chapter 3

Applications to Research in Biology and Medicine

CONTENTS

Boxes

Figures

Tables

Chapter 3

Applications to Research in Biology and Medicine

"[physical and genetic maps 1 will certainly be very useful [but] you have to interpret that sequence, and that's going to be a lot of work. It will be like having a whole history of the world written in a language you can't read. "

Joseph Gall
American Scientist 76:17-18,
February 1988

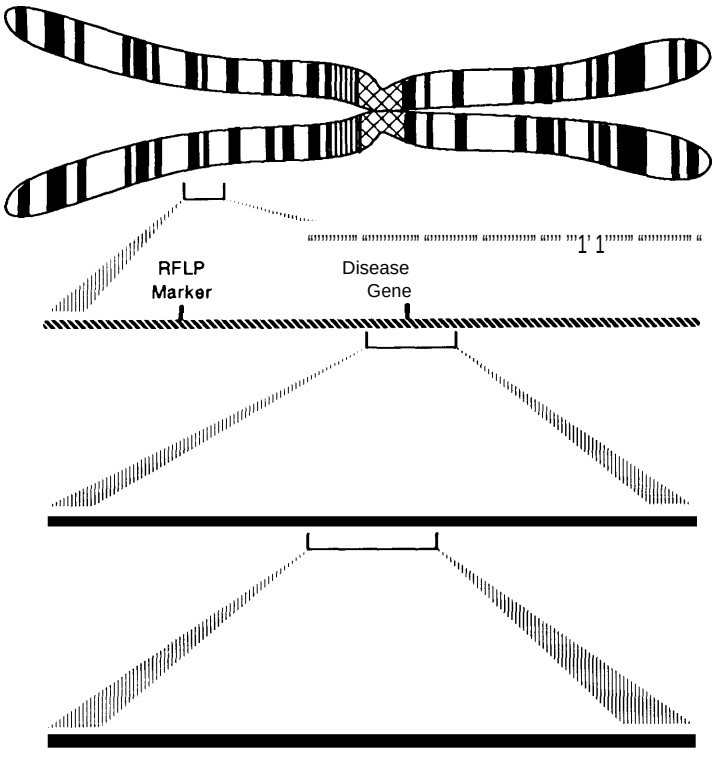
INTRODUCTION

Research efforts aimed at creating genetic linkage and physical maps of chromosomes or entire genomes are collectively referred to in this report as genome projects (see chs. 1 and 2). The goals of genome projects are to develop technologies and tools for mapping and sequencing DNA and to complete maps of human and other genomes. Proponents expect that the products and processes generated from genome projects will enable researchers to answer important questions in biology and medicine. Meeting this objective, however, will depend on the success of concurrent projects aimed at *analyzing* the information generated from mapping genomes. Interpreting genome maps will require the combined efforts of individuals with expertise in structural biology, cell biology, population biology, biochemistry, genetics, computer science, and other fields.

Biology and medicine have already benefited from efforts to map and sequence specific genes from human and other organisms. Some questions might be addressed sooner or better, however, if more extensive genetic linkage maps, cDNA maps, contig maps, and DNA sequences were available (figure 3-1). (See ch. 2 for detailed discussion of the types of genetic linkage and physical maps,) Research on inherited and nongenetic diseases, the physiology and development of organisms, the molecular basis of evolution, and other fundamental problems in biology could all be facilitated in the long run by genome projects.

Scientists continue to debate about which applications depend on information from maps of entire genomes and which require only maps of specific regions. The value of a complete DNA sequence of a reference human genome is the most hotly contended scientific issue (see box 3-A). Focused research has been the mode of molecular genetics to date: Scientists have targeted specific regions of genomes for intensive study. Many of the potential applications of genome mapping summarized in this chapter have already been and will continue to be achieved by targeted research projects. Wherever possible, therefore, this chapter attempts to differentiate between the uses for which extensive maps will be necessary and those for which partial maps are adequate.

Figure 3-1 .—Mapping at Different Levels of Resolution



MAPPING GENES TO WHOLE CHROMOSOMES

Chromosome Banding
DNA Hybridization to Somatic Cell Hybrids or Sorted Chromosomes
In Situ Hybridization
Karyotyping

GENETIC LINKAGE MAPPING

PHYSICAL MAPPING OF LARGE DNA FRAGMENTS

Yeast Artificial Chromosome Cloning (Up to 1 Million Base Pairs)
Pulsed Field Gel Electrophoresis (Up to 10 Million Base Pairs)

PHYSICAL MAPPING OF SMALL DNA FRAGMENTS

Cosmid Cloning (Up to *45,000* Base Pairs)
Plasmid and Phage Cloning (Up to 23,000 Base Pairs)

APPLICATIONS IN MEDICINE

Genome projects have accelerated the production of new technologies, research tools, and basic knowledge. At current or perhaps increased levels of effort, they may eventually make possible control of many human diseases—first through more effective methods of detecting disease, then, in some cases, through development of effective therapies based on improved understanding of disease mechanisms. Advances in human genetics and molecular biology have already provided insight into the origins of such diseases as hemophilia, sickle cell disease, and hypercholesterolemia.

The new technologies for genetics research will also help in the assessment of public health needs. Techniques for sequencing DNA rapidly, for example, should permit the detection of mutations following exposure to radiation or environmental agents. Susceptibilities to environmental and work place toxins might be identified as more detailed genetic linkage maps are developed, and special methods of surveillance can be used to monitor individuals at risk. By providing tools for determining the presence or absence of pathogens (e.g., bacteria and viruses) in large numbers of individuals as well as identifying genetic factors that render some human beings more susceptible to infection than others, genome projects might also yield methods for tracking epidemics through populations.

Developing Diagnostic Tools

The use of DNA hybridization probes for detecting changes, such as restriction fragment length polymorphisms (RFLPs), in the DNA of in-

Box 3-A.—Why Sequence Entire Genomes?

Rapid advances in technology have made it feasible to sequence the entire genome of an organism, at least a small one such as bacteria or yeast. Researchers do not yet agree, however, on the value of a complete DNA sequence of a genome the size of the human genome. Several types of arguments have been made in favor of sequencing entire genomes:

- The information in a genome is the fundamental description of a living system—it is what the cell uses to construct a copy of itself—and so is of fundamental concern to biologists.
- Genome sequences provide a conceptual framework within which much future research in biology will be structured. Questions concerning control of gene expression (signals for control of gene expression, genome replication, development mechanisms, and so on) ultimately depend on knowing genome sequences.
- The genomes of some higher organisms, including those of human beings, have repeated DNA sequences, sequences of unknown function, and some sequences which are likely to have no function, comprising nearly 90 percent of the total DNA content. Without the complete DNA sequence of several genomes, it will be impossible to determine whether such sequences have meaning or are ancestral "junk" sequences.
- Genome sequences are important for addressing questions concerning evolutionary biology. The reconstruction of the history of life on this planet, the definition of gene families (also critical to other areas of biology), and the search for a universal ancestor all require an understanding of the organization of genomes.
- Genomes are natural information storage and processing sytems; unraveling them may be of general interest to computer and physical scientists.

Other scientists would argue that these possible applications can be derived from sequences of single genes or larger regions of chromosomes. They believe it is a waste of time and money to sequence the entire human genome, particularly because some regions have no known or essential function. Many of these researchers favor sequencing only those regions believed to be clinically or scientifically important, including expressed sequences and sequences involved in the control of gene expression, and putting the others off indefinitely.

SOURCES:
National Research Council, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).
C. Woese, LJniversity of Illinois, Urbana, personal communication, June 1987.

Table 3-1.—Examples of Single-Gene Diseases

Disease	Description	Genetic marker identified	Gene cloned	Protein identified
Duchenne muscular dystrophy	Progressive muscle deterioration	Yes	Yes	Yes
Cystic fibrosis	Lung and gastrointestinal degeneration	Yes	No	No
Huntington's disease	Late-onset disorder with progressive physical and mental deterioration	Yes	No	No
Sickle cell anemia	Deformed red blood cells block blood flow	Yes	Yes	Yes
Hemophilia	Defect in clotting factor VIII causes uncontrolled bleeding	Yes	Yes	Yes
Beta-Thalassemia	Failure to produce sufficient hemoglobin	Yes	Yes	Yes
Chronic granulomatous disease	Frequent bacterial and fungal infections involving lungs, liver, and other organs	Yes	Yes	Yes—tentative
Pherrylketonuria	Enzyme deficiency that causes brain damage and mental retardation	Yes	Yes	Yes
Polycystic kidney disease	Pain, hypertension, kidney failure in half of victims	Yes	No	No
Retinoblastoma	Cancer of the eye	Yes	Yes	Yes

SOURCE: Office of Technology Assessment, 198S.

