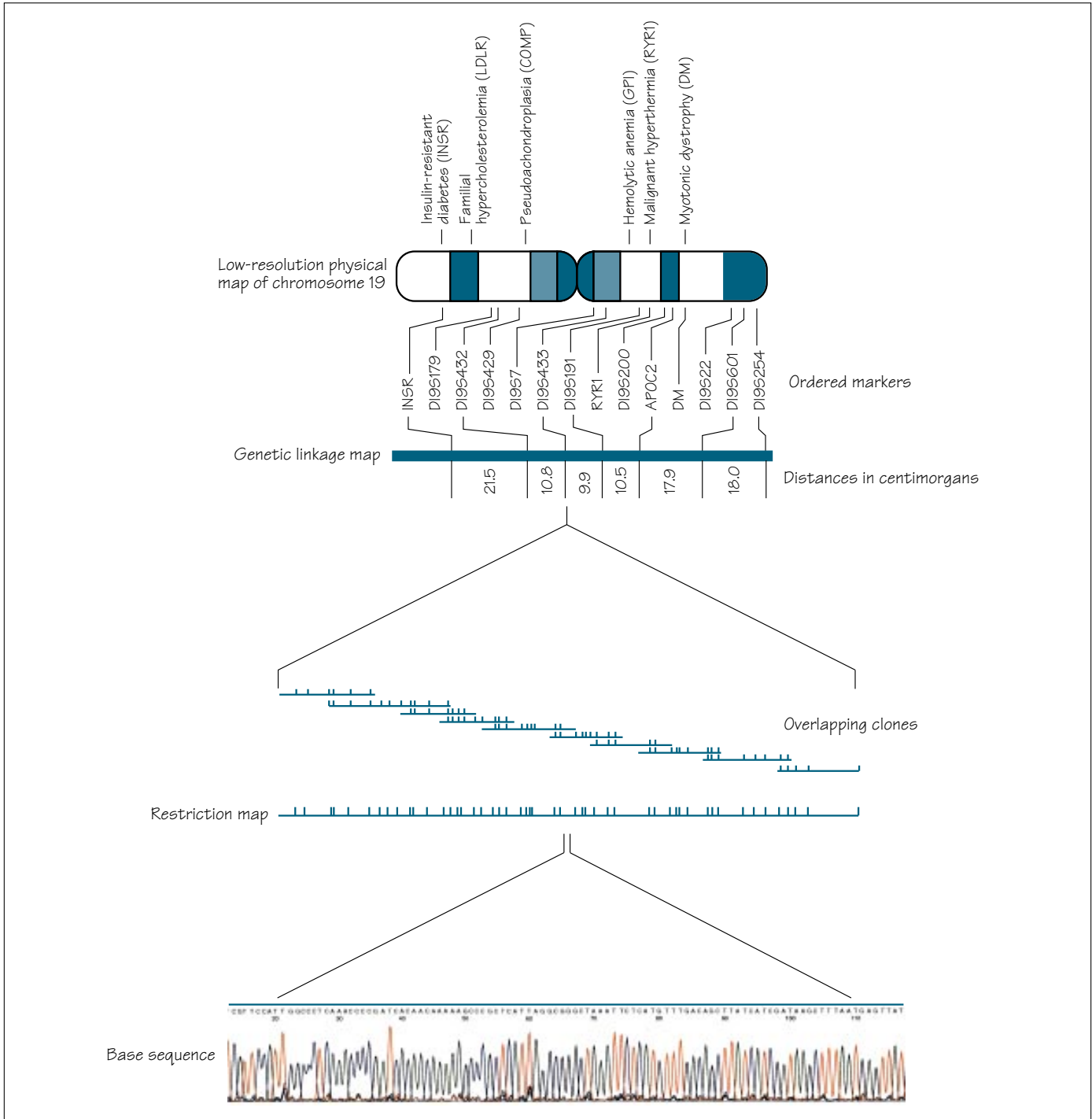


FIGURE 3. GENOMIC GEOGRAPHY. The human genome can be mapped in a number of ways. The familiar and reproducible banding pattern of the chromosomes constitutes one kind of physical map, and in many cases, the positions of genes or other heritable markers have been localized to one band or another. More useful are genetic linkage maps, on which the relative positions of markers have been established by studying how frequently the markers are separated during a natural process of chromosomal shuffling called genetic recombination. The cryptically coded ordered markers near the top of this figure are physically mapped to specific regions of chromosome 19; some of them also constitute

a low-resolution genetic linkage map. (Hundreds of genes and other markers have been mapped on chromosome 19; only a few are indicated here. See Figure 5 for a display of mapped genes.) A higher-resolution physical map might describe, as shown here, the cutting sites (the short vertical lines) for certain DNA-cleaving enzymes. The overlapping fragments that allow such a map to be constructed are then the resources for obtaining the ultimate physical map, the base-pair sequence for the human genome. At the bottom of this figure is an example of output from an automatic sequencing machine.

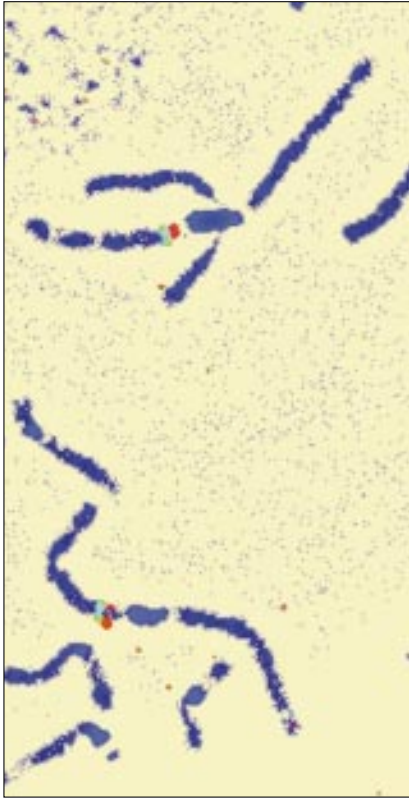


FIGURE 4. FISHING FOR GENES. Fluorescence in situ hybridization (FISH) probes are strands of DNA that have been labeled with fluorescent dye molecules. The probes bind uniquely to complementary strands of chromosomal DNA, thus pinpointing the positions of target DNA sequences. In this example, one probe, whose fluorescence signal is shown in red, binds specifically to a gene (DSRAD) that codes for an important RNA-modifying enzyme. A second probe, whose signal appears in green, binds to a marker sequence whose location was already known. The previously unknown location of the DSRAD gene was thus accurately mapped to a narrow region on the long arm of chromosome 1.

label, which can then be detected and thus pinpointed on a specific region of the chromosome. Figure 4 shows some results of fluorescence in situ hybridization (FISH). Of particular interest are probes known as cDNA (for *complementary DNA*), which are synthesized by using molecules of messenger RNA as templates. These molecules of cDNA thus hybridize to “expressed” chromosomal regions—regions that directly dictate the synthesis of proteins. However, a physical map that depended only on in situ hybridization would be a fairly coarse one. Fluorescent tags on intact chromosomes cannot be resolved into separate spots unless they are two to five million base pairs apart.

Fortunately, means are also available to produce physical maps of much higher resolution—analogueous to large-scale county maps that show every village and farm road, and indicate distances at a similar level of detail. Just such a detailed physical map is one that emerges from the use of *restriction enzymes*—DNA-cleaving enzymes that serve as highly selective microscopic scalpels (see “Tools of

the Trade,” pages 17–19). A typical restriction enzyme known as *EcoRI*, for example, recognizes the DNA sequence GAATTC and selectively cuts the double helix at that site. One use of these handy tools involves cutting up a selected chromosome into small pieces, then cloning and ordering the resulting fragments. The *cloning*, or copying, process is a product of recombinant DNA technology, in which the natural reproductive machinery of a “host” organism—a bacterium or a yeast, for example—replicates a “parasitic” fragment of human DNA, thus producing the multiple copies needed for further study (see “Tools of the Trade”). By cloning enough such fragments, each overlapping the next and together spanning long segments (or even the entire length) of the chromosome, workers can eventually produce an ordered library of clones. Each contiguous block of ordered clones is known as a *contig* (a small one is shown in Figure 3), and the resulting map is a contig map. If a gene can be localized to a single fragment within a contig map, its physical location is thereby accurately pinned down. Further, these conveniently sized clones become resources for further studies by researchers around the world—as well as the natural starting points for systematic sequencing efforts.

TWO GIANT STEPS: CHROMOSOMES 16 AND 19

One of the signal achievements of the DOE genome effort so far is the successful physical mapping of chromosomes 16 and 19. The high-resolution chromosome 19 map, constructed at the Lawrence Livermore National Laboratory, is based on restriction fragments cloned in *cosmids*, synthetic cloning “vectors” modeled after bacteria-infecting viruses known as bacteriophages. Like a phage, a cosmid hijacks the cellular machinery of a bacterium to mass-produce its own genetic material, together with any “foreign” human DNA that has been smuggled into it. The foundation of the chromosome 19 map is a large set of cosmid contigs that were assembled by automated analysis of overlapping

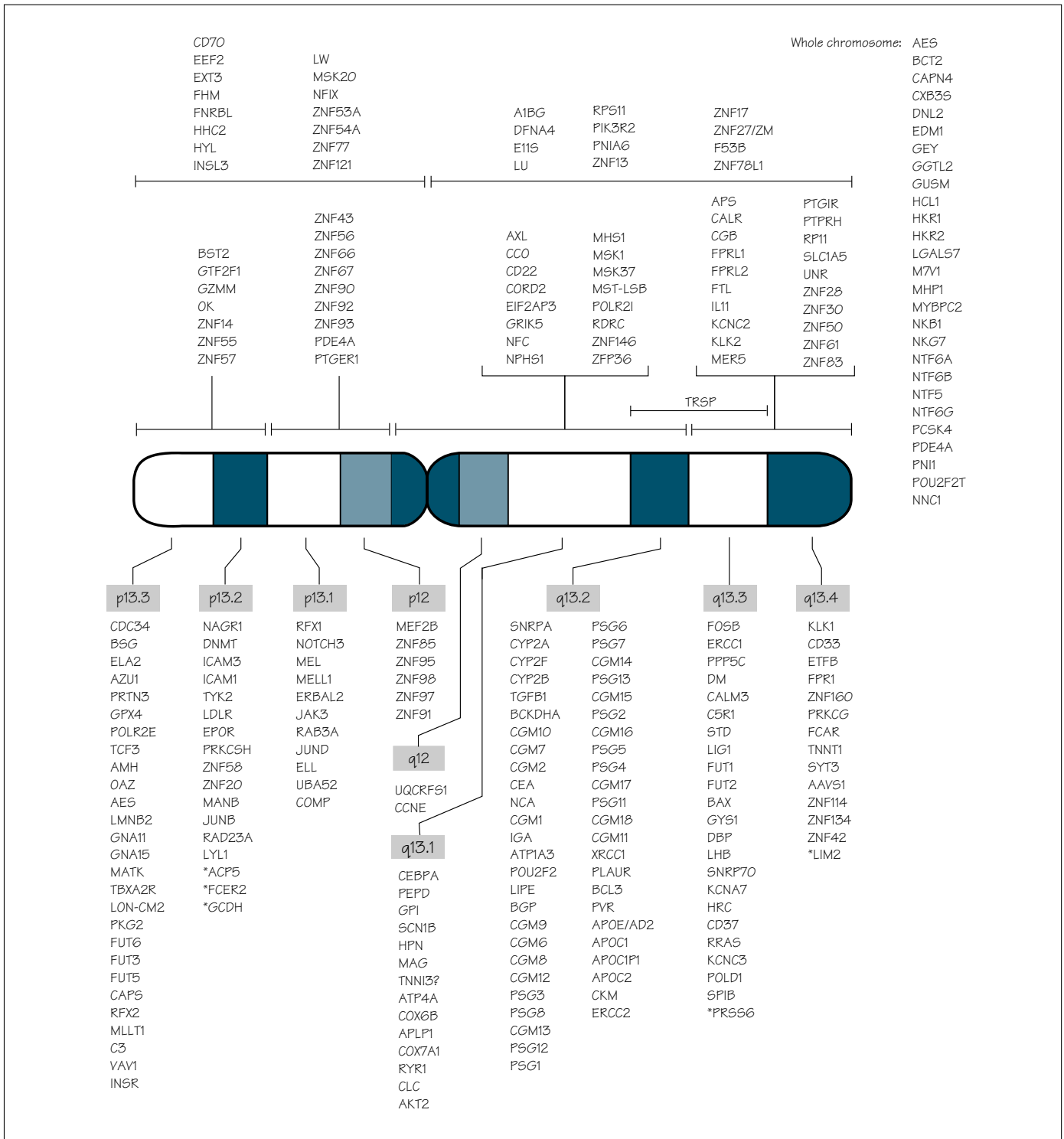


FIGURE 5. AN EMERGING GENE MAP. More than 250 genes have already been mapped to chromosome 19. Those listed on the lower half of this illustration have been assigned to specific cosmids and (except for those marked with asterisks) have been ordered on the Livermore physical map. Their positions are therefore known with far greater accuracy than shown here. The genes listed above the chromosome have been mapped to larger regions of the chromo-

some—or merely localized to chromosome 19 generally—and have not yet been assigned to cosmids in the Livermore database. The text mentions several of the most important genes mapped so far. Others include INSR, which codes for an insulin receptor and is involved in adult-onset diabetes; LDLR, a gene for a low-density lipoprotein receptor involved in hypercholesterolemia; and ERCC2, a DNA repair gene implicated in one form of xeroderma pigmentosum.

but unordered restriction fragments. These contigs span an estimated 54 million base pairs, more than 95 percent of the chromosome, excluding the centromere.

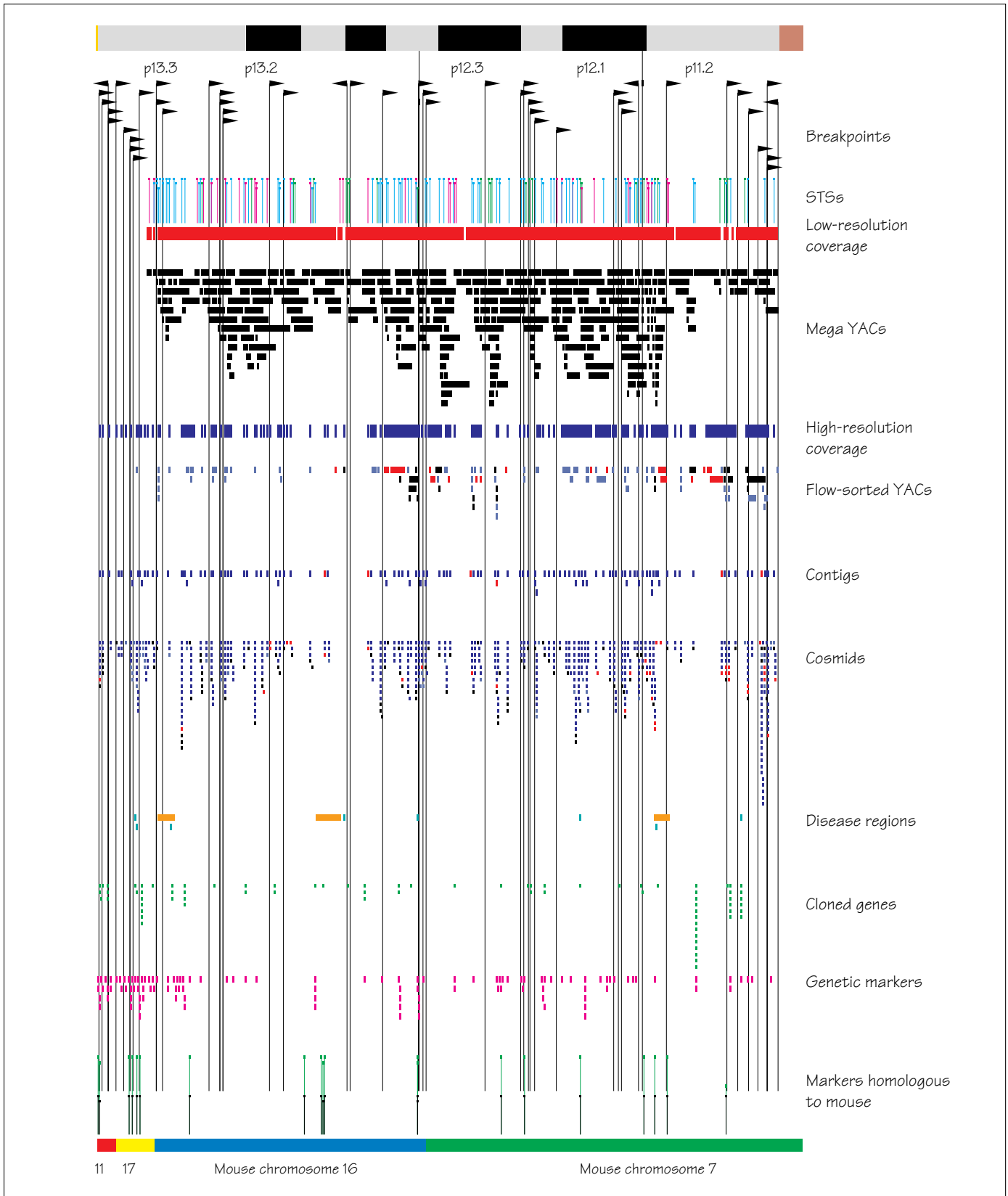
Most of the contigs have been mapped by fluorescence in situ hybridization to visible chromosomal bands. Further, more than 200 cosmids have been more accurately ordered along the chromosome by a high-resolution FISH technique in which the distances between cosmids are determined with a resolution of about 50,000 base pairs. This ordered FISH map, with cosmid reference points separated by an average of 230,000 base pairs, provides the essential framework to which other cosmid contigs can be anchored. Moreover, the *EcoRI* restriction sites have been mapped on more than 45 million base pairs of the overall cosmid map. Over 450 genes and genetic markers have also been localized on this map, of which nearly 300 have been incorporated into the ordered map. Figure 5 shows the locations of the mapped genes. Among these genes is the one responsible for the most common form of adult muscular dystrophy (DM), which was identified in 1992 by an international consortium that included Livermore scientists. A second important disease gene (COMP), responsible for a form of dwarfism known as pseudoachondroplasia, has also been identified. And yet another gene, one linked to a form of congenital kidney disease, has been localized to a single contig spanning one million base pairs, but has not yet been precisely pinpointed. About 2000 other genes are likely to be found eventually on chromosome 19.

In a similar effort, the Los Alamos National Laboratory Center for Human Genome Studies has completed a highly integrated map of chromosome 16, a chromosome that contains genes linked to blood disorders, a second form of kidney disease, leukemia, and breast and prostate cancers. A readable display of this integrated map covers a sheet of paper more than 15 feet long; a portion of it, much reduced and showing only some of its central features, is reproduced here as Figure 6. The framework

for the Los Alamos effort is yet another kind of map, a "cytogenetic breakpoint map" based on 78 lines of cultured cells, each a hybrid that contains mouse chromosomes and a fragment of human chromosome 16. Natural breakpoints in chromosome 16 are thus identified, leading to a breakpoint map that divides the chromosome into segments whose lengths average 1.1 million base pairs. Anchored to this framework are a low-resolution contig map based on YAC clones and a high-resolution contig map based largely on cosmids (for more on YACs, yeast artificial chromosomes, see "Tools of the Trade," pages 17–19). The low-resolution map, comprising 700 YACs from a library constructed by the Centre d'Etude du Polymorphisme Humain (CEPH), provides practically complete coverage of the chromosome, except the highly repetitive DNA in the centromere region. The high-resolution map comprises some 4000 cosmid clones, assembled into about 500 contigs covering 60 percent of the chromosome. In addition, it includes 250 smaller YAC clones that have been merged with the cosmid contig map. The cosmid contig map

FIGURE 6. MAPPING CHROMOSOME 16.

This much-reduced physical map of the short arm of human chromosome 16 summarizes the progress made at Los Alamos toward a complete map of the chromosome. A legible, fully detailed map of the chromosome is more than 15 feet long; only a few features of the map can be described here. Just below the schematic chromosome, the black arrowheads and the vertical lines extending the full length of the page signify "breakpoints" and indicate the portions of the chromosome maintained in separate cell cultures. The cultured portions typically extend from a breakpoint to one end of the chromosome. These breakpoints establish the framework for the Los Alamos mapping effort. Within this framework, some 700 megaYACs (shown in black) provide low-resolution coverage for essentially the entire chromosome. Smaller flow-sorted YACs (light blue, red, and black), together with about 4000 cosmids, assembled into about 500 cosmid contigs (blue and red), establish high-resolution coverage for 60% of the chromosome. Sequence-tagged sites (STSs) are shown as colored vertical lines above the megaYACs, and genes (green) and genetic markers (pink) that have been localized only to the breakpoint map are shown near the bottom. Also shown are cloned and uncloned disease regions, as well as those markers whose analogs have been identified among mouse chromosomes (see "The Mighty Mouse," pages 24–25).



is an especially important step forward, since it is a “sequence-ready” map. It is based on bacterial clones that are ideal substrates for DNA sequencing, and further, these clones have been restriction mapped to allow identification of a minimum set of overlapping clones for a large-scale sequencing effort.

These maps are mere stepping stones to the string of three billion characters – A’s, T’s, C’s, and G’s – that defines our species.

The high- and low-resolution maps have been tied together by sequence-tagged sites (STSs), short but unique stretches of DNA sequence. They have also been integrated into the breakpoint map, and with genetic maps developed at the Adelaide Children’s Hospital and by CEPH. The integrated map also includes a transcription map of 1000 sequenced *exons* (expressed fragments of genes) and more

than 600 other markers developed at other laboratories around the world.