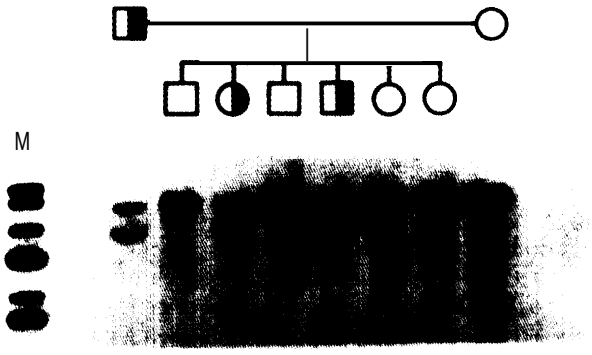


Photo credit: Ray White, The University of Utah Medical Center, Salt Lake **City,** *UT. Repr/nted with* **permission *from* American Journal of Human Genetics 39:3(B 306, 1986.**

Identification of a genetic marker showing linkage between high levels of low-density lipoprotein (LDL) cholesterol and the genetic locus for the LDL receptor gene using restriction fragment length polymorphism (RFLP) analysis of the LDL receptor genes from a multigenerational family with inherited hypercholesterolemia. A radioactively labeled DNA fragment from the cloned LDL receptor gene was used as a probe to observe differences among affected and unaffected individuals in the numbers of electrophoretically separated DNA fragments after cutting the DNA with a restriction enzyme. Individuals without the polymorphism are represented as unfilled squares (males) or circles (females) and show only one DNA fragment. Hal f-f illed symbols represent individuals with one **allele** for the defective gene and one for the normal gene and show two DNA fragments. The lane marked "M" is a set of DNA fragment size markers.

dividuals with genetic diseases is described in detail in chapter 2. Such methods of DNA analysis offer several advantages over traditional approaches to the study of human disease. Knowing the organization of genes on chromosomes and their DNA sequences could enable clinicians to detect mutant genes before a disease manifests itself in the form of damaged cells or tissues and will eventually lead to a more complete understanding of the pathogenesis of human disease [Friedmann, see app. A].

The study of randomly selected RFLP markers in human families has revealed linkages to a number of genetic diseases (table 3-1) (1,3,6, 10,16,17, 23)25,29,32,37,38,41,42,46,52,55,56), As the chromosomal locations of more disease-causing genes are identified, more probes for diagnosing genetic diseases will become available. **A genetic linkage map saturated with RFLP markers (or one with other polymorphic markers) is viewed by many molecular geneticists as crucial to the development of diagnostic reagents for there maining human genetic diseases** [Friedmann, see app. A]. (See table 3-2 for a list of companies developing diagnostic probes for such diseases.)

It is important to recognize that DNA probes for RFLP markers are not always reliable tools for diagnosing genetic diseases before the onset of symptoms. Without enough data from relatives of potential disease carriers, it may not be possible to confirm the linkage between a particular RFLP marker and a genetic disease. The main limitation to reliable diagnosis of most genetic diseases is the lack of an adequate number of DNA samples from several generations of affected and unaffected individuals.

Many available RFLP markers can be used only in a few families, and the RFLP marker map is a cumulative one that aggregates the data from many families. The largest standard data set is derived from the Center for the Study of Human Polymorphism (CEPH) in Paris (see ch. 7 on international efforts in genome mapping). The data collected by CEPH are taken from 40 families around the world, most of which do not have any known genetic disease. Materials from these families are used to locate RFLP and other polymorphic markers. Once markers have been identified, they can be tested for linkage to a particular genetic disease in families known to have that disease. The

Table 3-2.—Some Companies Developing DNA Probes for Diagnosis of Genetic Diseases

Company	Probes under development
California Biotechnology (Mountain View, CA)	Susceptibility to heart disease
Cetus Corporation (Emeryville, CA)	sickle cell anemia
Collaborative Research (Bedford, MA)	Cystic fibrosis Duchenne muscular dystrophy Polycystic kidney disease
Integrated Genetics (Framingham, MA)	Cystic fibrosis Hemophilia B Huntington's disease Polycystic kidney disease Sickle cell anemia
Lifecodes (Elmsford, NY)	Cystic fibrosis Down's syndrome Polycystic kidney disease Sickle cell anemia

SOURCE: Office of Technology Assessment, 198S.

Photo credit: The Bettmann Archive, New York, NY

A large New England family of the early 1900s spanning three generations. Samples of genomic DNA from members of such families are very useful for constructing genetic linkage maps, such as a RFLP map.

CEPH families are large, selected to enable scientists to trace DNA markers through at least three generations.

Isolating Genes Associated With Disease

Some inherited human diseases arise from or cause differences in detectable proteins that circulate in the blood, such as human growth hormone and insulin. A research scheme called forward genetics has been used to isolate the genes encoding these proteins. In this strategy, a gene is cloned after the altered protein product has been characterized. Other genetic diseases, such as retinoblastoma, chronic granulomatous disease, and Duchenne muscular dystrophy, involve protein products that were not identified before the corresponding gene was cloned. An experimental approach called reverse genetics was used to find these genes. First the gene containing the mutation responsible for the disease is linked to a RFLP or other polymorphic marker, then the gene and its protein product are isolated and characterized [Friedmann, see app. A].

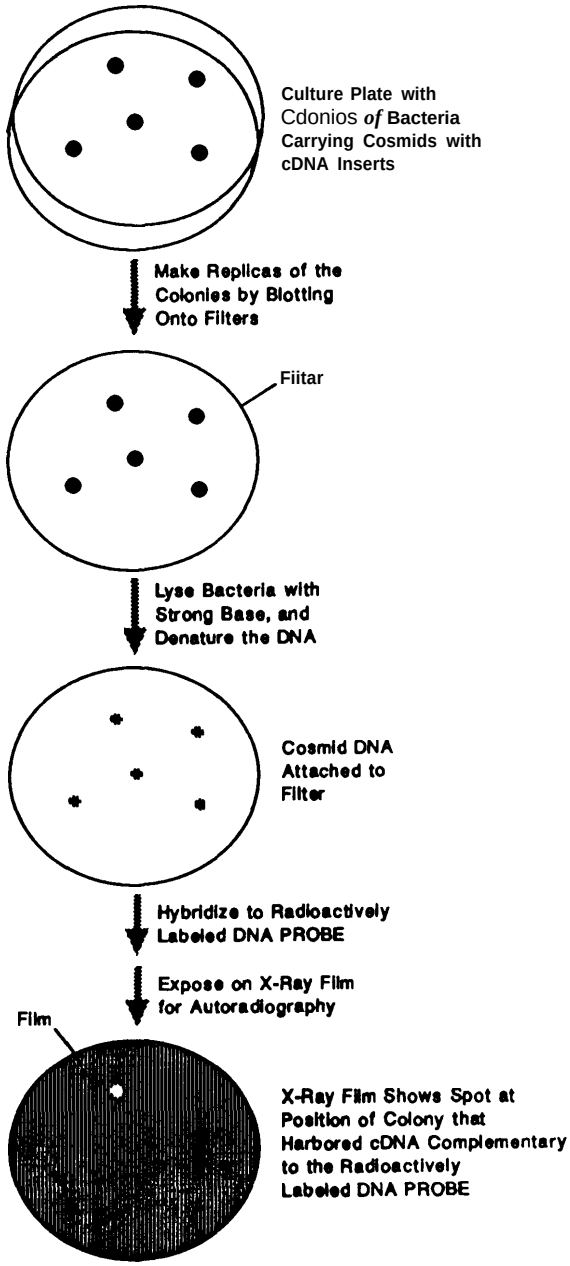
Forward Genetics

Until recently, most methods for cloning disease-associated genes required prior characterization of the biochemical defect responsible for the disease. Using the forward genetics approach, researchers identify the mutant gene product—a protein—then isolate a clone of the gene from a library of cDNA clones (clones made from DNA copies of the mRNA transcripts of genes—see ch. 2). If the protein has been purified, antibodies can be made and used to select for clones of cells expressing the product. Alternatively, if part of the protein's amino acid sequence is known, synthetic DNA prohes complementary to the *exons,* or protein-coding sequences, of the gene can be designed, based on the genetic code (figure 3-2). Once the cDNA clones are isolated, they can be used as DNA probes to pick out clones from genomic libraries.

83-360 0 - 88 - 3 QL 3

Figure 3=2.-The Use of Synthetic DNA Probes To Clone Genes

Amino Acid Sequence of a Small Region of a Protein	Methionine	Tyrosine	Arginine	Methionine	Glutamine	Leucine	Serine	Cysteine
An mRNA Sequence Predicted from Amino Acid Codon Usage Frequencies	AUG	UAC	AGG	AUG	CAA	CUG	UCU	UGC
DNA PROBE Sequence Complementary to Predicted mRNA Sequence	TAC	ATG	TCC	ATC	GTT	GAC	AGA	ACG



The difference between cDNA copies of genes and genes on chromosomes is that the latter have both exons and introns (noncoding sequences interrupting protein-coding sequences). The genes in the human genome range from fewer than 1,000 base pairs to more than 2 million base pairs in size and are thus typically too large to be contained on standard cloning vectors (table 3-3). The cDNA clones, which are smaller because they contain only exons, are useful because they can be introduced into bacteria, yeast, or mammalian tissue culture cells and transcribed and translated into protein. The resulting proteins can be used in studies of the physiology of diseases or in some cases as human therapeutics.

The utility of the various types of physical maps in the forward genetics strategy depends on the purpose of isolating the gene. If only cDNA copies of a particular gene are needed for making large quantities of the protein product, then extensive genomic maps would not be necessary. If the cDNA copy of the gene is to be used as a DNA probe for isolating the whole gene from a collection of genomic DNA clones, or for studying the organization of the genome in the region of interest, then a contig map illustrating the order of DNA segments from the relevant portion of the genome would be very useful.