GETTING DOWN TO DETAILS: SEQUENCING THE GENOME

Ultimately, though, these physical maps and the clones they point to are mere stepping stones to the most visible goal of the genome project, the string of three billion characters – A’s, T’s, C’s, and G’s – 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 stretches of DNA whose functions we don’t yet know (but which may be involved in such little-understood processes as orchestrating gene expression in different parts of our bodies, at different times of our lives). Should anyone undertake to print it all out, the result would fill several hundred volumes the size of a big-city phone book.

Only the barest start has been made in taking this dramatic step in the Human Genome Project. Several hundred million base pairs have been sequenced and archived in databases, but the great majority of these

are from short “sequence tags” on cloned fragments. Only about 30 million base pairs of human DNA (roughly one percent of the total) have been sequenced in longer stretches, the longest being about 685,000 base pairs long. Even more daunting is the realization that we will eventually need to sequence many parts of the genome many times, thus to reveal differences that indicate various forms of the same gene.

Hence, as with so many human enterprises, the challenge of sequencing the genome is largely one of doing the job cheaper and faster. At the beginning of the project, the cost of sequencing a single base pair was between \$2 and \$10, and one researcher could produce between 20,000 and 50,000 base pairs of continuous, accurate sequence in a year. Sequencing the genome by the year 2005 would therefore likely cost \$10–20 billion and require a dedicated cadre of at least 5000 workers. Clearly, a major effort in technology development was called for—an effort that would drive the cost well below \$1 per base pair and that would allow automation of the sequencing process. From the beginning, therefore, the DOE has emphasized programs to pave the way for expeditious and economical sequencing efforts—programs to develop new technologies, including new cloning vectors, and to establish suitable resources for sequencing, including clone libraries and libraries of expressed sequences.

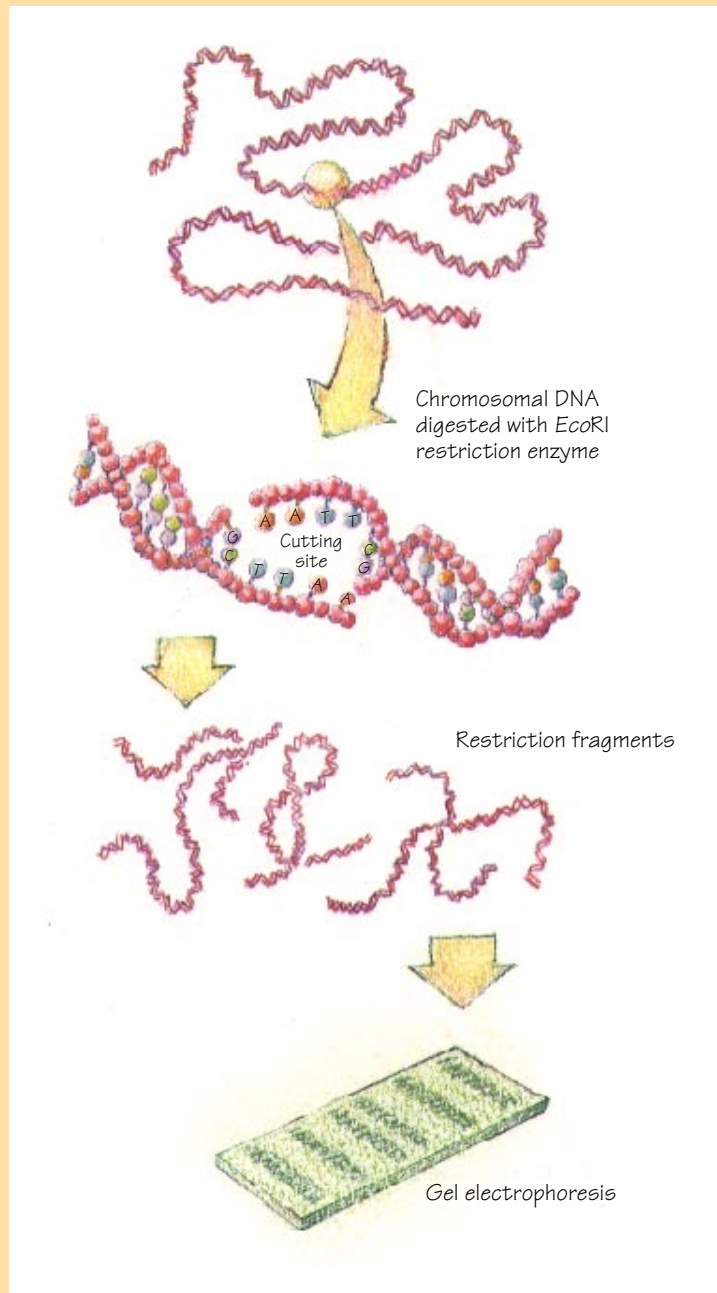
Efforts to develop new cloning vectors have been especially productive. YACs remain a classic tool for cloning large fragments of human DNA, but they are not perfect. Some regions of the genome, for example, resist cloning in YACs, and others are prone to rearrangement. New vectors such as bacterial artificial chromosomes (BACs), P1 phages, and P1-derived artificial cloning systems (PACs) have thus been devised to address these problems. These new approaches are critical for ensuring that the entire genome can be faithfully represented in clone libraries, without the danger of deletions, rearrangements, or spurious insertions. *Continues on p. 20*

Tools of the Trade

Over the next decade, as molecular biologists tackle the task of sequencing the human genome on a massive scale, any number of innovations can be expected in mapping and sequencing technologies. But several of the central tools of molecular genetics are likely to stay with us—much improved perhaps, but not fundamentally different. One such tool is the class of DNA-cutting 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. For example, a restriction enzyme called *EcoRI* recognizes the single-strand sequence GAATTC and invariably cuts the double helix as shown in the illustration on the right.

When digested with a particular restriction enzyme, then, identical segments of human DNA yield identical sets of restriction fragments. On the other hand, DNA from the same genomic region of two different people, with their subtly different genomic sequences, can yield dissimilar sets of fragments, which then produce different patterns when sorted according to size.

This leads directly to discussion of a second essential tool of modern molecular genetics, *gel electrophoresis*, for it is by electrophoresis that DNA fragments of different sizes are most often separated. 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. In 1984 a further advance was made with the invention of pulsed-field gel electrophoresis, in which the strength and direction of the applied field is varied rapidly, thus allowing DNA strands of more than 50,000 base pairs to be separated.



DIGESTING DNA. Isolated from various bacteria, restriction enzymes serve as microscopic scalpels that cut DNA molecules at specific sites. The enzyme *EcoRI*, for example, cuts double-stranded DNA only where it finds the sequence GAATTC. 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 susceptibilities to certain diseases. Knowledge of the cutting sites also yields a kind of physical map known as a restriction map.

A third necessary tool is some means of DNA “amplification.” The classic example is the *cloning vector*, which may be circular DNA molecules derived from bacteria or from bacteriophages (viruslike parasites of bacteria), or artificial chromosomes constructed from yeast or bacterial 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 chromosome—DNA sequences that initiate replication, sequences that mark the ends of the chromosomes, and sequences required for chromosome separation

One of the important achievements of the project has been to establish several libraries of cloned fragments covering the entire human genome.

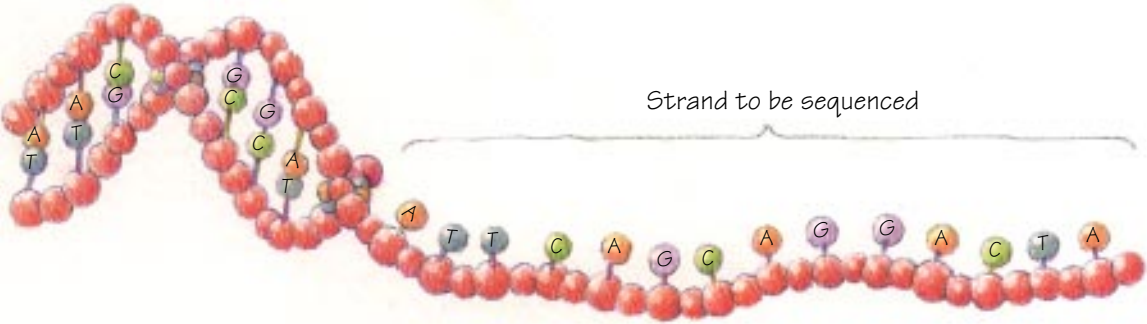
during cell division—then splicing in a fragment of human DNA. This engineered chromosome is then reinserted into a yeast cell, which reproduces the YAC during cell division, as if it were part of the yeast’s normal complement of chromosomes. The result is a colony of yeast cells, each containing a copy, or clone, of the same fragment of human DNA. One of the important achievements of the Human Genome Project has been to establish several libraries of such cloned fragments, using several different vectors (bacterial artificial chromosomes, P1 phages, and P1-derived cloning systems), that cover the entire human genome.

Another way of amplifying DNA is the *polymerase chain reaction*, or PCR. This enzymatic replication technique requires that initiators, or PCR primers, be attached as short complementary strands at the ends of the separated DNA fragments to be replicated. An enzyme then completes the synthesis of the complementary strands, thus dou-

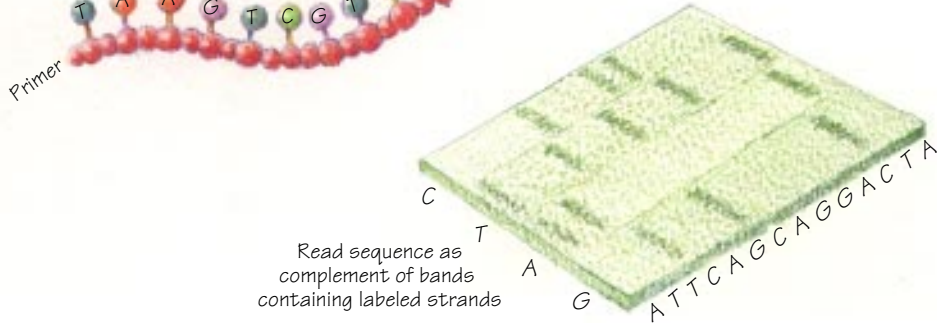
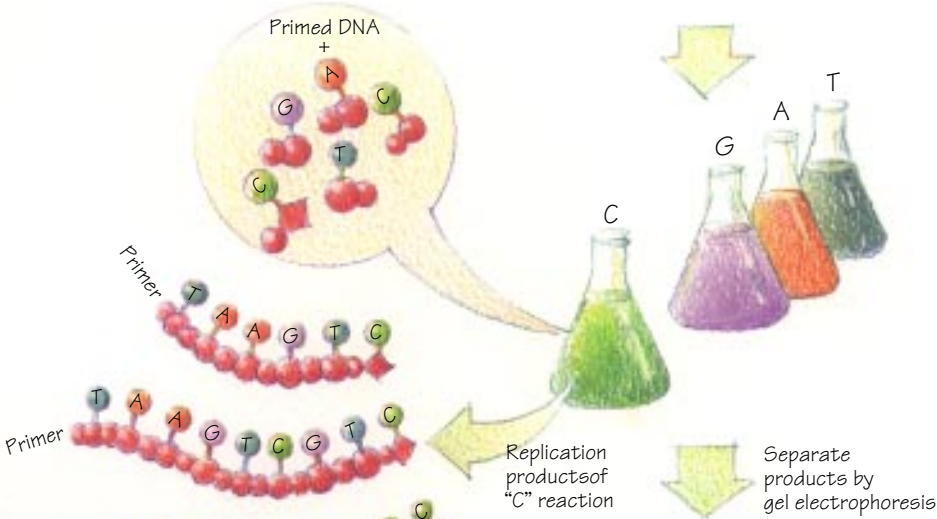
bling the amount of DNA originally present. Again and again, the strands can be separated and the polymerase reaction repeated—so effectively, in fact, that DNA can be amplified by 100,000-fold in less than three hours. As with cloning vectors, the result is a large collection of copies of the original DNA fragment.

When a clone library can be ordered—that is, when the relative positions on the human chromosomes can be established for all the fragments—one then has the perfect resource for achieving the project’s central goal, sequencing the human genome. How the sequencing is actually done can be illustrated by the most popular method in current use, the Sanger procedure, which is depicted schematically on the facing page. The first step is to prime each identical DNA strand in a preparation of cloned fragments. The preparation is then divided into four portions, each of which contains a different reaction-terminating nucleotide, together with the usual reagents for replication. In one batch, the replication reaction always produces complementary strands that end with A; in another, with G; and so on. Gel electrophoresis is used to sift the resulting products according to size, allowing one to infer the exact nucleotide sequence for the original DNA strand. ❖

SPELLING OUT THE ANSWER. In the much-automated Sanger sequencing method, the single-stranded DNA to be sequenced is “primed” for replication with a short complementary strand at one end. This preparation is then divided into four batches, and each is treated with a different replication-halting nucleotide (depicted here with a diamond shape), together with the four “usual” nucleotides. Each replication reaction then proceeds until a reaction-terminating nucleotide is incorporated into the growing strand, whereupon replication stops. Thus, the “C” reaction produces new strands that terminate at positions corresponding to the G’s in the strand being sequenced. (Note that when long strands are being sequenced the concentration of the reaction-terminating nucleotide must be carefully chosen, so that a “normal” C is usually paired with a G; otherwise, replication would typically stop with the first or second G.) Gel electrophoresis—one lane per reaction mixture—is then used to separate the replication products, from which the sequence of the original single strand can be inferred.



Prepare four reaction mixtures; include in each a different replication-stopping nucleotide



Marked progress is also evident in the development of sequencing technologies, though all of those in widespread current use are still based on methods developed in 1977 by Allan Maxam and Walter Gilbert and by Frederick Sanger and his coworkers (see “Tools of the Trade,” pages 17–19). Both of these methods rely on gel-based electrophoresis systems to separate DNA fragments, and recent advances in commercial systems include increasing the number of gel lanes, decreasing run times, and enhancing the accuracy of base identification. As a result of such improvements, a standard sequencing machine can now turn out raw, unverified sequences of 50,000 to 75,000 bases per day.

Equally important to the sequencing goals of the genome project is a rational system for organizing and distributing the material to be sequenced. The DOE’s commitment to such resources dates back to 1984, when it organized the National Laboratory Gene Library Project. Based on cell- and chromosome-sorting technologies developed at Livermore and Los Alamos, libraries of clones were established for each of the human chromosomes, and the individual clones are widely available for mapping and for isolating genes. These clones were invaluable in

Advances have brought much nearer the day when “production sequencing” can begin.

such notable “gene hunts” as the successful searches for the cystic fibrosis and Huntington disease genes. More recently, as more efficient vectors have become available, complete human DNA libraries have been established using BACs, PACs, and YACs.

Another critical resource is being assembled in an effort known as I.M.A.G.E. (Integrated Molecular Analysis of Genomes and their Expression), cofounded by the Livermore Human Genome Center. The aim is a master set of mapped and sequenced human cDNA, representing the expressed parts of the human genome. By early 1996,

I.M.A.G.E. had distributed over 250,000 partial and complete cDNA clones, most of them with one or both ends sequenced to provide unique identifiers. These identifiers, *expressed sequence tags* (ESTs), are usually 300–500 base pairs each. Twenty-five hundred genes have also been newly mapped as part of this coordinated effort.

SHOTGUNS AND TRANSPOSONS

Such advances as these, in both technology development and the assembly of resource libraries, have brought much nearer the day when “production sequencing” can begin. A great deal of variety remains, however, in the approaches available to sequencing the human genome, and it is not yet clear which will prove the most efficient and most cost-effective way to read long stretches of DNA over the next decade. One of the available choices, for example, is between “shotgun” and “directed” strategies. Another is the degree of redundancy—that is, how many times must a given strand be sequenced to ensure acceptable confidence in the result?

Shotgun sequencing derives its name from the randomly generated DNA fragments that are the objects of scrutiny. Many copies of a single large clone are broken into pieces of perhaps 1500 base pairs, either by restriction enzymes or by physical shearing. Each fragment is then separately cloned, and a convenient portion of it sequenced. A computational assembly process then compares the terminal sequences of the many fragments and, by finding overlaps that indicate neighboring fragments, constructs an ordered library for the parent clone. The members of this ordered library can then be sequenced from end to end to yield a complete sequence for the parent. The statistics involved in taking this approach require that many copies of the original clone be randomly fragmented, if no gaps are to be tolerated in the final sequence. A benefit is that the final sequence is highly reliable; the main disadvantage is that the same sequence must be done many times (in the many overlapping fragments). Nevertheless, shotgun

sequencing has been the primary means for generating most of the genomic sequence data in public DNA databases. This includes the longest contiguous fragment of sequenced human DNA, from the human T-cell receptor beta region, of about 685,000 base pairs—a product of DOE-supported work at the University of Washington.

The shotgun strategy is also being used at the Genome Therapeutics Corporation and The Institute for Genomic Research (TIGR), as part of the DOE-supported Microbial Genome Initiative. Genome Therapeutics has sequenced 1.8 million base pairs of *Methanobacterium thermoautotrophicum*, a bacterium important in energy production and bioremediation, and TIGR has successfully sequenced the complete genomes of three free-living bacteria, *Haemophilus influenzae* (1,830,137 base pairs; an effort supported mostly by private funds), *Mycoplasma genitalium* (580,070 base pairs), and *Methanococcus jannaebii* (1,739,933 base pairs).

The alternative to shotgun sequencing is a directed approach, in which one seeks to sequence the target clone from end to end with a minimum of duplication. The essence of this approach is embodied in a technique known as *primer walking*. Starting at one end of a single large fragment, one replicates a stretch of DNA—say, 400 base pairs long—that can be sequenced in one run. With the sequence for this first segment in hand, the next stretch of DNA, just overlapping the first, is then tackled in the same way. In principle, one can thus “walk” the entire length of the original clone. Unfortunately, this conceptually simple approach has been historically beset with disadvantages, mainly the expense and inconvenience of custom-synthesizing a primer as the necessary starting point for each sequencing step. The widely automated Sanger sequencing method involves a DNA replication step that must be “primed” by a DNA fragment that is complementary to 15 to 20 base pairs of the strand to be sequenced (see “Tools of the Trade,” pages 17–19). Until recently, making these primers was an expensive and time-consuming business, but recent innovations have made

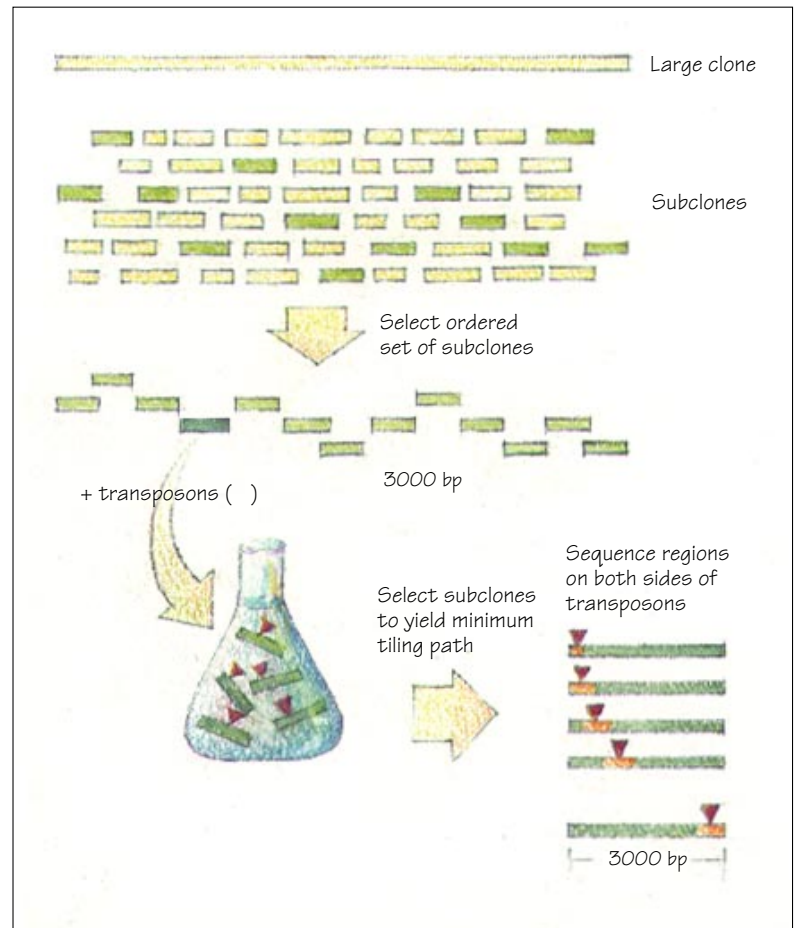


FIGURE 7. TAKING A DIRECTED APPROACH. One directed sequencing strategy exploits a naturally occurring genetic element known as a transposon. The starting point is an ordered set of subclones, each about 3000 base pairs long, derived from a much larger clone (say, a YAC). For each subclone, a preparation is then made in which transposons insert themselves randomly into the subclone—on average, one transposon in each 3000-base-pair strand. The positions of the transposons are mapped, and a set of strands is selected such that the insertion points are about 300 base pairs apart. Sequencing then proceeds in both directions from the transposon insertion points, using the known transposon sequence as a primer. The full set of overlapping regions yields the sequence for the entire subclone, and the sequences of the full set of subclones yield the sequence for the larger original clone.

primer walking, and similar directed strategies, more and more economically feasible.

One way to deal with the primer bottleneck, for example, is to use sets of very short fragments to prime the next sequencing step. As an illustration, the four nucleotides (A, T, C, and G) can be ordered in more than 68 billion ways to create an 18-base primer, an imposing set of possibilities. But it is eminently practical to create a library of the 4096 possible 6-base primers. Three of these “6-mers” can be matched to the end of the

fragment to be sequenced, thus serving as an 18-base primer. This modular primer technology, developed at the Brookhaven National Laboratory, is currently being applied to *Borrelia burgdorferi*, the organism that causes Lyme disease; a 34,000-base-pair fragment has already been sequenced.