Reverse Genetics

Reverse genetics has made it possible to isolate genes associated with inherited diseases for which no specific biochemical defect has been established. To do this, the genetic disease is usually linked first to a particular chromosome by studying inheritance patterns at the DNA level. The general region of the gene on the chromosome is identified using DNA probes for RFLP markers on that chromosome. Samples of DNA from families of individuals afflicted with the genetic disease are tested with a set of DNA probes which hybridize to markers spaced throughout the chromosome until a linkage between the mutant gene that causes the disease and the RFLP is detected. The location of the gene is then identified more pre-

Table 3.3.—The Size of Human Genes

Gene	Gene size (in thousands of nucleotides)	mRNA size (in thousands of nucleotides)	Number of introns
Small:			
Alpha-globin	0.8	0.5	2
Beta-globin	1.5	0.6	2
Insulin	1.7	0.4	2
Apoliproprotein E	3.6	1.2	3
Parathyroid	4.2	1.0	2
Protein kinase C	11.0	1.4	7
Medium:			
Collagen I			
Pro-alpha-l(1)	18.0	5.0	50
Pro-alpha-2(1)	38.0	5.0	50
Albumin	25.0	2.1	14
High-mobility group COA reductase	25.0	4.2	19
Adenosine deaminase	32.0	1.5	11
Factor IX	34.0	2.8	7
Catalase	34.0	1.6	12
Low-de nsity-lipoprotei n receptor	45.0	5.5	17
Large:			
Phenylalanine hydroxylase	90.0	2.4	12
Factor Vll	186.0	9.0	25
Thyroglobulin	300.0	8.7	36
Very *large:*			
Duchenne muscular dystrophy	2,000<0	17.0	50

SOURCE: Victor McKusick, The Johns Hopkins University, Baltimore, MD.

cisely by using additional probes for closely spaced markers that cover the region of interest.

In order to distinguish the gene locus that actually causes the disease from nearby, but unrelated, genes, it is generally necessary to demonstrate that the identified gene is expressed abnormally in tissues from patients with the disease. A genomic region 1 million base pairs in length, for example, could contain as many as 100 genes. In such cases, it is necessary to use biochemical methods to identify the gene that is responsible for the disease. Techniques for detecting messenger RNA transcripts or proteins can be used to search for differences in amounts of geneprociuct in the tissues of affected and unaffected individuals; these differences can then be correlated with an alteration in a particular gene. Retinal cells were analyzed in this way as part of the search for the retinoblastoma gene (18,28), as were muscle cells in individuals with and without Duchenne muscular dystrophy (see box 3-B). Once the gene product has been identified, it is possible to study the physiology of a particular disease with the aim of identifying a therapy or preventive treatment.

Although reverse genetics is generic in concept, the amount of effort involved in isolating and characterizing genes using genetic and physical map data varies. Over 100 person-years have been spent searching for the gene that causes cystic fibrosis–an effort that has led to localizing the gene on a small region of chromosome 7 but not to finding the gene itself or determining how it causes the disease (17). On the other hand, researchers identified and isolated the gene for chronic granulomatous disease in far fewer person-years (38). The existence of DNA probes for RFLP markers has also made possible the identification of the genes for Duchenne muscular dystrophy (32) and retinoblastoma (18,28) [Friedmann, see app. A].

The technical difficulty involved in locating the gene responsible for a particular disease by reverse genetics usually depends on the physical map distance between the nearest RFLP marker and the linked gene. Existing RFLP maps of the human genome have a resolution of only about 10 centimorgans (approximately 10 million base pairs). **A map with markers spaced every 1 centimorgan would make it much less timeconsuming to locate the genes by reverse genetics (7,33).** Such a map would be constructed using a pool of several thousand DNA probes that detect RFLP markers spaced about every 1 million base pairs throughout the genome. **A library of clones made from overlapping segments of the genome and a contig map illustrating the relative position of each clone with respect to its neighbors would also be useful in reverse genetics.** These tools would spare researchers the labor-intensive step of isolating and characterizing all of the genomic clones between the marker and the gene of interest; the only work remaining would be to associate the characteristics of the disease with the correct clone or clones.

Identification of Genes Involved in Polygenic Disorders

Genetic linkage maps of the human genome are also useful for characterizing inherited diseases caused by more than one factor, often referred to as polygenic *disorders*. Among the diseases for which more than one gene is likely to be responsible are certain cancers, diabetes, and coronary heart disease (27). For example, in a complex disorder such as coronary heart disease, blood plasma lipoproteins, the coagulation system, and elements of the arterial walls all play a role, so the number of genes involved can be very large (40). Some scientists argue that the RFLP maps currently available, with markers spaced an average of 10 centimorgans apart, area sufficient starting point for studies of polygenic diseases (11). **Highenresolution RFLP maps, such as a 1-centimorgan map, would no doubt simplify the job of identifying the genes responsible for polygenic disorders.**

Developing Human Therapeutics

Forward genetics has yielded important results in the area of drug development. As stated earlier, the ability to use cDNA clones has been crucial to the development of commercial products such as human growth hormone and insulin and to potential human therapeutics such as tumor necrosis factor and interleukin-2, therapeutics that would not otherwise be available in the quantity or quality necessary for effective use (table 3-4)

Box 3-B.—Duchenne and Becker's Muscular Dystrophies

Duchenne muscular dystrophy (DMD) is a genetic disease that affects 1 in 3,000 to 1 in 3,500 male infants born. Becker's muscular dystrophy is a similar but milder disorder with much lower incidence. Both diseases begin in childhood and lead to muscle wasting. DMD typically results in death before age 20. The search for the gene causing these diseases and the protein encoded by that gene has been an exciting story of molecular biology in the 1980s. The effort in many ways typifies modern genetics, with extensive international collaboration, study of nonhuman species, and creative use of molecular methods.

The gene causing these diseases had been known for some time to be on the X chromosome because of inheritance patterns. Duchenne and Becker's muscular dystrophies affect primarily boys, who have only one X chromosome, inherited from their mothers. Girls have two X chromosomes and therefore must, as a rule, receive abormal genes from *both* parents in order to develop Duchenne or Becker's muscular dystrophy–a much less likely occurrence.

The search for the gene started with studies of families. DNA from persons with DMD, including several girls and one boy, was collected in an effort to find a common area of the X chromosome that had been lost or altered. once the correct region of the X chromosome had been identified (its absence was found to cause DMD), DNA from that region was obtained and cloned. The clones were used as DNA probes for complementary mRNA sequences in muscle tissue from affected and unaffected individuals. The purpose was to identify the mRNA gene transcript that was present in unaffected individuals but altered in persons with DMD. The mRNA was located and subsequently shown to encode a large protein called dystrophin found in muscle cells.

The DMD search has uncovered some extraordinary facts. Duchenne and Becker's muscular dystrophies are caused by different changes in the same gene. That gene is the largest found to date, spanning over 2 million base pairs (table 3-3). It is broken into at least 60 exons.

The scientific collaboration that led to the discovery of dystrophin was notably efficient. One paper alone listed 77 authors from 24 research institutions in 8 countries. Molecular probes, clones, and materials from affected patients were openly exchanged, hastening researchers in their quest for the culprit gene.

SOURCES

K.H Fischbeck, A.W Ritter, D.L Tirschwell, et al., "Recombination With pERT87 (DXS164) in Families With X-Linked Muscular Dystrophy, *Lancet* 2(July) 104, 1986

E P Hoffman, A P. Monaco, C.C. Feenel, et al, "Conservation of the Duchenne Muscular Dystrophy Gene in Mice and Humans" *Science* 238347-3.50, 1987

E.P Hoffman, R.H Brown, and L.M Kunkel, "Dystrophin: The Protein Product of the Duchenne Muscular Dystrophy Locus," *Cell* .51 919.928, 1987

M Koenig, E P Hoffman, C.J. Bertelson, et al, "Complete Cloning of the Duchenne Muscular Dystrophy (DMD) cDNA and Preliminary Genomic Organization of the DMD Gene in Normal and Affected Individuals, " *Cell* 50:509-517, 1987.

L. M. Kunkel et al., "Analysis of Deletions From Patients With Becker and Duchenne Muscular Dystrophy," *Nature 322 "73-77,* 1986

A P Monaco, R .1,. Neve, C. Colletti-Feener, et al., "Isolation of Candidate cDNAs for Portions of the Duchenne Muscular Dystrophy Gene "*Nature* 323646-650, *1986.*

A.P Monaco, C.J. Bertelson, C. Colletti-Feener, et al., "Localization and Cloning of Xp21 Deletion Breakpoints Involved in Muscular Dystrophy "*Human Genetics* 75:221-227, 1987.