Another directed approach uses a naturally occurring genetic element called a *transposon*, which insinuates itself more or less randomly in longer DNA strands. This predilection for random insertion and the fact that the transposon's DNA sequence is well known are the keys to the sequencing strategy depicted schematically in Figure 7. The largest clones are broken into smaller subclones (each of about 3000 base pairs), which then become the targets of the transposons. Multiple copies of each subclone are exposed to the transposons, and reaction

conditions are controlled to yield, on average, a single insertion in each 3000-base-pair strand. The individual strands are then analyzed to yield, for each, the approximate position of the inserted transposon. By mapping these positions, a "minimum tiling path" can be determined for each subclone—that is, a set of strands can be identified whose transposon insertions are roughly 300 base pairs apart. In this set of strands, the region around

each transposon is then sequenced, using the inserted transposons as starting points. The known transposon sequence allows a single primer to be used for sequencing the full set of overlapping regions.

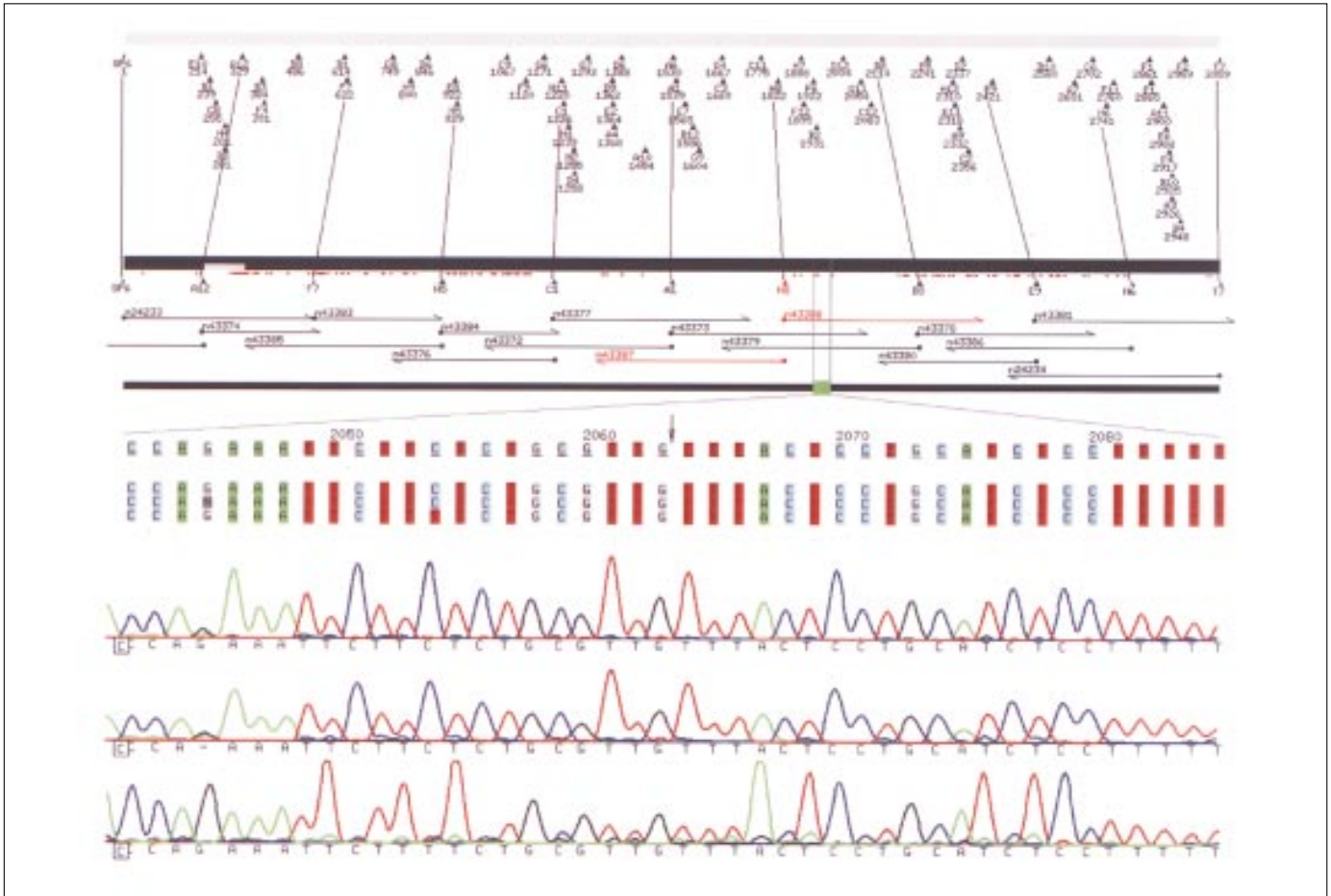
At the Lawrence Berkeley National Laboratory, this technique has been used to sequence over 1.5 million base pairs of DNA on human chromosomes 5 and 20, as well as over three million base pairs from the fruit fly *Drosophila melanogaster*. On chromosome 5, interest focuses on a region of three million base pairs that is rich in growth factor and receptor genes; whereas, on chromosome 20,

Berkeley researchers are interested in a region of about two million base pairs that is implicated in 15 to 20 percent of all primary breast carcinomas. As an example of the kind of output these efforts produce, Figure 8 shows a stretch of sequence data from chromosome 5.

Researchers supported by the DOE at the University of Utah are also pursuing the use of directed sequencing. In addition, they have developed a methodology for "multiplex" DNA sequencing, which offers a way of increasing throughput with either shotgun or directed approaches. By attaching a unique identifying sequence to each sequencing sample in a mixture of, say, 50 such samples, the entire mixture can be analyzed in a single electrophoresis lane. The 50 samples can be resolved sequentially by probing, first, for bands containing the first identifier, then for bands containing the second, and so forth. In a similar way, multiplexing can also be used for mapping. The Utah group is now able to map almost 5000 transposons in a single experiment, and they are using multiplexing in concert with a directed sequencing strategy to sequence the 1.8 million base pairs of the thermophilic microbe *Pyrococcus furiosus* and two important regions of human chromosome 17.

The completed physical maps of chromosomes 16 and 19, with their extensive coverage in many different kinds of cloning vectors, are especially ripe for large-scale sequencing. Los Alamos scientists have therefore begun sequencing chromosome 16, focusing special effort on locating the estimated 3000 expressed genes on that chromosome and using those sites as starting points for directed genomic sequencing. A region of 60,000 base pairs has already been sequenced around the adult polycystic kidney gene, and good starts have been made in mapping other genes. Interestingly, even random sequencing has led to the identification of gene DNA in over 15 percent of the samples, confirming the apparent high density of genes on this chromosome. Between chromosome 16 and the short arm of chromosome 5, another Los Alamos target, the genome center there

The completed physical maps of chromosomes 16 and 19 are especially ripe for large-scale sequencing.



has produced almost two million base pairs of human DNA sequence.

A parallel effort is under way at Livermore on chromosome 19 and other targeted genomic regions. Using a shotgun approach, researchers there have completed over 1.3 million bases of genomic sequence. Initially, they are attacking two major regions of chromosome 19: one of about two million base pairs, containing several genes involved in DNA repair and replication, and another of approximately one million base pairs, containing a kidney disease gene. The Livermore scientists are making use of the I.M.A.G.E. cDNA resource to sequence the cDNA from these regions, along with the associated segments of the genome. In addition, Livermore scientists have targeted DNA repair gene regions throughout the genome and, in many cases, have done comparative sequencing of these genes in other

Continues on p. 26

FIGURE 8. SEQUENCE DATA: THE FINAL PRODUCT. The ultimate description of the genome, though only a prelude to full understanding, is the base-pair sequence. This computer display shows results from the use of transposons at Berkeley. The array of triangles represents the transposons inserted into a 3000-base-pair subclone; the 11 selected by the computer to build a minimum tiling path are shown below the heaviest black line. The subclone segments sequenced by using these 11 starting points are depicted by the horizontal lines; the arrowheads indicate the sequencing directions. The expanded region between bases 2042 and 2085 is covered by three sequencing reactions, which produced the three traces at the bottom of the figure. Above the traces, the results are summarized, together with a consensus sequence (just below the numbers).

The Mighty Mouse

The human genome is not so very different from that of chimpanzees or mice, and it even shares many common elements with the genome of the lowly fruit fly. Obviously, the differences are critical, but so are the similarities. In particular, genetic experiments on other organisms can illuminate much that we could not otherwise learn about *homologous* human genes—that is, genes that are basically the same in the two species.

In some cases, the connection between a newly identified human gene and a known health disorder can be quickly established. More often, however, clear links between cloned genes and human hereditary diseases or disease susceptibilities are extremely elusive. Diseases that are modified by other genetic predispositions, for example, or by environment, diet, and lifestyle can be exceedingly difficult to trace in human families. The same holds for very rare diseases and for genetic factors contributing to birth defects and other developmental disorders. By contrast, disorders such as these can sometimes be followed relatively easily in animal systems, where uniform genetic backgrounds and controlled breeding schemes can be used to avoid the variability that often confounds human population studies. As a consequence, researchers looking for clues to the causes of many complex health problems are focusing more and more attention on model animal systems.

Among such systems, which range in complexity from yeast and bacteria to mammals, the most prominent is the mouse. Because of its small size, high fertility rate, and experimental manipulability, the mouse offers great promise in studying the genetic causes and pathological progress of ailments, as well as understanding the genetic role in disease susceptibility. In pursuing such studies, the DOE is exploiting several resources, among them the experimental mouse genetics facility at the Oak Ridge National Laboratory. Initially

established for genetic risk assessment and toxicology studies, the Oak Ridge facility is one of the world's largest. Mutant strains there express a variety of inherited developmental and health disorders, ranging from dwarfism and limb deformities to sickle cell anemia, atherosclerosis, and unusual susceptibilities to cancer.