G J van Omenn, J .M. Verkerk, M.H. Hofker, et al., "A Physical Map of 4 Million Base Pairs Around the Duchenne Muscular Dystrophy Gene on the X Chromosome," *Cell* 47:499-504, 1986

(53). The cDNA clones isolated by forward genetics could be used to make a cDNA map that illustrates the chromosomal locations of expressed regions of DNA. **This cDNA map, plus a library of previously ordered clones of genomic DNA, would be valuable tools for studying the role of certain genes in the manifestation of disease.** Knowledge of the mechanisms directing normal cellular functions will probably lead to important sources of new therapies for human diseases: natural human proteins made from isolated human genes, engineered proteins, and conventionally synthesized drugs designed from a knowledge of the structure of the proteins they target. **Advances in the development of human therapeutic products will be made more rapidly if research in the areas of protein engineering, the relationship of protein structure to function, rational drug design, and others parallels genome mapping efforts.**

Photo credit: Nancy Wexler, Columbia University, New York, NY

A Venezuelan man with Huntington's disease, a rare, late-onset genetic disease that causes degeneration of nerve cells in the brain.

Prospects for Human Gene Therapy

Clinical use of human genetic linkage and physical maps, now largely restricted to diagnosis, may eventually include the insertion of normal DNA directly into human cells to correct a particular genetic defect (54). This practice is called *human gene therapy*. Advances in gene therapy will depend on development of ways to insert DNA into cells safely and to ensure that the' inserted DNA corrects the defect (54). Gene mapping will not improve gene therapy directly, and for most diseases the ability to make a diagnosis will precede the availability of an effective treatment. The knowledge gained through use of gene maps will, however, enhance knowledge about the function of genes and thus indirectly improve the prospects for gene therapy [Friedmann, see app. Al.

Table 3=4.—Some Human Gene Products With Potential as Therapeutic Agents

Gene product • Actual or potential therapeutic application
Atrial Natluretic Factor Ž **Possible applications in treatment of hypertension and other** blood pressure disorders, and for some kidney diseases affecting excretion of salts and water.
Alpha Interferon[a] • Approved for treatment of hairy cell leukemia; possible broader applications in other cancers.
Beta Interferon • Inhibits viral infections and may be useful as an anticancer treatment.
Epidermal Growth Factor . Expected to have applications in wound healing, including burns, and for cataract surgery.
Erythropoietin . Anticipated treatment use for anemia resulting from chronic kidney disease.
Factor Vlll:C[b] • Prevents bleeding in patients with hemophilia A after injury.
Flbroblast Growth Factor . Possible use in wound healing and treating burns.
Gamma Interferon . Possible treatment for scleroderma and arthritis.
Granulocyte Colony Stimulating Factor . Possible treatment for Acquired Immune Deficiency Syndrome (AIDS) and leukemia.
Human Growth Hormone[a] • Approved as a treatment for child hood dwarfism; expected to have broader therapeutic potential in treatment for short stature resulting from Turner's syndrome and for wound healing.
Insulin[a] . **Approved for treatment of diabetes.**
Interleukin.2 • **Possible treatment for various cancers.**
Microphage Colony Stimulating Factor . **Potential applications are for treatment of infectious diseases, primarily parasites; possible cancer therapy.**
Superoxide Dismutase . **Possible preventive treatment for damage caused by oxygen-rich** blood entry into oxygen-deprived tissues (e.g., during organ transplants).
Tissue Plasminogen Activator . **Approved as treatment for dissolving blood clots associated with heart attacks.**
Tumor Necrosis Factor • Possible anti-tumor therapy.

[a]Approved for commercial sale in the United States by the Food and Drug Administration.
[b]Non-recombinant DNA version has been approved for sale, but the cloned gene product has not.
SOURCE: Office of Technology Assessment, 1988.

APPLICATIONS IN HUMAN PHYSIOLOGY AND DEVELOPMENT

Studies aimed at understanding the molecular basis of inherited diseases may yield information that can be generalized to other physiological processes. Knowledge of the structure and function of genes associated with Alzheimer's disease, for example, might give important clues to the cellular mechanisms regulating aging of brain tissue.

The organization of genes in genomes is another fundamental issue in biology. Is it important for genes to exist on a particular chromosome in a particular order? Comparisons of physical maps of the chromosomes of higher organisms could shed some light on the extent to which gene organization is associated with gene expression and gene function.

The nucleotide sequences of human genes have been and will continue to be important research tools for understanding the basic cellular processes underlying physiology and development. Nevertheless, knowing the DNA sequences of genes and how they translate into the amino acid sequences of protein products is not sufficient to establish how such genes are controlled or how the gene products function in a particular cell or in the organism as a whole. The genetic code that guides the translation of DNA sequence into protein sequence offers only the first step in unraveling the mysteries of the human genome. Understanding the relationship between protein structure and function is the crucial next step, but it faces the greatest number of technical bottlenecks (see box 3-C).

Identification of Protein-Coding Sequences

Individual efforts to clone particular genes will not be eliminated by the availability of genetic linkage and physical maps; rather they will be redirected toward localizing a particular gene within a region of a chromosome or within the DNA sequence of that region. Because human genes are more often interrupted by introns, the identification of the exon and regulatory sequences in and around genes has proved more difficult in human beings than in lower organisms. The most reliable method of identifying exons is to know the amino acid sequence of the protein product and, using the genetic code, find the corresponding DNA sequence by inspecting the whole gene sequence. DNA sequences can be determined at a faster rate than proteins can be isolated and sequenced, however, so computer-assisted methods offer a more practical approach.

There is a variety of computer software available for predicting exon sequences, some of it more reliable than others (12,14,48) [Mount, see app, A]. As more DNA sequences become available, methods for predicting exons can continue to be refined. **Computer scientists argue that the analysis phase of whole genome sequencing projects will progress efficiently only if the development of new computational and other theory-based predictive methods that can accommodate large sequences is emphasized.**

Photo credit: Shirley Ti/ghman, Princeton University, Princeton, NJ

Electron micrograph revealing an intron sequence interrupting the protein-coding sequences of the mouse beta-g lobin gene. DNA containing the gene (including intron sequences) was allowed to hybridize (base pair) with beta-globin mRNA that had been isolated from cells in its mature form with no intron sequences. A loop appears in the region of the intron where no complementary sequences exist between the two molecules (see arrow).

Box 3-C.—From Gene Structure to Protein Structure: The Protein-Folding Problem

"Protein-folding is the genetic code expressed in three dimensions," according to Fetrow and co-workers. How does the linear sequence of amino acids code for a protein's structure? How does the three dimensional conformation of a protein drive its function? Sometimes the amino acid sequence of a protein with an unknown function is similar to that of a protein with a known function; in many such cases, the similarity is a valid indicator of comparable jobs. In other cases, the threedimensional structure of a protein (the amino acid sequence folded into the actual structure of the protein) gives more reliable clues about function. It is therefore important to develop experimental and theoretical means for determining the three-dimensional structures of proteins. Because proteins are so large, often consisting of multiple domains (discrete portions) with different functions, this generally involves analysis of how each part of a protein contributes to its overall structure.

There is experimental evidence that certain structural domains can serve similar functions in a number of different proteins. It is the combination of domains that gives a protein its unique overall function. A stretch of amino acids in one protein can be nearly identical in sequence to that in another protein, but if the surrounding amino acid sequences are different, then the sequences might fold into domains with quite different threedimensional structures. At present, scientists cannot predict with certainty how the linear sequence of amino acids in a protein will fold into the protein's threedimensional structure—thus the protein-folding problem. As genome mapping projects make more gene sequences available, the problem will take on even greater significance. The National Academy of Sciences in a recent report called protein folding "the most fundamental problem at the chemistry-biology interface, and its solution has the highest long-range priority."

Most predictions of three-dimensional structure are based on theories of the behavior of amino acids in certain chemical and physical environments and on information gleaned from viewing the atomic structures of proteins through X-ray diffraction. (X-ray diffraction of protein crystals is an important tool in structural biology-the field dedicated to the study of proteins and other macromolecular structures. It is the most important technique for determining the threedimensional structure of large proteins at the atomic level.) Existing methods for predicting structure are not reliable for all proteins or protein domains, because structural data are available for only about zoo proteins and for even fewer classes of proteins. There are few membrane proteins in the structure database, for example, and thus little experimental basis for testing predictions about how such important proteins will fold. More structures of proteins need to be determined, using X-ray crystallographic and other biophysical technologies, in order to provide a solid foundation for protein-folding theories. Once the protein-folding problem is solved, the road from gene sequence to gene function will be considerably shortened and will lead, in some cases, toward the development of promising new human therapeutic products.

SC) URCES:

T. Blundell, B.L. Sibanda, M.J.E. **Sternberg, and** J.M. **Thornton, "Knowledge-Based Prediction of Protein Structures and the Design** of **Novel Molecules, "** *Nature* **326:347-352,** *1987.*

J.S. Fetrow, M.H. Zehfus, **and** G.D. **Rose, "Protein Folding: New Twists, "** *Bio/Technology* 6:167-171, **1988.**

T Koetzle, **Brookhaven National Laboratory, Upton, NY, personal communication, March 1987,**

National Academy of Sciences, *Research Briefings* **(Washington, DC: 1986, National Academy Press, 1986).**

Approaches to Understanding Gene Function

Isolating a gene is not nearly as difficult as determining how the gene and its products function in the cell. The following are some experimental approaches to solving this problem:

- Ž to modify or inactivate the normal function of a gene by replacing it with a modified version,
- to inhibit the function of a gene's mRNA or protein product by using antibodies to the protein or an RNA complementary to the mRNA, and

- to compare the DNA sequence of a cloned gene of unknown function with those of genes whose functions are known.