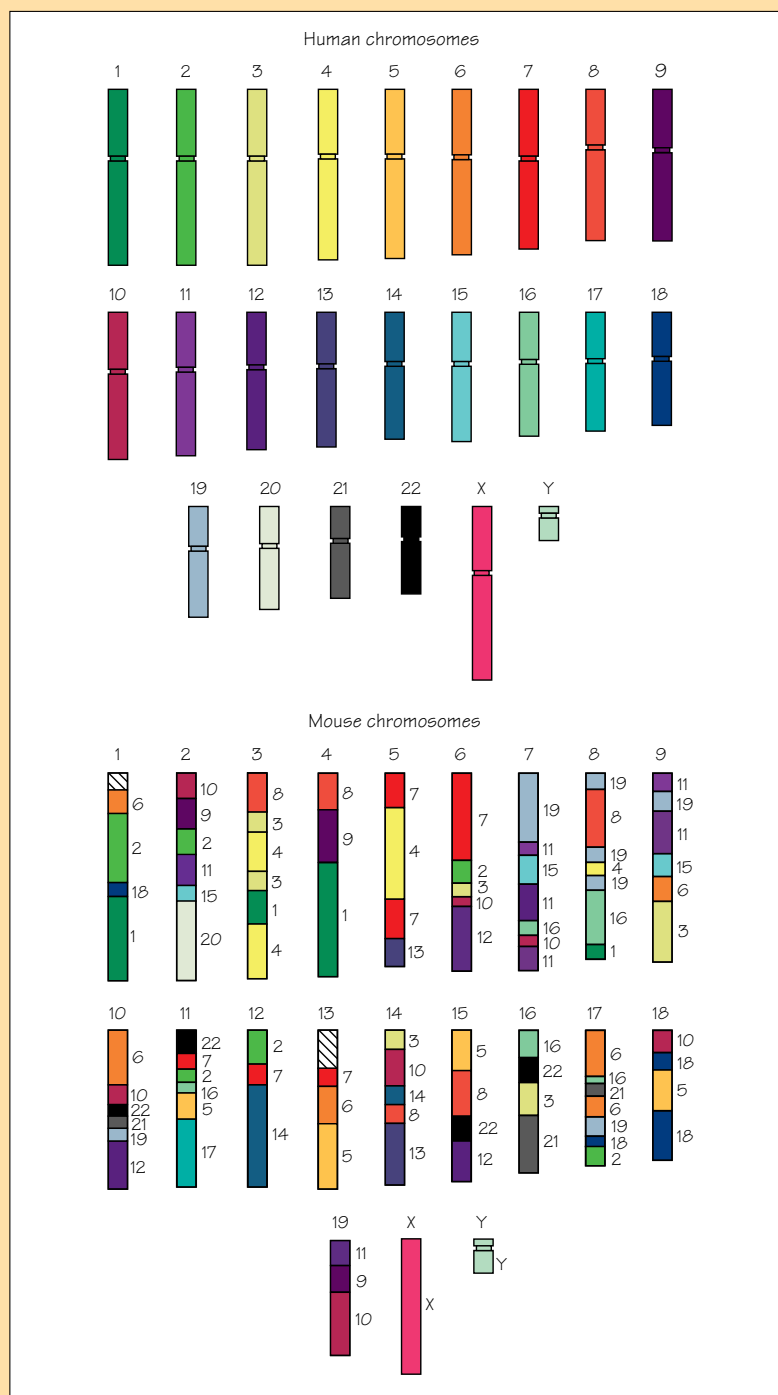
Most of these existing mutant strains have arisen from random alterations of genes, caused by the same processes that occur naturally in all living populations. However, other, more directed means of gene alteration are also available. So-called *transgenic* methods, which have been developed and refined over the past 15 years, allow DNA sequences engineered in the laboratory to be introduced directly into the genomes of mouse embryos. The embryos are subsequently transferred to a foster mother, where they develop into mice carrying specifically designed alterations in a particular gene. The differences in form, basic health, fertility, and longevity produced by these “designer mutations” then allow researchers to study the effects of genetic defects that can mimic those found in human patients. The payoff can be clues that aid in the design of drugs and other treatments for the human diseases.

The Human Genome Center at Berkeley is using mice for similar purposes. In vivo libraries of overlapping human genome fragments (each 100,000 to 1,000,000 base pairs long) are being propagated in transgenic mice. The region of chromosome 21 responsible for Down syndrome, for example, is now almost fully represented in a panel of transgenic mice. Such libraries have several uses. For example, the precise biochemical means by which identified genes produce their effects can be studied in detail, and new genes can be recognized by analyzing the effects of particular genome fragments on the transgenic animals. In such ways, the promise of the massive effort to map and sequence the human genome can be translated into the kind of biological

knowledge coveted by pharmaceutical designers and medical researchers.

Adding to the potential value of mutant mice as models for human genetic disease is growing evidence of similarities between mouse and human genes. Indeed, practically every human gene appears to have a counterpart in the mouse genome. Furthermore, the related mouse and human genes often share very similar DNA sequences and the same basic biological function. If we imagine that the 23 pairs of human chromosomes were shattered into smaller blocks—to yield a total of, say, 150 pieces, ranging in size from very small bits containing just a few genes to whole chromosome arms—those pieces could be reassembled to produce a serviceable model of the mouse genome. This mouse genome jigsaw puzzle is shown to the right. Thanks to this mouse-human genomic homology, a newly located gene on a human chromosome can often lead to a confident prediction of where a closely related gene will be found in the mouse—and vice versa.

Thus, a crippling heritable muscle disorder in mice maps to a location on the mouse X chromosome that is closely analogous to the map location for the X-linked human Duchenne muscular dystrophy gene (DMD). Indeed, we now know that these two similar diseases are caused by the mouse and human versions of the same gene. Although mutations in the mouse *mdx* gene produce a muscle disease that is less severe than the heartbreaking, fatal disease resulting from the DMD mutation in humans, the two genes produce proteins that function in very similar ways and that are clearly required for normal muscle development and function in the corresponding species. Likewise, the discovery of a mouse gene associated with pigmentation, reproductive, and blood cell defects was the crucial key to uncovering the basis for a human disease known as the piebald trait. Owing to such close human-mouse relationships as these, together with the benefits of transgenic technologies, the mouse offers enormous potential in identifying new human genes, deciphering their complex functions, and even treating genetic diseases. ❖



OF MICE AND MEN. The genetic similarity (or homology) of superficially dissimilar species is amply demonstrated here. The full complement of human chromosomes can be cut, schematically at least, into about 150 pieces (only about 100 are large enough to appear in this illustration), then reassembled into a reasonable approximation of the mouse genome. The colors of the mouse chromosomes and the numbers alongside indicate the human chromosomes containing homologous segments. This piecewise similarity between the mouse and human genomes means that insights into mouse genetics are likely to illuminate human genetics as well.

species, especially the mouse. Such comparative sequencing has identified conserved sequence elements that might act as regulatory regions for these genes and has also assisted in the identification of gene function (see “The Mighty Mouse,” pages 24–25).

HOW GOOD IS GOOD ENOUGH?

The goal of most sequencing to date has been to guarantee an error rate below 1 in 10,000, sometimes even 1 in 100,000. However, the difference between one human being and another is more like one base pair in five hundred, so most researchers now agree that one error in a thousand is a more reasonable standard. To assure a higher level of confidence, and perhaps to uncover important individual differences, the most biologically or medically important regions would still be sequenced more exhaustively, but using this lowered standard would greatly reduce the cost of acquiring sequence data for the bulk of human DNA.

With this philosophy in mind, Los Alamos scientists have begun a project to determine the cost and throughput of a low-redundancy sequencing strategy known as *sample sequencing* (SASE, or “sassy”). Clones are selected from the high-resolution Los Alamos cosmid map, then physically broken into 3000-base-pair subclones—much as in other sequencing approaches. In contrast to, say, shotgun sequencing, though, only a small random set of the subclones is then selected for sequencing. Sequence fragments already known—end sequences, sequence-tagged sites, and so forth—are used as the starting points. The result is sequence coverage for about 70 percent of the original cosmid clone, enough to allow identification of genes and ESTs, thus pinpointing the most critical targets for later, more thorough sequencing efforts. Further, the SASE-derived sequences provide enough information for researchers elsewhere to pursue just such comprehensive efforts, using whole genomic DNA. In addition, the cost of SASE sequencing is only one-tenth the cost of obtaining a complete sequence, and a genomic region can be “sampled” ten times as fast.

As the first major target of SASE analysis, Los Alamos scientists chose a cosmid contig of four million base pairs at the end (the *telomere*) of the short arm of chromosome 16. By early 1996, over 1.4 million base pairs had been sequenced, and a gene, EST, or suspected coding region had been located on every cosmid sampled.

In addition, Los Alamos is building on the SASE effort by using SASE sequence data as the basis for an efficient primer walking strategy for detailed genomic sequencing. The first application of this strategy, to a telomeric region on the long arm of chromosome 7, proved to be as efficient as typical shotgun sequencing, but it required only two- to threefold redundancy to produce a complete sequence, in contrast to the seven- to tenfold redundancy required in shotgun approaches. The resulting 230,000-base-pair sequence is the second-longest stretch of contiguous human DNA sequence ever produced.



In a sense, though, even a complete genome sequence—the ultimate physical map—is only a start in understanding the human genome. 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 very different functions with the same genetic program, and how these and countless other cell types arise in the first place from a single undifferentiated egg cell. 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. 🌟



Beyond Biology

INSTRUMENTATION AND INFORMATICS

FROM THE START, it has been clear that the Human Genome Project would require advanced instrumentation and automation if its mapping and sequencing goals were to be met. And here, especially, the DOE's engineering infrastructure and tradition of instrumentation development have been crucial contributors to the international effort. Significant DOE resources have been committed to innovations in instrumentation, ranging from straightforward applications of automation to improve the speed and efficiency of conventional laboratory protocols (see, for example, Figure 9a) to the development of technologies on the cutting edge—technologies that might potentially increase mapping and sequencing efficiencies by orders of magnitude.

On the first of these fronts, genome researchers are seeing significant improvements in the rate, efficiency, and economy of large-scale mapping and sequencing efforts as a result of improved laboratory automation tools. In many cases, commercial robots have simply been mechanically reconfigured and reprogrammed to perform repetitive tasks, including the replication of large clone libraries, the pooling of libraries as a prelude to various assays, and the arraying of clone libraries for hybridization studies. In other cases, custom-designed instruments have proved more efficient. A notable illustration is the world's fastest cell and chromosome sorter, developed at Livermore and now being commercialized, which is used to sort human chromosomes for chromosome-specific libraries. Other examples

include a high-speed, robotics-compatible thermal cycler developed at Berkeley, which greatly accelerates PCR amplifications, and instruments developed at Utah for automated hybridization in multiplex sequencing schemes.

SMALLER IS BETTER — AND OTHER DEVELOPMENTS

Beyond “mere” automation are efforts aimed at more fundamental enhancements of established techniques. In particular, a number of DOE-supported efforts aim at improved versions of the automated gel-based Sanger sequencing technique. For example, in place of the conventional slab gels, ultrathin gels, less than 0.1 millimeter thick, can be used to obtain 400 bases of sequence from each lane in a hour's run, a fivefold improvement in throughput over conventional systems. Even faster speedups are seen when arrays of 0.1-millimeter capillaries are used as the separation medium. Both of these approaches exploit higher electric field strengths to increase DNA mobility and to reduce analysis times. And Livermore scientists are looking beyond even capillaries, to sequencing arrays of rigid glass microchannels, supplemented by automated gel and sample loading.

The capillary approach is especially ripe for further development. Challenges include providing uniform excitation over

The project will require advanced instrumentation and automation if its goals are to be met.

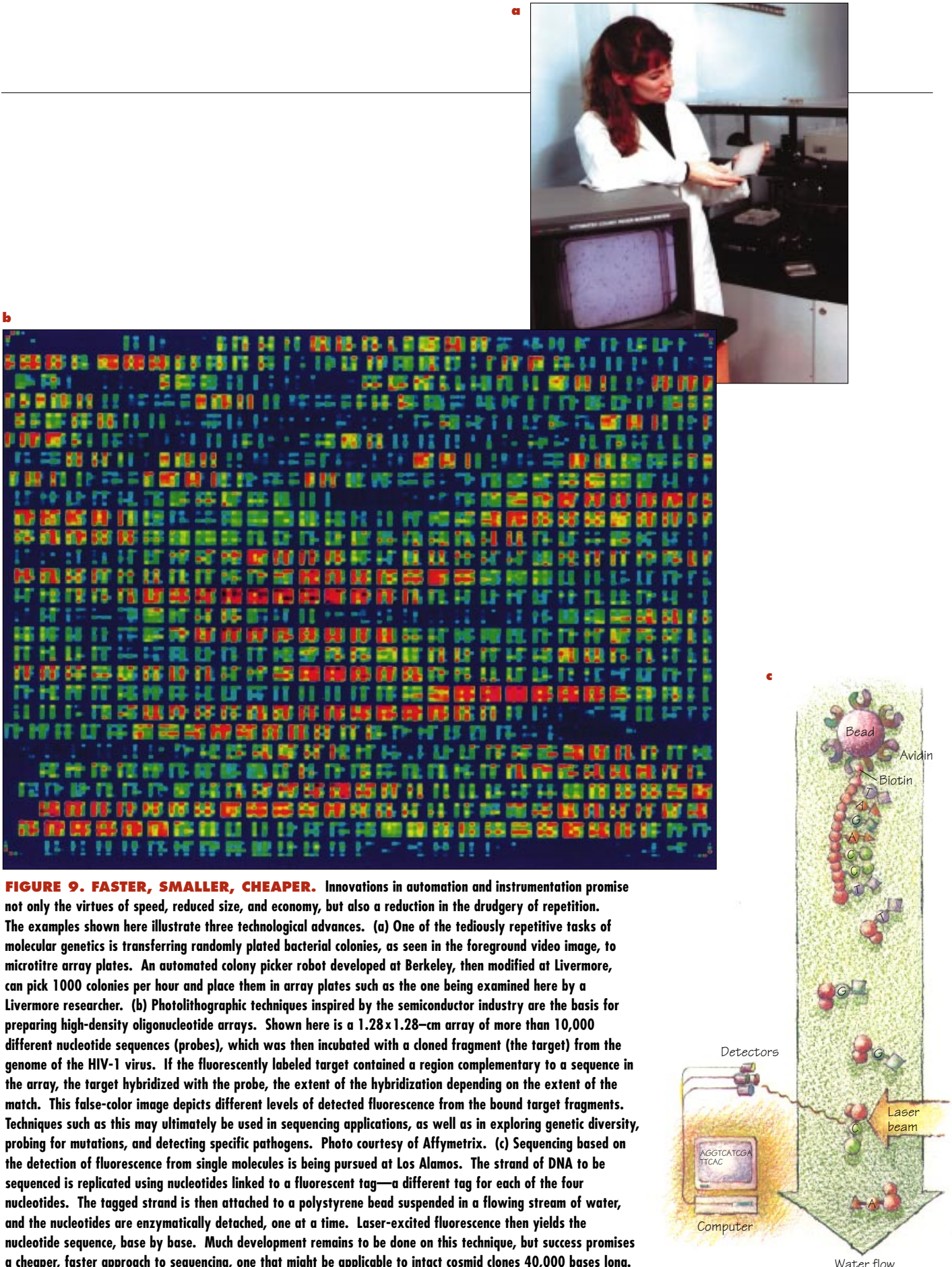


FIGURE 9. FASTER, SMALLER, CHEAPER. Innovations in automation and instrumentation promise not only the virtues of speed, reduced size, and economy, but also a reduction in the drudgery of repetition. The examples shown here illustrate three technological advances. (a) One of the tediously repetitive tasks of molecular genetics is transferring randomly plated bacterial colonies, as seen in the foreground video image, to microtitre array plates. An automated colony picker robot developed at Berkeley, then modified at Livermore, can pick 1000 colonies per hour and place them in array plates such as the one being examined here by a Livermore researcher. (b) Photolithographic techniques inspired by the semiconductor industry are the basis for preparing high-density oligonucleotide arrays. Shown here is a 1.28x1.28-cm array of more than 10,000 different nucleotide sequences (probes), which was then incubated with a cloned fragment (the target) from the genome of the HIV-1 virus. If the fluorescently labeled target contained a region complementary to a sequence in the array, the target hybridized with the probe, the extent of the hybridization depending on the extent of the match. This false-color image depicts different levels of detected fluorescence from the bound target fragments. Techniques such as this may ultimately be used in sequencing applications, as well as in exploring genetic diversity, probing for mutations, and detecting specific pathogens. Photo courtesy of Affymetrix. (c) Sequencing based on the detection of fluorescence from single molecules is being pursued at Los Alamos. The strand of DNA to be sequenced is replicated using nucleotides linked to a fluorescent tag—a different tag for each of the four nucleotides. The tagged strand is then attached to a polystyrene bead suspended in a flowing stream of water, and the nucleotides are enzymatically detached, one at a time. Laser-excited fluorescence then yields the nucleotide sequence, base by base. Much development remains to be done on this technique, but success promises a cheaper, faster approach to sequencing, one that might be applicable to intact cosmid clones 40,000 bases long.

arrays of 50 to 100 capillaries and then efficiently detecting the fluorescence emitted by labeled samples. Technologies under investigation include fiber-optic arrays, scanning confocal microscopy, and cooled CCD cameras. Some of this effort has already been transferred to the private sector, and tenfold improvements in speed, economy, and efficiency are projected in future commercial instruments.

The move toward miniaturization is afoot elsewhere as well. Building on experiences in the electronics industry, several DOE-supported groups are exploring ways to adapt high-resolution photolithographic methods to the manipulation of minuscule quantities of biological reagents, followed by assays performed on the same “chip.” Current thrusts of this “nanotechnology” approach include the design of microscopic electrophoresis systems and ultrasmall-volume, high-speed thermal cycling systems for PCR. A miniaturized, computer-controlled PCR device under development at Livermore operates on 9-volt batteries and might ultimately lead to arrays of thousands of individually controlled micro-PCR chambers.

Another miniaturization effort aims at the fabrication of high-density combinatorial arrays of custom *oligomers* (short chains of nucleotides), which would make feasible large-scale hybridization assays, including sequencing by hybridization. This innovative technique uses short oligomers that pair up with corresponding sequences of DNA. The oligomers are placed on an array by a process similar to that of making silicon chips for electronics. Successful matches between oligomers and genomic DNA are then detected by fluorescence, and the application of sophisticated statistical analyses reassembles the target sequence. This same technology has already been used for genetic screening and cDNA fingerprinting. Figure 9b illustrates a DOE-supported application of high-density oligonucleotide arrays to the detection of mutations in the HIV-1 genome. Similar approaches can be envisioned to understand differences in patterns of gene expression: Which genes are active (which

are producing mRNA) in which cells? Which are active at different times during an organism’s development? Which are active, or inactive, in disease?

Sequencing by hybridization is only one of several forward-looking ideas for revolutionizing sequencing technology. In spite of continuing improvements to sequencers based on the classic methods, it is nonetheless desirable to explore altogether new approaches, with an eye to simplifying sample preparation, reducing measurement times, increasing the length of the strands that can be analyzed in a single run, and facilitating interpretation of the results. Over the course of the past few years, several alternative approaches to direct sequencing have been explored, including atomic-resolution molecular scanning, single-molecule detection of individual bases, and mass spectrometry of DNA fragments.

All of these alternatives look promising in the long term, but mass spectrometry has perhaps demonstrated the greatest near-term potential. Mass spectrometry measures the masses of ionized DNA fragments by recording their time-of-flight in vacuum. It would therefore replace traditional gel electrophoresis as the last step in a conventional sequencing scheme. Routine application of this technique still lies in the future, but fragments of up to 500 bases have been analyzed, and practical systems based on high-resolution mass separations of DNA fragments of fewer than 100 bases are currently being developed at several universities and national laboratories.

Another innovative sequencing method is under investigation at Los Alamos. As depicted in Figure 9c, each of the four bases (A, T, C, G) in a single strand of DNA receives a different fluorescent label, then the bases are enzymatically detached, one at a time. The characteristic fluorescence is detected by a laser system, thereby yielding the sequence, base by base. This approach is beset by major technical challenges, and direct

In spite of improvements to sequencers based on the classic methods, it is nonetheless desirable to explore altogether new approaches.

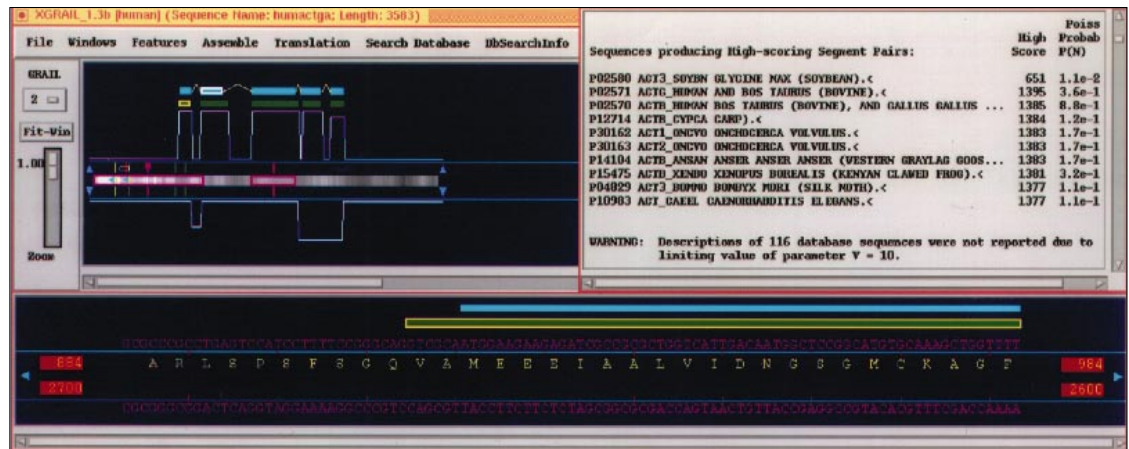


FIGURE 10. GENE HUNTS. Genes, the regions that actually code for proteins, constitute only a small fraction, perhaps 10%, of the human genome. Thus, even with sequence in hand, finding the genes is yet another daunting step away. One tool developed to help in the hunt is GRAIL, a computer program developed at Oak Ridge that uses heuristics based on existing data, together with artificial neural networks, to identify likely genes. Coding and noncoding regions of the genome differ in many subtle respects—for example, the frequency with which certain short sequences appear. Further, particular landmarks are known to characterize the boundaries of many genes. In the example shown here, GRAIL has searched for likely genes in both strands of a 3583-base-pair sequence. The results are shown at the upper left. The upper white trace indicates five possible exons (coding regions within a single gene) in one strand, whereas the lower white trace suggests two possible exons in the other strand. However, the lower trace scores worse on other tests, leading to a candidate set of exons shown by the five green rectangles. By refining this set further, GRAIL then produces the final gene model shown in light blue. The lower part of the figure zeros in on the end of the candidate exon outlined in yellow, thus providing a detailed look at one of the differences between the preliminary and final models. The sequence is shown in violet, together with the amino acids it codes for, in yellow. The preliminary model thus begins with the sequence GTCGCA. . . , which codes for the amino acids valine and alanine. In fact, though, almost all genes begin with the amino acid methionine, a feature of the final gene model. At the upper right, GRAIL displays the results of a database search for sequences similar to the final five-exon gene model. Close matches were found among species as diverse as soybean and the nematode *Caenorhabditis elegans*.

sequencing has not yet been achieved. But the potential benefits are great, and much of the instrumentation for sensitive detection of fluorescence signals has already proved useful for molecular sizing in mapping applications.