Using the first two methods, scientists have studied the function of gene products by identifying alterations in the biochemical or physical characteristics of the affected cell or organism (2,15, 24,26,30,36,39,44,50,51). The third strategy is theoretical, using sequence data accumulated from previously characterized gene products to predict a function for a newly identified gene. Such predictions can then be tested experimentally.

Probably the most widely used first step in determining the role of a gene is to find similarities between its DNA sequence and those of genes from other organisms. Yeast, for example, shares with animal cells many of the molecules and processes that are being studied intensively in modern cell biology, including the factors modulating cell structure and dynamics, the components of the machinery that modulates protein secretion from cells, the constituents of basic chemical pathways, and analogs of several mammalian *oncogenes* (genes involved in controlling the rate of cell growth). Many of these factors and processes were first identified or characterized, or both, in higher organisms, but the application of them to yeast genetics has provided new insights (57).

A recent study reported the use of yeast cells to isolate a human gene that can substitute for a yeast gene in regulating the yeast cell's life cycle (34). Plasmid vectors carrying cDNA were introduced into yeast cells to find a human gene product that was similar enough to a yeast gene to replace it in the regulation of the yeast cell's life cycle. This was accomplished by mutating the yeast's copy of the gene and then finding cells that, upon introduction of the appropriate human gene, appeared to regain their normal function. The human cDNA clone identified by this technique is thus a candidate for a protein that regulates life cycles in human cells. Studies of the genomic version of the human gene will be necessary to definitively establish the role of this product in human cell cycle regulation. This example illustrates how yeast genetics and biochemistry can be used to identify human genes with important functions (57).

Photo credit: Oonald Riddle, University of Missouri, Columbia, MO

Micrograph comparing the appearance of a short, fat nematode mutant called "dumpy" (above) with that of a normal nematode (below). The mutant grows to only two-thirds of the normal body length because of a mutation in a gene for a type of collagen (a protein) needed for normal development (magnified 80 times).

In fruit flies, genetic research and the tools of recombinant DNA have made it clear that certain DNA sequences are involved in regulating the development of the organism. Different sets of genes appear to be expressed at different times in the course of development, causing the patterns observed in the developing embryo. How gene expression is regulated to create developmental patterns is a central question in biological studies of many organisms. Fruit flies are easy to dissect and to manipulate genetically, and much is known about their development; they have therefore proven to be a very useful model. One DNA sequence, called the horneo *box,* was first identified in the genes of fruit flies and later in those of higher organisms, including human beings (31). There is substantial evidence that the homeo box, a short stretch of nucleotides of nearly identical sequence in the genes that contain it, determines when the expression of particular groups of genes is turned on and off during development of the fruit fly (35). As more gene sequences from fruit flies, human beings, and other organisms are determined, more knowledge about the signals governing developmentally expressed genes is likely to be acquired [Mount, see app. Al.

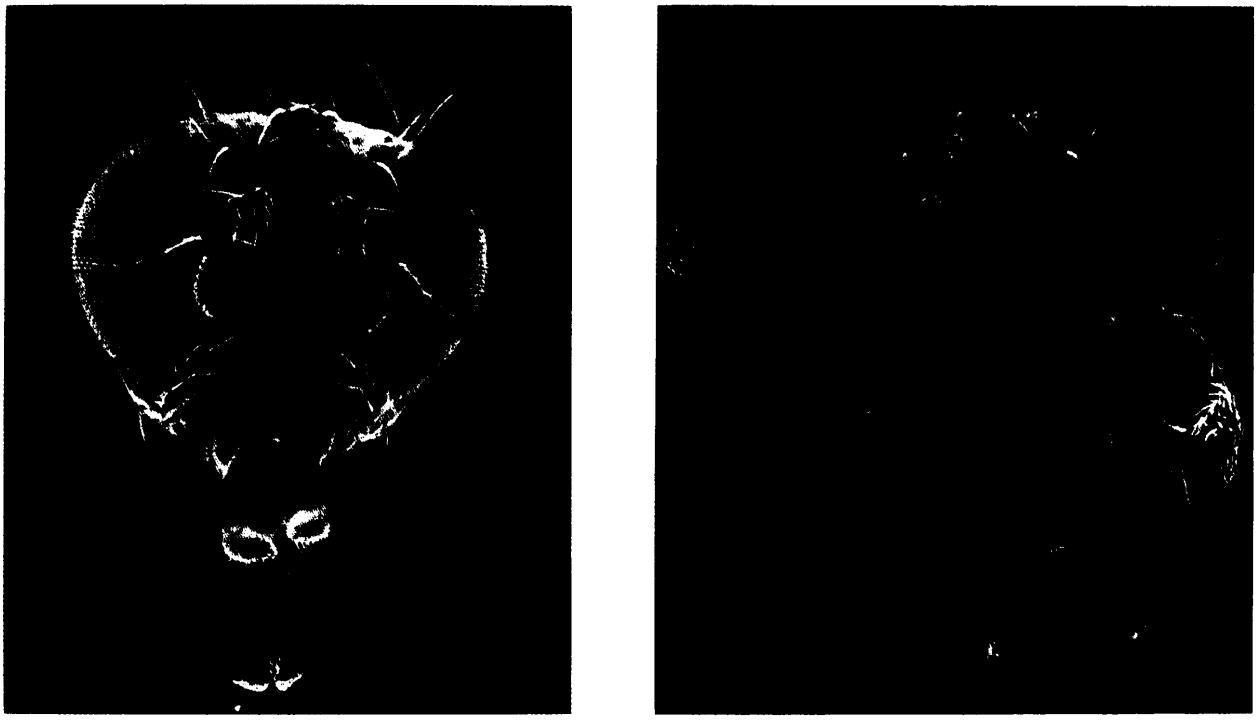


Photo credit: John Post/ethwalt, University of Oregon, Eugene

Mutations in the gene *Antennapedia,* a homeotic gene, cause the fruit fly to develop an extra pair of antennas. Pictured at left is the normal fruit fly, and at right a fly with the mutation. Homeotic genes have counterparts in humans and vertebrates; each gene has a characteristic DNA sequence within its protein-coding sequences called the homeo box.

APPLICATIONS IN MOLECULAR EVOLUTION

The disciplines of population biology, genetics, molecular biology, and cellular biology merge in the study of how species evolve, constituting the field of molecular evolution. The construction of a physical map of the human genome will permit molecular analysis of several questions fundamental to evolution, including how genomes change and what factors cause them to change, as well as how small-scale changes relate to the overall evolution of the organism (45).

Species with different degrees of relatedness can be usefully compared because their genes, and thus the proteins encoded by those genes, will have differing rates of sequence divergence. The course of human evolution can be read in the sequences of proteins (14). Comparisons of human and mouse DNA sequences are probably the most useful in the identification of genes unique to higher organisms because mice genes are more homologous to human genes than are the genes of any other well-characterized organism, Comparisons of human DNA sequences with those of lower organisms such as the fruit fly or nematode are most useful in the identification of genes encoding proteins that are essential to all multicellular organisms. Finally, since yeasts are single-celled *eukaryotes* (cells whose chromosomes are contained in nuclei), their sequences are most useful in the identification of genes that make proteins whose functions are essential to the life of all eukaryotic cells because such proteins would be least likely to have undergone major changes in the course of their evolution IMount, see app. A]. Table 3-5 shows how human proteins can be classified by their period of invention: from ancient, to middle-age, to modern (14).