DEALING WITH THE DATA

Among the less visible challenges of the Human Genome Project is the daunting prospect of coping with all the data that success implies. Appropriate information systems are needed not only during data acquisition, but also for sophisticated data analysis and for the management and public distribution of unprecedented quantities of biological information. Further, because much of the challenge is interpreting genomic data and making the results available for scientific and technological applications, the challenge extends not just to the Human Genome


Project, but also to the microbial genome program and to public- and private-sector programs focused on areas such as health effects, structural biology, and environmental remediation. Efforts in all these areas are the mandate of the DOE genome informatics program, whose products are already widely used in genome laboratories, general molecular biology and medical laboratories, biotechnology companies, and biopharmaceutical companies around the world.

The roles of laboratory data acquisition and management systems include the construction of genetic and physical maps, DNA sequencing, and gene expression analysis. These systems typically comprise databases for tracking biological materials and experimental procedures, software for controlling robots or other automated systems, and software for acquiring laboratory data and presenting it in useful form. Among such systems are physical mapping databases

developed at Livermore and Los Alamos, robot control software developed at Berkeley and Livermore, and DNA sequence assembly software developed at the University of Arizona. These systems are the keys to efficient, cost-effective data production in both DOE laboratories and the many other laboratories that use them.

The interpretation of map and sequence data is the job of data analysis systems. These systems typically include task-specific computational engines, together with graphics and user-friendly interfaces that invite their use by biologists and other non-computer scientists. The genome informatics program is the world leader in developing automated systems for identifying genes in DNA sequence data from humans and other organisms, supporting efforts at Oak Ridge National Laboratory and elsewhere. The Oak Ridge-developed GRAIL system, illustrated in Figure 10, is a world-standard gene identification tool. In 1995 alone, more than 180 million base pairs of DNA were analyzed with GRAIL.

A third area of informatics reflects, in a sense, the ultimate product of the Human Genome Project—information readily available to the scientific and lay communities.

Public resource databases must provide data and interpretive analyses to a worldwide research and development community. As this community of researchers expands and as the quantity of data grows, the challenges of maintaining accessible and useful databases likewise increase. For example, it is critical to develop scientific databases that “interoperate,” sharing data and protocols so that users can expect answers to complex questions that demand information from geographically distributed data resources. As the genome project continues to provide data that interlink structural and functional biochemistry, molecular, cellular, and developmental biology, physiology and medicine, and environmental science, such interoperable databases will be the critical resources for both research and technology development. The DOE genome informatics program is crucial to the multiagency effort to develop just such databases. Systems now in place include the Genome Database of human genome map data at Johns Hopkins University, the Genome Sequence DataBase at the National Center for Genome Resources in Santa Fe, and the Molecular Structure Database at Brookhaven National Laboratory. 



Ethical, Legal, and Social Implications

AN ESSENTIAL DIMENSION OF GENOME RESEARCH

THE HUMAN GENOME PROJECT is rich with promise, but also fraught with social implications. We expect to learn the underlying causes of thousands of genetic diseases, including sickle cell anemia, Tay-Sachs disease, Huntington disease, myotonic dystrophy, cystic fibrosis, and many forms of cancer—and thus to predict the likelihood of their occurrence in any individual. Likewise, genetic information might be used to predict sensitivities to various industrial or environmental agents. The dangers of misuse and the potential threats to personal privacy are not to be taken lightly.

Both the DOE and the NIH devote a portion of their resources to studies of ethical, legal, and social implications.

In recognition of these important issues, both the DOE and the National Institutes of Health devote a portion of their resources to studies of the ethical, legal, and social implications (ELSI) of human genome research. Perhaps the most critical of social issues are the questions of privacy and fair use of genetic information. Most observers agree that personal knowledge of genetic susceptibility can be expected to serve us well, opening the door to more accurate diagnoses, preventive intervention, intensified screening, lifestyle changes, and early and effective treatment. But such knowledge has another side, too: the risk of anxiety, unwelcome changes in personal relationships, and the danger of

stigmatization. Consider, for example, the impact of information that is likely to be incomplete and indeterminate (say, an indication of a 25 percent increase in the risk of cancer). And further, if handled carelessly, genetic information could threaten us with discrimination by potential employers and insurers. Other issues are perhaps less immediate than these personal concerns, but they are no less challenging. How, for example, are the “products” of the Human Genome Project to be patented and commercialized? How are the judicial, medical, and educational communities—not to mention the public at large—to be effectively educated about genetic research and its implications?

To confront all these issues, the NIH-DOE Joint Working Group on Ethical, Legal, and Social Implications of Human Genome Research was created in 1990 to coordinate ELSI policy and research between the two agencies. One focus of DOE activity has been to foster educational programs aimed both at private citizens and at policy-makers and educators. Fruits of these efforts include radio and television documentaries, high school curricula and other educational material, and science museum displays. In addition, the DOE has concentrated on issues associated with privacy and the confidentiality of genetic information, on workplace and commercialization issues (especially screening for susceptibilities to environmental or workplace agents), and on the implications of research findings regarding the interactions among multiple genes and environmental influences.

Whereas the issues raised by modern genome research are among the most challenging we face, they are not unprecedented. Issues of privacy, knotty questions of how knowledge is to be commercialized, problems of dealing with probabilistic risks, and the imperatives of education have all been confronted before. As usual, defensible perspec-

tives and reasonable arguments, even precious rights, exist on opposing sides of every issue. It is a balance that must be sought. Accordingly, further study is needed, as well as continuing efforts to promote public awareness and understanding, as we strive to define policies for the intelligent use of the profound knowledge we seek about ourselves. 🌟



THE AGE OF DISCOVERY was the age of da Gama, Columbus, and Magellan, an era when European civilization reached out to the Far East and thus filled many of the voids in its map of the world. But in a larger sense, we have never ceased from our exploration and discovery. Science has been unstinting over the ages in its efforts to complete our intellectual picture of the universe. In this century, our explorations have extended from the subatomic to the cosmic, as we have mapped the heavens to their farthest reaches and charted the properties of the most fleeting elementary particles. Nor have we neglected to look inward, seeking, as it were, to define the topography of the human body. Beginning with the first modern anatomical studies in the sixteenth century, we have added dramatically to our picture of human anatomy, physiology, and biochemistry. The Human Genome Project is thus the next stage in an epic voyage of discovery—a voyage that will bring us to a profound understanding of human biology.

In an important way, though, the genome project is very different from many of our exploratory adventures. It is spurred by a conviction of practical value, a certainty that human benefits will follow in the wake of success. The product of the Human Genome Project will be an enormously rich biological

database, the key to tracking down every human gene—and thus to unveiling, and eventually to subverting, the causes of thousands of human diseases. The sequence of our genome will ultimately allow us to unlock the secrets of life's processes, the biochemical underpinnings of our senses and our memory, our development and our aging, our similarities and our differences.

It has further been said that the Human Genome Project is *guaranteed* to succeed: Its goal is nothing more assuming than a sequence of three billion characters. And we have a very good idea of how to read those characters. Unlike perilous voyages or searches for unknown subatomic particles, this venture is assured of its goal. But beyond a detailed picture of human DNA, no one can predict the form success will take. The genome project itself offers no promises of cancer cures or quick fixes for Alzheimer's disease, no detailed understanding of genius or schizophrenia. But if we are *ever* to uncover the mysteries of carcinogenesis, if we are *ever* to know how biochemistry contributes to mental illness and dementia, if we *ever* hope to really understand the processes of growth and development, we must first have a detailed map of the genetic landscape. That's what the Human Genome Project promises. In a way, it's a rather prosaic step, but what lies beyond is breathtaking. 🌟

The World Wide Web offers the easiest path to current news about the Human Genome Project. Good places to start include the following:

- DOE Human Genome Program—http://www.er.doe.gov/production/ohcr/hug_top.html
- NIH National Center for Human Genome Research—<http://www.nchgr.nih.gov>
- Human Genome Management Information System at Oak Ridge National Laboratory—http://www.ornl.gov/TechResources/Human_Genome/home.html
- Lawrence Berkeley National Laboratory Human Genome Center—<http://www-hgc.lbl.gov/GenomeHome.html>
- Lawrence Livermore National Laboratory Human Genome Center—<http://www-bio.llnl.gov/bbrp/genome/genome.html>
- Los Alamos National Laboratory Center for Human Genome Studies—<http://www-ls.lanl.gov/LSwelcome.html>
- The Genome Database at Johns Hopkins University School of Medicine—<http://gdbwww.gdb.org/>
- The National Center for Genome Resources—<http://www.ncgr.org/>

ACKNOWLEDGMENTS

This booklet was prepared at the request of the U.S. Department of Energy, Office of Health and Environmental Research, as an overview of the Human Genome Project, especially the role of the DOE in this international, multiagency effort. Though edited and produced at the Lawrence Berkeley National Laboratory, this account aims to reflect the full scope of the DOE-sponsored effort. In pursuit of this goal, the contributions of many have been essential. Within the Department of Energy, David A. Smith deserves special mention. He managed the DOE Human Genome Program until his retirement this year, and he was the principal catalyst of this effort to summarize its achievements. Also contributing program descriptions, illustrations, advice, and criticism: at DOE, Daniel W. Drell; at Berkeley, Michael Palazzolo, Christopher H. Martin, Sylvia Spengler, David Gilbert, Joseph M. Jaklevic, Eddy Rubin, Kerrie Whitelaw, and Manfred Zorn; at Lawrence Livermore National Laboratory, Anthony Carrano, Gregory G. Lennon, and Linda Ashworth; at Los Alamos National Laboratory, Robert K. Moyzis and Larry Deaven; at Oak Ridge National Laboratory, Lisa Stubbs; at the National Center for Genome Resources, Christopher Fields; and at Affymetrix, Robert J. Lipshutz. Behind the scenes, many others no doubt had a hand.

DOUGLAS VAUGHAN

Editor


Design: Debra Lamfers Design

Illustrations: Marilee Bailey

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, 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 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 product, 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.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098. PUB-773/July 1996.

 Printed on recycled paper.