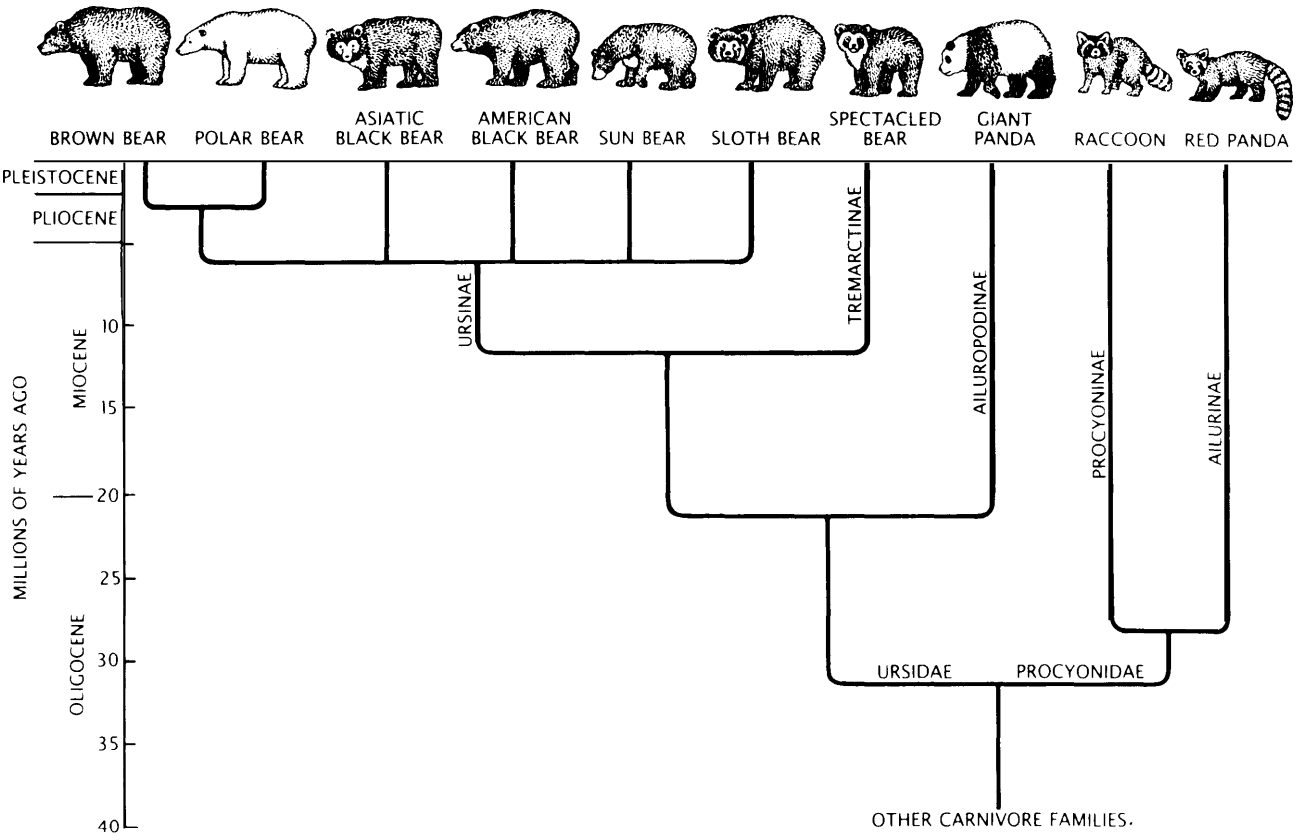


Photo credit: Stephen O'Brien, The National Cancer Institute, Frederick, MD. Reprinted with permission from Scientific American, *November 1987, pp. 102-107*

A phylogenic tree based on data obtained from modern molecular genetic methods places the giant panda in the Ursidae, or bear family. The red panda is left in the Procyonidae, or raccoon family. Molecular analysis of the chromosomes of these pandas suggests that the raccoon and bear families diverged from a common carnivorous ancestor about 35 to 40 million years ago.

Table 3-5.—Classification of Human Proteins by Invention Period

1. **Ancient proteins**
   - A. *First editions.* **Direct-line descendancy to human and contemporary prokaryotes. Mostly** enzymes in**volved** in metabolism.
   - B. *Second editions.* Homologous sequences in human and prokaryotic proteins, but apparently different functions.

II. Middle-age proteins
Proteins found in most eukaryotes but prokaryotic counterparts are as yet unknown.

III. Modern proteins
- C. *Recent vintage.* Proteins found in animals or plants but not both. Not found in prokaryotes.
- D. Very *recent inventions.* Proteins found in vertebrate animals but not elsewhere.
- E. *Recent mosiacs.* Modern proteins clearly the result of shuffling exons.

SOURCE: Adapted from Doolittle, R F., Feng, D, F., Johnson, MS., and McClure, M.A , "Relationships of Human Protein Sequences to Those of **Other Organisms,"** *Cold Spring Harbor Symposia on Quantitative Biology 51 "447-456, 1986.*

Physical map and sequence data accumulated from many species over the past 10 years have led scientists to recognize patterns of genome change quite different from those proposed earlier. Now, molecular evolutionists are beginning to understand such patterns as the duplication and acquisition of new genes and their corresponding functions, differences in the use of the genetic code among different organisms (48), and differences in the occurrence of gene families in different species (45).

important questions in molecular evolution arise from the fact that the genes of *prokaryotes* (organisms without nuclei, e.g., bacteria), as well as many genes in yeast and some multicellular organisms, are not interrupted by introns:

- Are the introns found in genes today descendants of extra or unused DNA from bacteria and eukaryotes such as yeast?
- Did prokaryotes rid themselves of intron sequences, or did they never have them?
- How did intron sequences get into the genes that code for modern proteins (14)?

By sequencing similar genes from many species, scientists have found that some introns have been in place for very long evolutionary periods and that the positions of introns within genes divide the genes in ways that correspond to the distinct functional domains of the proteins' structures (5,22,49). These observations have led to new models of molecular evolution (8,13,19-21,47). The *availability of more gene sequence data should facilitate the assessment of theories about the evolution of genes and gene structures.*

Sequences of more human genes, high-level understanding of variations in genomic organization among individuals, and analyses of differences between human beings and other organisms should aid in the evaluation of molecular evolutionary theories on how species originate (see boxes 3-D) 3-E) and 3-F). Recognition of differences in rates of nucleotide substitution, recombination, and other mechanisms responsible for variation in the human genome will lead to a better understanding of the molecular basis of these processes and of the constraints on each. Evaluation of proposed models for the propagation and evolution of multigene families, such as certain classes of cell surface receptors, requires a detailed knowledge not only of the relatedness of the DNA sequences in these genes, but also of their locations in the genome and the DNA sequences of the regions surrounding them (45).

Box 3-D.—Constructing the Evolutionary Tree: Morphology v. Molecular Genetics in the Search for Human Origins

Ever since Linnaeus, biologists have classified animals according to similarities and differences in form and structure. When the concept of evolution took root, these morphological features were used to establish phylogenies—trees or lineages that indicate the evolutionary relationships among species. New and sophisticated methods of genetic analysis have challenged morphology as the prime determinant of family trees. Recent debates about human origins have revealed the potential power of genetic techniques for evolutionary studies.

For the past two decades or so, anthropologists and biologists studying the problem of primate evolution have agreed that chimpanzees and gorillas are closely related enough to be classified in the same family, while humans stand alone in a separate, more distant family. Morphological evidence favors this view. Both chimps and gorillas, for example, walk on their knuckles; humans do not, and the fossils of their most direct ancestors show no features associated with knuckle walking. Chimps and gorillas also share similarities in the thickness and structure of their tooth enamel which suggest a common ancestry separate from humans.

Analyses of the DNA of chimps, apes, and human beings contradict this view, Scientists recently examined comparable segments of DNA in the region of the beta -globin gene from human beings, chimpanzees, gorillas, and orangutans. They sequenced 4,900 base pairs of DNA from this region in each organism, then appended data for nearby regions for which sequences had already been published. In all, they compared a 7,100-base-pair region and concluded that chimpanzee and human gene sequences were the least divergent. The most parsimonious explanation of the data was that human and chimpanzee are more closely related to each other than either is to the gorilla.

The beta-globin study, while it strongly suggests that chimpanzees are the closest cousins of human beings, does not conclusively end the search for human origins. Contradictions remain in the evidence gathered from comparative anatomy and from genetic analysis; studies of other gene loci will be necessary to settle the matter.

SOURCES:
R.L. Cann, "in Search of Eve," *The Sciences* (September/October) :30-37, 1987.
R. Lewin, '<My Close Cousin the Chimpanzee," *Science* 239:273-275, 1987.
M.M. Miyamoto et al., "Phylogenetic Relations of Humans and African Apes From DNA Sequences in the Globin Region," *Science* 239:369-373, 1987

Box 3-E. —The Origin of Human Beings: Clues From the Mitochondrial Genome

For more than a century, archaeologists, anthropologists, and biologists have been digging through layers of dirt and rock, sieving fossils and artifacts, in an attempt to figure out when, where, and how human beings differentiated from other primates to become a unique species. These scientists have relied on a variety of tools, everything from the picks and axes used to dig up fossils to sophisticated techniques for determining the age of the bones they have unearthed. Unfortunately, archaeological digs do not always yield perfect clues: Even well-preserved fossil remains are generally incomplete, and there are still missing links, cases in which fossils that could hint at the genealogy of several precursor species have not been found. Thus, it has been difficult to determine exactly when human beings diverged from prehistoric ancestors to become the species now known as *Homo sapiens*.

The development of molecular genetic techniques for analyzing DNA offers a new source of evidence in the ongoing debate about human origins. Techniques for mapping and sequencing DNA allow researchers to compare different species and different individuals from the same species at the most basic le~'el. These comparisons can aid evolutionary studies.

one promising approach is the study of the DNA sequences of mitochondria, small structures that are found in the cells of all multicellular organisms. Mitochondria are the power plants of eukaryotic cells. They produce energy for life processes by providing a site for the combination of oxygen and food molecules. Without them, cells would depend on less efficient processes of energy production and could not survi~~e in an environment containing oxygen. Mitochondria have much in common with bacteria: They are similar in size and shape, they both contain DNA, and they each reproduce by dividing in two.

The DNA in mitochondria can be more useful for some evolutionary studies than the DNA in cell nuclei, for several reasons. First, since it lies outside the cell nucleus and sexual recombination occurs only within the nuclei of sperm and egg cells, mitochondrial DNA is not recombined during sexual reproduction. It is inherited only from the mother. Consequently, changes in the nucleotide sequences are due only to mutation and not to the natural shuffling of DNA that occurs during reproduction. Second, DNA in the mitochon - dria is not protected as well as DNA in the nucleus, nor does it have the same kinds of mechanisms for repair. Thus, mitochondrial DNA mutates about 10 times as fast as the chromosomal DNA in the cell's nucleus, which means that the mitochondrial genome has evolved more rapidly than the chromosomal genome. Finally, mitochondria are relatively small: They contain approximately 16,000 base pairs, considerably fewer than the 3 billion base pairs in the entire set of human chromosomes, making them easier to analyze.

These three characteristics of mitochondrial DNA—absence of sexual recombination, a high natural mutation rate, and small size—have helped scientists construct a "molecular clock" that can be used to help establish the approximate time and place of human origins.By calculating the rate at which mitochon - drial DNA changes and then comparing the DNA sequences of mitochondria from many individuals, researchers have begun to formulate genealogical trees. For example, scientists sequenced samples of mitochon - drial DNA from 140 people around the world and used the information to propose that the first *Homo sapiens* lived .200,000 years ago on the African continent. Prior to these findings, anthropologists speculated that human beings originated nearly1 million years ago. Debate continues among scientists about the validitv and proper application of mitochondrial DNA sequences in evolutionary studies, but it is clear that molecu - lar genetics will play a growing role in this area.

SOURCES

B. Alberts, D Bray, J. Lewis, et al., "The Evolution of the Cell," in *Molecular Biology of the Cell* (New York, NY: Garland Publishing 1983)
R.L Cann, "In Search of Eve," *The Sciences* (September 'October): 30-37, 1987.
R.L. Cann, M. Stoneking, and A C, Wilson, "Mitochondrial DNA and Human Evolution," *Nature* 325:31-36, 1986.
R Lewin, "Molecular Clocks Turn a Quarter Century," *Science* 239:561-563, 1988.
J Tierney, L Wright, and K Springen, "The Search for Adam and Eve," *Newsweek*, Jan. 11, 1988, pp. 46-52.
J Wainscot, "Out of the *Garden* of Eden," *Nature* 32,5.13, 1986.
J D. Watson, N H Hopkins, J W. Roberts, et al ,,$ *tolecular Biology of the Gene* (Menlo Park, NJ: Benjamin Cummings Publishing, 1987)

Box 3-F.—Molecular Anthropology

Anthropologists working in a central Florida bog recently discovered 8,000-year-old human skeletons with well-preserved brains, some of which have provided the oldest available samples of human DNA. Before this discovery, samples of DNA had been available only from the dried tissue remains of archaeological specimens from more arid regions. The fact that DNA can be preserved in other than excessively dry conditions greatly increases the number of archaeological sites at which more ancient DNA samples may be discovered. DNA fragments have also been prepared from Egyptian mummies, from an extinct animal called a quagga, and from a 35,0()()-year-old bison from Alaska. Biologists have been trying to clone these DNA fragments for use in studies of evolution. The sample from the extinct bison is likely to be old enough for comparison with modern buffalo DNA; this comparison may provide clues to the mechanisms of genome evolution. The human DNA samples, although important discoveries, are too recent to be particularly informative in studies of molecular evolution. As methods for working with the DNA extracted from these ancient species are improved, and as more specimens are uncovered, the application of gene mapping and sequencing technologies to anthropology and archaeology will be more feasible.

SOURCES:

B. **Bower, "Human DNA Intact After 8,000 Years,"** *Science* **News, Nov. 8, 1986, p, 293,**

G,H. Doran, D. A'. Dickel, W.E. Ballinger, et al., "Anatomical, Cellular and Molecular Analysis of 8,000-Year-Old Brain Tissue From the Windover Archaeological Site," *Nature 323:803-806, 1986.*

APPLICATIONS IN POPULATION BIOLOGY

Population biologists study populations by analyzing many individuals. They are interested in similarities and differences among individuals, among groups, among varieties, and among species. **To address such questions as how geography and environment affect inheritance patterns of certain traits, a physical map and a complete sequence of a single reference genome are not particularly valuable.** It would be more useful to have corresponding sequence information from widely diverse geographical areas, from various religious and ethnic subgroups, and from all races (9).

Population geneticists studying human beings, plants, or animals make great use of molecular markers—RFLPs and, increasingly, sequences of specific regions—to assess the extent of genetic variability (see box 3-G). Information on the same small chromosomal region (e.g., a gene or a region important for gene expression) from many individuals might be more useful than information on larger chromosomal regions from a few persons (43), Genes for rare diseases are not all found in a single human genome: Sickle cell hemoglobin, for instance, might not have been discovered if only Northern Europeans had been studied (4,9).

Problems in population genetics that bear on public health involve finding means for estimating human mutation rates,[1] for studying susceptibility to pathogens such as the virus responsible for acquired immune deficiency syndrome (AIDS), and for assessing possible environmental influences on these phenomena. The mechanisms generating physical variability among human beings are by no means well understood and involve not only genetic factors, but, among other things, a complex set of environmental factors. **DNA sequences from representative portions of many human genomes would also be of more immediate use than whole genome sequences for monitoring the effects of specific environmental factors on the structure of the human genome (9).**

[1] 'An OTA report assesses these scientific issues: U.S. Congress, Office of Technology Assessment, ***Technologies for Detecting Heritable Mutations in Human*** Beings, OTA-H-298 (Washington, DC: U.S. Government Printing Office, September 1986).

BOX 3-G.—Implications of Genome Mapping for Agriculture

Since the dawn of agriculture, people have manipulated plants to enhance desired traits simply by observing the results of breeding, with no true understanding of the genetic principles involved. Many scientists working in the field of plant molecular biology believe that genome projects will have important implications for agriculture, by increasing knowledge about the genes that control or influence yield, time to maturation, nutritional content, resistance to disease, insects, and drought, and other factors in the production of crops.

The first gene maps ever constructed were assembled as a result of a series of painstakingly detailed crosses of pea plants and statistical analyses of data carried out by Austrian monk Gregor Mendel. Mendel was the first to recognize that some traits could be transmitted according to regular hereditary patterns. All modern genetics and much of modern biology build upon the foundation laid by Mendel.

Construction of RFLP marker maps has begun for corn, tomatoes, cabbage, and other crop plants. Such genetic maps give plant breeders the ability to use gene structure rather than observable characteristics to develop new varieties of plants. This ability should facilitate the development of intricate strategies for manipulating complex traits controlled by multiple, interacting genes.

The availability of RFLP maps makes it possible to select for several unrelated traits simultaneously or to manipulate traits controlled by clusters of genes that interact in complex ways. Researchers have mapped three genes that control efficiency of water use (drought tolerance) and five genes that have a major impact on flavor and soluble solids in tomatoes. Three genes that make a major contribution to insect resistance in tomatoes have also been mapped. A group of genes that influences yield has been found in corn. RFLP maps of genes influencing equally important traits are being developed for alfalfa, azaleas, cucumbers, onions, roses, sugar beets, and grasses.

Recently, there has been renewed interest in a small flowering plant called *Arabidopsis thaliana,* a duckweed in the mustard family. Although this plant has no obvious economic or nutritional value, it is a valuable research tool for plant molecular biologists. The *Arabidopsis* genome, at about 70 million base pairs, is about 10 percent or less the size of some of the major crop plant genomes, such as cotton, tobacco, or wheat. The small size of this plant makes it an important model system for studying general mechanisms of gene regulation that may be directly applicable to economically important but genetically less tractable plants. For these reasons, work has already begun on making complete genetic linkage and contig maps of the *Arabidopsis* genome.

SOURCES:
H. Bollinger, Native Plants Incorporated, personal communication, Jan 19, 1988.
Judson, see app A.
Mount, see app. A
S J O'Brien, Genetic *Maps1987* (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1987).
P.P Pang, and E.M.Meyerowitz, "Arabidopsis Thaliana: A Model System for Plant Molecular Biology," *Biotechnology* 5:1177.1181, 1987.
M. Walton, and T. Helentjaris, "Application of Restriction Fragment Length Polymorphism (RFLP) Technology to Maize Breeding," unpublished manuscript, 1988.
J.D Watson, N.H. Hopkins, J.W. Roberts, et al, in "Molecular Biology of the Gene," vol. I, (Menlo Park, NJ: Benjamin/Cummings Publishing, 1987)

CHAPTER 3 REFERENCES

1. Barker, D., Wright, E., Nguyen, K., et al., "Gene for von Recklinghausen Neurofibromatosis Is in the Pericentromeric Region of Chromosome 17," *Science* 236:1100-1102, 1987.
2. Bass, B. L., and Weintraub, H., "A Developmentally Regulated Activity That Unwinds RNA Duplexes" Cell 48:607-613, 1987,
3. Bodmer, W. F., Bailey, C. J., Bodmer, J., et al., "Localization of the Gene for Familial Adenomatous Polyposis in Chromosome 5)" *Nature* 328:614-616, 1987.
4. Bowman, J. E., Department of Pathology, The University of Chicago, personal communication, December 1987.
5. Branden, C .-1., "Anatomy of a/b Proteins," *Current Communications in Molecular Biology: Computer Graphics and Molecular Modeling,* R. Fletterick and M. Zoner (eds.) (Cold Spring Harbor, NY: Cold

Spring Harbor Laboratory, 1986), pp. 45-51.
6. Cavenee, W. K., Hansen, M. F., Nordernskjoid, M., et al., "Genetic Origin of Mutations Predisposing to Retinoblastoma," Science 228:501-503, 1985.
7. Costs of Human Genome Projects, OTA workshop, Aug. 7, 1987.
8. Craik, C, S., Rutter, W .J., and Fletterick, R., "Splice Junctions: Association With Variation in Protein Structure," *Science* 204:264-271, 1983.
9. Crow, J. F., Laboratory of Genetics, University of Wisconsin, Madison, personal communication, May 1987.
10. Davies, K. E., Pearson, P. L., Harper, P. S., et al., '(Linkage Analysis of Two Cloned Sequences Flanking the Duchenne Muscular Dystrophy Locus on the Short Arm of the Human X Chromosome, " *Nucleic Acids Research 11:2302-2312, 1983.*
11. Donis-Keller, H., Green, P., Helms, C., et al., "A Genetic Linkage Map of the Human Genome," *Cell 51:319-337, 1987.*
12. Doolittle, R. F., *Of URFS and ORFs: A Primer on How To Analyze Derived Amino Acid Sequences (Mill Valley, CA: University Science Books, 1987).*
13. Doolittle, R. F., '(Genes in Pieces: Were They Ever Together?" *Nature 272:581-582, 1978.*
14, Doolittle, R. F., Feng, D. F., Johnson, M. S., et al., "Relationships of Human Protein Sequences to Those of Other Organisms," *Molecular Biology of Homo Sapiens: Cold Spring Harbor Symposium on Quantitative Biology* 51(part 1):447-455, 1986.
15. Ecker, J. R., and Davis, R. W., "Inhibition of Gene Expression in Plant Cells by Expression of Antisense RNA, " *Proceedings of the National Academy of Sciences USA 83:5372-5376, 1986.*
16. Egeland, J., Gerhard, D., Pauls, D., et al., "Bipolar Affective Disorders Linked to DNA Markers on Chromosome 11," *Nature 325:783-787, 1987.*
17. Estivill, X., Farrall, M., Scambler, P., et al., "A Candidate for the Cystic Fibrosis Locus Isolated by Selection for Methylation-Free Island," *Nature 326:840-845, 1987.*
18. Friend, S. H., Bernards, R., Rogelj, S., et al., "A Human DNA Segment With Properties of the Gene That Predisposes to Retinoblastoma and Osteosarcoma, " *Nature 323:643-646, 1987.*
19. Gilbert, W., "Genes in Pieces Revisited," *Science* 228:823-824, 1985.
20. Gilbert, W., "Why Genes in *Pieces? ''Nature 271:501, 1978.*
21. Gilbert, W., Marchionni, M., and McKnight, G., "On the Antiquity of Introns, " Cell 46:151-154, 1986.
22. Go, M., "Correlation of DNA Exonic Regions With Protein Structural Units in Hemoglobin," *Nature 271:90-92, 1981.*
23. Gusella, J., Wexler, N., Conneally, P., et al., "A Polymorphic DNA Marker Genetically Linked to Huntington's Disease," *Nature 306:234-238, 1983.*
24. Herskowitz, I., "Functional Inactivation of Genes by Dominant Negative Mutations," *Nature 329:219-222, 1987.*
25 Knowlton, R. G., Cohen-Haguenauer, O., Van Cong, N., et al., "A Polymorphic DNA Marker Linked to Cystic Fibrosis Is Located on Chromosome 7,"*Nature 318:381-382, 1985.*
26 Kucherlapti, R., "Gene Replacement by Homologous Recombination in Mammalian Cells)" *Somatic Cell and Molecular Genetics 13:447-449, 1987.*
27. Lander, E. S., and Green, P., "Construction of Multilocus Genetic Linkage Maps in Humans, " *Proceedings of the National Academy of Sciences USA 84:2363-2367, 1987,*
28 Lee, W.-H., Bookstein, R., Hong, F., et al., "Human Retinoblastoma Gene: Cloning, Identification, and Sequence)" *Science 235:1394-1399, 1987.*
29 Mathew, C. G. P., Chin, K. S., Easton, D. F., et al., "A Linked Genetic Marker for Multiple Endocrine Neoplasia Type 2a on Chromosome 10," *Nature 328:527-528, 1987.*
30. McGarry, T.J., and Lindquist, S., "Inhibition of Heat Shock Protein Synthesis by Heat-Inducible Antisense RNA," *Proceedings of the National Academy of Sciences USA 83:399-403, 1986.*
31. McGinnis, W., Garber, R. L., Wirz, J., et al., "A Homologous Protein-Coding Sequence in *Drosophila* Homeotic Genes and Its Conservation in Other Metazoans," Cell 37:403-408, 1984.
32 Monaco, A., Neve, R., Colletti-Feener, C., et al., "Isolation of Candidate cDNAs for Portions of the Du - chenne Muscular Dystrophy Gene," *Nature 323: 646-650, 1986.*
33. National Research Council,*Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).
34. Nurse, P., and Lee, M. G., "Complementation Used To Clone a Human Homologue of the Fission Yeast Cell Cycle Control Gene*cdc2, " Nature 327:31-35, 1987.*
35. Patrusky, B., "Homeoboxes: A Biological Rosetta Stone," *Mosaic* 18:26-35, 1987.
36. Rebagliati, M. R., and Melton, D. A., "Antisense RNA Injections in Fertilized Frog Eggs Reveal an RNA Duplex Unwinding Activity," Cell 48:599-605, **1987.**
37. Reeders, S, T., Breuning, M. H., Davies, K. E., et al., "A Highly Polymorphic DNA Marker Linked to Adult Polycystic Kidney Disease on Chromosome 16)" *Nature 317:542-544, 1985.*
38. Royer-Pokora, B., Kunkel, L. M., Monaco, A. P., et al., '(Cloning the Gene for an Inherited Human

Disorder---- Chronic Granulomatous Disease---on the Basis of Its Chromosomal Location," *Nature 322:32-38, 1986.*

39. Salmons, B., Groner, B., Friis, R. et al., "Expression of Antisense mRNA in h-ras Transected NIH-3T3 Cells Does Not Suppress the Transformed Phenotype, " *Gene* **45:215-220, 1986.**

40. Scott, J., "Molecular Genetics of Common Diseases, " *British Medical Journal 295:769-771, 1987.*

41. Seizinger, B. R., Rouleau, G. A., Ozelius, L. J., et al., "Genetic Linkage of von Recklinghausen Neurofibromatosis to the Nerve Growth Factor Receptor Gene, " *Gene* 49:589-594, 1987.

42. Simpson, N. E., Kidd, K. K., Goodfellow, P. J., et al., '(Assignment of Multiple Endocrine Neoplasia Type 2a to Chromosome 10 by Linkage, ''Nature 328:528-530, 1987.

43. Siniscalco, M., "On the Strategies and Priorities for Sequencing the Human Genome: A Personal View," *Trends In Genetics 3:182-184, 1987.*

44. Smithies, O., Gregg, R. G., Boggs, S. S., et al., *Nature 317:230-234, 1985.*

45. Stephens, J. C., Human Gene Mapping Library, Howard Hughes Medical Institute, New Haven, CT, personal communication, 1987.

46. St. George-Hyslop, P. H., Tanzi, R. E., and Polinsky, R.J., "The Genetic Defect Causing Familial Alzheimer's Disease Maps on Chromosome 21," *Science* 235:885-890, 1987.

47. Stein, J. P., Catterall, J. F., Kristo, P., et al., "Ovomucoid Intervening Sequences Specify Functional Domains and Generate Protein Polymorphism, "*Cell 21:681-687, 1980.*

48. Stormo, G. 1987; 'Identifying Coding Sequences, " in *Nucleic Acid and Protein Sequence Ana{vsis: A Practical Approach,* M.J. Bishop, and C.J. Rollings (eds.) (Oxford: IRL Press, 1987).

49. Sudhof, T. C., Russell, D. W., Goldstein, J. L., et al., "Cassette of Eight Exons Shared by Genes for LDL Receptor and EGF Precursor," *Science* 228:893-895, 1985.

50. Thomas, K. R., and Capecchi, M. R., *Nature 324:34-38, 1986.*

51. Thomas, K. R., Folger, K. R., and Capecchi, M. R., *Cell* 44:419-428, 1986.

52. Tsui, L.-C., Buchwald, M., Barker, D., et al., "Cystic Fibrosis Locus Defined by a Genetically Linked Polymorphic DNA Marker," *Science* 230:1054-1057, 1985.

53. U.S. Congress, Office of Technology Assessment, New *Developments in Biotechnology, 4: U.S. Investment in Biotechnology, OTA-BA-360* (Washington, DC: U.S. Govenment Printing Office, in press).

54. U.S. Congress, Office of Technology Assessment, *Human Gene Therapy, OTA-BP-BA-32* (Washington, DC: U.S. Government Printing Office, December 1984).

55. Wainwright, B.J., Scambler, P. J., Schmidtke, J., et al., "Localization of Cystic Fibrosis Locus to Human Chromosome 7cen-q22)" *Nature 318:384-386, 1985.*

56. White, R., Woodward, S., Leppert, M., et al., '(A Closely Linked Genetic Marker for Cystic Fibrosis, " *Nature 318:382-384, 1985.*

57. Wise, J. A., Department of Biochemistry, University of Illinois, Urbana, personal communication, June 1987.